



Aldino José Martins Viegas

Licenciado

Molecular Determinants of Ligand Specificity in Carbohydrate-Binding Modules: an NMR and X-ray crystallography integrated study

Dissertação para obtenção do Grau de Doutor em
Bioquímica – Ramo Bioquímica Estrutural

Orientador: Eurico José da Silva Cabrita, Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Co-orientadores: Maria dos Anjos Lopez de Macedo, Professora Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Ana Luísa Moreira de Carvalho, Investigadora Auxiliar do Laboratório Associado Requimte - Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutora Ana Isabel Nobre Martins Aguiar Oliveira Ricardo

Arguentes: Prof. Doutor Jesús Jiménez Barbero
Prof. Doutor Shabir Husein Najmudin

Vogais: Prof. Doutor Marta Bruix Bayés
Prof. Doutor Carlos Mendes Godinho de Andrade Fontes
Prof. Doutora Maria João Lobo de Reis Madeira Crispim Romão



Março, 2012

Molecular determinants of ligand specificity in carbohydrate-binding modules: an NMR
and X-ray crystallography integrated study

Copyright 2012 Aldino José Martins Viegas

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Acknowledgements

I'd like to start by acknowledging and expressing my gratitude to my supervisors Prof. Eurico Cabrita, Prof. Maria dos Anjos Lopes Macedo and Dr Ana Luísa Carvalho, for allowing me to carry out my PhD in their groups, for providing me all the conditions for a successful work and, above all, for their friendship.

To Prof. Maria João Romão for taking me in her laboratory and providing me with all the conditions for the development of my work.

To Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa and Associate Laboratory REQUIMTE for providing the resources that allowed me to carry out my work.

To Prof. Carlos Fontes and Prof. José Prates of Faculdade de Medicina Veterinária da Universidade Técnica de Lisboa for letting me use their laboratory and expertise for all the molecular biology-related work and for all the assistance provided.

To my colleagues Dr Marta Corvo, Dr Maria Manuel Marques, Dr Rui Almeida, Filipe Freire, Dr Cristiano Mota and, most recently Dr Filipa Marcelo, Dr João Sardinha and Dr Ângelo Figueiredo for all the good times, all the discussions, all the help... Thank you!

To all of my colleagues from the Xtal group for helping me whenever I needed, for the very useful discussions and for your friendship.

To Dr Marta Bruix from IQFR, CSIC, Madrid, for her help in several aspects and, above all, for her affection and always good humor.

To Fundação para a Ciência e Tecnologia for funding and support, through grant SFRH/BD/35992/2007, and projects PTDC/QUI/68286/2006, PTDC/QUI-BIQ/100359/2008 and PTDC/BIA-PRO/103980/2008.

To the Portuguese Nuclear Magnetic Resonance Network for funding and support.

To my beloved family without whom I wouldn't be here. Thank you for all your support and unconditional love.

Finally, to my dear Sofia. Thank you for everything. I love you!

Resumo

A degradação da parede celular vegetal por parte de microrganismos é um dos processos mais importantes para a renovação do dióxido de carbono atmosférico. O trabalho apresentado nesta tese aborda os celulosomas de *Clostridium thermocellum* e *Bacteroides cellulosolvens*, essenciais para o processo de degradação da celulose, e visa o estudo de alguns dos componentes envolvidos na sua arquitetura (coesinas e doquerinas) e eficiência (Carbohydrate-Binding Modules - CBMs). Para isso utilizei uma combinação de técnicas de Ressonância Magnética Nuclear (RMN), cristalografia de raios-X e modelação computacional. O meu objetivo era contribuir para a racionalização dos determinantes moleculares de especificidade de CBMs, nomeadamente os C_tCBM das famílias 11, 30 e 44, e dos mecanismos de reconhecimento molecular entre coesinas e doquerinas. No capítulo I faço uma introdução geral ao tema da degradação da parede celular vegetal com especial atenção ao celulosoma e aos seus componentes. No capítulo II discuto as características estruturais do C_tCBM11 tendo como base estruturas obtidas por RMN a 25 e a 50 °C e a estrutura obtida por cristalografia. Os resultados mostram que as estruturas apesar de semelhantes, apresentam algumas diferenças, nomeadamente no que respeita à área do sítio de ligação, o que explica os resultados negativos obtidos por co-cristalização. Nos capítulos III e IV descrevo o estudo acerca dos determinantes moleculares de especificidade dos módulos C_tCBM11, 30 e 44, com base em estudos de RMN e de modelação computacional. Observei que os átomos de carbono-oligosacáridos mais importantes para a ligação a estes módulos estão nas posições 6 e 2 das unidades centrais dos ligandos. Caracterizei também os mecanismos responsáveis pela seleção e ligação destes módulos aos vários substratos. Verifiquei que a ligação ocorre por um mecanismo de seleção conformacional onde a disposição dos resíduos da proteína, a conformação do ligando e o número de unidades de glucose, desempenham um papel fundamental. Os capítulos V e VI dizem respeito à determinação da estrutura 3D dos complexos coesina-módulo X-doquerina de *C. thermocellum* e coesina-doquerina de *B. cellulosolvens*, respetivamente. Ambos os complexos pertencem ao tipo II e a sua análise permitiu extrair informações importantes acerca das características estruturais que definem a interação coesina-doquerina. A estrutura de *C. thermocellum* revelou que o módulo X é fundamental para a estabilidade do complexo. Por outro lado, foi a primeira vez que foi determinada a estrutura 3D de um complexo coesina-doquerina de *B. cellulosolvens*. Neste complexo a doquerina aparece rodada 180° quando comparada com outros complexos. Esta característica confere plasticidade ao celulosoma. Nos capítulos finais apresento as técnicas de RMN e cristalografia de raios-X que utilizei ao longo do trabalho. Por fim apresento algumas conclusões gerais sobre todo o trabalho realizado.

Palavras Chave: Celulosoma, coesina, doquerina, C_tCBM11, C_tCBM30, C_tCBM44

Abstract

The microbial plant cell wall degradation is one of the most important processes in the global turnover of atmospheric carbon dioxide. The work presented in this thesis addressed the cellulosomes of *Clostridium thermocellum* and *Bacteroides cellulosolvens*, essential to the process of cellulose degradation, and aimed to study some of the components involved in their architecture (cohesins and dockerins) and efficiency (Carbohydrate-Binding Modules - CBMs). For this I used a combination of Nuclear Magnetic Resonance (NMR), X-ray crystallography and computer modeling techniques. My objective was to help rationalize the molecular determinants of specificity of CBMs, including the *Ct*CBMs of families 11, 30 and 44, and the mechanisms of molecular recognition between cohesins and dockerins. In Chapter I, I present a general introduction to the theme of degradation of plant cell walls, with special attention to the cellulosome and its components. In Chapter II, I discuss the structural characteristics of the *Ct*CBM11 based on the structures obtained by NMR at 25 and 50 °C and the structure obtained by crystallography. I found that although similar, the structures show some differences, particularly regarding the binding cleft area, which explains the negative results obtained by co-crystallization. In Chapter III and IV I study the molecular determinants of specificity in modules *Ct*CBM11, 30 and 44, based on NMR and computer modeling data. I found that the atoms of the celooligosaccharides most important for binding are the ones at positions 2 and 6 of the central units of the ligands. Moreover, I characterized the mechanisms responsible for selection and binding of these modules to various substrates. I established that binding occurs by a mechanism for conformational selection, where the topology of the residues of the protein, the conformation of the ligand and the number of glucose units, play a fundamental role. Chapters V and VI reveal the determination of the 3D structure of the cohesin-module X-dockerin complex of *C. thermocellum* and the cohesin-dockerin complex of *B. cellulosolvens*, respectively. Both complexes belong to the type II and their analysis allowed obtaining important information about the structural features that define the cohesin-dockerin interaction. The structure belonging to *C. thermocellum* revealed that the module X is essential for the stability of the complex. Moreover, for the first time the 3D structure of a cohesin-dockerin complex from *B. cellulosolvens* was determined. In this complex the dockerin is rotated 180° when compared to other complexes. This gives the cellulosome plasticity. In the final chapters, I present the NMR and X-ray crystallography techniques I used throughout the study. Finally, I draw some general conclusions about all the work done.

Keywords: Cellulosome, cohesin, dockerin, *Ct*CBM11, *Ct*CBM30, *Ct*CBM44

Table of Contents

Acknowledgements	i
Resumo.....	iii
Abstract	v
Table of Contents	vii
List of Figures	xvii
List of Tables.....	xxiii
Nomenclature	xxv
CHAPTER I <i>INTRODUCTION - THE IMPORTANCE OF THE RESEARCH</i>	1
Summary	3
I.1 Introduction	3
I.2 The plant cell wall	6
I.2.1 Cellulose.....	7
I.2.2 Xyloglucan	8
I.3 Plant cell wall hydrolysis	9
I.3.1 Enzymatic hydrolysis: The cellulosome.....	10
I.4 The cellulosome of <i>Clostridium thermocellum</i>: architecture and function .	14
I.5 The cohesin-dockerin interaction	17
I.6 Carbohydrate-binding modules	19
I.6.1 Nomenclature of CBMs	21
I.6.1.1 Type A CBMs – surface-binding.....	24
I.6.1.2 Type B CBMs – glycan-chain-binding	24
I.6.1.3 Type C CBMs – small sugar-binding	25
I.6.2 Molecular determinants of binding	25
I.6.3 Utilization of CBMs	27
I.7 Objectives and outline of the thesis	29

I.8	References	30
------------	-------------------------	-----------

<i>CHAPTER II</i>		<i>STRUCTURE OF THE FAMILY 11 CARBOHYDRATE-BINDING MODULE</i>
	<i>FROM CLOSTRIDIUM THERMOCELLUM (CTCBM11)</i>	<i>37</i>
	Summary	39
II.1	Introduction	40
II.2	Results and Discussion	42
II.2.1	Structure of <i>CtCBM11</i>	42
II.2.1.1	The crystal structure of <i>CtCBM11</i> without the histidine tail	42
II.2.1.2	The solution structure of <i>CtCBM11</i>	45
II.2.1.3	Comparison between the X-ray and NMR structures	47
II.3	Conclusions	48
II.4	Materials and methods	49
II.4.2	Molecular biology	49
II.4.2.1	Recombinant protein production.....	49
II.4.2.2	Double labeled (¹³ C and ¹⁵ N) protein expression and purification.....	49
II.4.3	X-ray crystallography.....	52
II.4.3.1	Protein crystallization and data collection	52
II.4.3.2	Phasing, model building and refinement	53
II.4.4	NMR spectroscopy	53
II.4.4.1	Data acquisition	53
II.4.4.2	Resonance assignment and structure calculation.....	54
II.4.4.2.1	Resonance assignment.....	54
II.4.4.2.2	Structure calculation.....	55
II.4.4.2.3	Structure validation	56
II.7	References	57

<i>CHAPTER III</i>	<i>MOLECULAR DETERMINANTS OF LIGAND SPECIFICITY IN</i>	
	<i>CTCBM11</i>	61
	Summary	64
III.1	Introduction	65
III.2	Results and Discussion	68
III.2.1	Characterization of the sugars	68
III.2.2	Molecular determinants of ligand specificity	72
III.2.2.1	Co-crystallization studies	73
III.2.2.2	Influence of calcium in the structure of cellobiose	73
III.2.2.3	Linebroadening studies	74
III.2.2.4	Saturation-transfer difference NMR (STD-NMR)	75
III.2.2.5	Diffusion studies (DOSY)	82
III.2.2.6	Interaction studies with cellobiosaccharides	83
III.2.2.7	Computational studies	86
III.4.4.2.1	Docking experiments with the crystallographic structure	87
III.4.4.2.1	Docking experiments with the NMR solution structure	91
III.2.3	Molecular dynamics	94
III.2.4.1	Relaxation data, diffusion tensor and hydrodynamic calculations	95
III.2.4.2	Internal mobility	99
III.2.4.3	Estimation of the conformational entropy from NMR Relaxation data	101
III.2.4.4	Amide proton exchange	102
III.3	Conclusions	104
III.4	Materials and methods	106
III.4.1	Sources of sugars	106

III.4.2	Molecular biology	106
III.4.2.1	Recombinant protein production.....	106
III.4.2.2	Transformation, expression and purification of <i>Ct</i> CBM11 with the 6-histidine tail	106
III.4.2.3	Transformation, expression and purification of the double labeled ¹³ C and ¹⁵ N) <i>Ct</i> CBM11 with the 6-histidine tail ...	107
III.4.3	X-ray crystallography.....	107
III.4.3.1	Co-crystallization studies.....	107
III.4.4	NMR spectroscopy	107
III.4.4.1	Data acquisition	107
III.4.4.2	Characterization of the sugars.....	108
III.4.4.3	Influence of calcium in the structure of cellobiose.....	109
III.4.4.4	Linebroadening studies	109
III.4.4.5	STD-NMR studies	110
III.4.4.6	Diffusion studies (DOSY)	111
III.4.4.7	<i>Ct</i> CBM11 titration	112
III.4.4.8	Combined chemical shift	113
III.4.4.9	Determination of the association constant (K_a).....	114
III.4.4.10	Determination of the thermodynamic parameters.....	116
III.4.4.11	¹⁵ N backbone relaxation measurements	116
III.4.4.12	Relaxation data processing and analysis.....	117
III.4.4.13	Estimation of the molecular diffusion tensor.....	118
III.4.4.14	Hydrodynamic calculations	118
III.4.4.15	Calculation of the model free dynamics parameters.....	119
III.4.4.16	Estimation of the conformational entropy from NMR relaxation data.....	119
III.4.4.17	Amide proton exchange	120
III.4.5	Computational studies	122

II.4.5.1	Docking experiments with the crystallographic structure and molecular dynamics	122
II.4.5.2	Docking experiments with the NMR solution structure and molecular dynamics	123
III.7	References	124
<i>CHAPTER IV</i>	<i>MOLECULAR DETERMINANTS OF LIGAND SPECIFICITY IN CTCBM30 AND CTCBM44.....</i>	<i>131</i>
	Summary	133
IV.1	Introduction	134
IV.2	Results and Discussion	137
IV.2.1	Molecular determinants of ligand specificity	137
IV.2.1.1	Saturation transfer difference NMR (STD-NMR)	138
IV.2.1.2	Docking models of the interaction of <i>Ct</i> CBM30 and <i>Ct</i> CBM44 with cellooligosaccharides	145
IV.2.1.2.1	Model of <i>Ct</i> CBM30 bound to cellotetraose ..	145
IV.2.1.2.2	Model of <i>Ct</i> CBM30 bound to cellohexaose..	147
IV.2.1.2.3	Model of <i>Ct</i> CBM44 bound to cellohexaose..	148
IV.2.1.2.4	Model of <i>Ct</i> CBM44 bound to cellopentaose.	151
IV.2.1.2.4	Model of <i>Ct</i> CBM44 bound to cellotetraose ..	151
IV.2	Conclusions	152
IV.3	Materials and methods	153
IV.3.1	Sources of sugars.....	153
IV.3.2	Molecular biology	154
IV.4.2.1	Recombinant protein production.....	154
IV.4.2.2	Protein expression and purification	154
IV.3.3	NMR spectroscopy	155
IV.4.3.1	Data acquisition	155

IV.4.4.2	STD-NMR studies	155
IV.4.5	Docking studies	155
IV.4.5.1	Preparation of the ligand pdb files	155
IV.4.5.2	Docking models of the interaction of <i>Ct</i> CBM30 and <i>Ct</i> CBM44 with celooligosaccharides	156
IV.7	References	156
<i>CHAPTER V</i>	<i>THE ORF2 TYPE II COHESIN-XDOCKERIN COMPLEX FROM C.</i> <i>THERMOCELLUM</i>	<i>159</i>
	Summary	161
V.1	Introduction	161
V.2	Results and Discussion	165
V.2.1	Architecture of the Orf2 type II Coh-XDoc complex from <i>C. thermocellum</i>	165
V.2.1.1	Type II Coh structure in the complex	168
V.2.1.2	Type II XDoc structure in the complex	168
V.2.1.3	The complex interface	173
V.3	Conclusions	177
V.4	Materials and methods	178
V.4.1	Molecular biology	178
V.4.2.1	Transformation, expression, purification and quantification	178
V.4.2	X-ray crystallography.....	179
V.4.2.1	Protein crystallization and data collection	179
V.4.2.2	Phasing, model building and refinement	179
V.5	References	180

<i>CHAPTER VI</i>	<i>THE SCAA TYPE II COHESIN-DOCKERIN COMPLEX FROM B.</i>	
	<i>CELLULOSOLVENS</i>	183
	Summary	185
VI.1	Introduction	186
VI.2	Results and Discussion	187
VI.2.1	Architecture of the SdbA type II Coh-Doc complex from <i>B. cellulosolvens</i>	187
VI.2.1.1	Type II Coh structure in the complex	189
VI.2.1.2	Type II Doc structure in the complex	190
VI.2.1.3	The complex interface – an alternative binding mode	192
VI.3	Conclusions	196
VI.4	Materials and methods	196
VI.4.1	Molecular biology	196
VI.4.2.1	Transformation, expression, purification and quantification	196
VI.4.2	X-ray crystallography.....	197
VI.4.2.1	Protein crystallization and data collection	197
VI.4.2.2	Phasing, model building and refinement	197
VI.5	References	198
 <i>CHAPTER VII</i>	 <i>PROTEIN NMR SPECTROSCOPY</i>	 201
	Summary	204
VII.1	Introduction	204
VII.2	Protein NMR	207
VII.2.1	Chemical Shift.....	207
VII.2.1.1	Spin-spin coupling and spin systems	208
VII.2.2	Relaxation.....	212
VII.2.2.1	The Bloch equations	213

VII.2.2.2	T_1 relaxation.....	214
VII.2.2.3	T_2 relaxation.....	215
VII.2.2.4	Dipole-dipole relaxation	217
VII.2.2.5	Chemical shift anisotropy relaxation	218
VII.2.3	The protein's fingerprint – ^{15}N - ^1H -HSQC.....	219
VII.2.4	Nuclear Overhauser effect.....	221
VII.3	Protein structure determination.....	229
VII.3.1	Three-dimensional experiments	231
VII.3.1.1	Experiments for backbone assignments.....	232
VII.3.1.1.1	HNCO	232
VII.3.1.1.2	HN(CA)CO	234
VII.3.1.1.3	HN(CO)CACB	235
VII.3.1.1.4	HNCACB	237
VII.3.1.1.5	Angular restraints	239
VII.3.1.2	Experiments for side-chain assignments.....	240
VII.3.1.2.1	(H)CCH-TOCSY.....	240
VII.3.1.2.2	HNHA	241
VII.3.1.3	Experiments for NOE measurement	242
VII.3.1.3.1	$^{15}\text{N}/^{13}\text{C}$ -NOESY-HSQC	242
VII.3.1.3.2	Distance restraints	243
VII.3.2	Structure validation	244
VII.4	Protein dynamics by NMR.....	245
VII.4.1	Theory of spin relaxation in proteins	246
VII.4.2	Protein motions and relaxation.....	248
VII.4.2.1	Reduced spectral density mapping.....	248
VII.4.2.2	Rotational diffusion tensor.....	249
VII.4.2.3	The Lipari-Szabo Model-free formalism.....	251

VII.4.2.3.1	Relationship between the generalized order . parameter, S^2 , and the conformational entropy, ΔS_{conf}	253
VII.4.2.4	Amide proton exchange	254
VII.5	Study of protein-ligand complexes	256
VII.5.1	Saturation-transfer difference	256
VII.5.2	Diffusion ordered spectroscopy	261
VII.6	References	265
<i>CHAPTER VIII</i>	<i>X-RAY CRYSTALLOGRAPHY</i>	<i>273</i>
Summary	275
VIII.1	Introduction	275
VIII.2	Crystal systems: symmetry operations and space groups.....	277
VIII.3	Protein crystallization	281
VIII.3.1	Matthews' volume.....	283
VIII.4	Structure determination.....	284
VIII.4.1	X-ray diffraction and data collection.....	284
VIII.4.1.1	Synchrotron radiation	286
VIII.4.2	Model building and refinement	287
VIII.4.2.1	Molecular replacement	288
VIII.4.2.1.1	Patterson function	290
VIII.4.2.2	Model building.....	291
VIII.4.2.3	Model refinement.....	294
VIII.4.3	Structure validation	297
VIII.5	References	299
<i>FINAL CONCLUSIONS</i>	<i>301</i>
<i>APPENDIX A</i>	<i>305</i>
<i>APPENDIX B</i>	<i>311</i>
<i>APPENDIX C</i>	<i>323</i>

List of Figures

Figure I.1: From biomass to biofuels	4
Figure I.2: Plant cell wall structure	7
Figure I.3: Structure of cellulose.....	8
Figure I.4: Simplified structure and abbreviated names of xyloglucan oligosaccharides.....	8
Figure I.5: Cellulosomes at the surface of <i>Clostridium thermocellum</i>	12
Figure I.6: Schematic representation of the <i>Clostridium thermocellum</i> cellulosome..	15
Figure I.7: The cohesin-dockerin complex..	18
Figure I.8: Classification of CBMs..	21
Figure I.9: The binding-site platforms of the three types of CBMs.....	26
Figure I.10: Applications of hybrid CBMs..	27
Figure II.1: 3D structure of <i>Ct</i> CBM11 obtained by X-ray crystallography.....	39
Figure II.2: Amino acid sequence of <i>Ct</i> CBM11.....	40
Figure II.3: Coordination of the two calcium ions in <i>Ct</i> CBM11.....	41
Figure II.4: Ribbon representation of <i>Ct</i> CBM11 packing in the two different crystal forms, $P2_12_12$ and $P2_1$	43
Figure II.5: Superposition of the <i>Ct</i> CBM11 structures determined with and without the histidine tail (structures depicted in blue and grey, respectively).	44
Figure II.6: Ribbon representation of the NMR-determined 20-structure ensemble of <i>Ct</i> CBM11 at 25 °C (A) and 50 °C (B).....	46
Figure II.7: Comparison between the X-ray and NMR structures.....	48
Figure II.8: SDS-PAGE gel of the purified <i>Ct</i> CBM11 fractions.....	51
Figure II.9: Crystals of <i>Ct</i> CBM11 with no engineered 6-His tail.....	53
Figure III.1: Highlight of the binding cleft of <i>Ct</i> CBM11 with the bound C-terminal histidine tail of a symmetry related molecule.	66
Figure III.2: Structure and ^1H spectra of cellobiose.	69
Figure III.3: Structure and ^1H spectra of cellotetraose.	69
Figure III.4: Structure and ^1H spectra of cellohexaose.....	70
Figure III.5: Structure and ^1H spectra of laminarihexaose.	70
Figure III.6: Titration of cellohexaose with CaCl_2	74
Figure III.7: Line broadening studies	75
Figure III.8: STD-NMR of cellobiose with <i>Ct</i> CBM11.....	77
Figure III.9: STD-NMR and epitope mapping of cellotetraose bound to <i>Ct</i> CBM11.	78

Figure III.10: STD-NMR and epitope mapping of celohexaose bound to <i>CtCBM11</i>	79
Figure III.11: STD-NMR of laminarihexaose with <i>CtCBM11</i>	81
Figure III.12: DOSY spectra for the calculation of the association constant for the celohexaose/ <i>CtCBM11</i> interaction	82
Figure III.13: Backbone amide chemical shift variations between <i>CtCBM11</i> and A) celohexaose at 25°C; B) celohexaose at 50 °C and C) cellotetraose at 25 °C.....	84
Figure III.14: Representation of the conformations of the three-dimensional structure of binding of the different ligands obtained by docking.	88
Figure III.15: Representation of the most important interactions between the β -cellotetraose (A) and β -celohexaose (B) with the <i>CtCBM11</i> binding cleft.....	89
Figure III.16: Schematic representation of the main interaction between the pentasaccharide with the <i>CfCBM4</i> (pdb entry: 1GU3) (A) and the hexasaccharide with <i>CtCBM11</i> (B).....	90
Figure III.17: Docking models of <i>CtCBM11</i> with celohexaose at 25 °C (A) and 50 °C (B) and cellotetraose at 25 °C (C).	92
Figure III.18: Graphical superposition of the $\{^1\text{H}\}$ - ^{15}N -NOE of <i>CtCBM11</i> in the free (black) and bound state (red) at 25 (top) and 50 °C (bottom).	96
Figure III.19: Effect of binding and temperature on the R_2/R_1 ratio.	97
Figure III.20: Effect of binding (left) and temperature (right) on the S^2 order parameter.	101
Figure III.21: Effect of binding in the (A) amide hydrogen/deuterium exchange rates and (B) free energy of structural opening for the free and bound protein at 25 °C.....	103
Figure IV.1: 3D structure of <i>CtCBM30</i> (A) and <i>CtCBM44</i> (B) obtained by X-ray crystallography.....	133
Figure IV.2: Solvent-exposed tryptophan residues at the surface of <i>CtCBM30</i> (A) and <i>CtCBM44</i> (B).	135
Figure IV.3: STD-NMR of cellobiose with <i>CtCBM30</i> and <i>CtCBM44</i>	138
Figure IV.4: STD-NMR of cellotetraose with <i>CtCBM30</i> and <i>CtCBM44</i>	139
Figure IV.5: STD-NMR of celohexaose with <i>CtCBM30</i> and <i>CtCBM44</i>	141
Figure IV.6: STD-NMR of laminarihexaose with <i>CtCBM30</i> and <i>CtCBM44</i>	143
Figure IV.7: Model of the structure of <i>CtCBM30</i> in complex with cellotetraose.	145
Figure IV.8: Model of the structure of <i>CtCBM30</i> in complex with celohexaose.....	147
Figure IV.9: Model of the structure of <i>CtCBM44</i> in complex with celohexaose.....	149
Figure IV.10: Model of the structure of <i>CtCBM44</i> in complex with cellopentaose.....	151
Figure IV.11: Model of the structure of <i>CtCBM44</i> in complex with cellotetraose.	152
Figure IV.12: SDS-PAGE gel of the purified <i>CtCBM44</i> fractions.....	154

Figure V.1: Crystal structure of the Orf2 type II cohesin-modules X-dockerin complex (CohII-XDocII) from <i>C. thermocellum</i> (PDB code: 2vt9).....	161
Figure V.2: Schematic representation of the <i>Clostridium thermocellum</i> cellulosome.....	162
Figure V.3: The dual binding mode of type I cohesin-dockerin complexes.	164
Figure V.4: Comparison of the structure of the Orf2 type II Coh-XDoc with the Structure of the SdbA type II Coh-XDoc.....	167
Figure V.5: Ribbon representation of the structure of the type II cohesin module of the Orf2 type II Coh-X-Doc complex.	168
Figure V.6: Structure of the type II X-dockerin module of the Orf2 type II Coh-X-Doc complex.	169
Figure V.7: The XDoc and X-Coh interface hydrogen bonds in the type II Orf2 and type II SdbA complexes.	171
Figure V.8: The Coh-Doc and X-Coh interface hydrogen bonds in the type II Orf2 and type II SdbA complexes.	174
Figure V.9: Sequence alignment of the type II dockerins from the native Orf2 and SdbA complexes and the type I dockerin module.....	175
Figure V.10: Ribbon representation of the native and 180°-rotated type II Orf2 dockerin modules.	176
Figure VI.1: Crystal structure of the type II cohesin-dockerin complex (Coh-Doc) from <i>B. cellulosolvens</i> (PDB code: 2y3n).....	185
Figure VI.2: Schematic representation of the <i>Bacteroides cellulosolvens</i> cellulosome (A) and phylogenetic relationships of the ScaA and ScaB cohesins (B).	187
Figure VI.3: Sequence alignment showing the dyad symmetry within the dockerin sequence	188
Figure VI.4: Ribbon representation of the structure of the type II cohesin module of the ScaA type II Coh ₁₁ -Doc complex.....	190
Figure VI.5: Ribbon representation of the structure of the type II dockerin module of the ScaA Coh ₁₁ -Doc complex.	191
Figure VI.6: The Coh-Doc interface hydrogen bonds in the type II ScaA complex.....	193
Figure VI.7: Alternative binding mode in the <i>B. cellulosolvens</i> Coh-Doc complex and internal symmetry of the dockerin.	195

Figure VII.1: Yearly and annual growth of structures solved by NMR.....	205
Figure VII.2: ^{13}C - ^{13}C TOCSY pattern of the 20 standard amino acids.....	210
Figure VII.3: ^1H - ^1H TOCSY and COSY pattern of the 20 standard amino acids.....	211
Figure VII.4: Peptide torsion angles	212
Figure VII.5: The inversion recovery process.	214
Figure VII.6: Effect of the correlation time, τ_c , in the relaxation time T_1	215
Figure VII.7: The spin-echo refocuses magnetization dephased by field inhomogeneity.	216
Figure VII.8: Effect of the correlation time, τ_c , in the relaxation time T_2	217
Figure VII.9: The ^{15}N - ^1H -HSQC (A) and ^{13}C - ^1H -HSQC (B) magnetization transfer.	219
Figure VII.10: ^{15}N - ^1H -HSQC spectrum of the 52 amino acid (5.677 Da) protein rubredoxin from the sulfate-reducing bacterium <i>Desulfovibrio gigas</i> (pdb code: 1rdg).	220
Figure VII.11: Irradiation of resonance A leads to an increase of peak intensity of the neighboring spin C (positive NOE) or to a decrease of peak intensity (negative NOE).	221
Figure VII.12: Energy level diagram for a two homonuclear spin system $-\frac{1}{2}$ nuclei, I and S, showing definitions of transition probabilities and spin states.....	222
Figure VII.13: Schematic representation of the origin of the NOE in a homonuclear two $\frac{1}{2}$ nuclei spin system.	223
Figure VII.14: Variation of the spectral density with the molecular motion as a function of the frequency.....	225
Figure VII.15: Schematic representation of the relaxation pathways that lead to direct and indirect contributions to the NOE enhancement of spin I upon S saturation in a multispin system.....	228
Figure VII.16: Process of 3D solution structure calculation from NMR data.	230
Figure VII.17: Anatomy of a 3D NMR experiment.....	231
Figure VII. 18: Scalar coupling constants between the different nuclei in amino acids.....	231
Figure VII.19: The HNC0 magnetization transfer.	233
Figure VII.20: Identifying the CO_i resonance.	233
Figure VII.21: The HN(CA)CO magnetization transfer.	234
Figure VII.22: Identifying the CO_i resonance.....	235
Figure VII.23: The HN(CO)CACB magnetization transfer.....	236
Figure VII.24: Identifying the CA_{i-1} and CB_{i-1} resonances.....	236
Figure VII.25: The HNCACB magnetization transfer.....	237
Figure VII.26: Identifying the CA_i and CB_i resonances.....	238
Figure VII.27: Sequential assignment of the protein backbone resonances based on the HNCACB spectrum.....	239
Figure VII.28: The (H)CCH-TOCSY magnetization transfer.	240
Figure VII.29: The HNHA magnetization transfer.	241

Figure VII.30: The ^{15}N - ^1H -HSQC-NOESY (A) and ^{13}C - ^1H -HSQC-NOESY (B) magnetization transfer.....	243
Figure VII.31: Protein motion time scales and NMR techniques used to study each time scale...246	246
Figure VII.32: Representation of an amide vector in a protein.	250
Figure VII.33: Interpretation of the generalized order parameter, S^2 , in a diffusion-in-a-cone model.....	252
Figure VII.34: Scheme of the STD-NMR experiment.....	256
Figure VII.35: STD amplification factor as a function of the saturation time (A) and ligand concentration (B).	259
Figure VII.36: The Stejskal and Tanner pulsed field gradient NMR sequence.	262
Figure VIII.1: Flowchart of the main steps involved in a 3D structure determination by X-ray crystallography.....	275
Figure VIII.2: Yearly and total growth of structures solved by X-ray crystallography.....	276
Figure VIII.3: Crystal architecture.	277
Figure VIII.4: The Miller indices.	278
Figure VIII.5: The 14 Bravais lattices.	280
Figure VIII.6: Solubility curve of a protein as a function of the precipitant concentration.....	282
Figure VIII.7: Obtaining crystals by the hanging drop method.....	283
Figure VIII.8: Bragg's Law.	285
Figure VIII.9: The Molecular Replacement method.	290
Figure VIII.10: Patterson map derived from a crystal with three atoms.	291
Figure VIII.11: Criteria for assessment of the quality of crystallographic models of macromolecular structures.	298

List of Tables

Table I.1: List of cellulosomal components of <i>C. thermocellum</i> (http://www.cazy.org).....	16
Table I.2: Classification of CBM fold families	23
Table I.3: Classification of CBM types	23
Table II.1: X-ray data and structure quality statistics for <i>Ct</i> CBM11.	44
Table II.2: Structural statistics for the NMR structures of <i>Ct</i> CBM11.	46
Table II.3: NMR experiments and acquisition details for the <i>Ct</i> CBM11 resonance assignment.....	54
Table II.4: Short-range distances in the secondary structure elements.	56
Table III.1: Quantitative assessment of <i>Ct</i> CBM11 binding to oligosaccharides and polysaccharides as determined by ITC.....	65
Table III.2: Binding of wild type <i>Ct</i> CBM11 and its mutant derivatives to soluble polysaccharides quantified by affinity gel electrophoresis (AGE).....	66
Table III.3: ¹ H chemical shifts of cellobiose, cellotetraose, cellohexaose and laminarihexaose in D ₂ O.	71
Table III.4: ¹³ C chemical shifts of cellobiose, cellotetraose, cellohexaose and laminarihexaose in D ₂ O.	71
Table III.5: Linewidths at half-height for the different protons of cellohexaose during the titration experiment.	75
Table III.6: Amplification factors and epitope mapping for the interaction between <i>Ct</i> CBM11 and cellotetraose and cellohexaose.....	80
Table III.7: Self diffusion coefficients measured for the mixture of sugars with and without the protein.	83
Table III.8: Quantitative assessment of <i>Ct</i> CBM11 binding to cellohexaose and cellotetraose, using the NH resonance of Tyr129 as a probe.	86
Table III.9: Average relaxation data and estimation of total correlation time (τ_m) taken from R_2/R_1 ratios.....	95
Table III.10: Characterization of the diffusion tensor obtained for <i>Ct</i> CBM11 at the different experimental conditions, obtained with Tensor2.0 and HYDRONMR.....	98
Table III.11: Average order parameter (S^2) and dynamic model used to fit the data of the different experimental conditions, obtained with Tensor2.0.....	100
Table III.12: Estimation of the conformational entropy from NMR relaxation data.....	102

Table III.13: Series of ^{15}N - ^1N -HSQC spectra acquired in order to analyze the decay of the amide proton signal intensities due to hydrogen exchange with D_2O for the free and bound <i>CtCBM11</i> at 298 K	121
Table IV.1: Quantitative assessment of <i>CtCBM30</i> and <i>CtCBM44</i> binding to oligosaccharides and polysaccharides as determined by ITC	136
Table IV.1: Amplification factors and epitope mapping for the interaction between <i>CtCBM30</i> and <i>CtCBM44</i> with cellotetraose, cellohexaose and laminarihexaose	144
Table V.1: X-ray data and structure quality statistics for the <i>Clostridium thermocellum</i> Orf2 type II Coh–XDoc complex.....	166
Table V.2: Calcium coordination in the dockerin domain	170
Table V.3: XDoc interface hydrogen bonds and salt bridges	172
Table V.4: X-Coh contacts	173
Table V.5: Coh-Doc interface hydrogen bonds	174
Table V.6: Coh-Doc interface hydrogen bonds in the 180° -rotated complex	176
Table VI.1: X-ray data and structure quality statistics for the <i>B. cellulosolvens</i> type II Coh–Doc complex.....	188
Table VI.2: Calcium coordination in the dockerin domain.....	191
Table VI.3: Coh-Doc interface hydrogen bonds	193
Table VI.4: Coh-Doc interface hydrogen bonds in the 180° -rotated complex	195
Table VII.1: A summary of some key developments that have had a major influence on the practice and application of high-resolution NMR spectroscopy in chemical research ..	206
Table VII.2: Random coil chemical shifts for common amino acids.....	207
Table VII.3: Typical spin coupling constants in amino acids	209
Table VII.4: Pulse sequences typically used for protein structure determination as described in this chapter	232
Table VII.5: Different models that can be used in a model-free analysis of relaxation rates ..	253
Table VII.6: Dissociation rates for known K_d values assuming that k_{on} is diffusion controlled.....	260
Table VIII.1: Space groups in proteins	281

Nomenclature

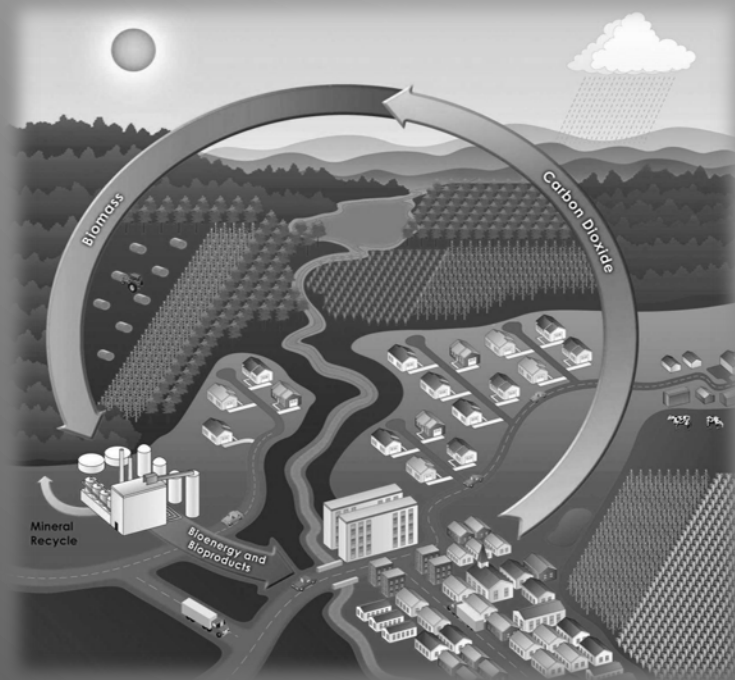
σ_0^{corr}	Corrected standard deviation to zero
^{13}C - ^{15}N - CtCBM11	Double-labeled (^{13}C and ^{15}N) CtCBM11
$^{15}\text{NH}_4\text{Cl}$	^{15}N -labeled ammonium chloride
A/Ala	Alanine
Abs	Absorbance
Ac	<i>Acetivibrio cellulolyticus</i>
AGE	Affinity Gel Electrophoresis
A_{STD}	STD amplification factor
<i>B. cellulosolvens</i>	<i>Bacteroides cellulosolvens</i>
Bc	<i>Bacteroides cellulosolvens</i>
BCA	Bicinchoninic acid
BSA	Bovine Serum Albumin
<i>C. fimi</i>	<i>Cellulomonas fimi</i>
<i>C. thermocellum</i>	<i>Clostridium thermocellum</i>
C/Cys	Cysteine
Ca	<i>Clostridium acetobutylicum</i>
CBF	Cellulose-Binding Factor
CBM	Carbohydrate-Binding Module
Cc	<i>Clostridium cellulolyticum</i>
Cc	<i>Clostridium cellulovorans</i>
CE	Carbohydrate Esterase
Cel	Cellulase
Cf	<i>Cellulomonas fimi</i>
CfCBM2	Family 2 Carbohydrate-Binding Module from <i>Cellulomonas fimi</i>
CfCBM4	Family 4 Carbohydrate-Binding Module from <i>Cellulomonas fimi</i>
CipA	Cellulosome-integrating protein A
Cj	<i>Clostridium josui</i>
Cj	<i>Cellvibrio japonicus</i>
CjCBM10	Family 10 Carbohydrate-Binding Module from <i>Cellvibrio japonicus</i>
COSY	Correlation Spectroscopy
Cp	<i>Clostridium papyrosolvens</i>

<i>Coh</i>	Cohesin
<i>Ct</i>	<i>Clostridium thermocellum</i>
<i>CtCBM11</i>	Family 11 Carbohydrate-Binding Module from <i>Clostridium thermocellum</i>
<i>CtCBM3</i>	Family 3 Carbohydrate-Binding Module from <i>Clostridium thermocellum</i>
<i>CtCBM30</i>	Family 30 Carbohydrate-Binding Module from <i>Clostridium thermocellum</i>
<i>CtCBM44</i>	Family 44 Carbohydrate-Binding Module from <i>Clostridium thermocellum</i>
<i>Cthe</i>	<i>Clostridium thermocellum</i>
D/Asp	Aspartate
D₂O	Deuterium oxide
DOSY	Diffusion Ordered Spectroscopy
Doc	Dockerin
<i>E. coli</i>	<i>Escherichia coli</i>
E/Glu	Glutamate
<i>Ec</i>	<i>Erwinia chrysanthemi</i>
<i>EcCBM5</i>	Family 5 Carbohydrate-Binding Module from <i>Erwinia chrysanthemi</i>
ESRF	European Synchrotron Radiation facility
F/Phe	Phenylalanine
F1	Direct dimension
F2	Indirect dimension
FCUP	Faculdade de Ciências da Universidade do Porto
G/Gly	Glycine
GH	Glycoside Hydrolase
GH26	Family 26 glycoside hydrolase
GH44	Family 44 glycoside hydrolase
GH5	Family 5 glycoside hydrolase
GH9	Family 9 glycoside hydrolase
GT	Glycosyltransferase
H/His	Histidine
H₂O	Water

HEPES	4-(2-Hydroxyethyl)-1-piperazine-ethanesulfonic acid
HetNOE	Heteronuclear steady-state NOE
HSQC	Heteronuclear Single Quantum Coherence
I/Ile	Isoleucine
IPTG	Isopropyl 1-thio- β -D-galactopyranoside
ITC	Isothermal Titration Calorimetry
K/Lys	Lysine
K_a	Equilibrium affinity constant
K_d	Equilibrium dissociation constant
kHz	Kilohertz
K_M	Michaelis constant
L/Leu	Leucine
LB	Luria-Bertani growth medium
Lic	Lichenase
LMW	Low Molecular Weight
M/Met	Methionine
M9	Minimal medium
MAD	multiwavelength anomalous diffraction
MD	Molecular Dynamics
MM	Molecular Mechanics
MT	Mega tons
<i>Mv</i>	<i>Micromonospora viridifaciens</i>
<i>Mv</i>CBM32	Family 32 Carbohydrate-Binding Module from <i>Micromonospora viridifaciens</i>
N/Asn	Asparagine
NaCl	Sodium chloride
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
OD	Optical density
OH	hydroxyl
OlpA	Outer-layer protein component A
OlpB	Outer -layer protein component B

OlpC	Outer -layer protein component C
Orf2	Open reading frame 2
P/Pro	Proline
PDB	Protein Data Bank
PEG	Polyethyleneglycol
PKD	Polycystic kidney disease
PL	Polysaccharide Lyase
ppm	Parts per million
Q/Gln	Glutamine
R/Arg	Arginine
R₁	Longitudinal relaxation rate
R₂	Transverse relaxation rate
Ra	<i>Ruminococcus albus</i>
Rf	<i>Ruminococcus flavefaciens</i>
rmsd	Root Mean Square Deviation
rpm	Rotations per minute
S/Ser	Serine
SdbA	Scaffoldin dockerin binding protein A
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
Sl	<i>Streptomyces lividans</i>
SlCBM13	Family 13 Carbohydrate-Binding Module from <i>Streptomyces lividans</i>
SLH	S-Layer Homology
STD-NMR	Saturation Transfer Difference Nuclear Magnetic Resonance
T/Thr	Threonine
T₁	Longitudinal relaxation
T_{1ρ}	Spin-lock filter
T₂	Transverse relaxation
TLS	Translation, Libration and Screw-rotation
Tm	<i>Thermotoga maritima</i>
T_{max}	Maximal temperature
TmCBM9-2	Family 9 Carbohydrate-Binding Module from <i>Thermotoga maritima</i>
T_{min}	Minimal temperature

TOCSY	Total Correlation Spectroscopy
<i>T_{opt}</i>	Optimum temperature
<i>Tr</i>	<i>Trichoderma reesei</i>
TrCBM1	Family 1 Carbohydrate-Binding Module from <i>Trichoderma reesei</i>
TSP	Trimethylsilyl propionate
<i>Tt</i>	<i>Tachypleus tridentatus</i>
TtCBM14	Family 14 Carbohydrate-Binding Module from <i>Tachypleus tridentatus</i>
<i>Ud</i>	<i>Urtica dioica</i>
UdCBM18	Family 18 Carbohydrate-Binding Module from <i>Urtica dioica</i>
V/Val	Valine
W/Trp	Tryptophan
XDoc	Module X-dockerin
Y/Tyr	Tyrosine
γ	Magnetogyric ratio
ΔG	Binding Gibbs energy
ΔH	Binding enthalpy
ΔS	Binding entropy
$\Delta\delta_{\text{comb}}$	Combined chemical shift
$\Delta\nu_{1/2}$	Half-height linewidth



Chapter I: Introduction - The Importance of the Research

In this chapter I give an introduction to the plant cell wall degradation theme, explaining how some microorganisms master this task. I will provide an overview on the cellulosome and on the modules responsible for its assembly and architecture (cohesin and dockerin) and efficiency (carbohydrate-binding modules). In the end I will show some biotechnological applications that can result from understanding how this nanomachines work at the molecular level.

Table of Contents

Summary	3
I.1 Introduction	3
I.2 The plant cell wall	6
I.2.1 Cellulose.....	7
I.2.2 Xyloglucan.....	8
I.3 Plant cell wall hydrolysis.....	9
I.3.1 Enzymatic hydrolysis: The cellulosome.....	10
I.4 The cellulosome of <i>Clostridium thermocellum</i> : architecture and function	14
I.5 The cohesin-dockerin interaction	17
I.6 Carbohydrate-binding modules	19
I.6.1 Nomenclature of CBMs	21
I.6.1.1 Type A CBMs – surface-binding	24
I.6.1.2 Type B CBMs – glycan-chain-binding	24
I.6.1.3 Type C CBMs – small sugar-binding.....	25
I.6.2 Molecular determinants of binding	25
I.6.3 Utilization of CBMs.....	27
I.7 Objectives and outline of the thesis.....	29
I.8 References	30

Summary

In this introductory chapter I will give an introduction on the plant cell wall degradation theme, explaining how some microorganisms master this task (*Sections I.2 and I.3*). A special attention will be given to the cellulosome of the bacterium *Clostridium thermocellum* (*C. thermocellum*, *Ct* – *Section I.4*) and its constituents, namely on the modules responsible for cellulosome assembly and architecture (cohesin and dockerin – *Section I.5*) and efficiency (carbohydrate-binding modules – CBMs – *Section I.6*). In the end I will show some biotechnological applications that can result from understanding how this nanomachines work at the molecular level. Finally I will explain the objectives of the work and make a small outline of the thesis.

I.1 Introduction

The plant cell wall is composed mainly of cellulose and hemicellulose (15-40% and 30-40%, respectively)¹ and its degradation is one of the most important steps in the global turnover process of atmospheric CO₂, therefore, of considerable biological and biotechnological importance.² Regardless of its abundance in nature, cellulose is a particularly difficult polymer to degrade, as it is insoluble and is present as hydrogen-bonded crystalline fibers, coated with hemicellulose chains and pectin all “glued” into an intricate 3D network (*see Section I.2*).³ At the present time, biomass accounts for about 10% of the world’s primary energy consumption. The other 90% is made up of nonrenewable fossil fuels (80%), hydroelectricity (2%), nuclear energy (6%), and renewable solar energies (2%).¹

Both the cellulose and hemicellulose fractions are polymers of sugars, and thereby a potential source of fermentable sugars that can be used for ethanol production (**Figure I.1**) and other products of economic interest like acetone, alcohols and volatile fatty acids.^{1,2} Economic production of ethanol from cellulosic biomass on commercial scales will help reduce our dependence on fossil fuels. Ethanol produced from biological sources can efficiently be used as a gasoline replacement or additive and, when compared to fossil fuels, presents many advantages, namely²:

- Unblended ethanol burns more cleanly and more efficiently,
- Has a higher octane rating,
- It is thought to produce smaller amounts of ozone precursors (thus decreasing urban air pollution),

- Has a low net CO₂ put into the atmosphere,
- It is significantly less toxic to humans than gasoline,
- Reduces smog formation because of low volatility,
- Its high heat of vaporization, high octane rating, and low flame temperature yield good engine performance.

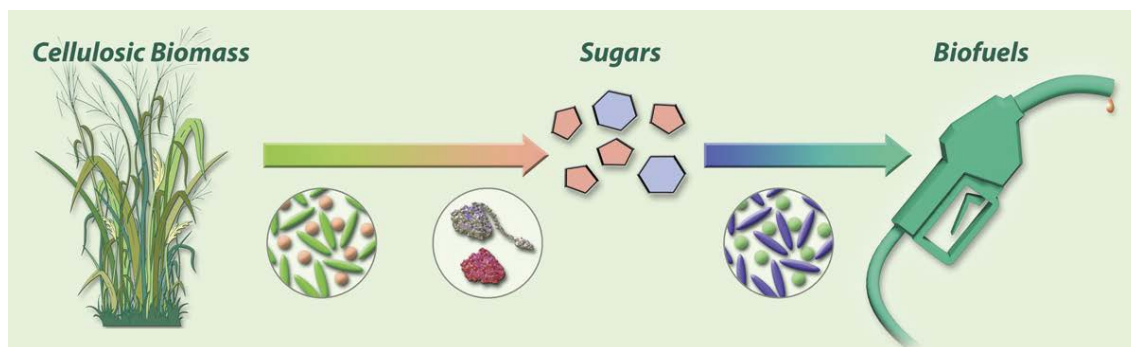


Figure I.1: From biomass to biofuels.

The goal is to develop crops dedicated to biofuels production. The biomass would then be broken down into fermentable sugars by microbes (for instance *C. thermocellum*) that would convert them into biofuel. Adapted from: <http://genomics.energy.gov>.

Furthermore, ethanol produced by fermentation offers a more favorable trade balance and a major opportunity for a depressed agricultural economy. Nevertheless, due to the complexity of the plant cell wall, most methods for producing biofuel from biomass are still relatively expensive when compared to fossil fuels.

Efficient methods for degrading cellulose chains have been intensively investigated worldwide in the last decades.^{1,4-8} The degradation of plant cell wall polysaccharides into soluble sugars has been found to be possible either by chemical means or by certain microorganisms.² The latter method has become the most attractive due to economic and efficiency reasons. The potential quantity of ethanol that could be produced from cellulose is over an order of magnitude larger than that producible from corn. As a result, microorganisms that metabolize cellulose have gained prominence in recent years.^{2,4,7,9} One of these microorganisms is the anaerobic cellulolytic thermophilic bacterium, *Clostridium thermocellum*.¹⁰⁻¹² *Clostridium thermocellum* produces an extracellular complex - **cellulosome**^{11,13} (see Section I.3.1) - capable of hydrolyzing the cell wall with the formation of cellobiose* and other cellodextrins† as main products that can be further utilized by the organism. The final products are ethanol, acetic acid, lactic acid, hydrogen, and carbon dioxide.²

* Cellobiose is a disaccharide composed of two glucose units linked by a β -1,4 glycosidic bond. As each glucose unit is rotated 180° relative to the previous, cellobiose is the structural subunit of cellulose.

† Cellodextrins are glucose polymers of varying length resulting from the breakdown of cellulose. They are classified by the degree of polymerization (DP): DP=2 – cellobiose; DP=3 – cellotriose; DP=4 – cellotetraose; etc

In fact, there are several advantages of using *C. thermocellum* for ethanol fermentation from biomass:²

- The cellulolytic and ethanogenic nature, allowing saccharification and fermentation in a single step,
- The anaerobic nature, avoiding the need for expensive oxygen transfer,
- Low cell growth yield, favoring ethanol conversion,
- The thermophilic nature, facilitating ethanol removal and recovery and reducing cooling cost,
- Thermophilic fermentation being less prone to contamination,
- Thermophilic biomass-degrading enzymes enhancing protein stability.

In order to efficiently hydrolyze the plant cell wall, these mega-Dalton extracellular machines are composed of a huge paraphernalia of enzymes and non-catalytic modules (*see Section I.4*). The enzymes present reflect the composition and complexity of the plant cell wall¹⁴ and, in order to increase their catalytic activity, most enzymes are linked to one or more non-catalytic **carbohydrate-binding modules** (CBMs).¹⁵ These modules, as reflected by their name, bind to carbohydrates and have a fundamental role in the enzymatic degradation of plants and in polysaccharide storage due to their high specificity and substrate recognition mechanisms. Due to their key importance in recycling carbon from plant biomass, these enzyme systems have a considerable biotechnological potential (*see Section I.6.3*). Profound knowledge about the cellulosome assembly and, more important, about the specificity of the different CBMs, will bring a relevant contribution to the possible engineering of more efficient catalysts. Furthermore, the rationalization of the molecular recognition mechanisms that determine the specificity of these proteins opens the way for the creation of efficient and low cost mechanisms for the conversion of biomass into ethanol.

Cellulosomes are bound to the bacterial cell wall via the type II cohesin-dockerin interaction (*see Section I.5 and Chapters V and VI*).^{16,17} This interaction promotes the close contact between the microbe and the substrate enabling the ready uptake of simple sugars resulting from polysaccharide hydrolysis and thus, representing an evolutionary advantage.^{9,10} On the other hand, the various catalytic subunits are incorporated into the cellulosome complex by virtue of a key non-catalytic polypeptide, called scaffoldin, which bears a collection of type I cohesin modules for this purpose. Each type I cohesin binds a single dockerin domain located on the enzymes, thereby generating the fully assembled cellulosome.^{18,19} The arrangement of these modules on the scaffoldin subunit and their specificity for the modular counterpart dictates the overall architecture of the cellulosome (*see Section I.5*).²⁰ The specificity displayed between

type I and type II cohesin-dockerin interactions is thus of major importance to cellulosome assembly and attachment.

I.2 The plant cell wall

Among all the features that distinguish plant cells from animal cells, the presence of a plant cell wall is the most distinctive. Its presence is the basis of many of the characteristics of plants as organisms. The plant cell walls are not simply an outer, inactive shell of the plant cell itself but rather dynamic structures that play critical roles such as:

- Structural support allowing the organism to build and hold its shape
- Protection against mechanical stress
- Limits the entry of large molecules that may be toxic to the cell acting as a filtering mechanism
- Creates a stable osmotic environment preventing enlargement of the plant cell and osmotic lysis
- It's involved in absorption, transport and secretion of substances in plants
- Cell-cell interactions
- Source of biological signaling molecules

Plants can have two types of cell walls: primary and secondary. Primary cell walls surround growing and dividing plant cells, providing mechanical strength but allowing the cells to expand. They are composed of cellulose microfibrils that are extensively cross-linked by hemicellulose polysaccharide chains and pectin all woven into an intricate network (**Figure I.2**).²¹ In contrast, secondary walls are much thicker and stronger and are deposited only when cells have ceased growing. In some higher plants, the secondary walls are strengthened by the incorporation of lignin. Lignin is the general name for a group of polymers of aromatic alcohols that are hard and give considerable strength to the structure of the secondary wall preventing biochemical degradation and physical damage by fungi or bacteria but its structure and organization within the cell wall are poorly understood. The association of cellulose, hemicellulose and lignin is named lignocellulose and its quantitative composition depends on the plant species, age and growth conditions.

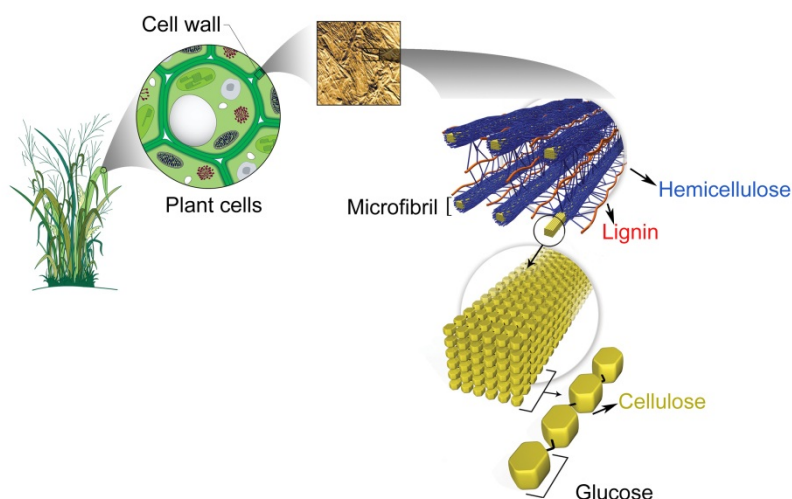


Figure I.2: Plant cell wall structure.

Adapted from: <http://genomics.energy.gov>.

I.2.1 Cellulose

Cellulose is the structural component of the primary cell wall of green plants, but it is also found in many forms of algae, bacteria and the oomycetes[‡]. About 33% of all plant matter is cellulose, which makes this polymer the most common organic compound on Earth.²² Cellulose is a linear polymer composed of several hundred to over ten thousand of β -1,4-D-glucopyranose units in 4C_1 conformation (**Figure I.3**). Each glycosyl residue is oriented at an angle of 180° to the next residue of the chain, which makes cellobiose (a disaccharide) the repeating structural unit. The glycosyl residues form one covalent bond at $C1\beta-C4'$ plus intramolecular hydrogen bonds at $O3-H\rightarrow O5'$ and $O6\rightarrow H-O2'$ and intermolecular $O6-H\rightarrow O3'$.²³

This extensive hydrogen bond network keeps the strands tightly bound and gives rise to complex three-dimensional structures. The chains of cellulose associate with other polymers to form linear structures of high tensile strength known as microfibrils which consist of up to 40 cellulose chains and have about 10 to 20 nm in diameter. This complex structure, allied with tightly intercalated lignin and hemicellulose leads to a structural resistance that prevents enzymes (cellulases and hemicellulases) from attacking cellulose.^{3,23} Therefore, pretreatment of biomass (with acids for instance) is necessary to remove the surrounding matrix of hemicellulose and lignin prior to cellulose hydrolysis.

[‡] Oomycetes - distinct phylogenetic lineage of fungus-like eukaryotic microorganisms (Protists).

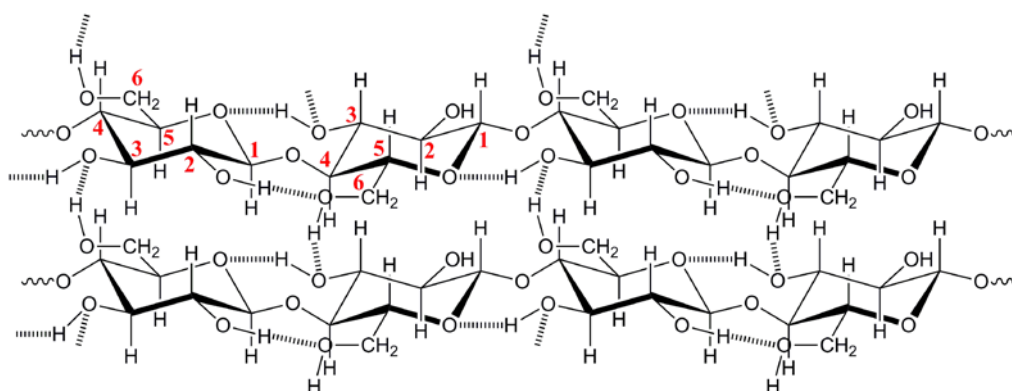
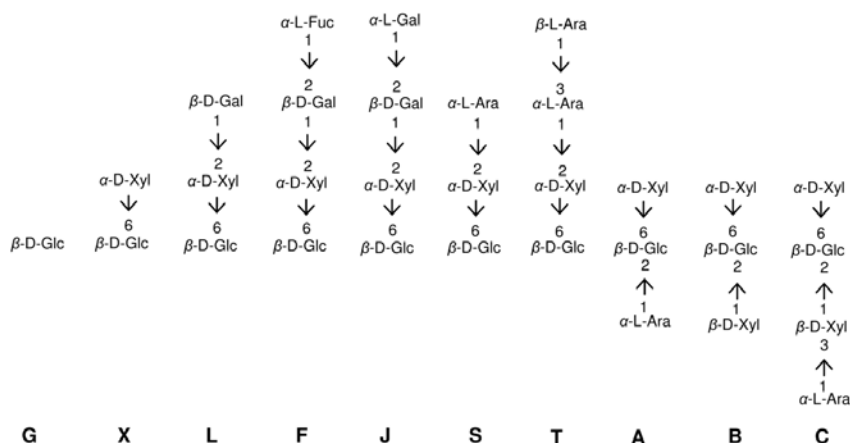


Figure I.3: Structure of cellulose.

The picture shows two adjacent cellulose chains and the glycosidic and hydrogen bonds holding them together. Note the parallel arrangement with the reducing ends aligned in the same direction.

I.2.2 Xyloglucan

Hemicellulose is collective term used to describe a family of polysaccharides composed of different sugars such as xylose, mannose, galactose, rhamnose and arabinose, among others and xyloglucan is the most abundant polysaccharide of the hemicellulose present on the primary cell wall in many dicotyledonous. It consists of α -1,6-D-xylosyl residues along a β -1,4-glucan backbone with additional branching of α -L-arabinose or β -D-galactose in a species-dependent manner. Because the β -1,4-glucan backbone binds to the cellulose microfibrils via hydrogen bonds, xyloglucan confers rigidity to the cell wall by cross-linking adjacent microfibrils. In fact, microfibrils are covered in xyloglucan, which is located both on and between microfibrils.¹ A single-letter nomenclature is used to simplify the xyloglucan nomenclature according to the substituent. For instance: a G represents an unbranched glucose unit, an X represents a glucose unit with a 1,6-linked xylose, an F represents a glucose residue with a fucose-containing trisaccharide and so on (**Figure I.4**).²⁴



I.3 Plant cell wall hydrolysis

Lignocellulosic biomass is composed of cellulose, hemicellulose and lignin and is the most abundant renewable natural resource on Earth with a global production of about 1×10^{10} MT.^{2,8} Because the cellulose and hemicellulose fractions are polymers of sugars they can be used as a source of fermentable sugars for conversion into fuels. Lignocellulose is inexpensive, plentiful and renewable. The hemicellulose fraction can be easily hydrolyzed under mild acid or alkaline conditions whilst cellulose requires more rigorous treatment since it is more resistant. Cellulose is a very stable molecule, with a half-life of several million years for spontaneous β -glycosidic bond cleavage at room temperature. This means that practically all cellulose degradation in Nature is accomplished by enzymatic action.¹ The general protocol for conversion of lignocellulosic biomass into fermentable sugars involves three steps:^{4,6}

1. An initial milling step to grind the raw materials and increase the surface area;
2. A pretreatment process to make the cellulose microfibrils accessible. In this step hydrolysis of hemicellulose may occur (depending on the process conditions) as well as separation of the lignin fraction (for production of chemicals, combined heat and power production or other purposes);
3. Enzymatic cellulose hydrolysis to liberate the monosaccharides.

Current research is focused on converting biomass into its constituents in a market competitive and environmentally sustainable way and an improvement of pretreatment technologies and enzymatic hydrolysis gives scope for numerous ongoing research projects.

Pretreatment methods can be chemical, thermal, physical or any combination of the three. To achieve higher efficiency a combination of physical and chemical means is required. Physical methods (often called size reduction) are used to trim down biomass physical size. Chemical methods remove the chemical barriers allowing enzymes to hydrolyze cellulose.²⁵ The pretreatment step is one of the most expensive ones for the extractions of sugars from biomass. Over the years a “wish list” of pretreatment attributes has been developed. As a result, a successful pretreatment should:^{4,26}

- Maximize the enzymatic convertibility and minimize the loss of sugars
- Maximize the production of other valuable by-products, e.g. lignin
- Not require the addition of toxic chemicals
- Minimize the use of energy, chemicals and capital equipment
- Be scalable to industrial size.

Nevertheless, full accomplishment of all the above issues is very difficult, with the last two points being fundamental for economical and practical viability of the entire process.

I.3.1 Enzymatic hydrolysis: The cellulosome

As referred above, despite its chemical homogeneity, cellulose is a very stable molecule and no single enzyme is able to hydrolyze it.⁹ Efficient hydrolysis of cellulose requires the synergistic action of several enzymes that can be divided into three classes:

- **endo-1,4- β -D-glucanases** (EC 3.2.1.4), which randomly hydrolyze internal β -1,4-glucosidic bonds in the cellulose chain to produce new termini available to exoglucanase attack;
- **exo-1,4- β -D-glucanases** (EC 3.2.1.91), which move along the cellulose chain and progressively cleave off cellobiose units at the reducing and non-reducing ends;
- **1,4- β -D-glucosidases** (3.2.1.21), which hydrolyze cellobiose to glucose and cleave of glucose units from cellooligosaccharides.

These enzymes work together in a synergistic way to hydrolyze cellulose by creating accessible sites for each other and reducing product inhibition.^{1,4} Furthermore, in the plant cell wall there are also hemicelluloses with their many different side groups which significantly increase its complexity. Among the enzymes responsible for degradation of hemicellulose there are:⁴

- **endo-1,4- β -D-xylanases** (EC 3.2.1.8), which hydrolyze internal bonds in the xylan chain;
- **1,4- β -D-xylosidases** (EC 3.2.1.37), which attack xylooligosaccharides from the non-reducing end and liberate xylose;
- **endo-1,4- β -D-mannanases** (EC 3.2.1.78), which cleave internal bonds in mannan;
- **1,4- β -D-mannosidases** (EC 3.2.1.25), which cleave mannooligosaccharides to mannose.
- The side groups are removed by a number of enzymes:
 - **α -D-galactosidases** (EC 3.2.1.22);
 - **α -L-arabinofuranosidases** (EC 3.2.1.55);
 - **α -glucuronidases** (EC 3.2.1.139);
 - **acetyl xylan esterases** (EC 3.1.1.72);
 - **feruloyl and *p*-cumaric acid esterases** (EC 3.1.1.73).

All these hydrolytic enzymes are relatively expensive and difficult to produce in large amounts and, therefore, significant reduction of production costs is important for their commercial use. Currently, most commercially available enzymes are produced by genetically engineered strains of filamentous fungi, particularly *Trichoderma reesei*.² However, the enzymatic hydrolysis of cellulose is generally a slow and incomplete process. On the other hand, in Nature, microorganisms have evolved in order to profit from this highly abundant source of energy. In some cases, microorganisms directly explore these polysaccharides from decaying plant matter while in other cases, in a symbiotic way, they assist higher animals (e.g. ruminants) in the conversion of the polysaccharides into digestible compounds. While aerobic microorganisms produce large amounts of relevant enzymes (e.g. cellulases and hemicellulases), the mechanism of biosynthetic anaerobic organisms is simpler with respect to the production of such enzymes. In this context, it is thought that the anaerobic environment presents a great selective pressure on the evolution of highly efficient machinery for extracellular degradation of cell wall components.²⁰ Consequently, anaerobic organisms tend to adopt alternative strategies to degrade material plant.

Anaerobic organisms secrete a large range of plant cell wall hydrolases, which are organized in multi-enzyme complexes termed cellulosomes (**Figure I.5**).^{9,13,14,20,27-30} The cellulosome was first described by Lamed *et al*^{13,16} and defined as “a discrete, cellulose binding, multienzyme complex for the degradation of cellulosic substrates” pointing to the molecular ordering of the cellulosome components. The initial cellulosome concept was based on studies in the cellulase system of the anaerobic cellulolytic thermophilic bacterium, *Clostridium thermocellum*^{10,11} (see Section I.4) and it was believed that it solely degraded cellulose (hence the initial term “cellulose-binding factor – CBF”).¹⁰ Early on it became clear that this multienzyme complex contained more than cellulases.^{16,31} Throughout the years there’s been a great effort in order to fully understand and characterize these mega-Dalton complexes. It is now clear that cellulosomes actively degrade other plant cell wall components by incorporating polysaccharide lyases, carbohydrate esterases and glycoside hydrolases in the multienzyme complex.²⁰ Cellulosome attachment to the bacterial surface enables the ready uptake of simple sugars resulting from polysaccharide hydrolysis and represents an evolutionary advantage by maintaining the microbe into close proximity with the extracellular substrates and resulting by-products.^{9,10}

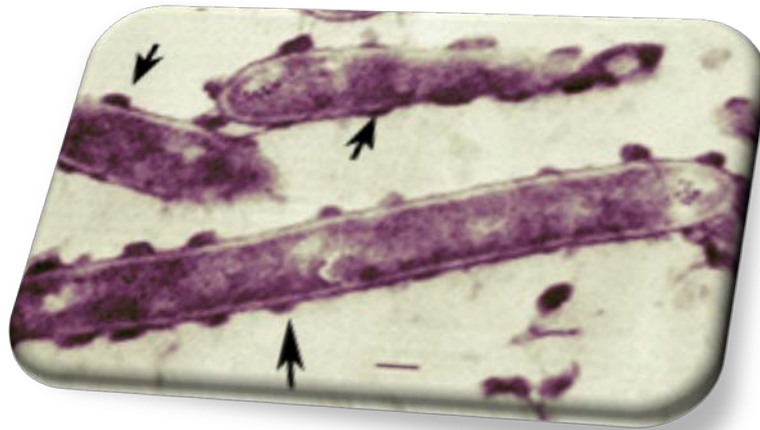


Figure I.5: Cellulosomes at the surface of *Clostridium thermocellum*.³²

The cellulosomes are indicated by the black arrows.

Basically, cellulosomes are composed of five different components (**Figure I.6**):

- **The scaffoldin subunit:** The scaffoldin subunit is a non-catalytic protein that contains one or more cohesin modules connected to other types of functional modules. Depending on the scaffoldin protein, the referred modules include a cellulose-specific carbohydrate-binding module, a dockerin, an X module of unknown function, an S-layer homology (SLH) module or a sortase anchoring motif.^{14,27} The scaffoldin is responsible for organizing the different subunits into the complex, therefore, shaping the overall architecture of the cellulosome.^{16,20} Motional freedom of the scaffoldin subunit allows precise positioning of the catalytic modules according to the topography of the substrate.³³
- **The cohesin modules:** Cohesin modules are the major building blocks of the scaffoldin subunit and are responsible for organizing the cellulolytic subunits into the multi-enzyme complex (see Section I.5).²⁷ Cohesins are classified into three groups: type I, type II and (recently) type III³⁴, according to their phylogenetic similarity.³⁴ type I cohesins are located in the scaffoldin subunit and are responsible for incorporating the different catalytic subunits; type II cohesins are located at the cell surface and are responsible for anchoring the multienzyme complex into the cell wall; type III cohesins still have an unclear function¹⁴.
- **The dockerin modules:** Dockerins are non-catalytic proteins with approximately 70 amino acids that contain two duplicated segments of about 22 residues and display internal two-fold symmetry, consisting of a duplicated F-hand calcium-binding motif (see Section I.5).^{18,35,36} Dockerins specifically bind to determined type of cohesin and,

therefore, they are named after them.²⁰ As a result we have type I, II and III dockerins that bind to type I, II and III cohesins, respectively. Essentially, the dockerin modules act as anchors: they anchor the catalytic subunits to the scaffoldin protein (type I) and anchor the scaffoldin protein to the cell wall (type II). The function of type III dockerins is still unknown. Although structurally related, type I cohesins and dockerins were shown to be different from type II and do not cross react.³⁷

- **The catalytic modules:** Cellulosomes contain an amazing diversity of enzymes that is proportional to the complexity of plant cell wall. In this sense, the array of polysaccharides presented by the plant cell walls is matched by the complexity and diversity of the cellulosomal catalytic machinery.¹⁴ The catalytic modules include glycoside hydrolases (GHs), glycosyltransferases (GTs), carbohydrate esterases (CE) and polysaccharide lyases (PL).
- **The carbohydrate-binding modules:** Carbohydrate-binding modules (CBMs) are non-catalytic proteins that bind to a wide range of poly- and oligosaccharides (*see Section I.6*).¹⁵ Their main function is to increase the activity of the associated catalytic modules by maintaining the enzyme in the proximity of the substrate through their sugar-binding activity. Furthermore, they are also responsible for anchoring the cellulosome to the substrate (targeting function) and for breaking the substrate (disruptive function).^{15,38}

To date, cellulosomes have been identified in several bacteria: *Acetivibrio cellulolyticus*³⁹, *Bacteroides cellulosolvens*^{40,41} (*see Chapter VI*), *Clostridium acetobutylicum*⁴², *Clostridium cellulolyticum*⁴³, *Clostridium cellulovorans*⁴⁴, *Clostridium josui*⁴⁵, *Clostridium papyrosolvens*⁴⁶, *Clostridium thermocellum*¹¹, *Ruminococcus albus*⁴⁷, *Ruminococcus flavefaciens*⁴⁸, and fungi¹⁴ of the genera: *Neocalimastix*, *Piromyces*, and *Orpinomyces*.

Due to the efficiency of cellulosomes in degrading the plant cell wall there's been an extensive effort in order to understand how these mega-Dalton cell-degrading nanomachines work and how they could be used to obtain valuable products from low-cost biomass or agricultural waste.^{2,7,14,49,50} Recombinant DNA technology allows the construction of engineered cellulosomes that can be specifically tuned^{9,14,20} and improved enzyme systems and self-assembling chimeric protein constituents with high potential for biotechnological and nanotechnological applications.^{14,20,51}

I.4 The cellulosome of *Clostridium thermocellum*: architecture and function

The cellulosome was first discovered in the anaerobic cellulolytic thermophilic bacterium, *Clostridium thermocellum*^{10,11} (**Figure I.6**) and much of the understanding of catalytic components, architecture and mechanisms of action derive from its study (**Table I.1**).^{14,20,28} The cellulosome of *C. thermocellum* is one of the most complex and, at the same time, one of the most studied (**Table I.1**). Its main component is the scaffoldin protein termed cellulosome-integrating protein A – CipA.⁵² CipA is a large enzyme-integrating protein composed of several modules (**Figure I.6**):

- **Nine type I cohesins:** the nine type I cohesins specifically recognize the type I dockerins in the catalytic subunits. The arrangement of these modules on the scaffoldin subunit and their specificity for the modular counterpart dictates the overall architecture of the cellulosome (*see Section I.5*).²⁰
- **A carbohydrate-binding module from family 3 (CBM3)**⁵³: the scaffoldin Type A CBM3 binds strongly to crystalline cellulose ($K_a=0.4 \mu M$),⁵⁴ therefore, mediating the attachment of the cellulosome (and its enzymes) to the cellulosic substrate. The topology of the binding interface of CBM3 rules out their interaction with single β -1,4 glucan chains, which adopt a more helical conformation.¹⁴
- **A C-terminal type II dockerin:** the C-terminal type II dockerin specifically recognizes the type II cohesins at the cell surface and is, therefore, responsible for the attachment of CipA to the bacterial cell wall (*see Section I.5 and Chapter V*).⁵⁵
- **An X module:** the X module is usually present at the N-terminal site of type II dockerins and its function is still unclear (*see Section I.5*). However it has been demonstrated that the presence of this module is fundamental for the type II cohesin-dockerin interaction (*see Chapter V*).^{17,56}

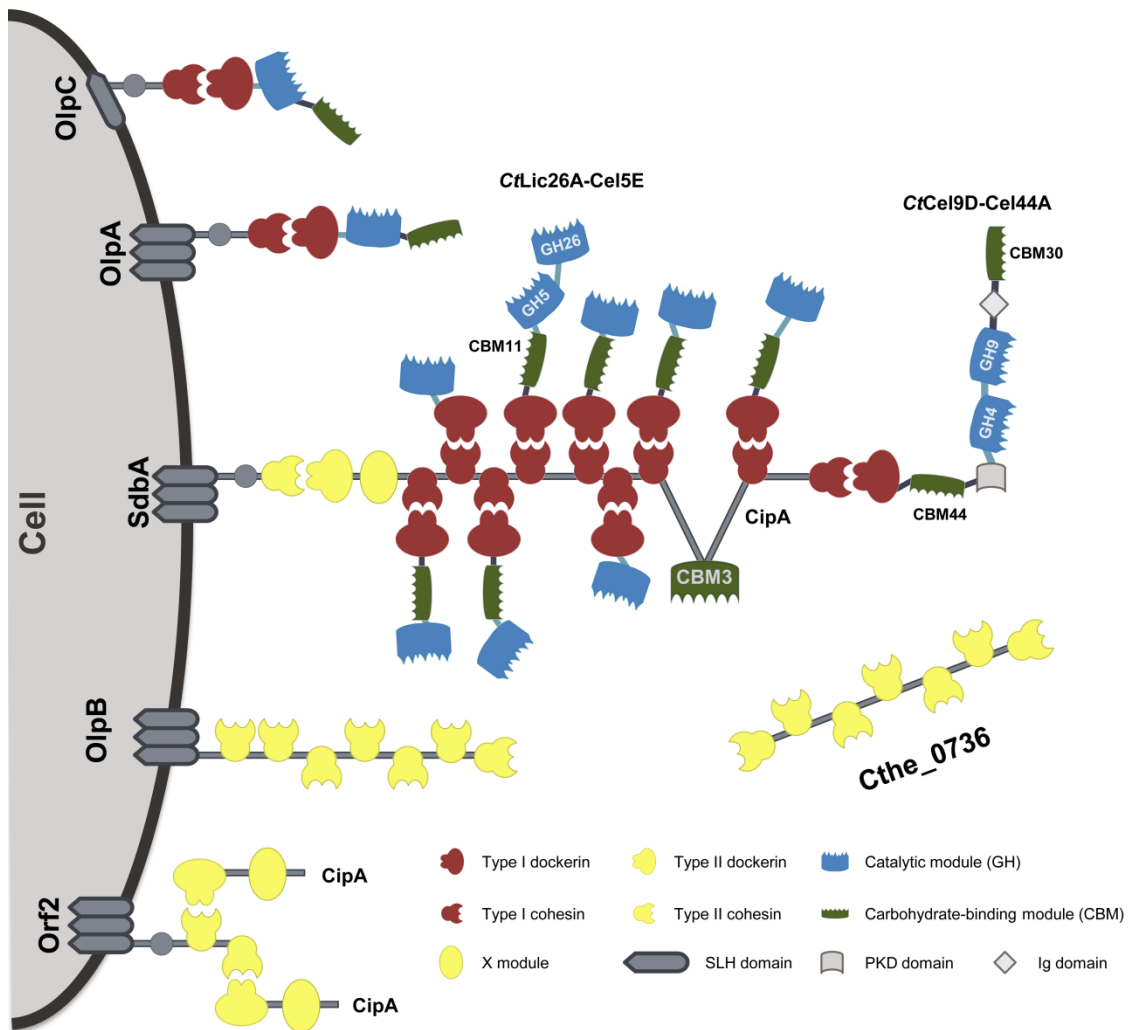


Figure I.6: Schematic representation of the *Clostridium thermocellum* cellulosome.

The cellulosome of *C. thermocellum* is composed of five SLH domains for anchoring the complex to the bacterial cell wall (Orf2, OlpA, OlpB, OlpC and SdbA) through cohesin-dockerin interactions, (type II in the case of Orf2, OlpB and SdbA and type I for OlpA and OlpC), and free scaffoldins (Cthe_0736) that do not bind the cell wall. The main component of the cellulosome of *C. thermocellum* is the scaffoldin protein CipA. This scaffoldin consists of nine type I cohesins, a CBM3, an X module and C-terminal type II dockerin that recognizes type II cohesins at the cell surface. The binding of the enzymes to specific positions is hypothetical, as is the linear orientation of the scaffoldin. The scaffoldins bound to Orf2 and OlpB are only sketched partially. All cellulosome components are not drawn to scale. Adapted from Fontes *et al*, 2010.¹⁴

The assembly of *C. thermocellum* cellulosome onto the bacterial surface is coordinated by five proteins, Orf2, OlpA, OlpB, OlpC and SdbA, which are presumed to be bound onto the *C. thermocellum* cell wall via N-terminal SLH domains.¹⁹ SdbA, Orf2p and OlpB contain type II cohesins, which bind to the type II dockerin present at the *C-terminus* of CipA and recruit the cellulosome onto the surface of the cell wall (**Figure I.6**). Furthermore there are also free scaffoldins (Cthe_0736) that do not bind to the cell wall.¹⁴ The multiple type II cohesin domains present in OlpB, Orf2, and Cthe_0736 contribute to the formation of polycellulosomes that may contain up to 63 catalytic subunits. Alternatively, cellulosomal enzymes may adhere directly to

the bacterium cell surface by binding the single type I cohesin domain found in OlpA and OlpC.¹⁴

Table I.1: List of cellulosomal components of *C. thermocellum* (<http://www.cazy.org>).

GH Family	1	2	3	5	8	9	10	11	13	15	16	18	23
Number of sequences	2	1	2	10	1	16	6	1	2	1	2	4	2
GH Family (cont.)	26	30	39	43	44	48	51	53	74	81	94	124	126
Number of sequences	3	2	1	6	1	2	1	1	1	1	3	1	1

Glycosyl Transferase Family	1	2	4	5	8	26	28	32	35	39	51	84	NC*
Number of sequences	4	9	12	2	1	1	3	1	1	1	1	1	1

Polysaccharide Lyase Family	1	9	11
Number of sequences	2	1	1

Carbohydrate Esterase Family	1	2	3	4	7	8	9	12	14	NC*
Number of sequences	3	1	2	3	1	1	1	2	1	1

CBM Family	3	4	6	9	11	13	16	22	25	30
Number of sequences	24	7	11	2	1	2	4	5	3	1
CBM Family (cont.)	32	34	35	42	44	48	50	54	62	
Number of sequences	1	1	7	4	1	1	15	1	1	

* : Non classified

An essential part of the cellulosome of *C. thermocellum* (and any cellulosome) is the catalytic machinery. As said above, cellulosomes contain several types of enzymes, such as: glycoside hydrolases, glycosyl transferases, carbohydrate esterases, polysaccharide lyases among many others.²⁸ Altogether, these cellulases and hemicellulases are able to fully degrade the plant cell wall, including crystalline forms of cellulose such as cotton and Avicel.² As in the free enzymes, cellulosomal cellulases and hemicellulases are modular entities.²⁰ Most of cellulosomal enzymes are composed of a dockerin domain, one or two catalytic units and one or more CBMs [for instance CtCBM11⁵⁷ (see Chapters II and III), CtCBM44⁵⁸ (see Chapter IV) and CtCBM30 (see Chapter IV)] whose primary function is to increase the catalytic efficiency

of the carbohydrate-active enzymes against soluble and/or insoluble substrates (see Section I.6).^{15,59} *C. thermocellum* produces 72 cellulosome-associated components that can be arranged in 72⁹ different manners (as CipA comprises nine enzyme receptors - cohesins).¹⁴ This amazing plasticity may reflect the need to adapt to the changeable composition and complexity of different plant cell walls. Furthermore *C. thermocellum* expresses cell associated β -glucosidases (at least four exoglucanases and more than ten different endoglucanases) which act in a synergistic manner in order to hydrolyze to glucose the products released by the cellulosome activity.¹⁴ For all this aspects, *C. thermocellum* exhibits one of the highest rates of cellulose utilization known.²

I.5 The cohesin-dockerin interaction

The cellulosome architecture is defined by high affinity ($K_d > 10^{-9}$ to 10^{-12} M)^{16,60} protein-protein interactions between cohesins and dockerins (**Figure I.6**). Dockerin and cohesin domains have been identified as conserved homologous sequence elements of the proteins that make up the cellulosome scaffold and enzymatic subunits.

Dockerins are non-catalytic proteins of approximately 60-70 amino acids that recognize cohesin domains and mediate the assembly of the cellulolytic subunits into the scaffoldin subunit and of the latter to the bacterial cell wall.^{14,20} The dockerin sequence is highly conserved and made up of two 22-residue sequence repeats separated by a linker region of about 9-18 residues.^{18,35} They fold into three α -helices, with helices 1 and 3 comprising the repeated segments. Within each duplicated sequence there is a 12-residue segment with sequence similarity to the calcium-binding loop of the EF-hand motif, in which all the calcium binding residues (i.e. aspartic acid and asparagines) are highly conserved.³⁶ However, because the EF-hand motif homology is restricted to the calcium-binding loop and the F-helix, structural data points to an F-hand motif instead.⁶¹ The residues that coordinate calcium (aspartate or asparagine) are conserved in loop positions 1, 3, 5, 9, and 12 of nearly all dockerins. The presence of the duplicated segment suggests that both halves of the dockerin are able to interact with the cohesin in very a similar manner.¹⁸ This means that there may be plasticity in cohesin recognition by the dockerin with either the *N*- or *C*-terminal helix. This plasticity allows, in principle, the simultaneous binding of two cohesins by a single dockerin. Such an interaction would not only provide a higher level of structure to the cellulosome but might also allow the crosslinking of two scaffoldins through a single dockerin.^{18,62} Nevertheless, the stoichiometry of type I cohesin-dockerin binding is, invariably 1:1, suggesting that the two binding sites are not able to bind simultaneously.¹⁴ Thus, it remains unclear the biological significance of the dual

binding mode in dockerins. NMR studies have showed that stability and function of the cohesin modules is calcium dependent. In fact, in the absence of calcium cohesins and dockerins were shown not to interact.⁶³

Cohesins are 150-residue modules, usually present as tandem repeats in scaffoldins. They are elongated, conical molecules that comprise a jelly-roll topology that folds into a nine-stranded β -sandwich. The cohesin modules are the main components of the scaffoldin subunit and are responsible for organizing the cellulolytic subunits into the cellulosome.²⁷ According to their phylogenetic relationship, cohesins have been separated into three distinct types: type I, type II (Figure I.7) and type III.^{14,34} By definition, the dockerins that interact with each type of cohesin are of the same type. Most of the glycosyl hydrolases contain a C-terminal type I dockerin domain which binds type I cohesins found in the scaffold. The type II interaction is used for anchoring the scaffoldins to the cell wall (type II cohesins at the cell surface interact with their dockerin counterparts at the C-terminal of the scaffoldin subunit). The function of the type III interaction is still unclear¹⁴.

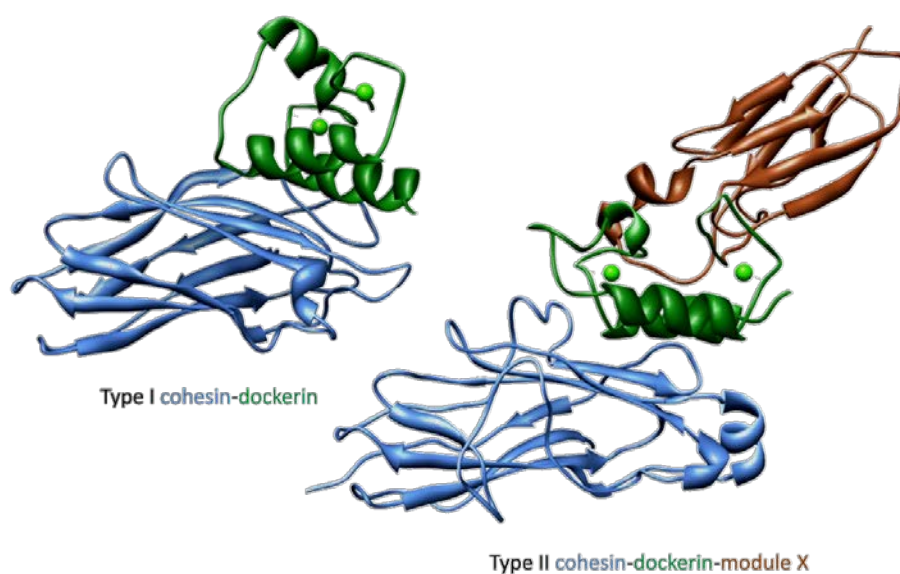


Figure I.7: The cohesin-dockerin complex.

In both complexes, cohesin-dockerin recognition is dominated by hydrophobic interactions, amplified through an extensive hydrogen-bonding network. Cohesin modules are depicted in blue, dockerin modules are depicted in green and the X module is depicted in brown. The light green spheres represent calcium ions (Ca^{2+}) bound to the dockerins. The structures represented are from *C. thermocellum*. The type I complex¹⁸ (PDB code: 1ohz) and the type II complex (PDB code: 2vt9 - see Chapter V) were determined by X-ray crystallography.

Type II dockerins are usually present at the *C-terminus* side of a module of unknown function termed X module.⁵⁶ The importance of this module in the type II cohesin-dockerin interaction was recently demonstrated¹⁷ through the resolution of the structure of the cohesin-dockerin-X module complex. The type II dockerin, which displays a fold similar to its type I

counterpart, establishes an extensive range of interactions with the X module that adopts an immunoglobulin-like fold.

Although structurally related, type I cohesins and dockerins were shown to be different from type II (15-25% identity) and do not cross react³⁷. In fact, comparison of the primary structure of *C. thermocellum* cohesins and dockerins shows a small degree of similarity between them, consistent with the lack of cross-specificity between type I and type II cohesin–dockerin pairs.¹⁹ Several studies show that type I cohesins of *C. thermocellum* recognize almost all of type I dockerins present on the enzymatic subunits^{16,55} but, interestingly, type I and type II cohesin/dockerins partners do not interact, ensuring a clear distinction between the mechanism for cellulosome assembly and cell-surface attachment.⁵⁵ Furthermore, it was also shown that, although type I cohesins/dockerins from one species do not interact with other type I cohesins/dockerins from other species,^{61,64} type II cohesins/dockerins demonstrate a rather extensive cross-species plasticity.⁶⁵ The biological relevance of this cross-species interaction is still uncertain. The fact that type I cohesins in the enzymatic units recognize nearly all the type I dockerins in the scaffoldin unit suggests that, within a given species, the arrangement of the several enzymes occurs randomly along the scaffoldin, reflecting, perhaps, the complexity of the substrate in the microbial environment.¹⁴

I.6 Carbohydrate-binding modules

In order to degrade the highly complex plant cell wall, microorganisms have developed a specialized complex (cellulosome) composed of multiple enzymes and non-catalytic modules. Many carbohydrate-active enzymes are modular proteins bound to one or more non-catalytic carbohydrate-binding modules (CBMs) that function in an independent manner.^{15,59} These modules were first described in 1988^{66,67} and named as cellulose-binding domains based on the discovery of several modules that bound cellulose. Later, with the discovery of other modules with specificities other than cellulose the name was changed to CBM (*see the Section I.6.1*). A CBM is defined as a continuous amino acid sequence within a carbohydrate-active enzyme with a separate fold having carbohydrate-binding activity.⁶⁸ To date several hundred putative CBM sequences have been identified experimentally in more than 50 species and they have been classified into 64 different families according to their sequence similarity. (Carbohydrate Active Enzymes database - <http://www.cazy.org>).⁶⁹ CBMs are composed of 30 to 200 amino acids and they occur as a single, double or triple domain in one protein. They can be found at the C- or N-terminal of the catalytic protein and, invariably, their key role is to recognize and specifically bind to the several different carbohydrates found in the plant cell wall.^{15,38,59} This specific

recognition and binding to the carbohydrates of the plant cell wall has considerable biological consequences such as:¹⁵

- Anchoring the multienzyme complex to the substrate;
- Bringing the catalytic domain in close proximity to the substrate and, therefore, enhancing the hydrolysis of insoluble substrates through an effective increase of the concentration of cellulase on the surface of the substrate;
- Disrupting the structure of the polysaccharides.

The first studies on the cellulosome of the bacterium *C. thermocellum*^{10,13} have shown that it was tightly bound to cellulose but, at that time, the reason for that was still unclear. Later it was shown that this strong adherence to cellulose was mediated by a family 3 carbohydrate-binding module (CtCBM3) belonging to the scaffoldin protein (CipA).⁵³ The first studies of CBM-cellulose interaction also showed that removal of the CBM from the cellulase or from the scaffoldin dramatically reduces the enzymatic activity.^{66,70} Furthermore, it was shown that adding a CBM to a carbohydrate-active enzyme results in increased hydrolytic activity.⁷¹ Besides this proximity function, some CBMs also have a non-catalytic disruptive function which is thought to also enhance the hydrolytic capacity of the catalytic modules.^{15,38} Studies have revealed that the mechanism involved in carbohydrate disruption involves modification of the hydrogen bond network in cellulose.⁷² Binding of CBMs to carbohydrates is seldom irreversible as their mobility is fundamental for relocation of the enzymes to new regions of the substrate. Conversely, there are examples of such kind of interaction (for instance CMB2a from *C. fimi*)⁷³ although its biological significance remains uncertain and, at the same time, senseless, as the enzyme activity is unlikely to be enhanced (proximal cleavage sites accessible to the enzyme's active site will be quickly exhausted).

Our knowledge on these systems has grown considerably over the last years as a result of structural information provided by NMR spectroscopic and X-ray crystallographic studies^{15,59,74} deepening our understanding on the biological functions of CBMs. In addition to plant cell wall carbohydrate recognition, CBMs are involved in a large number of other processes such, pathogen defense, polysaccharide biosynthesis, virulence, plant development, etc.³⁸ Therefore, understanding of the CBMs properties and mechanisms of ligand binding and recognition is imperative for the development of new carbohydrate-recognition technologies and for providing the basis for fine manipulation of the carbohydrate–CBM interactions.

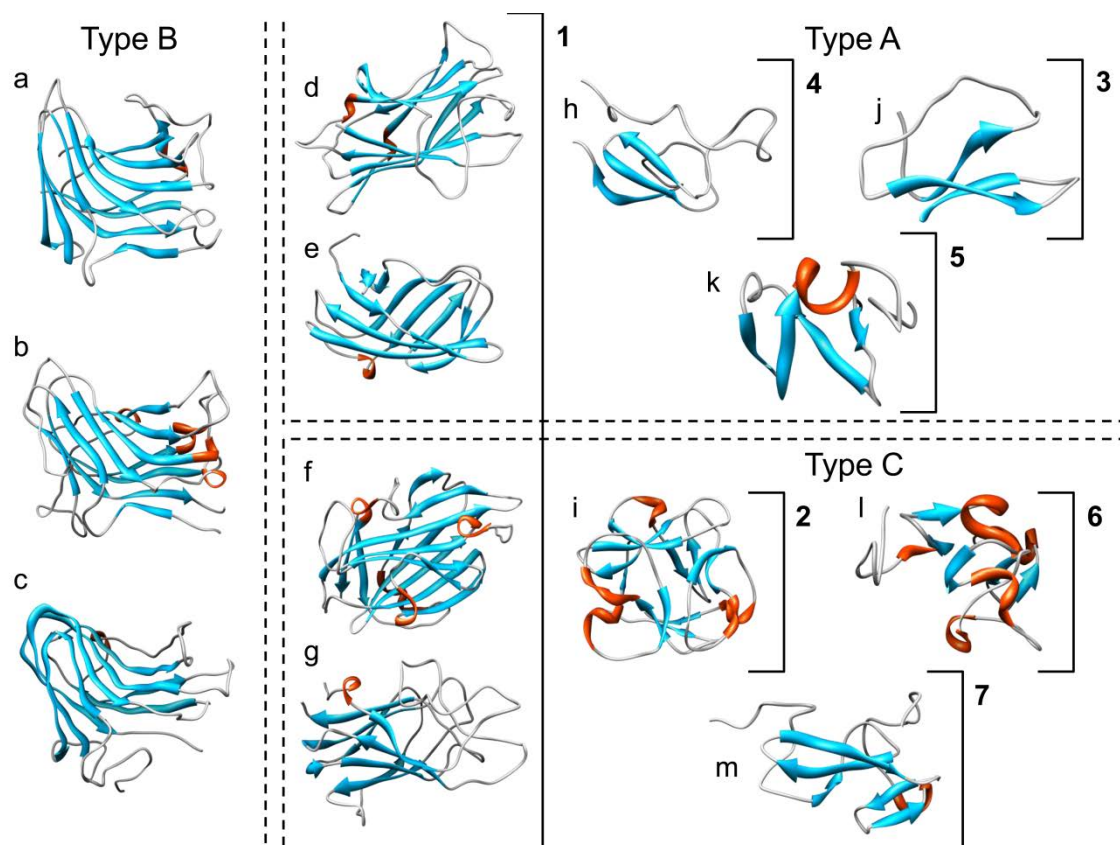


Figure I.8: Classification of CBMs.

Dotted boxes surround examples of CBMs belonging to the functional Types A, B, and C. Brackets with numbers indicate examples of CBMs belonging to fold families 1–7 (see the sections below and tables I.3 and I.4). CBMs shown are as follows: (a) family 11 CBM, CtCBM11, from *Clostridium thermocellum* (PDB code 1v0a – see Chapter II)⁵⁷; (b) family 30 CBM, CtCBM30, from *Clostridium thermocellum* (not deposited – see Chapter IV); (c) family 44 CBM, CtCBM44, from *Clostridium thermocellum* (PDB code 2c4x – see Chapter III)⁵⁸; (d) family 3 CBM, CtCBM3, from *Clostridium thermocellum* (PDB code 1nbc)⁷⁵; (e) family 2 CBM, CfCBM2, from *Cellulomonas fimi* (PDB code 1exg)⁷⁶; (f) family 9 CBM, TmCBM9-2, from *Thermotoga maritima* (PDB code 1l82)⁷⁷; (g) family 32 CBM, MvCBM32, from *Micromonospora viridifaciens* (PDB code 1euu)⁷⁸; (h) family 5 CBM, EcCBM5, from *Erwinia chrysanthemi* (PDB code 1aiw)⁷⁹; (i) family 13 CBM, SlCBM13, from *Streptomyces lividans* (PDB code 1mc9)⁸⁰; (j) family 1 CBM, TrCBM1, from *Trichoderma reesei* (PDB code 1cbh)⁸¹; (k) family 10 CBM, CjCBM10, from *Cellvibrio japonicus* (PDB code 1e8r)⁸²; (l) family 18 CBM, UdCBM18, from *Urtica dioica* (PDB code 1en2)⁸³; (m) family 14 CBM, TrCBM14, from *Tachypleus tridentatus* (PDB code 1dq)⁸⁴. Bound ligands or metal ions are not shown. Adapted from Boraston *et al.*, 2004¹⁵

I.6.1 Nomenclature of CBMs

When they were first described, carbohydrate-binding modules were designated as cellulose-binding domains, CBDs, due to their ability to bind cellulose.^{66,67} This terminology lasted until 1999, at which point, due to the finding of non-catalytic modules that bound to carbohydrates other than cellulose, the name was changed to carbohydrate-binding modules (CBMs).^{15,85} The conventions for the naming of CBMs were adopted by following the nomenclature system of the glycosyl hydrolases.¹⁵ Therefore, CBMs are divided into families according to the primary

sequence similarities. So far, CBMs have been grouped into 64 families (<http://www.cazy.org>).⁶⁹ In this way a given CBM, let's say for instance belonging to family 11, will be denominated CBM11. Furthermore, the name can also include the organism from which the CBM originates. So, CBM11 from *Clostridium thermocellum* can be named as CtCBM11. If the enzymes contain more than one CBM from the same family, a number, corresponding to the position of the CBM in the enzyme with respect to the *N*-terminus is included. This simple nomenclature eliminates the need to memorize arbitrary names and, because it is complementary to the naming system of glycosyl hydrolases, it keeps these two fields linked.¹⁵

Another way of classifying CBMs is based on the fold similarities between the different families (as an analogy to the catalytic modules' superfamilies).^{15,38,59} By grouping the several CBM families according to their fold similarities it was possible to identify seven fold superfamilies (**Table I.2**): β -sandwich, β -trefoil, cysteine knot, unique, OB fold, hevein fold and hevein-like fold.¹⁵ By far, the dominant fold among CBMs is the β -sandwich (fold family 1). CBMs belonging to this family fold as a β -jelly roll with two β -sheets, each consisting of three to six antiparallel β -strands.¹⁵ In most cases β -sandwich CBMs have bounded metal ions (usually calcium) which have a structural role. With the exception of CBMs 6⁸⁶ and 32⁷⁸, the binding site in these CBMs is localized in the concave side of the β -barrel. The β -trefoil fold family (fold family 2) is generally associated with ricin toxin β -chain.¹⁵ CBMs belonging to this fold contain twelve β -sheet strands that form six hairpin turns. Six of the β -strands form a β -barrel structure attendant with three hairpin turns. The other three hairpins form a triangular cap on one end of the β -barrel denominated "hairpin triplet".¹⁵ As a consequence of this fold, the molecule has a pseudo-3-fold axis.⁸⁷ The three functional binding sites are an advantage as they lead to significantly enhanced affinities.^{59,87} CBMs from fold families 3 to 5 are small amino acid polypeptides (30-60 amino acids) that contain only β -sheet and coil (**Figure I.8**). They appear to be specialized in binding cellulose and/or chitin. The majority of these CBMs have planar surfaces, complementary to the surface of the crystalline polysaccharides. Fold families 6 and 7 contain small CBMs with approximately 40 amino acids, originally identified in plants as chitin-binding proteins. This fold is dominated by coil with two small β -sheets and a α -helix. The minimal hevein fold is found in family 18 CBMs and is classified as fold family 6. The family 14 CBMs also incorporates a hevein fold but it's fused with a small β -sheet structure which justifies its inclusion onto a different fold family – fold family 7.¹⁵

Table I.2: Classification of CBM fold families.^{15,38,59}

<i>Fold family</i>	<i>Fold</i>	<i>CBM families</i>
1	β -Sandwich	2, 3, 4, 6, 9, 11, 15, 16, 17, 20, 21, 22, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 39, 40, 41, 42, 44, 47, 48, 51, 57, 61
2	β -Trefoil	13, 42
3	Cysteine knot	1
4	Unique	5, 12
5	OB fold	10
6	Hevein fold	18
7	Unique: contains hevein-like fold	14

Despite the advantages of the previous classification systems, they do not give any idea about the function of CBMs. Therefore, based on structural and functional similarities of CBMs, three types have been proposed (**Table I.3**).¹⁵

Table I.3: Classification of CBM types^{15,38,59}

<i>Type</i>	<i>Fold family</i>	<i>CBM family</i>
A	1, 3, 4, 5	1, 2a, 3, 5, 10
B	1	2b, 4, 6, 11, 15, 16, 17, 20, 21, 22, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 39, 41, 44, 47
C	1, 2, 4, 6, 7	9, 12, 13, 14, 18, 32, 40, 42, 43, 50

Type A CBMs, or “surface-binding”, present a flat exposed binding surface, complementary to the planar surface of the crystalline polysaccharides. In contrast, the “glycan-chain-binding” Type B CBMs show a recessed binding cleft, usually described as groove or cleft that binds to soluble polysaccharide chains. Finally, Type C, or “small sugar-binding” CBMs display lectin-like binding to mono-, di-, or tri-saccharides and lack the extended binding cleft found in Type B CBMs. Within these three CBM types are seven structural fold families (**Table I.3**) which cover the 64 CBM families known to date. Further details on the three types of CBMs will be given below (see Sections I.6.1.1 to I.6.1.3).

I.6.1.1 Type A CBMs – surface-binding

Type A CBMs range in size from 35 to 140 amino acids and include CBMs from families 1, 2a, 3, 5, and 10 (**Table I.3**). They bind to insoluble, highly crystalline cellulose and/or chitin and show minor affinity for soluble carbohydrates.^{75,88,89} It has been shown that these type of CBMs bind to the hydrophobic 110 face of crystalline cellulose.⁹⁰ The interaction of type A modules with crystalline cellulose is associated with positive entropy, which is relatively unique among carbohydrate-binding proteins.⁹¹ It has been proposed that the water molecules released from the protein and ligand when CBMs bind to their target carbohydrates increases the entropy of the system. In the case of soluble saccharides it is postulated to be more than counterbalanced by the conformational restriction of the bound ligand leading to a net reduction in entropy.^{15,91} However, the molecular basis for the thermodynamic forces that drive protein–carbohydrate interactions remains a highly hot area, particularly with respect to the role of water molecules and the loss of entropy through conformational restriction. Structurally, all Type A CBMs have a flat platform of aromatic residues (tryptophan, tyrosine, and occasionally histidine and phenylalanine) aligned along one face of the globular polypeptide that is thought to be complementary to the flat surfaces presented by cellulose or chitin crystals (**Figure I.8**).^{75,88,92} These aromatic residues are often involved in the binding of the type A CBMs to cellulose.^{75,88,93,94}

I.6.1.2 Type B CBMs – glycan-chain-binding

Type B CBMs are usually described as glycan-chain-binders as their binding affinity depends on the degree of polymerization of the carbohydrate chain – they show increased affinity for ligands up to six moieties (hexasaccharides) and little or no affinity for ligands with three or less.¹⁵ They bind to a large variety of substrates, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose. The substrate binding sites of Type B CBMs are described as grooves and can vary from very shallow to being able to accommodate the entire pyranose ring (**Figure I.8**). As with Type A CBMs, aromatic residues (tryptophan, tyrosine and, less commonly, phenylalanine) play a pivotal role in ligand binding and recognition, and the orientation of these amino acids is a key determinant of specificity.^{59,74,95} Although, as in Type A CBMs, the carbohydrate moieties are recognized by aromatic residues, in Type B CBMs, the side chains of these residues can form planar, twisted or sandwich platforms for substrate binding.^{57,96} Unlike Type A CBMs, direct hydrogen bonds are also fundamental in defining the affinity and ligand specificity of Type B glycan chain binders.^{15,97} These stacking/hydrophobic interactions

between the sugar rings and the aromatic residues along with the conformational fitting of the glycan chains play a fundamental role in ligand recognition. The thermodynamics of the interaction of this type of CBMs is invariably enthalpically driven with an unfavorable entropic contribution. The role of water-mediated hydrogen bonds to the binding of Type B CBMs to their target ligands is still very controversial⁹⁷ with very few examples of its importance (*see Chapter III*). Structurally all Type B CBMs known to date belong to the β -sandwich fold family (fold family 1 - **Table I.3**). CBMs from families 11, 44 and 30 from *Clostridium thermocellum* will be discussed in more detail in chapters II, III and IV.

I.6.1.3 Type C CBMs – small sugar-binding

Type C CBMs demonstrate lectin-like binding properties, having high affinity to simple sugars, soluble or insoluble (mono-, di- or trisaccharides).¹⁵ Therefore the epithet: “small-sugar-binding”. These binding modules come from a variety of sources, including animals, plants, crustaceans and microbes. They differ from Type B by lacking the characteristic extended binding clefts, although distinguishing between the two types can be very difficult.^{98,99} However, in a good agreement with their lectin-like properties, the protein-ligand hydrogen bond network is more extensive in Type C than in Type B CBMs.¹⁵

I.6.2 Molecular determinants of binding

Data obtained from all the determined CBM structures indicate that different families are structurally similar and that their carbohydrate binding capacity can be attributed in great extent to several aromatic amino acids (tryptophan, tyrosine, and occasionally histidine and phenylalanine) that constitute the hydrophobic surface (**Figure I.9**)^{75,88,90,92}. These amino acids are often involved in stacking/hydrophobic interactions between the sugar rings and aromatic residues conferring specificity and stability to the protein-carbohydrate complex.⁹⁷ The relative importance of direct hydrogen bonds depends on the CBM Type. In Type A CBMs, it was shown that mutation to alanine of residues involved in direct hydrogen bonds has little effect on affinity, suggesting that, in these proteins, hydrogen bonds play only a minor role in ligand recognition.⁹² In Type B and Type C CBMs, replacement of direct hydrogen-bonding residues with alanine can lead to significant losses in affinity to complete abolition of binding.⁵⁷ However, it must be noted that in some of these cases, it is uncertain if the loss in affinity is exclusively due to the loss of the hydrogen bond or if subtle structural changes in the binding sites are the responsible for the decrease or loss of ligand affinity.

Furthermore, as seen above, the topology of the binding site also displays a key role of binding specificity. For instance, CBMs with the β -sandwich fold, the positioning of the aromatic residues and the loop arrangement shape the binding sites in order to accommodate the substrate.¹⁵ The aromatic amino acid side chains pack onto the sugar rings forming a sandwich like platform.⁵⁷ Moreover, the binding sites of Type B CBMs can adopt other conformations according to their specificity. In CBMs of families 2b, 15, 17, 27, 29, 34 and 36, the binding sites can be twisted due to the rotation of the planes of two to three aromatic amino acid side chains relative to one another.¹⁵ On their own, these two types of platforms are able to confer specificity to the CBM-carbohydrate recognition as different sugars may have a rather linear shape (for instance cellulose) or a more curved shape (for instance xylan). CBMs seem to adopt conformations that mirror the substrate conformations in solution, therefore minimizing the energy of binding.^{15,95}

On the other hand, the flat platform, distinctive of Type A CBMs (**Figure I.9**), specifically recognize the flat surfaces presented by the crystalline substrates. Tyrosines and tryptophans are often separated by a distance corresponding to the length of the repeating unit (10.3 Å is the length of one cellobiose unit) and the aromatic ring interacts with the pyranose rings of the polysaccharides.⁸⁹ This interaction may be supplemented by few hydrogen bonds mediated by polar residues located at the binding interface.¹⁵

Another possible factor for ligand recognition and binding is calcium. It is well established that calcium plays a major role in CBM stability¹⁰⁰ but only recently its influence on CBM-carbohydrate interaction has been demonstrated.^{101,102} However there are only a few examples of this type of behavior, so it does not seem to be a rule regarding carbohydrate recognition.

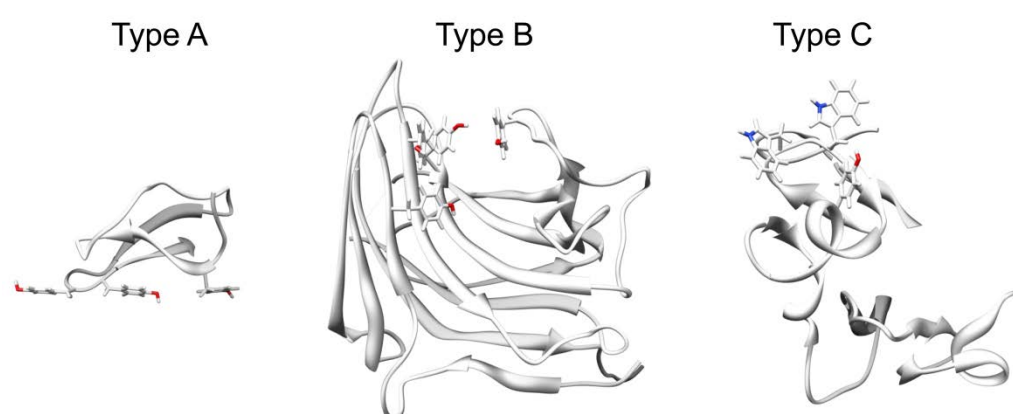


Figure I.9: The binding-site platforms of the three types of CBMs.

The Type A CBM (*TrCBM1* – PDB code: 1cbh)⁸¹ shows a flat platform complementary to the flat surfaces presented by the crystalline substrates; The Type B CBM (*CtCBM11* – PDB code: 1v0a – see Chapter II)⁵⁷ presents a sandwich platform, or cleft, appropriated for binding soluble single glycan chains from four to six units; The Type C CBM (*UdCBM18* – PDB code: 1en2)⁸³ shows a small platform able to bind only to mono-, di- or trisaccharides.

I.6.3 Utilization of CBMs

Carbohydrate recognition is an essential step of many biological and biotechnological processes and CBMs, due to their properties, are becoming the perfect candidates for many applications. The basic properties that make CBMs such good candidates are mainly three:⁶⁸

- They are independent units that can function by their own in chimeric proteins;
- The substrates are abundant and inexpensive and have excellent chemical and physical properties;
- The binding specificities can be controlled, and therefore the right solution can be adapted to an existing problem.

Given that the large-scale recovery and purification of biologically active molecules continues to be a limiting step for many biotechnological purposes, the main application of CBMs is, probably, **bioprocessing**. CBMs have been used as low-cost, high-capacity purification tags for the isolation of biologically active target peptides (**Figure I.10**). Cellulose is a very economical support-matrix for the industry when compared with other immobilization systems,⁶⁸ while CBM tags allow the development of secure and quick purification protocols.

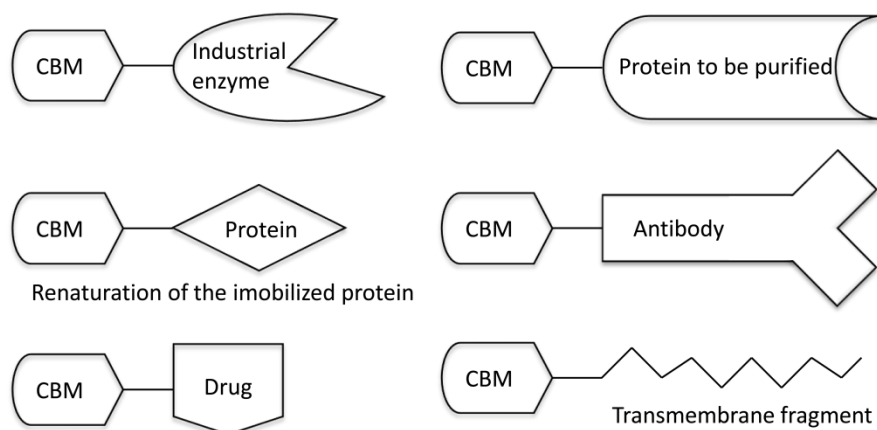


Figure I.10: Applications of hybrid CBMs (adapted from Volkov *et al*, 2004¹⁰³).

The main direction of biotechnological research is immobilization of hybrid proteins, composed of commercially important enzymes and CBMs, on cellulose. Immobilized enzymes can be used, for instance, for continuous hydrolysis in flow reactors.¹⁰³ Furthermore, as CBMs can be attached to proteins without altering their biological activity¹⁰⁴ they can be used for improving enzyme activity or in high-level expression vectors for the production of CBM-fused proteins.^{68,103,105} Production of recombinant proteins in plants has been recently accepted as one

of the most cost-effective production systems and CBMs have been used with success in the production of chimeric proteins. In this system the plant produces both the target protein and its purification matrix (cellulose).⁶⁸ Hybrid CBMs can also be applied to immunochemistry for the purification or detection of interesting chemical compounds using antibodies (**Figure I.10**). A CBM-antibody chimera immobilized in cellulose could be used for efficient purification of target compounds.¹⁰³ Another interesting application of this hybrid CBMs is for renaturation of proteins (**Figure I.10**). Matrix-assisted refolding of recombinant proteins aims to prevent the aggregation of protein during the course of renaturation and, so far, only histidine and arginine tags have fitted this purpose as they stay bound to the matrix under denaturing conditions. CtCBM3 has been used successfully as the attachment support for matrix-assisted refolding of a single-chain antibody expressed in *E. coli*.¹⁰⁶ CtCBM3 can bind cellulose in the presence of 6 M urea and provide a threefold increase in protein yield compared with standard refolding procedures.

Another area of high interest is **biofuel production** from biomass. As referred above, efficient hydrolysis of cellulose is very difficult due to the complex composition of the plant cell wall. Because of the high variety of binding specificities that CBMs have, they can be used to construct high affinity CBM-cellulase chimeras fitted for the proficient breakdown of the cellulosic biomass to sugars, which can then be converted to liquid fuel, namely bioethanol.²

The textile industry has also been exploring the CBM technology, mainly for the recycling of several products or for changing the properties of specific fabrics. Because most of textiles have cellulose as a major component, CBMs can be used for **targeting** specific components. For instance, CBMs can be linked to enzymes in laundry powders, increasing the affinity for the cellulose substrate and improving the enzyme performance.⁶⁸ Additional substances, such as fragrance-bearing particles, can also be linked to CBMs and added to laundry-powder, decreasing the amount needed in the product.⁶⁸

CBMs can also be applied as tools for **research and diagnosis**. For instance, conjugation of a CBM with a bacterium-binding protein can be used for detecting pathogenic microbes in food samples.⁶⁸

The examples presented above are only a small sample of all the applications found for CBMs until now. The utilization of CBMs in different field of biotechnology is perfectly established and the tendency is for further applications to emerge. Due to their properties (**Figure I.10**) CBMs are the perfect candidates for solving an enormous variety of problems and will certainly occupy an important place in the inventory of biotechnological tools. The potential for these molecules for improving life in many aspects cannot be overstated.

I.7 Objectives and outline of the thesis

The work presented in this thesis aims to understand the molecular interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their substrates. The CBMs under study belong to families 11, 30 and 44 from *C. thermocellum*. The crystal and NMR solution structures of *Ct*CBM11 will be addressed in **Chapter II**. The molecular determinants of ligand specificity of *Ct*CBM11 will be discussed in **Chapter III** while the ones from *Ct*CBM30 and *Ct*CBM44 will be discussed in **Chapter IV**. Although structurally similar, these modules have distinct specificities in terms of ligand recognition. Using NMR spectroscopy, X-ray crystallography and computational studies, supported by techniques of molecular biology, I aimed to identify the structural features of both ligand and protein that determine the selective recognition and binding. The knowledge gained about the molecular interactions that define the specificity of these modules is fundamental for future work involving the deployment of nano-molecular machines, capable of efficiently degrading the cell wall. Thus, this work will be an important contribution to the implementation of sustainable processes with potential impact on several aspects.

On the other hand, the assembly of the enzymatic components into the cellulosome complex and the attachment of the last to the bacterial cell wall are also of great significance for the overall process of plant cell wall degradation. In order to understand this mechanism, the elucidation of the molecular determinants responsible for recognition is fundamental. In this sense I have used X-ray crystallography to determine the crystal structures of two type II complexes from *C. thermocellum* (**Chapter V**) and *B. cellulosolvens* (**Chapter VI**) and gain some insights into the structural characteristics that define the cohesin-dockerin interaction.

In **Chapter VII** and **Chapter VIII** I will discuss the theory and methods from the NMR and X-ray crystallography techniques, respectively, used to describe the structural characteristics observed in the previous chapters.

The results obtained represent a significant improvement in the understanding of the factors that determine the specificity and the mode of action of Type B CBMs, namely *Ct*CBM11, *Ct*CBM30 and *Ct*CBM44, at the molecular level. Moreover, structures of the two Type II cohesin–dockerin complexes provide valuable information about the atomic interactions that mediate complex assembly. Altogether the work presented represents an important contribution to the understanding of this phenomenal mega-Dalton machine termed cellulosome.

Finally, I will make an overall discussion on the results obtained and draw some future perspectives from this work.

I.8 References

1. Buckeridge, M. S.; Goldman, G. H., *Routes to Cellulosic Ethanol*. Springer: New York, 2011.
2. Demain, A. L.; Newcomb, M.; Wu, J. H., Cellulase, clostridia, and ethanol. *Microbiology and molecular biology reviews : MMBR* **2005**, *69* (1), 124.
3. O'Sullivan, A., Cellulose: the structure slowly unravels. *Cellulose* **1997**, *4* (3), 173.
4. Jorgensen, H.; Kristensen, J. B.; Felby, C., Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuel Bioprod Bior* **2007**, *1* (2), 119.
5. Sun, Y.; Cheng, J. Y., Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource Technol* **2002**, *83* (1), 1.
6. Harmsen, P.; Huijgen, W.; Bermudez, L.; Bakker, R., Literature review of physical and chemical pretreatment processes for lignocellulosic biomass. *Wageningen UR Food & Biobased Research* **2010**.
7. Bayer, E. A.; Lamed, R.; Himmel, M. E., The potential of cellulases and cellulosomes for cellulosic waste management. *Curr Opin Biotech* **2007**, *18* (3), 237.
8. Sanchez, O. J.; Cardona, C. A., Trends in biotechnological production of fuel ethanol from different feedstocks. *Bioresource Technol* **2008**, *99* (13), 5270.
9. Schwarz, W. H., The cellulosome and cellulose degradation by anaerobic bacteria. *Appl Microbiol Biot* **2001**, *56* (5-6), 634.
10. Bayer, E. A.; Kenig, R.; Lamed, R., Adherence of Clostridium-Thermocellum to Cellulose. *J Bacteriol* **1983**, *156* (2), 818.
11. Lamed, R.; Setter, E.; Bayer, E. A., Characterization of a Cellulose-Binding, Cellulase-Containing Complex in Clostridium-Thermocellum. *J Bacteriol* **1983**, *156* (2), 828.
12. Freier, D.; Mothershed, C. P.; Wiegel, J., Characterization of Clostridium-Thermocellum Jw20. *Appl Environ Microb* **1988**, *54* (1), 204.
13. Lamed, R.; Setter, E.; Kenig, R.; Bayer, E. A., The Cellulosome - a Discrete Cell-Surface Organelle of Clostridium-Thermocellum Which Exhibits Separate Antigenic, Cellulose-Binding and Various Cellulolytic Activities. *Biotechnol Bioeng* **1984**, 163.
14. Fontes, C. M. G. A.; Gilbert, H. J., Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry*, Vol 79 **2010**, *79*, 655.
15. Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **2004**, *382* (Pt 3), 769.
16. Uversky, V. N.; Kataeva, I. A., *Cellulosome*. Nova Science Publishers: New York, 2006; p xiii.
17. Adams, J. J.; Currie, M. A.; Ali, S.; Bayer, E. A.; Jia, Z. C.; Smith, S. P., Insights into Higher-Order Organization of the Cellulosome Revealed by a Dissect-and-Build Approach: Crystal Structure of Interacting Clostridium thermocellum Multimodular Components. *Journal of Molecular Biology* **2010**, *396* (4), 833.
18. Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A., Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *P Natl Acad Sci USA* **2003**, *100* (24), 13809.
19. Carvalho, A. L.; Pires, V. M. R.; Gloster, T. M.; Turkenburg, J. P.; Prates, J. A. M.; Ferreira, L. M. A.; Romao, M. J.; Davies, G. J.; Fontes, C. M. G. A.; Gilbert, H. J., Insights into the structural determinants of cohesin dockerin specificity revealed by the crystal structure of the type II cohesin from Clostridium thermocellum SdbA. *Journal of Molecular Biology* **2005**, *349* (5), 909.
20. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R., The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* **2004**, *58*, 521.
21. Lodish, H. F., *Molecular cell biology*. 4th ed.; W.H. Freeman: New York, 2000; p 973.

22. Kamide, K., *Cellulose and cellulose derivatives : molecular characterization and its applications*. 1st ed.; Boston : Elsevier: Amsterdam, 2005; p 630.
23. Hayashi, T.; Kaida, R., Hemicelluloses as Recalcitrant Components for Saccharification in Wood. In *Routes to Cellulosic Ethanol*, Buckeridge, M. S.; Goldman, G. H., Eds. Springer New York: 2011; pp 45.
24. Fry, S. C.; York, W. S.; Albersheim, P.; Darvill, A.; Hayashi, T.; Joseleau, J. P.; Kato, Y.; Lorences, E. P.; Maclachlan, G. A.; Mcneil, M.; Mort, A. J.; Reid, J. S. G.; Seitz, H. U.; Selvendran, R. R.; Voragen, A. G. J.; White, A. R., An Unambiguous Nomenclature for Xyloglucan-Derived Oligosaccharides. *Physiol Plantarum* **1993**, 89 (1), 1.
25. Mosier, N.; Wyman, C.; Dale, B.; Elander, R.; Lee, Y. Y.; Holtzapple, M.; Ladisch, M., Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresource Technol* **2005**, 96 (6), 673.
26. Yang, B.; Wyman, C. E., Pretreatment: the key to unlocking low-cost cellulosic ethanol. *Biofuel Bioprod Bior* **2008**, 2 (1), 26.
27. Bayer, E. A.; Morag, E.; Lamed, R., The Cellulosome - a Treasure-Trove for Biotechnology. *Trends Biotechnol* **1994**, 12 (9), 379.
28. Doi, R. H.; Kosugi, A., Cellulosomes: Plant-cell-wall-degrading enzyme complexes. *Nat Rev Microbiol* **2004**, 2 (7), 541.
29. Vodovnik, M.; Logar, R. M., Cellulosomes - Promising Supramolecular Machines of Anaerobic Cellulolytic Microorganisms. *Acta Chim Slov* **2010**, 57 (4), 767.
30. Gilbert, H. J.; Stalbrand, H.; Bruner, H., How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation. *Curr Opin Plant Biol* **2008**, 11 (3), 338.
31. Morag, E.; Bayer, E. A.; Lamed, R., Relationship of Cellulosomal and Noncellulosomal Xylanases of Clostridium-Thermocellum to Cellulose-Degrading Enzymes. *J Bacteriol* **1990**, 172 (10), 6098.
32. Bayer, E. A.; Lamed, R., Ultrastructure of the Cell-Surface Cellulosome of Clostridium-Thermocellum and Its Interaction with Cellulose. *J Bacteriol* **1986**, 167 (3), 828.
33. Hammel, M.; Fierober, H. P.; Czjzek, M.; Kurkal, V.; Smith, J. C.; Bayer, E. A.; Finet, S.; Receveur-Brechot, V., Structural basis of cellulosome efficiency explored by small angle X-ray scattering. *Journal of Biological Chemistry* **2005**, 280 (46), 38562.
34. Alber, O.; Noach, I.; Rincon, M. T.; Flint, H. J.; Shimon, L. J. W.; Lamed, R.; Frolow, F.; Bayer, E. A., Cohesin diversity revealed by the crystal structure of the anchoring cohesin from Ruminococcus flavefaciens. *Proteins-Structure Function and Bioinformatics* **2009**, 77 (3), 699.
35. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Heckman, M. P.; Wu, J. H. D., Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. *Journal of Molecular Biology* **2001**, 307 (3), 745.
36. Chauvaux, S.; Beguin, P.; Aubert, J. P.; Bhat, K. M.; Gow, L. A.; Wood, T. M.; Bairoch, A., Calcium-Binding Affinity and Calcium-Enhanced Activity of Clostridium-Thermocellum Endoglucanase-D. *Biochemical Journal* **1990**, 265 (1), 261.
37. Gilbert, H. J., Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol Microbiol* **2007**, 63 (6), 1568.
38. Guillen, D.; Sanchez, S.; Rodriguez-Sanoja, R., Carbohydrate-binding domains: multiplicity of biological roles. *Appl Microbiol Biot* **2010**, 85 (5), 1241.
39. Ding, S. Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R., A novel cellulosomal scaffoldin from Acetivibrio cellulolyticus that contains a family 9 glycosyl hydrolase. *J Bacteriol* **1999**, 181 (21), 6720.
40. Ding, S. Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R., A scaffoldin of the Bacteroides cellulosolvens cellulosome that contains 11 type II cohesins. *J Bacteriol* **2000**, 182 (17), 4915.
41. Xu, Q.; Bayer, E. A.; Goldman, M.; Kenig, R.; Shoham, Y.; Lamed, R., Architecture of the Bacteroides cellulosolvens cellulosome: Description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *J Bacteriol* **2004**, 186 (4), 968.

42. Nolling, J.; Breton, G.; Omelchenko, M. V.; Makarova, K. S.; Zeng, Q. D.; Gibson, R.; Lee, H. M.; Dubois, J.; Qiu, D. Y.; Hitti, J.; Wolf, Y. I.; Tatusov, R. L.; Sabathe, F.; Doucette-Stamm, L.; Soucaille, P.; Daly, M. J.; Bennett, G. N.; Koonin, E. V.; Smith, D. R.; Finishing, G. S. C. P., Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J Bacteriol* **2001**, *183* (16), 4823.
43. Pages, S.; Belaich, A.; Fierobe, H. P.; Tardif, C.; Gaudin, C.; Belaich, J. P., Sequence analysis of scaffolding protein CipC and ORFXp, a new cohesin-containing protein in *Clostridium cellulolyticum*: Comparison of various cohesin domains and subcellular localization of ORFXp. *J Bacteriol* **1999**, *181* (6), 1801.
44. Shoseyov, O.; Takagi, M.; Goldstein, M. A.; Doi, R. H., Primary Sequence-Analysis of *Clostridium-Cellulovorans* Cellulose Binding Protein-A. *P Natl Acad Sci USA* **1992**, *89* (8), 3483.
45. Kakiuchi, M.; Isui, A.; Suzuki, K.; Fujino, T.; Fujino, E.; Kimura, T.; Karita, S.; Saki, K.; Ohmiya, K., Cloning and DNA sequencing of the genes encoding *Clostridium josui* scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. *J Bacteriol* **1998**, *180* (16), 4303.
46. Pohlschroder, M.; Canaleparola, E.; Leschine, S. B., Ultrastructural Diversity of the Cellulase Complexes of *Clostridium Papyrosolvans* C7. *J Bacteriol* **1995**, *177* (22), 6625.
47. Lamed, R.; Naimark, J.; Morgenstern, E.; Bayer, E. A., Specialized Cell-Surface Structures in Cellulolytic Bacteria. *J Bacteriol* **1987**, *169* (8), 3792.
48. Ding, S. Y.; Rincon, M. T.; Lamed, R.; Martin, J. C.; McCrae, S. I.; Aurilia, V.; Shoham, Y.; Bayer, E. A.; Flint, H. J., Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. *J Bacteriol* **2001**, *183* (6), 1945.
49. Xu, J. C.; Crowley, M. F.; Smith, J. C., Building a foundation for structure-based cellulosome design for cellulosic ethanol: Insight into cohesin-dockerin complexation from computer simulation. *Protein Sci* **2009**, *18* (5), 949.
50. Bayer, E. A.; Lamed, R.; White, B. A.; Flint, H. J., From Cellulosomes to Cellulosomics. *Chem Rec* **2008**, *8* (6), 364.
51. Heyman, A.; Barak, Y.; Caspi, J.; Wilson, D. B.; Altmana, A.; Bayer, E. A.; Shoseyov, O., Multiple display of catalytic modules on a protein scaffold: Nano-fabrication of enzyme particles. *J Biotechnol* **2007**, *131* (4), 433.
52. Fujino, T.; Beguin, P.; Aubert, J. P., Organization of a *Clostridium-Thermocellum* Gene-Cluster Encoding the Cellulosomal Scaffolding Protein Cipa and a Protein Possibly Involved in Attachment of the Cellulosome to the Cell-Surface. *J Bacteriol* **1993**, *175* (7), 1891.
53. Poole, D. M.; Morag, E.; Lamed, R.; Bayer, E. A.; Hazlewood, G. P.; Gilbert, H. J., Identification of the Cellulose-Binding Domain of the Cellulosome Subunit-S1 from *Clostridium-Thermocellum* Ys. *FEMS Microbiology Letters* **1992**, *99* (2-3), 181.
54. Morag, E.; Lapidot, A.; Govorko, D.; Lamed, R.; Wilchek, M.; Bayer, E. A.; Shoham, Y., Expression, Purification, and Characterization of the Cellulose-Binding Domain of the Scaffoldin Subunit from the Cellulosome of *Clostridium-Thermocellum*. *Appl Environ Microb* **1995**, *61* (5), 1980.
55. Leibovitz, E.; Beguin, P., A new type of cohesin domain that specifically binds the dockerin domain of the *Clostridium thermocellum* cellulosome-integrating protein CipA. *J Bacteriol* **1996**, *178* (17), 5335.
56. Adams, J. J.; Pal, G.; Jia, Z. C.; Smith, S. P., Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. *P Natl Acad Sci USA* **2006**, *103* (2), 305.
57. Carvalho, A. L.; Goyal, A.; Prates, J. A. M.; Bolam, D. N.; Gilbert, H. J.; Pires, V. M. R.; Ferreira, L. M. A.; Planas, A.; Romao, M. J.; Fontes, C. M. G. A., The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *Journal of Biological Chemistry* **2004**, *279* (33), 34785.
58. Najmudin, S.; Guerreiro, C. I. P. D.; Carvalho, A. L.; Prates, J. A. M.; Correia, M. A. S.; Alves, V. D.; Ferreira, L. M. A.; Romao, M. J.; Gilbert, H. J.; Bolam, D. N.; Fontes, C.

- M. G. A., Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *Journal of Biological Chemistry* **2006**, 281 (13), 8815.
59. Hashimoto, H., Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **2006**, 63 (24), 2954.
 60. Miras, I.; Schaeffer, F.; Beguin, P.; Alzari, P. M., Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. *Biochemistry* **2002**, 41 (7), 2115.
 61. Pages, S.; Belaich, A.; Belaich, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A., Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins* **1997**, 29 (4), 517.
 62. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A.; Gilbert, H. J., Evidence for a dual binding mode of dockerin modules to cohesins. *P Natl Acad Sci USA* **2007**, 104 (9), 3089.
 63. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Wu, J. H. D., Secondary structure and calcium-induced folding of the *Clostridium thermocellum* dockerin domain determined by NMR spectroscopy. *Arch Biochem Biophys* **2000**, 379 (2), 237.
 64. Mechaly, A.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Lamed, R.; Shoham, Y.; Bayer, E. A., Cohesin-dockerin interaction in cellulosome assembly - A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *Journal of Biological Chemistry* **2001**, 276 (13), 9883.
 65. Haimovitz, R.; Barak, Y.; Morag, E.; Voronov-Goldman, M.; Shoham, Y.; Lamed, R.; Bayer, E. A., Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. *Proteomics* **2008**, 8 (5), 968.
 66. Tomme, P.; Vantilbeurgh, H.; Pettersson, G.; Vandamme, J.; Vandekerckhove, J.; Knowles, J.; Teeri, T.; Claeysens, M., Studies of the Cellulolytic System of *Trichoderma reesei* Qm-9414 - Analysis of Domain Function in 2 Cellobiohydrolases by Limited Proteolysis. *Eur J Biochem* **1988**, 170 (3), 575.
 67. Gilkes, N. R.; Warren, R. A. J.; Miller, R. C.; Kilburn, D. G., Precise Excision of the Cellulose Binding Domains from 2 *Cellulomonas fimi* Cellulases by a Homologous Protease and the Effect on Catalysis. *Journal of Biological Chemistry* **1988**, 263 (21), 10401.
 68. Shoseyov, O.; Shani, Z.; Levy, I., Carbohydrate binding modules: Biochemical properties and novel applications. *Microbiology and Molecular Biology Reviews* **2006**, 70 (2), 283.
 69. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **2009**, 37, D233.
 70. Bolam, D. N.; Ciruela, A.; McQueen-Mason, S.; Simpson, P.; Williamson, M. P.; Rixon, J. E.; Boraston, A.; Hazlewood, G. P.; Gilbert, H. J., *Pseudomonas* cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochemical Journal* **1998**, 331, 775.
 71. Limon, M. C.; Margolles-Clark, E.; Benitez, T.; Penttila, M., Addition of substrate-binding domains increases substrate-binding capacity and specific activity of a chitinase from *Trichoderma harzianum*. *FEMS Microbiology Letters* **2001**, 198 (1), 57.
 72. Gao, P. J.; Wang, L. S.; Zhang, Y. Z., A novel function for the cellulose binding module of cellobiohydrolase I. *Sci China Ser C* **2008**, 51 (7), 620.
 73. McLean, B. W.; Boraston, A. B.; Brouwer, D.; Sanaie, N.; Fyfe, C. A.; Warren, R. A. J.; Kilburn, D. G.; Haynes, C. A., Carbohydrate-binding modules recognize fine substructures of cellulose. *Journal of Biological Chemistry* **2002**, 277 (52), 50245.
 74. Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008**, 275 (10), 2524.

75. Tormo, J.; Lamed, R.; Chirino, A. J.; Morag, E.; Bayer, E. A.; Shoham, Y.; Steitz, T. A., Crystal structure of a bacterial family-III cellulose-binding domain: A general mechanism for attachment to cellulose. *Embo J* **1996**, *15* (21), 5739.
76. Xu, G. Y.; Ong, E.; Gilkes, N. R.; Kilburn, D. G.; Muhandiram, D. R.; Harris-Brandts, M.; Carver, J. P.; Kay, L. E.; Harvey, T. S., Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. *Biochemistry* **1995**, *34* (21), 6993.
77. Notenboom, V.; Boraston, A. B.; Kilburn, D. G.; Rose, D. R., Crystal structures of the family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A in native and ligand-bound forms. *Biochemistry* **2001**, *40* (21), 6248.
78. Gaskell, A.; Crennell, S.; Taylor, G., The 3 Domains of a Bacterial Sialidase - a Beta-Propeller, an Immunoglobulin Module and a Galactose-Binding Jelly-Roll. *Structure* **1995**, *3* (11), 1197.
79. Brun, E.; Moriaud, F.; Gans, P.; Blackledge, M. J.; Barras, F.; Marion, D., Solution structure of the cellulose-binding domain of the endoglucanase Z secreted by *Erwinia chrysanthemi*. *Biochemistry* **1997**, *36* (51), 16074.
80. Notenboom, V.; Boraston, A. B.; Williams, S. J.; Kilburn, D. G.; Rose, D. R., High-resolution crystal structures of the lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A with bound substrates reveal a novel mode of xylan binding. *Biochemistry* **2002**, *41* (13), 4246.
81. Kraulis, P. J.; Clore, G. M.; Nilges, M.; Jones, T. A.; Pettersson, G.; Knowles, J.; Gronenborn, A. M., Determination of the 3-Dimensional Solution Structure of the C-Terminal Domain of Cellobiohydrolase-I from *Trichoderma-Reesei* - a Study Using Nuclear Magnetic-Resonance and Hybrid Distance Geometry Dynamical Simulated Annealing. *Biochemistry* **1989**, *28* (18), 7241.
82. Raghothama, S.; Simpson, P. J.; Szabo, L.; Nagy, T.; Gilbert, H. J.; Williamson, M. P., Solution structure of the CBM10 cellulose binding module from *Pseudomonas xylanase A*. *Biochemistry* **2000**, *39* (5), 978.
83. Saul, F. A.; Rovira, P.; Boulot, G.; Van Damme, E. J. M.; Peumans, W. J.; Truffa-Bachi, P.; Bentley, G. A., Crystal structure of *Urtica dioica* agglutinin, a superantigen presented by MHC molecules of class I and class II. *Struct Fold Des* **2000**, *8* (6), 593.
84. Suetake, T.; Tsuda, S.; Kawabata, S.; Miura, K.; Iwanaga, S.; Hikichi, K.; Nitta, K.; Kawano, K., Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. *Journal of Biological Chemistry* **2000**, *275* (24), 17929.
85. Boraston, A. B.; McLean, B. W.; Kormos, J. M.; Alam, M.; Gilkes, N. R.; Haynes, C. A.; Tomme, P.; Kilburn, D. G.; Warren, R. A. J., Carbohydrate-binding modules: diversity of structure and function. *Recent Advances in Carbohydrate Bioengineering* **1999**, (246), 202.
86. Henshaw, J. L.; Bolam, D. N.; Pires, V. M. R.; Czjzek, M.; Henrissat, B.; Ferreira, L. M. A.; Fontes, C. M. G. A.; Gilbert, H. J., The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities. *Journal of Biological Chemistry* **2004**, *279* (20), 21552.
87. Ribeiro, T.; Santos-Silva, T.; Alves, V. D.; Dias, F. M. V.; Luis, A. S.; Prates, J. A. M.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A., Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities. *Bba-Proteins Proteom* **2010**, *1804* (10), 2054.
88. Nagy, T.; Simpson, P.; Williamson, M. P.; Hazlewood, G. P.; Gilbert, H. J.; Orosz, L., All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *Febs Letters* **1998**, *429* (3), 312.
89. Tomme, P.; Warren, R. A. J.; Miller, R. C.; Kilburn, D. G.; Gilkes, N. R., Cellulose-binding domains: Classification and properties. *Acs Sym Ser* **1995**, *618*, 142.
90. Lehtio, J.; Sugiyama, J.; Gustavsson, M.; Fransson, L.; Linder, M.; Teeri, T. T., The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *P Natl Acad Sci USA* **2003**, *100* (2), 484.

91. Creagh, A. L.; Ong, E.; Jervis, E.; Kilburn, D. G.; Haynes, C. A., Binding of the cellulose-binding domain of exoglucanase Cex from *Cellulomonas fimi* to insoluble microcrystalline cellulose is entropically driven. *P Natl Acad Sci USA* **1996**, *93* (22), 12229.
92. McLean, B. W.; Bray, M. R.; Boraston, A. B.; Gilkes, N. R.; Haynes, C. A.; Kilburn, D. G., Analysis of binding of the family 2a carbohydrate-binding module from *Cellulomonas fimi* xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues. *Protein Eng* **2000**, *13* (11), 801.
93. Bray, M. R.; Johnson, P. E.; Gilkes, N. R.; McIntosh, L. P.; Kilburn, D. G.; Warren, R. A. J., Probing the role of tryptophan residues in a cellulose-binding domain by chemical modification. *Protein Sci* **1996**, *5* (11), 2311.
94. Poole, D. M.; Hazlewood, G. P.; Huskisson, N. S.; Virden, R.; Gilbert, H. J., The Role of Conserved Tryptophan Residues in the Interaction of a Bacterial Cellulose Binding Domain with Its Ligand. *FEMS Microbiology Letters* **1993**, *106* (1), 77.
95. Simpson, P. J.; Xie, H. F.; Bolam, D. N.; Gilbert, H. J.; Williamson, M. P., The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *Journal of Biological Chemistry* **2000**, *275* (52), 41137.
96. McCartney, L.; Blake, A. W.; Flint, J.; Bolam, D. N.; Boraston, A. B.; Gilbert, H. J.; Knox, J. P., Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. *P Natl Acad Sci USA* **2006**, *103* (12), 4765.
97. Pell, G.; Williamson, M. P.; Walters, C.; Du, H. M.; Gilbert, H. J.; Bolam, D. N., Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry* **2003**, *42* (31), 9316.
98. Boraston, A. B.; Notenboom, V.; Warren, R. A. J.; Kilburn, D. G.; Rose, D. R.; Davies, G., Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains. *Journal of Molecular Biology* **2003**, *327* (3), 659.
99. Pires, V. M. R.; Henshaw, J. L.; Prates, J. A. M.; Bolam, D. N.; Ferreira, L. M. A.; Fontes, C. M. G. A.; Henrissat, B.; Planas, A.; Gilbert, H. J.; Czjzek, M., The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5A in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *Journal of Biological Chemistry* **2004**, *279* (20), 21560.
100. Abou-Hachem, M.; Karlsson, E. N.; Simpson, P. J.; Linse, S.; Sellers, P.; Williamson, M. P.; Jamieson, S. J.; Gilbert, H. J.; Bolam, D. N.; Holst, O., Calcium binding and thermostability of carbohydrate binding module CBM4-2 of Xyn10A from *Rhodothermus marinus*. *Biochemistry* **2002**, *41* (18), 5720.
101. Jamal-Talabani, S.; Boraston, A. B.; Turkenburg, J. P.; Tarbouriech, N.; Ducros, V. M. A.; Davies, G. J., Ab initio structure determination and functional characterization of CBM36: A new family of calcium-dependent carbohydrate binding modules. *Structure* **2004**, *12* (7), 1177.
102. Nakamura, S.; Yazawa, R.; Takakura, J.; Sakata, T.; Ihsanawati; Yatsunami, R.; Fukui, T.; Kumasaka, T.; Tanaka, N., A Calcium-Dependent Xylan-Binding Domain of Alkaline Xylanase from Alkaliphilic *Bacillus* sp Strain 41M-1. *Biosci Biotech Bioch* **2011**, *75* (2), 379.
103. Volkov, I. Y.; Lunina, N. A.; Velikodvorskaya, G. A., Prospects for the practical application of substrate-binding modules of glycosyl hydrolases. *Appl Biochem Micro+* **2004**, *40* (5), 427.
104. Tomme, P.; Boraston, A.; McLean, B.; Kormos, J.; Creagh, A. L.; Sturch, K.; Gilkes, N. R.; Haynes, C. A.; Warren, R. A. J.; Kilburn, D. G., Characterization and affinity applications of cellulose-binding domains. *J Chromatogr B* **1998**, *715* (1), 283.
105. Gunnarsson, L. C.; Karlsson, E. N.; Albrekt, A. S.; Andersson, M.; Holst, O.; Ohlin, M., A carbohydrate binding module as a diversity-carrying scaffold. *Protein Eng Des Sel* **2004**, *17* (3), 213.

106. Berdichevsky, Y.; Lamed, R.; Frenkel, D.; Gophna, U.; Bayer, E. A.; Yaron, S.; Shoham, Y.; Benhar, I., Matrix-assisted refolding of single-chain Fv-cellulose binding domain fusion proteins. *Protein Express Purif* **1999**, *17* (2), 249.



Chapter II: Structure of the Family 11 Carbohydrate-Binding Module from Clostridium thermocellum (CtCBM11)

In this chapter I describe the 3D structure of CtCBM11 as determined by X-ray crystallography and NMR spectroscopy. The data here presented is part of a published paper (Viegas et al, 2008)¹ and from a manuscript in preparation.

Table of Contents

Summary	39
II.1 Introduction	40
II.2 Results and Discussion.....	42
II.2.1 Structure of CtCBM11	42
II.2.1.1 The crystal structure of CtCBM11 without the histidine tail	42
II.2.1.2 The solution structure of CtCBM11.....	45
II.2.1.3 Comparison between the X-ray and NMR structures	47
II.3 Conclusions	48
II.4 Materials and methods	49
II.4.1 Molecular biology	49
II.4.1.1 Recombinant protein production.....	49
II.4.1.2 Double labeled (¹³ C and ¹⁵ N) protein expression and purification.....	49
II.4.2 X-ray crystallography.....	52
II.4.2.1 Protein crystallization and data collection	52
II.4.2.2 Phasing, model building and refinement.....	53
II.4.3 NMR spectroscopy.....	53
II.4.3.1 Data acquisition	53
II.4.3.2 Resonance assignment and structure calculation.....	54
II.5 References	57

Summary

The focus of this chapter is on the 3D structure of the family 11 carbohydrate-binding module from *C. thermocellum* – CtCBM11 (**Figure II.1**).^{1,2} The native structure of CtCBM11 was determined in 2004² to a resolution of 1.98 Å and is deposited in the PDB under the code: 1v0a. Its structure suggested that the contacts between residues Ser59, Asp99, Tyr53, Arg126, Tyr129 and Tyr152 and the histidine tail of a symmetry-related molecule could impair ligand binding and thus co-crystallization and soaking experiments.

To tackle this problem I have determined the crystal structure of CtCBM11 without the histidine tag. The new crystals belong to the $P2_1$ space group and comparison of the two structures reveals no major differences at the main-chain level and the two structurally relevant calcium atoms are conserved.

Moreover, I have also determined the NMR solution structure of CtCBM11 at 25 and 50 °C. Both structures are very similar to each other, which is indicative of a very stable protein as one would expect from a thermophilic organism. Additionally, the solution structures are also very similar to the crystal structure. However, a careful comparison between the structures shows that in the NMR structures the binding cleft area is larger than in the crystal structure. The smaller size of the cleft in the crystal structure, probably imposed by the crystal packing, may be the reason for the lack of binding with different cellooligosaccharides in co-crystallization experiments. This result denotes the importance of the geometry of the binding cleft for the binding of cellooligosaccharides and points to a conformation-selection mechanism of ligand recognition and binding for CtCBM11.

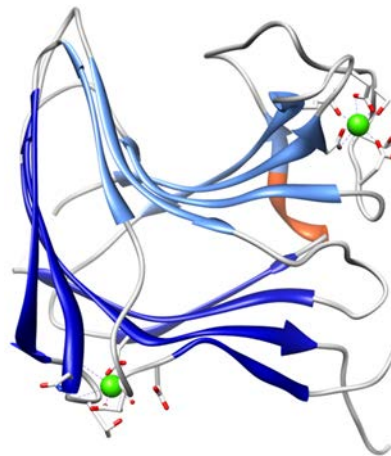


Figure II.1: 3D structure of CtCBM11 obtained by X-ray crystallography.²

The CtCBM11 structure reveals a classical distorted β -jelly roll fold consisting of two six-stranded anti-parallel β -sheets, which form a convex side (β -strands depicted in light blue) and a concave side (β -strands depicted in dark blue). The two calcium ions are (Ca1 – top – and Ca2 – bottom) depicted as green spheres and the residues that bind to calcium are depicted as sticks. The α -helix is depicted in red.

II.1 Introduction

CtCBM11 belongs to a bifunctional enzyme, Lic26A-Cel5E, which contains two glycoside hydrolase (GH) domains - GH5 and GH26 - each one with a CBM11, that display β -1,4- and β -1,3-1,4-mixed linked endoglucanase activity, respectively.² CtCBM11 belongs to the Type B subfamily (see Chapter I - Section I.6.1.2) and it binds to a single polysaccharide chain that can be either β -1,4- or β -1,3-1,4-mixed linked, thus reflecting the specificity of the associated catalytic domains.² Carvalho *et al* (2004)² showed that CtCBM11 has only one binding site that can accommodate at least four sugar units, which is consistent with Type B CBMs.

The native structure of CtCBM11 was determined in 2004² to a resolution of 1.98 Å and is deposited in the PDB under the code: 1v0a. The structure belongs to the $P2_12_12$ space group. CtCBM11 is composed of 172 amino acids (**Figure II.2**), excluding the histidine tag (6 histidines), and has a molecular weight of approximately 20 kDa. Its structure consists of a distorted β -barrel that folds into a β -jelly roll composed of two six-stranded anti-parallel β -sheets, which form a convex side and a concave side (**Figure II.1**). The concave side of CtCBM11 forms a cleft defined by polypeptide stretches Gly20-Glu25, Asp51-Ser59, Glu84-Glu91, Gly98-Ile107, Phe123-Gly133 and Asp146-Asn154 (**Figure II.2**). Furthermore, this depression is occupied by the side chains of residues Tyr22, Asp51, Tyr53, Ser59, Arg86, Met88, Asp99, His102, Ser106, Arg126, Asp128, Tyr129, Asp146, Ser147, His149, Met151 and Tyr152. The core of the β -barrel is extremely hydrophobic and includes seven phenylalanine and six tryptophan residues. Residues Phe120, Ser121, and Ser122 define a 3_{10} -helix.² Due to symmetry constraints, the reported structure exhibits a binding cleft occupied by the *C-terminus* histidine tag of a symmetry-related molecule.

10	20	30	40	50
MASAVGEKML	DDFEGVNLWG	SYSGE GAKVS	TKIVSGKTGN	GMEVSYTGTT
60	70	80	90	100
DGYWGT TVYSL	PDGDWSKWLK	ISFDIKSVDG	SANE IRFMIA	EKSINGVGDG
110	120	130	140	150
EHWVYSI TPD	SSWKTIEIP F	SSFRRRLDYQ	PPGQDMSGTL	DLDNID SIHF
160	170	178		
MYAN NKSGKF	VVDNIK LIGA	LEHHHHHH		

Figure II.2: Amino acid sequence of CtCBM11.

The residues that form the concave side (binding cleft) are colored in blue. The residues that define the 3_{10} -helix are colored in red and the C-terminal histidine tail used is colored in light grey.

As many β -sandwich structures, CtCBM11 has bound calcium ions (two in this case) that are distant from the carbohydrate-binding cleft, thus suggesting a structural role. The coordination of

the two calcium ions is illustrated in **Figure II.3**. The first calcium ion (Ca1) is coordinated in an octahedral fashion by the side chain oxygen atoms of Glu91 (O ϵ 1 and O ϵ 2), Glu101 (O ϵ 1), Asp135 (O δ 1 and O δ 2), Ser137 (O γ), Asp141 (O δ 2), and the main chain oxygen atom of Thr139. The second calcium ion (Ca2) also shows an octahedral coordination and is bound to the main chain oxygen atoms of residues Asp12, Thr38, and Asn40 and to the side chain oxygen atoms of Glu14 (O ϵ 1) and Asp163 (O δ 1 and O δ 2). One water molecule completes the Ca2 coordination sphere. The distances between the ligands and the calcium ions vary from 2.3 to 2.6 Å. Both calcium ions are solvent inaccessible, which represents further evidence for their structural role.

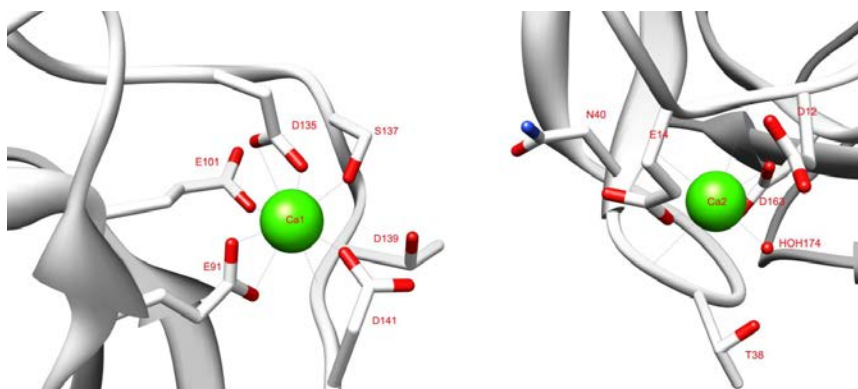


Figure II.3: Coordination of the two calcium ions in CtCBM11.

Both calcium ions show an octahedral coordination and are bound to main chain and side chain oxygens. The calcium ions are represented as green spheres and the residues that bind to calcium are represented as sticks colored by heteroatom. The rest of the protein is represented as ribbons colored in grey.

The main function of CBMs is to increase the catalytic efficiency of the enzymes by putting the substrate and the enzyme into prorogated and close contact.^{3,4} Type B CBMs bind to a large variety of substrates, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose. These proteins disrupt the structure of cellulose fibers through two major mechanisms: (i) by the action of aromatic amino acids, like tryptophan and tyrosine, that are thought to pack onto the sugar rings^{1,3-5}, (ii) and by the conformational fitting of the glycan chains in the binding cleft³. Therefore, stacking/hydrophobic interactions between the sugar rings and aromatic residues in the CBMs and conformational fitting of the glycan chains, that confer additional specificity and stability to the protein-carbohydrate complex, seem to play a key role in ligand recognition.^{1,2,6-8} In spite of these findings, a detailed molecular and mechanistic understanding of CBM-carbohydrate interaction and of the molecular determinants for CBM/ligand recognition is still an open question and a major topic of research.

In order to achieve my goal - *understand the molecular interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their*

substrates – a fundamental requirement is the three dimensional structure of the protein. In this chapter I describe the crystal structure of CtCBM11 without the engineered histidine tail and the solution structure of the same protein at 25 and 50 °C. The newly determined crystal structure reveals no major differences with respect to the one previously determined (PDB code: 1v0a²), with a root mean square deviation (rmsd) of only 0.6 Å for 167 α -carbon atoms. Regarding the NMR-determined solution structures at 25 and 50°C they are similar to each other with and rmsd of 1.24 Å (for 120 C α atoms) between the ensemble representative NMR solution structures. Both structures are also similar to the X-ray structure, with a rmsd of 1.24 Å (for 121 C α atoms) for the structure at 25°C and 1.12 Å (for 86 C α atoms) for the structure at 50°C. The main differences between all three structures are localized at the loop regions and suggest a key role of the geometry of the binding cleft in the interaction with cellooligosaccharides.

II.2 Results and Discussion

II.2.1 Structure of CtCBM11

In order to get a deeper understanding on the molecular determinants that defines ligand specificity CtCBM11, a fundamental requirement is the three dimensional structure of the protein. In a first approach, a new protocol was developed in which, after the protein was over expressed, the tail was removed (*see Section II.4.1*). With this new protein, crystals were obtained for the subsequent structure determination. Because I was also interested in understanding the internal dynamic processes that occur upon binding and on how the structure is affected by temperature, in a second approach, the solution structure of the apo form of CtCBM11 was determined by Nuclear Magnetic Resonance (NMR) at 25 and 50°C. Experimental details of all the technique applied are explained in Materials and methods (*Section II.4*).

II.2.1.1 The crystal structure of CtCBM11 without the histidine tail

The structure of CtCBM11 with the histidine tail suggests that residues Ser59, Asp99, Tyr53, Arg126, Tyr129 and Tyr152 might be involved in binding mechanisms of possible ligands. However, the presence of the histidine tail seems to have impaired crystal soaking and co-crystallization experiments with candidate ligands (*see Chapter III*). To overcome this problem I have determined the crystal structure of CtCBM11 without the histidine tag. The crystallization conditions of the newly purified protein are different from the tagged one (*see Section II.4.2*), and the new crystals belong to a different space group (**Figure II.4**). The previously determined (with the histidine tail) structure belongs to space group $P2_12_12$, while, in

the absence of the 6-histidine tail, CtCBM11 crystals grew in the $P2_1$ space group. Comparison of the two structures reveals no major differences at the main-chain level, with an rmsd of 0.6 Å for 167 α -carbon atoms (**Figure II.5**) and with the two structurally relevant calcium atoms conserved. In contrast with the previously characterized model, this new model includes residues Asp79, Gly80 and Ser81, which were absent due to loop disorder. In the model with the histidine tail this loop was solvent exposed while in the new structure it has restricted movement as a consequence of the absence of the *C-terminus* histidines.

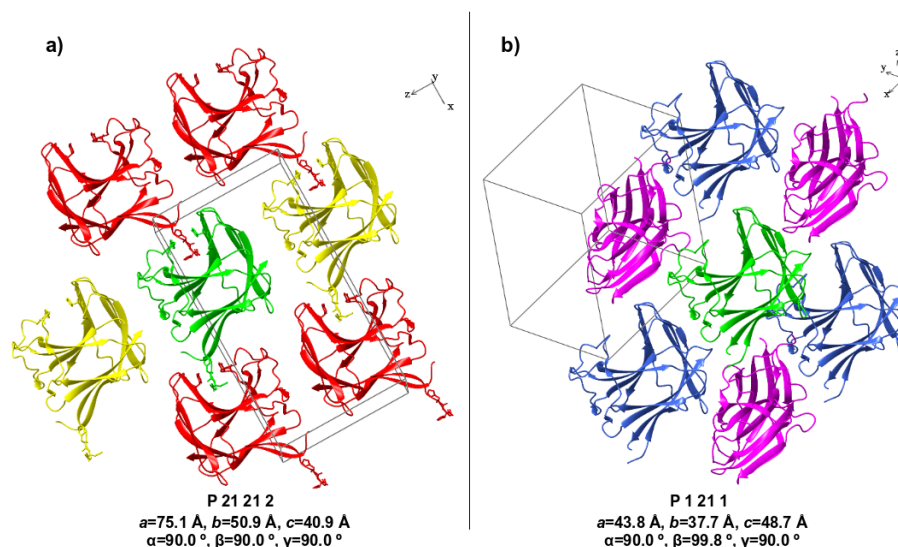


Figure II.4: Ribbon representation of CtCBM11 packing in the two different crystal forms, $P2_12_12$ and $P2_1$.

The $P2_1$ packing is a consequence of the histidine tag removal. This tag (depicted as stick model) was occupying the putative ligand-binding cleft of each symmetry-related molecule. The asymmetric unit is represented in green, while other molecules are colored according to equivalent symmetry operations. In the $P2_12_12$ crystal form, the two tyrosine residues (Tyr53 and Tyr129), flanking the symmetry-related histidine tail, are also shown as stick model and colored accordingly.

Although crystals of the protein without the histidine tail were obtained, the engineered tag seems to be important for crystallization, since the crystals, in the absence of these extra residues, were comparatively more fragile and exhibited a lower diffraction quality. This is intuitive from the observation of the crystal packing (**Figure II.4**). Binding of the three histidines to the substrate recognition site strengthens the intermolecular contacts, favoring crystal stability.

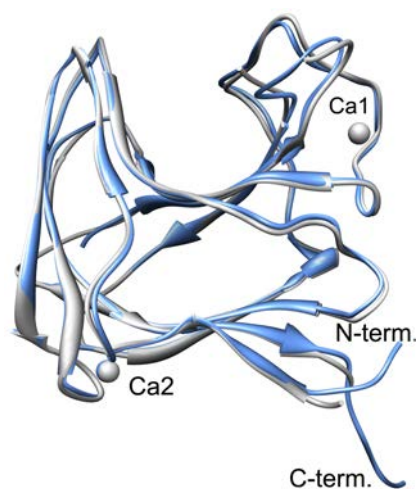


Figure II.5: Superposition of the CtCBM11 structures determined with and without the histidine tail (structures depicted in blue and grey, respectively).

residues from the *C-terminus* end. X-ray data collection and final refinement statistics are shown in **Table II.1**.

The structure of CtCBM11 without the histidine tail was solved by molecular replacement (see Chapter VIII, Section VIII.4.2.1) using the software PHASER⁹ from the CCP4 suite¹⁰ and the previous structure (1v0a) as a model. I used ARPwARP¹⁰ to perform initial building of the complex into the electron density and COOT¹¹ to build the remaining residues. The refinement was performed with REFMAC5.¹² Water molecules were added and final refinement included translation, libration and screw-rotation groups (TLS).^{13,14} The final model has *R-value* = 23.5% and *R_{free}* = 29.5% (see Chapter VIII, Section VIII.4.2.3) and includes 59 water molecules and two calcium ions. Due to disorder, three residues are missing from the *N-terminus*, as well as two

Table II.1: X-ray data and structure quality statistics for CtCBM11.

<i>Data collection</i>	<i>CtCBM11 with no HisTag</i>
Space group	$P2_1$
Cell parameters	$a=43.8 \text{ \AA}$, $b=37.7 \text{ \AA}$, $c=48.7 \text{ \AA}$ $\alpha=90.0^\circ$, $\beta=99.8^\circ$, $\gamma=90.0^\circ$
Wavelength, \AA	1.5418
Resolution of data (outer shell), \AA	20.00 – 2.40 (2.53 – 2.40)
R_{merge} (outer shell), % ^a	31.1 (44.9)
Mean $I/\sigma(I)$	3.8 (2.1)
Completeness (outer shell), %	99.1 (99.9)
Redundancy	3.4
<i>Structure refinement</i>	
No. protein atoms	1357
No. solvent waters	143
Resolution used in refinement, \AA	20.00 – 2.40
Reflections	5629

R-value / R_{free} (%)^b	23.5 / 29.5
Rms deviation 1-2 bonds (Å)	0.011
Rms deviation 1-3 bonds (degrees)	1.637
Rms deviation chiral volume (Å³)	0.159
Average B factors (Å²)	
main-chain	29.1
side-chain	28.5
Ca²⁺ (1)	48.6
Ca²⁺ (2)	39.1
water molecules	41.2

^a $R_{\text{merge}} = \frac{\sum |I - \langle I \rangle|}{\sum \langle I \rangle}$, where I is the observed intensity, and $\langle I \rangle$ is the statistically weighted average intensity of multiple observations.

^b $R\text{-value} = \frac{\sum ||F_{\text{calc}}| - |F_{\text{obs}}||}{\sum |F_{\text{obs}}|} \times 100$, where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively (R_{free} is calculated for a randomly chosen 5% of the reflections).

II.2.1.2 The solution structure of CtCBM11

C. thermocellum grows at T_{opt} of 60 °C and has T_{max} of 69 °C and a T_{min} above 28 °C¹⁵. In order to investigate the influence of temperature in the protein structure and dynamics I have determined the NMR solution structure of the protein at 25 and 50 °C following a standard triple resonance approach using double labeled (¹³C and ¹⁵N) CtCBM11 (see Chapter VII, Section VII.3).^{16,17} For both temperatures, the NH of residue Gly39 was not observed in the ¹⁵N-¹H-HSQC. At 25 °C, the NH of residues Met1, Ser3, Ala4, Val5, Lys67 and Leu69 were not assigned and at 50 °C, residues Met1, Ser3, Ala4, Val5, Thr50, Lys67 and Asn155 were also not assigned. In both data sets, the resonances of the *C-terminal* histidine tag were not used for the calculation of the structures. The coordinates of the structures determined at 25 and 50 °C were deposited in the BMRB data bank (<http://www.bmrwisc.edu>) (18388 and 18389, for the structures at 25 and 50 °C, respectively) and in the PDB (<http://www.pdb.org/pdb>) (2lro and 2lrp, for the structures at 25 and 50 °C, respectively). **Table II.2** lists the structural statistics for the deposited NMR structures and **Figure II.6** shows the energy minimized representative structures of CtCBM11 at 25 °C and 50 °C (A and B, respectively). Using the software MolProbity¹⁸ (<http://molprobity.biochem.duke.edu/>) for analyzing the NMR structures I got that at 25 °C, 92.6% of the residues lie in the favored regions (99.4% in allowed regions), while at 50°C, 92.3% of the residues lie in the favored regions (99.2% in allowed regions) of the Ramachandran plot.

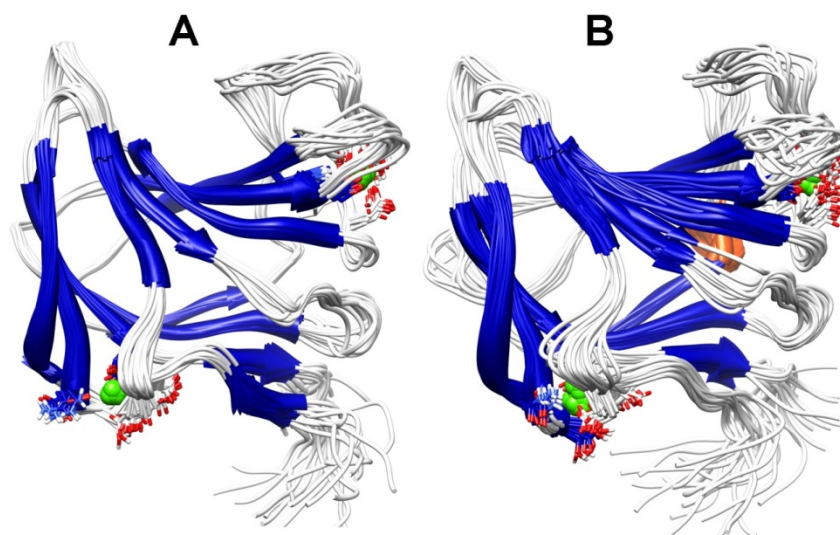


Figure II.6: Ribbon representation of the NMR-determined 20-structure ensemble of CtCBM11 at 25 °C (A) and 50 °C (B).

The calcium ions are depicted as green spheres and the residues that bind to calcium are depicted as sticks and colored by heteroatom. β -sheets are depicted in blue, α -helix is depicted in red and random coil is depicted in grey.

The two calcium ions (**Figure II.6** – green spheres) were added at the final stages of the structure calculation by adding a new residue in the amino acid sequence (*see Materials and methods, Section II.4.3.2*). The coordination of both ions is identical to the one seen in the crystal structure with the exception of the water molecules which were not included in the calculation. Also in these structures the distances between the ligands and the calcium ions vary from 2.3 to 2.6 Å.

Table II.2: Structural statistics for the NMR structures of CtCBM11.

	CtCBM11	
	25 °C	50 °C
NMR distance and dihedral constraints		
Distance constraints		
Total distance restraints from NOEs	2559	1398
Short range ($ i-j \leq 1$)	1658	873
Medium-range ($1 < i-j < 5$)	207	109
Long-range ($ i-j \geq 5$)	694	416
Total dihedral angle restraints	772	708
phi	300	292
psi	197	191

chi	225	225
Structure statistics		
Violations (mean and s.d.)		
Distance constraints (Å)	0.0252	0.0388
Dihedral angle constraints (°)	1.5336	1.7652
Max. dihedral angle violation (°)	2.0185	2.3475
Max. distance constraint violation (Å)	0.0361	0.0578
Average pairwise rmsd for residues 12-160 (Å)		
Heavy	0.78	0.93
Backbone	1.16	1.59
Cyana target function (Å ²)	6.39	4.75
Ramachandran's plot analysis		
Favored regions %	92.6	92.3
Allowed regions %	99.4	99.2

II.2.1.3 Comparison between the X-ray and NMR structures

As can be seen in **Figure II.7** both structures are very similar to each other, with an rmsd of 1.03 Å between the ensemble representative NMR solution structures (**Figure II.7 - C**). This is indicative of a very stable protein, as one would expect from a thermophilic organism.

Both structures are also similar to the X-ray structure, with rmsd of 1.20 Å for the structure at 25°C and 1.10 Å, for the structure at 50°C (**Figure II.7 - D**). However, a careful comparison between the NMR solution structures and the crystal structure shows that the β -sheet elements superpose quite well, whereas the loop regions superpose less well (**Figure II.7 - C and D**). This is especially true in the loop formed between residues R125-Q134, which has the largest rmsd value. Interestingly, this makes the binding cleft area larger in the NMR structure than in the crystal structure (approximately 3700 and 3760 Å² for the structures at 25 and 50 °C versus 3225 Å²). The closed conformation of the binding cleft imposed by the crystal packing, as displayed in the X-ray structure, may impair the binding of cellooligosaccharides and the difference found between the solution and the X-ray structure might explain the failed attempts for co-crystallizing CtCBM11 with several ligands. This result reveals a key role of the geometry of the binding cleft in the interaction with cellooligosaccharides that is in good agreement with other reported results.¹⁹ In this sense, NMR provides a more accurate description of the solution structure of CtCBM11 as it accounts for the conformational modifications of the binding cleft that allow ligand binding. The results indicate that significant

changes in the binding cleft may occur do to the crystal packing and this is important information to consider when using X-ray structures for binding studies, especially molecular docking studies.

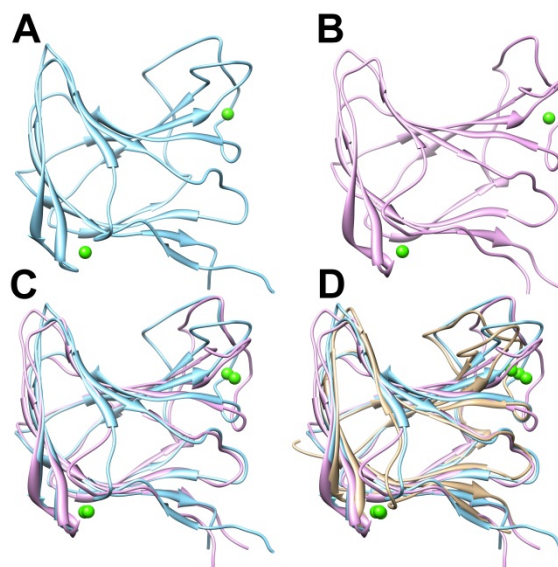


Figure II.7: Comparison between the X-ray and NMR structures.

A) Structure determined at 25 °C; **B)** Structure determined at 50 °C. **C)** Superposition of the structures determined at 25 (light blue) and 50 °C (pink). **D)** Superposition of the X-ray structure (brown) with the NMR solution structures determined at 25 (dark blue) and 50 °C (cyan).

II.3 Conclusions

The crystals of CtCBM11 without the histidine tag grew in the $P2_1$ space group contrasting with the previous $P2_12_12$. The absence of the histidine tag seems to be important for crystallization, since the crystals obtained in these conditions were comparatively more fragile and exhibit a lower diffraction quality than the previous ones. Comparison of the two structures reveals no major differences at the main-chain level. Furthermore, this new model includes residues Asp79, Gly80 and Ser81, which were absent in the previous one due to loop disorder.

Besides the crystal structure of CtCBM11 without the histidine tag, the NMR solution structure was also determined at 25 and 50 °C. The calculated solution structures were almost identical at both temperatures revealing a very stable protein, as expected from a thermophilic organism. Comparison of the protein solution structure with the crystal structure revealed that the binding cleft area in the solution structure is larger than in the crystal structure (~ 3700 and 3760 \AA^2 for the structures at 25 and 50 °C versus 3225 \AA^2). The smaller size of the cleft in the crystal structure, probably imposed by the crystal packing, may explain of the failed co-

crystallization attempts with different cellooligosaccharides. This result denotes the importance of the geometry of the binding cleft for the binding of cellooligosaccharides and points to a conformation-selection mechanism of ligand recognition and binding for CtCBM11.

II.4 Materials and methods

II.4.1 Molecular biology

II.4.1.1 Recombinant protein production

To express CtCBM11 in *Escherichia coli*, I used a vector kindly provided by Professor Carlos Fontes (Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa). For the production of CtCBM11 with the histidine tag the region of the Lic26A-Cel5A gene (lic26A-cel5A) encoding the internal family 11 CBM was amplified from *C. thermocellum* as described elsewhere². The excised CtCBM11 encoding gene was cloned into the vector pET21a (Novagen) to generate pAG1. The recombinant plasmids contain the clostridial gene under the control of the T7 promoter allowing very high expression levels (*see Appendix A, Section A.2*).

This part of the work as well as the production, expression, purification and quantification of the protein without the histidine tag was performed at Faculdade de Medicina Veterinária from Universidade Técnica de Lisboa prior to the beginning of my PhD.

II.4.1.2 Double labeled (¹³C and ¹⁵N) protein expression and purification

Double labeled CtCBM11 (¹³C/¹⁵N-CtCBM11) was produced by first transforming the pAG1 expression vector into competent *E. coli* BL21 (DE3) cells (Novagen). For the transformation, 3 µL of pAG1 were added to 100 µL of *E. coli* BL21 cells and then incubated 30 min in ice. Then the cells were incubated at 42 °C during 45 s and transferred to ice where they rested for 5 min. 1 mL of sterile Luria-Bertani medium (*see Appendix A, Table A.1*) pre-warmed at 37 °C was added to the cells and incubated at the same temperature for 1 h. 100 µL of cells were spread in a LB-agar plate containing 100 µg/mL of ampicillin. The plate was incubated overnight at 37°C.

Initially 5 mL of sterile LB medium containing 100 µg/ml of ampicillin was inoculated with a single colony from the plate and let to grow overnight at 37 °C, at 180 rpm. From the resulting culture, 500 µL were used to produce a glycerol stock, which was kept at -80 °C. The remaining culture was used to inoculate 1 L of sterile M9 minimal medium containing 100 µg/ml

ampicillin, $^{15}\text{NH}_4\text{Cl}$ and $^{13}\text{C}_6$ glucose (see Appendix A, Tables A.2 and A.3). The culture was grown at 37 °C at 200 rpm until the optical density at 595 nm reached 0.6 ($\text{OD}_{595}=0.6$), at which point isopropyl- β -D-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM to induce the gene expression (see Appendix A, Section A.2). The culture was then incubated overnight (~ 17 h) at 30°C and 200 rpm. These conditions are a result of an optimization of the induction time that led to a yield increase of about 10-fold. The cells were collected by centrifugation (5000 rpm, for 15 min at 4 °C), and the cell pellet was resuspended in a 50 mM sodium HEPES buffer, pH 7.5, containing 1 M NaCl, 10 mM imidazole and 5mM CaCl_2 (see Appendix A, Table A.4). The cells were then lysed by sonication (10 x 1 min pulses with 1 min pause between pulses) and put in a 60 °C bath for 30 min to remove the majority of the *E. coli* proteins. The cell residues were removed by centrifugation (7000 rpm, for 30 min at 4 °C) and the supernatant was filtered (0.45 μm membrane pore) and kept at 4 °C for further protein purification.

The protein was purified by ion metal affinity chromatography (IMAC). The protein extract was loaded onto a Ni-NTA-agarose column (QIAGEN) previously washed with 2 column volumes of distilled water, charged with 2 column volumes NiSO_4 and washed again with 2 column volumes of working buffer (see Appendix A, Table A.4). When charged with Ni^{2+} ions, the column will selectively retain proteins if complex-forming amino acids residues, in particular histidines, are exposed on the surface of the protein. Histidine tagged proteins can be desorbed with buffers containing imidazole.²⁰ CtCBM11 was loaded into the column and washed with 2 column volumes of washing buffer (50 mM sodium HEPES buffer, pH 7.5, containing 1 M NaCl and 10 mM imidazole – see Appendix A, Table A.4). The purified protein was then desorbed in a discontinuous way by loading 5 column volumes of elution buffer, consisting of 50 mM sodium HEPES buffer, pH 7.5, 1 M NaCl and 300 mM imidazole and collecting the outflow (see Appendix A, Table A.5). The purified protein was buffer exchanged, in PD-10 Sephadex G-25M gel filtration columns (Amersham Pharmacia Biosciences), into water to remove the imidazole. The column was first washed with 25 mL of distilled water and loaded with 2.5 mL of sample. The resulting outflow was discarded and the protein was eluted with 3.5 mL of distilled water. This procedure was repeated until all the sample was buffer exchanged.

The purity of the protein was then confirmed by running a sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) on the collected fractions (**Figure II.8**). Samples of 40 μL of each collected fraction were boiled with 10 μL of 5x sample buffer for 5 min before loading 18 μL of each into the gel. The gel was stained with Coomassie brilliant blue for 20 min and then destained with a mixture of 10% methanol/10% acetic acid in water (see Appendix A, Tables A.6, A.7, A.8 and A.9).

The purified protein was then concentrated with Amicon centricons with 10-kDa molecular-mass centrifugal membranes (Millipore, Billerica, MA, USA) by centrifuging at 5000 rpm at 4°C. The final concentration of the protein was kept around 1 mM.

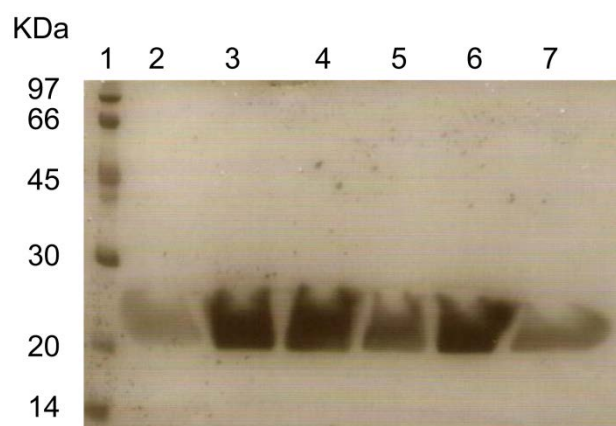


Figure II.8: SDS-PAGE gel of the purified CtCBM11 fractions.

Lane 1 – LMW markers; Lanes 2-7 purified fractions

The concentration of the protein was determined with the Bicinchoninic acid method (BCA) from Sigma Aldrich. The BCA assay primarily relies on two reactions. Firstly the peptide bonds in protein reduce Cu^{2+} ions from the cupric sulfate to Cu^+ (a temperature dependent reaction). The amount of Cu^{2+} reduced is proportional to the amount of protein present in the solution. Secondly, two molecules of bicinchoninic acid chelate with each Cu^{1+} ion and form a purple-colored complex that has a maximum absorbance at a wavelength of 562 nm. The bicinchoninic acid Cu^{1+} complex is aided in protein samples by the presence of cysteine, tyrosine, and tryptophan side chains. As the absorbance is directly proportional to protein concentration, the amount of protein present in a solution can be quantified by measuring the absorption spectra and comparing with protein solutions with known concentrations.²¹

For the application of the BCA assay, first the working reagent was prepared by adding the two BCA reagents, A (sodium bicinchoninate) and B (cupric sulfate), in a proportion of 1:20 (v/v) in water. Then the standard samples were prepared by adding increasing amounts (0, 10, 20, 30, 40 and 50 μL) of bovine serum albumin (BSA) in a concentration of 1 $\text{mg}/\mu\text{L}$ to decreasing amounts (50, 40, 30, 20, 10 and 0 μL) of buffer (water in this case) and 1 mL of the BCA working reagent. The sample tubes were prepared by adding 1, 2 and 5 μL of the protein sample to 49, 48 and 45 μL of water and 1 mL of the BCA working reagent (*see Appendix A, Tables A.10 and A.11*). All the samples were then gently mixed and incubated at 37 °C for 30 min. After the incubation the absorbance was read at 562 nm and the standard curve was constructed by plotting Abs_{562} versus protein concentration. The concentration of the unknown

samples was determined using the equation of the previously determined curve. The yields obtained were around 10 mg/L of protein.

Using the determined concentration, the molar extinction coefficient (ϵ) was determined by UV-visible spectroscopy by reading the absorbance at 280 nm (using a 1.5 mL cuvette with 1 cm path length) and applying the Lambert-Beer law:

$$A = \epsilon cl$$

II.1

where A is the absorbance (read at 280 nm), ϵ is the molar extinction coefficient and l is the path length.

For CtCBM11 the determined molar extinction coefficient was **32449 M⁻¹.cm⁻¹**.

II.4.2 X-ray crystallography

II.4.2.1 Protein crystallization and data collection

Crystals of CtCBM11 without the 6-His tail were grown by vapor diffusion using the hanging drop method and obtained by mixing an equal volume (1 μ L) of protein (50 mg/ml in water) and reservoir solution (30% (m/v) polyethyleneglycol (PEG) 4000, 0.1 M Tris-HCl, pH 8.5, and 0.2 M magnesium chloride)²². In approximately three days, the crystals reach maximal dimensions of 0.3x0.3x0.1 mm³ (**Figure II.9**). Single crystals were harvested in a solution containing 35% (m/v) PEG 4000 and 0.2 M magnesium chloride, and flash-frozen in a liquid nitrogen stream at 100K, using 30% (v/v) glycerol as cryoprotectant²³. Crystal characterization and diffraction data collection were performed in-house, using CuK α X-ray radiation from an Enraf-Nonius rotating anode generator operated at 5 kW, with a MAR-Research image-plate detector. The wavelength of the radiation used was 1.5418 Å and 200 images were collected with an exposure time of 15 minutes per frame. Diffraction data were processed and scaled, respectively, with programs MOSFLM²⁴ and SCALA²⁵ from the CCP4 suite¹⁰. Diffraction experiments showed that, in the absence of the engineered 6-histidine tail, CtCBM11 crystallized in the $P2_1$ space group. The unit cell dimensions are $a=43.8$ Å, $b=37.7$ Å, $c=48.7$ Å and $\beta=99.8^\circ$ and the crystals diffracted beyond 2.4 Å resolution. Solvent calculations revealed a Matthews coefficient of 2.2 Å³Da⁻¹, which corresponds to 44% solvent, with one CtCBM11 molecule in the asymmetric unit.

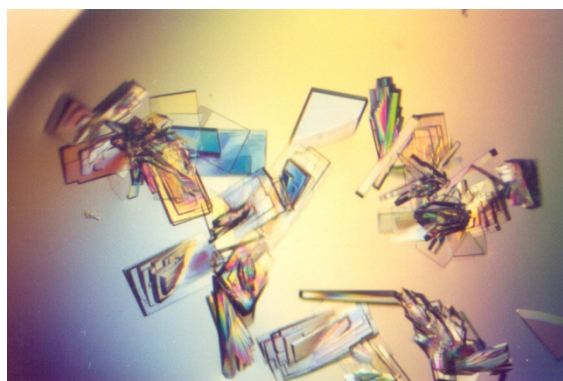


Figure II.9: Crystals of CtCBM11 without the engineered 6-His tail.

II.4.2.2 Phasing, model building and refinement

Considering the calculated Matthews coefficient, molecular replacement attempts were performed searching for one molecule of CtCBM11 in the monoclinic $P2$ cell. The previously described and available structure of CtCBM11, with accession code 1v0a², was used as a search model for molecular replacement. The Patterson search was done with program PHASER⁹, implemented in the CCP4 interface¹⁰, and a clear solution was found in space group $P2_1$, with a z-score of 15.2, against a z-score of 3.1 for the $P2$ alternative space group. Model building was performed interactively using program COOT¹¹. Model refinement and electron density map calculations were done with program REFMAC5¹² from the CCP4 suite¹⁰ to a final R -factor of 23.5% and R_{free} of 29.5%. The final model contains 167 amino-acid residues from an expected number of 172 residues in a single polypeptide chain. Due to disorder, three residues are missing from the N -terminus, as well as two residues from the C -terminus end. The model also includes two calcium ions and 59 water molecules. X-ray data collection and final refinement statistics are included in **Table II.1**.

II.4.3 NMR spectroscopy

II.4.3.1 Data acquisition

All NMR spectra were acquired in a 600 MHz Bruker AvanceIII spectrometer (Bruker, Wissembourg, France) equipped with a 5 mm inverse detection triple-resonance z-gradient cryogenic probehead (CP TCI). All data was processed in Bruker TopSpin2.1 (Bruker).

II.4.3.2 Resonance assignment and structure calculation

In order to assign all the resonances of CtCBM11 and determine its solution structure I have followed a standard triple resonance-based protocol (see Chapter VII, Section VII.3).^{26,27} Because *C. thermocellum* is a thermophilic organism I also acquired data at 50°C. The resonances were assigned with CARA1.8.4.2²⁸ and the structure calculation was performed with CYANA2.1²⁹. For the CtCBM11 resonance assignment I used a double labeled protein sample (¹³C-¹⁵N-CtCBM11) at a concentration of 0.7 mM in 90% H₂O / 10% D₂O. Data were collected in the Bruker Avance III 600 MHz spectrometer at 25 and 50 °C.

II.4.3.2.1 Resonance assignment

Two-dimensional ¹⁵N-¹H- and ¹³C-¹H-edited heteronuclear single quantum coherence (HSQC) and three-dimensional HNCO, HN(CA)CO, HN(CO)CACB, HNCACB and (H)CCH-TOCSY experiments were performed to obtain the chemical shift assignments of backbone atoms. Additional three-dimensional ¹⁵N- and ¹³C-NOESY-HSQC (mixing time 60 and 80 ms, respectively), both in the aliphatic and aromatic regions and HNHA experiments were acquired for complete side chain resonance assignment and NOE measurements (see Chapter VII, Section VII.3). **Table II.3** summarizes the acquisition parameters for the different experiments.

The assignment of the ¹H, ¹³C, and ¹⁵N signals in spectra was performed in CARA1.8.4.2.²⁸ For the semiautomatic protein backbone assignment, I have used the AutoLink module³⁰ integrated into the CARA program.

Table II.3: NMR experiments and acquisition details for the CtCBM11 resonance assignment.

	<i>Complex points</i>			<i>Spectral width (Hz)</i>			<i>Number of scans</i>
	¹ H	¹⁵ N	¹³ C	¹ H	¹⁵ N	¹³ C	
Backbone assignment							
2D							
¹⁵ N/ ¹ H-HSQC	2048	256	-	12019	2311	-	8
¹³ C/ ¹ H-HSQC (aliph.)	2048	-	512	12019	-	24999	32
¹³ C/ ¹ H-HSQC (aro.)	2048	-	1024	9014	-	11495	32
3D							
NHCO	2048	40	128	9615	2311	2777	16
HN(CA)CO	2048	40	128	9615	2311	2777	16
NH(CO)CACB	2048	40	128	9615	2311	11320	16
NHCACB	2048	40	128	9615	2311	11320	16

Side-chain assignment

	¹ H	¹³ C	¹³ C	¹ H	¹³ C	¹³ C	
(H)CCH-TOCSY	2048	48	180	9615	11364	11364	16
	¹ H	¹⁵ N	¹ H	¹ H	¹⁵ N	¹ H	
HNHA	2048	128	40	9615	2311	9615	16

NOE measurement

¹⁵ N-NOESY-HSQC	2048	40	256	9615	2311	9615	16
	¹ H	¹³ C	¹ H	¹ H	¹³ C	¹ H	
¹³ C-NOESY-HSQC (aliph.)	2048	60	256	10000	11363	8333	16
¹³ C-NOESY-HSQC (aro.)	2048	60	256	10000	11363	8333	16

II.4.3.2.2 Structure calculation

After assignment completion, CYANA2.1²⁹ analyzed peak data derived from the NOESY spectra in a semi-automated iterative manner²⁶. I have used CARA1.8.4.2²⁸ to automatically generate the NOE coordinates and intensities for the analysis. The input data consisted of the amino acid sequence, assigned chemical shift list, peak volume list and backbone dihedral angles (Φ and Ψ) derived from TALOS³¹ (see Chapter VII, Section VII.3.1.1.5). The unambiguous NOEs were converted into upper limits by CYANA2.1²⁹ using the macro *calibrate* (see Chapter VII, Section VII.3.1.3.2). No stereospecific assignments were introduced initially. In the final steps, 43 and 23 pairs of stereospecific limits were introduced by CYANA for the structures at 25 and 50 °C, respectively. To ensure that the peak lists are faithful representatives of the NOESY spectra, the chemical shift positions of the NOESY cross-peaks must be correctly calibrated to fit the chemical shift lists within the chemical shift tolerances. As a result I have used 0.02 ppm for the direct and indirect dimensions and 0.40 for the heavy atom dimension (¹⁵N and ¹³C). CYANA2.1²⁹ used the given input to compute seven cycles of NOE cross-peak assignment and structure calculation, each with 100 starting structures, from which the 20 best were kept. After the first few rounds of calculations, I analyzed the spectra again to identify additional cross-peaks consistent with the structural model and to remove misidentified peaks. I have applied 97 hydrogen bond constraints at a late stage of the structure calculation for identifiable characteristic NOE patterns observed for α -helices or β -strands according to **Table II.4** (89 for β -strands and 8 for α -helices for both structures). The calcium ions were finally included in the calculations by adding a new residue in the amino acid sequence. This residue is formed from a chain of dummy atoms that have their van der Waals

radii set to zero so they can freely penetrate into the protein and one atom, which mimics the calcium ion. Atoms O ϵ 1 and O ϵ 2 from Glu91, O ϵ 1 from Glu101, O δ 1 and O δ 2 from Asp135, O γ from Ser137, O δ 2 from Asp141 and main-chain O from Thr139 were linked to the first calcium ion through upper and lower distance limits of 2.4 and 2.2 Å, respectively. Atoms O ϵ 1 from Glu14, main-chain O from Asp12, Asp 38 and Asn40 and O δ 1 and O δ 2 from Asp163 Thr139 were linked to the second calcium ion through the same upper and lower distance limits. This approach does not impose any fixed orientation of the ligands with respect to the calcium ion. Input data and structure calculation statistics are summarized in **Table II.2**.

Table II.4: Short-range distances in the secondary structure elements.³²

<i>Distance</i>	<i>α-helix</i>	<i>3₁₀-helix</i>	<i>β-sheet (A)</i>	<i>β-sheet (P)</i>
d$_{\alpha N}$	3.5	3.4	2.2	2.2
d$_{\alpha N(i,i+2)}$	4.4	3.8		
d$_{\alpha N(i,i+3)}$	3.4	3.3		
d$_{\alpha N(i,i+4)}$	4.2	3.3		
d$_{NN}$	2.8	2.6	4.3	4.2
d$_{NN(i,i+2)}$	4.2	4.1		

The 20 conformers with the lowest final CYANA target function values were further subjected to restrained energy-minimization in a water shell by using the AMBER 9.0 package³³ using the all atom force field ff99SB³⁴. The structures were immersed in an octahedric box using the TIP3P water model³⁵, with a thickness of 10 Å. A total of 8 sodium counter ions were also included to neutralize charge. The simulation was performed by using periodic boundary conditions and the particle-mesh Ewald approach to account for the electrostatic interactions.³⁶ The restrained energy minimization was performed in three stages. In the first stage, the solvent molecules were minimized by MM keeping the solute fixed with the positional restraint of 500 Kcal mol⁻¹ Å⁻² followed by the relaxing of the entire system after restraint removal. In the last stage, a maximum of 1500 steps of restrained energy minimization and a combination of the steepest descent and conjugate gradient algorithms were applied by using a parabolic or linear penalty function for the NOE upper distance bonds and torsion-angle restraints.

I have used CHIMERA³⁷ and PyMOL1.4.1³⁸ to visualize the structures, calculate accessibilities, and to prepare the diagrams of the molecules.

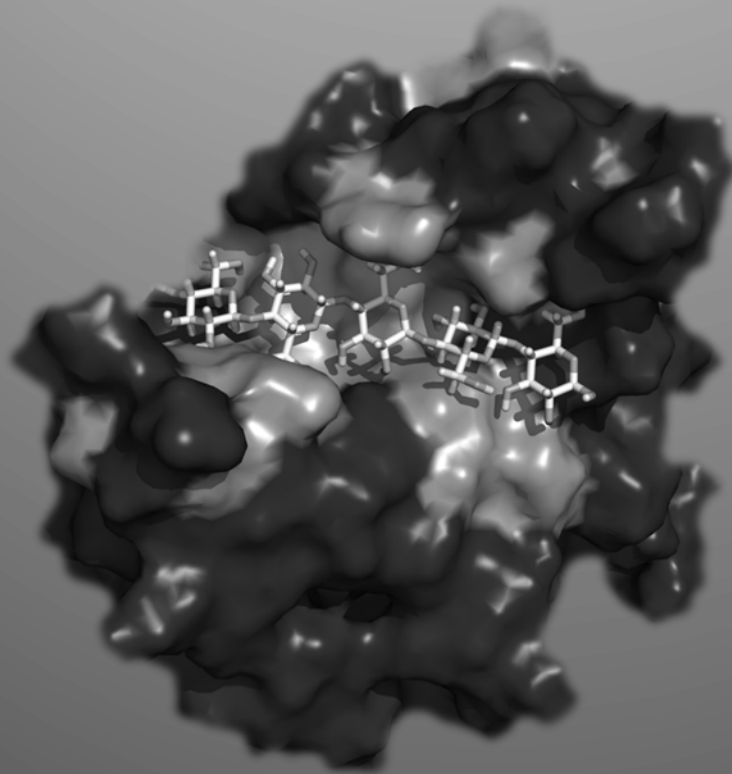
II.4.3.2.3 Structure validation

The quality of the CtCBM11 ensembles (at 25 and 50 °C) was evaluated by their agreement with the quality scores as determined by the software MolProbity¹⁸ (<http://molprobity.biochem.duke.edu/>).

II.5 References

- Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008**, 275 (10), 2524.
- Carvalho, A. L.; Goyal, A.; Prates, J. A. M.; Bolam, D. N.; Gilbert, H. J.; Pires, V. M. R.; Ferreira, L. M. A.; Planas, A.; Romao, M. J.; Fontes, C. M. G. A., The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *Journal of Biological Chemistry* **2004**, 279 (33), 34785.
- Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **2004**, 382 (Pt 3), 769.
- Hashimoto, H., Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **2006**, 63 (24), 2954.
- Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R., The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* **2004**, 58, 521.
- Tsukimoto, K.; Takada, R.; Araki, Y.; Suzuki, K.; Karita, S.; Wakagi, T.; Shoun, H.; Watanabe, T.; Fushinobu, S., Recognition of celooligosaccharides by a family 28 carbohydrate-binding module. *Febs Letters* **2010**, 584 (6), 1205.
- Pell, G.; Williamson, M. P.; Walters, C.; Du, H. M.; Gilbert, H. J.; Bolam, D. N., Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry* **2003**, 42 (31), 9316.
- Xie, H. F.; Bolam, D. N.; Nagy, T.; Szabo, L.; Cooper, A.; Simpson, P. J.; Lakey, J. H.; Williamson, M. P.; Gilbert, H. J., Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry* **2001**, 40 (19), 5700.
- McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr D* **2005**, 61, 458.
- Bailey, S., The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallogr D* **1994**, 50, 760.
- Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **2004**, 60, 2126.
- Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* **1997**, 53, 240.
- Painter, J.; Merritt, E. A., Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D* **2006**, 62, 439.
- Painter, J.; Merritt, E. A., TLSMD web server for the generation of multi-group TLS models. *J Appl Crystallogr* **2006**, 39, 109.
- Freier, D.; Mothershed, C. P.; Wiegel, J., Characterization of *Clostridium-Thermocellum* Jw20. *Appl Environ Microb* **1988**, 54 (1), 204.
- Yamazaki, T.; Lee, W.; Arrowsmith, C. H.; Muhandiram, D. R.; Kay, L. E., A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of ¹⁵N, ¹³C, ²H Labeled Proteins with High Sensitivity. *Journal of the American Chemical Society* **1994**, 116 (26), 11655.
- Shan, X.; Gardner, K. H.; Muhandiram, D. R.; Rao, N. S.; Arrowsmith, C. H.; Kay, L. E., Assignment of ¹⁵N, ¹³C α , ¹³C β , and HN Resonances in an ¹⁵N,¹³C,²H Labeled 64 kDa Trp Repressor-Operator Complex Using Triple-Resonance NMR Spectroscopy and ²H-Decoupling. *Journal of the American Chemical Society* **1996**, 118 (28), 6570.
- Chen, V. B.; Arendall, W. B., III; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure

- validation for macromolecular crystallography. *Acta Crystallographica Section D* **2010**, 66 (1), 12.
19. Czjzek, M.; Bolam, D. N.; Mosbah, A.; Allouch, J.; Fontes, C. M. G. A.; Ferreira, L. M. A.; Bornet, O.; Zamboni, V.; Darbon, H.; Smith, N. L.; Black, G. W.; Henrissat, B.; Gilbert, H. J., The location of the ligand-binding site of carbohydrate-binding modules that have evolved from a common sequence is not conserved. *Journal of Biological Chemistry* **2001**, 276 (51), 48580.
 20. Heijbel, A.; Andersson, K.; Bell, P.; Gustafsson, C., Purification of poly(His)-tagged recombinant proteins using HisTrap(TM). *Faseb J* **1996**, 10 (6), 743.
 21. Smith, P. K.; Krohn, R. I.; Hermanson, G. T.; Mallia, A. K.; Gartner, F. H.; Provenzano, M. D.; Fujimoto, E. K.; Goeke, N. M.; Olson, B. J.; Klenk, D. C., Measurement of Protein Using Bicinchoninic Acid. *Analytical Biochemistry* **1985**, 150 (1), 76.
 22. Jancarik, J.; Kim, S. H., Sparse-Matrix Sampling - a Screening Method for Crystallization of Proteins. *J Appl Crystallogr* **1991**, 24, 409.
 23. Garman, E. F.; Mitchell, E. P., Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions. *J Appl Crystallogr* **1996**, 29, 584.
 24. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletters on Protein Crystallography* **1992**, 26.
 25. Evans, P. R., Scaling of MAD data. In *Proceedings of the CCP4 Study Weekend. Recent advances in phasing*, Winn, M., Ed. 1997; Vol. 33, pp 22.
 26. Wuthrich, K., NMR studies of structure and function of biological macromolecules. *Bioscience Rep* **2003**, 23 (4), 119.
 27. Cavanagh, J., *Protein NMR spectroscopy : principles and practice*. 2nd ed.; Academic Press: Amsterdam ; Boston, 2007; p 885.
 28. Keller, R. The Computer Aided Resonance Assignment Tutorial. The Swiss Federal Institute of Technology, Zurich, 2004.
 29. Guntert, P., Automated NMR structure calculation with CYANA. *Methods Mol Biol* **2004**, 278, 353.
 30. Masse, J. E.; Keller, R., AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. *J Magn Reson* **2005**, 174 (1), 133.
 31. Cornilescu, G.; Delaglio, F.; Bax, A., Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol Nmr* **1999**, 13 (3), 289.
 32. Wüthrich, K., *NMR of proteins and nucleic acids*. Wiley: New York, 1986; p 292.
 33. Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, H. M.; Pearman, D. A.; Crowley, M.; Walker, R. C.; Zhang, B.; Wang, S.; Havik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S. R.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Morgan, J.; Hornak, R.; Cui, G.; Beroza, P.; Mathew, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*, San Francisco, 2006.
 34. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, 65 (3), 712.
 35. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, 79 (2), 926.
 36. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, 98 (12), 10089.
 37. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **2004**, 25 (13), 1605.
 38. Schrödinger, LLC *The PyMOL Molecular Graphics System*, 1.4.1; 2010.



Chapter III: Molecular Determinants of Ligand Specificity in CtCBM11

*The focus of this chapter is on the molecular determinants that define ligand specificity and binding in the family 11 carbohydrate-binding module from *C. thermocellum* – CtCBM11. Using a X-ray crystallography, NMR and molecular docking combined approach, I was able to identify the atoms of the ligand and the residues of the protein responsible for binding and the mechanisms involved in ligand recognition. The data presented in this chapter is part of a published paper (Viegas et al, 2008)¹, a book chapter (Viegas et al, 2010)² and a manuscript in preparation*

Table of Contents

Summary	64
III.1 Introduction	65
III.2 Results and Discussion.....	68
III.2.1 Characterization of the sugars	68
III.2.2 Molecular determinants of ligand specificity.....	72
III.2.2.1 Co-crystallization studies	73
III.2.2.2 Influence of calcium in the structure of celohexaose	73
III.2.2.3 Linebroadening studies	74
III.2.2.3 Saturation transfer difference NMR (STD-NMR)	75
III.2.2.4 Diffusion studies (DOSY).....	82
III.2.2.5 Interaction studies with celooligosaccharides	83
III.2.2.6 Computational studies	86
III.2.3 Molecular dynamics	94
III.2.3.1 Relaxation data, diffusion tensor and hydrodynamic calculations.....	95
III.2.3.2 Internal mobility.....	99
III.2.3.3 Estimation of the conformational entropy from NMR relaxation data	101
III.2.3.4 Amide proton exchange	102
III.3 Conclusions	104
III.4 Materials and methods	106
III.4.1 Sources of sugars.....	106
III.4.2 Molecular biology	106
III.4.2.1 Recombinant protein production.....	106
III.4.2.2 Transformation, expression, purification and quantification of CtCBM11 with the 6-histidine tail.....	106
III.4.2.2 Transformation, expression, purification and quantification of double labeled (¹³ C and ¹⁵ N) CtCBM11 with the 6-histidine tail	107

III.4.3	X-ray crystallography.....	107
III.4.3.3	Co-crystallization studies	107
III.4.4	NMR spectroscopy.....	107
III.4.4.1	Data acquisition.....	107
III.4.4.2	Characterization of the sugars	108
III.4.4.3	Influence of calcium in the structure of cellobiose	109
III.4.4.4	Linebroadening studies	109
III.4.4.5	STD-NMR studies.....	110
III.4.4.6	Diffusion studies (DOSY).....	111
III.4.4.7	CtCBM11 titration.....	112
III.4.4.8	Combined chemical shift.....	113
III.4.4.9	Determination of the association constant (K_a).....	114
III.4.4.10	Determination of the thermodynamic parameters	116
III.4.4.11	^{15}N backbone relaxation measurements	116
III.4.4.12	Relaxation data processing and analysis	117
III.4.4.13	Estimation of the molecular diffusion tensor	118
III.4.4.14	Hydrodynamic calculations.....	118
III.4.4.15	Calculation of the model free dynamics parameters	119
III.4.4.16	Estimation of the conformational entropy from NMR relaxation data	119
III.4.4.17	Amide proton exchange	120
III.4.5	Computational studies	122
III.4.5.1	Docking experiments with the crystallographic structure and molecular dynamics	122
III.4.5.2	Docking experiments with the NMR solution structure and molecular dynamics	123
III.5	References	124

Summary

The direct conversion of plant cell wall polysaccharides into soluble sugars is one of the most important reactions on earth, and is performed by certain microorganisms such as *Clostridium thermocellum*. These organisms produce extracellular multi-subunit complexes, called cellulosomes that include a consortium of enzymes, which contain non-catalytic carbohydrate-binding modules (CBM) that increase the activity of the catalytic module.

In this chapter, I describe a combined approach by X-ray Crystallography, NMR and Computational Chemistry, in order to gain further insight into the binding mode of different carbohydrates (cellobiose, cellotetraose and cellohexaose) to the binding pocket of the family 11 CBM^{1,2}. Since the structure with a bound substrate could not be obtained, protein titration experiments and computational studies with cellobiose, cellotetraose and cellohexaose were carried out in order to understand the molecular recognition of glucose polymers by *CtCBM11*. These studies provided information on the residues of the protein involved in ligand recognition and on the influence of the length of the saccharide chain on binding. A cluster of aromatic residues has been found to be important for guiding and packing of the polysaccharide. Linebroadening, STD-NMR and DOSY experiments allowed screening the binding activity of the several ligands and identifying the atoms of the ligands closer to the protein upon binding (epitope mapping). The binding cleft of *CtCBM11* interacts more strongly with the central glucose-units of cellotetraose and cellohexaose, mainly through interactions with the OH groups at position 2 and 6 of the central sugar units.

The models of the *CtCBM11*/cellohexaose and *CtCBM11*/cellotetraose complexes obtained by docking allowed a detailed inspection of the main protein ligand interactions. CH- π and Van der Waals interactions were found to be important for the stability of the complexes and to the specificity of the protein. Protein relaxation data analyzed in terms of the model-free approach revealed that the protein behaves as an axial symmetric rotor of the oblate type, independently of the state (bound or free) or temperature. Moreover, thermodynamic data extracted from the titration experiments at 25 and 50 °C and from the general order parameter, S^2 , indicate that binding of cellooligosaccharides to *CtCBM11* must occur by a “conformational selection” mechanism where the disposition of the residues in the binding cleft and interactions with specific groups of the ligand act as determinants of specificity in *CtCBM11*.

Altogether, the results presented allow an atomistic rationalization of the molecular determinants of ligand specificity in *CtCBM11* and the mechanism by which this protein is able to distinguish and select its ligands.

III.1 Introduction

CtCBM11 binds to a single polysaccharide chain that can be either β -1,4- or β -1,3-1,4-mixed linked, reflecting the specificity of the associated catalytic domains (**Table III.1**)³. Quantitative binding studies by ITC showed that the β -1,3-1,4-mixed glucans possess the highest affinity, whereas no affinity for β -1,3 glucans was observed, indicating that not all the sugar-binding sites can accommodate β -1,3-linked glucose residues^{1,3}. The affinity for the mixed linkage tetraoligosaccharide Glc- β -1,4-Glc- β -1,4- Glc- β -1,3-Glc was approximately four times higher than for cellotetraose, corroborating the hypothesis that the protein displays a preference for a β -1,3-linked glucose in at least one subsite. The introduction of another β -1,3 linkage drastically reduces the affinity, suggesting that the protein may only be able to accommodate a single β -1,3-linked glucose.^{3,4}

Table III.1: Quantitative assessment of CtCBM11 binding to oligosaccharides and polysaccharides as determined by ITC.³

<i>Ligand</i>	<i>Temp.</i> (K)	$K_a \times 10^4$ (M^{-1})	ΔG ($kcal\ mol^{-1}$)	ΔH ($kcal\ mol^{-1}$)	ΔTS ($kcal\ mol^{-1}$)	n^a
Lichenan	298.15	30.1 ± 0.4^b	-7.5 ± 0.1	-10.4 ± 0.2	-2.9 ± 0.2	1.0 ± 0.0
Lichenan	333.15	5.3 ± 0.1	-7.2 ± 0.0	-13.0 ± 0.0	-5.8 ± 0.0	1.1 ± 0.0
β-Glucan	298.15	27.1 ± 0.5	-7.4 ± 0.1	-11.2 ± 0.3	-3.8 ± 0.2	1.0 ± 0.0
Cellohexaose	298.15	7.8 ± 0.1	-6.6 ± 0.0	-9.5 ± 0.2	-2.9 ± 0.2	0.8 ± 0.0
Cellopentaose	298.15	5.9 ± 0.3	-6.5 ± 0.0	-8.7 ± 0.3	-2.2 ± 0.3	0.9 ± 0.0
Cellotetraose	298.15	4.4 ± 0.8	-6.3 ± 0.1	-9.8 ± 0.1	-3.5 ± 0.1	1.0 ± 0.0
G4G4G3G^c	298.15	19.2 ± 1.5	-7.2 ± 0.1	-10.2 ± 0.1	-3.0 ± 0.1	1.1 ± 0.0

^a n is the number of binding sites on the protein.

^b The values are given with the standard deviations of replicate titrations.

^c Mixed linkage glucotetraoligosaccharide: Glc- β -1,4-Glc- β -1,4-Glc- β -1,3-Glc.

Determination of the crystallographic structure of the protein with a *C-terminus* histidine tag revealed that, due to symmetry constraints, the binding cleft is occupied by the tag of a symmetry-related molecule (**Figure III.1**). Direct contacts to the histidine tail residues are established by residues Tyr53, Arg126, Tyr129 and Tyr152, suggesting that these residues may contribute to the accommodation and orientation of ligands in the cleft. Residue Asp99 contacts the *C-terminus* tail by means of a water molecule bound to the side chain atoms O δ 1 and O δ 2, in a bidentate way. The side chain Oy of Ser59 is in proximity to the side-chain of the symmetry-related His172 and a possible contact may be mediated by a water molecule, although its location is not clear in the electron density map. This data suggested that residues Ser59,

Asp99, Tyr53, Arg126, Tyr129 and Tyr152 might be involved in binding mechanisms of possible ligands. Further mutagenesis studies (**Table III.2**) confirmed the importance of residues Tyr22, Tyr53 and Tyr129. For all the tested ligands, upon mutation of these residues the affinity dropped dramatically.³

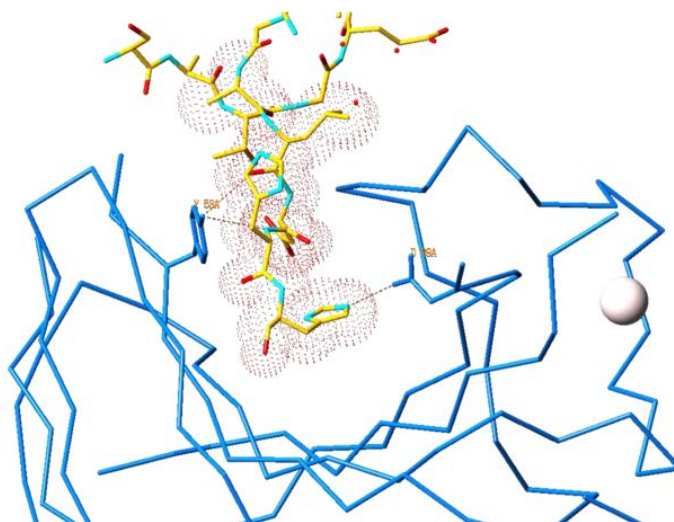


Figure III.1: Highlight of the binding cleft of CtCBM11 with the bound C-terminal histidine tail of a symmetry related molecule.

The histidine tail of the symmetry related molecule is depicted as sticks and coloured by heteroatom. The Van der Waals surface of the histidine tail is depicted as red dots. The calcium ion is depicted as a white sphere.

Table III.2: Binding of wild type CtCBM11 and its mutant derivatives to soluble polysaccharides quantified by affinity gel electrophoresis (AGE).^{3,4}

<i>Ligand</i>	<i>K_a</i> (w/v)				
	Wild type	Y22A	Y53A	Y129A	Y152A
β-Glucan	1194.9	15.5	74.2	56.2	1080.4
Lichenan	701.6	9.7	89.1	68.1	690.1
Hydroxyethyl cellulose	24.4	NB ^a	NB	NB	53.6
Glucomannan	25.9	NB	NB	NB	20.4
Oat spelt xylan	17.5	NB	NB	NB	ND ^b

^a *K_a* below 2.

^b Not determined.

The main function of CBMs is to increase the catalytic efficiency of the associated enzymes by putting the substrate and the enzyme into prorogated and close contact.^{5,6} Type B CBMs bind to a large variety of substrates, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose. These proteins disrupt the structure of cellulose fibers through two major mechanisms: (i) the

action of aromatic amino acids, like tryptophan and tyrosine, that are thought to pack onto the sugar rings^{1,5-7}, (ii) and the conformational fitting of the glycan chains in the binding cleft⁵. Therefore, stacking/hydrophobic interactions between the sugar rings and aromatic residues in the CBMs and conformational fitting of the glycan chains, that confer additional specificity and stability to the protein-carbohydrate complex, seem to play a key role in ligand recognition.^{1,3,8-10} In spite of these findings, a detailed molecular and mechanistic understanding of CBM-carbohydrate interaction and of the molecular determinants for CBM/ligand recognition is still an open question and a major topic of research, because of its importance to fully rationalize the complex mechanism of biomass hydrolysis.

In order to deepen the current knowledge concerning the molecular interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their substrates, I used X-ray Crystallography, NMR and Computational Chemistry approaches to identify the molecular determinants of ligand specificity of CtCBM11. Unfortunately, crystal soaking and co-crystallization of CtCBM11 with candidate ligands was unsuccessfully attempted, as concluded from the observation of difference electron density maps, calculated after diffraction experiments. Confronted with these negative results from the crystallographic approach, I have considered complementary experiments by NMR and computational calculations. The strategy included two complementary ways: (i) one focused on the structure of the ligand and the atoms responsible for binding to the proteins (epitope mapping*), (ii) and the other focused in the identification of the protein residues responsible for ligand recognition. Using saturation transfer difference NMR (STD-NMR) and line broadening studies I have shown that CtCBM11 does not interact (or has a very low affinity) with cellobiose and displays very low affinity (most likely unspecific) for laminarihexaose. Moreover, experiments with cellotetraose and cellohexaose show that the protein interacts more strongly with the central glucose-units, mainly through interactions with positions 2 and 6 of the sugar units. In order to identify the residues of the proteins responsible for recognition and binding, I titrated the protein with several ligands and followed the variations in the amide chemical shifts by NMR. This allowed pinpointing the residues involved in ligand recognition and identifying key features in ligand recognition. This information was complemented with docking and molecular dynamics studies that gave localized structural information on the pocket site of CtCBM11. Furthermore, I have also studied the influence of temperature and binding in the structure of the protein by analyzing the backbone dynamics of CtCBM11 and amide exchange rates in the presence and absence of ligand and at 25 and 50°C. ¹⁵N longitudinal relaxation rates R_1 , transverse relaxation rates, R_2 , and steady state-state heteronuclear {1H}-

* In this context, epitopes are the atoms of the ligand that are closer to the protein when the complex is formed.

^{15}N - NOEs have been determined and analyzed in terms of the model-free formalism of molecular dynamics, using both isotropic and axially symmetric diffusion of the molecule, to determine the overall rotational correlation time (τ_m), the generalized order parameter (S^2), the effective correlation time for internal motions (τ_e), and amide exchange broadening contributions (R_{ex}) for each residue.

The results presented allow a better understanding, at the molecular level, of the interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their substrates.

III.2 Results and Discussion

III.2.1 Characterization of the sugars

Prior to the identification of the atoms of the ligand closer to the protein upon binding to the protein (epitope mapping) by NMR it is necessary to assign all the resonances of the different ligands so that I can later epitope map them. The assigned proton spectra of the select sugars as well as their structures are represented in **Figure III.2** to **Figure III.5**. When assigning the resonances of these sugars it is fundamental to have in mind that there is directionality in the chains as both extremities are different: there is a reducing end and a non-reducing end. Furthermore, the reducing end can exist in two conformations - α or β conformation. The designation ' α -' means that the hydroxyl group attached to C1 and the $-\text{CH}_2\text{OH}$ group at C5 lies on opposite sides of the ring's plane (a *trans* arrangement), while ' β -' means that they are on the same side of the plane (a *cis* arrangement). The α and β conformations exist in an approximately 40:60 ratio.¹¹

The assignment of the ^1H and ^{13}C NMR spectra was achieved through the analysis of the ^1H , ^{13}C , COSY, HSQC, HSQC-TOCSY and 1D selTOCSY spectra and the paper by Sugiyama *et al*¹² (see *Materials and methods*, Section II.4.4.2).

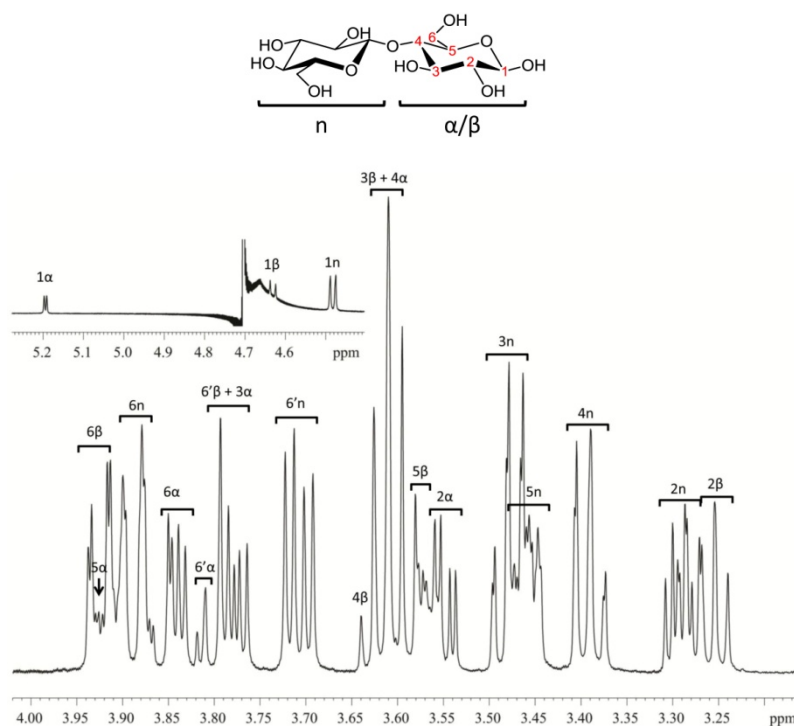


Figure III.2: Structure and ^1H spectra of cellobiose.

The spectrum was acquired with 1 mM solutions (100% D_2O) at 600 MHz at 298 K with 32 scans.

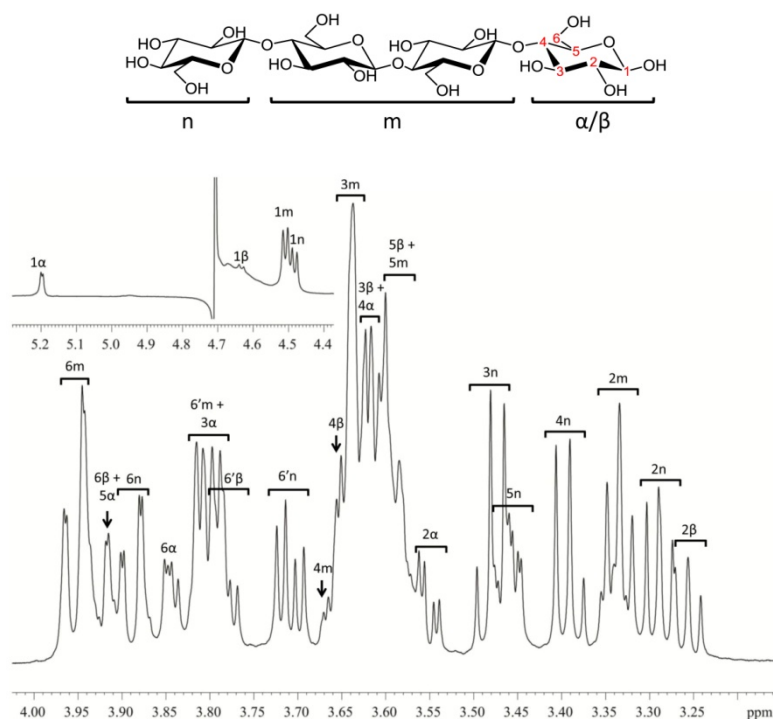


Figure III.3: Structure and ^1H spectra of cellotetraose.

The spectrum was acquired with 1 mM solutions (100% D_2O) at 600 MHz at 298 K with 32 scans.

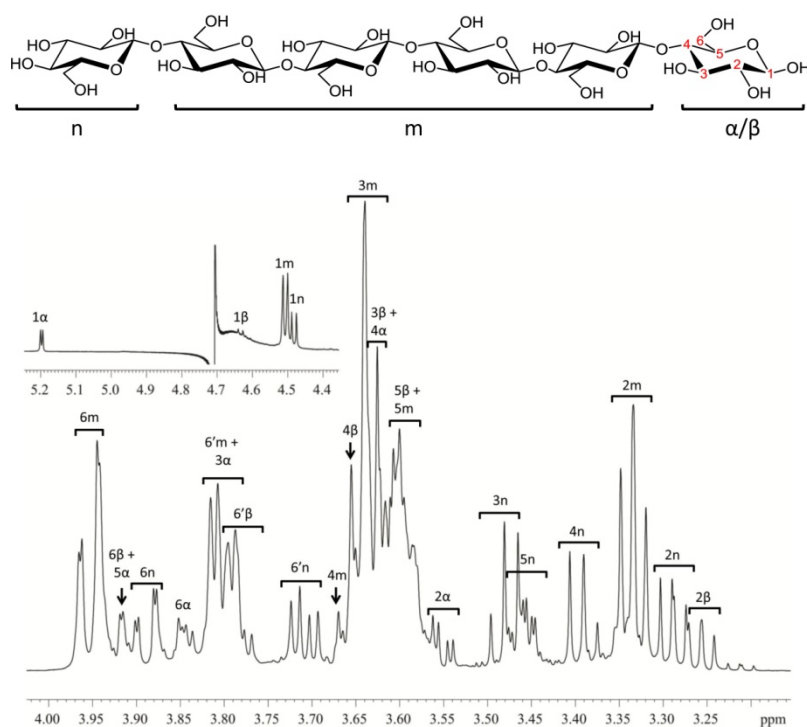


Figure III.4: Structure and ^1H spectra of cellohexaose.

The spectrum was acquired with 1 mM solutions (100% D_2O) at 600 MHz at 298 K with 32 scans.

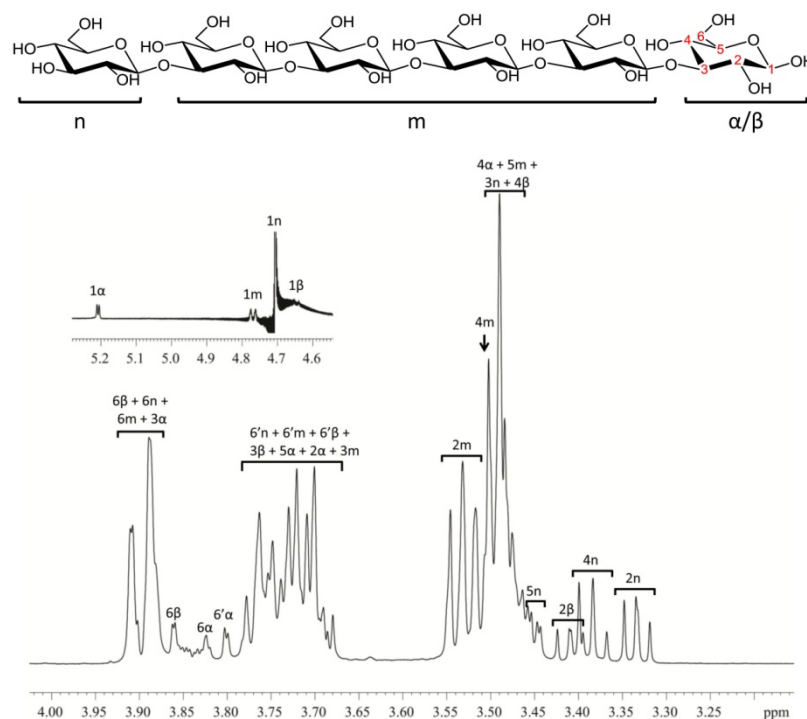


Figure III.5: Structure and ^1H spectra of laminarihexaose.

The spectrum was acquired with 1 mM solutions (100% D_2O) at 600 MHz at 298 K with 32 scans.

The complete ^1H and ^{13}C resonance assignment of cellobiose, cellotetraose, cellohexaose and laminarihexaose is summarized in **Table III.3** and **Table III.4**.

Table III.3: ^1H chemical shifts of cellobiose, cellotetraose, cellohexaose and laminarihexaose in D_2O .

	1	2	3	4	5	6	6'
Cellobiose							
α	5.20 (3.8)	3.55 (3.9, 9.8)	3.80 (9.6)	3.61	3.92	3.85	3.82 (5.3)
β	4.63	3.25 (8.6)	3.61 (9.3)	3.64	3.58	3.93 (2.2, 12.2)	3.78 (5.1, 12.3)
n	4.48 (8.6)	3.29	3.48	3.39	3.45	3.89 (12.0)	3.71 (5.9, 12.5)
Cellotetraose							
α	5.20 (3.8)	3.55 (3.8, 9.8)	3.80 (9.5)	3.62 (9.5)	3.92 (9.7)	3.85	3.83
β	4.63 (8.0)	3.26 (8.7)	3.61 (9.6)	3.65	3.57	3.93 (11.0)	3.78 (5.0, 12.2)
m	4.51 (8.0)	3.33 (8.6)	3.64 (8.4)	3.68	3.60	3.95 (2.0, 12.3)	3.80 (5.0, 12.4)
n	4.48 (8.0)	3.29 (8.7)	3.48 (9.1)	3.39 (9.5)	3.44	3.89 (12.3)	3.71 (5.9, 12.4)
Cellohexaose							
α	5.20 (3.8)	3.55 (4.0, 9.7)	3.80 (9.5)	3.62 (9.5)	3.92 (10.1)	3.85	3.83
β	4.63 (7.9)	3.26 (8.6)	3.62 (9.4)	3.65	3.57	3.93 (11.4)	3.78 (5.1, 12.2)
m	4.51 (7.9)	3.33 (8.4)	3.64	3.66	3.59	3.95 (10.9)	3.80 (4.8, 12.5)
n	4.48 (7.9)	3.29 (9.0)	3.48 (9.1) <i>t</i>	3.40 (9.5)	3.45	3.89 (10.9)	3.71 (6.1, 12.4)
Laminarihexaose							
α	5.21(3.8)	3.70 (3.7, 9.7)	3.89 (9.3)	3.49 (9.5)	3.75 (4.9, 12.3)	3.84	3.80
β	4.65 (8.3)	3.41 (8.7)	3.71	3.49 (9.0)	3.46	3.87 (10.3)	3.71
m	4.77 (8.1)	3.53 (8.4)	3.76	3.50	3.49	3.90 (11.0)	3.72 (5.2, 11.6)
n	4.73	3.33 (8.7)	3.50	3.38 (9.5)	3.46 (4.1, 10.4)	3.89 (11.5)	3.70 (5.3, 12.3)

Table III.4: ^{13}C chemical shifts of cellobiose, cellotetraose, cellohexaose and laminarihexaose in D_2O .

	1	2	3	4	5	6
Cellobiose						
α	91.80	71.21	71.32	78.62	70.09	59.85
β	95.73	73.87	74.27	78.63	74.78	60.06
n	102.53	73.16	75.49	69.46	75.97	60.56
Cellotetraose						
α	91.72	71.17	71.34	78.29	70.04	59.85
β	95.76	73.92	74.09	78.29	74.73	59.85
m	102.24	72.95	74.09	78.29	74.73	59.85
n	102.58	73.12	75.38	69.39	75.87	60.50
Cellohexaose						
α	91.88	71.17	71.34	78.29	70.04	59.85
β	95.77	73.92	73.92	78.29	74.73	59.85

<i>m</i>	102.24	72.90	73.93	78.24	74.74	59.85
<i>n</i>	102.53	73.10	75.40	69.39	75.90	60.50
Laminarihexaose						
<i>α</i>	92.04	71.01	82.17	68.16	73.28	60.50
<i>β</i>	95.60	73.76	84.44	68.16	75.99	60.66
<i>m</i>	102.40	73.28	84.12	68.13	75.54	60.66
<i>n</i>	102.73	73.44	68.10	69.56	75.99	60.66

III.2.2 Molecular determinants of ligand specificity

The strategy followed in order to understand how these proteins distinguish and select their substrates includes two complementary ways: (i) one focused on the structure of the ligand and the atoms responsible for binding to the proteins, (ii) and the other focused in the identification of the protein residues responsible for ligand recognition. Concerning the first approach, I have applied several techniques that could give insight about the atoms of the ligand that were in close contact with the protein upon binding. As I had the crystals of the protein without the histidine tag (were the binding cleft was not occupied by the C-terminal tail of a symmetry related molecule), I first tried to obtain co-crystals of the protein with ligands of interest (Section III.2.2.1). Due to negative results¹ I used NMR to identify and map the ligand epitopes^{1,2}. For this purpose linebroadening studies¹ (Section III.2.2.3), saturation transfer difference NMR (STD-NMR)^{1,2} (Section III.2.2.4) and diffusion ordered spectroscopy (DOSY)¹ (Section III.2.2.5) were applied. The interaction between CtCBM11 and cellobiose, cellotetraose, cellohexaose was used as a model to study the interaction between the protein and cellulose and accessing the influence of the length of the polysaccharide chain. Laminarihexaose was used to infer about the specificity of CtCBM11.²

NMR was also the tool chosen for tackling the second approach - *identification of the protein residues responsible for ligand recognition*. In this sense I studied the interaction between CtCBM11 and cellohexaose and cellotetraose by titrating ¹³C-¹⁵N- labeled CtCBM11 with the ligands and following the chemical shift perturbations by NMR (Section III.2.2.5). I have also studied the influence of temperature in binding by performing the titrations at 25 and 50 °C. Using either the crystallographic structure of CtCBM11 or NMR solution structures obtained (Chapter II) and the data derived from STD-NMR and titration studies I have calculated computational models of the CtCBM11-cellobiose, CtCBM11-cellotetraose CtCBM11-cellohexaose complexes (see Section III.2.2.6). Experimental details of all the techniques applied are explained in Materials and methods, Section III.4.

III.2.2.1 Co-crystallization studies

Since the crystals of the protein with the histidine tail had the binding site occupied with the *C-terminus* residues of a symmetry related molecule (**Figure III.1**), thus preventing the attempts to incubate the protein crystals with ligands of interest, I attempted to co-crystallize the protein without the histidine tail with cellohexaose. The first attempts were done under the conditions previously established³ and in which crystals were already obtained, but there were no positive results. Thus I tested new crystallization conditions (*see Appendix B, Table B.1*). Of the 80 crystallization conditions and different temperatures (4 and 20 °C) tested none produced positive results. The results obtained were mainly precipitate.¹ As seen in the previous chapter, the smaller size of the cleft in the crystal structure, probably imposed by the crystal packing, may be the cause for the failed co-crystallization attempts with different celooligosaccharides.

III.2.2.2 Influence of calcium in the structure of cellohexaose

As seen in Chapter II, CtCBM11 has two calcium-binding sites (similar to what happens with other CBMs). These calcium ions are thought to have a structural role, helping stabilizing the tertiary structure of the protein.³ Nonetheless, in some CBM families (for instance the family 36 CBM from *Paenibacillus polymyxa*¹³ or the family 35 CBM of the *Cellvibrio japonicus*¹⁴) the carbohydrate recognition is calcium-dependent. Despite in CtCBM11 the two calcium-binding sites are distant from the ligand binding site, it is known that calcium may alter the conformation of carbohydrates.¹⁵ Therefore, I wanted to check if the presence of calcium ions would affect the conformation of cellohexaose. For that I titrated a solution of cellohexaose with calcium chloride (CaCl₂) and followed the titration by ¹H-NMR (**Figure III.6**). The data shows that calcium does not interact with cellohexaose as the linewidth of the signals is not altered (*see Chapter VII, Section VII.2.2.3*). Only for very high concentrations of calcium (6 equivalents - **Figure III.6 - F**) I started to see some broadening of the signals of cellohexaose meaning that, at this concentration the calcium may be interacting with the sugar. Nonetheless, as this only happens for very high concentrations it is safe to say that calcium does not influence ligand binding and has only a structural role.

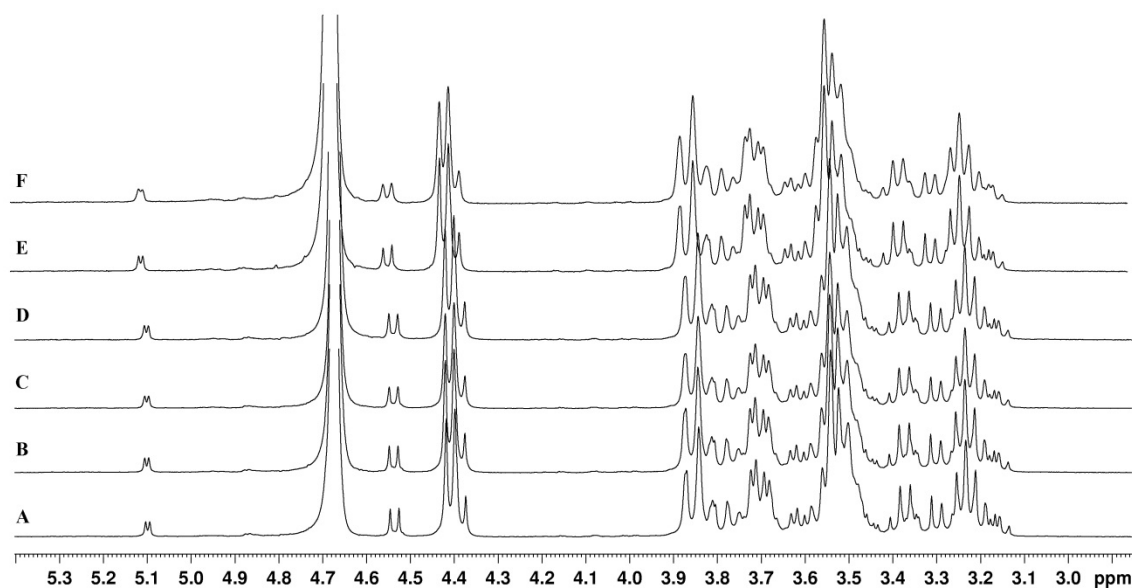


Figure III.6: Titration of cellohexaose with CaCl_2 .

A) Reference spectrum of 4 mM cellohexaose; B to F) 0.5, 1.0, 2.0, 3.0 and 6.0 equivalents of calcium, respectively.

III.2.2.3 Linebroadening studies

The simple measure or estimation of line widths may serve as a basis to deduce the occurrence of binding or recognition (*see Chapter VII, Section VII.2.2.3*). Since the relaxation properties of the oligosaccharides will be affected upon protein binding due to their dependence on molecular motion, I have studied the linebroadening effects (related to transverse relaxation - T_2) of cellohexaose resonances upon addition of CtCBM11¹. The spectra were acquired at 298 K in a Bruker ARX spectrometer, operating at a frequency of 400 MHz (*see Materials and methods, Section III.4.4.4*).

In general, a progressively line broadening of all the cellohexaose protons was observed during titration with increased amounts of protein, which can be understood as a result of loss of local mobility caused by the binding of the sugar to the protein. Chemical shifts are only slightly affected suggesting fast equilibrium between free ligand and protein bound forms. The cellohexaose proton resonances can be identified in **Figure III.4**. A detailed comparison of the cellohexaose spectra showed that the most significant linebroadening was observed for protons 6 and 2, from the central glucose units (**Figure III.7**) indicating that the corresponding hydroxyl groups are involved in protein binding.

The results for the linebroadening measurements of anomeric proton of the reducing end in the *alpha* and *beta* configurations, H1 α and H1 β , plotted in **Figure III.7-I** and **IV**, showed that these protons are hardly affected by protein binding, as would be expected for protons on the terminal end of the sugar, located out of the binding cavity. However, for H1 β a slight effect can

be detected when compared to H1 α , which can be indicative of a higher affinity of the protein for the β form. Furthermore, proton 4 from subunit n (non-reducing end) also shows a significant broadening (Table III.5). This indicates that, although the non-reducing end lay outside the binding cleft, some contacts with the protein may occur that restrict its mobility. Moreover, the overall loss of mobility of the whole cellobiose molecule will also lead to a general broadening of all resonances.

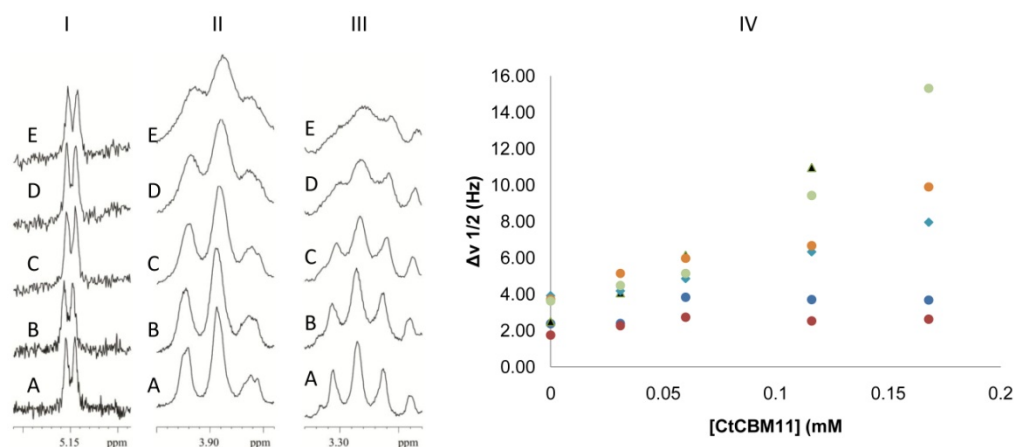


Figure III.7: Line broadening studies.¹

I, II, and III - series of spectral regions of a solution of cellobiose 0.80 mM in D₂O, corresponding to protons α H1, H2m and H6m, respectively, acquired at 298K as a function of peptide (CtCBM11) concentration (A = 0.0 mM, B = 0.031 mM, C = 0.060 mM, D = 0.116 mM and E = 0.168 mM). **IV** - Linewidths ($\Delta\nu_{1/2}$) of selected cellobiose protons, determined after spectral deconvolution, as a function of peptide (CtCBM11) concentration: ● - H1 α , ▲ - H1m, ● - H1 β , ● - H2m, ◆ - H6m, ● - H6'm+6'β.

Table III.5: Linewidths at half-height for the different protons of cellobiose during the titration experiment.

[CBM11] (mM)	Proton							
	H1 α	H1 β	H1m	H1n	H6m	H6'm+6'β	H3n	H4n
0.000	1.75	2.38	2.50	2.06	3.91	3.72	1.87	2.22
0.031	2.27	2.39	4.08	2.46	4.18	5.14	3.49	2.81
0.060	2.73	3.83	6.15	3.64	4.86	5.97	3.46	3.18
0.116	2.53	3.70	10.97	6.14	6.33	6.67	5.32	3.87
0.168	2.62	3.67	-	-	7.95	9.90	6.24	4.40

III.2.2.3 Saturation transfer difference NMR (STD-NMR)

In order to understand how CtCBM11 distinguishes and selects the different ligands it is extremely important to identify which atoms of the ligand are closer to the protein when the

complex is formed (epitope mapping). Identification and mapping of the epitopes was achieved using a NMR technique, known as Saturation Transfer Difference (STD-NMR [*see Chapter VII, Section VII.5.1*]). The ability to detect binding of low molecular weight compounds to large biomolecules using the STD-NMR technique has already been demonstrated.^{1,2,16-18} This technique offers several advantages over other methods to detect binding activity:

1. The binding component can usually be directly identified, even from a substance mixture, allowing it to be utilized in screening for ligands with dissociation constants K_d ranging from ca. 10^{-3} to 10^{-8} M.
2. The atoms of the ligand having the strongest contact to the protein show the most intense NMR signals, enabling the mapping of the ligand's binding epitope.
3. Very important for a NMR-based detection system, its high sensitivity allows using as little as 1 nmol of protein with a molecular weight >10 kDa.¹⁶

STD-NMR spectroscopy was applied to analyze the binding of cellobiose (**Figure III.8**), cellotetraose (**Figure III.9**), cellohexaose (**Figure III.10**) and laminarihexaose (**Figure III.11**) to CtCBM11. All the spectra were acquired at 298 K in a Bruker AvanceIII spectrometer, operating at a frequency of 600 MHz with a 100-fold excess of ligand over the protein (*see Materials and methods, Section III.4.4.5*).

The STD-NMR spectrum of cellobiose is presented in **Figure III.8**, along with the sugar's reference spectrum. The absence of signals in the STD-NMR spectrum is a clear indication that either there is no interaction between CtCBM11 and cellobiose or it is very weak. These results are in accordance with previous data^{3,4} where the ITC-determined affinity constant (K_a) was reported to be around $1.3 \times 10^3 \text{ M}^{-1}$, which is in the lower limit of STD detection capabilities (10^3 to 10^8 M^{-1}).¹⁸

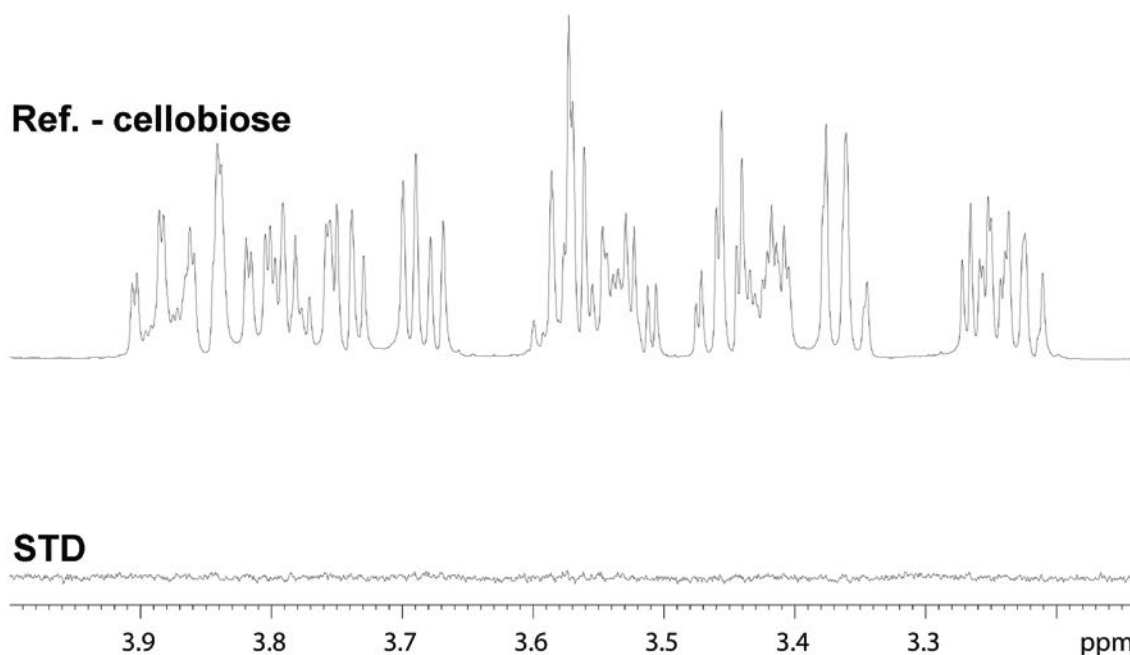


Figure III.8: STD-NMR of cellobiose with CtCBM11.

Top - Reference ^1H -NMR cellobiose spectrum. Bottom - STD-NMR spectra of the solution of cellobiose (2 mM) with the protein (20 μM). No signals appear in the STD-NMR spectrum, indicating that either there is no interaction between cellobiose and CtCBM11 or it has a very low affinity.

Unlike cellobiose, the STD-NMR spectrum with cellotetraose clearly shows some signals (**Figure III.9**). This is a clear indication that CtCBM11 binds to this ligand. Moreover, comparison of the reference with the STD-NMR spectrum shows that the relative intensity of the peaks is different, therefore allowing to epitope-map the ligand. The binding epitope is created by the comparison of the STD intensity relative to the reference one and this is described by the STD amplification factor (A_{STD}) shown in **Equation III.1** (see also *Materials and methods*, Section III.4.4.5).

$$A_{\text{STD}} = \frac{I_0 - I_{\text{SAT}}}{I_0} \times \text{ligand excess} = \frac{I_{\text{STD}}}{I_0} \times \text{ligand excess}$$

III.1

were A_{STD} is the STD amplification factor, I_0 , I_{SAT} and I_{STD} are the intensities of the reference (off resonance), saturated (on resonance) and difference (STD-NMR) respectively. The differences in A_{STD} for the different protons can be quantitatively expressed by analyzing the relative STD effects at a given saturation time - epitope mapping of the ligand. Provided that all the ligand protons have similar relaxation rates, then the differences in the relative STD response (I_{STD}/I_0 or A_{STD}) reflect the relative proximity of that proton to the receptor binding site. The procedure is simple, for a given saturation time the relative STD (or A_{STD}) with the

highest intensity is set to 100 %, and all other STD signals are calculated accordingly. **Table III.6** shows the calculated A_{STD} values and the epitope mapping of all possible protons.

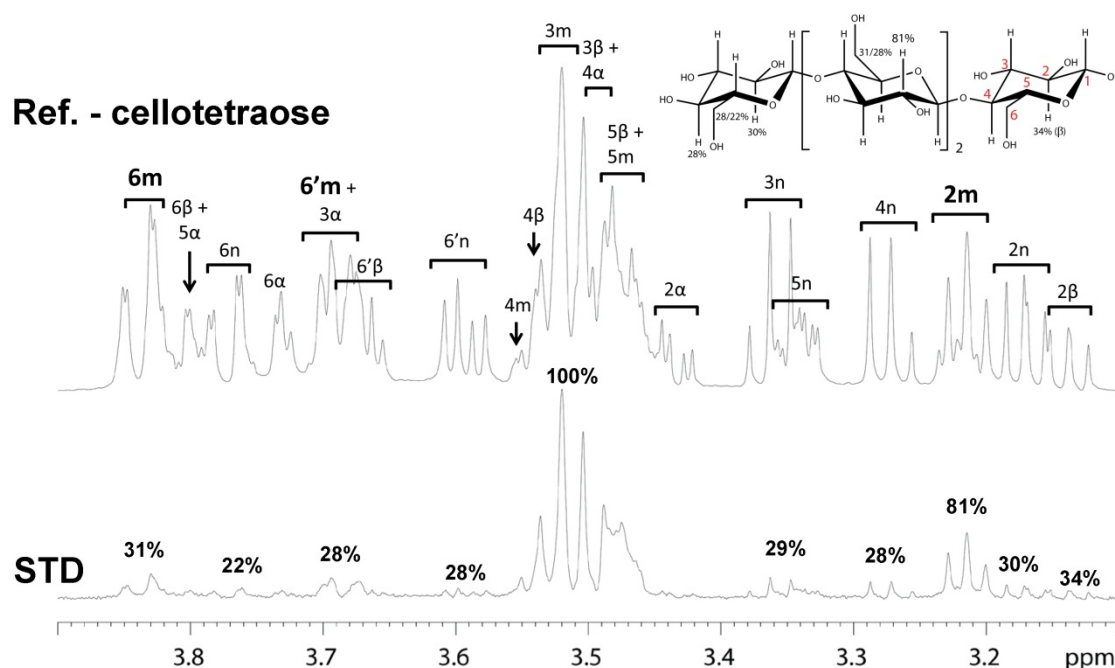


Figure III.9: STD-NMR and epitope mapping of cellotetraose bound to CtCBM11.

Top - Reference $^1\text{H-NMR}$ cellotetraose spectrum. Bottom - STD-NMR spectra of the solution of cellotetraose (2 mM) with the protein (20 μM). The binding epitope for the interaction of cellotetraose with CtCBM11 is shown above each peak and mapped in the structure of the sugar.

For cellotetraose the maximum intensity is found for the peaks in the region between 3.45 and 3.56 ppm. These peaks correspond to protons H4m, H4 β , H3m, H3 β , H4 α , H5 β and H5m and their higher intensity means that these protons, or at least some of them, are the ones closer to the protein upon complex formation. Unfortunately, due to signal overlapping it is not possible to distinguish the individual contributions. The other protons that show a high intensity are the ones bound to C2 in the central glucose units (H2m) with 81% relative intensity. This indicates that these protons are also very close to the protein when the complex is formed and may be key for binding and recognition. All other protons have relative intensities around 30%, meaning that they are more distant from the protein when the complex is formed. In general all glucose units show some degree of saturation indicating that the whole molecule is in contact with CtCBM11. This is in good agreement with previous data that showed that the binding cleft of this protein can accommodate at least 4 sugar units.³ The STD epitope map of cellotetraose upon binding to CtCBM11 is shown in **Figure III.9** and summarized in **Table III.6**.

Regarding the interaction of cellohexaose with CtCBM11 (**Figure III.10**), it can be seen that it is very similar to the one with cellotetraose.

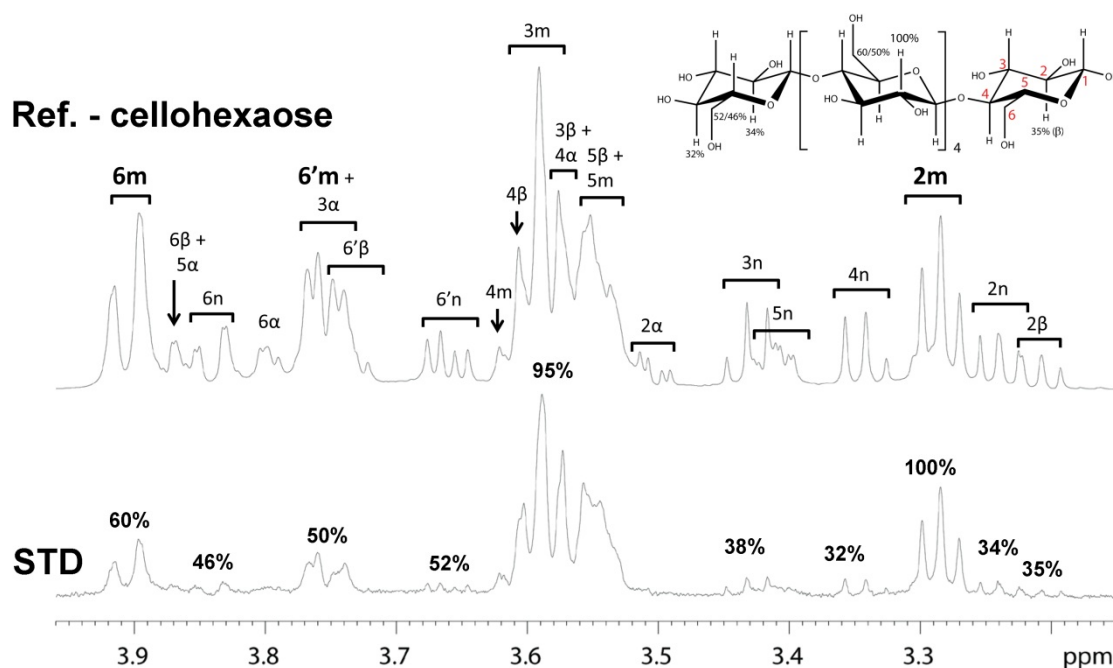


Figure III.10: STD-NMR and epitope mapping of cellohexaose¹ bound to CtCBM11.

Top - Reference ¹H-NMR cellohexaose spectrum. Bottom - STD-NMR spectra of the solution of cellohexaose (2 mM) with the protein (20 μM). The binding epitope for the interaction of cellotetraose with CtCBM11 is shown above each peak and mapped in the structure of the sugar.

Comparison of the reference and STD-NMR spectra clearly shows that the residues of the hexasaccharide are differently involved in binding. It can be seen from **Figure III.10** that the more intense signals are those corresponding to H2 from central glucose units (H2m) indicating that, when the complex is formed, these protons are the ones closer to the protein. As in the case of cellotetraose, the signals located at the central region of the spectrum (H4m, H4β, H3m, H3β, H4α, H5β and H5m) also show a very high degree of saturation (95%), again indicating that at least some of them are close to the protein upon complex formation. Due to signal overlapping it is not possible to distinguish the individual contributions of these protons. Additionally protons from the methylene groups (H6 and H6'), particularly the ones of the central glucose units (H6m and H6'm), also display a relative high degree of saturation (60 and 50%, respectively). The fact that one of the diastereotopic protons from the methylene groups shows a relative more intense peak in the STD spectrum is indicative of a precise orientation of the methylene groups upon binding to the protein. With respect to reducing and non-reducing ends (α/β and n, respectively), the observed signals in the STD-NMR spectrum show that they should not contribute significantly to the binding as the relative degrees of saturation are low (**Table III.6**). Nonetheless, some contact still exists between the protein and the extremities of the hexasaccharide. These contacts occur with all protons of the non-reducing end and with

protons H2 and H4 of the reducing end and may be responsible for stabilizing the complex as the extremities of cellohexaose lay outside the binding cleft. In the absence of these relatively weak contacts the entropy of the cellohexaose molecule could lead to a decrease in the affinity.

ITC studies³ (**Table III.1**) showed that the affinity of CtCBM11 for cellohexaose is higher than for cellotetraose (~2-fold). The possible mechanism for the tighter binding of ligands that extend beyond the hydrophobic platform may be related to the more extended interchain hydrogen bonding network afforded by these longer ligands that stabilizes the conformation adopted by the oligosaccharide in the binding cleft.¹⁹ Alternatively, the flexible anomeric configuration adopted by the O1 of the reducing end glucose may reduce binding affinity, and thus these CBMs bind optimally to internal regions of glucan chains. These results indicate that the binding cleft of CtCBM11 interacts more strongly with the central glucose-units, mainly through interactions with position 2 and 6 of the sugar units, which is consistent with the ligands accommodated by other Type B CBMs.^{8,20-22}

Table III.6: Amplification factors and epitope mapping for the interaction between CtCBM11 and cellotetraose and cellohexaose.

<i>A_{STD}</i> / Epitope mapping (%)							
	1	2	3	4	5	6	6'
CtCBM11/Cellotetraose							
<i>α</i>	-	-	0.67 / 28 ^c	2.40 / 100 ^b	-	0.47 / 19	-
<i>β</i>	-	0.81 / 34	2.40 / 100 ^b	2.40 / 100 ^b	2.40 / 100 ^b	-	0.67 / 28 ^c
<i>m</i>	-	1.94 / 81	2.40 / 100 ^b	2.40 / 100 ^b	2.40 / 100 ^b	0.75 / 31	0.67 / 28 ^c
<i>n</i>	-	0.72 / 30	0.70 / 29 ^a	0.67 / 28	0.70 / 29 ^a	0.54 / 22	0.68 / 28
CtCBM11/Cellohexaose							
<i>α</i>	-	-	0.89 / 50 ^f	1.85 / 95 ^e	-	-	-
<i>β</i>	-	0.63 / 35	1.85 / 95 ^e	1.85 / 95 ^e	1.85 / 95 ^e	-	0.89 / 50 ^f
<i>m</i>	-	1.79 / 100	1.85 / 95 ^e	1.85 / 95 ^e	1.85 / 95 ^e	1.07 / 60	0.89 / 50 ^f
<i>n</i>	-	0.61 / 34	0.69 / 38 ^d	0.58 / 32	0.69 / 38 ^d	0.83 / 46	0.92 / 52

a, b, c, d, e, f – These peaks are overlapped

Regarding the STD-NMR results with laminarihexaose (**Figure III.11**), because previous studies indicated that CtCBM11 didn't bind to β -1,3 linked glucans (as is the case of laminarihexaose)³, no signals were expected. Nonetheless, as seen in **Figure III.11**, some signals (although very weak) appear in the STD-NMR spectrum, indicating some degree of interaction may occur, despite being possibly non-specific. The low *A_{STD}* values determined for laminarihexaose (0.15 for proton H2n, 0.22 for proton H4n, 0.43 for proton H2m, 0.38 for protons H6'n, H6'm, H6' β , H3 β , H5 α , H2 α and H3m and 0.41 for protons H6 β , H6n, H6m and H3 α) are a good indication of this low affinity interaction. The affinity of CtCBM11 for several ligands, including laminarin, was previously determined by affinity gel electrophoresis

(AGE).^{3,4} In these studies it was shown that CtCBM11 displays the highest affinity for β -1,3-1,4-mixed glucans while exhibiting significantly weaker binding to hydroxyethyl cellulose, glucomannan and oat spelt xylan and no affinity for arabinan, galactomannan, laminarin, rhamnogalacturan, glucuronoxylan and or rye-arabinoxylan, which contrasts with the results obtained by STD-NMR. The range of association constants that can be determined by affinity gel electrophoresis goes from about 10^2 to 10^5 M^{-1} ,²³ which, in principle, should be enough to detect the binding of laminarihexaose as it was detected by STD-NMR whose detection interval ranges from 10^3 to 10^8 M^{-1} . Nevertheless, the lower limit of AGE is determined by the concentration of ligand in the gel and by the ability to measure small migration changes.²³ For low affinity ligands, the mobility of the protein won't be as affected if not enough ligand is in the gel. In order to detect this type of binding an increase in the ligand concentration in the gel is needed.²³ Therefore, the fact that no binding was detected for laminarin may only indicate that its affinity is too low for AGE detection in the conditions used. My results show that, though CtCBM11 is not specific to β -1,3-linked saccharides, it still retains some activity towards laminarihexaose. Whether this low affinity has a biological meaning or not is still unknown.

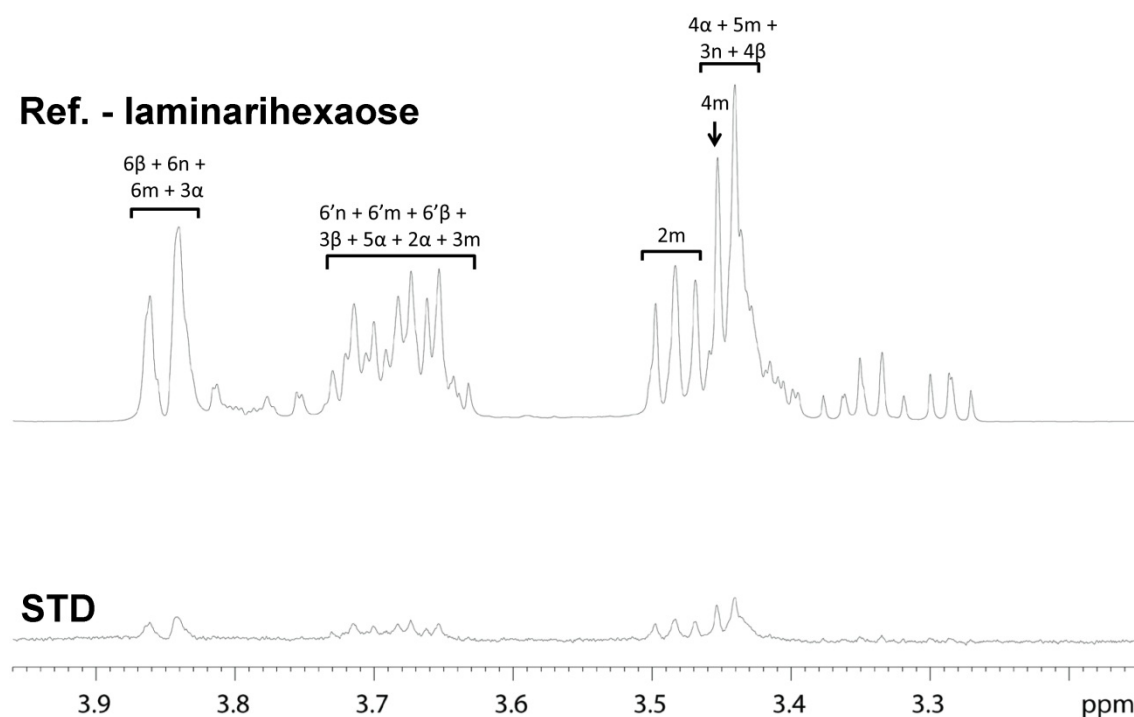


Figure III.11: STD-NMR of laminarihexaose with CtCBM11.

Top - Reference 1H -NMR laminarihexaose spectrum. Bottom - STD-NMR spectra of the solution of laminarihexaose (2 mM) with the protein (20 μM). Despite previous studies indicated that CtCBM11 didn't bind to β -1,3-linked glucans, some signals appear in the STD-NMR spectrum, indicating some degree of interaction.

III.2.2.4 Diffusion studies (DOSY)

Another way to study molecular interaction in solution is through the NMR technique, known as Diffusion Ordered Spectroscopy, DOSY (see Chapter VII – Section VII.5.2). The DOSY technique aims identifying the molecular components of a mixture acquiring, at the same time, information on their size and is based on the self-diffusion coefficient.^{2,24,25} Self-diffusion is the random translational motion of molecules driven by their internal kinetic energy.²⁴ Self-diffusion coefficients and the structural properties of a molecule are connected by the dependence of the self-diffusion coefficients on molecular size and shape. Therefore, it is not surprising that the determination of molecular self-diffusion coefficients by NMR has become a valuable methodology for studies of molecular interaction in solution. The concept behind the application of diffusion NMR techniques for binding and screening studies is very simple and is based on the fact that the diffusion coefficient of a small molecule is altered upon binding to a large receptor.

With this experiment I intended to determine the association constant (K_d) for the cellohexaose/CtCBM11 interaction and to confirm if binding of laminarihexaose to CtCBM11 could be detected by DOSY. **Figure III.12** shows the DOSY spectrum of the mixture of cellohexaose and laminarihexaose before adding the protein (A) and after (B).

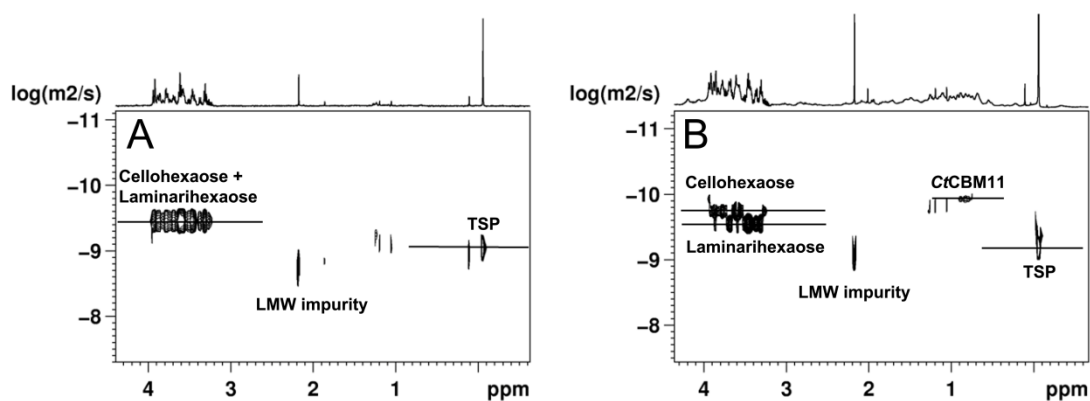


Figure III.12: DOSY spectra for the calculation of the association constant for the cellohexaose/CtCBM11 interaction.²

A) DOSY spectrum from the mixture of cellohexaose and laminarihexaose, 40 μM in D_2O with TSP, B) DOSY spectrum from the mixture of cellohexaose, laminarihexaose and CtCBM11, 40 μM in D_2O with TSP. The spectra were acquired in a Bruker Avance II 600 MHz spectrometer, at 298K, with 512 scans in 32 steps and a spectral width of 12376 Hz in the direct dimension centered in the solvent frequency. The duration of the encoding/decoding gradient was 1.5 ms in A and 1.1 ms in B. The diffusion time was 400 ms in A and 800 ms in B. LMW: Low Molecular Weight.

From these results it is possible to say, only by direct observation of the DOSY spectra, that there is an interaction between cellohexaose and the protein whereas laminarihexaose does not interact (the diffusion coefficient of cellohexaose decreases when the protein is added to the

mixture of sugars and the one from laminarihexaose remains the same). This is in good agreement with the STD-NMR results and confirms that binding of laminarihexaose to CtCBM11 is non-specific. The protein and carbohydrate diffusion coefficient values are listed in **Table III.7** and were extracted directly using the variable gradient fitting routines in Bruker TopSpin2.2 software.

Table III.7: Self diffusion coefficients measured for the mixture of sugars with and without the protein.

<i>Self-Diffusion Coefficients, D (m²/s)</i>	
Sugar sample	
Cellohexaose	3.55×10^{-10}
Laminarihexaose	3.55×10^{-10}
TSP	8.85×10^{-10}
Mixture sample	
Cellohexaose	1.82×10^{-10}
Laminarihexaose	2.82×10^{-10}
CtCBM11	1.15×10^{-10}
TSP	6.52×10^{-10}

Using **Equation III.6** (see *Material and methods - Section III.4.4.6*) and the data in **Table III.7** I was able to calculate the association constant for the binding of cellohexaose to CtCBM11: $K_a = 6.33 \times 10^4 \text{ M}^{-1}$. This result is in agreement with previous studies³ (**Table III.1**).

III.2.2.5 Interaction studies with cellooligosaccharides

Through linebroadening and STD-NMR studies I was able to identify the atoms of the ligands involved in binding and to distinguish between the ones closer to the protein when the complex is formed and the ones more distant. Nonetheless, so far I had no experimental information about the residues responsible for ligand binding and recognition.

In order to characterize the residues responsible for binding of CtCBM11 to cellooligosaccharides, I titrated a 0.1 mM sample of double-labeled protein with cellohexaose and cellotetraose and acquired a ¹⁵N-¹H-HSQC at each titration. Besides the length of the cellooligosaccharide chain, I have also studied the influence of temperature by performing the titrations at 25 and 50 °C (**Figure III.13**).

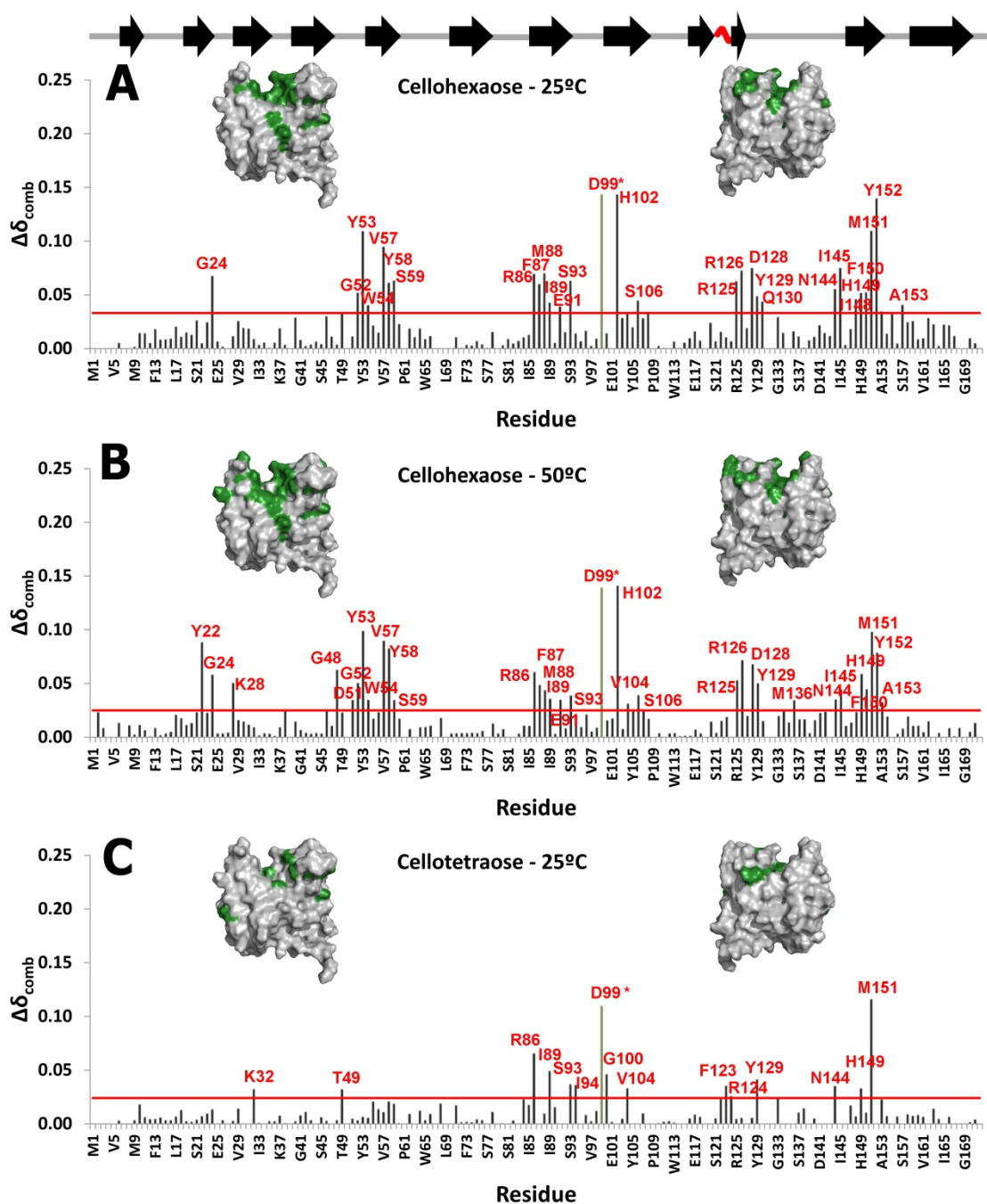


Figure III.13: Backbone amide chemical shift variations between *CtCBM11* and **A)** cellohexaose at 25°C; **B)** cellohexaose at 50 °C and **C)** cellotetraose at 25 °C.

Chemical shifts variations larger than the corrected standard deviation to zero²⁶ were considered as significant. Green bars mark residues that disappear during the titration. Above each plot is depicted the surface of the solution structure of *CtCBM11* in light grey with the residues that show significant chemical variations depicted in green.

Several protein protons substantially changed their chemical shifts upon addition of increasing amounts of cellohexaose and cellotetraose which allowed pinpointing of the binding cleft of *CtCBM11* (**Figure III.13**). In order to better represent the distribution of affected and

non-affected residues I have calculated the combined chemical shift perturbation, $\Delta\delta_{comb}$, and determined a cut-off line²⁶ (see *Materials and methods - Section III.4.4.8*).

The interaction with cellohexaose clearly shows that most changes occur for residues Tyr53-Ser59; Arg86-Ser93; Asp99-Ser106, Arg125-Tyr129, Ans144, Ile145 and His149-Ala153, independently of the temperature (**Figure III.13 - A and B**). Upon addition of only 0.3 equivalents of cellohexaose, the amide signals of residues Asp99 and Tyr152 disappear from the ¹H-¹⁵N-HSQC spectra, most probably due to conformational broadening, suggesting an important role in ligand binding/recognition.

The interaction with cellotetraose shows that this smaller ligand interacts with fewer residues of the protein and preferentially with one side of the binding cleft. Residues Arg86-Ile94, Asp99-Val104, Phe123-Tyr129, Ile145 and His149-Ala153 are the most affected by binding (**Figure III.13 - C**). Interestingly, although cellohexaose and cellotetraose share the same binding cleft, the interaction pattern is very distinctive. While cellohexaose interacts with both sides of the binding cleft, cellotetraose seems to interact preferentially with one side and, as I said above, with fewer residues. This difference is related to the smaller size of cellotetraose and is reflected in the affinity displayed towards the different ligands (**Table III.8**). Nonetheless, independently of the ligand and temperature, all resonances that undergo large chemical shift changes on binding are located in and around the putative binding cleft³ of CtCBM11 (**Figure III.13**), confirming this region as the binding site. In addition, several of the identified residues were already recognized by site directed mutagenesis³ (Tyr22, Tyr53 and Tyr129) and molecular docking studies¹ (Asp99, Arg126, Asp128 and Asp146) as key for the binding process.

The observed effects on the chemical shifts indicate that the interaction is fast in the NMR time scale. Thus, the alterations in chemical shifts can be used to determine the equilibrium association constants.^{26,27} From the titration data, I saw that Tyr129 interacts with both cellohexaose and cellotetraose and from previous mutation studies^{1,3} I knew that this residue is essential for ligand binding. Due to this fact and because Tyr129 NH resonance is fairly well resolved, I followed its chemical shift as a function of the concentration of ligand to obtain binding constants (**Table III.8**). The results yielded a K_a of $5.20 \pm 1.10 \times 10^4$ and $1.83 \pm 0.33 \times 10^4$ M⁻¹ for the interaction with cellohexaose at 25 and 50 °C, respectively. For the interaction with cellotetraose at 25 °C a K_a value of $2.33 \pm 0.56 \times 10^4$ M⁻¹ was obtained (**Table III.8**). A full list of the calculated affinity constants and thermodynamic parameters from the interaction of CtCBM11 with cellohexaose and cellotetraose is given in Appendix C, **Tables C1 and C2**, respectively. The determined K_a values for both ligands at 25 °C are in good agreement with previous ITC results³ as one can see in **Table III.8** and with the results obtained by DOSY (6.33×10^4 M⁻¹ for cellohexaose). The lower affinity of cellotetraose when compared to

cellohexaose is most likely due to the loss of several key contacts with the protein as seen from the titration experiments.

Table III.8: Quantitative assessment of CtCBM11 binding to cellohexaose and cellotetraose, using the NH resonance of Tyr129 as a probe.

	$K_a \times 10^4 (M^{-1})$	$\Delta G (kcal.mol^{-1})$ (25°C)	$\Delta H (kcal.mol^{-1})$	$T\Delta S (kcal.mol^{-1})$ (25°C)
Cellohexaose – 25 °C (NMR)	5.20±1.10	-6.43±0.22	-7.99±0.26	-1.57±0.01
Cellohexaose – 50 °C (NMR)	1.83±0.33			
Cellohexaose – 25 °C (ITC)³	7.8±0.1	-6.6±0.0	-9.5±0.2	-2.9±0.2
Cellotetraose – 25 °C (NMR)	2.33±0.56	-5.95±0.25	-	-
Cellotetraose – 25 °C (ITC)³	4.4±0.8	-6.3±0.1	-9.8±0.1	-3.5±0.1

The thermodynamic parameters, ΔH and ΔS of the residues involved in binding were calculated from the K_a values determined from the titration experiments using a van't Hoff plot of $\ln(K_a)$ vs. $1/T$. For the binding of cellohexaose to CtCBM11 a ΔH of -6.43 ± 0.22 kcal.mol⁻¹ and a binding entropy, $T\Delta S$, of -1.57 ± 0.01 kcal.mol⁻¹ (T=298K) were obtained.

The thermodynamic parameters of binding presented in **Table III.8** show that the K_a , ΔH and $T\Delta S$ values determined based on the chemical shift perturbation of Tyr129 are in good agreement with the literature values determined by ITC³. These values show that the association of CtCBM11 with cellohexaose is enthalpically driven (i.e., exothermic) with an unfavorable entropic contribution ($\Delta G = -6.43 \pm 0.22$, $\Delta H = -7.99 \pm 0.26$ and $T\Delta S = -1.57 \pm 0.01$ kcal.mol⁻¹). This is common to the majority of carbohydrate-binding modules²⁸. However, when considering the thermodynamic parameters determined with all the residues perturbed, we see that the ΔG value does not change considerably, but the entropy term becomes positive and the enthalpy less negative ($\Delta G = -5.95 \pm 0.62$, $\Delta H = -3.03 \pm 1.84$ and $T\Delta S = 2.92 \pm 0.01$ kcal.mol⁻¹). This raises the question about the individual contributors to the thermodynamic parameters, such as the role of favorable direct CBM-saccharide interactions, conformational rearrangements of the oligosaccharide, thermodynamic favorable structural rearrangements of the protein backbone, etc.

III.2.2.6 Computational studies

Since the X-ray structure of the CtCBM11 with a bound substrate was not available, it is difficult to evaluate the importance and function of each residue at the CtCBM11 cleft in the binding process of carbohydrates. Consequently, computational studies were used to deduce this

kind of information and complement the NMR studies. These studies provided localized structural information of the binding pocket of the CtCBM11 helping to interpret all the NMR data.

The first attempt to obtain the CtCBM11/ligand models was performed by my colleges at Faculdade de Ciências da Universidade do Porto (Dr. Natércia Brás, Prof. Nuno Cerqueira, Prof. Pedro Alexandrino Fernandes and Prof. Maria João Ramos).¹ In their calculations they used the crystal structure of the protein with the histidine tail (1v0a)³ instead of the one without the tag (*see Chapter II*) because the first was acquired at higher resolution and no significant structural differences are observed between the two. Moreover, at that time the NMR solution structure was not available. These studies were conducted using only the STD-NMR information.

Later, using the experimental information about the residues that are most affected by binding, together with the NMR solution structure of the protein, in combination with the previously obtained information obtained by STD-NMR concerning the ligand, I have recalculated a model of the CtCBM11-cellohexaose/cellotetraose complex. The two approaches are discussed below.

III.2.2.6.1 Docking experiments with the crystallographic structure

Calculations were performed with cellobiose, cellotetraose and cellohexaose. Moreover, for each ligand the α and β isomers were considered.¹ The ligands were built independently and the structure was optimized using the AMBER force field²⁹.

The first results that came from the initial simulations were quite disappointing since the conformations of some residues near the binding pocket, namely Tyr22, Tyr53, Tyr129 and Tyr152, gave rise to a steric obstacle, and were precluding an efficient binding of the ligands. To overcome this issue my colleagues used the software MADAMM³⁰ that allows a certain degree of protein flexibility in standard docking processes. The process tries to mimic a conformational binding model, in which the receptor is assumed to pre-exist in a number of energetically similar conformations. Accordingly, the ligand binds preferentially to one of these conformers displacing the equilibrium towards this particular conformer and increasing in this way its proportion relatively to the total protein population. In this study the flexibilization was applied to Tyr22, Tyr53, Tyr129 and Tyr152. At the end of this process a group of complexes was obtained, with optimized affinities between the CtCBM11 and each studied ligand. In order to refine these results, molecular dynamics simulations were performed on the best solution. This process was repeated for all the studied ligands, including the α and β isomers.

The simulations showed that all ligands have common binding poses at the CtCBM11 cavity, near the aromatic amino acids that were flexibilized. Furthermore, the ligands bind in an equidistant mode at the CtCBM11 cleft, suggesting an apparent symmetry at the binding cavity. Most of the interactions between the CtCBM11 cleft and each carbohydrate occur through hydrogen bonds, namely with the equatorial OH groups of the glucose monomers, and also by several van de Waals contacts that are promoted by the aliphatic side chains present at the interface, namely with, Tyr22, Tyr53, Tyr129 and Tyr152. The only exception was cellobiose that showed no specificity and different binding poses at the CtCBM11 cleft could be observed (**Figure III.14**). This is in agreement with the experimental work, where no specific interaction could be detected with this ligand (**Figure III.8**).

The docking results obtained with MADAMM, have also revealed there is no substantial differences between the α and β conformations of carbohydrates. However, in some carbohydrates, the C1 terminal of the α conformation is turned towards the left hand side of the binding cavity, whereas in the β conformation is in the opposite direction. Keeping in mind that the monomers that constitute the ligands are equal among themselves, this change in the orientation is not of great importance to the establishment of the binding interactions between the ligand and the CtCBM11, and this kind of behavior should occur commonly in nature.

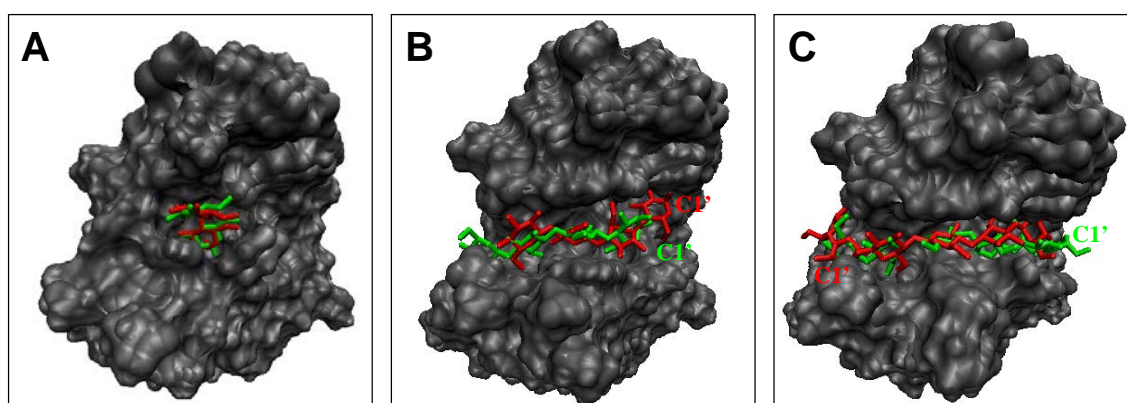


Figure III.14: Representation of the conformations of the three-dimensional structure of binding of the different ligands obtained by docking.

A) α - (red) and β -cellobiose (green); **B)** α - (red) and β -cellotetraose (green); **C)** α - (red) and β -cellotetraose (green).

From the studied carbohydrates, cellotetraose was the one that fitted perfectly inside the binding cleft of the CtCBM11. In the case of β -cellotetraose, the hydrogen bonds were established with the amino acids Glu25, Asp99, Arg126, Asp128, Asp146 and Ser147 (**Figure III.15**), that closely match the amino acids that interact with the α isomer, differing only in Glu25 residue. In the case of β -cellohexaose ligand the carbohydrate oligomer interacts mainly with the amino acids: Asp51, Trp54, Thr56, Gly96, Gly98, Asp99, Arg126, Asp128 and

Asp146. In the case of the α -isomer some hydrogen bonds with amino acids Tyr22, Thr50 and Ala153 can also be observed, but not with Trp54, Gly96 and Gly98.

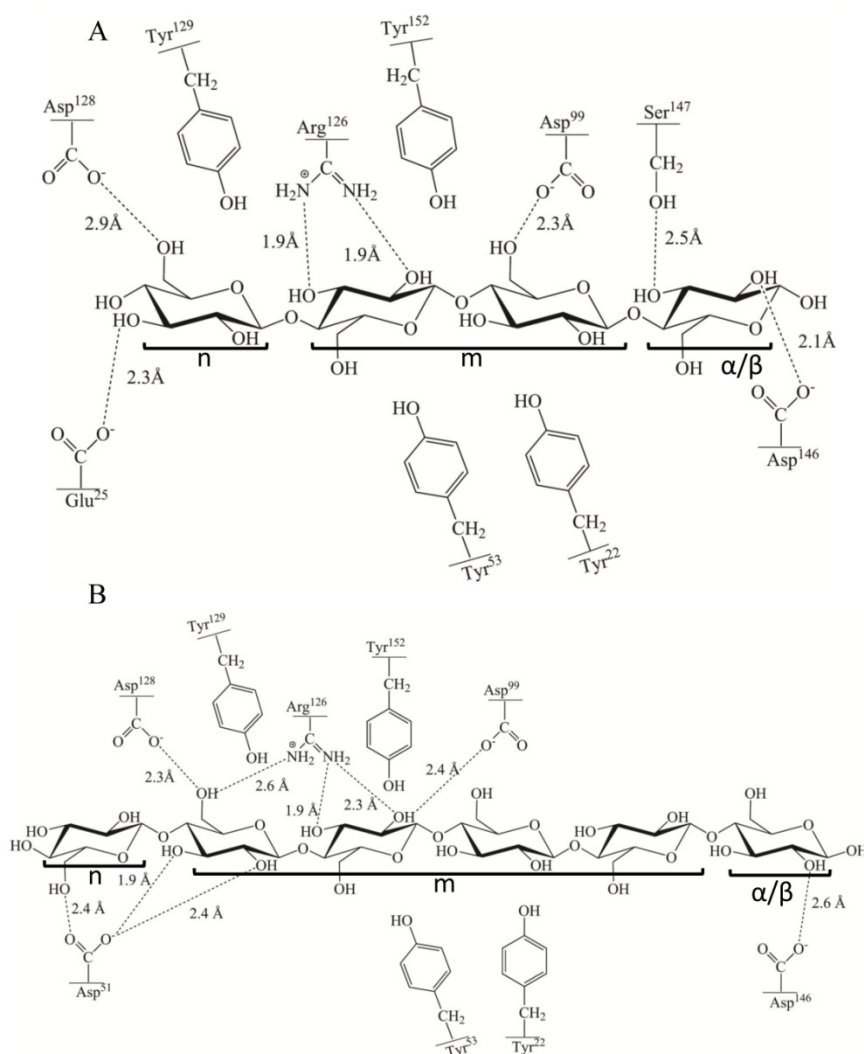


Figure III.15: Representation of the most important interactions between the β -cellotetraose (**A**) and β -cellohexaose (**B**) with the CtCBM11 binding cleft.

Comparison of all the simulated complexes shows that there is a common binding site at the CtCBM11 cleft and all the studied polysaccharides make several contacts with Asp99, Arg126, Asp128 and Asp146 amino acids. Most of the hydrogen bonds occur via the hydroxyl groups associated to the C2 and C6 carbon atoms of each glucose ring, which is in agreement with the results obtained experimentally with STD-NMR and linebroadening studies (**Figure III.7**, **Figure III.9** and **Figure III.10**).

From the above data it can be seen that the central glucose units interact closely with several tyrosine residues. These residues are also involved in the stabilization of the complex through an important dispersive component, between the hydrogens of the sugar and the aromatic ring of the tyrosine residues, which give rise to three so-called non-conventional hydrogen bonds that

help the stabilization of the complex (CH- π interactions).^{5,31,32} The initial conformations adopted by these residues were responsible for the unsatisfactory results of the initial docking trials. Only after exploring the configurational space of these residues, through a multi stage docking with an automated molecular modeling protocol (MADAMM software³⁰), more reliable results were obtained in agreement with the experimental data. Previous site-directed mutagenic experiments have shown that mutating these residues to alanine, causes a significant drop in the activity of the associated enzymes.³ Considering these observations, it was hypothesized that the main function of these residues is to guide the polysaccharide chain and direct it to a specific polar region in the protein populated with several aspartate residues. This would disconnect the chain from other attached polysaccharide chains such as crystalline cellulose.

We have also compared the computational results with another type B CBM that was crystallized in complex with a pentasaccharide (**Figure III.16**).

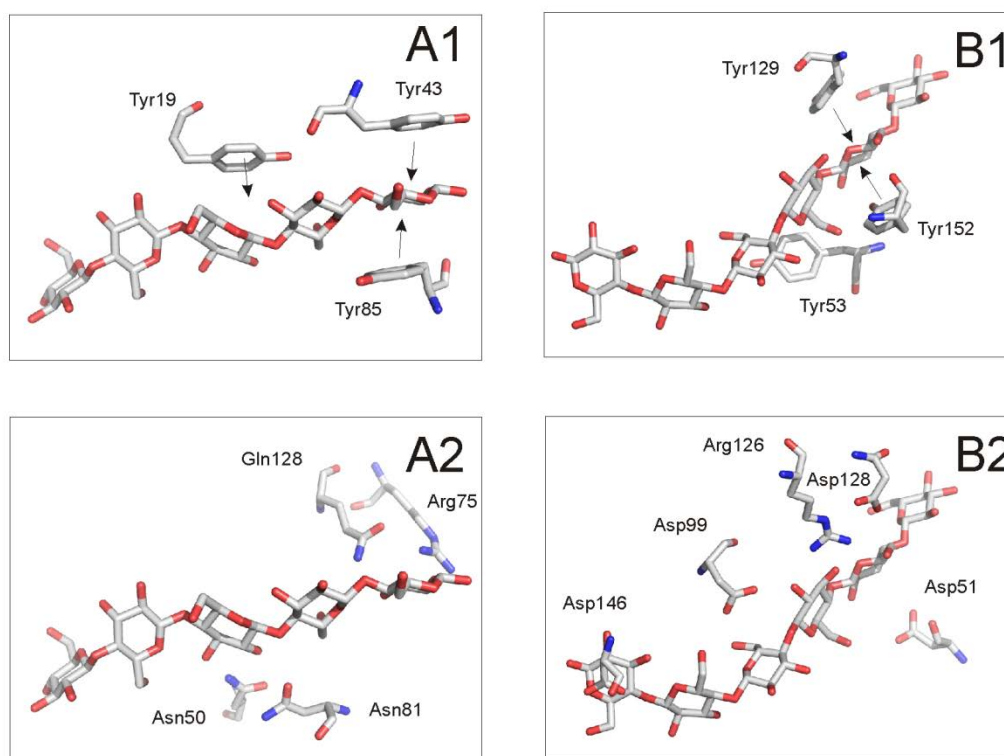


Figure III.16: Schematic representation of the main interaction between the pentasaccharide with **A)** *CfCBM4* (pdb entry: 1GU3³³) and **B)** the hexasaccharide with *CtCBM11*.

A1 and **B1**: interactions involving neighbor tyrosine residues. **A2** and **B2**: residues that establish several hydrogen bonds with the equatorial hydroxyl groups of the glucose units.

Many similarities were found both in the binding region that comprises a flat platform of the CBM, and in the type of interactions between the carbohydrates and *CtCBM11*. Generally, regardless of the CBM, the central carbohydrate interacts with aromatic residues and several charged amino acids that are located at the border of the CBM cleft. In the particular case of

CtCBM11, close interactions with several tyrosines (Tyr22, Tyr53, Tyr129 and Tyr152), one arginine (Arg126) and several aspartate residues (Asp99, Asp128 and Asp146) were observed that closely resemble what it is found in CfCBM4 (**Figure III.16**). These common contacts are responsible for the reorientation of the carbohydrate chain directing it to the regions that are populated with aspartate residues.

III.2.2.6.2 Docking experiments with the NMR solution structure

Using the experimental information about the residues that are most affected by binding, together with the NMR solution structure of the protein, in combination with the previously obtained information obtained by STD-NMR concerning the ligand, I have recalculated a model of the CtCBM11-cellohexaose/cellotetraose complex in a molecular docking approach. The docking procedure was driven with HADDOCK^{34,35}, using the representative NMR solution structure of the ensembles at 25 and 50 °C and the sugar parameters obtained from Glycam Web³⁶ (see *Materials and methods – Section III.4.5.2*). **Figure III.17** shows the obtained models for the interaction of CtCBM11 with cellohexaose at 25 and 50 °C (A and B, respectively) and cellotetraose at 25 °C (C). The models are similar to the ones previously obtained and in good agreement with the experimental data (previous STD-NMR data¹ and titration experiments) and allow a better rationalization of the results.

Because in the NMR-determined structures the binding cleft of CtCBM11 is wider than in the crystal structure there was no need to flexibilize any residue as previously.¹ As can be seen in **Figure III.17**, the models for the interaction of CtCBM11 with cellohexaose at 25 and 50 °C are very similar. For both temperatures, cellohexaose lies equidistant from the two sides of the binding cleft and binding occurs mainly with the four central glucose units (as seen previously). This binding mode is a common feature among CBMs^{5,37,38} that bind ligands that extend over the binding cleft. The similarity of the docked models for both temperatures agrees well with the similarity found in the chemical shift perturbation data from the titration experiments (**Figure III.13 - A and B**).

The majority of the residues perturbed in the titration experiments do indeed interact directly with cellohexaose. For the model at 25 °C only residues Gly24, Trp54, Phe87, Ser93, Ser106, Arg125, Asn144, Ile145 and Phe150 (10 out of 29) do not interact directly with the ligand, while at 50°C, Gly24, Lys28, Gly48, Ile89, Asp51, Gly52, Trp54, Phe87, Ile89, Ser106, Arg125, Met136, Asn144 and Ile145 (14 out of 33) do not interact directly with the ligand. These residues seem to be affected by their directly interacting neighbors.

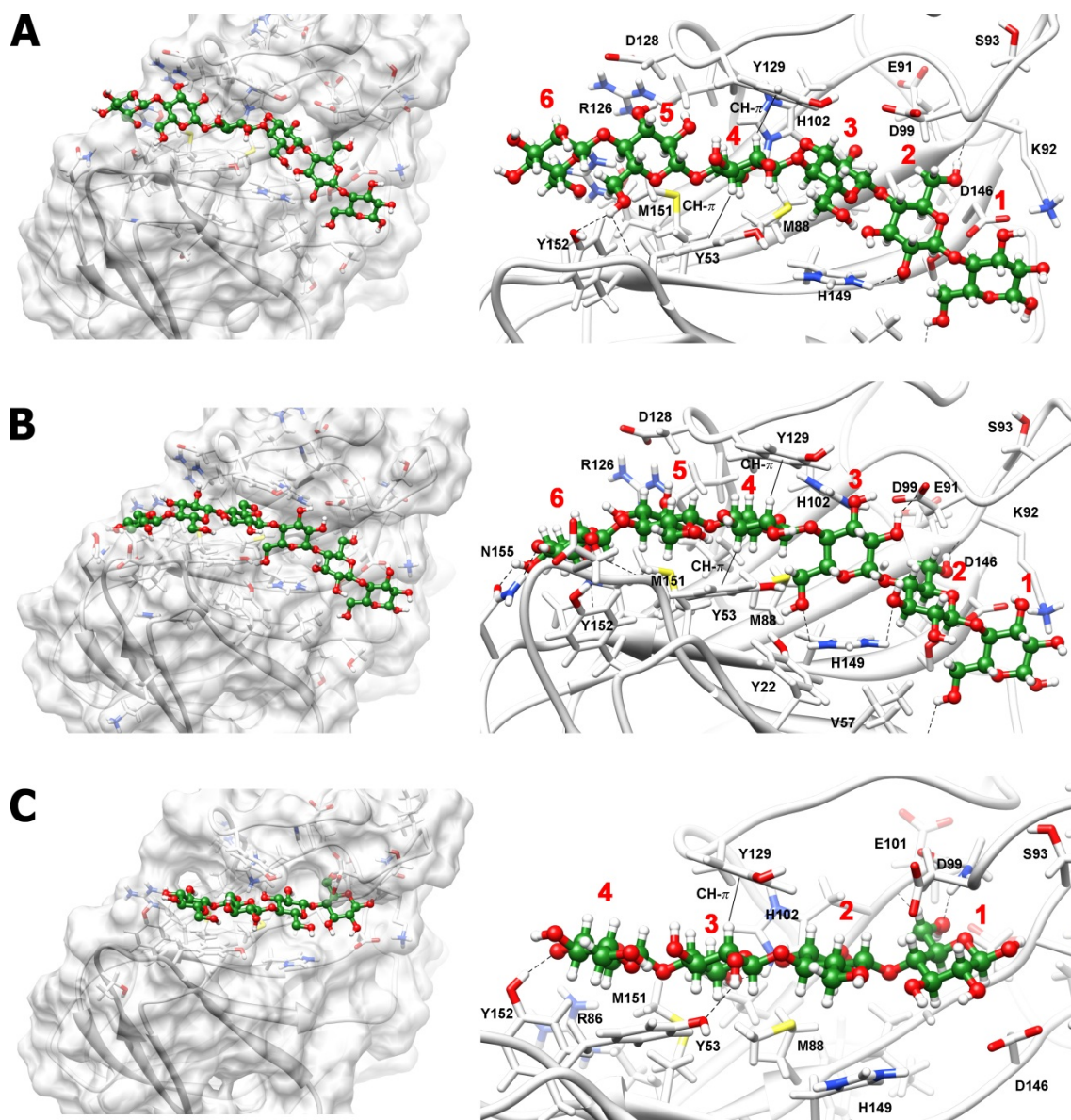


Figure III.17: Docking models of CtCBM11 with cellohexaose at 25 °C (A) and 50 °C (B) and cellotetraose at 25 °C (C).

In the left panel the protein is depicted as a white surface and the ligand as green balls-and-sticks and colored by heteroatom. The right panel shows a highlight of the cleft of the complex. The protein is represented as white ribbons with the interacting residues represented as sticks and the ligand represented as green balls-and-sticks, colored by heteroatom.

Of the residues directly or indirectly affected by binding to cellohexaose, some belong to the loop that binds the first calcium ion. We have Glu91, which is directly bound to the calcium ion and makes a direct hydrogen bond with the ligand; H102 that is the sequential partner of Glu101 and makes a hydrogen bond with cellohexaose and Met136 that lays in between Asp135 and Ser136 and that is only indirectly affected by binding. Therefore, although I previously showed that calcium does not interact with cellohexaose (**Figure III.6**) it seems that its presence is fundamental for the correct positioning of key residues for ligand binding and recognition, thus confirming its structural role.

Looking at the models, we see also that one characteristic of this interaction is the very high number of contacts between the ligand and the protein. For the model at 25 °C the interactions include seven hydrogen bonds involving residues Gly52, Try58, Ile89, Glu91, His102, His149 and Tyr152 and two CH- π interactions between H2 and H3 of sugar unit 3 and residues Try129 and Try53, respectively. For the model at 50 °C, there are ten hydrogen bonds, which involve residues Gly52, Try58, Glu91, Asp99, Asp146, His149, Tyr152 and Asn155 and the same two CH- π contacts between H2 and H3 of sugar unit 3 and residues Try129 and Try53, respectively.

When comparing the model obtained for CtCBM11 and cellotetraose (**Figure III.17 - C**), with that of cellohexaose, we see that as a consequence of the shorter length of the oligosaccharide there is a large decrease in the number of contacts between the protein and the ligand. From the model we see that residues Lys32, Thr49, Arg86, Ile94, Phe123, Arg124 and Asn144 (7 out of 15) whose chemical shift is perturbed by the addition of ligand, do not interact directly with the ligand. Moreover, and in agreement with the perturbation map of **Figure III.13 - C**, we see that cellotetraose interacts preferentially with one side of the cleft. This seems to be a consequence of the fact that the CH- π contact between Tyr53 and the H3 of a sugar unit, as seen for cellohexaose, is lost in the case of cellotetraose. However, the OH group of Tyr53 still interacts with the oligosaccharide through a hydrogen bond with a C2 hydroxyl. From the comparison of the models obtained for cellohexaose with that of cellotetraose, we see also that the total number of hydrogen bonds does not decrease much. In fact, six hydrogen bonds are found between residues Try53, Glu91, Gly100, His102 and Try152 and the sugar.

The large number of protein-ligand interactions, as observed in **Figure III.17**, stabilizes the conformation of cellohexaose in the binding cleft and their careful inspection provide an explanation why this CBM displays a higher affinity for larger ligands when compared to those with the minimal length to fit the binding cleft. As seen, a reduction of the size of the oligosaccharide is accompanied by the loss of several contacts with the protein, including the CH- π interaction with Tyr53, but the overall number of hydrogen bonds is very similar. This fact shows that CH- π interactions and Van der Waals interactions are determinant for increasing the stability of the complexes.

As seen by STD-NMR and with the previous models, a characteristic of the interaction of CtCBM11 with the cellooligosaccharides is the interaction through the hydroxyl groups attached to carbons 2 and 6 from the central glucose units.¹ The models obtained with the crystal structure and the ones obtained with the NMR solution structure (**Figure III.17**) show that these groups make several contacts with the protein, including a number of hydrogen bonds whose presence may dictate the specificity of the protein as it does for other CBMs^{22,39}. For instance, ligands that lack the methylene group (e.g. xylose), have the C2 hydroxyl group in a different position or have any of these positions substituted (e.g. arabinoxylan, galactomannan or carboxymethylcellulose) cannot bind to CtCBM11.⁴ Similarly, β -1,3-linked glucans (as the

case of laminarin - **Figure III.11** and **Figure III.12**) should not bind to CtCBM11⁴ as the orientation of the C2 and C6 hydroxyl groups is different from the β -1,4-linked glucans. Nonetheless, there is still some promiscuity in ligand recognition as shown by Najmudin *et al.*¹⁹ These authors showed that CtCBM11 is capable of binding to xyloglucan, a hemicellulosic polysaccharide composed by a backbone of β -1,4-linked glucose residues which has up to 75% of these residues substituted at O6 with mono-, di-, or triglycosyl side chains.^{40,41} Our experimental results and models show that binding to xyloglucan is only possible if the ramifications of the β -1,4-linked glucose backbone leaves at least four sequential glucose units unsubstituted, thus minimizing any possible sterical clash with the protein. This could explain the low affinity displayed towards xyloglucan – only $0.6 \times 10^4 \text{ M}^{-1}$.¹⁹

In the three models obtained, the same orientation of the ligand (cellohexaose or cellotetraose) in the cleft is maintained and some interactions are conserved; Try129 contacts with the α -face of a sugar unit, while Try53 contacts with the β -face of the same unit. Additionally, the non-reducing end of the sugar is always facing the same side of the protein.

Comparing the models obtained with the crystal structure and with the NMR solution structures we see that, they provide essentially the same conclusions, despite some differences. These differences are mainly in the hydrogen bond network of the different models and on the conformation of the sugars. Nonetheless one has to have in mind that, first, all these are just models that, despite based on experimental data, may not reflect the exact details of the complexes; second, simple rotations on the OH groups for instance are enough to form, change or impair the formation of hydrogen bonds and we are analyzing a single snapshot of this highly dynamic complex; third different starting structures were used (crystal structure and NMR solution structures at 25 and 50 °C) and finally, different software were used for the calculation of the models.

III.2.3 Molecular dynamics

To gain insight into the backbone dynamics of CtCBM11 in solution I measured the longitudinal (R_1) and transverse (R_2) relaxation rates as well as ^1H - ^{15}N steady state NOE for the free and bound protein (with cellohexaose) at 25 and 50 °C. Relaxation parameters (R_1 , R_2 and $\{^1\text{H}\}$ - ^{15}N -NOE) allow to characterize the overall dynamic behavior of the protein in terms of the total correlation time and properties of the diffusion tensor, and internal dynamics in terms of order parameters (S^2) and internal dynamic models. Moreover, it has been shown that order parameters (S^2) derived from NMR relaxation data are related to conformational entropy and can be used to estimate changes in conformational entropy.^{42,43}

The parameters R_1 and R_2 are sensitive to different motional frequencies: R_1 values provide information about motional properties with a frequency of approximately 10^8 – 10^{12} s⁻¹, whereas R_2 values, in addition to depending on motions occurring at these frequencies, are also sensitive to dynamics on the micro-millisecond time scale.^{44,45} Hence, by measuring both R_1 and R_2 , it is feasible to obtain dynamic information over a large motional regime. $\{^1\text{H}\}$ - ^{15}N -NOE relaxation data is highly sensitive to motions of the polypeptide backbone on a pico to nanosecond time scale. NOE values smaller than 0.65 indicate large amplitude backbone fluctuations. Furthermore I have used the model-free approach^{43,46,47} and hydrodynamic⁴⁸ calculations to describe the parameters that characterize internal mobility (S^2 , τ_e and R_{ex}) for the free and bound states at 25 and 50 °C.

In order to better understand the mechanism of ligand recognition/binding of CtCBM11 I have performed hydrogen/deuterium exchange experiments which provided information on the thermodynamics of the structural opening reaction that allows the hydrogen/deuterium exchange process.

III.2.3.1 Relaxation data, diffusion tensor and hydrodynamic calculations

Longitudinal (R_1) and transverse (R_2) relaxation rates as well as ^1H - ^{15}N steady state NOE ($\{^1\text{H}\}$ - ^{15}N -NOE) values were obtained for the free and cellohexaose-bound protein at 25 and 50 °C and **Table III.9** summarizes the average relaxation rates (R_1 and R_2) and the $\{^1\text{H}\}$ - ^{15}N -NOE values obtained under the different experimental conditions as well as the estimation of the total correlation time (τ_m) of the protein from the average R_2/R_1 ratio, excluding values that fail the selection criteria described by Tjandra *et al*⁴⁹ (see Chapter VII – Section VII.4.1).

Table III.9: Average relaxation data and estimation of total correlation time (τ_m) taken from R_2/R_1 ratios.

	25 °C		50 °C	
	<i>Free</i>	<i>Bound</i>	<i>Free</i>	<i>Bound</i>
R_1 (s ⁻¹)	1.34±0.01	1.31±0.02	1.84±0.01	2.04±0.04
R_2 (s ⁻¹)	11.84±0.13	11.37±0.23	7.78±0.15	7.00±0.19
<i>NOE</i>	0.80±0.01	0.78±0.06	0.79±0.01	0.79±0.06
τ_m (ns)	9.11±0.02	8.78±0.04	4.25±0.03	3.43±0.05

The full set of the calculated values is given in Appendix C, **Tables C.3, C.4, C.5** and **C.6** and represented in **Figures C.1, C.2, C.3** and **C.4**. On average, at 25 °C, the values of R_1 do not change significantly upon ligand binding (1.34 ± 0.01 and 1.31 ± 0.02 s⁻¹, respectively) whereas at 50 °C the R_1 values for the complex are higher than for the free protein (1.84 ± 0.01 and 2.04 ± 0.04 s⁻¹, respectively). Concerning the effect of the temperature on the average R_1 values it can be seen that higher temperatures correspond to higher R_1 values independently of the state.

Regarding the transverse relaxation rate, R_2 , ligand binding only causes a very slight decrease at both temperatures, while increasing the temperature leads to a significant decrease in the average R_2 . At 25 °C the average R_2 values are 11.84 ± 0.13 and 11.37 ± 0.23 s⁻¹ for the free and bound protein, respectively, while at 50 °C the average R_2 values are 7.78 ± 0.15 and 7.00 ± 0.19 s⁻¹ for the free and bound protein.

The $\{^1\text{H}\}$ - ^{15}N -NOE values remain fairly constant throughout the amino acid sequence with the exception of some regions that show NOE values well below the average. The residues in these regions belong mainly to loops and are the ones involved (or sequential neighbors) in carbohydrate recognition (**Figure III.18**).

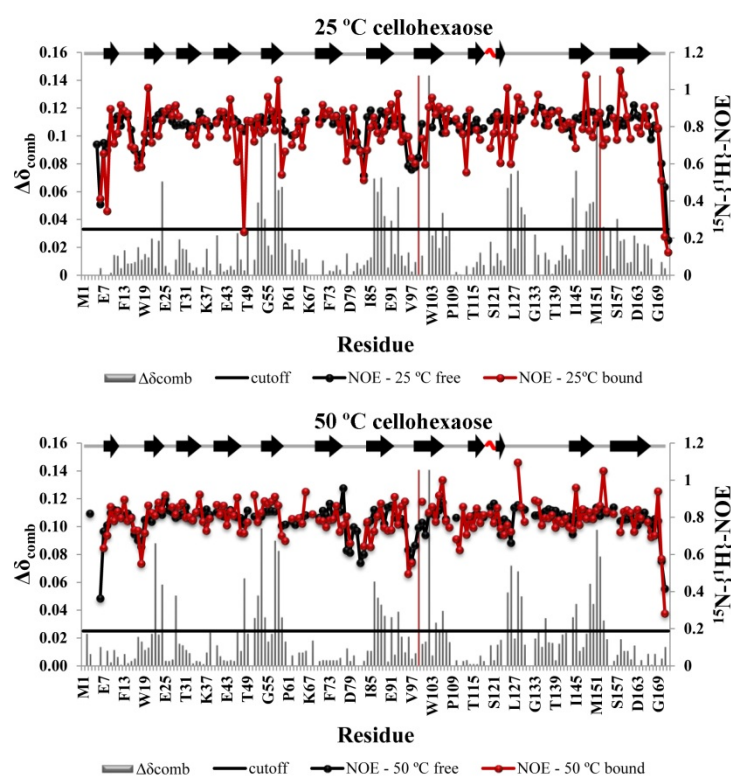


Figure III.18: Graphical superposition of the $\{^1\text{H}\}$ - ^{15}N -NOE of CtCBM11 in the free (black) and bound state (red) at 25 (top) and 50 °C (bottom).

The combined chemical shift is represented as light grey bars.

An initial estimate of the total correlation time τ_m can be obtained from the ratio R_2/R_1 (**Table III.9** and **Figure III.19**) if there are none or only few fast internal motions in the range of the picoseconds and using data from residues that do not undergo any conformational and/or solvent exchange processes ($\text{NOE} < 0.65$).^{43,49} For the free protein at 25 °C the R_2/R_1 ratio was calculated using 134 residues out of 178 and yielded a value of 9.11 ± 0.02 ns. For the bound protein at 25 °C, I have used 115 residues and obtained a value of 8.78 ± 0.04 ns. At 50 °C, the correlation times were 4.25 ± 0.03 (using 125 residues) and 3.43 ± 0.05 ns (using 128 residues) for the free and bound protein, respectively.

As expected based on the Stokes-Einstein relationship, R_2/R_1 ratios and τ_m values decrease with temperature (**Table III.9** and **Figure III.19 - right**), reflecting the reduction in solvent viscosity as a function of the increased temperature. Furthermore, it can be seen that, at 25 °C, the binding of cellobiohexose to the protein does not seem to affect much the total correlation time, while at 50 °C the binding is accompanied by a reduction of about 20% in the total correlation time (**Table III.9** and **Figure III.19 - left**). Because the variation relative to the average for the values of R_2 is much larger than that for the R_1 values, the R_2/R_1 values that deviate from the average belong mainly to the same residues as those that deviate from the average R_2 .

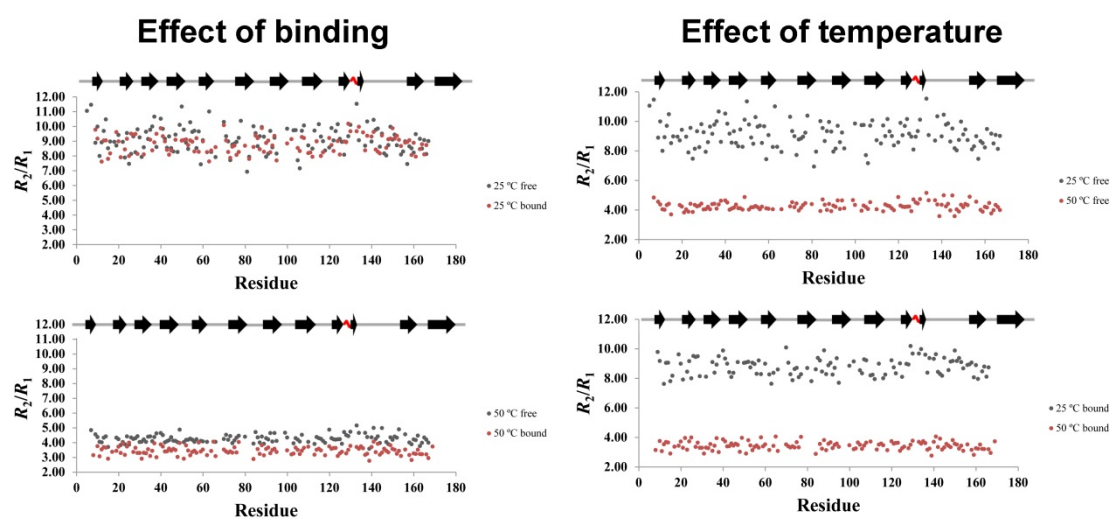


Figure III.19: Effect of binding and temperature on the R_2/R_1 ratio.

The left panel illustrates the effect of binding in the R_2/R_1 ratio whereas the right panel illustrates the effect of the temperature.

Using the software Tensor2.0⁵⁰ and the energy minimized representative conformers of the NMR derived solution structures I have further optimized the total correlation times and calculated the rotational diffusion tensors for the free and bound protein at 25 and 50 °C (*see Appendix C, Table C.7*). The results obtained using the different models are summarized in **Table III.10**. Binding of cellobiohexose is accompanied by a decrease in the overall correlation

time (**Table III.10**). While at 25 °C the variation is very small (9.02 ± 0.05 ns free and 8.88 ± 0.06 ns bound) at 50 °C there is a 15% reduction (5.65 ± 0.05 ns free and 4.83 ± 0.04 ns bound). The structures obtained by docking show that the oligosaccharide fills the binding cleft completely and, for this reason, the complex acquires a more spherical shape than the free protein (**Figure III.17**). The reduction in the correlation time could then be associated with a faster rotation in solution caused by a reduction in friction due to the filling of the binding cleft. The effect is more pronounced at 50 °C. This is in agreement with the structures obtained by docking that show a more intimate contact between the protein and the oligosaccharide at this temperature (**Figure III.17 - B**).

The overall rotational diffusion of CtCBM11 is best described by an axially symmetric model of rotational diffusion (*see Chapter VII – Section VII.4.2.2*), independently of the temperature or the state - bound or unbound. For the unbound protein at 25 °C the diffusion tensors yield a $D_{\parallel}/D_{\perp} = D_{\text{ratio}}^{25^{\circ}\text{C}, \text{free}}$ of 0.87 ± 0.06 (**Table III.10**) which is very similar to the one obtained for the bound protein at the same temperature, $D_{\text{ratio}}^{25^{\circ}\text{C}, \text{bound}} = 0.90 \pm 0.07$. The same behavior of D_{\parallel}/D_{\perp} is obtained at 50 °C, for the unbound protein the $D_{\text{ratio}}^{50^{\circ}\text{C}, \text{free}} = 0.87 \pm 0.08$ and for the bound protein $D_{\text{ratio}}^{50^{\circ}\text{C}, \text{bound}} = 0.88 \pm 0.08$. A $D_{\text{ratio}} < 1$ indicates that the protein behaves as an oblate.

Table III.10: Characterization of the diffusion tensor obtained for CtCBM11 at the different experimental conditions, obtained with Tensor2.0⁵⁰ and HYDRONMR⁴⁸.

		25 °C		50 °C	
		<i>Unbound</i>	<i>Bound</i>	<i>Unbound</i>	<i>Bound</i>
τ_m (ns)	<i>Experimental</i>	9.02 ± 0.05	8.88 ± 0.06	5.65 ± 0.04	4.83 ± 0.04
	<i>HYDRONMR</i>	8.82	-	5.39	-
D_{\parallel}/D_{\perp}	<i>Experimental</i>	0.87 ± 0.06	0.90 ± 0.07	0.87 ± 0.08	0.88 ± 0.08
	<i>HYDRONMR</i>	0.94		0.87	

The program HYDRONMR⁴⁸ was used to perform hydrodynamic calculations assuming a rigid model relaxing only through dipole-dipole and chemical shift anisotropy mechanisms. According to the observation from Bernadó *et al*⁵¹ the inclusion of residues in flexible regions can negatively influence the outcome of hydrodynamic calculations, therefore I removed the first 5 residues of the *C-terminus* and the last 10 (including the 6-residue histidine tail) from the calculation. The energy minimized representative NMR structure at 25 and 50 °C were used for the calculations and the results are summarized in **Table III.10**. The calculated correlation times (8.82 and 5.39 ns for the structures at 25 and 50 °C, respectively) and axial anisotropy diffusion

tensor ratios (0.94 and 0.87 for the structures at 25 and 50 °C, respectively) are in good agreement with the ones derived from the analysis of NMR data. Additionally, both methods also agree about the anisotropy of the rotational diffusion, indicating that the free molecule behaves as an oblate (axially symmetric) rotor.

III.2.3.2 Internal mobility

I used the software Tensor2.0⁵⁰ to determine the parameters characterizing the internal mobility (S^2 , τ_e and R_{ex}) of CtCBM11 in the free and bound states at 25 and 50 °C. The full set of the calculated values is given in Appendix C, **Tables C.8, C.9, C.10 and C.11**. Throughout the analysis, the energy minimized representative NMR solution structures (either at 25 or at 50 °C) was used and the data was fitted into one of five possible dynamic models^{43,46,47,52} (see *Chapter VII – Section VII.4.2.3*). **Table III.11** summarizes the number of residues assigned to each dynamic model for all conditions studied. For the free protein, most residues (99 and 61 for the data at 25 and 50 °C, respectively) were fitted using model 4 (S^2 , τ_e , R_{ex}), meaning that the internal dynamics of those residues is only explainable taking in account a conformational exchange term (R_{ex}) and assuming that they have very fast correlation times ($\tau_e < 500$ ps). For 25 and 27 residues of the protein at 25 and 50 °C, respectively, the data was fitted to model 5 (S^2_{ss} , S^2_{fs} , τ_m), which assumes two time scales for internal motions (fast and slow) and no conformational exchange term. For the free protein at 25° C, 10 residues were fitted with model 2 (S^2 , τ_m) and 20 with model 3 (S^2 , R_{ex}), while for the protein at 50 °C, 54 residues were fitted with model 2 and only two with model 3. Interestingly, none of the residues for the free protein at 25 °C and only three at 50 °C were fitted to the simplest model (model 1 – S^2). This behavior clearly changes upon binding – 24 and 39 residues are fitted with model 1 for the structure as 25 and 50 °C, respectively. The number of residues fitted with model 4 drops to about half for both temperatures (45 and 33 for 25 and 50 °C, respectively) but the number of residues fitted by model 3 increases (45 and 20 for 25 and 50 °C, respectively). In all models, some residues could not be fitted by any of the proposed models.

The order parameter, S^2 reports on the amplitudes of conformational fluctuations on time scales faster than overall rotational diffusion (ps-ns time scale) and ranges from 0 for unrestricted motions to 1 for fully restricted motions (see *Chapter VII – Section VII.4.4*).^{43,53} As seen in **Table III.11**, S^2 has average values greater than 0.8 for all the conditions tested, showing that CtCBM11 has very little internal mobility. Solvent-exposed loops have also high S^2 values but slightly below the average, as expected. The full set of the calculated S^2 values is represented in Appendix C, **Figure C.5**.

Table III.11: Average order parameter (S^2) and dynamic model used to fit the data of the different experimental conditions, obtained with Tensor2.0⁵⁰

	25 °C		50 °C	
	<i>Unbound</i>	<i>Bound</i>	<i>Unbound</i>	<i>Bound</i>
S^2	0.84±0.01	0.82±0.02	0.85±0.01	0.84±0.03
<i>Dynamic model</i>	<i>Number of residues assigned to each model</i>			
^a				
1 (S^2)	0	24	3	39
2 (S^2, τ_m)	10	10	54	33
3 (S^2, R_{ex})	20	45	2	20
4 (S^2, τ_e, R_{ex})	99	45	61	33
5 (S^2_s, S^2_f, τ_e)	25	20	27	13
NA	2	1	3	6
Total	156	145	150	144

^a S^2 is the square of the generalized order parameter; τ_m is the effective correlation time for the internal motions; R_{ex} , is the exchange contribution to T_2 , and the subscripts f and s indicate fast and slow time scales, respectively.

Upon binding, there are a significant number of residues that change their dynamical model to be explained by the simplest dynamic model (model 1) at the expense of more complicated models, particularly models 4 and 5. The obtained results agree well with the previous observation of a more isotropic protein upon binding. Most interestingly, the majority of these residues are the ones identified as affected by binding or their sequential neighbors (16 out of 24 at 25 °C and 24 out of 39 at 50°C). This is also consistent with the structural data at 25 and 50 °C (very similar 3D structures at both temperatures) and with the small variation of the R_2/R_1 ratios along the protein sequence. Furthermore, this shows that both the free and the bound protein are well defined with very little conformational changes. This seems to be inconsistent with the thermodynamic data. Because of the negative $T\Delta S$ value for complex formation, one would expect a more flexible free state and a higher rigidity in the bound state. However, binding is accompanied by a slight decrease on the average S^2 values, denoting a more flexible backbone (**Figure III.20**).

Figure III.20 shows the effect of binding (left panel) and temperature (right panel) on the order parameter, S^2 . Ligand binding causes a decrease in the S^2 values for the majority of the residues at both temperatures, indicating that the protein becomes slightly more flexible upon binding. Regarding the effect of temperature, increasing the temperature leads to an increase in the S^2 value of the majority of the residues.

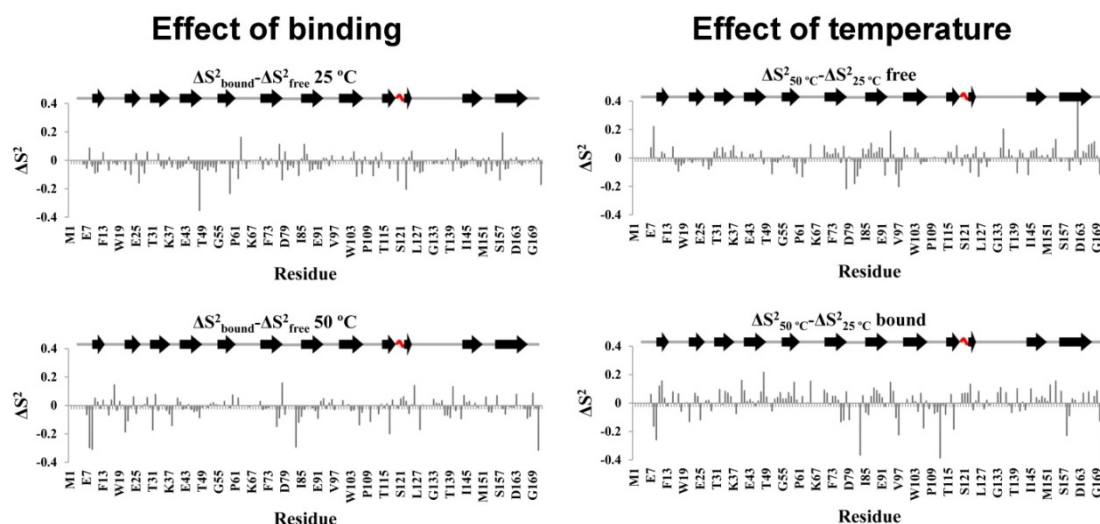


Figure III.20: Effect of binding (left) and temperature (right) on the S^2 order parameter.

III.2.3.3 Estimation of the conformational entropy from NMR relaxation data

The conformational entropy (S_{conf}) can be calculated from the internal mobility-derived order parameters (S^2)^{42,43}, assuming that the motion of the NH bond vector is confined to a cone (see *Materials and methods, Equation III.24 and Chapter VII – Section VII.4.2.3.1*). In general, an increase in the order parameter results in loss of entropy and vice versa. Despite the attractiveness of these approach, one must bear in mind that it comes with several shortcomings (see *Chapter VII – Section VII.4.2.3.1*). Thus, we have to consider that *i*) S^2 values may not be available for all residues and *ii*) of the ones available, only those less than 0.95 can be used; *iii*) the motion of the vectors may not be truly independent; *iv*) the order parameters do not reflect motions outside the ns-ps timescale, and (*v*) solvent ordering (disordering) is not included.^{54,55} For all these reasons, entropy values calculated from order parameters should be considered carefully and used as upper limits of the entropy component (due to the possibility of correlated motions).⁴²

Table III.12 summarizes the average conformational entropy values calculated for the different models (see *Appendix C, Table C.12 for the full set of the calculated values*). Conformational entropy values were extracted accounting for the influence of binding at 25 and 50 °C (143 and 137 residues, respectively) and for the influence of temperature (145 and 131 residues, for the free and bound protein, respectively). As seen in **Table III.12**, the average conformational entropy associated with binding, independently of the temperature is slightly positive while the conformational entropy relative to the increase in the temperature is slightly

negative, independently of the state of the protein. This result is independent of whether all residues are considered or only those involved in binding.

Table III.12: Estimation of the conformational entropy from NMR relaxation data

	$S_{conf25^{\circ}C}^{bound} - S_{conf25^{\circ}C}^{free}$	$S_{conf50^{\circ}C}^{bound} - S_{conf50^{\circ}C}^{free}$	$S_{conf50^{\circ}C}^{free} - S_{conf25^{\circ}C}^{free}$	$S_{conf50^{\circ}C}^{bound} - S_{conf25^{\circ}C}^{bound}$
ΔS_{conf}	$J.mol^{-1}.K^{-1}$			
All	1.50±0.03	0.83±0.05	-0.87±0.02	-1.21±0.06
Cleft	1.53±0.03	0.20±0.04	-0.88±0.01	-2.11±0.05

This means that binding does not occur through an “induced-fit” mechanism with a loss of conformational entropy⁴² but is governed by a conformational selection mechanism, where ligand conformation is determinant for recognition by a rigid protein. These results show that the contribution for the negative binding entropy must originate in the loss of conformational entropy of the ligand. The occupation of the binding cleft, in the free state, by ordered water molecules that act as mobility restrictors could explain the rigidity of this form. The binding event would replace these water molecules by groups of the ligand, thus maintaining the overall rigidity of the protein. In fact, evidence that dehydration effects are involved in the binding process were already postulated before^{5,56,57}.

III.2.3.4 Amide proton exchange

In order to further probe the local environment in the binding cleft I have performed hydrogen/deuterium exchange experiments. These experiments allowed me to identify the residues that are either solvent-exposed (fast exchange rate) or buried or hydrogen-bonded (slow exchange) and provided information on the thermodynamics of the structural opening reaction that allows the hydrogen/deuterium exchange process. Exchange rates were determined as described in the experimental section (*Section III.4.4.17*)

For the free protein, of the 165 assigned amide groups, 58 have very fast exchange rates that could not be determined by this method. From the remaining 107 amide protons, exchange rates were determined only for 51 as for the others the exchange rates are too slow for the experimental time used (about 27h). For the bound protein, of the 154 assigned amide groups, 59 have very fast exchange rates. Of the remaining 95, exchange rates were only determined for 52, as for the remaining 43 the exchange rates are too slow. In both structures, the amide protons that show very fast exchange rates belong mostly to solvent-exposed loops and the ones showing very slow exchange rates belong mostly to β -strand core of the protein (**Figure III.21-**

A). Overall, the different amide groups in CtCBM11 have a wide range of exchange rates, varying from milliseconds to several hours/days.

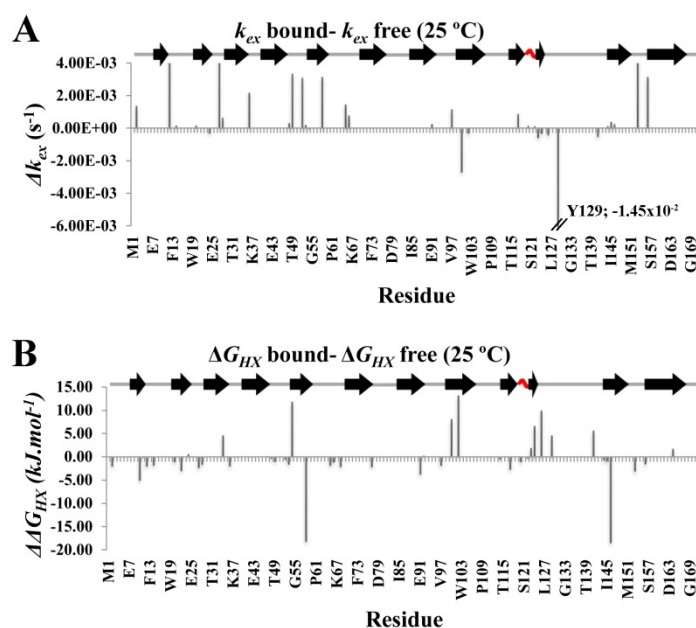


Figure III.21: Effect of binding in the (A) amide hydrogen/deuterium exchange rates and (B) free energy of structural opening for the free and bound protein at 25 °C.

For the great majority of the residues the exchange rate increases upon binding indicating that they become more solvent-exposed. Nonetheless, some residues become more protected. These residues correspond mostly to the ones assumed to be involved in binding or their sequential neighbors.

The free energy of exchange (ΔG_{HX}) of the amide protons was calculated according to **Equation III.25** (see *Materials and methods section III.4.4.17*) assuming an EX2 limit condition (see *Chapter VII – Section VII.4.5 for further details*). These values can provide information on the thermodynamics of the structural opening reaction that allows the hydrogen/deuterium exchange process (the higher the ΔG_{HX} value, the more protected the amide group is).^{58,59} The difference between the measured ΔG_{HX} for the free and bound protein (**Figure III.21 - B**) shows that upon binding, although some residues become less protected (i.e., solvent exposed), residues involved in binding or their sequential neighbors become more protected (**Figure III.21 - A**). This is especially clear for residues Gly100 ($\Delta K_{ex}=2.69 \times 10^{-3} \text{ s}^{-1}$) and for Tyr129 ($\Delta K_{ex}=1.45 \times 10^{-2} \text{ s}^{-1}$). This data is consistent with the formation of hydrophobic interactions between the ligand and the protein and in is good agreement with the dehydration effects pointed earlier. Observing the CtCBM11/cellohexaose models (**Figure III.17**), we see that in fact the amide groups of these two residues make direct contacts with the sugar subunits. The fact that some residues become more solvent exposed may indicate that some parts of the protein need to go through some degree of rearrangement in order to bind to the ligand. This is agrees well with the internal mobility data and thermodynamics of binding. By averaging the

ΔG_{HX} values obtained in the absence and presence of ligand, we see that they remain essentially the same (27.6 and 28.1 kJ.mol⁻¹ for the free and bound protein, respectively). This shows that, although there are local variations in the protection of determined amide groups, the overall net effect is minimal. The complete set of amide proton/deuterium exchange rates and the free energy of the structural opening reaction for free and cellobiose-bound CtCBM11 at 25 °C is given in Appendix C, **Table C.13**.

III.3 Conclusions

X-ray Crystallography, NMR and Computational Chemistry have been shown to be complementary methodologies to study the interaction of carbohydrate-modules with target ligands at an atomic level. When combined, the several techniques here applied can give a deep insight into the mechanisms ruling ligand recognition and binding of CBMs, thus contributing to the global understanding on the exceptional nanomachine that is the cellulosome. By tackling the question in two complementary ways: (i) one focused on the structure of the ligand and the atoms responsible for binding to the proteins, (ii) and the other focused in the identification of the protein residues responsible for ligand recognition I have obtained a full understanding at an atomistic level of the structural and dynamic features that define ligand specificity in CtCBM11 and the mechanism by which this protein is able to distinguish and select its ligands.

From the ligand point-of-view, the absence of signals in the STD-NMR spectrum of the solution of cellobiose with the protein (**Figure III.8**) is a clear indication that either there is no interaction or it is very weak, which is in accordance with previous data.^{3,4} Regarding the interaction with cellotetraose and cellobiose, linebroadening studies and STD-NMR experiments showed that CtCBM11 interacts more strongly with protons H2 and H6 of the central glucose units of both sugars (**Figure III.9** and **Figure III.10**). This is consistent with the binding mode of other Type B CBMs.^{22,39} Moreover, due to the small number of signals for the extremities of cellobiose, it is likely that these sugar units lay outside the binding cleft upon complex formation. This is in good agreement with previous data that showed that the binding cleft of this protein can accommodate at least 4 sugar units.³ However, some contact still exists between the protein and the extremities of the hexasaccharide. These contacts are responsible for stabilizing the complex as the extremities of cellobiose lay outside the binding cleft. In the absence of these relatively weak contacts the entropy of the cellobiose molecule could lead to a decrease in the affinity. These results are in good agreement with the docking experiments (**Figure III.17 – A and B**).

The structural models of cellohexaose bound to the protein were obtained by docking and their analysis reveal a large number of protein-ligand interactions, including CH- π interactions with Tyr53 and Tyr129, that stabilize the conformation of ligands in the binding cleft and should contribute in decreasing the ligand's entropy. Furthermore, the models show that the extremities lay outside the binding cleft but make several contacts with the residues flanking the cleft. These interactions explain why this CBM displays a higher affinity for larger ligands when compared to those with the minimal length to fit the binding cleft. Additionally, the models show that the C2 and C6 OH groups of the central glucose units make several contacts with the protein, including a number of hydrogen bonds whose presence may dictate the specificity of the protein as it does for other CBMs^{22,39}. These contacts, allied to the rigid conformation of the cleft seem to be the specificity determinants of the protein. Therefore, only ligands with a methylene group at C5, with the OH group at C2 in an equatorial position and displaying the typical twisted conformation of β -1,4-linked glucans can bind to this protein. The fact that only one of the diastereotopic protons H6/H6' from the methylene groups shows a relevant peak in the STD spectrum is indicative of a precise orientation of the methylene groups upon binding to the protein. However, this is not clear from the docking models. The docking experiments showed no significant differences in the binding conformations between the α and β isomers.

From the protein's point-of-view, chemical shift perturbation data obtained from ligand titration experiments in combination with the docking studies allowed the identification of the main residues involved in binding in the putative binding cleft. These residues include Tyr22, Tyr53, Asp99, Arg126, Asp128, Tyr129 and Asp146. When using cellotetraose instead of cellohexaose (**Figure III.13**) there is a significant loss of contacts with the protein, including the CH- π interaction with Tyr53 (**Figure III.17**), which is in good agreement with the experimental determined decrease in affinity (**Table III.8**). This fact shows that CH- π interactions and Van der Waals interactions are determinant for increasing the stability of the complexes.

The binding entropy was calculated from binding constants determined from chemical shift perturbation data at both temperatures and showed that the association of CtCBM11 with cellohexaose is enthalpically driven with an unfavorable entropic contribution, which is in good agreement with previous results³ (**Table III.8**). On the other hand, the conformational backbone entropy change associated with binding, as estimated from order parameters (S^2) obtained from relaxation data, resulted in small but positive entropy variation (**Table III.12**). These results suggest that binding does not occur through an "induced-fit" and further support a conformational selection mechanism, where ligand conformation is determinant for recognition by a rigid protein. The contribution for the negative binding entropy must therefore originate in the loss of conformational entropy of the ligand upon complexation with the protein. The

structural models obtained with cellohexaose (**Figure III.17**) bound to the protein reveal a large number of protein-ligand interactions, including CH- π interactions with Tyr53 and Tyr129, which stabilize the conformation of ligands in the binding cleft and should contribute to the decrease in the ligand's entropy.

Overall, I have shown through several experiments that binding of cellooligosaccharides to CtCBM11 must occur primarily by a conformational selection mechanism. This mechanism is common to other CBMs⁶⁰ and is the main determinant of ligand selection for CtCBM11. Because CtCBM11 is topologically similar and structurally homologous to CBMs of families 4, 6, 15, 17, 22, 27 and 29⁶, we can infer that the binding mechanism of these CBMs to their substrates should be also very similar to that of CtCBM11.

Altogether, the results presented allow an atomistic rationalization of the molecular determinants of ligand specificity in CtCBM11 and the mechanism by which this protein is able to distinguish and select its ligands.

III.4 Materials and methods

III.4.1 Sources of sugars

All the sugars (cellobiose, cellotetraose, cellohexaose and laminarihexaose) were obtained from Seikagaku Corporation (Tokyo, Japan) and used without further purification.

III.4.2 Molecular biology

III.4.2.1 Recombinant protein production

The recombinant protein production was done as described in Chapter II.

III.4.2.2 Transformation, expression, purification and quantification of CtCBM11 with the 6-histidine tail

To express CtCBM11 in *E. coli* I have used the same expression vector (pAG1) and transformation procedure as in Chapter II. Furthermore, the colony plating and initial 5 mL culture procedures were the same.

The resulting culture was used to inoculate 1 L of sterile LB medium containing 100 µg/ml of ampicillin. From this point on, all the steps are the same as described in Chapter II. The yields obtained were around 10 mg/L of protein.

III.4.2.2 Transformation, expression, purification and quantification of double labeled (^{13}C and ^{15}N) CtCBM11 with the 6-histidine tail

The transformation, expression, purification and quantification of $^{13}\text{C}/^{15}\text{N}$ -CtCBM11 were done as described in Chapter II.

III.4.3 X-ray crystallography

III.4.3.3 Co-crystallization studies

Attempts to co-crystallize CtCBM11 with candidate cellulosic substrates, involved the addition of excess amounts (1:10 ratio of protein to ligand) of each ligand (cellohexaose and cellotetraose) to the established crystallization conditions³. Crystals grew in these conditions with the same morphology as described before. Crystal characterization and diffraction data collection were performed in-house as described in Chapter II. Diffraction data were processed and scaled, respectively, with programs MOSFLM⁶¹ and SCALA⁶² from the CCP4 suite⁶³. Unfortunately, observation of the electron density maps revealed no ligand binding to the protein's cleft. Due to the negative results obtained I tested new crystallization conditions. I used the hanging drop method (*see Chapter VIII, Section VIII.3*) and the drops were prepared in proper crystallization plates (Nextal Biotechnologie) and were composed by 1 µL of protein and 1 µL of precipitant solution. Of the 80 crystallization conditions⁶⁴ (*see Appendix B, Table B.1*) and different temperatures (4 and 20 °C) tested none produced crystals.

III.4.4 NMR spectroscopy

III.4.4.1 Data acquisition

All NMR spectra were acquired in one of the three spectrometers:

- 400 MHz Bruker ARX spectrometer (Bruker, Wissembourg, France) equipped with a conventional inverse 5 mm probehead with z-gradients (QNPZ);

- 400 MHz Bruker AvanceIII spectrometer (Bruker, Wissembourg, France) equipped with a conventional inverse 5 mm probehead with z-gradients (TXI);
- 600 MHz Bruker AvanceIII spectrometer (Bruker, Wissembourg, France) equipped with a 5 mm inverse detection triple-resonance z-gradient cryogenic probehead (CP TCI).

All data was processed in Bruker TopSpin1.3 or Bruker TopSpin2.2 or Bruker TopSpin3.1 (Bruker).

III.4.4.2 Characterization of the sugars

I prepared solutions of 2 mM of the several ligands (cellobiose, cellotetraose, cellohexaose and laminarihexaose) in 100% D₂O. The assignment of the ¹H and ¹³C NMR spectra was achieved through the analysis of the ¹H, ¹³C, COSY, HSQC, HSQC-TOCSY and 1D selTOCSY spectra and the paper by Sugiyama *et al*¹². All spectra were acquired in a 600 MHz Bruker AVANCE III spectrometer at 298 K.

The ¹H-NMR spectra were acquired in a spectral window of 6002.40 Hz centered at 2824.81 Hz with 32 transients, 64 K data points and a relaxation delay of 1.0 second. The solvent suppression was performed using an excitation sculpting scheme with gradients⁶⁵ in which the solvent signal was irradiated with a selective pulse (Squa100.1000) with a length of 2 ms.

The ¹³C-NMR spectra were acquired in a spectral window of 36057.69Hz centered at 15089.81 Hz with 8192 transients, 64 K data points and a relaxation delay of 2.0 seconds.

The COSY spectra were acquired with 2 transients in a matrix with 4096 data points in F2 in a spectral window of 6009.62 Hz, centered at 2817.40 Hz and 512 increments in F1 with a relaxation delay of 1.0 s.

The HSQC spectra were acquired with 2 transients in a matrix with 2048 data points in F2 in a spectral window of 6009.62 Hz centered at 2824.81 Hz and with 256 increments in F1 in a spectral window of 24998.93 Hz centered at 11314.05 Hz and with a relaxation delay of 1.5 seconds. A delay of 1,72 ms was used for the evolution of the 1 bond CH coupling calculated for ¹J_{C,H} = 145 Hz.

The HSQC-TOCSY spectra were acquired with 4 transients in a matrix with 1024 data points in F2 in a spectral window of 6009.62 Hz, centered at 2824.81 Hz and 256 increments in F1 in a spectral window of 25000.00 Hz centered at 11314.05 Hz with a relaxation delay of 1.5 s. A delay of 1,72 ms was used for the evolution of the 1 bond CH coupling calculated for ¹J_{C,H} = 145 Hz. A delay of 45 ms was used as the mixing time. A delay of 3.45 ms was used for multiplicity selection (CH, CH3 positive, CH2 negative).

The 1D selTOCSY⁶⁶⁻⁶⁸ spectra were acquired in a spectral window of 6002.40 Hz with 32 transients, 32 K data points and a relaxation delay of 1.0 second. The selective irradiation of the different sugar units was performed by using a Gaus1_180r.1000 shaped pulse with a length of 80 ms for centered at the frequencies of the different anomeric proton signals. The TOCSY mixing time was set to 400 ms and a trim pulse with a length of 2.5 ms was used to eliminate unwanted solvent signals.

The ¹H and ¹³C resonance assignments of cellobiose, cellotetraose, cellohexaose and laminarihexaose are summarized on **Table III.3** and **Table III.4**, respectively.

III.4.4.3 Influence of calcium in the structure of cellohexaose

To study the influence of calcium to the structure of cellohexaose I have prepared 6 solutions in which the concentration of the sugar was maintained at 4 mM and the concentration of CaCl₂ increased from 0 to 6 equivalents (0; 0.5; 1.0; 2.0; 3.0 and 6.0). The solutions were prepared in 90% H₂O / 10% D₂O and I have acquired ¹H-NMR spectrum for each solution. The spectra were acquired in a 400 MHz Bruker ARX spectrometer (Bruker, Wissembourg, France) equipped with a conventional inverse 5 mm probehead with z-gradients (QNPZ) at 298 K in a spectral window of 6.636.4 Hz centered at 1879.8 Hz with 128 transients, 64 K data points and a relaxation delay of 1.0 second. The data was processed with TopSpin1.3 (Bruker).

III.4.4.4 Linebroadening studies

The broadening studies were performed at 400 MHz (Bruker ARX) at 298 K, by titration of a solution of cellohexaose 0.80 mM prepared in D₂O with CtCBM11 (1.6 mM). A first spectrum of the pure sugar was acquired. Then the peptide was added in 5 µl and 10 µl volumes to obtain the titration plots. The peptide concentrations were: 0.0, 0.031, 0.060, 0.116, 0.168 and 0.217 mM. All the spectra were acquired with 128 scans in a spectral window with 1991.6 Hz, centred at the solvent frequency (1881.0 Hz). The spectra were deconvoluted into individual Lorentzian lines to determine the full linewidth at half-height. **Table III.5** contains the linewidths at half-height for the different protons of cellohexaose during the titration experiment. Due to the very large broadening of the cellohexaose signals upon the last addition of protein, it was not possible to measure the linewidths at half-height. The data was processed with TopSpin2.1 (Bruker).

III.4.4.5 STD-NMR studies

The interaction between CtCBM11 and cellobiose, cellotetraose, cellohexaose and laminarihexaose was studied by saturation transfer difference NMR (STD-NMR) using the pulse sequence from the Bruker library (stddiffesgp.3)^{16,65}. The pseudo 2D spectra were performed using a solution of 2 mM of sugar and 20 μ M protein in D₂O. All the spectra were recorded at 600 MHz with 16 scans repeated 16 times in a matrix with 32 k points in t₂ in a spectral window of 6410.26 Hz centered at 2733.30 Hz. Excitation sculpting with gradients⁶⁵ was employed to suppress the water proton signals. A spin lock filter ($T_{1\rho}$) with a 2 kHz field and a length of 50 ms was applied to suppress protein background. Selective saturation of protein resonances was performed by irradiating at 0.6 ppm (on resonance spectrum) using a series of 40 Eburp2.1000 shaped 90° pulses (50 ms, 1 ms delay between pulses), for a total saturation time of 2.0 s. For the reference spectrum (off resonance) I irradiated at 20 ppm. To obtain the 1D STD-NMR spectra I subtracted the on resonance spectra from the off resonance using the Topspin2.2 (Bruker, Wissembourg, France) software. The difference spectrum corresponds to the STD-NMR spectrum and the intensity of its signals is proportional the proximity of the corresponding protons to the protein.

The STD was analyzed using the amplification factor (A_{STD}).⁶⁹ The STD amplification factor is obtained by multiplying the relative STD effect of a given hydrogen (I_{STD}/I_0) at a given ligand concentration ($[L]_T$) with the molar ratio of ligand in excess relative to the protein ($[L]_T/[P]$), according to **Equation III.1** (same as in Section III.2.2.3)¹⁸:

$$A_{STD} = \frac{I_0 - I_{SAT}}{I_0} \times \text{ligand excess} = \frac{I_{STD}}{I_0} \times \text{ligand excess}$$

III.1

were A_{STD} is the STD amplification factor, I_0 , I_{SAT} and I_{STD} are the intensities of the reference (off resonance), saturated (on resonance) and difference spectra (STD-NMR) respectively.

For a determined saturation time the A_{STD} can also be depicted as the average number of ligand molecules saturated per molecule of receptor. In principle the longer the saturation time and the more ligand used the stronger the STD and the higher the A_{STD} due to ligand turn over at the binding site. In order to get the epitope mapping information from the amplification factor for a given saturation time, the relative STD (or A_{STD}) with the highest intensity is set to 100 %, and all other STD signals are calculated accordingly (see Chapter VII – Section VII.5.1 for a complete explanation of the STD-NMR experiment).

III.4.4.6 Diffusion studies (DOSY)

The interaction between CtCBM11 and cellohexaose and laminarihexaose was studied by diffusion ordered spectroscopy (DOSY) using the pulse sequence from the Bruker library (ledbpgppr2s)⁷⁰. The pulse scheme uses stimulated an echo and LED (longitudinal eddy current delay), bipolar gradient pulses for diffusion, 2 spoil gradients and with presaturation during relaxation delay (see Chapter VII – Section VII.5.2 for a complete explanation of the DOSY experiment). All the spectra were recorded at 600 MHz with 512 scans in a matrix with 32 k points in t2 in a spectral window of 12335.526 Hz centered at 2817.10 Hz at 298 K. 32 gradient steps were acquired with the gradient strengths augmented linearly from 5% to 95% (100% \equiv 56 G/cm). It is important to start with a gradient strength bigger than 0, because one may get unwanted echoes when not applying a gradient. Furthermore, it is recommended the highest power to be 95 % to make sure that there is no non-linear behavior of the gradient amplifier at the end of the amplification range (but one may go up to 100 %).²

A first solution, with both carbohydrates at a concentration of 40 μ M in D₂O with 0.1% Trimethylsilyl propionate (TSP - to account for viscosity changes²⁴) was prepared in order to extract the self-diffusion coefficients for the free carbohydrates. The duration of the encoding/decoding gradient (little delta - δ) was calibrated to 1.5 ms and the diffusion time (big delta - Δ) was calibrated to 400 ms. The duration of the spoil gradients was set to 600 μ s. A second solution containing the mixture of both carbohydrates and CtCBM11 at a concentration of 40 μ M in D₂O with 0.1% TSP was prepared in order to get the self-diffusion coefficients for the carbohydrates in the presence of the protein and of the protein (it was assumed that the diffusion coefficient of the protein when bound is the same as when the protein is in the free state). The duration of the encoding/decoding gradient (little delta - δ) was calibrated to 1.1 ms and the diffusion time (big delta - Δ) was calibrated to 800 ms. The duration of the spoil gradients was set to 600 μ s. The data were analyzed using the variable gradient fitting routines in Bruker TopSpin2.2 software. All the peak intensities were fitted using a mono-exponential decay:

$$I = I_0 \exp \left[-D(\gamma g \delta)^2 \left(\Delta - \frac{\delta}{3} - \frac{\tau}{2} \right) \right]$$

III.2

where I_0 is the resonance amplitude at zero gradient strength, γ is the magnetogyric ratio of the proton ($2.675 \times 10^8 \text{ rad.T}^{-1}.\text{s}^{-1}$), g and δ are the strength and duration of the gradient, respectively, Δ is the diffusion time and τ is the gradient pulse recovery time.

From the data in **Table III.7** I was able to quantify the interaction in terms of the association constant (K_a) using the following equations:

$$K_a = \frac{[PL]}{[P][L]} \quad \text{III.3}$$

$$D_{obs} = f_L D_L + f_{PL} D_{PL} \quad \text{III.4}$$

where K_a is the association constant and $[PL]$, $[P]$ and $[L]$ are the equilibrium concentrations of the protein-ligand complex, protein and ligand, respectively, f_L and f_{PL} are the molar fractions of the free and bound protein, respectively and D_L , D_{obs} and D_{PL} are the diffusion coefficients of the free ligand, the ligand when bound to the protein and the protein when bound to the ligand, respectively, divided by the diffusion coefficient of the TSP to account for viscosity changes²⁴.

From **Equation III.4** we get:

$$f_{PL} = \frac{D_L - D_{obs}}{D_L - D_{PL}} \quad \text{III.5}$$

If it is assumed that D_{PL} is the same as the measurable diffusion of the free protein (D_P), then f_{PL} can be easily determined (D_P , D_L and D_{obs} can be extracted from the DOSY spectrum).

Accounting for mass balance and combining **Equations III.3** and **III.5** we get the expression for the association constant:

$$K_a = \frac{f_{PL}}{(1 - f_{PL}) \cdot ([P]_0 - f_{PL}[L]_0)} \quad \text{III.6}$$

where $[P]_0$ and $[L]_0$ represent the total concentrations of protein and ligand, respectively.

III.4.4.7 CtCBM11 titration

I have studied the interaction between CtCBM11 and cellohexaose and cellotetraose by NMR chemical shift perturbations by titrating double-labeled CtCBM11 with cellohexaose and cellotetraose. For the titration experiment, I have acquired a series of six ¹⁵N-¹H-HSQC spectra

in which the concentration of protein was maintained at 0.1 *mM* and the concentration of ligand varied from 0 to 2 equivalents (0; 0.3; 0.5; 1; 1.5 and 2). The spectra were acquired with 2048 × 256 points and 32 scans. The spectral widths were 9615.38 Hz for ¹H and 2311.07 Hz for ¹⁵N. The central frequency for proton was set on the solvent signal (2817.40 Hz) and for nitrogen was set on the center of the amide region (7175.66 Hz). The spectra relative to the interaction CBM11-cellohexaose were acquired at 25 and 50°C whilst the ones relative to the interaction CBM11-cellobetraose were acquired only at 25°C.

III.4.4.8 Combined chemical shift

For the evaluation of the behavior of individual amino acids upon addition of increasing amounts of ligand I have calculated the combined amide proton and nitrogen chemical shift differences using the following equation^{26,71}:

$$\Delta\delta_{comb} = \sqrt{(\Delta\delta_H)^2 + (w_i\Delta\delta_N)^2}$$

III.7

where $\Delta\delta_H$ and $\Delta\delta_N$ are the chemical shifts of proton and nitrogen, respectively and w_i is a weighting factor which accounts for differences in sensitivity of different resonances in an amino acid (e.g. amide ¹H and ¹⁵N). When chemical shifts are expressed in ppm a suitable estimate for the weighting factors is given by²⁶:

$$w_i = \frac{|\gamma_i|}{|\gamma_H|}$$

III.8

with γ_i and γ_H the magnetogyric ratio of nucleus i and the proton, respectively.

In order to decide whether a given residue belongs to the class of interacting or non-interacting residues I have calculated a cutoff value. In a first approximation, the chemical shift distributions of the non-interacting residues can be assumed to follow a normal distribution with a mean of zero. Therefore, the standard deviation to zero, σ_o , for the class of non-interacting residues is a reasonable measure to predict if a residue belongs to the class of interacting residues or not.²⁶ Nevertheless, if the values of all residues are used, the obtained result will be strongly biased by the large chemical shift changes of the interacting residues. Therefore, I have used an iterative procedure that successively removes outliers to calculate a corrected standard deviation to zero σ_0^{corr} that is used in the following as cutoff value.²⁶

III.4.4.9 Determination of the association constant (K_d)

Based on the fact that the variation in the chemical shift of the amide proton and nitrogen upon titration with ligand acts as marker for the binding equilibrium, I have used the combined chemical shifts to obtain the dissociation constant (K_d).^{26,27}

For a system in fast exchange the association constant is given by:

$$K_d = \frac{[P] \times [L]}{[PL]}$$

III.9

where $[P]$, $[L]$ and $[PL]$ are the concentrations of free protein, free ligand and the complex, respectively. Because:

$$[L]_0 = [L] + [PL] \text{ and } [P]_0 = [P] + [PL]$$

III.10

where, $[P]_0$ and $[L]_0$ are the total concentrations of protein and ligand, respectively. We get:

$$K_d = \frac{([P]_0 - [PL]) \times ([L]_0 - [PL])}{[PL]}$$

III.11

Rearranging in order to $[PL]$ we get:

$$[PL] = \frac{(K_d + [L]_0 + [P]_0) - \sqrt{(K_d + [L]_0 + [P]_0)^2 - (4[P]_0[L]_0)}}{2}$$

III.12

Because the system is in fast exchange, the NMR response – variation in the chemical shift – is given by:

$$\Delta\delta_{comb} = f_P \times \Delta\delta_P + f_{PL} \times \Delta\delta_{PL}$$

III.13

were $\Delta\delta_{comb}$ is the combined chemical shift, f_p and f_{PL} are the molar fractions of free and bound protein, respectively, and $\Delta\delta_p$ and $\Delta\delta_{PL}$ are the combined chemical shifts for the free and bound protein, respectively. As the molar fraction of bound protein, f_{PL} , is given by:

$$f_{PL} = \frac{[PL]}{[P]_0}$$

III.14

and, in the limit $f_p = 0$, rearranging **Equation III.13**, we get:

$$\Delta\delta_{comb} = \frac{[PL]}{[P]_0} \times \Delta\delta_{max}$$

III.15

were $\Delta\delta_{max}$ is maximum chemical shift of the bound protein (i.e. $\Delta\delta_{PL}$ in the limit). Rearranging in order to $[PL]$ we get:

$$[PL] = \frac{\Delta\delta_{comb} \times [P]_0}{\Delta\delta_{max}}$$

III.16

By replacing in **Equation III.12**, we finally get:

$$\Delta\delta_{comb} = \Delta\delta_{max} \frac{(K_D + [L]_0 + [P]_0) - \sqrt{(K_D + [L]_0 + [P]_0)^2 - (4[P]_0[L]_0)}}{2[P]_0}$$

III.17

Titration of ligand into protein so that the ligand eventually finishes in excess, thus saturating the protein binding site, is the only way to perform this study. Little useful information would come out of a protocol where the ligand concentration never exceeded that of the protein. Neither will much useful information come from a system where the ligand concentration vastly exceeds the protein concentration unless the binding event is very weak.

III.4.4.10 Determination of the thermodynamic parameters

Using the binding constants (K_a) determined ($K_a=1/K_d$) I have calculated the equilibrium thermodynamic parameters ΔH and ΔS using a van't Hoff plot⁷² according to the following equation:

$$-RT\ln(K_a) = \Delta G = \Delta H - T\Delta S$$

III.18

where R is the gas constant ($8.314472 \text{ J.K}^{-1}.\text{mol}^{-1}$), T is the temperature (either 298 or 323 K) and ΔH and ΔS are the enthalpy and entropy, respectively.

III.4.4.11 ¹⁵N backbone relaxation measurements

To gain insight into the backbone dynamics of CtCBM11 in solution I have measured the relaxation parameters R_1 , R_2 and $\{^1\text{H}\}$ -¹⁵N-NOE (HetNOE) for the free and bound protein (with cellohexaose) at 25 and 50 °C. I used a double labeled protein sample (¹³C-¹⁵N-CtCBM11) at a concentration of 0.7 mM for the free protein and 0.3 mM with 2 equivalents of cellohexaose for the bound protein. The solutions were prepared in 90% H₂O / 10% D₂O. All data were collected in the Bruker Avance III 600 MHz spectrometer.

Backbone relaxation rates, R_1 and R_2 , were determined by acquiring pseudo-3D spectra consisting in a series of 2D heteronuclear ¹H-¹⁵N-HSQC experiments⁷³⁻⁷⁵ where the relaxation period varied. For the ¹⁵N longitudinal relaxation rates (R_1), 13 time points were collected (50ms; 0.1s; 0.2s; 0.4s; 0.6s; 1s; 1.5s; 2s; 2.5s; 3s; 3.5s and 4s). The spectrum was acquired with 2048 points in ¹H indirect dimension and 40 points in the ¹⁵N direct dimension and 16 scans. The spectral width was 9615.39 Hz in the ¹H dimensions and 2311.08 Hz in the ¹⁵N dimension and the relaxation delay was 5s. The central frequency for proton was set on the solvent signal (2817.40 Hz) and for nitrogen was set on the center of the amide region (7175.66 Hz). For the ¹⁵N transverse relaxation rate (R_2) 8 time points were collected (0.016s; 0.032s; 0.065s; 0.097s; 0.129s; 1.161s, 1.194s and 0.258s). The spectrum was acquired in the same conditions as the above and the relaxation delay was 2.5s.

The $\{^1\text{H}\}$ -¹⁵N-NOE steady-state NOE^{76,77} experiments were recorded with a relaxation delay of 5 s, with 32 transients in a matrix with 2048 data points in F2 and 128 or 256 increments in F1 (for the free and bound protein, respectively) in an interleaved manner, with alternating proton-presaturated and non-presaturated spectra. The central frequency for proton was set on the solvent signal (2817.40 Hz) and for nitrogen was set on the center of the amide region (7175.66

Hz) and the spectral width was 9615.39 Hz in the ^1H dimensions and 2311.08 Hz in the ^{15}N dimension. The interleaved spectra were separated by a Bruker standard macro.

III.4.4.12 Relaxation data processing and analysis

The data was processed with the software TopSpin2.2 (Bruker) and analyzed in CARAM1.8.4.2⁷⁸. In order to correctly read the data in CARAM, all the T_1 set, T_2 set and both HetNOE spectra (saturated and unsaturated) were processed in TopSpin2.2 with same intensity scaling factor (nc_proc). T_1 and T_2 relaxation data peak intensities were fitted with the software OriginPro 8 (OriginLab, Northampton, MA) into **Equations III.19** and **III.20**, respectively:⁷⁹

$$I_t = I_0 \left(1 - e^{-\frac{\tau}{T_1}} \right) \quad \text{III.19}$$

$$I_t = I_0 e^{-\frac{\tau}{T_2}} \quad \text{III.20}$$

where I_t is the intensity at time τ and I_0 is the intensity at equilibrium. The errors were extracted directly from the fitting. The HetNOE values are defined as the ratios of peak intensities with and without proton saturation:

$$NOE = \frac{I_{sat}}{I_{unsat}} \quad \text{III.21}$$

where I_{sat} and I_{unsat} are the peak intensities with and without proton saturation, respectively. I have calculated the uncertainties of HetNOE values, σNOE , using the well-established method⁷⁴:

$$\frac{\sigma NOE}{NOE} = \sqrt{\left(\frac{\sigma I_{sat}}{I_{sat}} \right)^2 + \left(\frac{\sigma I_{unsat}}{I_{unsat}} \right)^2} \quad \text{III.22}$$

where I_{sat} and I_{unsat} are the peak intensities with and without proton saturation, respectively. Their uncertainties (σ) were determined from the root mean-square noise in the background

regions. A table with all the measured R_1 ($R_1=1/T_1$), R_2 ($R_2=1/T_2$) and NOE values is given in Appendix C (Tables C.3, C.4, C.5 and C.6).

III.4.4.13 Estimation of the molecular diffusion tensor

An initial estimate of the magnitude and orientation of the diffusion tensor of the free and bound protein at each temperature was obtained from the ratio R_2/R_1 .^{43,50} In order to obtain a reliable estimate of overall rotational diffusion tensor residues with large amplitude fast internal motions have to be excluded from the calculation (NOE<0.65) because their change in T_1 is much larger than the T_2 variation. Among the remaining residues, those with significant conformational exchange on the microsecond/millisecond time scale were also excluded according to the following condition⁵³:

$$\frac{\langle T_2 \rangle - T_{2,n}}{\langle T_2 \rangle} - \frac{\langle T_1 \rangle - T_{1,n}}{\langle T_1 \rangle} > 1.5 \times SD$$

III.23

where $\langle T_2 \rangle$ and $\langle T_1 \rangle$ are the average values of T_2 and T_1 , respectively, $T_{2,n}$ and $T_{1,n}$ are the T_2 and T_1 values of residue n , respectively. SD is the standard deviation of **Equation III.23**. The residues that do not fulfill these criteria often experience additional linebroadening, commonly described by the exchange term R_{ex} .⁴³

III.4.4.14 Hydrodynamic calculations

A theoretical estimation of the diffusion parameters and NMR relaxation data has been performed by using the program HYDRONMR⁴⁸ based on the bead-model method. All the calculations were made using the energy minimized representative conformers of the NMR solution structure of CtCBM11 at a temperature of 298 and 323 K and solvent viscosity of 0.00911 and 0.00557 poise, respectively, corresponding to a 90%/10% H₂O/D₂O mixture. The radius of the atomic elements (AERs) used was 2.2. According to the observation from Bernadó *et al*⁵¹ the inclusion of residues in flexible regions can negatively influence the outcome of hydrodynamic calculations, therefore the first 5 residues of the *C-terminus* and the last 10 (including the 6-residue histidine tail) have been excluded from the calculation.

III.4.4.15 Calculation of the model free dynamics parameters

After the initial estimation of the global correlation time as described above, the model-free formalism⁷⁶ was used to further refine the rotational correlation time, τ_m (see Appendix C, **Table C.7**) and to describe the motions of the protein in terms of an order parameter (S^2), conformational exchange (R_{ex}) and effective internal correlation time (τ_e). The model-free analysis was carried out with the Tensor2.0 software.⁵⁰ I used a N-H bond length of 1.02 Å and a chemical shift anisotropy (CSA) of -172 ppm for the ¹⁵N backbone spins.⁸⁰ The appropriate models for internal dynamics parameters were chosen using an iterative fitting procedure and statistical significance tests.⁴³ Five different models were tested to characterize the internal dynamics of the N-H groups; each model included optimization of different micro dynamic parameters (S^2 , τ_e , R_{ex}). The five models are described in detail in Chapter VII – Section VII.4.2.3. I have used the energy-minimized representative NMR structure of the two ensembles throughout the analysis and the same residues as for the initial estimations of the correlation time. All the calculated internal mobility parameters (S^2 , τ_e and R_{ex} and the dynamic model used to fit the data) can be found in Appendix C, **Tables C.8, C.9, C.10** and **C.11**

III.4.4.16 Estimation of the conformational entropy from NMR relaxation data

The conformational entropy arising from ps timescale motion of the NH bond vectors, assuming the bond motion to be confined to a cone was calculated for the several states considered (free and bound at 25 and 50 °C) using **Equation III.24**:⁵⁵

$$\Delta S_{conf} = k \sum_{j=1}^N \ln \left[\frac{3 - (1 + 8S_{j,final})^{1/2}}{3 - (1 + 8S_{j,initial})^{1/2}} \right]$$

III.24

where ΔS_{conf} is the change in conformational entropy, k is the Boltzmann constant and S_j is the order parameter for the residue j in the final ($S_{j,final}$) and initial state ($S_{j,initial}$).

This equation assumes that the NH bond motion is confined to a cone and that the motions of the individual NH vectors are independent, which may lead to an overestimate of the entropy value. Furthermore, the above equation is valid when the value of $S^2 < 0.95$ (see Chapter VII – Section VII.4.2.3.1 for further details). A full list of the ΔS_{conf} values calculated can be found in Appendix C, **Table C.12**

III.4.4.17 Amide proton exchange

In order to analyze the decay of the amide proton signal intensities due to hydrogen exchange with D₂O I have used a lyophilized double labeled (¹³C and ¹⁵N) protein sample with and without cellohexaose. For the data acquisition the samples were dissolved in 75 mM phosphate-buffered D₂O at pD = 7.5 to a final concentration of 1 mM (1:2 protein/ligand ratio). The dissolved sample was immediately placed into the NMR spectrometer, previously tuned and shimmed with a sample of the buffer used. For the free protein, the time required between dissolving the sample and starting the acquisition of the first spectrum was 1min and 46 s, whilst for the mixture it was 1 min and 14 s. For both experiments, a series of 30 ¹H-¹⁵N-HSQC spectra were acquired with 1024 × 128 complex points, in a spectral window of 9615.39 × 2311.08, in F2 and F1, respectively. The ¹H-¹⁵N-HSQC spectra series were acquired with an increasing number of scans – (**Table III.13**) – due to the loss of signal intensity and consequent decrease of the signal/noise ratio. Details on the theory of amide proton exchange are given in Chapter VII – Section VII.4.2.4 and a full list of rates by residue is presented in Appendix C – **Table C.13**

The data was processed with the software TopSpin2.2 (Bruker) and analyzed in CARA1.8.4.2⁷⁸. In order to correctly read the data in CARA, all the spectra were processed in TopSpin2.2 with same intensity scaling factor (nc_proc). The cross-peak volumes obtained from CARA were normalized to the number of scans of each experiment. To determine the exchange rates of the individual amide protons, the normalized peak volumes were plotted as a function of the elapsed time[†] and fitted to a three-parameter single-exponential decay function:⁵⁸

$$I(t) = I_0 e^{-k_{ex} \cdot t} + C$$

III.25

where $I(t)$ is the intensity at time t , I_0 is intensity at time 0, k_{ex} is the exchange constant, t is the time elapsed and C is the final amplitude.

The protection factors (Pf) for the several amide protons were estimated according to **Equation III.26**:⁸¹

$$Pf = \frac{k_{rc}}{k_{ex}}$$

III.26

[†] The elapsed time is defined as the period from the suspension of the sample in the D₂O phosphate buffer to half of acquisition time of an experiment.

where k_{rc} and k_{ex} represent the exchange rates of the protein in the random coil and native conformations states, respectively.

The hydrogen-exchange rates of amide protons in non-structured peptides, k_{rc} , were estimated using the software SPHERE⁸² (<http://www.fccc.edu/research/labs/roder/sphere>) with the default activation energies ($E_{a,s}$): Acid E_{aH} : 15.0 kcal/mol, Base E_{aOH} : 2.6 kcal/mol. The exchange media was set to D₂O, the temperature was set to 25 °C and the pH was set to 7.5. The reference data was set to poly-DL-alanine.⁸¹ The remaining parameters were kept with the defaults values.

The free energy of exchange of the amide protons was calculated according to the following equation:

$$\Delta G_{ex} = -RT \ln \frac{k_{ex}}{k_{rc}} = -RT \ln \frac{1}{Pf}$$

III.27

where R is the gas constant (8.314472 J.K⁻¹.mol⁻¹) and T is the absolute temperature at which the exchange was monitored (298K). The calculated ΔG_{ex} values for the free and bound protein are given in Appendix C, **Table C.13**.

Table III.13: Series of ¹⁵N-¹H-HSQC spectra acquired in order to analyze the decay of the amide proton signal intensities due to hydrogen exchange with D₂O for the free and bound CtCBM11 at 298 K.

<i>Exp.</i>	<i>N° of scans</i>	<i>Time elapsed - free (s)</i>	<i>Time elapsed - bound (s)</i>	<i>Exp.</i>	<i>N° of scans</i>	<i>Time elapsed - free (s)</i>	<i>Time elapsed - bound (s)</i>
-	-	106.0	74.0	16	16	9480.0	9448.0
1	2	250.0	218.0	17	16	10609.0	10577.0
2	2	394.0	362.0	18	16	11738.0	11706.0
3	4	683.0	651.0	19	16	12867.0	12835.0
4	4	972.0	940.0	20	16	13996.0	13964.0
5	4	1261.0	1229.0	21	16	15125.0	15093.0
6	4	1550.0	1518.0	22	16	16254.0	16222.0
7	4	1839.0	1807.0	23	32	18503.5	18471.5
8	4	2128.0	2096.0	24	32	20753.0	20721.0
9	4	2417.0	2385.0	25	32	23002.5	22970.5
10	4	2706.0	2674.0	26	32	25252.0	25220.0

11	16	3835.0	3803.0	27	32	27501.5	27469.5
12	16	4964.0	4932.0	28	32	29751.0	29719.0
13	16	6093.0	6061.0	29	64	34241.0	34209.0
14	16	7222.0	7190.0	30	64	38731.0	38699.0
15	16	8351.0	8319.0				

III.4.5 Computational studies

III.4.5.1 Docking experiments with the crystallographic structure and molecular dynamics

The 1v0a PDB deposited structure of CtCBM11³ was used as the starting point for all the computational studies. All waters and sulphate ions (SO₄²⁻) were deleted and only the protein atoms were kept. Furthermore, all selenium atoms were substituted by sulphur atoms.

The protein is composed of 172 amino acids but the crystallographic file lacks 3 amino acids in a loop between Val78 and Ala82. These residues were modeled with the help of the software Insight II⁸³, to generate the correct sequence. Once the structure was ready, hydrogen atoms were added using InsightII⁸³, with all residues in their physiological protonation state.

In order to evaluate the CtCBM11 selectivity to saccharides several ligands were designed, namely, cellobiose, cellotetraose and cellohexaose. As glucose can exist in two forms, α -glucose and β -glucose and these monomers have the ability to change between these two forms very easily, each ligand was modeled in both forms.

All geometry optimizations and molecular dynamics were performed with the parameterization adopted in Amber 8⁸⁴, using the GAFF, the general AMBER force field^{29,85}, for the protein and the Glycam-04 parameters for the carbohydrates.^{11,86,87} In all simulations an explicit solvation model was used with a truncated octahedral box of 12 Å with pre-equilibrated TIP3P water molecules using periodic boundaries.⁸⁵

In the initial stage, the structure was minimized in two stages. In the first stage the protein was kept fixed, just minimizing the position of the water molecules and ions. In the second stage the full system was minimized. Subsequently, 2 ns molecular dynamics (MD) simulations were performed with the optimized structures. All simulations were carried out using the Sander module, implemented in the Amber 8 simulations package, with the Cornell force field.²⁹ Bond lengths involving hydrogens were constrained using the SHAKE algorithm⁸⁸ and the equations of motion were integrated with a 2 fs time-step using the Verlet leapfrog algorithm and the non-bonded interactions truncated with a 10 Å cutoff. The temperature of the system was regulated by the Langevin thermostat to maintain it at 333.15 K.⁸⁹⁻⁹¹ This

temperature was chosen because it is the temperature of the microbial niche occupied by variants of the enzyme CelE in the bacterium *Clostridium thermocellum*.⁹²

III.4.5.2 Docking experiments with the NMR solution structure and molecular dynamics

Models of the CtCBM11-cellohexaose and CtCBM11-cellobetraose complexes were calculated using the software HADDOCK (high ambiguity-driven protein docking) under the WeNMR Grid-enabled server^{34,35} using the energy minimized representative conformers of the NMR derived solution structures at 25 and 50 °C. The ambiguous interaction restraints (AIRs), i.e., active residues, were derived from the NMR titration data and the passive ones were chosen automatically (6.5 Å around the active residues). The HADDOCK docking protocol was performed as described elsewhere.^{35,93} The rigid body docking stage was performed 5 times, and the best resulting structure was saved. 1000 structures were generated at the rigid body docking stage, the best 200 of which were selected for further semiflexible refinement and refinement in explicit water. Non-bonded energies were calculated using the OPLSX non-bonded parameters.⁹⁴ Parameters for the ligands were obtained from Glycam Web.³⁶ The resulting solutions were clustered using a 2Å cut off and analyzed with the software PyMol1.4.1⁹⁵. Because all the structures in a given cluster were very similar, only the first one was subjected to molecular dynamics.

Molecular mechanics (MM) calculations and molecular dynamics (MD) simulations were performed with Amber11⁹⁶, using the ff99 (parm99)⁹⁷ and GLYCAM 06⁸⁷ force fields to parameterize both protein and carbohydrates, respectively. The carbohydrate ligand molecules were constructed with the “Glycam Biomolecule Builder” available online from the website of Woods group³⁶. The ligands were then minimized by molecular mechanics, through 1000 steps of the steepest descent method, followed by the conjugate gradient method until a convergence criterion of 0.0001 Kcal.mol⁻¹ was achieved. The complexes were immersed in isometric truncated octahedron TIP3P-water boxes of 12 Å and the proper number of counter ions was added using LeaP.

The MD simulations were performed using periodic boundary conditions following a five-step protocol: The first step consisted in a 20000 cycles of minimization to remove any possible unfavorable contacts between solvent and complexes. The first 3000 cycles of the minimization were performed with the steepest descent method, followed by the conjugate gradient method. In this step, the solute is restrained in the cartesian space using a harmonic potential (weight 500 kcal mol⁻¹.Å⁻²). Subsequently, a 10000 cycles of minimization (3000 steps of steepest descent and 7000 steps of conjugate gradient method) without restraints was performed. The systems

were then heated up to 298 K for 50 ps using a NVT ensemble and a weak positional restraint ($10 \text{ mol}^{-1} \cdot \text{\AA}^{-2}$) on the solute, to avoid wild structural fluctuations, using the Langevin thermostat with a collision frequency of 1 ps^{-1} . The positional restraints were removed and a molecular dynamics run in a NPT ensemble at 298 K for 500 ps was performed for equilibration at 1 atm with isotropic scaling and a relaxation time of 2 ps. Finally, NPT data production runs were carried out for 4 ns and the snapshots were saved to a trajectory file every 0.2 ps.

All bonds involving hydrogen atoms were constrained with the SHAKE algorithm⁸⁸ allowing the use of a 2 fs time step. The Particle Mesh Ewald method⁹⁸ was used to treat the long-range electrostatic interactions and the non-bonded van der Waals interactions were truncated with a 12 Å cut-off. The structural collected data were analyzed with the PTRAJ module as implemented in the AMBER package. The MD trajectories were also clustered by RMSD similarity using the average-linkage clustering algorithm.⁹⁹ As a representative co-conformation of a given simulation, the snapshot of the cluster with larger population was taken. Their structures were used to illustrate the structural features discussed in the main text.

III.5 References

1. Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008**, 275 (10), 2524.
2. Viegas, A.; Macedo, A. L.; Cabrita, E. J., Ligand-Based Nuclear Magnetic Resonance Screening Techniques. In *Ligand-macromolecular interactions in drug discovery : methods and protocols*, Roque, A. C. A., Ed. Springer: New York, 2010; pp 81.
3. Carvalho, A. L.; Goyal, A.; Prates, J. A. M.; Bolam, D. N.; Gilbert, H. J.; Pires, V. M. R.; Ferreira, L. M. A.; Planas, A.; Romao, M. J.; Fontes, C. M. G. A., The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *Journal of Biological Chemistry* **2004**, 279 (33), 34785.
4. Pires, V. M. Estrutura e função de módulos não catalíticos envolvidos na degradação da parede celular vegetal: o efeito de enzimas exógenas na valorização nutritiva de dietas à base de *Lupinus albus* para leitões. Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, 2008.
5. Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **2004**, 382 (Pt 3), 769.
6. Hashimoto, H., Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **2006**, 63 (24), 2954.
7. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R., The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* **2004**, 58, 521.
8. Tsukimoto, K.; Takada, R.; Araki, Y.; Suzuki, K.; Karita, S.; Wakagi, T.; Shoun, H.; Watanabe, T.; Fushinobu, S., Recognition of cellooligosaccharides by a family 28 carbohydrate-binding module. *Febs Letters* **2010**, 584 (6), 1205.

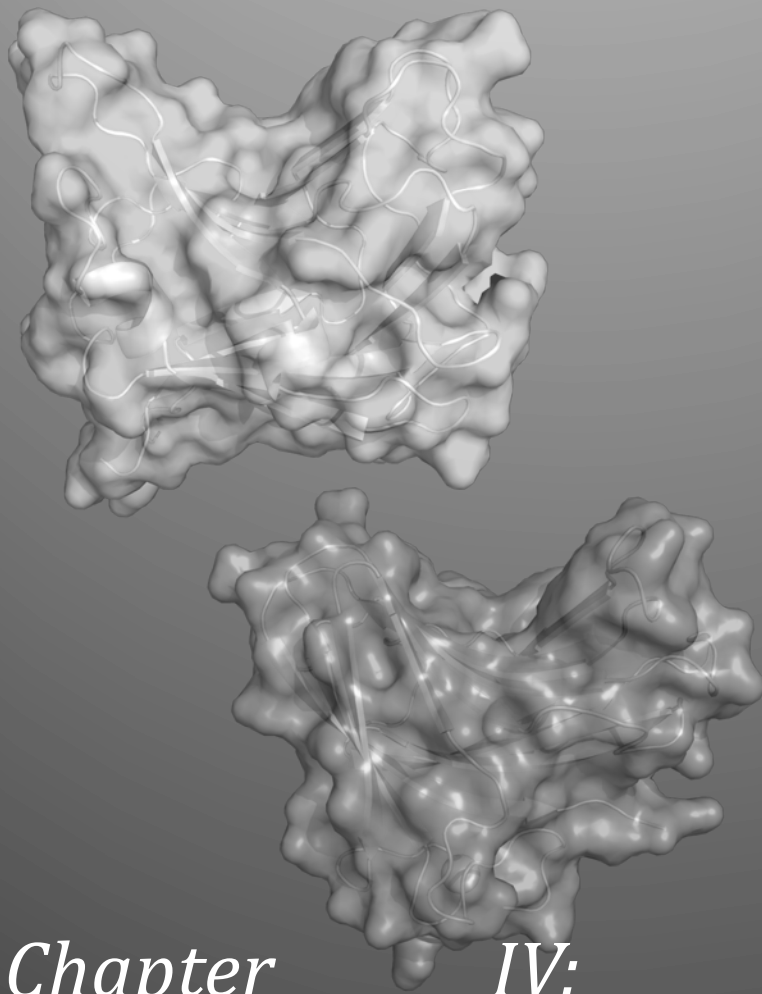
9. Pell, G.; Williamson, M. P.; Walters, C.; Du, H. M.; Gilbert, H. J.; Bolam, D. N., Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry* **2003**, *42* (31), 9316.
10. Xie, H. F.; Bolam, D. N.; Nagy, T.; Szabo, L.; Cooper, A.; Simpson, P. J.; Lakey, J. H.; Williamson, M. P.; Gilbert, H. J., Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry* **2001**, *40* (19), 5700.
11. Kirschner, K. N.; Woods, R. J., Solvent interactions determine carbohydrate conformation. *P Natl Acad Sci USA* **2001**, *98* (19), 10541.
12. Sugiyama, H.; Hisamichi, K.; Usui, T.; Sakai, K.; Ishiyama, J., A study of the conformation of beta-1,4-linked glucose oligomers, cellobiose to cellohexaose, in solution. *Journal of Molecular Structure* **2000**, *556* (1-3), 173.
13. Jamal-Talabani, S.; Boraston, A. B.; Turkenburg, J. P.; Tarbouriech, N.; Ducros, V. M. A.; Davies, G. J., Ab initio structure determination and functional characterization of CBM36: A new family of calcium-dependent carbohydrate binding modules. *Structure* **2004**, *12* (7), 1177.
14. Bolam, D. N.; Xie, H. F.; Pell, G.; Hogg, D.; Galbraith, G.; Henrissat, B.; Gilbert, H. J., X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *Journal of Biological Chemistry* **2004**, *279* (22), 22953.
15. Cook, W. J.; Bugg, C. E., Effects of Calcium Interactions on Sugar Conformation - Crystal-Structure of Beta-D-Fructose Calcium Bromide Dihydrate. *Acta Crystallogr B* **1976**, *32* (Feb15), 656.
16. Mayer, M.; Meyer, B., Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie-International Edition* **1999**, *38* (12), 1784.
17. Stockman, B. J.; Dalvit, C., NMR screening techniques in drug discovery and drug design. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2002**, *41* (3-4), 187.
18. Viegas, A.; Manso, J. o.; Nobrega, F. L.; Cabrita, E. J., Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *J Chem Educ* **2011**.
19. Najmudin, S.; Guerreiro, C. I. P. D.; Carvalho, A. L.; Prates, J. A. M.; Correia, M. A. S.; Alves, V. D.; Ferreira, L. M. A.; Romao, M. J.; Gilbert, H. J.; Bolam, D. N.; Fontes, C. M. G. A., Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *Journal of Biological Chemistry* **2006**, *281* (13), 8815.
20. Boraston, A. B.; Chiu, P.; Warren, R. A. J.; Kilburn, D. G., Specificity and affinity of substrate binding by a family 17 carbohydrate-binding module from *Clostridium cellulovorans* cellulase 5A. *Biochemistry* **2000**, *39* (36), 11129.
21. Flint, J.; Bolam, D. N.; Nurizzo, D.; Taylor, E. J.; Williamson, M. P.; Walters, C.; Davies, G. J.; Gilbert, H. J., Probing the mechanism of ligand recognition in family 29 carbohydrate-binding modules. *Journal of Biological Chemistry* **2005**, *280* (25), 23718.
22. Notenboom, V.; Boraston, A. B.; Chiu, P.; Frelove, A. C. J.; Kilburn, D. G.; Rose, D. R., Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: An X-ray crystallographic, thermodynamic and mutagenic study. *Journal of Molecular Biology* **2001**, *314* (4), 797.
23. Tomme, P.; Boraston, A.; Kormos, J. M.; Warren, R. A.; Kilburn, D. G., Affinity electrophoresis for the identification and characterization of soluble sugar binding by carbohydrate-binding modules. *Enzyme Microb Technol* **2000**, *27* (7), 453.
24. Brand, T.; Cabrita, E. J.; Berger, S., Intermolecular interaction as investigated by NOE and diffusion studies. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2005**, *46* (4), 159.
25. Johnson, C. S., Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1999**, *34* (3-4), 203.

26. Schumann, F. H.; Riepl, H.; Maurer, T.; Gronwald, W.; Neidig, K. P.; Kalbitzer, H. R., Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *J Biomol Nmr* **2007**, *39* (4), 275.
27. Fielding, L., NMR methods for the determination of protein-ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2007**, *51* (4), 219.
28. Boraston, A. B., The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven. *Biochemical Journal* **2005**, *385*, 479.
29. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* **1995**, *117* (19), 5179.
30. Cerqueira, N. M. F. S. A.; Bras, N. F.; Fernandes, P. A.; Ramos, M. J., MADAMM: A multistaged docking with an automated molecular modeling protocol. *Proteins-Structure Function and Bioinformatics* **2009**, *74* (1), 192.
31. Fernandez, M. D.; Canada, F. J.; Jimenez-Barbero, J.; Cuevas, G., Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate-aromatic interactions. *Journal of the American Chemical Society* **2005**, *127* (20), 7379.
32. Nagy, T.; Simpson, P.; Williamson, M. P.; Hazlewood, G. P.; Gilbert, H. J.; Orosz, L., All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *Febs Letters* **1998**, *429* (3), 312.
33. Boraston, A. B.; Nurizzo, D.; Notenboom, V.; Ducros, V.; Rose, D. R.; Kilburn, D. G.; Davies, G. J., Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *Journal of Molecular Biology* **2002**, *319* (5), 1143.
34. de Vries, S. J.; van Dijk, M.; Bonvin, A. M. J. J., The HADDOCK web server for data-driven biomolecular docking. *Nat. Protocols* **2010**, *5* (5), 883.
35. Dominguez, C.; Boelens, R.; Bonvin, A. M., HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **2003**, *125* (7), 1731.
36. WoodsGroup (2005-2012) GLYCAM Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (<http://www.glycam.com>).
37. Gilbert, H. J., The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiol* **2010**, *153* (2), 444.
38. Montanier, C.; Flint, J. E.; Bolam, D. N.; Xie, H. F.; Liu, Z. Y.; Rogowski, A.; Weiner, D. P.; Ratnaparkhe, S.; Nurizzo, D.; Roberts, S. M.; Turkenburg, J. P.; Davies, G. J.; Gilbert, H. J., Circular Permutation Provides an Evolutionary Link between Two Families of Calcium-dependent Carbohydrate Binding Modules. *Journal of Biological Chemistry* **2010**, *285* (41), 31742.
39. Bae, B.; Ohene-Adjei, S.; Kocherginskaya, S.; Mackie, R. I.; Spies, M. A.; Cann, I. K. O.; Nair, S. K., Molecular basis for the selectivity and specificity of ligand recognition by the family 16 carbohydrate-binding modules from *Thermoanaerobacterium polysaccharolyticum* ManA. *Journal of Biological Chemistry* **2008**, *283* (18), 12415.
40. Fry, S. C.; York, W. S.; Albersheim, P.; Darvill, A.; Hayashi, T.; Joseleau, J. P.; Kato, Y.; Lorences, E. P.; Maclachlan, G. A.; Mcneil, M.; Mort, A. J.; Reid, J. S. G.; Seitz, H. U.; Selvendran, R. R.; Voragen, A. G. J.; White, A. R., An Unambiguous Nomenclature for Xyloglucan-Derived Oligosaccharides. *Physiol Plantarum* **1993**, *89* (1), 1.
41. Del Bem, L. E.; Vincentz, M. G., Evolution of xyloglucan-related genes in green plants. *BMC Evol Biol* **2010**, *10*, 341.
42. Stone, M. J., NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Accounts Chem Res* **2001**, *34* (5), 379.
43. Jarymowycz, V. A.; Stone, M. J., Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem Rev* **2006**, *106* (5), 1624.
44. Chi, Y. H.; Kumar, T. K. S.; Chiu, I. M.; Yu, C., N-15 NMR relaxation studies of free and ligand-bound human acidic fibroblast growth factor. *Journal of Biological Chemistry* **2000**, *275* (50), 39444.

45. Lu, J. Y.; VanHalbeek, H., Molecular motions of a glycopeptide from human serum transferrin studied by C-13 nuclear magnetic resonance. *Biophys J* **1997**, 72 (1), 470.
46. Lipari, G.; Szabo, A., Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity. *Journal of the American Chemical Society* **1982**, 104 (17), 4546.
47. Lipari, G.; Szabo, A., Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .2. Analysis of Experimental Results. *Journal of the American Chemical Society* **1982**, 104 (17), 4559.
48. Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B., HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J Magn Reson* **2000**, 147 (1), 138.
49. Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A., Rotational diffusion anisotropy of human ubiquitin from ¹⁵N NMR relaxation. *Journal of the American Chemical Society* **1995**, 117 (50), 12562.
50. Dosset, P.; Hus, J. C.; Blackledge, M.; Marion, D., Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J Biomol Nmr* **2000**, 16 (1), 23.
51. Bernado, P.; Garcia de la Torre, J.; Pons, M., Interpretation of ¹⁵N NMR relaxation data of globular proteins using hydrodynamic calculations with HYDRONMR. *J Biomol Nmr* **2002**, 23 (2), 139.
52. Mandel, A. M.; Akke, M.; Palmer, A. G., Backbone Dynamics of Escherichia-Coli Ribonuclease Hi - Correlations with Structure and Function in an Active Enzyme. *Journal of Molecular Biology* **1995**, 246 (1), 144.
53. Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A., Rotational diffusion anisotropy of human ubiquitin from N-15 NMR relaxation. *Journal of the American Chemical Society* **1995**, 117 (50), 12562.
54. Sahu, S. C.; Bhuyan, A. K.; Udgaonkar, J. B.; Hosur, R. V., Backbone dynamics of free barnase and its complex with barstar determined by N-15 NMR relaxation study. *J Biomol Nmr* **2000**, 18 (2), 107.
55. Yang, D.; Kay, L. E., Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding. *Journal of Molecular Biology* **1996**, 263 (2), 369.
56. Creagh, A. L.; Ong, E.; Jervis, E.; Kilburn, D. G.; Haynes, C. A., Binding of the cellulose-binding domain of exoglucanase Cex from *Cellulomonas fimi* to insoluble microcrystalline cellulose is entropically driven. *P Natl Acad Sci USA* **1996**, 93 (22), 12229.
57. Boraston, A. B.; Warren, R. A. J.; Kilburn, D. G., beta-1,3-glucan binding by a thermostable carbohydrate-binding module from *Thermotoga maritima*. *Biochemistry* **2001**, 40 (48), 14679.
58. Chi, Y. H.; Kumar, T. K. S.; Kathir, K. M.; Lin, D. H.; Zhu, G. A.; Chiu, I. M.; Yu, C., Investigation of the structural stability of the human acidic fibroblast growth factor by hydrogen-deuterium exchange. *Biochemistry* **2002**, 41 (51), 15350.
59. Raschke, T. M.; Marqusee, S., Hydrogen exchange studies of protein structure. *Curr Opin Biotechnol* **1998**, 9 (1), 80.
60. Shan, X.; Gardner, K. H.; Muhandiram, D. R.; Rao, N. S.; Arrowsmith, C. H.; Kay, L. E., Assignment of ¹⁵N, ¹³C α , ¹³C β , and HN Resonances in an ¹⁵N,¹³C,²H Labeled 64 kDa Trp Repressor-Operator Complex Using Triple-Resonance NMR Spectroscopy and 2H-Decoupling. *Journal of the American Chemical Society* **1996**, 118 (28), 6570.
61. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletters on Protein Crystallography* **1992**, 26.
62. Evans, P. R., Scaling of MAD data. In *Proceedings of the CCP4 Study Weekend. Recent advances in phasing*, Winn, M., Ed. 1997; Vol. 33, pp 22.
63. Bailey, S., The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallogr D* **1994**, 50, 760.

64. Jancarik, J.; Kim, S. H., Sparse-Matrix Sampling - a Screening Method for Crystallization of Proteins. *J Appl Crystallogr* **1991**, *24*, 409.
65. Hwang, T. L.; Shaka, A. J., Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J Magn Reson Ser A* **1995**, *112* (2), 275.
66. Kessler, H.; Bermel, W.; Griesinger, C.; Kolar, C., The Elucidation of the Constitution of Glycopeptides by the Nmr Spectroscopic Coloc Technique. *Angew Chem Int Edit* **1986**, *25* (4), 342.
67. Stonehouse, J.; Adell, P.; Keeler, J.; Shaka, A. J., Ultrahigh-Quality Noe Spectra. *Journal of the American Chemical Society* **1994**, *116* (13), 6037.
68. Stott, K.; Stonehouse, J.; Keeler, J.; Hwang, T. L.; Shaka, A. J., Excitation Sculpting in High-Resolution Nuclear-Magnetic-Resonance Spectroscopy - Application to Selective Noe Experiments. *Journal of the American Chemical Society* **1995**, *117* (14), 4199.
69. Meyer, B.; Peters, T., NMR Spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie-International Edition* **2003**, *42* (8), 864.
70. Wu, D. H.; Chen, A. D.; Johnson, C. S., An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar-Gradient Pulses. *J Magn Reson Ser A* **1995**, *115* (2), 260.
71. Geyer, M.; Herrmann, C.; Wohlgemuth, S.; Wittinghofer, A.; Kalbitzer, H. R., Structure of the Ras-binding domain of RalGEF and implications for Ras binding and signalling. *Nat Struct Biol* **1997**, *4* (9), 694.
72. Atkins, P. W.; De Paula, J., *Atkins' Physical chemistry*. 9th ed.; Oxford University Press: Oxford ; New York, 2010.
73. Kay, L. E.; Torchia, D. A.; Bax, A., Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **1989**, *28* (23), 8972.
74. Farrow, N. A.; Muhandiram, R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Formankay, J. D.; Kay, L. E., Backbone Dynamics of a Free and a Phosphopeptide-Complexed Src Homology-2 Domain Studied by N-15 Nmr Relaxation. *Biochemistry* **1994**, *33* (19), 5984.
75. Barbato, G.; Ikura, M.; Kay, L. E.; Pastor, R. W.; Bax, A., Backbone Dynamics of Calmodulin Studied by N-15 Relaxation Using Inverse Detected 2-Dimensional Nmr-Spectroscopy - the Central Helix Is Flexible. *Biochemistry* **1992**, *31* (23), 5269.
76. Farrow, N. A.; Zhang, O. W.; Szabo, A.; Torchia, D. A.; Kay, L. E., Spectral Density-Function Mapping Using N-15 Relaxation Data Exclusively. *J Biomol Nmr* **1995**, *6* (2), 153.
77. Mulder, F. A. A.; van Tilborg, P. J. A.; Kaptein, R.; Boelens, R., Microsecond time scale dynamics in the RXR DNA-binding domain from a combination of spin-echo and off-resonance rotating frame relaxation measurements. *J Biomol Nmr* **1999**, *13* (3), 275.
78. Keller, R. The Computer Aided Resonance Assignment Tutorial. The Swiss Federal Institute of Technology, Zurich, 2004.
79. Teng, Q., *Handbook of structural biology : practical NMR applications*. Kluwer Academic/Plenum Publishers: New York, 2005.
80. Kroenke, C. D.; Rance, M.; Palmer, A. G., Variability of the N-15 chemical shift anisotropy in Escherichia coli ribonuclease H in solution. *Journal of the American Chemical Society* **1999**, *121* (43), 10119.
81. Bai, Y. W.; Milne, J. S.; Mayne, L.; Englander, S. W., Protein Stability Parameters Measured by Hydrogen-Exchange. *Proteins* **1994**, *20* (1), 4.
82. Zhang, Y.-Z. Protein and peptide structure and interactions studied by hydrogen exchange and NMR. Ph.D. Thesis, University of Pennsylvania, Philadelphia, 1995.
83. *InsightIII v. 2.3.0*, v. 2.3.0; Biosym Technologies: San Diego, 1993.
84. Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, H. M.; Wang, B.; Pearman, D. A.; Crowley, M.; S., B.; Tsui, V.; Gohlke, H.; Mongan, J.; Homak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*, San Francisco, 2004.

85. Asensio, J. L.; Jimenezbarbero, J., The Use of the Amber Force-Field in Conformational-Analysis of Carbohydrate Molecules - Determination of the Solution Conformation of Methyl Alpha-Lactoside by Nmr-Spectroscopy, Assisted by Molecular Mechanics and Dynamics Calculations. *Biopolymers* **1995**, *35* (1), 55.
86. Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J., Solvated ensemble averaging in the calculation of partial atomic charges. *J Comput Chem* **2001**, *22* (11), 1125.
87. Kirschner, K. N.; Woods, R. J., Quantum mechanical study of the nonbonded forces in water-methanol complexes. *J Phys Chem A* **2001**, *105* (16), 4150.
88. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* **1977**, *23* (3), 327.
89. Pastor, R. W.; Brooks, B. R.; Szabo, A., An Analysis of the Accuracy of Langevin and Molecular-Dynamics Algorithms. *Mol Phys* **1988**, *65* (6), 1409.
90. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W., Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers* **1992**, *32* (5), 523.
91. Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D., Langevin stabilization of molecular dynamics. *J Chem Phys* **2001**, *114* (5), 2090.
92. Freier, D.; Mothershed, C. P.; Wiegel, J., Characterization of Clostridium-Thermocellum Jw20. *Appl Environ Microb* **1988**, *54* (1), 204.
93. Tomaselli, S.; Ragona, L.; Zetta, L.; Assfalg, M.; Ferranti, P.; Longhi, R.; Bonvin, A. M. J. J.; Molinari, H., NMR-based modeling and binding studies of a ternary complex between chicken liver bile acid binding protein and bile acids. *Proteins: Structure, Function, and Bioinformatics* **2007**, *69* (1), 177.
94. Linge, J. P.; Williams, M. A.; Spronk, C. A. E. M.; Bonvin, A. M. J. J.; Nilges, M., Refinement of protein structures in explicit solvent. *Proteins: Structure, Function, and Bioinformatics* **2003**, *50* (3), 496.
95. Schrödinger, LLC *The PyMOL Molecular Graphics System*, 1.4.1; 2010.
96. Case, D. A.; Darden, T.; Cheatham III, T. E.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvai, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*, University of California: San Francisco, 2010.
97. Wang, J. M.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* **2000**, *21* (12), 1049.
98. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98* (12), 10089.
99. Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J Chem Theory Comput* **2007**, *3* (6), 2312.



Chapter IV: Molecular determinants of ligand specificity in CtCBM30 and CtCBM44

In this chapter I characterize the interaction of CtCBM30 and CtCBM44 with several ligands through STD-NMR and molecular docking. The results presented allowed a better understanding of the interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they can recognize and select their ligands. The results here presented are part of a manuscript in preparation.

Table of Contents

Summary	133
IV.1 Introduction	134
IV.2 Results and discussion.....	137
IV.2.1 Molecular determinants of ligand specificity	137
IV.2.1.1 Saturation transfer difference NMR (STD-NMR)	138
IV.2.1.2 Docking models for the interaction of CtCBM30 and CtCBM44 with cellooligosaccharides	145
IV.3 Conclusions.....	152
IV.4 Materials and methods	153
IV.4.1 Sources of sugars.....	153
IV.4.2 Molecular biology	154
IV.4.2.1 Recombinant protein production	154
IV.4.2.2 Protein expression and purification.....	154
IV.4.3 NMR spectroscopy.....	155
IV.4.3.1 Data acquisition.....	155
IV.4.3.2 STD-NMR studies.....	155
IV.4.4 Docking studies.....	155
IV.4.4.1 Preparation of the ligand pdb files	155
IV.4.4.2 Docking models for the interaction of CtCBM30 and CtCBM44 with cellooligosaccharides	156
IV.5 References.....	156

Summary

The focus of this chapter is on the family 30 and 44 carbohydrate-binding modules from *C. thermocellum* – CtCBM30 and CtCBM44 (**Figure IV.1**).^{1,2} These carbohydrate-binding modules belong to the bifunctional modular cellulase CtCel9D-Cel44A, which is one of the largest components of the cellulosome of *C. thermocellum*. The crystal structure of both proteins has been previously solved in the apo form and binding studies with several ligands provided some hints on the mechanism by which these proteins are able to select and bind to different substrates, namely xyloglucan. Nonetheless, no information could be obtained regarding the structure of the several complexes. In this chapter, I use STD-NMR and molecular docking to identify the molecular determinants of ligand specificity in CtCBM30 and CtCBM44 and to obtain models of both proteins in complex with several ligands (cellobiose, cellotetraose, cellopentaose, cellohexaose and laminarihexaose).

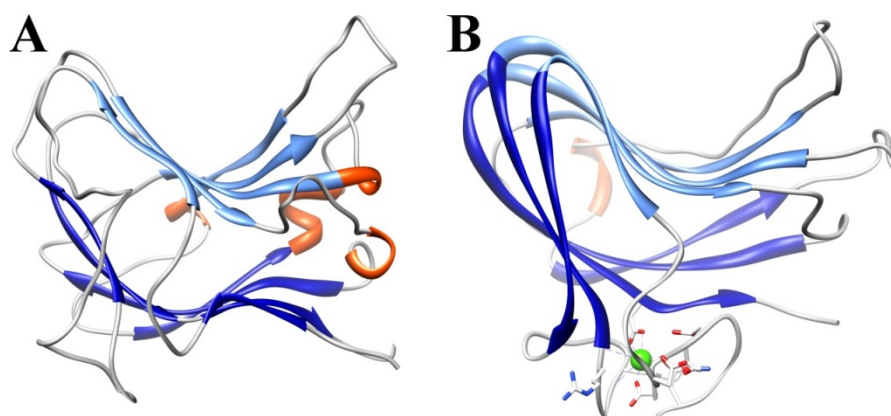


Figure IV.1: 3D structure of CtCBM30 (**A**) and CtCBM44 (**B**) obtained by X-ray crystallography.

Both the CtCBM30 (PDB code: 2c24) and CtCBM44 (PDB code: 2c26) structures reveal a classical distorted β -jelly roll that forms a convex side (light blue) and a concave side (dark blue). In the case of CtCBM44 the structure has one calcium ion, depicted as a green sphere (the residues that bind to calcium are depicted as sticks). The α -helical regions are depicted in red.

These studies revealed that the accommodation of branched ligands in the cleft of these proteins is dependent on the spatial arrangement of three solvent-exposed tryptophan residues in each protein (Trp27, Trp68 and Trp78 in CtCBM30 and Trp289, Trp194 and Trp198 in CtCBM44) and on the interactions that some polar residues make with the ligand. I found that in the case of CtCBM30 the two hydrogen bonds that Arg110 makes with the methylene hydroxyl group of the sugar unit at site $n+2$ provide an absolute requirement for an unsubstituted glucose moiety as does the presence of the sidechain of Lys112 near site n . Moreover, in CtCBM44 the hydrogen bonds that both Gln231 and Glu148 make with methylene hydroxyl group of the

sugar unit at site $n+3$ and the presence of the sidechain of Gln233 near site $n+1$, along with the hydrogen bond between N ϵ 1 of Trp198 and the methylene OH group at the same site also imply the presence of unsubstituted glucose moieties. In all other binding sites the methylene hydroxyl groups face the solvent, thus allowing these proteins to bind xyloglucan. These studies also showed that the optimal number of glucose units that can be accommodated by the cleft of these proteins is 4 in the case of CtCBM30 and 6 in the case of CtCBM44. Additionally, I have shown that the higher affinity that these proteins display for ligands longer than what they can accommodate may be related to the interaction of sugar units outside the binding cleft with polar residues of the protein.

IV.1 Introduction

CtCBM30 and CtCBM44 are part of the largest catalytic component of the cellulosome of *C. thermocellum*, designated CtCel9D-Cel44A.² This is a modular enzyme composed by an N-terminal family 30 carbohydrate-binding module (CtCBM30), two internal glycoside hydrolase domains (GH9 and GH44), a type I dockerin, a polycystic kidney-disease (PKD) module and the C-terminal family 44 carbohydrate-binding module (CtCBM44). CBM30, displays affinity for β -1,4-glucopolymers and plays a significant role in the function of GH9, a typical processive endoglucanase, whereas GH44 was assigned as displaying endo-xylanase activity.³

Both proteins belong to the Type B family (*see Chapter I, Section I.6.1.2*) and fold as a classical distorted β -jelly roll that forms a convex side and a concave side (**Figure IV.1**). In both proteins the concave side forms the sugar binding cleft and closely resembles the binding clefts in other Type B CBMs. In CtCBM30 this cleft is decorated by the residues Trp27, Trp68, Ile70, Leu72, Trp78, Asn79, Arg110, Lys112, Glu121, Asp123, Thr125, Ser166, and Arg168.¹ In the case of CtCBM44 the cleft is decorated with the side chains of Thr111, Ser113, Thr115, Glu144, Thr146, Glu148, Lys150, Asp152, Gln179, Tyr181, Met183, His185, Trp189, Trp194, Ser196, Trp198, Gln227, Gln231, and Gln233.¹ A closer inspection to both binding clefts shows that residues Trp27, Trp68 and Trp78 from CtCBM30 and Trp189, Trp194 and Trp198 from CtCBM44 form a solvent-exposed hydrophobic platform (**Figure IV.2**). In CtCBM30 this platform is about 20 Å in length while in CtCBM44 it is about 24 Å. Given the position of the aromatic residues and the length of both binding clefts, CtCBM30 is able to accommodate sugars with up to four units, binding at sites n , $n+1$ and $n+3$, whereas CtCBM44 is able to accommodate sugars with up to six units, binding at sites n , $n+2$ and $n+4$.

The importance of these residues was shown by producing CtCBM30 and CtCBM44 mutants (W27A, W68A and W78A for CtCBM30 and W27A, W68A and W78A for

CtCBM44). In these mutants the aromatic residues were changed to alanine and their biochemical properties investigated by affinity gel electrophoresis (AGE) and ITC (**Table IV.1**).

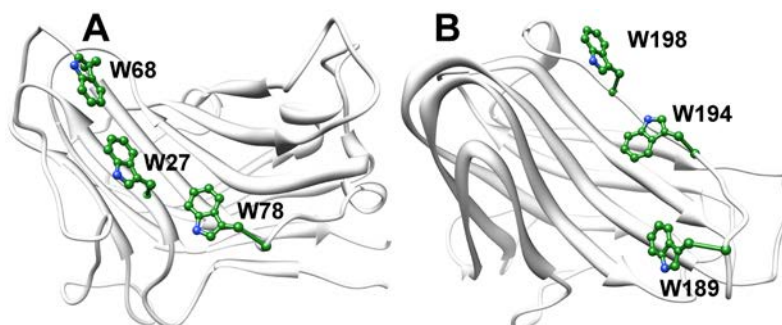


Figure IV.2: Solvent-exposed tryptophan residues at the surface of *CtCBM30* (A) and *CtCBM44* (B).

The secondary structural elements are shown as ribbons and depicted in white and the aromatic residues involved in ligand binding are shown in *ball* and *stick* and colored by heteroatom. PDB codes: 2c24 and 2c26 for *CtCBM30* and *CtCBM44*, respectively.

For *CtCBM30* it was shown that W27A and W68A displayed no affinity for decorated or undecorated ligands while W78A showed only reduced, but still significant affinity.¹ These results confirmed the involvement of these residues in ligand recognition. For *CtCBM44* the W194A mutant displayed no significant affinity for ligand while W189A and W198 showed a relatively modest decrease in affinity. Because Trp194 is the central aromatic residue of the binding site, it is possible that it makes a stronger hydrophobic interaction with the glucan than the flanking tryptophans, therefore, justifying the higher loss in affinity.¹ ITC studies with several ligands showed that the CBMs from *CtCel9D-Cel44A* recognize with equal efficiency linear and branched β -1,4-glucosidic ligands, such as cellulose and xyloglucan (**Table IV.1**).¹ The observation that both CBMs bind to xyloglucan provided the first evidence that these modules are able to accommodate the side chains of this decorated glucan. Neither of the CBMs displays affinity for galactomannan which may be because the axial O2 of mannose makes steric clashes with the protein at one or more sugar-binding sites.¹ Also, both proteins show reduced affinity for xylan, possibly pointing to the need for a direct interaction between the O6 of glucose and the protein, although the fact that the orientation of the aromatic platform in the binding site may act as a discriminative feature against ligands that adopt the 3-fold helical conformation displayed by the xylose polymer is also a possibility.⁴ Just like *CtCBM11*⁵, *CtCBM44* and *CtCBM30* also show increasing affinity (K_d) for the series cellotetraose, cellopentaose and celohexaose and a binding stoichiometry of 1. Moreover, the interaction of these modules with oligo- and polysaccharides is also enthalpy-driven (i.e., exothermic), with entropy making an unfavorable contribution to ligand binding.¹ As discussed in the previous

chapter, this is typical of the binding of proteins to soluble saccharides.⁶⁻⁹ The PKD module at the *N-terminus* of *CtCBM44* does not contribute to carbohydrate recognition as demonstrated by affinity gel electrophoresis (AGE) experiments with *CtCBM44* alone and attached to the PKD module.

Table IV.1: Quantitative assessment of *CtCBM30* and *CtCBM44* binding to oligosaccharides and polysaccharides as determined by ITC.¹

<i>Protein</i>	<i>Ligand</i>	$K_a \times 10^4$ (M^{-1})	ΔG ($kcal\ mol^{-1}$)	ΔH ($kcal\ mol^{-1}$)	ΔTS ($kcal\ mol^{-1}$)	n^a
CBM30	Cellohexaose ^b	6.4 ± 0.8	-6.2 ± 0.1	-8.0 ± 0.5	-1.8	1.2 ± 0.1
CBM44	Cellohexaose	72.8 ± 7.2	-8.0 ± 0.1	-15.9 ± 0.3	-7.9 ± 0.4	1.1 ± 0.1
CBM30	Cellopentaose ^b	1.2 ± 0.8	-5.3 ± 0.3	-6.9 ± 0.5	-1.7	1.3 ± 0.1
CBM44	Cellopentaose	6.6 ± 1.3	-6.6 ± 0.1	-14.5 ± 0.5	-7.9 ± 0.6	1.0 ± 0.0
CBM30	Xyloglucan	7.2_1.4	-6.6 ± 0.1	-10.4 ± 0.3	-3.8 ± 0.2	1.0 ± 0.0
CBM44	Xyloglucan	81.6 ± 9.8	-8.1 ± 0.1	-16.3 ± 0.6	-8.2 ± 0.7	1.0 ± 0.0
CBM30	HEC	4.5 ± 0.5	-6.3 ± 0.1	-10.0 ± 0.2	-3.7 ± 0.3	1.0 ± 0.0
CBM44	HEC	12.2 ± 3.3	-6.9 ± 0.2	-12.5 ± 0.5	-5.6 ± 0.7	1.0 ± 0.0
CBM30	β-Glucan	2.8 ± 0.3	-6.1 ± 0.1	-11.2 ± 0.2	-5.1 ± 0.3	1.0 ± 0.0
CBM44	β-Glucan	22.5 ± 3.6	-7.3 ± 0.1	-17.7 ± 0.6	-10.4 ± 0.7	1.0 ± 0.0
CBM30	Lichenan	3.6 ± 0.4	-6.2 ± 0.1	-11.5 ± 0.4	-5.3 ± 0.4	1.0 ± 0.0
CBM44	Lichenan	12.3 ± 2.8	-6.9 ± 0.1	-22.6 ± 1.3	-15.7 ± 1.4	1.0 ± 0.0
CBM30	Glucomannan	~0.4 ^c	-	-	-	-
CBM44	Glucomannan	9.0 ± 2.0	-6.7 ± 0.1	-15.9 ± 0.8	-9.2 ± 0.9	1.0 ± 0.0

^a Number of binding sites on the protein.

^b Data are from Arai *et al.*¹⁰

^c Value is an estimate because affinity was too low to obtain accurate value.

Interestingly, contrary to most Type B CBMs, *CtCBM30* does not contain any calcium ion in its structure,¹ showing that, although calcium is a common feature in these thermostable proteins, it is not fundamental for their stability. On the other hand, the structure of *CtCBM44* reveals the presence of one calcium ion with octahedral coordination bound to residues Asn101, Lys130, and Arg133 (main chain O atoms), Asp96 (Oε1), Glu103 (Oδ1), and Asp245 (bidentate coordination from Oε1 and Oε2). As in *CtCBM11*, the calcium ion has a structural role as it is solvent-inaccessible and its removal decreases the protein's melting temperature by 23 °C.¹

The CBMs from *CtCel9D-Cel44A* recognize undecorated and highly branched β-1,4-glucosidic ligands, yet, the structural determinants that may allow the binding of these CBMs (and all other CBMs in general), at a single binding site, to such different polysaccharides remain unknown. Xyloglucan is the most abundant hemicellulosic polysaccharide in the

primary walls of dicots and non-graminaceous monocots and may account for 20-40% of the dry weight of the primary wall. Xyloglucan has a backbone composed of β -1,4-linked glucose residues and up to 75% of these residues are substituted at O6 with mono-, di-, or triglycosyl side chains.^{11,12}

In order to understand the structural properties that determine the promiscuity in ligand recognition by these CBMs, I used an NMR approach combined with computational studies, to identify the molecular determinants of ligand specificity of *CtCBM30* and *CtCBM44*. I have used the STD-NMR technique to identify the atoms of the ligands (cellobiose, cellotetraose, cellohexaose, cellopentaose and laminarihexaose) that make intimate contact to the proteins upon binding and epitope map them in the ligand structures. Using the obtained STD-NMR information and the previously determined crystal structures of both proteins I calculated the models of *CtCBM30* bound to cellotetraose and cellohexaose and *CtCBM44* bound to cellotetraose, cellopentaose and cellohexaose. All the obtained models are in good agreement with the STD-NMR results. These studies provided localized structural information about the binding pocket of both *CtCBM30* and *CtCBM44* allowing a better understanding of the interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they are able to recognize and select linear and decorated β -1,4-glucans.

IV.2 Results and discussion

IV.2.1 Molecular determinants of ligand specificity

One of the key unresolved issues with respect to *CtCBM30* and *CtCBM44* is how these proteins interact with highly decorated polysaccharides; xyloglucan has a backbone composed of β -1,4-linked glucose residues and up to 75% of these residues are substituted at O6 with mono-, di-, or triglycosyl side chains.^{11,12}

In order to understand the structural properties that govern the promiscuity in ligand recognition by *CtCBM30* and *CtCBM44*, I used STD-NMR to study the interaction of these proteins with several ligands (cellobiose, cellotetraose, cellohexaose and laminarihexaose). This allowed me to ascertain about the influence of the length of the cellooligosaccharide chain (2, 4 or 6 glucose units) and the presence of β -1,3 glycosidic bonds in the binding. Furthermore, for the ligands that interacted with the proteins, I was able to identify which ligand atoms are more important for the complex formation.

Using this information and the X-ray structures of *CtCBM30* and *CtCBM44* (PDB codes: 2c24 and 2c26 for *CtCBM30* and *CtCBM44*, respectively), I calculated a model of the several

protein/ligand complexes. The docking procedure was driven with HADDOCK.^{13,14} Examination of the several CBM-carbohydrate complexes provided the first hints of how highly decorated polysaccharides can be accommodated by these xyloglucan-binding modules.

Experimental details of all the techniques applied are explained in Materials and methods, Sections IV.4.3 and IV.4.4 and further explanation of the theory behind the STD-NMR experiments is given in Chapter VII, Section VII.5.1.

IV.2.1.1 Saturation transfer difference NMR (STD-NMR)

STD-NMR spectroscopy was applied to analyze the binding of cellobiose (**Figure IV.3**), cellotetraose (**Figure IV.4**), cellohexaose (**Figure IV.5**) and laminarihexaose (**Figure IV.6**) to *CtCBM30* and *CtCBM44*. All the spectra were acquired at 298 K in a Bruker AvanceIII spectrometer, operating at a frequency of 600 MHz with a 100-fold excess of ligand over the protein.

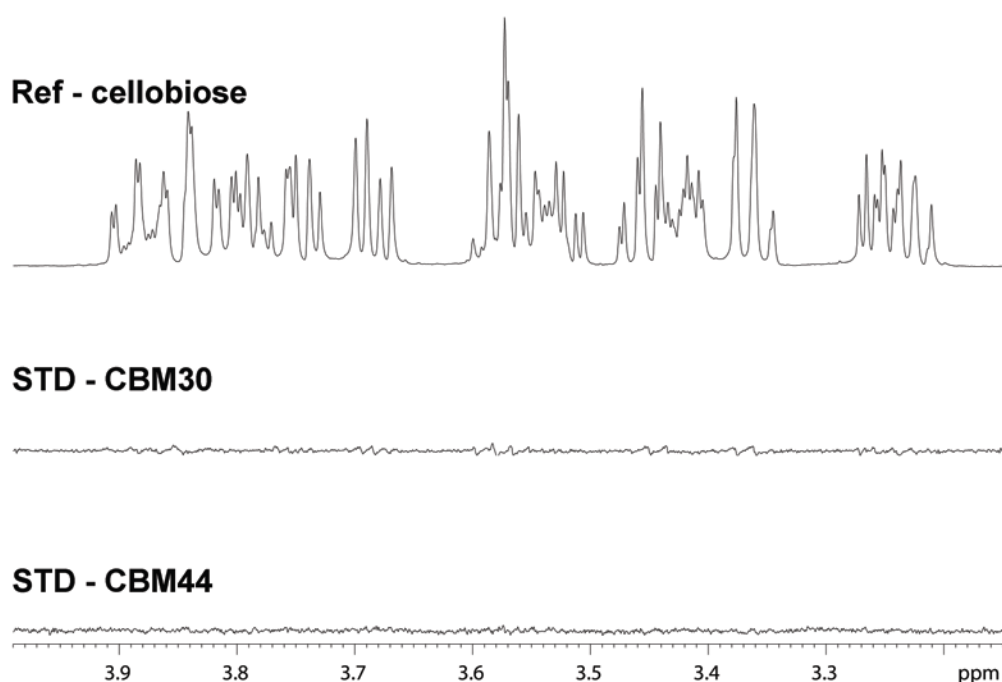


Figure IV.3: STD-NMR of cellobiose with *CtCBM30* and *CtCBM44*.

Top - Reference ¹H-NMR cellobiose spectrum. Middle - STD-NMR spectra of the solution of cellotetraose (3 mM) with *CtCBM30* (30 μM). Bottom - STD-NMR spectra of the solution of cellotetraose (2 mM) with *CtCBM44* (20 μM). No signals appear in both the STD-NMR spectra, indicating that there is no interaction between cellobiose and either of the proteins.

The STD-NMR spectrum of the cellobiose with *CtCBM30* and *CtCBM44* is presented in **Figure IV.3** along with the sugar's reference spectrum. Similar to the results obtained with *CtCBM11*, there is an absence of signals in both STD-NMR spectra. This absence of signals could be the result of an extremely strong and almost irreversible complex or an indication that

there is no interaction between these proteins and cellobiose or that it is a very weak interaction ($K_a < 10^3 \text{ M}^{-1}$). The last hypothesis seems to be the more plausible and is in good agreement with absence of affinity displayed by these proteins to cellotriose^{1,2} and the general lack of specificity of Type B CBMs towards small sugars^{15,16} (see Chapter I).

Unlike the data with cellobiose, the STD-NMR spectrum of CtCBM30 with cellotetraose (Figure IV.4 - middle) clearly shows some signals which is an indication that cellotetraose binds to this module.

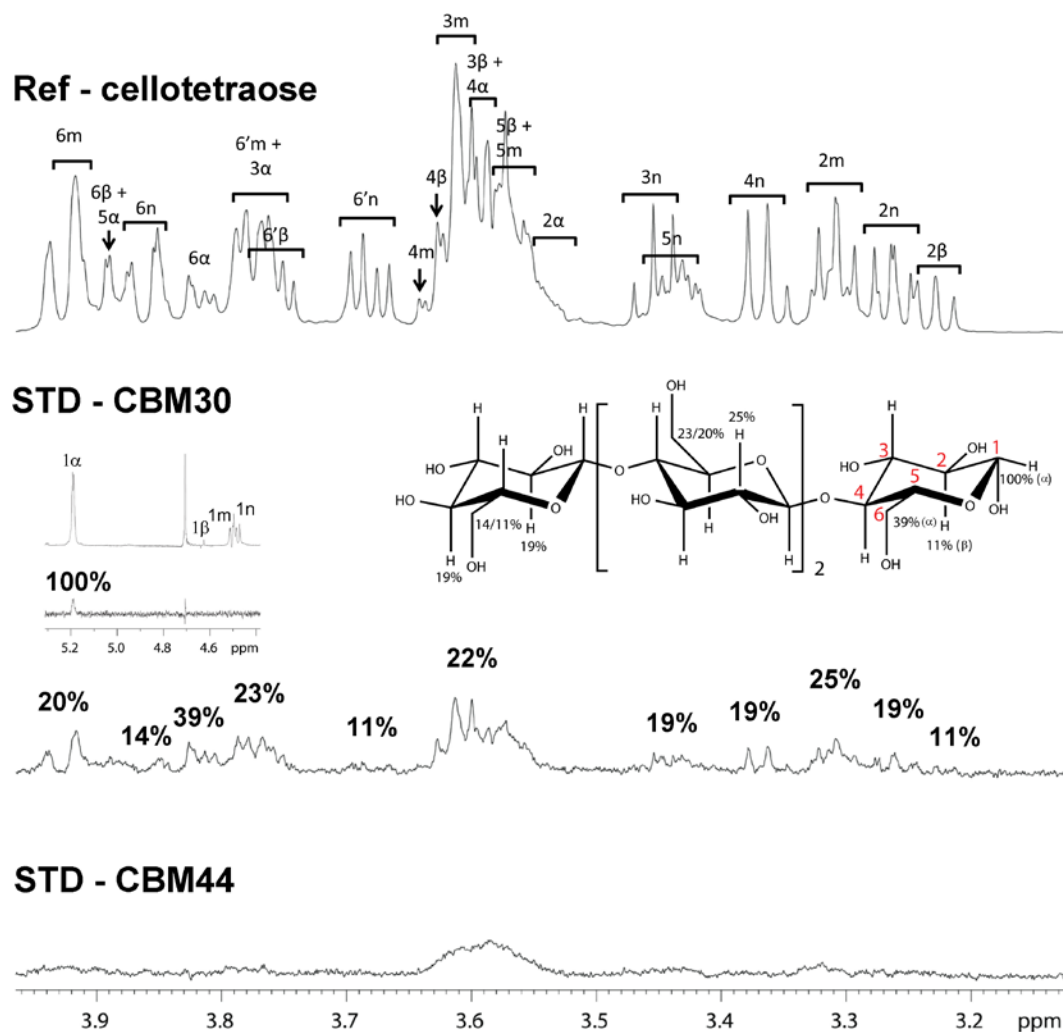


Figure IV.4: STD-NMR of cellotetraose with CtCBM30 and CtCBM44.

Top - Reference ^1H -NMR cellotetraose spectrum. Middle - STD-NMR spectra of the solution of cellotetraose (3 mM) with CtCBM30 (30 μM). Bottom - STD-NMR spectra of the solution of cellotetraose (2 mM) with CtCBM44 (20 μM). The binding epitope for the interaction of cellotetraose with CtCBM30 is shown above each peak and mapped in the structure of the sugar.

In a similar way as for the interaction with CtCBM11 (see Chapter III) it is possible to epitope map the interaction in the ligand structure. In general, in spite of the low values, all cellotetraose glucose units show some degree of saturation indicating that the whole molecule is

in contact with CtCBM30. Interestingly, for the interaction of cellotetraose with CtCBM30 the maximum A_{STD} value is found for the anomeric proton of the reducing end of the sugar in the α -conformation (H1 α). Moreover, the second highest A_{STD} value (39%) is also found for the methylene protons of the reducing end in the α -conformation (H6 α). This, together with the low A_{STD} value for the β -conformation, may indicate that this protein displays a favored affinity for the sugar in the α -conformation. For the protons of the central glucose units the A_{STD} values are between 20 and 25%, indicating a lower contribution for binding. The STD epitope map of cellotetraose upon binding to CtCBM30 is shown in **Figure IV.4 – middle** and resumed in **Table IV.2**.

The fact that CtCBM30 displays a preference for the reducing end of cellotetraose in the α -conformation may be related to the topology of the protein's binding site. As shown in **Figure IV.2 - A**, the three solvent-exposed tryptophan residues (Trp27, Trp68 and Trp78) at the surface of CtCBM30 form a platform that faces the ligand. In the α -conformation, the anomeric hydroxyl group of the reducing end of cellotetraose will stay in a privileged position to interact with the indole ring of either of the flanking tryptophan residues (Trp68 or Trp78) through hydrophobic contacts (*see Section IV.2.1.2*). This interaction should stabilize the complex with the α -conformation of cellotetraose, thus promoting binding with CtCBM30. Using only STD-NMR it is not possible to identify which tryptophan residue is interacting with the non-reducing end of cellotetraose.

Regarding the interaction of cellotetraose with CtCBM44 (**Figure IV.4 - bottom**), the STD-NMR spectrum shows only a very weak transfer of saturation. Due to the large broadening and overlapping of the signals it is not possible to distinguish the protons involved in this interaction. This broad signal corresponds to protons H4m, H4 β , H3m, H3 β , H4 α , H5 β and H5m and its presence means that CtCBM44 does recognize cellotetraose but binding is very weak, which is in accordance with previous results.¹ This lack of a significant interaction results from the disposition of the solvent-exposed tryptophan residues (Trp189, Trp194 and Trp198). Looking at the three-dimensional arrangement of these residues we see that they can bind sugars units at sites n , $n+2$ and $n+4$.¹ Consequently, for the interaction with cellotetraose only two units would participate in binding, thus justifying the low affinity. This means that CtCBM44 is only able to bind to cellooligosaccharides with five or more units, which is in accordance with the obtained STD-NMR results.

The interaction of cellohexaose with CtCBM30 is very similar to the interaction of cellotetraose (**Figure IV.5 – middle** and **Table IV.2**). The main difference is that, in this case, there are no signals arising from the interaction of the reducing end of cellohexaose with the protein.

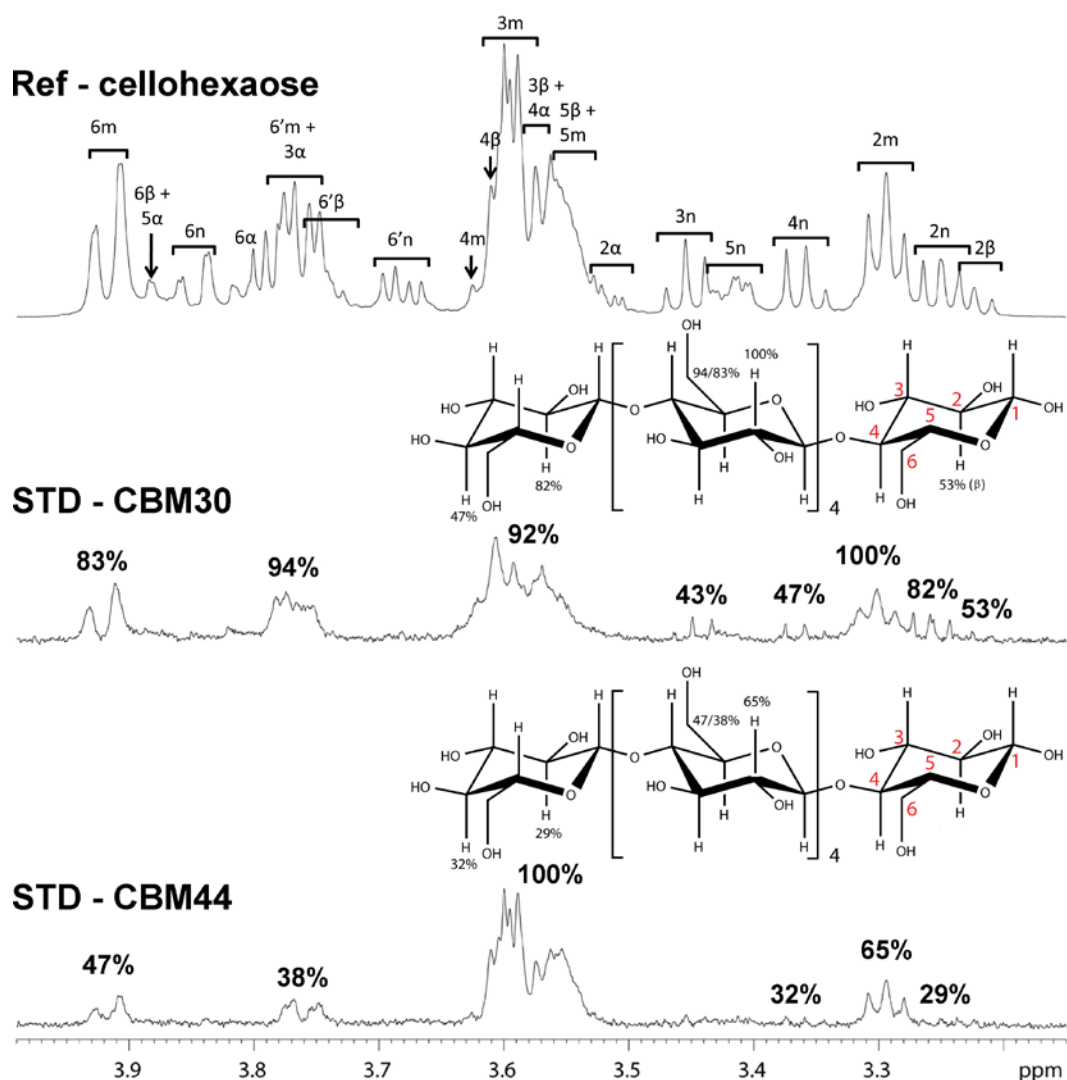


Figure IV.5: STD-NMR of cellohexaose with *CtCBM30* and *CtCBM44*.

Top - Reference $^1\text{H-NMR}$ cellohexaose spectrum. Middle - STD-NMR spectra of the solution of cellohexaose (3 mM) with *CtCBM30* (30 μM). Bottom - STD-NMR spectra of the solution of cellohexaose (2 mM) with *CtCBM44* (20 μM). The binding epitope for the interaction of cellohexaose with *CtCBM30* is shown above each peak and mapped in the structure of the sugar.

Considering that the binding cleft of *CtCBM30* can only accommodate up to four sugar units, these results indicate that, for longer saccharide chains, the reducing end rests outside the binding cleft. In this case two hypothesis arise: i) either the reducing end and the preceding unit stay outside the cleft or ii) both ends stay outside the cleft and the protein binds only to the central units. For this interaction the maximum A_{STD} value is found for protons H2 from the central glucose units (H2m – 100%) and the methylene protons H6 and H6' of the same units (H6m and H6'm – 83 and 94%, respectively). As seen for the interaction of *CtCBM11* with cellotetraose and cellohexaose, high values of A_{STD} are also obtained for the protons whose signals appear in the region between 3.50 and 3.64 (H4m, H4 β , H3m, H3 β , H4 α , H5 β and H5m). Again, due to extensive overlapping it is not possible to distinguish the individual

contributions of these protons. Additionally, protons H2 and H4 of the non-reducing end also show STD signals, although very weak (29 and 32%, respectively). Considering the first hypothesis, I would expect a much higher intensity of the resonances corresponding to the non-reducing end, as seen for the interaction with cellotetraose where the intensity of these signals is similar to the ones of the central glucose units. However, if I consider the second hypothesis, the protein will bind to the central glucose units leaving the extremities outside the binding cleft but still close enough to receive some degree of saturation, thus explaining the low A_{STD} values displayed by protons H2n and H4n. The binding to the central sugar units is a common feature among CBMs^{15,17,18} and may be the mechanism by which they are able to bind ligands that extend outside the binding cleft.¹ Another characteristic that this interaction shares with the interaction of *CtCBM11* with cellotetraose and cellohexaose is the fact that one of the diastereotopic methylene protons shows a relatively more intense peak in the STD spectrum than the other (about 10%).¹⁶ This is indicative of a precise orientation of the methylene groups upon binding to the protein.

Concerning the interaction of cellohexaose with *CtCBM44*, the STD-NMR spectrum (**Figure IV.5 - bottom** and **Table IV.2**) is clearly different from the one obtained with cellotetraose. To begin with, there is an obvious different response from the several sugar units that allow epitope-mapping the interaction.

The maximum A_{STD} value is obtained for the protons whose signals appear in the region between 3.64 and 3.50 ppm (protons H4m, H4 β , H3m, H3 β , H4 α , H5 β and H5m), which cannot be resolved. The other signals that appear correspond to protons H2 and H6 of the central glucose units (47 and 38%) and protons H2 and H4 from the non-reducing end (H2n – 29% and H4n – 32%). No individual signals were detected for protons of the reducing end of the saccharide, which may indicate that this unit does not contribute significantly to binding. Moreover, experiments with mutant proteins¹ showed that removal of the two flanking tryptophan residues (Trp189 and Trp198) caused only a relatively modest decrease in the affinity. This is in accordance with the low A_{STD} values obtained for protons of the non-reducing end.

As was observed for the interaction of cellohexaose with *CtCBM11* and *CtCBM30*, also here the diastereotopic protons of the methylene groups of the central glucose units show different relative STD intensities (H6m – 47% and H6'm – 38%) suggesting that the predicted well-defined geometry upon binding is a common feature of these proteins. This defined geometry may act as a determinant of specificity by discriminating against ligands that do not adopt this conformation.

The tighter binding of *CtCBM44* to cellohexaose than to cellotetraose¹ (**Table III.1**) is related to the geometry of the binding cleft, as the extra two sugar units promote the formation

of hydrophobic interactions with all the three solvent-exposed tryptophan residues (*see Section IV.2.1.1*).

Regarding the STD-NMR results with laminarihexaose (**Figure III.8**), only very low intensity signals appear in the STD spectra, as depicted from the A_{STD} values in **Table IV.2**, which are about 75% lower than the corresponding ones for cellobiohexaose.

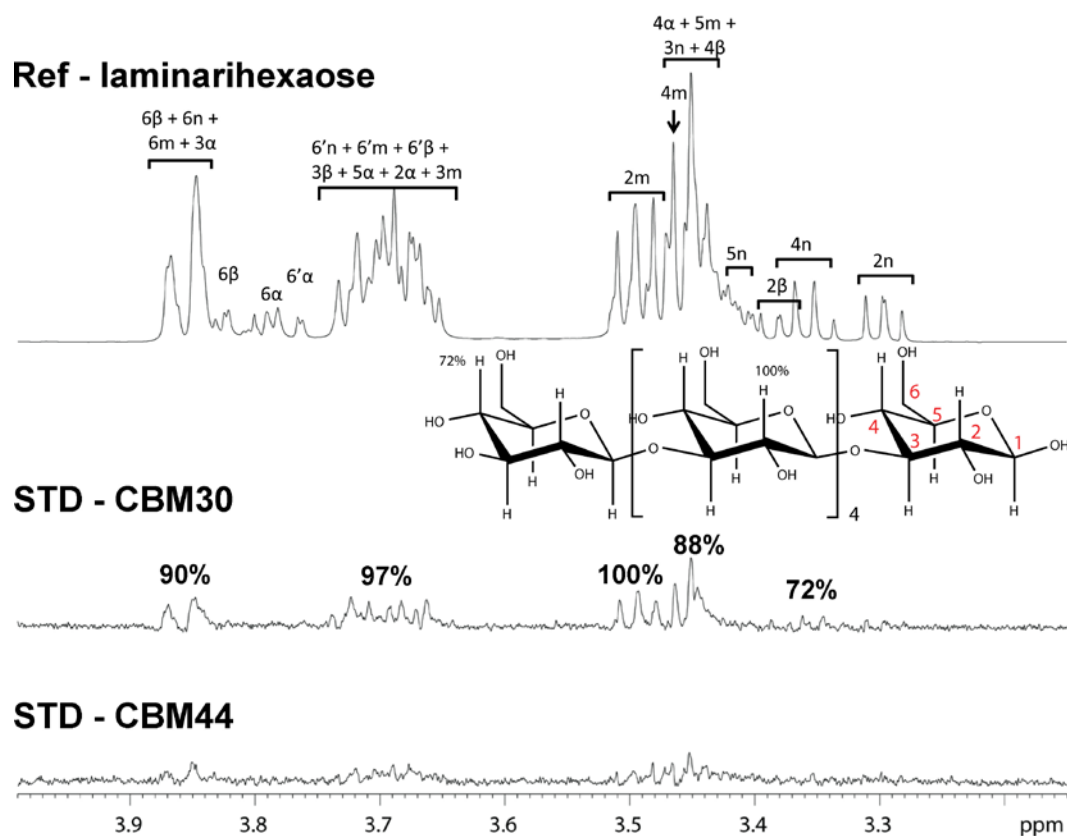


Figure IV.6: STD-NMR of laminarihexaose with CtCBM30 and CtCBM44.

Top - Reference $^1\text{H-NMR}$ laminarihexaose spectrum. Middle - STD-NMR spectra of the solution of laminarihexaose (3 mM) with CtCBM30 (30 μM). Bottom - STD-NMR spectra of the solution of laminarihexaose (2 mM) with CtCBM44 (20 μM). Only very low intensity signals, probably deriving from non-specific contacts, appear in the spectrum.

These signals can emerge from non-specific contacts between the proteins and laminarihexaose and may not be indicative of specific binding. Unfortunately, this is a major limitation of the STD-NMR technique as it is not able to distinguish specific from non-specific binding^{19,20} (*as explained in Chapter VII*). Because of the wide area of the binding cleft in both proteins it is possible that some contacts between laminarihexaose and the aromatic residues are established, giving rise to the observed signals in the STD-NMR spectra. This interaction is stronger for CtCBM30 than for CtCBM44 (for which A_{STD} values couldn't be measured due to the very weak intensities of the signals). For CtCBM30 it is even possible to do some epitope mapping (**Figure IV.6 – middle** and **Table IV.2**). Due to an extensive overlapping of the

resonances of laminarihexaose, the only signal that can be isolated belongs to protons H2 from the central glucose units (H2m). Moreover, this is also the signal with the highest STD intensity (100%). This is similar to what happens with the β -1,4-linked saccharides, indicating that this unspecific binding may occur in a similar fashion to the natural binding. Nonetheless, the hydrophobic platform formed by the tryptophan residues can only engage with slightly twisted ligands of β -1,4-linked saccharides and not with helical β -1,3-glucans. Therefore, it is my opinion that this affinity for β -1,3-linked ligands does not have any biological meaning.

Table IV.2: Amplification factors and epitope mapping for the interaction between CtCBM30 and CtCBM44 with cellotetraose, cellohexaose and laminarihexaose.

		<i>A_{STD} / Epitope mapping (%)</i>						
		1	2	3	4	5	6	6'
		<i>CtCBM30/Cellotetraose</i>						
α	2.56 / 100	-	0.60 / 23 ^c	0.56 / 22 ^b	-	0.99 / 39	-	
β	-	0.29 / 11	0.56 / 22 ^b	0.56 / 22 ^b	0.56 / 22 ^b	-	0.60 / 23 ^c	
<i>m</i>	-	0.64 / 25	0.56 / 22 ^b	0.56 / 22 ^b	0.56 / 22 ^b	0.52 / 20	0.60 / 23 ^c	
<i>n</i>	-	0.48 / 19	0.48 / 19 ^a	0.48 / 19	0.48 / 19 ^a	0.35 / 14	0.29 / 11	
		<i>CtCBM30/Cellohexaose</i>						
α	-	-	0.79 / 94 ^f	0.78 / 92 ^e	-	-	-	
β	-	0.45 / 53	0.78 / 92 ^e	0.78 / 92 ^e	0.78 / 92 ^e	-	0.79 / 94 ^f	
<i>m</i>	-	0.84 / 100	0.78 / 92 ^e	0.78 / 92 ^e	0.78 / 92 ^e	0.70 / 83	0.79 / 94 ^f	
<i>n</i>	-	0.69 / 82	0.37 / 43 ^d	0.48 / 47	0.37 / 43 ^d	-	-	
		<i>CtCBM44/Cellohexaose</i>						
α	-	-	1.15 / 38 ⁱ	3.02 / 100 ^h	-	-	-	
β	-	-	3.02 / 100 ^h	3.02 / 100 ^h	3.02 / 100 ^h	-	1.15 / 38 ⁱ	
<i>m</i>	-	1.97 / 65	3.02 / 100 ^h	3.02 / 100 ^h	3.02 / 100 ^h	1.42 / 47	1.15 / 38 ⁱ	
<i>n</i>	-	0.88 / 29	1.30 / 43 ^g	0.98 / 32	1.30 / 43 ^g	-	-	
		<i>CtCBM30/Laminarihexaose</i>						
α	-	-	0.17 / 90 ^l	0.16 / 88 ^j	0.18 / 97 ^k	-	-	
β	-	-	0.18 / 97 ^k	0.16 / 88 ^j	-	0.17 / 90 ^l	0.18 / 97 ^k	
<i>m</i>	-	0.19 / 100	0.18 / 97 ^k	0.16 / 88 ^j	0.16 / 88 ^j	0.17 / 90 ^l	0.18 / 97 ^k	
<i>n</i>	-	-	0.16 / 88 ^j	0.13 / 72	-	0.17 / 90 ^l	0.18 / 97 ^k	

a, b, c, d, e, f, g, h, i, j, k, l – These peaks are overlapped

IV.2.1.2 Docking models for the interaction of CtCBM30 and CtCBM44 with cellooligosaccharides

Since no structures of CtCBM30 or CtCBM44 with a bound ligand are available, in order to better interpret the STD-NMR results I have used the software HADDOCK^{13,14} to calculate models of the CtCBM30/cellotetraose, CtCBM30/cellohexaose, CtCBM44/cellotetraose, CtCBM44/cellohexaose and CtCBM44/cellopentaose complexes (see *Materials and methods*, Section IV.4.4.2). For the docking experiments I used the X-ray structures of CtCBM30 and CtCBM44 (PDB codes: 2c24 and 2c26, respectively) and the sugar parameters obtained from Glycam Web²¹ (see *Materials and methods*, Section IV.4.4.1). These studies provided localized structural information of the binding pocket of both CtCBM30 and CtCBM44 allowing a better understanding of how these proteins recognize and bind to their substrates. All the obtained models are in good agreement with the STD-NMR results.

IV.2.1.2.1 Model of CtCBM30 bound to cellotetraose

According to the STD-NMR results for the interaction of CtCBM30 with cellotetraose (**Figure IV.4 – middle**), the α -conformation of the sugar is preferred against the naturally more abundant β -conformation. Therefore this was the one used in the docking experiments. The model of the structure of CtCBM30 in complex with cellotetraose is shown in **Figure IV.7**.

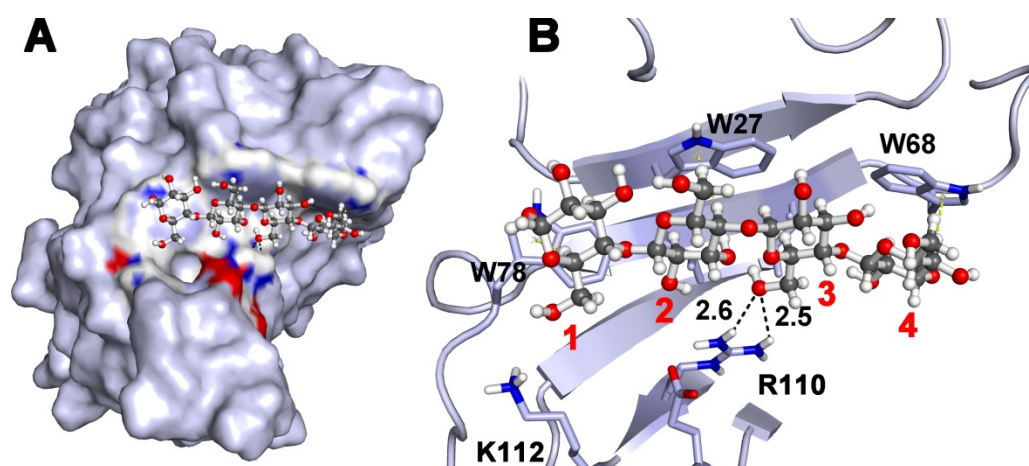


Figure IV.7: Model of the structure of CtCBM30 in complex with cellotetraose.

A) Surface representation of CtCBM30 bound to cellotetraose. B) Ribbon representation of the CtCBM30 binding cleft bound to cellotetraose. The ligand is depicted in ball-and-stick and the interacting residues are depicted as sticks. Atoms are colored by heteroatom. Hydrogen bonds are represented as black dashes and CH- π interactions are represented as yellow dashes. Glucosyl moieties (in red) are numbered as recommended by IUPAC-IUB JCBN (1983).²⁴

The structure shows that, as predicted by the arrangement of the solvent-exposed tryptophan residues¹, the sugar binds in units n , $n+1$ and $n+3$. These residues are placed along one face of the ligand-binding cleft and engage in hydrophobic interactions with all of the oligosaccharide units (**Figure IV.7 - B**), which is in good agreement with the obtained STD-NMR results. Regarding the orientation of the sugar in the binding cleft, STD-NMR alone can't give a straight answer. Four hypotheses are possible depending on which face of the sugar and tryptophan residue are interacting: either the β -face* of the reducing end is interacting with i) Trp68 or ii) Trp78 or the α -face of the reducing end is interacting with iii) Trp68 or with iv) Trp78.

Docking experiments showed that all orientations are possible and give very similar results with only little energy differences amongst them. In fact, the absence of a specific orientation in the ligand chain has already been seen^{25,26} and predicted²⁷ in other CBMs. Therefore, I selected the orientation that best described my experimental results and with the lowest energy – the α -face of the reducing end interacting with Trp78. Nonetheless I should stress out that, in principle, all four hypotheses can occur in solution as there is no impairment for any of them. All of the four possible models contradict the previous supposition that all three tryptophans would bind the same face of the sugars¹. Moreover, all the docking-obtained solutions, regardless of the orientation of the ligand chain, interact with one face of glucosyl ring 1 and with the opposite face of glucosyl rings 2 and 4. According to the orientation I have chosen, Trp78 interacts with the α -face of the glucosyl ring 1 (reducing end) while Trp68 and Trp27 interact with the β -face of the glucosyl ring 2 and 4. As predicted, in the α -conformation the anomeric hydroxyl group of the reducing end of cellotetraose makes strong hydrophobic contact with the NH group of the indole ring of Trp78. Again, this is in good agreement with the STD-NMR data, and justifies why this proton is the one that receives the highest degree of saturation. Additionally, proton H5 points into the π -electron cloud of the aromatic ring, suggesting a CH- π interaction²⁸.

Regarding units 2 and 4, there is also the probability of CH- π interactions (protons H4 and H4) with the sidechain rings of Trp27 and Trp68, respectively. A feature of the CtCBM30 interaction is a low number of direct hydrogen bonds to the protein. Indeed, there are only two hydrogen bonds in the interaction between CtCBM30 and cellotetraose (**Figure IV.7** – black dashed lines). The OH of the methylene group of sugar ring 3 makes a 2.4 Å hydrogen bond to the NH2 of Arg110 and a 2.6 Å hydrogen bond with the NH1 of the same residue. This explains the reduced affinity for xylan, pointing to the need for a direct interaction between the O6 of glucose and the protein. Besides giving these results it can be predicted that this site would display an absolute requirement for an unsubstituted glucose moiety. Also in this sense,

* The α -face of the glucosyl ring is defined as the face at which the numbering of atoms in the ring (C1, C2, C3, C4, O5) appear clockwise, and the β -face is the face at which the numbering is anticlockwise.^{20,21}

substitution in the glucose in site 1 (reducing end) would impair the interaction that arises between the N ζ group of the lysine side chain with the OH of the methylene group of sugar ring. Moreover, Ile70 and Leu72, located in the same face of the cleft as the tryptophan residues and Glu121, located in the opposite face also contact sugar units 2, 3 and 4. Due to the lack of any other significant interactions, possible substituents at other sites can be displaced away from the binding cleft and accommodated with no obvious energetic penalty. These substituents can even make additional interactions with the protein, further stabilizing the complex and thus explaining the higher affinity for branched ligands when compared to unbranched¹.

Taken together, structural analysis, STD-NMR and docking studies show that the interaction of CtCBM30 with branched ligands, namely xyloglucan, is coupled with both the orientation of the residues in the binding cleft¹ and the orientation of the ligand. The orientation of the solvent-exposed tryptophans selects ligands that display the twisted conformation exhibited by cello-oligosaccharides in solution and the orientation of some of the C6 hydroxyl groups towards the solvent provides an explanation for the ability of this protein to bind xyloglucan.

IV.2.1.2.2 Model of CtCBM30 bound to cellohexaose

For the interaction of CtCBM30 with cellohexaose the same problem regarding the orientation of the sugar on the cleft was considered. Once more I chose the structure that best described the STD-NMR data and with the lowest energy. In this structure the ligand is positioned as previously (**Figure IV.8**).

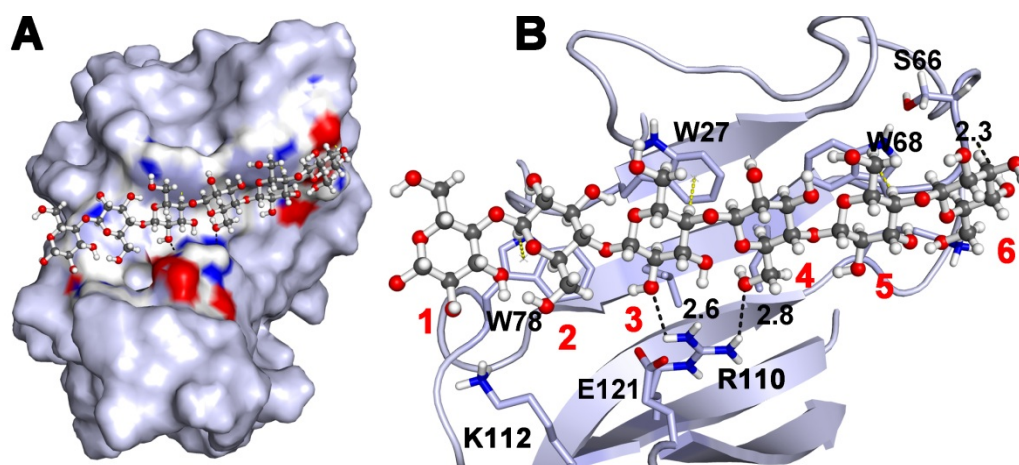


Figure IV.8: Model of the structure of CtCBM30 in complex with cellohexaose.

A) Surface representation of CtCBM30 bound to cellohexaose. **B)** Ribbon representation of the CtCBM30 binding cleft bound to cellohexaose. The ligand is depicted in ball-and-stick and the interacting residues are depicted as sticks. Atoms are colored by heteroatom. Hydrogen bonds are represented as black dashes and CH- π interactions are represented as yellow dashes. Glucosyl moieties (in red) are numbered as recommended by IUPAC-IUB JCBN (1983).²⁴

The interaction of CtCBM30 with longer saccharides, namely cellohexaose, is very similar to the one with cellotetraose. Likewise, Trp78 interacts with the α -face of the glucosyl moiety 2 while Trp68 and Trp27 interact with the β -face of the glucosyl moieties 3 and 5. Essentially, the three tryptophans interact with the central glucose units and the extremities lay outside the binding cleft (**Figure IV.8**). This is in good agreement with the STD-NMR results, explaining the low A_{STD} values observed for the non-reducing end and the absence of STD signals for the reducing end and the higher intensity STD of the central glucose units. Trp78, Trp27 and Trp68 make CH- π interactions with sugar rings at sites 2 (H1), 3 (H4) and 5 (H4), respectively. Lys112, Glu121 and Arg110, located along one side of the binding cleft contact with the ligand at sites 1, 3 and 4, respectively and Lys167, Ser66, Ile70 and Leu72, located along the opposite face of the cleft, contact the sugar residues at sites 6, 4 and 3. Similarly to the interaction with cellotetraose, there is a 2.8 Å hydrogen bond between the C6 hydroxyl of sugar ring 4 (corresponding to unit 3 of cellotetraose) and the NH2 of Arg110. This highlights the previous assumption that this site would display an absolute requirement for an unsubstituted glucose moiety. The NH1 of Arg121 makes another hydrogen bond, not with the C6 OH group as in the case of cellotetraose, but with the C2 OH group of unit 3 (2.6 Å). Compared to the interaction with cellotetraose, the sidechain of Lys112 changes its conformation to interact with sugar unit 1 (that stays outside the binding cleft) instead of sugar unit 2.

Interestingly, although the binding cleft of CtCBM30 can ideally accommodate 4 sugar units, it displays higher affinity for longer ligands (**Table IV.1**). This was attributed to a possible more extensive hydrogen bonding network between these longer ligands and the protein, possibly stabilizing the conformation adopted by the oligosaccharides in the binding cleft.¹ Surprisingly, the number of contacts between CtCBM30 and cellohexaose does not increase significantly when compared to cellotetraose. Nonetheless, the formation of a hydrogen bond between the C4 OH group of unit 6 and the backbone oxygen of Ser66 together with the conformational alteration of the sidechain of Lys112 may be sufficient to further stabilize the interaction of cellohexaose, thus increasing the affinity.

Overall, the obtained model for the CtCBM30/cellohexaose complex is in good agreement with the STD-NMR data and provides an explanation on the mechanism behind the higher affinity that this CBMs display towards longer ligands.

IV.2.1.2.3 Model of CtCBM44 bound to cellohexaose

For the model of CtCBM44 bound to cellohexaose (**Figure IV.9**), I chose the orientation where the solvent-exposed tryptophan residues 198, 194 and 189 make CH- π interactions with the α -face of sugar moieties 1 (H4), 3 (H4) and with the β -face of sugar moiety 6 (H5),

respectively. This is the one with the lowest energy and has more contacts between the sugar and the protein, when compared to all other possible orientations from HADDOCK. Moreover, it is also in good agreement with the STD-NMR data. However, the interaction of Trp189 with the β -face of the glucosyl ring 6 is unexpected. Due to the arrangement of the three tryptophan residues, it was predicted that they would bind to sugar units n , $n+2$ and $n+4$. Nonetheless, looking at the model we see that for this to happen it would require a different conformation of the sidechain of Trp189. This different conformation would clash with the sidechain of Met183. Therefore, instead of binding optimally to sugars with at least five units as previously proposed¹, I suggest that *CtCBM44* binds optimally to sugars with 6 units, at sites n , $n+2$ and $n+5$. This would explain the much higher binding affinity of *CtCBM44* to cellohexaose than for cellopentaose (Table IV.1). To test this hypothesis I have also performed docking calculation with the pentasaccharide, cellopentaose – see below, Section IV.2.1.2.5.

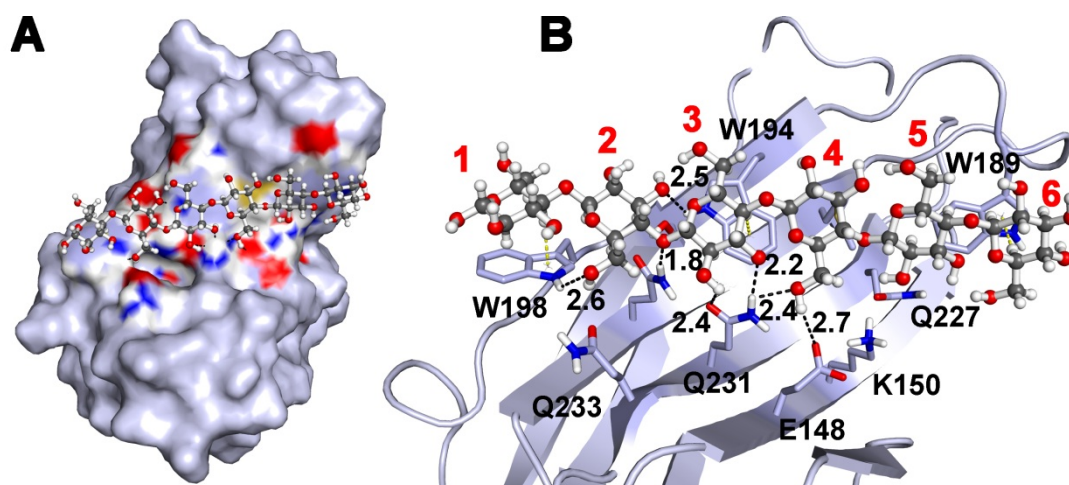


Figure IV.9: Model of the structure of *CtCBM44* in complex with cellohexaose.

A) Surface representation of *CtCBM44* bound to cellohexaose. **B)** Ribbon representation of the *CtCBM44* binding cleft bound to cellohexaose. The ligand is depicted in ball-and-stick and the interacting residues are depicted as sticks. Atoms are colored by heteroatom. Hydrogen bonds are represented as black dashes and CH- π interactions are represented as yellow dashes. Glucosyl moieties (in red) are numbered as recommended by IUPAC-IUB JCBN (1983).²⁴

Concerning the interaction with cellohexaose, the data shows that residues Gln233, Gln231, Glu148 and Lys150, located in one side of the cleft, make several contacts with the four central sugar units. These contacts include 3 hydrogen bonds between the O ϵ 1 and N ϵ 2 groups of Gln231 and OH group 2 of sugar unit 3 (2.4 Å) and the OH groups 4 (2.2 Å) and 6 (2.4 Å) of sugar units 3 and 4, respectively. The formation of these 3 hydrogen bonds explains why mutation of Gln231 for an alanine caused a 7-fold decrease in affinity towards cellohexaose¹ and highlights the importance of this residue for ligand recognition and binding. It also suggests that, similar to *CtCBM30*, site 4 requires an unsubstituted glucose moiety. Also the OH of the methylene group of unit 2 makes a 2.6 Å hydrogen bond with the NH of the indole ring of

Trp198 and polar contacts with Oε1 of the sidechain of Gln233, possibly impairing a substitution also at this site. On the opposite side of the cleft, residues Gln179, Ser196, Met183 and Gln227 make additional contacts with the central glucose units. Of these contacts, there is a 1.8 Å hydrogen bond between the Nε2 of Gln179 and the O4 of sugar unit 2. The OH of the methylene group of glucose units 2, 3 and 4 makes a large number of contacts with several residues of the protein, thus justifying the relatively high A_{STD} values obtained (**Table IV.2**). The same is true for the C2 OH groups of the central glucose units, which is in good agreement with the STD-NMR data. The high number of contacts between the ligand and several residues of the protein may help to explain the much higher affinity of CtCBM44 towards cellohexaose when compared to CtCBM30.

Above, I proposed that the absence of STD-NMR signals for the protons of the reducing end of cellohexaose could indicate that this unit didn't contribute significantly for binding. However, according to the obtained model we see that this is not true as this unit can make several interactions with Trp198, including a CH- π interaction. Nevertheless, this unit faces the tryptophan ring with its α -face, meaning that protons H4 and H2 are the ones pointing to the ring. Because the signals of these protons appear in the crowded central area of the STD-NMR spectrum (the one with the highest A_{STD} value), possible STD-NMR signals arising from these protons cannot be distinguished from other signals appearing in the same region. As for the non-reducing end of cellohexaose, the fact that it faces Trp189 with its β -face (protons H1, H3 and H5) is in good agreement with the STD-NMR results.

Overall, STD-NMR data and docking studies showed that similar to CtCBM30, the binding of CtCBM44 to branched ligands is coupled both with the orientation of the residues in the binding cleft and the orientation of the ligand. As for CtCBM30, the orientation of the solvent-exposed tryptophans selects ligands that display the twisted conformation exhibited by cello-oligosaccharides in solution and the orientation of some of the C6 hydroxyl groups towards the solvent provides an explanation for the ability of this protein to bind xyloglucan. Additionally, the docking model allowed to propose a minimal length for the oligosaccharide chain (6 units) different from the one previously suggested¹ (5 units) which could explain the much higher affinity displayed for the interaction with cellohexaose when compared to cellopentaose. Given the high similarities between CtCBM44 and CtCBM30, this fact may also be responsible for the much higher affinity of CtCBM44 to cellohexaose when compared with CtCBM30 - $72.8 \times 10^4 \pm 7.2$ and $6.4 \times 10^4 \pm 0.8 \text{ M}^{-1}$, respectively (**Table IV.1**). The larger platform offered by the three tryptophan residues could promote a higher stabilization of the ligand, thus increasing the affinity.

IV.2.1.2.4 Model of *CtCBM44* bound to cellopentaose

In order to test the hypothesis that *CtCBM44* binds optimally to a minimum of 6 sugar units instead of the 5 previously predicted¹ I have calculated the model with cellopentaose (**Figure IV.10**). The obtained model is almost identical to the one with celohexaose and, as predicted, the non-reducing end of cellopentaose, although close to Trp189, does not make the CH- π interaction. All other interactions are maintained in the complex with cellopentaose. The loss of the CH- π interaction with Trp189 introduces flexibility at the non-reducing end of the ligand destabilizing it and thus causing a decrease in the affinity when compared to celohexaose.

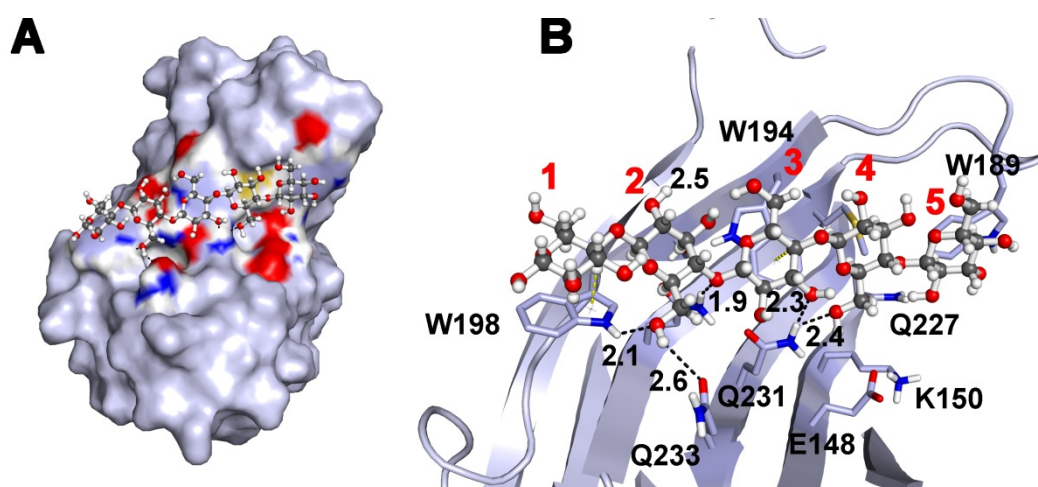


Figure IV.10: Model of the structure of *CtCBM44* in complex with cellopentaose.

A) Surface representation of *CtCBM44* bound to cellopentaose. **B)** Ribbon representation of the *CtCBM44* binding cleft bound to cellopentaose. **C)** Superposition of the models with celohexaose (grey) and cellopentaose (orange). The ligand is depicted in ball-and-stick and the interacting residues are depicted as sticks. Atoms are colored by heteroatom. Hydrogen bonds are represented as black dashes and CH- π interactions are represented as yellow dashes. Glucosyl moieties (in red) are numbered as recommended by IUPAC-IUB JCBN (1983).²⁴

IV.2.1.2.5 Model of *CtCBM44* bound to cellotetraose

Although the STD-NMR study indicated that the interaction of *CtCBM44* with cellotetraose was very weak, I decided to calculate a model for this complex in order to get a possible explanation for the reason of this weak interaction. The model of *CtCBM44* bound to cellotetraose is shown in **Figure IV.11**.

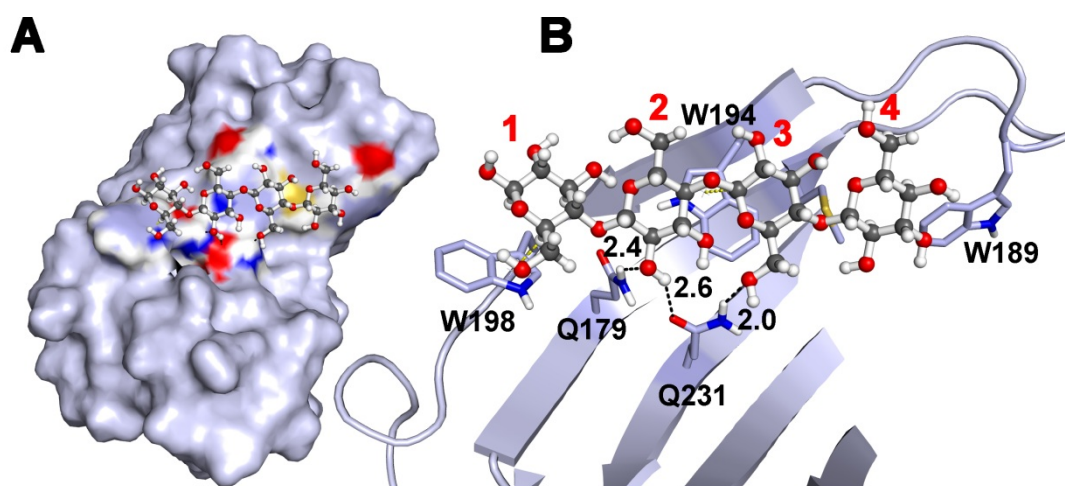


Figure IV.11: Model of the structure of CtCBM44 in complex with cellotetraose.

A) Surface representation of CtCBM44 bound to cellotetraose. **B)** Ribbon representation of the CtCBM44 binding cleft bound to cellotetraose. The ligand is depicted in ball-and-stick and the interacting residues are depicted as sticks. Atoms are colored by heteroatom. Hydrogen bonds are represented as black dashes and CH- π interactions are represented as yellow dashes. Glucosyl moieties (in red) are numbered as recommended by IUPAC-IUB JCBN (1983).²⁴

According to this model the majority of the interactions with the protein are lost and just two of the three tryptophan residues interact with the ligand and make CH- π interactions. Besides the tryptophan residues, cellotetraose only interacts with Met183, Gln179 and Gln231. These two last residues make three hydrogen bonds with the hydroxyl groups 2 (2.4 and 2.6 Å) and 6 (2.0 Å) of units 2 and 3, respectively. The interaction with two of the three solvent-exposed tryptophans and these two hydrogen bonds, together with the lack of any other significant interaction justifies the weak, but still present interaction of CtCBM44 with cellotetraose.

IV.3 Conclusions

The plant cell wall is composed mainly of cellulose and hemicellulose and its degradation is one of the most important steps in the global turnover process of atmospheric CO₂. Regardless of its abundance in nature, cellulose is a particularly difficult polymer to degrade, as it is insoluble and is present as hydrogen-bonded crystalline fibers, coated with hemicellulose chains and pectin all “glued” into an intricate 3D network. For the cellulolytic microorganisms (like *C. thermocellum*) the ability for degrading this paraphernalia is conferred by the plasticity displayed by their cellulases. These proteins are able to recognize and cleave a wide range of β -1,4-glucosidic bonds in a variety of polysaccharides (*e.g.* cellulose, xyloglucan, glucomannan, and mixed-linked β -1,4- β -1,3-glucans). A fundamental piece for this are the non-catalytic

carbohydrate-binding modules whose specificity mimics that of the attached catalytic module and whose function is mainly to target the enzymes to their substrates.

In this chapter I have studied the interaction of *CtCBM30* and *CtCBM44* with cellobiose, cellotetraose, cellopentaose and cellohexaose through a combination of STD-NMR and molecular docking. Both experimental and theoretical results are in good agreement and indicate that a combination between the arrangement of the three solvent-exposed tryptophan residues in each protein and interactions of polar residues with the C6 hydroxyl group of the central glucose units are key for defining ligand specificity. The twisted arrangement of the tryptophan residues selects against ligands that do not have this geometry, while the interaction with some C6 OH groups selects against substituted (or without this group) glucose units. It is my belief that this mechanism is common for CBMs that bind to highly decorated ligands but further experimental work is required.

Moreover, I have shown that the higher affinity that these proteins display against ligands longer than they can accommodate in the binding cleft may be related to the interaction of sugar units that lay outside the binding cleft with polar residues of the protein. These residues flank the binding cleft and make hydrogen bonds with the sugar units at the extremities, thus stabilizing the conformation adopted by these ligands in the binding cleft.

Docking experiments showed that the platform designed by the three tryptophan residues in *CtCBM44* can ideally accommodate ligands with up to six glucose units and not five as previously thought. Given the structural similarities between *CtCBM44* and *CtCBM30*, this fact may explain the much higher affinity of *CtCBM44* to cellohexaose when compared with *CtCBM30* - $72.8 \times 10^4 \pm 7.2$ and $6.4 \times 10^4 \pm 0.8 \text{ M}^{-1}$, respectively (**Table IV.1**). The larger platform designed by the three tryptophan residues could promote a higher stabilization of the ligand, thus increasing the affinity.

IV.4 Materials and methods

IV.4.1 Sources of sugars

All the sugars (cellobiose, cellotetraose, cellohexaose and laminarihexaose) were obtained from Seikagaku Corporation (Tokyo, Japan) and were used without further purification.

IV.4.2 Molecular biology

IV.4.2.1 Recombinant protein production

To express CtCBM30 and CtCBM44 in *Escherichia coli*, I used two vectors, pCG1 and pCG3, respectively, kindly provided by Professor Carlos Fontes of Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa. For the production of CtCBM30 and CtCBM44 with the histidine tag, DNA encoding for both proteins was amplified from the *C. thermocellum* CtCel9D-Cel44A gene as described elsewhere.^{1,2} The excised CtCBM30 and CtCBM44 encoding genes were cloned into the vector pET21a (Novagen) to generate pCG1 and pCG3, respectively. The recombinant plasmids contain the clostridial gene under the control of the T7 promoter (see Appendix A for supporting information on the pET system and pET21a plasmid and T7 promoter).

IV.4.2.2 Protein expression and purification

CtCBM30 and CtCBM44 were produced by first transforming the pCG1 and pCG3 expression vectors into competent *E. coli* BL21 cells (Novagen). All the procedure for transformation, expression, purification and quantification of both proteins was the same as for CtCBM11 – see Chapter II. The yields obtained were around 50 mg/L of CtCBM30 and 12 mg/L of CtCBM44 and the final concentration of the protein was kept around 1 mM. **Figure IV.12** shows the SDS-PAGE gel of the purified CtCBM44. Unfortunately no picture of the SDS-PAGE gel of the purified CtCBM30 was taken.

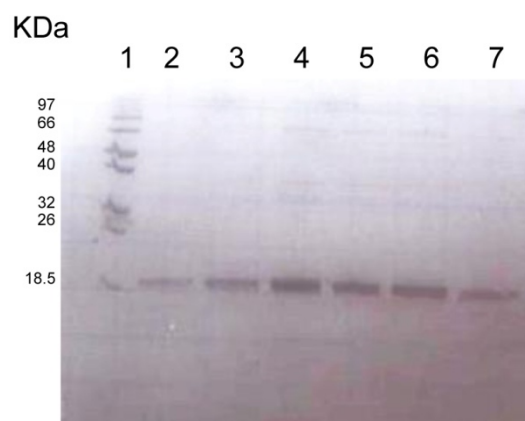


Figure IV.12: SDS-PAGE gel of the purified CtCBM44 fractions.

Lane 1 – LMW markers; Lanes 2-7 purified fractions

IV.4.3 NMR spectroscopy

IV.4.3.1 Data acquisition

All NMR spectra were acquired with a 600 MHz Bruker AvanceIII spectrometer (Bruker, Wissembourg, France) equipped with a 5 mm inverse detection triple-resonance z-gradient cryogenic probehead (CP TCI) and processed in Bruker TopSpin3.1 (Bruker).

IV.4.3.2 STD-NMR studies

The interaction between CtCBM44 and cellobiose, cellotetraose, cellohexaose and laminarihexaose was studied by saturation transfer difference NMR (STD-NMR) using the pulse sequence from the Bruker library (stddiffesgp.3).^{29,30} The pseudo 2D spectra were acquired using a solution of 2 mM ligand and 20 μ M protein in D₂O for the case of CtCBM44 and 3 mM ligand and 30 μ M protein in D₂O for CtCBM44. All the spectra were recorded at 600 MHz with 16 scans repeated 16 times in a matrix with 32 k points in t₂ in a spectral window of 12019.23 Hz centered at 2814.60 Hz. Excitation sculpting with gradients²⁹ was employed to suppress the water proton signals. A spin lock filter ($T_{1\rho}$) with a 2 kHz field and a length of 20 ms was applied to suppress protein background. Selective saturation of protein resonances was performed by irradiating at 0.8 ppm for CtCBM44 and 7.0 ppm for CtCBM30 (on resonance spectrum) using a series of 40 Eburp2.1000 shaped 90° pulses (50 ms, 1 ms delay between pulses), for a total saturation time of 2.0 s. For the reference spectrum (off resonance), I irradiated at 20 ppm.

To obtain the 1D STD-NMR spectra I subtracted the on resonance spectra from the off resonance using the Topspin3.1 (Bruker, Wissembourg, France) software. The difference spectrum corresponds to the STD-NMR spectrum and, at the correct saturation time, the intensity of its signals gives information on the proximity of the corresponding protons to the protein. To calculate the STD amplification factors (**Table III.2**) I have proceeded as for CtCBM11 (see Chapter III – Section II.4.4.5)

IV.4.4 Docking studies

IV.4.4.1 Preparation of the ligand pdb files

The carbohydrate ligand molecules were constructed with the “Glycam Biomolecule Builder” available online from the website of Woods group²¹. The ligands were then minimized

by molecular mechanics, through 1000 steps of the steepest descent method, followed by the conjugate gradient method until a convergence criterion of 0.0001 Kcal.mol⁻¹ was achieved.

IV.4.4.2 Docking models for the interaction of CtCBM30 and CtCBM44 with cellooligosaccharides

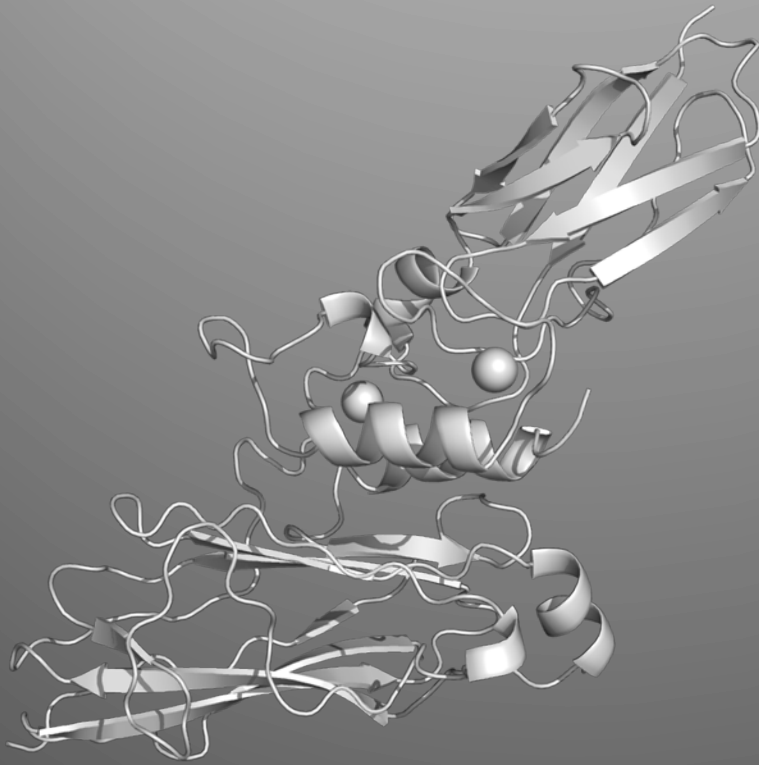
Models of the CtCBM30/cellotetraose, CtCBM30/cellohexaose, CtCBM44/cellotetraose, CtCBM44/cellopentaose and CtCBM44/cellohexaose complexes were calculated using the software HADDOCK (high ambiguity-driven protein docking) under the WeNMR Grid-enabled server^{13,14} using the previously determined X-ray structures (PDB codes: 2c24 and 2c26 for CtCBM30 and CtCBM44, respectively). For the ambiguous interaction restraints (AIRs), i.e., active residues, only the solvent-exposed tryptophan residues (Trp27, Trp68 and Trp78 for CtCBM30 and Trp189, Trp194 and Trp198 for CtCBM44) were chosen. The passive residues were selected automatically (6.5 Å around the active residues). The HADDOCK docking protocol was performed as described elsewhere.^{14,31} The rigid body docking stage was performed 5 times, and the best resulting structure was saved. 1000 structures were generated at the rigid body docking stage, the best 200 of which were selected for further semiflexible refinement and refinement in explicit water. Non-bonded energies were calculated using the OPLSX non-bonded parameters.³² Parameters for the ligands were obtained from Glycam Web as described above.²¹ The resulting solutions were clustered using a 2Å cut off and analyzed with the software PyMol1.4.1³³. The best structure of the cluster with the lowest energy was compared against the STD-NMR data and used for subsequent analysis without further refinement.

IV.5 References

1. Najmudin, S.; Guerreiro, C. I. P. D.; Carvalho, A. L.; Prates, J. A. M.; Correia, M. A. S.; Alves, V. D.; Ferreira, L. M. A.; Romao, M. J.; Gilbert, H. J.; Bolam, D. N.; Fontes, C. M. G. A., Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *Journal of Biological Chemistry* **2006**, *281* (13), 8815.
2. Najmudin, S.; Guerreiro, C. I. P. D.; Ferreira, L. M. A.; Romao, M. J. C.; Fontes, C. M. G. A.; Prates, J. A. M., Overexpression, purification and crystallization of the two C-terminal domains of the bifunctional cellulase ctCel9D-Cel44A from *Clostridium thermocellum*. *Acta Crystallogr F* **2005**, *61*, 1043.
3. Hashimoto, H., Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **2006**, *63* (24), 2954.
4. Atkins, E. D. T.; Parker, K. D., Helical Structure of a Beta-D-1,3-Xylan. *J Polym Sci Polym Chem* **1969**, (28PC), 69.

5. Carvalho, A. L.; Goyal, A.; Prates, J. A. M.; Bolam, D. N.; Gilbert, H. J.; Pires, V. M. R.; Ferreira, L. M. A.; Planas, A.; Romao, M. J.; Fontes, C. M. G. A., The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *Journal of Biological Chemistry* **2004**, 279 (33), 34785.
6. Boraston, A. B.; Revett, T. J.; Boraston, C. M.; Nurizzo, D.; Davies, G. J., Structural and thermodynamic dissection of specific mannan recognition by a carbohydrate binding module, TmCBM27. *Structure* **2003**, 11 (6), 665.
7. Charnock, S. J.; Bolam, D. N.; Nurizzo, D.; Szabo, L.; McKie, V. A.; Gilbert, H. J.; Davies, G. J., Promiscuity in ligand-binding: The three-dimensional structure of a *Piromyces* carbohydrate-binding module, CBM29-2, in complex with cello-and mannohexaose. *P Natl Acad Sci USA* **2002**, 99 (22), 14077.
8. Notenboom, V.; Boraston, A. B.; Chiu, P.; Freelove, A. C. J.; Kilburn, D. G.; Rose, D. R., Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: An X-ray crystallographic, thermodynamic and mutagenic study. *Journal of Molecular Biology* **2001**, 314 (4), 797.
9. Xie, H. F.; Bolam, D. N.; Nagy, T.; Szabo, L.; Cooper, A.; Simpson, P. J.; Lakey, J. H.; Williamson, M. P.; Gilbert, H. J., Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry* **2001**, 40 (19), 5700.
10. Arai, T.; Araki, R.; Tanaka, A.; Karita, S.; Kimura, T.; Sakka, K.; Ohmiya, K., Characterization of a cellulase containing a family 30 carbohydrate-binding module (CBM) derived from *Clostridium thermocellum* CelJ: Importance of the CBM to cellulose hydrolysis. *J Bacteriol* **2003**, 185 (2), 504.
11. Fry, S. C.; York, W. S.; Albersheim, P.; Darvill, A.; Hayashi, T.; Joseleau, J. P.; Kato, Y.; Lorences, E. P.; Maclachlan, G. A.; Mcneil, M.; Mort, A. J.; Reid, J. S. G.; Seitz, H. U.; Selvendran, R. R.; Voragen, A. G. J.; White, A. R., An Unambiguous Nomenclature for Xyloglucan-Derived Oligosaccharides. *Physiol Plantarum* **1993**, 89 (1), 1.
12. Del Bem, L. E.; Vincentz, M. G., Evolution of xyloglucan-related genes in green plants. *BMC Evol Biol* **2010**, 10, 341.
13. de Vries, S. J.; van Dijk, M.; Bonvin, A. M. J. J., The HADDOCK web server for data-driven biomolecular docking. *Nat. Protocols* **2010**, 5 (5), 883.
14. Dominguez, C.; Boelens, R.; Bonvin, A. M., HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **2003**, 125 (7), 1731.
15. Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J., Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **2004**, 382 (Pt 3), 769.
16. Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008**, 275 (10), 2524.
17. Viegas, A.; Macedo, A. L.; Cabrita, E. J., Ligand-Based Nuclear Magnetic Resonance Screening Techniques. In *Ligand-macromolecular interactions in drug discovery : methods and protocols*, Roque, A. C. A., Ed. Springer: New York, 2010; pp 81.
18. Johnson, P. E.; Tomme, P.; Joshi, M. D.; McIntosh, L. P., Interaction of soluble cellooligosaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC .2. NMR and ultraviolet absorption spectroscopy. *Biochemistry* **1996**, 35 (44), 13895.
19. Viegas, A.; Manso, J. o.; Nobrega, F. L.; Cabrita, E. J., Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *J Chem Educ* **2011**.
20. Meyer, B.; Peters, T., NMR Spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie-International Edition* **2003**, 42 (8), 864.
21. WoodsGroup (2005-2012) GLYCAM Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (<http://www.glycam.com>).

22. Nand K, V., Atomic features of protein-carbohydrate interactions. *Curr Opin Struc Biol* **1991**, *1* (5), 732.
23. Divne, C.; Ståhlberg, J.; Teeri, T. T.; Jones, T. A., High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *Journal of Molecular Biology* **1998**, *275* (2), 309.
24. Symbols for Specifying the Conformation of Polysaccharide Chains. *Eur J Biochem* **1983**, *131* (1), 5.
25. Notenboom, V.; Boraston, A. B.; Kilburn, D. G.; Rose, D. R., Crystal structures of the family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A in native and ligand-bound forms. *Biochemistry* **2001**, *40* (21), 6248.
26. Johnson, P. E.; Brun, E.; MacKenzie, L. F.; Withers, S. G.; McIntosh, L. P., The cellulose-binding domains from *Cellulomonas fimi* β -1,4-glucanase CenC bind nitroxide spin-labeled celooligosaccharides in multiple orientations. *Journal of Molecular Biology* **1999**, *287* (3), 609.
27. Szabo, L.; Jamal, S.; Xie, H.; Charnock, S. J.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J., Structure of a family 15 carbohydrate-binding module in complex with xylopentaose. Evidence that xylan binds in an approximate 3-fold helical conformation. *J Biol Chem* **2001**, *276* (52), 49061.
28. Diaz, M. D.; Fernandez-Alonso, M. D.; Cuevas, G.; Canada, F. J.; Jimenez-Barbero, J., On the role of aromatic-sugar interactions in the molecular recognition of carbohydrates: A 3D view by using NMR. *Pure Appl Chem* **2008**, *80* (8), 1827.
29. Hwang, T. L.; Shaka, A. J., Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J Magn Reson Ser A* **1995**, *112* (2), 275.
30. Mayer, M.; Meyer, B., Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie-International Edition* **1999**, *38* (12), 1784.
31. Tomaselli, S.; Ragona, L.; Zetta, L.; Assfalg, M.; Ferranti, P.; Longhi, R.; Bonvin, A. M. J. J.; Molinari, H., NMR-based modeling and binding studies of a ternary complex between chicken liver bile acid binding protein and bile acids. *Proteins: Structure, Function, and Bioinformatics* **2007**, *69* (1), 177.
32. Linge, J. P.; Williams, M. A.; Spronk, C. A. E. M.; Bonvin, A. M. J. J.; Nilges, M., Refinement of protein structures in explicit solvent. *Proteins: Structure, Function, and Bioinformatics* **2003**, *50* (3), 496.
33. Schrödinger, LLC *The PyMOL Molecular Graphics System*, 1.4.1; 2010.



Chapter V: The Orf2 Type II Cohesin-XDockerin complex from C. thermocellum

In this chapter I have used X-ray crystallography to determine the 3D structure of the Orf2 Type II Cohesin-Module X-Dockerin complex from Clostridium thermocellum. The data obtained allowed a better understanding on the mechanisms of cell wall attachment in anaerobic bacteria and provided insights on the mechanism the rule cohesin-dockerin interaction. The results here presented are part of a manuscript currently in preparation.

Table of Contents

Summary	161
V.1 Introduction	161
V.2 Results and Discussion.....	165
V.2.1 Architecture of the Orf2 type II Coh-XDoc complex from <i>C. thermocellum</i>	165
V.2.1.1 Type II Coh structure in the complex.....	168
V.2.1.2 Type II XDoc structure in the complex.....	168
V.2.1.3 The complex interface.....	173
V.3 Conclusions	177
V.4 Materials and methods	178
V.4.1 Molecular biology	178
V.4.1.1 Transformation, expression, purification and quantification	178
V.4.2 X-ray crystallography.....	179
V.4.2.1 Protein crystallization and data collection.....	179
V.4.2.2 Phasing, model building and refinement.....	179
V.5 References	180

Summary

The assembly of cellulosomes of *C. thermocellum* onto the bacterial surface is orchestrated by five proteins, SdbA, Orf2, OlpA, OlpB and OlpC which are presumed to be tethered onto the bacterial cell wall via N-terminal SLH domains (see Chapter I). These proteins contain type II cohesins, which recruit the cellulosome onto the surface of the bacterial membrane by binding to the type II dockerin present at the C-terminus of CipA. In order to better understand this mechanism I have solved the crystal structure of the Orf2 type II Coh-XDoc from *C. thermocellum* (Figure V.1) to a resolution of 1.98 Å. The structure

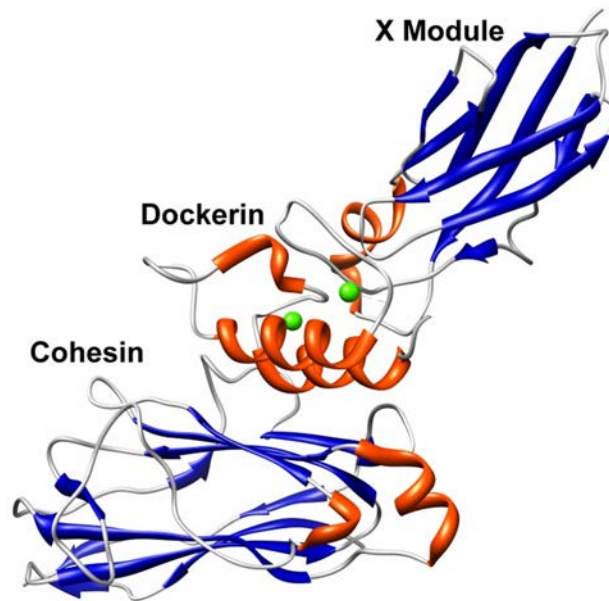


Figure V.1: Crystal structure of the Orf2 Type II cohesin-modules X-dockerin complex (CohII-XDocII) from *C. thermocellum* (PDB code: 2vt9)

The two calcium ions are depicted as green spheres; α -helical regions are depicted in red; β -sheet regions are depicted in blue and random coil regions are depicted in grey.

obtained is very similar to the previously determined SdbA Coh-XDoc structure¹ and reveals that both helix 1 and helix 3 of the dockerin are involved in the interaction with the cohesin. Furthermore, the solved structure confirms that the X module displays an important role in dockerin stabilization and cohesin recognition. The multiple contacts made with the cohesin module by both helices and the lack of symmetry of type II dockerin amino acids at the interface indicates that the module is unlikely to display the dual binding mode exhibited by the corresponding type I module.

V.1 Introduction

Protein-protein recognition plays a pivotal role in an array of biological processes and cellulosic mass degradation is not an exception. The plant cell wall is the major source of organic carbon on the planet and the recycling of photosynthetically fixed carbon is a crucial

microbial process. This process is critical to the cycling of carbon between microbes, herbivores and plants. In anaerobes, the degradation of this composite structure is carried out by a high molecular weight multifunctional complex termed the **cellulosome** (see Chapter I).² The architecture of the cellulosome is defined by high affinity protein-protein interactions ($K_d > 10^{-9}$ to 10^{-12} M)^{3,4} between cohesins (Coh) and dockerins (Doc) and promotes the enhanced substrate degradation by these megadalton complexes (Figure V.1).⁵⁻⁷ These interactions are among the strongest protein-protein interactions in nature and are vital to cellulosome assembly.

The cellulosome of *C. thermocellum* is composed of two types of cohesin-dockerin partners (Coh-Doc): type I, (usually) responsible for the assembly of the several enzymes to the scaffoldin protein (CipA), and type II, (usually) responsible for anchoring of the cellulosome complex to the bacterial cell wall (Figure V.2) via binding to specific domains found in the cell-surface proteins (OlpC, OlpA, SdbA, OlpB and Orf2).⁷

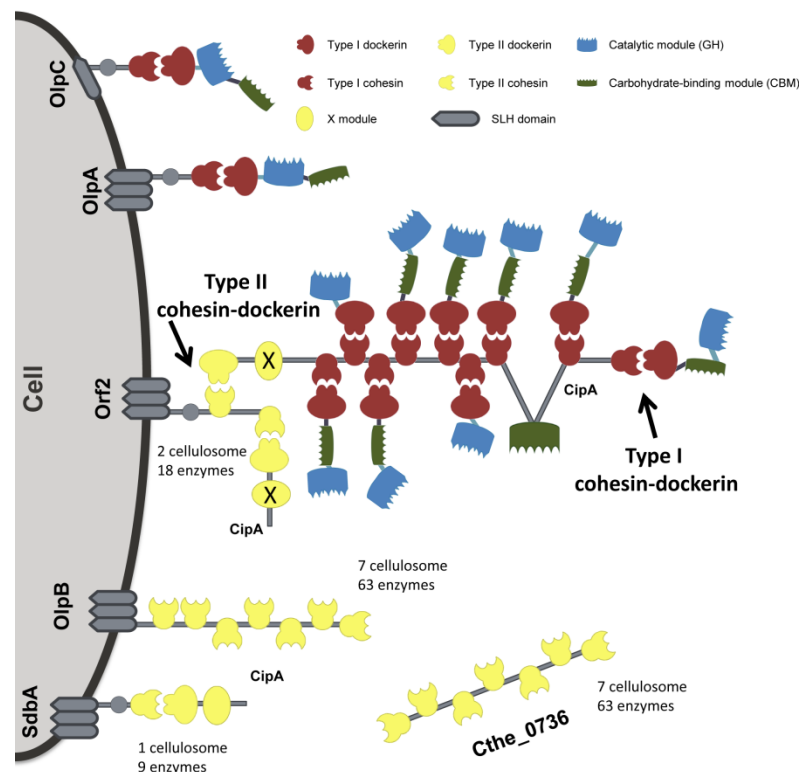


Figure V.2: Schematic representation of the *Clostridium thermocellum* cellulosome (adapted from Fontes *et al.*, 2010⁷).

The cellulosome of *C. thermocellum* is composed of five SLH domains for anchoring the complex to the bacterial cell wall (Orf2, OlpA, OlpB, OlpC and SdbA) through cohesin-dockerin interactions. The fact that Orf2, OlpB and the extracellular Cthe_0736 have more than one associated type II cohesin (2 and 7, respectively) contributes to the presence of polycellulosomes. The binding of the enzymes to specific positions is hypothetical, as is the linear orientation of the scaffoldin. The scaffoldins are only sketched partially. All cellulosome components are not drawn to scale.

These proteins are tethered onto the *C. thermocellum* cell wall via N-terminal SLH domains. The fact that these proteins may contain more than one type II cohesin domains contributes to the presence of polycellulosomes.⁸

Structural and mutagenesis studies have previously focused mainly on the type I interaction, while the knowledge regarding the attachment of the cellulosome to the bacterial cell surface through the type II interaction is more limited. Although structurally very similar, type I and type II cohesin-dockerins share only 15-25% sequence similarity which is consistent with the lack of cross-specificity between type I and type II cohesin-dockerin pairs.^{7,8} This ensures a clear distinction between the mechanism for cellulosome assembly and cell-surface attachment. On the other hand, it was also shown that, although type I cohesins/dockerins from one species do not interact with other type I cohesins/dockerins from other species,^{9,10} the type II pairs demonstrate a rather extensive cross-species plasticity.¹¹ The biological relevance of this cross-species interaction is still uncertain. Interestingly, type I dockerins in the enzymatic units recognize nearly all type I cohesins in the scaffoldin unit suggesting that, within a given species, the arrangement of the several enzymes occurs randomly along the scaffoldin, reflecting, perhaps, the complexity of the substrate in which the microbe is.⁷

High-resolution structures of cohesins and dockerins have already been determined^{1,5,6,8,10,12-14} individually or in complexes, providing insights into the molecular mechanisms that define the cellulosome assembly and cell surface attachment. For both complexes, recognition is dominated by hydrophobic interactions, augmented through an extensive hydrogen bonding network. In the type I complex it was shown that this extensive hydrogen-bonding network was dominated by the highly conserved Ser-Thr pair located in helix 3 of the dockerin, conferring species specificity among the type I dockerins.^{1,5} In *C. thermocellum* these residues are conserved among all type I dockerins, which is consistent with the inability of type I dockerins to distinguish between the 9 type I cohesins in CipA.⁷ In fact, the dockerin sequence is highly conserved and made up of two 22-residue sequence repeats separated by a linker region of about 9-18 residues.⁶ They fold into three α -helices, with helices 1 and 3 comprising the repeated segments such that residues 1-22 of helix 1 overlay with residues 35-56 of helix 3 and vice-versa (**Figure V.3**). Mutagenesis studies have shown that the internal two-fold symmetry displayed by dockerins gives both duplicated regions the potential to bind to the cohesin but, so far, only interaction through helix 3 has been found.¹⁴ The biological significance of the dual binding still remains unclear but, in principle, could provide a higher level of structure to the cellulosome or allow the crosslinking of two scaffoldins through a single dockerin.^{6,14} Moreover, there is evidence that this dual binding feature is not exclusive to *C. thermocellum*. In *C. cellulolyticum* the same internal structural symmetry is observed for type I dockerins, indicating that there is little, if any specificity between type I cohesin-dockerin partners in the cellulosome of this bacterium.^{7,15} Nonetheless, dockerins from nonclostridial species display a

higher degree of variation in the amino acid sequence between the two segments. Thus, it is not clear if this dual binding feature is invariant in all cellulosomes.⁷

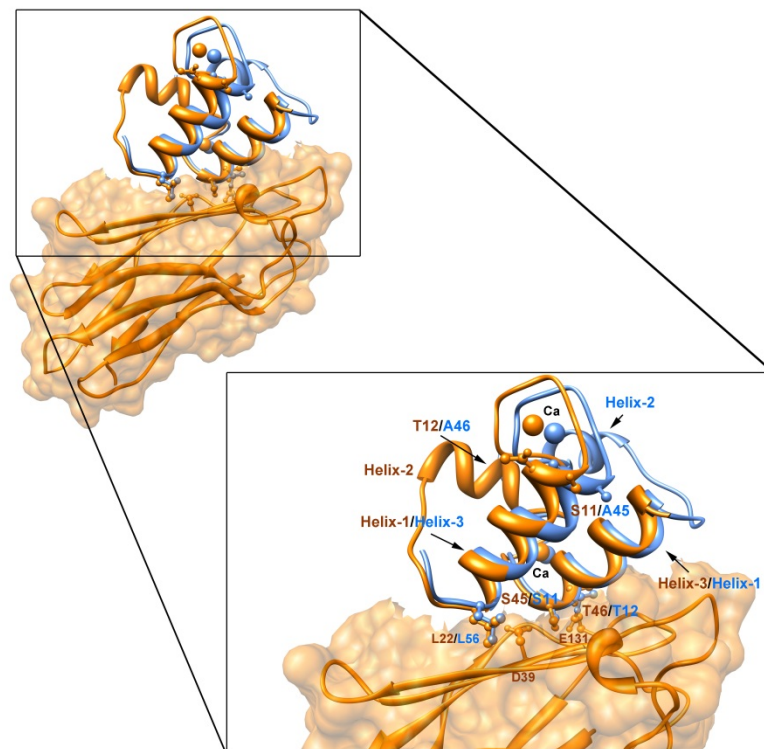


Figure V.3: The dual binding mode of type I cohesin-dockerin complexes.

Ribbon representation of the superposition of the dockerin modules of native type I Coh-Doc complex (PDB code: 1ohz⁶) (orange) with the S45A-T46A mutant (PDB code: 2ccl¹⁴) (blue) in *C. thermocellum*. The superposition was done taking the cohesin module of the native structure as the reference. For simplification only one cohesin module is represented (the one from the native complex). The inset shows a more detailed view of the cohesin-dockerin contacts and of the almost perfect superposition of helices 1 and 3 of both complexes. In the mutant complex, helix-1 (containing Ser-11 and Thr-12) dominates binding whereas, in the native complex, helix-3 (containing Ser-45 and Thr-46) plays a key role in ligand recognition. Ser-11, Thr-12, Ser-45, and Thr-46, which interact with the cohesin module, are depicted as ball-and-stick models and the calcium ions are depicted as spheres.

The type II Coh-Doc complex displays structural similarities with the corresponding type I complex¹ and, is typically responsible for recruiting the cellulosome onto the bacterial cell wall. This is done through high affinity interactions between the type II cohesins attached to one of the five SLH domains (**Figure V.2**) and the type II dockerins present at the C-terminal of the scaffoldin protein. The only exceptions to this are the type II dockerins of *B. cellulosolvens*, which are present in the primary scaffoldin (*see Chapter VI*). Usually, type II dockerins are present at the C-terminus of an Ig-like module termed X module.^{1,16} These modules are found in the scaffoldins of thermophilic and mesophilic bacteria and in cellulolytic enzymes.^{17,18} Although the function of the X module is still unknown, its importance was demonstrated through several biophysical studies¹⁶ and the resolution of the structure of the type II SdbA cohesin-dockerin-X module complex¹ (Coh-XDoc) and type II cohesin-dockerin-X module-type

I cohesin complex⁵ (CohII-XDoc-CohI₉). These studies suggest that the X module may be involved in the stabilization of the Coh-XDoc complex as well as in the formation of polycellulosomes, act as a solubility enhancer and be involved in the cellulosome attachment to the bacterial cell-wall. In the structures determined it is possible to identify an extensive range of interactions between the type II dockerin and the X module that are thought to help stabilize the complex.^{1,5} It was also possible to see that the cohesin-dockerin interaction surface is much larger than in the type I complex and that both the N- and C-terminal helices of the dockerin participate in the interaction. Furthermore, due to the presence of the X module, the type II cohesin-dockerin interaction is much more hydrophobic and tighter than the corresponding type I interaction. In light of these facts, it is thought that, although there is a considerable degree of symmetry in the type II dockerin, there will be no dual binding.^{3,7}

Even with the high-resolution structures determined so far,^{5,8,11} the molecular determinants responsible for the type II interaction and specificity are still not completely known, however they are key to understanding the mechanism of cellulosome assembly and activity. Towards this goal, I have solved the structure of a multimodular heterodimeric complex from *Clostridium thermocellum* composed of the type-II cohesin module of the cell surface protein Orf2 bound to a bimodular C-terminal fragment of the scaffoldin subunit CipA (X module bound to type II dockerin - XDoc) to a resolution of 1.98 Å (PDB code: 2vt9).

V.2 Results and Discussion

V.2.1 Architecture of the Orf2 type II Coh-XDoc complex from *C. thermocellum*

I have solved the crystal structure of the Orf2 type II Coh-XDoc complex from *C. thermocellum* (**Figure V.1**) by molecular replacement (MR – see *Chapter VII, Section VIII.4.2.1*) using as model the SdbA type II Coh-XDoc complex (PDB code: 2b59¹). The data was refined at 1.98 Å resolution and the final statistics are summarized in **Table V.1**. The final model has $R_{crist} = 18.7\%$ and $R_{free} = 24.7\%$ and includes 322 water molecules and two calcium ions. The residues Met1 and Ala1 of the Coh module (chain A), Met1, Asn2, Asn3, Asp4, Ser5 and Thr6 of the X module (chain B) and Leu160, Pro161, Ser162, Arg163 and Tyr164 from the Doc module (chain B) are disordered and, hence, not observed. The side chains of residues Arg73, Lys158 and terminal His-tag from the Coh module and Glu63 and Lys85 from the X module are also disordered and, therefore, not observed. The structure is deposited in the Protein Data Bank under the accession code: 2vt9.

Table V.1: X-ray data and structure quality statistics for the *Clostridium thermocellum* Orf2 type II Coh–XDoc complex.

<i>Data quality</i>	<i>Type II Coh-XDoc</i>
Cell dimensions, Å	$a = 116.67$ $b = 78.63$ $c = 35.80$ $\beta = 95.87^\circ$
Space group	C2
X-ray source	European Synchrotron Radiation Facility, ID14-EH4
Wavelength, Å	0.9735
Resolution of data (outer shell), Å	39.31– 1.98 (2.09 - 1.98)
R_{pim} (outer shell), %	5.8 (10.4)
R_{merge} (outer shell), %*	9.4 (17.3)
Mean $I/\sigma(I)$	13.1 (5.5)
Completeness (outer shell), %	97.8 (97.2)
Multiplicity (outer shell)	3.6 (3.7)
Structure quality	
No. protein atoms	5185
No. ligand atoms	2
No. solvent waters	322
Resolution used in refinement, Å	1.98
R_{cryst}/R_{free} (%) [†]	18.7 /24.7
rms deviation bonds, Å	0.01
rms deviation angles, °	1.2
Average cohesin B , Å ²	16.6
Average dockerin B , Å ²	16.6
Average module X B , Å ²	17.7
Average solvent B , Å ²	35.6

* $R_{merge} = \sum |I - \langle I \rangle| / \sum \langle I \rangle$, where I is the observed intensity, and $\langle I \rangle$ is the statistically weighted average intensity of multiple observations.

[†] $R_{work} = \sum ||F_{calc}| - |F_{obs}|| / \sum |F_{obs}| \times 100$, where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively (R_{free} is calculated for a randomly chosen 5% of the reflections).

All polypeptide chains are well defined in the electron density map (with the exception of the residues mentioned above) with average B factors of 16.6 Å² for the cohesin module, 16.7 Å² for the X module and 16.6 Å² for the dockerin module. The high degree of similarity of this

structure when compared to the SdbA type II complex¹ is reflected by the low rmsd values between them - 1.12 Å for 166 Ca atoms of the whole complex, 0.86 Å for 156 Ca atoms of the Coh alone, 0.87 Å for 127 Ca atoms of the XDoc module, 0.77 Å for 83 Ca atoms of the X module alone and 0.78 Å for 44 Ca atoms of the Doc alone (**Figure V.4**).

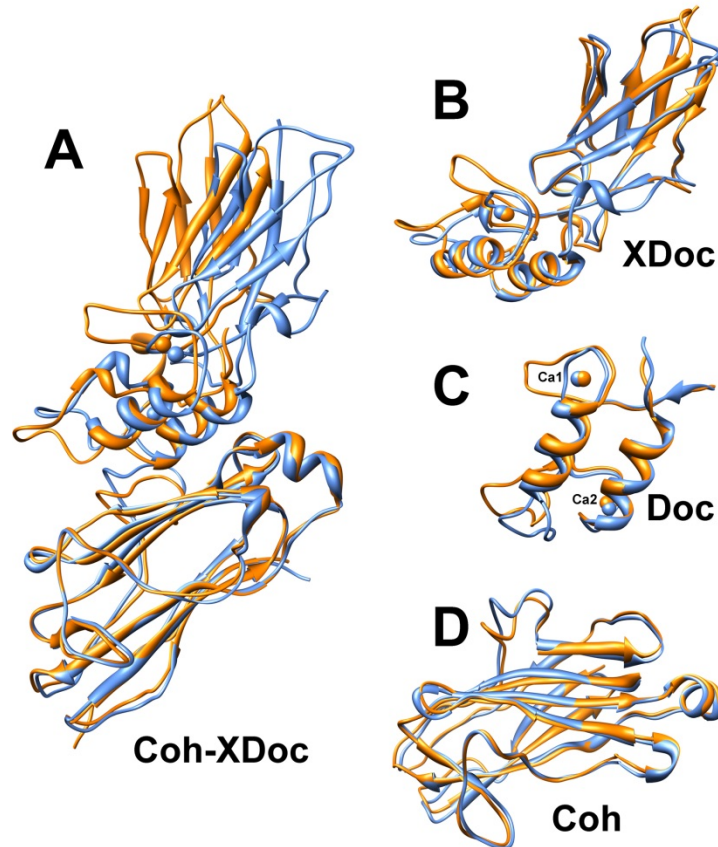


Figure V.4: Comparison of the structure of the Orf2 type II Coh-XDoc with the structure of the SdbA type II Coh-XDoc (PDB code: 2b59¹).

The structure of the Orf2 type II Coh-XDoc is represented as orange ribbons and the one from SdbA type II Coh-XDoc is represented as blue ribbons. **A)** Superposition of the entire complex taking the cohesin module as the reference; **B)** superposition of the XDoc complex; **C)** superposition of the Doc modules; **D)** superposition of the Coh modules.

One major difference between the Orf2 cohesin module and the one from SdbA is in the first calcium-binding loop (**Figure V.4 - C**), which is much larger in the Orf2 module (13 residues *versus* 8 residues in the Orf2 and SdbA modules, respectively). This has implications at the Coh-Doc interface level as it causes the loss of several Coh-Doc interactions (*see below, Section V.2.1.3*).

V.2.1.1 Type II Coh structure in the complex

The cohesin domain of the Coh-XDoc complex (**Figure V.5**) forms a flattened, elongated 9-stranded β -barrel with a jelly-roll topology. The nine β -strands define two β -sheets – the first β -sheet is defined by strands 8-3-6-5 (front face) and the second is defined by strands 9-1-2-7 (back face). Its core is highly hydrophobic. The α -helical crowning observed between strands 6 and 7 and the two β -flap regions that disrupt the normal progression of strands 4 and 8 are a common feature in this type of structure.^{1,8,13,19} The β -flap that disrupts the route of β -strand 8 forms a 12-residue raised loop (residues Glu132 to Gly143) that delimits the posterior face of the complex and forms several contacts with the dockerin (mainly with the second calcium-binding site). These β -flap regions are thought to be involved in the type II interaction and specificity (*see Section V.2.1.3*) but further experimental work is required in order to fully understand their role.^{8,13} Comparing this structure with the unbound SdbA type II Coh (PDB code 2bm3⁸) shows that, as in other related structures, the cohesin does not undergo significant conformational changes upon binding as revealed by the rmsd of 0.77 Å² (for 155 C α atoms) between both structures.

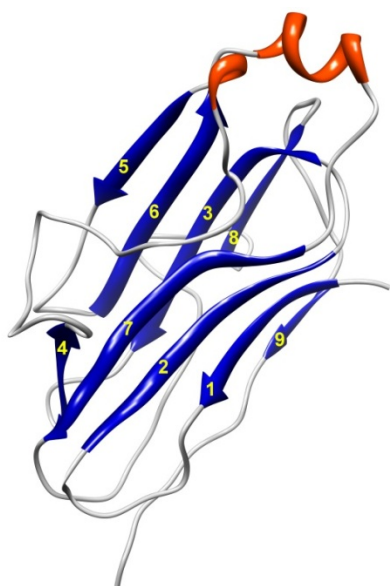


Figure V.5: Ribbon representation of the structure of the type II cohesin module of the Orf2 type II Coh-X-Doc complex.

The structure forms a flattened, elongated β -barrel with a jelly-roll topology and is composed by nine β -strands that form two β -sheets (8-3-6-5 and 9-1-2-7). The β -strands are depicted in blue, the helices are depicted in red and the random coil regions are depicted in grey.

V.2.1.2 Type II XDoc structure in the complex

The XDoc module (**Figure V.6 - A and B**) was modeled as one single polypeptide chain (chain B) of 164 amino acids (the first 98 belonging to the X module and the remaining to the dockerin). The X module subunit is composed of seven β -strands arranged into two β -sheets (1-4-7 and 2-3-5-6) and a small α -helix connecting stands 1 and 2. The overall fold of this subunit

and the β -sheet topology are similar to Ig-like module of avian carboxypeptidase D domain II (PDB code: 1qmu)²⁰ with a backbone rmsd of 0.96 Å² to this module.

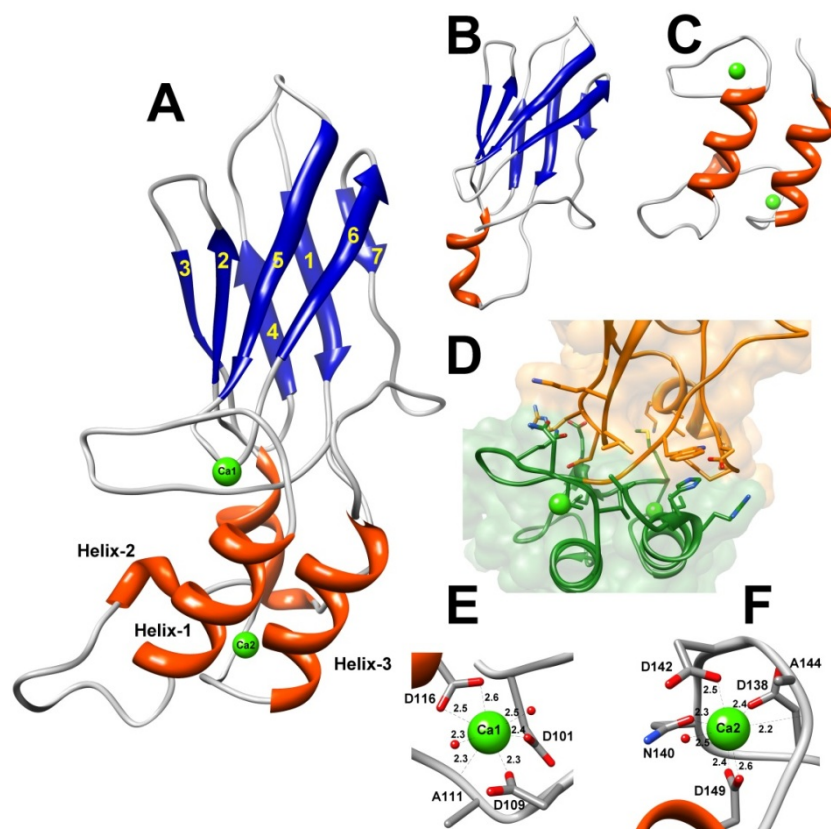


Figure V.6: Structure of the type II X-dockerin module of the Orf2 type II Coh-X-Doc complex.

A) The structure of type II XDoc module. The X module is formed by seven β -strands arranged into two β -sheets (1-4-7 and 2-3-5-6) and a small α -helix connecting strands 1 and 2. The dockerin module forms a typical EF-hand motif with a linker of 23 residues connecting helices 1 and 3 and has two calcium ions, coordinated in a typical octahedral geometry; **B)** Ribbon representation of the X module alone; **C)** Ribbon representation of the dockerin module alone; **D)** Highlight of the interaction surface between the X module and the dockerin. The residues involved in domain contacts are shown as sticks; **E)** and **F)** Coordination of the two calcium ions in the dockerin module. The residues involved in domain contacts are shown as sticks and the distances are indicated. The β -strands are depicted in blue, the helices are depicted in red, the random coil regions are depicted in grey and the calcium ions are represented as green spheres.

The type II dockerin domain (residues 99-164) forms two loop-helix motifs, named EF-hand motifs⁹, separated by a 23-residue linker that also forms a small helix (**Figure V.6 - A** and **C**). Helix 1 is defined by residues Met114 to Val 123, helix 2 (the one in the linker) is defined by residues Ala135 to Asp138 and helix 3 is formed by residues Leu147 to His156. Helices 1 and 3 are arranged in an antiparallel orientation that places the two calcium ions in opposite sides of the Doc module (**Figure V.6 - A** and **C**), similar to that observed for the type I Doc.⁶ Nonetheless, the linker in type II Doc is less structured than in the type I Doc, comprising only one turn in contrast with the three turns in the type I structures. The EF-hand motif loops bind to two calcium ions (**Figure V.6 - E** and **F**) coordinated in a typical octahedral geometry. The first

calcium ion, Ca1, is located near the C-terminus of the X module and is coordinated by residues Asp101 (O γ 1), Asp109 (O γ 1), Ala111 (backbone carbonyl), Asp116 (O γ 1 and O γ 2) and two water molecules (274 and 283). The second calcium, Ca2, is coordinated by residues Asp138 (O γ 1), Asn140 (O γ 1), Asp142 (O γ 1), Ala144 (backbone carbonyl), Asp149 (O γ 1 and O γ 2) and a water molecule (316). The residues involved in calcium coordination and the distances are given in **Table V.2**. These calcium ions are fundamental for the folding stabilization of the dockerin and for cohesin recognition. Furthermore, in the absence of the cohesin subunit, it was shown that binding of calcium to the XDoc module induces homodimerization.¹²

Table V.2: Calcium coordination in the dockerin domain

<i>Calcium ion</i>	<i>Residues/Atom</i>	<i>Distance (Å)</i>
<i>Ca1</i>	Asp101 - O δ 1	2.45
	Asp109 - O δ 1	2.33
	Ala111 - O	2.28
	Asp116 - O δ 1	2.49
	Asp116 - O δ 2	2.57
	H ₂ O274 - O	2.52
	H ₂ O283 - O	2.33
<i>Ca2</i>	Asp138 - O δ 1	2.36
	Asn140 - O δ 1	2.31
	Asp142 - O δ 1	2.49
	Ala144 - O	2.25
	Asp149 - O δ 1	2.60
	Asp149 - O δ 2	2.39
	H ₂ O316 - O	2.51

The X module and the dockerin form an intimate hydrophobic interface (**Figure V.6 - D** and **Figure V.7- A**) involving residues Asp18, Phe19, Asp20, Tyr21, Pro22, Glu24, Ser25, Lys28, Ile29, Lys70, Arg71, Asn72, Tyr73, Leu74, Lys75, Leu97 and Trp98 from the X module and residues Ala99, Gly100, Asp 101, Val102, Glu103, Gln108, Asn110, Ile112, Val134, Glu136, Leu137, Leu139, Asn140, Met141, Asp142, Ile152, Arg155, His156, Asn158 and Ala159 from the dockerin. These interactions include 10 direct hydrogen bonds, 4 water-mediated hydrogen bonds and 5 salt bridges (**Table V.3**). The contacts were calculated using the PISA server (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html).^{21,22}

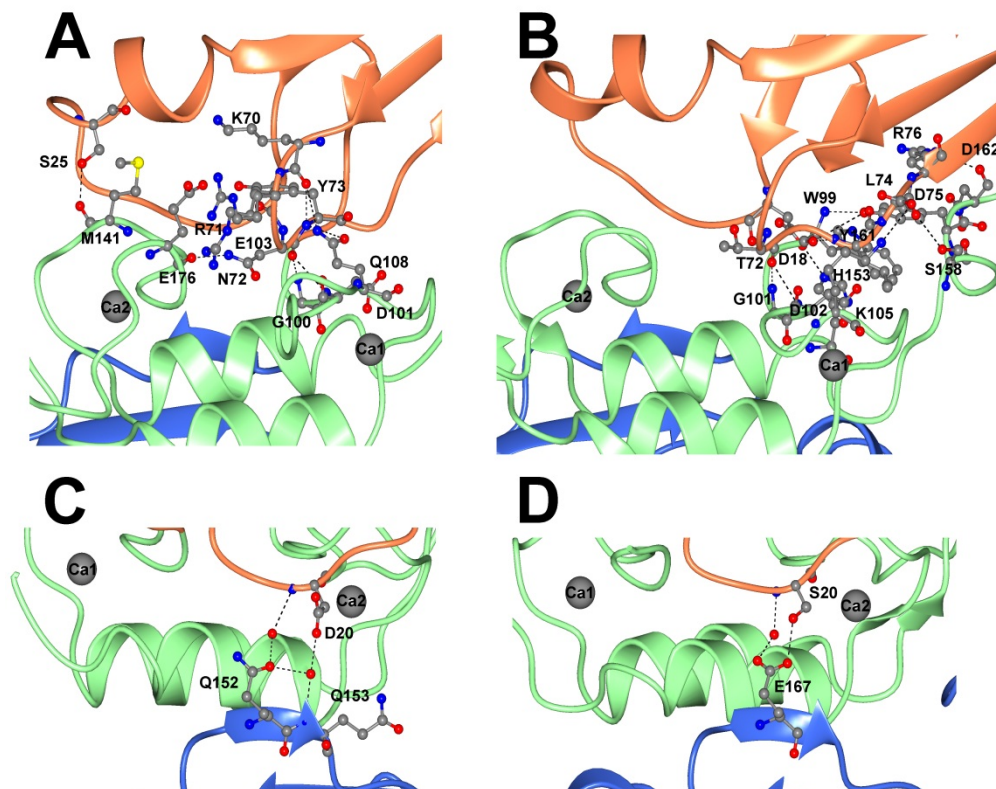


Figure V.7: The XDoc and X-Coh interface hydrogen bonds in the type II Orf2 and type II SdbA complexes.

A and B) The XDoc interface contacts in the type II Orf2 and SdbA complexes, respectively. **C and D)** The X-Coh interface contacts in the type II Orf2 and SdbA complex, respectively. The cohesin, doquerin and X module are represented as blue, green and orange ribbons, respectively; interface residues are represented as ball and sticks and depicted by heteroatom; calcium ions are represented as grey spheres and hydrogen bonds are represented as dashed lines.

Several studies^{1,12,16,18,23} have proposed a key role in structure stability and solubility of the cellulosome components for the X module. The highly hydrophobic interface between the X module and the dockerin and the extensive hydrogen bond network was pointed as the reason for the increased affinity of the type II cohesin for the XDoc modular pair¹ when compared to the interaction with the doquerin module alone²⁴ ($K_{a[\text{Coh-XDoc}]} = 1.44 \times 10^{10} \text{ M}^{-1}$ and $K_{a[\text{Coh-Doc}]} = 5.6 \times 10^8 \text{ M}^{-1}$). These contacts will help stabilize the dockerin module, therefore potentiating the Coh recognition and favoring the formation of the Coh-XDoc complex. When comparing the XDoc interface in the Orf2 complex with the one from the SdbA (**Figure V.7 A and B**) it is clear that there is a more extensive network of contacts in the Orf2 complex than in the SdbA one. While in the SdbA complex the X module only interacts with the first calcium-binding loop, in the Orf2 complex the interaction occurs with both loops. These contacts are probably needed for the dockerin stability and correct folding.

Moreover, the higher number of contacts in the Orf2 complex results in a more rigid complex that, as previously suggested, would reduce the entropic cost arising from a tightening of the isolated type II Doc structure upon type II Coh binding.¹

Table V.3: XDoc interface hydrogen bonds and salt bridges

<i>Direct hydrogen bonds</i>							
#	<i>Module X</i>		<i>Distance (Å)</i>	<i>Dockerin</i>			
	Residue	Atom		Residue	Atom		
1	Ser25	O γ	2.62	Met141	O		
2	Lys70	O	3.76	Gln108	N ϵ 1		
3	Arg71	N ϵ	2.98	Glu103	O ϵ 1		
4	Asn72	O	2.88	Glu103	O ϵ 1		
5	Asn72	O	3.14	Asp101	N		
6	Asn72	O	2.76	Gly100	N		
7	Asn72	O δ 1	3.00	Glu103	N		
8	Asn72	N δ 2	3.05	Glu136	O		
9	Tyr73	N	3.83	Asp101	O		
10	Tyr73	N	3.85	Gln108	O ϵ 1		

<i>Water-mediated hydrogen bonds</i>								
#	<i>X module</i>		<i>Distance (Å)</i>	<i>H₂O</i>		<i>Distance (Å)</i>	<i>Dockerin</i>	
	Residue	Atom		Residue	Atom		Residue	Atom
1	Asp18	O δ 1	2.81	H ₂ O270	O	3.13	Ala99	N
2	Lys70	O	3.58	H ₂ O282	O	2.74	Gln108	N ϵ 2
3	Lys70	O	3.58	H ₂ O282	O	3.51	Gln108	O ϵ 1
4	Asn72	O δ 1	2.92	H ₂ O244	O	2.14	Glu103	O

<i>Salt bridges</i>							
#	<i>Module X</i>		<i>Distance (Å)</i>	<i>Dockerin</i>			
	Residue	Atom		Residue	Atom		
1	Asp18	O δ 1	2.96	His156	N ϵ 2		
2	Asp18	O δ 2	3.61	His156	N δ 1		
3	Asp18	O δ 2	3.34	His156	N ϵ 2		
4	Arg71	N ϵ	2.98	Glu103	O ϵ 1		
5	Arg71	N η 1	3.40	Glu103	O ϵ 1		

The importance of the X module for the complex stability is further demonstrated by the presence of water-mediated hydrogen bonds with the cohesin module (**Figure V.7 - C and D** and **Table V.4**). In the SdbA complex, the two hydrogen bonds between Ser20 of the X-module and Glu167 of cohesin increase the cohesin–dockerin binding affinity by 2 orders of magnitude.^{1,24} Furthermore, molecular dynamics (MD) simulations of the cohesin-dockerin complex in the presence and absence of the X module have shown that the dockerin, when not

connected to the X module, becomes unstable and deviates largely from the crystal structure. MD simulations have also revealed that, in the absence of the X module, helix-1 from the dockerin is moved away from the cohesin, contrasting with the relatively fixed position when the X module is present.¹⁷ Another consequence upon removal of the X module is the structural fluctuation of the two calcium-binding loops which are vital for cohesin recognition. Based on these studies it was suggested that the X module is able to keep the binding sites of the dockerin in place by restricting its flexibility and orientation and this is the key for the enhanced affinity verified for the cohesin-dockerin affinity in the presence of the X module.¹⁷

Table V.4: X-Coh contacts.

<i>Water-mediated hydrogen bonds</i>								
#	<i>X module</i>		<i>Distance (Å)</i>	<i>H₂O</i>		<i>Distance (Å)</i>	<i>Cohesin</i>	
	Residue	Atom		Residue	Atom		Residue	Atom
<i>1</i>	Asp20	N	2.94	H ₂ O186	O	2.98	Gln152	Oε1
<i>2</i>	Asp20	Oδ2	3.15	H ₂ O153	O	3.18	Gln152	Oε1
<i>3</i>	Asp20	Oδ2	3.15	H ₂ O153	O	2.82	Gln153	N

In light of these observations and our results we can say that the probable reason why it was not possible to crystallize the Orf2 type II complex without the X module is due to the loss of all the above mentioned contacts that led to destabilization of the Coh-Doc complex, thus impairing its crystallization.

V.2.1.3 The complex interface

Contrary to the type I interaction but similar to the type II, both helix 1 and 3 of the dockerin domain in the Orf2 complex interact with the cohesin. The interaction surface is defined by residues Gly35, Ile36, Gln37, Asn76, Leu78, Thr80, Ala81, Val82, Asp84, Asn91, Tyr92, Ala93, Ser94, Cys95, Tyr96, Val97, Tyr98, Trp99, Arg135, Phe136, Pro138, Asn139, Leu145, Val146, Ile147, Tyr150, and Gly151 from the 8-3-6-5 face and loop region leading to the crowning helix between strands 6 and 7 of the cohesin module and residues Met114, Val117, Met118 and Ser121, from helix-1, residues Phe124, Gly125, Thr126, Arg127, Asp142, Gly143, Ala144 and Asn146 from the linker region and residues Leu147, Phe148, Ile150, Ala151, Ile154, Arg155 and Phe157 from helix-3 of the dockerin module (**Figure V.8**). Furthermore, residues Asp18 and Asp20 from the X module also make hydrophobic and water-mediated hydrogen bonds with residues Gln152 and Gln153 of the cohesin (**Figure V.7 C and D** and

Table V.4). The contacts were calculated using the PISA server (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html).^{21,22}

Interestingly, contrary to the SdbA type II complex¹ (PDB code: 2b59), in this case there is not a significant hydrogen bonding network at the interface. In fact, only two hydrogen bonds can be identified (**Table V.5**). This could lead to a weaker association between these two modules. However, in order to test this hypothesis, further experiments (like ITC) are required and are under way.

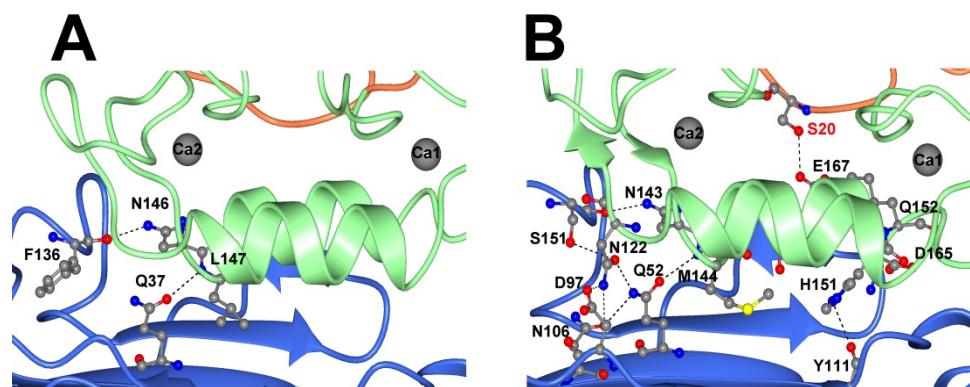


Figure V.8: The Coh-Doc and X-Coh interface hydrogen bonds in the type II Orf2 and type II SdbA complexes.

A and B) The Coh-Doc interface contacts in the type II Orf2 and SdbA complexes, respectively. The hydrophobic interface is defined between the 8-3-6-5 face of the cohesin and helices 1 and 3 of the dockerin. The cohesin, doquerin and X module are represented as blue, green and orange ribbons, respectively; interface residues are represented as ball and sticks and depicted by heteroatom; calcium ions are represented as grey spheres and hydrogen bonds are represented as dashed lines.

Table V.5: Coh-Doc interface hydrogen bonds

<i>Hydrogen bonds</i>					
#	<i>Cohesin</i>		<i>Distance (Å)</i>	<i>Dockerin</i>	
	Residue	Atom		Residue	Atom
1	Gln37	Oε1	2.84	Asn146	Nδ2
2	Phe136	O	2.99	Leu147	N

In the SdbA complex it was seen that residues at positions 10 and 11 of the calcium-binding loops make several contacts with residues of the cohesin. This was used to explain the complete abolition of Coh recognition when residues at both positions 10 and 11 of the second calcium-binding loop were mutated (Met → Ser and Gln → Ser, respectively).^{1,25} Concerning the Orf2 complex, we see that only Asp142 and Ala144, both belonging to the second calcium-binding loop, contact with the cohesin, namely with residues Pro138 and Asn139 from the β -flap region that disrupts the normal progression of β -strand 8 (**Table V.2**). This happens because in this

complex helix-1 of the dockerin lays a little bit further from the cohesin, enough to abolish any contacts from the calcium-binding loop with the cohesin.

V.2.1.3.1 Plasticity in the type II Coh-Doc complex

Upon determination of the first structure of the *C. thermocellum* type I dockerin it was seen that it displayed a near-perfect internal two-fold symmetry, such that residues 11-22 of helix-1 overlay with residues 35-56 of helix-3, and vice-versa (**Figure V.9 A**).⁶ Based on these observations it was proposed that a 180° rotation of the dockerin would result in cohesin recognition by helix-1 instead of the recognition by helix-3 (observed in the crystals), in which residues Ser11 and Thr12 would take the place of Ser45 and Thr46. This hypothesis was later confirmed by mutagenesis studies where residues Ser45 and Thr46 were mutated to alanine residues¹⁴ (Coh-DocS45A-T46A). In these experiments it was seen that the correct folding of the dockerin was retained and that residues in helix-1 (Ser11 and Thr12) were the ones dominating the interaction with the cohesin, proving that the observed internal symmetry was not just structural but also functional. This dual binding mode is thought to confer flexibility to the cellulosome function and assembly.

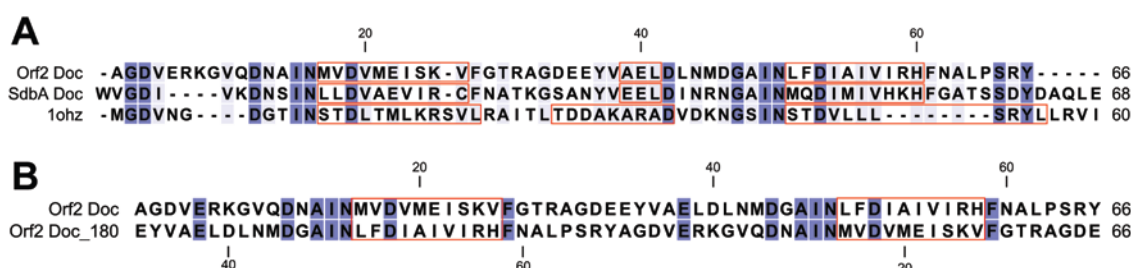


Figure V.9: Sequence alignment of the type II dockerins from the native Orf2 and SdbA complexes and the type I dockerin module

A) Sequence alignment of the type II dockerins from the native Orf2 (Orf2 Doc) and SdbA (SdbA Doc) complexes and the type I dockerin module (PDB code: 1ohz⁶). **B)** Sequence alignment of the native and 180°-rotated (Orf2 Doc_180) type II dockerin of the Orf2 complex. Residues are colored by similarity, where the darker the background, the higher the similarity. Red boxes delimit the three helices. The sequence alignment was performed with the software CLC Main Workbench 6.4 (CLC Bio, Denmark).

In the Orf2 type II dockerin, although there is a high internal structural similarity (**Figure V.10**), with an rmsd of only 0.26 Å² between the native and 180°-rotated structures, there is very little sequence similarity (**Figure V.9 B**). When comparing the native Coh-XDoc with the one with the XDoc module rotated 180° (the structure was built from the native structure by superposing helix-3 of the dockerin with its helix-1) we see that, interestingly, more potential hydrogen bonds are created (**Table V.6**). Nonetheless, in the rotated complex the water-mediated hydrogen bonds between the X module and the cohesin (thought to help stabilize the complex) are lost and several steric clashes between the Coh-Doc interfacing residues are found.

Table V.6: Coh-Doc interface hydrogen bonds in the 180°-rotated complex

<i>Hydrogen bonds</i>					
#	<i>Cohesin</i>		<i>Distance (Å)</i>	<i>Dockerin</i>	
	Residue	Atom		Residue	Atom
1	Gln37	Nε2	3.58	Asn113	Oδ1
2	Cys95	Sγ	2.31	Met118	Sδ
3*	Phe136	O	2.52	Leu147	N
4*	Gln37	Oε1	2.93	Asn146	Nδ2
5	Gly151	O	2.36	Lys122	Nζ

* Hydrogen bonds marked in red are the ones present in the native complex.

Moreover, when compared to the type I interaction, the affinity of the type II Coh-Doc interaction is much higher ($K_{a[\text{type I Coh-Doc}]} = 6.2 \times 10^6 \text{ M}^{-1}$, $K_{a[\text{type II Coh-Doc}]} = 5.6 \times 10^8 \text{ M}^{-1}$ and $K_{a[\text{type II Coh-XDoc}]} = 1.44 \times 10^{10} \text{ M}^{-1}$).^{1,6,24} The fact that in the type II complex both helices participate in the interaction with the cohesin, allied with the lack of internal symmetry in the type II Coh-Doc complexes, suggests that there is no dual binding mode in these complexes.

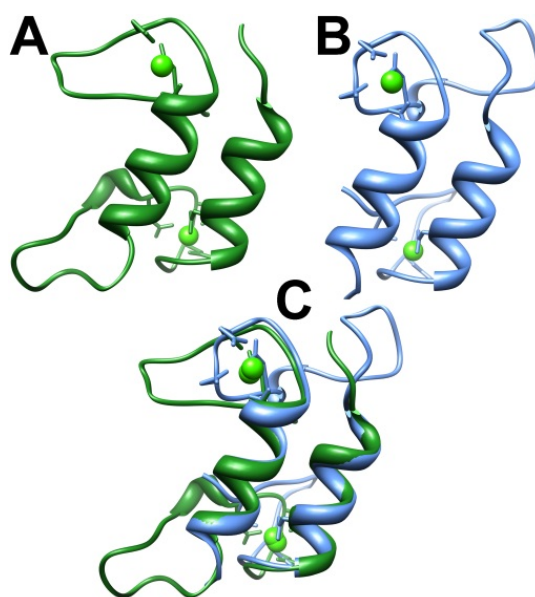


Figure V.10: Ribbon representation of the native and 180°-rotated type II Orf2 dockerin modules.

A) Native dockerin module. **B)** 180°-rotated dockerin module. **C)** Superposition of the native and 180°-rotated dockerin modules.

The absence of plasticity in the type II Coh-Doc interaction is thought to be related with selection between binding of the cellulosome catalytic modules and cell-surface attachment. This plasticity in the type I interaction confers increased flexibility in the quaternary architecture of the cellulosome and, conceivably, it may be required for the correct assembly of

the catalytic modules towards the different substrates. On the other hand, for the cell-surface attachment this feature is not a requirement and, thus there is no biological need for a dual binding mode. This fact, associated with the promiscuous inter-species cohesin-dockerin interaction⁷, suggests an evolutionary path where type II cohesins might have appeared first and were a common feature in cellulolytic organisms. Later they may have evolved into the type I modules, developing in the process ligand and species specificity according to their ecological niche.

V.3 Conclusions

The assembly of the enzymatic components into the cellulosome complex and the attachment of the last to the bacterial cell wall are of great significance for the overall process of plant cell wall degradation. In order to better understand this mechanism I have solved the crystal structure of the Orf2 type II Coh-XDoc from *C. thermocellum* (**Figure V.1**) to a resolution of 1.98 Å. The obtained structure is very similar to the SdbA type II Coh-XDoc structures determined by Adams *et al* (2006)¹, which is reflected in the low rmsd values between them - 1.12 Å for 166 C α atoms of the whole complex, 0.86 Å for 156 C α atoms of the Coh alone, 0.87 Å for 127 C α atoms of the XDoc module, 0.77 Å for 83 C α atoms of the X module alone and 0.78 Å for 44 C α atoms of the Doc alone (**Figure V.4**). The cohesin domain of the Coh-XDoc complex (**Figure V.5**) forms the typical flattened, elongated 9-stranded β -barrel with a jelly-roll topology and comparison of this structure with other cohesins shows that the cohesin does not undergo significant conformational changes.

The X module subunit is composed of seven β -strands arranged into two β -sheets and a small α -helix connecting stands 1 and 2 and its overall fold is similar to Ig-like module of avian carboxypeptidase D domain II²⁰. The type II dockerin domain forms the classical EF-hand motifs⁹, separated by a 23-residue linker that also forms a small helix (**Figure V.6 - A and C**). The interface of the XDoc complex is characterized by a high number of hydrophobic contacts (**Figure V.6 - D and Figure V.7- A**) which include 10 direct hydrogen bonds, 4 water-mediated hydrogen bonds and 5 salt bridges (**Table V.3**). These contacts occur with both calcium-binding loops of the dockerin module which are probably needed for the dockerin stability and correct folding. These observations suggest that the probable reason why it was not possible to crystallize the Orf2 type II complex without the X module is due to the loss of all the above mentioned contacts that led to destabilization of the Coh-Doc complex, thus impairing its crystallization.

Concerning the Coh-Doc interface, the obtained structure shows that both helix 1 and 3 of the dockerin domain in the Orf2 complex interact with the cohesin. This fact, allied to the lack of internal symmetry between helices 1 and 2 of the dockerin (as verified for type I dockerins) suggests that there is no dual binding mode in these complexes. This is thought to be related with selection between binding of the cellulosome catalytic modules and cell-surface attachment.

V.4 Materials and methods

V.4.1 Molecular biology

V.4.1.1 Transformation, expression, purification and quantification

The Orf2 type II Coh-XDoc complex of *C. thermocellum* was produced by first transforming the pET21a_Xdoc2_Orf2C2 expression vector into competent *E. coli* BL21 cells (Novagen). Recombinant *E. coli* cells were grown in LB media supplemented with ampicillin in a similar fashion as for CtCBMs 11, 30 and 44 (see Chapters II, III and IV). The complex was purified in three steps using an AKTA FPLC machine. The first step was IMAC purification in a HisTrapTM HP 5 ml column (GE Healthcare). The column was equilibrated with 50 mM NaHepes buffer, pH 7.5, containing 1 M NaCl, 10 mM imidazole and 5 mM CaCl₂. Proteins were eluted from the column in a gradient flow of the equilibration buffer and 50 mM sodium Hepes (NaHepes) buffer, pH 7.5, containing 1 M NaCl, 300mM Imidazole and 5 mM CaCl₂. The fractions containing the protein-protein complexes were selected by following native gel electrophoresis and SDS-PAGE.

Because the complex is usually co-purified with unbound cohesin, a control consisting exclusively of purified cohesin should be incorporated in the native gel to allow an easy identification of the complex band. The IMAC-purified proteins were then buffer-exchanged in PD-10 Sephadex G25M gel filtration columns (GE Healthcare) into 20 mM Tris-HCl buffer, pH 8.0, and 5 mM CaCl₂ (as previously – see Chapter II). The proteins were then subjected to another purification step by anion exchange chromatography using a column loaded with Source 30Q media (GE Healthcare). The separation of the individual proteins from the complex was achieved through the application of a 0-1 M NaCl elution gradient. Prior to filtration chromatography the protein fractions were buffer-exchanged into 20 mM NaHepes buffer, pH 7.5, containing 200 mM NaCl and 2 mM CaCl₂. The purity of the protein was confirmed by running a native gel electrophoresis and SDS-PAGE on the collected fractions. The purified

protein was concentrated with Amicon centricons with 10-kDa molecular-mass centrifugal membranes (Millipore, Billerica, MA, USA) by centrifuging at 5000 rpm at 4°C. The final concentration of the protein was kept around 20 mg/ml.

Protein expression, purification and complex crystallization (described below) were performed by Professor Carlos Fontes' group at the Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa and the protein kindly provided to us.

V.4.2 X-ray crystallography

V.4.2.1 Protein crystallization and data collection

The Type II complex Coh-XDoc was crystallized at 293K by the hanging drop vapor diffusion method and obtained by mixing an equal volume (1 μ l) of protein (20 mg/ml in water) and reservoir solution (12% (m/v) polyethyleneglycol (PEG) 3350, 4% tacsimate, pH 5.0). Single crystals were harvested and flash-frozen in a liquid nitrogen stream at 100K, using 30% (vol/vol) of glycerol as a cryoprotectant.

The data was collected at a wavelength 0.9735 in the European Synchrotron Radiation Facility (ESRF), ID14-4 (Grenoble, France) to 1.98 Å resolution at 100 K. Diffraction data were processed and scaled, respectively, with programs MOSFLM²⁶ and SCALA²⁷ from the CCP4 suite²⁸. The Matthews coefficient of the Orf2 type II Coh-XDoc crystal is 2.2 Å³ Da⁻¹ for one heterodimer in the asymmetric unit, with a solvent content of 43.13%. The space group was determined to be C121 with unit cell dimensions: $a = 116.67$ Å, $b = 78.63$ Å, $c = 35.80$ Å, with $\beta = 95.87^\circ$ (Table V.1).

V.4.2.2 Phasing, model building and refinement

Considering the calculated Matthews coefficient, molecular replacement attempts were performed searching for one Coh-XDoc complex in the monoclinic C121 cell (see Appendix B, Section B.2.2). The previously described and available crystal structure of the Orf2 type II Coh-XDoc complex from *C. thermocellum*, with accession code 2b59¹, was used as a search model for molecular replacement (see Chapter VIII, Section VIII.4.2.1). The Patterson search was done with program PHASER²⁹, implemented in the CCP4 interface²⁸, and a clear solution was found in space group C121. Initial building of the complex into the electron density was performed using ARP/wARP^{28,30} and the remaining residues were built interactively using program COOT³¹. Model refinement and electron density map calculations were done with program REFMAC5³² from the CCP4 suite²⁸. The final model has $R_{cryst} = 18.7\%$ and $R_{free} = 24.7\%$ and

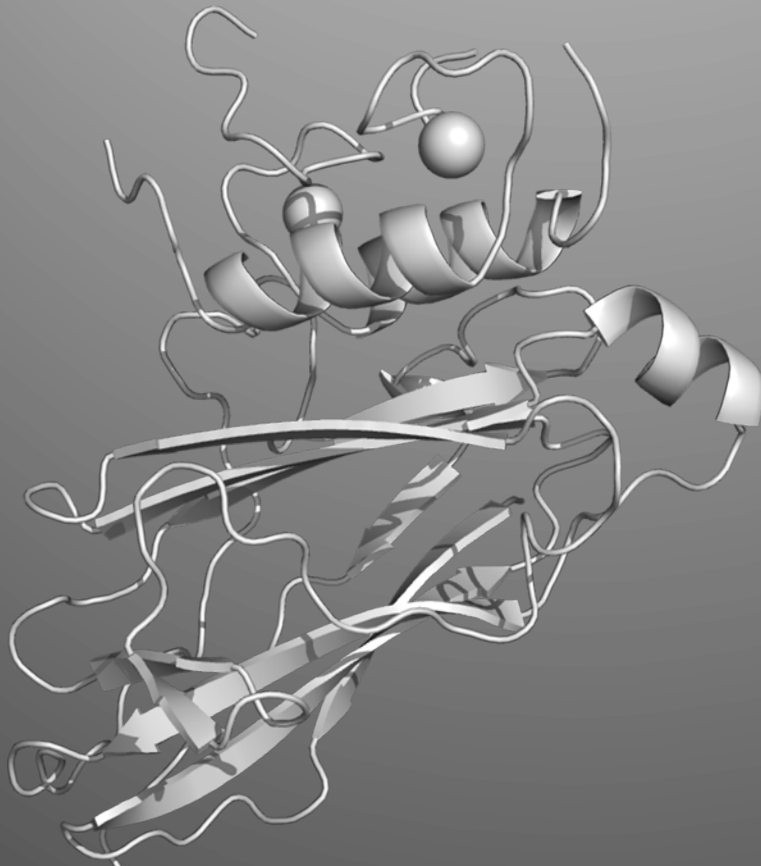
includes 322 water molecules and two calcium ions. Due to disorder, residues Met1 and Ala1 of the Coh module (chain A), Met1, Asn2, Asn3, Asp4, Ser5 and Thr6 of the X module (chain B) and Leu160, Pro161, Ser162, Arg163 and Tyr164 from the Doc module (chain B), as well as the side chains of residues Arg73, Lys158 and terminal His-tag from the Coh module and Glu63 and Lys85 from the X module were not observed. The structure is deposited in the Protein Data Bank under the accession code: 2vt9.

V.5 References

1. Adams, J. J.; Pal, G.; Jia, Z. C.; Smith, S. P., Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. *P Natl Acad Sci USA* **2006**, *103* (2), 305.
2. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R., The cellulosomes: Multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* **2004**, *58*, 521.
3. Uversky, V. N.; Kataeva, I. A., *Cellulosome*. Nova Science Publishers: New York, 2006; p xiii.
4. Miras, I.; Schaeffer, F.; Beguin, P.; Alzari, P. M., Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. *Biochemistry* **2002**, *41* (7), 2115.
5. Adams, J. J.; Currie, M. A.; Ali, S.; Bayer, E. A.; Jia, Z. C.; Smith, S. P., Insights into Higher-Order Organization of the Cellulosome Revealed by a Dissect-and-Build Approach: Crystal Structure of Interacting *Clostridium thermocellum* Multimodular Components. *Journal of Molecular Biology* **2010**, *396* (4), 833.
6. Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A., Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *P Natl Acad Sci USA* **2003**, *100* (24), 13809.
7. Fontes, C. M. G. A.; Gilbert, H. J., Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry*, Vol 79 **2010**, *79*, 655.
8. Carvalho, A. L.; Pires, V. M. R.; Gloster, T. M.; Turkenburg, J. P.; Prates, J. A. M.; Ferreira, L. M. A.; Romao, M. J.; Davies, G. J.; Fontes, C. M. G. A.; Gilbert, H. J., Insights into the structural determinants of cohesin dockerin specificity revealed by the crystal structure of the type II cohesin from *Clostridium thermocellum* SdbA. *Journal of Molecular Biology* **2005**, *349* (5), 909.
9. Pages, S.; Belaich, A.; Belaich, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A., Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins* **1997**, *29* (4), 517.
10. Mechaly, A.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Lamed, R.; Shoham, Y.; Bayer, E. A., Cohesin-dockerin interaction in cellulosome assembly - A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *Journal of Biological Chemistry* **2001**, *276* (13), 9883.
11. Haimovitz, R.; Barak, Y.; Morag, E.; Voronov-Goldman, M.; Shoham, Y.; Lamed, R.; Bayer, E. A., Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. *Proteomics* **2008**, *8* (5), 968.

12. Adams, J. J.; Webb, B. A.; Spencer, H. L.; Smith, S. P., Structural characterization of type II dockerin module from the cellulosome of *Clostridium thermocellum*: Calcium-induced effects on conformation and target recognition. *Biochemistry* **2005**, *44* (6), 2173.
13. Noach, I.; Frolow, F.; Jakoby, H.; Rosenheck, S.; Shimon, L. J. W.; Lamed, R.; Bayer, E. A., Crystal structure of a type-II cohesin module from the *Bacteroides cellulosolvens* cellulosome reveals novel and distinctive secondary structural elements. *Journal of Molecular Biology* **2005**, *348* (1), 1.
14. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A.; Gilbert, H. J., Evidence for a dual binding mode of dockerin modules to cohesins. *P Natl Acad Sci USA* **2007**, *104* (9), 3089.
15. Pinheiro, B. A.; Proctor, M. R.; Martinez-Fleites, C.; Prates, J. A.; Money, V. A.; Davies, G. J.; Bayer, E. A.; Fontesm, C. M.; Fierobe, H. P.; Gilbert, H. J., The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. *J Biol Chem* **2008**, *283* (26), 18422.
16. Mosbah, A.; Belaich, A.; Bornet, O.; Belaich, J. P.; Henrissat, B.; Darbon, H., Solution structure of the module X2 1 of unknown function of the cellulosomal scaffolding protein CipC of *Clostridium cellulolyticum*. *J Mol Biol* **2000**, *304* (2), 201.
17. Xu, J.; Smith, J. C., Probing the mechanism of cellulosome attachment to the *Clostridium thermocellum* cell surface: computer simulation of the Type II cohesin-dockerin complex and its variants. *Protein Eng Des Sel* **2010**, *23* (10), 759.
18. Kataeva, I. A.; Uversky, V. N.; Brewer, J. M.; Schubot, F.; Rose, J. P.; Wang, B. C.; Ljungdahl, L. G., Interactions between immunoglobulin-like and catalytic modules in *Clostridium thermocellum* cellulosomal cellobiohydrolase CbhA. *Protein Eng Des Sel* **2004**, *17* (11), 759.
19. Noach, I.; Lamed, R.; Xu, Q.; Rosenheck, S.; Shimon, L. J. W.; Bayer, E. A.; Frolow, F., Preliminary X-ray characterization and phasing of a type II cohesin domain from the cellulosome of *Acetivibrio cellulolyticus*. *Acta Crystallogr D* **2003**, *59*, 1670.
20. Gomis-Ruth, F. X.; Companys, V.; Qian, Y.; Fricker, L. D.; Vendrell, J.; Aviles, F. X.; Coll, M., Crystal structure of avian carboxypeptidase D domain II: a prototype for the regulatory metallo-carboxypeptidase subfamily. *Embo J* **1999**, *18* (21), 5817.
21. Krissinel, E.; Henrick, K. Protein interfaces, surfaces and assemblies service PISA at European Bioinformatics Institute. http://www.ebi.ac.uk/pdbe/prot_int/pistart.html.
22. Krissinel, E.; Henrick, K., Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **2007**, *372* (3), 774.
23. Schubot, F. D.; Kataeva, I. A.; Chang, J.; Shah, A. K.; Ljungdahl, L. G.; Rose, J. P.; Wang, B. C., Structural basis for the exocellulase activity of the cellobiohydrolase CbhA from *Clostridium thermocellum*. *Biochemistry* **2004**, *43* (5), 1163.
24. Jindou, S.; Kajino, T.; Inagaki, M.; Karita, S.; Beguin, P.; Kimura, T.; Sakka, K.; Ohmiya, K., Interaction between a type-II dockerin domain and a type-II cohesin domain from *Clostridium thermocellum* cellulosome. *Bioscience, biotechnology, and biochemistry* **2004**, *68* (4), 924.
25. Schaeffer, F.; Matuschek, M.; Guglielmi, G.; Miras, I.; Alzari, P. M.; Beguin, P., Duplicated dockerin subdomains of *Clostridium thermocellum* endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity. *Biochemistry* **2002**, *41* (7), 2106.
26. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletters on Protein Crystallography* **1992**, *26*.
27. Evans, P. R., Scaling of MAD data. In *Proceedings of the CCP4 Study Weekend. Recent advances in phasing*, Winn, M., Ed. 1997; Vol. 33, pp 22.
28. Bailey, S., The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallogr D* **1994**, *50*, 760.
29. McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr D* **2005**, *61*, 458.

30. Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A., Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **2008**, 3 (7), 1171.
31. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **2004**, 60, 2126.
32. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* **1997**, 53, 240.



Chapter VI: The ScaA type II Cohesin-Dockerin complex from B. cellulosolvens

In this chapter I have used X-ray crystallography to determine the 3D structure of the ScaA Type II Cohesin-Dockerin complex from Bacteroides cellulosolvens. At this time, this is the first cohesin-dockerin complex ever determined from B. cellulosolvens. Moreover, the data shows also for the first time the 3D structure of a B. cellulosolvens dockerin and evidences the possibility for an alternate binding mode, similar to the one proposed for C. thermocellum. The results here presented are part of a manuscript currently in preparation.

Table of Contents

Summary	185
VI.1 Introduction	186
VI.2 Results and Discussion.....	187
VI.2.1 Architecture of the type II Coh-Doc complex from <i>B. cellulosolvens</i>	187
VI.2.1.1 Type II Coh structure in the complex.....	189
VI.2.1.2 Type II Doc structure in the complex.....	190
VI.2.1.3 The complex interface – an alternative binding mode	192
VI.3 Conclusions.....	196
VI.4 Materials and methods	196
VI.4.1 Molecular biology	196
VI.4.1.1 Transformation, expression, purification and quantification	196
VI.4.2 X-ray crystallography.....	197
VI.4.2.1 Protein crystallization and data collection.....	197
VI.4.2.2 Phasing, model building and refinement.....	197
VI.5 References	198

Summary

In this chapter I describe the 1.9 Å crystal structure of the ScaA type II cohesin-dockerin (Coh-Doc) of *Bacteroides cellulosolvens* (*B. cellulosolvens*, *Bc*) as determined by X-ray crystallography using molecular replacement (**Figure VI.1**).

At the time of writing, this is the first cohesin-dockerin complex ever determined from *B. cellulosolvens*. Furthermore, for the first time, it reveals the 3D structure of a type II dockerin of this organism and it shows the possibility of an alternate binding mode between the cohesin and the dockerin, in a similar way to that proposed for the type I interaction in *C. thermocellum*. The cohesin

domain in the complex is similar to the free domain as shown by the low rmsd between both structures (rmsd = 0.66 Å for 166 Ca atoms). The structure of the dockerin domain is very similar to the type I dockerins from *C. thermocellum* with the main differences in helix 2, which has a high degree of disorder in this complex. As in those structures, there is an internal two-fold symmetry between helix 1 and 3. This internal symmetry is shown by the low rmsd values between both helices (0.62 Å for 24 Ca atoms). Remarkably, in this complex the dockerin is rotated 180° when compared to other native cohesin-dockerin complexes determined so far^{8,9,12}. This represents the first native complex in which the predicted dual binding mode^{13,15} is verified. This feature confers a large degree of plasticity to the complex and has profound implications at the level of the current understanding of cellulosome architecture and assembly.

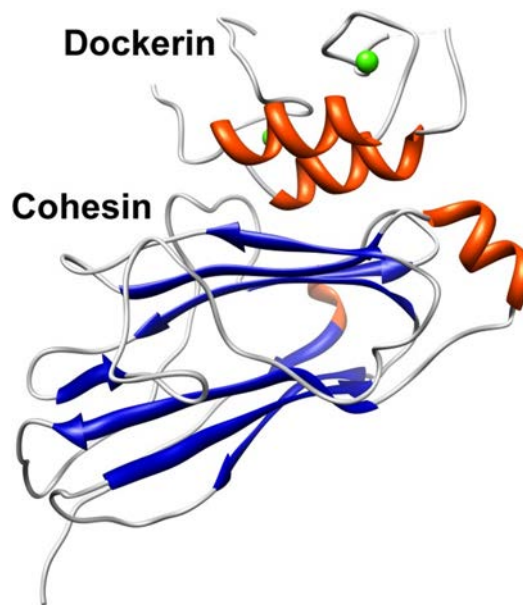


Figure VI.1: Crystal structure of the type II cohesin-dockerin complex (Coh-Doc) from *B. cellulosolvens* (PDB code: 2y3n)

The two calcium ions are depicted as green spheres; α -helical regions are depicted in red; β -sheet regions are depicted in blue and random coil regions are depicted in grey. Residues 32-37 are missing and represented as a light grey dashed line.

VI.1 Introduction

Recycling of photosynthetically fixed carbon is a crucial microbial process, critical to the cycling of carbon between microbes, herbivores and plants. *Bacteroides cellulosolvens* (*B. cellulosolvens*, *Bs*) is a mesophilic, anaerobic bacterium capable of degrading crystalline cellulose.^{1,2} Like *C. thermocellum*, *B. cellulosolvens* produces extracellular cellulolytic complexes – **cellulosomes** – responsible for the degradation of the plant cell wall.³

The outstanding capabilities of cellulosomes have drawn a great deal of attention in the past years for biotechnological applications. Thus, understanding the properties of this mega Dalton complex, its architecture and assembly via the cohesin-dockerin interactions is fundamental before any technological advance can be made. In traditional cellulosomes (as is the case of the one from *C. thermocellum*), assembly of the different enzymes and non-catalytic modules to the scaffoldin subunit is mediated by type I cohesin-dockerin interactions whereas the anchoring of the scaffoldin to the bacterial cell wall is mediated by type II cohesin-dockerin interactions^{4,5} (see Chapter V). In *B. cellulosolvens*, the sequencing of the primary scaffoldin subunit (initially termed CipBc and latter termed ScaA) revealed the presence of 11 type II, rather than type I, cohesins.³ Furthermore, these type II cohesins lacked the associated X module (**Figure VI.2**). Phylogenetic analysis further confirms that the ScaA cohesins are indeed type II and places them in close proximity to the type II cohesins from *C. thermocellum* anchoring proteins (**Figure VI.2 - B**).^{3,6} In a similar way to other scaffoldin proteins, the ScaA scaffoldin carries a dockerin domain at its *C-terminus*. This dockerin domain is similar to type I dockerins from *C. thermocellum*, with a near-perfect internal symmetry and the proposed recognition dyads in positions 10 and 11.^{3,6}

The sequence of secondary scaffoldin of *B. cellulosolvens* (ScaB) confirmed its participation in cell-surface anchoring through its SLH domain (**Figure VI.2**) and the type I character of its cohesins. ScaB is composed of 10 sequential type I cohesins followed by an X module and an SLH domain, which are closely associated, with little or no detectable linker sequence.⁶ Overall, the cellulosome of *B. cellulosolvens* comprises a total of 110 enzymes and shows typical features of a powerful cellulolytic complex (large variety of cell wall degrading enzymes, substrate recognition modules and synergistic effects).

Even though several high-resolution structures of cohesins and dockerins have already been determined⁸⁻¹⁴, only the structure of the 11th type II cohesin module (cohesin₁₁) of *B. cellulosolvens* has been reported so far (PDB code: 1tyj)¹¹. This type II cohesin module shows an overall fold similar to the Orf2 type II cohesin from *C. thermocellum* with the characteristic α -helical crown and the two singular β -flaps that flank the protein (see Chapter V).

In this chapter I report the crystal structure of the multimodular heterodimeric SdbA type II cohesin₁₁-dockerin from *Bacteroides cellulosolvans* to a resolution of 1.90 Å (PDB code: 2y3n).

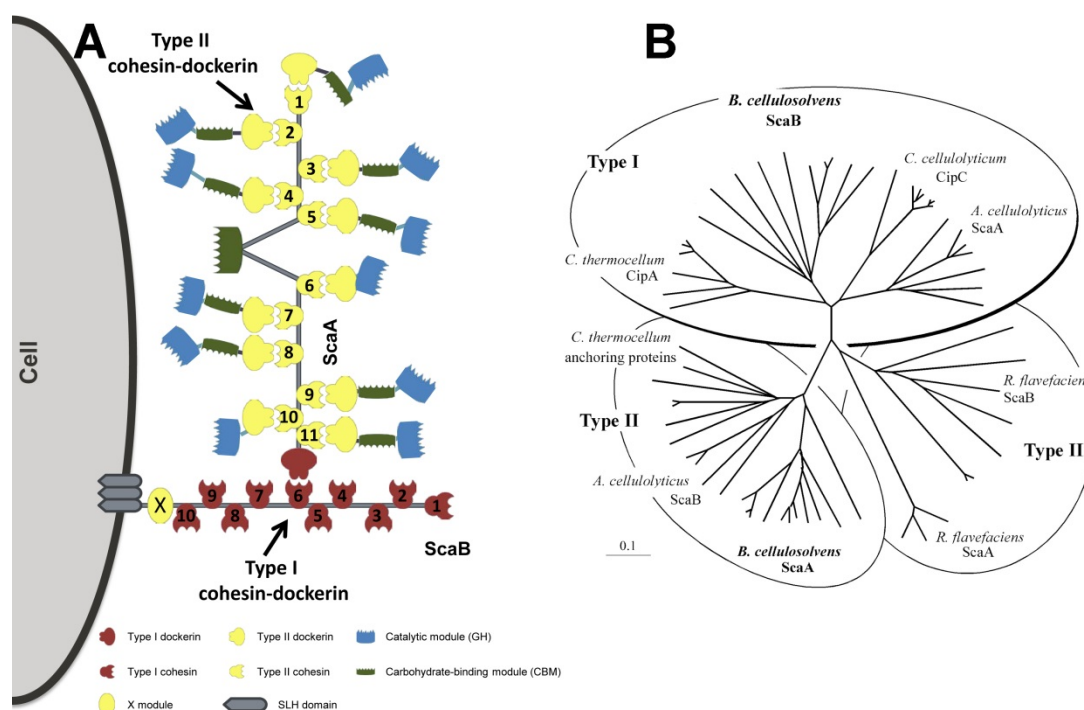


Figure VI.2: Schematic representation of the *Bacteroides cellulosolvans* cellulosome (A) and phylogenetic relationships of the ScaA and ScaB cohesins (B).

The primary scaffoldin in the *B. cellulosolvans* cellulosome (ScaA) is organized into 11 type II cohesin domains, an internal CBM3 module and a C-terminal dockerin. Phylogenetic relationships place the ScaA cohesins in close proximity to the type II cohesins from *C. thermocellum* anchoring proteins and the ScaB cohesins close to the type I cohesins from *C. thermocellum*. The binding of the enzymes to specific positions is hypothetical, as is the linear orientation of the scaffoldin. The scaffoldins are only sketched partially. All cellulosome components are not drawn to scale. Adapted from Noach *et al.*, 2005⁷ and Xu *et al.*, 2004⁶.

VI.2 Results and Discussion

VI.2.1 Architecture of the type II Coh-Doc complex from *B. cellulosolvans*

I have solved the crystal structure of the 11th ScaA type II cohesin-dockerin (Coh₁₁-Doc) complex from *B. cellulosolvans* (Figure VI.1) by molecular replacement (MR) using as model the ScaA type II Coh complex (PDB code: 1tyj⁷) which yielded a solution with two cohesins in the asymmetric unit. The dockerin modules were built using the software ARP/wARP.¹⁵⁻¹⁷ The data was refined at 1.90 Å resolution and the final statistics are summarized in Table VI.1. The

final model has $R_{cryst} = 16.3\%$ and $R_{free} = 22.5\%$ and includes 299 water molecules and four calcium ions. Due to disorder, residues Met1, Ala2, and the 6 C-terminal histidines of chain A (cohesin), Gly32-Asn37 and Ala66-Phe71 of chain B (dockerin), Met1, Ala2, Leu174, Glu175 and the 6 C-terminal histidines of chain C (cohesin) and Ala30-Asn44 and Ser65-Phe71 of chain D (dockerin) could not be built. The structure is deposited in the Protein Data Bank under the accession code: 2y3n. The great extent of the polypeptide chain is well defined in the electron density map (with the exception of the residues mentioned above) with average B factors of 22.8 and 22.6 \AA^2 for the cohesin modules in chain A and C, respectively and 31.9 and 45.8 \AA^2 for the dockerin modules in chains B and D, respectively. For the calcium ions the B factors are 25.6 and 22.9 \AA^2 for the dockerin in chain B and 27.3 and 61.8 \AA^2 for the dockerin in chain D. The high temperature factor of calcium 2 in chain D reflects the disorder of the calcium-binding residues in that area.

Curiously, in this complex the dockerin module is rotated 180° when compared to other Coh-Doc structures determined so far^{8,9,12}. Despite the inherent difficulties in interpreting an electron density of a protein with a dyad symmetry, some sequence differences, for instance, Gly11/Asn44, Met17/Ser50 and Ser26/Phe59, allowed the correct and unambiguous protein orientation and assignment. The implications of this binding mode are discussed below (*Section VI.2.1.3*). **Figure VI.3** illustrates the internal symmetry found in this dockerin module. This internal symmetry is reflected by the low rmsd values between both helices (0.62 \AA for 24 Ca atoms).



Figure VI.3: Sequence alignment showing the dyad symmetry within the dockerin sequence

The sequence alignment was performed with the software CLC Main Workbench 6.4 (CLC Bio, Denmark).

Table VI.1: X-ray data and structure quality statistics for the *B. cellulosolvens* type II Coh–Doc complex.

<i>Data quality</i>	<i>BcCoh-DocII</i>
Cell dimensions, \AA	$a = 43.4$ $b = 116.1$ $c = 45.2$
Space group	$\beta = 112.5^\circ$ $P2_1$
X-ray source	European Synchrotron Radiation Facility, ID14-EH1
Wavelength, \AA	0.934

Resolution of data (outer shell), Å	41.74-1.90 (2.00-1.90)
R_{pim} (outer shell), %	0.073 (0.278)
R_{merge} (outer shell), %*	0.090-0.051 (0.329)
Mean I/σ (I)	3.9627
Completeness (outer shell), %	83.9 (66.4)
Multiplicity (outer shell)	2.40 (2.2)
Structure quality	
N° of atoms (AU)	3765
N° ligand atoms	4
N° solvent waters	299
Resolution used in refinement, Å	1.90
R_{cryst}/R_{free} (%) [†]	16.3/22.5
Ramachandran's plot analysis	
Favorable %	96.1
Allowed %	3.6
Outlier %	0.2

* $R_{merge} = \sum |I - \langle I \rangle| / \sum \langle I \rangle$, where I is the observed intensity, and $\langle I \rangle$ is the statistically weighted average intensity of multiple observations.

[†] $R_{work} = \sum ||F_{calc}| - |F_{obs}|| / \sum |F_{obs}| \times 100$, where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively (R_{free} is calculated for a randomly chosen 5% of the reflections).

VI.2.1.1 Type II Coh structure in the complex

The cohesin domain of the type II Coh₁₁-Doc complex of *B. cellulosolvens* shows the typical flattened, elongated 9-stranded β -barrel jelly-roll topology (**Figure VI.4**). Similar to the *C. thermocellum* structure (see Chapter V), the nine β -strands define two β -sheets – the first β -sheet is defined by strands 8-3-6-5 (front face) and the second is defined by strands 9-1-2-7 (back face). Its core is highly hydrophobic. The common α -helical crowning observed between strands 6 and 7 and the two β -flap regions that disrupt the normal progression of strands 4 and 8 are maintained. Comparing this structure with the unbound ScaA type II Coh₁₁ (PDB code: 1tyj¹¹) shows that the cohesin does not undergo significant conformational changes upon binding as revealed by the low rmsd value (0.66 Å for 166 C α atoms) between both structures.

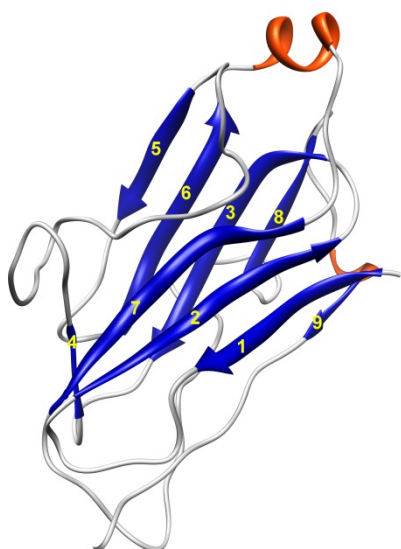


Figure VI.4: Ribbon representation of the structure of the type II cohesin module of the ScaA type II Coh₁₁-Doc complex.

The structure forms a flattened, elongated β -barrel with a jelly-roll topology and is composed of nine β -strands that form two β -sheets (8-3-6-5 and 9-1-2-7). The β -strands are depicted in blue, the helices are depicted in red and the random coil regions are depicted in grey.

VI.2.1.2 Type II Doc structure in the complex

The dockerin domain of the type II Coh₁₁-Doc complex of *B. cellulosolvens* reveals a classic structure^{12,14,18-20} (**Figure VI.5 - A**), composed of two loop-helix motifs, named EF-hand motifs²¹, separated by a 12-residue unstructured linker. Helix 1 is formed by residues Asn16 to Ser26 and helix 3 is formed by residues Ser50 to Phe60. Helices 1 and 3 are arranged in an antiparallel orientation that places the two calcium ions in opposite sides of the dockerin module, similar to that observed for other dockerins. The EF-hand motif loops bind to two calcium ions (**Figure VI.5 - B and C**) coordinated in a typical octahedral geometry. The first calcium ion, Ca1, is located near the *N*-terminus of the dockerin and is coordinated by residues Asp8 (O δ 1), Asn10 (O δ 1), Asp12 (O δ 1), Val14 (backbone carbonyl), Asp19 (O δ 1 and O δ 2) and a water molecule. The second calcium, Ca2, is coordinated by Asp41 (O δ 1), Asn43 (O δ 1), Asp45 (O δ 1), Val47 (backbone carbonyl), Asp52 (O δ 1 and O δ 2) and a water molecule. The residues involved in calcium coordination and the distances are given in **Table VI.2**. The 12-residue linker region between helices 1 and 3 (Phe27-Asn49) shows a large degree of mobility impairing the building of several residues (*see Section VI.2.1*). This is more evident in the dockerin in chain D, where the calcium ion has a temperature factor of 61.8 Å² and residues Asp41 and Asn43, both participating in calcium coordination, could not be built. Since the type II complex of *B. cellulosolvens* lacks the X module, which is thought to help stabilize the cohesin-dockerin interaction, it is possible that for the correct assembly of this complex, the presence of other(s) module(s) of the enzyme is (are) required. In order to further investigate this possibility, the solution of more structures is essential.

When comparing this dockerin module with both a CipA type I (Coh₂ - PDB code: 1ohz¹²) and the SdbA type II (PDB code: 2vt9) dockerins from *C. thermocellum* (**Figure VI.5 - D and E**, respectively) we see that the main differences are at the level of the linker region, which is

not structured. At the level of helices 1 and 3, the similarity with both type I and type II dockerins is high as shown by the low rmsd values (0.84 Å for 23 C α atoms and 0.70 31 C α atoms for the type I and type II modules, respectively). Interestingly, the first calcium-binding region is shorter than in the SdbA complex, bringing it closer to the type I structure. Moreover the internal symmetry between helices 1 and 3 is not only structural, as in the case of the SdbA type II complex (see Chapter V) but also sequential, as in the type I module (Figure VI.3). This feature opens the door for the alternative binding mode verified in this complex (see Section VI.2.1.3).

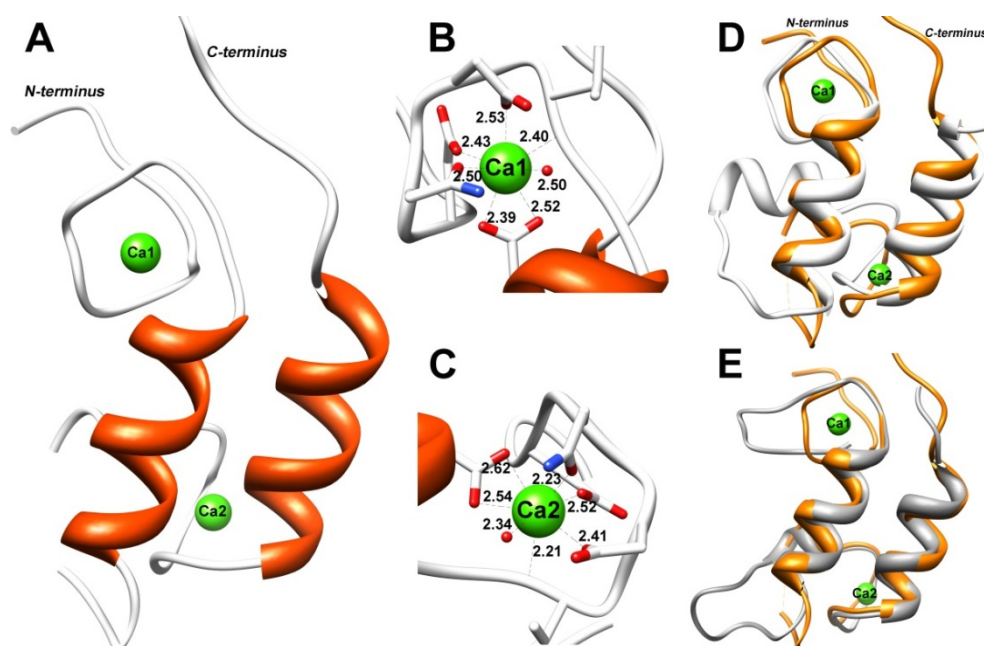


Figure VI.5: Ribbon representation of the structure of the type II dockerin module of the ScaA Coh₁₁-Doc complex.

A) Ribbon representation of the structure of the type II dockerin module of the ScaA type II Coh₁₁-Doc complex; B and C) Coordination of the calcium ions 1 and 2, respectively; D and E) Superposition of the dockerin module of the ScaA type II Coh₁₁-Doc complex (orange) with a CipA type I dockerin (Coh₂ - white) and the Orf2 type II dockerin from *C. thermocellum* (grey), respectively.

Table VI.2: Calcium coordination in the dockerin domain

Calcium ion	Residues/Atom	Distance (Å)
<i>Ca1</i>	Asp8 - O δ 1	2.43
	Asn10 - O δ 1	2.50
	Asp12 - O δ 1	2.53
	Val14 - O	2.50
	Asp19 - O δ 1	2.39
	Asp19 - O δ 2	2.52
	H ₂ O7 - O	2.50

	Asp41 - O δ 1	2.52
	Asn43 - O δ 1	2.23
	Asp45 - O δ 1	2.41
Ca2	Val47 - O	2.21
	Asp52 - O δ 1	2.62
	Asp52 - O δ 2	2.54
	H ₂ O74 - O	2.34

VI.2.1.3 The complex interface - an alternative binding mode

The SdbA type II cohesin-dockerin interface comprises mainly one face of the cohesin (defined by strands 8-3-6-5) and helices 1 and 3 of the dockerin. The interaction surface is defined by residues Phe33, Ser34, Gly35, Tyr36, Gln37, Asn75, Thr77, Asp78, Met79, Ser80, Lys81, Asn90, Phe91, Gly92, Arg93, Leu94, Met96, Asn97, Leu98, Ser99, Arg102, Ser138, Ser139, Met140, Asn141, Asn142, Met148, Phe150, As153, Gly154, Asn155 and Met156 of the cohesin module and residues Val14, Ile15, Asn16, Met17 (from the *N-terminus* of the dockerin), Ala18, Val20, Met21, Leu23, Ala24 (from helix 1), Gln25, Phe27 (from the linker region), Ser50, Ala53, Leu56, Ala57, Tyr59 (from helix 3) and Phe60, Gly61, Lys62 and Thr63 from the *C-terminus* of the dockerin. The contacts were calculated using the PISA server (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html).^{22,23} Among these interactions there are several hydrogen bond contacts (**Figure VI.6** and **Table VI.3**) between the cohesin and the dockerin. These hydrogen bonds occur mainly between helix 1 and the cohesin. This indicates a preferential helix for the formation of the complex as in the case of the type I *C. thermocellum* complex. When compared to other Coh-Doc complexes, namely with the type I (PDB code: 1ohz¹²) and type II (PDB code: 2vt9) complexes from *C. thermocellum* we see that the position of the dockerin relative to the cohesin lays in between both structures: with respect to helix 1 of the type I dockerin, the type II is rotated by about 40° while the present structure is rotated only by about 25°. As a consequence, this helix (that in this complex is helix 3) forms fewer contacts with the cohesin than in the type II complex but a few more than in the type I. Moreover, the residues of helix 1 that make direct and water-mediated hydrogen bonds with the cohesin are conserved in the internal sequence duplication of the dockerin (with the exception of Met17 whose symmetry related residue is Ser50). Altogether, these facts are indicative of a possible dual binding mode in this complex and justify why the position of the dockerin relative to the cohesin in this complex is rotated by 180° when compared to other Coh-Doc complexes (**Figure**

VI.7). The observation of this alternative binding mode provides significant clues concerning the overall assembly and architecture of the cellulosome of *B. cellulolyticus*.

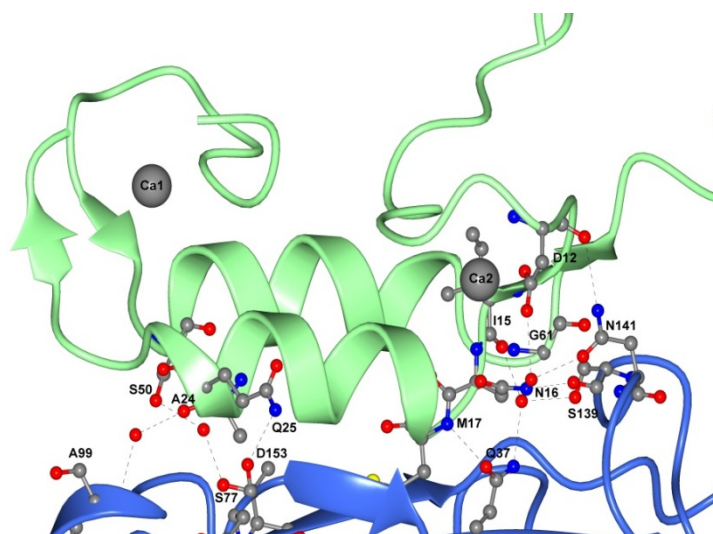


Figure VI.6: The Coh-Doc interface hydrogen bonds in the type II ScaA complex.

The hydrophobic interface is defined between the 8-3-6-5 face of the cohesin and helices 1 and 3 of the dockerin. The cohesin and the dockerin are represented as blue and green ribbons, respectively; interface residues are represented as ball-and-stick and depicted by heteroatom; calcium ions are represented as grey spheres and hydrogen bonds are represented as dashed lines.

Table VI.3: Coh-Doc interface hydrogen bonds

<i>Direct hydrogen bonds</i>								
#	<i>Cohesin</i>		<i>Distance (Å)</i>			<i>Dockerin</i>		
	Residue	Atom		Residue	Atom			
1	Gln37	Oε1	2.74	Met17			N	
2	Ser139	O	2.96	Asn16			Nδ2	
3	Asn141	Nδ2	3.52	Asp12			O	
4	Asp153	O	2.82	Gln25			Nε2	
<i>Water-mediated hydrogen bonds</i>								
#	<i>Cohesin</i>		<i>Distance (Å)</i>	<i>H₂O</i>		<i>Distance (Å)</i>	<i>Dockerin</i>	
	Residue	Atom		Residue	Atom		Residue	Atom
1	Gln37	Nε2	3.09	H ₂ O11	O	3.17	Gly61	N
2	Gln37	Nε2	3.09	H ₂ O11	O	2.70	Ile15	O
3	Ser99	N	3.58	H ₂ O50	O	2.80	Ala24	O
4	Asn141	Oδ1	2.73	H ₂ O100	O	2.71	Asn16	Nδ2
5	Asn141	Oδ1	2.73	H ₂ O100	O	3.11	Asn12	Oδ2
6	Thr77	Oγ1	2.78	H ₂ O239	O	2.73	Ser50	Oγ

V.2.1.3.1 Plasticity in the type II Coh-Doc complex

The dockerin in the structure of the type II Coh₁₁-Doc complex of the primary scaffoldin (ScaA) of *B. cellulosolvens* is bound to the cohesin in a symmetry-related manner when compared to other Coh-Doc modules¹², with helices 1 and 3 are rotated 180° with respect to each other and overlapping almost perfectly (**Figure VI.7**). The low rmsd value between both helices (0.62 Å for 24 Cα atoms) reflects this internal symmetry. Moreover, when comparing the native Coh-Doc with a structure where the Doc module was rotated 180° (the structure was built from the native structure by superposing helix 3 to helix 1), we see that all direct hydrogen bonds are maintained (**Table VI.4**) and no significant clashes are found. The contacts were calculated using the PISA server^{22,23} (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html)^{22,23} and the clash analysis was performed with Molprobit server²⁴ (<http://molprobit.biochem.duke.edu/>). This means that, in principle, similar to the type I complex in *C. thermocellum*^{12,13}, both halves of the dockerin can interact with the cohesin.

Given that these two different binding modes have been verified for some complexes¹³ and not for others (for instance the type I and type II complexes of *C. thermocellum*, respectively) it is likely that the mode a dockerin binds to the cohesin depends on the particular Coh-Doc pair. For instance, the type I complex of *C. thermocellum* is involved in assembly of the different enzymes in the scaffoldin subunit, thus it makes sense that a certain degree of flexibility is necessary for avoiding overlapping of enzymes and for maximizing the plant cell wall degradation by being able to fine tune the cellulolytic properties of a given cellulosome. Furthermore, the efficiency of the cellulosome function may require the switching of the enzyme subunits to optimize the synergy between specific enzymes. On the other hand, the type II complex of the same organism is thought not to display a dual binding mode²³ (*see Chapter V*). This may indicate that flexibility is required for binding of the catalytic subunits but is selected against in the anchorage of cellulosomes to the cell surface.⁵ Regarding the present complex, although it belongs to the type II, it behaves as a type I, thus it makes sense that it is also capable of displaying a dual binding mode.

In the light of these results, we can postulate that the dual binding mode shown by some Coh-Doc pairs is not dependent on their type but on their function; complexes involved in cell surface attachment won't display a dual binding mode whereas complexes involved in enzyme assembly will. However, whether this is a common feature of all Coh-Doc pairs requires further investigation.

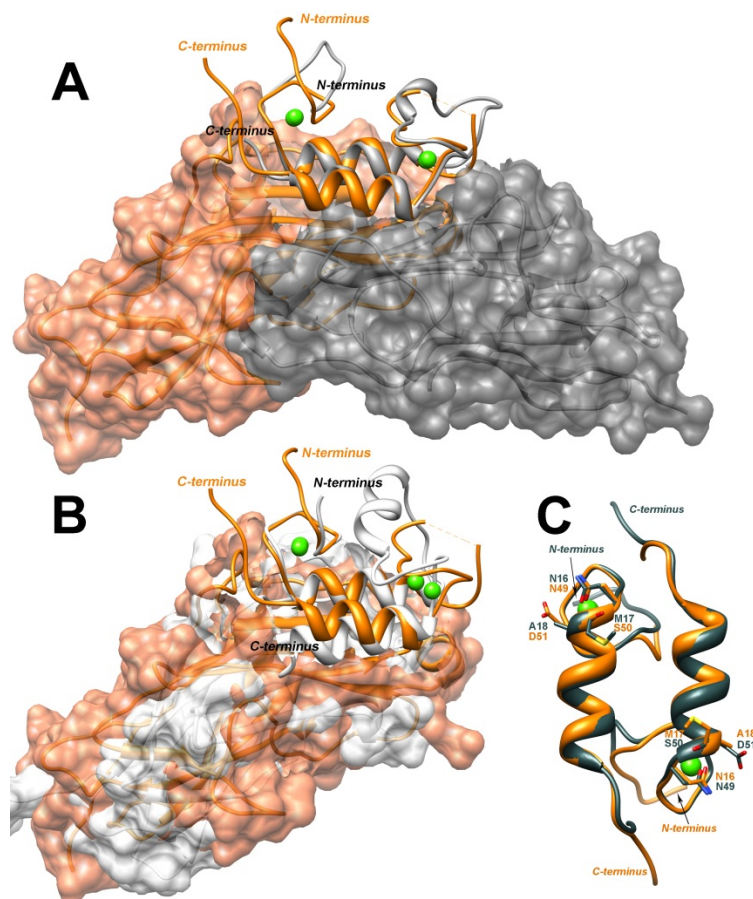


Figure VI.7: Alternative binding mode in the *B. cellulosolvens* Coh-Doc complex and internal symmetry of the dockerin.

Superposition of the type II Coh-Doc complex of *B. cellulosolvens* (orange) with: **A**) the type II Coh-Doc (grey; PDB code: 2vt9) and **B**) the rotated type I complexes from *C. thermocellum* (white; PDB code: 2cc1¹³). **C**) Superposition of the dockerin of *B. cellulosolvens* (orange) with its symmetry-related image (dark green). The cohesin modules are represented as surface, the dockerin modules are represented as ribbons, atoms are represented as sticks and the calcium ions are represented as green spheres.

Table VI.4: Coh-Doc interface hydrogen bonds in the 180°-rotated complex

<i>Direct hydrogen bonds</i>					
#	<i>Cohesin</i>		<i>Distance (Å)</i>	<i>Dockerin</i>	
	Residue	Atom		Residue	Atom
1	Gln37	Oε1	2.71	Ser50	N
2	Ser139	O	2.79	Asn49	Nδ2
3	Asn141	Nδ2	3.08	Asp45	O
4	Gly54	O	3.19	Gln58	Nε2

VI.3 Conclusions

I have solved the crystal structure of the 11th ScaA type II cohesin-dockerin (Coh₁₁-Doc) complex from *B. cellulosolvens* (**Figure VI.1**) to a resolution of 1.90 Å. At the time of writing, this is the first cohesin-dockerin complex ever determined from *B. cellulosolvens*. Also for the first time, it reveals the 3D structure of a type II dockerin of this organism and, more important, it indicates the possibility of an alternate binding mode between the cohesin and the dockerin, in a similar way to that proposed for the type I interaction in *C. thermocellum*.

The cohesin domain of the type II Coh₁₁-Doc complex of *B. cellulosolvens* shows the typical flattened, elongated β -barrel jelly-roll topology (**Figure VI.4**) and the α -helical crowning and two β -flap regions observed for other type II cohesin modules. Comparison of this structure with the unbound ScaA type II Coh₁₁ (PDB code: 1tyj¹¹) shows that the cohesin does not undergo significant conformational changes upon binding.

The structure of the dockerin domain (**Figure VI.5 - A**) reveals the classic EF-hand motifs²¹, separated by a 12-residue unstructured linker and it is very similar to the type I dockerins from *C. thermocellum*. The main differences are in helix 2 that has a high degree of disorder in this complex which impaired building of several residues. Since the type II complex of *B. cellulosolvens* lacks the X module, it is possible that other modules of the enzyme are required for stabilizing the complex. In order to further investigate this possibility, the solution of more structures is essential. As in other dockerin domains, there is an internal two-fold symmetry between helix 1 and 3. This internal symmetry is shown by the low rmsd values between both helices. Most remarkable is the fact that in this complex the dockerin is rotated 180° when compared to other native cohesin-dockerin complexes determined so far^{8,9,12}. This represents the first native complex in which the predicted dual binding mode^{13,15} is verified. This feature confers a large degree of plasticity to the complex and has profound implications at the level of the current understanding of cellulosome architecture and assembly.

VI.4 Materials and methods

VI.4.1 Molecular biology

VI.4.1.1 Transformation, expression, purification and quantification

The ScaA type II Coh-Doc complex of *B. cellulosolvens* was produced by first transforming the BcpET21a_Coh11-DocCel48 expression vector into competent *E. coli* BL21 cells

(Novagen). Expression, purification and quantification of the complex were performed as explained in Chapter V.

All the molecular biology work was done by Professor Carlos Fontes' group at the Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa and the protein kindly provided to us.

VI.4.2 X-ray crystallography

VI.4.2.1 Protein crystallization and data collection

The Type II complex Coh-Doc of *B. cellulosolvens* was crystallized at 293K by the hanging drop vapor diffusion method and obtained by mixing an equal volume (1 μ L) of protein (50 mg/ml in water) and reservoir solution (30% (m/v) polyethyleneglycol (PEG) 2000, 0.2 M ammonium sulfate, 0.1 M sodium acetate tri-hydrate, pH 4.6). Single crystals were harvested in a solution containing 35% (m/v) PEG 2000 and 0.2 ammonium sulfate, and flash-frozen in a liquid nitrogen stream at 100K, using 30% (vol/vol) of glycerol as a cryoprotectant.

The data were collected at wavelength 0.934 Å in the European Synchrotron Radiation Facility (ESRF), ID14-EH1 (Grenoble, France) to 1.90 Å resolution at 100 K. Diffraction data were processed and scaled, respectively, with programs MOSFLM²⁵ and SCALA²⁶ from the CCP4 suite¹⁶. The Matthews coefficient of the ScaA type II Coh-Doc crystal is 1.91 Å³ Da⁻¹ for two heterodimer in the asymmetric unit, with a solvent content of 35.67%. The space group was determined to be $P12_1$ with unit cell dimensions: $a = 43.4$ Å, $b = 116.1$ Å, $c = 45.2$ Å, with $\beta = 112.45^\circ$ (Table VI.1).

VI.4.2.2 Phasing, model building and refinement

Considering the calculated Matthews coefficient and because there was no dockerin structure available from the cellulosome of *B. cellulosolvens* molecular replacement attempts were performed searching for just two copies of the cohesin module in the monoclinic $P12_1$ cell (see Appendix B, Section B.2.3). The previously described and available crystal structure of the ScaA type II cohesin module from *B. cellulosolvens*, with accession code 1tyj⁷, was used as a search model for molecular replacement (see Chapter VIII, Section VIII.4.2.1). The Patterson search was done with program PHASER²⁹, implemented in the CCP4 interface²⁸, and a clear solution with two cohesins in the asymmetric unit was found in space group $P12_1$.

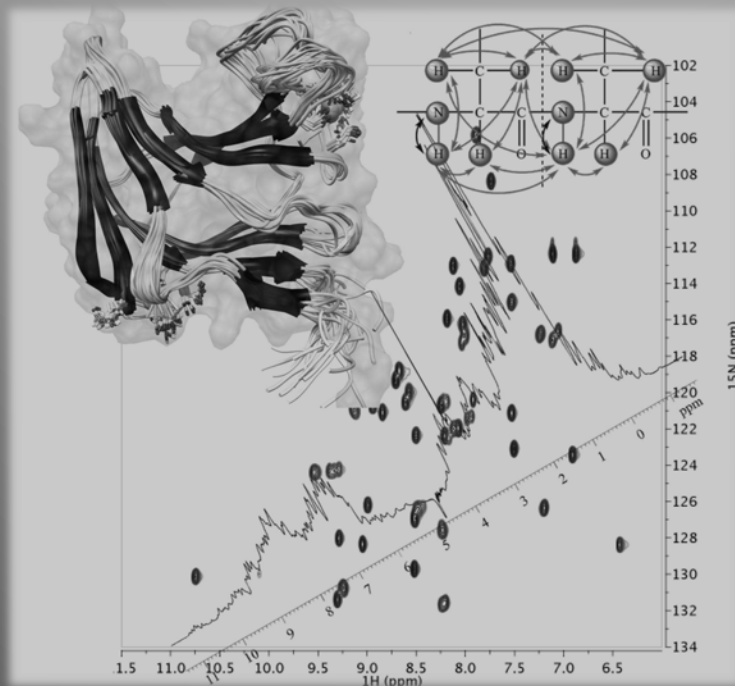
Initial building of the structures into the electron density as well as building of the dockerin modules was performed using the software ARP/wARP^{28,30} and any remaining residues were built interactively using program COOT³¹. Model refinement and electron density map

calculations were done with program REFMAC5³² from the CCP4 suite²⁸. The final model has $R_{cryst} = 16.3\%$ and $R_{free} = 22.5\%$ and includes 299 water molecules and four calcium ions. Due to disorder, residues Met1, Ala2, and the 6 C-terminal histidines of chain A (cohesin), Gly32-Asn37 and Ala66-Phe71 of chain B (dockerin), Met1, Ala2, Leu174, Glu175 and the 6 C-terminal histidines of chain C (cohesin) and Ala30-Asn44 and Ser65-Phe71 of chain D (dockerin) could not be built. The structure is deposited in the Protein Data Bank under the accession code: 2y3n

VI.5 References

1. Giuliano, C.; Khan, A. W., Cellulase and Sugar Formation by *Bacteroides cellulosolvens*, a Newly Isolated Cellulolytic Anaerobe. *Appl Environ Microbiol* **1984**, *48* (2), 446.
2. Giuliano, C.; Khan, A. W., Conversion of cellulose to sugars by resting cells of a mesophilic anaerobe, *Bacteroides cellulosolvens*. *Biotechnol Bioeng* **1985**, *27* (7), 980.
3. Ding, S. Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R., A scaffoldin of the *Bacteroides cellulosolvens* cellulosome that contains 11 type II cohesins. *J Bacteriol* **2000**, *182* (17), 4915.
4. Beguin, P.; Lemaire, M., The cellulosome: An exocellular, multiprotein complex specialized in cellulose degradation. *Crit Rev Biochem Mol* **1996**, *31* (3), 201.
5. Fontes, C. M. G. A.; Gilbert, H. J., Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry*, Vol 79 **2010**, *79*, 655.
6. Xu, Q.; Bayer, E. A.; Goldman, M.; Kenig, R.; Shoham, Y.; Lamed, R., Architecture of the *Bacteroides cellulosolvens* cellulosome: Description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *J Bacteriol* **2004**, *186* (4), 968.
7. Noach, I.; Frolow, F.; Jakoby, H.; Rosenheck, S.; Shimon, L. J. W.; Lamed, R.; Bayer, E. A., Crystal structure of a type-II cohesin module from the *Bacteroides cellulosolvens* cellulosome reveals novel and distinctive secondary structural elements. *Journal of Molecular Biology* **2005**, *348* (1), 1.
8. Adams, J. J.; Currie, M. A.; Ali, S.; Bayer, E. A.; Jia, Z. C.; Smith, S. P., Insights into Higher-Order Organization of the Cellulosome Revealed by a Dissect-and-Build Approach: Crystal Structure of Interacting *Clostridium thermocellum* Multimodular Components. *Journal of Molecular Biology* **2010**, *396* (4), 833.
9. Adams, J. J.; Pal, G.; Jia, Z. C.; Smith, S. P., Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. *P Natl Acad Sci USA* **2006**, *103* (2), 305.
10. Mechaly, A.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Lamed, R.; Shoham, Y.; Bayer, E. A., Cohesin-dockerin interaction in cellulosome assembly - A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *Journal of Biological Chemistry* **2001**, *276* (13), 9883.
11. Adams, J. J.; Webb, B. A.; Spencer, H. L.; Smith, S. P., Structural characterization of type II dockerin module from the cellulosome of *Clostridium thermocellum*: Calcium-induced effects on conformation and target recognition. *Biochemistry* **2005**, *44* (6), 2173.
12. Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A., Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *P Natl Acad Sci USA* **2003**, *100* (24), 13809.

13. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A.; Gilbert, H. J., Evidence for a dual binding mode of dockerin modules to cohesins. *P Natl Acad Sci USA* **2007**, *104* (9), 3089.
14. Carvalho, A. L.; Pires, V. M. R.; Gloster, T. M.; Turkenburg, J. P.; Prates, J. A. M.; Ferreira, L. M. A.; Romao, M. J.; Davies, G. J.; Fontes, C. M. G. A.; Gilbert, H. J., Insights into the structural determinants of cohesin dockerin specificity revealed by the crystal structure of the type II cohesin from *Clostridium thermocellum* SdbA. *Journal of Molecular Biology* **2005**, *349* (5), 909.
15. Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A., Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **2008**, *3* (7), 1171.
16. Bailey, S., The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallogr D* **1994**, *50*, 760.
17. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D* **1997**, *53*, 240.
18. Shimon, L. J. W.; Bayer, E. A.; Morag, E.; Lamed, R.; Yaron, S.; Shoham, Y.; Frolow, F., A cohesin domain from *Clostridium thermocellum*: The crystal structure provides new insights into cellulosome assembly. *Structure* **1997**, *5* (3), 381.
19. Tavares, G. A.; Beguin, P.; Alzari, P. M., The crystal structure of a type I cohesin domain at 1.7 angstrom resolution. *Journal of Molecular Biology* **1997**, *273* (3), 701.
20. Spinelli, S.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Henrissat, B.; Cambillau, C., Crystal structure of a cohesin module from *Clostridium cellulolyticum*: Implications for dockerin recognition. *Journal of Molecular Biology* **2000**, *304* (2), 189.
21. Pages, S.; Belaich, A.; Belaich, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A., Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins* **1997**, *29* (4), 517.
22. Krissinel, E.; Henrick, K. Protein interfaces, surfaces and assemblies service PISA at European Bioinformatics Institute. http://www.ebi.ac.uk/pdbe/prot_int/pistart.html.
23. Krissinel, E.; Henrick, K., Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **2007**, *372* (3), 774.
24. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D* **2010**, *66*, 12.
25. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletters on Protein Crystallography* **1992**, *26*.
26. Evans, P. R., Scaling of MAD data. In *Proceedings of the CCP4 Study Weekend. Recent advances in phasing*, Winn, M., Ed. 1997; Vol. 33, pp 22.
27. McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr D* **2005**, *61*, 458.
28. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **2004**, *60*, 2126.



Chapter VII: Protein NMR Spectroscopy

In this chapter I describe some fundamental principles and concepts of NMR spectroscopy applied to the determination of 3D structures of proteins and to the study of protein/ligand interactions. The section dedicated to the saturation-transfer difference experiment is part of a published paper (Viegas et al, 2011)¹ and the section dedicated to the diffusion ordered spectroscopy is part of a published book chapter (Viegas et al, 2010)² and from a manuscript in preparation.

Table of Contents

Summary	204
VII.1 Introduction.....	204
VII.2 Protein NMR.....	207
VII.2.1 Chemical Shift.....	207
VII.2.1.1 Spin-spin coupling and spin systems.....	208
VII.2.2 Relaxation.....	212
VII.2.2.1 The Bloch equations.....	213
VII.2.2.2 T_1 relaxation	214
VII.2.2.3 T_2 relaxation	215
VII.2.2.4 Dipole-dipole relaxation.....	217
VII.2.2.5 Chemical shift anisotropy relaxation.....	218
VII.2.3 The protein's fingerprint – ^{15}N - ^1H -HSQC.....	219
VII.2.4 Nuclear Overhauser effect.....	221
VII.3 Protein structure determination	229
VII.3.1 Three-dimensional experiments	231
VII.3.1.1 Experiments for backbone assignments	232
VII.3.1.2 Experiments for side-chain assignments	240
VII.3.1.3 Experiments for NOE measurement.....	242
VII.3.2 Structure validation	244
VII.4 Protein dynamics by NMR.....	245
VII.4.1 Theory of spin relaxation in proteins	246
VII.4.2 Protein motions and relaxation.....	248
VII.4.2.1 Reduced spectral density mapping.....	248
VII.4.2.2 Rotational diffusion tensor	249
VII.4.2.3 The Lipari-Szabo Model-free Formalism.....	251
VII.4.2.4 Amide proton exchange	254

VII.5	Study of protein-ligand complexes	256
VII.5.1	Saturation transfer difference	256
VII.5.2	Diffusion ordered spectroscopy.....	261
VII.6	References	265

Summary

In this chapter I describe some fundamental principles and concepts of NMR spectroscopy applied to the determination of 3D structures of proteins and to the study of protein/ligand interactions. I start by giving a general introduction to some aspects of protein NMR spectroscopy (*Section VII.2*), fundamental for a comprehensive interpretation of the data like chemical shift, spin systems, coupling constants and relaxation. In the same section I also introduce the ^{15}N - ^1H -HSQC spectrum as the protein's fingerprint and explain some theoretical aspects of the nuclear Overhauser effect (NOE).

Section VII.3 is dedicated to the NMR techniques I used to determine the solution structure of CtCBM11 (*Chapter II*). In this section I start by explaining in some detail the experiments used and then I show how they can be applied for determining a 3D solution structure.

In Section VII.4 I will focus on molecular motions of proteins (as they are not static entities) and on the importance of these motions (that occur in different time scales) for the interpretation of structural data and binding studies. I start by giving a general explanation of the concepts behind protein dynamics I then explain how this useful information can be extracted from NMR data.

Finally, Section VII.5 covers the techniques I used to study the interaction of the several CBMs with target ligands, namely saturation-transfer difference¹ (STD) and diffusion ordered spectroscopy² (DOSY).

VII.1 Introduction

Over the last 60 years the field of nuclear magnetic resonance (NMR) spectroscopy, explicitly macromolecular NMR, has experienced an explosive growth and has emerged as one of the main techniques of structural biology³⁻⁵ (**Figure VII.1**). Instrumental improvements in recent years have contributed significantly to this development. Digital recording, cryogenic probes, auto-samplers, and higher magnetic fields shorten the time for data acquisition and improve the spectral quality. In addition, new experiments and pulse sequences⁶⁻¹⁵ make a vast amount of information available for the use of NMR for the characterization of **structure** and **dynamics** of biological molecules in solution and in the **drug discovery** process (**Table VII.1**¹⁶). From the initial observation of proton magnetic resonance in water¹⁷ and in paraffin¹⁸, NMR has evolved to become one of the leading analytical methods available. Although macromolecular NMR has been always limited by the molecular weight of the proteins (20-40 KDa), the recent advances mentioned above allied to new recombinant protein expression

protocols (^{15}N , ^{13}C , ^2H and selective methyl labeling)¹⁹⁻²³ allow the study of large complexes, thereby extending the molecular weight limit of systems that can be studied up to **100 kDa**²⁴.

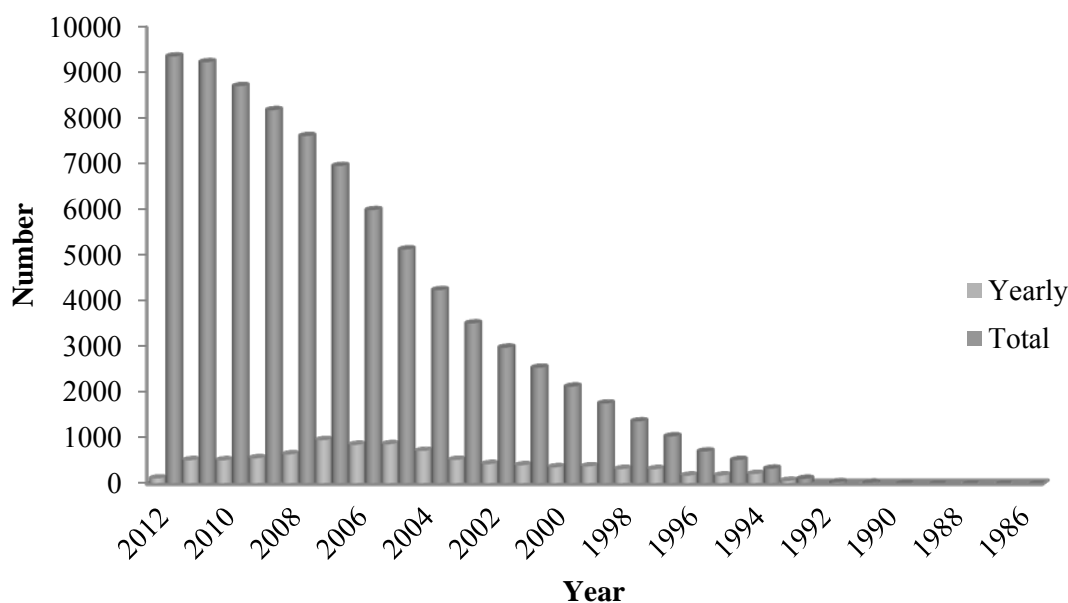


Figure VII.1: Yearly and annual growth of structures solved by NMR.

Data was taken from the Protein Data Bank (<http://www.pdb.org/pdb>)

Together with X-ray crystallography, NMR spectroscopy is one of the techniques that can provide high-resolution structures of biomolecules and both techniques can be used in conjunction.²⁵⁻²⁷ NMR spectroscopy can be used for investigating time-dependent chemical phenomena that provide information about **conformational dynamics**²⁸⁻³³, **exchange processes**^{34,35} and **kinetics**^{36,37} of biomolecules at timescales ranging from picoseconds to seconds, and is very efficient in determining **ligand binding**^{1,25,38-41} and mapping **interaction surfaces** of protein/ligand complexes^{25,42-45}. It allows the visualization of single atoms and molecules in various media in solution as well as in solid state and it is nondestructive, giving molar responses that allow **structure elucidation** under near physiological conditions or membrane mimetic environments⁴⁶⁻⁴⁹ and **quantification** simultaneously⁵⁰. Since crystals are not needed, **protein folding** studies can be done by monitoring NMR spectra upon folding or under denaturing conditions in real time⁵¹⁻⁵³, making this method one of the most powerful for these studies. By exploring the fact that upon complex formation between a target molecule and a ligand, significant perturbations can be observed in NMR sensitive parameters of both target and ligand, NMR spectroscopy has become an essential tool in the pharmaceutical industry^{1,2,54-59}. These perturbations can be used qualitatively to detect ligand binding and screen for novel compounds during the process of lead generation or quantitatively to assess the strength of the binding interaction and provide structural information useful for lead optimization during the

latter stages of a drug discovery program.^{1,2,54-60} For all of the above, NMR has become a sophisticated and powerful analytical technology, with a large variety of applications in many disciplines of scientific research, medicine, and various industries.

Table VII.1: A summary of some key developments that have had a major influence on the practice and application of high-resolution NMR spectroscopy in chemical research.¹⁶

<i>Decade</i>	<i>Notable advances</i>
1940s	First observation of NMR in solids and liquids (1945)
1950s	Development of chemical shifts and spin–spin coupling constants as structural tools
1960s	Use of signal averaging for improving sensitivity Application of the pulse-FT approach The NOE employed in structural investigations
1970s	Use of superconducting magnets and their combination with the FT approach Computer controlled instrumentation 2D NMR
1980s	Development of multipulse NMR First solution structure of a small protein – BPTI – from NOE restraints (1985) Automated spectroscopy 3D NMR + ¹³ C and ¹⁵ N isotope labeling of recombinant proteins (resolution)
1990s	Routine application of pulsed field gradients for signal selection Development of coupled analytical methods, e.g. LC-NMR New long range structural parameters: Residual dipolar couplings (RDCs) TROSY (molecular weight up to 100 kDa)
2000–	Use of high-sensitivity cryogenic probes Routine availability of actively shielded magnets for reduced stray fields Development of microscale tube and flow probes
2010+	Adoption of fast and parallel data acquisition methods

FT, Fourier transformation; LC-NMR, liquid chromatography nuclear magnetic resonance.

VII.2 Protein NMR

VII.2.1 Chemical Shift

Chemical shifts communicate in a very simple way detailed molecular information that almost any chemist can understand. They have long been used as tools for structural analysis, giving information on several parameters such as non-covalent structure, solvent interactions, ionization constants, ring orientations, hydrogen bond interactions, among other.⁶¹⁻⁶⁴

In structural biology, chemical shifts are most often used to predict regions of secondary structure, to refine complex structures or to characterize binding.^{25,61-66} The NMR spectra of proteins provide unique fingerprints (*see Section VII.2.3*) suggesting that chemical shifts carry enough information to determine their structures at high resolution.⁶²⁻⁶⁴ In fact, due to the increasing number of NMR structures deposited it is now possible to calculate the probability of amino acid types⁶⁷ from a set of chemical shift values through the use of the BioMagResBank⁶⁸ (BMRB), a databank of chemical shifts from assigned proteins (<http://www.bmrwisc.edu/>). The chemical shifts of certain atomic nuclei in proteins ($^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and ^{13}CO) are dependent on whether or not the amino acid residue is part of a secondary structure (α -helix, β -sheet), and if so, whether it is helix or sheet.^{66,69,70} **Table VII.2** shows the random coil chemical shifts for common amino acids.⁶³ By calculating the difference between the random coil chemical shift and the observed one it is possible to predict the secondary structure of proteins. If the obtained difference is greater than 0 it is given the value 1; otherwise it is given the value -1. The secondary structure is established following this designation. Thus:

- Alpha helix is defined when four or more "-1" $\text{H}\alpha$ and/or "1" $\text{C}\alpha/\text{CO}$ are sequentially found.
- A beta-strand is defined when three or more "1" $\text{H}\alpha$ and/or "-1" $\text{C}\alpha/\text{CO}$ are sequentially found.
- All other regions are designated as coil.

Table VII.2: Random coil chemical shifts for common amino acids.^{63,68}

<i>Amino acid</i>	^1HN	^{15}N	$^1\text{H}\alpha$	$^{13}\text{C}\alpha$	$^1\text{H}\beta$	$^{13}\text{C}\beta$	^{13}CO
Ala	8.20	123.2	4.26	53.1	1.35	19.0	177.7
Cys(r)	8.39	120.1	4.66	58.2	2.95/2.89	32.6	174.9
Cys(o)	8.43	118.6	4.71	55.4	3.25/2.99	41.1	174.6
Asp	8.31	120.7	4.59	54.7	2.72/2.66	40.9	176.4
Glu	8.33	120.7	4.25	57.3	2.02/1.99	30.0	176.9

Phe	8.36	120.5	4.63	58.1	3.00/2.94	40.0	175.4
Gly	8.33	109.7	3.97/3.90	45.4	—	—	173.9
His	8.25	119.7	4.61	55.0	3.10/3.04	29.0	175.2
				56.3 (pH 9)		30.8 (pH 9)	
Ile	8.28	121.5	4.18	61.6	1.78	38.6	175.8
Lys	8.19	121.1	4.27	57.0	1.78/1.74	32.8	176.6
Leu	8.23	121.9	4.32	55.6	1.62/1.52	42.3	177.0
Met	8.26	119.6	4.41	56.1	2.03/1.99	33.0	176.2
Asn	8.34	120.1	4.67	53.5	2.81/2.75	38.7	175.3
Pro	—	134.0	4.40	63.3 (trans)	2.08/2.00	31.9 (trans)	176.7
				62.8 (cis)		34.5 (cis)	
Gln	8.22	119.9	4.27	56.6	2.05/2.01	29.2	176.3
Arg	8.24	120.8	4.30	56.8	1.79/1.76	30.7	176.4
Ser	8.28	116.3	4.48	58.7	3.88/3.85	63.8	174.6
Thr	8.24	115.4	4.46	62.2	4.17	69.7	174.5
Val	8.29	121.1	4.19	62.5	1.98	32.7	175.6
Trp	8.29	121.7	4.68	57.7	3.19/3.12	30.0	176.1
Tyr	8.32	120.5	4.63	58.1	2.91/2.84	39.3	175.4

Note that Cys(r) refers to cysteine and Cys(o) refers to cystine.

VII.2.1.1 Spin-spin coupling and spin systems

The chemical shift is not the only indicator used to structurally characterize a molecule. Nuclei themselves possess a small magnetic field, that affect each other, changing the energy and hence frequency of nearby nuclei as they resonate - **spin-spin coupling** or **scalar coupling**⁷¹. This interaction between two nuclei occurs through chemical bonds, and can typically be seen up to three bonds (**Table VII.3**).⁷² The strength of the interaction is measured by the scalar coupling constant, ${}^nJ_{IS}$, in which n is the number of covalent bonds between the nuclei I and S and its magnitude is given in *Hz*. Depending on whether the low energy state is favored or not, J can assume either positive or negative values. The low energy state is that in which the magnetic moments of nuclei I and S are in antiparallel configuration to the magnetic moments of their bonding electrons.⁷³ As we will see below (*Section VII.3.1*), these scalar couplings allow the transfer of magnetization between the several nuclei present in amino acids (${}^1\text{H}$, ${}^{13}\text{C}$ and ${}^{15}\text{N}$) and are the basis for all the experiments required for the complete assignment of protein resonances.

Table VII.3: Typical spin coupling constants in amino acids.^{72,74}

Spin coupling	Typical J value (Hz)
$^1J_{C-H}$	140
$^1J_{N-H^N}$	92
$^1J_{C_\alpha-CO}$	55
$^1J_{C_\alpha-C_\beta}$	35
$^1J_{C_\beta-C_\gamma}$	35
$^1J_{N-CO_{i-1}}$	15
$^2J_{N-CO}$	< 1
$^1J_{N-C_\alpha}$	11
$^2J_{N-C_{\alpha_{i-1}}}$	7
$^2J_{N-H_\alpha}$	19
$^1J_{CO-H_\alpha}$	4-7
$^3J_{H^N-C_{\alpha_{i-1}}}$	5.5

If a group of spins are connected to each other by scalar spin-spin couplings, they are said to belong to the same **spin system** (Figure VII.2 and Figure VII.3). In a simple way, if a set of nuclei are coupled with a large chemical shift separation, $\Delta\nu$ (weak coupling, $\Delta\nu \gg J$), the spin system is said to be an **AX** or **AMX** system. In contrast, when the frequencies of the coupling nuclei are on the same order of magnitude as J coupling (strong coupling, $\Delta\nu \approx J$), nuclei are labeled with adjacent letters of the alphabet (**AB**, **ABC** or **XYZ**). If groups of nuclei are magnetically equivalent, they are labeled **A_nB_n**, etc, where n is the number of equivalent nuclei (for instance, CH₃ groups are A₃, or X₃). A group of magnetically equivalent nuclei must have identical chemical shifts, and all members of the group must be coupled equally to nuclei outside the group. If nuclei are chemically equivalent but not magnetically equivalent, then they are labeled **AA'**, **BB'B''** or **XX'**. These relationships are very useful when assigning the resonances of a protein as the patterns formed by the several spin systems of the different amino acids can be easily identified and by themselves, allow unambiguous recognition of some residues such as glycine, alanine, threonine, valine, isoleucine and leucine (Figure VII.2 and Figure VII.3).

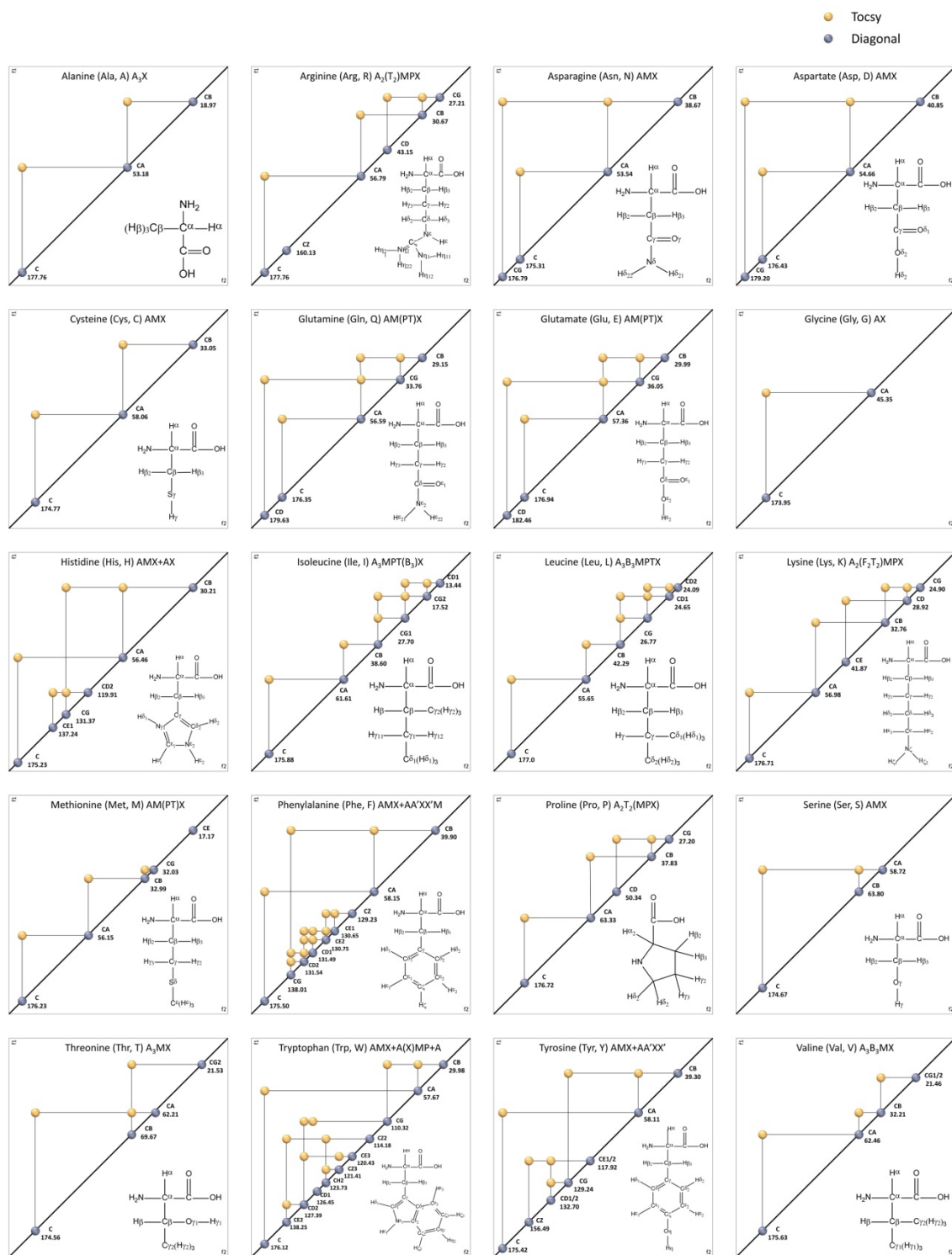


Figure VII.2: ^{13}C - ^{13}C TOCSY pattern of the 20 standard amino acids.

The chemical shift values for the different protons are an average value calculated from the full BMRB database. The calculated statistics are derived from a total of 5129743 chemical shifts.

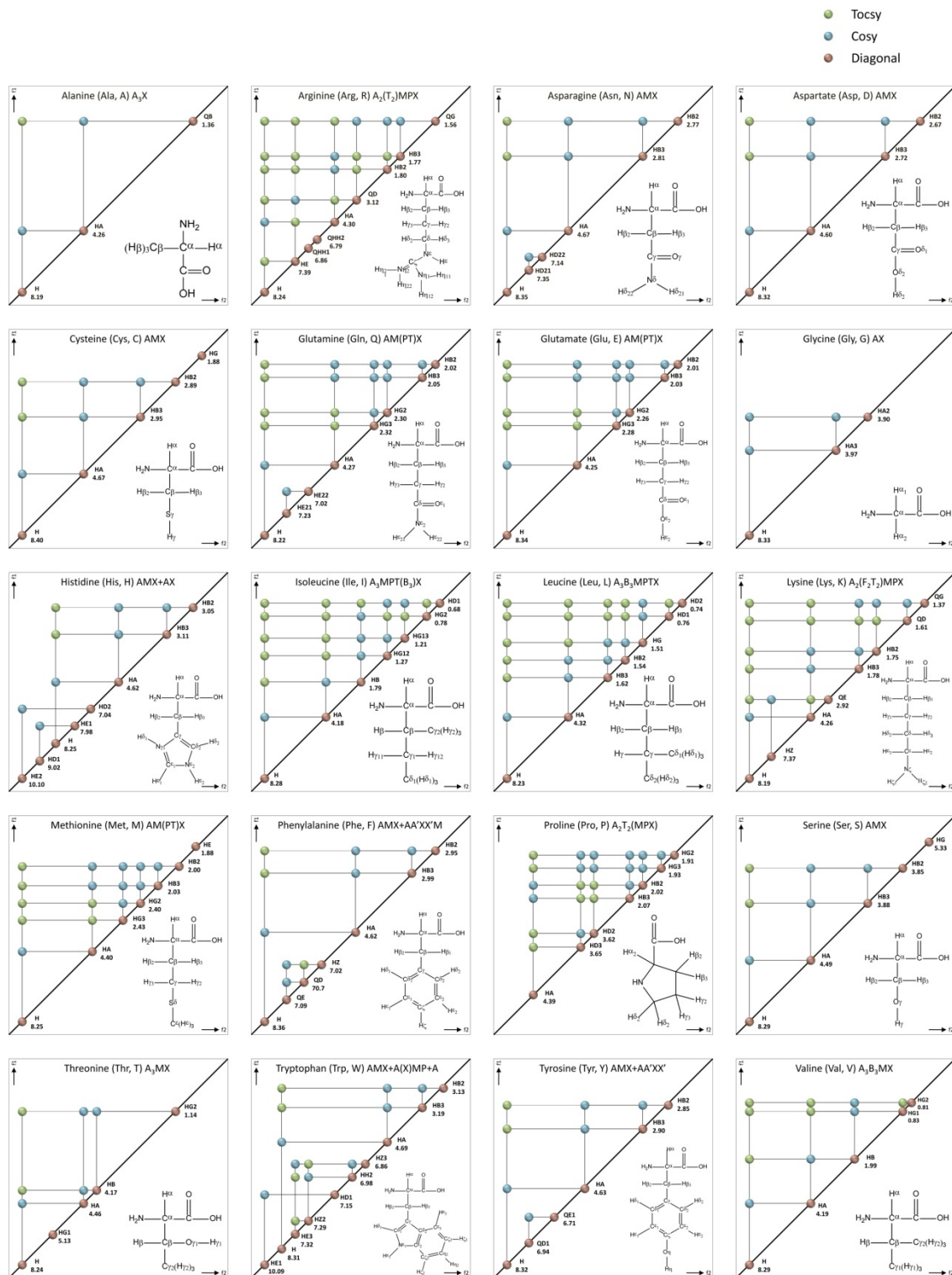


Figure VII.3: ¹H-¹H TOCSY and COSY pattern of the 20 standard amino acids.

The chemical shift values for the different protons are an average value calculated from the full BMRB database. The calculated statistics are derived from a total of 5129743 chemical shifts.

Furthermore, the vicinal scalar coupling constant between the H-H separated by three-bond (${}^3J_{HH}$) has a relationship with the relative orientation of the coupled protons that provides geometrical information between atoms in a molecule.

Using the *Karplus* equation (**Equation VII.1**) it is possible to determine the **dihedral torsion angles**.⁷³

$$J(\theta) = A\cos^2\theta + B\cos\theta + C$$

VII.1

where J is the 3J coupling constant, A, B, and C are constants that depend on the specific coupled nuclei and θ is the dihedral angle. By studying the relationship of ${}^3J_{H^N H_\alpha}$ to the dihedral angle φ for the structure of ubiquitin, Wang and Bax⁷⁵ have obtained the values of constants A, B and C:

$$J(\theta) = 6.98\cos^2\theta - 1.38\cos\theta + 1.72$$

VII.2

where $\theta = \varphi - 60$. The rigid peptide dihedral angle, ω is always close to 180 degrees. The dihedral angles φ and ψ can have a certain range of possible values. These angles function as the internal degrees of freedom of a protein, and control the protein's conformation (**Figure VII.4**). They are restrained by geometry to allowed ranges typical for

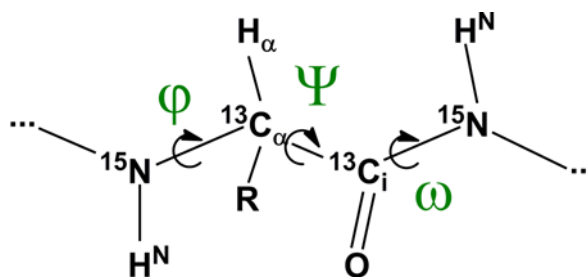


Figure VII.4: Peptide torsion angles

particular secondary structure elements. φ and ψ dihedral angles can be represented in a **Ramachandran** plot. This type of plot is a way to visualize backbone dihedral angles ψ against φ of amino acid residues in protein structure and is a way of validating the structure (*see Section VII.3.2*).

VII.2.2 Relaxation

An rf pulse applied onto a sample at thermal equilibrium causes a perturbation on the nuclear spins removing them from the rest state. After this pulse the system will try to return to the equilibrium, losing the excess energy. Nevertheless, due to the low transition energies associated with magnetic resonance, the lifetime of the excited states is extremely long (from a

few seconds to minutes). These long lifetimes are fundamental for NMR spectroscopy as they result in relative narrow lines (as a consequence of the Heisenberg Uncertainty Principle).¹⁶ Furthermore, they allow the manipulation of the spin systems permitting the acquisition of complicated pulse schemes. There are mainly two ways this can happen: either by **spin-lattice relaxation** (also known as T_1 relaxation or longitudinal relaxation) or by **spin-spin relaxation** (also known as T_2 relaxation or transverse relaxation). T_1 relaxation corresponds to the process of re-establishing the normal population distribution of α and β spin states in the magnetic field (acts along the static magnetic field direction - z) and T_2 is the loss of phase coherence among nuclei and acts on the transverse plane (x - y), perpendicular to the static magnetic field.⁷³ Since the return of magnetization to the z -direction inherently causes loss of magnetization in the x - y plane T_2 is always less than or equal to T_1 . In an NMR experiment the linewidth of a signal is determined by T_2 - short T_2 give broader lines (see Section VII.2.2.3). The maximum repetition rate during acquisition of an NMR signal is governed by T_1 - short T_1 means a spectrum can be acquired faster. Relaxation rates of nuclear spins can also be related to aspects of molecular structure and behavior such as internal molecular motions (see Section VII.4).³²

VII.2.2.1 The Bloch equations

The **Bloch equations** were introduced by Felix Bloch in 1946¹⁷ and are used to calculate the nuclear magnetization as a function of the relaxation times T_1 and T_2 . In a simple way, given a $\frac{1}{2}$ spin, the Bloch equations can be written as:

$$\frac{dM_x(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_x - \frac{M_x}{T_2} \tag{VII.3}$$

$$\frac{dM_y(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_y - \frac{M_y}{T_2} \tag{VII.4}$$

$$\frac{dM_z(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_z - \frac{M_0 - M_z}{T_1} \tag{VII.5}$$

where γ is the magnetogyric ratio, $\mathbf{M}(t)$ is the nuclear magnetization vector (with components $M_x(t)$, $M_y(t)$, and $M_z(t)$), M_0 is the equilibrium magnetization (when $t \rightarrow \infty$) and $\mathbf{B}(t)$ is the applied magnetic field (consisting of the static and rf fields).⁷¹

VII.2.2.2 T_1 relaxation

T_1 relaxation is the mechanism by which the system reestablishes the equilibrium populations. In order to measure T_1 relaxation, the most often applied experiment is the so called **inversion recovery**. In this experiment the first step is inverting the population by applying 180° pulse. The magnetization vector, initially aligned with the $-z$ axis, will recover only along the z axis as there is no x - y magnetization. The recovery is monitored by placing the vector back in the x - y plane with a 90° pulse after a suitable period, τ , following the initial inversion (**Figure VII.5**).

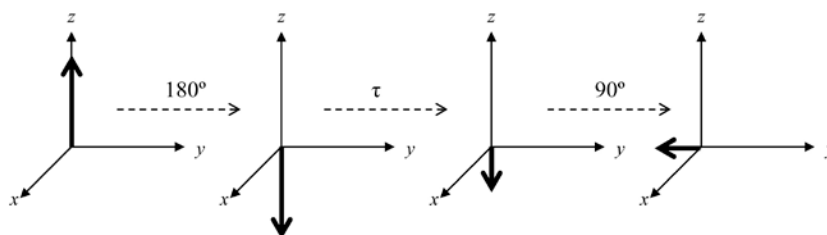


Figure VII.5: The inversion recovery process.

In these conditions, the solution of the Bloch equation for the M_z magnetization can be written as:

$$M_z(t) = M_0 \left(1 - 2e^{-\frac{t}{T_1}} \right)$$

VII.6

The relaxation time can be determined by fitting the signal intensity (measured at different times, τ) to this equation. This relaxation time is dependent on the magnetogyric ratio, γ , of the nucleus and on the mobility of the molecule. As mobility increases, the vibrational and rotational frequencies increase, making it more likely to stimulate the transition from high to low energy states. However, at extremely high mobilities, the probability decreases as the vibrational and rotational frequencies no longer correspond to the energy gap between states (**Figure VII.6**). Only frequencies that influence the population distribution (thus have a component in the z axis) will influence T_1 relaxation. **Equation VII.7** translates this behavior:

$$\frac{1}{T_1} = \gamma^2 \bar{H}^2 \frac{\tau_c}{1 + (2\pi\nu_0\tau_c)^2}$$

VII.7

where γ is the magnetogyric ratio, τ_c is the correlation time, ν_0 is the Larmor frequency and $\overline{H^2}$ is the mean-square average of the local magnetic fields.

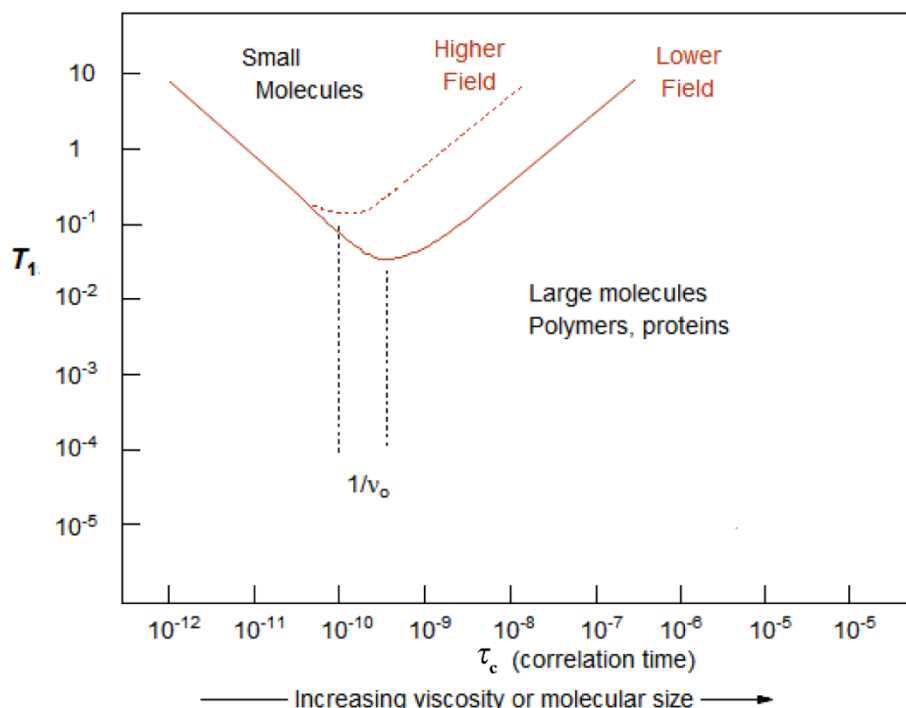


Figure VII.6: Effect of the correlation time, τ_c , in the relaxation time T_1 .

Adapted from: <http://www.chem.wisc.edu/areas/reich/nmr/08-tech-01-relax.htm>.

VII.2.2.3 T_2 relaxation

T_2 relaxation is the mechanism by which the transverse component of the magnetization, M_{xy} , exponentially decays towards the equilibrium. This happens by loss of coherence among the different spins, caused mainly by differences in the magnetic field experienced by the different nuclei. Only small differences in the magnetic field will make some spins experience a slightly greater local field while others experience smaller one resulting in the loss of magnetization on the transverse plane. These magnetic field differences arise mainly from two sources: i) inhomogeneity of the static magnetic field and ii) the local magnetic fields arising from intramolecular and intermolecular interactions in the sample.¹⁶ The first is an instrumental imperfection that can be minimized, for instance, with a good shimming; the second represents the natural transverse relaxation process. In order to get only the natural T_2 relaxation contribution a **spin-echo** sequence is often used (**Figure VII.7**). In this experiment, the first step is a 90° pulse that places the magnetization in the x - y plane. The magnetizations will then lose coherence due to field inhomogeneity during a time period, τ . The second step is to apply a 180° pulse. This will rotate the vectors towards the $-y$ axis and, after a second time period, τ (equal to

the first) the magnetization will be refocused. However, during the 2τ time period, some loss of phase coherence by natural transverse relaxation also occurs, and this is not refocused by the spin-echo since, the acting mechanisms are random. This means that at the time of the echo, the intensity of the observed magnetization will have decayed according to the natural T_2 time constant, independent of field inhomogeneity.¹⁶

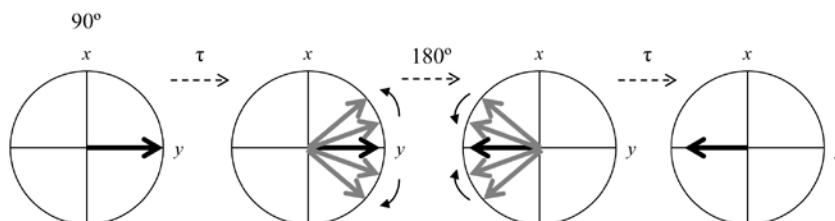


Figure VII.7: The spin-echo refocuses magnetization dephased by field inhomogeneity.

In these conditions, the solution of the Bloch equation for the M_z magnetization can be written as:

$$M_{x,y}(t) = M_0 e^{-\frac{t}{T_2}}$$

VII.8

Again, the relaxation time can be determined by fitting the signal intensity (measured at different times, τ) to this equation. Nonetheless, the determination of T_2 by this method (or by any other available) is not straightforward as homonuclear couplings are not refocused by the spin-echo and hence will impose additional phase modulations on the detected signals. Still, from the experimental point-of-view, exact T_2 values are not important and the value of T_2^* (which may be calculated from linewidths) has far greater significance:

$$\nu_{1/2} = \frac{1}{\pi T_2^*}$$

VII.9

where $\nu_{1/2}$ is the linewidth at half height and T_2^* is the combination of the natural and experimental T_2 relaxation times. T_2^* determines the rate of decay of the transverse magnetization, thus defining how long an experiment can be before the system has decayed to such an extent that there is no longer any signal left to detect.¹⁶

The return of magnetization to the z -direction inherently causes loss of magnetization in the x - y plane, making T_2 always less than or equal to T_1 . Therefore, all aspects that influence T_1 will

also indirectly influence T_2 . Moreover, all other frequencies acting on the x-y plane will also act on T_2 . **Equation VII.10** translates this behavior:

$$\frac{1}{T_2} = \gamma^2 \bar{H}^2 \frac{\tau_c}{1 + (2\pi\nu_0\tau_c)^2} + \tau_c$$

VII.10

where γ is the magnetogyric ratio, τ_c is the correlation time, ν_0 is the Larmor frequency and \bar{H}^2 is the mean-square average of the local magnetic fields. NMR resonance linewidths in solution are, generally speaking, inversely proportional to the T_2 , relaxation time, which decreases with increasing molecular size and tumbling time, τ_c (**Figure VII.8**).

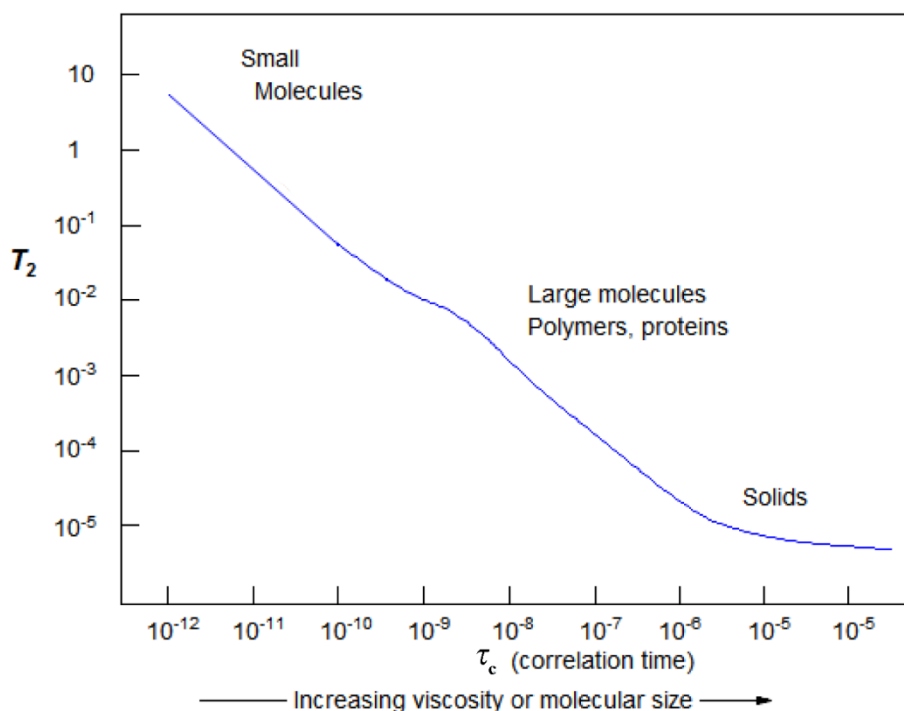


Figure VII.8: Effect of the correlation time, τ_c , in the relaxation time T_2 .

Adapted from: <http://www.chem.wisc.edu/areas/reich/nmr/08-tech-01-relax.htm>.

VII.2.2.4 Dipole-dipole relaxation

Dipole-dipole interaction is probably the most important mechanism of relaxation pathway for protons in molecules containing contiguous protons and for carbons with directly attached protons. This is also the source of the **Nuclear Overhauser Effect** (NOE) and further details of this mechanism are given in Section VII.2.4. Dipolar coupling occurs when the magnetic field generated by one nuclear dipole affects the magnetic field at another nucleus and depends

essentially on the distance between nuclei, the angular relationship between the magnetic field and the internuclear vectors and the magnetic moment of the involved spins. This type of coupling does not require connecting bonds; it takes place **through-space**.^{16,71}

This mechanism is often the dominant relaxation process for protons that rely on their neighbors as a source of magnetic dipoles. As the molecule tumbles in solution the dipole-dipole coupling is constantly changing as the vector relationships change creating a fluctuating magnetic field at each nucleus. To the extent that these fluctuations occur at the Larmor precession frequency, they can cause nuclear relaxation. As such, protons that don't have near neighbors relax more slowly all have longer T_1 times than more crowded groups. If T_1 data are available, then protons with long relaxation times can be predicted to be remote from others in the molecule. Since the proton has the highest magnetic dipole of common nuclei, it is the most effective nucleus for causing dipole-dipole relaxation.¹⁶

Besides the internuclear distance, dipole-dipole relaxation also depends on the correlation time, τ_c , of the molecules such that, for small molecules tumbling very fast (short τ_c), the dipole-dipole relaxation is not very efficient, thus, the longer T_1 times will be; large molecules (e.g. proteins) are usually moving too slowly (τ_c is too long), and they have the opposite relationship between molecular motion and T_1 (i.e., the faster the molecule tumbles, the more effective the relaxation).⁷⁶

VII.2.2.5 Chemical shift anisotropy relaxation

Because the electron distribution in chemical bonds is asymmetric or anisotropic, the local field experienced by a nucleus (therefore its chemical shift) will depend on the relative orientation of the bond to the applied static field. This is referred to as **chemical shift anisotropy** (CSA). In solution this effect is averaged so that only one frequency is observed for each chemically distinct site. Nonetheless, this fluctuating field can stimulate relaxation if sufficiently strong (e.g. ^{19}F , ^{31}P and many metals).¹⁶

The CSA interaction is the only one requiring the presence of a magnetic field, and it makes a stronger contribution to relaxation as the magnetic field increases. Its dependence on the square of the applied field has greater significance at higher B_0 . For some spin $\frac{1}{2}$ nuclei with large chemical shift ranges, lines become sufficiently broadened by CSA relaxation at high field to cause loss of coupling information.

This mechanism is never seen for protons, and is seen for carbon only when there are no attached protons (e.g., carbonyl compounds). This is of great importance for instance when choosing (if possible) between different magnetic fields for acquiring HNCO or HN(CA)CO experiments (*see sections VII.3.3.1.1 and VII.3.3.1.2*). In these cases lower fields may give

better results. Another consequence of CSA relaxation is the reduction or loss of NOE effects when protons are decoupled.⁷⁶

VII.2.3 The protein's fingerprint - ¹⁵N-¹H-HSQC

The **H**eteronuclear **S**ingle-**Q**uantum **C**oherence (HSQC) experiment^{71,73,77,78} correlates the ¹⁵N or ¹³C nuclei with the attached ¹H via the one-bond scalar coupling J_{N-H} or J_{C-H} , respectively (**Figure VII.9**). The result is a two-dimensional spectrum with one axis for ¹H and the other for the heteronucleus (¹⁵N or ¹³C). Thus, in the ¹⁵N-¹H-HSQC one signal is expected for each amino acid residue with the exception of proline which has no amide-hydrogen due to the cyclic nature of its backbone. Tryptophan side chain N_ε-H_ε group and asparagine and glutamine side chains N_δ-H_{δ2} and N_ε-H_{ε2}, respectively, also give rise to additional signals. The arginine N_ε-H_ε peaks are in principle also visible, but because the N_ε chemical shift is outside the region usually recorded, the peaks are folded. If working at low pH the Arg N_η-H_η and Lys N_ζ-H_ζ groups can also be visible, but are also folded.⁷⁹ In the ¹³C-¹H-HSQC each C-H will give a crosspeak.

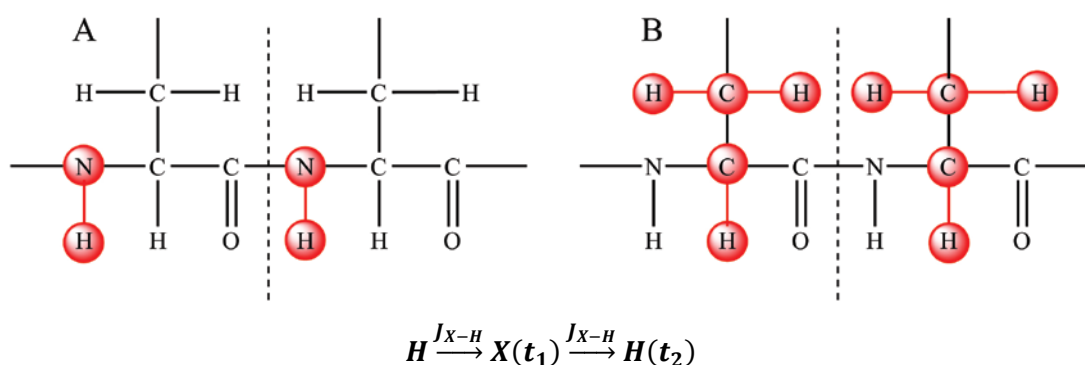


Figure VII.9: The ¹⁵N-¹H-HSQC (A) and ¹³C-¹H-HSQC (B) magnetization transfer.

Magnetization is first transferred from ¹H to the X nuclei (either ¹⁵N or ¹³C) with a standard INEPT sequence via $^1J_{X-H}$ scalar coupling and then transferred back to the proton for detection. Proton magnetization is detected (during t_2 - detection time) while the X nuclei evolves during the evolution time - t_1 . Because of the detection of the high frequency nuclei, this sequence is very sensitive.

As each protein has a unique pattern of signal positions, the ¹⁵N-¹H-HSQC is often referred to as the fingerprint of a protein. Because of this characteristic it is typically the first experiment to be measured with an isotope-labeled protein. Analysis of the ¹⁵N-¹H-HSQC allows researchers to make an initial assessment of several parameters such as:

- Whether the protein is well folded or unfolded;
- Whether the expected number of peaks is present and thus identifying possible problems due to multiple conformations or sample heterogeneity;
- Whether it is feasible to do subsequent longer, more expensive, and more elaborate experiments, thus saving time and money

Although it is not possible to assign peaks to specific atoms from the heteronuclear single quantum correlation alone, due to some specific chemical shift values it is possible to narrow some amino acid types as shown in **Figure VII.10**. Furthermore, because the ^{15}N - ^1H -HSQC acts as a fingerprint of the protein (as said above) it is very useful for detecting interactions with ligands, such as other proteins or drugs as the chemical shift of the residues that are interacting will change. By comparing the ^{15}N - ^1H -HSQC of the free protein with the one bound to the ligand, it is possible to identify the binding interface, as seen in Chapter III. If the entire signals are assigned, by titrating the protein with ligand, one can also calculate equilibrium affinity constants or, if it is done at different temperatures, calculate thermodynamic parameters (*see Chapter III*).

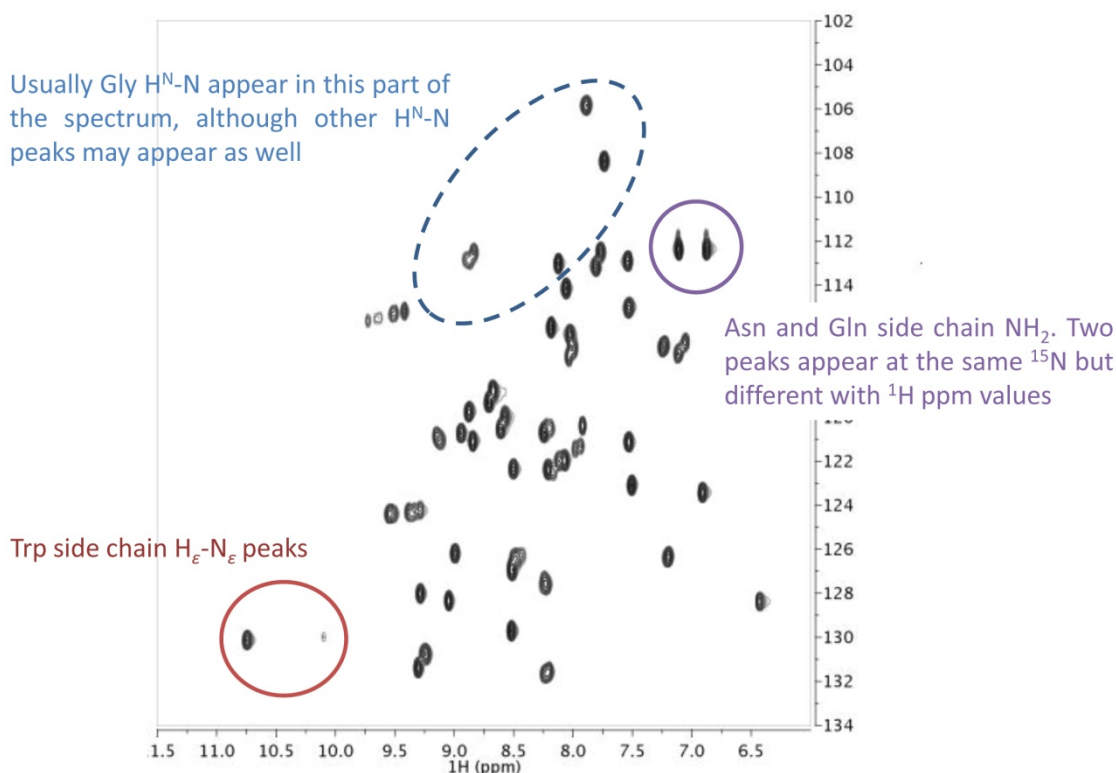


Figure VII.10: ^{15}N - ^1H -HSQC spectrum of the 52 amino acid (5.677 Da) protein rubredoxin from the sulfate-reducing bacterium *Desulfovibrio gigas*⁸⁰ (pdb code: 1rdg).

Each peak in the spectrum represents a bonded N-H pair belonging to the amide group of the amino acids or to the side chains of tryptophans (brown circle), asparagines or glutamines (purple circle).

VII.2.4 Nuclear Overhauser effect

When the resonance of a spin is perturbed by saturation or inversion of the magnetization, it may cause the spectral intensities of other resonances in the spectrum to change. This phenomenon is known as the **Nuclear Overhauser Effect** (NOE).⁸¹ NOE based methodologies are an essential part of routine NMR spectroscopy used for assignment, structure elucidation and conformational analysis. It is also one of the most important experimental methods for the structural analysis of biological macromolecules.

The NOE may be defined as the change in intensity of one resonance when the spin transitions of another are somehow perturbed from their equilibrium populations.¹⁶ For instance: if one resonance A is irradiated, an increase (positive NOE) or decrease (negative NOE) of signal intensity of other resonances is observed when the spins are close in space (**Figure VII.11**):

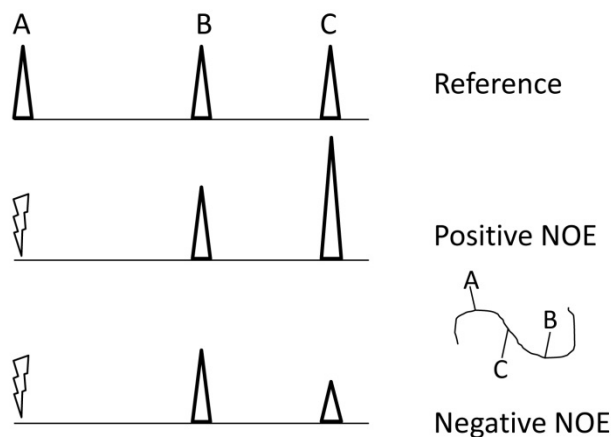


Figure VII.11: Irradiation of resonance A leads to an increase of peak intensity of the neighboring spin C (positive NOE) or to a decrease of peak intensity (negative NOE).

Coupled with information from scalar spin-spin couplings, the NOE effect is **the** method for elucidation of 3D structural features and stereochemistry.^{41,82,83} The magnitude is expressed as a relative intensity change between the equilibrium intensity, I_0 , and that in the presence of the NOE, I , such that:

$$NOE \equiv f_I\{S\} = \frac{I - I_0}{I_0}$$

VII.11

where $f_I\{S\}$ indicates the fractional change in the signal intensity upon irradiation for spin I when spin S is perturbed and I and I_0 are the signal intensity with and without irradiation, respectively.^{82,84}

In order to understand the origin of the NOE and the factors that dictate its sign and magnitude let's consider a system comprising only two homonuclear spin $-1/2$ nuclei, I and S , which exist in a rigid molecule. The two nuclei do not share a scalar coupling ($J_{IS} = 0$) but are sufficiently close to share a **dipolar coupling**. This is the direct, through-space magnetic interaction between the two spins such that one spin is able to sense the presence of its dipolar-coupled partner. An energy diagram is shown in **Figure VII.12**:

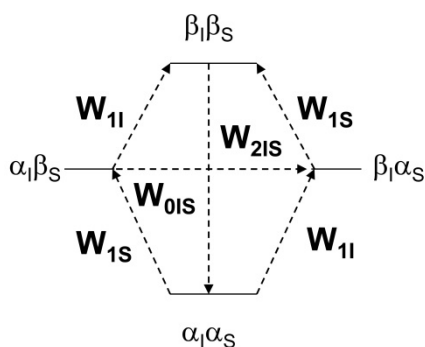


Figure VII.12: Energy level diagram for a two homonuclear spin system $-1/2$ nuclei, I and S , showing definitions of transition probabilities and spin states.¹

In **Figure VII.12** spin states are written with the state of I first and S second (e.g., $\alpha\beta$ means spin I in state α (low energy – aligned with B_0) and spin S in state β (higher energy – aligned against B_0)). The W labels represent the transition probabilities for each spin. The two other transitions, $\alpha\beta\text{-}\beta\alpha$ and $\alpha\alpha\text{-}\beta\beta$, involve the simultaneous flipping of both S and I spins. The $\alpha\beta\text{-}\beta\alpha$ W_0 process is referred to as the **zero quantum transition**, whereas the $\alpha\alpha\text{-}\beta\beta$ W_2 process is the **double-quantum transition**.^{1,59} Both are able to act as relaxation pathways and, in fact, it is only these two that are responsible for the NOE itself. Collectively, they are referred to as **cross relaxation pathways**, a term suggestive of the simultaneous participation of both spins. Because we are considering a homonuclear system, the energies of the I and S transitions will be essentially identical (chemical shift differences are negligible relative to Larmor frequencies), and we can therefore assume that the populations of the $\alpha\beta$ and $\beta\alpha$ states are equal at equilibrium (**Figure VII.13 - A**). According to the Boltzmann distribution, there will be an excess of nuclei in the lower energy orientation, and a deficit in the higher energy $\beta\beta$ state.¹

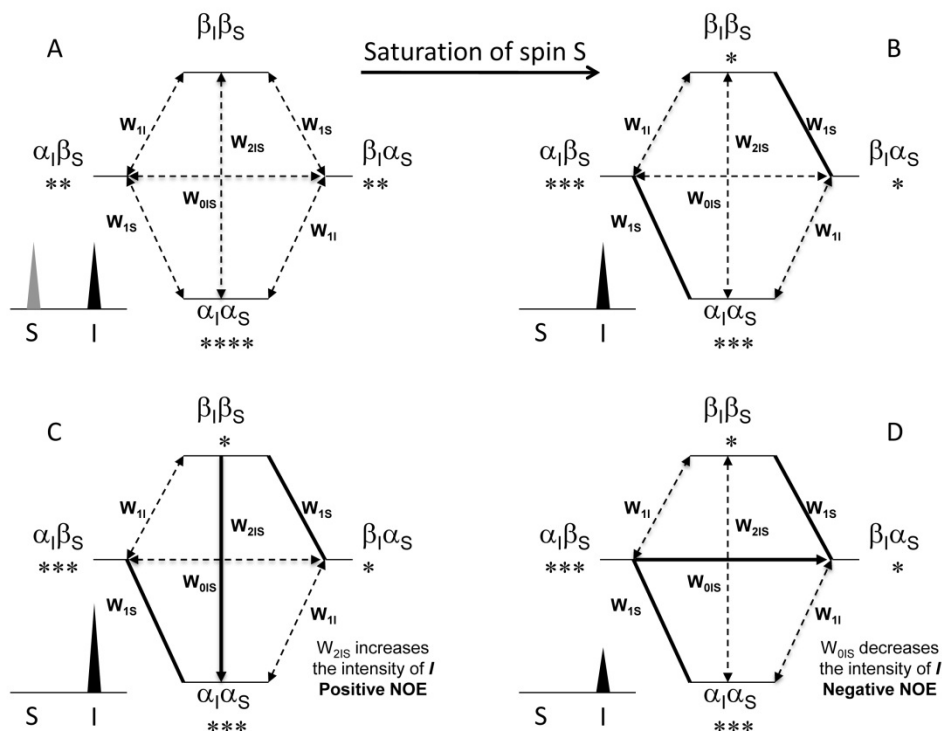


Figure VII.13: Schematic representation of the origin of the NOE in a homonuclear two $\frac{1}{2}$ nuclei spin system.

A) Equilibrium situation; B) Condition after saturation of S resonance; C) Effect of W_2 relaxation during saturation of S and origin of the positive NOE enhancement; D) Effect of W_0 relaxation during saturation of S and origin of the negative NOE enhancement. The “*” represent spin populations.

Saturating the S resonance will force the population differences across the S transitions to zero, i.e. the populations are equalized such that $\alpha_i\alpha_s = \alpha_i\beta_s$ and $\beta_i\alpha_s = \beta_i\beta_s$ (**Figure VII.13 - B**). Therefore, transitions between these states are no longer possible. The only way for the system to return to the equilibrium state is by altering its spin populations via W_{0IS} and W_{2IS} (**Figure VII.13 – C and D**). The frequencies associated with the transition probabilities are:

- W_{1I} ($\alpha_i\alpha_s \leftrightarrow \beta_i\alpha_s$ and $\alpha_i\beta_s \leftrightarrow \beta_i\beta_s$) is associated with ω_I
- W_{1S} ($\alpha_i\alpha_s \leftrightarrow \alpha_i\beta_s$ and $\beta_i\alpha_s \leftrightarrow \beta_i\beta_s$) is associated with ω_S
- W_{2IS} ($\alpha_i\alpha_s \leftrightarrow \beta_i\beta_s$) is associated with $(\omega_I + \omega_S)$, approximately $2\omega_I$
- W_{0IS} ($\alpha_i\beta_s \leftrightarrow \beta_i\alpha_s$) is associated with $(\omega_I - \omega_S)$, approximately zero

As can be seen from **Figure VII.13**, the W_2 process will act to remove spins from the $\beta\beta$ state and transfer them to the $\alpha\alpha$ state in an attempt to recover the population differences across the S transitions. In doing so, this will increase the population difference across the two I transitions. Thus, relaxation via the W_2 process will result in a net increase in the I spin resonance intensities in the spectrum; this is then a **positive NOE**. Likewise, the W_0 process will act to transfer spins from the $\alpha\beta$ to the $\beta\alpha$ state, again in an attempt to recover the population

differences across the S transitions. In this case, the result will be a decrease in the population difference across the two I transitions so that relaxation via the W_0 process will result in a net reduction in the I spin resonance intensities in the spectrum; this is then a **negative NOE**.^{16,59,85} The magnitude of the steady state NOE enhancement of I after saturating S can be calculated according to the following equation:^{16,82}

$$f_I\{S\} = \frac{\gamma_S}{\gamma_I} \frac{W_{2IS} - W_{0IS}}{W_{0IS} + 2W_{1I} + W_{2IS}}$$

VII.12

where $W_{2IS}-W_{0IS}$ describes the rate of the dipole-dipole transitions and is called the cross-relaxation rate, σ_{IS} , and $W_{0IS} + 2W_{1I} + W_{2IS}$ is the longitudinal relaxation rate constant of spin I, ρ_{IS} (auto-relaxation).

Whether the final result is a positive or negative NOE depends on the relative contribution of each type of relaxation to the total relaxation. The individual contribution of each transition for the total relaxation depends on the probability of a molecular motion having the same frequency as the transition. If the frequency of the motion matches the difference of two energy levels it can induce changes in their populations. In order to analyze how a molecule tumble one can use a **correlation function**, $G(t)$ (**Equation VII.13**), defined as the average of the molecular orientation at a certain time (t), and a little while after that ($t + \tau$) and, for isotropic rotational diffusion of a rigid rotor is given by:⁸⁶

$$G(t) = G(0)e^{-t/\tau_c}$$

VII.13

where, τ_c is the **correlation time** (decay time of the correlation function). When considering isotropic molecular tumbling, τ_c is related with the time taken for the molecule to rotate by 1 radian about any axis. Therefore, rapidly tumbling molecules will have short correlation times while slowly tumbling molecules will have long correlation times. It is possible to relate correlation times with the size and shape of the molecules and a very rough estimate of τ_c for molecule of mass M_w is given by:^{16,82}

$$\tau_c \cong 10^{-12}M_w$$

VII.14

Usually, τ_c is of the order of picoseconds for small molecules and in the order of nanoseconds for large molecules in aqueous solution.⁷¹ The power available within a molecular

system to induce transitions by virtue of its correlation time is referred to as the **spectral density** $J(\omega)$ and is obtained after a Fourier transformation of the correlation function:¹⁶

$$J(\omega) = \frac{2\tau_c}{1 + \omega^2\tau_c^2}$$

VII.15

Since the correlation time is affected by the motion regime (related to the molecular size), the spectral density function can be analyzed as a function of slow, intermediate or fast motion (**Figure VII.14**). Accordingly, for a molecule with a short τ_c (rapid tumbling – small molecule), there is an almost equal chance of finding components at both high and low frequencies, up to about $1/\tau_c$ at which point the probability drops rapidly. On the other hand, for molecules that possess a long τ_c (slow tumbling – large molecules), the probability of generating rapidly oscillating fields is very small, so the corresponding spectral density is concentrated into a smaller frequency window. These curves therefore predict how the relaxation rates will vary with correlation time.^{16,85}

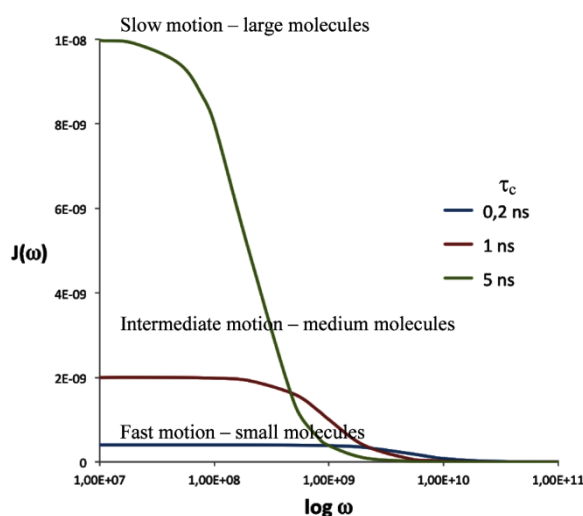


Figure VII.14: Variation of the spectral density with the molecular motion as a function of the frequency.⁸⁵

Since W_{0S} transitions are associated with small energy differences and low frequencies ($\omega_I - \omega_S$), they will be favored by large molecules, tumbling slow in solution. On the other hand W_{2S} transitions are associated to higher frequencies ($\omega_I + \omega_S$) and will be favored by fast tumbling molecules. Therefore, small molecules are associated with positive NOEs and large molecules are associated with negative NOEs.

The rate constants for the transitions mentioned above can be expressed in terms of the spectral density $J(\omega)$ and the distance r_{IS} between the two spins according to:^{16,82}

$$W_{I0S} = \frac{3\mu_0^2 h^2 \gamma_I^2 \gamma_S^2 J(\omega_I - \omega_S)}{1280\pi^4 r_{IS}^6}$$

VII.16

$$W_{1I} = \frac{3\mu_0^2 h^2 \gamma_I^2 \gamma_S^2 J(\omega_I)}{2560\pi^4 r_{IS}^6}$$

VII.17

$$W_{2IS} = \frac{3\mu_0^2 h^2 \gamma_I^2 \gamma_S^2 J(\omega_I + \omega_S)}{640\pi^4 r_{IS}^6}$$

VII.18

where μ_0 is the magnetic permeability of vacuum, h is Planck's constant, γ_I and γ_S are the gyromagnetic ratios of the spins I and S , respectively and r_{IS} is the internuclear distance between the two spins.

The inverse-sixth relationship means the NOE falls away very rapidly with distance, so in practice significant NOEs will only develop between protons that are within approximately **5Å** of each other (even if they are far apart in the bonding network). These measured distances are used to determine accurate three-dimensional structures of proteins and nucleic acids. Furthermore, because it also depends on the square of the magnetogyric ratios of the two spins involved, for heteronuclear systems very distinct rates may occur depending on the participating spins. These rate constants can be used in **Equation VII.12** for the calculation of the NOE effect:⁸²

$$f_I\{S\} = \frac{\gamma_S}{\gamma_I} \frac{6J(\omega_I + \omega_S) - J(\omega_I - \omega_S)}{6J(\omega_I + \omega_S) + 3J(\omega_I) + J(\omega_I - \omega_S)}$$

VII.19

According to the above equation, for an ideal two spin system the steady state NOE is not dependent on the internuclear distance. In the homonuclear case, where $\gamma_I = \gamma_S$ and there is only one frequency $\omega_I \cong \omega_S \cong \omega_0$, so that $(\omega_I - \omega_S)$ is always much less than one, $f_I\{S\}$ simplifies to:

$$f_I\{S\} = \frac{5 + \omega_0^2\tau_c^2 - \omega_0^4\tau_c^4}{10 + 23\omega_0^2\tau_c^2 + 4\omega_0^4\tau_c^4}$$

VII.20

When a molecule tumbles so rapidly in solution such that $\omega\tau_c \ll 1$ (small molecules in non-viscous solvent), all terms in these expressions containing ω become negligible. Under these conditions the NOE has a maximum value of 0.5 (50%). This is known as **extreme narrowing condition**. In the intermediate region, as the molecular motions slow down, the NOE approaches zero (when $W_{2IS} = W_{0IS}$) and then changes sign to reach a new negative maximum (-100%) for molecules that tumble very slow in solution (large molecules in viscous solvent). The zero NOE cross-over point (when $\omega_0\tau_c = \sqrt{5/4} \cong 1.12$) occurs for medium sized molecules (1000-2000 Daltons) is highly sensitive to molecular motion and is also dependent on the spectrometer field strength, which determines ω_0 .

As said above, this is only true for an ideal two spin system relaxing exclusively by dipolar interactions. Nonetheless, in real cases the steady-state NOE depends on the molecular geometry and the equations need to be extended to realistic multispin systems. Under these conditions, spin I will be relaxed not only by spin S but also potentially by all other spins (X) in the molecule collectively depending on their distances to I . Another important factor when considering NOE in multispin systems is the possibility of indirect effects. Once spins close to S have received NOE enhancements, their own populations are no longer at equilibrium and this disturbs the balance of their cross relaxation with their own neighboring spins, thereby creating additional NOE enhancements often called **indirect enhancements**. In this situation, and assuming that the spin system is part of a rigid molecule tumbling isotropically in solution at a rate described by τ_c and that relaxation is entirely dipole-dipole*, in the homonuclear case we have:⁸²

$$f_I\{S\} = \eta_{max} \left[\frac{r_{IS}^{-6} - \sum_X f_X\{S\} r_{IX}^{-6}}{r_{IS}^{-6} + \sum_X r_{IX}^{-6}} \right]$$

VII.21

This equation relates steady-state NOE enhancements with internuclear distances and predicts two types of contributions to the steady state enhancements: the direct and indirect. The direct contributions are related to the proportion of cross relaxation of spin I with the saturated

* Other relaxation mechanisms include, for instance, intermolecular dipole-dipole, quadrupolar relaxation, chemical shift anisotropy or spin-rotation and are referred to as “leakage”. Their effect is the reduction of the NOE enhancement by “diluting” the contribution of intramolecular dipole-dipole relaxation to the total relaxation of spin I .

spin S and the indirect contribution corresponds to all the enhancements at I that have arrived via cross-relaxation of S with some third-spin X followed by cross-relaxation of X with I , over any number of intermediates (**Figure VII.15**).

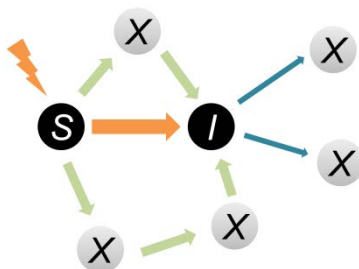


Figure VII.15: Schematic representation of the relaxation pathways that lead to direct and indirect contributions to the NOE enhancement of spin I upon S saturation in a multispin system.⁸⁵

The population disturbance initially present only at spin S spreads through the molecule by cross-relaxation from spin to spin and at steady state all spins are affected to a greater or lesser extent. This process is referred to as spin-diffusion and has different properties in the positive ($\eta_{\max} > 0$) or negative ($\eta_{\max} < 0$) NOE regime. For small molecules (provided that the extreme narrowing limit can be assumed) direct enhancements are positive and the shorter the internuclear distance, r_{IS} , the larger the corresponding enhancement. The presence of other spins that cross-relax with I will diminish the enhancement $f_I\{S\}$ and the effect will be more pronounced the closer these other spins are to I . In the positive NOE regime, where W_2 predominates over W_0 , indirect enhancements transmitted over one intermediate spin are negative; those transmitted over two intermediate spins are positive, and so on. Luckily, the transmission of enhancements down a chain of spins is a relatively inefficient process and in practice effects transmitted over more than one intermediate spin are almost never observed in small molecules.⁸⁵

On the other hand, for homonuclear experiments with large molecules all enhancements (direct and indirect) are negative and the predominance of W_0 transitions means that indirect enhancements can be transmitted efficiently down a chain of spins. In this situation, saturation of any spin in a homonuclear multispin system will, at steady state, cause a -100% enhancement of every other spin, leading to a saturated spectrum where no resonances can be seen. Because of this, steady-state NOEs in the negative NOE regime are useless to provide reliable distance or proximity information. However the internuclear distance affects the rate at which steady state is reached so in order to extract distance information, it becomes necessary to consider the rate at which NOEs build-up between spins.⁸⁵

VII.3 Protein structure determination

The potential of solution NMR spectroscopy for determining *de novo* structures of biological macromolecules such as proteins, DNA and RNA has been widely demonstrated.^{3,4,21,48,61,64,71,87} However, although there are more than 9000 NMR structures deposited in the Protein Data Bank, no standard procedures have been developed for NMR structure determination of proteins, and different laboratories use a variety of different approaches.^{61,88-90}

The first step for NMR solution structure determination is obtaining the protein, which can be either non-labeled or isotope labeled with ^{13}C , ^{15}N , ^1H , $^{13}\text{C}/^{15}\text{N}$ or $^{13}\text{C}/^{15}\text{N}/^1\text{H}$, depending on what one wants and on the size of the protein. The protocols used for expressing $^{13}\text{C}/^{15}\text{N}$ -labeled and non-labeled proteins are given in Chapters II and III, respectively. After obtaining a pure protein sample one can start acquiring the data. The type of spectra and the time required for their acquisition will depend on the protein size and concentration. The approach I have followed consists in:

- 1) Acquire a 1D ^1H spectrum in order to check the protein purity, folding/unfolding state, confirm that the concentration is good enough and calibrate the required pulses and solvent suppression scheme;
- 2) Acquire 2D $^{13}\text{C}/^{15}\text{N}$ - ^1H -HSQC spectra. These spectra are vital as they allow researchers to make an initial assessment of several parameters (*see Section VII.2.3*). The ^{13}C - ^1H -HSQC is acquired both in the aliphatic and aromatic regions.
- 3) Acquire the 3D set of experiments that will allow me to assign all the resonances of the protein. According to the protocol I have followed, these spectra are:
 - a) For backbone assignment:
 - i. HNCO
 - ii. HN(CA)CO
 - iii. HN(CO)CACB
 - iv. HNCACB
 - b) For sidechain assignment
 - i. (H)CCH-TOCSY
 - ii. HNHA
 - c) Distance calculation
 - i. ^{15}N - ^1H -NOESY-TOCSY
 - ii. ^{13}C - ^1H -NOESY-TOCSY (aliphatic)
 - iii. ^{13}C - ^1H -NOESY-TOCSY (aromatic)

All the spectra referred above are explained in more detail in Section VII.3.1. When compared with 2D experiments, triple resonance experiments¹¹ provide better signal dispersion and, therefore, less ambiguities in chemical shift assignment.

After acquiring and processing all data it is necessary to assign all peaks, identify all the spin systems and sequentially link them. I have done this using the software CARA1.8.4.2⁹¹. When all backbone and sidechain peaks are assigned (or at least most of them) it is necessary to get the distance information key to structure calculation. In order to get this information I have used the NOESY data (Section VII.3.1.3.2). The peaks were peaked and integrated using CARA1.8.4.2⁹¹. The volumes were converted into upper limits (UPLs) by CYANA2.1⁹² using the macro *calibrate* (Section VII.3.1.3.2). Besides the distance constrains I also used angular constraints. These were obtained using the software TALOS+⁹³ (Section VII.3.1.1.5). This data, along with the chemical shifts of the peaks were finally introduced into the software CYANA2.1⁹² for the structure calculation (see Chapter II). The assignments were then evaluated and, in accordance to the previously determined structure, new assignments were found and then a new structure was calculated. This process was repeated until good statistics were obtained (Section VII.3.2) (Figure VII.16). The final ensemble of structures (20) was refined in AMBER⁹⁴ and validated using PROCHECK-NMR⁹⁵ (Section VII.3.2).

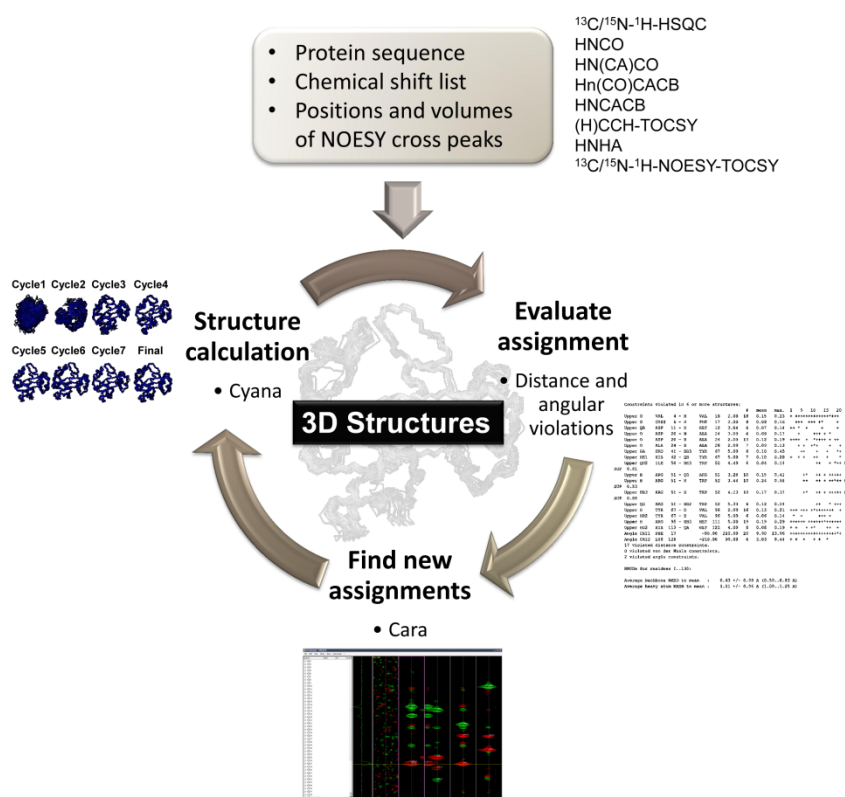


Figure VII.16: Process of 3D solution structure calculation from NMR data.

VII.3.1 Three-dimensional experiments

A three dimensional NMR experiment can be constructed from a two dimensional one simply by inserting an additional indirect evolution time and a second mixing period between the first mixing period and the data acquisition (**Figure VII.17**). Each of the different indirect time periods (t_1 , t_2) is incremented separately.

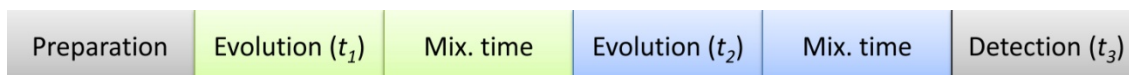


Figure VII.17: Anatomy of a 3D NMR experiment.

The green rectangles represent the additional evolution and mixing times, necessary for constructing a 3D experiment from a 2D one and the blue rectangles represent the additional evolution and mixing times, necessary for constructing a 3D experiment from a 2D one.

Triple-resonance experiments, involving ^{15}N , ^{13}C and ^1H spins, are the method of choice to provide consistent and robust protein resonance sequential assignments. The addition of a third dimension reduces tremendously the signal overlapping. These experiments rely on the fact that one-bond and some two-bond scalar couplings, 1J and 2J (**Table VII.3** and **Figure VII. 18**) are relatively larger than the linewidth of the nuclei under consideration ($J > \Delta\nu_{1/2}$), consequently, the transfer via these couplings remains highly efficient even for relatively large molecules.⁷² Furthermore, 1J couplings are independent of the conformation of the protein. The major drawback is that one needs double-labeled (^{15}N and ^{13}C) protein which is often expensive.

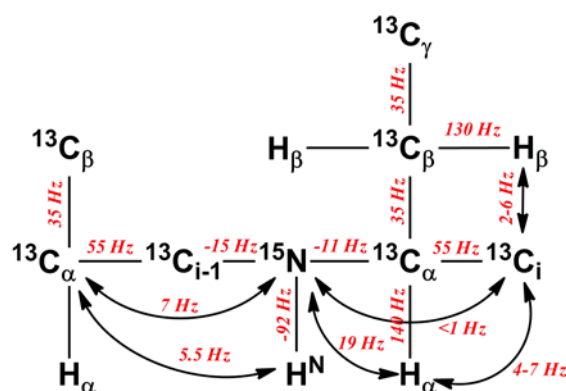


Figure VII. 18: Scalar coupling constants between the different nuclei in amino acids.⁷²

All of the one-bond scalar couplings are basically independent of the secondary structure whilst two-bond couplings are not. Note that the two-bond coupling between the amide nitrogen and its own carbonyl carbon is essentially zero, thus it is only practical to directly correlate the amide nitrogen shift with the carbonyl shift of the preceding residue.⁸⁷

The triple-resonance experiments that I used for protein structure determination are listed in **Table VII.4** and briefly explained in the following sections.

Table VII.4: Pulse sequences typically used for protein structure determination as described in this chapter.

<i>Experiment</i>	<i>Nuclei observed</i>	<i>Relative S/N (%)^a</i>	<i>Section</i>
<i>HNCO</i>	H _i , N _i , CO _{i-1}	100	VII.3.1.1.1
<i>HN(CA)CO</i>	H _i , N _i , CO _i , CO _{i-1}	13/4 α/β	VII.3.1.1.2
<i>HN(CO)CACB</i>	H _i , N _i , C _{αi-1} , C _{βi-1}	13/9 α/β	VII.3.1.1.3
<i>HNCACB</i>	H _i , N _i , C _{α} , C _{β} , C _{αi-1} , C _{βi-1}	4/1.7 α/β (<i>i</i>)	VII.3.1.1.4
<i>(H)CCH-TOCSY</i>	H ^{aliph} , C ^{aliph}		VII.3.1.2.1
<i>HNHA</i>	H _i , N _i , H _{αi}		VII.3.1.2.2
<i>¹⁵N/¹³C-NOESY-HSQC</i>	H _i , N _i /C _i , H _j , N _j /C _j		VII.3.1.3.1

^a The sensitivity of backbone assignment experiments is relative to the HNCO experiment

VII.3.1.1 Experiments for backbone assignments

VII.3.1.1.1 HNCO

The HNCO experiment^{12,71-74,87} is one of the simplest 3D experiments and, at the same time, the most sensitive[†]. It correlates the amide group chemical shift with the carbonyl carbon (CO₋₁) of the preceding residue by using the one-bond J_{N-H^N} and J_{N-CO} coupling constants, as shown in **Figure VII.19**. In addition, asparagine and glutamine side-chain correlations are also visible and the CO chemical shifts obtained can be used to help predict secondary structure. **Figure VII.20** shows an example of how such assignment is done for the protein CtCBM11 (*see Chapter II*). The HNCO can also be useful for backbone assignment in conjunction with the HN(CA)CO, if the CBCANNH and CBCA(CO)NNH spectra are of bad quality. When acquiring this type of spectra it is necessary to have in mind that, due to **carbonyl CSA relaxation**, high magnetic fields may give worse results than lower fields.

[†] Since the coherence transfer rate between two spins is proportional to their mutual coupling constant, the most efficient three-dimensional NMR experiments take advantage of coherence transfer between spins coupled with the largest J values

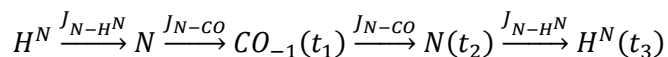
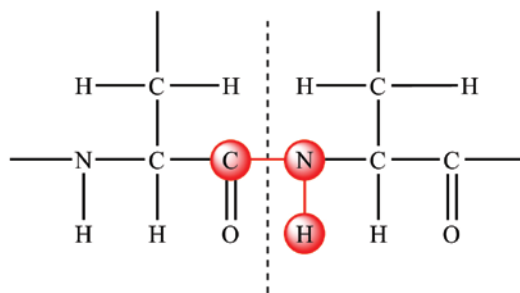


Figure VII.19: The HNCO magnetization transfer.

Magnetization is passed from $^1H^N$ to ^{15}N with a standard INEPT sequence via J_{N-H^N} coupling. Then it is selectively transferred to the carbonyl ^{13}CO via the J_{N-CO} coupling. Magnetization is then passed back via ^{15}N to $^1H^N$ for detection. The chemical shift is evolved on all three nuclei resulting in a three-dimensional spectrum.⁷⁹

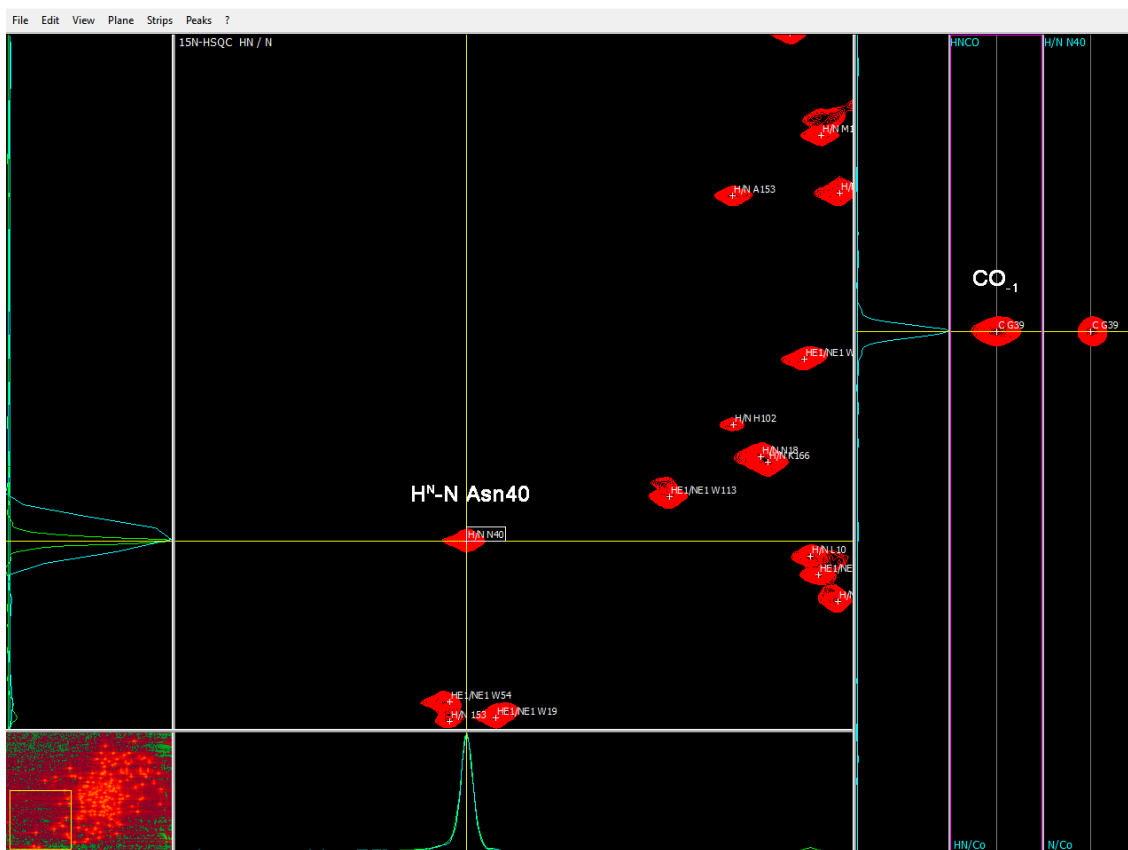


Figure VII.20: Identifying the CO_{-1} resonance.

In this window, the central panel shows a part of the ^{15}N - 1H -HSQC spectrum of CrCBM11 and the two right strips show H^N -CO (labeled HN/Co) and N-CO (labeled N/Co) planes of the HNCO spectrum at the frequency of the selected amide group (in this case its Asn40).

VII.3.1.1.2 HN(CA)CO

The HN(CA)CO experiment^{13,71-74} correlates the amide group chemical shift with its own carbonyl carbon (CO) and with the one of the preceding residue by using the J_{N-H^N} , J_{N-C_α} and $J_{C_\alpha-CO}$ coupling constants, as it is shown in **Figure VII.21**.

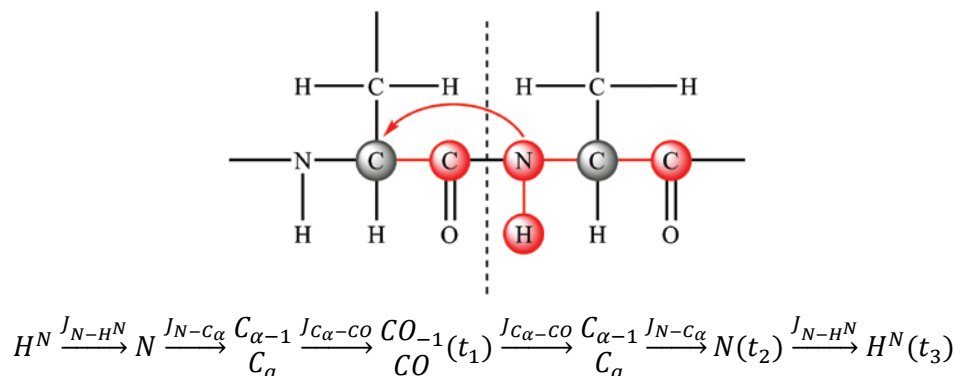


Figure VII.21: The HN(CA)CO magnetization transfer.

Magnetization is passed from $^1H^N$ to ^{15}N with a standard INEPT sequence via J_{N-H^N} coupling. Then, via a second INEPT sequence, the magnetization from the amide ^{15}N is transferred to the $^{13}C_\alpha$ (red arrow) using the J_{N-C_α} coupling constant. From there it is transferred to the ^{13}CO via the $J_{C_\alpha-CO}$ coupling constant. For detection the magnetization is transferred back the same way: from ^{13}CO to $^{13}C_\alpha$, ^{15}N and finally $^1H^N$. The chemical shift is only evolved on ^{13}CO (t_1), ^{15}N (t_2) and $^1H^N$ (t_3) and not on the $^{13}C_\alpha$. The result is a three-dimensional spectrum. Because the amide nitrogen is coupled both to the C_α of its own residue and that of the preceding residue, both these transfers occur and transfer to both ^{13}CO nuclei occurs. Thus for each NH group, two carbonyl groups are observed in the spectrum. But because the coupling between N_i and $C_{\alpha i}$ (11 Hz) is stronger than that between N_i and $C_{\alpha i-1}$ (7 Hz), the $H_i-N_i-CO_i$ peak generally ends up being more intense than the $H_i-N_i-CO_{i-1}$ peak.^{71,73,79}

When used in conjunction with the HNC(O) experiment (see previous section), this experiment provides a method for **sequentially** assigning the amide 1H , ^{15}N , and ^{13}CO resonances. The main limitation of the HN(CA)CO experiment is the low sensitivity that results from the (i) rapid relaxation of the transverse $^{13}C_\alpha$ magnetization during the delays and the (ii) the weaker $N_i-C_{\alpha i-1}$ coupling in relation to $N_i-C_{\alpha i}$. As a consequence, a fraction of the correlations may not be observed in the experiment. **Figure VII.22** shows an example of how the CO_i resonance is assigned for the protein CtCBM11 (see Chapter II).

Again, when acquiring this type of experiment, one must have in mind that as in the case of the previous one, high magnetic fields may give worse results than lower fields due to **carbonyl CSA relaxation**.

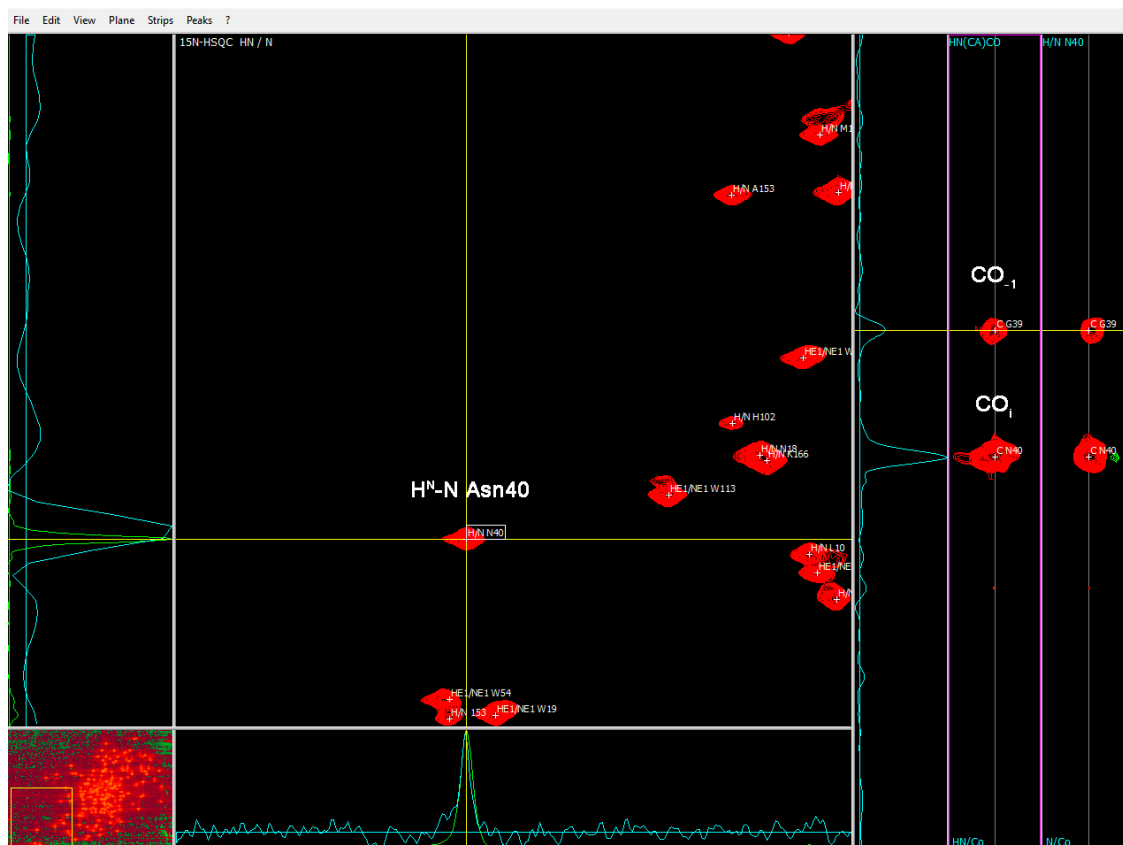


Figure VII.22: Identifying the CO_i resonance.

In this window, the central panel shows a part of the ^{15}N - ^1H -HSQC spectrum of *CtCBM11* and the two right strips show H^{N} -CO (labeled HN/Co) and N-CO (labeled N/Co) planes of the HN(CA)CO spectrum at the frequency of the selected amide group (in this case its Asn40).

VII.3.1.1.3 HN(CO)CACB

The HN(CO)CACB experiment¹² correlates the amide group chemical shift with both the alpha (C_α) and beta (C_β) carbons of the preceding residue via the intervening ^{13}C spin by means of the $J_{\text{N}-\text{H}^{\text{N}}}$, $J_{\text{N}-\text{CO}}$, $J_{\text{C}_\alpha-\text{CO}}$ and $J_{\text{C}_\alpha-\text{C}_\beta}$ coupling constants, as it is shown in **Figure VII.23**. The resonances of C_α and C_β provide information about the amino acid type of the preceding residue in addition to the sequential connectivity: for instance, for threonine and serine the C_β usually appears at higher ppm values than the C_α ; another thing is that it is very easy to identify glycines as they don't have C_β . The main limitation of this experiment is its limited sensitivity due to the fast transverse relaxation rate of $^{13}\text{C}_\alpha$. **Figure VII.24** shows an example of how the CA_{-1} and CB_{-1} resonances are assigned for the protein *CtCBM11* (see *Chapter II*). Note that the absence of a CB_{-1} peak indicates that the residue prior to Asn40 is a glycine (which is correct).

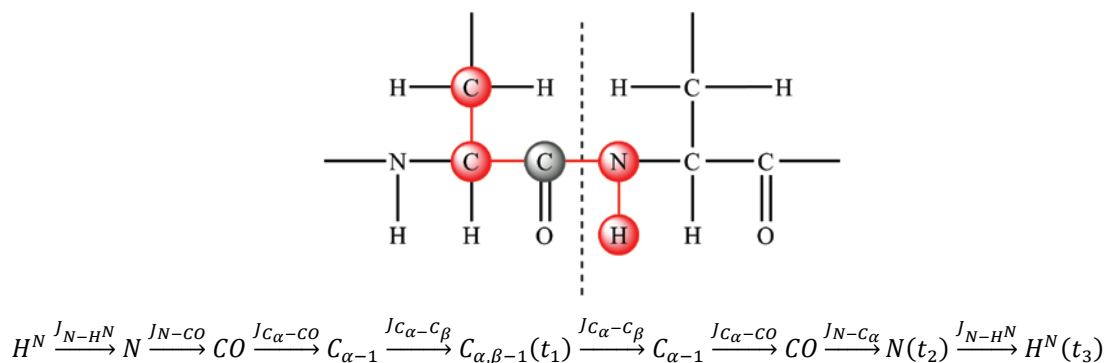


Figure VII.23: The HN(CO)CACB magnetization transfer.

Magnetization is passed from $^1\text{H}^{\text{N}}$ to ^{15}N with a standard INEPT sequence via $J_{\text{N}-\text{H}^{\text{N}}}$ coupling. Then, via a second INEPT sequence, the magnetization from the amide ^{15}N is transferred to the ^{13}CO using the $J_{\text{N}-\text{CO}}$ coupling constant. From there it is transferred to the $^{13}\text{C}_\alpha$ and $^{13}\text{C}_\beta$ via the $J_{\text{C}_\alpha-\text{CO}}$ and $J_{\text{C}_\alpha-\text{C}_\beta}$ coupling constants, respectively, where the chemical shifts evolve during t_1 . From there the magnetization returns to $^{13}\text{C}_\alpha$. From here it is transferred first to ^{13}CO , then to ^{15}N , where the chemical shifts evolve during t_2 . Finally, the magnetization is transferred to $^1\text{H}^{\text{N}}$ for detection with an evolution of the chemical shifts during t_3 . Because the chemical shift is evolved simultaneously on $^{13}\text{C}_\alpha$ and $^{13}\text{C}_\beta$, these signals appear in one dimension. The chemical shifts evolved in the other two dimensions are ^{15}N and $^1\text{H}^{\text{N}}$. The chemical shift is not evolved on ^{13}CO . In this spectrum the $^{13}\text{C}_\alpha$ signal appears with a positive phase while the $^{13}\text{C}_\beta$ appears with a negative one.⁷⁹

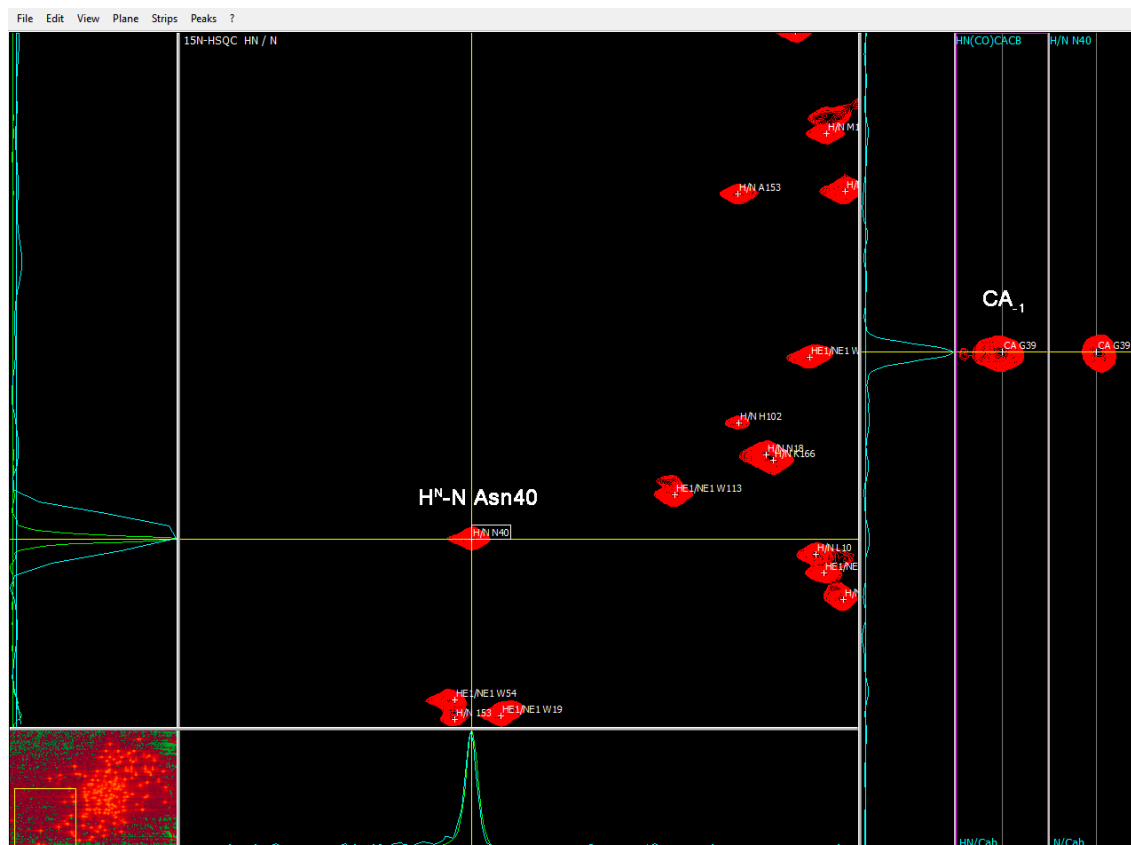


Figure VII.24: Identifying the CA₁ and CB₁ resonances.

In this window, the central panel shows a part of the ^{15}N - ^1H -HSQC spectrum of *Ct*CBM11 and the two right strips show $\text{H}^{\text{N}}\text{-}^{13}\text{C}$ (labeled HN/Cab) and $\text{N-}^{13}\text{C}$ (labeled N/Cab) planes of the HN(CO)CACB spectrum at the frequency of the selected amide group (in this case its Asn40). Note that the absence of a CB₁ peak indicates that the residue prior to Asn40 is a glycine (which is correct).

VII.3.1.1.4 HNCACB

The HNCACB experiment¹² correlates the amide group chemical shift with both the alpha (C_α) and beta (C_β) carbons of the own and preceding residues by means of the J_{N-H^N} , J_{N-C_α} and $J_{C_\alpha-C_\beta}$ coupling constants, as it is shown in **Figure VII.25**. As the experiment HN(CO)CACB, the HNCACB experiment, besides the chemical shifts of C_α , $C_{\alpha-1}$, C_β and $C_{\beta-1}$ also provides the same useful information about the amino acid type. **Figure VII.26** shows an example of how the CA_i and CB_i resonances are assigned for the protein CtCBM11 (see Chapter II). In combination, the experiments HNCO, HN(CA)CO, HN(CO)CAC and HNCACB can provide complete sequential assignments of the $^1H^N$, ^{15}N , $^{13}C_\alpha$, and $^{13}C_\beta$ and ^{13}CO resonances for proteins up to about 20 kDa. **Figure VII.27** shows the general procedure for the sequential assignment of the protein backbone resonances. This is done by simply linking the correspondent i and $i-1$ resonances. Because it is rare that any ambiguities remain in the backbone assignments after consideration of all the assigned chemical shifts, use of these four experiments makes data analysis so straightforward that the backbone can often be assigned automatically by programs such AutoLink⁹⁶, which was the one I used.

A major limitation of the HNCACB and HN(CO)CACB experiments, however, is that they are relatively insensitive due to fast transverse relaxation rate of $^{13}C_\alpha$.⁷¹

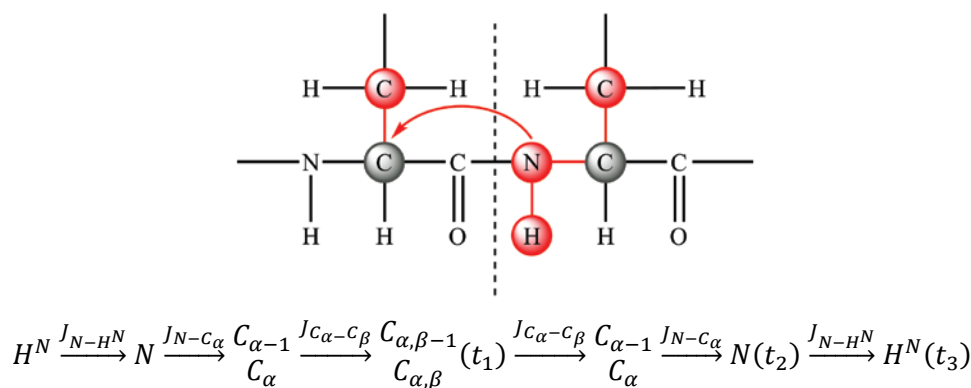


Figure VII.25: The HNCACB magnetization transfer.

Magnetization is passed from $^1H^N$ to ^{15}N with a standard INEPT sequence via J_{N-H^N} coupling. Then, via a second INEPT sequence, the magnetization from the amide ^{15}N is transferred to the $^{13}C_\alpha$ using the J_{N-C_α} coupling constant (red arrow) and from there to the $^{13}C_\beta$ via the $J_{C_\alpha-C_\beta}$. The $^{13}C_{\alpha\beta}$ chemical shift evolves during this period, t_1 . From there the magnetization returns to $^{13}C_\alpha$. From here it is transferred to ^{15}N , where the chemical shifts evolve during t_2 and then to $^1H^N$ for detection with an evolution of the chemical shifts during t_3 . Transfer from $C_{\alpha-1}$ can occur both to $^{15}N_{i-1}$ and $^{15}N_i$. Thus for each NH group there are two C_α and C_β peaks visible. The chemical shift is evolved simultaneously on $^{13}C_\alpha$ and $^{13}C_\beta$, so these appear in one dimension. The chemical shifts evolved in the other two dimensions are ^{15}N and $^1H^N$. In this spectrum the $^{13}C_\alpha$ signal appears with a positive phase while the $^{13}C_\beta$ appears with a negative one.⁷⁹

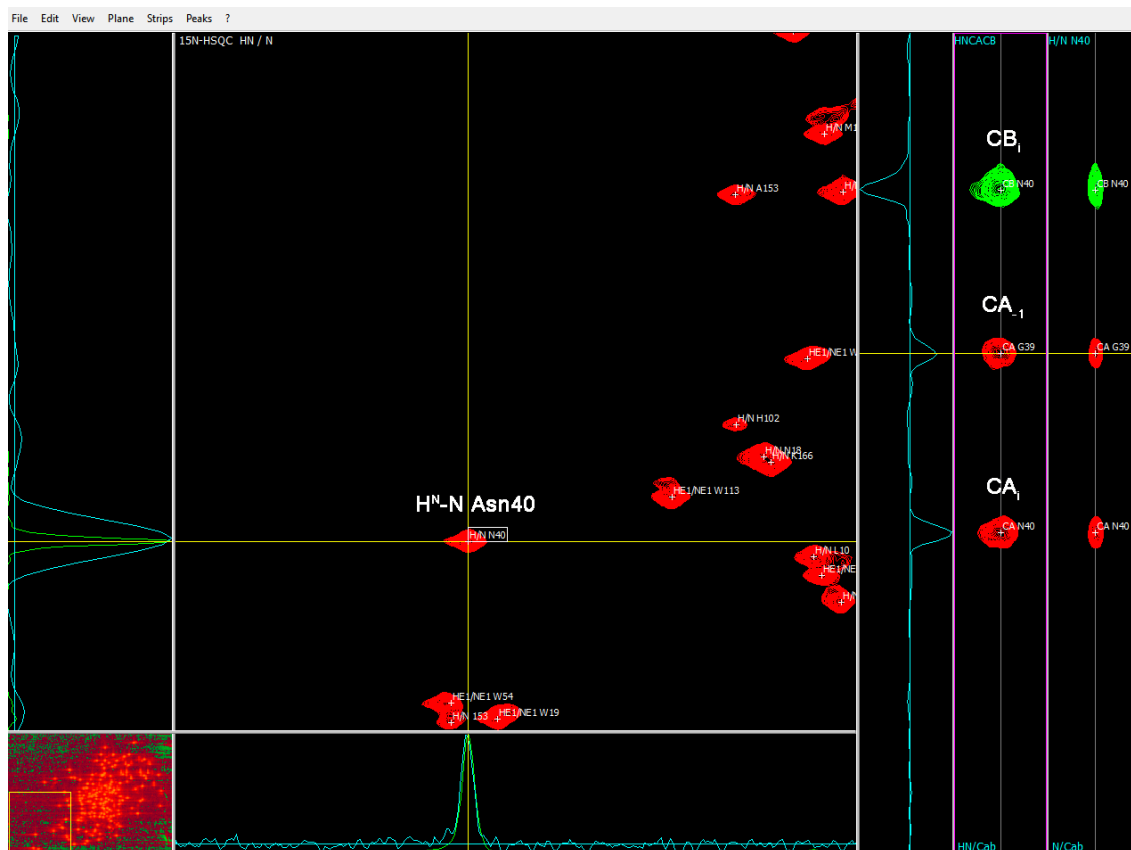


Figure VII.26: Identifying the CA and CB resonances.

In this window, the central panel shows a part of the ^{15}N - ^1H -HSQC spectrum of $C7CBM11$ and the two right strips show $\text{H}^{\text{N}}\text{-}^{13}\text{C}$ (labeled HN/Cab) and $\text{N-}^{13}\text{C}$ (labeled N/Cab) planes of the HNCACB spectrum at the frequency of the selected amide group (in this case its Asn40).

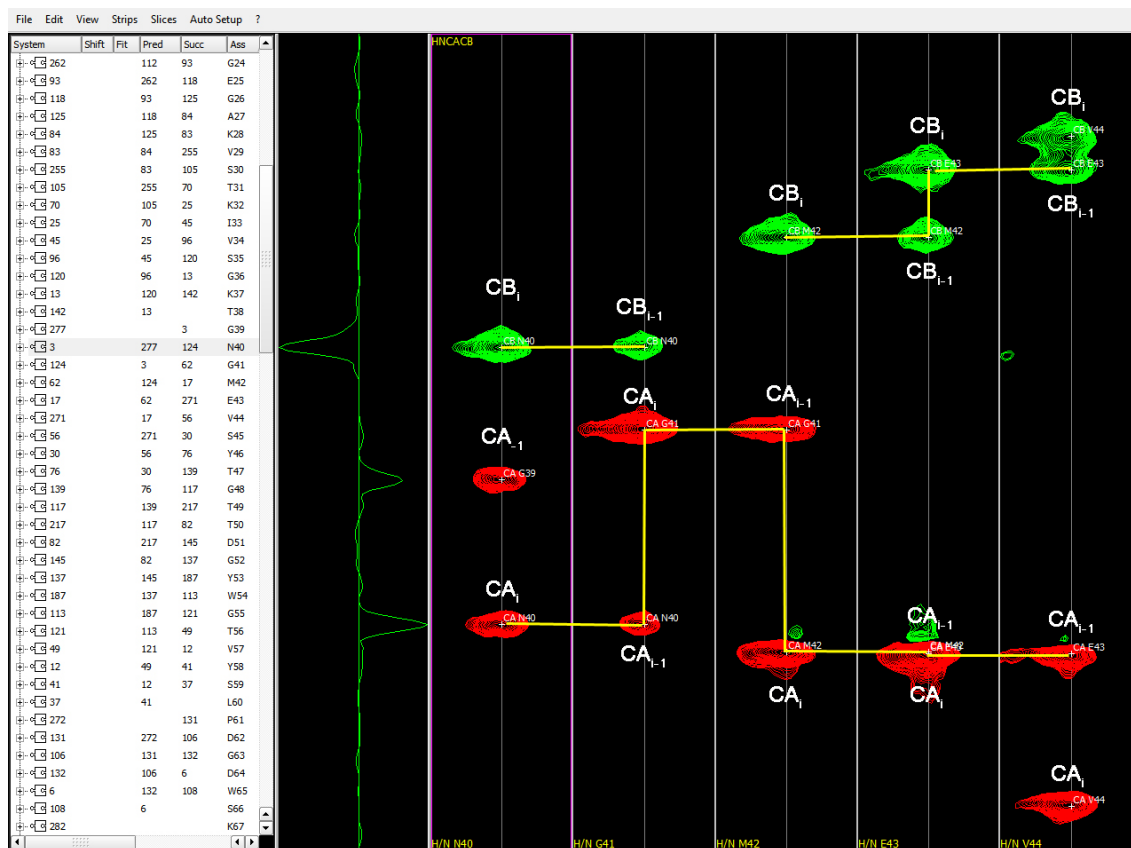


Figure VII.27: Sequential assignment of the protein backbone resonances based on the HNCACB spectrum.

VII.3.1.1.5 Angular restraints

Once the backbone assignments are complete, much useful information is already in hand. Besides information on the secondary structures that can be calculated as explained in Section VII.2.1, dihedral angles can be predicted on the basis of backbone chemical shifts. Programs like TALOS⁺⁹³ are based on torsion angle likelihood obtained from shift and sequence similarity and use a database of proteins for which both chemical shifts and high-resolution X-ray crystal structures are known. The prediction is based on the observation that similar amino acid sequences with similar backbone chemical shifts have similar backbone torsion angles. For each set of three consecutive amino acids in the target protein, the database is searched for the closest matches based on $^1\text{H}_\alpha$, $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ and ^{13}CO chemical shifts and sequence similarity. The torsion angles for the central residue from the best 10 matches are chosen as the predicted torsion angles for the residue, which are used as backbone dihedral angles in the structure calculation. The error in TALOS predictions is around 3%, however mistakes can be identified during structure calculations by the inconsistency of a constraint with NOEs or other types of data.

VII.3.1.2 Experiments for side-chain assignments

VII.3.1.2.1 (H)CCH-TOCSY

The (H)CCH-TOCSY experiment^{97,98} correlates side-chain aliphatic proton and ¹³C resonances within the spin system via J_{C-H} and J_{C-C} coupling constants (**Figure VII.28**). This experiment provides practically the complete assignments of all aliphatic ¹H and ¹³C resonances, with the exception of some resonances of the long aliphatic side chains (as lysine or arginine) for which substantial overlap may remain. In (H)CCH-TOCSY there will be two carbon dimension and a third proton dimension.

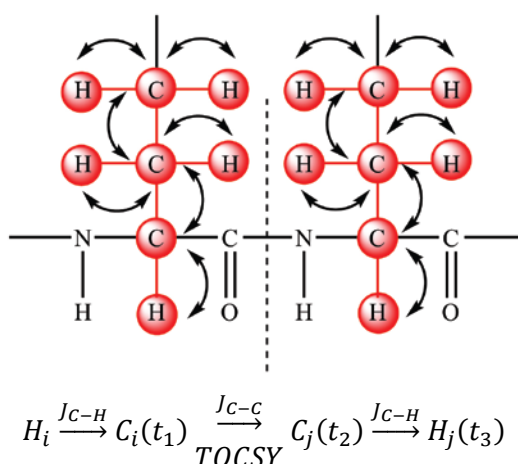


Figure VII.28: The (H)CCH-TOCSY magnetization transfer.

Magnetization is transferred from the side-chain ¹H nuclei to their attached ¹³C nuclei via the J_{C-H} coupling constant. After a ¹³C chemical shift evolution period, t_1 , there is an isotropic ¹³C mixing period that transfers magnetization along the ¹³C side chain via J_{C-C} . The ¹³C chemical shift evolves during t_2 and is transferred back to the side-chain hydrogen atoms for detection.⁷³ Black arrows indicate INEPT transfers.

When acquiring this type of experiment (or almost every heteronuclear two and three-dimensional experiment) it will be necessary to use composite pulse decoupling (CPD). These pulses are used for broadband decoupling of ¹H during acquisition or for spin-locking. In the (H)CCH-TOCSY experiment for instance, when the coherence is transferred from the ¹H to a directly bonded ¹³C (t_1), a CPD sequence is applied that spin-locks the appropriate ¹³C spins.^{73,99} The need for these pulses represents a problem when going to higher magnetic fields as the requirement of shorter pulses at high powers in order to obtain the needed excitation profile may result in **sample heating**, particularly for samples that are at high ionic strength. Furthermore, off-resonance effects also become more serious. For ¹H, because its chemical shift range is fairly narrow, the correct bandwidth profile is relatively easy to achieve, and the major concern

is to insure that only as much power is applied as is required. For ^{13}C decoupling things get more complicated as ^{13}C chemical-shift ranges are large and increase with field. Because of this, simple rectangular pulse decoupling can be problematic at higher fields.⁹⁹

VII.3.1.2.2 HNHA

Even though the TOCSY experiment can identify all of the protons of a spin-system, it cannot automatically differentiate between the types of proton (i.e. H_α , H_β , H_γ ...), an important consideration for amino acid the spin-system assignment. H_α protons will give stronger crosspeaks, but the actual intensity of the crosspeaks will depend on the individual J-couplings throughout the residue. These protons can be unambiguously identified in the HNHA experiment.^{71,87,93,99} The HNHA experiment correlates the amide group chemical shift with alpha proton (H_α) via the $J_{\text{N}-\text{H}^{\text{N}}}$ and $J_{\text{N}-\text{H}_\alpha}$ coupling constants, as shown in **Figure VII.29**.

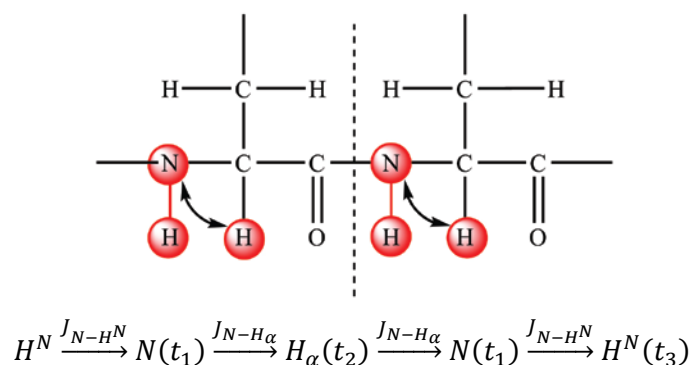


Figure VII.29: The HNHA magnetization transfer.

Magnetization is transferred from $^1\text{H}^{\text{N}}$ to ^{15}N creating zero- and double quantum coherence. In addition, the transverse $^1\text{H}^{\text{N}}$ magnetization dephases due to homonuclear $^1J_{\text{N}-\text{H}_\alpha}$ coupling. Chemical shift evolution of the ^{15}N spins occurs during t_1 in a constant manner. The magnetization is then transferred to the H_α where it evolves during t_2 . During the following rephasing period, the ^{15}N chemical shift evolution is continued for an additional period (t_1) and then converted back to observable $^1\text{H}^{\text{N}}$ magnetization. Black arrows indicate an INEPT transfer.

This experiment deals with the overlap problems by spreading the signals to an additional dimension according to the ^{15}N -frequency. Furthermore, it allows the determination of the coupling constant, J from the ratio between the intensities of the diagonal and cross-peak.⁸⁷

$$\frac{I_{\text{cross}}}{I_{\text{diag}}} = -\tan^2(\pi J 2\zeta)$$

where I_{cross} and I_{diag} are the intensities of the cross- and diagonal peaks and ζ denotes the time allowed for the transfer of magnetization between the two protons.

The relationship between the observed coupling constant and the peptide φ angle is given by the Karplus relationship (**Equation VII.1**). As it was said above, the ${}^3J_{HN-H\alpha}$ J coupling constants are an important source of information on the secondary structure and improve convergence and accuracy of the structure calculation. Nevertheless, accurate determination of the ${}^3J_{HN-H\alpha}$ couplings is complicated by their small size relative to the natural proton line width. A direct measurement is possible only for very small peptides.^{71,87}

VII.3.1.3 Experiments for NOE measurement

VII.3.1.3.1 ${}^{15}\text{N}/{}^{13}\text{C}$ -NOESY-HSQC

Chemical shifts should carry enough information to determine protein structures at high resolution.⁶²⁻⁶⁴ In fact, there are already some tools that allow precisely that, namely: SHIFTX¹⁰⁰ (<http://shiftx.wishartlab.com/>), CS23D¹⁰¹ (<http://www.cs23d.ca/>) and CS-ROSETTA^{102,103} (<http://spin.niddk.nih.gov/bax/software/CSROSETTA/>). Nevertheless, these tools are not completely developed yet and are most useful to refine structures (e.g. secondary structure and dihedral angles) rather than calculate them. For NMR-based structure determination, the most important parameters are still the **${}^1\text{H}$ - ${}^1\text{H}$ distances** derived from NOE intensities, and **dihedral angles** which are obtained from 3J coupling constants (that can be obtained from the HNHA experiment and/or from the ${}^1\text{H}_\alpha$, ${}^{13}\text{C}_\alpha$, ${}^{13}\text{C}_\beta$ and ${}^{13}\text{CO}$ chemical shifts as seen above).

The 3D HSQC-NOESY experiment^{104,105} is specifically designed to obtain X-edited ($X = {}^{15}\text{N}$ or ${}^{13}\text{C}$) NOESY spectra of labeled biomolecules from which homonuclear ${}^1\text{H}$ - ${}^1\text{H}$ NOEs can be clearly assigned even in overcrowded regions. The mechanism involves a ${}^1\text{H}$ - ${}^1\text{H}$ NOE step and heteronuclear transfer via J_{X-H} coupling (**Figure VII.30**).

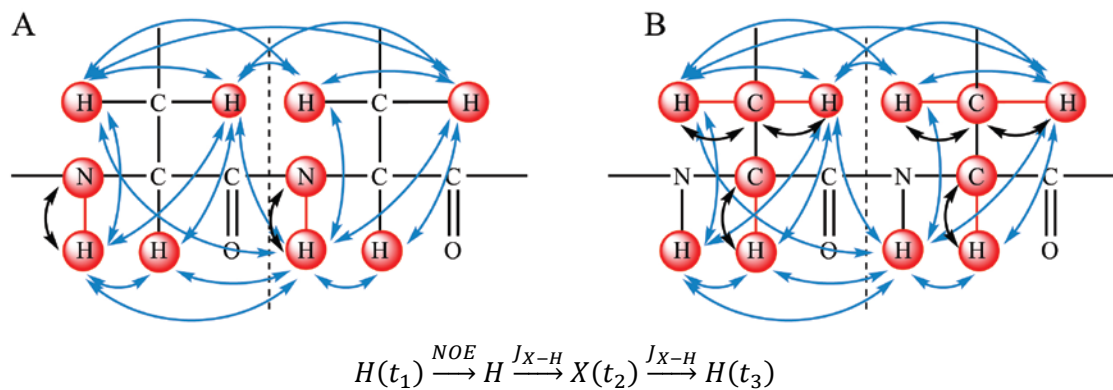


Figure VII.30: The ^{15}N - ^1H -HSQC-NOESY (A) and ^{13}C - ^1H -HSQC-NOESY (B) magnetization transfer.

First, all ^1H are excited and their chemical shift is labeled in t_1 evolution. After the evolution of ^1H chemical shifts, the magnetization is transferred to vicinal protons by cross relaxation (NOE) during the NOESY mixing period. The magnetization is then transferred to the X nuclei with a standard INEPT sequence via J_{X-H} coupling. The chemical shift of the X nuclei is labeled during t_2 . Finally, through a reverse INEPT sequence, the magnetization returns to the ^1H nuclei for detection. In the ^{13}C - edited HSQC-NOESY the transfer either occurs to/from the aliphatic ^{13}C nuclei or to/from the aromatic ^{13}C nuclei (but not both) depending on the ^{13}C frequency used during the pulse sequence.⁷⁹ Black arrows indicate an INEPT transfer and blue arrows indicate a NOE transfer.

VII.3.1.3.2 Distance restraints

After assigning the peaks in the NOESY spectra, distance restraints can be extracted. In order to do this, the first step is to integrate the peaks in the NOESY spectra in order to get their intensities. The intensity of a NOESY peak is proportional to the distance to the minus 6th power, so the distance is determined according to intensity of the peak according to **Equation VII.23**:

$$d_i = d_{ref} \left(\frac{I_{ref}}{I_i} \right)^{-6}$$

VII.23

were d_{ref} and d_i are the inter proton distances and I_{ref} and I_i are the cross-peak intensities, for a reference and observed cross-peak, respectively. However, this relation is only valid for short mixing times (assuming that dipole-dipole interaction is the only mechanism for cross-relaxation) as for longer ones the intensities of NOESY cross-peaks are no longer directly proportional to the cross-relaxation rate constants between the interacting spins, as seen in Section VII.2.4, **Figure VII.15**.⁷¹ As a consequence precise ^1H - ^1H separations cannot be determined and NOE cross-peaks are commonly grouped on the basis of their intensities into three categories - strong, medium, and weak. Each category is associated with an upper bound (upper limit - **UPL**) separation between the interacting spins given by the volume of the cross

peak. The lower bound (lower limit - **LOL**) separations for pairs of protons are set to the sum of the van der Waals radii ($\sim 1.8\text{\AA}$).

In my case the volumes were converted into upper limits (UPLs) by CYANA2.1⁹² using the macro *calibrate*. This macro uses a center averaging protocol for comparing the distances between atoms. In center averaging pseudoatoms are created at the mean position of the protons involved. The distance in the evolving structure is calculated from this pseudoatom to the other proton of interest.¹⁰⁶ A pseudoatom correction has then to be applied to the upper limit of the distance restraint involving the pseudoatom:

- Multiplicity correction is applied by dividing the peak volume by the numbers of ^1H spins in pseudoatoms assigned to the peak. For instance, the volume of a cross peak between a Leu QQD pseudo atom and a Tyr QD pseudo atom is divided by a factor of $6 \times 2 = 12$ prior to applying the calibration function. The resulting UPL is subject to the upper and lower cutoffs.
- Distance correction is applied by adding a distance between the pseudoatom and its constituent spins. It is applied after the application of upper and lower cutoffs. For example, for a Tyr QD pseudoatom this correction is equal to half the distance between the HD1 and HD2 spins.

VII.3.2 Structure validation

A fundamental aspect of any structure determination is its final structure quality. This can be measured essentially by two parameters: the average target function and the final positional uncertainty in the molecular coordinates (rmsd).

The CYANA target function is defined such that it is zero if and only if all experimental distance constraints and torsion angle constraints are fulfilled and all non-bonded atom pairs satisfy a check for the absence of steric overlap¹⁰⁷. Therefore, the objective of refinement is to lower this value as much as possible.

Regarding the rmsd, a low value indicates that the calculated structures are close to the average structure, which represents a high precision of the structure calculation. The precision of NMR structures is related to the precision of the experimental data. Errors in the measurements will affect the precision in the estimation of distance and angular restraints derived from the data. In general, an increase in the number of experimental restraints will improve the precision of the calculated structures. However, the precision of the structure determination does not guarantee the accuracy of the NMR structures.¹⁰⁸ For instance, if the distances derived from NOE are wrongly calibrated, the calculated structures will be

significantly different from the structures that would be obtained with the correct distance restraints. Therefore, the accuracy of NMR structures is required to be examined with additional criteria. It is thought that an accurate structure should not have substantial violations in Ramachandran diagrams¹⁰⁹ and covalent bond geometry. Programs such as PROCHECK¹¹⁰ and WHATIF¹¹¹, have been developed for checking the values of bond lengths and angles, the number and scale of violations of experimental restraints, potential energy, and other parameters¹⁰⁸. Structures with poor scores do not necessarily indicate errors in the structure, but they require attention to locate possible miss-assigned experimental data. On the other hand, structures with high scores also do not assure the accuracy of the calculation.¹⁰⁸ In general, a high resolution structure will have:

- backbone rmsd $\leq \sim 0.8$ Å, heavy atom rmsd $\leq \sim 1.5$ Å;
- low rmsd from restraints (good agreement with restraints – target function $< \sim 10$ Å²)
- good stereochemical quality:
 - ideally $> 90\%$ of residues in most favorable regions of Ramachandran plot
 - very few “unusual” side chain angles and rotamers
 - low deviations from idealized covalent geometry

VII.4 Protein dynamics by NMR

Proteins are not static objects but rather highly dynamic entities whose motion range varies from very fast fluctuations, normally associated with individual atoms (on the picosecond timescale) and/or loop and domain motions (on the nanosecond timescale), to conformational exchange or rearrangements (on the millisecond to second (or even hours or days) timescale). These motions are involved in several key functions such as catalysis and ligand recognition and binding^{28,30,32,112} (**Figure VII.31**) and give rise to the conformational ensemble that characterizes protein structure by NMR. Consequently, the classical “lock-and-key” model for molecular interaction is incorrect.³¹

Despite the several techniques available for detailed characterization of molecular motions (NMR, simulation, temperature jump, stop flow, fluorescent microscopy), NMR spectroscopy is the leading tool due to its versatility and precision. From NMR relaxation experiments it is possible to extract the frequency of the motion (i.e. how fast the motion is – correlation time, τ) and the amplitude of the motion (i.e. how far the atoms move from an average position – order parameter, S^2). Moreover, it is possible to characterize the different motions that a certain atom undergoes just by changing the NMR experiment, thus allowing a complete motional characterization of the system in a per-residues basis.³¹⁻³³

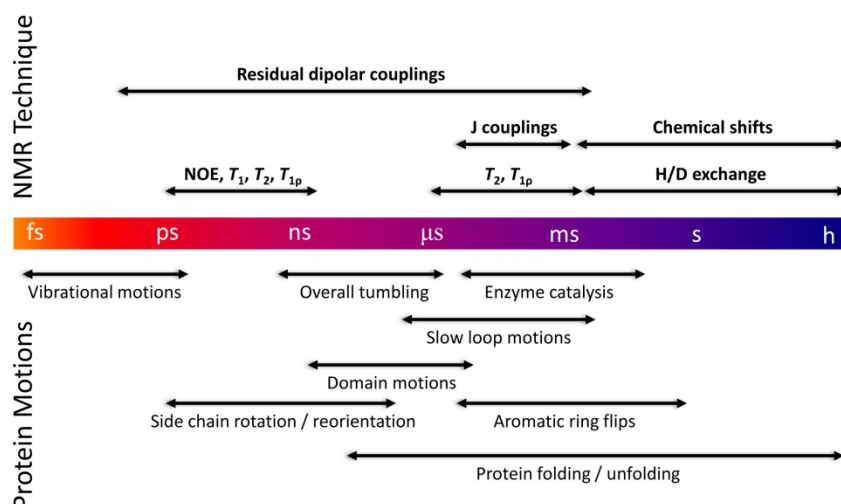


Figure VII.31: Protein motion time scales and NMR techniques used to study each time scale.

VII.4.1 Theory of spin relaxation in proteins

Each observable process in NMR involves transitions between magnetic quantized energy levels. Such transitions are stimulated by magnetic fields that oscillate at the transition frequencies. Thus, the relaxation rates are determined by the probability that the relevant nuclei experience appropriate oscillating magnetic fields. In proteins these fields result from the movements of magnetic nuclei relative to each other or relative to the overall permanent field of the NMR magnet.³² As a result, relaxation is exquisitely sensitive to molecular motion.

Spin relaxation can occur mainly by two mechanisms: either by T_1 relaxation or by T_2 relaxation (see Sections VII.2.2.2 and VII.2.2.3, respectively). The rates by which these phenomena happen R_1 ($R_1=1/T_1$) and R_2 ($R_2=1/T_2$) contain information on the pico to nanosecond time scale and are affected primarily by dipole-dipole interactions and chemical shift anisotropy, CSA (see Sections VII.2.2.4 and VII.2.2.5, respectively).³²

For the study of protein backbone dynamics the ^{15}N nucleus is of particular interest. For an isolated NH spin system, the relaxation rate constants of the ^{15}N spin caused by the dipolar interaction of the ^{15}N spin with the ^1H spin and by the magnetic shielding arising from the CSA interaction are given by⁷³:

$$R_1 = R_1^D + R_1^{CSA} = \frac{d^2}{4} [6J(\omega_H + \omega_N) + J(\omega_H - \omega_N) + 3J(\omega_N)] + c^2J(\omega_N)$$

VII.24

$$R_2 = R_2^D + R_2^{CSA}$$

$$= \frac{d^2}{8} [6J(\omega_H + \omega_N) + 6J(\omega_H) + J(\omega_H - \omega_N) + 3J(\omega_N) + 4J(0)]$$

$$+ \frac{c^2}{6} [3J(\omega_N) + 4J(0)] + R_{ex}$$

VII.25

$$\sigma_{NH} = \frac{d^2}{4} [6J(\omega_H + \omega_N) + J(\omega_H - \omega_N)] = R_1(NOE - 1) \frac{\gamma_N}{\gamma_H}$$

VII.26

with:

$$d = \frac{\mu_0 h \gamma_N \gamma_H r^{-3}}{8\pi^2} \approx -7.21 \times 10^4; \quad c = \frac{\omega_N \Delta\sigma}{\sqrt{3}} \approx -3.53 \times 10^4; \quad \mu_0 = 4\pi \times 10^{-7} TmA; \quad r = 102 \text{ nm};$$

$$\Delta\sigma = -170 \text{ ppm}; \quad h = 6.626 \times 10^{-34} J.Hz^{-1}; \quad \gamma_H = 2.6752 \times 10^8 \text{ MHz}.T^{-1};$$

$$\gamma_N = -2.712 \times 10^7 \text{ MHz}.T^{-1}; \quad \omega_N = \gamma_N B_0; \quad \omega_H = \gamma_H B_0;$$

where: R_1 , R_2 and σ_{NH} are the spin-lattice, spin-spin relaxation and cross-relaxation rates, respectively. NOE is the resonance line intensity change caused by dipolar cross-relaxation from neighboring spins with the perturbed energy level populations. All are dependent on the spectral density functions (remember **Equation VII.15**) evaluated at five frequencies ($\omega_H + \omega_N$, ω_H , $\omega_H - \omega_N$, ω_N , and 0). μ_0 is the permeability of vacuum; h is Planck's constant; r_{NH} is the NH bond length; γ_N and γ_H are the gyromagnetic ratios; c is the nitrogen chemical shift anisotropy (CSA) with the assumption that the chemical shift tensor is axially symmetrical, which has been demonstrated to be valid for the peptide bond ^{15}N with $\Delta\sigma = -160 \sim -170 \text{ ppm}$ ¹¹³. It should be kept in mind though that, as shown in **Equation VII.26**, conformational exchange on a μs to ms time scale leads to a modulation of the chemical shift of the affected nuclei, resulting in an increased contribution (R_{ex}) to the effective R_2 transverse relaxation rate.

The R_1 and R_2 rate constants can be determined experimentally whereas σ_{NH} is determined from the steady-state $\{^1\text{H}\}$ - ^{15}N -NOE via **Equation VII.26**¹¹⁴. The parameters R_1 and R_2 are sensitive to different motional frequencies: R_1 values provide information about motional properties with a frequency of approximately 10^8 - 10^{12} s^{-1} , whereas R_2 values, in addition to depending on motions occurring at these frequencies, are also sensitive to dynamics on the micro-millisecond time scale. Hence, by measuring both R_1 and R_2 , it is feasible to obtain dynamic information over a large motional range.¹¹⁵ $\{^1\text{H}\}$ - ^{15}N -NOE relaxation data is highly sensitive to motions of the polypeptide backbone on a pico to nanosecond time scale. NOE values smaller than 0.65 indicate large amplitude backbone fluctuations^{32,116}.

VII.4.2 Protein motions and relaxation

VII.4.2.1 Reduced spectral density mapping

As seen from **Equations VII.24, VII.25 and VII.26**, the several relaxation rates are dependent on the spectral density functions evaluated at five frequencies ($\omega_H + \omega_N$, ω_H , $\omega_H - \omega_N$, ω_N , and 0).^{117,118} Without any assumptions, the spectral density functions at these five frequencies cannot be determined from the three experimentally determined relaxation rate constants by measuring T_1 , T_2 , and NOE. However, because the spectral density function at high frequencies does not fluctuate much, the high frequency terms ($J(\omega_H + \omega_N)$ and $J(\omega_H - \omega_N)$) can be replaced by a single average term $(0.87\omega_H)^{119}$, thus enabling the mapping of the spectral density functions using only the R_1 , R_2 and σ_{NH} relaxation rates:

$$\begin{bmatrix} J(0) \\ J(\omega_N) \\ J(0.87\omega_H) \end{bmatrix} = \begin{bmatrix} \frac{-3}{4(3d^2 + c^2)} & \frac{3}{2(3d^2 + c^2)} & \frac{-9}{10(3d^2 + c^2)} \\ \frac{1}{(3d^2 + c^2)} & 0 & \frac{-7}{5(3d^2 + c^2)} \\ 0 & 0 & \frac{1}{5d^2} \end{bmatrix} \times \begin{bmatrix} R_1 \\ R_2 \\ \sigma_{NH} \end{bmatrix}$$

VII.27

This approach is referred to as reduced spectral density mapping¹¹⁹. Given these equations, the following points are noteworthy: i) because only $J(0)$ depends on the transverse relaxation rate, contributions of μ s-ms time-scale exchange processes will cause an increase of only $J(0)$; ii) fast local motions on the ps-ns timescale will be reflected in a decrease of mainly $J(0)$ with a corresponding increase of $J(0.87\omega_H)$; iii) anisotropic rotational diffusion will lead to fluctuations in the spectral density at all three frequencies, however, due to its low value the effect on $J(0.87\omega_H)$ is less pronounced; iv) increased motions close to either the ^{15}N or ^1H Larmor frequency enhance R_1 relaxation of the ^{15}N nucleus, although R_1 is more sensitive to changes in $J(\omega_N)$.^{32,120}

An important warning concerning the reduced spectral density approach involves the influence of slow conformational exchange on transverse ^{15}N (or ^{13}C) relaxation. If conformational exchange is present, measured R_2 values will contain contributions from both $J(0)$ and $J(\omega_N)$ and the exchange broadening term R_{ex} . Consequently, if exchange broadening is assumed to be absent, the calculated $J(\omega)$ values will be incorrect. In particular, $J(0)$ values will be overestimated. To separate these two effects, relaxation data can be measured at multiple magnetic field strengths. In the absence of such data, $J(0)$ should be interpreted as representing a combination of slow motions (such as molecular tumbling) and conformational exchange.³²

VII.4.2.2 Rotational diffusion tensor

For NH bond vectors subject only to low-amplitude and fast intramolecular motions, the ratio between the transversal and longitudinal ^{15}N relaxation rates (R_2/R_1) is approximately independent of intramolecular dynamics and only depends on the rotational diffusion of the protein.^{116,120} Thus, for a ^{15}N - ^1H vector for a spherical molecule with radius r that tumbles in a solution of viscosity η , the correlation function can be assumed to be a simple exponential that decays with the rotational correlation time τ_m , so that the spectral density function is given by:

$$J(\omega) = \frac{2}{5} \frac{\tau_c}{1 + (\omega\tau_c)^2} \quad \text{with} \quad \tau_c = \frac{1}{6D_{iso}} \quad \text{VII.28}$$

where D_{iso} is the isotropic rotational diffusion constant, given by:

$$D_{iso} = \frac{kT}{8\pi\eta r^3} \quad \text{VII.29}$$

By rewriting **Equations VII.24, VII.25 and VII.26** using a fixed value for the ^{15}N chemical shift anisotropy, $\Delta\sigma_N$ (-170 ppm) and NH bond length, r_{NH} (1.02 Å) it is possible to extract the isotropic overall rotational correlation time, τ_m from the R_2/R_1 ratios^{32,116,120}.

Nonetheless, proteins cannot always be described as spherical entities, but rather as an asymmetric top with an anisotropic tensor. In this case the spectral density function for an amide vector is given by:

$$J_i(\omega) = \frac{2}{5} \sum_{j=1}^5 A_{ij} \frac{\tau_j}{1 + (\omega\tau_j)^2} \quad \text{VII.30}$$

where τ_j are the time constants that depend on the diffusion constants D_{xx} , D_{yy} and D_{zz} and the coefficients A_{ij} are functions of the diffusion constants and the angles θ and φ that define the orientation of the amide vector with respect to the rotational diffusion tensor (**Figure VII.32**).¹²⁰

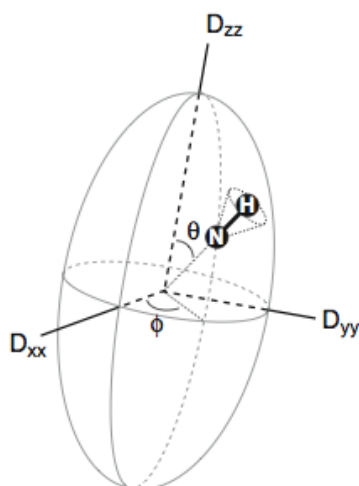


Figure VII.32: Representation of an amide vector in a protein.¹²⁰

For this hypothetical protein the rotational diffusion behavior can be described by a rod-shaped diffusion tensor with tensor components D_{xx} , D_{yy} and D_{zz} . The orientation of the amide bond with respect to the diffusion tensor is defined by the angles θ and φ .

In absence of a proper structural model the diffusion tensor components D_{xx} , D_{yy} and D_{zz} can be estimated from the distribution of R_2/R_1 ratios¹¹⁶. However, for an accurate value, residues with large amplitude and fast internal motions have to be excluded from the calculation ($\text{NOE} < 0.65$). Among the remaining residues, those with significant conformational exchange on the microsecond/millisecond time scale have also to be excluded according to the following condition¹¹⁶:

$$\frac{\langle T_2 \rangle - T_{2,n}}{\langle T_2 \rangle} - \frac{\langle T_1 \rangle - T_{1,n}}{\langle T_1 \rangle} > 1.5 \times SD$$

VII.31

where $\langle T_2 \rangle$ and $\langle T_1 \rangle$ are the average values of T_2 and T_1 , respectively, $T_{2,n}$ and $T_{1,n}$ are the T_2 and T_1 values of residue n , respectively. SD is the standard deviation of **Equation VII.31**.

When an accurate 3D structure exists, the anisotropic diffusion tensor can be determined from a subset of R_2/R_1 ratios¹¹⁶, using for example the program Tensor2¹²¹. An accurate description of anisotropic diffusion of a protein in solution can be obtained from a full hydrodynamic analysis as performed by the program HYDRONMR¹²², where a bead shell model of the 3D protein structure is used (*see Chapter III*).

VII.4.2.3 The Lipari-Szabo Model-free Formalism

The spectral density functions shown above (**Equation VII.27**) give us probabilities with which a bond vector is oscillating at each specified frequency. However, they do not directly indicate whether these oscillations are associated with global molecular rotation or the local internal motions affecting the bond vector. In order to gain a detailed understanding of the protein dynamics it is fundamental to distinguish between internal dynamics and global motions. In 1982 Lipari and Szabo developed a method for characterizing fast motions (ps-ns), the so-called “model-free” formalism.^{123,124}

Rather than fitting the experimental data to any specific physical models, Lipari and Szabo showed that the fast motion of atoms is easily described by three parameters: 1) a global rotational correlation time, τ_m , which describes the overall tumbling of the molecule; 2) a local correlation time, τ_e , which describes any ps-ns motion present at a specific location, and 3) an order parameter, S^2 , which describes the amplitude and rate of internal dynamics for individual chemical bond vectors (e.g., peptide NH bonds) giving us the percentage of motion coming from the global tumbling compared to the local motion.^{73,125} $1-S^2$ gives the percentage derived from the local fluctuations. The order parameter can have a value between 0 and 1, in which lower values indicate larger amplitudes of internal motions.

Because the values of these three parameters do not depend on a model, this approach was named “model-free” and, 30 years after being developed it is still one of the most widely used methods for the relaxation data analysis of proteins. The basic idea of the model-free formalism is that the internal motions of bond vectors in proteins are independent of the overall rotational diffusion of the molecule as a whole. In addition, the rotational diffusion of the molecule influences each bond vector identically (for isotropic rotation) or in a manner that is related through the relative orientations of the bond vectors in the molecule (for non-isotropic rotation), whereas the internal motions of any two bond vectors are independent of each other or at least unrelated in any predictable way. In this conditions, the simple model-free equation gives $J(\omega)$ as a the sum of two Lorentzian functions:

$$J(\omega) = \frac{2}{5} \left(S^2 \frac{\tau_m}{1 + (\omega\tau_m)^2} + (1 - S^2) \frac{\tau}{1 + (\omega\tau)^2} \right)$$

VII.32

$$\tau = \left(\frac{1}{\tau_m} + \frac{1}{\tau_e} \right)^{-1}$$

VII.33

where τ_m is the overall rotational correlation time, S^2 is the generalized order parameter and τ_e is the local correlation time which is related to τ through **Equation VII.33**.

The generalized order parameter S^2 measures the degree of spatial restriction of the bond vector in a molecular frame, providing information about the angular amplitude of the internal motions of bond vectors. If the bond vector diffuses in a cone with an angle θ defined by the diffusion tensor and the equilibrium orientation of the bond vector, S^2 is highly sensitive to the cone angle in the range from 0° to 75° and decreases dramatically as the cone angle increases⁷³. The value of θ may vary from 1 when the bond is rigid to 0 when the internal motion is completely isotropic (**Figure VII.33**).

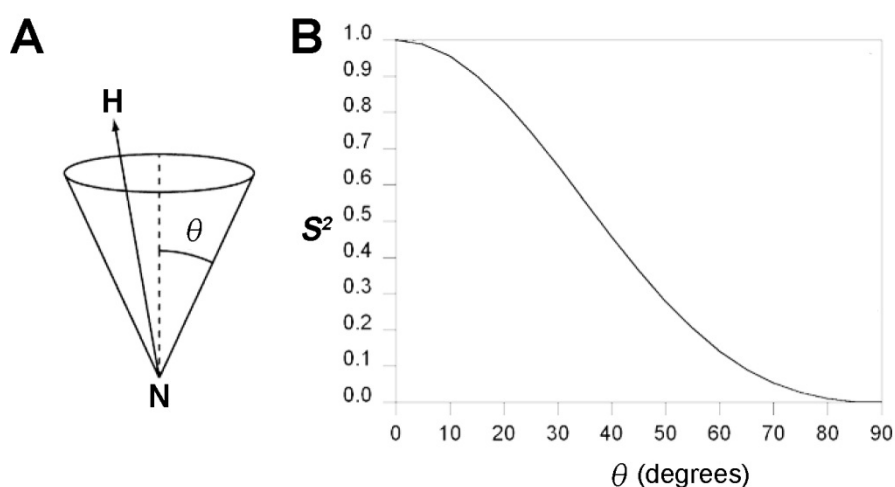


Figure VII.33: interpretation of the generalized order parameter, S^2 , in a diffusion-in-a-cone model.

A) The N-H bond vector is assumed to diffuse freely within a cone defined by semi-angle θ ; **B)** relationships of the generalized order parameter (S^2) to the cone semi-angle (θ). Adapted from Jarymowycz and Stone, 2006.³²

In **Table VII.5** there are listed the five commonly used models for the spectral density function used to analyze ^{15}N relaxation data using the model-free approach. If τ_e is small ($\tau_e \ll \tau_m$), the dynamics can be described entirely by S^2 (**Table VII.5** – Model 1). In the presence of slow motion events (ms- μ s) Models 1 and 2 (**Table VII.5**) can be extended by a chemical exchange factor or line-broadening term, R_{ex} , (Models 3 and 4). These models can be used for residues that have high R_2 relaxation rates due to a possible contribution of μ s-ms conformational exchange (R_{ex}). An extended form of the model-free spectral density function has been developed¹²⁶ to describe internal motions that take place on two distinct time scales, τ_f and τ_s (**Table VII.5** – Model 5). In this model, it is assumed that the contribution of the faster of the two motions can be neglected ($\tau_f \approx 0$). Therefore, while the faster motion contributes to the overall S^2 ($S^2 = S_f^2 S_s^2$), the term containing the fast effective correlation time τ_f is left out.

The spectral density functions in **Table VII.5** all assume isotropic rotational diffusion, but can be extended to allow for axial or complete anisotropic rotational diffusion¹²⁷, as is for example included in the program Tensor2¹²¹.

Table VII.5: Different models that can be used in a model-free analysis of relaxation rates.³²

	Model	$J(\omega)$	Parameters	Assumptions
1	Simplified model-free (with isotropic tumbling)	$\frac{2}{5} \left(\frac{S^2 \tau_m}{1 + (\omega \tau_m)^2} \right)$	S^2	$\tau_e \ll \tau_m$ $R_{ex} \approx 0$
2	Original model-free (slow isotropic tumbling with faster, spatially restricted internal motions)	$\frac{2}{5} \left(\frac{S^2 \tau_m}{1 + (\omega \tau_m)^2} + \frac{(1 - S^2) \tau}{1 + (\omega \tau)^2} \right)$	S^2, τ_e	$\tau_e < 500$ ps $R_{ex} \approx 0$
3	Like 1 plus conformational exchange term, R_{ex}	$\frac{2}{5} \left(\frac{S^2 \tau_m}{1 + (\omega \tau_m)^2} \right)$	S^2, R_{ex}	$\tau_e \ll \tau_m$
4	Like 2 plus conformational exchange term, R_{ex}	$\frac{2}{5} \left(\frac{S^2 \tau_m}{1 + (\omega \tau_m)^2} + \frac{(1 - S^2) \tau}{1 + (\omega \tau)^2} \right)$	S^2, τ_e, R_{ex}	$\tau_e < 500$ ps
5	Extended model-free (two time scales of internal motion with isotropic tumbling)	$\frac{2}{5} \left(\frac{S^2 \tau_m}{1 + (\omega \tau_m)^2} + \frac{(S_f^2 - S^2) \tau_s}{1 + (\omega \tau_s)^2} \right)$	S_s^2, S_f^2, τ_s $\tau^{-1} = \tau_m^{-1} + \tau_s^{-1}$	$\tau_f \ll \tau_m$ $\tau_s \geq 500$ ps $R_{ex} \approx 0$

VII.4.2.3.1 Relationship between the generalized order parameter, S^2 , and conformational entropy, ΔS_{conf}

The generalized order parameter as calculated with the above equations can be associated with the apparent entropy of the bond vector reorientation. Backbone or side chain flexibility can either decrease or increase upon binding. Decreases are often associated with ‘enthalpy-entropy compensation’ and ‘induced fit’, whereas increased flexibility leads to an entropic stabilization of the complex.¹²⁰ In order to relate NMR derived order parameters with conformational entropy, a proper model that describes the motional behavior needs to be chosen. Three groups have independently developed methods for accomplishing this goal¹²⁸⁻¹³⁰ whose main difference is the partition function. The most frequently employed method is based on the diffusion-in-a-cone model¹³⁰, according to which, the change in conformational entropy (ΔS_{conf}) can be calculated from the order parameters in final (F) and initial (B) states as given by

Equation VII.34:

$$\Delta S_{conf} = k \sum_{j=1}^N \ln \left[\frac{3 - (1 + 8S_{j,final})^{1/2}}{3 - (1 + 8S_{j,initial})^{1/2}} \right]$$

VII.34

where ΔS_{conf} is the change in conformational entropy, k is the Boltzmann constant and S_j is the order parameter for the residue j in the final ($S_{j,final}$) and initial state ($S_{j,initial}$).

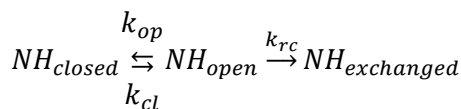
Despite the attractiveness of this simple approach, it presents several limitations^{32,131,132}:

- i. NMR relaxation measurements are limited to only a subset of vectors within the protein (e.g. only backbone amide bond vectors or side chain methyl axes);
- ii. The three NMR relaxation parameters, R_1 , R_2 and NOE, used to extract S^2 , are generally not sensitive to rotational motions slower than molecular diffusion (a few nanoseconds); although R_2 can be influenced by microsecond to millisecond time scale conformational exchange, the order parameter does not reflect these motions;
- iii. The order parameter is only sensitive to motions that reorient the bond vector involved;
- iv. Possible correlations between motions of different bond vectors are not taken into account.

However, mainly the fast motions contribute to conformational entropy and even the slowest vibrational modes of proteins tend to fall within the ps-ns time window, which implies that limitation (ii) may not be severe. Additionally, caveats (i) to (iii) tend to result in a reduction of the conformational entropy while caveat (iv) results in an increase in the estimated entropy, thus, there may be some cancelation between these systematic errors. Furthermore, the agreement between entropy contributions estimated based on calorimetric measurements and NMR derived conformational entropy values, supports the validity of the above-mentioned approach.¹³³

VII.4.2.4 Amide proton exchange

Because non-hydrogen bonded protons are in constant exchange with the solvent, their exchange rates depend on their protection level and bond strength. Protons participating in hydrogen bonds, for instance, will be more protected and thus have lower exchange rates than those which are solvent exposed. Since deuterium (^2H) has an integer spin number (1), is it invisible in a ^1H - ^{15}H -HSQC experiment, meaning that this experiment is suitable for monitoring amide proton exchange (H/D) as the signal decays over time, which can vary from seconds to hours or even days. The hydrogen exchange rate of a certain amide group in a protein depends on the opening and closing rates according to¹³⁴:



VII.35

where k_{op} and k_{cl} refer to the opening and closing rate of protected hydrogen groups, respectively, k_{rc} is the intrinsic exchange rate in an unfolded polypeptide chain, which is affected by neighboring residues, pH and buffer conditions^{34,35,135}. The hydrogen exchange mechanism can either follow an EX₁ ($k_{cl} \ll k_{rc}$) or EX₂ ($k_{cl} \gg k_{rc}$) mechanism.³⁵ However, the apparent exchange rate is heavily dependent on the pH in the solution and can be altered by adjusting buffer conditions to fit a convenient laboratory timescale.³⁵

To determine the exchange rates of the individual amide protons, the normalized peak volumes are plotted as a function of the elapsed time and fitted to a three-parameter single-exponential decay function:¹³⁶

$$I(t) = I_0 e^{-k_{ex} \cdot t} + C$$

VII.36

where $I(t)$ is the intensity at time t , I_0 is intensity at time 0, k_{ex} is the exchange constant, t is the time elapsed and C is the final amplitude. The protection factors (Pf) for the several amide protons are estimated according to **Equation VII.37**.³⁵

$$Pf = \frac{k_{rc}}{k_{ex}}$$

VII.37

where k_{rc} and k_{ex} represent the exchange rates of the protein in the random coil and native conformations states, respectively. The hydrogen-exchange rates of amide protons in non-structured peptides, k_{rc} , can be estimated using the software SPHERE³⁴ (<http://www.fccc.edu/research/labs/roder/sphere>).

The free energy of exchange of the amide protons was calculated according to the following equation¹³⁷:

$$\Delta G_{ex} = -RT \ln \frac{k_{ex}}{k_{rc}} = -RT \ln \frac{1}{Pf}$$

VII.38

where R is the gas constant (8.314472 J.K⁻¹.mol⁻¹) and T is the absolute temperature at which the exchange was monitored.

VII.5 Study of protein-ligand complexes

VII.5.1 Saturation transfer difference

Nuclear Magnetic Resonance (NMR) spectroscopy (**Figure VII.34**) is a unique tool to study molecular interactions in solution, and has become an essential technique to characterize events of molecular recognition and obtain information about the interactions of small ligands with biologically relevant macromolecules (proteins and/or nucleic acids).⁵⁷ Ligand-based NMR screening and the NMR determination of the bound conformation of a ligand are nowadays important tools in the rational drug discovery process.^{42,54,56,59}

In this context, the Saturation Transfer Difference (STD-NMR) experiment has emerged as one of the most popular ligand-based NMR techniques for the study of protein-ligand interactions.^{1,2,41} The success of this technique is a consequence of its robustness and the fact that it is focused on the signals of the ligand, without any need of processing NMR information about the receptor and only using small amounts of non-labeled macromolecule.

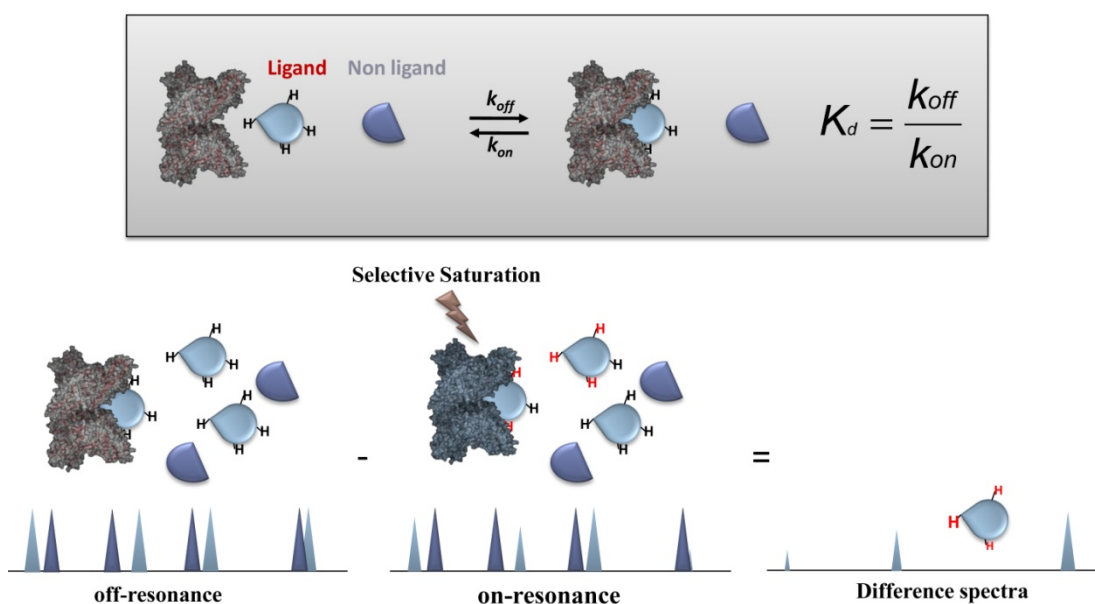


Figure VII.34: Scheme of the STD-NMR experiment.

The exchange between free and bound ligand allows intermolecular transfer of magnetization from the receptor to the bound small molecule

The STD-NMR experiment is based on the Nuclear Overhauser Effect (NOE – see section VII.2.4) and in the observation of the ligand resonance signals. It can be used as a **screening** technique, for identification of lead structures, or for **mapping the binding epitope** (useful for identifying ligand moieties important for binding).^{1,2,26,41}

This technique involves the acquisition and subtraction of two spectra and relies on the fact that, for a weak-binding ligand (K_d ranging from 10^{-8} M to 10^{-3} M – *see below*) there is exchange between the bound and the free ligand state.^{1,41}

As seen above (*section VII.2.4*), large molecules (like proteins) tumble slowly (large correlation time, τ_c) while small molecules tumble fast (small correlation time, τ_c). Furthermore, for large molecules the spin diffusion is very efficient, meaning that if some resonance is selectively saturated, in a short amount of time the whole protein is also saturated. While in contact with the protein a ligand is subject to the same NMR properties as the protein as a result of the slow tumbling of the complex. Saturation applied to the protein spreads to the ligand via **dipolar interactions**.⁴¹ The spectrum containing the information about the ligand binding is recorded with selective saturation of the receptor resonances. In these conditions, the exchange between free and bound ligand allows intermolecular transfer of magnetization from the receptor to the bound small molecule (via spin diffusion, through dipolar interactions) during the time used for the receptor saturation, which in turn is moved into solution where it is detected.

Basically, an STD experiment involves subtracting a spectrum in which the protein was selectively saturated (*on-resonance* spectrum obtained by irradiating at a region of the spectrum that contains only resonances of the receptor/protein such as 0 ppm to -1 ppm) from one recorded without protein saturation (*off-resonance* spectrum). It is important that the choice of the on-resonance irradiation frequency does not overlap with any of the ligand resonances. In the difference spectrum only the signals of the ligand(s) that received saturation transfer from the protein will remain. Other compounds that may be present but do not bind to the receptor will not receive any saturation transfer, their signals will be of equal intensity on the *on-resonance* and the *off-resonance* spectra and, as a consequence, after subtraction no signals will appear in the difference spectrum from the non-binding small molecule(s) (**Figure VII.34**).^{35,38}

The time interval used to saturate the receptor and the K_d of the ligand control the efficiency of the magnetization transfer process. The protein-to-ligand saturation transfer will affect the intensity of the ligand resonance signals in the spectrum obtained with selective receptor saturation (I_{SAT}), and when compared to a spectrum acquired without saturation transfer (I_0), the difference in intensity due to saturation transfer can be quantified ($I_{STD} = I_0 - I_{SAT}$) and constitutes an indication of binding (**Figure VII.34**).

Moreover, for a molecule that binds to the receptor, only the signals of the protons that are in close contact to the protein ($\leq 5 \text{ \AA}$) and receive magnetization transfer will appear in the difference spectrum and from those, the ones that are closer to the protein will have more intense signals, due to a more efficient saturation transfer. Therefore the STD can be used qualitatively to detect ligand binding or quantitatively to assess the strength of the binding

interaction and identify which part of the ligand is in close contact with the protein (epitope mapping).

STD is ideally suited to receptors with large masses (>30 KDa). Receptors with large molecular masses possess large rotational correlation time, τ_c that enhance spin diffusion and, consequently, saturation transfer within the receptor and to the ligand. In general, the intensity of the detected STD-NMR signal depends not only on the efficiency of the receptor-to-ligand saturation transfer but also on the number of ligand molecules in solution that received saturation from the receptor. Because ligand exchange is in place during the saturation time, long saturation times (up to 3 seconds)¹³⁸ or high ligand excess (10 to 100 fold), allow transfer of saturation from one receptor molecule to much more than one molecule of ligand (**Figure VII.35**). This can be used to benefit the experiment since it increases sensitivity and allows the use of very diluted protein solutions (in the micro-molar range), which is usually the critical factor in this type of studies. Normally, for a determinate system the ligand-to-protein ratio and the saturation time have to be tuned up and both have to be selected according to the expected K_d (*see below*).

The STD can best be analyzed if the amplification factor (A_{STD}) is used.⁴¹ The STD amplification factor is obtained by multiplying the relative STD effect of a given hydrogen (I_{STD}/I_0) at a given ligand concentration ($[L]_0$) with the molar ratio of ligand in excess relative to the protein ($[L]_0/[P]_0$), according to **Equation VII.39**:¹

$$A_{STD} = \frac{I_0 - I_{SAT}}{I_0} \times \text{molar ratio} = \frac{I_{STD}}{I_0} \times \text{molar ratio}$$

VII.39

were A_{STD} is the STD amplification factor, I_0 , I_{SAT} and I_{STD} are the intensities of the reference (off resonance), saturated (on resonance) and difference spectra (STD-NMR) respectively. I_{STD}/I_0 is the steady state STD response, η_{STD} .¹³⁹

For a determined saturation time the A_{STD} can also be depicted as the average number of ligand molecules saturated per molecule of receptor. In principle the longer the saturation time and the more ligand used the stronger the STD and the higher the A_{STD} due to ligand turn over at the binding site. In order to get the epitope mapping information from the amplification factor for a given saturation time, the relative STD (or A_{STD}) with the highest intensity is set to 100 %, and all other STD signals are calculated accordingly.^{1,41}

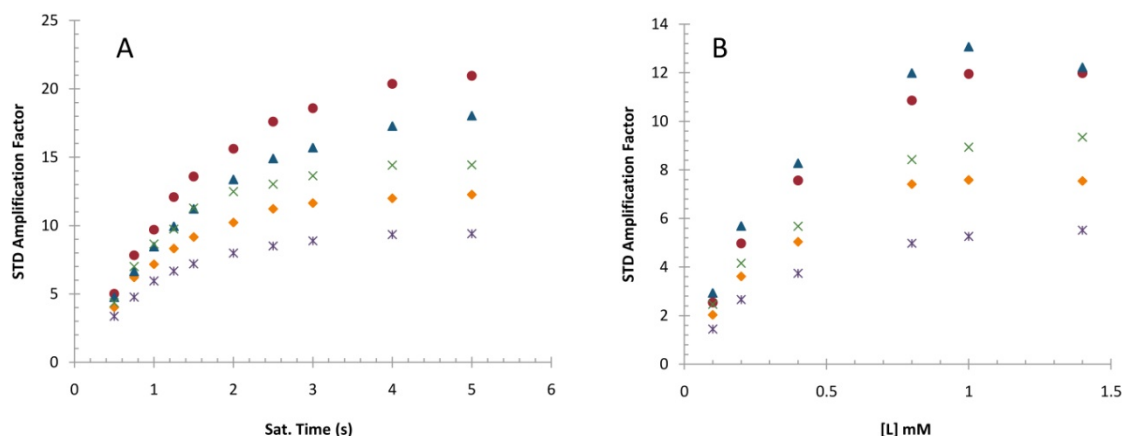
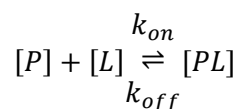


Figure VII.35: STD amplification factor as a function of the saturation time (A) and ligand concentration (B).¹

The saturation transfer takes place only to molecules bound to the protein with a rate that depends on the protein mobility, ligand/protein complex lifetime, and geometry. Some knowledge and understanding of the relative timescales of several important events is crucial for setting-up a successful STD NMR experiment and to understand its limitations. Let's consider a system where a protein, P , with a single binding site is in fast exchange with a ligand, L , yielding a protein/ligand complex, PL :



VII.40

where $[P]$, $[L]$ and $[PL]$ are the concentrations of free protein, free ligand and the complex, respectively. For this system the binding of the ligand to the receptor can be characterized by an *off* (k_{off}) and an *on* rate (k_{on}), and the corresponding thermodynamic equilibrium dissociation constant, K_d , given by:

$$K_d = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}}$$

VII.41

Assuming a purely diffusion controlled mechanism for the association reaction forming the complex, k_{on} would be about $10^7 \text{ s}^{-1}\text{M}^{-1}$. From this the dissociation rate k_{off} can be calculated as shown in **Table VII.6**:

Table VII.6: Dissociation rates for known K_d values assuming that k_{on} is diffusion controlled.

K_d [M]	k_{off} [s^{-1}]
1×10^{-3}	1×10^4
1×10^{-6}	10
1×10^{-9}	1×10^{-2}

The residence time, t_r^B in the binding pocket is:^{1,140}

$$t_r^B = \frac{1}{k_{off}}$$

VII.42

For a successful STD-NMR experiment it is desirable that the exchange between free and bound ligand is fast enough to allow the build-up of a population of saturated ligand in solution. For that reason k_{off} should be large enough to allow this amplification to occur, but not so high that it does not allow the ligand to remain in the binding site for enough time to receive the saturation from the receptor. If one makes the above assumptions, then it has been shown that the upper limit for K_d in a STD experiment will be controlled by the minimum residence time needed for saturation transfer, leading to a maximum K_d of 10^{-3} M.⁴¹ Before indicating the lower limit of K_d for the STD experiment, we have to consider the kinetics of another important NMR process; the rate at which the magnetization relaxes back to equilibrium. This rate is small for a small molecule and large for a large molecule.¹⁶ When the small ligand is bound to the large receptor it behaves as part of the receptor and therefore its relaxation rate is much faster than in the free-state. As a consequence, the ligand has to dissociate faster than the magnetization relaxation rate, otherwise relaxation occurs and the magnetization is lost. This represents a problem for tight binding ligands and sets a maximum residence time for the ligand in the binding site of the receptor, determined by the relaxation rate of the large receptor. As before, assuming a diffusion limited k_{on} rate, the lower limit for the K_d for normal STD experiments was determined to be 10^{-8} M. The K_d range of the STD-NMR experiment is then between $10^{-8} < K_D < 10^{-3}$ M.

Because the STD-NMR response arises directly from the protein/ligand complex the STD amplification factor, A_{STD} (**Equation VII.39**) can be used to determine the equilibrium dissociation constant K_d .¹ Given the equilibrium represented by **Equation VII.40** and a system in **fast exchange**,⁵⁷ a similar analysis to the one performed for the determination of the association constant from chemical shift data (*see Chapter III – Section III.4.4.9*) can be done. Thus, for the determination of the dissociation constant from the A_{STD} we have:

$$A_{STD} = \alpha_{max} \frac{(K_D + [L]_0 + [P]_0) - \sqrt{(K_D + [L]_0 + [P]_0)^2 - 4[P]_0[L]_0}}{2[P]_0}$$

VII.43

From the definition of A_{STD} , it follows that the amplification factor can be understood as the average number of ligand molecules saturated per molecule of receptor, and as a result it is expected that A_{STD} will increase with increasing $[L]_0$, until a maximum amplification (α_{STD}) is reached (when $[L]_0 \gg K_D$ and the receptor binding site is saturated). However, after the point of receptor saturation ($[L]_0 \gg K_D$), A_{STD} will decrease with increasing $[L]_0$ as seen in **Figure VII.35 – B**. This behavior of A_{STD} has to do with the fact that after this point I_{STD}/I_0 decreases with increasing ligand concentration, since I_0 is proportional to $[L]_0$. Therefore, provided that $[L]$ approximates to $[L]_0$ the STD data obtained for different ligand concentrations can be fitted with the equation above and used to estimate the values of K_D and α_{STD} .¹

VII.5.2 Diffusion ordered spectroscopy

Diffusion ordered spectroscopy (DOSY) is a method developed by Morris and Johnson.¹⁴¹ DOSY aims at identifying the molecular components of a mixture and to obtain at the same time information on their size. This information may be accessed by measuring the self-diffusion. Self-diffusion is the random translational motion of molecules or ions and it is driven by their internal kinetic energy.⁸⁴ Self-diffusion coefficients are related to the structural properties of a molecule by the dependence of the self-diffusion coefficients on the physical properties of the molecule (e.g. size, charge and shape). Furthermore, the self-diffusion coefficients also depend on the characteristics of the surrounding medium (e.g. temperature and viscosity).

For a spherical molecule moving in an unconstrained environment, the Stokes–Einstein law predicts a correlation between the hydrodynamic radius r and the self-diffusion coefficient D :

$$D = \frac{kT}{6\pi\eta r}$$

VII.44

where k is the Boltzmann constant, T is the temperature and η is the medium viscosity.

The diffusion of molecules is measured by evaluating the attenuation of a spin echo signal using **pulsed-field gradients** (PFG).¹⁴² A field gradient is a pulse or a period during which the static magnetic field (B_0) becomes deliberately heterogeneous.¹⁴³ In an experiment like this

(**Figure VII.36**) a first 90° pulse puts the magnetization aligned with the x plane (perpendicular to the applied static magnetic field – B_0). This field is then perturbed by the first gradient (PFG) of length δ and strength g . During the PFG, the field intensity varies linearly along the main axis of the sample introducing a dephasing of the bulk NMR signal. This causes a spatial phase encoding which depends on the spin position along the z-axis. The magnetic fields produced by the gradient create a situation where the magnetic field strength is added to the top of the sample and subtracted from the bottom, or vice-versa. At the end of the PFG, a magnetization helix is thus observed. A 180° pulse changes the direction of the precession and creates an echo that removes any contribution of the chemical shift to the evolution. The final PFG has an equal magnitude as the first one and will cancel its effects and refocus all spins. Because there is a time interval between the two PFGs (Δ – diffusion time), when the second gradient is applied the nuclei will not be in the same position as initially (due to diffusion) and, therefore, their intensity will not be fully recovered. The measurement of the diffusion is carried out by observing the attenuation of the NMR signal. **Figure VII.36** illustrates the most simple diffusion experiment – the Stejskal and Tanner sequence¹⁴⁴.

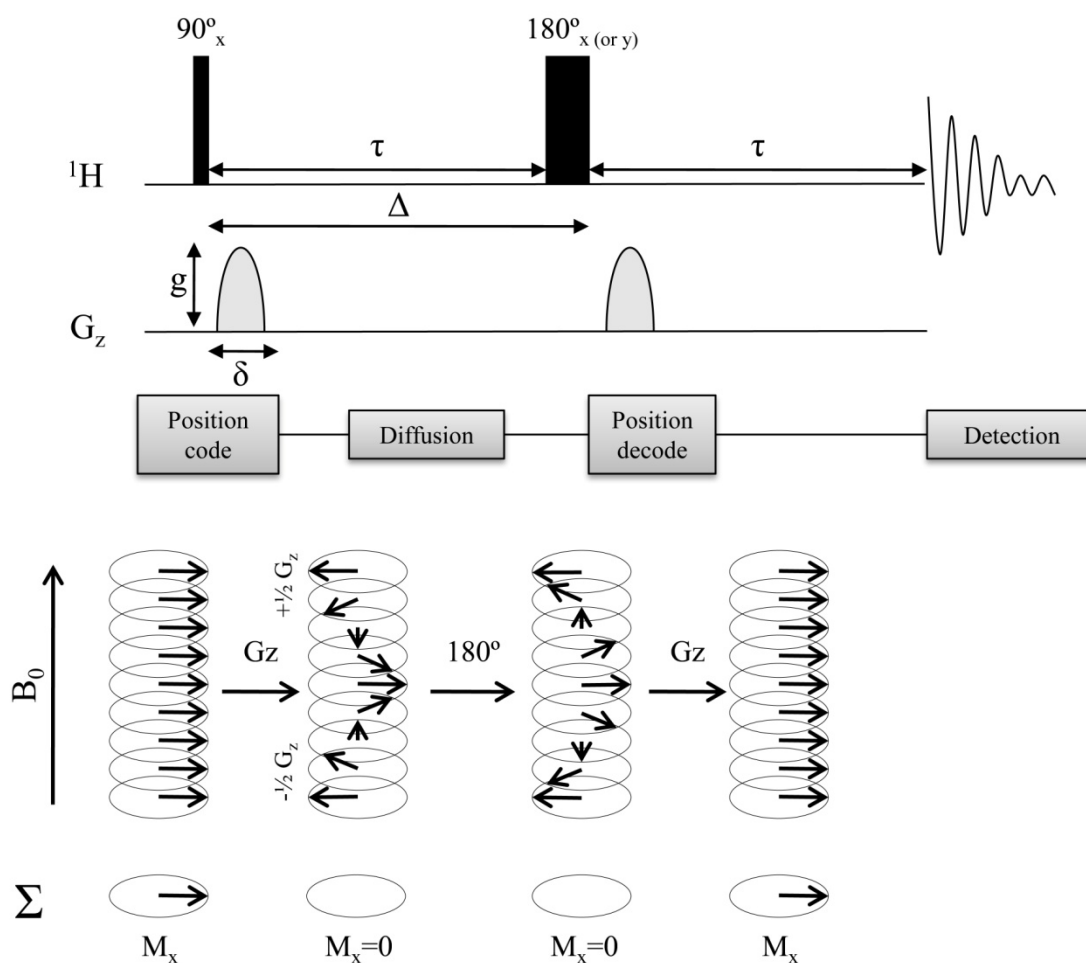


Figure VII.36: The Stejskal and Tanner pulsed field gradient NMR sequence.¹⁴⁴

In the experiment illustrated in **Figure VII.36** the degree of attenuation is a function of the magnetic gradient pulse amplitude (g) and occurs at a rate proportional to the diffusion coefficient (D) of the molecule according to:

$$I = I_0 \exp \left[-D(\gamma g \delta)^2 \left(\Delta - \frac{\delta}{3} \right) \right]$$

VII.45

where I_0 is the resonance amplitude at zero gradient strength, γ is the magnetogyric ratio of the proton ($2.675 \times 10^8 \text{ rad.T}^{-1}.\text{s}^{-1}$), g and δ are the strength and duration of the gradient, respectively and Δ is the diffusion time.

Nowadays there are several NMR experiments used in DOSY acquisition¹⁴⁵⁻¹⁴⁷ but, the most often used is the **BiPolar Pulse with Longitudinal Eddy current Delays** - BPPLIED pulse sequence. This sequence allows eddy currents to decay by storing the magnetization along the z-axis while all the generated fluctuations die away before the acquisition and uses bipolar gradients (i.e., applied in two opposite pulses, sandwiching the 180 ° pulse) which enable double effective strength as well as compensation for imperfections. These two optimizations have the same purpose: reduce the intensity of the eddy current generated by the PFG and to minimize its impact on the observed signal. If bipolar gradients are used, a correction for the time between those gradients, τ , has to be applied. In this situation **Equation VII.45** becomes:

$$I = I_0 \exp \left[-D(\gamma g \delta)^2 \left(\Delta - \frac{\delta}{3} - \frac{\tau}{2} \right) \right]$$

VII.46

where τ is the gradient pulse recovery time.

Although the potential of the DOSY technique for the analysis of complex mixtures is vast, several difficulties can be encountered. In order to achieve the most reliable results we must reduce or eliminate any experimental artifacts. Let's start by identifying what a good data set must have:

- Good registration of resonances
- No gradient-dependent spectral phase distortion or broadening
- No baseline artifacts
- Pure exponential decays with good differentiation in decay among the components

Having this in mind we see that, for a proper interpretation of the diffusion data one must have a good control of data acquisition. The first thing to consider is that the two gradient pulses have to be identical. Therefore, the gradient driver should be stable enough to deliver reproducible gradients within 1 part in 10^5 in order to measure diffusion as slow as $10^{-13} \text{ m}^2 \cdot \text{s}^{-1}$.¹⁴⁷ Furthermore, eddy current delays, due to fast switching (on and off) of gradients coils produce a magnetic field that can be experienced by the sample and cause distortion in the spectra. This can be avoided by placing two gradient pulses with opposite polarity with a 180° pulse between them (which is the case of the pulse BPPLIED pulse sequence). This creates a self-compensated composite where, usually, the distortion created by the first gradient is canceled by the second gradient. The 180° pulse assures that the magnetization continues to dephase in the same direction during both gradients.

One of the most difficult problems to solve is the temperature control. Temperature variations will cause a temperature gradient. Depending on the viscosity of the solvent used, temperature gradients along the axis of the tube can cause convection currents to establish. This adds a velocity term to the diffusion and will perturb the ideal decay in a PGSE NMR experiment which results into errors when analyzing the diffusions. Most NMR spectrometers introduce air through the bottom of the sample region that travels along the length of the tube and exits near the top. Due to the very little distance of the coils from the sample, a situation where the bottom of the tube experiences a different temperature than the top is easily created. There are two ways to overcome this issue: i) spin the sample; ii) reduce the diameter of the sample tube. Spinning the sample tube may suppress the effects of the temperature gradients but create problems at the level of sample vibration. The best results are obtained by reducing the sample tube diameter from 5 mm to 3 mm. As the more standard probes are optimized for 5 mm tubes, a 3 mm tube gives more room for the gas flow and, therefore, results into a more homogeneous temperature around the sample. Of course reducing the sample tube diameter will have consequences for the *S/N*.

As we saw above, despite the general limitations of the DOSY experiment, continuous improvements at the hardware and pulse sequence level allow overcoming most difficulties, making this technique an exceptional tool for mixture analysis. Extremely fine differentiation may be achieved if high quality data is used.

VII.6 References

1. Viegas, A.; Manso, J. o.; Nobrega, F. L.; Cabrita, E. J., Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding. *J Chem Educ* **2011**.
2. Viegas, A.; Macedo, A. L.; Cabrita, E. J., Ligand-Based Nuclear Magnetic Resonance Screening Techniques. In *Ligand-macromolecular interactions in drug discovery : methods and protocols*, Roque, A. C. A., Ed. Springer: New York, 2010; pp 81.
3. Wuthrich, K., NMR studies of structure and function of biological macromolecules. *Bioscience Rep* **2003**, *23* (4), 119.
4. Wagner, G., An account of NMR in structural biology. *Nat Struct Biol* **1997**, *4 Suppl*, 841.
5. Wuthrich, K., The second decade--into the third millenium. *Nat Struct Biol* **1998**, *5 Suppl*, 492.
6. Dalvit, C.; Fogliatto, G.; Stewart, A.; Veronesi, M.; Stockman, B., WaterLOGSY as a method for primary NMR screening: Practical aspects and range of applicability. *J Biomol Nmr* **2001**, *21* (4), 349.
7. Bax, A.; Davis, D. G., Mlev-17-Based Two-Dimensional Homonuclear Magnetization Transfer Spectroscopy. *J Magn Reson* **1985**, *65* (2), 355.
8. Hwang, T. L.; Shaka, A. J., Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *J Magn Reson Ser A* **1995**, *112* (2), 275.
9. Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A* **1997**, *94* (23), 12366.
10. Salzmann, M.; Pervushin, K.; Wider, G.; Senn, H.; Wuthrich, K., TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc Natl Acad Sci U S A* **1998**, *95* (23), 13585.
11. Ikura, M.; Kay, L. E.; Bax, A., A novel approach for sequential assignment of 1H, 13C, and 15N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **1990**, *29* (19), 4659.
12. Grzesiek, S.; Dobeli, H.; Gentz, R.; Garotta, G.; Labhardt, A. M.; Bax, A., 1H, 13C, and 15N NMR backbone assignments and secondary structure of human interferon-gamma. *Biochemistry* **1992**, *31* (35), 8180.
13. Clubb, R. T.; Thanabal, V.; Wagner, G., A Constant-Time 3-Dimensional Triple-Resonance Pulse Scheme to Correlate Intraresidue H-1(N), N-15, and C-13(') Chemical-Shifts in N-15-C-13-Labeled Proteins. *J Magn Reson* **1992**, *97* (1), 213.
14. Bax, A.; Grzesiek, S., Methodological Advances in Protein Nmr. *Accounts Chem Res* **1993**, *26* (4), 131.
15. Cordier, F.; Rogowski, M.; Grzesiek, S.; Bax, A., Observation of through-hydrogen-bond (2h)J(HC ') in a perdeuterated protein. *J Magn Reson* **1999**, *140* (2), 510.
16. Claridge, T. D. W., *High-resolution NMR techniques in organic chemistry*. 2nd ed.; Elsevier: Amsterdam ; Boston, 2009; p 383.
17. Bloch, F.; Hansen, W. W.; Packard, M., Nuclear Induction. *Phys Rev* **1946**, *69* (3-4), 127.
18. Purcell, E. M.; Torrey, H. C.; Pound, R. V., Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys Rev* **1946**, *69* (1-2), 37.
19. Guo, C.; Zhang, D.; Tugarinov, V., An NMR experiment for simultaneous TROSY-based detection of amide and methyl groups in large proteins. *J Am Chem Soc* **2008**, *130* (33), 10872.
20. Guo, C.; Tugarinov, V., Selective 1H- 13C NMR spectroscopy of methyl groups in residually protonated samples of large proteins. *J Biomol Nmr* **2010**, *46* (2), 127.

21. LeMaster, D. M., Deuterium labelling in NMR structural analysis of larger proteins. *Q Rev Biophys* **1990**, *23* (2), 133.
22. McIntosh, L. P.; Dahlquist, F. W., Biosynthetic incorporation of ¹⁵N and ¹³C for assignment and interpretation of nuclear magnetic resonance spectra of proteins. *Q Rev Biophys* **1990**, *23* (1), 1.
23. Tugarinov, V.; Kay, L. E., Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *J Am Chem Soc* **2003**, *125* (45), 13868.
24. Salzmann, M.; Pervushin, K.; Wider, G.; Senn, H.; Wüthrich, K., NMR Assignment and Secondary Structure Determination of an Octameric 110 kDa Protein Using TROSY in Triple Resonance Experiments. *Journal of the American Chemical Society* **2000**, *122* (31), 7543.
25. Viegas, A., cbm11 -2.
26. Viegas, A.; Bras, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Prates, J. A. M.; Fontes, C. M. G. A.; Bruix, M.; Romao, M. J.; Carvalho, A. L.; Ramos, M. J.; Macedo, A. L.; Cabrita, E. J., Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *Febs J* **2008**, *275* (10), 2524.
27. Bardiaux, B.; Favier, A.; Etzkorn, M.; Baldus, M.; Böckmann, A.; Nilges, M.; Malliavin, T. E., Simultaneous use of solution, solid-state NMR and X-ray crystallography to study the conformational landscape of the Crh protein during oligomerization and crystallization. *Advances and applications in bioinformatics and chemistry* **2010**, *3*, 25
28. Kay, L. E., Protein dynamics from NMR. *Biochem Cell Biol* **1998**, *76* (2-3), 145.
29. Palmer, A. G., A topical issue: NMR investigations of molecular dynamics. *J Biomol Nmr* **2009**, *45* (1-2), 1.
30. Kay, L. E., NMR studies of protein structure and dynamics. *J Magn Reson* **2005**, *173* (2), 193.
31. Mittermaier, A.; Kay, L. E., New tools provide new insights in NMR studies of protein dynamics. *Science* **2006**, *312* (5771), 224.
32. Jarymowycz, V. A.; Stone, M. J., Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem Rev* **2006**, *106* (5), 1624.
33. Palmer, A. G., 3rd, NMR characterization of the dynamics of biomacromolecules. *Chem Rev* **2004**, *104* (8), 3623.
34. Zhang, Y.-Z. Protein and peptide structure and interactions studied by hydrogen exchange and NMR. Ph.D. Thesis, University of Pennsylvania, Philadelphia, 1995.
35. Bai, Y. W.; Milne, J. S.; Mayne, L.; Englander, S. W., Protein Stability Parameters Measured by Hydrogen-Exchange. *Proteins* **1994**, *20* (1), 4.
36. Ciobanu, L.; Jayawickrama, D. A.; Zhang, X.; Webb, A. G.; Sweedler, J. V., Measuring reaction kinetics by using multiple microcoil NMR spectroscopy. *Angew Chem Int Ed Engl* **2003**, *42* (38), 4669.
37. Olsen, S. N.; Lumby, E.; McFarland, K.; Borch, K.; Westh, P., Kinetics of enzymatic high-solid hydrolysis of lignocellulosic biomass studied by calorimetry. *Appl Biochem Biotechnol* **2011**, *163* (5), 626.
38. Viegas, A., cbm44.
39. Wang, Y. S.; Liu, D. J.; Wyss, D. F., Competition STD NMR for the detection of high-affinity ligands and NMR-based screening. *Magn Reson Chem* **2004**, *42* (6), 485.
40. Ji, Z.; Yao, Z.; Liu, M., Saturation transfer difference nuclear magnetic resonance study on the specific binding of ligand to protein. *Analytical Biochemistry* **2009**, *385* (2), 380.
41. Meyer, B.; Peters, T., NMR Spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie-International Edition* **2003**, *42* (8), 864.
42. Carlomagno, T., Ligand-target interactions: What can we learn from NMR? *Annu Rev Bioph Biom* **2005**, *34*, 245.

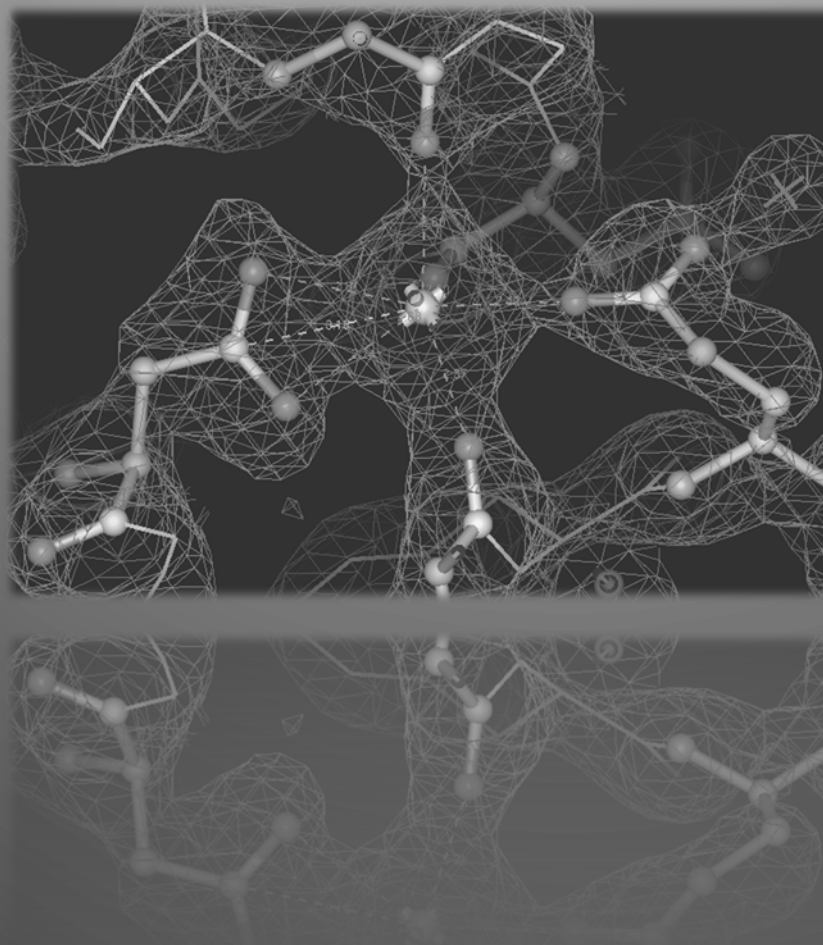
43. Schumann, F. H.; Riepl, H.; Maurer, T.; Gronwald, W.; Neidig, K. P.; Kalbitzer, H. R., Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *J Biomol Nmr* **2007**, *39* (4), 275.
44. Wuthrich, K., Protein recognition by NMR. *Nat Struct Biol* **2000**, *7* (3), 188.
45. Shortridge, M. D.; Hage, D. S.; Harbison, G. S.; Powers, R., Estimating Protein-Ligand Binding Affinity Using High-Throughput Screening by NMR. *J Comb Chem* **2008**, *10* (6), 948.
46. Biverstahl, H.; Andersson, A.; Graslund, A.; Maler, L., NMR solution structure and membrane interaction of the N-terminal sequence (1-30) of the bovine prion protein. *Biochemistry* **2004**, *43* (47), 14940.
47. Fernandez, C.; Hilty, C.; Wider, G.; Wuthrich, K., Lipid-protein interactions in DHPC micelles containing the integral membrane protein OmpX investigated by NMR spectroscopy. *Proc Natl Acad Sci U S A* **2002**, *99* (21), 13533.
48. Fernandez, C.; Wuthrich, K., NMR solution structure determination of membrane proteins reconstituted in detergent micelles. *Febs Letters* **2003**, *555* (1), 144.
49. Ader, C.; Pongs, O.; Becker, S.; Baldus, M., Protein dynamics detected in a membrane-embedded potassium channel using two-dimensional solid-state NMR spectroscopy. *Biochim Biophys Acta* **2010**, *1798* (2), 286.
50. Fielding, L., NMR methods for the determination of protein-ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2007**, *51* (4), 219.
51. Dobson, C. M.; Hore, P. J., Kinetic studies of protein folding using NMR spectroscopy. *Nat Struct Biol* **1998**, *5 Suppl*, 504.
52. Balbach, J.; Forge, V.; van Nuland, N. A.; Winder, S. L.; Hore, P. J.; Dobson, C. M., Following protein folding in real time using NMR spectroscopy. *Nat Struct Biol* **1995**, *2* (10), 865.
53. Dyson, H. J.; Wright, P. E., Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv Protein Chem* **2002**, *62*, 311.
54. Lepre, C. A.; Moore, J. M., *Fragment-Based NMR Screening in Lead Discovery*. Springer: 2007.
55. Betz, M.; Saxena, K.; Schwalbe, H., Biomolecular NMR: a chaperone to drug discovery. *Current Opinion in Chemical Biology* **2006**, *10* (3), 219.
56. Sun, C. H.; Huth, J. R.; Hajduk, P. J., NMR in pharmacokinetic and pharmacodynamic profiling. *Chembiochem* **2005**, *6* (9), 1592.
57. Lepre, C. A.; Moore, J. M.; Peng, J. W., Theory and applications of NMR-based screening in pharmaceutical research. *Chem Rev* **2004**, *104* (8), 3641.
58. Erlanson, D. A.; Wells, J. A.; Braisted, A. C., Tethering: Fragment-based drug discovery. *Annu Rev Bioph Biom* **2004**, *33*, 199.
59. Salvatella, X.; Giralt, E., NMR-based methods and strategies for drug discovery. *Chem Soc Rev* **2003**, *32* (6), 365.
60. Jahnke, W.; Widmer, H., Protein NMR in biomedical research. *Cellular and Molecular Life Sciences* **2004**, *61* (5), 580.
61. Wüthrich, K., *NMR of proteins and nucleic acids*. Wiley: New York, 1986; p 292.
62. Wishart, D. S.; Sykes, B. D., Chemical-Shifts as a Tool for Structure Determination. *Method Enzymol* **1994**, *239*, 363.
63. Wishart, D. S.; Case, D. A., Use of chemical shifts in macromolecular structure determination. *Methods Enzymol* **2001**, *338*, 3.
64. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M., Protein structure determination from NMR chemical shifts. *P Natl Acad Sci USA* **2007**, *104* (23), 9615.
65. Wishart, D. S.; Sykes, B. D.; Richards, F. M., Relationship between Nuclear-Magnetic-Resonance Chemical-Shift and Protein Secondary Structure. *Journal of Molecular Biology* **1991**, *222* (2), 311.
66. Wishart, D. S.; Sykes, B. D.; Richards, F. M., The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **1992**, *31* (6), 1647.

67. Marin, A.; Malliavin, T. E.; Nicolas, P.; Delsuc, M. A., From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. *J Biomol Nmr* **2004**, *30* (1), 47.
68. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L., BioMagResBank. *Nucleic Acids Res* **2008**, *36* (Database issue), D402.
69. Wishart, D. S.; Sykes, B. D., The ¹³C chemical-shift index: a simple method for the identification of protein secondary structure using ¹³C chemical-shift data. *J Biomol Nmr* **1994**, *4* (2), 171.
70. Avbelj, F.; Kocjan, D.; Baldwin, R. L., Protein chemical shifts arising from alpha-helices and beta-sheets depend on solvent exposure. *Proc Natl Acad Sci U S A* **2004**, *101* (50), 17394.
71. Cavanagh, J., *Protein NMR spectroscopy : principles and practice*. 2nd ed.; Academic Press: Amsterdam ; Boston, 2007; p 885.
72. Sattler, M.; Schleucher, J.; Griesinger, C., Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1999**, *34* (2), 93.
73. Teng, Q., *Structural biology : practical NMR applications*. Springer: New York, 2005; p 295.
74. Berger S., B. S., *200 and More NMR Experiments: A Practical Course*. Wiley-VCH: 2004.
75. Wang, A. C.; Bax, A., Determination of the backbone dihedral angles phi in human ubiquitin from reparametrized empirical Karplus equations. *Journal of the American Chemical Society* **1996**, *118* (10), 2483.
76. Reich, H. J. Chem 605 - Structure Determination Using Spectroscopic Methods.
77. Palmer, A. G.; Cavanagh, J.; Wright, P. E.; Rance, M., Sensitivity Improvement in Proton-Detected 2-Dimensional Heteronuclear Correlation Nmr-Spectroscopy. *J Magn Reson* **1991**, *93* (1), 151.
78. Bodenhausen, G.; Ruben, D. J., Natural Abundance N-15 Nmr by Enhanced Heteronuclear Spectroscopy. *Chem Phys Lett* **1980**, *69* (1), 185.
79. Higman, V. A. Protein NMR - A Practical Guide. <http://www.protein-nmr.org.uk/>.
80. Frey, M.; Sieker, L.; Payan, F.; Haser, R.; Bruschi, M.; Pepe, G.; LeGall, J., Rubredoxin from *Desulfovibrio gigas*. A molecular model of the oxidized form at 1.4 Å resolution. *J Mol Biol* **1987**, *197* (3), 525.
81. Kaiser, R., *Use of the Nuclear Overhauser Effect in the Analysis of High-Resolution Nuclear Magnetic Resonance Spectra*. AIP: 1963; Vol. 39, p 2435.
82. Neuhaus, D.; Williamson, M. P., *The nuclear Overhauser effect in structural and conformational analysis*. 2nd ed.; Wiley: New York, 2000; p xxvii.
83. Mo, H. P.; Pochapsky, T. C., Intermolecular interactions characterized by nuclear Overhauser effects. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1997**, *30*, 1.
84. Brand, T.; Cabrita, E. J.; Berger, S., Intermolecular interaction as investigated by NOE and diffusion studies. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2005**, *46* (4), 159.
85. Cabrita, E. J., The Nuclear Overhauser Effect. In *IX Curso Avanzado RMN, GERMN-RSEQ*: 2011; pp 211.
86. Keeler, J., *Understanding NMR spectroscopy*. 2nd ed.; John Wiley and Sons: Chichester, U.K., 2010; p 511.
87. Rule, G. S.; Hitchens, T. K., *Fundamentals of protein NMR spectroscopy*. Springer: Dordrecht, 2006; p 530.
88. Liu, G.; Shen, Y.; Atreya, H. S.; Parish, D.; Shao, Y.; Sukumaran, D. K.; Xiao, R.; Yee, A.; Lemak, A.; Bhattacharya, A.; Acton, T. A.; Arrowsmith, C. H.; Montelione, G. T.; Szyperski, T., NMR data collection and analysis protocol for high-throughput protein structure determination. *P Natl Acad Sci USA* **2005**, *102* (30), 10487.

89. Billeter, M.; Wagner, G.; Wuthrich, K., Solution NMR structure determination of proteins revisited. *J Biomol Nmr* **2008**, *42* (3), 155.
90. Shan, X.; Gardner, K. H.; Muhandiram, D. R.; Rao, N. S.; Arrowsmith, C. H.; Kay, L. E., Assignment of ¹⁵N, ¹³C α , ¹³C β , and HN Resonances in an ¹⁵N,¹³C,²H Labeled 64 kDa Trp Repressor–Operator Complex Using Triple-Resonance NMR Spectroscopy and 2H-Decoupling. *Journal of the American Chemical Society* **1996**, *118* (28), 6570.
91. Keller, R. The Computer Aided Resonance Assignment Tutorial. The Swiss Federal Institute of Technology, Zurich, 2004.
92. Guntert, P., Automated NMR structure calculation with CYANA. *Methods Mol Biol* **2004**, *278*, 353.
93. Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A., TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol Nmr* **2009**, *44* (4), 213.
94. Case, D. A.; Darden, T.; Cheatham III, T. E.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvai, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 11*, University of California: San Francisco, 2010.
95. Laskowski, R. A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M., AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol Nmr* **1996**, *8* (4), 477.
96. Masse, J. E.; Keller, R., AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. *J Magn Reson* **2005**, *174* (1), 133.
97. Olejniczak, E. T.; Xu, R. X.; Fesik, S. W., A 4D HCCH-TOCSY experiment for assigning the side chain ¹H and ¹³C resonances of proteins. *J Biomol Nmr* **1992**, *2* (6), 655.
98. Bax, A.; Clore, G. M.; Gronenborn, A. M., ¹H--¹H correlation via isotropic mixing of ¹³C magnetization, a new three-dimensional approach for assigning ¹H and ¹³C spectra of ¹³C-enriched proteins. *Journal of Magnetic Resonance (1969)* **1990**, *88* (2), 425.
99. Pochapsky, T. C.; Pochapsky, S. S., *NMR for physical and biological scientists*. Taylor & Francis: New York, 2007; p xxii.
100. Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S., Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *J Biomol Nmr* **2003**, *26* (3), 215.
101. Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G., CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Res* **2008**, *36*, W496.
102. Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A., Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A* **2008**, *105* (12), 4685.
103. Shen, Y.; Vernon, R.; Baker, D.; Bax, A., De novo protein structure generation from incomplete chemical shift assignments. *J Biomol Nmr* **2009**, *43* (2), 63.
104. Marion, D.; Driscoll, P. C.; Kay, L. E.; Wingfield, P. T.; Bax, A.; Gronenborn, A. M.; Clore, G. M., Overcoming the overlap problem in the assignment of ¹H NMR spectra of larger proteins by use of three-dimensional heteronuclear ¹H-¹⁵N Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1 beta. *Biochemistry* **1989**, *28* (15), 6150.
105. Zuiderweg, E. R.; Fesik, S. W., Heteronuclear three-dimensional NMR spectroscopy of the inflammatory protein C5a. *Biochemistry* **1989**, *28* (6), 2387.
106. Wüthrich, K.; Billeter, M.; Braun, W., Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-

- proton distance constraints with nuclear magnetic resonance. *Journal of Molecular Biology* **1983**, 169 (4), 949.
107. Downing, A. K., *Protein NMR techniques*. 2nd ed.; Humana Press: Totowa, N.J., 2004; p xi.
 108. Teng, Q., *Handbook of structural biology : practical NMR applications*. Kluwer Academic/Plenum Publishers: New York, 2005.
 109. Ramachandran, G. N.; University of Madras., *Crystallography and crystal perfection*. Academic Press: London, New York,, 1963; p 374.
 110. Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M., Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **1993**, 26, 283.
 111. Vriend, G., WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **1990**, 8 (1), 52.
 112. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007**, 450 (7172), 964.
 113. Hiyama, Y.; Niu, C. H.; Silverton, J. V.; Bavoso, A.; Torchia, D. A., Determination of ¹⁵N chemical shift tensor via ¹⁵N-2H dipolar coupling in Boc-glycylglycyl[¹⁵N glycine]benzyl ester. *Journal of the American Chemical Society* **1988**, 110 (8), 2378.
 114. Kay, L. E.; Torchia, D. A.; Bax, A., Backbone dynamics of proteins as studied by ¹⁵N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **1989**, 28 (23), 8972.
 115. Chi, Y. H.; Kumar, T. K. S.; Chiu, I. M.; Yu, C., N-15 NMR relaxation studies of free and ligand-bound human acidic fibroblast growth factor. *Journal of Biological Chemistry* **2000**, 275 (50), 39444.
 116. Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A., Rotational diffusion anisotropy of human ubiquitin from ¹⁵N NMR relaxation. *Journal of the American Chemical Society* **1995**, 117 (50), 12562.
 117. Peng, J. W.; Wagner, G., Mapping of Spectral Density-Functions Using Heteronuclear Nmr Relaxation Measurements. *J Magn Reson* **1992**, 98 (2), 308.
 118. Peng, J. W.; Wagner, G., Mapping of the Spectral Densities of N-H Bond Motions in Eglin-C Using Heteronuclear Relaxation Experiments. *Biochemistry* **1992**, 31 (36), 8571.
 119. Farrow, N. A.; Zhang, O. W.; Szabo, A.; Torchia, D. A.; Kay, L. E., Spectral Density-Function Mapping Using N-15 Relaxation Data Exclusively. *J Biomol Nmr* **1995**, 6 (2), 153.
 120. Houben, K. Studies on protein dynamics: Development and application of NMR relaxation measurements. Utrecht University 2004.
 121. Dosset, P.; Hus, J. C.; Blackledge, M.; Marion, D., Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J Biomol Nmr* **2000**, 16 (1), 23.
 122. Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B., HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J Magn Reson* **2000**, 147 (1), 138.
 123. Lipari, G.; Szabo, A., Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity. *Journal of the American Chemical Society* **1982**, 104 (17), 4546.
 124. Lipari, G.; Szabo, A., Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .2. Analysis of Experimental Results. *Journal of the American Chemical Society* **1982**, 104 (17), 4559.
 125. Doucleff, M.; Hatcher-Skeers, M.; Crane, N. J., *Pocket Guide to Biomolecular NMR*. Springer: 2011.
 126. Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M., Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society* **1990**, 112 (12), 4989.
 127. Mikhailov, D. V.; Washington, L.; Voloshin, A. M.; Daragan, V. A.; Mayo, K. H., Angular variances for internal bond rotations of side chains in GXG-based tripeptides

- derived from (^{13}C) -NMR relaxation measurements: Implications to protein folding. *Biopolymers* **1999**, *49* (5), 373.
128. Akke, M.; Brueschweiler, R.; Palmer, A. G., NMR order parameters and free energy: an analytical approach and its application to cooperative calcium(2+) binding by calbindin D9k. *Journal of the American Chemical Society* **1993**, *115* (21), 9832.
 129. Li, Z.; Raychaudhuri, S.; Wand, A. J., Insights into the local residual entropy of proteins provided by NMR relaxation. *Protein Sci* **1996**, *5* (12), 2647.
 130. Yang, D.; Kay, L. E., Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding. *Journal of Molecular Biology* **1996**, *263* (2), 369.
 131. Cavanagh, J.; Akke, M., May the driving force be with you--whatever it is. *Nat Struct Biol* **2000**, *7* (1), 11.
 132. Wand, A. J., Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nat Struct Biol* **2001**, *8* (11), 926.
 133. Stone, M. J., NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Accounts Chem Res* **2001**, *34* (5), 379.
 134. Berger, A.; Linderstrom-Lang, K., Deuterium exchange of poly-DL-alanine in aqueous solution. *Arch Biochem Biophys* **1957**, *69*, 106.
 135. Maity, H.; Lim, W. K.; Rumbley, J. N.; Englander, S. W., Protein hydrogen exchange mechanism: local fluctuations. *Protein Sci* **2003**, *12* (1), 153.
 136. Chi, Y. H.; Kumar, T. K. S.; Kathir, K. M.; Lin, D. H.; Zhu, G. A.; Chiu, I. M.; Yu, C., Investigation of the structural stability of the human acidic fibroblast growth factor by hydrogen-deuterium exchange. *Biochemistry* **2002**, *41* (51), 15350.
 137. Krishna, M. M.; Hoang, L.; Lin, Y.; Englander, S. W., Hydrogen exchange methods to study protein folding. *Methods* **2004**, *34* (1), 51.
 138. Yan, J. L.; Kline, A. D.; Mo, H. P.; Shapiro, M. J.; Zartler, E. R., The effect of relaxation on the epitope mapping by saturation transfer difference NMR. *J Magn Reson* **2003**, *163* (2), 270.
 139. Mayer, M.; Meyer, B., Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *Journal of the American Chemical Society* **2001**, *123* (25), 6108.
 140. Angulo, J.; Enriquez-Navas, P. M.; Nieto, P. M., Ligand-Receptor Binding Affinities from Saturation Transfer Difference (STD) NMR Spectroscopy: The Binding Isotherm of STD Initial Growth Rates. *Chem-Eur J* **2010**, *16* (26), 7803.
 141. Morris, K. F.; Johnson, C. S., Diffusion-Ordered 2-Dimensional Nuclear-Magnetic-Resonance Spectroscopy. *Journal of the American Chemical Society* **1992**, *114* (8), 3139.
 142. Price, W. S., Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part II. Experimental aspects. *Concept Magnetic Res* **1998**, *10* (4), 197.
 143. Keeler, J.; Clowes, R. T.; Davis, A. L.; Laue, E. D., Pulsed-Field Gradients - Theory and Practice. *Method Enzymol* **1994**, *239*, 145.
 144. Stejskal, E. O.; Tanner, J. E., Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *J Chem Phys* **1965**, *42* (1), 288.
 145. Holzgrabe, U.; Wawer, I.; Diehl, B., *NMR Spectroscopy in Pharmaceutical Analysis*. Elsevier: 2008.
 146. Pelta, M. D.; Barjat, H.; Morris, G. A.; Davis, A. L.; Hammond, S. J., Pulse sequences for high-resolution diffusion-ordered spectroscopy (HR-DOSY). *Magn Reson Chem* **1998**, *36* (10), 706.
 147. Antalek, B., Using pulsed gradient spin echo NMR for chemical mixture analysis: How to obtain optimum results. *Concept Magnetic Res* **2002**, *14* (4), 225.



Chapter VIII: X-Ray Crystallography

In this chapter, I give a general overview of protein crystallography, focusing on the crystallization of proteins, the basic theory behind the method, the main steps involved in solving a crystal structure, and the criteria used to validate the structural models.

Table of Contents

Summary	275
VIII.1 Introduction.....	275
VIII.2 Crystal systems: symmetry operations and space groups.....	277
VIII.3 Protein crystallization.....	281
VIII.3.1 Matthews' volume.....	283
VIII.4 Structure determination.....	284
VIII.4.1 X-ray diffraction and data collection.....	284
VIII.4.1.1 Synchrotron radiation.....	286
VIII.4.2 Model building and refinement.....	287
VIII.4.2.1 Molecular replacement.....	288
VIII.4.2.2 Model building.....	291
VIII.4.2.3 Model refinement.....	294
VIII.4.3 Structure validation.....	297
VIII.5 References.....	299

Summary

In this chapter, I describe some fundamental principles and aspects of protein crystallography, giving emphasis on protein crystallization (*Section VIII.3*) and the basic theory behind the method, the main steps involved in solving a crystal structure (*Section VIII.3*), and the criteria used to validate the structural models (*Section VIII.4.3*). **Figure VIII.1** shows a flowchart of the main steps involved in a 3D structure determination by X-ray crystallography.

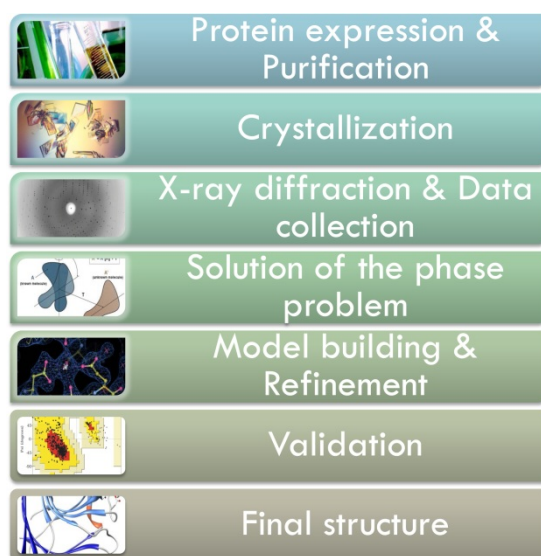


Figure VIII.1: Flowchart of the main steps involved in a 3D structure determination by X-ray crystallography.

VIII.1 Introduction

Protein crystal structures began to be determined in the late 50's, beginning with the structure of myoglobin¹ (at a resolution of 6 Å) by Max Perutz and Sir John Kendrew, for which they were awarded the Nobel Prize in chemistry in 1962. Since then X-ray crystallography has been the most common experimental method to obtain atomic resolution structures of macromolecules with 14 Nobel prizes in chemistry or medicine awarded to protein crystallographers². The development of highly sophisticated X-ray sources (synchrotron beam lines), advanced software tools, and superior workstations makes structure determination by X-ray crystallography a very powerful tool for structural biologists. Currently (May 2012), there are more than 70,000 protein and nucleic acid structure solved by X-ray crystallography deposited in the Protein Data Bank (PDB - <http://www.pdb.org/pdb>) (**Figure VIII.1**). This

clearly contrasts with the other two techniques able to produce tridimensional structures of proteins: NMR (with ~ 9300) and electron microscopy, EM (~ 420). In principle, it has become possible to solve the 3D crystal structure of any molecular entity, may it be as small as water in ice crystals or as large as complete ribosomes³ (contrasting with NMR, whose limit is around 100 kDa), providing detailed information which includes positions of the atoms, bond angles and distances and other structural parameters. Elucidation of these properties is fundamental for understanding the processes that take place in living organisms and, in a more practical application drug design and development⁴. However, the high accuracy of crystallography comes with a price: good crystals must be found and limited information about the molecule's dynamic behavior in solution is available from one single diffraction experiment.

In this chapter I will describe some theoretical principles and experimental techniques I used for determining the crystal structures of *CtCBM11* and the type II cohesin-dockerin complexes presented in Chapter II and Chapters V and VI, respectively and address some fundamental principles of protein crystallography.

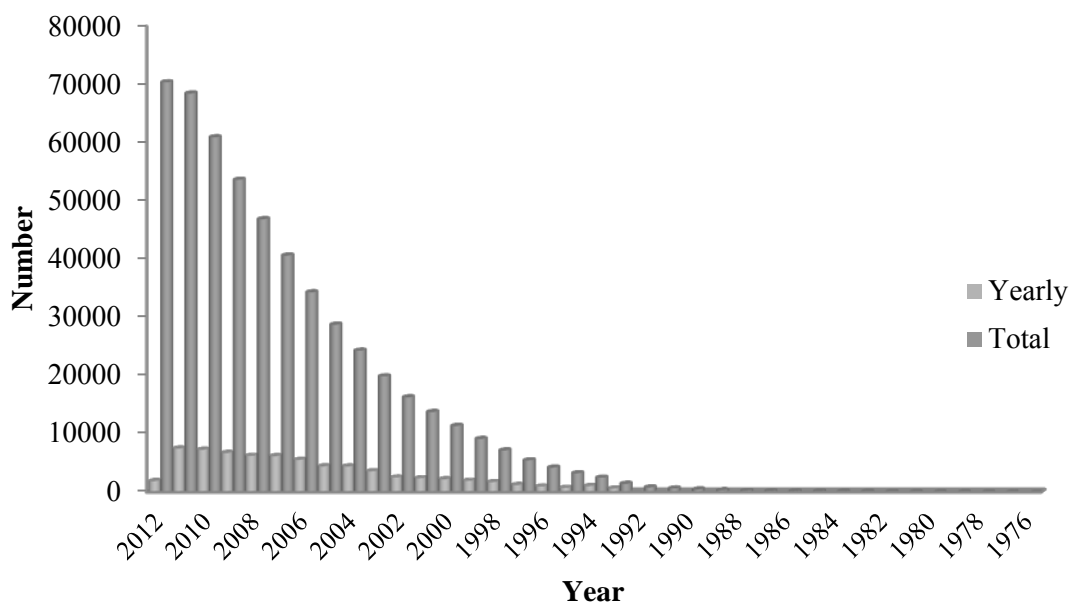


Figure VIII.2: Yearly and total growth of structures solved by X-ray crystallography.

Data was taken from the Protein Data Bank (<http://www.pdb.org/pdb>)

VIII.2 Crystal systems: symmetry operations and space groups

Before getting into protein crystallization and structure determination, it is necessary to have some knowledge about crystal systems, symmetry operations and space groups. By definition crystals are three-dimensional, ordered and periodical structures of molecules that are arranged in a repeating pattern, extending in all three spatial dimensions. The smallest repeating unit that, when duplicated and translated, can generate the entire crystal, it's called a **unit cell** and it may have a number of shapes, depending on the angles between the cell edges and the relative lengths of the edges (**Figure VIII.3**). The **asymmetric unit** is the smallest portion of the crystal that, when duplicated and moved by crystal symmetry operations, can produce the unit cell of the crystal (**Figure VIII.3**). A crystal asymmetric unit can contain one biological entity, only a part of a biological entity or multiple biological entities.

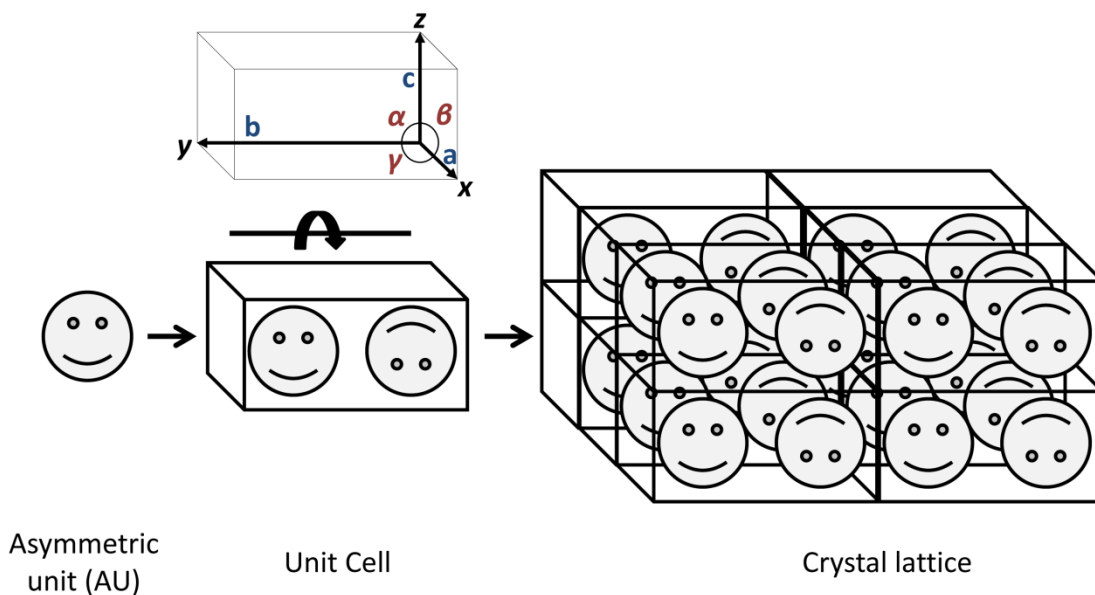


Figure VIII.3: Crystal architecture.

The dimension of the unit cell is given by three vectors, a , b and c and by three angles, α , β and γ (**Figure VIII.3**). The location of each atom in the unit cell is then defined by tridimensional coordinates, x , y and z , with the origin of one of the vertices as the origin of the coordinate system. By definition the direction x of the crystalline network corresponds to the direction of vector a , the direction y corresponds to the direction of vector b and the direction z corresponds to the direction of vector c . In crystallography, it is useful to describe the relationship between a crystal face and its counterpart in the crystal lattice. These and all other

regularly spaced planes that can be drawn through lattice points can be thought of as sources of diffraction and can be designated by a set of three numbers called **Miller indices** (h,k,l).⁵ Three indices h , k and l identify a particular set of equivalent, parallel planes. The index h gives the number of planes in the set per unit cell in the x direction or, equivalently, the number of parts into which the set of planes cut the a edge of each cell. The indices k and l are related with the division of b and c , respectively. Hence, if the first plane encountered cuts the a edge at some fraction $1/n_a$ of its length, and the same plane cuts the b edge at some fraction $1/n_b$ of its length, then the h index is n_a and the k index is n_b (**Figure VIII.4**). If a set of planes is parallel to an axis, that particular index is 0. Therefore, the unit cell is bounded by the planes (100), (010), and (001). The application of Miller indices allows crystal faces to be labeled in a consistent fashion, which together with accurate measurements of the angles between crystal faces, allows the morphology of crystals to be described in a reproducible way.

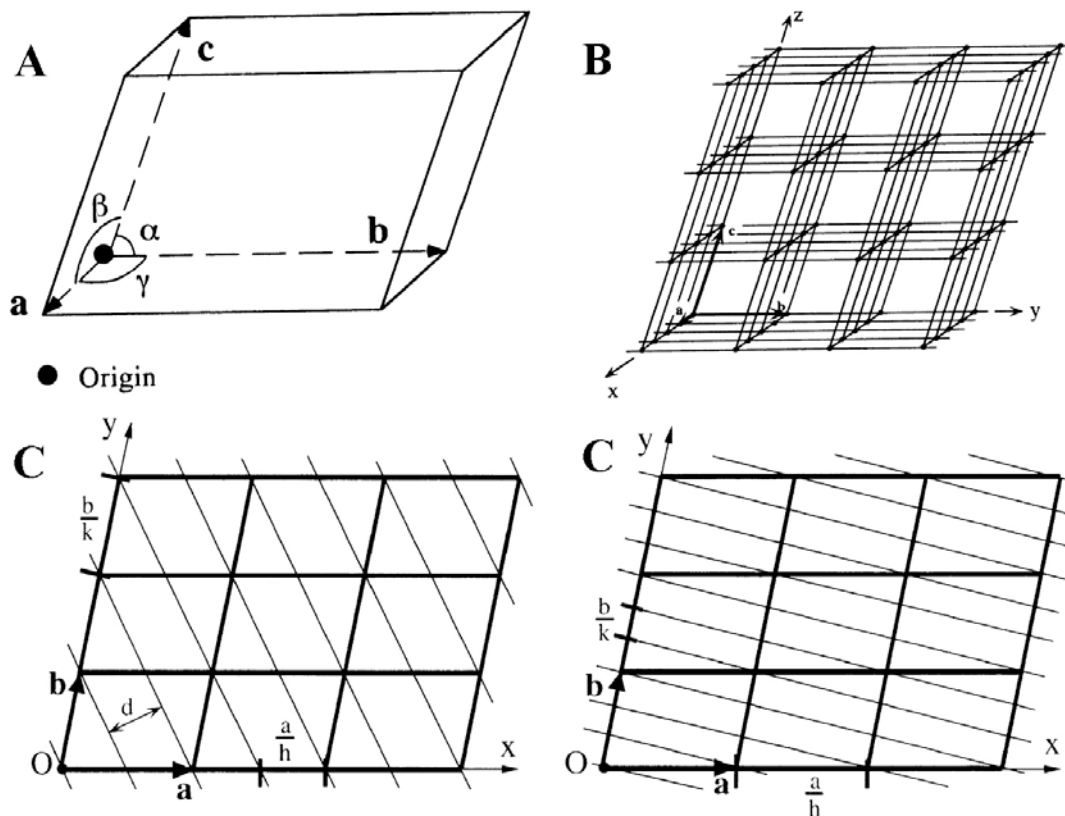


Figure VIII.4: The Miller indices.

A) One unit cell in the crystal lattice. **B)** A crystal lattice in a 3D stack of unit cells. **C)** Lattice planes in a 2D lattice with $h = 2$ and $k = 1$ and **C)** $h = 1$ and $k = 3$. Adapted from Drenth J. *et al* (2007)⁶

The choice of the unit cell in the crystal has to follow certain rules. If there are no symmetry considerations, the following rules must be followed⁶:

1. The axis system should be right-handed;
2. The basis vectors should coincide as much as possible with directions of highest symmetry;
3. The cell taken should be the smallest one that satisfies condition 2. This condition sometimes leads to the preference of a face-centered (A, B, C, or F) or a body-centered (I) cell over a primitive (P) smallest cell. Primitive cells have only one lattice point per unit cell, whereas non-primitive cells contain two or more lattice points per unit cell. These cells are designated A, B, or C if one of the faces of the cell is centered: It has extra lattice points on opposite faces of the unit cell, respectively, on the *bc* (A), *ac* (B), or *ab* (C) faces. If all faces are centered, the designation is F.
4. Of all lattice vectors, none is shorter than *a*;
5. Of those not directed along *a*, none is shorter than *b*;
6. Of those not lying in the *a, b* plane none is shorter than *c*;
7. The three angles between the base vectors *a, b*, and *c* are either all acute or all obtuse.

Crystals can have three basic types of symmetry: **rotation** (1-, 2-, 3-, 4- and 6-), **mirror** (*m*) **and inversion** and **translation**. For a crystal with only rotational symmetry, every molecule in the crystal can be obtained by rotating a copy of itself by a specific angle about a particular axis. Allowed rotational symmetries are 1-fold, 2-fold (180°), 3-fold (120°), 4-fold (90°), and 6-fold (60°). 5-fold symmetry is not allowed in crystals, nor is 7-fold symmetry or higher because it is physically impossible to build up a repeating tridimensional array that is based on 5-fold or 7-fold symmetry.⁷ Mirror and inversion symmetry is not possible in protein crystals as they imply changing the hand of objects and proteins are chiral. Finally, translation can be combined with rotations or mirror planes to give **screw axes** or glide plans, respectively. The screw axis is noted by a number, *n*, where the angle of rotation is $360^\circ/n$. The degree of translation is then added as a subscript showing how far along the axis the translation is, as a portion of the parallel lattice vector. For instance, 2_1 denotes a 180° (2-fold) rotation, followed by a translation of $\frac{1}{2}$ of the lattice vector.

The different combinations of symmetry operations that characterize a crystal define a **space group**. The space group can be defined as a set of symmetry operations that allow converting the asymmetric unit in the crystal lattice. The allowed symmetry operations are restricted by two conditions: i) they should be compatible with the infinite translational repetition of the crystal

lattice and ii) they cannot induce a different symmetry than the one of the asymmetric unit. If the space group contains a 4-fold axis, then the unit cell parallelepiped must have a 4-fold axis; if the space group relating the asymmetric units has a 3-fold axis, then a 3-fold axis is required to be present in the unit cell, and so on.⁸ The combination of all symmetry operations with the translational elements gives 230 possible space groups, divided by seven lattice types (triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic - **Table VIII.1**). By combining one of these seven lattice systems with one of the lattice centerings (**P** – primitive; **C** – centered on the a, b or A,B face; **I** – body centered; **R** – rhombohedral; **F**- face centered) we obtain the 14 Bravais lattices⁹ (**Figure VIII.5**).

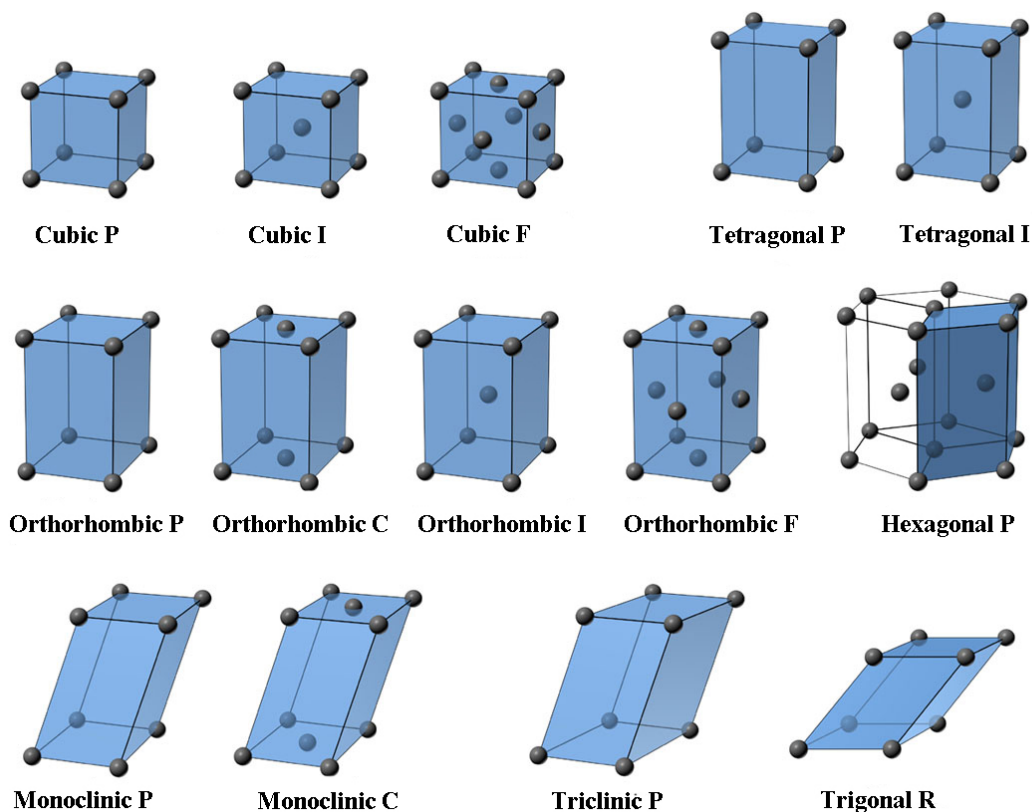


Figure VIII.5: The 14 Bravais lattices.

Adapted from: <http://people.tribe.net/scottthesculptor/photos/53c3eae8-d1d1-44a9-83d4-12269c50676f>

The characteristics of each space group are described in the International Tables for Crystallography¹⁰. In the particular case of biological molecules, because they are chiral, the number of possible space groups is reduced to 65 (**Table VIII.1**). The precise space group in which a protein will crystallize is impossible to predict and the same protein, given different crystallization conditions, can crystallize in different space groups.

Table VIII.1: Space group in proteins

Lattice type	Class	Space group	Cell restrictions	Angular restrictions
Triclinic	1	P1	$a \neq b \neq c$	-
Monoclinic*	2	P2, P2 ₁ , C2	$a \neq b \neq c$	$\alpha = \beta = 90^\circ$
Orthorhombic	222	P222, P222 ₁ , P2 ₁ 2 ₁ 2, P2 ₁ 2 ₁ 2 ₁	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90^\circ$
		C222, C222 ₁ F222, I222, I2 ₁ 2 ₁ 2 ₁		
Tetragonal	4	P4, P4 ₁ , P4 ₂ , P4 ₃ , I4, I4 ₁	$a = b \neq c$	$\alpha = \beta = \gamma = 90^\circ$
	422	P422, P4 ₂ 2, P4 ₁ 22, P4 ₃ 22 P4 ₂ 2 ₁ 2, P4 ₃ 2 ₁ 2, P4 ₂ 22, P4 ₂ 2 ₁ 2, I422, I4 ₁ 22		
Trigonal	3	P3, P3 ₁ , P3 ₂ , R3	$a = b \neq c$	$\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$
	32	P312, P321, P3 ₁ 2 ₁ , P3 ₂ 2 ₁ , P3 ₁ 12, P3 ₂ 12, R32	For R: $a = b = c$	$\alpha = \beta = \gamma < 120^\circ$
Hexagonal	6	P6, P6 ₁ P6 ₂ , P6 ₃ , P6 ₄ , P6 ₅	$a = b \neq c$	$\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$
	622	P622, P6 ₁ 22, P6 ₅ 22, P6 ₂ 22 P6 ₄ 22, P6 ₃ 22		
Cubic	23	P23, F23, I23, P2 ₁ 3, I2 ₁ 3	$a = b = c$	$\alpha = \beta = \gamma = 90^\circ$
	432	P432, P4 ₁ 32, P4 ₃ 32, P4 ₂ 32 F432, F4 ₁ 32, I432, I4 ₁ 32		

* See Appendix B, Section B.2

P – primitive; **C** – centered on the a, b or A,B face; **I** – body centered; **R** – rhombohedral; **F**- face centered;

VIII.3 Protein crystallization

The first step to obtain a crystallographic structure is to obtain a crystal (and a good one!). First of all, why crystals? Well, in the first place it would be impossible to measure the diffraction of a single molecule as it would be too weak and full of noise and second, the molecule would be burned up by the X-rays. Obtaining good crystals can be considered the bottle neck of solving a structure by X-ray crystallography as growing single crystals of good diffraction quality represents a major challenge⁶. Protein crystallization is mainly a trial-and-error procedure in which the protein slowly precipitates until it forms crystals (**Figure VIII.2**). When considering the conditions that may affect protein crystallization one has to consider several factors such as: pH (determined by the buffer), ionic strength, temperature, protein concentration and purity (a pure protein sample is fundamental – approximately 97%), which

precipitant and at which concentration, additives (see Appendix B, Table B.1 and B.2), etc. Any of these factors can make the difference between a good crystal and no crystal at all. Moreover, the conditions that worked for one crystal won't necessarily work for a different one. A good crystal is characterized by a high purity and order and large enough to provide a diffraction pattern when exposed to X-rays.⁵

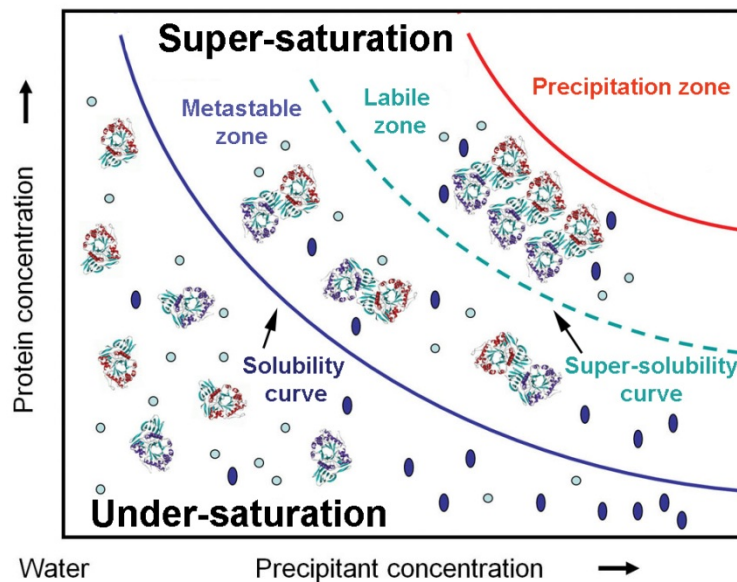


Figure VIII.6: Solubility curve of a protein as a function of the precipitant concentration.

The solubility curve (blue line) divides a phase separation into regions that support crystallization processes (super-saturated solutions) from those where crystals will dissolve (under-saturated solutions). The super-solubility curve (green dashed) further divides the super-saturated region into higher super-saturation conditions where nucleation and growth compete (labile zone) and lower levels where only crystal growth will occur (metastable zone). Adapted from Rupp, B. (2010)¹¹.

Another aspect one has to consider in order to obtain crystals is the technique to be used. The two most often used methods for obtaining crystals, hanging drop (**Figure VIII.4**) and sitting drop, are based on the vapor diffusion principle. In the hanging drop method, which is the one I used in all my crystallization experiments, drops containing a mixture of usually 1-2 μL of protein with the same volume of the precipitant solution are prepared in a glass slide which is then placed upside down over the reservoir containing the precipitant solution ($\sim 500 \mu\text{L}$). The chamber is sealed by applying silicone in the borders of reservoir before the glass slide is put into place. Because the concentration of the protein and precipitant are reduced to half, water evaporates from the drop to the reservoir until equilibrium is reached, thus slowly increasing the concentration of both protein and precipitant in the drop. Then, hopefully, (good) crystals will grow.

In order to maximize the chances of obtaining suitable crystals, it is necessary to test several different conditions (see Appendix B, Table B.1). The drawback of this approach is that usually large amounts of protein are required before a good crystal is obtained. However, nowadays

crystallization robots are becoming a standard piece of equipment in crystallography laboratories for screening and optimization of crystallization conditions. The main advantage of robots is the small sample size required, thus allowing to test many conditions with minimal protein volumes.

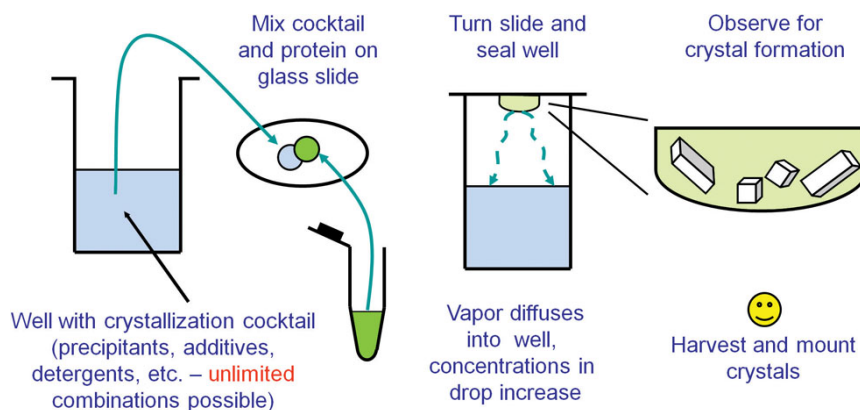


Figure VIII.7: Obtaining crystals by the hanging drop method.

A few microliters of protein solution are mixed with an equal volume of the precipitant solution. A drop of this mixture is put on a glass slide which covers the reservoir. Because the protein/precipitant mixture in the drop is less concentrated than the reservoir solution, water evaporates from the drop into the reservoir, resulting in a slow increase of the concentration until crystals (may) form. Adapted from Rupp, B. (2010)¹¹.

VIII.3.1 Matthews' volume

Protein crystals are fragile due to their high content in water. The ratio between the solvent content and the macromolecule in a given asymmetric unit is given by the parameter V_M ($\text{\AA}^3/\text{Da}$), designated Matthews' coefficient¹²:

$$V_M = \frac{V_{unit\ cell}}{(Z \times M_{prot})}$$

VIII.1

where $V_{unit\ cell}$ (\AA^3) is the volume of the unit cell, M_{Prot} (Da) is the molecular mass of the protein in the unit cell and Z is the number of asymmetric units in the cell (i. e. the number of symmetry operations of the space group).

VIII.4 Structure determination

Considering we were successful in obtaining a good crystal, it's time to acquire the data. The crystal is placed in an intense beam of X-rays, usually of a single wavelength (monochromatic X-rays), producing the regular pattern of reflections. Based on the diffraction pattern obtained from X-ray scattering off the periodic assembly of molecules in the crystal, the electron density can be reconstructed. Additional phase information must be extracted either from the diffraction data or from supplementing diffraction experiments to complete the reconstruction (**the phase problem** in crystallography – see Section VIII.4.2). A model is then progressively built into the experimental electron density, refined against the data and the result is an accurate molecular structure – a crystal structure – which will then be deposited in the Protein Data Bank (PDB - <http://www.pdb.org/pdb>). The following sections (VIII.4.1 to VIII.4.3) explore the different steps from the data collection to the deposition of the structure.

VIII.4.1 X-ray diffraction and data collection

X-ray crystallography is an experimental technique that exploits the fact that X-rays are diffracted by crystals. X-rays have the proper wavelength (in the Ångström range, $\sim 10^{-8}$ cm) to be scattered by the electron cloud of an atom of comparable size. However, only the scattered waves that interfere constructively (according to Bragg's law – **Equation VIII.1**) give rise to a diffracted beam, registered as diffraction spots (reflections) on a detector (**Figure VIII.8**). According to the **Bragg's law**, an X-Ray beam will only be diffracted when it impinges upon a set of planes in a crystal, defined by the Miller indices (hkl), if the geometry of the situation fulfills **Equation VIII.1**⁵:

$$n\lambda = 2d_{hkl} \sin \theta$$

VIII.2

where n is an integer, λ is the wavelength of the X-ray beam, d_{hkl} is the interplanar spacing and θ is the diffraction angle or **Bragg's angle** (**Figure VIII.8**). Thus, for a planar interspacing d_{hkl} and an incident angle θ , constructive interference occurs when the path difference between the waves with wavelength λ is equal to an integral number n . The maximum θ angle corresponds to the minimum distance $d_{hkl, \min}$ in the crystal that can be resolved, and is called the **resolution of the diffraction pattern**: $d_{hkl, \min} = \lambda / \sin \theta_{\max}$ ⁴

For a given crystal there is an infinite number of sets of atom planes, and Bragg's law applies to all of them and if the crystal is rotated, each set of planes will diffract the radiation when the value of $\sin \theta$ becomes appropriate. This is the reason why diffraction data is collected for the whole of the crystal. The precise pattern made by the scattered X-ray beams is called the **diffraction pattern**.

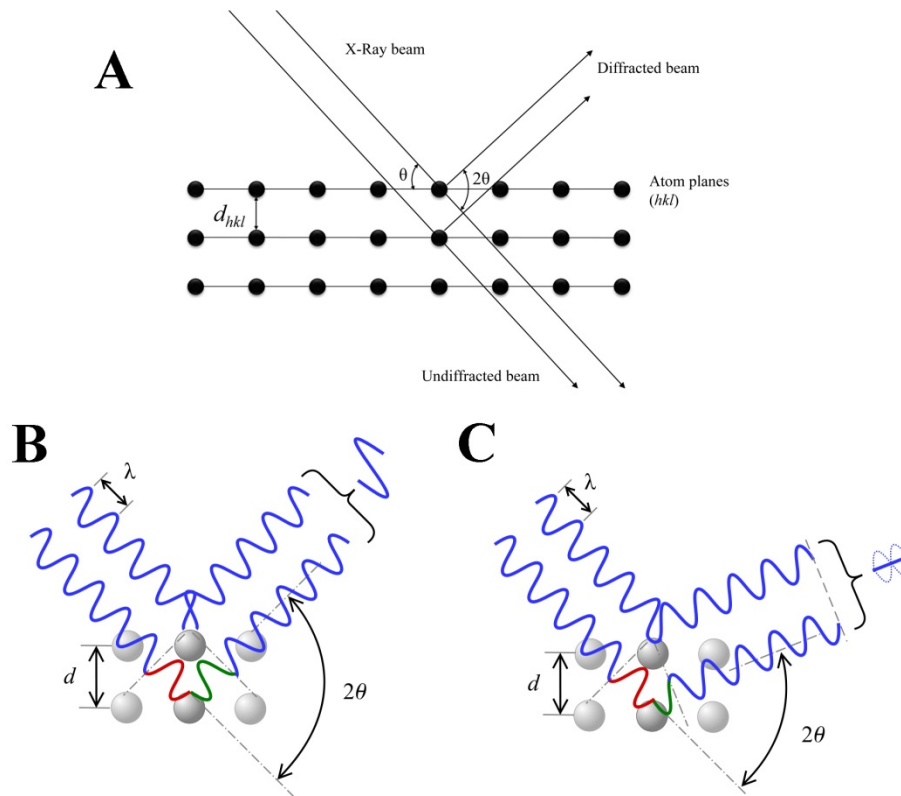


Figure VIII.8: Bragg's Law.

A) Two beams with identical wavelength and phase approach a crystalline solid and are scattered off two different atoms within it. The lower beam traverses an extra length of $2d \sin \theta$. According to the 2θ deviation, the phase shift causes constructive (B) or destructive (C) interferences. Adapted from Tilley, R. J. D. (2006)¹³ and http://en.wikipedia.org/wiki/Bragg's_law

According to Bragg's law (**Equation VIII.2**), by increasing the wavelength, the total diffracted intensity becomes less sensitive to the spacing or to changes in angle. This means that the diffraction pattern becomes less sensitive to the fine details. Moreover, the angle of diffraction, θ , is inversely related to the interplanar spacing d_{hkl} ($\sin \theta$ is proportional to $1/d_{hkl}$) which implies that large unit cells give small angles of diffraction and hence produce many reflections that fall within a convenient angle from the incident beam. On the other hand, small unit cells give large angles of diffraction, producing fewer measurable reflections.⁵ Because of this inverse relationship between the spacing in the object and the angle of diffraction, the diffraction space is called "**reciprocal space**" whereas the diffraction pattern is called the "**real space**".

In the reciprocal space, each point of coordinates (h,k,l) corresponds to a family of planes hkl in real space. The center of the diffraction pattern corresponds to the origin of the reciprocal space, which is reflection (000). The d_{hkl} values for any crystal can be calculated from knowledge of the lattice parameters. The Bragg equation, applied to diffraction data, results in a list of d_{hkl} values for a compound. It is possible, by putting these two data sets together, to determine the size of the unit cell of the material producing the diffraction pattern. This means allocating a value hkl to each diffracted beam, a process called **indexing**.⁵ The process of indexing can be done with the software MOSFLM¹⁴, from the CCP4 suite of programs¹⁵. The next step is to scale and merge the data set in order to produce a file containing the averaged intensities for each reflection, which is done with software SCALA¹⁶, also from the CCP4 suite of programs¹⁵

VIII.4.1.1 Synchrotron radiation

The success of the X-ray crystallography methodology depends on the ability to generate sufficiently strong X-ray beams that provide measurable diffraction images.¹⁷ The two main sources of X-rays used for collecting the diffraction data are rotating anodes and synchrotron radiation. In the first, X-rays are generated by electrons from heated filament (cathode) and accelerated by a magnetic field that collide with a metal target, usually copper or molybdenum (anode). When electrons collide with the anode they withdraw electrons from the lower energy orbitals of the anode. The electrons of the higher energy orbitals then tend to occupy the lower energy levels and in the process emit X-rays with a specific wavelength: radiation K_α , originated from a transition from the L to the K layer and K_β , originated from a transition from the M to the K layer. The wavelengths for the K_α and K_β transitions are 1.54Å (the one at the Crystallography laboratory at Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa) and 1.39Å, respectively for copper and 0.71Å and 0.63Å, respectively for molybdenum.¹⁸ Such sources were generally sufficient for studies of the crystals of comparatively small molecules, but collection of data for macromolecules such as proteins would often require many days or weeks.

Synchrotron radiation is nowadays one of the most common source of X-rays and its importance for macromolecular crystallography lies in i) the high brilliance of the beam (much smaller crystals can be used than in conventional X-ray crystallography), ii) the high intensity (allows data collection that previously took hours or days to be done in minutes, generating a significant increase in throughput), iii) tunability of the wavelength in the relevant range from 0.5 to 3.0 Å (which allows Multiple Anomalous Dispersion (MAD) techniques) and iv) the highly focused beams, which allow the structures of very large molecules to be obtained.¹⁹

However, high-energy photons of X-rays may have a harmful effect on crystals of biological macromolecules that undergo radiation damage if exposed to X-rays - **radiation damage**. The photons cause the formation of radicals, which leads to subsequent chemical reactions that progressively destroy the crystalline order. Moreover, some of these radicals may diffuse and exercise their destructive effects at other sites in the crystal. Nonetheless, the radiation damage problem can be reduced with modern, sensitive X-ray detectors that allow relatively short exposure times, and, mainly, by cooling the crystals to cryogenic temperatures (100 - 120 K). At these temperatures, radicals are still created by the X-ray photons, but their diffusion through the crystal is eliminated. This allows for most biological macromolecules to collect a complete dataset on one crystal. Nonetheless, even at cryotemperature, specific groups in the protein are damaged. For instance, disulfide bonds are especially prone to be damaged, leading to bond cleavage; carboxylic acids can be decarboxylated; cysteine, methionine and tyrosine can also suffer.⁶

VIII.4.2 Model building and refinement

Each reflection in the diffraction pattern is produced by a wave that can be described as the sum of the contributions of all scatterers in the unit cell and is characterized by its wavelength, λ (that of the X-rays), amplitude ($|F_{hkl}|$), and phase (α). Each one of these waves can be mathematically described as a Fourier series by the so-called structure factor (F_{hkl}) equation (**Equation VIII.3**). F_{hkl} has associated frequency, amplitude, and phase that can be formulated as a function of the electron density $\rho(x, y, z)$ of all atoms in the unit cell^{4,5}:

$$F_{hkl} = \int_V \rho(x, y, z) e^{[2\pi i(hx, ky, lz)]} dV$$

VIII.3

where, V is the volume of the unit cell and $\rho(x, y, z)$ is the electron density at position (x, y, z) in the unit cell. Each volume element contributes to F_{hkl} with a phase determined by its coordinates (x, y, z) . Because the Fourier transform operation is reversible, the electron density is in turn the transform of the structure factors, as follows:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx, ky, lz)}$$

VIII.4

Because:

$$F_{hkl} = |F_{hkl}| e^{i\alpha_{hkl}}$$

VIII.5

where α_{hkl} is the phase of the diffracted beam, we can rewrite **Equation VIII.4** such that:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i(hx, ky, lz)} + i\alpha_{hkl}$$

VIII.6

The amplitude can be obtained experimentally from the intensities of the reflections, I_{hkl} :

$$|F_{hkl}| \propto \sqrt{I_{hkl}}$$

VIII.7

However, the phase angles α_{hkl} cannot be derived straightforwardly from the diffraction pattern. This is commonly known as **the phase problem**, which will be discussed below (*Section VIII.4.2.1*) in terms of the method of **Molecular replacement**^{20,21} which was the only one I used in the work presented in this thesis.

From the above equations it's clear that reflections hkl and $-h-k-l$ have the same intensity, $I_{hkl} = I_{-h-k-l}$, the same structure factor, $|F_{hkl}| = |F_{-h-k-l}|$ but opposite phase angles, $\alpha_{hkl} = -\alpha_{-h-k-l}$ (assuming there is no anomalous diffraction). Therefore, **Equation VIII.6** can be rewritten to:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \cdot \cos[2\pi(hx, ky, lz) - \alpha_{hkl}]$$

VIII.8

and only the reflections hkl are considered (not the $-h-k-l$). This expression does not contain any imaginary term. The reflections hkl and $-h-k-l$ are called **Friedel pairs** (or **Bijvoet pairs** when anomalous diffraction occurs and hkl and $-h-k-l$ are different).⁶

VIII.4.2.1 Molecular replacement

In order to get the electron density map, necessary for the determination of the 3D structure, it is necessary to get phase information. Several approaches exist in order to solve this problem⁸:

- Deconvolution of the Patterson map
- The heavy atom method
- Isomorphous replacement
- Anomalous scattering
- Molecular replacement
- Direct methods

However, I will only focus on the molecular replacement method.

In crystallography it is possible to use the phases from structure factors of a known protein as initial phases for a new protein given that both share a common folding and at least 30% sequence identity²². This method was used to solve up to 70% of the deposited macromolecular structures and at its best has the advantages of being fast, cheap and highly automated.²³ Moreover even NMR derived structures can be used. This method is based on the **Patterson function**²⁴ which will be mentioned during the discussion of the method and explained in Section VIII.4.2.1.1.

The principle behind molecular replacement is very simple: using a model that we assume is similar to the unknown structure and a set of measured diffraction intensities, we try all possible orientations and positions of the model in the unknown crystal and find where the predicted diffraction best matches the observed diffraction. Then we use the phases of the model and the observed intensities to build an initial electron density map. Then, it's just a question of crystallographic refinement. The molecular replacement method is a three-step process^{20,21} (**Figure VIII.9**):

1. Rotation - the model is rotated and for each orientation a Patterson map is calculated and compared to the Patterson map calculated from the structure factors of the unknown structure (obtained in the diffraction experiment). The correct orientation is found based on the maximum-likelihood method;
2. Translation – the correctly orientated model is translated within the asymmetric unit to the correct coordinates. This is accomplished by moving the model, calculating a new Patterson map, and comparing it to the unknown-derived Patterson map.
3. Phase determination – using **Equation VIII.8** a set of initial phases can be determined using the coordinates of the models as determined by (1) and (2) and the experimentally measured intensities.

These phases, of course, will only be approximate because the molecules are not truly identical, yet, because they are structurally similar, the calculated phases may provide adequate

estimates and a starting point for improvement and refinement of the unknown molecules in both real and reciprocal space. However, this information should be used with some care; despite low resolution data can be used, as these are the ones that influence the most the Patterson function, high resolution data is important in order to avoid model bias.

Molecular replacement can be performed using, for instance, the software PHASER²⁵, MOLREP²⁶, or BALBES²⁷, which are part of the CCP4¹⁵ suite of programs.

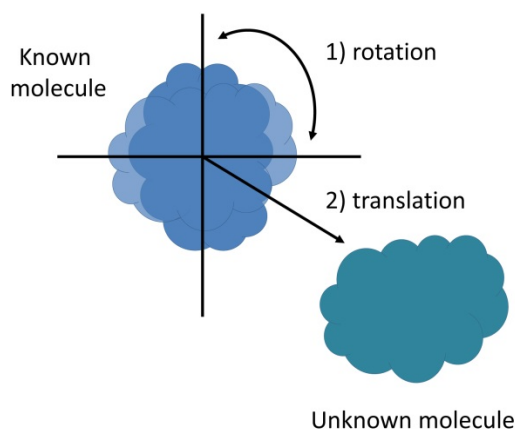


Figure VIII.9: The molecular replacement method.

VIII.4.2.1.1 Patterson function

The Patterson function²⁴ was introduced in 1935 by Arthur Lindo Patterson as a method for localizing the position of atoms without previous knowledge of the phase angles (only for small molecules). The Patterson function, $P(u, v, w)$ is given by:

$$P(u, v, w) = \frac{1}{V} \sum_{hkl} |F_{hkl}|^2 \cdot \cos[2\pi(hu, kv, lw)]$$

VIII.9

where u , v and w are relative coordinates in the unit cell of volume V . Note that the coefficients in the summations are $|F_{hkl}|^2$, not $|F_{hkl}|$ as in **Equation VIII.8**, which are proportional to the intensity (see **Equation VIII.7**) and, because all phase angles are zero in the Patterson function, it can be calculated without any previous knowledge of the structure. In practical terms, the Patterson map exhibits peaks resulting from the vectors connecting the atoms in the unit cell (**Figure VIII.10**). Given a crystal space (where the atoms are) defined by the value of the electron density function, ρ at every point in the unit cell given by the coordinates x , y , z , the Patterson space (also periodic and defined by a unit cell identical to the crystal unit cell) is defined by generic coordinates (u, v, w) in such a way that any pair of atoms in the crystal,

located at (x_1, y_1, z_1) and (x_2, y_2, z_2) , will be shown in the Patterson map by a maximum with coordinates: $u = x_1 - x_2$; $v = y_1 - y_2$; $w = z_1 - z_2$ (**Figure VIII.10**). Moreover, the Patterson map is centro-symmetric, which means that for each vector u, v, w a vector $-u, -v, -w$ exists. Another characteristic of the Patterson map is that the height of the peaks is proportional to the product of atomic numbers of the atoms involved, which provides a great advantage in detecting the heavier atoms in a structure. As can be seen from **Figure VIII.10**, the number of peaks in the Patterson map is much greater than the number of atoms. For n atoms in a unit cell there are n^2 peaks in the Patterson map, from which n correspond to self-vectors at the origin, thus, in a Patterson map there are $n^2 - n$ peaks. If the unit cell of a protein crystal contains for instance 5000 non-hydrogen atoms, then the number of Patterson peaks would be 25×10^6 , which clearly gives an uninterpretable Patterson map.

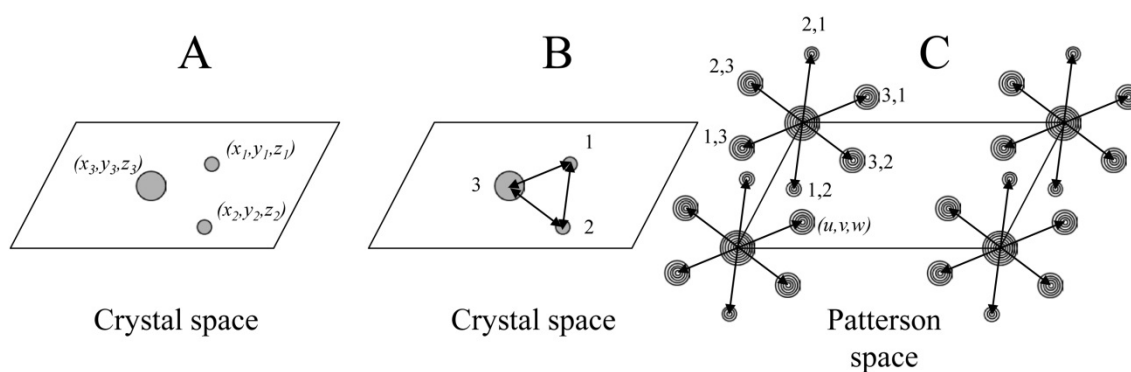


Figure VIII.10: Patterson map derived from a crystal with three atoms.

To obtain this function graphically from the known structure of a crystal (A) all interatomic vectors are plotted (B) and moved parallel to themselves to the origin of the unit cell of the Patterson space (C). The ends of these vectors correspond with the maximum values of the Patterson Function, whose heights are proportional to the product of the atomic numbers of the involved atoms. The positions of these maxima (with coordinates u, v, w) represent the differences between the coordinates of each pair of atoms in the crystal: $u = x_1 - x_2$, $v = y_1 - y_2$, $w = z_1 - z_2$. At the origin (at the corners of the Patterson cell), there is a high maximum corresponding to the interatomic vectors of each atom with itself, that is with coordinates $(0, 0, 0)$. Adapted from: <http://www.xtal.iqfr.csic.es/Cristalografia/index-en.html>

VIII.4.2.2 Model building

After determination of an initial set of phases, an electron density map is calculated. If the initial phases are good, clear secondary structure features can be identified. However, even if the interpretation is easy, model building is still a laborious task. Model building requires fitting, as carefully as possible, the polypeptide chain into the strongest density in the map while maintaining chemical reality, geometric and stereo chemical properties, and using simple common sense. Currently, model building is done using software like COOT²⁸, ARP/wARP²⁹ or with the AutoBuild³⁰ module from the software PHENIX³¹.

The first electron density map is calculated using the experimentally obtained amplitudes, $|F_{obs}|$ and the calculated phases, α_{calc} according to⁵:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l w_{hkl} |F_{obs}| e^{-2\pi i(hx, ky, lz)} + \alpha_{calc}$$

VIII.10

where, w_{hkl} is a weighting factor that accounts for the quality of the determined phases and varies from 0 to 1. A bad phase will have a low weighting factor while a good phase will have a high one. Accordingly, the desired electron-density function is a Fourier sum in which term hkl has amplitude $|F_{obs}|$, which equals the square root of the measured intensity I_{hkl} from the native data set, $(I_{hkl})^{1/2}$. The phase α_{calc} of the same term is calculated from molecular replacement data. The term is weighted by the factor w_{hkl} . This Fourier sum is called an F_{obs} or F_o synthesis.⁵

However, experimentally determined electron density maps are never perfect due to imperfections in the data and phases. As a consequence, even the best of models will have errors in atomic positions, errors in dihedral angles, improper rotamers for side chains, or unacceptable contacts between atoms or chemical groups. In order to improve these initial phases, procedures like **solvent flattening**, **histogram matching**, and **non-crystallographic symmetry** (NCS) averaging are the main techniques used to improve the phases in a process called **density modification**.⁴ For a detailed explanation see Messerschmidt, A (2007)⁹ but, basically⁴:

- **Solvent flattening** – This method relies on the fact that protein crystals typically contain 30–70% solvent, forming channels through the crystal lattice. It works by removing the negative electron density and setting the value of the electron density of solvent regions to a fixed value. Automatic methods are used to define a protein–solvent boundary^{32,33}.
- **Histogram matching** - The density histogram is a probability distribution of values of the electron density sampled at regular intervals (grid points) throughout the three-dimensional map. The histogram matching method calculates the density histogram from the initial set of phases and modifies it so that it takes the form of an expected density histogram.
- **NCS averaging** - When two or more copies of the same molecule are present in the asymmetric unit, NCS averaging can be used. This method averages the density of equivalent positions imposing the same value for each symmetrical molecule.

Even though density modification may provide better phases, a bad map will not get better.

At this point, with a clearer map, we can start to build the molecular model of the protein (**map fitting**) using, for instance, COOT²⁸. The resulting model will most certainly contain many errors and undefined regions. The objective at this point is to correct as many of these errors as possible by “walking” through the amino acid sequence and checking residue-by-residue. Of course, due to the errors mentioned above, this can only be done to some extent.

A problem that arises when building a structure using calculated phases is the possible introduction of errors due to the influence of the model – **model bias**. As phases from the model begin to be the most reliable, they begin to dominate the Fourier sum. In the extreme, the series would contain amplitudes purely from the intensity data and phases purely from the model. In order to compensate for the increased influence of the model phases, these can either be used in conjunction with the measured amplitudes or combined with the experimental phases in order to calculate a new electron density map. Two types of maps can be calculated that reduce the overall model influence by subtracting the calculated structure factor amplitudes ($|F_{calc}|$ or $|F_c|$) to some multiple (usually 1 or 2) of the observed amplitudes ($|F_{obs}|$ or $|F_o|$)^{4,5}:

- the electron density difference map (usually referred to as $F_o - F_c$ map – **Equation VIII.11**),

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l ([|F_o| - |F_c|]) e^{-2\pi i(hx, ky, lz)} + \alpha_{calc}$$

VIII.11

- and the double difference map ($2F_{obs} - F_{calc}$ or $2F_o - F_c$ map – **Equation VIII.12**),

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l (2[|F_o| - |F_c|]) e^{-2\pi i(hx, ky, lz)} + \alpha_{calc}$$

VIII.12

The $F_o - F_c$ map will have both positive and negative density depending on whether the contribution of the observed intensities to the density function, ρ , are larger or smaller than the contribution of the model. In practical terms this means that the map tells us if the model should be adjusted to increase the electron density in a certain region, by adding atoms (in the case of positive density) or, on the other hand (in the case of negative density), if we have to delete some atoms in order to decrease the electron density. For instance, if an amino acid side chain in the model is in the wrong conformation, the $F_o - F_c$ map will exhibit negative density coincident

with the erroneous model side chain and a nearby positive density indicating the correct position. Therefore, the $F_o - F_c$ map emphasizes errors in the current model and removes the influence of the current model so that the original data can “indicate” where the model is wrong. However, if the model still contains many errors, the $F_o - F_c$ map becomes very noisy, full of small positive and negative peaks, difficult to interpret. In order to minimize this, the double difference $2F_o - F_c$ maps are used (**Equation VIII.12**). These are regular electron density maps of the protein, but with reduced bias from the model. Unless the model contains extremely serious errors, this map is positive everywhere, and contours at carefully chosen electron densities resemble a molecular surface.

The newly obtained model phases can be combined with the previous phases and a further model-building cycle can be started with such new and improved electron density maps, thus improving the quality of the maps. After several cycles of model building and crystallographic refinement the atomic model will be complete and the biochemical interpretation can be started.

VIII.4.2.3 Model refinement

Once we have a preliminary model we can refine it against our data, which will improve the phases, thus resulting in clearer maps and therefore better models. We typically repeat this cycle several times until little or no further improvements are obtained. Refinement is the process of systematically altering the model so that the observed and calculated data agree more and more closely- everything goes back to those original reflection intensities, which give us our $|F_{obs}|$ values.

The exact mathematical relationship that connects the model, with the diffraction data is the structure factor given in **Equation VIII.4**. The input to this equation is a set of atomic coordinates - the model - and the output is a set of F_{hkl} . In order to systematically improve the model, the simplest method is the **least-squares refinement**. This method of refinement consists in the minimization of a function, F , which is the sum of the differences between the observed and calculated amplitudes:

$$F = \sum_{hkl} w_{hkl} (|F_o| - |F_c|)_{hkl}^2$$

VIII.13

where, $(|F_o| - |F_c|)_{hkl}^2$ is the squared difference between the calculated and observed amplitudes for the reflection hkl and w_{hkl} is the weighting factor applied to each difference, which is defined as:

$$w_{hkl} = \frac{1}{\sigma_{hkl}^2}$$

VIII. 14

where σ is the standard deviation calculated from the multiple measures of $|F_o|$, thus depending on the reliability of the corresponding measured intensity. However, the data do not usually contain enough measurements of each reflection to determine its standard deviation - for each atom, one refines its position (x,y,z), its temperature factor, ***B-factor***, and its occupancy.

The **temperature factor**, B_j is a measure of how much an atom oscillates around the position specified by the model. Atoms at side chain termini are expected to have a higher degree of freedom of movement than those in the main chain. This movement affects diffraction, thus is it realistic to refine these values. From the temperature factors computed during refinement we gain some insight into the dynamics of our largely static model and also into errors in the model-building process as wrongly placed atoms will exhibit higher *B-factors*, when compared to neighboring atoms.^{4,5}

The **occupancy**, n_j of an atom defines the fraction of asymmetric units where the atom is actually present in its mean position and ranges from 0.0 to 1.0, where intermediate values indicate that it does not occupy this position in all asymmetric units. This parameter can be used to define alternate conformations of amino acid side chains. Like the *B-factor*, the occupancy gives additional information about the dynamics of the protein molecule in the crystal.^{4,5}

Moreover, additional information is incorporated by using certain restraints like bond length, bond angle, and close contact restrictions. These restraints allow variation within a certain limit and are obtained from ideal values established from high-resolution structures of small molecules³⁴⁻³⁶.

The **maximum likelihood** method^{4,9,37} is a more modern approach to fit the data and refine the structure. This method evaluates the probability that the observations (the experimental data) will occur, given a certain model. The model fitting has to be performed so that the probability of the observed data is maximized. Maximum likelihood refinement is particularly useful for incomplete models because it produces residuals that are less biased by the current model than those obtained by least squares⁹. Moreover, it also provides a rigorous formulation for all forms of error in both the model and the observations, and allows incorporation of additional forms of prior knowledge (such as additional phase information) into the probability distributions.

The likelihood (L) of a model represented by a set of observations is the product of the probabilities (P) of all the observations (F_o) of the given model and is defined by:

$$L = \prod_{hkl} P(F_o; F_c)_{hkl}$$

VIII.15

where F_c is the calculated model structure factor. This expression is usually transformed in its logarithmic form which is more tractable:

$$\log L = \sum_i \log P(F_o; F_c)$$

VIII.16

This function will have its maximum when F_o and F_c are equal.

Regardless of the refinement method, the difference between the observed amplitudes of the modified model $|F_o|$ and the calculated $|F_c|$, and thus the quality of the crystallographic model, are expressed by the **R-factor**:

$$R = \frac{\sum_{hkl} w_{hkl} (|F_o| - k|F_c|)}{\sum_{hkl} F_o}$$

VIII.17

where w_{hkl} is the weight applied to the difference and k is a scaling factor. As the model converges to the correct structure, the difference between the amplitudes decreases, as does the *R-factor*. Values of R range from zero, for perfect agreement of calculated and observed intensities, to about 0.6 when a set of measured amplitudes is compared with a set of random amplitudes.⁵

However, the *R-factor* can be artificially decreased by simply increasing the number of adjustable parameters, independently of how many of those parameters are correct – **over fitting**. For instance, a typical problem arises when too many water molecules are fitted to the diffraction data, thus compensating for errors in the model or the data. A related issue is the over interpretation of models by placing too much faith in the accuracy of atomic positions at the particular resolution of the diffraction data.⁹

In order to overcome this situation, Brünger (1992)³⁸ suggested improving this situation with the introduction of a free *R-factor*, R_{free} , which is unbiased by the refinement process. In this method, a random subset of reflections (usually 5-10%), *test set*, T , is set aside from the rest of the reflections, the *working set*, W . It is fundamental to ensure that reflections in the test set are not correlated with reflections in the working test, for instance due to non-crystallographic symmetry (NCS). The refinement is carried out with the working set only, and the R_{free} is calculated with the test set of reflections only:

$$R_{free} = \frac{\sum_{hkl \in T} w_{hkl} (|F_o| - k|F_c|)}{\sum_{hkl \in T} F_o}$$

VIII.18

where $hkl \in T$ means all reflections belonging to test set T . The R value for the reflections in the working set will almost always decrease during refinement, but if the model is truly improving, then the R_{free} for the test set should also decrease.⁶ A comparison of the two parameters, called **cross-validation**, can indicate problems of model over-fitting.

VIII.4.3 Structure validation

Having obtained a reliable structural model of a protein, the next step is to validate the structure. Validation of the macromolecular models is a crucial part of structure determination³⁹. It is important both during structure refinement and at the final stages of data deposition in the PDB. The quality of a structure can be assessed based on a number of indicators such as: ***R-factor*** and **R_{free} ($R_{free}-R$)**, **root-mean-square deviations** from stereochemical standards (rmsd), **Ramachandran plots** and peptide planarity.

The *R-factor* and R_{free} are good indicators of how well the model fits the data. The *R-factor* combines the error inherent in the experimental data and the deviation of the model from reality. Good protein structures should have an *R-factor* < 20%. When the *R-factor* approaches 30% (**Figure VIII.11**), the structure should be regarded with a high degree of reservation because at least some parts of the model may be incorrect.²

The R_{free} is an important validation parameter and should not exceed the *R-factor* by more than ~ 5% (**Figure VIII.11**)². A high R_{free} value may indicate over-fitting of the experimental data, or may result from a serious model defect. For instance, addition of an unreasonable number of water molecules into the noisy features of the solvent region will always lower the ordinary *R-factor*, but will not improve R_{free} .²

In addition to the *R-factor* and R_{free} values, it is necessary to observe various structural parameters that indicate whether the model is chemically, stereochemically, and conformationally reliable. This monitorization is done by the root-mean-square deviations (rmsd) of all the model's bond lengths and angles from geometrical parameters that are considered typical, or represent chemical common sense based on previous experience.^{40,41} Good quality models are expected to have a rmsd_(bond) of ~0.02Å (**Figure VIII.11**). When the rmsd becomes too high (> 0.03Å), it may be indicative that something is wrong with the model.²

Another validation tool (probably one of the most important) is the information on the planarity of the backbone peptides. The peptide planes are usually under very tight

stereochemical restraints and their conformation should be verified by a Ramachandran plot⁴² where the dihedral angles ϕ (defined by the atoms $C_{i-1}-N_i-C\alpha_i-C_i$) and ψ (defined by the atoms $N_i-C\alpha_i-C_i-N_{i+1}$) are plotted against each other for each residue. The data points should lie in the allowed regions of the plot which correspond to energetically favorable secondary structures such as α -helices, β -sheets and defined turn structures. Exceptions are glycine residues, which may occur at any position in the Ramachandran plot due to the lack of a side chain.

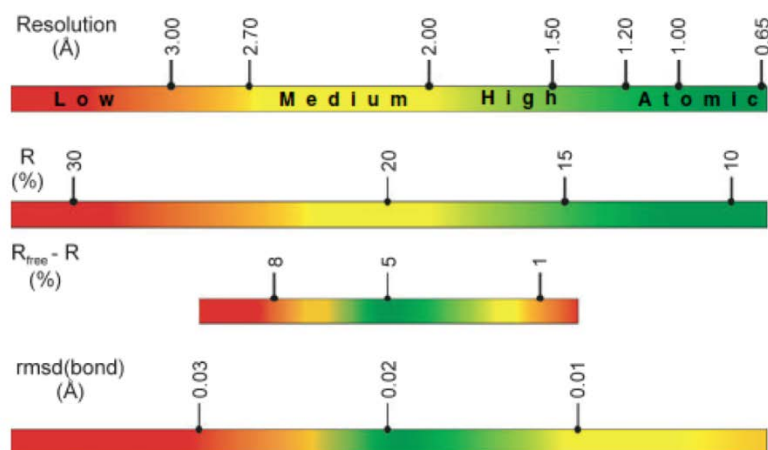


Figure VIII.11: Criteria for assessment of the quality of crystallographic models of macromolecular structures.

For the resolution and R criteria, the lower the value, the better. For $R_{free} - R$ and rmsd there is some optimal value (green area) and severe errors in both directions, although for different reasons. When the difference between R_{free} and R exceeds 7%, it indicates possible over-interpretation of the experimental data. But if it is very low (say below 2%), it strongly suggest that the test data set is not truly 'free', for example, because the test reflections have been compromised in a round of refinement. When rmsd (bonds) is very high, it is an obvious signal of model errors. However, when it is very low (e.g. 0.004Å), it indicates that through too tight restraints the model underwent geometry optimization, rather than refinement driven by the experimental diffraction data. There are different opinions about how rigorous the stereochemical restraints should be, however, because the 'ideal' bond lengths themselves suffer from errors in the order of 0.02Å, it is reasonable to require the model to adhere to them also only at this level. Adapted from Wlodawer, A *et al* (2008).²

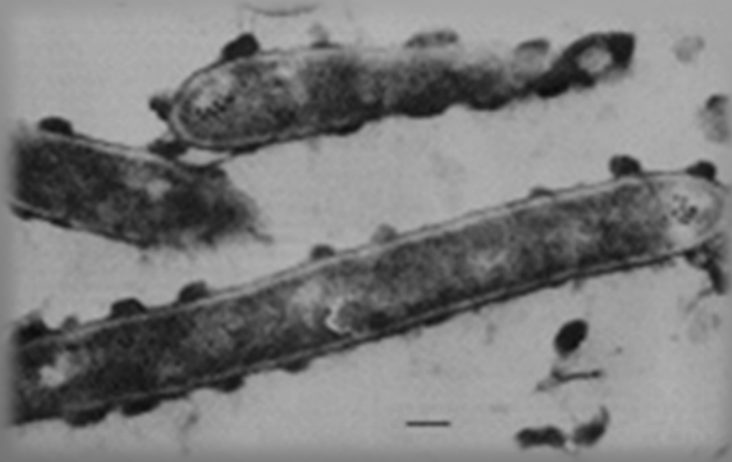
All these tests can be performed by either standalone programs (PROCHECK⁴³, WHAT IF⁴⁴) or Web servers [MolProbity⁴⁵ (<http://molprobity.biochem.duke.edu/>)], which can output highly detailed information that can help correct the model to its best final state.

Once the model is finalized and has passed all the validation tests it is ready to be deposited in the Protein Data Bank (PDB - <http://www.pdb.org/pdb>), where it is further evaluated and validated.

VIII.5 References

1. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181* (4610), 662.
2. Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M., Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *Febs J* **2008**, *275* (1), 1.
3. Moore, P. B.; Steitz, T. A., The ribosome revealed. *Trends Biochem Sci* **2005**, *30* (6), 281.
4. Carvalho, A. L.; Trincao, J.; Romão, M. J., X-Ray Crystallography in Drug Discovery. In *Ligand-macromolecular interactions in drug discovery : methods and protocols*, Roque, A. C. A., Ed. Springer: New York, 2010; pp 31.
5. Rhodes, G., *Crystallography made crystal clear : a guide for users of macromolecular models*. 3rd ed.; Elsevier/Academic Press: Amsterdam ; Boston, 2006; p xxv.
6. Drenth, J.; Mesters, J., *Principles of protein x-ray crystallography*. 3rd ed.; Springer: New York, 2007; p xiv.
7. Lattman, E.; Loll, P., *Protein crystallography : a concise guide*. Johns Hopkins University Press: Baltimore, 2008; p viii.
8. McPherson, A., *Introduction to macromolecular crystallography*. 2nd ed.; Wiley-Blackwell: Hoboken, N.J., 2009; p x.
9. Messerschmidt, A., *X-ray crystallography of biomacromolecules : a practical guide*. Wiley-VCH: Weinheim, 2007; p xiii.
10. Hahn, T., *International Tables for Crystallography -Volume A: Space-Group Symmetry*. 5th ed.; Springer-Verlag: New York, 2005; Vol. A.
11. Rupp, B., *Biomolecular crystallography : principles, practice, and application to structural biology*. Garland Science: New York, 2010; p xxi.
12. Matthews, B. W., Solvent content of protein crystals. *Journal of Molecular Biology* **1968**, *33* (2), 491.
13. Tilley, R. J. D., *Crystals and crystal structures*. John Wiley: Hoboken, NJ, 2006; p xiii.
14. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EACBM Newsletters on Protein Crystallography* **1992**, *26*.
15. Bailey, S., The Ccp4 Suite - Programs for Protein Crystallography. *Acta Crystallogr D* **1994**, *50*, 760.
16. Evans, P., Scaling and assessment of data quality. *Acta Crystallogr D* **2006**, *62*, 72.
17. Dauter, Z.; Jaskolski, M.; Wlodawer, A., Impact of synchrotron radiation on macromolecular crystallography: a personal view. *J Synchrotron Radiat* **2010**, *17* (4), 433.
18. Stout, G. H.; Jensen, L. H., *X-ray structure determination : a practical guide*. 2nd ed.; Wiley: New York, 1989; p xv.
19. Helliwell, J. R., *International Tables for Crystallography -Volume F: Crystallography of Biological Macromolecules*. 5th ed.; Springer-Verlag: New York, 2005; Vol. F.
20. Rossmann, M. G., The molecular replacement method. *Acta Crystallogr A* **1990**, *46* (Pt 2), 73.
21. Rossmann, M. G.; Blow, D. M., The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr* **1962**, *15* (1), 24.
22. DiMaio, F.; Terwilliger, T. C.; Read, R. J.; Wlodawer, A.; Oberdorfer, G.; Wagner, U.; Valkov, E.; Alon, A.; Fass, D.; Axelrod, H. L.; Das, D.; Vorobiev, S. M.; Iwai, H.; Pokkuluri, P. R.; Baker, D., Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **2011**, *473* (7348), 540.
23. Evans, P.; McCoy, A., An introduction to molecular replacement. *Acta Crystallogr D Biol Crystallogr* **2008**, *64* (Pt 1), 1.

24. Patterson, A. L., A Direct Method for the Determination of the Components of Interatomic Distances in Crystals. *Z. Krist* **1935**, *A90*, 517.
25. McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr D* **2005**, *61*, 458.
26. Vagin, A.; Teplyakov, A., MOLREP: an Automated Program for Molecular Replacement. *J Appl Crystallogr* **1997**, *30* (6), 1022.
27. Long, F.; Vagin, A. A.; Young, P.; Murshudov, G. N., BALBES: a molecular-replacement pipeline. *Acta Crystallographica Section D* **2008**, *64* (1), 125.
28. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr D* **2004**, *60*, 2126.
29. Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A., Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* **2008**, *3* (7), 1171.
30. Terwilliger, T. C.; Grosse-Kunstleve, R. W.; Afonine, P. V.; Moriarty, N. W.; Zwart, P. H.; Hung, L.-W.; Read, R. J.; Adams, P. D., Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D* **2008**, *64* (1), 61.
31. Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D* **2010**, *66* (2), 213.
32. Wang, B. C., Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol* **1985**, *115*, 90.
33. Leslie, A., A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Crystallogr A* **1987**, *43* (1), 134.
34. Engh, R. A.; Huber, R., Structure quality and target parameters. In *International Tables for Crystallography*, John Wiley & Sons, Ltd: 2006.
35. Hendrickson, W. A., Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol* **1985**, *115*, 252.
36. Priestle, J., Improved dihedral-angle restraints for protein structure refinement. *J Appl Crystallogr* **2003**, *36* (1), 34.
37. Ten Eyck, L. F.; Watenpugh, K. D., Introduction to refinement. In *International Tables for Crystallography*, John Wiley & Sons, Ltd: 2006.
38. Brünger, A. T., Free R value: Cross-validation in crystallography. In *Methods in Enzymology*, Charles W. Carter Jr, R. M. S., Ed. Academic Press: 1997; Vol. Volume 277, pp 366.
39. Kleywegt, G., Validation of protein crystal structures. *Acta Crystallographica Section D* **2000**, *56* (3), 249.
40. Engh, R.; Huber, R., *Structure quality and target parameters. International Tables for Crystallography -Volume F: Crystallography of Biological Macromolecules*. 5th ed.; Springer-Verlag: New York, 2005; Vol. F.
41. Engh, R. A.; Huber, R., Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A* **1991**, *47* (4), 392.
42. Ramachandran, G. N.; University of Madras., *Crystallography and crystal perfection*. Academic Press: London, New York,, 1963; p 374.
43. Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M., Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J Appl Crystallogr* **1993**, *26*, 283.
44. Vriend, G., WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **1990**, *8* (1), 52.
45. Chen, V. B.; Arendall, W. B., III; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D* **2010**, *66* (1), 12.



Final conclusions

Final conclusions

The general aim of this thesis was to contribute to the understanding of the molecular interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their substrates. Using NMR spectroscopy, X-ray crystallography and computational studies, the CBMs belonging to families 11, 30 and 44 from *C. thermocellum* were systematically studied in order to establish a relationship between structure and specificity. The use of X-ray crystallography and NMR as complementary techniques allowed several questions to be addressed both from the viewpoint of the ligand or the protein, thus enabling a more comprehensive and complete analysis. The results obtained represent a significant improvement in understanding the factors that determine the specificity and the mode of action of Type B CBMs, namely *CtCBM11*, *CtCBM30* and *CtCBM44*, at the molecular level (Chapters II, III and IV).

One of the key findings concerning the structure of *CtCBM11* was the smaller size of the cleft in the crystal structure, when compared to the NMR solution structure. This is probably imposed by the crystal packing and seems to be in the origin of the failed co-crystallization attempts of *CtCBM11* with different cellooligosaccharides. This result shows the importance of the geometry of the binding cleft pointing to a conformation-selection mechanism of ligand recognition and binding for *CtCBM11*. The importance of the geometry and size of the binding cleft and its relation with specific protein/sugar interactions was emphasized by the data that showed that the binding cleft of this protein can accommodate at least 4 sugar units, and that the number of sugar units is fundamental to stabilize the complex. In fact, protein/oligosaccharide contacts are detected for the extremities of cellohexaose that lay outside of the binding cleft and are thought to be responsible for stabilizing the complex when compared to cellotetraose. In the absence of these relatively weak contacts, the entropy of the cellohexaose molecule could lead to a decrease in the affinity. This type of interactions seems to be common in type B CBMs since the same was found for *CtCBM44* and *CtCBM30*. The higher affinity that these proteins display against ligands longer than they can accommodate in the binding cleft seems to be related to the interaction of sugar units that lay outside the binding cleft with polar residues of the protein. These residues flank the binding cleft and make hydrogen bonds with the sugar units at the extremities, thus stabilizing the conformation adopted by these ligands in the binding cleft.

Concerning specific protein/sugar interactions in the binding cleft, the experimentally derived structural models of cellohexaose bound to *CtCBM11* revealed a large number of protein-ligand interactions, including CH- π interactions with Tyr53 and Tyr129. These are

fundamental to stabilize the conformation of ligands in the binding cleft. Additionally, the models show that the C2 and C6 OH groups of the central glucose units make several contacts with the protein, including a number of hydrogen bonds whose presence may dictate the specificity of the protein as it does for other CBMs. These contacts, allied to the rigid conformation of the cleft seem to be determinant to the specificity of the protein. Therefore, only ligands with a methylene group at C5, with the OH group at C2 in an equatorial position and displaying the typical twisted conformation of β -1,4-linked glucans can bind to this protein.

The importance of the tryptophan residues for ligand selection and recognition was also demonstrated for C7CBM44 and C7CBM30. Docking experiments and STD NMR results showed that a combination between the arrangement of the three solvent-exposed tryptophan residues in each protein and interactions of polar residues with the C6 hydroxyl group of the central glucose units are key for defining ligand specificity. The twisted arrangement of the tryptophan residues selects against ligands that do not have this geometry, while the interaction with some C6 OH groups selects against substituted (or without this group) glucose units. It is my belief that this mechanism is common for CBMs that bind to highly decorated ligands.

The association of cellooligosaccharides to CBMs is enthalpically driven with an unfavorable entropic contribution. In this thesis the work performed with C7CBM11 allowed an estimate of a positive variation in protein conformational entropy upon ligand binding to made, therefore, supporting a conformational selection mechanism where ligand conformation is determinant for recognition by a rigid protein. Thus, the origin of the negative binding entropy should be due to the loss of conformational entropy upon complexation with the protein.

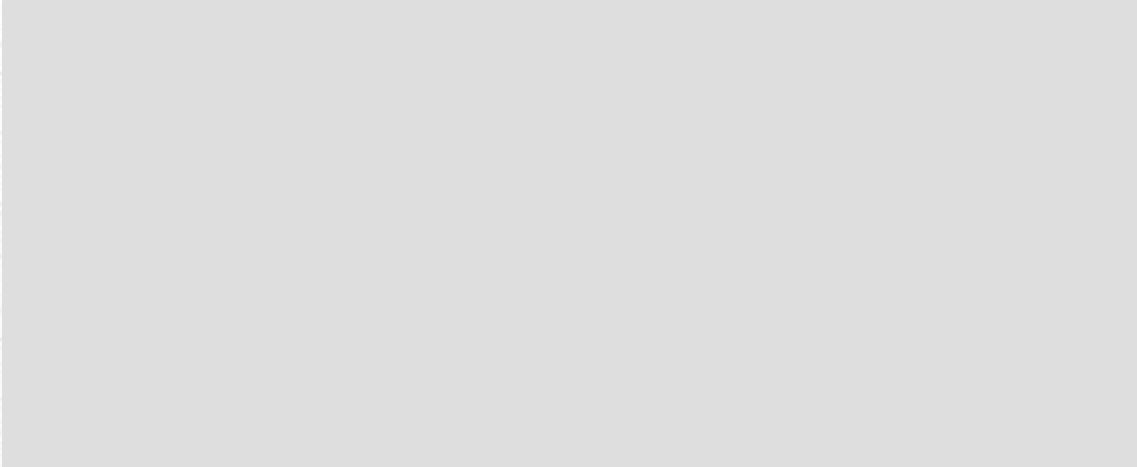
Overall, I have shown through several experiments that binding of cellooligosaccharides to CBMs must occur primarily by a conformational selection mechanism that results from a combination of specific protein/ligand interactions and a rigid protein cleft. This mechanism is common to other CBMs and should be the main determinant of ligand selection. Altogether, the results presented allow an atomistic rationalization of the molecular determinants of ligand specificity and the mechanism by which these proteins are able to distinguish and select its ligands.

In order for this outstanding nanomachine to work properly and at its full capacity, the assembly of the enzymatic components into the cellulosome and the attachment of the latter to the bacterial cell wall are of great significance. To better understand this mechanism I have solved the crystal structure of two type II cohesin-dockerin complexes: one from *C. thermocellum* (Chapter V) and the other from *B. cellulosolvens* (Chapter VI). The first complex is composed by a cohesin bound to a module X and to a dockerin (Coh-XDoc) and it was solved to a resolution of 1.98 Å. The overall structure is very similar to the SdbA type II Coh-XDoc structures (PDB code: 2b59) and it reveals that both helix 1 and helix 3 of the dockerin interact with the cohesin module. This, allied to the lack of internal symmetry of both helices indicates

that this complex does not show the dual binding mode predicted for other complexes. The structure also exposed the possible role of module X. The high number of contacts it makes with both the dockerin and the cohesin indicates that its presence is fundamental for the stability of the complex.

The structure of 11th SdbA type II cohesin-dockerin (Coh11-Doc) complex from *B. cellulosolvens*, was solved to a resolution of 1.90 Å and is the first cohesin-dockerin complex ever determined from *B. cellulosolvens*. Also for the first time, it reveals the 3D structure of a type II dockerin from this organism and, more important, it indicates the possibility of an alternate binding mode between the cohesin and the dockerin, in a similar way to what is proposed for the type I interaction in *C. thermocellum*. Like other dockerins that show a dual binding mode, the type II dockerin of *B. cellulosolvens* also shows an internal two-fold symmetry between helix 1 and 3. Most remarkable is the fact that in this complex the dockerin is rotated 180° when compared to other native cohesin-dockerin complexes determined so far. This feature confers a large degree of plasticity to the complex and has profound implications at the level of the current understanding of cellulosome architecture and assembly.

Taken together, the structures of the two type II cohesin–dockerin complexes provide valuable information about the atomic interactions that mediate complex assembly. Altogether our findings represent an important development on the overall understanding of this phenomenal mega-Dalton machine termed Cellulosome.



Appendix A

A.1 Molecular biology reagents

Table A.1: Preparation of the Luria-Bertani (LB) medium.

<i>Component</i>	<i>Quantity for 1 L (g)</i>
Yeast extract	5
Sodium chloride (NaCl)	10
Bactotryptone	10

Table A.2: M9 minimal medium composition

<i>Component</i>	<i>Quantity</i>
M9 salt solution	100 mL/L
¹³C Glucose	0.4% (4g/L)
Magnesium sulphate (MgSO₄·7H₂O)	2 mL/L of a 1M solution
Iron chloride (FeCl₃·7H₂O)	2 mL/L of a 12 mg/L solution
Thiamine	1 mL/L of a 1mg/L solution

Table A.3: M9 salt solution

<i>Component</i>	<i>Quantity</i>
¹⁵N Ammonium chloride (¹⁵NH₄Cl)	1g/L
Potassium dihydrogen phosphate (KH₂PO₄)	60g/L
Disodium hydrogen phosphate (Na₂KPO₄)	120g/L

Adjust the pH to 7.5

Filter with 0.45 µm membrane pore filters

Keep at 4°C

Table A.4: Preparation of the working/washing buffer.

<i>Component</i>	<i>Quantity</i>
Hepes	11,92g/L (50mM)
NaCl	58,44g/L (1M)
Imidazole	0,681g/L (10mM)
CaCl₂	0,55g/L (5mM)

Adjust the pH with NaOH (pH=7.5)

Filter with 0.45 µm membrane pore filters

Keep at 4°C

Table A.5: Preparation of the elution buffer.

<i>Component</i>	<i>Quantity</i>
Hepes	11,9155g/L (50mM)
NaCl	58,44g/L (1M)
Imidazole	20.42g/L (300mM)
CaCl₂	0,555g/L (5mM)

Adjust the pH with NaOH (pH=7.5)

Filter with 0.45 µm membrane pore filters and keep at 4°C

Table A.6: Composition of the SDS-PAGE stacking gel

<i>Component</i>	<i>Stacking gel</i>		
	<i>2,5 ml</i>	<i>5ml</i>	<i>10 ml</i>
H₂O (ml)	1,4	2,8	5,6
Acrylamide 30% (ml)	0,35	0,7	1,4
SDS 10% (ml)	0,25	0,5	1
Tris buffer, pH=6,8 (ml)	0,5	1	2
Ammonium persulfate (NH₄)₂S₂O₈ 10% (µL)	25	50	100
TEMED (µL)	5	10	20

Table A.7: Composition of the SDS-PAGE resolution gel

<i>Component</i>	<i>Resolution gel</i>					
	<i>7%</i>	<i>7,50%</i>	<i>8%</i>	<i>10%</i>	<i>12%</i>	<i>14%</i>
H₂O (ml)	1,2	1,1	1,025	0,7	0,35	0,05
Acrylamide 30% (ml)	1,15	1,25	1,325	1,5	2	2,3
SDS 10% (ml)	0,5	0,5	0,5	0,5	0,5	0,5
Glycerol 50% (ml)	0,5	0,5	0,5	0,5	0,5	0,5
Tris buffer, pH=8,8 (ml)	1,65	1,65	1,65	1,65	1,65	1,65
Ammonium persulfate (NH₄)₂S₂O₈ 10% (µl)	50	50	50	50	50	50
TEMED (µL)	5	5	5	5	5	5

Table A.8: SDS-PAGE 5x sample buffer

<i>Component</i>	<i>Quantity</i>
Tris-HCl, pH 7	200 mM
Glycerol	20% (v/v)
SDS	10% (w/v)
Bromophenol blue	0.05% (w/v)

β-mercapto-ethanol	10 mM
--	-------

Table A.9: Coomassie brilliant blue

<i>Component</i>	<i>Quantity</i>
Coomassie brilliant blue	0.1%
Acetic acid	10%
Methanol	50%

Table A.10: BCA assay working reagent.

<i>N° of assays</i>	<i>Reagent A (ml)</i>	<i>Reagent B (μL)</i>	<i>Total volume (ml)</i>
4	4	80	4.08
8	8	160	8.16
9	9.5	190	9.69
12	12.5	250	12.75

Volume of Working Reagent is dependent on how many blanks, BSA protein standards and unknown samples are to be assayed.

Table A.11: BCA assay standards.

<i>Tube n°</i>	<i>BSA 1mg/ml (μL)</i>	<i>Buffer (μL)</i>	<i>BCA working reagent (mL)</i>	<i>[BSA] (μg/mL)</i>
1	0	50	1	0
2	10	40	1	200
3	20	30	1	400
4	30	20	1	600
5	40	10	1	800
6	50	0	1	1000

A.2 The T7lac promoter

The pET System^{1,2} is a very powerful system developed for the cloning and expression of recombinant proteins in *E. coli* where target genes are cloned in pET plasmids under control of strong bacteriophage T7 transcription. The expression is induced by providing a source of T7 RNA polymerase in the host cell. This polymerase is so selective and active that, when fully induced, almost all of the cell's resources are converted to target gene expression; the desired product can comprise more than 50% of the total cell protein a few hours after induction.²

Because T7 RNA polymerase is extremely promoter-specific and transcribes only DNA downstream of a T7 promoter, one way to control expression is to use vectors that contain a lac operator sequence just downstream of the T7 promoter and the natural promoter and coding sequence for the lac repressor (*lacI*), oriented so that the T7lac and *lacI* promoters diverge. Like this, transcription of the T7 RNA polymerase gene by the host polymerase is repressed by the *lac* repressor and, at the same time, transcription of the target gene by any T7 RNA polymerase that is made is also blocked by the T7lac promoter. Because the host cells do not contain the T7 polymerase, in order to initiate the expression process it is necessary to add IPTG, which will induce the expression of the T7 polymerase, which rapidly begins to transcribe the desired gene.

A.3 The pET21a vector

A pET vector is a bacterial plasmid designed to enable the quick production of a large quantity of any desired protein when activated. This plasmid (**Figure A.1**) contains several important elements - a *lacI* gene which codes for the lac repressor protein, a T7 promoter which is specific to only T7 RNA polymerase (not bacterial RNA polymerase) and also does not occur anywhere in the prokaryotic genome, a lac operator which can block transcription, an ampicillin resistance gene, and a ColE1 origin of replication³.

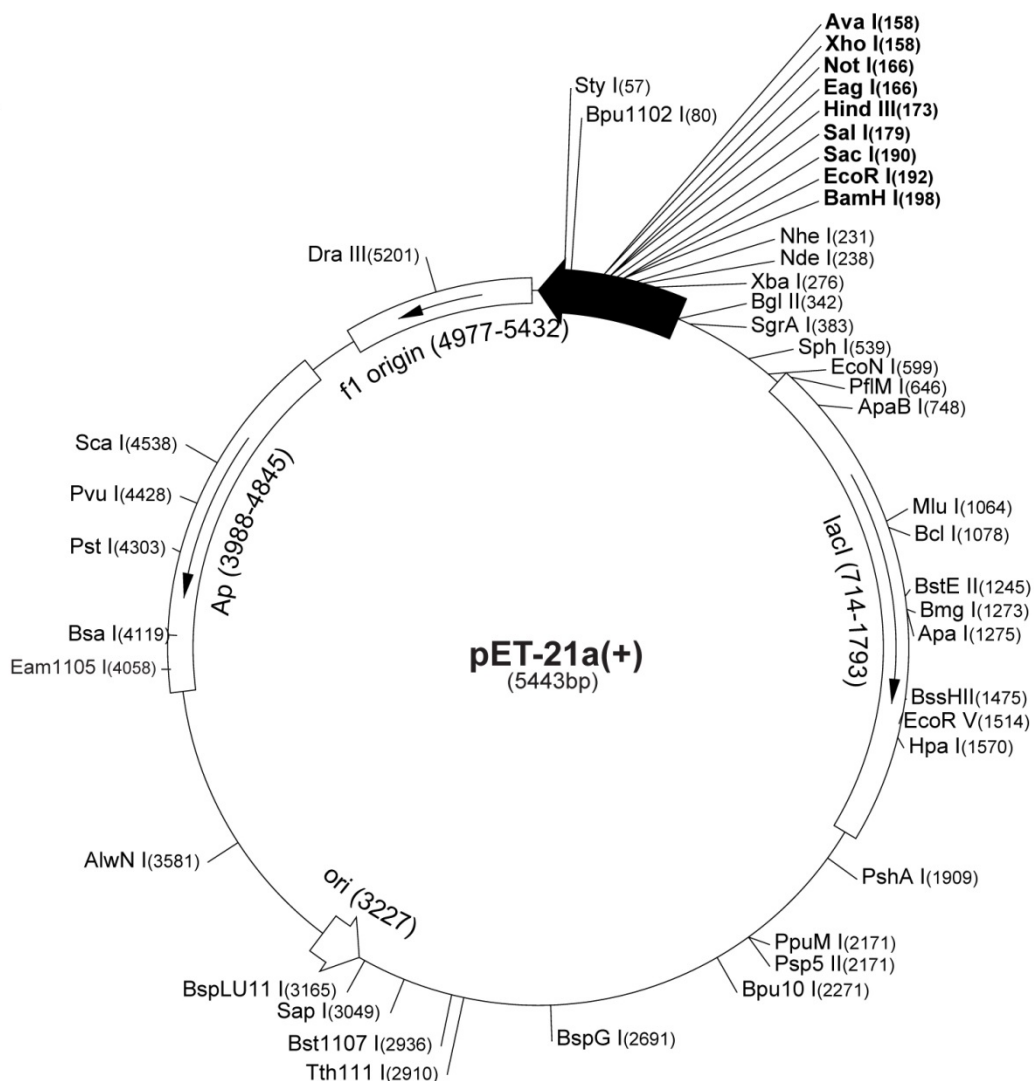
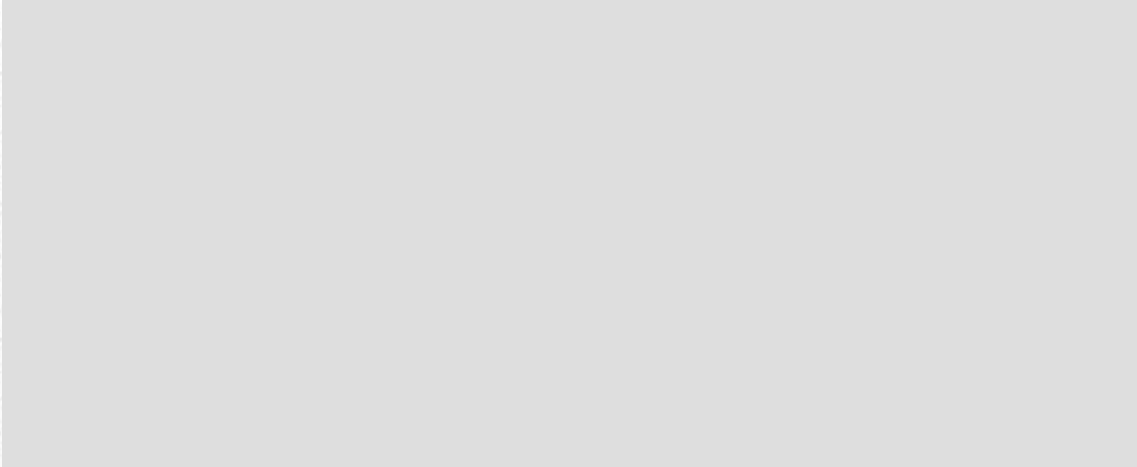


Figure A.1: The pET-21a(+) vector.

A.4 References

1. Studier, F. W.; Moffatt, B. A., Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology* **1986**, *189* (1), 113.
2. Novagen, pET System Manual, 11th edition. 2006.
3. Blaber, M. Molecular Biology and Biotechnology - Lecture 25. <http://www.mikeblaber.org/oldwine/bch5425/bch5425.htm>.



Appendix B

B.1 Screening and cryo-protectant solutions

Table B. 1: Set of solutions used in an initial screening, according to the method developed by Jancarik & Kim, (1991)¹

<i>N°</i>	<i>Salt</i>	<i>Buffer</i>	<i>Precipitant</i>	<i>pH</i>
1	0.2M CaCl ₂	0.1M Acetate	30% MPD	4.4
2	1.0M Na/K-tartrate	0.1M MES	-----	6.7
3	-----	-----	0.4M Ammonium phosphate	6.5
4	-----	0.1M Tris HCl	3.0M Ammonium sulfate	7.2
5	0.2M Sodium citrate	0.1M HEPES	30% MPD	7.2
6	0.2M MgCl ₂	0.1M Acetate	30% PEG 4000	4.3
7	1.2M Sodium citrate	0.1M HEPES	-----	7.7
8	0.2M Sodium citrate	-----	2.0M Ammonium sulfate	5.5
9	0.2M Ammonium acetate	0.1M Citrate	30% PEG 400	6.4
10	-----	0.1M Acetate	1.5M Ammonium phosphate	5.9
11	0.2M Ammonium sulfate	0.1M HEPES	2.0M Na/K-phosphate	6.2
12	0.2M Sodium citrate	0.1M Tris HCl	20% PEG 400	8.7
13	0.2M CaCl ₂	0.1M HEPES	25% PEG 4000	7.2
14	0.1M MgCl ₂	0.1M MES	30% PEG 8000	6.4
15	0.2M Lithium sulfate	0.1M Citrate	30% PEG 4000	5.9
16	1.0M Lithium sulfate	0.2M Acetate	-----	4
17	0.2M Ammonium phosphate	0.1M Tris HCl	30% MPD	7.4
18	0.2M Ammonium acetate	0.1M Tris HCl	2.0M Na/K-phosphate	6.3
19	0.1M Ammonium sulfate	0.1M Citrate	30% PEG 8000	6.1
20	-----	0.1M MES	30% MPD	6.4
21	0.2M MgCl ₂	0.1M HEPES	30% PEG 400	7
22	0.2M Sodium Acetate	0.1M Tris HCl	30% PEG 4000	8.9
23	-----	0.1M Tris HCl	1.0M Na/K-tartrate	8.9
24	0.2M CaCl ₂	0.1M Tris HCl	-----	8.5
25	0.5M Ammonium acetate	0.1M Citrate	30% MPD	6.5
26	2.0M Sodium Acetate	0.1M MES	-----	4.5
27	0.2M Na/K-tartrate	0.1M MES	30% PEG 8000	6.6
28	1.0M Na/K-tartrate	0.1M HEPES	-----	7.6
29	0.2M Ammonium sulfate	0.1M Acetate	30% PEG 400	4.6
30	0.1M Ammonium sulfate	0.1M HEPES	20% PEG 4000	6.7
31	2.0M Ammonium sulfate	0.1M MES	-----	6.8
32	0.2M NaCl	0.1M MES	30% Ethanol	6.3

33	0.2M MgCl ₂	0.1M HEPES	30% Ethanol	7
34	0.2M Ammonium acetate	0.1M Tris HCl	30% Ethanol	8.2
35	0.2M CaCl ₂	0.1M Acetate	30% Ethanol	4.4
36	0.2M Sodium Acetate	0.1M HEPES	30% Ethanol	7.2
37	0.2M MgCl ₂	0.1M HEPES	30% 2-propanol	7.2
38	-----	0.1M Cacodylate	30% MPD	6.5
39	-----	0.1M Acetate	2.0M Sodium formate	5.4
40	0.2M Citrate	0.1M Cacodylate	40% 2-propanol	6.8
41	-----	0.1M HEPES	20% PEG 4000 / 10% 2-propanol	7.4
42	-----	0.1M HEPES	1.0M Lithium sulfate	7.7
43	0.2M Lithium sulfate	0.1M Tris HCl	30% PEG 4000	8.8
44	0.2M Ammonium sulfate	0.1M Cacodylate	30% PEG 6000	6.4
45	-----	0.1M Acetate	1.5M Sodium Acetate	6.3
46	0.1M Citrate	-----	-----	6.3
47	-----	-----	1.0M Ammonium phosphate	7.3
48	-----	0.1M HEPES	4.0M Sodium formate	7.9
49	-----	-----	1.2M citrate	7.9
50	-----	-----	0.4M Na/K-tartrate	3.5
51	-----	0.1M Cacodylate	30% PEG 4000	6.8
52	0.2M Ammonium acetate	0.1M Citrate	1.4M Sodium Acetate	6.1
53	0.2M Ammonium acetate	0.1M Acetate	30% PEG 4000	5.5
54	0.2M CaCl ₂	0.1M HEPES	28% PEG 400	7.2
55	0.2M Ammonium sulfate	0.1M Cacodylate	30% PEG 8000	6.8
56	0.2M Acetate-Mg	0.1M Cacodylate	20% PEG 8000	6.3
57	0.2M Ammonium acetate	0.1M Tris HCl	30% 2-propanol	8.3
58	0.2M Ammonium sulfate	0.1M Acetate	25% PEG 4000	4.7
59	0.2M Acetate-Mg	0.1M Cacodylate	30% MPD	6.5
60	0.2M CaCl ₂	0.1M Acetate	20% 2-propanol	4.3
61	-----	0.1M Imidazole	1.0M Sodium Acetate	7.4
62	0.2M Sodium citrate	0.1M HEPES	20% 2-propanol	7.5
63	0.2M Sodium Acetate	0.1M Cacodylate	30% PEG 8000	6.7
64	0.2M Ammonium sulfate	-----	30% PEG 8000	6.1
65	0.2M Ammonium sulfate	-----	30% PEG 4000	5.8
66	-----	0.1M HEPES	1.6M Na/K-phosphate	6.8
67	-----	0.1M Tris HCl	8% PEG 8000	8.3
68	-----	0.1M Acetate	8% PEG 4000	4.5
69	-----	0.1M HEPES	2% PEG 400 / 2.0M Ammonium sulfate	7.7
70	-----	0.1M Citrate	20% 2-propanol / 20% PEG 4000	5.8
71	0.05M Potassium phosphate	-----	20% PEG 8000	8.9

72	-----	-----	30% PEG 1500	~2.9
73	-----	-----	0.2M formate-Mg	
74	0.2M Acetate-Zn	0.1M Cacodylate	18% PEG 8000	5.6
75	0.2M Acetate-Ca	0.1M Cacodylate	18% PEG 8000	6.2
76	-----	0.1M Acetate	2.0M Ammonium sulfate	4.6
77	-----	0.1M Tris HCl	2.0M Ammonium sulfate	8.4
78	1.0M Lithium sulfate	-----	2% PEG 8000	5.6
79	1.0M Lithium sulfate	-----	15% PEG 8000	
80	0.2M Ammonium acetate	0.1M Citrate	20% PEG 4000 / 20% 2-propanol	

Table B.2: Possible cryo-protectant solutions as developed by Garman & Mitchell, (1996)²

<i>N^o</i>	<i>Salt</i>	<i>Buffer</i>	<i>Precipitant</i>	<i>Glycerol (v/v)</i>
1	0.02M CaCl ₂	0.1M Acetate	30% MPD	0
2	-----	-----	0.4M Na/K-tartrate	35
3	-----	-----	0.4M Ammonium phosphate	35
4	-----	0.1M Tris HCl	2.0M Ammonium sulfate	25
5	0.2M Sodium citrate	0.1M HEPES	30% MPD	0
6	0.2M MgCl ₂	0.1M Tris HCl	30% PEG 4000	20
7	-----	0.1M Cacodylate	1.4M Sodium Acetate	30
8	0.2M Sodium citrate	0.1M Cacodylate	30% 2-propanol	30
9	0.2M Ammonium acetate	0.1M Citrate	30% PEG 4000	15
10	0.2M Ammonium acetate	0.1M Acetate	30% PEG 4000	15
11	-----	0.1M Citrate	1.0M Ammonium phosphate	30
12	0.2M MgCl ₂	0.1M HEPES	30% 2-propanol	10
13	0.2M Sodium citrate	0.1M Tris HCl	30% PEG 400	0
14	0.1M CaCl ₂	0.1M HEPES	28% PEG 400	5
15	0.2M Ammonium sulfate	0.1M Cacodylate	30% PEG 8000	15
16	-----	0.1M HEPES	1.5M Lithium sulfate	25
17	0.2M Lithium sulfate	0.1M Tris HCl	30% PEG 4000	15
18	0.2M Acetate-Mg	0.1M Cacodylate	20% PEG 8000	20
19	0.1M Ammonium acetate	0.1M Tris HCl	30% 2-propanol	20
20	0.2M Ammonium sulfate	0.1M Acetate	25% PEG 4000	20
21	0.2M Acetate-Mg	0.1M Cacodylate	30% MPD	0
22	0.2M Sodium Acetate	0.1M Tris HCl	30% PEG 4000	15
23	0.2M MgCl ₂	0.1M HEPES	30% PEG 400	0
24	0.2M CaCl ₂	0.1M Acetate	20% 2-propanol	30
25	-----	0.1M Imidazole	1.0M Sodium Acetate	30
26	2.0M Ammonium acetate	0.1M Citrate	30% MPD	0

27	0.2M citrate-Na	0.1M MES	20% 2-propanol	30
28	0.2M Sodium Acetate	0.1M Cacodylate	30% PEG 8000	15
29	-----	0.1M HEPES	0.8M Na/K-tartrate	35
30	0.2M Ammonium sulfate	-----	30% PEG 8000	15
31	0.2M Ammonium sulfate	-----	30% PEG 8000	15
32	-----	-----	2.0M Ammonium sulfate	25
33	-----	-----	4.0M Sodium formate	10
34	-----	0.1M Acetate	2.0M Sodium formate	30
35	-----	0.1M HEPES	1.6M Na/K-phosphate	25
36	-----	0.1M Tris	8% PEG 8000	35
37	-----	0.1M Acetate	8% PEG 4000	30
38	-----	0.1M HEPES	1.4M citrate-Na	10
39	-----	0.1M HEPES	2% PEG 400 / 2.0M Ammonium sulfate	15
40	-----	0.1M Citrate	20% 2-propanol / 20% PEG 4000	5
41	-----	0.1M HEPES	20% PEG 4000 / 10% 2-propanol	15
42	0.05M Potassium phosphate	-----	20% PEG 8000	20
43	-----	-----	30% PEG 1500	20
44	-----	-----	0.2M formate-Mg	50
45	0.2M Acetate-Zn	0.1M Cacodylate	18% PEG 8000	20
46	0.2M Acetate-Ca	0.1M Cacodylate	18% PEG 8000	20
47	-----	0.1M Acetate	2.0M Ammonium sulfate	20
48	-----	0.1M Tris HCl	2.0M Ammonium phosphate	20
49	1.0M Lithium sulfate	-----	2% PEG 8000	20
50	0.5M Lithium sulfate	-----	15% PEG 8000	20

B.2 Space groups

B.2.1 $P2_1$

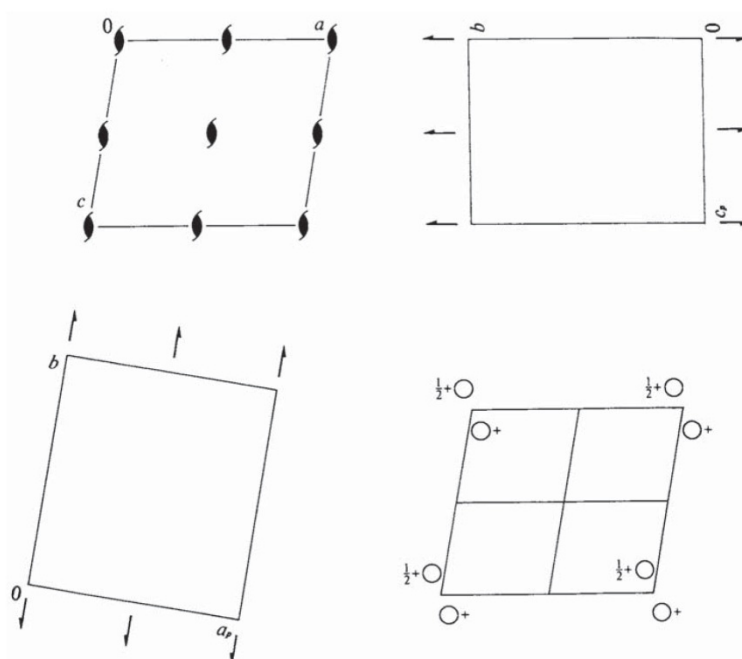
From the International Tables for Crystallography – Volume A³, among several useful information, we see that the $P2_1$ space group is the number 4 (International Union of Crystallography, IUCr number), the Laue class is $2/m$ and belongs to the monoclinic crystal system. Moreover, we see that this space group is non-centrosymmetric with a primitive Bravais lattice (P) and the Patterson symmetry is $P1\ 2/m\ 1$. The symmetry operations are (x,y,z) and $(-x, \frac{1}{2}+y, -z)$, yielding the general multiplicity of 2 ($Z=2$). Additionally, due to the space group symmetry, we find that for space group $P2_1$, reflections $0k0$ with k odd are absent.

 $P2_1$ C_2^2

2

Monoclinic

No. 4

 $P12_11$ Patterson symmetry $P12_1/m1$ UNIQUE AXIS b Origin on 2_1 Asymmetric unit $0 \leq x \leq 1; 0 \leq y \leq 1; 0 \leq z \leq \frac{1}{2}$

Symmetry operations

- (1) 1 (2) $2(0, \frac{1}{2}, 0) \ 0, y, 0$

CONTINUED

No. 4

 $P2_1$ **Generators selected** (1); $t(1,0,0)$; $t(0,1,0)$; $t(0,0,1)$; (2)**Positions**Multiplicity,
Wyckoff letter,
Site symmetry

Coordinates

Reflection conditions

2 a 1(1) x, y, z (2) $\bar{x}, y + \frac{1}{2}, \bar{z}$ General:
 $0k0 : k = 2n$ **Symmetry of special projections**Along [001] $p1g1$ $\mathbf{a}' = \mathbf{a}$ $\mathbf{b}' = \mathbf{b}$ Origin at $0, 0, z$ Along [100] $p11g$ $\mathbf{a}' = \mathbf{b}$ $\mathbf{b}' = \mathbf{c}$ Origin at $x, 0, 0$ Along [010] $p2$ $\mathbf{a}' = \mathbf{c}$ $\mathbf{b}' = \mathbf{a}$ Origin at $0, y, 0$ **Maximal non-isomorphic subgroups****I** [2] $P1(1)$ 1**IIa** none**IIb** none**Maximal isomorphic subgroups of lowest index****IIc** [2] $P12_11$ ($\mathbf{c}' = 2\mathbf{c}$ or $\mathbf{a}' = 2\mathbf{a}$ or $\mathbf{a}' = \mathbf{a} + \mathbf{c}$, $\mathbf{c}' = -\mathbf{a} + \mathbf{c}$) ($P2_1, 4$); [3] $P12_11$ ($\mathbf{b}' = 3\mathbf{b}$) ($P2_1, 4$)**Minimal non-isomorphic supergroups****I** [2] $P2_1/m$ (11); [2] $P2_1/c$ (14); [2] $P222_1$ (17); [2] $P2_12_12$ (18); [2] $P2_12_12$ (19); [2] $C222_1$ (20); [2] $Pmc2_1$ (26); [2] $Pca2_1$ (29); [2] $Pmn2_1$ (31); [2] $Pna2_1$ (33); [2] $Cmc2_1$ (36); [2] $P4_1$ (76); [2] $P4_2$ (78); [3] $P6_1$ (169); [3] $P6_3$ (170); [3] $P6_3$ (173)**II** [2] $C121$ ($C2, 5$); [2] $A121$ ($C2, 5$); [2] $I121$ ($C2, 5$); [2] $P121$ ($\mathbf{b}' = \frac{1}{2}\mathbf{b}$) ($P2, 3$)

Figure B.1: Information of the space group $P2_1$, taken from the International Tables for Crystallography – Volume A.³

B.2.2 C2

From the International Tables for Crystallography – Volume A³, among several useful information, we see that the $C121$ space group (Hermann-Mauguin notation) is the number 5 (IUCr number), the Laue class is $2/m$ and belongs to the monoclinic crystal system. Moreover, we see that this space group is non-centrosymmetric with a C-face-centered Bravais lattice and the Patterson symmetry is $C1\ 2/m\ 1$. The symmetry operations are (x, y, z) and $(-x, y, -z)$. The coordinates of the first molecule are $(0, 0, 0)$ and the one of the second molecule are $(\frac{1}{2}, \frac{1}{2}, 0)$. Because these molecules are duplicated by a C-face-centered translation, the general multiplicity is 4 ($Z=4$). Additionally, due to Bravais centering, we find that for space group $C121$, only reflections in which h and k are simultaneously even or odd are observed. Moreover, if h becomes negative, those reflections are not observed as well.

C_2

C_2^3

2

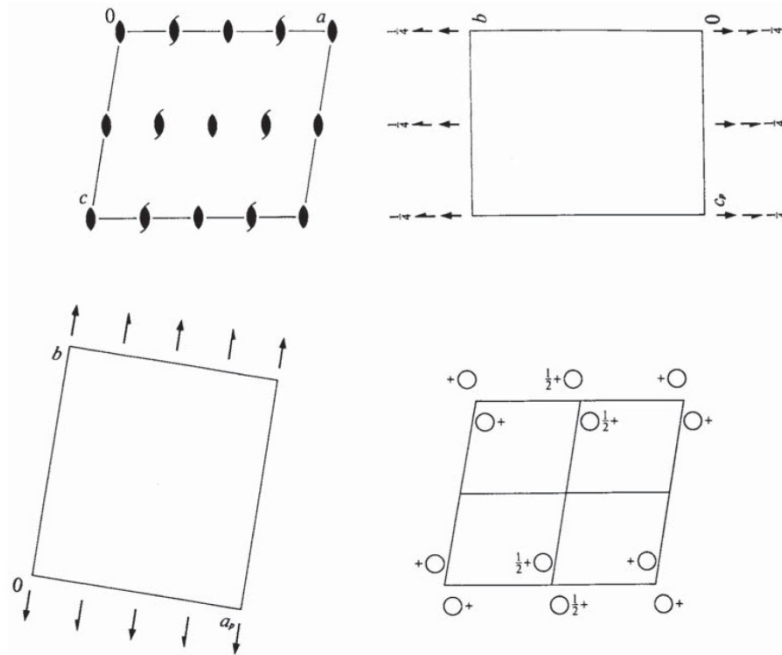
Monoclinic

No. 5

$C121$

Patterson symmetry $C12/m1$

UNIQUE AXIS b , CELL CHOICE 1



Origin on 2

Asymmetric unit $0 \leq x \leq \frac{1}{2}; 0 \leq y \leq \frac{1}{2}; 0 \leq z \leq 1$

Symmetry operations

For $(0,0,0)+$ set

(1) 1 (2) $2\ 0,y,0$

For $(\frac{1}{2},\frac{1}{2},0)+$ set

(1) $i(\frac{1}{2},\frac{1}{2},0)$ (2) $2(0,\frac{1}{2},0)\ \frac{1}{2},y,0$

CONTINUED

No. 5

C2

Generators selected (1); $t(1,0,0)$; $t(0,1,0)$; $t(0,0,1)$; $t(\frac{1}{2}, \frac{1}{2}, 0)$; (2)

Positions

Multiplicity, Wyckoff letter, Site symmetry	Coordinates	Reflection conditions
	$(0, 0, 0) + (\frac{1}{2}, \frac{1}{2}, 0) +$	General:
4 <i>c</i> 1	(1) x, y, z (2) \bar{x}, y, \bar{z}	$hkl : h + k = 2n$ $h0l : h = 2n$ $0kl : k = 2n$ $hk0 : h + k = 2n$ $0k0 : k = 2n$ $h00 : h = 2n$
		Special: no extra conditions
2 <i>b</i> 2	$0, y, \frac{1}{2}$	
2 <i>a</i> 2	$0, y, 0$	

Symmetry of special projections

Along [001] $c1m1$
 $\mathbf{a}' = \mathbf{a}$ $\mathbf{b}' = \mathbf{b}$
 Origin at $0, 0, z$

Along [100] $p11m$
 $\mathbf{a}' = \frac{1}{2}\mathbf{b}$ $\mathbf{b}' = \mathbf{c}$
 Origin at $x, 0, 0$

Along [010] $p2$
 $\mathbf{a}' = \mathbf{c}$ $\mathbf{b}' = \frac{1}{2}\mathbf{a}$
 Origin at $0, y, 0$

Maximal non-isomorphic subgroups

I	[2] $C1(P1, 1)$	1+
IIa	[2] $P12_1(P2_1, 4)$ [2] $P121(P2, 3)$	1; $2 + (\frac{1}{2}, \frac{1}{2}, 0)$ 1; 2
IIb	none	

Maximal isomorphic subgroups of lowest index

IIc [2] $C121(\mathbf{c}' = 2\mathbf{c} \text{ or } \mathbf{a}' = \mathbf{a} + 2\mathbf{c}, \mathbf{c}' = 2\mathbf{c})(C2, 5)$; [3] $C121(\mathbf{b}' = 3\mathbf{b})(C2, 5)$

Minimal non-isomorphic supergroups

I	[2] $C2/m(12)$; [2] $C2/c(15)$; [2] $C22_2(20)$; [2] $C222(21)$; [2] $F222(22)$; [2] $I222(23)$; [2] $I2_12_1(24)$; [2] $Amm2(38)$; [2] $Aem2(39)$; [2] $Ama2(40)$; [2] $Aea2(41)$; [2] $Fmm2(42)$; [2] $Fdd2(43)$; [2] $Imm2(44)$; [2] $Iba2(45)$; [2] $Ima2(46)$; [2] $I4(79)$; [2] $I4_1(80)$; [2] $I4_2(82)$; [3] $P312(149)$; [3] $P321(150)$; [3] $P3_112(151)$; [3] $P3_21(152)$; [3] $P3_212(153)$; [3] $P3_21(154)$; [3] $R32(155)$
II	[2] $P121(\mathbf{a}' = \frac{1}{2}\mathbf{a}, \mathbf{b}' = \frac{1}{2}\mathbf{b})(P2, 3)$

Figure B.2: Information of the space group $C121$, taken from the International Tables for Crystallography – Volume A.³

B.2.3 P2

From the International Tables for Crystallography – Volume A³, we see that the $P121$ space group (Hermann-Mauguin notation) is the number 3 (IUCr number), the Laue class is $2/m$ and belongs to the monoclinic crystal system. This space group is also non-centrosymmetric with a primitive Bravais lattice (P) and the Patterson symmetry is $P1\ 2/m\ 1$. The symmetry operations are the same as for the $P2_1$ space group, which is (x, y, z) and $(-x, \frac{1}{2} + y, -z)$, again yielding a multiplicity of 2 ($Z=2$). There are no conditions limiting reflections neither due to Bravais centering nor due to space group symmetry.

$P2$

C_2^1

2

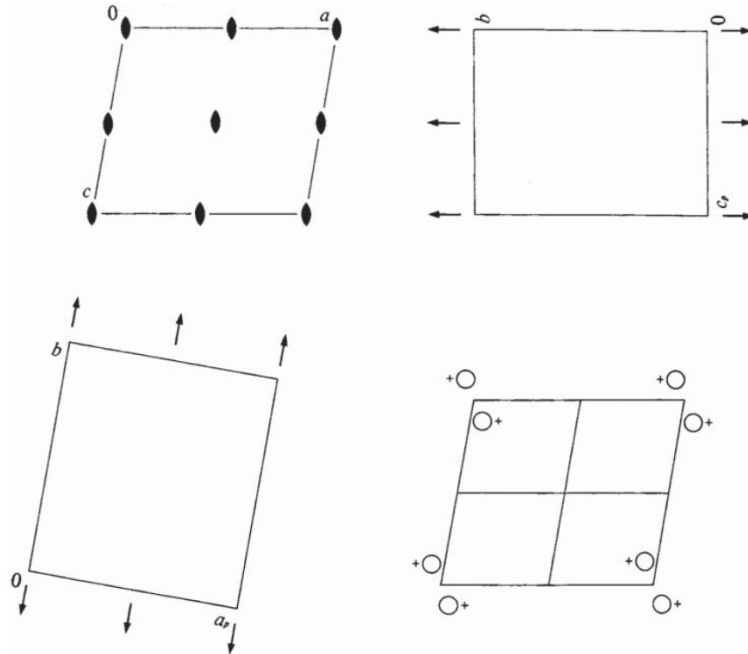
Monoclinic

No. 3

$P121$

Patterson symmetry $P12/m1$

UNIQUE AXIS b



Origin on 2

Asymmetric unit $0 \leq x \leq 1; 0 \leq y \leq 1; 0 \leq z \leq \frac{1}{2}$

Symmetry operations

(1) 1 (2) 2 $0, y, 0$

CONTINUED

No. 3

P 2

Generators selected (1); $t(1,0,0)$; $t(0,1,0)$; $t(0,0,1)$; (2)**Positions**

Multiplicity, Wyckoff letter, Site symmetry		Coordinates	Reflection conditions
2	<i>e</i>	1) x, y, z 2) \bar{x}, y, \bar{z}	General: no conditions Special: no extra conditions
1	<i>d</i>	$\frac{1}{2}, y, \frac{1}{2}$	
1	<i>c</i>	$\frac{1}{2}, y, 0$	
1	<i>b</i>	$0, y, \frac{1}{2}$	
1	<i>a</i>	$0, y, 0$	

Symmetry of special projections

Along [001] $P1m1$ $\mathbf{a}' = \mathbf{a}_z$ $\mathbf{b}' = \mathbf{b}$ Origin at 0, 0, z	Along [100] $P11m$ $\mathbf{a}' = \mathbf{b}$ $\mathbf{b}' = \mathbf{c}_y$ Origin at $x, 0, 0$	Along [010] $P2$ $\mathbf{a}' = \mathbf{c}$ $\mathbf{b}' = \mathbf{a}$ Origin at 0, $y, 0$
--	--	--

Maximal non-isomorphic subgroups

I	[2] $P1(1)$ 1
IIa	none
IIb	[2] $P12_1$ ($\mathbf{b}' = 2\mathbf{b}$) ($P2$, 4); [2] $C121$ ($\mathbf{a}' = 2\mathbf{a}, \mathbf{b}' = 2\mathbf{b}$) ($C2$, 5); [2] $A121$ ($\mathbf{b}' = 2\mathbf{b}, \mathbf{c}' = 2\mathbf{c}$) ($C2$, 5); [2] $F121$ ($\mathbf{a}' = 2\mathbf{a}, \mathbf{b}' = 2\mathbf{b}, \mathbf{c}' = 2\mathbf{c}$) ($C2$, 5)

Maximal isomorphic subgroups of lowest index

IIc	[2] $P121$ ($\mathbf{b}' = 2\mathbf{b}$) ($P2$, 3); [2] $P121$ ($\mathbf{c}' = 2\mathbf{c}$ or $\mathbf{a}' = 2\mathbf{a}$ or $\mathbf{a}' = \mathbf{a} + \mathbf{c}, \mathbf{c}' = -\mathbf{a} + \mathbf{c}$) ($P2$, 3)
------------	--

Minimal non-isomorphic supergroups

I	[2] $P2/m$ (10); [2] $P2/c$ (13); [2] $P222$ (16); [2] $P222$ (17); [2] $P2_12_12_1$ (18); [2] $C222$ (21); [2] $Pmm2$ (25); [2] $Pcc2$ (27); [2] $Pma2$ (28); [2] $Pnc2$ (30); [2] $Pba2$ (32); [2] $Pnn2$ (34); [2] $Cmm2$ (35); [2] $Ccc2$ (37); [2] $P4$ (75); [2] $P4_1$ (77); [2] $P4$ (81); [3] $P6$ (168); [3] $P6_1$ (171); [3] $P6_2$ (172)
II	[2] $C121$ ($C2$, 5); [2] $A121$ ($C2$, 5); [2] $I121$ ($C2$, 5)

Figure B.3: Information of the space group $P121$, taken from the International Tables for Crystallography – Volume A.³

B.3 References

1. Jancarik, J.; Kim, S.-H., Sparse matrix sampling: a screening method for crystallization of proteins. *J Appl Crystallogr* **1991**, *24* (4), 409.
2. Garman, E. F.; Mitchell, E. P., Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions. *J Appl Crystallogr* **1996**, *29*, 584.
3. Hahn, T., *International Tables for Crystallography -Volume A: Space-Group Symmetry*. 5th ed.; Springer-Verlag: New York, 2005; Vol. A.



Appendix C

C.1 Interaction with cellooligosaccharides

Table C.1: Affinity constants (K_a) and thermodynamic parameters (ΔG , ΔH and $T\Delta S$) determined for the interaction of CtCBM11 with celohexaose at 25 and 50 °C.

<i>Res.</i>	$K_a \times 10^{-4} (M^{-1})$ 25°C	ΔG ($kcal.mol^{-1}$) 25°C	$K_a \times 10^{-4} (M^{-1})$ 50°C	ΔG ($kcal.mol^{-1}$) 50°C	$\Delta H (kcal.mol^{-1})$	$T\Delta S$ ($kcal.mol^{-1}$) (25°C)
Y22			1.75±0.25	-6.27±0.14		
G24	2.36±0.91	-5.96±0.40	1.53±0.42	-6.18±0.28	3.32±0.92	2.64±0.00
T49	3.31±0.32	-6.16±0.10	2.35±0.52	-6.46±0.23	2.60±0.99	3.55±0.00
G52	4.42±0.73	-6.33±0.17	2.87±0.07	-6.59±0.03	3.29±1.07	3.04±0.00
Y53	2.99±0.51	-6.10±0.17	1.92±0.16	-6.33±0.08	3.38±0.67	2.72±0.00
W54	0.71±0.29	-5.24±0.44	1.19±0.20	-6.02±0.17	-3.99±2.11	9.24±0.01
G55			3.37±0.54	-6.69±0.16		
T56			2.97±0.55	-6.61±0.19		
V57	0.71±0.67	-5.25±1.72	1.23±0.23	-6.04±0.19	-4.19±2.98	9.44±0.04
Y58			2.19±0.14	-6.41±0.07		
S59	0.22±0.22	-4.57±2.33	1.18±0.19	-6.01±0.17	-12.66±16.57	17.22±0.05
R86	5.26±2.08	-6.43±0.42	2.19±0.56	-6.41±0.26	6.69±1.21	-0.26±0.00
F87	3.03±0.96	-6.11±0.33	0.78±0.09	-5.75±0.12	10.35±1.62	-4.24±0.00
M88	2.90±0.82	-6.08±0.29	0.93±0.15	-5.86±0.16	8.73±0.99	-2.65±0.00
I89	3.22±0.66	-6.14±0.21	1.36±0.04	-6.10±0.03	6.62±1.34	-0.48±0.00
E91	1.42±0.01	-5.66±0.01	2.22±0.63	-6.42±0.29	-3.42±2.19	9.08±0.01
S93	4.74±1.87	-6.37±0.42	1.32±0.39	-6.09±0.30	9.76±0.90	-3.39±0.00
H102	1.74±1.65	-5.78±1.84	1.36±0.20	-6.10±0.15	1.89±0.87	3.89±0.04
R125	3.31±0.54	-6.16±0.16	1.52±0.16	-6.18±0.10	5.95±0.46	0.21±0.00
R126	5.00±0.62	-6.40±0.12	1.06±0.22	-5.95±0.21	11.86±0.66	-5.46±0.00
D128	0.24±0.23	-4.59±3.46	1.02±0.20	-5.92±0.20	-11.18±1.84	15.77±0.08
Y129	5.20±1.10	-6.43±0.22	1.83±0.33	-6.29±0.18	7.99±0.26	-1.57±0.00
Q130	3.99±0.56	-6.27±0.14	3.24±0.44	-6.66±0.14	1.58±0.05	4.69±0.00
N144	3.34±1.35	-6.16±0.43	0.94±0.37	-5.87±0.42	9.68±0.06	-3.52±0.00
I145	4.98±2.18	-6.40±0.47	2.42±0.59	-6.47±0.25	5.53±1.66	0.87±0.00
I148	3.52±2.58	-6.20±0.93	4.50±1.61	-6.87±0.37	-1.88±4.28	8.07±0.01
H149	1.79±0.59	-5.80±0.34	1.19±0.16	-6.02±0.14	3.11±1.59	2.69±0.00
F150	4.03±0.52	-6.27±0.13	1.88±0.13	-6.31±0.07	5.83±0.44	0.45±0.00
M151	2.57±0.43	-6.01±0.17	1.35±0.17	-6.10±0.12	4.92±0.33	1.08±0.00
Y152			1.84±0.17	-6.30±0.09		

Table C.2: Affinity constants (K_a) and binding energy (ΔG) determined for the interaction of CrCBM11 with cellotetraose at 25 °C.

<i>Res.</i>	$K_a \times 10^{-4} (M^{-1})$ 25°C	$\Delta G (kcal.mol^{-1})$ 25°C
K32	0.62±0.01	-5.17±0.02
T49	1.49±0.14	-5.69±0.18
R86	5.13±3.57	-6.42±1.72
S93	1.85±1.27	-5.81±1.69
I94	0.69±0.19	-5.23±0.58
G100	5.45±1.25	-6.45±0.47
V104	0.07±0.01	-3.89±0.15
F123	0.72±0.24	-5.25±0.71
R124	1.62±0.78	-5.74±1.05
Y129	2.33±0.56	-5.95±0.49
N144	2.50±0.98	-5.99±0.82
H149	0.73±0.22	-5.26±0.62
I148	0.67±0.15	-5.21±0.47
M151	0.62±0.01	-5.17±0.02

C.2 Relaxation data

Table C.3: Longitudinal and transverse relaxation rates, ^1H - ^{15}N steady state NOE values and R_2/R_1 ratios for the free *CtCBM11* at 25 °C

Residue	25 °C free							
	NOE		R_1		R_2		$\tau_{c,i} (R_2/R_1)$	
	NOE	Error	R_1	Error	R_2	Error	R_2/R_1	Error
M1								
A2								
S3								
A4								
V5	0.71	0.05	0.83	0.02	9.21	0.41	11.04	0.06
G6	0.38	0.01	1.38	0.01	8.31	0.11		
E7	0.71	0.01	1.33	0.00	15.19	0.14	11.46	0.01
K8	0.67	0.03	1.49	0.01	6.41	0.09		
M9	0.80	0.01	1.36	0.01	12.05	0.12	8.89	0.01
L10	0.84	0.02	1.27	0.01	12.71	0.13	9.98	0.02
D11	0.86	0.01	1.32	0.01	10.51	0.13	7.99	0.02
D12	0.81	0.01	1.37	0.01	12.19	0.14	8.92	0.02
F13	0.88	0.01	1.31	0.01	12.73	0.10	9.69	0.01
E14	0.85	0.01	1.29	0.00	13.49	0.10	10.47	0.01
G15	0.78	0.01	1.27	0.00	10.80	0.07	8.52	0.01
V16	0.67	0.01	1.21	0.00	10.86	0.06	8.97	0.01
L17	0.61	0.01	1.33	0.00	9.22	0.07		
N18	0.65	0.01	1.35	0.01	12.04	0.12	8.95	0.01
W19	0.72	0.01	1.22	0.01	10.65	0.11	8.69	0.01
G20	0.83	0.01	1.27	0.01	11.95	0.13	9.41	0.02
S21	0.76	0.01	1.31	0.00	8.95	0.09		
Y22	0.84	0.01	1.42	0.01	12.72	0.08	8.95	0.01
S23	0.87	0.01	1.30	0.00	10.25	0.08	7.89	0.01
G24	0.88	0.01	1.27	0.01	12.11	0.13	9.55	0.01
E25	0.87	0.01	1.44	0.01	10.77	0.12	7.46	0.02
G26	0.87	0.01	1.46	0.01	11.90	0.10	8.14	0.01
A27	0.83	0.01	1.34	0.01	12.52	0.11	9.32	0.01
K28	0.81	0.01	1.38	0.01	11.38	0.08	8.24	0.01
V29	0.81	0.01	1.34	0.00	13.14	0.10	9.80	0.01
S30	0.81	0.01	1.28	0.01	11.87	0.07	9.25	0.01
T31	0.82	0.01	1.28	0.00	10.16	0.08	7.93	0.01
K32	0.76	0.01	1.33	0.00	11.40	0.07	8.58	0.01
I33	0.81	0.01	1.27	0.00	12.20	0.07	9.63	0.01
V34	0.80	0.01	1.29	0.00	11.78	0.07	9.14	0.01
S35	0.88	0.01	1.25	0.00	12.23	0.09	9.76	0.01
G36	0.84	0.01	1.32	0.01	13.23	0.13	9.99	0.01

K37	0.81	0.02	1.33	0.01	14.16	0.20	10.66	0.02
T38	0.80	0.01	1.29	0.01	11.07	0.13	8.60	0.02
G39								
N40	0.84	0.01	1.21	0.01	12.71	0.15	10.52	0.02
G41	0.85	0.01	1.28	0.00	12.54	0.11	9.83	0.01
M42	0.87	0.01	1.38	0.01	12.66	0.12	9.15	0.01
E43	0.81	0.01	1.41	0.01	12.08	0.10	8.55	0.01
V44	0.84	0.01	1.33	0.01	12.42	0.13	9.32	0.02
S45	0.84	0.01	1.36	0.01	11.64	0.12	8.53	0.02
Y46	0.83	0.01	1.39	0.01	11.55	0.10	8.29	0.01
T47	0.80	0.01	1.27	0.00	11.95	0.12	9.44	0.01
G48	0.78	0.01	1.33	0.01	10.96	0.09	8.27	0.01
T49	0.83	0.01	1.28	0.01	12.90	0.11	10.10	0.01
T50	0.83	0.02	1.19	0.01	13.54	0.27	11.34	0.03
D51	0.79	0.01	1.37	0.00	13.42	0.11	9.77	0.01
G52	0.81	0.01	1.28	0.00	11.37	0.08	8.89	0.01
Y53	0.85	0.01	1.35	0.01	11.47	0.12	8.47	0.02
W54	0.83	0.01	1.39	0.01	14.29	0.17	10.27	0.02
G55	0.83	0.01	1.36	0.01	13.11	0.14	9.63	0.02
T56	0.89	0.01	1.37	0.01	13.25	0.14	9.70	0.02
V57	0.83	0.01	1.37	0.01	11.66	0.15	8.52	0.02
Y58	0.88	0.02	1.36	0.01	13.09	0.17	9.66	0.02
S59	0.83	0.01	1.32	0.01	9.81	0.11	7.43	0.02
L60	0.78	0.01	1.31	0.01	11.89	0.16	9.06	0.02
P61								
D62	0.75	0.01	1.29	0.00	10.61	0.09	8.25	0.01
G63	0.76	0.01	1.11	0.00	12.26	0.06	11.01	0.01
D64	0.81	0.01	1.44	0.00	11.86	0.07	8.26	0.01
W65	0.82	0.02	1.21	0.01	14.31	0.24		
S66	0.88	0.02	1.50	0.01	11.90	0.23	7.92	0.03
K67								
W68								
L69								
K70	0.84	0.03	1.18	0.02	12.15	0.30	10.29	0.04
I71	0.84	0.01	1.45	0.01	13.57	0.14	9.33	0.02
S72	0.87	0.01	1.40	0.01	12.62	0.10	8.99	0.01
F73	0.87	0.01	1.46	0.00	11.22	0.08	7.70	0.01
D74	0.82	0.01	1.35	0.00	11.79	0.09	8.76	0.01
I75	0.86	0.01	1.36	0.00	12.01	0.09	8.83	0.01
K76	0.82	0.01	1.44	0.00	12.93	0.10	8.95	0.01
S77	0.81	0.01	1.25	0.00	11.56	0.10	9.22	0.01
V78	0.85	0.02	1.42	0.01	14.67	0.15	10.37	0.02
D79	0.72	0.01	1.66	0.01	10.45	0.08		
G80	0.70	0.01	1.47	0.01	9.94	0.10		

S81	0.77	0.01	1.67	0.01	11.60	0.09	6.93	0.01
A82	0.67	0.01	1.37	0.01	10.88	0.11	7.94	0.02
N83	0.54	0.01	1.42	0.00	9.72	0.06		
E84	0.85	0.01	1.25	0.01	12.50	0.18	10.00	0.02
I85	0.89	0.02	1.29	0.00	12.30	0.09	9.54	0.01
R86	0.81	0.01	1.28	0.01	11.04	0.13	8.65	0.02
F87	0.80	0.01	1.22	0.01	11.39	0.14	9.33	0.02
M88	0.89	0.01	1.39	0.01	13.84	0.15	9.95	0.02
I89	0.87	0.02	1.31	0.01	12.20	0.16	9.30	0.02
A90	0.80	0.01	1.38	0.01	10.93	0.12	7.94	0.02
E91	0.83	0.02	1.35	0.01	11.03	0.14	8.15	0.02
K92	0.85	0.02	1.43	0.01	13.81	0.15	9.69	0.02
S93	0.82	0.02	1.31	0.01	11.28	0.10	8.58	0.01
I94	0.85	0.02	1.16	0.01	10.70	0.20	9.23	0.03
N95	0.74	0.01	1.31	0.00	10.68	0.08	8.15	0.01
G96	0.59	0.01	1.43	0.01	10.25	0.10		
V97	0.57	0.01	1.32	0.00	9.15	0.07		
G98	0.58	0.01	1.49	0.01	10.29	0.10		
D99	0.63	0.01	1.22	0.00	10.70	0.09		
G100	0.82	0.02	1.27	0.01	12.54	0.19	9.84	0.02
E101								
H102								
W103	0.80	0.02	1.33	0.01	12.92	0.15	9.74	0.02
V104	0.89	0.02	1.36	0.01	13.20	0.21	9.74	0.02
Y105	0.86	0.01	1.37	0.01	10.55	0.13	7.69	0.02
S106	0.77	0.01	1.40	0.01	9.99	0.11	7.16	0.02
I107	0.84	0.01	1.37	0.01	13.24	0.12	9.70	0.01
T108	0.83	0.01	1.28	0.00	12.87	0.12	10.06	0.01
P109								
D110	0.83	0.01	1.38	0.01	12.03	0.10	8.71	0.01
S111	0.78	0.01	1.47	0.01	9.93	0.11		
S112	0.76	0.01	1.33	0.00	11.33	0.05	8.49	0.01
W113	0.86	0.01	1.31	0.00	12.69	0.08	9.71	0.01
K114	0.77	0.01	1.35	0.01	12.10	0.09	9.00	0.01
T115	0.81	0.01	1.27	0.01	12.86	0.17	10.14	0.02
I116	0.84	0.01	1.35	0.01	12.57	0.11	9.31	0.01
E117	0.78	0.01	1.35	0.00	12.13	0.09	8.99	0.01
I118	0.79	0.01	1.28	0.01	10.68	0.14	8.36	0.02
P119								
F120	0.83	0.03	1.26	0.01	10.83	0.19	8.57	0.03
S121	0.86	0.01	1.38	0.01	13.60	0.10	9.89	0.01
S122	0.88	0.01	1.31	0.01	11.83	0.12	9.05	0.02
F123	0.84	0.02	1.46	0.01	14.17	0.13	9.68	0.01
R124	0.85	0.02	1.22	0.01	12.42	0.13	10.14	0.01

R125	0.81	0.01	1.24	0.01	11.40	0.15	9.16	0.02
R126	0.85	0.02	1.27	0.01	11.41	0.22	8.95	0.03
L127	0.84	0.02	1.35	0.01	10.93	0.18	8.08	0.02
D128	0.84	0.01	1.39	0.01	12.46	0.13	8.95	0.01
Y129	0.86	0.01	1.36	0.01	12.71	0.13	9.33	0.01
Q130	0.89	0.02	1.22	0.01	12.34	0.20	10.10	0.03
P131								
P132								
G133	0.88	0.02	1.33	0.01	15.34	0.24	11.53	0.02
Q134	0.91	0.02	1.29	0.01	11.73	0.22	9.06	0.03
D135	0.90	0.02	1.28	0.01	8.28	0.33		
M136	0.81	0.03	1.37	0.02				
S137	0.87	0.01	1.35	0.01	11.75	0.13	8.69	0.02
G138	0.89	0.02	1.36	0.01	14.08	0.20	10.32	0.02
T139								
L140	0.82	0.02	1.37	0.01	12.32	0.20	8.97	0.02
D141	0.83	0.02	1.15	0.01	11.98	0.25	10.43	0.03
L142	0.79	0.02	1.26	0.01	12.35	0.18	9.77	0.02
D143	0.80	0.01	1.45	0.00	14.25	0.12	9.85	0.01
N144	0.75	0.01	1.42	0.01	12.26	0.10	8.62	0.01
I145	0.85	0.02	1.26	0.01	11.95	0.19	9.50	0.03
D146	0.84	0.01	1.31	0.01	10.68	0.15	8.17	0.02
S147	0.86	0.01	1.37	0.01	11.00	0.13	8.03	0.02
I148	0.83	0.01	1.39	0.01	12.15	0.12	8.74	0.02
H149	0.88	0.02	1.32	0.01	13.21	0.15	9.97	0.02
F150	0.88	0.01	1.28	0.01	11.61	0.14	9.10	0.02
M151	0.82	0.01	1.28	0.01	12.08	0.14	9.44	0.02
Y152	0.87	0.02	1.47	0.01	12.83	0.22	8.71	0.03
A153	0.81	0.03	1.26	0.01	9.08	0.27		
N154	0.82	0.01	1.14	0.00	9.26	0.11	8.09	0.02
N155	0.90	0.03	1.27	0.01	10.54	0.17	8.30	0.02
K156	0.85	0.01	1.28	0.00	11.50	0.09	8.98	0.01
S157	0.80	0.01	1.29	0.01	9.62	0.14	7.44	0.02
G158	0.83	0.01	1.37	0.01	11.85	0.10	8.65	0.01
K159	0.81	0.01	1.40	0.01	11.93	0.11	8.51	0.01
F160	0.85	0.01	1.36	0.00	11.19	0.06	8.22	0.01
V161	0.85	0.01	1.28	0.00	12.12	0.09	9.47	0.01
V162	0.92	0.02	1.60	0.01	6.86	0.10		
D163	0.86	0.01	1.39	0.01	12.74	0.11	9.15	0.01
N164	0.85	0.01	1.30	0.00	11.03	0.12	8.49	0.01
I165	0.85	0.01	1.40	0.01	12.71	0.06	9.06	0.01
K166	0.86	0.01	1.33	0.01	10.80	0.13	8.13	0.02
L167	0.73	0.01	1.29	0.01	11.64	0.12	9.00	0.01
I168	0.81	0.03	1.30	0.01	15.30	0.20		

G169	0.79	0.04	1.41	0.02	17.42	0.46
A170	0.60	0.01	1.34	0.01	13.44	0.11
L171	0.48	0.01	1.64	0.01	7.98	0.06
E172	0.19	0.00	1.53	0.00	6.75	0.05

Table C.4: Longitudinal and transverse relaxation rates, ^1H - ^{15}N steady state NOE values and R_2/R_1 ratios for the bound *Ct*CBM11 at 25 °C

Residue	25 °C bound							
	NOE		R_1		R_2		$\tau_{c,i}(R_2/R_1)$	
	NOE	Error	R_1	Error	R_2	Error	R_2/R_1	Error
M1								
A2								
S3								
A4								
V5								
G6	0.41	0.08	1.17	0.05	6.70	0.20		
E7	0.66	0.04	1.32	0.01	16.67	0.54		
K8	0.35	0.08	1.30	0.02	12.08	0.24		
M9	0.90	0.06	1.30	0.02	12.66	0.29	9.78	0.04
L10	0.71	0.05	1.25	0.02	11.48	0.18	9.17	0.03
D11	0.77	0.04	1.25	0.01	9.48	0.14		
D12	0.92	0.05	1.38	0.01	10.51	0.14	7.62	0.02
F13	0.89	0.04	1.42	0.02	12.83	0.20	9.03	0.03
E14	0.87	0.04	1.31	0.01	11.83	0.17	9.04	0.02
G15	0.69	0.03	1.26	0.01	9.82	0.09	7.80	0.02
V16	0.69	0.03	1.25	0.01	10.20	0.09	8.16	0.01
L17	0.58	0.04	1.40	0.01	9.55	0.12		
N18	0.58	0.06	1.23	0.02	9.61	0.22		
W19	0.76	0.05	1.20	0.02	11.52	0.16	9.62	0.03
G20	1.01	0.08	1.24	0.02	11.13	0.24	8.98	0.04
S21	0.72	0.04	1.28	0.02	10.08	0.14	7.90	0.03
Y22								
S23	0.75	0.03	1.33	0.01	11.21	0.11	8.44	0.02
G24	0.84	0.05	1.29	0.02	11.79	0.23	9.16	0.04
E25	0.88	0.06	1.39	0.02	11.13	0.16	7.99	0.03
G26	0.90	0.07	1.28	0.03	12.12	0.25	9.50	0.04
A27	0.86	0.03	1.33	0.01	12.54	0.14	9.44	0.02
K28	0.92	0.05	1.29	0.01	12.28	0.12	9.48	0.02
V29	0.86	0.04	1.31	0.01	10.66	0.10	8.13	0.02
S30								
T31								
K32	0.75	0.05	1.33	0.01	10.76	0.18	8.10	0.03
I33	0.79	0.04	1.31	0.01	9.75	0.10		
V34	0.70	0.04	1.33	0.02				

S35	0.83	0.03	1.29	0.02	11.75	0.15	9.09	0.02
G36	0.84	0.05	1.31	0.01	10.99	0.14	8.37	0.02
K37	0.83	0.10	1.34	0.04	11.95	0.44	8.92	0.07
T38	0.75	0.06	1.22	0.02	11.56	0.21	9.49	0.03
G39								
N40	0.82	0.05	1.20	0.02	11.87	0.19	9.87	0.03
G41	0.88	0.05	1.26	0.01	11.73	0.17	9.33	0.03
M42	0.88	0.05	1.31	0.01	11.75	0.23	8.98	0.03
E43	0.74	0.04	1.40	0.01	11.78	0.18	8.41	0.03
V44	0.95	0.09	1.39	0.02	11.20	0.27	8.08	0.04
S45	0.82	0.03	1.39	0.01	11.64	0.09	8.36	0.02
Y46	0.61	0.06	1.40	0.02	13.24	0.32		
T47	0.79	0.04	1.23	0.01	11.19	0.16	9.09	0.02
G48	0.23	0.16	1.28	0.06	7.63	0.38		
T49	0.83	0.05	1.24	0.01	13.05	0.19		
T50	0.83	0.07	1.18	0.02	10.66	0.42	9.03	0.06
D51	0.72	0.03	1.31	0.01	11.90	0.11	9.06	0.02
G52	0.85	0.05	1.27	0.02	11.60	0.15	9.12	0.02
Y53	0.77	0.05	1.30	0.02	11.35	0.27	8.73	0.04
W54	0.78	0.04	1.28	0.03	11.56	0.22	9.05	0.04
G55	0.96	0.07	1.33	0.02	11.94	0.23	8.95	0.04
T56	0.89	0.05	1.34	0.02	11.15	0.17	8.30	0.03
V57	0.78	0.08	1.31	0.03	12.08	0.28	9.19	0.04
Y58	1.05	0.14	1.36	0.05	11.24	0.45	8.28	0.08
S59	0.54	0.12	1.07	0.03	8.74	0.41		
L60	0.67	0.06	1.25	0.02	11.19	0.22	8.98	0.03
P61								
D62	0.70	0.04	1.27	0.01	10.47	0.13	8.25	0.02
G63	0.76	0.02	1.29	0.01	9.86	0.06	7.62	0.01
D64	0.83	0.04	1.40	0.01	11.74	0.14	8.37	0.02
W65	0.69	0.07	1.18	0.02	13.89	0.36		
S66	0.84	0.05	1.36	0.02	11.69	0.20	8.59	0.03
K67								
W68								
L69								
K70	0.81	0.09	1.21	0.03	12.24	0.43	10.07	0.06
I71	0.92	0.06	1.36	0.02	12.30	0.26	9.05	0.04
S72	0.86	0.04	1.47	0.01	12.14	0.15	8.26	0.02
F73	0.89	0.05	1.45	0.02	11.39	0.14	7.86	0.03
D74	0.85	0.04	1.38	0.01	11.88	0.15	8.62	0.02
I75	0.86	0.04	1.37	0.01	11.88	0.14	8.70	0.02
K76	0.78	0.05	1.36	0.01	10.94	0.14	8.06	0.02
S77	0.89	0.07	1.38	0.02	11.40	0.17	8.26	0.03
V78	0.62	0.08	1.43	0.03	11.85	0.30		

D79	0.72	0.06	1.37	0.02	9.01	0.18		
G80	0.90	0.20	1.27	0.07	11.05	0.79	8.69	0.13
S81	0.72	0.06	1.38	0.03	9.40	0.19		
A82	0.67	0.13	1.35	0.03	10.53	0.39	7.80	0.06
N83	0.51	0.04	1.34	0.01	9.49	0.10		
E84	0.68	0.08	1.20	0.02	10.28	0.25	8.60	0.05
I85	0.80	0.08	1.31	0.02	12.29	0.24	9.37	0.03
R86	0.85	0.06	1.40	0.02	12.89	0.25	9.20	0.04
F87	0.77	0.05	1.33	0.03	11.22	0.20	8.45	0.04
M88	0.73	0.06	1.32	0.02	13.05	0.26	9.88	0.04
I89	0.78	0.05	1.26	0.02	10.41	0.18	8.28	0.03
A90	0.88	0.07	1.32	0.02	11.67	0.28	8.87	0.04
E91	0.92	0.10	1.29	0.03	11.70	0.35	9.04	0.05
K92	0.80	0.07	1.26	0.02	10.39	0.26	8.25	0.04
S93	0.98	0.10	1.32	0.02	11.52	0.24	8.70	0.03
I94	0.72	0.07	1.32	0.02	12.36	0.42	9.35	0.05
N95	0.75	0.04	1.36	0.01	10.49	0.11	7.69	0.02
G96	0.75	0.05	1.37	0.02	9.72	0.13		
V97	0.63	0.04	1.34	0.01	9.12	0.11		
G98	0.60	0.06	1.41	0.02	9.30	0.21		
D99								
G100	0.74	0.08	1.34	0.03	11.87	0.32	8.84	0.05
E101	0.60	0.06	1.13	0.03				
H102	0.91	0.12	1.27	0.04	11.63	0.53	9.13	0.08
W103	0.96	0.08	1.34	0.02	9.95	0.18		
V104	0.87	0.10	1.38	0.03	10.21	0.25		
Y105	0.91	0.05	1.31	0.01	11.15	0.17	8.54	0.03
S106	0.89	0.06	1.35	0.02	11.21	0.17	8.32	0.03
I107	0.77	0.06	1.25	0.02	11.58	0.19	9.27	0.03
T108	0.90	0.09	1.30	0.02	11.98	0.19	9.21	0.03
P109								
D110	0.84	0.06	1.36	0.02	11.12	0.27	8.19	0.04
S111	0.80	0.06	1.28	0.02	10.53	0.08	8.23	0.02
S112	0.76	0.04	1.31	0.01	10.39	0.09	7.93	0.02
W113	0.56	0.04	1.33	0.01	12.03	0.13		
K114	0.89	0.07	1.30	0.02	11.65	0.17	8.99	0.03
T115								
I116	0.78	0.05	1.36	0.02	10.90	0.20	7.99	0.03
E117	0.73	0.04	1.35	0.01	11.09	0.11	8.24	0.02
I118								
P119								
F120	0.69	0.08	1.28	0.02	11.75	0.28	9.21	0.04
S121	0.77	0.04	1.39	0.01	12.16	0.20	8.73	0.03
S122	0.86	0.04	1.35	0.02	11.29	0.19	8.33	0.03

F123	0.61	0.04	1.30	0.01	12.57	0.22		
R124	0.77	0.05	1.30	0.01	11.73	0.21	9.00	0.03
R125	1.01	0.07	1.33	0.02	10.76	0.17	8.09	0.03
R126	0.60	0.06	1.35	0.03	12.06	0.36		
L127	0.75	0.07	1.31	0.03	11.77	0.30	8.97	0.05
D128	0.96	0.08	1.34	0.02	11.96	0.46	8.93	0.06
Y129	0.92	0.04	1.23	0.01	12.53	0.18	10.17	0.03
Q130	0.89	0.05	1.19	0.02	11.55	0.31	9.67	0.04
P131								
P132								
G133	0.82	0.07	1.34	0.02	12.99	0.23	9.66	0.03
Q134	0.97	0.06	1.31	0.02	13.01	0.24	9.97	0.03
D135								
M136	0.87	0.07	1.34	0.02	12.83	0.32	9.59	0.04
S137	0.81	0.03	1.40	0.01	11.68	0.11	8.37	0.02
G138	0.87	0.06	1.38	0.03	12.87	0.33	9.36	0.04
T139								
L140	0.89	0.06	1.28	0.02	11.94	0.24	9.30	0.03
D141	0.79	0.07	1.24	0.02	11.89	0.34	9.62	0.04
L142	0.81	0.06	1.27	0.02	10.71	0.22	8.43	0.04
D143	0.82	0.02	1.35	0.01	11.23	0.10	8.29	0.01
N144	0.82	0.05	1.41	0.01	11.48	0.17	8.16	0.03
I145	0.69	0.09	1.32	0.02	12.41	0.30	9.38	0.04
D146								
S147	0.79	0.06	1.43	0.02	12.20	0.24	8.55	0.04
I148	1.08	0.09	1.38	0.03	12.59	0.30	9.12	0.04
H149	0.78	0.07	1.29	0.02	11.85	0.31	9.17	0.04
F150	0.75	0.05	1.25	0.02	12.37	0.17	9.87	0.03
M151	0.84	0.05	1.28	0.02	11.77	0.22	9.16	0.03
Y152	0.88	0.09	1.22	0.05	11.43	0.54	9.37	0.08
A153	0.70	0.09	1.25	0.05	11.39	0.58	9.09	0.09
N154	0.73	0.06	1.08	0.02	9.35	0.15	8.66	0.03
N155								
K156	0.85	0.04	1.25	0.01	11.13	0.10	8.90	0.02
S157	0.73	0.06	1.31	0.01	11.71	0.14	8.94	0.02
G158	1.10	0.09	1.30	0.02	11.74	0.29	9.05	0.04
K159	0.97	0.07	1.32	0.02	10.67	0.19	8.06	0.03
F160	0.86	0.06	1.34	0.02	11.94	0.15	8.89	0.02
V161	0.74	0.03	1.37	0.01	10.88	0.15	7.95	0.02
V162								
D163	0.80	0.06	1.40	0.02	11.89	0.25	8.46	0.03
N164	0.76	0.04	1.30	0.01	11.44	0.18	8.79	0.03
I165	0.90	0.05	1.35	0.01	10.91	0.13	8.10	0.02
K166	0.81	0.04	1.39	0.02	12.17	0.19	8.73	0.03

L167						
I168	0.91	0.10	1.27	0.02	17.08	0.46
G169	0.80	0.13	1.38	0.04	15.07	0.53
A170	0.51	0.04	1.43	0.01	10.33	0.14
L171	0.21	0.03	1.42	0.01	7.03	0.11
E172	0.12	0.03	1.36	0.01	3.40	0.06

Table C.5: Longitudinal and transverse relaxation rates, ^1H - ^{15}N steady state NOE values and R_2/R_1 ratios for the free *CtCBM11* at 50 °C

Residue	50 °C free							
	NOE		R_1		R_2		$\tau_{c,i} (R_2/R_1)$	
	NOE	Error	R_1	Error	R_2	Error	R_2/R_1	Error
M1								
A2								
S3	0.82	0.03	1.62	0.02	8.55	0.96		
A4								
V5								
G6	0.36	0.03	1.48	0.03				
E7	0.73	0.01	1.82	0.00	8.77	0.08	4.83	0.01
K8	0.76	0.01	1.73	0.01	8.89	0.18		
M9	0.82	0.01	1.78	0.01	8.09	0.19	4.55	0.03
L10	0.84	0.01	1.76	0.01	7.70	0.10	4.37	0.02
D11	0.84	0.01	1.84	0.01	7.42	0.11	4.04	0.02
D12	0.82	0.01	1.89	0.01	7.57	0.08	4.01	0.01
F13	0.80	0.01	1.95	0.01	8.45	0.10	4.32	0.02
E14	0.82	0.01	1.85	0.00	8.13	0.08	4.40	0.01
G15	0.80	0.01	1.74	0.00	6.41	0.06	3.69	0.01
V16	0.71	0.01	1.48	0.00	4.94	0.05		
L17	0.68	0.01	1.64	0.00	5.56	0.08		
N18	0.64	0.01	1.65	0.01	6.15	0.16		
W19	0.79	0.01	1.61	0.00	6.97	0.07	4.32	0.01
G20	0.80	0.01	1.79	0.01	7.56	0.10	4.22	0.02
S21	0.78	0.01	1.72	0.00	6.56	0.12	3.82	0.02
Y22	0.82	0.01	1.95	0.01	7.84	0.08	4.03	0.01
S23	0.86	0.01	1.89	0.00	7.33	0.06	3.88	0.01
G24	0.81	0.01	1.88	0.01	8.27	0.12	4.41	0.02
E25	0.90	0.01	2.03	0.01	7.82	0.15	3.85	0.02
G26	0.84	0.01	1.84	0.01	7.95	0.37	4.33	0.05
A27	0.83	0.01	1.95	0.00	8.25	0.08	4.23	0.01
K28	0.80	0.01	1.86	0.00	7.99	0.10	4.29	0.01
V29	0.81	0.01	1.81	0.01	8.01	0.09	4.42	0.01
S30	0.82	0.01	1.94	0.01	7.81	0.09	4.03	0.01
T31	0.86	0.01	1.91	0.00	7.72	0.08	4.05	0.01
K32	0.80	0.01	1.86	0.01	7.82	0.09	4.20	0.01

I33	0.79	0.01	1.79	0.00	7.24	0.07	4.04	0.01
V34	0.83	0.01	1.91	0.01	8.37	0.09	4.38	0.01
S35	0.83	0.01	1.80	0.01	7.95	0.11	4.40	0.02
G36	0.81	0.01	1.84	0.01	7.53	0.14	4.09	0.02
K37	0.84	0.01	1.91	0.02	8.34	0.23	4.36	0.04
T38	0.82	0.01	1.80	0.01	8.29	0.11	4.62	0.02
G39								
N40	0.83	0.01	1.80	0.01	8.36	0.11	4.63	0.02
G41	0.83	0.01	1.84	0.01	8.26	0.10	4.49	0.02
M42	0.81	0.01	1.90	0.01	7.62	0.11	4.00	0.02
E43	0.82	0.01	1.98	0.01	8.05	0.08	4.08	0.01
V44	0.84	0.01	1.93	0.01	7.99	0.15	4.15	0.02
S45	0.82	0.01	1.86	0.01	8.10	0.08	4.35	0.01
Y46	0.80	0.01	1.98	0.01	8.14	0.14	4.11	0.02
T47	0.80	0.01	1.92	0.01	7.87	0.12	4.10	0.02
G48	0.76	0.01	1.80	0.01	7.22	0.14	4.01	0.02
T49	0.84	0.01	1.82	0.01	8.87	0.11	4.87	0.02
T50								
D51	0.81	0.01	1.71	0.01	6.86	0.16	4.01	0.03
G52	0.79	0.01	1.78	0.00	7.43	0.09	4.17	0.01
Y53	0.81	0.01	1.78	0.01	7.46	0.09	4.18	0.02
W54	0.83	0.01	1.90	0.01	8.09	0.17	4.26	0.03
G55	0.83	0.01	1.93	0.01	8.05	0.12	4.17	0.02
T56	0.83	0.01	1.95	0.01	7.88	0.11	4.04	0.02
V57	0.83	0.01	1.93	0.01	8.12	0.12	4.22	0.02
Y58								
S59	0.76	0.01	1.80	0.01	7.36	0.09	4.10	0.02
L60	0.76	0.01	1.58	0.01	6.41	0.08	4.07	0.02
P61								
D62	0.78	0.01	1.67	0.00	6.75	0.08	4.04	0.01
G63	0.76	0.01	1.51	0.01	5.09	0.10		
D64								
W65	0.80	0.01	1.77	0.01	9.13	0.14		
S66	0.81	0.01	2.00	0.01	8.10	0.10	4.04	0.02
K67								
W68								
L69								
K70	0.80	0.01	1.86	0.01	7.86	0.11	4.23	0.02
I71	0.84	0.01	1.98	0.01	8.96	0.11	4.51	0.02
S72	0.81	0.01	1.99	0.01	8.33	0.10	4.18	0.01
F73	0.87	0.01	2.00	0.01	8.85	0.10	4.43	0.01
D74	0.79	0.01	1.98	0.01	7.81	0.09	3.95	0.01
I75	0.83	0.01	1.99	0.01	8.21	0.09	4.14	0.01
K76	0.83	0.01	1.96	0.01	8.47	0.12	4.32	0.02

S77	0.96	0.02	2.17	0.02	9.66	0.15	4.45	0.03
V78	0.62	0.01	1.59	0.01	5.88	0.11		
D79	0.61	0.01	1.71	0.01	6.09	0.16		
G80	0.75	0.04	1.73	0.03	7.57	0.47	4.39	0.08
S81	0.73	0.01	1.82	0.01	5.53	0.17		
A82	0.56	0.05	1.59	0.05	5.79	0.58		
N83	0.60	0.01	1.54	0.01	5.41	0.10		
E84	0.77	0.01	1.76	0.01	6.90	0.11	3.91	0.02
I85	0.78	0.01	1.92	0.01	9.01	0.15	4.69	0.02
R86	0.84	0.01	1.91	0.01	8.30	0.18	4.35	0.03
F87	0.82	0.01	1.98	0.01	7.92	0.13	4.00	0.02
M88	0.82	0.01	1.98	0.01	8.71	0.12	4.40	0.02
I89	0.77	0.01	1.90	0.01	7.40	0.13	3.90	0.02
A90	0.85	0.01	1.96	0.01	8.43	0.12	4.29	0.02
E91	0.86	0.01	1.88	0.01	7.90	0.17	4.20	0.03
K92	0.80	0.01	1.73	0.01	8.02	0.16	4.63	0.03
S93	0.79	0.01	1.75	0.01	7.10	0.08	4.06	0.02
I94	0.79	0.01	1.93	0.01	8.92	0.22	4.63	0.03
N95	0.76	0.01	1.86	0.01	7.62	0.12	4.10	0.02
G96	0.63	0.01	1.71	0.01	5.59	0.10		
V97	0.56	0.01	1.65	0.00	5.08	0.06		
G98	0.65	0.01	1.68	0.01	6.06	0.15		
D99	0.75	0.01	1.72	0.01	8.01	0.07	4.65	0.01
G100	0.77	0.01	1.79	0.01	7.74	0.13	4.33	0.02
E101	0.71	0.01	1.79	0.01	5.26	0.13		
H102	0.82	0.03	1.84	0.02	10.64	0.61		
W103	0.85	0.01	1.90	0.01	7.48	0.13	3.94	0.02
V104	0.81	0.02	1.95	0.01	8.35	0.18	4.28	0.03
Y105	0.84	0.01	1.93	0.01	8.07	0.11	4.19	0.02
S106	0.88	0.01	1.79	0.01	7.27	0.14	4.06	0.02
I107	0.79	0.01	1.87	0.01	8.43	0.13	4.51	0.02
T108								
P109								
D110	0.80	0.01	1.89	0.01	7.63	0.11	4.04	0.02
S111								
S112								
W113	0.77	0.01	1.54	0.01	5.99	0.12	3.90	0.02
K114	0.80	0.01	1.83	0.01	7.97	0.13	4.36	0.02
T115	0.81	0.01	1.77	0.01	7.59	0.10	4.29	0.02
I116	0.82	0.01	1.92	0.01	8.13	0.14	4.24	0.02
E117	0.81	0.01	1.80	0.00	7.37	0.09	4.10	0.02
I118	0.80	0.01	1.88	0.00	7.46	0.08	3.97	0.01
P119								
F120	0.86	0.01	1.87	0.01	9.10	0.15	4.86	0.02

S121	0.88	0.01	1.95	0.01	8.34	0.16	4.29	0.02
S122	0.79	0.01	1.85	0.01	7.79	0.15	4.21	0.02
F123	0.71	0.01	1.76	0.01	7.38	0.08	4.19	0.01
R124	0.79	0.01	1.82	0.01	7.68	0.20	4.21	0.03
R125	0.77	0.01	1.84	0.01	7.77	0.14	4.22	0.02
R126	0.66	0.01	1.57	0.01	6.14	0.09	3.90	0.02
L127	0.85	0.02	1.92	0.01	9.01	0.32	4.70	0.04
D128	0.87	0.01	1.93	0.01	8.68	0.23	4.49	0.03
Y129	0.86	0.01	1.76	0.01	7.80	0.13	4.42	0.02
Q130	0.84	0.01	1.75	0.01	8.25	0.17	4.72	0.03
P131								
P132								
G133	0.81	0.01	1.94	0.01	10.01	0.22	5.15	0.03
Q134	0.81	0.01	1.92	0.01	8.93	0.18	4.64	0.02
D135	0.82	0.01	1.91	0.01	10.55	0.31		
M136	0.83	0.02	1.91	0.01	13.06	0.38		
S137	0.84	0.01	1.99	0.01	9.22	0.23	4.64	0.03
G138	0.84	0.01	2.00	0.01	8.71	0.28	4.36	0.04
T139	0.84	0.01	2.03	0.01	7.23	0.47	3.57	0.07
L140	0.79	0.01	1.73	0.01	7.50	0.11	4.34	0.02
D141	0.82	0.01	1.77	0.01	8.81	0.19	4.98	0.03
L142	0.78	0.01	1.82	0.01	7.72	0.12	4.23	0.02
D143	0.81	0.01	1.90	0.01	8.39	0.09	4.41	0.01
N144	0.71	0.01	1.73	0.01	6.86	0.08	3.96	0.02
I145	0.80	0.01	1.78	0.01	8.85	0.10	4.97	0.02
D146	0.84	0.01	2.00	0.01	7.12	0.17	3.57	0.03
S147	0.81	0.01	1.99	0.01	8.01	0.11	4.03	0.02
I148	0.83	0.01	1.94	0.01	7.56	0.10	3.89	0.02
H149	0.80	0.01	1.92	0.01	8.35	0.14	4.35	0.02
F150	0.82	0.01	1.82	0.01	7.62	0.08	4.18	0.01
M151	0.85	0.01	1.89	0.01	8.32	0.13	4.40	0.02
Y152	0.81	0.01	1.94	0.01	8.36	0.17	4.32	0.03
A153	0.84	0.02	1.86	0.02	9.08	0.42	4.89	0.06
N154	0.83	0.01	1.59	0.01	5.54	0.12		
N155								
K156	0.85	0.01	1.75	0.00	7.96	0.09	4.54	0.01
S157								
G158	0.78	0.01	1.94	0.01	7.75	0.14	3.99	0.02
K159	0.79	0.01	1.82	0.01	7.05	0.12	3.88	0.02
F160	0.79	0.01	2.02	0.01	8.49	0.10	4.21	0.01
V161	0.81	0.01	1.93	0.01	8.03	0.11	4.16	0.02
V162	0.78	0.01	1.91	0.01	8.52	0.11	4.46	0.02
D163	0.76	0.01	1.95	0.00	7.33	0.09	3.75	0.01
N164	0.79	0.01	1.90	0.01	7.54	0.12	3.97	0.02

I165	0.83	0.01	2.00	0.00	8.66	0.03	4.33	0.00
K166	0.79	0.01	2.07	0.01	8.69	0.14	4.19	0.02
L167	0.73	0.01	1.94	0.01	7.75	0.14	3.98	0.02
I168	0.79	0.01	1.94	0.01	12.00	0.16		
G169	0.78	0.01	1.93	0.01	11.12	0.27		
A170	0.57	0.01	1.66	0.00	7.02	0.09		
L171	0.42	0.01	1.52	0.00	4.31	0.05		
E172								

Table C.6: Longitudinal and transverse relaxation rates, ^1H - ^{15}N steady state NOE values and R_2/R_1 ratios for the bound *CtCBM11* at 50 °C

Residue	50 °C bound							
	NOE		R_1		R_2		$\tau_{c,i} (R_2/R_1)$	
	NOE	Error	R_1	Error	R_2	Error	R_2/R_1	Error
M1								
A2								
S3								
A4								
V5								
G6								
E7	0.64	0.04	1.95	0.03	6.43	0.12		
K8	0.70	0.05	2.28	0.03	7.13	0.11	3.13	0.03
M9	0.86	0.34	1.24	0.32				
L10	0.78	0.05	2.06	0.03	7.71	0.11	3.74	0.03
D11	0.83	0.04	2.09	0.03	6.39	0.11	3.06	0.03
D12	0.80	0.03	2.03	0.02	7.15	0.10	3.53	0.02
F13	0.90	0.04	2.15	0.02	7.90	0.10	3.68	0.02
E14	0.78	0.03	2.02	0.01	7.34	0.09	3.64	0.02
G15	0.80	0.03	1.97	0.02	5.70	0.07	2.89	0.02
V16	0.73	0.04	1.52	0.03	4.40	0.16		
L17	0.73	0.02	1.95	0.02	6.56	0.07	3.37	0.02
N18	0.55	0.04	1.94	0.05	6.19	0.26		
W19	0.72	0.03	1.83	0.02	6.29	0.08	3.43	0.02
G20	0.86	0.05	1.94	0.02	7.21	0.10	3.71	0.02
S21	0.83	0.14	1.55	0.17	5.11	0.85	3.29	0.27
Y22	0.81	0.07	1.85	0.05	7.31	0.17	3.95	0.05
S23	0.88	0.03	2.16	0.03	6.84	0.09	3.16	0.03
G24	0.84	0.04	2.14	0.03	7.98	0.14	3.73	0.03
E25	0.92	0.09	1.80	0.10	7.15	0.51	3.98	0.13
G26								
A27	0.86	0.03	2.20	0.03	7.37	0.11	3.35	0.03
K28	0.82	0.04	2.04	0.03	7.41	0.09	3.63	0.03
V29	0.86	0.03	2.10	0.03	6.89	0.10	3.28	0.03
S30	0.88	0.04	2.07	0.03	6.81	0.09	3.29	0.03

T31	0.80	0.04	2.02	0.04	5.84	0.17	2.89	0.05
K32	0.80	0.05	2.20	0.03	7.26	0.10	3.30	0.03
I33	0.78	0.03	2.06	0.02	6.72	0.09	3.27	0.03
V34	0.83	0.04	2.05	0.02	7.25	0.10	3.54	0.03
S35	0.92	0.06	2.11	0.05	6.49	0.20	3.08	0.05
G36	0.78	0.03	1.97	0.02	6.73	0.09	3.42	0.03
K37	0.73	0.09	1.86	0.13	7.33	0.59	3.94	0.15
T38	0.79	0.05	2.14	0.04	6.40	0.16	2.99	0.04
G39								
N40	0.87	0.04	2.04	0.03	7.91	0.15	3.87	0.03
G41	0.81	0.04	2.08	0.03	7.73	0.12	3.72	0.03
M42	0.86	0.05	2.04	0.03	6.91	0.08	3.38	0.03
E43	0.76	0.03	2.15	0.02	7.30	0.06	3.39	0.02
V44	0.83	0.06	2.12	0.03	7.59	0.15	3.58	0.03
S45	0.81	0.03	1.95	0.02	6.71	0.08	3.44	0.02
Y46	0.91	0.08	1.99	0.05	7.08	0.20	3.56	0.05
T47	0.72	0.04	2.02	0.02	7.03	0.09	3.48	0.03
G48	0.72	0.05	1.92	0.03	5.80	0.14	3.02	0.04
T49	0.77	0.04	1.96	0.03	7.78	0.14	3.96	0.03
T50								
D51	0.92	0.06	2.43	0.04	7.04	0.13	2.90	0.03
G52	0.78	0.03	1.96	0.02	7.45	0.10	3.80	0.03
Y53	0.82	0.04	2.02	0.02	6.96	0.08	3.45	0.02
W54	0.89	0.04	2.20	0.03	6.86	0.13	3.12	0.03
G55	0.87	0.05	2.12	0.03	7.10	0.16	3.36	0.04
T56	0.89	0.04	2.10	0.02	7.23	0.10	3.44	0.02
V57	0.91	0.05	2.18	0.04	7.19	0.12	3.30	0.03
Y58	0.86	0.06	2.18	0.05	8.70	0.22	4.00	0.05
S59	0.70	0.05	1.93	0.03	6.53	0.13	3.39	0.04
L60	0.67	0.05	1.95	0.03	6.85	0.11	3.51	0.03
P61								
D62	0.79	0.05	1.84	0.04	5.76	0.19	3.13	0.06
G63								
D64	0.80	0.05	2.14	0.07	7.30	0.32	3.41	0.08
W65	0.77	0.05	2.00	0.03	8.12	0.15	4.05	0.03
S66	0.94	0.04	2.24	0.03	6.89	0.08	3.07	0.02
K67								
W68	0.82	0.05	1.89	0.02	10.27	0.23		
L69								
K70	0.78	0.05	2.14	0.03	7.52	0.12	3.51	0.03
I71	0.78	0.04	2.12	0.02	7.16	0.09	3.38	0.02
S72	0.75	0.03	2.17	0.02	7.17	0.09	3.31	0.02
F73	0.79	0.03	2.20	0.03	7.45	0.10	3.38	0.03
D74	0.79	0.03	2.16	0.02	7.14	0.08	3.31	0.02

I75	0.86	0.03	2.15	0.02	7.88	0.08	3.67	0.02
K76	0.72	0.07	1.85	0.04	6.14	0.18	3.32	0.05
S77	0.76	0.08	1.94	0.06	7.79	0.21	4.02	0.06
V78	0.81	0.09	2.09	0.07	7.35	0.32		
D79	0.66	0.12	1.69	0.09	4.74	0.49		
G80								
S81								
A82								
N83	0.64	0.08	1.80	0.06	3.97	0.24		
E84	0.78	0.07	2.19	0.04	6.29	0.16	2.87	0.04
I85	0.64	0.06	2.06	0.04	6.90	0.14		
R86	0.73	0.04	2.11	0.03	7.51	0.13	3.55	0.03
F87	0.81	0.04	2.15	0.03	6.92	0.11	3.22	0.03
M88	0.88	0.05	2.12	0.04	8.08	0.14	3.81	0.03
I89	0.75	0.04	2.07	0.03	6.46	0.12	3.12	0.04
A90	0.73	0.04	2.14	0.03	7.63	0.13	3.56	0.03
E91	0.73	0.06	1.93	0.05	7.06	0.18	3.66	0.05
K92	0.91	0.06	2.05	0.04	6.69	0.15	3.27	0.04
S93	0.76	0.07	1.97	0.02	6.86	0.08	3.48	0.02
I94	0.86	0.07	2.00	0.04	7.15	0.17	3.57	0.04
N95	0.89	0.05	2.05	0.05	6.88	0.18	3.36	0.05
G96	0.50	0.08	1.80	0.10	6.12	0.57		
V97	0.56	0.04	1.96	0.04	5.23	0.14		
G98								
D99								
G100	0.89	0.07	2.00	0.04	6.93	0.14	3.46	0.04
E101								
H102	0.83	0.08	2.04	0.05	6.51	0.23	3.19	0.06
W103	0.86	0.06	1.99	0.04	6.45	0.11	3.24	0.03
V104	0.78	0.08	2.02	0.04	7.37	0.18	3.64	0.05
Y105	0.92	0.05	2.09	0.03	7.32	0.11	3.49	0.03
S106	1.00	0.16	1.66	0.17	5.34	0.58	3.21	0.21
I107	0.78	0.05	1.98	0.03	6.80	0.12	3.44	0.03
T108	0.75	0.03	2.01	0.03	7.29	0.11	3.63	0.03
P109								
D110	0.68	0.11	1.83	0.09	6.81	0.31	3.71	0.09
S111	0.62	0.10	1.67	0.09	6.43	0.48		
S112	0.86	0.05	2.17	0.03	5.47	0.19		
W113	0.71	0.06	2.19	0.04	6.93	0.20	3.16	0.05
K114	0.81	0.05	2.01	0.03	7.04	0.13	3.51	0.03
T115	0.74	0.04	2.07	0.05	7.39	0.19	3.57	0.05
I116	0.85	0.04	2.08	0.03	7.21	0.12	3.46	0.03
E117	0.77	0.05	1.97	0.05	4.94	0.16		
I118	0.81	0.03	2.04	0.02	6.38	0.08	3.13	0.02

P119								
F120	0.82	0.05	2.01	0.03	7.50	0.14	3.73	0.03
S121	0.77	0.10	2.24	0.12	8.02	0.59	3.57	0.13
S122	0.85	0.03	2.14	0.02	7.73	0.08	3.61	0.02
F123	0.74	0.04	2.02	0.03	7.31	0.14	3.63	0.03
R124	0.71	0.05	2.11	0.05	6.52	0.17	3.08	0.05
R125	0.76	0.04	2.18	0.04	7.09	0.15	3.25	0.04
R126	0.72	0.05	2.06	0.03	6.90	0.17	3.36	0.04
L127								
D128	1.10	0.21	1.75	0.20	5.68	1.12	3.26	0.31
Y129	0.86	0.04	1.96	0.02	6.84	0.12	3.50	0.03
Q130	0.78	0.04	1.93	0.03	7.27	0.14	3.76	0.04
P131								
P132								
G133	0.89	0.07	2.15	0.05	7.94	0.32	3.69	0.07
Q134	0.88	0.05	2.11	0.04	7.55	0.20	3.58	0.04
D135	0.76	0.05	2.12	0.05	7.93	0.23	3.73	0.05
M136	0.80	0.05	2.18	0.04	7.83	0.24	3.60	0.05
S137	0.80	0.04	2.28	0.03	7.20	0.15	3.16	0.04
G138	0.81	0.04	2.23	0.05	7.23	0.21	3.24	0.05
T139	0.75	0.04	2.36	0.03	6.50	0.13	2.75	0.03
L140	0.79	0.05	2.20	0.04	8.10	0.19	3.69	0.04
D141	0.77	0.05	1.88	0.04	7.60	0.18	4.04	0.04
L142	0.79	0.04	2.00	0.02	6.69	0.10	3.34	0.03
D143	0.75	0.04	1.86	0.04	7.27	0.19	3.90	0.05
N144	0.73	0.04	2.06	0.03	6.53	0.09	3.17	0.03
I145	0.96	0.07	2.05	0.03	6.78	0.12	3.31	0.03
D146	0.76	0.05	2.16	0.04	7.66	0.13	3.54	0.03
S147	0.85	0.04	2.11	0.02	7.23	0.12	3.43	0.03
I148	0.79	0.04	2.18	0.04	7.25	0.11	3.33	0.03
H149	0.83	0.05	2.02	0.04	7.69	0.16	3.81	0.04
F150	0.79	0.04	2.04	0.02	7.25	0.07	3.56	0.02
M151	0.84	0.04	2.08	0.02	7.30	0.10	3.51	0.03
Y152	0.86	0.05	2.31	0.04	7.48	0.16	3.24	0.04
A153	1.05	0.10	1.99	0.09	6.80	0.37	3.43	0.10
N154	0.83	0.06	1.89	0.03	5.83	0.09	3.08	0.03
N155								
K156	0.82	0.04	2.07	0.02	7.30	0.13	3.52	0.03
S157								
G158	0.72	0.05	2.12	0.04	6.67	0.14	3.14	0.04
K159	0.83	0.04	2.11	0.02	5.93	0.10	2.81	0.03
F160	0.84	0.06	2.16	0.03	7.03	0.11	3.25	0.03
V161	0.83	0.03	2.10	0.02	6.78	0.08	3.22	0.02
V162	0.72	0.06	2.08	0.02	7.30	0.10	3.50	0.03

D163	0.84	0.02	2.16	0.01	6.93	0.06	3.21	0.02
N164	0.80	0.04	2.10	0.03	6.70	0.12	3.20	0.03
I165								
K166	0.77	0.04	2.16	0.03	6.90	0.13	3.19	0.03
L167	0.69	0.04	2.03	0.04	5.95	0.14	2.94	0.04
I168	0.70	0.22	1.93	0.12	8.57	0.55		
G169	0.94	0.07	2.28	0.05	8.50	0.21	3.73	0.05
A170	0.58	0.08	1.81	0.06	5.18	0.27		
L171	0.28	0.08	1.86	0.09	6.45	0.28		
E172								

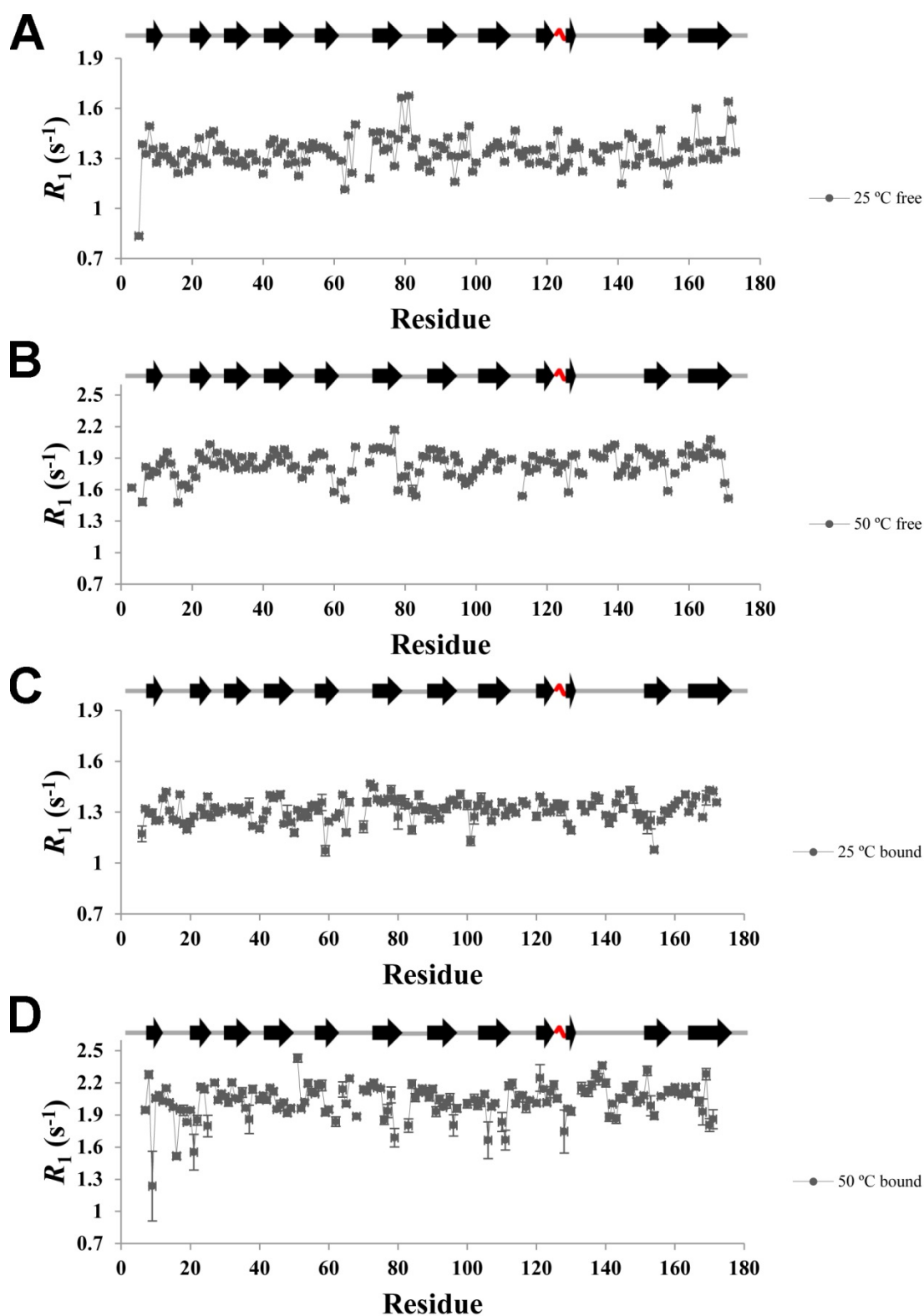


Figure C.1: R_1 relaxation rates determined for the free and bound *CtCBM11* at 25 and 50 °C.

A) Free *CtCBM11* at 25 °C; B) Free *CtCBM11* at 50 °C; C) Bound *CtCBM11* at 25 °C; D) Bound *CtCBM11* at 50 °C.

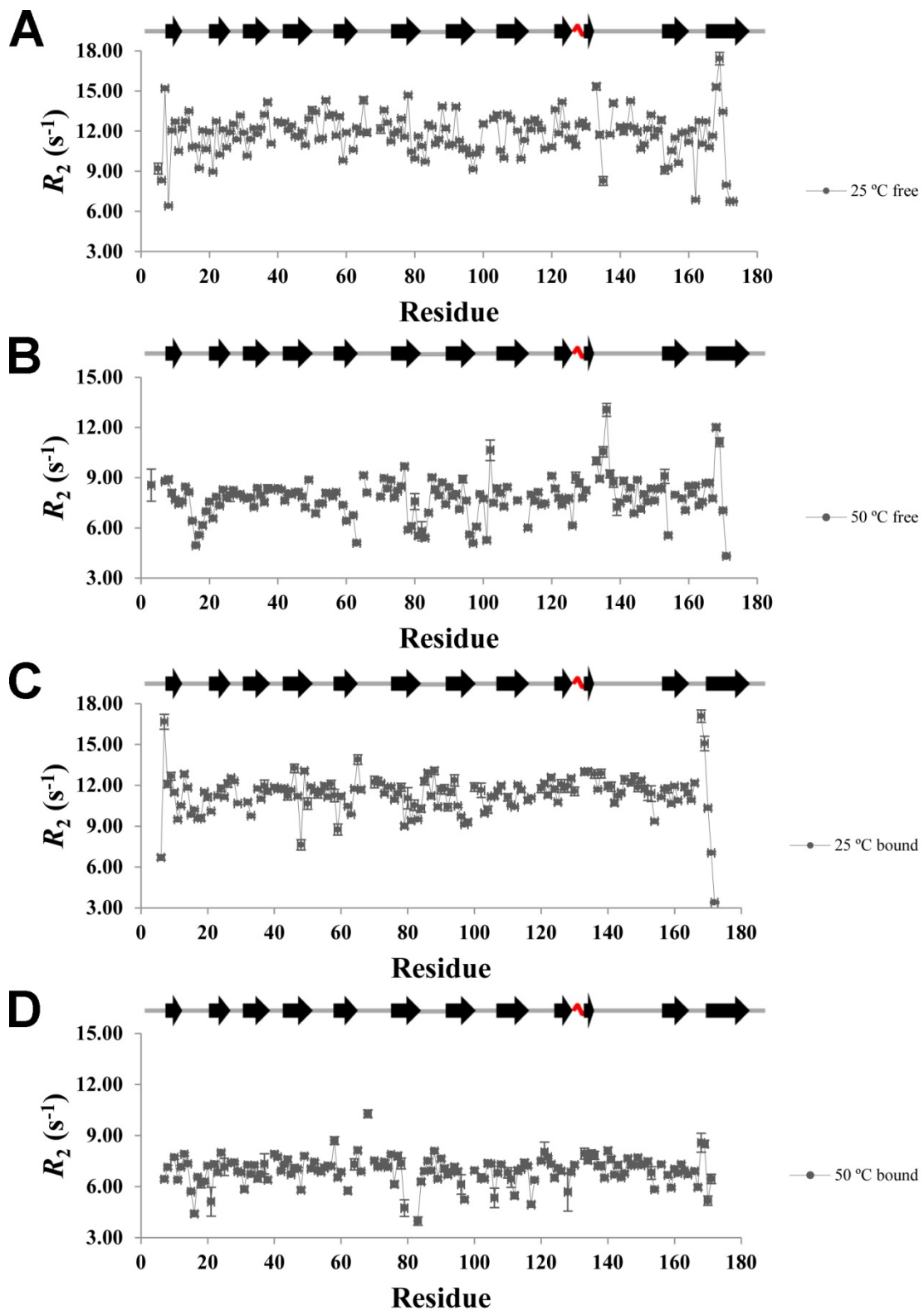


Figure C.2: R_2 relaxation rates determined for the free and bound CtCBM11 at 25 and 50 °C.

A) Free CtCBM11 at 25 °C; B) Free CtCBM11 at 50 °C; C) Bound CtCBM11 at 25 °C; D) Bound CtCBM11 at 50 °C.

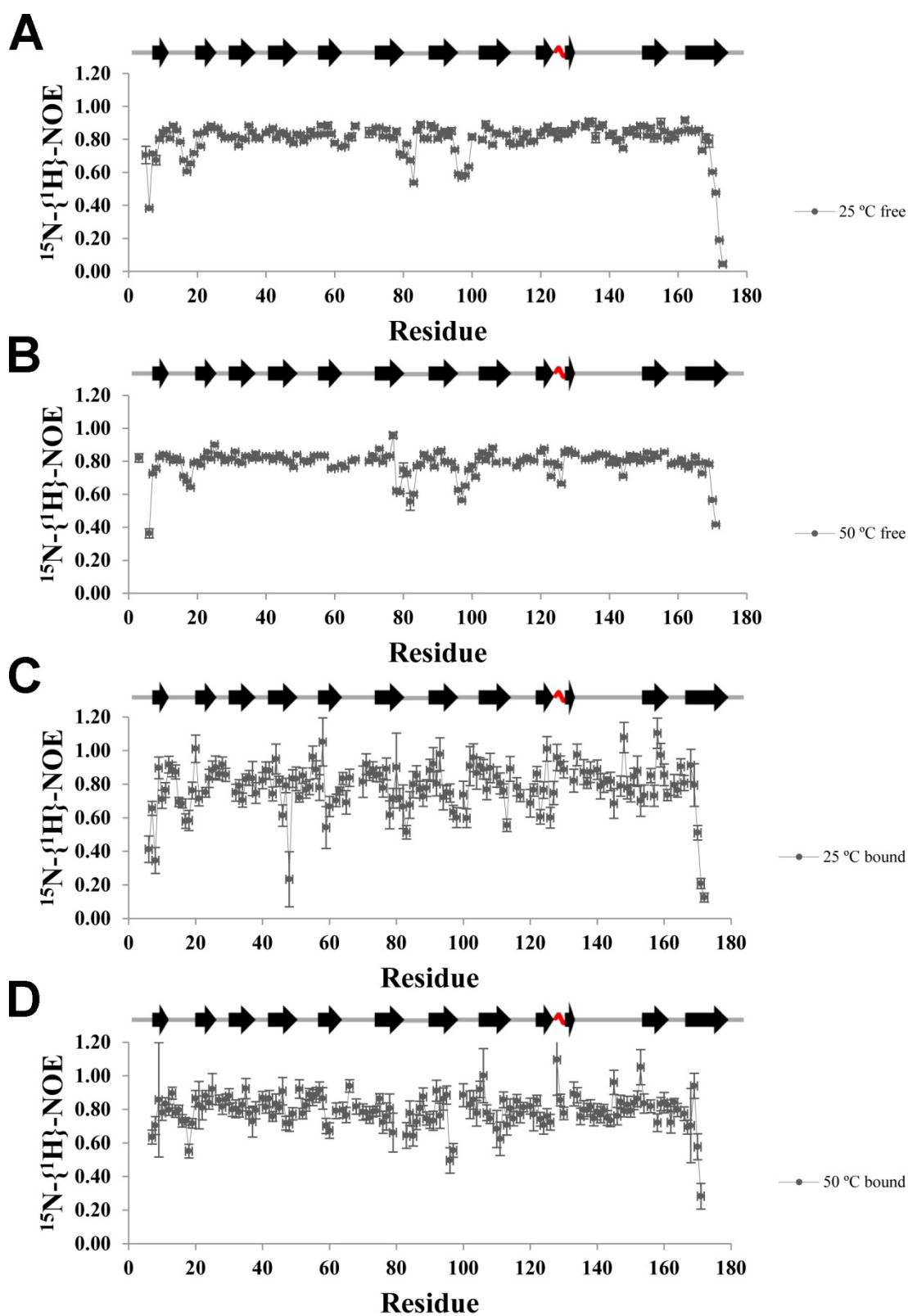


Figure C.3: $^1\text{H}\text{-}^{15}\text{N}$ steady state NOE values determined for the free and bound CtCBM11 at 25 and 50 °C.

A) Free CtCBM11 at 25 °C; B) Free CtCBM11 at 50 °C; C) Bound CtCBM11 at 25 °C; D) Bound CtCBM11 at 50 °C.

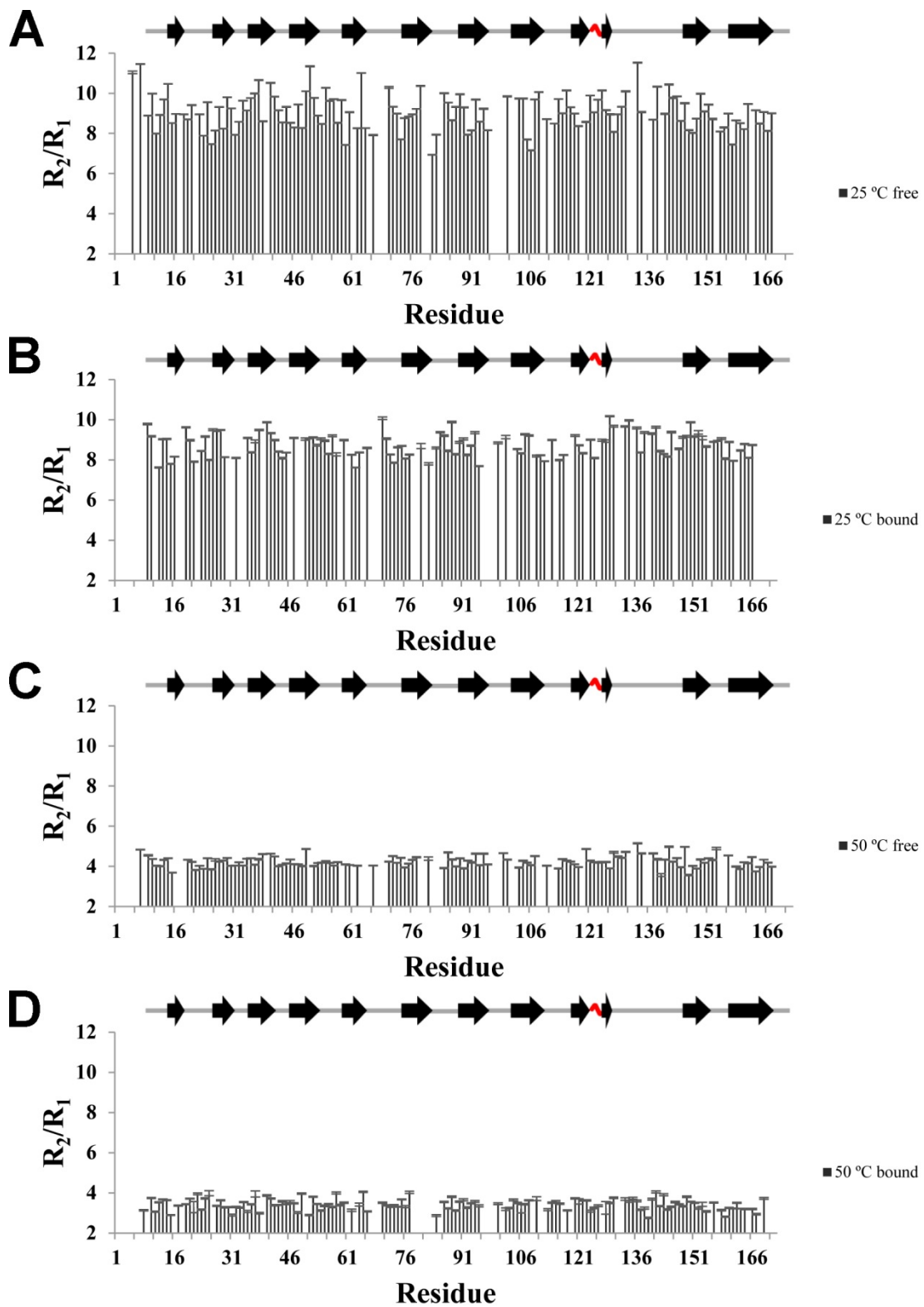


Figure C.4: R_2/R_1 values determined for the free and bound *CtCBM11* at 25 and 50 °C.

A) Free *CtCBM11* at 25 °C; B) Free *CtCBM11* at 50 °C; C) Bound *CtCBM11* at 25 °C; D) Bound *CtCBM11* at 50 °C.

Table C.7: Characterization of the diffusion tensor obtained for CrCBM11 at the different experimental conditions, obtained with Tensor2.0¹.

		25 °C		50 °C	
		Free	Bound	Free	Bound
Isotropy	τ_c (s)	9.017e-09 ±	8.883e-09 ±	5.653e-09 ±	4.826e-09 ±
		5.362e-11	5.975e-11	4.581e-11	4.097e-11
	$\text{Chi}^2_{\text{exp}}$	6.2599e+01	2.5730e+01	2.7462e+01	4.2989e+01
	$\text{Chi}^2_{\text{Mc}} (0.10)$	1.5077e+02	1.2873e+02	1.3946e+02	1.4177e+02
	$\text{Chi}^2_{\text{Mc}} (0.05)$	1.5245e+02	1.3430e+02	1.4635e+02	1.4856e+02
	<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>	
	Phi (°)	-40.790 ± 18.404	-45.246 ± 25.790	-46.552 ± 16.881	61.443 ± 32.718
	Teta (°)	43.356 ± 27.243	21.752 ± 27.518	29.291 ± 19.624	88.202 ± 36.660
Axial Symmetry 1st Minimum	D_{\perp} (1e8 s ⁻¹)	1.898e-01 ±	1.930e-01 ±	3.066e-01 ±	3.359e-01 ±
		2.835e-03	3.877e-03	5.675e-03	1.117e-02
	D_{\parallel} (1e8 s ⁻¹)	1.723e-01 ±	1.752e-01 ±	2.662e-01 ±	3.625e-01 ±
		5.423e-03	7.587e-03	1.091e-02	2.027e-02
	D_{\parallel}/D_{\perp} (1e8 s ⁻¹)	0.91 ± 0.05	0.91 ± 0.06	0.87 ± 0.06	1.08 ± 0.09
	$\text{Chi}^2_{\text{exp}}$	5.8584e+01	2.2971e+01	2.2792e+01	4.2056e+01
	$\text{Chi}^2_{\text{Mc}} (0.10)$	1.4646e+02	1.2624e+02	1.3476e+02	1.3970e+02
	$\text{Chi}^2_{\text{Mc}} (0.05)$	1.5123e+02	1.3393e+02	1.4039e+02	1.4568e+02
		<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>
		Phi (°)	64.753 ± 25.543	40.897 ± 31.203	42.256 ± 20.650
	Teta (°)	-11.025 ± 26.808	30.587 ± 38.427	65.214 ± 29.593	16.241 ± 41.235
Axial Symmetry 2nd Minimum	D_{\perp} (1e8 s ⁻¹)	1.793e-01 ±	1.829e-01 ±	2.805e-01 ±	3.523e-01 ±
		3.765e-03	4.082e-03	8.389e-03	1.114e-02
	D_{\parallel} (1e8 s ⁻¹)	1.948e-01 ±	1.958e-01 ±	3.193e-01 ±	3.298e-01 ±
		6.763e-03	6.640e-03	1.449e-02	2.043e-02
	D_{\parallel}/D_{\perp} (1e8 s ⁻¹)	1.09 ± 0.06	1.07 ± 0.06	1.14 ± 0.08	0.94 ± 0.09
	$\text{Chi}^2_{\text{exp}}$	5.9989e+01	2.4347e+01	2.4604e+01	4.2102e+01
	$\text{Chi}^2_{\text{Mc}} (0.10)$	1.4574e+02	1.2547e+02	1.3435e+02	1.4003e+02
	$\text{Chi}^2_{\text{Mc}} (0.05)$	1.5185e+02	1.3342e+02	1.4005e+02	1.4639e+02
		<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>
		Alpha (°)	-61.785 ± 30.343	81.145 ± 35.203	68.660 ± 29.754
	Beta (°)	58.691 ± 26.139	43.528 ± 20.266	43.472 ± 16.983	61.523 ± 33.141
	Gamma (°)	-4.436 ± 43.215	31.869 ± 47.459	56.485 ± 37.346	84.334 ± 38.012
Full Asymmetry	D_x (1e8 s ⁻¹)	1.725e-01 ±	1.748e-01 ±	2.639e-01 ±	3.286e-01 ±
		4.879e-03	5.689e-03	1.086e-02	1.088e-02
	D_y (1e8 s ⁻¹)	1.870e-01 ±	1.900e-01 ±	2.966e-01 ±	3.432e-01 ±
		4.108e-03	4.353e-03	8.942e-03	9.600e-03

	Dz (1e8 s⁻¹)	1.925e-01 ± 4.646e-03	1.962e-01 ± 4.934e-03	3.177e-01 ± 1.147e-02	3.619e-01 ± 1.291e-02
	Chi²_{exp}	5.8385e+01	2.2720e+01	2.2061e+01	4.1768e+01
	Chi²_{Mc} (0.10)	1.4316e+02	1.2410e+02	1.3266e+02	1.3775e+02
	Chi²_{Mc} (0.05)	1.4975e+02	1.3225e+02	1.3856e+02	1.4430e+02
		<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>	<i>Accepted model</i>
	F_{exp}	2.9356e-01 (P=7.4598e-01)	9.4974e-01 (P=3.8885e-01)	2.8468e+00 (P=6.0752e-02)	5.9292e-01 (P=5.5383e-01)
	F_{exp} < 0.2	2.1771e+00	2.2499e+00	2.3093e+00	1.6529e+00
FTest	FTest	<i>No Significant improvement</i>	<i>No Significant improvement</i>	<i>Significant improvement</i>	<i>No Significant improvement</i>
	F_{exp} < 0.1	3.0982e+00	3.1042e+00	3.5301e+00	2.4430e+00
	FTest	<i>No Significant improvement</i>	<i>No Significant improvement</i>	<i>No Significant improvement</i>	<i>No Significant improvement</i>

C.3 Internal mobility

Table C.8: Internal mobility parameters (S^2 , τ_e and R_{ex}) for the free protein at 25 °C

Residue	25 °C free						Model
	S^2		$\tau_{e,i}$		K_{ex}		
	S^2	Error	$\tau_{e,i}$	Error	K_{ex}	Error	
M1							
A2							
S3							
A4							
V5	0.53	1.20E-02	1.38E-11	3.88E-12	2.24	0.40	4
G6	0.61	8.30E-03	6.07E-10	1.59E-11	0.00	0.00	5
E7	0.78	3.50E-03	4.68E-11	2.57E-12	5.87	0.14	4
K8	0.57	1.27E-02	1.26E-09	1.19E-10	0.00	0.00	5
M9	0.84	4.30E-03	3.29E-11	4.32E-12	1.85	0.13	4
L10	0.83	6.70E-03	1.84E-11	5.28E-12	2.04	0.15	4
D11	0.81	4.40E-03	1.07E-11	3.35E-12	0.93	0.14	4
D12	0.83	3.50E-03	3.01E-11	3.41E-12	2.24	0.15	4
F13	0.91	3.40E-03	0.00E+00	0.00E+00	0.65	0.10	3
E14	0.86	3.20E-03	1.66E-11	3.52E-12	2.22	0.11	4
G15	0.79	2.70E-03	2.82E-11	2.01E-12	0.95	0.07	4
V16	0.78	2.30E-03	5.18E-11	1.61E-12	0.54	0.07	4
L17	0.76	5.80E-03	6.31E-10	1.90E-11	0.00	0.00	5
N18	0.85	4.20E-03	1.12E-10	6.91E-12	0.79	0.13	4
W19	0.78	3.80E-03	4.19E-11	2.60E-12	0.54	0.13	4
G20	0.84	4.50E-03	2.11E-11	3.95E-12	0.91	0.14	4
S21	0.87	9.30E-03	7.25E-10	5.26E-11	0.00	0.00	5
Y22	0.94	4.30E-03	5.34E-11	1.24E-11	0.64	0.09	4
S23	0.94	6.60E-03	1.44E-09	3.21E-10	0.00	0.00	5
G24	0.88	3.30E-03	0.00E+00	0.00E+00	0.41	0.13	3
E25	0.89	1.11E-02	2.39E-09	1.47E-09	0.00	0.00	5
G26	0.91	4.20E-03	1.94E-11	6.35E-12	1.07	0.12	4
A27	0.92	3.20E-03	4.53E-11	6.27E-12	0.00	0.00	2
K28	0.94	1.21E-02	4.20E-10	1.19E-10	0.00	0.00	5
V29	0.89	4.00E-03	4.36E-11	5.03E-12	1.50	0.11	4
S30	0.85	4.10E-03	3.07E-11	3.26E-12	0.77	0.08	4
T31	0.88	5.40E-03	1.23E-09	7.94E-11	0.00	0.00	5
K32	0.84	3.40E-03	4.81E-11	2.86E-12	0.72	0.08	4
I33	0.82	2.90E-03	2.64E-11	2.43E-12	1.70	0.08	4
V34	0.85	2.70E-03	3.64E-11	2.70E-12	0.50	0.07	4
S35	0.83	2.70E-03	0.00E+00	0.00E+00	1.58	0.08	3
G36	0.81	3.50E-03	1.42E-11	2.66E-12	3.73	0.14	4
K37	0.83	7.60E-03	2.70E-11	4.69E-12	4.03	0.19	4

T38	0.85	4.80E-03	3.15E-11	3.84E-12	0.00	0.00	2
G39							
N40	0.79	4.90E-03	1.44E-11	2.54E-12	2.46	0.15	4
G41	0.86	3.30E-03	1.69E-11	3.25E-12	1.21	0.10	4
M42	0.91	4.70E-03	1.81E-11	6.60E-12	1.12	0.14	4
E43	0.89	4.70E-03	4.79E-11	4.68E-12	1.10	0.12	4
V44	0.87	5.00E-03	2.47E-11	4.88E-12	1.35	0.15	4
S45	0.87	4.00E-03	2.24E-11	3.50E-12	0.85	0.14	4
Y46	0.92	5.90E-03	9.34E-10	1.12E-10	0.00	0.00	5
T47	0.82	4.00E-03	2.95E-11	3.38E-12	1.40	0.14	4
G48	0.88	7.20E-03	8.37E-10	7.76E-11	0.00	0.00	5
T49	0.85	4.20E-03	2.33E-11	3.65E-12	1.62	0.11	4
T50	0.80	5.10E-03	1.75E-11	3.92E-12	3.00	0.30	4
D51	0.90	3.30E-03	6.89E-11	6.66E-12	1.55	0.12	4
G52	0.84	3.60E-03	2.82E-11	3.23E-12	0.40	0.08	4
Y53	0.86	4.00E-03	1.80E-11	4.31E-12	0.72	0.12	4
W54	0.90	6.60E-03	4.17E-11	7.23E-12	2.84	0.21	4
G55	0.88	5.30E-03	3.15E-11	5.48E-12	1.97	0.15	4
T56	0.90	4.70E-03	0.00E+00	0.00E+00	1.84	0.17	3
V57	0.88	5.40E-03	2.96E-11	5.54E-12	0.65	0.16	4
Y58	0.88	6.50E-03	0.00E+00	0.00E+00	2.00	0.18	3
S59	0.89	8.60E-03	1.26E-09	3.10E-10	0.00	0.00	5
L60	0.84	5.20E-03	4.07E-11	4.63E-12	1.20	0.17	4
P61							
D62	0.90	7.60E-03	5.00E-10	6.21E-11	0.00	0.00	5
G63	0.74	1.70E-03	2.39E-11	1.05E-12	2.41	0.06	4
D64	0.91	3.40E-03	5.99E-11	5.41E-12	0.43	0.08	4
W65	0.76	7.60E-03	1.65E-11	4.32E-12	4.98	0.27	4
S66	0.91	1.94E-02	2.92E-09	2.74E-09	0.00	0.00	5
K67							
W68							
L69							
K70	0.78	1.10E-02	0.00E+00	0.00E+00	2.30	0.32	3
I71	0.89	6.00E-03	2.95E-11	6.82E-12	2.94	0.15	4
S72	0.89	4.50E-03	1.47E-11	5.50E-12	1.82	0.12	4
F73	0.89	3.20E-03	1.37E-11	4.12E-12	0.62	0.10	4
D74	0.83	3.60E-03	2.56E-11	2.69E-12	1.81	0.10	4
I75	0.89	3.50E-03	1.86E-11	4.63E-12	0.66	0.09	4
K76	0.96	3.90E-03	1.57E-10	3.35E-11	0.39	0.11	4
S77	0.83	3.10E-03	2.85E-11	2.32E-12	0.56	0.10	4
V78	0.97	8.80E-03	1.36E-10	1.27E-10	1.58	0.20	4
D79	0.70	1.04E-02	1.20E-09	2.57E-11	0.58	0.13	4
G80	0.83	8.80E-03	6.50E-10	4.54E-11	0.00	0.00	5
S81	0.81	1.04E-02	1.09E-09	3.09E-11	1.45	0.13	4

A82	0.83	4.80E-03	8.22E-11	4.96E-12	0.56	0.12	4
N83	0.79	5.80E-03	3.75E-10	2.12E-11	0.00	0.00	5
E84	0.77	5.20E-03	9.70E-12	3.53E-12	3.22	0.19	4
I85	0.89	3.10E-03	0.00E+00	0.00E+00	0.38	0.09	3
R86	0.84	5.10E-03	2.81E-11	3.96E-12	0.00	0.00	2
F87	0.81	3.90E-03	2.43E-11	2.66E-12	0.82	0.15	4
M88	0.88	5.10E-03	0.00E+00	0.00E+00	3.24	0.15	3
I89	0.81	5.50E-03	0.00E+00	0.00E+00	2.50	0.17	3
A90	0.84	4.40E-03	3.36E-11	4.78E-12	0.75	0.13	4
E91	0.83	8.40E-03	1.99E-11	5.60E-12	1.17	0.18	4
K92	0.97	7.20E-03	9.99E-11	7.82E-11	0.95	0.19	4
S93	0.84	4.50E-03	2.45E-11	4.98E-12	0.84	0.10	4
I94	0.72	8.10E-03	7.47E-12	2.70E-12	2.11	0.24	4
N95	0.83	2.80E-03	5.37E-11	2.68E-12	0.00	0.00	2
G96	0.78	7.20E-03	5.50E-10	2.72E-11	0.00	0.00	5
V97	0.78	6.80E-03	4.98E-10	2.26E-11	0.00	0.00	5
G98	0.79	7.30E-03	4.88E-10	2.16E-11	0.00	0.00	2
D99	0.72	2.90E-03	4.70E-11	1.73E-12	2.07	0.10	4
G100	0.81	6.00E-03	2.23E-11	3.78E-12	2.42	0.18	4
E101							
H102							
W103	0.81	6.00E-03	2.78E-11	4.14E-12	3.23	0.14	4
V104	0.87	6.50E-03	0.00E+00	0.00E+00	2.50	0.22	3
Y105	0.96	1.01E-02	9.80E-10	7.74E-10	0.00	0.00	5
S106	0.87	9.40E-03	7.89E-10	7.31E-11	0.00	0.00	5
I107	0.92	4.80E-03	4.14E-11	7.62E-12	1.01	0.13	4
T108	0.86	4.00E-03	2.55E-11	4.54E-12	1.39	0.12	4
P109							
D110	0.85	3.90E-03	2.55E-11	3.87E-12	1.88	0.10	4
S111	0.89	1.07E-02	6.91E-10	8.55E-11	0.00	0.00	5
S112	0.86	2.90E-03	5.46E-11	2.74E-12	0.16	0.06	4
W113	0.88	3.00E-03	1.82E-11	3.90E-12	1.21	0.09	4
K114	0.90	4.20E-03	8.00E-11	8.23E-12	0.00	0.00	2
T115	0.86	4.10E-03	3.43E-11	3.73E-12	1.20	0.17	4
I116	0.85	4.10E-03	2.09E-11	3.72E-12	2.17	0.13	4
E117	0.86	3.70E-03	4.65E-11	4.70E-12	1.20	0.11	4
I118	0.77	3.10E-03	2.32E-11	2.34E-12	1.58	0.15	4
P119							
F120	0.94	1.88E-02	7.11E-10	1.42E-09	0.00	0.00	5
S121	0.92	4.10E-03	2.85E-11	6.43E-12	1.66	0.11	4
S122	0.82	5.00E-03	0.00E+00	0.00E+00	2.02	0.14	3
F123	0.91	4.80E-03	4.12E-11	1.11E-11	3.01	0.15	4
R124	0.81	5.40E-03	1.28E-11	4.22E-12	1.81	0.15	4
R125	0.78	3.70E-03	2.07E-11	3.32E-12	1.69	0.18	4

R126	0.83	8.10E-03	1.54E-11	5.84E-12	0.95	0.24	4
L127	0.94	1.53E-02	8.49E-10	8.41E-10	0.00	0.00	5
D128	0.88	4.10E-03	2.85E-11	5.39E-12	1.63	0.15	4
Y129	0.91	4.50E-03	2.58E-11	7.30E-12	0.80	0.13	4
Q130	0.84	7.50E-03	0.00E+00	0.00E+00	1.05	0.22	3
P131							
P132							
G133	0.90	6.40E-03	0.00E+00	0.00E+00	3.45	0.29	3
Q134	0.83	7.10E-03	0.00E+00	0.00E+00	1.49	0.20	3
D135	0.68	5.12E-02	8.71E-09	2.46E-09	0.00	0.00	6
M136	0.91	1.30E-02	5.75E-11	2.08E-11	13.83	0.59	4
S137	0.87	3.10E-03	0.00E+00	0.00E+00	1.05	0.16	3
G138	0.93	6.70E-03	0.00E+00	0.00E+00	1.83	0.25	3
T139							
L140	0.92	6.20E-03	5.32E-11	1.11E-11	0.00	0.00	2
D141	0.78	5.90E-03	1.35E-11	3.70E-12	1.65	0.23	4
L142	0.81	7.30E-03	2.84E-11	4.24E-12	2.05	0.20	4
D143	0.89	3.90E-03	5.13E-11	4.15E-12	3.48	0.13	4
N144	0.89	4.90E-03	9.26E-11	9.48E-12	1.04	0.10	4
I145	0.80	7.10E-03	0.00E+00	0.00E+00	2.13	0.20	3
D146	0.80	4.40E-03	1.44E-11	3.21E-12	1.26	0.18	4
S147	0.84	4.40E-03	1.14E-11	3.83E-12	0.95	0.12	4
I148	0.89	5.50E-03	3.64E-11	5.69E-12	1.03	0.14	4
H149	0.86	4.00E-03	0.00E+00	0.00E+00	2.33	0.15	3
F150	0.86	4.30E-03	0.00E+00	0.00E+00	0.35	0.14	3
M151	0.87	5.00E-03	3.03E-11	4.24E-12	0.48	0.16	4
Y152	0.94	1.29E-02	1.32E-09	3.82E-10	0.00	0.00	2
A153	0.83	2.23E-02	1.40E-09	8.38E-10	0.00	0.00	5
N154	0.69	2.80E-03	1.18E-11	1.96E-12	1.05	0.13	4
N155	0.79	4.00E-03	0.00E+00	0.00E+00	1.22	0.16	3
K156	0.86	3.30E-03	1.69E-11	4.16E-12	0.00	0.00	2
S157	0.89	1.08E-02	8.34E-10	1.27E-10	0.00	0.00	5
G158	0.89	4.10E-03	3.24E-11	4.74E-12	0.50	0.13	4
K159	0.92	4.30E-03	6.37E-11	8.83E-12	0.00	0.00	2
F160	0.94	5.10E-03	8.92E-10	1.15E-10	0.00	0.00	5
V161	0.84	3.40E-03	1.54E-11	2.33E-12	1.37	0.09	4
V162	0.35	7.70E-02	8.71E-09	2.00E-09	0.00	0.00	6
D163	0.89	4.20E-03	1.90E-11	5.79E-12	1.66	0.10	4
N164	0.82	3.60E-03	1.41E-11	3.07E-12	1.05	0.12	4
I165	0.88	3.30E-03	2.21E-11	4.22E-12	2.01	0.07	4
K166	0.82	3.90E-03	1.13E-11	2.87E-12	0.94	0.13	4
L167	0.77	4.10E-03	3.71E-11	2.83E-12	2.56	0.14	4
I168	0.81	7.50E-03	2.45E-11	6.59E-12	5.42	0.22	4
G169	0.89	1.36E-02	5.84E-11	1.90E-11	6.29	0.51	4

A170	0.83	3.70E-03	1.14E-10	5.55E-12	2.74	0.12	4
L171	0.54	9.30E-03	7.93E-10	1.02E-11	0.63	0.10	4
E172	0.41	7.10E-03	6.34E-10	6.47E-12	0.81	0.08	4

Table C.9: Internal mobility parameters (S^2 , τ_e and R_{ex}) for the bound protein at 25 °C

Residue	25 °C bound						Model
	S^2		$\tau_{c,i}$		K_{ex}		
	S^2	Error	$\tau_{c,i}$	Error	K_{ex}	Error	
M1							
A2							
S3							
A4							
V5							
G6	0.58	3.42E-02	6.89E-10	8.81E-11	0.00	0.00	5
E7	0.73	9.90E-03	4.87E-11	7.65E-12	8.47	0.58	4
K8	0.66	1.96E-02	1.02E-10	1.57E-11	4.37	0.30	4
M9	0.80	1.45E-02	0.00E+00	0.00E+00	3.25	0.32	3
L10	0.74	1.77E-02	3.72E-11	1.06E-11	2.59	0.26	4
D11	0.73	9.80E-03	2.40E-11	7.04E-12	1.15	0.21	4
D12	0.80	6.70E-03	0.00E+00	0.00E+00	1.55	0.15	3
F13	0.97	9.50E-03	0.00E+00	0.00E+00	0.00	0.00	1
E14	0.87	7.50E-03	0.00E+00	0.00E+00	0.63	0.17	3
G15	0.72	8.30E-03	3.80E-11	4.77E-12	1.37	0.12	4
V16	0.79	4.80E-03	5.08E-11	5.74E-12	0.00	0.00	2
L17	0.75	1.40E-02	6.15E-10	8.70E-11	0.00	0.00	5
N18	0.82	2.19E-02	3.89E-10	1.13E-10	0.00	0.00	5
W19	0.78	1.21E-02	2.93E-11	9.47E-12	1.38	0.22	4
G20	0.84	1.27E-02	0.00E+00	0.00E+00	0.00	0.00	1
S21	0.80	8.70E-03	3.71E-11	9.70E-12	0.00	0.00	2
Y22							
S23	0.84	9.50E-03	5.02E-11	9.29E-12	0.55	0.16	4
G24	0.88	1.20E-02	0.00E+00	0.00E+00	0.00	0.00	1
E25	0.94	1.64E-02	2.54E-09	3.46E-09	0.00	0.00	5
G26	0.75	1.54E-02	0.00E+00	0.00E+00	3.82	0.34	3
A27	0.88	7.80E-03	0.00E+00	0.00E+00	1.26	0.19	3
K28	0.85	8.20E-03	0.00E+00	0.00E+00	1.34	0.15	3
V29	0.95	1.20E-02	8.74E-10	3.16E-09	0.00	0.00	5
S30							
T31							
K32	0.84	1.03E-02	4.74E-11	1.46E-11	0.00	0.00	2
I33	0.87	1.58E-02	9.31E-10	7.11E-10	0.00	0.00	5
V34	0.81	1.39E-02	5.89E-11	1.17E-11	13.43	0.52	4
S35	0.77	1.04E-02	1.54E-11	7.03E-12	2.69	0.18	4
G36	0.77	8.00E-03	0.00E+00	0.00E+00	2.25	0.17	3

K37	0.85	2.52E-02	0.00E+00	0.00E+00	1.44	0.46	3
T38	0.80	1.55E-02	3.74E-11	1.36E-11	1.03	0.27	4
G39							
N40	0.73	9.90E-03	0.00E+00	0.00E+00	3.25	0.23	3
G41	0.81	7.30E-03	0.00E+00	0.00E+00	1.64	0.20	3
M42	0.86	8.10E-03	0.00E+00	0.00E+00	0.62	0.26	3
E43	0.87	1.34E-02	7.05E-11	1.91E-11	0.96	0.21	4
V44	0.90	1.23E-02	0.00E+00	0.00E+00	0.00	0.00	1
S45	0.85	8.30E-03	2.99E-11	1.02E-11	1.46	0.13	4
Y46	0.86	4.01E-02	2.55E-10	1.10E-10	1.79	0.62	4
T47	0.75	8.50E-03	2.10E-11	7.57E-12	2.10	0.20	4
G48	0.52	2.94E-02	4.36E-10	7.14E-11	0.00	0.00	2
T49	0.79	1.07E-02	0.00E+00	0.00E+00	3.21	0.22	3
T50	0.76	1.30E-02	0.00E+00	0.00E+00	1.20	0.40	3
D51	0.83	7.90E-03	5.95E-11	9.45E-12	1.09	0.15	4
G52	0.80	9.00E-03	0.00E+00	0.00E+00	1.90	0.18	3
Y53	0.82	1.52E-02	3.78E-11	1.35E-11	1.04	0.32	4
W54	0.82	1.89E-02	3.33E-11	1.34E-11	0.99	0.33	4
G55	0.88	1.29E-02	0.00E+00	0.00E+00	0.72	0.27	3
T56	0.88	8.60E-03	0.00E+00	0.00E+00	0.00	0.00	1
V57	0.86	1.74E-02	0.00E+00	0.00E+00	1.23	0.35	3
Y58	0.89	2.33E-02	0.00E+00	0.00E+00	0.00	0.00	1
S59	0.66	2.21E-02	4.45E-11	1.40E-11	0.00	0.00	2
L60	0.79	1.64E-02	5.72E-11	1.39E-11	0.94	0.28	4
P61							
D62	0.78	1.13E-02	4.56E-11	8.77E-12	0.83	0.17	4
G63	0.91	7.70E-03	4.63E-10	9.18E-11	0.00	0.00	5
D64	0.92	5.80E-03	0.00E+00	0.00E+00	0.00	0.00	1
W65	0.70	1.47E-02	3.32E-11	1.06E-11	5.41	0.42	4
S66	0.91	1.04E-02	0.00E+00	0.00E+00	0.00	0.00	1
K67							
W68							
L69							
K70	0.81	2.35E-02	0.00E+00	0.00E+00	1.80	0.50	3
I71	0.83	1.35E-02	0.00E+00	0.00E+00	2.58	0.29	3
S72	0.90	8.40E-03	0.00E+00	0.00E+00	1.49	0.17	3
F73	0.86	1.06E-02	0.00E+00	0.00E+00	1.55	0.19	3
D74	0.84	9.50E-03	0.00E+00	0.00E+00	1.94	0.18	3
I75	0.90	6.50E-03	0.00E+00	0.00E+00	0.46	0.18	3
K76	0.91	2.52E-02	5.86E-10	1.33E-09	0.00	0.00	5
S77	0.95	1.60E-02	3.44E-09	3.60E-09	0.00	0.00	5
V78	0.83	2.44E-02	6.10E-10	1.33E-10	0.00	0.00	2
D79	0.77	2.74E-02	9.86E-10	3.20E-10	0.00	0.00	5
G80	0.76	4.36E-02	0.00E+00	0.00E+00	2.28	0.97	3

S81	0.77	1.88E-02	4.53E-11	1.29E-11	0.64	0.28	4
A82	0.78	3.02E-02	6.08E-11	3.22E-11	1.30	0.48	4
N83	0.80	1.40E-02	3.27E-10	6.25E-11	0.00	0.00	5
E84	0.67	1.90E-02	3.19E-11	1.08E-11	2.79	0.32	4
I85	0.91	9.30E-03	0.00E+00	0.00E+00	0.00	0.00	1
R86	0.96	1.39E-02	0.00E+00	0.00E+00	0.00	0.00	1
F87	0.86	1.22E-02	5.85E-11	2.15E-11	0.00	0.00	2
M88	0.80	1.74E-02	4.70E-11	1.59E-11	3.22	0.33	4
I89	0.75	1.28E-02	2.42E-11	9.45E-12	1.50	0.22	4
A90	0.82	1.44E-02	0.00E+00	0.00E+00	1.86	0.35	3
E91	0.77	1.98E-02	0.00E+00	0.00E+00	3.02	0.43	3
K92	0.91	2.29E-02	8.05E-10	2.25E-09	0.00	0.00	5
S93	0.86	1.30E-02	0.00E+00	0.00E+00	0.70	0.28	3
I94	0.74	1.97E-02	3.56E-11	1.43E-11	4.18	0.45	4
N95	0.80	1.06E-02	3.96E-11	1.07E-11	1.20	0.16	4
G96	0.82	2.05E-02	9.45E-10	2.68E-10	0.00	0.00	5
V97	0.77	1.42E-02	6.65E-10	8.56E-11	0.00	0.00	5
G98	0.79	2.46E-02	5.06E-10	1.24E-10	0.00	0.00	5
D99							
G100	0.85	2.67E-02	5.94E-11	2.78E-11	1.12	0.45	4
E101							
H102	0.81	2.62E-02	0.00E+00	0.00E+00	1.68	0.56	3
W103	0.83	9.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
V104	0.93	2.98E-02	1.57E-09	3.68E-09	0.00	0.00	5
Y105	0.85	9.90E-03	0.00E+00	0.00E+00	0.51	0.20	3
S106	0.88	8.60E-03	0.00E+00	0.00E+00	0.00	0.00	1
I107	0.83	1.46E-02	3.89E-11	1.48E-11	0.49	0.24	4
T108	0.89	8.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
P109							
D110	0.82	1.21E-02	0.00E+00	0.00E+00	1.63	0.29	3
S111	0.78	8.90E-03	0.00E+00	0.00E+00	1.28	0.13	3
S112	0.89	1.43E-02	6.19E-10	1.59E-10	0.00	0.00	5
W113	0.83	1.24E-02	1.49E-10	1.76E-11	0.92	0.19	4
K114	0.96	1.67E-02	3.19E-09	3.66E-09	0.00	0.00	5
T115							
I116	0.84	1.56E-02	3.98E-11	1.85E-11	0.55	0.26	4
E117	0.81	1.01E-02	4.77E-11	1.02E-11	1.40	0.15	4
I118							
P119							
F120	0.80	2.50E-02	5.70E-11	2.15E-11	1.45	0.36	4
S121	0.92	8.40E-03	1.00E-10	2.93E-11	0.00	0.00	2
S122	0.84	1.03E-02	0.00E+00	0.00E+00	1.22	0.19	3
F123	0.71	1.02E-02	5.51E-11	7.57E-12	4.53	0.23	4
R124	0.84	1.39E-02	4.46E-11	1.58E-11	0.81	0.25	4

R125	0.86	8.90E-03	0.00E+00	0.00E+00	0.00	0.00	1
R126	0.75	2.00E-02	6.99E-11	1.51E-11	3.34	0.37	4
L127	0.89	1.59E-02	0.00E+00	0.00E+00	0.00	0.00	1
D128	0.79	1.41E-02	0.00E+00	0.00E+00	3.03	0.48	3
Y129	0.83	9.90E-03	0.00E+00	0.00E+00	1.51	0.21	3
Q130	0.84	1.08E-02	0.00E+00	0.00E+00	0.00	0.00	1
P131							
P132							
G133	0.88	1.40E-02	0.00E+00	0.00E+00	1.77	0.28	3
Q134	0.81	1.03E-02	0.00E+00	0.00E+00	3.25	0.26	3
D135							
M136	0.89	1.63E-02	0.00E+00	0.00E+00	1.40	0.37	3
S137	0.88	8.80E-03	4.19E-11	1.14E-11	0.76	0.15	4
G138	0.95	1.48E-02	0.00E+00	0.00E+00	0.00	0.00	1
T139							
L140	0.85	1.07E-02	0.00E+00	0.00E+00	1.09	0.27	3
D141	0.86	1.23E-02	0.00E+00	0.00E+00	0.00	0.00	1
L142	0.84	1.12E-02	0.00E+00	0.00E+00	0.00	0.00	1
D143	0.84	5.90E-03	2.44E-11	5.36E-12	1.02	0.11	4
N144	0.85	8.10E-03	0.00E+00	0.00E+00	1.53	0.20	3
I145	0.77	2.03E-02	5.20E-11	1.95E-11	3.20	0.34	4
D146							
S147	0.87	1.40E-02	0.00E+00	0.00E+00	1.93	0.27	3
I148	0.90	1.76E-02	0.00E+00	0.00E+00	1.23	0.41	3
H149	0.82	2.04E-02	3.50E-11	1.66E-11	1.37	0.44	4
F150	0.82	1.43E-02	4.39E-11	1.32E-11	1.52	0.24	4
M151	0.88	1.03E-02	0.00E+00	0.00E+00	0.00	0.00	1
Y152	0.85	2.42E-02	0.00E+00	0.00E+00	0.00	0.00	1
A153	0.86	2.65E-02	0.00E+00	0.00E+00	0.00	0.00	1
N154	0.62	1.42E-02	1.87E-11	6.48E-12	2.21	0.22	4
N155							
K156	0.82	6.20E-03	0.00E+00	0.00E+00	0.60	0.14	3
S157	0.75	1.25E-02	3.60E-11	1.29E-11	3.01	0.18	4
G158	1.09	2.62E-02	8.43E-09	3.47E-09	0.00	0.00	6
K159	0.85	1.00E-02	0.00E+00	0.00E+00	0.00	0.00	1
F160	0.88	1.00E-02	0.00E+00	0.00E+00	0.67	0.19	3
V161	0.86	8.50E-03	6.47E-11	1.53E-11	0.00	0.00	2
V162							
D163	0.92	1.02E-02	0.00E+00	0.00E+00	0.00	0.00	1
N164	0.80	1.24E-02	3.55E-11	1.07E-11	1.55	0.25	4
I165	0.84	7.30E-03	0.00E+00	0.00E+00	0.69	0.14	3
K166	0.81	1.10E-02	2.66E-11	1.23E-11	2.90	0.21	4
L167							
I168	0.80	1.10E-02	0.00E+00	0.00E+00	7.48	0.47	3

G169	0.91	2.53E-02	0.00E+00	0.00E+00	3.52	0.66	3
A170	0.82	1.07E-02	1.91E-10	3.87E-11	0.00	0.00	2
L171	0.57	1.30E-02	4.43E-10	2.57E-11	0.00	0.00	5
E172	0.24	8.40E-03	7.25E-10	1.94E-11	0.00	0.00	5

Table C.10: Internal mobility parameters (S^2 , τ_e and R_{ex}) for the free protein at 50 °C

Residue	50 °C free						Model
	S^2		$\tau_{c,i}$		K_{ex}		
	S^2	Error	$\tau_{c,i}$	Error	K_{ex}	Error	
M1							
A2							
S3	0.79	1.30E-02	2.09E-11	7.71E-12	0.00	0.00	2
A4							
V5							
G6	0.00	7.77E-02	1.21E-09	8.51E-11	0.00	0.00	6
E7	0.86	1.90E-03	1.04E-10	4.27E-12	0.76	0.07	4
K8	0.80	4.50E-03	4.65E-11	4.21E-12	1.81	0.19	4
M9	0.85	5.00E-03	3.03E-11	4.84E-12	0.46	0.19	4
L10	0.81	3.20E-03	1.68E-11	3.62E-12	0.74	0.11	4
D11	0.85	3.30E-03	2.44E-11	4.29E-12	0.00	0.00	2
D12	0.87	3.80E-03	3.82E-11	4.62E-12	0.00	0.00	2
F13	0.91	3.80E-03	8.14E-11	8.19E-12	0.49	0.10	4
E14	0.86	2.60E-03	3.47E-11	3.16E-12	0.62	0.08	4
G15	0.88	1.03E-02	8.49E-10	8.93E-11	0.00	0.00	5
V16	0.74	1.07E-02	9.28E-10	4.49E-11	0.00	0.00	5
L17	0.67	1.43E-02	1.00E-09	4.52E-11	0.00	0.00	5
N18	0.80	2.84E-02	4.43E-10	1.21E-10	0.00	0.00	5
W19	0.74	1.90E-03	2.52E-11	1.62E-12	0.45	0.07	4
G20	0.84	3.20E-03	4.12E-11	3.69E-12	0.00	0.00	2
S21	0.85	2.01E-02	8.16E-10	1.31E-10	0.00	0.00	5
Y22	0.90	3.30E-03	5.76E-11	6.35E-12	0.00	0.00	2
S23	0.93	9.50E-03	1.42E-09	2.71E-10	0.00	0.00	5
G24	0.87	3.90E-03	4.11E-11	5.03E-12	0.77	0.11	4
E25	0.88	2.58E-02	5.65E-09	1.34E-09	0.00	0.00	6
G26	0.85	6.90E-03	2.23E-11	5.84E-12	0.00	0.00	2
A27	0.91	2.70E-03	5.34E-11	5.49E-12	0.22	0.09	4
K28	0.86	2.90E-03	5.11E-11	4.93E-12	0.37	0.10	4
V29	0.84	3.20E-03	3.42E-11	3.18E-12	0.73	0.09	4
S30	0.90	3.00E-03	5.66E-11	6.08E-12	0.00	0.00	2
T31	0.96	1.20E-02	9.13E-10	3.48E-10	0.00	0.00	5
K32	0.86	3.30E-03	4.95E-11	5.25E-12	0.25	0.09	4
I33	0.90	1.23E-02	5.89E-10	1.11E-10	0.00	0.00	5
V34	0.90	3.10E-03	4.51E-11	5.40E-12	0.44	0.09	4
S35	0.84	2.70E-03	2.54E-11	2.49E-12	0.63	0.11	4

G36	0.86	2.90E-03	4.14E-11	3.81E-12	0.00	0.00	2
K37	0.92	7.50E-03	4.20E-11	1.35E-11	0.00	0.00	2
T38	0.86	3.60E-03	4.01E-11	4.66E-12	0.38	0.12	4
G39							
N40	0.84	3.30E-03	2.70E-11	3.47E-12	0.97	0.11	4
G41	0.87	3.20E-03	3.33E-11	3.99E-12	0.54	0.10	4
M42	0.91	1.85E-02	6.71E-10	1.90E-10	0.00	0.00	5
E43	0.92	3.30E-03	6.54E-11	7.34E-12	0.00	0.00	2
V44	0.90	4.80E-03	3.48E-11	8.89E-12	0.00	0.00	2
S45	0.86	3.40E-03	3.81E-11	3.94E-12	0.67	0.09	4
Y46	0.93	6.00E-03	1.18E-10	4.09E-11	0.00	0.00	2
T47	0.88	4.40E-03	5.88E-11	8.58E-12	0.30	0.13	4
G48	0.83	3.90E-03	5.76E-11	4.63E-12	0.00	0.00	2
T49	0.86	3.10E-03	2.50E-11	3.53E-12	1.35	0.11	4
T50							
D51	0.79	3.90E-03	2.74E-11	3.31E-12	0.00	0.00	2
G52	0.82	2.40E-03	3.80E-11	2.69E-12	0.31	0.09	4
Y53	0.83	3.10E-03	3.54E-11	2.85E-12	0.00	0.00	2
W54	0.89	4.90E-03	3.87E-11	5.71E-12	0.00	0.00	2
G55	0.90	4.30E-03	4.37E-11	8.05E-12	0.00	0.00	2
T56	0.91	3.20E-03	4.79E-11	7.46E-12	0.00	0.00	2
V57	0.90	4.80E-03	4.22E-11	8.16E-12	0.26	0.12	4
Y58							
S59	0.83	3.10E-03	5.88E-11	3.47E-12	0.00	0.00	2
L60	0.73	2.90E-03	3.02E-11	1.84E-12	0.00	0.00	2
P61							
D62	0.77	2.50E-03	3.19E-11	1.90E-12	0.00	0.00	2
G63	0.70	1.93E-02	1.43E-09	9.71E-11	0.00	0.00	5
D64							
W65	0.86	4.10E-03	4.67E-11	5.32E-12	1.15	0.14	4
S66	0.92	3.50E-03	7.72E-11	8.85E-12	0.24	0.10	4
K67							
W68							
L69							
K70	0.87	4.50E-03	5.26E-11	6.31E-12	0.00	0.00	2
I71	0.94	3.50E-03	6.99E-11	1.17E-11	0.64	0.11	4
S72	0.91	2.70E-03	7.62E-11	7.75E-12	0.44	0.10	4
F73	0.93	3.20E-03	0.00E+00	0.00E+00	0.90	0.10	3
D74	0.90	2.80E-03	8.17E-11	6.63E-12	0.00	0.00	2
I75	0.93	2.90E-03	7.32E-11	9.00E-12	0.00	0.00	2
K76	0.92	3.60E-03	6.20E-11	7.59E-12	0.27	0.12	4
S77	0.92	3.82E-02	5.65E-09	1.54E-09	1.14	0.25	6
V78	0.75	2.15E-02	5.57E-10	7.17E-11	0.00	0.00	5
D79	0.71	2.79E-02	6.50E-10	7.91E-11	0.00	0.00	5

G80	0.82	1.51E-02	5.68E-11	1.79E-11	0.00	0.00	2
S81	0.63	2.96E-02	1.39E-09	1.17E-10	0.00	0.00	5
A82	0.70	2.35E-02	8.19E-11	1.71E-11	0.00	0.00	2
N83	0.72	1.87E-02	5.95E-10	5.57E-11	0.00	0.00	5
E84	0.84	1.78E-02	8.29E-10	1.15E-10	0.00	0.00	5
I85	0.92	5.40E-03	1.42E-10	2.79E-11	0.48	0.15	4
R86	0.90	5.20E-03	3.57E-11	1.04E-11	0.00	0.00	2
F87	0.92	4.30E-03	7.21E-11	1.04E-11	0.00	0.00	2
M88	0.92	4.40E-03	7.44E-11	1.17E-11	0.71	0.12	4
I89	0.86	4.80E-03	7.38E-11	7.13E-12	0.00	0.00	2
A90	0.92	4.90E-03	3.36E-11	1.16E-11	0.31	0.12	4
E91	0.90	6.00E-03	0.00E+00	0.00E+00	0.00	0.00	1
K92	0.84	5.60E-03	4.07E-11	5.40E-12	0.00	0.00	2
S93	0.81	4.10E-03	3.48E-11	3.42E-12	0.00	0.00	2
I94	0.91	4.50E-03	9.71E-11	1.24E-11	0.69	0.22	4
N95	0.86	2.80E-03	7.94E-11	4.56E-12	0.00	0.00	2
G96	0.67	1.87E-02	8.07E-10	5.09E-11	0.00	0.00	5
V97	0.58	1.00E-02	8.48E-10	2.40E-11	0.00	0.00	5
G98	0.71	2.28E-02	7.90E-10	7.14E-11	0.00	0.00	5
D99	0.80	2.80E-03	5.02E-11	2.85E-12	0.90	0.07	4
G100	0.83	4.90E-03	5.26E-11	4.72E-12	0.25	0.14	4
E101	0.55	2.08E-02	1.47E-09	9.17E-11	0.00	0.00	5
H102	0.86	1.18E-02	3.51E-11	1.45E-11	3.10	0.58	4
W103	0.88	4.70E-03	2.41E-11	9.38E-12	0.00	0.00	2
V104	0.91	6.90E-03	7.27E-11	1.62E-11	0.37	0.18	4
Y105	0.92	3.20E-03	4.75E-11	8.24E-12	0.00	0.00	2
S106	0.84	3.50E-03	0.00E+00	0.00E+00	0.00	0.00	1
I107	0.90	4.40E-03	7.81E-11	9.69E-12	0.00	0.00	2
T108							
P109							
D110	0.86	4.90E-03	4.79E-11	6.99E-12	0.28	0.12	4
S111							
S112							
W113	0.87	2.74E-02	6.30E-10	1.86E-10	0.00	0.00	5
K114	0.87	4.20E-03	5.25E-11	5.65E-12	0.00	0.00	2
T115	0.91	1.68E-02	6.36E-10	1.72E-10	0.00	0.00	5
I116	0.89	4.00E-03	4.61E-11	6.04E-12	0.35	0.14	4
E117	0.82	2.60E-03	3.01E-11	2.81E-12	0.34	0.10	4
I118	0.86	2.50E-03	4.78E-11	3.91E-12	0.00	0.00	2
P119							
F120	0.89	3.40E-03	2.15E-11	6.40E-12	1.20	0.15	4
S121	0.94	2.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
S122	0.85	3.90E-03	4.93E-11	3.72E-12	0.43	0.15	4
F123	0.81	3.10E-03	7.62E-11	3.74E-12	0.00	0.00	2

R124	0.85	4.90E-03	4.86E-11	5.72E-12	0.00	0.00	2
R125	0.87	4.90E-03	7.03E-11	6.89E-12	0.00	0.00	2
R126	0.70	3.50E-03	4.76E-11	2.46E-12	0.00	0.00	2
L127	0.90	7.30E-03	2.69E-11	1.22E-11	1.09	0.33	4
D128	0.92	4.60E-03	0.00E+00	0.00E+00	0.56	0.22	3
Y129	0.85	3.70E-03	1.49E-11	3.74E-12	0.00	0.00	2
Q130	0.82	5.30E-03	1.75E-11	4.22E-12	0.95	0.18	4
P131							
P132							
G133	0.91	5.60E-03	7.01E-11	1.11E-11	1.98	0.23	4
Q134	0.91	4.60E-03	6.83E-11	1.01E-11	0.82	0.18	4
D135	0.89	5.60E-03	4.78E-11	8.48E-12	2.76	0.30	4
M136	0.92	7.40E-03	6.15E-11	1.98E-11	4.56	0.39	4
S137	0.93	5.40E-03	5.10E-11	1.27E-11	1.04	0.24	4
G138	0.95	6.80E-03	8.87E-11	5.25E-11	0.00	0.00	2
T139	0.78	7.09E-02	2.36E-09	7.59E-10	0.00	0.00	5
L140	0.82	4.20E-03	3.86E-11	3.70E-12	0.00	0.00	2
D141	0.84	4.80E-03	3.15E-11	5.28E-12	1.27	0.21	4
L142	0.85	3.70E-03	5.23E-11	5.09E-12	0.00	0.00	2
D143	0.88	3.00E-03	5.55E-11	4.24E-12	0.62	0.10	4
N144	0.77	2.70E-03	5.71E-11	2.60E-12	0.20	0.09	4
I145	0.86	4.10E-03	4.48E-11	5.78E-12	1.02	0.11	4
D146	0.86	2.68E-02	1.68E-09	5.21E-10	0.00	0.00	5
S147	0.92	4.10E-03	9.36E-11	1.23E-11	0.00	0.00	2
I148	0.90	1.42E-02	1.17E-09	2.20E-10	0.00	0.00	5
H149	0.89	5.10E-03	6.65E-11	8.36E-12	0.58	0.15	4
F150	0.86	3.50E-03	3.50E-11	4.81E-12	0.00	0.00	2
M151	0.89	3.60E-03	2.42E-11	5.71E-12	0.42	0.13	4
Y152	0.92	5.60E-03	8.26E-11	1.34E-11	0.00	0.00	2
A153	0.91	9.40E-03	3.82E-11	1.64E-11	0.00	0.00	2
N154	0.83	2.52E-02	1.63E-09	3.45E-10	0.00	0.00	5
N155							
K156	0.84	2.60E-03	1.43E-11	3.23E-12	0.45	0.09	4
S157							
G158	0.89	4.10E-03	7.98E-11	9.04E-12	0.00	0.00	2
K159	0.83	3.80E-03	4.19E-11	4.06E-12	0.00	0.00	2
F160	0.93	3.70E-03	1.41E-10	1.98E-11	0.41	0.11	4
V161	0.90	3.10E-03	5.80E-11	5.13E-12	0.00	0.00	2
V162	0.88	4.30E-03	7.59E-11	8.95E-12	0.80	0.12	4
D163	0.84	1.31E-02	7.38E-10	7.68E-11	0.00	0.00	5
N164	0.87	3.70E-03	6.15E-11	4.93E-12	0.00	0.00	2
I165	0.92	2.20E-03	6.19E-11	8.79E-12	0.74	0.03	4
K166	0.92	3.42E-02	4.06E-10	2.71E-10	0.73	0.24	4
L167	0.87	4.50E-03	1.29E-10	1.34E-11	0.00	0.00	2

I168	0.93	4.30E-03	1.56E-10	3.27E-11	3.47	0.17	4
G169	0.91	5.30E-03	1.11E-10	1.50E-11	2.89	0.28	4
A170	0.71	2.20E-03	8.53E-11	2.01E-12	0.78	0.09	4
L171	0.51	9.20E-03	6.93E-10	1.64E-11	0.00	0.00	5
E172							

Table C.11: Internal mobility parameters (S^2 , τ_e and R_{ex}) for the bound protein at 50 °C

Residue	50 °C bound						Model
	S^2		$\tau_{c,i}$		K_{ex}		
	S^2	Error	$\tau_{c,i}$	Error	K_{ex}	Error	
M1							
A2							
S3							
A4							
V5							
G6							
E7	0.80	1.66E-02	1.34E-10	8.09E-11	0.00	0.00	2
K8	0.50	1.94E-01	1.61E-09	1.81E-10	1.83	0.91	4
M9	0.54	1.40E-01	0.00E+00	0.00E+00	9.36	3.05	3
L10	0.87	1.47E-02	6.73E-11	3.38E-11	0.94	0.16	4
D11	0.89	2.66E-02	1.41E-09	1.43E-09	0.00	0.00	5
D12	0.85	1.18E-02	4.45E-11	1.85E-11	0.68	0.12	4
F13	0.95	1.13E-02	0.00E+00	0.00E+00	0.33	0.12	3
E14	0.87	8.50E-03	6.59E-11	1.90E-11	0.38	0.10	4
G15	0.81	2.43E-02	1.32E-09	5.30E-10	0.00	0.00	5
V16	0.78	5.26E-02	8.65E-10	3.08E-10	0.00	0.00	5
L17	0.82	6.10E-03	7.66E-11	1.25E-11	0.00	0.00	2
N18	0.77	3.01E-02	1.62E-10	9.66E-11	0.00	0.00	2
W19	0.78	7.90E-03	5.68E-11	1.09E-11	0.00	0.00	2
G20	0.85	9.10E-03	0.00E+00	0.00E+00	0.53	0.13	3
S21	0.67	5.68E-02	0.00E+00	0.00E+00	0.00	0.00	1
Y22	0.79	1.86E-02	0.00E+00	0.00E+00	1.20	0.23	3
S23	0.91	3.03E-02	4.86E-09	1.59E-09	0.00	0.00	6
G24	0.93	1.27E-02	0.00E+00	0.00E+00	0.65	0.16	3
E25	0.82	3.65E-02	0.00E+00	0.00E+00	0.00	0.00	1
G26							
A27	0.90	2.34E-02	9.26E-10	3.67E-10	0.00	0.00	2
K28	0.88	1.17E-02	0.00E+00	0.00E+00	0.61	0.13	3
V29	0.90	9.10E-03	0.00E+00	0.00E+00	0.00	0.00	1
S30	0.89	7.40E-03	0.00E+00	0.00E+00	0.00	0.00	1
T31	0.78	4.55E-02	1.53E-09	8.90E-10	0.00	0.00	5
K32	0.94	8.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
I33	0.86	9.10E-03	7.28E-11	2.15E-11	0.00	0.00	2
V34	0.91	8.60E-03	0.00E+00	0.00E+00	0.00	0.00	1

S35	0.85	6.46E-02	4.86E-09	1.69E-09	0.00	0.00	6
G36	0.83	1.16E-02	4.92E-11	1.55E-11	0.27	0.12	4
K37	0.86	4.97E-02	0.00E+00	0.00E+00	0.00	0.00	1
T38	0.72	3.22E-02	1.11E-09	1.51E-10	0.00	0.00	2
G39							
N40	0.90	1.31E-02	0.00E+00	0.00E+00	0.76	0.17	3
G41	0.90	1.41E-02	5.96E-11	3.28E-11	0.56	0.16	4
M42	0.88	8.40E-03	0.00E+00	0.00E+00	0.00	0.00	1
E43	0.90	3.41E-02	1.36E-10	1.86E-10	0.34	0.18	4
V44	0.91	1.26E-02	0.00E+00	0.00E+00	0.56	0.18	3
S45	0.83	9.50E-03	3.38E-11	1.53E-11	0.24	0.09	4
Y46	0.88	1.61E-02	0.00E+00	0.00E+00	0.00	0.00	1
T47	0.84	1.49E-02	9.57E-11	3.34E-11	0.43	0.15	4
G48	0.75	3.56E-02	9.17E-10	2.79E-10	0.00	0.00	5
T49	0.84	1.50E-02	5.66E-11	1.99E-11	1.04	0.18	4
T50							
D51	0.78	4.11E-02	4.86E-09	9.10E-10	0.00	0.00	2
G52	0.84	1.16E-02	5.35E-11	1.68E-11	0.79	0.12	4
Y53	0.86	1.05E-02	3.19E-11	1.93E-11	0.29	0.10	4
W54	0.90	3.66E-02	4.86E-09	1.72E-09	0.00	0.00	6
G55	0.91	9.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
T56	0.91	7.50E-03	0.00E+00	0.00E+00	0.00	0.00	1
V57	0.93	1.10E-02	0.00E+00	0.00E+00	0.00	0.00	1
Y58	0.94	2.20E-02	0.00E+00	0.00E+00	1.40	0.28	3
S59	0.81	1.59E-02	7.73E-11	5.59E-11	0.00	0.00	2
L60	0.81	1.50E-02	1.05E-10	3.16E-11	0.31	0.15	4
P61							
D62	0.83	5.06E-02	1.13E-09	1.13E-09	0.00	0.00	5
G63							
D64	0.93	2.40E-02	0.00E+00	0.00E+00	0.00	0.00	1
W65	0.86	1.30E-02	7.35E-11	3.80E-11	1.18	0.15	4
S66	0.91	2.80E-02	4.86E-09	1.73E-09	0.00	0.00	6
K67							
W68	0.83	1.14E-02	0.00E+00	0.00E+00	3.68	0.25	3
L69							
K70	0.90	4.18E-02	1.04E-10	2.13E-10	0.42	0.24	4
I71	0.91	1.07E-02	8.26E-11	1.20E-10	0.00	0.00	2
S72	0.89	4.85E-02	1.52E-10	2.03E-10	0.30	0.24	4
F73	0.91	4.14E-02	1.22E-10	2.54E-10	0.46	0.20	4
D74	0.90	1.07E-02	8.42E-11	3.40E-11	0.28	0.11	4
I75	0.92	7.50E-03	0.00E+00	0.00E+00	0.74	0.10	3
K76	0.78	1.65E-02	5.28E-11	2.45E-11	0.00	0.00	2
S77	0.83	2.72E-02	0.00E+00	0.00E+00	1.33	0.29	3
V78	0.92	2.35E-02	0.00E+00	0.00E+00	0.00	0.00	1

D79	0.65	1.26E-01	9.83E-10	1.23E-09	0.00	0.00	5
G80							
S81							
A82							
N83	0.43	6.84E-02	1.44E-09	4.81E-10	0.00	0.00	5
E84	0.72	3.61E-02	1.14E-09	2.19E-10	0.00	0.00	2
I85	0.84	2.36E-02	2.46E-10	1.38E-10	0.00	0.00	2
R86	0.88	8.29E-02	1.59E-10	3.08E-10	0.57	0.43	4
F87	0.91	9.10E-03	0.00E+00	0.00E+00	0.00	0.00	1
M88	0.91	1.48E-02	0.00E+00	0.00E+00	1.01	0.18	3
I89	0.84	1.20E-02	8.81E-11	7.66E-11	0.00	0.00	2
A90	0.89	1.27E-01	2.15E-10	4.36E-10	0.60	0.63	4
E91	0.81	2.29E-02	6.52E-11	3.25E-11	0.67	0.24	4
K92	0.88	3.97E-02	4.86E-09	1.78E-09	0.00	0.00	6
S93	0.86	6.60E-03	0.00E+00	0.00E+00	0.00	0.00	1
I94	0.89	1.52E-02	0.00E+00	0.00E+00	0.00	0.00	1
N95	0.88	1.51E-02	0.00E+00	0.00E+00	0.00	0.00	1
G96	0.71	4.70E-02	1.26E-10	1.55E-10	0.00	0.00	2
V97	0.55	2.66E-02	8.19E-10	7.96E-11	0.00	0.00	2
G98							
D99							
G100	0.88	1.20E-02	0.00E+00	0.00E+00	0.00	0.00	1
E101							
H102	0.87	1.70E-02	0.00E+00	0.00E+00	0.00	0.00	1
W103	0.84	1.09E-02	0.00E+00	0.00E+00	0.00	0.00	1
V104	0.87	1.77E-02	0.00E+00	0.00E+00	0.57	0.20	3
Y105	0.92	8.80E-03	0.00E+00	0.00E+00	0.00	0.00	1
S106	0.70	5.99E-02	0.00E+00	0.00E+00	0.00	0.00	1
I107	0.85	1.13E-02	5.74E-11	2.80E-11	0.00	0.00	2
T108	0.85	1.37E-02	8.32E-11	2.08E-11	0.49	0.15	4
P109							
D110	0.75	4.25E-02	6.04E-11	4.05E-11	1.06	0.43	4
S111	0.72	3.66E-02	6.15E-11	4.74E-11	0.00	0.00	2
S112	0.50	5.66E-02	4.86E-09	1.09E-09	0.00	0.00	6
W113	0.81	3.65E-02	9.72E-10	2.99E-10	0.00	0.00	2
K114	0.88	1.00E-02	0.00E+00	0.00E+00	0.00	0.00	1
T115	0.90	2.57E-02	1.37E-10	2.48E-10	0.00	0.00	2
I116	0.91	9.40E-03	0.00E+00	0.00E+00	0.00	0.00	1
E117	0.62	6.42E-02	1.90E-09	1.07E-09	0.00	0.00	5
I118	0.91	2.27E-02	7.67E-10	5.86E-10	0.00	0.00	5
P119							
F120	0.87	1.17E-02	0.00E+00	0.00E+00	0.74	0.14	3
S121	0.99	2.94E-02	0.00E+00	0.00E+00	0.00	0.00	1
S122	0.92	7.00E-03	0.00E+00	0.00E+00	0.73	0.10	3

F123	0.85	1.53E-02	8.65E-11	2.96E-11	0.54	0.19	4
R124	0.79	3.58E-02	6.33E-10	2.98E-10	0.00	0.00	2
R125	0.88	2.99E-02	5.69E-10	3.35E-10	0.00	0.00	2
R126	0.84	1.76E-02	9.47E-11	3.86E-11	0.44	0.19	4
L127							
D128	0.75	7.22E-02	0.00E+00	0.00E+00	0.00	0.00	1
Y129	0.86	7.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
Q130	0.82	1.63E-02	4.77E-11	1.87E-11	0.75	0.17	4
P131							
P132							
G133	0.96	1.93E-02	0.00E+00	0.00E+00	0.00	0.00	1
Q134	0.93	1.20E-02	0.00E+00	0.00E+00	0.00	0.00	1
D135	0.90	9.96E-02	1.70E-10	4.40E-10	0.64	0.55	4
M136	0.96	1.35E-02	0.00E+00	0.00E+00	0.00	0.00	1
S137	0.86	2.85E-02	1.47E-09	4.53E-10	0.00	0.00	2
G138	0.88	4.13E-02	1.05E-09	6.57E-10	0.00	0.00	2
T139	0.69	2.98E-02	2.18E-09	3.46E-10	0.00	0.00	2
L140	0.95	1.34E-02	0.00E+00	0.00E+00	0.64	0.24	3
D141	0.80	1.87E-02	4.46E-11	1.98E-11	1.20	0.24	4
L142	0.85	9.30E-03	5.29E-11	2.10E-11	0.00	0.00	2
D143	0.79	1.91E-02	4.90E-11	1.78E-11	0.99	0.24	4
N144	0.85	1.48E-02	9.04E-11	9.84E-11	0.00	0.00	2
I145	0.88	9.90E-03	0.00E+00	0.00E+00	0.00	0.00	1
D146	0.89	6.24E-02	1.21E-10	2.57E-10	0.82	0.35	4
S147	0.92	8.20E-03	0.00E+00	0.00E+00	0.00	0.00	1
I148	0.92	1.33E-02	1.31E-10	1.97E-10	0.00	0.00	2
H149	0.87	1.54E-02	0.00E+00	0.00E+00	0.91	0.20	3
F150	0.86	1.16E-02	5.43E-11	2.45E-11	0.55	0.09	4
M151	0.90	1.05E-02	0.00E+00	0.00E+00	0.30	0.14	3
Y152	0.98	1.21E-02	0.00E+00	0.00E+00	0.00	0.00	1
A153	0.86	3.11E-02	0.00E+00	0.00E+00	0.00	0.00	1
N154	0.78	1.05E-02	0.00E+00	0.00E+00	0.00	0.00	1
N155							
K156	0.91	7.70E-03	0.00E+00	0.00E+00	0.00	0.00	1
S157							
G158	0.86	1.95E-02	1.55E-10	1.47E-10	0.00	0.00	2
K159	0.76	3.71E-02	2.54E-09	1.36E-09	0.00	0.00	5
F160	0.92	7.60E-03	0.00E+00	0.00E+00	0.00	0.00	1
V161	0.88	7.20E-03	5.22E-11	2.34E-11	0.00	0.00	2
V162	0.87	8.87E-02	1.29E-10	2.37E-10	0.49	0.46	4
D163	0.93	1.76E-02	1.08E-09	1.12E-09	0.00	0.00	5
N164	0.88	1.13E-02	6.51E-11	3.04E-11	0.00	0.00	2
I165							
K166	0.90	1.67E-02	1.10E-10	1.53E-10	0.00	0.00	2

L167	0.78	4.93E-02	6.65E-10	2.68E-10	0.00	0.00	5
I168	0.85	5.82E-02	0.00E+00	0.00E+00	1.76	0.70	3
G169	1.00	1.11E-02	0.00E+00	0.00E+00	0.00	0.00	1
A170	0.69	2.22E-02	8.59E-11	2.45E-11	0.00	0.00	2
L171	0.19	2.13E-01	9.37E-10	2.94E-10	3.32	1.00	4
E172							

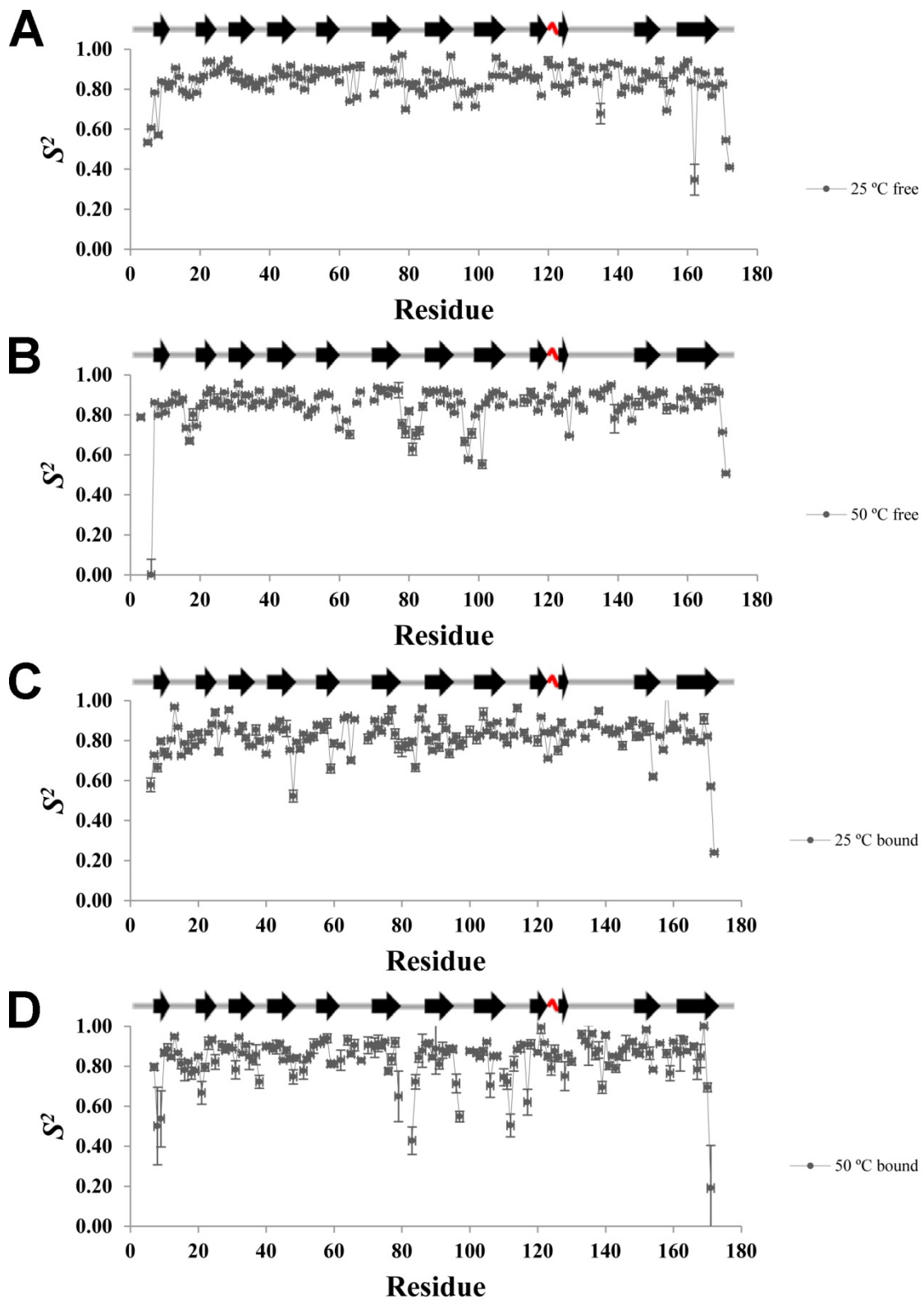


Figure C.5: S^2 values determined for the free and bound *CtCBM11* at 25 and 50 °C.

A) Free *CtCBM11* at 25 °C; B) Free *CtCBM11* at 50 °C; C) Bound *CtCBM11* at 25 °C; D) Bound *CtCBM11* at 50 °C.

Table C.12: Estimation of the conformational entropy from NMR relaxation data.

	$S_{conf_{25^{\circ}C}}^{bound} - S_{conf_{25^{\circ}C}}^{free}$	$S_{conf_{50^{\circ}C}}^{bound} - S_{conf_{50^{\circ}C}}^{free}$	$S_{conf_{50^{\circ}C}}^{free} - S_{conf_{25^{\circ}C}}^{free}$	$S_{conf_{50^{\circ}C}}^{bound} - S_{conf_{25^{\circ}C}}^{bound}$
	$\Delta S_{conf} (J.mol^{-1}.K^{-1})$			
M1				
A2				
S3				
A4				
V5				
G6	0.63 ± 0.07			
E7	2.01 ± 0.02	3.32 ± 0.02	-3.87 ± 0.01	-2.55 ± 0.03
K8	-2.27 ± 0.05	8.27 ± 0.39	-6.81 ± 0.03	3.73 ± 0.42
M9	1.96 ± 0.02	9.91 ± 0.27	-0.49 ± 0.01	7.47 ± 0.28
L10	3.77 ± 0.03	-3.04 ± 0.02	1.05 ± 0.01	-5.76 ± 0.04
D11	3.10 ± 0.02	-2.10 ± 0.03	-2.42 ± 0.01	-7.63 ± 0.04
D12	1.30 ± 0.01	1.27 ± 0.02	-2.03 ± 0.01	-2.05 ± 0.02
F13		-5.09 ± 0.02	-0.01 ± 0.01	
E14	-0.31 ± 0.01	-0.49 ± 0.01	0.16 ± 0.01	-0.02 ± 0.02
G15	2.55 ± 0.01	3.88 ± 0.04	-4.67 ± 0.02	-3.34 ± 0.04
V16	-0.29 ± 0.01	-1.67 ± 0.08	1.60 ± 0.02	0.22 ± 0.07
L17	0.56 ± 0.03	-5.42 ± 0.03	2.98 ± 0.03	-2.99 ± 0.03
N18	1.64 ± 0.03	1.47 ± 0.07	2.67 ± 0.04	2.50 ± 0.07
W19	0.24 ± 0.02	-1.26 ± 0.01	1.41 ± 0.01	-0.09 ± 0.03
G20	0.26 ± 0.02	-0.55 ± 0.01	0.27 ± 0.01	-0.55 ± 0.03
S21	3.53 ± 0.02	7.25 ± 0.11	0.74 ± 0.03	4.46 ± 0.10
Y22		6.53 ± 0.03	3.38 ± 0.01	
S23	8.17 ± 0.02	1.29 ± 0.04	1.44 ± 0.02	-5.44 ± 0.04
G24	-0.44 ± 0.02	-6.05 ± 0.02	0.70 ± 0.01	-4.92 ± 0.03
E25	-5.49 ± 0.03	3.35 ± 0.07	0.57 ± 0.04	9.41 ± 0.06
G26	8.55 ± 0.03		4.03 ± 0.01	
A27	3.36 ± 0.01	0.90 ± 0.03	0.66 ± 0.01	-1.80 ± 0.03
K28	8.18 ± 0.02	-0.92 ± 0.02	7.54 ± 0.02	-1.56 ± 0.02
V29		-4.13 ± 0.01	3.34 ± 0.01	
S30		0.91 ± 0.01	-3.44 ± 0.01	
T31				
K32	0.11 ± 0.02	-7.76 ± 0.01	-1.09 ± 0.01	-8.96 ± 0.02
I33	-2.99 ± 0.02	2.57 ± 0.02	-4.98 ± 0.02	0.58 ± 0.03
V34	2.09 ± 0.02	-0.74 ± 0.01	-2.92 ± 0.01	-5.74 ± 0.03
S35	2.56 ± 0.02	-0.54 ± 0.08	-0.41 ± 0.01	-3.51 ± 0.09
G36	1.29 ± 0.01	2.00 ± 0.02	-3.07 ± 0.01	-2.36 ± 0.02
K37	-1.40 ± 0.04	4.78 ± 0.07	-6.43 ± 0.02	-0.25 ± 0.09
T38	2.35 ± 0.03	6.26 ± 0.05	-1.11 ± 0.01	2.80 ± 0.06
G39				
N40	2.26 ± 0.02	-3.80 ± 0.02	-2.31 ± 0.01	-8.37 ± 0.03
G41	2.66 ± 0.01	-2.31 ± 0.02	-0.62 ± 0.01	-5.58 ± 0.02

M42	3.12 ± 0.01	2.16 ± 0.03	-0.29 ± 0.03	-1.25 ± 0.02
E43	1.47 ± 0.02	1.85 ± 0.04	-2.57 ± 0.01	-2.19 ± 0.05
V44	-2.38 ± 0.02	-0.99 ± 0.02	-2.47 ± 0.01	-1.08 ± 0.03
S45	1.19 ± 0.01	1.58 ± 0.02	0.64 ± 0.01	1.02 ± 0.02
Y46	4.44 ± 0.05	4.04 ± 0.02	-0.91 ± 0.01	-1.32 ± 0.06
T47	2.84 ± 0.02	2.35 ± 0.02	-3.22 ± 0.01	-3.71 ± 0.03
G48	12.05 ± 0.06	3.69 ± 0.05	2.53 ± 0.01	-5.84 ± 0.10
T49	3.06 ± 0.02	0.76 ± 0.02	-0.02 ± 0.01	-2.32 ± 0.03
T50	1.63 ± 0.02			
D51	4.70 ± 0.01	0.65 ± 0.06	6.60 ± 0.01	2.55 ± 0.06
G52	2.08 ± 0.02	-0.88 ± 0.02	1.23 ± 0.01	-1.73 ± 0.03
Y53	2.63 ± 0.02	-1.49 ± 0.02	1.78 ± 0.01	-2.35 ± 0.03
W54	5.02 ± 0.03	-0.91 ± 0.05	0.84 ± 0.01	-5.09 ± 0.06
G55	0.13 ± 0.02	-0.90 ± 0.02	-1.92 ± 0.01	-2.96 ± 0.03
T56	1.56 ± 0.02	-0.48 ± 0.01	-1.04 ± 0.01	-3.08 ± 0.02
V57	1.31 ± 0.03	-3.67 ± 0.02	-1.68 ± 0.01	-6.66 ± 0.03
Y58	-0.27 ± 0.03			-5.39 ± 0.05
S59	10.17 ± 0.04	0.82 ± 0.02	4.08 ± 0.01	-5.27 ± 0.05
L60	2.47 ± 0.03	-3.15 ± 0.02	4.51 ± 0.01	-1.11 ± 0.04
P61				
D62	7.29 ± 0.02	-2.61 ± 0.06	7.48 ± 0.01	-2.42 ± 0.08
G63	-9.13 ± 0.01		1.21 ± 0.03	±
D64	-0.92 ± 0.01			-1.06 ± 0.03
W65	1.90 ± 0.03	-0.09 ± 0.02	-4.72 ± 0.01	-6.70 ± 0.04
S66	0.76 ± 0.03	0.93 ± 0.03	-0.13 ± 0.03	0.04 ± 0.04
K67				
W68				
L69				
K70	-1.34 ± 0.04	-2.58 ± 0.05	-4.75 ± 0.02	-5.99 ± 0.08
I71	3.90 ± 0.02	3.34 ± 0.02	-4.61 ± 0.01	-5.17 ± 0.03
S72	-1.32 ± 0.01	1.83 ± 0.06	-2.40 ± 0.01	0.74 ± 0.06
F73	2.33 ± 0.02	1.74 ± 0.05	-3.46 ± 0.01	-4.05 ± 0.06
D74	-0.82 ± 0.02	0.16 ± 0.02	-4.63 ± 0.01	-3.66 ± 0.02
I75	-0.49 ± 0.01	0.59 ± 0.01	-3.77 ± 0.01	-2.69 ± 0.02
K76		9.36 ± 0.03		7.70 ± 0.05
S77		6.57 ± 0.07	-6.62 ± 0.05	
V78		-9.65 ± 0.05		-6.26 ± 0.05
D79	-2.26 ± 0.05	1.83 ± 0.23	-0.44 ± 0.05	3.65 ± 0.23
G80	2.87 ± 0.07		0.52 ± 0.03	
S81	1.48 ± 0.04		5.99 ± 0.06	
A82	2.25 ± 0.04		4.93 ± 0.04	
N83	-0.06 ± 0.02	6.77 ± 0.19	2.66 ± 0.03	9.49 ± 0.18
E84	3.47 ± 0.04	4.93 ± 0.07	-3.14 ± 0.03	-1.68 ± 0.08
I85	-1.62 ± 0.01	5.79 ± 0.03	-2.70 ± 0.01	4.71 ± 0.04
R86		2.12 ± 0.10	-4.47 ± 0.01	

F87	-2.61 ± 0.02	0.67 ± 0.01	-7.46 ± 0.01	-4.19 ± 0.02
M88	4.09 ± 0.03	0.25 ± 0.02	-3.35 ± 0.01	-7.19 ± 0.04
I89	2.61 ± 0.02	1.08 ± 0.02	-2.58 ± 0.01	-4.11 ± 0.03
A90	1.33 ± 0.02	3.38 ± 0.15	-6.17 ± 0.01	-4.12 ± 0.16
E91	2.63 ± 0.04	5.66 ± 0.03	-4.79 ± 0.02	-1.76 ± 0.05
K92		-2.26 ± 0.05		1.98 ± 0.07
S93	-1.28 ± 0.02	-2.97 ± 0.01	1.30 ± 0.01	-0.38 ± 0.02
I94	-0.63 ± 0.04	1.84 ± 0.02	-10.00 ± 0.02	-7.54 ± 0.04
N95	1.74 ± 0.02	-1.56 ± 0.02	-1.60 ± 0.01	-4.91 ± 0.03
G96	-1.67 ± 0.03	-1.37 ± 0.09	3.68 ± 0.04	3.98 ± 0.09
V97	0.26 ± 0.03	0.65 ± 0.07	5.86 ± 0.03	6.25 ± 0.07
G98	0.01 ± 0.04		3.03 ± 0.04	
D99			-2.94 ± 0.01	
G100	-1.80 ± 0.04	-2.46 ± 0.02	-1.24 ± 0.01	-1.90 ± 0.05
E101				
H102		-0.73 ± 0.03		-3.15 ± 0.05
W103	-0.97 ± 0.02	2.43 ± 0.02	-4.24 ± 0.01	-0.83 ± 0.02
V104	-5.84 ± 0.04	2.51 ± 0.03	-2.94 ± 0.02	5.41 ± 0.05
Y105	±	-0.54 ± 0.01		-5.76 ± 0.02
S106	-0.91 ± 0.02	5.52 ± 0.09	1.46 ± 0.02	7.90 ± 0.09
I107	6.67 ± 0.02	3.35 ± 0.02	2.23 ± 0.01	-1.09 ± 0.03
T108	-1.99 ± 0.01			2.72 ± 0.03
P109				
D110	1.43 ± 0.02	5.11 ± 0.06	-0.67 ± 0.01	3.00 ± 0.07
S111	5.99 ± 0.02			2.25 ± 0.06
S112	-2.00 ± 0.02			13.51 ± 0.13
W113	2.95 ± 0.02	3.31 ± 0.08	0.27 ± 0.03	0.63 ± 0.06
K114		-1.02 ± 0.02	2.67 ± 0.01	
T115		1.60 ± 0.05	-3.92 ± 0.02	
I116	0.40 ± 0.02	-1.52 ± 0.01	-2.88 ± 0.01	-4.80 ± 0.03
E117	2.99 ± 0.02	6.68 ± 0.11	2.29 ± 0.01	5.98 ± 0.12
I118		-3.48 ± 0.03	-4.61 ± 0.01	
P119				
F120	10.72 ± 0.05	1.41 ± 0.02	5.57 ± 0.02	-3.75 ± 0.04
S121	0.04 ± 0.01		-3.10 ± 0.01	
S122	-1.21 ± 0.02	-5.08 ± 0.01	-1.52 ± 0.01	-5.39 ± 0.02
F123	10.63 ± 0.02	-1.89 ± 0.02	6.71 ± 0.01	-5.81 ± 0.03
R124	-1.27 ± 0.02	2.80 ± 0.05	-1.80 ± 0.01	2.27 ± 0.06
R125	-3.54 ± 0.02	-0.71 ± 0.04	-4.11 ± 0.01	-1.29 ± 0.04
R126	3.13 ± 0.04	-5.72 ± 0.03	4.94 ± 0.01	-3.92 ± 0.05
L127	4.46 ± 0.03		3.57 ± 0.02	
D128	4.67 ± 0.02	9.92 ± 0.10	-3.70 ± 0.01	1.55 ± 0.11
Y129	5.28 ± 0.02	-0.60 ± 0.01	4.47 ± 0.01	-1.42 ± 0.02
Q130	0.35 ± 0.02	0.07 ± 0.03	1.01 ± 0.02	0.72 ± 0.03
P131				

P132				
G133	1.82 ± 0.02		-0.47 ± 0.01	
Q134	0.79 ± 0.02	-2.21 ± 0.02	-5.21 ± 0.01	-8.21 ± 0.03
D135		-1.41 ± 0.12	-9.16 ± 0.08	
M136	2.06 ± 0.03		-1.46 ± 0.02	
S137	-0.56 ± 0.01	5.97 ± 0.04	-5.66 ± 0.01	0.87 ± 0.04
G138	-2.76 ± 0.02	7.25 ± 0.05	-2.90 ± 0.01	7.11 ± 0.06
T139		2.95 ± 0.13		
L140	5.66 ± 0.02		7.30 ± 0.01	
D141	-3.97 ± 0.02	1.80 ± 0.03	-2.86 ± 0.01	2.92 ± 0.04
L142	-1.24 ± 0.02	-0.04 ± 0.02	-2.03 ± 0.01	-0.83 ± 0.02
D143	3.30 ± 0.01	5.11 ± 0.03	0.47 ± 0.01	2.28 ± 0.03
N144	2.52 ± 0.01	-3.51 ± 0.02	6.38 ± 0.01	0.35 ± 0.03
I145	1.10 ± 0.04	-1.62 ± 0.02	-2.76 ± 0.01	-5.48 ± 0.04
D146		-2.47 ± 0.10	-2.95 ± 0.04	
S147	-1.69 ± 0.02	0.33 ± 0.01	-5.81 ± 0.01	-3.79 ± 0.02
I148	-0.89 ± 0.03	-2.29 ± 0.03	-0.96 ± 0.02	-2.36 ± 0.03
H149	2.37 ± 0.03	0.98 ± 0.02	-1.57 ± 0.01	-2.96 ± 0.04
F150	2.32 ± 0.02	-0.30 ± 0.02	0.42 ± 0.01	-2.20 ± 0.03
M151	-1.09 ± 0.02	-0.47 ± 0.02	-1.97 ± 0.01	-1.34 ± 0.02
Y152	8.03 ± 0.04		2.97 ± 0.02	
A153	-1.41 ± 0.06	3.40 ± 0.05	-5.17 ± 0.04	-0.36 ± 0.07
N154	1.94 ± 0.03	2.17 ± 0.04	-5.27 ± 0.03	-5.04 ± 0.04
N155				
K156	2.20 ± 0.01	-5.26 ± 0.01	1.34 ± 0.01	-6.12 ± 0.02
S157	7.18 ± 0.03			
G158		1.49 ± 0.03	0.38 ± 0.01	
K159	4.70 ± 0.02	2.68 ± 0.05	6.21 ± 0.01	4.20 ± 0.06
F160	5.73 ± 0.02	0.39 ± 0.01	2.03 ± 0.01	-3.31 ± 0.02
V161	-1.05 ± 0.01	0.89 ± 0.01	-3.74 ± 0.01	-1.80 ± 0.02
V162		0.91 ± 0.11	-15.37 ± 0.23	
D163	-2.47 ± 0.02	-6.65 ± 0.03	3.00 ± 0.02	-1.17 ± 0.03
N164	0.70 ± 0.02	-0.39 ± 0.02	-2.97 ± 0.01	-4.06 ± 0.03
I165	2.12 ± 0.01		-3.51 ± 0.01	
K166	0.52 ± 0.02	2.02 ± 0.06	-6.80 ± 0.04	-5.31 ± 0.03
L167		4.66 ± 0.07	-5.31 ± 0.01	
I168	0.51 ± 0.02	6.31 ± 0.07	-8.55 ± 0.01	-2.75 ± 0.08
G169	-1.72 ± 0.04		-1.73 ± 0.02	
A170	0.28 ± 0.02	0.57 ± 0.04	4.39 ± 0.01	4.69 ± 0.04
L171	-0.57 ± 0.04	5.24 ± 1.14	0.76 ± 0.04	6.57 ± 1.14
E172	2.74 ± 0.05			

C.4 Amide proton exchange

Table C.13: Amide proton/deuterium exchange rates and free energy of the structural opening reaction for free and cellobiose-bound CtCBM11 at 25 °C.

	<i>free</i>		<i>bound</i>		k_{rc}^a	Δk_{ex}	$\delta\Delta G_{HX}$
	$k_{ex} (s^{-1})$	ΔG_{HX} (kJ.mol ⁻¹)	$k_{ex} (s^{-1})$	ΔG_{HX} (kJ.mol ⁻¹)			
M1					2.32E+02		
A2	1.14E-03	26.52	2.53E-03	24.54	5.07E+01	1.39E-03	-1.97E+00
S3					4.32E+01		
A4					4.32E+00		
V5					2.92E+01		
G6					9.89E+00		
E7					1.40E+01		
K8					2.79E+01		
M9					7.33E+00		
L10	3.69E-06	35.70	2.84E-05	30.65	6.69E+00	2.47E-05	-5.05E+00
D11	1.53E-05	32.35	1.64E-05	32.18	7.17E+00	1.13E-06	-1.76E-01
D12	3.87E-03	18.98	8.73E-03	16.97	8.23E+00	4.86E-03	-2.01E+00
F13			2.09E-05	31.74	7.68E+00		
E14	1.75E-04	29.73	3.61E-04	27.94	2.85E+01	1.86E-04	-1.79E+00
G15					6.38E+00		
V16					4.12E+00		
L17					4.12E+01		
N18					1.76E+01		
W19	1.32E-02	19.26			3.13E+01		
G20	3.82E-04	30.20	5.90E-04	29.12	7.50E+01	2.08E-04	-1.08E+00
S21					2.32E+01		
Y22	1.85E-05	37.01	6.15E-05	34.04	5.69E+01	4.30E-05	-2.98E+00
S23					8.04E+01		
G24	1.35E-03	22.05	1.05E-03	22.68	9.89E+00	-3.04E-04	6.32E-01
E25					2.85E+01		
G26					3.20E+01		
A27	2.95E-03	21.82	7.54E-03	19.50	1.97E+01	4.58E-03	-2.32E+00
K28	7.10E-04	22.27	1.36E-03	20.66	5.69E+00	6.51E-04	-1.61E+00
V29					3.67E+01		
S30	1.51E-05	36.43			3.67E+01		
T31					3.13E+01		
K32					5.31E+00		
I33					2.54E+00		
V34	1.70E-05	36.14	2.46E-06	40.92	3.67E+01	-1.46E-05	4.79E+00
S35					8.04E+01		

G36	1.82E-03	23.99	4.01E-03	22.03	2.92E+01	2.19E-03	-1.96E+00
K37					2.43E+01		
T38					6.38E+01		
G39					9.89E+01		
N40					8.41E+01		
G41					3.13E+01		
M42					8.62E+00		
E43					3.06E+00		
V44					3.67E+01		
S45					2.32E+01		
Y46			8.98E-07	42.00	2.07E+01		
T47	3.63E-05	35.63	3.32E-05	35.85	6.38E+01	-3.09E-06	2.21E-01
G48	2.35E-03	23.19	2.70E-03	22.83	2.72E+01	3.58E-04	-3.52E-01
T49	6.69E-03	20.76	1.00E-02	19.76	2.92E+01	3.35E-03	-1.01E+00
T50					1.72E+01		
D51			1.06E-03	25.10	2.66E+01		
G52	1.31E-02	17.79	1.62E-02	17.26	1.72E+01	3.12E-03	-5.30E-01
Y53	2.72E-04	25.90	5.14E-04	24.33	9.44E+00	2.42E-04	-1.57E+00
W54	5.86E-05	32.68	4.80E-07	44.58	3.13E+01	-5.81E-05	1.19E+01
G55					2.72E+01		
T56					6.84E+00		
V57					8.41E+00		
Y58	2.06E-06	42.45	3.16E-03	24.28	5.69E+01	3.16E-03	-1.82E+01
S59					1.14E+01		
L60	7.25E-05		1.33E-04			6.07E-05	
P61					6.24E+00		
D62					2.66E+01		
G63					1.60E+01		
D64					5.56E+00		
W65	1.41E-03	25.37	2.90E-03	23.58	3.94E+01	1.49E-03	-1.79E+00
S66	1.41E-03	25.37	2.24E-03	24.22	3.94E+01	8.27E-04	-1.14E+00
K67					1.11E+01		
W68	7.51E-06	32.92	1.74E-05	30.83	4.42E+00	9.94E-06	-2.09E+00
L69					1.22E+01		
K70					5.31E+00		
I71					2.99E+01		
S72					2.48E+01		
F73					1.25E+01		
D74					2.66E+00		
I75					1.16E+01		
K76					6.68E+01		
S77	3.51E-05	30.75	8.24E-05	28.63	8.61E+00	4.74E-05	-2.12E+00
V78					7.86E+00		
D79	1.26E-04	30.37			2.66E+01		

G80					7.50E+01		
S81					4.32E+01		
A82					6.68E+01		
N83			6.37E-05	30.48	1.40E+01		
E84					2.85E+00		
I85					1.53E+01		
R86					2.07E+01		
F87					2.43E+01		
M88					5.19E+00		
I89			3.33E-05	31.84	1.27E+01		
A90					6.69E+00		
E91	8.10E-05	29.88	3.63E-04	26.17	1.40E+01	2.82E-04	-3.71E+00
K92	1.47E-05	37.97	1.26E-05	38.37	6.68E+01	-2.17E-06	3.95E-01
S93					8.04E+00		
I94					3.94E+01		
N95					8.41E+01		
G96					6.38E+00		
V97	1.10E-03	25.23	2.30E-03	23.41	2.92E+01	1.20E-03	-1.83E+00
G98					1.60E+01		
D99					2.66E+01		
G100	2.79E-03	20.25	1.03E-04	28.42	9.89E+00	-2.69E-03	8.17E+00
E101	3.30E-04	26.40			1.40E+01		
H102	3.12E-04	26.34	1.46E-06	39.63	1.29E+01	-3.11E-04	1.33E+01
W103			1.79E-05	30.08	3.35E+00		
V104					8.41E+00		
Y105					5.69E+01		
S106					8.04E+00		
I107					1.08E+01		
T108							
P109					6.24E+00		
D110					3.35E+01		
S111					1.01E+02		
S112					1.68E+01		
W113					1.53E+01		
K114	2.28E-05	34.39	2.84E-05	33.85	2.43E+01	5.57E-06	-5.41E-01
T115			1.71E-05	31.79	6.38E+00		
I116					3.94E+00		
E117	4.73E-04	21.57	1.39E-03	18.89	2.85E+00	9.19E-04	-2.68E+00
I118							
P119					7.16E+00		
F120	3.05E-04	30.13	4.71E-04	29.05	5.82E+01	1.66E-04	-1.08E+00
S121					1.01E+02		
S122	1.07E-03	24.91	1.21E-03	24.60	2.48E+01	1.40E-04	-3.06E-01
F123	1.07E-03	25.37	4.84E-04	27.33	2.99E+01	-5.82E-04	1.96E+00

R124	3.42E-04	29.11	2.28E-05	35.82	4.32E+01	-3.19E-04	6.71E+00
R125					4.32E+01		
R126	4.17E-04	24.85	7.04E-06	34.96	9.44E+00	-4.10E-04	1.01E+01
L127					6.69E+00		
D128					7.68E+00		
Y129	1.70E-02	18.34	2.51E-03	23.08	2.79E+01	-1.45E-02	4.74E+00
Q130			1.06E-02				
P131							
P132					2.32E+01		
G133					3.67E+01		
Q134					1.72E+01		
D135					1.40E+01		
M136					6.53E+01		
S137					8.04E+01		
G138					2.72E+01		
T139					9.02E+00		
L140					6.69E+00		
D141	5.74E-04	21.77	5.73E-05	27.48	3.76E+00	-5.16E-04	5.71E+00
L142					6.69E+00		
D143	4.99E-03	22.52			4.42E+01		
N144	5.30E-04	23.96	6.90E-04	23.31	8.41E+00	1.60E-04	-6.54E-01
I145	9.42E-04	21.86	1.35E-03	20.96	6.39E+00	4.10E-04	-8.95E-01
D146	1.56E-07	47.54	2.68E-04	29.08	3.35E+01	2.67E-04	-1.85E+01
S147					8.04E+00		
I148					1.16E+01		
H149					1.91E+01		
F150					2.43E+01		
M151					1.50E+01		
Y152					2.43E+01		
A153	1.82E-03	26.05	6.25E-03	22.99	6.68E+01	4.43E-03	-3.06E+00
N154	1.43E-02	22.76			1.40E+02		
N155	7.81E-04	26.94			4.12E+01		
K156	3.80E-03	24.22	6.99E-03	22.71	6.68E+01	3.19E-03	-1.51E+00
S157					8.04E+01		
G158					2.92E+01		
K159					1.64E+01		
F160					4.96E+00		
V161					3.13E+00		
V162					7.86E+00		
D163					4.42E+01		
N164	2.71E-05	31.33	1.28E-05	33.19	8.41E+00	-1.43E-05	1.86E+00
I165					1.16E+01		
K166					7.50E+00		
L167					2.48E+00		

I168	2.37E+01
G169	3.20E+01
A170	5.69E+00
L171	4.12E+00
E172	1.40E+01

^a The hydrogen-exchange rates of amide protons in non-structured peptides, k_{rc} , were estimated using the software SPHERE² (<http://www.fccc.edu/research/labs/roder/sphere>) with the default activation energies (E_a s): Acid E_{aH} : 15.0 kcal/mol, Base E_{aOH} : 2.6 kcal/mol. The exchange media, temperature and pH were set to D₂O, 25 °C and 7.5, respectively. The reference data was set to poly-DL-alanine.³ The remaining parameters were kept with the defaults values.

C.5 References

1. Dosset, P.; Hus, J. C.; Blackledge, M.; Marion, D., Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J Biomol Nmr* **2000**, *16* (1), 23.
2. Zhang, Y.-Z. Protein and peptide structure and interactions studied by hydrogen exchange and NMR. Ph.D. Thesis, University of Pennsylvania, Philadelphia, 1995.
3. Bai, Y. W.; Milne, J. S.; Mayne, L.; Englander, S. W., Protein Stability Parameters Measured by Hydrogen-Exchange. *Proteins* **1994**, *20* (1), 4.