# Molecular determinants of ligand specificity in family 11 carbohydrate binding modules – an NMR, X-ray crystallography and computational chemistry approach

Aldino Viegas[1,*], Natércia F. Brás[2,*], Nuno M. F. S. A. Cerqueira[2,*], Pedro Alexandrino Fernandes[2], José A. M. Prates[3], Carlos M. G. A. Fontes[3], Marta Bruix[4], Maria João Romão[1], Ana Luísa Carvalho[1], Maria João Ramos[2], Anjos L. Macedo[1] and Eurico J. Cabrita[1]

1 REQUIMTE–CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
2 REQUIMTE, Departamento de Química, Faculdade de Ciências do Porto, Portugal
3 Centro Interdisciplinar de Investigação em Sanidade Animal, Faculdade de Medicina Veterinária, Lisbon, Portugal
4 Instituto de Química Física Rocasolano, CSIC, Madrid, Spain

**Keywords**
cellulosome; *Clostridium thermocellum*; *Ct*CBM11; STD-NMR molecular modelling; X-ray crystallography

**Correspondence**
E. J. Cabrita, REQUIMTE-CQFB,
Departamento de Química, Faculdade de
Ciências e Tecnologia, Universidade Nova
de Lisboa, 2829-516 Caparica, Portugal
Fax: +351 212948550
Tel: +351 212948358
E-mail: ejc@dq.fct.unl.pt
M. J. Ramos, REQUIMTE, Departamento de
Química, Faculdade de Ciências do Porto,
4169-007 Porto, Portugal
Fax: +351 226082959
Tel: +351 226082806
E-mail: mjramos@fc.up.pt
A. L. Carvalho, REQUIMTE-CQFB,
Departamento de Química, Faculdade de
Ciências e Tecnologia, Universidade Nova
de Lisboa, 2829-516 Caparica, Portugal
Fax: +351 212948550
Tel: +351 212948300
E-mail: alcarvalho@dq.fct.unl.pt

*These authors contributed equally to this work

(Received 7 February 2008, revised 7 March 2008, accepted 13 March 2008)

doi:10.1111/j.1742-4658.2008.06401.x

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* (*Ct*). These organisms produce extracellular multi-subunit complexes (i.e. cellulosomes) comprising a consortium of enzymes, which contain noncatalytic carbohydrate-binding modules (CBM) that increase the activity of the catalytic module. In the present study, we describe a combined approach by X-ray crystallography, NMR and computational chemistry that aimed to gain further insight into the binding mode of different carbohydrates (cellobiose, cellotetraose and cellohexaose) to the binding pocket of the family 11 CBM. The crystal structure of *C. thermocellum* CBM11 has been resolved to 1.98 Å in the apo form. Since the structure with a bound substrate could not be obtained, computational studies with cellobiose, cellotetraose and cellohexaose were carried out to determine the molecular recognition of glucose polymers by *Ct*CBM11. These studies revealed a specificity area at the *Ct*CBM11 binding cleft, which is lined with several aspartate residues. In addition, a cluster of aromatic residues was found to be important for guiding and packing of the polysaccharide. The binding cleft of *Ct*CBM11 interacts more strongly with the central glucose units of cellotetraose and cellohexaose, mainly through interactions with the sugar units at positions 2 and 6. This model of binding is supported by saturation transfer difference NMR experiments and linebroadening NMR studies.

**Abbreviations**
AMBER, assisted model building and energy refinement; CBM, carbohydrate-binding modules; *Ct*, *Clostridium thermocellum*; STD, saturation transfer difference.

The enzymatic degradation of insoluble polysaccharides and of cellulose, in particular, is one of the most important reactions on earth. This subject is currently under intense research because glucose derivatives can be obtained from degradation of polysaccharides. After fermentation processes, compounds such as glucose derivatives [1,2], acetone, alcohols and volatile fatty acids [3,4] can be obtained that are essential for biotech and pharmaceutical industries. Furthermore, the biofuel industry has a great interest in this field because ethanol can also be directly obtained from glucose monomers [2].

Efficient methods for degrading cellulose chains have been intensively investigated worldwide within the last decade. The degradation of plant cell wall polysaccharides into soluble sugars has been found to be possible either by chemical means or by certain microorganisms. The latter method has become the most attractive due to reasons of economy and efficiency [2].

However, the enzymatic degradation of this type of polysaccharide was shown to be relatively inefficient in most cases because their targets (i.e. the glycosidic bonds) are often inaccessible to the active site of the appropriate enzymes [5]. Even so, it was found that some microorganisms (e.g. *Clostridium thermocellum*) have evolved and improved their catalytic capabilities. These organisms have a consortium of enzymes associated together in high molecular weight cellulolytic multi-subunit complexes, normally called cellulosomes, which exist at the extracellular level [6]. The enzymes are generally modular proteins that contain noncatalytic carbohydrate-binding modules (CBM), which increase the activity of the catalytic module [7–9].

The catalytic mechanisms of the enzymes present in the cellulosome are well understood [2], but the function and behaviour of the noncatalytic modules have not yet been fully elucidated. It has been proposed that the latter may play different roles in the cellulosome consortium, including promotion of the association of the enzyme with the substrate and guiding the substrate to the catalytic site of the enzyme. Moreover, it is believed that it serves as an ‘anchor’ that promotes an increase in the concentration of the enzyme on the surface of the substrate polymers, leading to a faster degradation of the polysaccharide [5,8].

Generally, CBMs can be grouped into several families taking into account ligand specificity (http://afmb.cnrs-mrs.fr/CAZY), the conservation of the protein fold, and based on structural and functional similarities. In this last case, the protein modules have been grouped into three subfamilies: ‘surface-binding’ CBMs (type A), ‘glycan-chain-binding’ CBMs (type B), and ‘small sugar-binding’ CBMs (type C) [5].

The focus of the present study is on the noncatalytic modules present in *C. thermocellum*. In this organism, bifunctional cellulosomes are found that contain two catalytic modules (GH5 and GH26), each one with a family 11 CBM (*Ct*CBM11). This *Ct*CBM11 is part of the type B subfamily and is characterized by the binding of a single polysaccharide chain [10]. It has been observed that this type of CBM can bind to a diversity of ligands and its specificity depends mostly on the aromatic residues present in the binding cleft. Direct hydrogen bonds also play a key role in defining the affinity and ligand specificity of type B glycan chain binders [5,8,11–13].

Additionally, it has been shown that the specificity of *Ct*CBM11 is consistent with the type of substrates that are hydrolyzed by the associated catalytic domains [14].

To increase the current knowledge of the molecular interactions that define the ligand specificity in cellulosomal CBMs and the mechanism by which they recognize and select their substrates, we used X-ray crystallography, NMR and computational chemistry approaches to identify the molecular determinants of ligand specificity of *Ct*CBM11. By means of NMR studies, we have analyzed various cello-oligosaccharides of different sizes. This approach enabled us to identify a range of cello-oligosaccharides with an affinity for the binding cleft. This information was complemented with docking and molecular mechanics studies that allowed localized structural information to be obtained on the pocket site of *Ct*CBM11 and, in particular, the identification of the atoms of the ligand that are closer to the protein when the complex is formed. The ligands cellobiose, cellotetraose and cellohexaose were studied.

## Results and Discussion

### The crystal structure of *Ct*CBM11, the binding cleft and its ligand specificity

In a previous study [14], isothermal titration calorimetry of wild-type *Ct*CBM11 with oligosaccharides and polysaccharides was used to analyse and determine the binding affinities of *Ct*CBM11 for substrates such as lichenan, β-glucan, cellohexaose, cellotetraose, cellopentaose and G4G4G3G. *Ct*CBM11 exhibits a preference for β-1,3-1,4 glucans and a considerable affinity for *β*-1,4 linked glucose polymers. No affinity for β-1,3 glucans was observed. The same study also described the affinity gel electrophoresis results obtained from binding of wild-type *Ct*CBM11 and its mutant derivatives [14]. Tyrosines 22, 53 and 129 appear to play a central role in carbohydrate recognition.

The 3D structure of *Ct*CBM11 has been resolved to 1.98 Å resolution and is deposited in the protein databank under the accession code 1v0a. Its 3D structure has been fully characterized and a complete description of its fold has been performed, including a compilation of the residues that compose the binding cleft [14]. It folds as a β-jelly roll [8] of two six-stranded anti-parallel β-sheets that form a convex side (β-strands 1, 3, 4, 6, 9 and 12) and a concave side (β-strands 2, 5, 7, 8, 10 and 11). The concave side is decorated by the side chains of several residues, with a probable substrate recognition role. Most relevant is the presence of four tyrosine residues (numbers 22, 53, 129 and 152), as well as four aspartate, two arginine and two histidine residues. The cleft is also decorated by the side chains of three serine and two methionine residues. Due to symmetry constraints, the reported structure of 1v0a exhibits a binding cleft occupied by the C-terminus residues (an engineered six-histidine tail) of a symmetry-related molecule. The structure details of 1v0a suggest that residues Ser59, Asp99, Tyr53, Arg126, Tyr129 and Tyr152 might be involved in the binding mechanisms of possible ligands. However, the presence of the His-tag residues appears to have impaired crystal soaking and co-crystallization experiments with candidate ligands.

The hypothesis that the histidine tail was preventing ligand binding led us to design a new protein production strategy that would allow *Ct*CBM11 to be obtained with an unoccupied binding cleft. The crystallization conditions of the newly purified protein are different from those of the tagged one (data not shown), and the new crystals belong to a different space group. The deposited structure of 1v0a belongs to the $P2_12_12$ space group whereas, in the absence of the six-histidine tail, *Ct*CBM11 crystals grow in the $P2_1$ space group. However, crystal soaking and co-crystallization of *Ct*CBM11 with candidate ligands was unsuccessful. Nevertheless, the engineered six-histidine tag appears to be important for crystallization because the crystals, in the absence of these extra residues, are comparatively more fragile and exhibit a lower diffraction quality (data not shown).

Confronted with these negative results from the crystallographic approach, complementary experiments by NMR and computational calculations were considered.

## NMR interaction studies

Different information may be deduced for protein–carbohydrate complexes in solution by NMR spectroscopy. In the present study, we focused our attention on those methods that allow us to obtain information on the bound carbohydrate.

The identification and mapping of the ligand epitopes (i.e. atoms of the ligand that are closer to the protein when the complex is formed) was performed using the saturation transfer difference (STD)-NMR technique [15,16]. The interaction between cellohexaose and *Ct*CBM11 was used as a model to study the interaction between the soluble protein and cellulose because cellohexaose is the longest readily available cello-oligosaccharide that can be used to mimic the glucose chain of cellulose [17]. Line broadening effects on cellohexaose resonances upon addition of increasing amounts of *Ct*CBM11 were also explored as an aid to identify those sugar resonances that are more affected upon binding to the protein.

### Line broadening studies

The simple measure or estimation of linewidths may serve as a basis to deduce the occurrence of binding or recognition (a dynamic process). Because the relaxation properties of the oligosaccharides are affected upon protein binding due to their dependence on molecular motion, we studied the linebroadening effects (related to $T_2$ relaxation) of cellohexaose resonances upon addition of *Ct*CBM11.

In general, a progressive line broadening of all the cellohexaose protons was observed during titration with increasing amounts of protein, which can be understood as a result of the loss of local mobility caused by 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 are identified in Fig. 1I.

A detailed comparison of the cellohexaose spectra showed that the most significant linebroadening was observed for protons 6 and 2, from glucose units b to e (Fig. 1III–V), which could indicate that the corresponding hydroxyl groups are involved in protein binding.

The results for the linebroadening measurements of protons H1a in the alpha and beta configurations, αHa1 and βHa1 (Fig. 1II,V), showed that these protons are almost unaffected by protein binding, as would be expected for protons on the terminal end of the sugar located out of the binding cavity. However, a slight effect can be detected for βHa1 compared to αHa1, which may indicate a higher affinity of the protein for the β form.

### STD-NMR

To understand how *Ct*CBM11 distinguishes and selects the different ligands, it is extremely important to

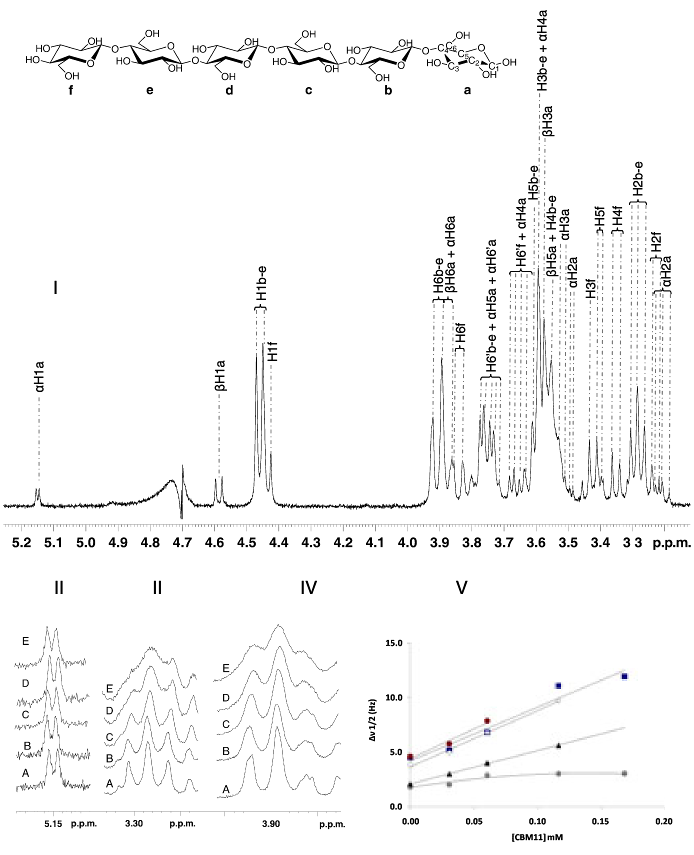

**Fig. 1.** Line broadening studies. (I) Spectral assignment of $^{1}$H NMR cellohexaose resonances. (II–IV) Series of spectral regions of a solution of cellohexaose 0.787 mM in $D_2O$, corresponding to protons αa1, 6 and 2, respectively, acquired at 298 K as a function of peptide (*Ct*CBM11) concentration (A, 0.0 mM; B, 0.031 mM; C, 0.060 mM; D, 0.116 mM; E, 0.168 mM). V, Linewidths ($\Delta\upsilon_{1/2}$) of selected cellohexaose protons, determined after spectral deconvolution, as a function of peptide (*Ct*CBM11) concentration: ●, αH1a; ▲, βH1a; ◇, H2b-e; ●, H6′b-e, βH6′a, αH5a; ■, H6b-e, βH6a, αH6a.

identify which atoms of the ligand are closer to the protein when the complex is formed (epitope mapping). Identification and mapping of the epitopes can be achieved using the STD-NMR technique. The ability of the STD-NMR technique to detect the binding of low molecular weight compounds to large biomolecules has been demonstrated previously [16,18–20]. This technique offers several advantages over other methods in detecting binding activity. First, 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 approximately $10^{-3}$ to $10^{-8}$ M. Second, the building block of the ligand having the strongest contact with the protein shows the most intense NMR signals, enabling mapping of the ligand's binding epitope. Finally, and most importantly for a NMR-based detection system, its high sensitivity allows the use of as little as 1 nmol of protein with a molecular mass > 10 kDa [16,18,21].

STD-NMR spectroscopy was used to analyze the binding of cellohexaose to *Ct*CBM11. The STD-NMR spectrum of the hexasaccharide in a 20-fold excess over *Ct*CBM11 is shown in Fig. 2 along with the cellohexaose reference spectrum. Comparison of both spectra clearly shows that the residues of the hexasaccharide are involved in the binding in different ways. From Fig. 2, it can be seen that the more intense signals are those corresponding to H2 and H6 from glucose units b to e, indicating that, when the complex is formed, these protons are those that are closer to the 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 the precise orientation of the methylene groups upon binding to the protein.

No STD signals could be detected for protons αH1a and βH1a, the anomeric protons of the reducing end of the oligosaccharide.

In the region between 3.63 and 3.52 p.p.m., despite of the presence of STD signals, the individual contributions of protons αH4a, βH3a, H4b-e and H5b-e to the binding cannot be determined due to signal overlap. Nevertheless, information concerning the relative binding contribution can be obtained by comparing the intensity of the signals in this region with that of protons H2 and H6. By comparison of the STD

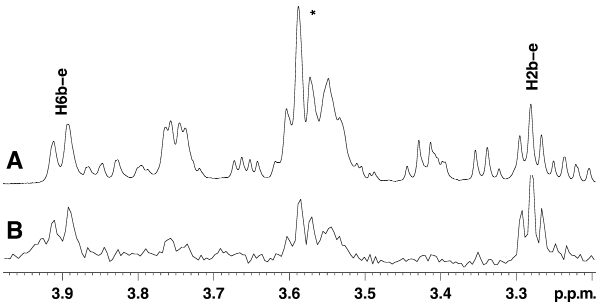

**Fig. 2.** STD-NMR of cellohexaose with *Ct*CBM11. (A) Reference $^1$H NMR cellohexaose spectrum. (B) STD spectra of the solution of cellohexaose (50 μM) with the protein (5 μM). Protons H6b-e and H2b-e show the more intense signals, indicating that these are the ones closer to the protein upon binding. In the region between 3.63 and 3.52 p.p.m. (*), the signal overlap does not allow determination of the individual contributions of protons αH4a, βH3a, H4b-e and H5b-e to the binding.

intensity relative to the reference, a binding epitope map can be created. This is described by the STD factor ($A_{\mathrm{STD}}$):

$$A_{\mathrm{STD}} = (I_0 - I_{\mathrm{sat}})/I_0 \times \text{ligand excess} \qquad (1)$$

The STD epitope map of cellohexaose binding to *Ct*CBM11 (Fig. 3) was obtained by normalizing the largest value to 100%.

From these data, it is clear that, regardless of the large number of protons in the region between 3.63 and 3.52 p.p.m. (16 protons), the relative intensity of their signal in the STD is smaller than that from protons H2 (four protons) and H6 (six protons). In this way, we can clearly distinguish between those protons very close to the protein (protons H2 and H6 from subunits b to e) and those other protons that, in spite of having a STD signal, are more distant from the protein.

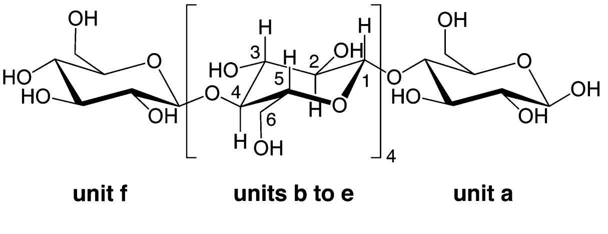

**Fig. 3.** Structure of cellohexaose. Relative degrees of saturation of the individual protons normalized to that of the proton H2b-e: H2b-e, 100%; H6b-e, 48.4% and 36.6% (two non-equivalent protons), determined from 1D STD NMR spectra at a 20-fold ligand excess. The concentrations of *Ct*CBM11 and cellohexaose were 18 μM and 364 μM, respectively.

Subunits a and f should not contribute significantly to the binding because the signals of its protons do not appear in the STD spectrum, meaning that their protons are more distant from the protein.

STD-NMR spectroscopy experiments were also performed with cellobiose and cellotetraose. With cellobiose, no STD signals could be detected, which is in accordance with a previous report demonstrating a weak binding of cellobiose to *Ct*CBM11 [14] in the limits of STD detection. The STD results obtained for cellotetraose are very similar to those obtained for cellohexaose. Again, not all protons give a STD signal and the maximum intensity is found for protons H2 and H6 of the central glucose units and α-H1 of the reducing end.

These results indicate that the binding cleft of *Ct*CBM11 interacts more strongly with the central glucose units, mainly through interactions with positions 2 and 6 of the sugar units, which is consistent with previous studies [14] and with the ligands accommodated by other type B CBMs. The fact that only one of the methylene protons at position 6 gives a STD signal, together with the presence of a STD signal from the anomeric proton, suggests a very well defined geometry upon binding.

## Computational studies

As the X-ray structure of *Ct*CBM11 with a bound substrate is not available, it is difficult to evaluate the importance and function of each residue at the *Ct*CBM11 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 can provide localized structural information about the binding pocket of *Ct*CBM11 and identify which atoms of the ligand and of *Ct*CBM11 interact preferentially. Calculations were performed with cellobiose, cellotetraose and cellohexaose carbohydrates. Moreover, for each ligand, the α and β isomers were considered.

Initial attempts to simulate the interaction between the carbohydrates and the *Ct*CBM11 cleft resorted to standard docking methodologies. The ligands were built independently and the structure was optimized using the assisted model building and energy refinement (AMBER) force field.

The results obtained from these simulations were, however, disappointing because the conformations of some residues near the binding pocket (i.e. Tyr22, Tyr53, Tyr129 and Tyr152) give rise to a steric obstacle, and precluded the efficient binding of the ligands. The importance of these residues in the binding process had already been noted in several previous studies [13,14], and confirms our own observations. To overcome this cornerstone issue, we used MADAMM software [22] that allows the introduction of 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 selectively binds preferentially to one of these conformers displacing the equilibrium towards this particular conformer and, in this way, increasing its proportion relatively to the total protein population. In the present study, the flexibilization was applied to Tyr22, Tyr53, Tyr129 and Tyr152. At the end of this process, a group of complexes is obtained, with optimized affinities between *Ct*CBM11 and each studied ligand.

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 *Ct*CBM11 cavity, near the aromatic amino acids that were flexibilized. Furthermore, the ligands bind in an equidistant mode at the *Ct*CBM11 cleft, which suggests an apparent symmetry at the binding cavity. Most of the interaction between the *Ct*CBM11 cleft and each carbohydrate occurs 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, which shows no specificity, and different binding poses at the *Ct*CBM11 cleft could be observed (Fig. 4). This is in agreement with the experimental work, where no specific interaction could be detected with this ligand.

The orientation of the $CH_2OH$ groups in all docked solutions did not change significantly, and they commonly appeared in alternate positions in the carbohydrate oligomers chain (above and below the plain of the sugar rings) even if the initial calculations were performed on a conformation in which all these groups were on the same plane.

The docking results obtained with MADAMM also revealed that there is no substantial differences between the α or β conformations of carbohydrates. However, we found that, in some carbohydrates, the C1-terminal of the α conformation is turned towards the left hand side of the binding cavity, whereas the β conformation is in the opposite direction. Considering that the monomers constituting the ligands are equal among themselves, this change in orientation is of no great importance for the establishment of the binding interactions between the ligand and *Ct*CBM11, and this kind of behaviour should occur commonly in nature.

From the studied carbohydrates, cellotetraose was the one that fitted perfectly inside the binding cleft of *Ct*CBM11. In the case of β-cellotetraose, the hydrogen bonds were established with the amino acids Glu25, Asp99, Arg126, Asp128, Asp146 and Ser147 (Fig. 5), which closely match the amino acids that interact with the α isomer, differing only in the 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.

Table 1 summarizes the most important interactions that occur between all the analyzed carbohydrate ligands, including the α and β isomers, and the neighbouring amino acids of the *Ct*CMB11 cleft. These average values were obtained after 2 ns of molecular dynamics simulations, with the best solution obtained with MADAMM as reference.

Comparing all the simulated complexes, it is clear that there is a common binding site at the *Ct*CBM11

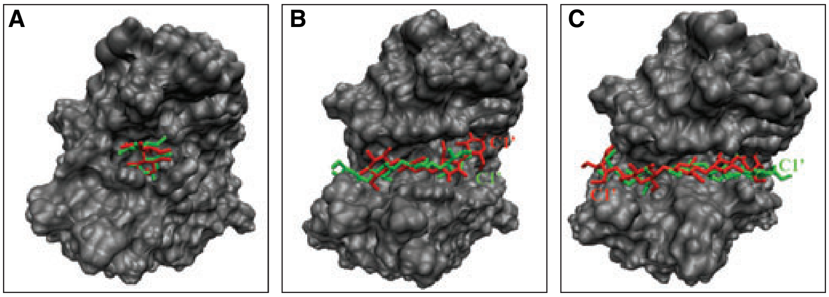

**Fig. 4.** Representation of the conformations of the 3D structure of binding of the different ligands obtained by docking. (A) α- (red) and β-cellobiose (green); (B) α- (red) and β-cellotetraose (green); (C) α- (red) and β-cellotetraose (green). The picture was constructed using the programme VMD 1.8.3. [26].

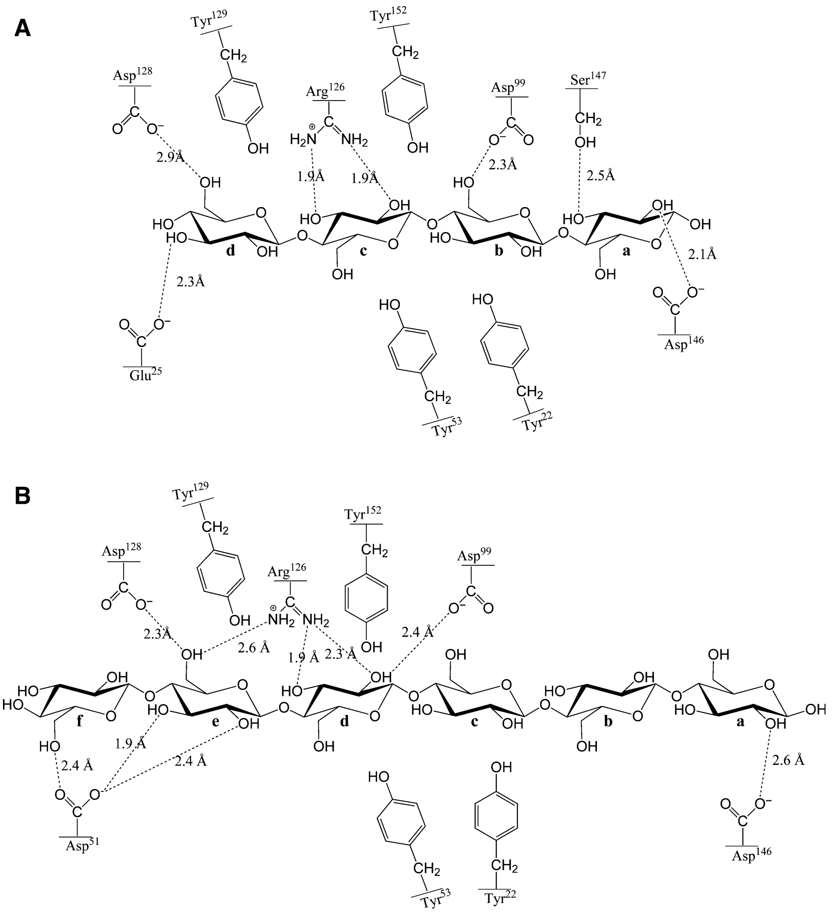

**Fig. 5.** (A,B) Representation of the most important interactions between the β-cellotetraose and β-cellohexaose with the *Ct*CBM11 binding cleft. The distances correspond to the average of the last 2 ns of the molecular dynamics simulations (for further details, see Table 1).

cleft and that all the studied polysaccharides make several hydrogen bonds with the Asp99, Arg126, Asp128 and Asp146 amino acids and, in the case of the larger ligands, with Asp51 as well. Most of the hydrogen bonds occur via the hydroxyl groups associated with the C2 and C6 carbon atoms of each glucose ring, which is in agreement with the results obtained experimentally by NMR.

We also found that the central glucose units interact closely with several tyrosine residues. The function of these residues appears to be more related to the guiding and packing of the carbohydrate ligands at the *Ct*CBM11 cleft, leading to the overall conformation of the bound carbohydrate chain. The same type of interaction also appears to control the overall carbohydrate conformation in the X-ray structures of CBM4 and CBM17 complexed with cellopentaose and cellohexaose, respectively [13,23]. The involvement of the tyrosine residues in the stabilization of the complex cannot be excluded because recent theoretical work, as well as NMR, has demonstrated the existence of an important dispersive component between the hydrogens of the sugar and the aromatic ring of the tyrosine residues, which gives rise to three so-called nonconventional hydrogen bonds that help stabilize the complex [24,25]. The initial conformations adopted by these residues were responsible for the unsatisfactory results of the initial docking trials, and only after exploring the configurational space of these residues, through a multi-stage docking with an automated molecular modelling protocol (MADAMM software), were more reliable results obtained that are 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. Considering these observations, we hypothesize 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 also compared the computational results with another type B CBM that was crystallized in complex with a pentasaccharide (Fig. 6).

**Table 1.** Summary of the distances involved in the main interactions between the carbohydrates and the neighbouring amino acids of the CBM cleft.

| Residue | α-Cellotetraose interaction d(Å) | | β-Cellotetraose interaction d(Å) | | α-Cellohexaose interaction d(Å) | | β-Cellohexaose interaction d(Å) | |
|---|---|---|---|---|---|---|---|---|
| Glu25 | | | COO$^-$↔OH (C3) Glc d | 2.2 | | | | |
| | | | COO$^-$↔OH (C2) Glc d | 2.3 | | | | |
| Asp51 | | | | | COO$^-$↔OH (C3) Glc b | 1.9 | COO$^-$↔OH (C2) Glc e | 2.4 |
| | | | | | | | COO$^-$↔OH (C3) Glc e | 1.9 |
| | | | | | | | COO$^-$↔OH (C6) Glc f | 2.4 |
| Asp99 | COO$^-$↔OH (C6) Glc b | 3.0 | COO$^-$↔OH (C6) Glc b | 2.3 | COO$^-$↔OH (C6) Glc e | 2.3 | COO$^-$↔OH (C2) Glc d | 2.4 |
| | COO$^-$↔H (C3) Glc a | 2.3 | | | | | | |
| | COO$^-$↔OH (C3) Glc a | 2.2 | | | | | | |
| Arg126 | NH$_2$↔OH (C2) Glc c | 1.9 | NH$_2$↔OH (C2) Glc c | 1.9 | NH$_2$↔H (C2) Glc d | 3.0 | NH$_2$↔OH (C2) Glc d | 2.3 |
| | NH$_2$↔OH (C3) Glc c | 2.0 | NH$_2$↔OH (C3) Glc c | 1.9 | | | NH$_2$↔OH (C3) Glc d | 1.9 |
| | NH$_2$↔OH (C6) Glc d | 2.8 | | | | | NH$_2$↔OH (C6) Glc e | 2.6 |
| Asp128 | COO$^-$↔OH (C6) Glc d | 1.9 | COO$^-$↔OH (C6) Glc d | 2.9 | COO$^-$↔H (C1) Glc c | 2.9 | COO$^-$↔OH (C6) Glc e | 2.3 |
| | | | | | COO$^-$↔H (C5) Glc c | 2.9 | | |
| Asp146 | COO$^-$↔OH (C1) Glc a | 2.7 | COO$^-$↔OH (C3) Glc a | 2.7 | COO$^-$↔OH (C2) Glc f | 2.4 | COO$^-$↔OH (C2) Glc a | 2.6 |
| | COO$^-$↔OH (C2) Glc a | 2.5 | COO$^-$↔OH (2) Glc a | 2.1 | COO$^-$↔OH (C3) Glc f | 2.1 | | |
| Ser147 | OH↔OH (C2) Glc a | 2.3 | OH↔OH (C3) Glc a | 2.5 | | | | |
| | NH↔OH (C3) Glc a | 2.7 | | | | | | |
| Tyr22 | | | | | Arom ring↔Glc b | 4.9 | Arom ring↔Glc c | 4.6 |
| Tyr53 | Arom ring↔Glc b | 4.5 | Arom ring↔Glc d | 3.9 | Arom ring↔Glc c | 3.7 | Arom ring↔Glc d | 6.5 |
| Tyr129 | Arom ring↔Glc c | 4.6 | Arom ring↔Glc c | 4.5 | Arom ring↔Glc c | 4.4 | Arom ring↔Glc e | 4.1 |
| Tyr152 | Arom ring↔Glc d | 3.6 | Arom ring↔Glc d | 5.8 | Arom ring↔Glc e | 6.1 | Arom ring↔Glc e | 4.4 |

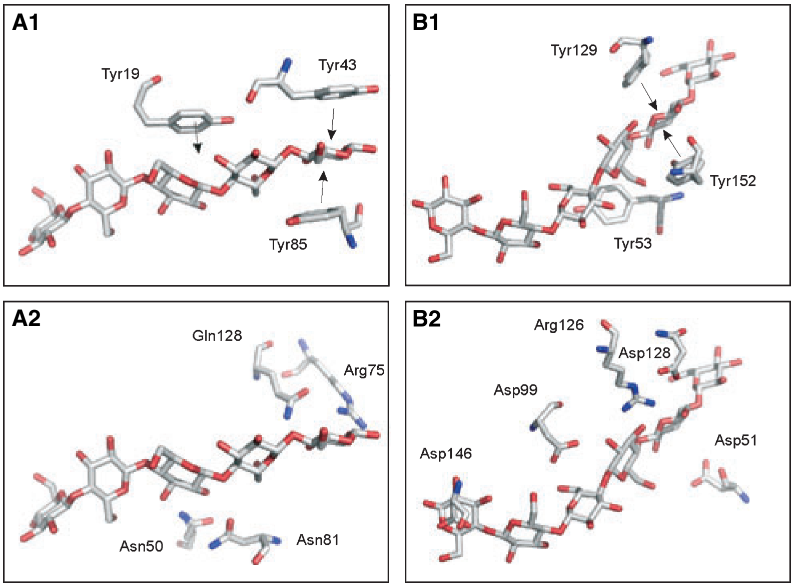

**Fig. 6.** Schematic representation of the main interaction between (A) the pentasaccharide with the *Cf*CBM4 (protein databank entry: 1GU3) [23] and (B) the hexasaccharide with *Ct*CBM11. Interactions involving neighbouring tyrosine residues are shown in (A1) and (B1). Residues that establish several hydrogen bonds with the equatorial hydroxyl groups of the glucose units are shown in (A2) and (B2).

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 *Ct*CBM11. Regardless of the CBM, generally, we have found that 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 *Ct*CBM11, 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 we found in *Cf*CBM4 (Fig. 6). The interaction leads to a slight alteration of the normal chain dihedral angles of the fifth glucose ring that is reflected on the overall conformation of the bounded oligosaccharide. We propose that this common CH-π stacking is responsible for the reorientation of the carbohydrate chain and directing it to the regions that are populated with aspartate residues. Accordingly, we propose that these residues have a preponderant role in the reorientation of the carbohydrate chain.

## Conclusions

X-ray crystallography, NMR and computational chemistry have been shown to comprise complementary methodologies. These techniques were combined to derive structural information on the binding interaction of cello-oligosaccharides and *Ct*CBM11 at the molecular and atomic levels because it is still unclear whether polysaccharides adopt their normal conformation when bound to CBMs or whether these proteins cause a change in the structure of the sugar chain upon binding.

In the present study, it was not possible to use cello-oligosaccharides longer than cellohexaose due to their limited solubility in aqueous buffers [17]. To overcome this limitation, we used cellobiose, cellotetraose and cellohexaose as model compounds.

Both the theoretical and experimental results suggest that all ligands interact mainly by hydrogen bonds, with a central area of *Ct*CBM11 containing the amino acids Asp99, Arg126, Asp128 and Asp146 and, in the case of the larger ligands, with Asp51. It is important to emphasize that most of the hydrogen bonds occur via the hydroxyl groups associated with the C2 and C6 carbon atoms of each ring of glucose. This model of binding is supported by the STD and linebroadening NMR studies performed with cellohexaose, which have shown that the protons of the central glucose units are closer to the protein than those from both ends. Our theoretical and experimental results are further supported by 3D structures of CBM–cellohexaose complexes, namely $\mathrm{CBD_{CBHI}}$, $\mathrm{CBD_{CBHII}}$, $\mathrm{CBD_{EGI}}$ [17], *Pe*CBM29-2 [27,28] and *Cf*CBM2a [29].

We also observed that there are key aromatic residues at the *Ct*CBM11 interface (i.e. Tyr22, Tyr53, Tyr129 and Tyr152) that appear to have a preponderant role in guiding and packing the carbohydrate chain and therefore in the binding process. The initial conformations of these residues were responsible for the negative results of the initial docking calculations, and only after exploring the configurational space of these residues, through a multi-stage docking with an automated molecular modelling protocol (MADAMM software), were more reliable results obtained that are in agreement with the experimental data. No significant differences in the binding conformations were detected regarding α and β isomers.

Moreover, we propose that these residues have a preponderant role in the reorientation of the carbohydrate chain, directing it to a specific polar region in the protein that is populated with aspartate residues.

Regarding the overall evaluation of the results obtained in the present study, we can infer a general mechanism for the interaction between *Ct*CBM11 and cellulose. A minimum number of glucose units in the polymer chain are necessary for a stable binding (four in this case). Another feature is the strong interaction of some residues in the putative binding site with the hydroxyl groups at positions 2 and 6 from the central glucose units of the ligand. The guiding and packing of the carbohydrates is achieved through the interaction of the oligosaccharide with tyrosine residues that direct it towards polar amino acids responsible for zipping the oligosaccharide at the CBM cleft. As *Ct*CBM11 is topologically similar and structurally homologous to CBMs of families 4, 6, 15, 17, 22, 27 and 29 [8], we can infer that the binding mechanism of these CBMs to their substrates should be very similar to that of *Ct*CBM11.

Because these residues are conserved in type B CBMs, a multidisciplinary NMR, molecular modelling and X-ray crystallography study is currently in progress to determine their role in the global mechanism of interaction for several CBMs.

## Experimental procedures

### Sources of sugars

Cellobiose, cellotetraose and cellohexaose, were obtained from (Seikagaku Corporation) (Tokyo, Japan) and were used without further purification.

### Protein expression and purification

To express *Ct*CBM11 in *Escherichia coli*, the region of the Lic26A-Cel5A gene (*lic26A-cel5A*) encoding the internal family 11 CBM was amplified from *C. thermocellum* as described previously [14]. The protein was purified by ion metal affinity chromatography. Fractions containing the purified protein were buffer exchanged, in PD-10 Sephadex G-25M gel filtration columns (Amersham Pharmacia Biosciences, Piscataway, NJ, USA), into water. The purified protein was then concentrated with Amicon 10 kDa molecular-mass centrifugal membranes (Millipore, Billerica, MA, USA).

## NMR spectroscopy

All NMR experiments were performed with a Bruker ARX 400 spectrometer or a Bruker Avance 600 or a Bruker Avance 400 spectrometer (Bruker, Wissembourg, France) and conducted at 300.4 K. All spectra were processed with the software TOPSPIN 2.0 (Bruker).

$^1$H spectrum of cellohexaose was acquired at 600 MHz with 16 scans and a spectral width of 6009.6 Hz, centered at 2820.93 Hz. The solution of the sugar was prepared in 90% $H_2O$ and 10% (v/v) $D_2O$.

The interaction between *Ct*CBM11 and cellohexaose was studied by STD-NMR (the pulse sequence from the Bruker library was used) and by broadening of the resonances of the $^1$H spectrum of the sugar [16]. The 1D STD-NMR was performed using a solution of cellohexaose 95 μM and *Ct*CBM11 5 μM in $D_2O$. The spectra were recorded at 600 MHz with 8192 scans in a spectral window with 8980 Hz centered at 2824.35 Hz. Selective saturation of protein resonances at 0.6 p.p.m. (12 p.p.m. for reference spectra) was performed using a series of 40 Gaussian shaped pulses (50 ms, 1 ms delay between pulses) for a total saturation time of 2.0 s. Subtraction of saturated spectra from reference spectra was performed by phase cycling. Measurement of enhancement intensities was performed by direct comparison of STD-NMR. The broadening studies were performed at 400 MHz by titration of a solution of cellohexaose 0.79 mM prepared in $D_2O$ with *Ct*CBM11. 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 concentration in the cellohexaose solution at the end of the titration was 0.23 mM. All the spectra were acquired with 128 scans in a spectral window with 1991.6 Hz, centered at 1881.0 Hz. The spectra were deconvoluted into individual Lorentzian lines to determine the full linewidth at half-height.

The interaction between calcium and cellohexaose was studied by titration of a solution of cellohexaose 8 mM prepared in $D_2O$ with $CaCl_2$ 0.16 M. A first $^1$H-NMR spectrum was acquired on the sugar alone. Five further spectra were acquired with 0.5, 1.0, 2.0, 3.0 and 6.0 equivalents of $CaCl_2$, respectively. All the spectra were acquired at 400 MHz, with 128 scans and a spectral width of 6636.36 Hz, centered at 1879.78 Hz.

## Molecular modelling

The 1v0a protein databank deposited structure of *Ct*CBM11 [14] was used as the starting point for all the computational studies. All waters and sulfate ions ($SO_4^{2-}$) were deleted and only the protein atoms were kept. Furthermore, all selenium atoms were substituted by sulfur atoms.

The protein is composed of 173 amino acids but the crystallographic file lacks three amino acids in a loop between Val78 and Ala82. These residues were modelled with the help of the software INSIGHT II [30] to generate the correct sequence (i.e. Val78, Asp79, Gly80, Ser81 and Ala82). Once the structure was ready, hydrogen atoms were added using INSIGHT II software, considering all residues in their physiological protonation state.

To evaluate *Ct*CBM11, selectivity to saccharides several ligands were designed, namely, cellobiose, cellotetraose and cellohexaose [14]. As glucose can exist in two forms, α-glucose and β-glucose, and as these monomers have the ability to change between these two forms very easily at the considered temperature (333 K), each ligand was modelled in both forms.

## Molecular docking

The six modelled substrates were initially docked in the structure of the unbound *Ct*CBM11, and the best docking solutions were taken as starting structures for the subsequent molecular dynamics simulations. The docking procedure resorted to GOLD [31], a program that calculates the docking modes of small molecules into protein binding sites. The program is based on a genetic algorithm that is used to place different ligand conformations in the protein binding site, recognized by a fitting points strategy. Two scoring functions are *a posteriori* available to rank the obtained solutions (i.e. GoldScore and ChemScore) [32]. In our calculations, we used GoldScore as the scoring function, which has four terms:

$$\text{GOLD GoldScore fitness} = S_{\text{hb\_ext}} + S_{\text{vdw\_ext}} + S_{\text{hb\_int}} + S_{\text{vdw\_int}} \tag{2}$$

in which $\mathrm{S}_{\text{hb\_ext}}$ is the protein–ligand hydrogen bond score and $S_{\text{vdw\_ext}}$ is the van der Walls score. $S_{\text{hb\_int}}$ is the contribution due to intramolecular hydrogen bonds and $S_{\text{vdw\_int}}$ is the sum of the intenal torsion strain energy and internal van der Walls terms in the ligand. In general, the GoldScore function appears to perform better binding energy predictions than the ChemScore function, which justifies our choice [5].

## Molecular dynamics

All geometry optimizations and molecular dynamics were performed with the parameterization adopted in AMBER 8, [33] using the general AMBER force field for the protein and the Glycam-04 parameters for the carbohydrates [34–36].

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 [37].

In the initial stage, the structure was minimized in two stages. In the first stage, we kept the protein fixed and only minimized the position of the water molecules and ions. In the second stage, the full system was minimized. Subsequently, 2 ns molecular dynamics simulations were performed with the optimized structures. All simulations presented were carried out using the sander module, implemented in the AMBER 8 simulations package, with the Cornell force field [38].

Bond lengths involving hydrogens were constrained using the SHAKE algorithm [39] and the equations of motion were integrated with a 2 fs time-step using the Verlet leap-frog algorithm and the nonbonded interactions truncated with a 10 Å cutoff. The temperature of the system was regulated by the Langevin thermostat to maintain the temperature of our system at 333.15 K [40–42]. This temperature was chosen because it is the temperature of the microbial niche occupied by variants of the enzyme CelE in the bacterium *C. thermocellum* [43].

## Acknowledgements

The authors would like to thank the research network REQUIMTE (Project Reqmol), as well as the Portuguese Science and Technology Foundation (FCT-MCTES), for financial support through project PTDC/QUI/68286/2006 and scholarships SFRH/BPD/27237/2006 and SFRH/BD/31359/2006.

## References

1 Energy UDo (2006) Genomics:GTL Bioenergy Research Centers White Paper.

2 Demain AL, Newcomb M & Wu JHD (2005) Cellulase, clostridia, and ethanol. *Microbiol Mol Biol Rev* **69**, 124–154.

3 Béguin P & Lemaire M (1996) The cellulosome: an exocellular, multiprotein complex specialized in cellulose degradation. *Crit Rev Biochem Mol Biol* **31**, 201–236.

4 Tuka K, Zverlov VV & Velikodvorskaya GA (1992) Synergism between *Clostridium thermocellum* cellulases cloned in *Escherichia coli*. *Appl Biochem Biotecnol* **37**, 201–207.

5 Boraston AB, Bolam DN, Gilbert HJ & Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769–778.

6 Ozkan M & Özcengiz G (2006) Primary structure of the carbohydrate-binding modules in various cellulolytic, thermophilic, anaerobic, ethanol-producing isolates. *Turk J Biol* **30**, 45–50.

7 Pires VMR, Henshaw JL, Prates JAM, Bolam DN, Ferreira LMA, Fontes CMGA, Henrissat B, Planas A, Gilbert HJ & Czjzek M (2004) The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5Å in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *J Biol Chem* **279**, 21560–21568.

8 Hashimoto H (2006) Recent structural studies of carbohydrate-binding modules. *CMLS, Cell Mol Life Sci* **63**, 2954–2967.

9 Divne C, Stahlberg J, Reinikainen T, Ruohonen L, Pettersson G, Knowles JKC, Teeri TT & Jones A (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase-I from *Trichoderma reesei*. *Science* **265**, 524–528.

10 Charnock SJ, Bolam DN, Turkenburg JP, Gilbert HJ, Ferreira LMA, Davies GJ & Fontes CMGA (2000) The X6 'thermostabilizing' domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. *Biochemistry* **39**, 5013–5021.

11 Pell G, Williamson MP, Walters C, Du HM, Gilbert HJ & Bolam DN (2003) Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry* **42**, 9316–9323.

12 Xie HF, Gilbert HJ, Charnock SJ, Davies GJ, Williamson MP, Simpson PJ, Raghothama S, Fontes CMGA, Dias FMV, Ferreira LMA *et al.* (2001) *Clostridium thermocellum* Xyn10B carbohydrate-binding module 22-2: the role of conserved amino acids in ligand binding. *Biochemistry* **40**, 9167–9176.

13 Notenboom V, Boraston AB, Chiu P, Freelove ACJ, Kilburn DG & Rose DR (2001) Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: An X-ray crystallographic, thermodynamic and mutagenic study. *J Mol Biol* **314**, 797–806.

14 Carvalho AL, Goyal A, Prates JAM, Bolam DN, Gilbert HJ, Pires VMR, Ferreira LMA, Planas A, Romão MJ & Fontes CMGA (2004) The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accomodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *J Biol Chem* **279**, 34785–34793.

15 Meyer B & Peters T (2003) NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie-International Edition* **42**, 864–890.

16 Mayer M & Meyer B (1999) Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie-International Edition* **38**, 1784–1788.

17 Mattinen ML, Linder M, Teleman A & Annila A (1997) Interaction between cellohexaose and cellulose binding domains from *Trichoderma reesei* cellulases. *FEBS Lett* **407**, 291–296.

18 Vogtherr M & Peters T (2000) Application of NMR based binding assays to identify key hydroxy groups for intermolecular recognition. *J Am Chem Soc* **122**, 6093–6099.

19 Stockman BJ & Dalvit C (2002) NMR screening techniques in drug discovery and drug design. *Prog Nucl Magn Reson Spectrosc* **41**, 187–231.

20 Klages J, Coles M & Kessler H (2006) NMR-based screening: a powerful tool in fragment-based drug discovery. *Mol Biosyst* **2**, 319–331.

21 Klein J, Meinecke R, Mayer M & Meyer B (1999) Detecting binding affinity to immobilized receptor proteins in compound libraries by HR-MAS STD NMR. *J Am Chem Soc* **121**, 5336–5337.

22 Cerqueira NMFSA, Brás NF, Fernandes PA & Ramos MJ (2008) MADAMM–Multi staged docking with a molecular modelling protocol (http://www.fc.up.pt/pessoas/nscerque/MADAMM.html).

23 Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG & Davies GJ (2002) Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J Mol Biol* **319**, 1143–1156.

24 Chavez MI, Andreu C, Vidal P, Aboitiz N, Freire F, Groves P, Asensio JL, Asensio G, Muraki M, Canada FJ *et al.* (2005) On the importance of carbohydrate-aromatic interactions for the molecular recognition of oligosaccharides by proteins: NMR studies of the structure and binding affinity of AcAMP2-like peptides with non-natural naphthyl and fluoroaromatic residues. *Chem-Eur J* **11**, 7060–7074.

25 Fernandez MD, Canada FJ, Jimenez-Barbero J & Cuevas G (2005) Molecular recognition of saccharides by proteins, insights on the origin of the carbohydrate-aromatic interactions. *J Am Chem Soc* **127**, 7379–7386.

26 Humphrey W, Dalke A & Schulten K (1996) VMD – Visual Molecular Dynamics. *J Mol Graph* **14**, 33–38.

27 Charnock SJ, Bolam DN, Nurizzo D, Szabo L, McKie VA, Gilbert HJ & Davies GJ (2002) Promiscuity in ligand-binding: the three-dimensional structure of a Piromyces carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc Natl Acad Sci U S A* **99**, 14077–14082.

28 Flint J, Bolam DN, Nurizzo D, Taylor EJ, Williamson MP, Walters C, Davies GJ & Gilbert HJ (2005) Probing the mechanism of ligand recognition in family 29 carbohydrate-binding modules. *J Biol Chem* **280**, 23718–23726.

29 Simpson PJ, Xie HF, Bolam DN, Gilbert HJ & Williamson MP (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem* **275**, 41137–41142.

30 Accelrys (1993) InsightII v. 2.3.0. Accelrys, San Diego, CA.

31 Jones G, Willett P, Glen RC, Leach AR & Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727–748.

32 Verdonk ML, Cole JC, Hartshorn MJ, Murray CW & Taylor RD (2003) Improved protein-ligand docking using GOLD. *Proteins-Struct Funct Genetics* **52**, 609–623.

33 Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz HM, Wang B, Pearlman DA *et al.* (2004) *AMBER 8*. University of California, San Francisco, CA.

34 Kirschner KN & Woods RJ (2001) Solvent interactions determine carbohydrate conformation. *Proc Natl Acad Sci U S A* **98**, 10541–10545.

35 Basma M, Sundara S, Calgan D, Venali T & Woods RJ (2001) Solvated ensemble averaging in the calculation of partial atomic charges. *J Comput Chem* **22**, 1125–1137.

36 Kirschner KN & Woods RJ (2001) Quantum mechanical study of the nonbonded forces in water-methanol complexes. *J Phys Chem A* **105**, 4150–4155.

37 Asensio JL & Jimenez-Barbero J (1995) 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* **35**, 55–73.

38 Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW & Kollman PAJ (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic molecules. *J Am Chem Soc* **117**, 5179–5197.

39 Ryckaert JP, Ciccotti G & Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints – molecular-dynamics of N-alkanes. *J Comput Phys* **23**, 327–341.

40 Pastor IWR, Brooks BR & Szabo AJ (1998) An analysis of the accuracy of Langevin and molecular-dynamics algorithms. *Mol Phys* **65**, 1409–1419.

41 Loncharich RJ, Brooks BR & Pastor RW (1992) Langevin dynamics of peptides – the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **32**, 523–535.

42 Izaguirre JA, Catarello DP, Wozniak JM & Skeel RDJ (2001) Langevin stabilization of molecular dynamics. *Chem Phys* **114**, 2090–2098.

43 Mosolova TP, Kalyuzhnyi SV, Varfolomeyev SD & Velikodvorskaya GA (1995) Characterization of 3 enzymes from clostridium-thermocellum cellulase complex - synergism in cellulose hydrolysis. *Biochemistry-Mosc* **60**, 569–574.