

**DEVELOPMENT OF NMR METHODS FOR  
THE STRUCTURAL ELUCIDATION OF  
LARGE PROTEINS**

**ZHENG YU**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2010**

**DEVELOPMENT OF NMR METHODS FOR  
THE STRUCTURAL ELUCIDATION OF  
LARGE PROTEINS**

**ZHENG YU**

(B.Sc., Xiamen University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2010

## *Acknowledgements*

I would like to express my sincere appreciation and gratitude to my enthusiastic supervisor Associate Professor Yang Daiwen, for his guidance, inspiration, patience, encouragement and trust throughout the project.

My special thanks to Prof. Ho, Chien from Department of Biological Sciences, Carnegie Mellon University for providing the HbCO A sample and Prof. Wyss, Daniel F. from Schering-Plough Research Institute for providing the AcpS sample. Without their kind support and efficient collaboration it would not have been possible for me to complete this project.

I would also like to express my appreciation to Dr. Mok, Yu-Keung and other QE committee members, for their helpful advice and critical suggestions. Thanks were also due to Dr. Xu, Yingqi and Dr. Fang, Jingsong for their assistance in NMR experiments and data analysis.

I wish to take this opportunity to express my gratitude to my fellow graduates, postdoctoral fellows, friends, brothers and sisters from department of biological sciences and other departments/institutes. Their friendship made my research life at the NUS a pleasant learning experience. In particular, I'd like to thank Lin Zhi, Li Kai, Dr. Ru Mingbo, Shi Jiahai, Siu Xiaogang, Xu Xingfu, Yang Shuai, Dr. Zhang Xu, Dr. Zhang Yonghong, and Zhang Yuning for many discussions and help on the subject of this thesis.

Although any words are not even enough to express my heartfelt gratitude to my family in China, I would still like to thank my parents for their sustaining family

love and support. Without this everlasting love, I would not have been able to accomplish or even start this thesis.

Lastly, the financial assistance in the form of a research scholarship provided by National University of Singapore is gratefully acknowledged.

# *Table of Contents*

|   |      |
|---|------|
| <b>Acknowledgements</b>   | i    |
| <b>Table of Contents</b>  | iii  |
| <b>Summary</b>  | ix   |
| <b>List of Tables</b>   | xii  |
| <b>List of Figures</b>  | xiii |
| <b>List of Abbreviations</b>                                      | xx   |
| <b>Chapter 1:</b>   | 1    |
| <b>Related background and previous work</b>                       |      |
| 1.1 Protein NMR in structural biology                             | 2    |
| 1.2 Protein structure determination by NMR spectroscopy           | 5    |
| 1.2.1 Protein sample preparation                                  | 7    |
| 1.2.2 NMR data Processing   | 7    |
| 1.2.3 Sequence-specific NMR resonance assignment                  | 8    |
| 1.2.4 Structural restraint extraction                             | 9    |
| 1.2.5 Structure calculation and refinement                        | 9    |
| 1.3 Introduction to sequence-specific NMR resonance assignment    | 10   |
| 1.3.1 Important role of sequence-specific resonance assignment    | 10   |
| 1.3.2 General strategy for sequence-specific resonance assignment | 13   |
| 1.3.2.1 <sup>1</sup> H homonuclear assignment strategy            | 14   |
| 1.3.2.2 Triple-resonance assignment strategy                      | 16   |
| 1.3.3 Limitations of the conventional strategies                  | 20   |
| 1.4 Previous works on large proteins                              | 21   |
| 1.4.1 Reducing protein transverse relaxation rate                 | 23   |

|  |    |
|--|----|
| 1.4.2 Reducing protein spectral crowding and chemical shift degeneration   | 25 |
| 1.5 Research objectives  | 26 |
| <b>Chapter 2:</b><br><b>Sequence-specific assignments of methyl groups in large proteins</b>                             | 28 |
| 2.1 Introduction   | 29 |
| 2.2 General strategy for sequence-specific assignments of methyl groups  | 30 |
| 2.3 Discussion   | 35 |
| 2.4 Conclusion   | 38 |
| 2.5 Materials and methods  | 38 |
| <b>Chapter 3:</b><br><b>Side-chain assignments of methyl-containing residues in large proteins</b>                       | 40 |
| 3.1 Introduction   | 41 |
| 3.2 General strategy for side-chain assignments of methyl-containing residues  | 44 |
| 3.2.1 Methyl assignments   | 44 |
| 3.2.2 Assignment of side-chain protons in methyl-containing residues   | 47 |
| 3.3 Conclusion   | 51 |
| 3.4 Materials and methods  | 51 |
| 3.4.1 MQ-(H)CCH-TOCSY experiment   | 51 |
| 3.4.2 H(C)C <sub>m</sub> H <sub>m</sub> -TOCSY experiment  | 53 |
| 3.4.3 Protein Samples and NMR Spectroscopy   | 53 |
| 3.4.4 Correction of <sup>13</sup> C chemical shifts  | 54 |
| <b>Chapter 4:</b><br><b>A new strategy for structure determination of large proteins in solution without deuteration</b> | 56 |
| 4.1 Introduction   | 57 |

---

|  |     |
|--|-----|
| 4.2 General strategy for sequence-specific assignments   | 58  |
| 4.2.1 General strategy for sequential assignment   | 58  |
| 4.2.1.1 Peak clusters  | 60  |
| 4.2.1.2 Spin-system identification and amino acid type determination   | 64  |
| 4.2.1.3 Assembly and mapping of connectivity fragments   | 68  |
| 4.2.1.4 Resolution of ambiguity in connectivity  | 69  |
| 4.2.2 Side-chain assignment  | 72  |
| 4.3 NOE assignment and structure determination   | 72  |
| 4.4 Discussion and conclusion  | 79  |
| 4.5 Materials and methods  | 81  |
| 4.5.1 Protein samples and NMR Spectroscopy   | 81  |
| 4.5.2 Identifying spin-systems   | 82  |
| 4.5.3 Structure calculation  | 83  |
| 4.5.4 Data deposition  | 84  |
| <b>Chapter 5:</b>  | 102 |
| <b>STARS: software for statistics on inter-atomic distances and torsion angles in protein secondary structures</b> |     |
| 5.1 Introduction   | 103 |
| 5.2 Overview of STARS  | 104 |
| 5.2.1 Composition of database  | 104 |
| 5.2.2 Definition   | 105 |
| 5.2.3 User interface   | 111 |
| 5.3 Results and discussion   | 113 |
| <b>Chapter 6:</b>  | 114 |
| <b>NMRspy: software package for NMR spectroscopy visualization, analysis and management</b>                        |     |
| 6.1 Introduction   | 115 |

|  |     |
|--|-----|
| 6.2 Feature and advantages of NMRspy                 | 117 |
| 6.2.1 Intrinsic capabilities                         | 117 |
| 6.2.2 Capability of analyzing Folded-spectrum        | 118 |
| 6.2.2.1 Proper frequency display of aliased peaks    | 118 |
| 6.2.2.2 Spectra synchronization & cursor correlation | 120 |
| 6.2.3 Multi-dimension-peakpicking capability         | 123 |
| 6.2.4 Project management capability                  | 125 |
| 6.2.5 Spectral view simplification capability        | 126 |
| 6.3 User's interface                                 | 129 |
| 6.3.1 Control panel                                  | 130 |
| 6.3.1.1 Spectrum menu                                | 132 |
| 6.3.1.2 DataSet menu                                 | 134 |
| 6.3.1.3 Project menu                                 | 134 |
| 6.3.1.4 Analysis menu                                | 135 |
| 6.3.1.5 Extensions menu                              | 138 |
| 6.3.2 Spectral display windows                       | 139 |
| 6.3.2.1 Spectrum control bar                         | 140 |
| 6.3.2.2 Mouse and keypad navigation                  | 144 |
| 6.3.2.3 Status bar                                   | 146 |
| 6.3.3 Spectral attribute windows                     | 147 |
| 6.3.3.1 File panel                                   | 148 |
| 6.3.3.2 View panel                                   | 150 |
| 6.3.3.3 Level panel                                  | 152 |
| 6.3.3.4 Peak & label panel                           | 153 |
| 6.3.4 Other dialogs & windows                        | 156 |



---

|   |     |
|---|-----|
| 6.3.4.1 Peak (label, grid) editor   | 156 |
| 6.3.4.2 Peak (label, grid) table  | 157 |
| 6.3.4.3 Peak auto-assign dialog   | 158 |
| 6.3.4.4 Peak identification dialog  | 159 |
| 6.4 Results and discussion  | 160 |
| <b>Chapter 7:</b>   | 162 |
| <b>XYZ4D: software plug-in for backbone assignment using the new NOESY-based strategy</b> |     |
| 7.1 Introduction  | 163 |
| 7.2 Interface and algorithms  | 166 |
| 7.2.1 The main application window   | 166 |
| 7.2.2 Project preparation module  | 168 |
| 7.2.3 Spectral calibration module   | 171 |
| 7.2.3.1 Main panel  | 172 |
| 7.2.3.2 Selection of isolated HSQC peaks  | 173 |
| 7.2.3.3 HNCA calibration (H, N)   | 174 |
| 7.2.3.4 HN(CO)CA calibration (H, N)   | 176 |
| 7.2.3.5 HN(CO)CA calibration (C)  | 176 |
| 7.2.3.6 4DNOE calibration (H, N)  | 177 |
| 7.2.3.7 4DNOE calibration (C)   | 178 |
| 7.2.3.8 CCH diagonal calibration (C, CH)  | 180 |
| 7.2.3.9 CCH calibration (H,C)   | 181 |
| 7.2.3.10 Results panel  | 183 |
| 7.2.4 Cluster identification module   | 184 |
| 7.2.4.1 Method  | 185 |
| 7.2.4.2 Main panel  | 188 |

---

|  |     |
|--|-----|
| 7.2.4.3 Cluster inspection panel         | 189 |
| 7.2.4.4 Results panel                    | 192 |
| 7.2.5 CCH & 4DNOE inspection module      | 193 |
| 7.2.5.1 Interface                        | 194 |
| 7.2.5.2 CCH water-peak elimination       | 196 |
| 7.2.5.3 CCH artificial -peak elimination | 197 |
| 7.2.5.4 NOE-peak collection              | 198 |
| 7.2.5.5 NOE-peak alias correction        | 198 |
| 7.2.6 Spin-system identification module  | 199 |
| 7.2.6.1 Methods                          | 200 |
| 7.2.6.2 Interface                        | 202 |
| 7.2.7 Cluster mapping module             | 205 |
| 7.2.7.1 Methods                          | 206 |
| 7.2.7.2 Interface                        | 214 |
| 7.2.8 Backbone assignment module         | 220 |
| 7.3 Results and discussion               | 221 |
| <b>References</b>                        | 223 |
| <b>Publications</b>                      | 234 |

## Summary

Protein structures are an important source of information for understanding biological function at the molecular level and provide the basis for many studies in research areas such as structure-based drug design and homology modelling. Currently the two main techniques for determining the three-dimensional structures of biological macromolecules are X-ray diffraction and NMR spectroscopy. In cases where proteins cannot be crystallized, NMR is the best, perhaps the only, method available to characterize the structures.

At present, ~15% of protein structures deposited in the protein data bank is determined by NMR, but only ~1% of the NMR structures are for proteins larger than 25 kDa. Additionally, most of the large proteins only have crude global folds based on backbone assignments and a few side chain assignments which are obtained using deuterated samples. Unfortunately, the preparation of deuterated or/and specific isotopic labelled protein samples is often challenging and places a bottleneck on the NMR study of large proteins.

In this thesis, I proposed several new NMR techniques and computational methods to obtain partial or complete sequence specific assignments and to further determine high-resolution structures of larger proteins, using both the simple and cheap non-deuterated protein samples.

Firstly, a new 3D multiple-quantum MQ-(H)CCmHm-TOCSY experiment is presented in chapter 2 to assign methyl resonances in high-molecular weight proteins, on the basis of spectral patterns and prior backbone assignments. The favorable relaxation properties of the multiple-quantum

coherences and the slow decays of in-phase methyl  $^{13}\text{C}$  magnetizations optimize performance of the proposed experiment for application to large proteins. In combination with the H(C)CmHm-TOCSY experiment, a strategy is presented in chapter 3 for assigning protons of methyl-containing residues of uniformly  $^{13}\text{C}$ -labeled large proteins.

Secondary, I present a novel strategy in chapter 4 to assign backbone and side chain resonances of large proteins without deuteration, with which one can obtain high resolution structures from  $^1\text{H}$ - $^1\text{H}$  distance restraints. The strategy uses information from through-bond correlation experiments to filter intra-residue and sequential correlations from through-space correlation experiments, and then matches the filtered correlations to obtain sequential assignment. The strategy extends the size limit for structure determination by NMR to 42 kDa for monomeric proteins and to 65 kDa for differentially labeled multimeric proteins without deuteration or selective labeling.

To assist the development of the new strategy mentioned above, a graphics package STARS was developed for performing statistics on interatomic distances and torsion angles in protein secondary structures from a protein crystal structure database. This graphics package shown in chapter 5 is also capable of facilitating assignment of ambiguous NOESY peaks, NMR structure determination, structure validation and comparison of protein folds.

In order to comply with the requirements of our new experiments and strategies, I present a new software package NMRspy in chapter 6 which can be used for NMR spectroscopy visualization, analysis and management. It provides a variety of function and analysis routines that facilitate the analysis of complex,

crowded and folded high-dimensional spectra. On the basis of this software platform, in chapter 7 I present a software extension XYZ4D for semi-automatic and automatic analysis of NMR data using the novel strategy shown in chapter 4. This software extension corresponds to the manual assignment steps of the new strategy but release users from tedious and time-consuming routines.

## List of Tables

|                   |   |     |
|-------------------|---|-----|
| <b>Table 1.1:</b> | Heteronuclear Experiments Used for protein sequence-specific resonance assignment.  | 17  |
| <b>Table 2.1:</b> | The relatively good dispersion of ( $^{13}\text{C}_\alpha$ , $^{13}\text{C}_\beta$ ) chemical shifts in large monomeric proteins. | 35  |
| <b>Table 3.1:</b> | Summary of assignment of non-methyl protons in methyl-containing residues of both $\alpha$ - and $\beta$ -chains of rHbCOA.       | 49  |
| <b>Table 4.1:</b> | Summary of clusters, spin-systems, dipeptide segments and assignments.  | 63  |
| <b>Table 4.2:</b> | Structural statistics for the final 10 conformers of MBP.   | 75  |
| <b>Table 4.3:</b> | Structural statistics for the final 10 conformers of HbCO A.  | 76  |
| <b>Table 4.4:</b> | Experimental parameters.  | 77  |
| <b>Table 5.1:</b> | Ten types of secondary structures defined in STARTS and their one-letter symbols.   | 106 |
| <b>Table 6.1:</b> | Icons in control bar.   | 140 |
| <b>Table 7.1:</b> | Statistic $^{13}\text{C}$ - $^1\text{H}$ chemical shift region.   | 199 |

## List of Figures

|                    |   |    |
|--------------------|---|----|
| <b>Figure 1.1:</b> | The flowchart of protein structure determination by NMR.  | 6  |
| <b>Figure 1.2:</b> | Schematic depiction of backbone assignment using the CBCANH and CBCA(CO)NH spectra.   | 18 |
| <b>Figure 1.3:</b> | Effects of protein size on NMR signals.   | 22 |
| <b>Figure 2.1:</b> | Pulse sequence for the MQ-(H)CC <sub>m</sub> H <sub>m</sub> -TOCSY experiment.  | 31 |
| <b>Figure 2.2:</b> | Representative slices from the MQ-(H)CC <sub>m</sub> H <sub>m</sub> -TOCSY spectrum used for methyl assignments.  | 33 |
| <b>Figure 2.3:</b> | CT <sup>13</sup> C- <sup>1</sup> H HSQC of the <sup>13</sup> C, <sup>15</sup> N-labeled AcpS. Cross-peaks are labeled with their assignments.                           | 34 |
| <b>Figure 2.4:</b> | Histograms of signal-to-noise ratios of correlations from MQ-(H)CC <sub>m</sub> H <sub>m</sub> -TOCSY and HCCH-TOCSY spectra acquired at 25 °C.                         | 37 |
| <b>Figure 2.5:</b> | Pulse scheme for the CC <sub>m</sub> H <sub>m</sub> -TOCSY experiment applied to <sup>2</sup> H, <sup>13</sup> C, <sup>1</sup> H <sub>m</sub> -labeled protein samples. | 39 |
| <b>Figure 3.1:</b> | Representative F1–F3 slices from the MQ-(H)CC <sub>m</sub> H <sub>m</sub> -TOCSY (A) and MQ-(H)CCH-TOCSY (B) spectra of <sup>13</sup> C-labeled α-chain of rHbCO A.     | 43 |
| <b>Figure 3.2:</b> | CT <sup>13</sup> C- <sup>1</sup> H HSQC of the <sup>13</sup> C-labeled α-chain and β-chain of rHbCO A.  | 46 |
| <b>Figure 3.3:</b> | Representative F1–F3 slices from the H(C)C <sub>m</sub> H <sub>m</sub> -TOCSY spectrum of <sup>13</sup> C-labeled β-chain of rHbCOA.                                    | 48 |

|                    |   |     |
|--------------------|---|-----|
| <b>Figure 3.4:</b> | F1-F3 slices taken from the spectra of H(C)C <sub>m</sub> H <sub>m</sub> -TOCSY, MQ-(H)CC <sub>m</sub> H <sub>m</sub> -TOCSY and MQ-(H)CCH-TOCSY experiments.                                 | 50  |
| <b>Figure 3.5:</b> | Pulse sequences for the MQ-(H)CCH-TOCSY (A) and H(C)C <sub>m</sub> H <sub>m</sub> -TOCSY (B) experiments.   | 52  |
| <b>Figure 4.1:</b> | Pulse sequence for recording 4D <sup>13</sup> C, <sup>15</sup> N-edited NOESY.  | 59  |
| <b>Figure 4.2:</b> | The middle region of a 2D TROSY-HSQC of fully protonated MBP recorded on an 800 MHz NMR at 30 °C.   | 61  |
| <b>Figure 4.3:</b> | Distributions of peak signal-to-noise (S/N) ratio for the 3D TROSY-HNCA experiments.  | 62  |
| <b>Figure 4.4:</b> | Identification of spin-systems.   | 65  |
| <b>Figure 4.5:</b> | Resolution of ambiguous connectivity between clusters.  | 67  |
| <b>Figure 4.6:</b> | Distribution of $\delta$ -NOE that reflects the difference in the number of common NOEs shared by two adjacent amide protons and those by two non-adjacent amides.                            | 70  |
| <b>Figure 4.7:</b> | Comparison of structures determined by NMR and x-ray methods.   | 74  |
| <b>Figure 4.8:</b> | Relative peak intensity ( $I(j,k)/I_{\text{ref}}$ ), as a function of overall correlation time ( $\tau_m$ ), calculated for different types of correlations in a number of 3D and 4D spectra. | 85  |
| <b>Figure 4.9:</b> | Detailed information on backbone assignments.   | 89  |
| <b>Figure 5.1:</b> | Definition of residues $i, J, j, K, k$ in antiparallel (a), parallel (b) and mixed parallel and antiparallel (c and d) $\beta$ -sheets.   | 107 |



---

|                    |   |     |
|--------------------|---|-----|
| <b>Figure 5.2:</b> | STARS user interface - Main window with the page for interatomic distance statistics in a single mode.                                      | 108 |
| <b>Figure 5.3:</b> | STARS user interface – (a) Window for selection of protein structures. (b) Page for torsion angle statistics in a single mode.              | 109 |
| <b>Figure 5.4:</b> | STARS user interface – (a) Page for interatomic distance statistics in a batch mode. (b) Page for torsion angle statistics in a batch mode. | 110 |
| <b>Figure 5.5:</b> | STARS user interface – Windows for result display and analysis.   | 111 |
| <b>Figure 6.1:</b> | Corresponding crosshairs in different windows.  | 122 |
| <b>Figure 6.2:</b> | Peak Resonance & DataHeight Adjustor.   | 124 |
| <b>Figure 6.3:</b> | Multiple spectral views with standard layout (a) and simple layout (b).   | 128 |
| <b>Figure 6.4:</b> | Overall Diagram of interfaces in NMRspy.  | 129 |
| <b>Figure 6.5:</b> | NMRspy Control Panel and its menus.   | 131 |
| <b>Figure 6.6:</b> | Project Manager Window.   | 131 |
| <b>Figure 6.7:</b> | Format Conversion Dialog.   | 133 |
| <b>Figure 6.8:</b> | Synchronize Views Panel.  | 135 |

---

|                     |  |     |
|---------------------|--|-----|
| <b>Figure 6.9:</b>  | Atom List Panel.                         | 136 |
| <b>Figure 6.10:</b> | Assignment Summarized Table.             | 137 |
| <b>Figure 6.11:</b> | NOE Calibration Panel.                   | 138 |
| <b>Figure 6.12:</b> | Spectral View (Spectral Display Window). | 139 |
| <b>Figure 6.13:</b> | Spectrum Printing Dialog.                | 143 |
| <b>Figure 6.14:</b> | Status Bar Setting Dialog.               | 147 |
| <b>Figure 6.15:</b> | Spectrum File Setting Panel.             | 149 |
| <b>Figure 6.16:</b> | Spectrum Reference Editor.               | 149 |
| <b>Figure 6.17:</b> | Spectral View Setting Panel.             | 151 |
| <b>Figure 6.18:</b> | Spectral Level Setting Panel.            | 151 |
| <b>Figure 6.19:</b> | Peak & Label Setting Panel.              | 155 |
| <b>Figure 6.20:</b> | Peak Editor Dialog.                      | 155 |

---

|                     |   |     |
|---------------------|---|-----|
| <b>Figure 6.21:</b> | Peak Assignment Dialog.   | 156 |
| <b>Figure 6.22:</b> | Peak Table.   | 158 |
| <b>Figure 6.23:</b> | Peak Auto-assign Dialog.  | 158 |
| <b>Figure 6.24:</b> | Peak Identification Dialog.   | 159 |
| <b>Figure 7.1:</b>  | Overall Diagram of interfaces in XYZ4D.   | 167 |
| <b>Figure 7.2:</b>  | Main application window of XYZ4D (a) and its pull-down menus (b).               | 168 |
| <b>Figure 7.3:</b>  | Graphic Interfaces of Project Preparation Module.                               | 169 |
| <b>Figure 7.4:</b>  | Over-edge peak.   | 170 |
| <b>Figure 7.5:</b>  | Main panel (a) and result summary panel (b) of the Spectral Calibration Module. | 172 |
| <b>Figure 7.6:</b>  | Isolated HSQC peak selection panel (a) and its correlated HSHC spectrum (b).    | 175 |
| <b>Figure 7.7:</b>  | Graphic interfaces for HNCA Calibration (H, N).                                 | 175 |
| <b>Figure 7.8:</b>  | Graphic interfaces for HN(CO)CA Calibration (C).                                | 177 |

---

|                     |   |     |
|---------------------|---|-----|
| <b>Figure 7.9:</b>  | Graphic interfaces for 4DNOE Calibration (H,N).   | 179 |
| <b>Figure 7.10:</b> | Graphic interfaces for 4DNOE Calibration (C).   | 180 |
| <b>Figure 7.11:</b> | Graphic interfaces for CCH Diagonal Calibration (C, CH).  | 181 |
| <b>Figure 7.12:</b> | Graphic interfaces for CCH Calibration (H, C).  | 182 |
| <b>Figure 7.13:</b> | Examples of cluster classification.   | 187 |
| <b>Figure 7.14:</b> | Main window (a) and result summary window (b) of Cluster Identification Module.                   | 189 |
| <b>Figure 7.15:</b> | Cluster inspection interface.   | 191 |
| <b>Figure 7.16:</b> | Control panels of (a) CCH-TOCSY and (b) 4D-NOESY Inspection.                                      | 195 |
| <b>Figure 7.17:</b> | Interfaces of (a) CCH Peak Navigator and (b) Cluster Navigator.                                   | 195 |
| <b>Figure 7.18:</b> | An example of artificial-peaks that surround strong peaks along the Y-axis in CCH-TOCSY spectrum. | 197 |
| <b>Figure 7.19:</b> | The graphic interface of spin-system identification.  | 204 |
| <b>Figure 7.20:</b> | Ten simulated annealing cooling schedules provide by XYZ4D.                                       | 212 |

---

|  |     |
|--|-----|
| <b>Figure 7.21:</b> Setting Panels of Energy Calculation Parameters.           | 214 |
| <b>Figure 7.22:</b> Control panel of Simulated Annealing-Monte Carlo approach. | 215 |
| <b>Figure 7.23:</b> Graphic interfaces for cluster mapping.                    | 218 |
| <b>Figure 7.24:</b> Protein Sequence Mapping.                                  | 219 |
| <b>Figure 7.25:</b> The panel of cluster mapping module.                       | 220 |
| <b>Figure 7.26:</b> Graphic interface of Backbone Assignment Module.           | 221 |

## *List of Abbreviations*

|                |   |
|----------------|---|
| <b>2D</b>      | two-dimensional                                   |
| <b>3D</b>      | three-dimensional                                 |
| <b>4D</b>      | four-dimensional                                  |
| <b>AcpS</b>    | Acyl Carrier Protein Synthase                     |
| <b>BMRB</b>    | Biological Magnetic Resonance Bank                |
| <b>COSY</b>    | Correlated Spectroscopy                           |
| <b>CSI</b>     | Chemical Shift Index                              |
| <b>DdCAD-1</b> | Ca <sup>2+</sup> -dependent cell adhesion protein |
| <b>FID</b>     | Free induction decay                              |
| <b>Hb A</b>    | Human normal adult haemoglobin                    |
| <b>HbCO A</b>  | Liganded Carbonmonoxy-Hb A                        |
| <b>HSQC</b>    | Heteronuclear Single Quantum Coherence            |
| <b>MBP</b>     | Maltose Binding Protein                           |
| <b>MQ</b>      | Multiple-quantum                                  |
| <b>MQF</b>     | Multiple Quantum Filtered                         |
| <b>NMR</b>     | Nuclear Magnetic Resonance                        |
| <b>NMRspy</b>  | NMR spectral pinpoint analysis system             |
| <b>NOE</b>     | Nuclear Overhauser Effect                         |
| <b>NOESY</b>   | Nuclear Overhauser Enhancement Spectroscopy       |
| <b>PDB</b>     | Protein Data Bank                                 |
| <b>ppm</b>     | Parts per million                                 |
| <b>rHbCO A</b> | Recombinant hemoglobin in the carbonmonoxy form   |
| <b>RMSD</b>    | Root-mean-square deviation                        |

|              |  |
|--------------|--|
| <b>SQ</b>    | Single-quantum   |
| <b>STARS</b> | Software tool for statistics on interatomic distances and dihedral angles in protein secondary structures  |
| <b>TOCSY</b> | Total Correlation Spectroscopy   |
| <b>TROSY</b> | Transverse Relaxation-Optimized Spectroscopy   |
| <b>XYZ4D</b> | Software tool that developed for Xu Yingqi, Yang Daiwen & Zheng Yu's novel strategy for solution structure determination of large proteins without deuteration using 4D NOESY and other 3D NMR spectra |

# **Chapter 1:**

## **Related background and previous work**

- 1.1 Protein NMR in structural biology
- 1.2 Protein structure determination by NMR spectroscopy
- 1.3 Introduction to sequence-specific NMR resonance assignment
- 1.4 Previous work on large proteins
- 1.5 Research objectives



## Chapter 1:

### Related background and previous work

#### 1.1 Protein NMR in structural biology

The dream of having genomes completely sequenced is now a reality. However, an even greater challenge, proteomics – the study of all the proteins coded by the genes under different conditions, awaits biologists to further unravel biological processes.

As one of the main categories in proteomics, structural proteomics, the determination and prediction of atomic resolution three-dimensional (3D) structures of proteins on a genome-wide scale for better understanding their structure-function relationships, has now provided a new rationale for structural biology and has become a major initiative in biotechnology. (Liu and Hsu 2005) In the field of protein structure determination, two instrumental methods have played dominant roles: X-ray crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy. These two main techniques can be used to determine the structures of macromolecules at atomic resolution.

Although X-ray crystallography is still the most powerful technique for structure determination, the throughput of structure determination using it remains unclear. It requires protein crystallization which is usually regarded as a slow, resource-intensive step with low success rates. In contrast, NMR spectroscopy does not require protein crystals, the experiments can be carried out in aqueous solution similar to the physiological conditions in which the protein normally functions. As NMR spectroscopy is an inherently insensitive technique,

NMR samples need not be as stringently pure as samples for crystallography, and it is relatively easy to explore a range of solution conditions (pH, temperature, salts) to find an optimum condition for data collection. The vast majority (~75%) of the NMR structures of proteins in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000) do not have the corresponding crystal structures, in large part because the proteins could not be crystallized. Another advantage of this crystal-free technique is that it avoids the crystallization process which may select a certain subset of conformers present under particular conditions. With the development of new techniques such as "In-cell NMR spectroscopy", it's now even possible to directly observe and analyze the conformational and functional properties of proteins inside living cells at atomic resolution. (Selenko and Wagner 2007)

Moreover, NMR spectroscopy has extended our ability to characterize protein dynamics and is a promising tool to study mechanisms by which these molecules might function. (Mittermaier and Kay 2006) Unlike the beautiful and static pictures of structures emerged from X-ray, proteins are in fact dynamic over a spectrum of time scales and we now know that there is an intimate relation between dynamics and molecular function. For example, protein dynamics contribute to the thermodynamic stability of functional states and play an important role in catalysis, where conformational rearrangements can juxtapose key catalytic residues; in ligand binding, which often involves the entry of molecules into areas that would normally be occluded; in molecular recognition processes, which are often fine-tuned by disorder-to-order transitions; and in allostery, where coupled structural fluctuations can transmit information between

distant sites in a protein. NMR spectroscopy is uniquely suited to study many of these dynamic processes. It has been developed to provide site-specific information about protein motions that cover various time scales, from rapid bond librations (picoseconds) to events that take seconds (Kay, Torchia et al. 1989; Palmer, Kroenke et al. 2001).

In addition, NMR spectroscopy is particularly valuable tool in investigation of protein interactions with other macromolecules or small molecules (Takeuchi and Wagner 2006). Such interactions play important roles in biological processes but often are weak and transient. The complexes of these interactions cannot be easily crystallized. The NMR's ability to characterize protein complexes under physiological conditions, even if the interactions are weak and transient, making it a good tool to understand the nature of these interactions. Thus, the development of new exchange-based NMR methods might provide an opportunity for studying large and more complex systems. (Post 2003)

NMR spectroscopy is also a prime tool for studying the structures and interactions of partially or fully unfolded proteins. It is predicted that 7-33% of bacterial proteins and 36-63% of eukaryotic proteins are intrinsically unfolded (Dunker and Obradovic 2001). Many proteins, such as those involved in gene expression, are natively unstructured and only structured upon forming specific complexes with other polypeptides or even small-molecule cofactors. Significant fractions of proteins may thus be partially or fully unfolded. Thus it is difficult to crystallize those proteins. NMR spectroscopy can determine if a protein contains extensive regions that are unfolded. More sophisticated analysis can be carried out using relaxation and heteronuclear NOE measurements to detect flexible

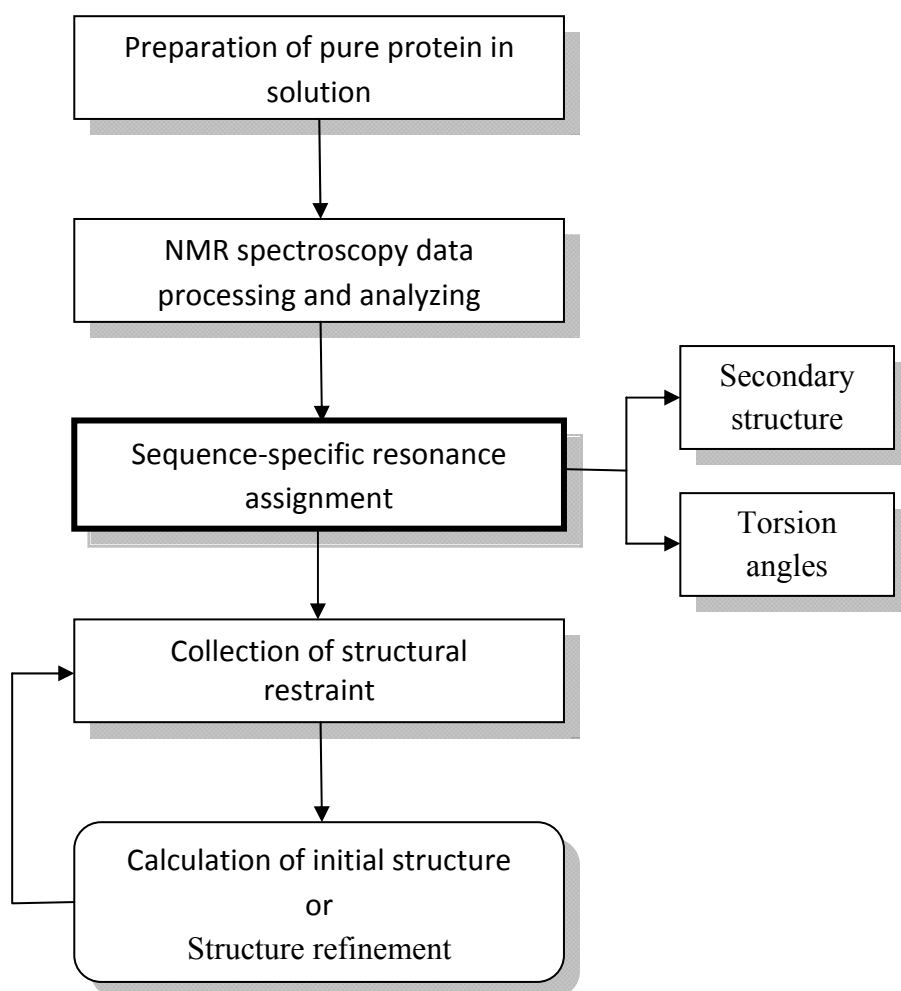
regions and obtain structural information. Therefore, NMR spectroscopy is also the preferred technique for the study of protein folding.

With these particular features, NMR not only provides structural and biophysical information that is complementary to X-ray crystallography, but also provides insights into structure–function relationships for a large number of proteins. The important role that NMR plays in structural biology is illustrated by far more than 6000 NMR protein solution structures deposited in the PDB.

NMR does not directly create an image of a protein. Rather, it is able to yield a wealth of indirect structural information from which the 3D structure can only be revealed by extensive data analysis and computer calculation. The typical strategy of a NMR structure determination follows a suite of steps, as described below.

## **1.2 Protein structure determination by NMR spectroscopy**

Figure 1.1 depicts the basic steps toward determining solution structures from NMR data set.



**Figure 1.1 The flowchart of protein structure determination by NMR.**

The sequence-specific resonance assignment that is emphasized by bold plays a key role in protein structure determination. Several strategies and software packages proposed in this thesis facilitate sequence-specific resonance assignment procedures on large proteins, as described in chapters 2, 3, 4, 6 and 7.

### 1.2.1 Protein sample preparation

Protein production using *Escherichia coli*-cell based expression systems has an established record of being the most successful approach to generate protein samples for structure study. It provides a cost-effective, flexible, reliable, and scalable way of sample preparation. In the case where the protein is not expressed well in *E.coli* or requires posttranslational modification (glycosylation, phosphorylation, etc.) , several eukaryotic options such as yeast, insect, and mammalian expression systems, or cell-free *in vitro* translation methods are available. Metabolic labelling of biomolecules with stable isotopes ( $^{15}\text{N}$ ,  $^{13}\text{C}$  and/or  $^2\text{H}$ ) for NMR spectroscopy was pioneered with *E.coli* expression systems and has been extended successfully to a few other systems (Kainosho 1997).

The higher the protein concentration, the faster the NMR data can be collected, provided that the protein does not aggregate. Practically, the sample concentration limits are about 200  $\mu\text{M}$  with ordinary probes and about 60  $\mu\text{M}$  with cryogenic probes. Depending on the length of the detection coil in the probe, a sample volume of 300 to 500  $\mu\text{L}$  is usually required. Some samples may be not stable over data collection period. Cryogenic probes together with higher magnetic field can shorten the time of each experiment, which makes it possible to investigate proteins that are less stable over time.

### 1.2.2 NMR data Processing

Normally NMR spectrometers produce resonance signals in 1D, 2D, 3D, and 4D spaces, which could reflect both the signature information of amino acid type and the adjacency information between amino acids. The general approach

in a biomolecular NMR study is to first convert time-domain data to frequency-domain spectra by Fourier transform. Then peaks are picked out from each spectrum. This identifies real resonance peaks that are generated from protein residues rather than noises.

Current protocols for processing NMR data set and peak picking use the programs NMRPipe (Fourier transformation) (Delaglio, Grzesiek et al. 1995), XEASY (peak picking and semi-automated assignment) (Bartels, Xia et al. 1995), NMRView (peak picking and spectrum data analysis as well as semi-automated assignment) (Johnson and Blevins 1994) and Sparky (peak picking and spectrum data analysis as well as semi-automated assignment) (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco).

### **1.2.3 Sequence-specific NMR resonance assignment**

Once NMR spectra are acquired, individual cross peaks in the experiments have to be assigned to sequence-specific positions in the primary sequence of protein before other structural restraints (e.g., the distance information between residues in the NOESY spectrum) can be fully interpreted. Sequence-specific NMR resonance assignment plays a key role in the whole process of structure determination.

As a major objective of my study is to improve and automate the resonance assignment procedures on large proteins, the detailed approach that is currently most widely used assignment procedure is depicted in latter sections of this chapter.

### 1.2.4 Structural restraint extraction

Structural restraints are obtained from the interpretation of data from one or more different classes of NMR experiments. Once all  $^1\text{H}$ ,  $^{15}\text{N}$ , and  $^{13}\text{C}$  resonances have been assigned, full analysis of one or more NOESY spectra, ‘NOE assignment’, provides the most important restraint,  $^1\text{H}$ - $^1\text{H}$  distance constraints ( $<5\text{\AA}$ ). Three-bond spin-spin coupling experiments provide torsion angle constraints, two dihedral angles associated with each peptide bond: angle  $\Phi$ , is the torsion angle between bond  $^{15}\text{N}$ - $^1\text{H}_\text{N}$  and  $\text{C}_\alpha$ - $\text{H}_\alpha$  while angle  $\Psi$  is another torsion angle between bond  $\text{C}_\alpha$ - $\text{H}_\alpha$  and  $\text{C}$ - $\text{O}$ . Besides, these torsion angles can also be predicted from the assigned chemical shifts of  $^{15}\text{N}$ ,  $\text{C}_\alpha$ ,  $\text{C}_\text{O}$ , and  $\text{C}_\beta$ , as described in program TALOS (Cornilescu, Delaglio et al. 1999). Additional hydrogen bond constraints are determined from hydrogen exchange experiments, chemical shifts, and/or trans-hydrogen-bond couplings (Cordier, Rogowski et al. 1999).

### 1.2.5 Structure calculation and refinement

NMR structures are obtained from constrained molecular dynamics simulations and energy minimization calculations, with the NOE-derived inter-proton distances being the primary experimental constraints as well as other available constraints. As a consequence of chemical shift degeneracy, many NOE cross peaks may have multiple assignment possibilities, and the results of preliminary structure calculations are used to eliminate unlikely candidates on the basis of inter-proton distances. Refinement continues in an iterative manner until a self-consistent set of experimental constraints produces an ensemble of



structures that also satisfies standard covalent geometry and steric overlap considerations.

Several structure calculation tools are available, such as CNS (Brunger, Adams et al. 1998), CYANA (Guntert 2004) and Autostructure (Zheng, Huang et al. 2003). A variety of computational approaches have been introduced either to support the interactive analysis of structure constrains by visualization and book-keeping or to provide automation for specific parts of an NMR structure determination, such as iterative NOE assignment tools ARIA (Linge, O'Donoghue et al. 2001; Habeck, Rieping et al. 2004) and CANDID (Herrmann, Guntert et al. 2002), automate NOE peak-picking tool ATNOS (Herrmann, Guntert et al. 2002).

## **1.3 Introduction to sequence-specific NMR resonance assignment**

Sequence-specific assignment has been an important role for protein structural analysis by NMR. A major objective of my study is to improve and automate the sequence-specific assignment procedures on large proteins. Before discussing the progress and challenges on this particular task when dealing with large proteins, I will describe the traditional assignment strategy for small and medium-sized proteins and their limitations.

### **1.3.1 Important role of sequence-specific resonance assignment**

As mentioned above, NMR spectra contain information about the structure of a molecule through the chemical shift which is sensitive to local

physicochemical environment, through spin-spin coupling constraints which is sensitive to dihedral angles, and through relaxation (NOE) which is sensitive to the positions of nearby spins. However, before any of this information can be put to use in determining the structure of a molecule, it must first be determined which resonances come from which spins. The process of associating specific spins in the molecule with specific resonances is called *sequence-specific assignment of resonances*, on which this thesis will mainly focus.

Sequence-specific resonance assignment is essential in: (1) the structure determination of proteins, (2) intermolecular interactions, and (3) protein dynamics.

Firstly, consider the determination of protein structure from NMR data. Protein chemical shifts may be used in at least four different ways in structural analysis including: (i) secondary structure mapping, (ii) generating structural constraints, (iii) three-dimensional structure generation, and (iv) three-dimensional structure refinement. Perhaps the most well-known application of chemical shift in biomolecular NMR is in the area of secondary structure identification and quantification (Szilagyi and Jardetzky 1989; Pastore and Saudek 1990; Spera and Bax 1991; Wishart, Sykes et al. 1992; Le and Oldfield 1994; Luginbuhl, Szyperski et al. 1995; Wishart and Nip 1998; Iwadate, Asakura et al. 1999; Hung and Samudrala 2003; Eghbalnia, Wang et al. 2005; Wang, Chen et al. 2007). It has been confirmed that  $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ , and  $^{13}\text{C}_\text{O}$  NMR chemical shifts for all 20 amino acids are sensitive to their secondary structure. The assigned chemical shifts provide more reliable information about the secondary structure of the protein than any other computational prediction

methods based on sequence similarity. Chemical shifts can also play a useful role in delineating three-dimensional structure of proteins. The structural information mainly derives from NOE cross peaks. A NOE peak correlating two hydrogen atoms is observed if these hydrogens are located at a distance shorter than 5 Å from each other. Combined with resonance assignment these distance constraints can be attributed to specific sites along the protein chain and therefore the three dimensional structure can be initialized. In addition, using other constraints derived from chemical shift assignment (e.g., dihedral angles) along with the constraints from NOE correlations, the protein tertiary structure can be formed and further refined.

The second application of sequence-specific resonance assignment is to study protein-protein interactions. Analysis of intermolecular interactions by solving the structures of protein-protein complexes using conventional NMR methodology presents a considerable technical challenge and is highly time-consuming. If the structures of the free proteins are already known at high resolution, and conformational changes upon forming complexes are either minimal or localized, it is possible to use conjoined rigid body/torsion angle dynamics (Clore and Bewley 2002) to solve the structure of the complex based solely on intermolecular inter-proton distance restraints, derived from isotope-edited NOE measurements. Nevertheless, unambiguous assignment of intermolecular NOEs is still difficult and time-consuming, particularly for large complexes. In contrast, the mapping of interaction surfaces by  $^1\text{H}_\text{N}/^{15}\text{N}$  chemical shift perturbation (Zuiderweg 2002) is a simple, rapid and most widely used NMR method to study protein interactions. In a nutshell, the  $^{15}\text{N}$ - $^1\text{H}$  or/and  $^{13}\text{C}$ -

$^1\text{H}$  HSQC spectrum of one protein is monitored when an unlabeled interaction partner is titrated in, and the perturbations of chemical shifts are recorded. The interaction causes environmental changes on the protein interfaces and, hence, affects the chemical shifts of the nuclei in this area. It is easy and straightforward to correlate these value-changed chemical shifts with specific residues according to sequence-specific resonance assignment and therefore, the interaction regions derived from the perturbation of chemical shifts can be identified.

NMR spectroscopy can also be used to monitor the dynamic behaviour of a protein at a multitude of specific sites, which is associated with the specific functions of the protein. Once again, resonance assignment is a prerequisite to determine the residues implicated in the analysis of structural dynamic from nuclear spin relaxation.

### **1.3.2 General strategy for sequence-specific resonance assignment**

The first structures of biological macromolecules determined by NMR spectroscopy were solely based on [ $^1\text{H},^1\text{H}$ ]-proton correlation experiments (Williamson, Havel et al. 1985), from which protein structures up to a size of 10 kDa can be obtained without any isotopic enrichment using  $^1\text{H}$  homonuclear assignment strategy. However, for larger proteins the increased spectral overlap and linewidths make structure determination increasingly difficult. The introduction of isotopic labelling and triple-resonance experiments have extended the molecular weight range to approximately 20 kDa by reducing resonance overlap through separation of the peaks along one or more heteronuclear

frequency dimensions. Since then, even for proteins smaller than 10 kDa the isotopic labelling and heteronuclear assignment strategy are applied to accelerate the structure determination process. Nearly all NMR structure determinations of proteins recombinantly expressed are nowadays carried out with isotopic labelling and heteronuclear assignment strategy. The only exceptions are proteins that are isolated from natural sources (snake and scorpion toxins, pheromones etc.). Although the  $^1\text{H}$  homonuclear assignment strategy is insufficient in modern protein structure elucidation, it gave us an idea on developing new strategy for sequence-specific resonance assignment of large proteins (Chapter 4).

In the following sections, a brief introduction of  $^1\text{H}$  homonuclear assignment strategy and a detailed depiction of the currently most widely used heteronuclear sequence-specific resonance assignment strategy will be given.

### 1.3.2.1 $^1\text{H}$ homonuclear assignment strategy

The principal process of determining  $^1\text{H}$  resonance assignments is developed by Wüthrich and co-workers (Wüthrich 1986). This strategy is based upon the following critical observation: with few exceptions, correlations resulting from  $^1\text{H}$ - $^1\text{H}$  scalar couplings normally are only observed between  $^1\text{H}$  nuclei separated by two or three bonds in proteins. Cross-peaks in  $^1\text{H}$  homonuclear correlation NMR spectra occur between  $^1\text{H}$  spins within the same amino acid residue or *spin system*. 2D experiments, such as COSY, MQF-COSY, MQ spectroscopy, and TOCSY are used to identify resonance positions within each amino acid spin system, and the NOESY experiment is used to sequentially connect the amino acid spin systems.

Initially,  $^1\text{H}$  resonances are categorized into backbone amide  $^1\text{H}_\text{N}$ , aromatic  $^1\text{H}$ , backbone  $^1\text{H}_\alpha$ , aliphatic side chain methine and methylene  $^1\text{H}$ , and methyl  $^1\text{H}$ , on the basis of their chemical shifts. The first stage of analysis makes use of scalar couplings to establish sets of  $^1\text{H}_\text{N}$ ,  $^1\text{H}_\alpha$ , and aliphatic side-chain resonances that belong to the same amino acid residue spin system. A protein of  $N$  residues has  $N$  distinct backbone-based spin systems. Each spin system is assigned to an amino acid type (or one of several possible types) based on the coupling topology and resonance chemical shifts.

In the second stage of the assignment process, spin systems are connected using through-space dipolar coupling (NOE) interactions to generate dipeptide segments. Statistical analysis of the proton positions inferred from X-ray-crystal structures of proteins has shown that the majority of short interproton distances between  $^1\text{H}_\text{N}$ ,  $^1\text{H}_\alpha$ , and  $^1\text{H}_\beta$  are between residues adjacent in the primary sequence (Billeter, Braun et al. 1982). Thus, identification of intense NOEs from  $^1\text{H}_\text{N}$ ,  $^1\text{H}_\alpha$ , and/or  $^1\text{H}_\beta$  of one spin system to  $^1\text{H}_\text{N}$  of a second spin system suggests that the two spin systems are adjacent in the primary sequence with the first spin system nearer to the N-terminus of the protein. As more dipeptide segments are generated, one or more fragments will eventually be established and uniquely mapped to protein sequence.

If the spin system types are well characterized (i.e. the majority of side-chain resonance positions have been identified), then a fragment consisting of four or five spin systems usually can be placed on the protein sequence and to achieve sequence-specific assignment. The ambiguity in the assignment process can be reduced by the identification of other sequential NOEs and the match of

sequential ordering between fragments and protein sequence. The assignments encompass all spin systems, and self-consistency is the best measure of the validity of the results.

### 1.3.2.2 Triple-resonance assignment strategy

Thanks to the introduction of protein isotopic labelling, sets of nuclei other than  $^1\text{H}$  that can be detected by NMR are available, heteronuclear NMR experiments become the dominating method in protein NMR. Since the late 1980s, a large number of 3D or 4D triple resonance NMR experiments have been developed and used for protein sequence-specific resonance assignment. Other than extract the inter-residue correlations from NOE-based experiments where through-space dipolar couplings contribute to the observed cross-peaks, the triple-resonance experiments offer an alternative strategy which establish the inter-residue correlations via the relatively uniform and well-resolved heteronuclear one-bond and two-bond couplings, without any prior knowledge of spin system types. By properly combining several triple resonance NMR experiments, it is possible to establish a sequential walk from one residue to the next. Potential errors that arise from misassignment of sequential and long-range connectivities in the NOE-based procedures are avoided because assignments are based solely on predictable through-bond scalar correlations.

The strategy that currently most widely used for obtaining complete and unambiguous sequence-specific assignment obtains backbone assignments from a pair of triple-resonance experiments CBCANH and CBCA(CO)NH, and obtains side-chain assignments from a set of TOCSY-based experiments,

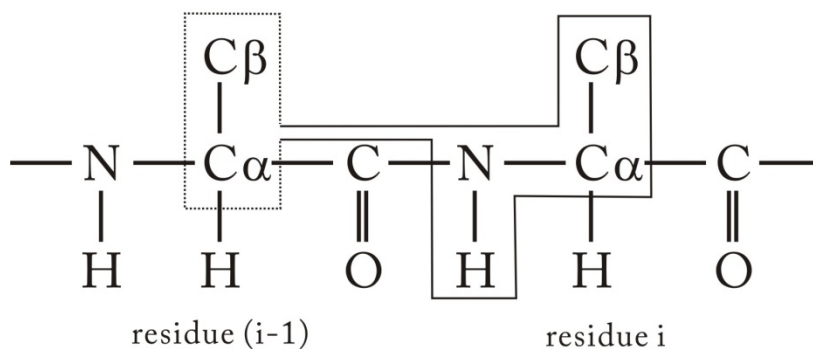
including a pair of triple-resonance experiments  $H(CC-CO)NH-TOCSY$ ,  $(H)C(C-CO)NH-TOCSY$  and a double-resonance experiment  $HCCH-TOCSY$ .

**Table 1.1 Heteronuclear Experiments Used for protein sequence-specific resonance assignment**

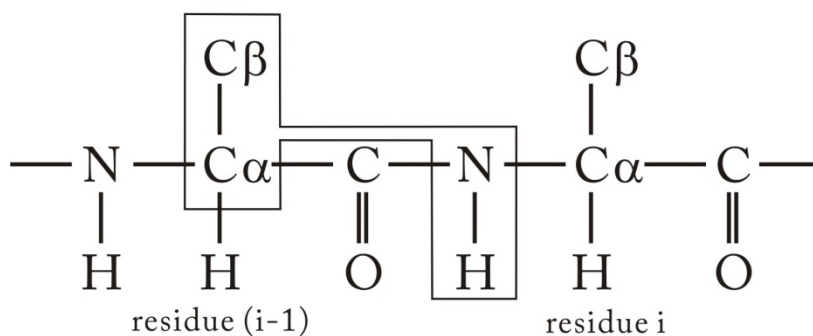
| Experiment           | Correlations observed  | Magnetization transfer |
|----------------------|--|------------------------|
| CBCANH               | $C\beta_{i-1}/C\alpha_{i-1} - N_i - H_i$<br>$C\beta_i/C\alpha_i - N_i - H_i$ |                        |
| CBCA(CO)NH           | $C\beta_{i-1}/C\alpha_{i-1} - N_i - H_i$                                     |                        |
| $(H)C(C-CO)NH-TOCSY$ | $C_{sc_{i-1}}/C\alpha_{i-1} - N_i - H_i$                                     |                        |
| $H(CC-CO)NH-TOCSY$   | $H_{sc_{i-1}}/H\alpha_{i-1} - N_i - H_i$                                     |                        |
| HCCH-TOCSY           | $H_{sc}/H\alpha_i - H_{sc}/H\alpha_i - C_{sc}/C\alpha_i$                     |                        |



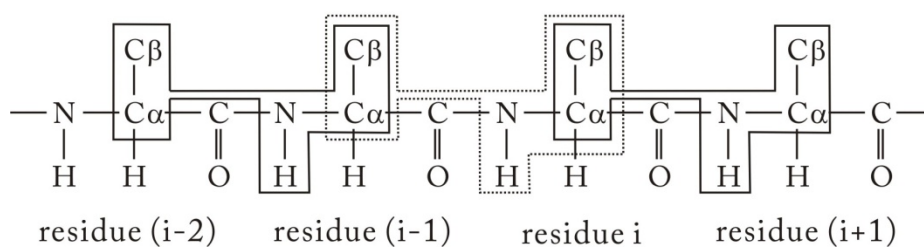
## CBCANH



## CBCA(CO)NH



## CBCANH + CBCA(CO)NH



**Figure 1.2 Schematic depiction of backbone assignment using the CBCANH and CBCA(CO)NH spectra.**

Each  $^1\text{H}_\text{N}$ - $^{15}\text{N}$  pair is correlated with the  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  of the previous residue using the CBCA(CO)NH and the  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  of its own residue using the CBCANH. To trace along the backbone, one searches for pairs of amides in which the  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  chemical shifts of one in the CBCANH are identical to the  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  chemical shifts observed from the second in the CBCA(CO)NH. Two such amides belong to adjacent residues in the protein. Because of the requirement for amide protons, the correlations are interrupted at proline residues.

### ***Backbone Assignment***

A pair of CBCANH and CBCA(CO)NH (Table 1.1) data sets give information about intra-residue correlations and the corresponding inter-residue correlations in the form of a three-dimensional spectrum with the amide proton ( $^1\text{H}_\text{N}$ ) chemical shift on one axis, the amide nitrogen ( $^{15}\text{N}$ ) on another, and the carbon chemical shift ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) on the third axis. By analyzing the data sets together, the  $^1\text{H}_\text{N}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  spins along the entire backbone (except for the prolines) can be correlated: the CBCANH correlates each  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  with the intra-residue  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$ , while the CBCA(CO)NH correlates each  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  with the  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  of the previous residue (Figure 1.2). The  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  chemical shifts can be used to identify amino acid types and thus to map segments of connected spin systems onto the sequence of the protein.

### ***Side-Chain Assignment***

Determination of  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  chemical shifts during backbone assignment provides a good starting point for side-chain assignment. The H(CC-CO)NH-TOCSY and (H)C(C-CO)NH-TOCSY experiments (Table 1.1) correlate proton or carbon resonances within a side chain to one another and to the amide resonances in the backbone. They are three-dimensional experiments in which the axes are the chemical shifts of  $^1\text{H}_\text{N}$ ,  $^{15}\text{N}$ , and side-chain protons or carbon spins. The amide correlations in these experiments make the spectra easy to interpret, since they contain much less ambiguity than an HCCH-TOCSY spectrum (Table 1.1), for example. Ambiguities may remain as to which proton is attached to which carbon, these can usually be resolved by the HCCH-TOCSY spectrum, since it correlates  $^1\text{H}$  resonances with their attached  $^{13}\text{C}$  resonances

using the well-resolved and strong one-bond  $^1\text{H}$ - $^{13}\text{C}$  and  $^{13}\text{C}$ - $^{13}\text{C}$  coupling to transfer magnetization along the side-chain.

### 1.3.3 Limitations of the conventional strategies

The conventional sequence-specific resonance assignment strategy can be applied to proteins with molecular weight up to  $\sim 25$  kDa, but for proteins larger than 30 kDa, the set of experiments used in the conventional strategy may not work because of several problems.

The first problem is spectral crowding. For larger proteins, some spectra (e.g. HCCH-TOCSY) could be severely overlapped due to the overwhelming number of resonances. In overlapped spectra, many of the resonances cannot be assigned unambiguously to individual nuclei. Indeed, even if a full sequential assignment is made, the identification of the interresidue NOEs usually proves too ambiguous to provide a sufficient number of restraints for a high resolution structure calculation.

The second problem is that large proteins tumble slower in solution, resulting in rapid transverse relaxation. During magnetization transfer periods in an NMR experiment, the signal decays away rapidly owing to the relaxation of the magnetization. Transverse relaxation is the main source for this signal loss. For large proteins the signal decays much faster, which causes poor sensitivity and line broadening of the spectrum (Figure 1.3 a vs. b). Especially for COSY and TOCSY experiments, the inefficiency of magnetization transfer from one spin to another spin caused greatly decreased sensitivity. Significant signal loss occurs during the relatively long mixing times required for magnetization transfer, as a

result of the reduced characteristic relaxation time when the correlation time increases.

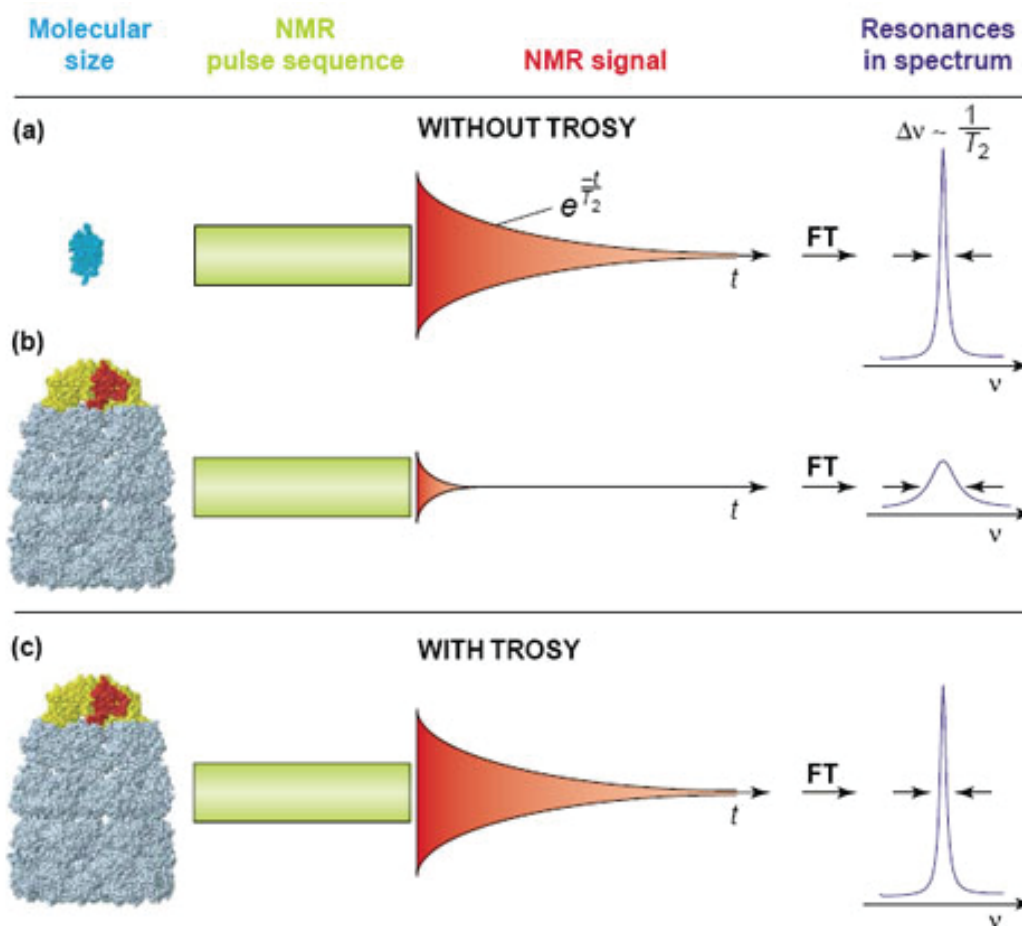
The third problem is degeneracy in chemical shift. Using heteronuclear backbone assignment as an example, in which the CBCA(CO)NH experiment correlates the  $^1\text{H}_\text{N}$ - $^{15}\text{N}$  pair of residue  $i$  with  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  shift of residue  $i-1$ . Thus, if all the  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  shifts are unique, the residues of the whole protein can be linked in a sequential order. If, however, there are degenerate  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  shifts, e.g. two residues have similar  $^{13}\text{C}_\alpha / ^{13}\text{C}_\beta$  shifts, the backbone assignment will fail at that point. As protein size increases (larger residue number) and spectrum resolution decreases, degeneracy like this becomes more and more common.

The desire to overcome these limitations for sequence-specific resonance assignment of large protein is the driving force behind the development of new NMR methodology, in sample preparation techniques and assignment strategies, which will be reviewed in the following section.

## 1.4 Previous works on large proteins

Significant advances in NMR technology over the past two decades have made it well suited for detailed analyses of macromolecular structure, dynamics and interactions of smaller proteins. (Foster, McElroy et al. 2007) With the availability of uniform  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeling and triple resonance experiments, it is almost a routine task to assign backbone and side-chain resonances for proteins with molecular weight below 25 kDa. Work on these proteins has been very fruitful and allowed us to learn much about structure-function relationships. This assignment strategy, however, is inherently limited, as majority of the

macromolecular complexes of biochemical interest are significantly larger than 25 kDa.



**Figure 1.3 Effects of protein size on NMR signals.**

(a) The NMR signal from small proteins has long transverse relaxation time ( $T_2$ ). This translates into narrow linewidth ( $\Delta\nu$ ) on the spectrum after Fourier transformation (FT). (b) By contrast, the signal from large proteins relaxes faster (shorter  $T_2$ ), resulting in weak signal detected after the pulse sequence and broad lines on the spectrum. (c) TROSY substantially reduces the effective relaxation of the detected signal, leading to improved spectral resolution and sensitivity for large proteins.

*Adopted from Ref (Fernandez and Wider 2003)*

NMR studies of large proteins are complicated by the increased spectral crowding, transverse relaxation rate and chemical shift degeneration. Several methodological advances have been developed to overcome these problems.

### 1.4.1 Reducing protein transverse relaxation rate

It had long been realized that substituting protons for deuterons would reduce the relaxation rates of the attached nuclei, leading to increased spectral resolution and significant gain in sensitivity. (Gardner and Kay 1998) Deuterated proteins can be produced in cultures with deuterated media (typically,  $^2\text{H}_2\text{O}$  and  $^2\text{H}$ -glucose provide the hydrogen and carbon atoms, and  $^{15}\text{N}$ -ammonium serves as the nitrogen source). Then, upon transfer of the protein into protonated solvents (i.e.,  $^1\text{H}_2\text{O}$ ), the exchangeable protons on the amides will be observed in a  $^1\text{H}$ - $^{15}\text{N}$  heteronuclear correlation spectrum, without being broadened by spin-spin interactions with carbon-bound protons.

Nevertheless, deuteration alone does not allow the application of protein NMR beyond 50 kDa. The major breakthrough in extending the size limit comes with the introduction of TROSY (transverse relaxation-optimized spectroscopy). (Pervushin, Riek et al. 1997) TROSY exploits the interference of two different relaxation mechanisms that affect the linewidths of certain NMR signals with opposite sign. In the optimal case, the two relaxation mechanisms cancel each other out and a very narrow line is observed in the NMR spectrum (Figure 1.3 c). For large proteins, TROSY works especially best at high field strength (700 to 900 MHz) with deuterated samples (Fernandez and Wider 2003). TROSY modules have been implemented in many of the triple resonance experiments, which allow the assignment of backbone and  $\text{C}_\beta$  resonances for proteins up to

100 kDa. (Pervushin, Riek et al. 1997; Yang and Kay 1999; Tugarinov, Muhandiram et al. 2002)

Unfortunately, the increase in size limit does come at a cost. The removal of aliphatic and aromatic protons by deuteration considerably reduces the number of NOEs which would otherwise provide valuable distance constraints for structure calculation. Although the global folds of a protein can be determined using only backbone NOEs and residual dipolar couplings in partially ordered medium (Giesen, Homans et al. 2003), such structural models always suffer from low resolution and the arrangements of many side chains cannot be defined precisely. Moreover, the preparation of such deuterated samples is always costly and time-consuming. Additionally, some proteins need to be unfolded in order to accelerate the exchange of amide  $^2\text{H}$  to  $^1\text{H}$ , and then subsequently refolded. This unfolding-refolding process is not trivial for most proteins.

Other approaches may also reduce the transverse relaxation rates of large proteins. One simple, albeit limited, solution is to increase the overall molecular tumbling rate by recording NMR spectra at elevated temperatures, which is only applicable for thermostable proteins (Hua, Dementieva et al. 2001; McElroy, Manfredo et al. 2002; Boomershine, McElroy et al. 2003). Another ingenious approach involves encapsulating hydrated proteins in low-viscosity solvents (Wand, Ehrhardt et al. 1998); while promising, this approach is not widely used, as the encapsulation process is technically challenging and system dependent.

### **1.4.2 Reducing protein spectral crowding and chemical shift degeneration**

An emerging approach to alleviate the problems of spectral crowding and chemical shift degeneration is to make use of protein splicing methods that allow for “segmental labelling” of specific regions of a protein with NMR active isotopes (Cowburn, Shekhtman et al. 2004). Because signals are not observed from the unlabeled segments of the molecules, this approach simplifies the NMR spectra without the loss of context that comes from study of an isolated domain. By labelling a different segment each time in a series of experiments, the structure of the entire protein can be studied.

In the case of multimeric protein complexes, spectral overlap can also be reduced by the use of subunit-selective isotope labelling. If a multimeric protein complex can be reconstituted from isotope-labelled and unlabelled binding partners, only the labelled protein is observed in the NMR spectrum, thus reducing the signal overlap.

For perdeuterated proteins, alternative labelling protocols involve the use of metabolic precursors allow for selective protonation and monitoring of specific groups (e.g., methyls) in concert with the backbone amides. The protonated methyl groups can be assigned with TOCSY-based experiments or the TROSY versions of these experiments, to provide many long range distance constraints as methyl groups are usually localized in the hydrophobic cores of proteins connecting secondary structural elements.



However, all specific isotopic labelling techniques are very costly and time-consuming, and may not be suitable for every protein. Despite several successful applications of these labelling strategy to large proteins (Tugarinov, Choy et al. 2005; Kainosho, Torizawa et al. 2006), the extremely high cost of the samples impedes the application of these techniques.

## 1.5 Research objectives

At present, ~15% of protein structures deposited in the protein data bank is determined by NMR, but only ~1% of the NMR structures are for proteins larger than 25 kDa (Kainosho, Torizawa et al. 2006). Additionally, most of the large proteins only have crude global folds based on backbone assignments and a few side chain assignments. The preparation of deuterated or/and specific isotopic labelled protein samples is often challenging and places a bottleneck on the NMR study of large proteins.

Because only  $^1\text{H}$  spins produce distance restraints that determine the quality of solution structures, the simplest and cheapest samples for obtaining high resolution structures are non-deuterated proteins. Unfortunately, most triple-resonance experiments for establishing resonance assignments do not work for uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled large proteins without deuteration, except for NOESY and multiple-quantum  $^{13}\text{C}$  total correlation spectroscopy (MQ-CCH-TOCSY) experiments (Chapter 2, 3 and 4).

Therefore, the objective of this thesis is to focus on developing new NMR techniques and computational means which can be used to obtain partial or

complete sequence specific assignments, or even high-resolution structures, of larger proteins, with:

- 1) uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled samples without the use of deuteration and
- 2) protein-size-insensitive NMR spectra (e.g. NOESY and MQ-CCH-TOCSY).

## **Chapter 2:**

# **Sequence-specific assignments of methyl groups in large proteins**

- 2.1 Introduction
- 2.2 General strategy for sequence-specific assignments of methyl groups
- 2.3 Discussion
- 2.4 Conclusion
- 2.5 Materials and methods

## Chapter 2:

# Sequence-specific assignments of methyl groups in large proteins

## 2.1 Introduction

Methyl groups are of particular interest in NMR studies of proteins since they occur frequently in the hydrophobic cores of these molecules (Janin, Miller et al. 1988) and thus are often sensitive reporters of structure and dynamics.

Thus, sequence-specific assignments of methyl resonances are important in applications that involve large proteins because of their favorable properties that facilitate the recording of NMR spectra with high sensitivity and resolution. First, the threefold degeneracy of methyl protons in  $^{13}\text{CH}_3$  isotopomers effectively increases the concentration of each group significantly beyond that for, say, backbone amides. Second, because methyl groups are localized at the peripheries of side chains, many tend to be dynamic (Nicholson, Kay et al. 1992); this leads to slower relaxation that can be exploited in studies of large systems. Third, distances between proximal methyl groups, established on the basis of NOEs, often connect regions of the molecule that are far removed in primary structure (Metzler, Leiting et al. 1996; Rosen, Gardner et al. 1996; Smith, Ito et al. 1996; Mueller, Choy et al. 2000). In addition, these moieties serve as probes in investigations of protein–ligand interactions (Hajduk, Augeri et al. 2000; Gross, Gelev et al. 2003), fast and slow timescale side-chain dynamics (Henry, Weiner et al. 1986; Muhandiram, Yamazaki et al. 1995; Ishima, Louis et al. 1999; Lee and Wand 2001; Mulder, Mittermaier et al. 2001; Skrynnikov, Mulder et al.

2001), dynamics of protein folding (Henry, Weiner et al. 1986; Muhandiram, Yamazaki et al. 1995; Ishima, Louis et al. 1999; Skrynnikov, Mulder et al. 2001; Choy, Shortle et al. 2003), and in the detection of proteins and complexes in in-cell NMR experiments (Serber, Straub et al. 2004).

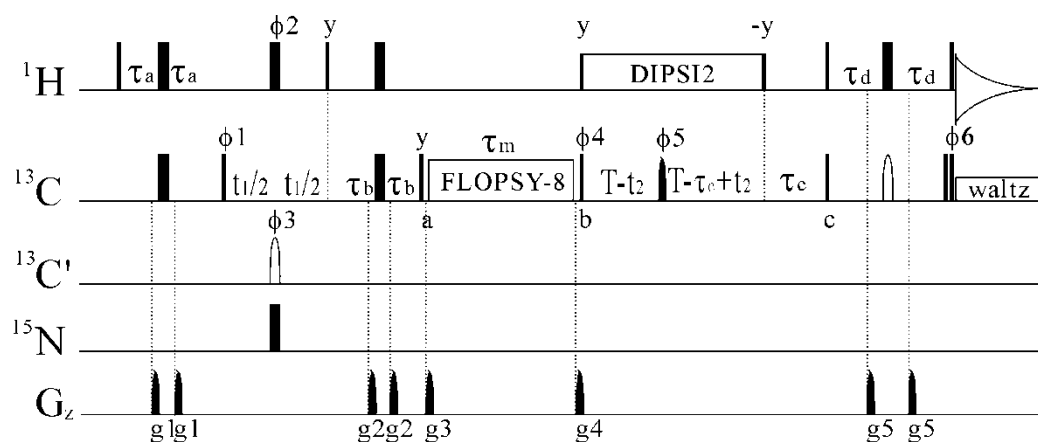
I propose here a novel 3D multiple-quantum (MQ) (H)CC<sub>m</sub>H<sub>m</sub>- TOCSY experiment for assignments of <sup>1</sup>H and <sup>13</sup>C resonances of methyl groups using uniformly <sup>13</sup>C-labeled samples. The new 3D MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment correlates chemical shifts of aliphatic carbon nuclei of amino acid side chains with those of the methyl <sup>13</sup>C<sub>m</sub> and <sup>1</sup>H<sub>m</sub> nuclei in the same residue in the protein sequence. On the basis of prior assignments of <sup>13</sup>C<sub>α</sub> and <sup>13</sup>C<sub>β</sub>, sequence-specific assignment of methyl resonances can be obtained.

## 2.2 General strategy for sequence-specific assignments of methyl groups

Figure 2.1 shows the pulse sequence of the 3D MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment. The magnetization transfer is shown schematically as follows:



The decay rates of the MQ coherences (H<sub>x</sub>C<sub>y</sub>) are normally significantly smaller than those of the single-quantum (SQ) coherences (Grzesiek, Kuboniwa et al. 1995; Shang, Swapna et al. 1997; Gschwind, Gemmecker et al. 1998). Thus, this experiment is more sensitive than its SQ version.



**Figure 2.1** Pulse sequence for the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment.

All narrow (wide) rectangular pulses have flip angles of  $90^\circ$  ( $180^\circ$ ). The  $^1\text{H}$  carrier is set at 4.7 ppm while the  $^{13}\text{C}$  carrier is centered at 41 ppm, until immediately prior to the  $^{13}\text{C}$  pulse of phase  $\Phi_4$  at which time the carrier is jumped to 17 ppm. All  $^1\text{H}$  pulses are applied with a 23 kHz field;  $^1\text{H}$  DIPS12-decoupling elements make use of a 6.25 kHz field. All  $^{13}\text{C}$  rectangular pulses employ a 16.8 kHz field, and the  $^{13}\text{C}$  shaped pulses have REBURP profiles. The first  $180^\circ$  shaped (filled)  $^{13}\text{C}$  pulse has a duration of  $400\ \mu\text{s}$  and is phase modulated by 24 ppm, while the second one (empty) has a duration of 1.5 ms. The  $^{13}\text{C}$  spin-lock field strength for FLOPSY is 7 kHz. A decoupling power of 1.25 kHz is used during acquisition. The  $180^\circ$  pulse on C' has a SEDUCE profile with a duration of  $250\ \mu\text{s}$  (center of excitation 176 ppm). The delays used are:  $\tau_a = 1.4\ \text{ms}$ ,  $\tau_b = 1.1\ \text{ms}$ ,  $\tau_c = 1.5\ \text{ms}$ ,  $\tau_d = 1.6\ \text{ms}$ ,  $\tau_m = 17\ \text{ms}$ , which is suitable for proteins with overall correlation times ranging from 20 to 30 ns as shown by numerical simulations;  $T = 14\ \text{ms}$ . The phase cycling employed is:  $\Phi_1 = 4(x), 4(-x)$ ;  $\Phi_2 = x, y, -x, -y$ ;  $\Phi_3 = 2(x), 2(-x)$ ;  $\Phi_4 = y$ ;  $\Phi_5 = 2(x), 2(y), 2(-x), 2(-y)$ ;  $\Phi_6 = 4(x), 4(-x)$ ;  $\text{rec} = x, -x, -x, x, -x, x, x, -x$ . The duration and strengths of the sine-shaped gradients are:  $g1 = (0.5\ \text{ms}, 20\ \text{G/cm})$ ;  $g2 = (0.3\ \text{ms}, 25\ \text{G/cm})$ ;  $g3 = (1\ \text{ms}, 25\ \text{G/cm})$ ;  $g4 = (1\ \text{ms}, 20\ \text{G/cm})$ ;  $g5 = (0.5\ \text{ms}, 20\ \text{G/cm})$ . Quadrature detection in F1 and F2 are achieved by States-TPPI of  $\Phi_1$  and  $\Phi_4$ , respectively.

During the first part of the constant-time period (point *b* to point *c*), proton decoupling is applied to maintain the slow decays of the transverse  $^{13}\text{C}_m$  magnetizations, enhancing experimental sensitivity. In the absence of proton decoupling, proton spin flip-flop rates dominate decays of the two inner

components of  $^{13}\text{C}_m$  quartets for fully protonated proteins and significantly elevate the apparent  $^{13}\text{C}_m$  decays. (Liu, Zheng et al. 2003)

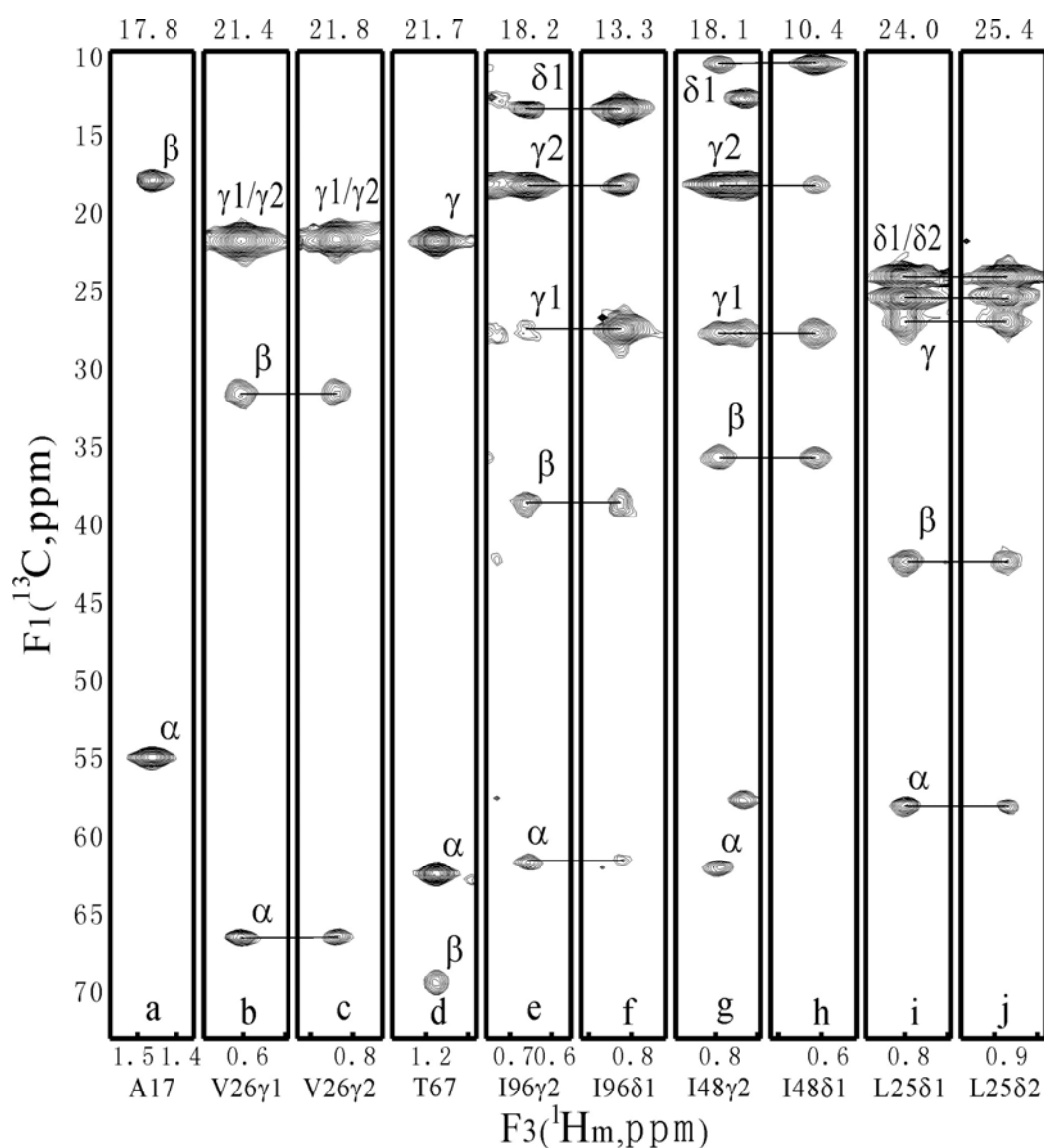
We have applied the 3D MQ-(H)CC $_m$ H $_m$ -TOCSY experiment to a  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled sample of acyl carrier protein synthase (AcpS) which consists of three subunits with a total molecular weight of 42 kDa. (Liu, Black et al. 2002)

As established from  $^{15}\text{N}$  relaxation data, AcpS has an overall rotational time of 26 ns at 25 °C (equivalent to a protein on the order of 60 kDa at 37 °C). Despite its large overall correlation time, all of the aliphatic  $^{13}\text{C}$  resonances were observed for most residues having methyl groups in the MQ-(H)CC $_m$ H $_m$ -TOCSY experiment.

Figure 2.2 shows a number of F1-F3 slices taken from a 3D MQ-(H)CC $_m$ H $_m$ -TOCSY spectrum. If the ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts of residues containing methyl groups are not degenerate with each other in a given protein, sequence-specific assignment of methyl resonances can be obtained from the MQ-(H)CC $_m$ H $_m$ -TOCSY spectrum on the basis of the assignment of ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ).

For example, according to the ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts and spectral pattern shown in slice e and the prior sequential assignment that was obtained using a uniformly  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled protein (Liu, Black et al. 2002), signals in this slice were assigned to attribute to I96 $\gamma_2$ .  $^{13}\text{C}_{\gamma_1}$  and  $^{13}\text{C}_{\delta_1}$  from the same slice can then be assigned. The assignment of  $^{13}\text{C}_{\delta_1}$  can be further confirmed from the slice taken at F2 frequency of  $^{13}\text{C}_{\delta_1}$  (slice f). Sometimes,  $^{13}\text{C}_\alpha$  or  $^{13}\text{C}_\beta$  resonances are not observable in the slice taken from  $^{13}\text{C}_{\delta_1}$  of Ile, but observable in the slice of  $^{13}\text{C}_{\gamma_2}$ . In this case, assignment of  $^{13}\text{C}_{\delta_1}$  and  $^1\text{H}_{\delta_1}$  can be done on the basis of the

assignment of  $^{13}\text{C}_{\gamma 2}$  and the matches of  $^{13}\text{C}_{\gamma 1}$ ,  $^{13}\text{C}_{\delta 1}$  and  $^{13}\text{C}_{\gamma 2}$  resonances between the two slices as shown in slices g and h. Similarly, the spectral information of two  $^{13}\text{C}_{\delta}$  in Leu or two  $^{13}\text{C}_{\gamma}$  in Val residues can be complementary to each other.



**Figure 2.2** Representative slices from the MQ-(H)CC $_m$ H $_m$ -TOCSY spectrum used for methyl assignments.

Each F1( $^{13}\text{C}$ )-F3( $^1\text{H}_m$ ) slice is labeled with the identity of the methyl-containing residue, and the F2( $^{13}\text{C}_m$ ) frequency in ppm is indicated at the top of each slice.





## 2.3 Discussion

For large monomeric proteins, one would expect that degeneracy in ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts may hinder the application of the method proposed here. In practice, the relatively good dispersion of ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts among the same types of residue allows one to assign most of the methyl resonances, according to our survey as shown in Table 2.1. Although the ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts of L18 and L25 are degenerate within a threshold of 0.3 ppm, L18 and L25 can be distinguished from each other from the differences of their ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts (0.15 ppm). Obviously, a larger number of methyl groups can be assigned with increasing spectral resolution in the F1 dimension. It is interesting to note that each type of amino acid displays a specific spectral pattern, and thus, ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shift degeneracy among different types of amino acids is not an issue for methyl assignments.

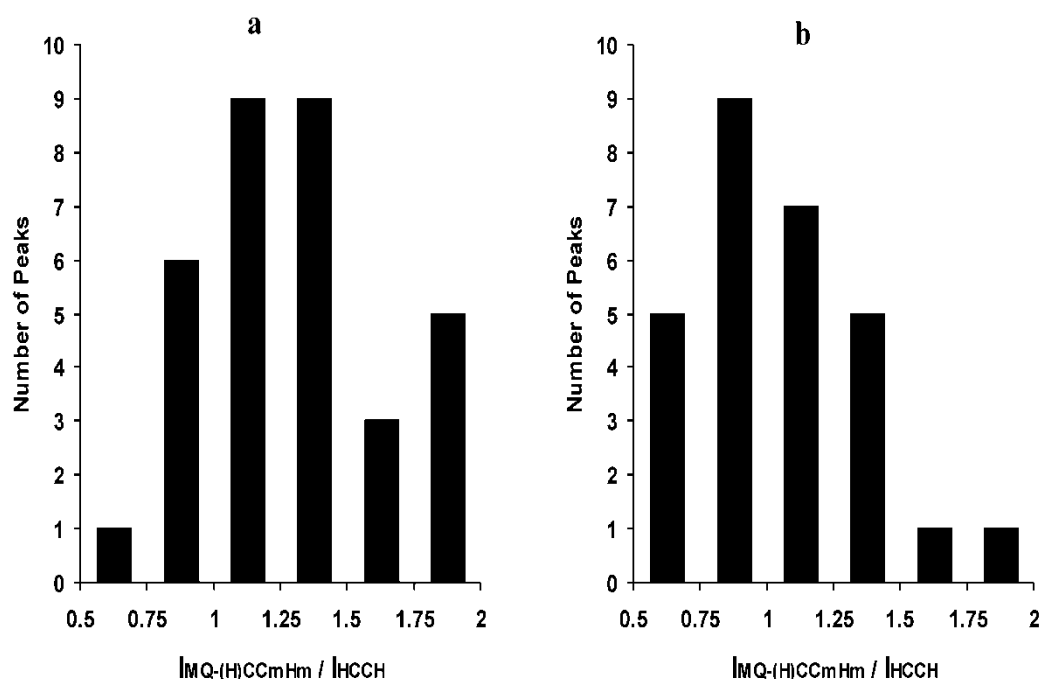
**Table 2.1 The relatively good dispersion of ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts in large monomeric proteins.**

| Protein / Residue number | ALA     | ILE     | LEU     | THR     | VAL     |
|--------------------------|---------|---------|---------|---------|---------|
| ACPS / 119               | 8 / 8*  | 9 / 11  | 6 / 9   | 5 / 5   | 7 / 7   |
| CBM28 / 204              | 16 / 20 | 8 / 8   | 12 / 14 | 7 / 12  | 12 / 12 |
| DFPase / 316             | 19 / 19 | 21 / 21 | 11 / 14 | 15 / 18 | 20 / 20 |
| MBP / 370                | 21 / 43 | 17 / 21 | 24 / 30 | 16 / 20 | 14 / 19 |
| MSG / 731                | 28 / 71 | 37 / 42 | 36 / 68 | 22 / 29 | 30 / 46 |

\*The ratio shows the number of specific types of methyl-containing residues (the difference is equal or larger than 0.3 ppm between two  $^{13}\text{C}_\alpha$  or  $^{13}\text{C}_\beta$  spins in the residues), to the total number of the same types of residue (with available ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ) chemical shifts from the BMRB database).

L9 and L51 displayed very weak  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  resonances in the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment. This is attributed to the strong scalar coupling interaction between  $^{13}\text{C}_\delta$  and  $^{13}\text{C}_\gamma$  spins in Leu residues. To achieve a high resolution in the F2 dimension, a constant-time acquisition mode was used to remove  $^{13}\text{C}$ - $^{13}\text{C}$  scalar coupling effects. For Leu, however, the strong scalar coupling interaction can be destructive to the refocus of  $^{13}\text{C}_\delta$  magnetization during the constant-time period which leads to a significant loss of sensitivity. Although a nonconstant-time version of the experiment that is similar to the HCCH-TOCSY should give better sensitivity, the resolution may not be sufficient to uniquely assign methyl resonances in the case where  $^{13}\text{C}$ - $^1\text{H}$  HSQC cross-peaks are not unique within a grid of 0.3 ppm ( $^{13}\text{C}$ )  $\times$  0.02 ppm ( $^1\text{H}$ ). Resolution in the F2 dimension is critical for methyl assignments of large and medium-sized proteins; e.g., only 62% of methyl groups in MBP are not degenerate within the 0.3 ppm  $\times$  0.02 ppm grid.

A comparison of the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment with the HCCH-TOCSY experiment (Figure 2.4) shows that the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY is more sensitive for most cross-peaks due to the gains from MQ line narrowing in the F1 dimension and from the slow decay of in-phase methyl  $^{13}\text{C}$  magnetizations during the constant-time  $t_2$  period. Most importantly, however, most methyl resonances cannot be assigned using the HCCH-TOCSY spectrum due to poor resolution in both the F1 and F2 dimensions and poor dispersion of ( $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ) and ( $^1\text{H}_\beta$ ,  $^{13}\text{C}_\beta$ ) chemical shifts. Compared to the C(CO)NH-TOCSY experiments, the experiment proposed here is much more sensitive ( $\sim 7$  times for  $^{13}\text{C}$ -labeled AcpS). The 4D HCCH-NOESY experiment (Fischer, Zeng et al. 1996) is also



**Figure 2.4** Histograms of signal-to-noise ratios of correlations from MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY and HCCH-TOCSY spectra acquired at 25 °C.

(a). Peak intensity ratios ( $I_{\text{MQ-(H)CC}_m\text{H}_m} / I_{\text{HCCH}}$ ) of ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) correlations from the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY spectrum to ( $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ,  $^1\text{H}_m$ ) correlations from the HCCH-TOCSY spectrum; (b) peak intensity ratios of ( $^{13}\text{C}_\beta$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) correlations from the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY spectrum to ( $^1\text{H}_\beta$ ,  $^{13}\text{C}_\beta$ ,  $^1\text{H}_m$ ) correlations from the HCCH-TOCSY spectrum. Only 34 ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) and 28 ( $^{13}\text{C}_\beta$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) cross-peaks were unambiguously identified in the HCCH-TOCSY data set according to the assignments of  $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_\beta$ ,  $^{13}\text{C}_m$  and  $^1\text{H}_m$ , which were obtained from the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY data set. On average, the relative sensitivity gains in ( $^{13}\text{C}_\alpha$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) versus ( $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ,  $^1\text{H}_m$ ) and ( $^{13}\text{C}_\beta$ ,  $^{13}\text{C}_m$ ,  $^1\text{H}_m$ ) versus ( $^1\text{H}_\beta$ ,  $^{13}\text{C}_\beta$ ,  $^1\text{H}_m$ ) correlations are  $1.28 \pm 0.37$  and  $1.06 \pm 0.31$ , respectively. Although about 50% of ( $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ,  $^1\text{H}_m$ ) and ( $^1\text{H}_\beta$ ,  $^{13}\text{C}_\beta$ ,  $^1\text{H}_m$ ) correlations can be identified, most methyl resonances cannot be assigned using the HCCH-TOCSY spectrum on the basis of prior assignments of  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  spins as a result of poor resolution in both the F1 and F2 dimensions and poor dispersion of ( $^1\text{H}_\alpha$ ,  $^{13}\text{C}_\alpha$ ) and ( $^1\text{H}_\beta$ ,  $^{13}\text{C}_\beta$ ) chemical shifts. Identical delays and the same FLOPSY-8 mixing scheme (mixing time of 17 ms and spin-lock field strength of 7 kHz) were used for the HCCH-TOCSY and the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiments.

less sensitive than the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment due to the inherently long NOE mixing time and the additional dimension involved. Similar to the HCCH-TOCSY spectrum, most methyl groups may not be assignable because of the low spectral resolution in the 4D and the poorly dispersed chemical shifts of (<sup>1</sup>H, <sup>13</sup>C) spin pairs in large proteins. Compared to TOCSY-based methods established previously (Gardner, Zhang et al. 1998; Hilty, Fernández et al. 2002), the method proposed here is more sensitive and efficient since methyl assignments require only a single experiment rather than two or three 3D experiments.

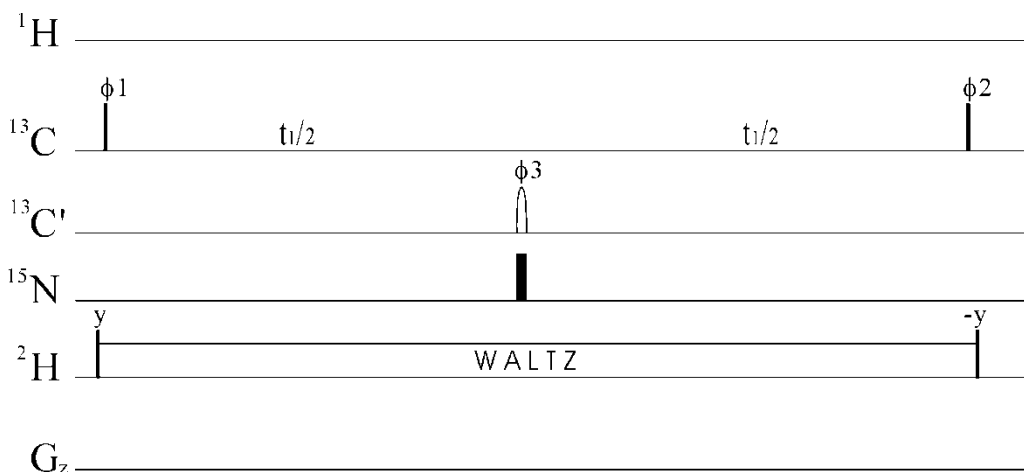
## 2.4 Conclusion

The experiment proposed here aims for only <sup>13</sup>C-labeled proteins that can be produced more easily than <sup>2</sup>H,<sup>13</sup>C,<sup>1</sup>H<sub>m</sub>-labeled proteins. If a <sup>2</sup>H,<sup>13</sup>C,<sup>1</sup>H<sub>m</sub>-labeled protein sample is available, one may use an alternative scheme (CC<sub>m</sub>H<sub>m</sub>-TOCSY) as shown in Figure 2.5. This experiment can be more sensitive than the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment when <sup>13</sup>CD *T*<sub>1</sub> can be effectively reduced by paramagnetic relaxation agents, especially for very large proteins (>50 kDa).

## 2.5 Materials and methods

We performed the experiment on the 42 kDa uniformly <sup>13</sup>C, <sup>15</sup>N-labeled AcpS homotrimer in <sup>1</sup>H<sub>2</sub>O:<sup>2</sup>H<sub>2</sub>O (95:5) solution (protein concentration 0.4 mM in the trimer, pH 7.5, 25 °C) on a Bruker Avance 500 MHz spectrometer equipped with a CryoProbe. 64(*t*<sub>1</sub>) × 70(*t*<sub>2</sub>) × 512(*t*<sub>3</sub>) complex points were collected, giving *t*<sub>1max</sub> = 7.9 ms, *t*<sub>2max</sub> = 27.4 ms and *t*<sub>3max</sub> = 64 ms. An interscan delay of 1s with 8 scans per increment was used, resulting in a total experimental

time of 47 h. The  $^{13}\text{C}$  and  $^{13}\text{C}_m$  time domains were doubled by forward-backward and mirror-image linear prediction respectively, prior to the application of cosine-squared window functions.



**Figure 2.5** Pulse scheme for the  $\text{CC}_m\text{H}_m\text{-TOCSY}$  experiment applied to  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^1\text{H}_m$ -labeled protein samples.

This pulse scheme provides the same correlations as the  $\text{MQ-(H)CC}_m\text{H}_m\text{-TOCSY}$  experiment. It is used to replace all pulses and gradients just prior to gradient pulse  $g_3$  (point a) in the pulse sequence for the  $\text{MQ-(H)CC}_m\text{H}_m\text{-TOCSY}$  experiment (Figure 2.1). Phase cycle used is:  $\Phi_1 = 4(x), 4(-x)$ ;  $\Phi_2 = x, -x$ ,  $\Phi_3 = 2(x), 2(-x)$ .

## **Chapter 3:**

# **Side-chain assignments of methyl-containing residues in large proteins**

- 3.1 Introduction
- 3.2 General strategy for side-chain assignments of methyl-containing residues
- 3.3 Conclusion
- 3.4 Materials and methods

## Chapter 3:

# Side-chain assignments of methyl-containing residues in large proteins

## 3.1 Introduction

Hemoglobin (Hb) is the iron-containing oxygen-transport metalloprotein inside the red cells of mammals and other animals (Dickerson and Geis 1983; Barrick, Lukin et al. 2004; Lukin and Ho 2004). Human normal adult hemoglobin (Hb A) is a tetramer with a molecular weight of about 65 kDa, consisting of two  $\alpha$ -chains and two  $\beta$ -chains. Each  $\alpha$ -chain contains 141 amino acids that coil into seven  $\alpha$ -helical regions and each  $\beta$ -chain contains 146 amino acids that form eight  $\alpha$ -helical regions.

The physiological function of Hb is to transport oxygen from the lungs to the tissues. Hb binds O<sub>2</sub>, CO, and NO reversibly and cooperatively, i.e., the binding of the first ligand enhances the binding of subsequent ligands. The ligand-binding affinity of Hb A is regulated by pH (the Bohr effect) and allosteric effectors, such as 2,3-bisphosphoglycerate (2,3-BPG).

The four subunits pack together through hydrophobic and hydrogen-bond interactions to form a quaternary structure. The arrangement of the subunits of Hb depends on the ligation state of the protein, and is related to the physiological function of Hb. X-ray crystallographic studies of deoxy- and liganded Hb A have found that there are at least three different quaternary structural forms (T, R, and



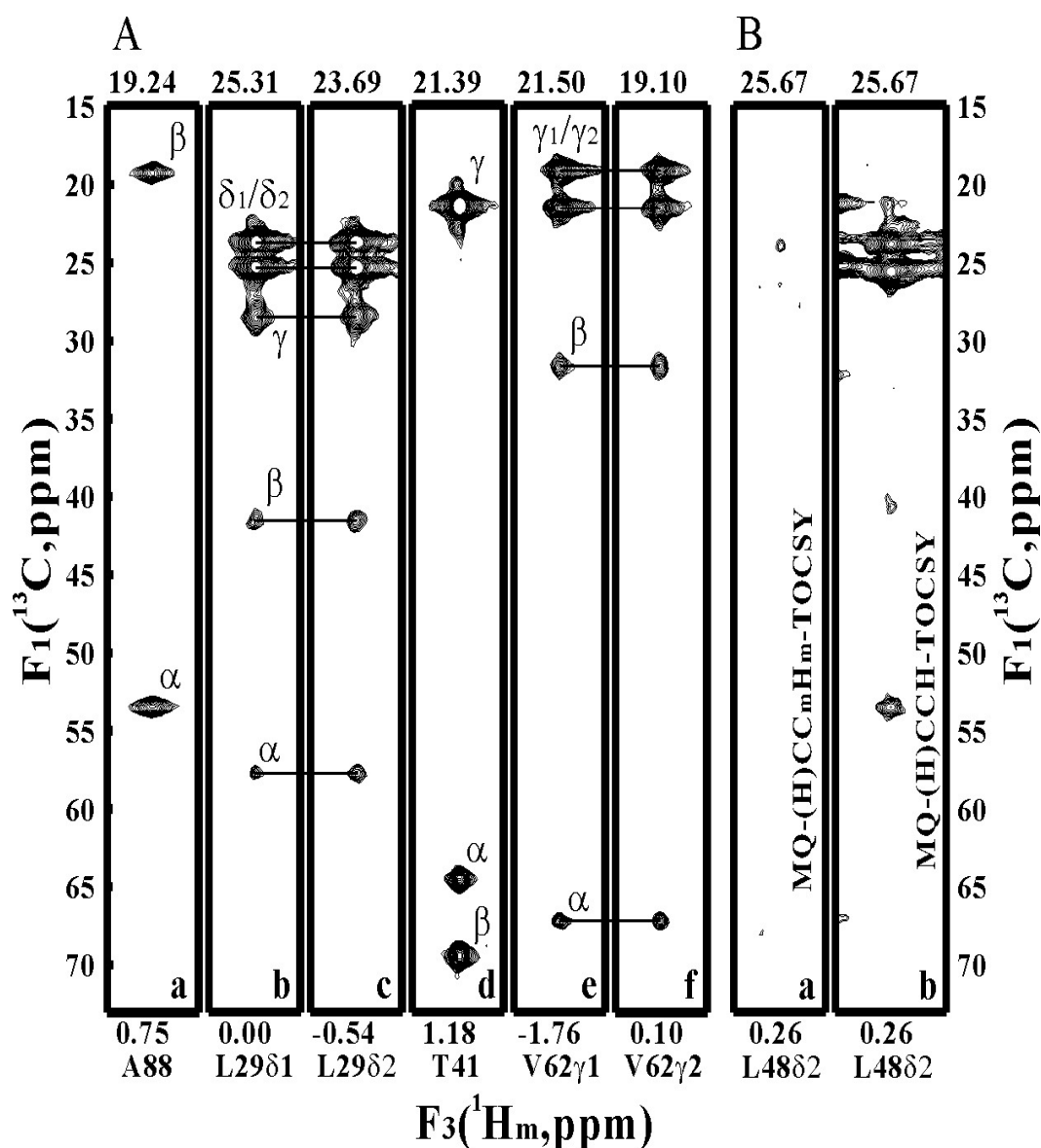
R2), in which the structures of each subunit are very similar; however, there are differences in the arrangement between  $\alpha_1\beta_1$  and  $\alpha_1\beta_2$  subunit interfaces (Silva, Rogers et al. 1992; Mueser, Rogers et al. 2000). T symbolizes the deoxy quaternary structure of crystalline deoxy-Hb A, and R and R2 represent the quaternary structures of crystalline liganded carbonmonoxy-Hb A (HbCO A) in high salt and low salt conditions, respectively.

More recently, by using NMR residual dipolar measurements on  $^{15}\text{N}$ -labeled recombinant HbCO A (rHbCO A), we have found that the solution structure of HbCO A is distinctly different from the previously determined R and R2 crystal structures and that the solution structure of HbCO A is a dynamic intermediate between R and R2 structures (Lukin, Kontaxis et al. 2003).

In spite of extensive studies on the Hb molecule, the detailed structure-function relationship in Hb is not fully understood and many aspects remain controversial. This may arise from a lack of information on the structure and dynamics of Hb under physiological conditions.

As a first step toward a detailed investigation of the structure and dynamics of Hb A in solution, side-chain resonances need to be assigned. We have developed an MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment for the assignment of methyl groups in uniformly  $^{13}\text{C}$ -labeled proteins (Chapter 2). In this study, we propose a strategy to assign side-chain  $^1\text{H}$  resonances of methyl-containing residues which applied the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment on the  $\alpha$ - and  $\beta$ -chains of rHbCO A, using only uniformly  $^{13}\text{C}$ -labeled protein. A non-constant-time MQ-

(H)CCH-TOCSY experiment is proposed to assign some Leu methyl groups that display very weak signals in the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment due to the strong coupling effect.



**Figure 3.1** Representative F1-F3 slices from the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY (A) and MQ-(H)CCH-TOCSY (B) spectra of <sup>13</sup>C-labeled  $\alpha$ -chain of rHbCO A.

Each slice is labeled with the identity of the methyl-containing residue, and the F2 (<sup>13</sup>C) frequency in ppm is indicated at the top of each slice.

## 3.2 General strategy for side-chain assignments of methyl-containing residues

We assigned methyl resonances using the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiments as described previously (Chapter 2). Several Leu methyl groups were assigned with 3D MQ-(H)CCH-TOCSY. We obtained the assignments of protons at methylene and methane positions in methyl-containing residues with 3D H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY experiments, and used MQ-(H)CCH-TOCSY experiment to confirm these assignments.

### 3.2.1 Methyl assignments

Figure 3.1A shows representative F1 – F3 slices from the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY spectrum of <sup>13</sup>C-labeled  $\alpha$ -chain of rHbCO A. Note that there is no Ile in HbCO A (Dickerson and Geis 1983).

For a given set of signals on one slice, firstly, the amino acid type of the residue contributing to these signals was determined from the spectral pattern, since each type of amino acid displays a characteristic pattern, as shown in Figure 3.1A. Secondly, the chemical shifts of (<sup>13</sup>C <sub>$\alpha$</sub> , <sup>13</sup>C <sub>$\beta$</sub> ) were measured from the same slice and compared with the corrected (<sup>13</sup>C <sub>$\alpha$</sub> , <sup>13</sup>C <sub>$\beta$</sub> ) shifts from prior sequential assignments. Lastly, if the (<sup>13</sup>C <sub>$\alpha$</sub> , <sup>13</sup>C <sub>$\beta$</sub> ) shifts uniquely matched the shifts of residue N, the set of signals was assigned to residue N. For example, slice *b* with 5 peaks at positions (<sup>13</sup>C<sub>*i*</sub>, <sup>13</sup>C<sub>*m*</sub>, <sup>1</sup>H<sub>*m*</sub>) in Figure 3.1A, where *i* =  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta_1$  and  $\delta_2$ , corresponds obviously to a Leu. The (<sup>13</sup>C <sub>$\alpha$</sub> , <sup>13</sup>C <sub>$\beta$</sub> ) shifts uniquely match the shifts of Leu29, within a threshold of 0.3 ppm. All aliphatic carbons of

Leu29 were subsequently obtained from this slice and the chemical shift of  $^1\text{H}_{\delta 1}$  for Leu29 is also measured from the F3 dimension.

With this procedure, we assigned 72 out of 92 methyl groups (excluding Met) for the  $\alpha$ -chain and 76 out of 94 for the  $\beta$ -chain. Due to the degeneracy of ( $^{13}\text{C}_{\alpha}$ ,  $^{13}\text{C}_{\beta}$ ) shifts within a threshold of 0.15 ppm, 14 and 16 methyl groups in the  $\alpha$ -chain and  $\beta$ -chain, respectively, could not be uniquely assigned. Respectively, 6 and 2 Leu methyls in the  $\alpha$ -chain and  $\beta$ -chain could not be assigned because of the absence of  $^{13}\text{C}_{\alpha}$  and  $^{13}\text{C}_{\beta}$  peaks in the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY spectra. Slice *a* in Figure 3.1B showed one such example. It was observed when the chemical shift difference between  $^{13}\text{C}_{\delta 1}/^{13}\text{C}_{\delta 2}$  and  $^{13}\text{C}_{\gamma}$  was smaller than or close to the  $^1\text{J}_{\text{CC}}$  value due to the strong coupling effect. To assign these methyl groups, a non-constant-time MQ-(H)CCH-TOCSY experiment was used.

Slice *b* in Figure 3.1B, which was taken from the non-CT MQ-(H)CCH-TOCSY spectrum, showed the correlations between  $^{13}\text{C}_{\delta}$  and all aliphatic carbons in the same Leu residue as shown in slice *a*. Using the ( $^{13}\text{C}_{\alpha}$ ,  $^{13}\text{C}_{\beta}$ ) chemical shifts obtained from this slice, the signals were assigned to Leu48 $\delta_2$ . The 6 and 2 Leu methyl groups in the  $\alpha$ - and  $\beta$ -chains, respectively, which showed very weak signals in the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment, were assigned with the MQ-(H)CCH experiments. Figure 3.2 shows the  $^1\text{H}$ - $^{13}\text{C}$  HSQC spectra of  $\alpha$ - and  $\beta$ -chains with assignments.

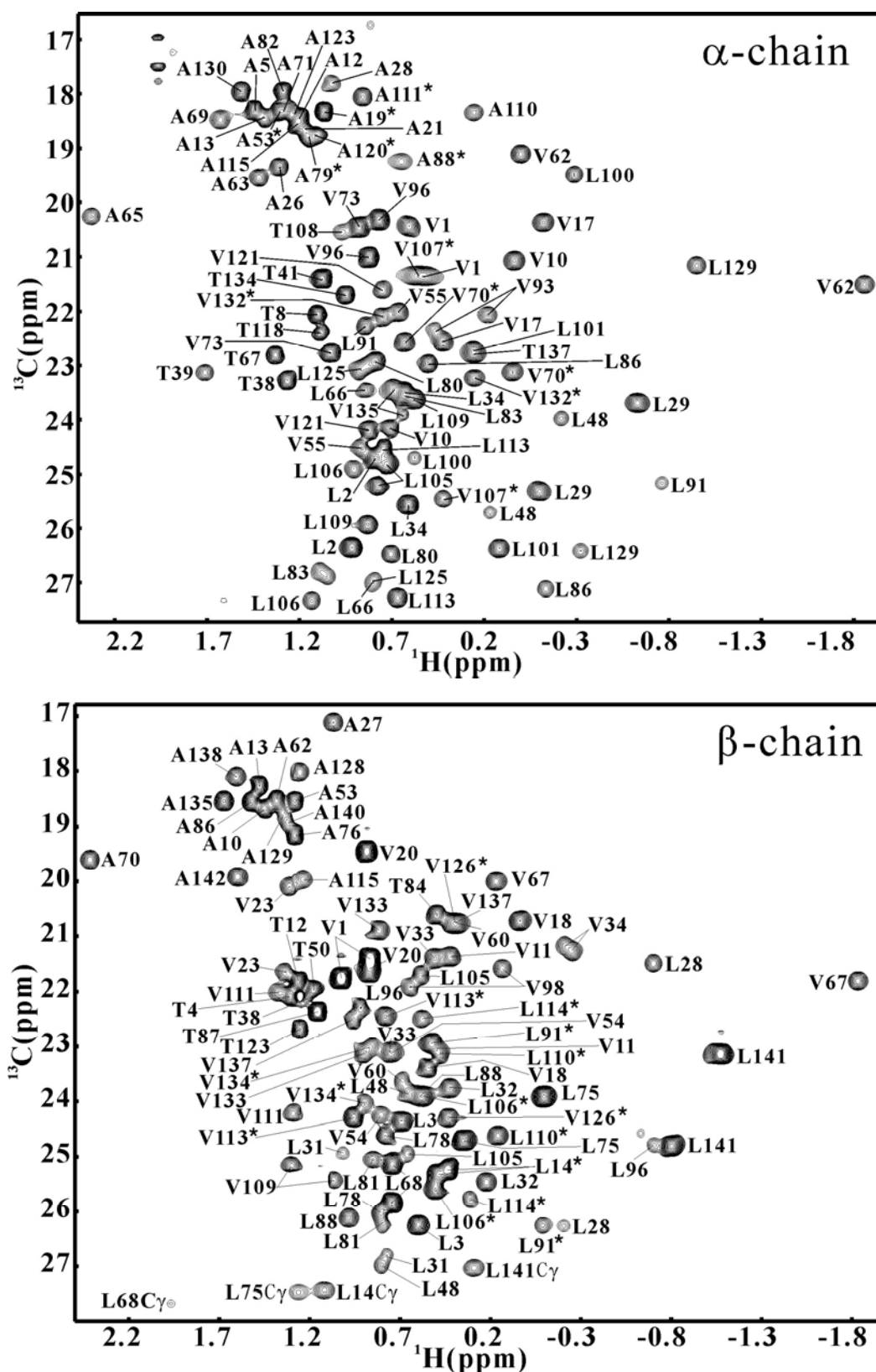


Figure 3.2 CT  $^{13}\text{C}$ - $^1\text{H}$  HSQC of the  $^{13}\text{C}$ -labeled  $\alpha$ -chain and  $\beta$ -chain of rHbCO A.

Cross-peaks are labeled with their assignments. Ambiguous assignments are indicated with asterisks (\*).

### 3.2.2 Assignment of side-chain protons in methyl-containing residues

In principle, one can assign the resonances of most side-chain  $^1\text{H}$  and  $^{13}\text{C}$  spins using the 3D HCCH-TOCSY or MQ-(H)CCH-TOCSY experiments. In practice, it is very difficult to do so because of poor resolutions in both indirect dimensions and poor dispersion of most  $^1\text{H}$ - $^{13}\text{C}$  correlations for large proteins. To assign side-chain protons in methyl-containing residues, an H(C) $\text{C}_m\text{H}_m$ -TOCSY experiment is proposed, which is similar to the HCCH-TOCSY experiment (Bax, Clore et al. 1990; Uhrin, Uhrinova et al. 2000). This procedure correlates methyl  $^1\text{H}$  with all aliphatic  $^1\text{H}$  spins in the same residue through a  $^{13}\text{C}$  TOCSY mixing scheme. The relatively good dispersion of methyl  $^1\text{H}$ - $^{13}\text{C}$  correlations and the slow decay of the methyl spins in uniformly  $^{13}\text{C}$ -labeled protein make  $^1\text{H}$  assignment possible (Liu, Zheng et al. 2003).

Figure 3.3 shows a number of slices taken from the H(C) $\text{C}_m\text{H}_m$ -TOCSY experiment. Each methyl  $^1\text{H}$  correlates with all aliphatic protons in the same residue. However, this experiment does not provide direct  $^1\text{H}$ - $^{13}\text{C}$  correlations of pairs of covalently bonded atoms. Assignment of  $^1\text{H}$  chemical shifts is based on empirical  $^1\text{H}$  chemical shift ranges of different protons and the assignment of methyl groups. In many cases, the assignment is straightforward. However, discrimination of  $^1\text{H}_\alpha$  and  $^1\text{H}_\beta$  spins for Thr and  $^1\text{H}_\beta$  and  $^1\text{H}_\gamma$  spins for Leu also needs spectral information from the MQ-(H)CCH-TOCSY experiment which provides direct  $^1\text{H}$ - $^{13}\text{C}$  correlations of pairs of covalently bonded atoms.

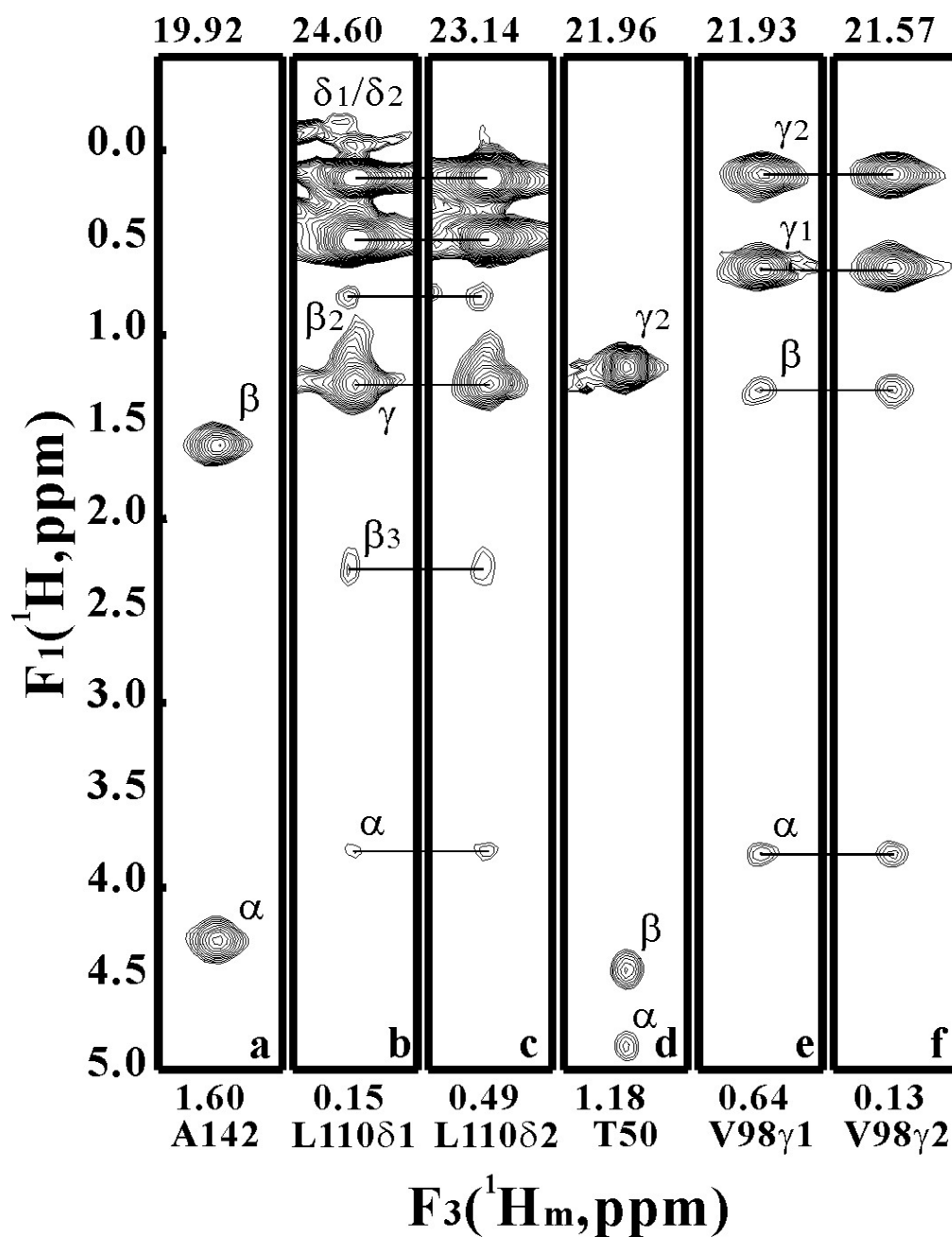


Figure 3.3 Representative F1–F3 slices from the H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY spectrum of <sup>13</sup>C-labeled β-chain of rHbCOA.

Each F1(<sup>1</sup>H)–F3(<sup>1</sup>H<sub>m</sub>) slice is labeled with the identity of the methyl-containing residue, and the F2 (<sup>13</sup>C<sub>m</sub>) frequency in ppm is indicated at the top of each slice.

For example, the assignment of T134H<sub>α</sub> of the α-chain can be confirmed from the correlations [<sup>13</sup>C<sub>i</sub>, <sup>13</sup>C<sub>α</sub>, <sup>1</sup>H<sub>α</sub>], where i = α, β, and γ, positioned at T134C<sub>α</sub> in the MQ-(H)CCH-TOCSY spectrum (slice *c*, Figure 3.4). On the basis of slices *b* and *c* in Figure 3.4, one can also assign T134H<sub>α</sub> and it seems that the H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY experiment is not necessary in this particular case. According to slices *e* and *f* in Figure 3.4, however, one cannot determine the chemical shift of V10H<sub>β</sub> because two or more protons in the range of 1.85 – 2.1 ppm correlate with sets of <sup>13</sup>C resonances with very similar chemical shifts to (<sup>13</sup>C<sub>α</sub>, <sup>13</sup>C<sub>β</sub>, <sup>13</sup>C<sub>γ</sub>) of V10 as shown on slice *f*. On the other hand, V10H<sub>β</sub> can be easily assigned from slice *d*.

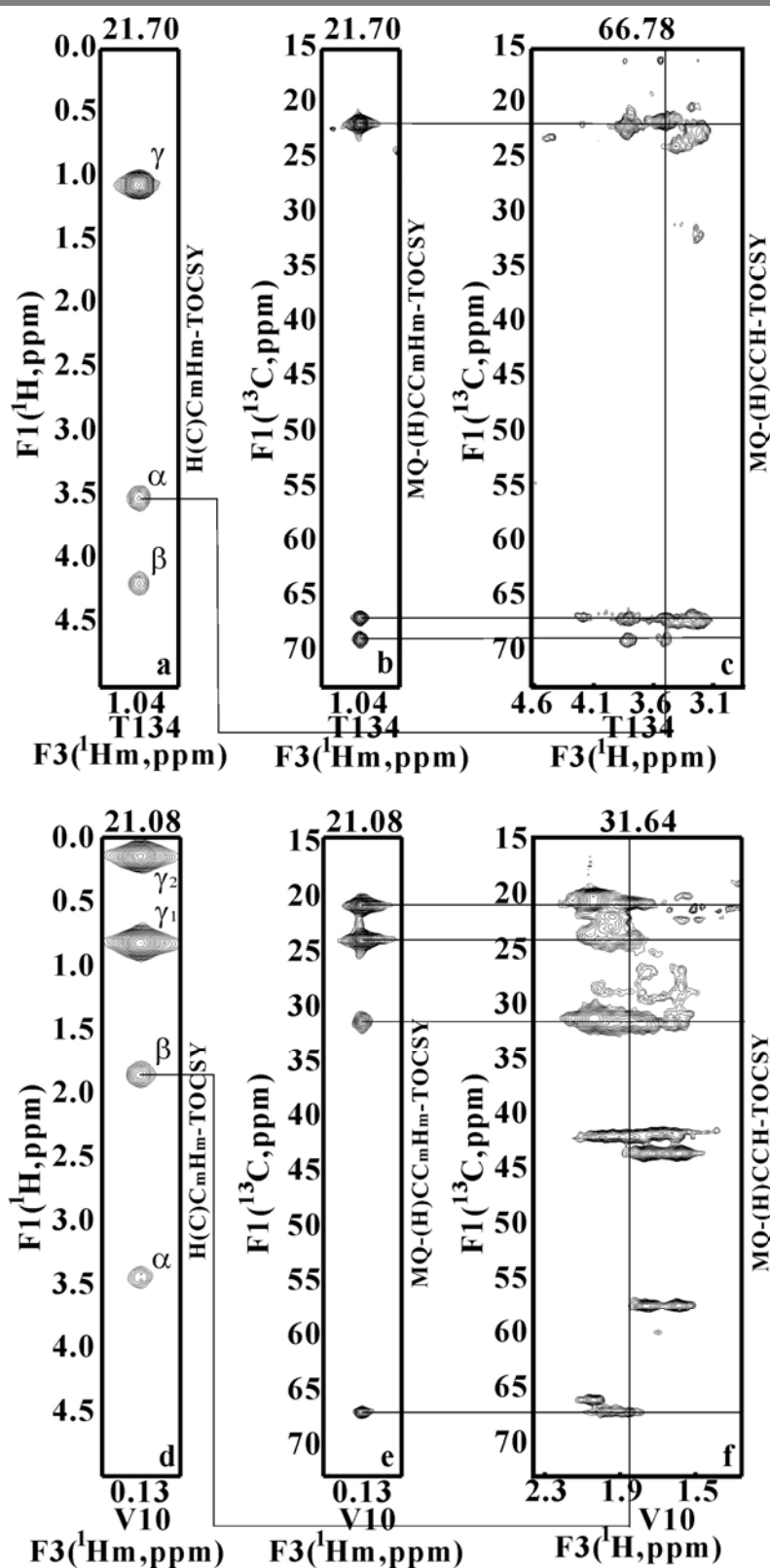
Actually, most <sup>1</sup>H resonances cannot be assigned without H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY data because the MQ-(H)CCH-TOCSY spectrum had poor resolutions in both indirect dimensions and poor dispersion of most <sup>1</sup>H-<sup>13</sup>C correlations. We have unambiguously assigned 90 out of 137 non-methyl protons in methyl-containing residues of the α-chain and 89 out of 137 non-methyl protons of the β-chain. Nearly all unassigned protons were <sup>1</sup>H<sub>β</sub> in Leu residues because the experimental sensitivity for CH<sub>2</sub> groups in which the two protons are magnetically different and two sets of correlations exist is lower than that for CH and CH<sub>3</sub> groups. The results are summarized in Table 3.1.

**Table 3.1 Summary of assignment of non-methyl protons in methyl-containing residues of both α- and β-chains of rHbCOA**

|            | rHbCOA α-chain *     |                      |                      | rHbCOA β-chain*      |                      |                      |
|------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| <b>Ala</b> | H <sub>α</sub> 13/21 |                      |                      | H <sub>α</sub> 14/15 |                      |                      |
| <b>Thr</b> | H <sub>α</sub> 9/9   | H <sub>β</sub> 9/9   |                      | H <sub>α</sub> 6/7   | H <sub>β</sub> 6/7   |                      |
| <b>Val</b> | H <sub>α</sub> 10/13 | H <sub>β</sub> 10/13 |                      | H <sub>α</sub> 15/18 | H <sub>β</sub> 15/18 |                      |
| <b>Leu</b> | H <sub>α</sub> 15/18 | H <sub>β</sub> 8/36  | H <sub>γ</sub> 16/18 | H <sub>α</sub> 13/18 | H <sub>β</sub> 8/36  | H <sub>γ</sub> 12/18 |

\* Number of assigned/total protons





**Figure 3.4** F1-F3 slices taken from the spectra of H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY, MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY and MQ-(H)CCH-TOCSY experiments.

The corresponding experiment for each slice is labeled beside the slice. The chemical shift of the F2 dimension is labeled on the top of each slice.

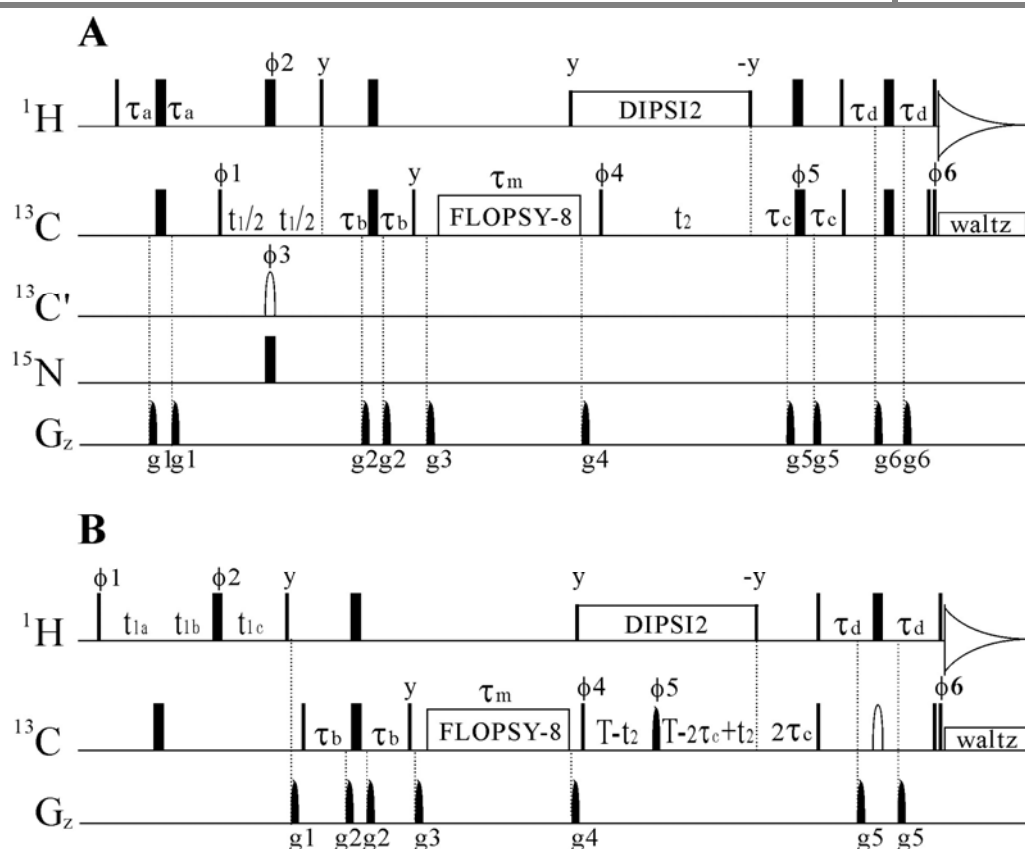
### 3.3 Conclusion

In summary, methyl resonances of large proteins, e.g., rHbCO A, can be assigned using uniformly  $^{13}\text{C}$ -labeled proteins with the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment. In addition, most side-chain  $^1\text{H}$  and  $^{13}\text{C}$  resonances of methyl-containing residues can be assigned with the H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY and MQ-(H)CCH-TOCSY experiments. The non-CT MQ-(H)CCH-TOCSY experiment is also complementary to the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment for the assignment of methyl groups in Leu residues. The strategy proposed here facilitates the study by NMR of structure, dynamics and structure-activity-relationship of large proteins, especially multimeric proteins, without specific isotope labeling.

### 3.4 Materials and methods

#### 3.4.1 MQ-(H)CCH-TOCSY experiment

Figure 3.5A shows the pulse sequence for establishing ( $^{13}\text{C}$ ,  $^{13}\text{C}$ ,  $^1\text{H}$ ) correlations through a TOCSY scheme, which is similar to the original HCCH-TOCSY (Bax, Clore et al. 1990; Fesik, Eaton et al. 1990) and MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY (Chapter 2) experiments. A non-constant-time acquisition mode in the  $t_2$  period is used to replace the constant-time (CT)  $t_2$  period in the MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment. In addition, a non-selective  $^{13}\text{C}$   $180^\circ$  pulse is used in the last INEPT period to allow the detection of all aliphatic  $^1\text{H}$  spins.



**Figure 3.5** Pulse sequences for the MQ-(H)CCH-TOCSY (A) and H(C) $C_m$ H $_m$ -TOCSY (B) experiments.

All narrow (wide) rectangular pulses have flip angles of  $90^\circ$  ( $180^\circ$ ). The  $^1\text{H}$  carrier is at 4.7 ppm while the  $^{13}\text{C}$  carrier is centered at 41 ppm. For scheme B, the  $^{13}\text{C}$  carrier is jumped to 17 ppm immediately prior to the  $g_4$  gradient pulse. All  $^1\text{H}$  pulses are applied with a 23 kHz field;  $^1\text{H}$  DIPS12-decoupling elements make use of a 6.25 kHz field. All  $^{13}\text{C}$  rectangular pulses employ a 16.8 kHz field and the  $^{13}\text{C}$  shaped pulses have REBURP profiles. The  $180^\circ$  shaped  $^{13}\text{C}$  pulse (filled) has a duration of  $400\ \mu\text{s}$  and is phase-modulated by 24 ppm while the second one (empty) has a duration of 1.5 ms. The  $^{13}\text{C}$  spin-lock field strength for FLOPSY is 7 kHz. A decoupling power of 1.25 kHz is used during acquisition. The  $180^\circ$  pulse on  $\text{C}'$  has a SEDUCE profile with a duration of  $250\ \mu\text{s}$  (center of excitation 176 ppm). The pulse on  $^{15}\text{N}$  is omitted for  $^{13}\text{C}$ -labeled samples. The delays used are:  $\tau_a = 1.4\ \text{ms}$ ;  $\tau_b = 1.1\ \text{ms}$ ;  $\tau_c = 1\ \text{ms}$  for scheme A and 0.75 ms for scheme B;  $\tau_d = 1.6\ \text{ms}$ ;  $\tau_m = 17\ \text{ms}$ ;  $T = 14\ \text{ms}$ ;  $t_{1a} = 1.4\ \text{ms} + t_1$ ;  $t_{1b} = t_1 - t_1'$ ;  $t_{1c} = 1.4\ \text{ms} - t_1'$ ;  $t_1' = 1.4\ \text{ms}/(n_i - 1)$  where  $n_i$  is the total complex points in the  $t_1$  dimension. The phase cyclings employed are:  $\Phi_1 = 4(x), 4(-x)$ ;  $\Phi_2 = x, y, -x, -y$ ;  $\Phi_3 = 2(x), 2(-x)$ ;  $\Phi_4 = y$ ;  $\Phi_5 = 2(x), 2(y), 2(-x), 2(-y)$ ;  $\Phi_6 = 4(x), 4(-x)$ ;  $\text{rec} = x, -x, -x, x, -x, x, x, -x$ . The duration and strengths of the sine-shaped gradients are:  $g_1 = (0.5\ \text{ms}, 20\ \text{G/cm})$ ;  $g_2 = (0.3\ \text{ms}, 25\ \text{G/cm})$ ;  $g_3 = (1\ \text{ms}, 25\ \text{G/cm})$ ;  $g_4 = (1\ \text{ms}, 20\ \text{G/cm})$ ;  $g_5 = (0.5\ \text{ms}, 20\ \text{G/cm})$ ;  $g_6 = (1\ \text{ms}, 10\ \text{G/cm})$ . Quadrature detection in F1 and F2 is achieved by State-TPPI of  $\Phi_1$  and  $\Phi_4$ , respectively.

The non-CT MQ-(H)CCH-TOCSY experiment proposed here has lower resolution in the F2 dimension than the CT MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY experiment due to  $J_{CC}$  couplings and a short acquisition time in this dimension. However, the former is significantly more sensitive, especially for Leu residues with strong scalar coupling interactions, because  $^{13}\text{C}$  magnetization cannot be refocused completely for strongly coupled spin systems during the CT period.

### 3.4.2 H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY experiment

Figure 3.5B shows the pulse scheme of 3D H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY experiment, which is similar to that of the MQ-(H)C<sub>m</sub>H<sub>m</sub>-TOCSY experiment. In this pulse sequence,  $^1\text{H}$  chemical shifts instead of  $^{13}\text{C}$  shifts were recorded in the  $t_1$  period in a single-quantum (SQ) mode. Although multiple-quantum (MQ) coherences ( $\text{H}_x\text{C}_y$ ) have longer relaxation times than SQ coherences ( $\text{H}_x$ ), the MQ mode involves additional signal loss from  $^{13}\text{C}$ - $^{13}\text{C}$  couplings in the  $t_1$  period. The SQ mode was more sensitive than the MQ mode without selective  $^{13}\text{C}$  decoupling in the  $t_1$  period and was thus used in this experiment.

### 3.4.3 Protein Samples and NMR Spectroscopy

Chain-specific  $^{13}\text{C}$ -labeled rHbCO A samples were prepared as described previously (Simplaceanu, Lukin et al. 2000). NMR experiments were performed on samples of  $\sim 1.0$  mM protein (in the tetramer), 20 mM sodium phosphate, pH 7.0, and 100% D<sub>2</sub>O at 34 °C. All spectra were recorded on a Bruker Avance 500 MHz NMR spectrometer equipped with pulse gradient units and an actively shielded cryoprobe.

The 3D MQ-(H)CC<sub>m</sub>H<sub>m</sub>-TOCSY data comprising  $64 \times 70 \times 512$  complex points with spectral widths of 8000, 2516, and 8000 Hz in F1, F2, and F3 dimensions (corresponding to acquisition times of 7.7, 27.4, and 64 ms, respectively) were collected with 4 scans and an inter-scan delay of 0.95 s for each FID, giving rise to a net experimental time of 24.5 h.

The 3D H(C)C<sub>m</sub>H<sub>m</sub>-TOCSY data set consisting of  $32 \times 70 \times 512$  complex points with spectral widths of 3500, 2516, and 8000 Hz in F1, F2, and F3 dimensions was acquired using 8 scans and a relaxation delay of 0.95 s for each increment, resulting in a total experimental time of 24.5 h.

The MQ-(H)CCH-TOCSY data comprising  $64 \times 30 \times 512$  complex points with spectral widths of 8000, 3774, and 8000 Hz in F1, F2, and F3 dimensions were collected with an inter-scan delay of 0.95 s and 8 scans per increment, resulting in a total experimental time of 20.5 h.

All data sets were apodized with a sine weighting function shifted by  $72^\circ$  in the direct proton dimension. The t1 and t2 domains were doubled by linear prediction prior to the application of a cosine-squared window function. After zero filling and Fourier transformation, all the final data sets comprised  $256 \times 256 \times 1024$  points along the F1, F2, and F3 dimensions, respectively. Processing of the spectra was carried out using NMRPipe and analyzed with NMRView.

#### 3.4.4 Correction of $^{13}\text{C}$ chemical shifts

Assignment of CH<sub>3</sub> groups relies on prior assignments of  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$  chemical shifts. For rHbCO A, sequential assignments were obtained from perdeuterated samples (Lukin, Kontaxis et al. 2004). Due to  $^2\text{H}$  isotope effects,

$^{13}\text{C}$  chemical shifts observed in a perdeuterated sample are smaller than those observed in a protonated sample. To make  $^{13}\text{C}$  chemical shifts consistent for both samples,  $^2\text{H}$  isotope effects are corrected according to the following equation (Venters, Farmer et al. 1996):

$$\delta\text{C}(\text{H}) = \delta\text{C}(\text{D}) - ({}^1\Delta\text{C}(\text{D}) * d_{1b} + {}^2\Delta\text{C}(\text{D}) * d_{2b} + {}^3\Delta\text{C}(\text{D}) * d_{3b}), \quad (1)$$

where  $\delta\text{C}(\text{H})$  and  $\delta\text{C}(\text{D})$  are the chemical shifts of a  $^{13}\text{C}$  spin in protonated and perdeuterated samples, respectively;  ${}^n\Delta\text{C}(\text{D})$  represents the n-bond isotope effect per deuteron; and  $d_{nb}$  is the number of deuterons n bonds removed from the  $^{13}\text{C}$  nucleus. Due to the negligible magnitude of  ${}^4\Delta\text{C}(\text{D})$  in saturated alkanes, Equation 1 has been restricted to isotope shifts over three bonds or fewer. The three  ${}^n\Delta\text{C}(\text{D})$  constants used are:  ${}^1\Delta\text{C}(\text{D}) = -0.29$ ;  ${}^2\Delta\text{C}(\text{D}) = -0.13$ ;  ${}^3\Delta\text{C}(\text{D}) = -0.07$  (Venters, Farmer et al. 1996).

## **Chapter 4:**

# **A new strategy for structure determination of large proteins in solution without deuteration**

- 4.1 Introduction
- 4.2 General strategy for sequence-specific assignments
- 4.3 NOE assignment and structure determination
- 4.4 Discussion and conclusion
- 4.5 Materials and methods

## Chapter 4:

# A new strategy for structure determination of large proteins in solution without deuteration

## 4.1 Introduction

So far high-resolution structure determination by NMR spectroscopy with uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled non-deuterated samples has been limited to proteins <25 kDa (Chapter 1). For proteins larger than 30 kDa, predeuterated or/and specific isotopic labelled samples are required to achieve sequential assignment or global fold determination. Although uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled non-deuterated samples can be used to obtain assignments of methyl groups (Chapter 2) and side-chain assignments of methyl-containing residues (Chapter 3) in large proteins, prior assignments of  $^{13}\text{C}_\alpha$  and  $^{13}\text{C}_\beta$ , which can be obtained from TROSY experiments with predeuterated samples are needed. Furthermore, high-resolution structure can only be achieved by constraining side chains of all or most residues using NOEs among side-chain protons. This requires not only the assignments of methyl groups or side-chain assignments of methyl-containing residues, but also a nearly completed sequential-specific resonance assignment of the whole protein.

In this chapter, we present a novel strategy to assign backbone and side chain resonances of large proteins without the use of deuterium and specific labeling. On the basis of the assignments, we determined high resolution structures from distance restraints derived from NOEs and dihedral restraints derived from chemical shifts. We demonstrated the strategy on three samples:

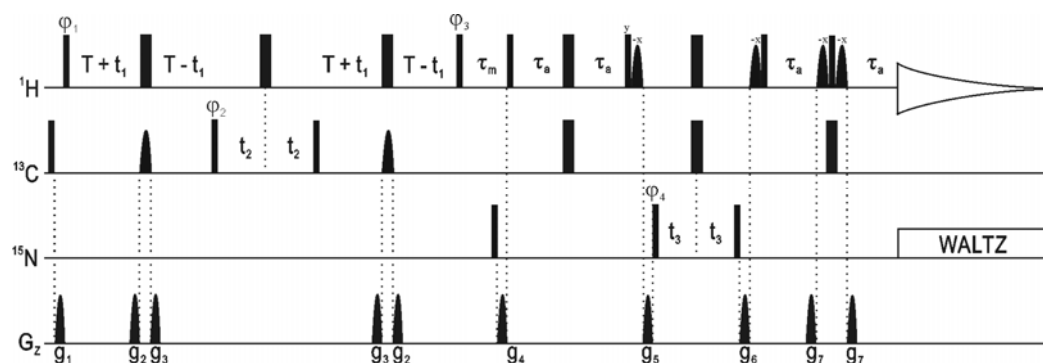


Ca<sup>2+</sup>-dependent cell adhesion protein (DdCAD-1, 214 residue, ~24 kDa, 3%  $\alpha$ -helices, 46%  $\beta$ -strands), maltose binding protein (MBP, 370 residues, ~42 kDa, 42%  $\alpha$ -helices, 16%  $\beta$ -strands), and human normal adult hemoglobin in the carbonmonoxy form (HbCO A, chain-specifically <sup>13</sup>C,<sup>15</sup>N-labeled, 141 residues for  $\alpha$ -chain, 146 residues for  $\beta$ -chain, ~65 kDa for tetramer, 77%  $\alpha$ -helices).

## 4.2 General strategy for sequence-specific assignments

### 4.2.1 General strategy for sequential assignment

The strategy consists of five steps. First, clusters are formed by grouping HC-NH NOE and C <sup>$\alpha$</sup> -NH (HNCA) correlations that have identical NH chemical shifts. Second, spin-systems are identified by separating out intra-residue and sequential HC-NH NOE correlations from other inter-residue NOEs observed in a four-dimensional 4D <sup>13</sup>C,<sup>15</sup>N-edited NOESY spectrum with the use of 3D TROSY-HNCA and MQ-CCH-TOCSY spectra. Third, spin-systems are classified by residue type based on <sup>1</sup>H and <sup>13</sup>C chemical shifts. Fourth, fragments are established from clusters by matching the intra-residue spin-system of one cluster with the sequential spin-system of another cluster. Fifth, fragments are mapped onto the protein sequence in a manner similar to the traditional triple-resonance approach (Ikura, Kay et al. 1990).



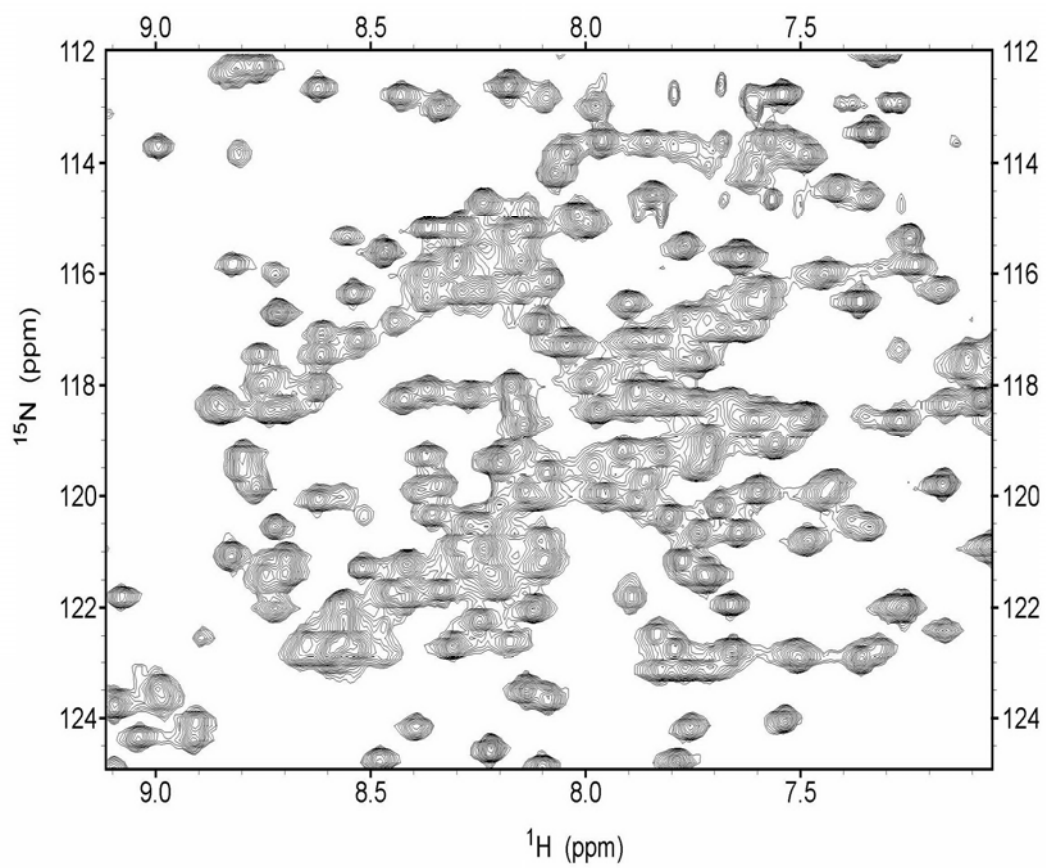
**Figure 4.1** Pulse sequence for recording 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY.

All narrow (wide) bars represent  $90^\circ$  ( $180^\circ$ ) rectangular pulses. The carriers are centered at 4.7, 65 and 119 ppm for  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$  respectively. Rectangular  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$  pulses are applied with field strengths of 25, 25 and 7.1 kHz, respectively. The  $^1\text{H}$  shaped  $90^\circ$  pulses have a sinc profile (1.4 ms, water-selective). The  $^{13}\text{C}$  shaped  $180^\circ$  pulses are ca-WURSTs ( $300 \times 800 / \gamma \mu\text{s}$ , for experiments recorded at  $\gamma$  MHz).  $^{15}\text{N}$ -decoupling is achieved with use of a 1.25 kHz WALTZ16 field. The delays used are:  $T = 1.5$  ms and  $\tau_a = 2.25$  ms. The durations and strengths of gradients are:  $g_1 = (1\text{ms}, 15\text{G/cm}, \text{sine-shaped})$ ,  $g_2 = (0.2\text{ms}, 10\text{G/cm}, \text{rectangle-shaped})$ ,  $g_3 = (0.2\text{ms}, 30\text{G/cm}, \text{rectangle-shaped})$ ,  $g_4 = (2\text{ms}, 22.5\text{G/cm}, \text{sine-shaped})$ ,  $g_5 = (2\text{ms}, 25\text{G/cm}, \text{sine-shaped})$ ,  $g_6 = (0.5\text{ms}, -25\text{G/cm}, \text{sine-shaped})$ ,  $g_7 = (0.5\text{ms}, 40\text{G/cm}, \text{sine-shaped})$ . The phase cycling employed is:  $\phi_1 = x$ ;  $\phi_2 = x, -x$ ;  $\phi_3 = 45^\circ$ ;  $\phi_4 = x, x, -x, -x$ ;  $\phi_{\text{ref}} = x, -x, -x, x$ . Quadrature detections in  $F_1$ ,  $F_2$  and  $F_3$  are achieved by States-TPPI of  $\phi_1$ ,  $\phi_2$  and  $\phi_4$ .

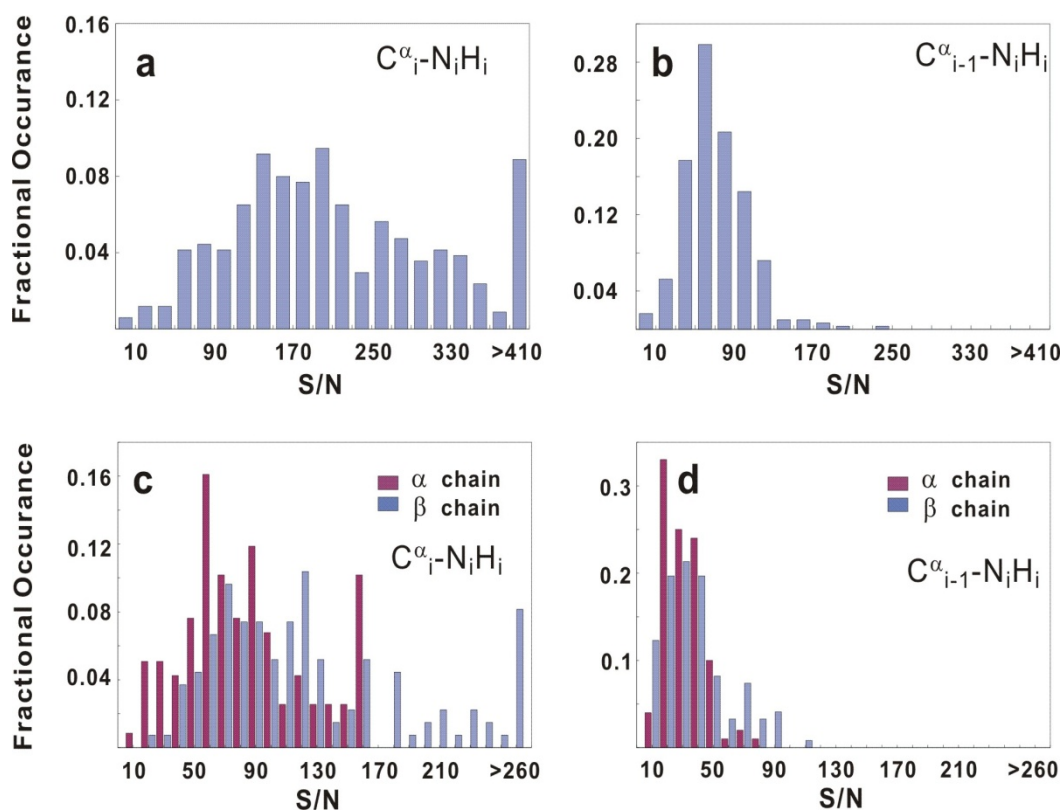
#### 4.2.1.1 Peak clusters

We grouped peaks in 3D TROSY-HNCA and 4D  $^{13}\text{C},^{15}\text{N}$ -edited NOESY spectra (Figure 4.1) to form clusters according to their common NH chemical shifts, where NH denotes amide spins  $^{15}\text{N}$  and  $^1\text{H}$ . Some clusters may comprise peaks from two or more amides which have degenerate NH chemical shifts. We could easily identify the number of amides involved in each cluster with TROSY-HNCO and HNCA spectra based on the fact that one amide gives rise to only one HNCO peak and no more than two HNCA peaks. When a NOE peak could be grouped into more than one cluster, we did not group it into any cluster in this step.

Most amide groups could be unambiguously distinguished in the two spectra for each sample, although the TROSY heteronuclear single-quantum correlation (HSQC) spectra were crowded in some regions (Figure 4.2). For the three proteins studied here, only a few clusters contained cross-peaks from more than one but less than four residues (Table 1). DdCAD-1 and MBP had 2 and 9 such clusters, respectively, and each HbCO A chain had 3 such clusters. Most clusters contained both intra-residue and sequential HNCA correlations (Figure 4.3), while ~10 - 15% clusters contained only one HNCA correlation.



**Figure 4.2** The middle region of a 2D TROSY-HSQC of fully protonated MBP recorded on an 800 MHz NMR at 30 °C.



**Figure 4.3** Distributions of peak signal-to-noise (S/N) ratio for the 3D TROSY-HNCA experiments.

(a-b) are the data from MBP. (c-d) are the data from HbCO A. For MBP, we observed 338 out of 348 expected intra-residue correlations ( $C^{\alpha}_i-N_iH_i$ ), while identified 305 sequential correlations ( $C^{\alpha}_{i-1}-N_iH_i$ ); eight non-proline residues did not display N-H HSQC peaks. For the  $\alpha$  chain of HbCO A, we observed 118 out of 133 expected intra-residue correlations and identified 110 sequential correlations; twelve nonproline residues underwent significant conformational exchange and gave rise to no N-H HSQC peaks. For the  $\beta$  chain of HbCO A, we observed 135 out of 138 expected intraresidue correlations, while detected 119 sequential correlations; two non-proline residues showed no N-H HSQC peaks. When an intra-residue HNCA peak was overlapped with a sequential HNCA peak, we considered the sequential peak as unidentified. Most unidentified sequential peaks resulted from signal overlap.

**Table 4.1 Summary of clusters, spin-systems, dipeptide segments and assignments**

|  | DdCAD-1<br>214/202/200 <sup>a</sup> | MBP<br>370/348/329 <sup>a</sup> | HbA $\alpha$ -chain<br>141/134/122 <sup>a</sup> | HbA $\beta$ -chain<br>146/139/138 <sup>a</sup> |          |
|--|-------------------------------------|---------------------------------|---|--|----------|
| Clusters   | 197+2(2) <sup>b</sup>               | 332+8(2)+1(3) <sup>b</sup>      | 118+3(2) <sup>b</sup>                           | 131+2(2)+1(3) <sup>b</sup>                     |          |
| Clusters lacking one HNCA peak   | 19                                  | 50                              | 16  | 15   |          |
| Clusters lacking one spin-system   | 3                                   | 26                              | 13  | 12   |          |
| Spin-systems identified  | 399                                 | 676                             | 231   | 258  |          |
| Spin-systems without NOEs  | 4                                   | 51                              | 10  | 24   |          |
| Spin-systems with only one H <sup><math>\alpha</math></sup> C <sup><math>\alpha</math></sup> spin-pair | 30                                  | 126                             | 79  | 66   |          |
|  | Gly                                 | 26/0/28                         | 55/0/56   | 13/0/14  | 24/0/26  |
|  | Ala                                 | 11/0/12                         | 83/1/88   | 34/0/42  | 29/0/30  |
| Spin-  | Thr                                 | 31/0/35                         | 37/0/41   | 15/0/17  | 6/0/11   |
| systems  | Val                                 | 42/2/42                         | 40/1/40   | 21/1/26  | 26/3/36  |
| typed as <sup>c</sup>  | Ile                                 | 17/0/17                         | 34/0/42   | 0/0/0  | 0/0/0    |
|  | Leu                                 | 16/1/16                         | 47/0/58   | 13/0/35  | 13/0/36  |
|  | others                              | 214/3/245                       | 161/0/352                                       | 45/4/126                                       | 57/7/132 |
| Dipeptide segments   | 164                                 | 217, (2) <sup>d</sup>           | 70  | 71   |          |
| Residues assigned from fragments unique in connectivity & mapping                                      | 174                                 | 176, (1) <sup>e</sup>           | 89  | 81   |          |
| Residues assigned based on uniquely mapped fragments   | 26                                  | 163, (1) <sup>f</sup>           | 28  | 54   |          |
| Uniquely assigned residues   | 200                                 | 339                             | 117   | 135  |          |
| Ambiguously assigned residues  | 0                                   | 0                               | 5   | 0  |          |
| Unassigned residues  | 2                                   | 9                               | 12  | 4  |          |

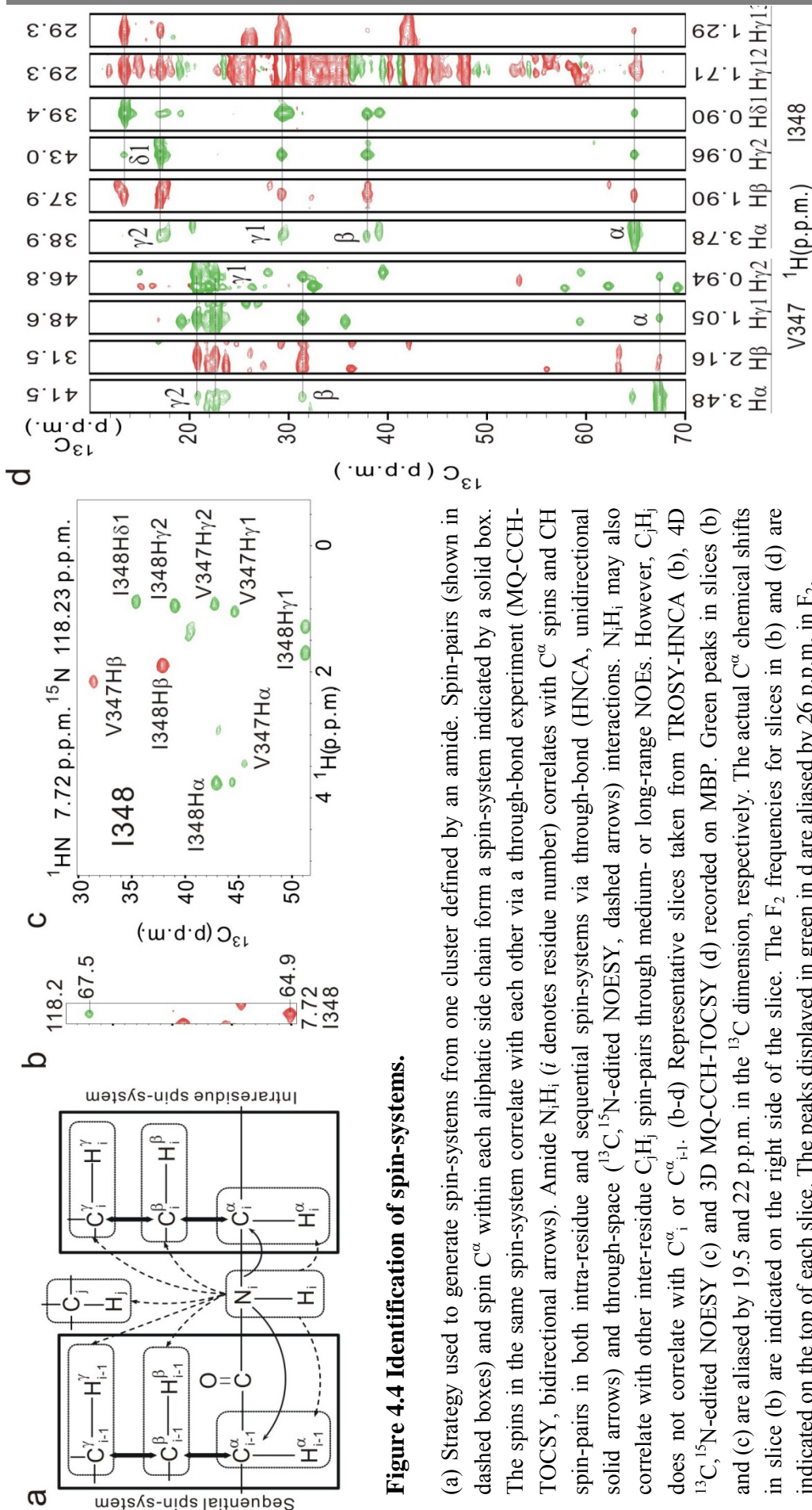
- total residue number/expected amide correlations/amides assigned with triple-resonance approach
- the second (third) term: number of the clusters containing information from two (three) amides or residues
- numbers of spin-systems correctly typed/wrongly typed/expected
- number of dipeptide segments formed by non-adjacent residues
- number of residues wrongly assigned in the first stage. The correctness of the assignments was assessed based on published results obtained with the triple-resonance approach.
- number of the wrong assignments in the first stage corrected in the second

#### 4.2.1.2 Spin-system identification and amino acid type determination

With the use of HNCA and MQ-CCH-TOCSY spectra, intra-residue and sequential HC-NH NOE correlations of each cluster can be separated out from other inter-residue HC-NH NOE correlations observed in the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectrum. Meanwhile, by grouping the intra-residue and sequential HC-NH NOE correlations into separate spin-system, an intra-residue spin-system and a sequential spin-system can be identified for each cluster (Figure 4.4 a).

The amino acid type of both spin-systems can subsequently be identified based on the dispersion of the  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts (Chapter 7). We could easily and unambiguously recognize most Glycine and methyl-containing residues since they contain characteristic spin-pairs. When Arginine, Lysine, Proline and Serine yielded nearly complete spin-systems, we could also identify them unequivocally. We classified the remaining amino acid residues into three groups: 1. Aspartate, Asparagine, Phenylalaline and Tyrosine (DNFY), 2. Glutamate, Glutamine and Methionine (EQM), and 3. Cysteine, Tryptophan and Histidine (CWH). If the characteristic information in a spin-system is not enough for determining the type of a residue, it will remain unclassified.

An example of identifying spin-systems from a given cluster is described below. First, we extracted the intra-residue and sequential  $\text{C}^\alpha$ -NH peaks to build two initial spin-systems from the HNCA slice defined by an amide in a cluster (Figure 4.4 b). Second, we found that two  $\text{H}^\alpha\text{C}^\alpha$ -NH NOE peaks from the NOESY slice located at the amide matched the two  $\text{C}^\alpha$ -NH peaks in  $\text{C}^\alpha$  chemical shifts (Figure 4.4 c). Third, we listed the TOCSY slices defined by the CH spin-pairs of individual HC-NH NOEs in Figure 4.4 c (Figure 4.4 d). According to



**Figure 4.4 Identification of spin-systems.**

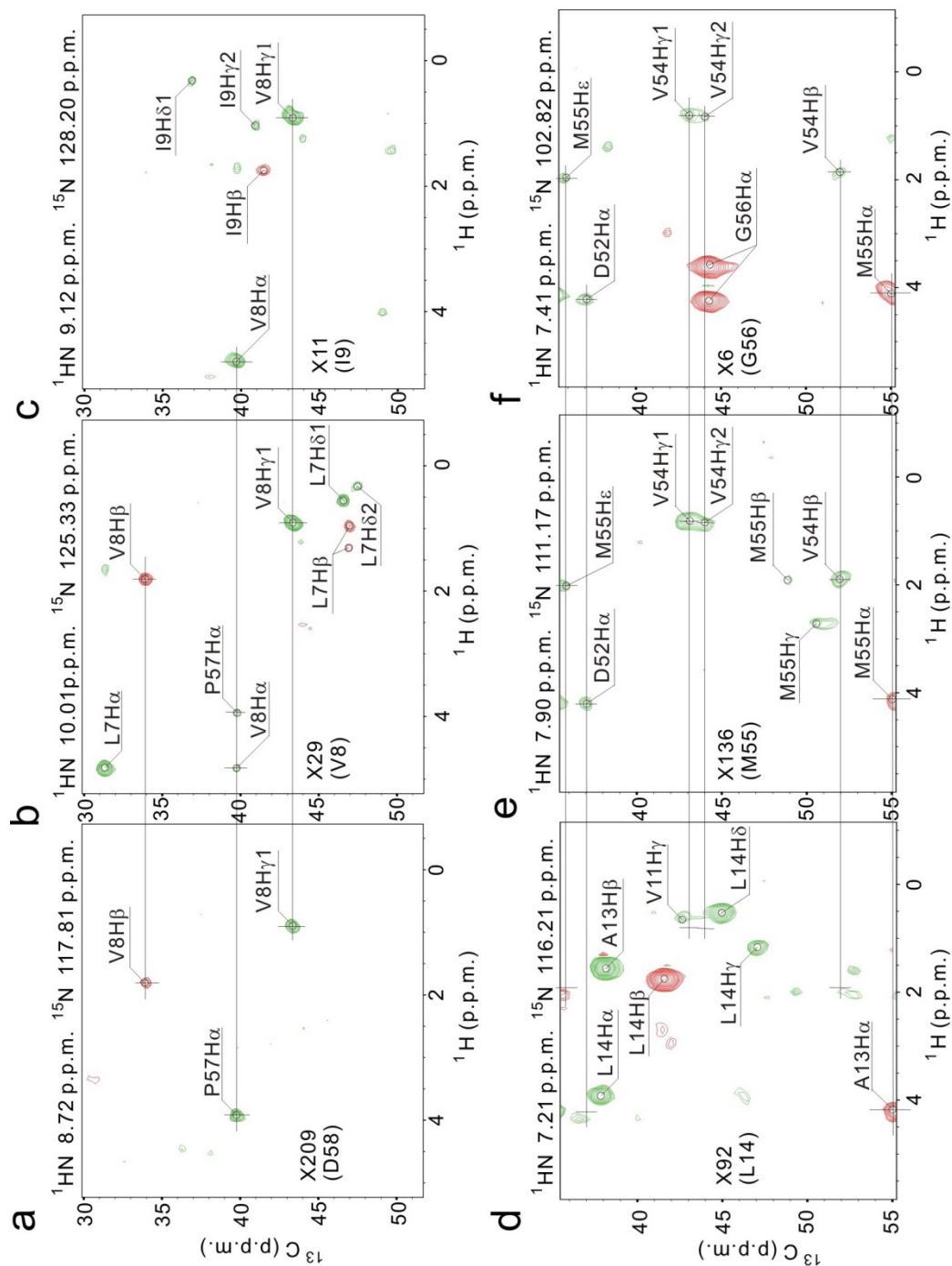
(a) Strategy used to generate spin-systems from one cluster defined by an amide. Spin-pairs (shown in dashed boxes) and spin  $C^\alpha$  within each aliphatic side chain form a spin-system indicated by a solid box. The spins in the same spin-system correlate with each other via a through-bond experiment (MQ-CCH-TOCSY, bidirectional arrows). Amide  $N_iH_i$  ( $i$  denotes residue number) correlates with  $C^\alpha$  spins and CH spin-pairs in both intra-residue and sequential spin-systems via through-bond (HNCA, unidirectional solid arrows) and through-space ( $^{13}C$ ,  $^{15}N$ -edited NOESY, dashed arrows) interactions.  $N_iH_i$  may also correlate with other inter-residue  $C_jH_j$  spin-pairs through medium- or long-range NOEs. However,  $C_jH_j$  does not correlate with  $C^\alpha_i$  or  $C^\alpha_{i-1}$ . (b-d) Representative slices taken from TROSY-HNCA (b), 4D  $^{13}C$ ,  $^{15}N$ -edited NOESY (c) and 3D MQ-CCH-TOCSY (d) recorded on MBP. Green peaks in slices (b) and (c) are aliased by 19.5 and 22 p.p.m. in the  $^{13}C$  dimension, respectively. The actual  $C^\alpha$  chemical shifts in slice (b) are indicated on the right side of the slice. The  $F_2$  frequencies for slices in (b) and (d) are indicated on the top of each slice. The peaks displayed in green in d are aliased by 26 p.p.m. in  $F_2$ .



TOCSY correlations of  $C^\alpha$ - $C^kH^k$  or  $C^k$ - $C^\alpha H^\alpha$ , where superscript  $k$  denotes the  $k^{\text{th}}$  side chain atom, we assigned three and five NOE peaks into the two spin-systems, respectively. The I348 $C^{\gamma11}H^{\gamma12}$  slice displayed a lot of noise as a result of the overlap of intense signals from lysine  $C^\delta H^\delta$ s. Nevertheless, its correlations with  $C^{\gamma2}$  and  $C^{\delta1}$  were strong and easily recognized. We determined the two spin-systems as valine and isoleucine, respectively, according to the chemical shifts of spins allocated to the spin-systems (Chapter 7). Following this procedure, we constructed many spin-systems (Table 4.1).

For DdCAD-1 (tumbling time 12.5 ns at 30 °C), almost all spin-systems contained one or more CH spin-pairs. For larger proteins such as MBP (~20 ns) and HbCO A (~30 ns), however, ~10% spin-systems contained no NOE correlations, and ~20% of spin-systems had no side-chain NOE correlations (Table 4.1). This resulted from poorer sensitivities of the 4D NOESY and 3D MQ-CCH-TOCSY toward larger proteins. Additionally, the MQ-CCH-TOCSY for MBP was very crowded in some regions, such as those corresponding to lysine residues, and we identified only <55% of the expected TOCSY correlations. Nevertheless, many spin-systems had enough characteristic chemical shift information, which allowed us to determine of amino acid types. For MBP and HbCO A, we typed ~60-70% of the spin-systems (Table 4.1 and Figure 4.9).

As the result of resonance degeneracy, this procedure led to a few medium- and long-range HC-NH NOEs being included erroneously as spin-system members; this happened in eight of the initial MBP spin-systems. Additionally, we initially assigned several spin-systems to wrong types because of missing characteristic peaks or inclusion of incorrect peaks (Table 4.1). As demonstrated below, these errors did not affect the correctness of the final assignments.



**Figure 4.5 Resolution of ambiguous connectivity between clusters.**

(a-f) Representative  $F_1F_2$  slices taken from the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectra recorded on MBP (a-c) and the  $\beta$ -chain of HbCO A (d-f). Each plane is labeled with its  $^{15}\text{N}$  and  $^1\text{HN}$  chemical shifts and the corresponding residue. The cluster number is also labeled inside each panel. All green peaks were aliased by 22 p.p.m. for MBP or 20 p.p.m. for HbCO A in the  $^{13}\text{C}$  dimension.

### 4.2.1.3 Assembly and mapping of connectivity fragments

When two spin-systems best matched each other in HC chemical shifts, we generated one dipeptide segment from them. In most cases, such a segment corresponded to one dipeptide fragment in the protein sequence. However, very few such segments might result from the connection of non-adjacent residues, as found in the study of MBP. Using all dipeptide segments, we have established several fragments (covering >48% of the residues) and uniquely mapped many of them to protein sequences (Table 4.1 and Figure 4.9). Owing to the presence of two wrong dipeptide segments for MBP, one mapped fragment initially contained one incorrect assignment at the C terminus of the fragment.

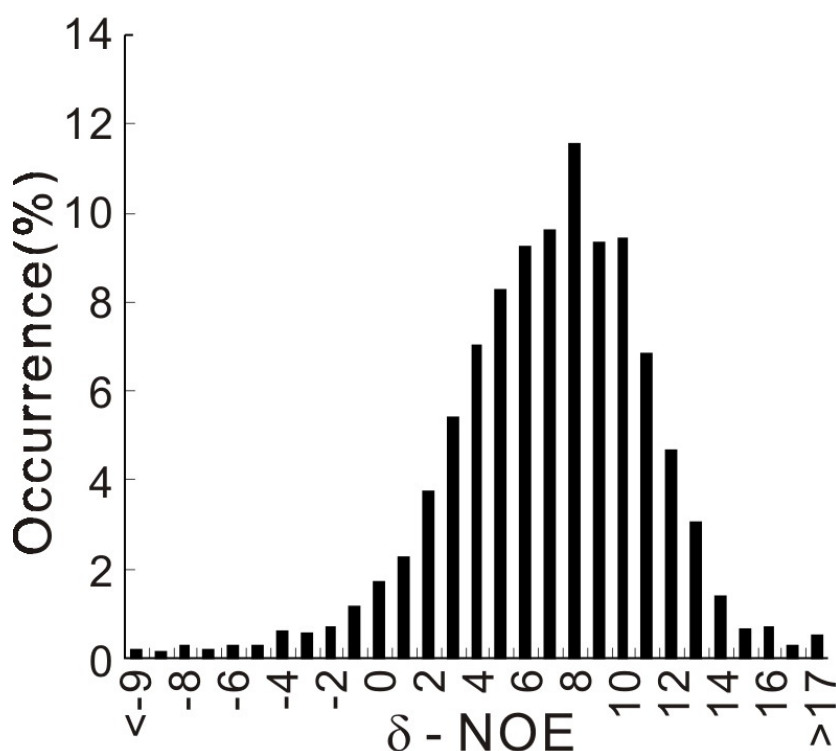
Here we describe the cause, identification and correction of the initial wrong assignment. The amide of Asp58 (cluster X209) displayed two long-range NOEs from its interactions with V8H<sup>1</sup> and V8H<sup>β</sup> (Figure 4.5 a). We initially identified these as sequential NOEs and incorrectly assigned them as the sequential spin-system members of Asp58, because coincidentally, Val8 and Pro57 have identical C<sup>α</sup> chemical shifts (Figure 4.5 a, b). Subsequently we incorrectly typed the spin-system as valine. The intra-residue spin-system of Val8 matched the spin-system of Asp58 better than that of Ile9 (Figure 4.5 a-c). Additionally, cluster X209 was initially assigned as Ile9 (note that Lys6-Leu7-Val8 was an assigned fragment). We did not detect the error immediately because X209 lacked intra-residue NOEs and was located at the fragment terminus. By checking the cluster of Ile9 (X11), we found that X11 contained two spin-systems definitely corresponding to valine and isoleucine. In the protein sequence, there are only two Val-Ile or Ile-Val segments: Val8-Ile9 and Val347-

Ile348. Because we had already assigned Val347 and Ile348, the previous assignment for cluster X209 was very likely incorrect. Moreover, the cluster of Ile9 could be connected with Trp10 and then extended to Asn12. Thus, the correct assignment for Ile9 was determined to be cluster X11. This example confirmed that errors made in the initial stage can be detected and corrected in the later stage of the sequential assignment.

#### 4.2.1.4 Resolution of ambiguity in connectivity

Owing to the presence of incomplete spin-systems, ambiguity on connectivity always exists (Figure 4.5 d-f). Fortunately, we can use the NOE peaks that cannot be identified as spin-system members to resolve this ambiguity, as shown below. Any two amides may simultaneously correlate with a set of CH, CH<sub>2</sub> and CH<sub>3</sub> groups via NOE interactions, that is, they may share a set of common NOEs. Inter-proton distances statistics done by our software (Chapter 5) indicated that amides *i* and *j* are more likely to have a sequential relationship when amide *i* shares a larger number of common NOEs with amide *j* than with other amides. The probability that two non-adjacent amides share the largest number of common NOEs and have matched C<sup>α</sup> chemical shifts in their HNCA correlations as well, is very low (<7%, Figure 4.6). Therefore, two clusters are most likely to have a sequential relationship when they share the largest number of common NOEs and also have matched C<sup>α</sup> shifts.

Using common NOEs, we resolved many ambiguities in connectivity, especially for the spin-systems containing only C<sup>α</sup>H<sup>α</sup> or no CH spin-pair (Figure 4.9). For instance, the intra-residue spin-system of Met55 of the β-chain of



**Figure 4.6** Distribution of  $\delta$ -NOE that reflects the difference in the number of common NOEs shared by two adjacent amide protons and those by two non-adjacent amides.

Two  $\delta$ -NOEs are defined for each amide:  $\delta_{i,i-1} = C_{i,i-1} - C_{i,k}^{\max}$  and  $\delta_{i,i+1} = C_{i,i+1} - C_{i,k}^{\max}$ , where  $C_{i,j}$  ( $j = i-1, i+1, k \neq i \pm 1$ ) is the number of CH spin-pairs that simultaneously correlate with amides  $i$  and  $j$  via NOE interactions, i.e., both the distances between CH and  $N_iH_i$  protons and between CH and  $N_jH_j$  protons are  $< 4.5 \text{ \AA}$ . We also denote  $C_{i,j}$  as the number of common NOEs shared by amides  $i$  and  $j$ .  $C_{i,k}^{\max}$  is the number of common NOEs shared by amides  $i$  and  $k$ , which is larger than those shared by amide  $i$  and all other non-adjacent amides. At the same time, residues  $i$  and  $k$  meet conditions:  $|C_{i-1}^{\alpha} - C_{k-1}^{\alpha}| < 0.5 \text{ ppm}$  for  $\delta_{i,i+1}$  and  $|C_{i-1}^{\alpha} - C_k^{\alpha}| < 0.5 \text{ ppm}$  for  $\delta_{i,i-1}$ , implying that amide  $i$  and amide  $k$  have matched  $C^{\alpha}$  chemical shifts in their HNCA correlations ( $C_{i-1}^{\alpha}-N_iH_i$  and  $C_{k-1}^{\alpha}-N_kH_k$  correlations for  $\delta_{i,i+1}$ ;  $C_{i-1}^{\alpha}-N_iH_i$  and  $C_k^{\alpha}-N_kH_k$  correlations for  $\delta_{i,i-1}$ ). If an amide is located in a terminus or adjacent to a proline, one  $\delta$ -NOE is calculated.  $\sim 93\%$  residues have a  $\delta$ -NOE value larger than zero. We derived the result from 3100 residues in 15 proteins whose x-ray or NMR structures and NMR resonance assignments are available.

HbCO A contained only one  $C^\alpha H^\alpha$  spin-pair as a result of no  $C^\alpha-H^\epsilon C^\epsilon$  or  $C^\alpha-H^\beta C^\beta$  TOCSY correlation. On the basis of  $C^\alpha$  and  $C^\alpha H^\alpha$  chemical shifts, the spin-system matched two clusters (X6 and X92) equally well (Figure 4.5 d-f). When considering all NOEs in the clusters, the correct path to X6 (Gly56) had six matched NOEs, whereas the incorrect one had only one.

On the basis of the mapped fragments, we assigned nearly all of the unmapped fragments and clusters with ambiguous connectivities in an iterative manner using sequence information and common NOEs (Table 4.1). In the end, the completeness of the backbone assignment was the same as that obtained with the conventional approach for DdCAD-1 (Lin, Huang et al. 2004). Only a few clusters for MBP remained unassigned because of very weak or no sequential HNCA correlations and missing HC-NH NOEs. Nevertheless, we assigned ten more residues for MBP than was previously done with triple-resonance experiments recorded on a  $^2H, ^{13}C, ^{15}N$ -labeled sample at 37°C (Gardner, Zhang et al. 1998). The backbone assignments for the  $\alpha$  and  $\beta$  chains of HbCO A obtained here were slightly less complete than those obtained using the conventional approach with a perdeuterated sample (Lukin, Kontaxis et al. 2004). Twelve and two residues displayed no N-H HSQC correlations for the  $\alpha$  and  $\beta$  chains of HbCO A, respectively, and thus could not be assigned here.

The intensity ratio of the sequential to intra-residue HNCA peak is normally  $<1$ . As found in the proteins examined here,  $\sim 1.5\%$  residues (Figure 4.9) showed intensity ratios  $>1$ , but no two adjacent residues simultaneously displayed such abnormal ratios. Thus, this can be used as a rule to assess the correctness of backbone assignments. If the amino-acid types determined from

the side-chain resonances of the residues are consistent with the actual residue types in the protein sequences, the assignments are further confirmed. This can be done during the process of side chain assignment.

### 4.2.2 Side-chain assignment

After obtaining backbone assignments, we assigned ~96%, ~91% and ~80% of aliphatic side chain resonances from the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY and 3D MQ-CCH-TOCSY spectra for DdCAD-1, MBP and HbCO A, respectively, using our previously established strategy (Lin, Xu et al. 2006). We obtained more assignments from  $^{13}\text{C}$ ,  $^{13}\text{C}$ -edited NOESY spectra based on initial NMR structures. We assigned many aromatic spins for MBP and HbCO A from the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited and  $^{13}\text{C}$ ,  $^{13}\text{C}$ -edited NOESY spectra. We also assigned some aromatic spins yielding weak or overlapped resonances in the 4D spectra with a more sensitive 3D  $^{13}\text{C}$ - or  $^{15}\text{N}$ -edited NOESY experiment (Lin, Xu et al. 2006).

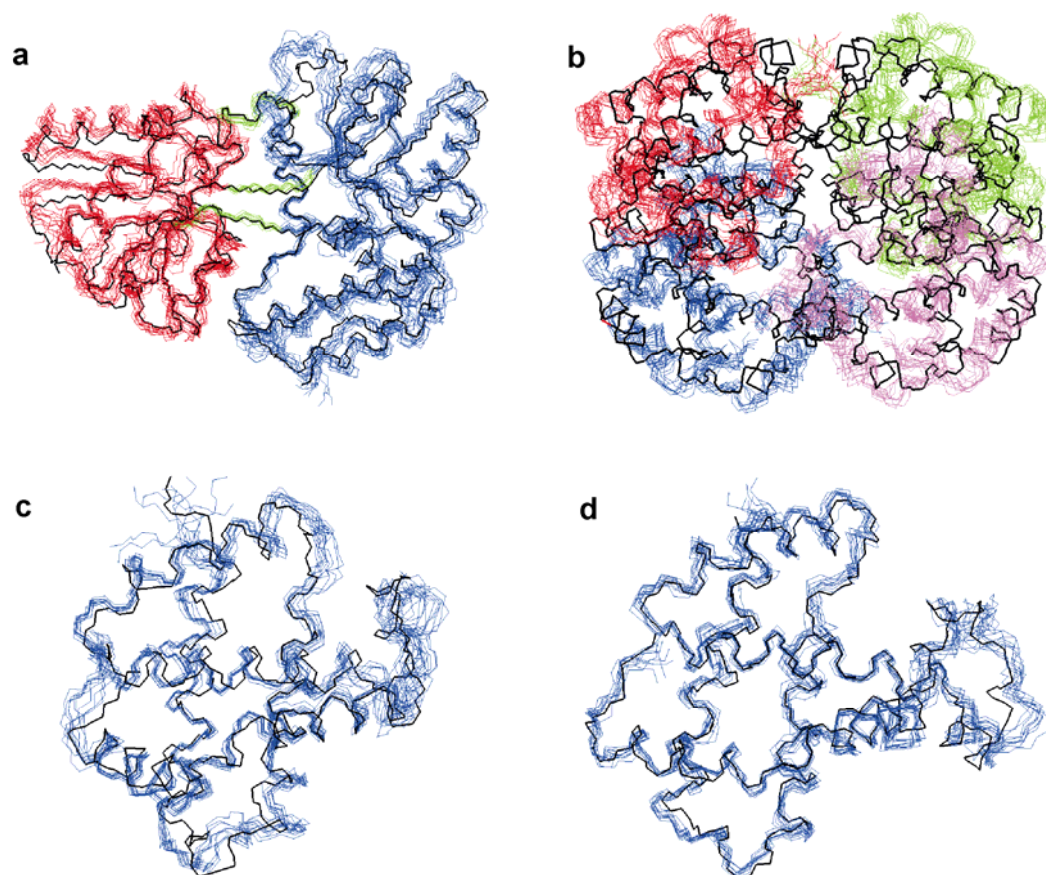
## 4.3 NOE assignment and structure determination

We unambiguously assigned 164  $\text{N}_i\text{H}_i\text{-C}_j\text{H}_j$ , 171  $\text{C}_i\text{H}_i\text{-C}_j\text{H}_j$  and 15  $\text{NH}_i\text{-NH}_j$  long-range NOEs ( $|i-j| \geq 5$ ) from the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited and  $^{13}\text{C}$ ,  $^{13}\text{C}$ -edited NOESY and 3D  $^{13}\text{C}$ - or  $^{15}\text{N}$ -edited NOESY spectra of MBP based on the assignments of backbone, aliphatic side chain and tryptophan  $\text{H}^{\epsilon 1}/\text{H}^{\delta 1}$  spins. The first-round structures determined from these NOEs (Table 4.2) had a pair-wise root-mean-square deviation (RMSD) of ~5.2 Å. We used the derived structure for the assignment of additional NOE peaks in an iterative process. In the end, we assigned 841 and 1561 medium- and long-range NOE distance restraints from

the NOESY spectra, and used them for the final structure calculation. The average RMSDs of the 10 structures to the mean structure were 1.52, 1.22, and 1.34 Å for all heavy atoms of the entire molecule, amino-terminal domain and carboxy-terminal domain, respectively (Figure 4.7 a). The RMSD for all heavy atoms between the mean structure and the X-ray structure was 1.95 Å (Table 4.2), indicating that the relative orientation of the two domains of MBP is slightly different in solution and crystal states (Skrynnikov, Goto et al. 2000). The quality of the MBP structures obtained here is comparable to the one very recently determined using an extremely expensive sample prepared with the stereo-array isotope labelling (SAIL) technique (Kainosho, Torizawa et al. 2006).

We also determined the structures of the  $\alpha$  chain,  $\beta$  chain and tetramer of HbCO A from intra- and intermolecular NOEs that were assigned in an iterative procedure as described above. The average RMSDs of the 10 lowest-energy structures to the mean structure were 1.84, 1.83 and 2.20 Å for all heavy atoms of the  $\alpha$  chain,  $\beta$  chain and tetramer, respectively (Figure 4.7 b and Table 4.3). The NMR structures of each chain and  $\alpha$ - $\beta$  dimer agreed very well with the x-ray R2 structure (Figure 4.7 c,d). Our tetramer structure differed notably from the x-ray structure notably (RMSD, 4.0 Å), but it was closer to the solution quaternary structure determined from X-ray tertiary structures and RDCs (Lukin, Kontaxis et al. 2003) (RMSD, 3.4 Å). To reveal the structural difference in solution and crystal states, further structure refinement with RDCs will be necessary.





**Figure 4.7 Comparison of structures determined by NMR and x-ray methods.**

The ten lowest-energy NMR structures of MBP (a) and the tetramer (b),  $\alpha$ -chain (c) and  $\beta$ -chain (d) of HbCO A, calculated on the basis of distance restraints and backbone torsion angle restraints, superimposed with x-ray structures. For MBP, bonds of the N-terminal domain (residues 1-109 and 264-309) are shown in red, C-terminal domain (residues 114-258 and 316-370) in blue and linkers (residues 110-113, 259-263 and 310-315) in green. For HbCO A tetramer,  $\alpha_1$ ,  $\beta_1$ ,  $\alpha_2$  and  $\beta_2$  chains are indicated in green, magenta, blue and red, respectively. The backbones of the x-ray structures (PDB codes, 1DMB for MBP and 1BBB for HbCO A) are shown as thick black lines.

**Table 4.2 Structural statistics for the final 10 conformers of MBP<sup>a</sup>**

|   |  |
|---|--|
| Distance restraints                                 |  |
| Intraresidue ( $i-j = 0$ )                          | 395 (119) <sup>b</sup>                           |
| Sequential ( $ i-j  = 1$ )                          | 929 (610)  |
| Medium range ( $2 \leq  i-j  \leq 4$ )              | 841 (196)  |
| Long range ( $ i-j  \geq 5$ )                       | 1561 (350)                                       |
| Total   | 3726 (1275)                                      |
| Dihedral angle restraints                           |  |
| $\phi$  | 235 (179)  |
| $\psi$  | 235 (179)  |
| Average rmsd to the mean structure (Å) <sup>c</sup> |  |
| Global  | $1.15 \pm 0.18$ ( $1.52 \pm 0.14$ ) <sup>d</sup> |
| N domain  | $0.78 \pm 0.07$ ( $1.22 \pm 0.07$ )              |
| C domain  | $0.87 \pm 0.12$ ( $1.34 \pm 0.09$ )              |
| rmsd of the mean structure to 1DMB <sup>e</sup>     |  |
| Global  | 1.41 (1.95)                                      |
| N domain  | 1.10 (1.80)                                      |
| C domain  | 1.12 (1.74)                                      |
| $\phi/\psi$ space <sup>f</sup>                      |  |
| Most favored region (%)                             | $81.2 \pm 1.7$                                   |
| Additionally allowed region (%)                     | $17.5 \pm 1.6$                                   |
| Generously allowed region (%)                       | $1.2 \pm 0.4$                                    |
| Disallowed region (%)                               | $0.1 \pm 0.1$                                    |
| rmsd from covalent geometry                         |  |
| Bonds (Å) <sup>g</sup>                              | $0.0009 \pm 0.0001$                              |
| Angles (deg.)                                       | $0.2629 \pm 0.0083$                              |
| Improper (deg.)                                     | $0.1063 \pm 0.0164$                              |
| rmsd from experimental restraints                   |  |
| NOEs (Å)  | $0.0062 \pm 0.0014$                              |
| Dihedral angles (deg.)                              | $0.0833 \pm 0.0374$                              |

<sup>a</sup> Selected from 40 calculated conformers according to overall energy. The structures have been deposited in PDB (code: 2H25).

<sup>b</sup> Number in parentheses refer to restraints used to calculate the initial structures, which were used to assign more NOE peaks and to decide whether the TALOS predictions could be used.

<sup>c</sup> Calculated with MOLMOL (Koradi, Billeter et al. 1996). The following residues were used in the rmsd calculation: Global: 6-235, 241-370; N domain: 6-109, 264-309; C-domain: 114-235, 241-258, 316-370.

<sup>d</sup> Averages are over heavy backbone atoms (all heavy atoms).

<sup>e</sup> ref (Sharff, Rodseth et al. 1993).

<sup>f</sup> Calculated with PROCHECK-NMR (Laskowski, Rullmann et al. 1996).

<sup>g</sup> Evaluated by CNS.

**Table 4.3 Structural statistics for the final 10 conformers of HbCO A<sup>a</sup>**

|   |  |                |
|---|--|----------------|
| Distance restraints                                 |  |                |
| subunit   | $\alpha$ -chain                                  | $\beta$ -chain |
| Intra-residue ( $i-j = 0$ )                         | 501  | 650            |
| Sequential ( $ i-j  = 1$ )                          | 361  | 458            |
| Medium range ( $2 \leq  i-j  \leq 4$ )              | 246  | 358            |
| Long range ( $ i-j  \geq 5$ )                       | 153  | 251            |
| Heme-subunit  | 55   | 50             |
| Total   | 1316   | 1767           |
| Inter-subunit                                       | 87   |                |
| Dihedral angle restraints                           |  |                |
| $\phi$  | 87   | 89             |
| $\psi$  | 87   | 89             |
| Average rmsd to the mean structure (Å) <sup>b</sup> |  |                |
| Global  | $1.58 \pm 0.27$ ( $2.20 \pm 0.24$ ) <sup>c</sup> |                |
| $\alpha$ -chain                                     | $1.12 \pm 0.16$ ( $1.84 \pm 0.16$ )              |                |
| $\beta$ -chain                                      | $1.00 \pm 0.19$ ( $1.83 \pm 0.19$ )              |                |
| rmsd of the mean structure to 1BBB                  |  |                |
| Global  | 4.02 (4.30)                                      |                |
| $\alpha$ -chain                                     | 1.59 (2.30)                                      |                |
| $\beta$ -chain                                      | 1.17 (1.78)                                      |                |
| $\alpha_1\beta_1$ two chains                        | 1.65 (2.22)                                      |                |
| $\phi/\psi$ space <sup>d</sup>                      |  |                |
| Most favored region (%)                             | 75.7   |                |
| Additionally allowed region (%)                     | 20.1   |                |
| Generously allowed region (%)                       | 3.5  |                |
| Disallowed region (%)                               | 0.7  |                |
| rmsd from covalent geometry                         |  |                |
| Bonds (Å)   | $0.0017 \pm 0.0000$                              |                |
| Angles (deg.)                                       | $0.2976 \pm 0.0069$                              |                |
| Impropers (deg.)                                    | $0.2528 \pm 0.0007$                              |                |
| rmsd from experimental restraints                   |  |                |
| NOEs (Å)  | $0.0143 \pm 0.0004$                              |                |
| Dihedral angles (deg.)                              | $0.2602 \pm 0.0297$                              |                |

<sup>a</sup> Selected from 100 calculated conformers according to overall energy. The structures have been deposited in PDB (code: 2H35).

<sup>b</sup> Calculated with MOLMOL(Koradi, Billeter et al. 1996). The following residues were used in the RMSD calculation: 4-138 for each  $\alpha$ -chain and 4-141 for each  $\beta$ -chain.

<sup>c</sup> Averages are over heavy backbone atoms (all heavy atoms).

<sup>d</sup> Calculated with PROCHECK-NMR(Laskowski, Rullmann et al. 1996).

**Table 4.4 Experimental parameters**

| TROSY-HNCA(Yang and Kay 1999)                                 |            |  |  |  |   |                           |
|---|------------|--|--|--|---|---------------------------|
| sample  | field      | F <sub>1</sub> <sup>13</sup> C<br>ni, sw, at | F <sub>2</sub> <sup>15</sup> N<br>ni, sw, at | F <sub>3</sub> <sup>1</sup> H<br>ni, sw, at  | scans/d1<br>total time                      |                           |
| DdCAD-1   | 800<br>MHz | 27,<br>3923.5 Hz,<br>6.6 ms                  | 65,<br>2594.9 Hz,<br>24.7 ms                 | 640,<br>11160.7 Hz,<br>57.3 ms               | 8/1.1 s<br>19 hr                            |                           |
| MBP   | 800<br>MHz | 24,<br>3923.5 Hz,<br>5.9 ms                  | 65,<br>2594.9 Hz,<br>24.7 ms                 | 640,<br>11160.7 Hz,<br>57.3 ms               | 16/1.1 s<br>33 hr                           |                           |
| HbCO A<br>α-chain   | 800<br>MHz | 25,<br>3923.5 Hz,<br>6.1 ms                  | 53,<br>2108.3 Hz,<br>24.7 ms                 | 640,<br>12820.5 Hz,<br>50.0 ms               | 24/1.0 s<br>39 hr                           |                           |
| HbCO A<br>β-chain   | 800<br>MHz | 25,<br>3923.5 Hz,<br>6.1 ms                  | 53,<br>2108.3 Hz,<br>24.71 ms                | 640,<br>12820.5 Hz,<br>50.0 ms               | 24/1.0 s<br>39 hr                           |                           |
| TROSY-HNCO(Yang and Kay 1999)                                 |            |  |  |  |   |                           |
| sample  | field      | F <sub>1</sub> <sup>13</sup> C<br>ni, sw, at | F <sub>2</sub> <sup>15</sup> N<br>ni, sw, at | F <sub>3</sub> <sup>1</sup> H<br>ni, sw, at  | scans/d1<br>total time                      |                           |
| MBP   | 800<br>MHz | 23,<br>2414.9 Hz,<br>9.1 ms                  | 65,<br>2594.7 Hz,<br>24.7 ms                 | 640,<br>12820.5 Hz,<br>50.0 ms               | 8/1.0 s<br>13 hr                            |                           |
| HbCO A<br>α-chain   | 800<br>MHz | 23,<br>2414.9 Hz,<br>9.1 ms                  | 53,<br>2108.3 Hz,<br>24.7 ms                 | 640,<br>12820.5 Hz,<br>50.0 ms               | 8/1.0 s<br>12 hr                            |                           |
| HbCO A<br>β-chain   | 800<br>MHz | 23,<br>2414.9 Hz,<br>9.1 ms                  | 53,<br>2108.3 Hz,<br>24.7 ms                 | 640,<br>12820.5 Hz,<br>50.0 ms               | 8/1.0 s<br>12 hr                            |                           |
| 4D <sup>13</sup> C, <sup>15</sup> N-edited NOESY (Figure 4.1) |            |  |  |  |   |                           |
| Sample<br>mixing<br>time                                      | field      | F <sub>1</sub> <sup>1</sup> H<br>ni, sw, at  | F <sub>2</sub> <sup>13</sup> C<br>ni, sw, at | F <sub>3</sub> <sup>15</sup> N<br>ni, sw, at | F <sub>4</sub> <sup>1</sup> H<br>ni, sw, at | scans/d1<br>total<br>time |
| DdCAD-1<br>75 ms  | 500<br>MHz | 39,<br>3300.0 Hz,<br>11.5 ms                 | 24,<br>4000.0 Hz,<br>5.8 ms                  | 20,<br>1444.8 Hz,<br>13.1 ms                 | 512,<br>8012.8 Hz,<br>63.8 ms               | 2/1.0 s<br>92 hr          |
| MBP<br>60 ms  | 800<br>MHz | 30,<br>6053.3 Hz,<br>4.8 ms                  | 24,<br>4426.7 Hz,<br>5.2 ms                  | 24,<br>1601.5 Hz,<br>14.4 ms                 | 640,<br>11160.7<br>Hz, 57.3<br>ms           | 4/0.9 s<br>160 hr         |
| HbCO A<br>α-chain<br>50 ms                                    | 500<br>MHz | 18,<br>3541.1 Hz,<br>4.8 ms                  | 17,<br>2515.7 Hz,<br>6.4 ms                  | 18,<br>1216.7 Hz,<br>14.0 ms                 | 512,<br>8012.8 Hz,<br>63.8 ms               | 8/1.0 s<br>110 hr         |
| HbCO A<br>β-chain<br>50 ms                                    | 500<br>MHz | 18,<br>3541.1 Hz,<br>4.8 ms                  | 17,<br>2515.7 Hz,<br>6.4 ms                  | 18,<br>1216.7 Hz,<br>14.0 ms                 | 512,<br>8012.8 Hz,<br>63.8 ms               | 8/1.0 s<br>110 hr         |
| 3D MQ-CCH-TOCSY(Zheng, Giovannelli et al. 2004)               |            |  |  |  |   |                           |
| sample  | field      | F <sub>1</sub> <sup>13</sup> C<br>ni, sw, at | F <sub>2</sub> <sup>13</sup> C<br>ni, sw, at | F <sub>3</sub> <sup>1</sup> H<br>ni, sw, at  | scans/d1<br>total time                      |                           |
| DdCAD-1   | 800<br>MHz | 75,<br>12878.3<br>Hz, 5.7 ms                 | 34,<br>5030.2 Hz,<br>6.6 ms                  | 640,<br>11160.7 Hz,<br>57.3 ms               | 8/0.9 s<br>23 hr                            |                           |
| MBP   | 800<br>MHz | 95,<br>13297.9                               | 38,<br>5231.5 Hz,                            | 640,<br>11160.7 Hz,                          | 8/1.0 s<br>36 hr                            |                           |

|  |            |   |  |  |   |                        |
|--|------------|---|--|--|---|------------------------|
|  |            | Hz, 7.1 ms                                | 7.1 ms                                       | 57.3 ms                                      |   |                        |
| HbCO A<br>$\alpha$ -chain  | 800<br>MHz | 105,<br>12073.6<br>Hz, 8.6 ms             | 35,<br>4024.1 Hz,<br>8.4 ms                  | 640,<br>11160.7 Hz,<br>57.3 ms               |   | 8/1.2 s<br>42 hr       |
| HbCO A<br>$\beta$ -chain   | 800<br>MHz | 105,<br>12073.6<br>Hz, 8.6 ms             | 35,<br>4024.1 Hz,<br>8.4 ms                  | 640,<br>11160.7 Hz,<br>57.3 ms               |   | 8/1.2 s<br>42 hr       |
| 4D $^{13}\text{C}$ , $^{13}\text{C}$ -edited NOESY(Clore, Kay et al. 1991)   |            |   |  |  |   |                        |
| Sample<br>mixing<br>time   | Field      | F <sub>1</sub> $^1\text{H}$<br>ni, sw, at | F <sub>2</sub> $^{13}\text{C}$<br>ni, sw, at | F <sub>3</sub> $^{13}\text{C}$<br>ni, sw, at | F <sub>4</sub> $^1\text{H}$<br>ni, sw, at | scans/d1<br>total time |
| MBP<br>50 ms   | 800<br>MHz | 30,<br>6053.3 Hz,<br>4.8 ms               | 24,<br>4426.7 Hz,<br>5.2 ms                  | 23,<br>4426.7 Hz,<br>5.0 ms                  | 512,<br>9615.4 Hz,<br>53.2 ms             | 4/1.0 s<br>162 hr      |
| HbCO A<br>$\alpha$ -chain<br>40 ms   | 800<br>MHz | 30,<br>6053.3 Hz,<br>4.8 ms               | 24,<br>4426.7 Hz,<br>5.2 ms                  | 23,<br>4426.7 Hz,<br>5.0 ms                  | 512,<br>9615.4 Hz,<br>53.2 ms             | 4/1.0 s<br>162 hr      |
| HbCO A<br>$\beta$ -chain<br>40 ms  | 800<br>MHz | 30,<br>6053.3 Hz,<br>4.8 ms               | 24,<br>4426.7 Hz,<br>5.2 ms                  | 23,<br>4426.7 Hz,<br>5.0 ms                  | 512,<br>9615.4 Hz,<br>53.2 ms             | 4/1.0 s<br>162 hr      |
| 3D $^{13}\text{C}/^{15}\text{N}$ -edited NOESY(Lin, Xu et al. 2006)  |            |   |  |  |   |                        |
| Sample<br>mixing<br>time   | Field      | F <sub>1</sub> $^1\text{H}$<br>ni, sw, at | F <sub>2</sub> $^{13}\text{C}$<br>ni, sw, at | F <sub>2</sub> $^{15}\text{N}$<br>ni, sw, at | F <sub>3</sub> $^1\text{H}$<br>ni, sw, at | scans/d1<br>total time |
| MBP<br>50 ms   | 800<br>MHz | 80,<br>9601.5 Hz,<br>8.2 ms               | 30,<br>5634.6 Hz,<br>5.1 ms                  | 30,<br>1601.5 Hz,<br>18.1 ms                 | 640,<br>11160.7Hz,<br>57.3 ms             | 8/1.3 s<br>60 hr       |
| HbCO A<br>$\alpha$ -chain<br>40 ms   | 800<br>MHz | 106,<br>9601.5 Hz,<br>10.9 ms             | 38,<br>5231.5 Hz,<br>7.1 ms                  | 38,<br>2107.9 Hz,<br>17.6 ms                 | 640,<br>13550.1Hz,<br>47.1 ms             | 4/1.2 s<br>43 hr       |
| HbCO A<br>$\beta$ -chain<br>40 ms  | 800<br>MHz | 106,<br>9601.5 Hz,<br>10.9 ms             | 38,<br>5231.5 Hz,<br>7.1 ms                  | 38,<br>2107.9 Hz,<br>17.6 ms                 | 640,<br>13550.1Hz,<br>47.1 ms             | 4/1.2 s<br>43 hr       |
| 3D $^{13}\text{C}/^{15}\text{N}$ F <sub>1</sub> filtered, F <sub>2</sub> -edited NOESY(Zwahlen, Legault et al. 1997) |            |   |  |  |   |                        |
| Sample<br>mixing<br>time   | Field      | F <sub>1</sub> $^1\text{H}$<br>ni, sw, at | F <sub>2</sub> $^{13}\text{C}$<br>ni, sw, at | F <sub>2</sub> $^{15}\text{N}$<br>ni, sw, at | F <sub>3</sub> $^1\text{H}$<br>ni, sw, at | scans/d1<br>total time |
| HbCO A<br>$\alpha$ -chain<br>50 ms   | 800<br>MHz | 55<br>9596.9 Hz,<br>5.6 ms                | 38,<br>5230.1 Hz,<br>7.1 ms                  | 38,<br>2107.9 Hz,<br>17.6 ms                 | 512,<br>12820.5Hz,<br>39.9 ms             | 16/1.0 s<br>82 hr      |
| HbCO A<br>$\beta$ -chain<br>50 ms  | 800<br>MHz | 55<br>9596.9 Hz,<br>5.6 ms                | 38,<br>5230.1 Hz,<br>7.1 ms                  | 38,<br>2107.9 Hz,<br>17.6 ms                 | 512,<br>12820.5Hz,<br>39.9 ms             | 16/1.0 s<br>82 hr      |

ni: complex points; sw: spectral width; at: maximum acquisition time; d1: interscan delay

For DdCAD-1, the  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY was recorded in a non-constant-time mode.

## 4.4 Discussion and conclusion

The assignment strategy described here used information from through-bond correlation experiments to filter intra-residue and sequential correlations obtained from through-space correlation experiments, and then used these filtered correlations to identify adjacent spin-systems in clusters and to classify them by residue type. The classical approach (Wagner and Wuthrich 1982) also uses a NOESY- and TOCSY-based strategy, but it does not separate sequential NOEs from other inter-residue NOEs. Also, the classical  $^1\text{H}$ - $^1\text{H}$  TOCSY experiment is insensitive for proteins larger than 10 kDa. The experiments used here for backbone assignment are quite sensitive for nondeuterated proteins with correlation times less than 30 ns as illustrated by numerical simulations (Figure 4.8) and our experimental results (Figure 4.3). Using these experimental strategy, one can identify most spin-systems and classify them by residue type, as most residues give rise to one or more TOCSY correlations for the proteins studied here.

In the application of this strategy to monomeric proteins larger than MBP, there may be more spin-systems lacking intra-residue or sequential correlations, mainly because of TOCSY signal overlap. When these spin-systems distribute randomly over the sequence, it is still possible to construct fragments from clusters based on sequential connectivities and common NOEs shared by two amides. Although some fragments contain less than four spin-systems, one can map such short fragments onto the protein sequence because the residue types of many spin-systems can be classified, as demonstrated on the proteins studied here. Additionally, one can use 3D TROSY-HN(CO)CA (Yang and Kay 1999),

as demonstrated in Chapter 7, to identify more sequential correlations and thus to improve assignments of monomeric proteins with greater than 400 residues.

Our approach may take longer time than the conventional approach for backbone assignment because the identification of spin-systems and assembly/mapping of clusters are time-consuming. One can overcome this drawback by using our automatic/semi-automatic software XYZ4D (Chapter 7). Compared with the conventional approach (Salzmann, Pervushin et al. 2000; Tugarinov, Choy et al. 2005), the time for data collection with our strategy can be substantially reduced as fewer experiments are required. The two 4D NOESY experiments are very time-consuming (each ~7 days), but in principle one can record the experiments simultaneously with a total experimental time less than 8 days using a time-sharing scheme (Lin, Xu et al. 2006) and a multiway decomposition method (Tugarinov, Kay et al. 2005). One can also apply our approach to smaller proteins by shortening experimental time by using fast multidimensional NMR spectroscopy techniques (Coggins, Venters et al. 2005; Tugarinov, Kay et al. 2005).

In summary, we have shown that backbone and side-chain assignments and structure determination of proteins as large as HbCO A and MBP can be achieved with uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled samples without the use of deuteration. Our strategy uses four indispensable experiments (4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY, 4D  $^{13}\text{C}$ ,  $^{13}\text{C}$ -edited NOESY, 3D TROSY-HNCA and 3D MQ-CCH-TOCSY) and two supplementary experiments (3D TROSY-HNCO and  $^{13}\text{C}/^{15}\text{N}$ -edited NOESY). Recording intermolecular NOEs is necessary for solving structures of protein-protein complexes. As these experiments are still sensitive enough to

obtain a structure for the tetrameric complex of HbCO A, it is expected that the approach demonstrated here can be applied to proteins or protein complexes smaller than 65 kDa. With the available assignments, one can measure RDCs from nondeuterated samples and then refine the structures derived from NOEs against RDCs. Because sample preparation is much less demanding, our approach will enhance the application of NMR spectroscopy to larger proteins.

## 4.5 Materials and methods

### 4.5.1 Protein samples and NMR spectroscopy

We performed experiments on uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled DdCAD-1 (~0.8 mM, 5%  $\text{D}_2\text{O}$ , pH 6.5), MBP (~1.2 mM, 10%  $\text{D}_2\text{O}$ , pH 7) and HbCO A (labeled  $\beta$  chain and unlabeled  $\alpha$  chain,  $\sim 2 \times 0.6$  mM  $\beta$ -chain; unlabeled  $\beta$  chain and labeled  $\alpha$  chain,  $\sim 2 \times 0.5$  mM  $\alpha$  chain, 5%  $\text{D}_2\text{O}$ ; pH 7) at 30 °C on a Bruker 500 MHz spectrometer equipped with a cryoprobe and a Bruker 800 MHz spectrometer equipped with a normal triple-resonance probe. We collected TROSY-HNCA, MQ-CCH-TOCSY,  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY (Figure 4.1),  $^{13}\text{C}$ ,  $^{13}\text{C}$ -edited NOESY and time-sharing  $^{13}\text{C}$ - or  $^{15}\text{N}$ -edited NOESY for each sample (experimental parameters are given in Table 4.4). We acquired  $^{13}\text{C}/^{15}\text{N}$   $F_1$ -filtered and  $F_2$ -edited NOESY for obtaining intermolecular NOE restraints for HbCO A. We processed NMR spectra with NMRPipe (Delaglio, Grzesiek et al. 1995) and analyzed with Sparky (Goddard, T.D. & Kneller, D.G. SPARKY 3, University of California, San Francisco) and NMRspy (Chapter 6). We used some in-house-written extensions in Sparky and XYZ4D (Chapter 8) to facilitate the assignment.



### 4.5.2 Identifying spin-systems

All the peaks, correlations or spins mentioned in this section are within the same cluster unless otherwise indicated. The aliphatic spins within the same residue form a spin-system (Figure 4.4 a). We used each HNCA correlation to construct one initial spin-system. When only one HNCA cross peak existed in one cluster because of overlap of intra-residue and sequential peaks, or because of a lack of the sequential peak, we built one spin-system in this step. We obtained other spin-system members for each spin-system from HC-NH NOE peaks based on the fact that all of the spins in the same spin-system correlate with each other via MQ-CCH-TOCSY. Using  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY, TROSY-HNCA and MQ-CCH-TOCSY experiments, we identified many intra-residue and sequential NOEs to form spin-systems in the following steps.

First, when only one of the NOE peaks matched one of the HNCA correlations in  $\text{C}^\alpha$  chemical shift, we considered this NOE peak as either an intra-residue or sequential  $\text{H}^\alpha\text{C}^\alpha$ -HN correlation (Xu, Lin et al. 2005). We placed the matched NOE and HNCA correlations in the same spin-system. In the case which only one HNCA peak existed in a given cluster, we designated one of the two matched NOEs to the initial spin-system in the cluster if the  $\text{C}^\alpha$  chemical shift of the HNCA peak matched the  $^{13}\text{C}$  shifts of only two NOE peaks, and we used the other matched NOE and the existing HNCA peak to create another spin-system for the cluster.

Second, if the side-chain  $\text{C}^k\text{H}^k$  spin-pair of a  $\text{H}^k\text{C}^k$ -NH NOE correlation and a  $\text{C}^\alpha$  spin in a given cluster belong to the same spin-system,  $\text{C}^k\text{H}^k$  should

yield a correlation with this  $C^\alpha$  spin in the MQ-CCH-TOCSY spectrum. The  $C^\alpha H^\alpha$  spin-pair should also correlate with spin  $C^k$ . When we observed  $C^\alpha$ - $C^k H^k$  or  $C^k$ - $C^\alpha H^\alpha$  TOCSY correlations, we considered that the  $C^k H^k$  and  $C^\alpha$  spins belonged to the same spin-system. Note that some  $C^k$ - $C^\alpha H^\alpha$  correlations may not be observable because of the faster decay of  $C^\alpha/H^\alpha$  magnetization or the overlap of  $H^\alpha$  with intense  $H_2O$  resonances.

Third, when we could not assign a  $H^N C^N$ -NH NOE peak to a spin-system because of missing  $C^N$ - $C^\alpha H^\alpha$  or  $C^\alpha$ - $C^N H^N$  correlations, we examined the TOCSY correlations between the  $C^N H^N$  spin-pair with the  $C^k H^k$  spin pairs that were already assigned to a spin-system in the second step. If one or more  $C^k$ - $C^N H^N$  correlations existed, we considered that the  $C^N H^N$  and  $C^k H^k$  spin-pairs were in the same spin-system. After these three steps, we could also assign an ungrouped HC-NH NOE peak to a spin-system by examining all possibilities.

Finally, all of the CH spin pairs within the same spin-system must display consistent C-CH TOCSY correlations. We removed inconsistent spin pairs from spin-systems. Because some intra-residue and sequential HC-NH NOEs were not observable in the 4D NOESY and some C-CH resonances were overlapped or were undetectable in the 3D TOCSY, many spin-systems contained fewer spins than expected, that is, they were incomplete.

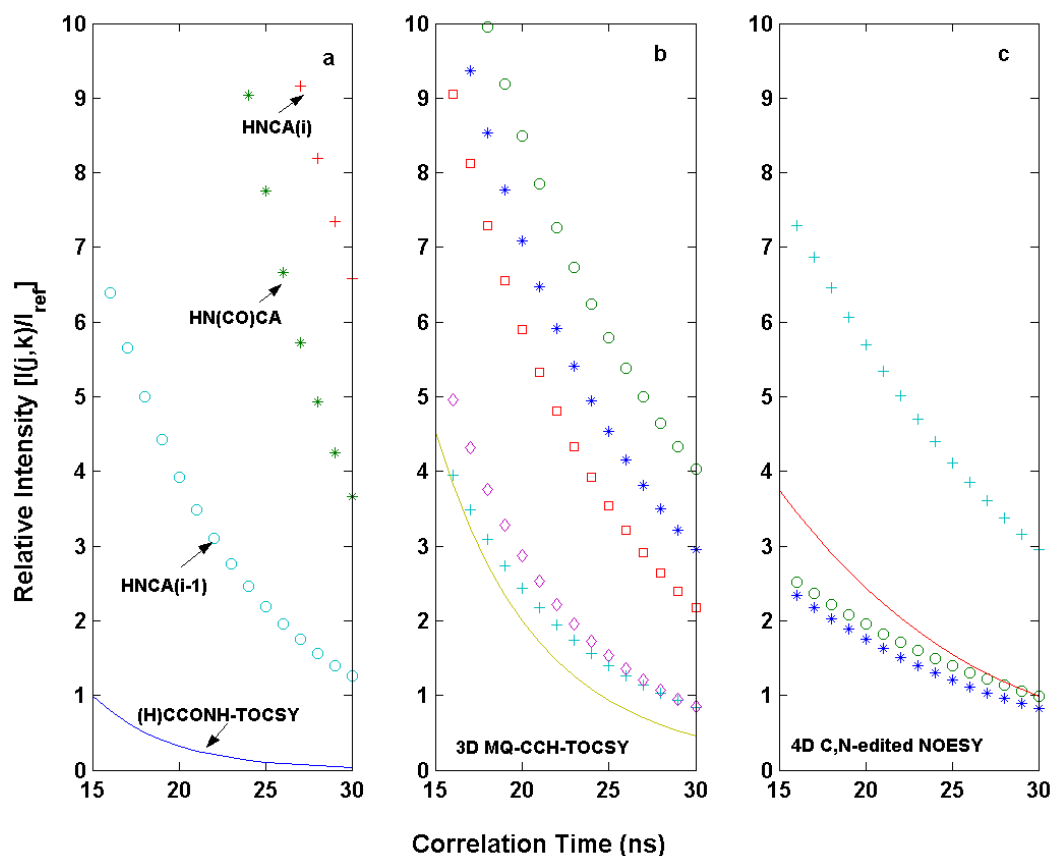
### 4.5.3 Structure calculation

We performed structure calculations for MBP with CNS (Brunger, Adams et al. 1998) using distance and dihedral restraints, which started from 40 extended conformers. We calculated structures of the  $\alpha$  chain,  $\beta$  chain and

tetramer of HbCO A with National Institute of Health X-PLOR software (Schwieters, Kuszewski et al. 2003) using distance and dihedral restraints and a combination of torsion-angle and Cartesian dynamics. We employed a calculation protocol similar to that described previously (Tugarinov, Choy et al. 2005). In the calculation of HbCO A tetramer, we used a special NCS (noncrystallographic symmetry) restraint to increase the rate of convergence. We predicted dihedral angles of backbone  $\phi$  and  $\psi$  using chemical shifts of  $C\alpha$ ,  $C\beta$ ,  $H\alpha$ , CO, and N of MBP and HbCO A with TALOS (Cornilescu, Delaglio et al. 1999). In the case of MBP, we removed the chemical shifts of MBP from the database. Only those angles for the residues in regions of secondary structures as predicted by chemical shift index (CSI) were used as restraints for initial structure calculations. Other predicted angles were also included for the final structure calculations, provided that they were consistent with the initial structures.

#### 4.5.4 Data deposition

We deposited the NMR spectroscopy assignments and coordinates of MBP and HbCO A in Biological Magnetic Resonance Bank (BMRB-7114 and BMRB-7125, respectively) and Protein Data Bank (2H25 and 2H35, respectively).



**Figure 4.8** Relative peak intensity ( $I(j,k)/I_{\text{ref}}$ ), as a function of overall correlation time ( $\tau_m$ ), calculated for different types of correlations in a number of 3D and 4D spectra.

$I_{\text{ref}}$  denotes the intensity of Leu  $C_{i-1}^{\delta}-N_iH_i$  correlation calculated for a 3D (H)CC(CO)NH-TOCSY experiment (Montelione, Lyons et al. 1992) recorded on a protein with a correlation time of 15 ns.  $I(j,k)$  represents the intensity of correlation  $j$  in experiment  $k$ . (a). Relative intensities for intraresidue (+) and sequential (o) HNCA correlations, HN(CO)CA correlation (\*) and Leu  $C_{i-1}^{\delta}-N_iH_i$  correlation (-). (b). Relative intensities for  $C^{\beta}(\text{CH})-C^mH^m$  (o),  $C^{\beta}(\text{CH}_2)-C^mH^m$  (□),  $C^{\alpha}-C^mH^m$  (\*),  $C^{\beta}(\text{CH})-C^{\alpha}H^{\alpha}$  (◇),  $C^{\beta}(\text{CH}_2)-C^{\alpha}H^{\alpha}$  (-) and  $C^m-C^{\alpha}H^{\alpha}$  (+) TOCSY correlations in a 3D MQ-CCH-TOCSY (Zheng, Giovannelli et al. 2004) spectrum, where superscript  $m$  denotes the methyl position. (c). Relative intensities for  $H^mC^m-NH$  (+),  $H^{\beta}C^{\beta}(\text{CH}_2)-NH$  (-),  $H^{\beta}C^{\beta}(\text{CH})-NH$  (o), and  $H^{\alpha}C^{\alpha}-NH$  (\*) NOE correlations in a 4D  $^{13}\text{C}, ^{15}\text{N}$ -edited NOESY spectrum. The proton distance between CH/CH<sub>2</sub>/CH<sub>3</sub> and NH groups were set to 3.0 Å. The two

protons in  $C^\beta H_2^\beta$  groups were assumed to have the identical chemical shift. The relaxation rates of NH and  $C^\alpha H^\alpha$  protons are 5%-10% larger for the residues in  $\beta$ -sheets than those in  $\alpha$ -helices of a fully protonated protein, implying that the experiments involving NH or  $C^\alpha H^\alpha$  spins are slightly more sensitive for  $\alpha$ -helical proteins. Therefore, only the results for residues in  $\beta$ -sheets are shown here. The fact that a 3D experiment is  $\sqrt{2}$  time more sensitive than a 4D experiment recorded with the same pulse sequence and experimental time was taken into consideration in the simulations.

$I(j,k)$  and  $I_{ref}$  were calculated by taking into account both J coupling and relaxation effects during each delay and each acquisition period of a pulse sequence. The J coupling constants used in the simulations were:  $^1J_{NH} = 92$  Hz;  $^1J_{NC\alpha} = 11$  Hz;  $^2J_{NC\alpha} = 6$  Hz;  $^1J_{NCO} = 15$  Hz;  $^1J_{C\alpha H\alpha} = 145$  Hz;  $^1J_{CmHm} = 125$  Hz (m denotes the methyl position);  $^1J_{CH} = 135$  Hz (for other CH/CH<sub>2</sub> groups);  $^1J_{C\alpha CO} = 55$  Hz; and  $^1J_{C\alpha C\beta} = 35$  Hz. The relaxation rates (including transverse relaxation rates, cross relaxation rate, and spin flip rates of amide protons) were computed using the well-known formulas in which the spectral density function has the model-free form. The order parameters ( $S^2$ ) of the backbone and side-chains were set as 0.86 and 0.4, respectively. The internal correlation time was set to 50 ps. The bond lengths for N-H and C-H were set to 1.02 and 1.09 Å, respectively. The distance between two atoms separated by two or more bonds in a protein was calculated from the high resolution x-ray structure. Parameters for chemical shift anisotropy (CSA) used were:  $\Delta\sigma_N = -170$  ppm;  $\phi_N = 18^\circ$  ( $\phi$  is the angle between the principal axis of the axially symmetric CSA tensor and the chemical bond N-H);  $\Delta\sigma_H = 13$  ppm;  $\phi_H = 15^\circ$ ;  $\Delta\sigma_C = 25$  ppm;  $\sigma_{COx} = 244$  ppm;  $\sigma_{COx} = 178$  ppm;  $\sigma_{COz} = 90$  ppm ( $\sigma_{COz}$ ,  $\sigma_{COz}$  and  $\sigma_{COz}$  are the principal components of the CO CSA tensor) (Teng, Iqbal et al. 1992). The relaxation rate of a given spin used in the peak intensity calculations was the average of the relaxation rates of this spin derived from 10 high resolution protein structures (PDB codes: 1IUAA, 1N55A, 1NWZA, 1PJXA, 1PQ7A, 1X8QA, 1W0NA, 1US0A, 1UCSA and 1R6JA). The TOCSY transfer efficiency was simulated using a FLOPSY8 mixing scheme with a spin-lock field strength of 8.3 kHz. In this simulation,  $^{13}C$  chemical shifts were the average values listed in the BMRB

database. In the absence of relaxation, the TOCSY transfer efficiency between  $C^\beta$  and  $C^\alpha$  was found to be  $\geq 0.15$  for all amino acids, and thus it was set as 0.15. The TOCSY transfer efficiency between other  $^{13}\text{C}$  spins in all amino acids was found to be  $\geq 0.08$  and thus it was set to 0.08. The relaxation effect during the TOCSY mixing period was approximated as  $\exp(-R_{2\text{av}}*T_{\text{mix}}/2)*\exp(-R_{1\text{av}}*T_{\text{mix}}/2)$ , where  $R_{2\text{av}}$  and  $R_{1\text{av}}$  are the average spin-spin and spin-lattice relaxation rates of the  $^{13}\text{C}$  spins involved in the TOCSY transfer spin-system; and  $T_{\text{mix}}$  is the mixing time ( $\sim 15$  ms). Note that the magnetization toggles between  $C_z$  and  $C_y$  during the TOCSY mixing period. The NOE transfer efficiency was calculated from the initial rate approximation, i.e.,  $I_{\text{ab}}/I_{\text{aa}} = R_{\text{c}}(\text{ab})*T_{\text{noemix}}$ , where  $I_{\text{ab}}$  and  $I_{\text{aa}}$  are the intensities of the NOESY cross-peak and diagonal peak, respectively;  $R_{\text{c}}(\text{ab})$  is the cross-relaxation rate between protons  $a$  and  $b$  which are separated by  $3.0 \text{ \AA}$ ; and  $T_{\text{noemix}}$  is the NOE mixing time, and was set as  $40*30/\tau_{\text{m}}$  ms, where  $\tau_{\text{m}}$  is expressed in nanosecond. The acquisition time in the direct observation dimension was set to 57 ms for all experiments. The acquisition times were 6.1 ms and 25 ms for the respective  $^{13}\text{C}$  and  $^{15}\text{N}$  dimensions in the TROSY-HNCA. The acquisition times in both indirect  $^{13}\text{C}$  dimensions were 7.3 ms in the 3D MQ-CCH-TOCSY experiment. The acquisition times were 5.9 ms and 14 ms for the respective  $^{13}\text{C}$  and  $^{15}\text{N}$  dimensions in the 4D  $^{13}\text{C},^{15}\text{N}$ -edited NOESY (supplementary Fig. S8). All simulations were performed under a proton frequency of 800 MHz and perfect pulses.

3D (H)CC(CO)NH-TOCSY is an insensitive experiment for relatively large proteins. Nevertheless, it was still sensitive enough to yield most correlations (including  $C^\delta$  methyls in Ile and Leu which have very low TOCSY transfer efficiency to  $C^\alpha$ ) for DdCAD-1 in the presence and absence of  $\text{Ca}^{2+}$  (protein concentration  $\sim 0.8$  mM, experimental time 60 hrs at a 800 MHz spectrometer with a normal probe). The apparent correlation times were  $\sim 12.5$  and  $\sim 15$  ns for the  $\text{Ca}^{2+}$ -free and  $\text{Ca}^{2+}$ -bound forms at  $30^\circ\text{C}$ , respectively (Dynamic dimerization in the presence of  $\text{Ca}^{2+}$  resulted in the correlation time difference of the two forms). Therefore, we used the intensity of the correlation Leu  $C_{i-1}^\delta\text{-N}_i\text{H}_i$  in the 3D (H)CC(CO)NH-TOCSY recorded on a 15 ns tumbling protein as a reference for different kinds of correlations obtained at various correlation times.

The 3D MQ-CCH-TOCSY experiment used here did not employ gradient selection to enhance sensitivity. Very recently, it has been shown that the sensitivity of this experiment can be enhanced by a factor of two with gradient selection in both indirect dimensions (Permi, Tossavainen et al. 2004).

Our simulations show that 50% partial deuteration reduces the intensities of NOEs between CH and NH groups for proteins with a correlation time  $<35$  ns, because the reduction of proton concentration (density) cannot be compensated for by the more favorable relaxation of the remaining protons (Nietlispach, Clowes et al. 1996). On the other hand, it can enhance the peak intensities of NOESY and TOCSY correlations involved in  $\text{CH}_2$  and  $\text{CH}_3$  groups for a protein with a correlation time  $>20$  ns. However, the enhancement may be canceled out by chemical shift heterogeneity due to the presence of numerous isotopomers (e.g., the  $^{13}\text{C}$  chemical shift difference between  $\text{CHD}_2$  and  $\text{CH}_2\text{D}$  is  $\sim 0.3$  ppm while the difference between  $\text{CHD}_2$  and  $\text{CH}_3$  is  $\sim 0.6$  ppm). Because partial deuteration reduces spin diffusion, longer NOE mixing time can be used, implying that the NOESY experiment may benefit from partial deuteration for proteins  $>20$  ns.

Figure 4.9 Detailed information on backbone assignments

A sample used for explaining Supplementary Figures 4.9 a-d.

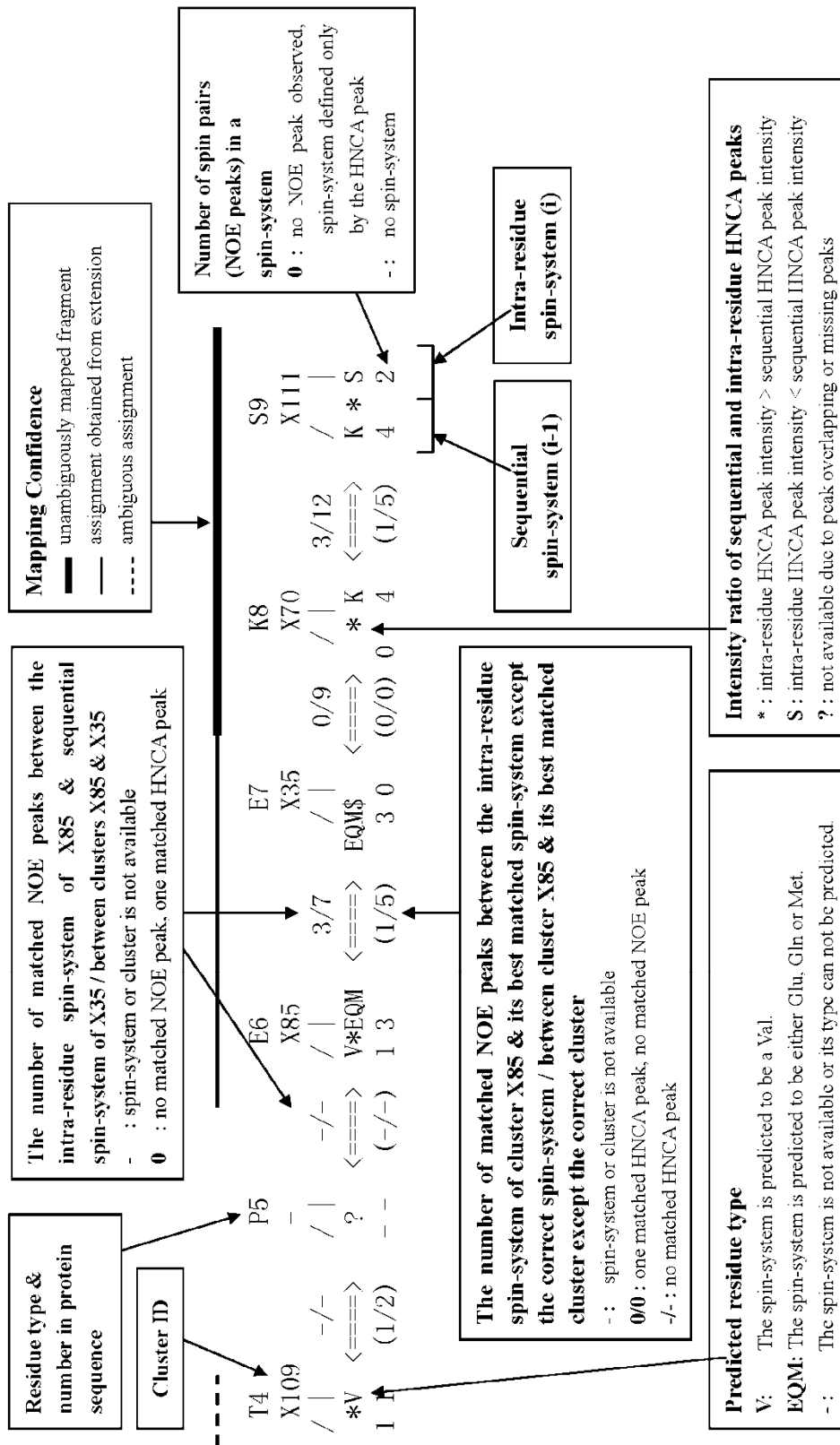




Figure 4.9 a Detailed information about spin-systems, clusters, fragments, and connectivities for DdCAD-1

|              |        |                |           |           |              |                |           |               |               |                 |              |
|--------------|--------|----------------|-----------|-----------|--------------|----------------|-----------|---------------|---------------|-----------------|--------------|
| G1           |        | S3             | V4        | D5        | A6           | N7             | R8        | N10           | F11           | F12             | F13          |
| -            |        | X148           | X101      | X13       | X59          | X265           | X134      | X34           | X44           | X182            | X121         |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| -2/          | -2/    | -2/            | -2/       | -2/       | -2/          | -2/            | -2/       | -2/           | -2/           | -2/             | -2/          |
| -            | -      | -              | -         | -         | -            | -              | -         | -             | -             | -               | -            |
|              |        |                |           |           |              |                |           |               |               |                 |              |
| G14          |        | N16            | C17       | T18       | G19          | E20            | S21       | E23           | Y24           | N25             | K26          |
| X346         | X118   | X195           | X329      | X345      | X351         | X100           | X227      | X82           | X115          | X110            | X137         |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| DNFY*G       | G*H    | R*DNFY<====>   | -*WHC     | <====>    | T*G          | G*EQM          | <====>    | EQ*DNFY<====> | EQ*DNFY<====> | DNFY*DNFY<====> | DNFY*H       |
| 3 2          | (-/-)  | 2 2 (1/1)      | 1 2 (0/1) | 1 3 (0/1) | 3 2 (1/2)    | 2 4 (-/-)      | 4 3 (1/1) | 3 4 (1/2)     | 4 3 (0/2)     | 2 2 (1/1)       | 2 6 (0/4)    |
|              |        |                |           |           |              |                |           |               |               |                 |              |
| G27          |        | T29            | V20       | F31       | F32          | N33            | N34       | D36           | K37           | W38             | N39          |
| X163         | X123   | X67            | X11       | X62       | X68          | X610           | X183      | X83           | X173          | X209            | X114         |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| R*G          | G*EQM  | EQ*H           | T*W       | V*H       | R*DNFY<====> | R*DNFY<====>   | <====>    | G*DNFY<====>  | <====>        | K*H             | -*DNFY<====> |
| 1 2          | (1/1)  | 2 5 (0/1)      | 2 4 (0/1) | 3 5 (0/1) | 4 3 (0/2)    | 2 1 (1/2)      | 2 2 (1/2) | 2 3 (1/1)     | 3 6 (0/2)     | 3 1 (1/1)       | 1 3 (0/1)    |
|              |        |                |           |           |              |                |           |               |               |                 |              |
| D40          |        | F42            | M43       | S44       | C45          | L46            | V47       | G48           | N50           | V51             | R52          |
| X107         | X88    | X102           | X35       | X322      | X190         | X24            | X40       | X156          | X120          | X226            | X237         |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| -2DNFY<====> | -2K    | <====>         | -*EQM     | <====>    | -*WHC        | WHC*H          | L*W       | V*G           | S*DNFY<====>  | DNFY*W          | V*H          |
| 1 2          | (1/1)  | 1 3 (1/2)      | 1 3 (0/1) | 2 1 (0/1) | 1 3 (0/1)    | 3 4 (0/2)      | 3 3 (0/1) | 2 2 (1/2)     | 2 3 (1/4)     | 3 3 (0/2)       | 4 4 (1/3)    |
|              |        |                |           |           |              |                |           |               |               |                 |              |
| C53          |        | N54            | W56       | E57       | H58          | N59            | B60       | I61           | T63           | F64             | T65          |
| X96          | X6     | X210           | X135      | X29       | X318         | X139           | -         | X152          | X292          | -               | X308         |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| -*WHC        | <====> | WHC*DNFY<====> | I*H       | WHC*H     | -2/          | WHC*DNFY<====> | -2/       | EQ*H          | I*DNFY<====>  | DNFY*H          | P*H          |
| 1 2          | (1/1)  | 2 3 (0/0)      | 2 6 (1/2) | 2 3 (0/1) | 3 2 (0/0)    | 3 2 (0/1)      | 3 2 (0/0) | 3 6 (-/-)     | 2 3 (1/2)     | 2 3 (1/2)       | 5 3 (0/0)    |
|              |        |                |           |           |              |                |           |               |               |                 |              |
| P66          |        | G67            | K66       | X66       | F69          | X57            | F69       | X57           | X280          | X280            | N78          |
| /            | /      | /              | /         | /         | /            | /              | /         | /             | /             | /               | /            |
| -2/          | -2/    | -2/            | -2/       | -2/       | -2/          | -2/            | -2/       | -2/           | -2/           | -2/             | -2/          |
| <====>       | <====> | <====>         | <====>    | <====>    | <====>       | <====>         | <====>    | <====>        | <====>        | <====>          | <====>       |
| (-/-)        | (-/-)  | 5 2 (1/1)      | 2 6 (1/1) | 3 5 (1/1) | 5 4 (0/2)    | 3 6 (1/2)      | 5 2 (0/1) | 2 4 (0/1)     | 4 1 (0/2)     | 1 1 (0/0)       | 1 3 (1/1)    |
|              |        |                |           |           |              |                |           |               |               |                 |              |



|                |               |                |               |              |                 |                   |                   |              |               |               |                |               |
|----------------|---------------|----------------|---------------|--------------|-----------------|-------------------|-------------------|--------------|---------------|---------------|----------------|---------------|
| G170           | S171          | V172           | Y173          | F174         | K175            | Y176              | S177              | P178         | T179          | T180          | G181           | Q182          |
| X277           | X181          | X91            | X43           | X154         | X136            | X21               | X164              | -            | X349          | X339          | X151           | X126          |
| /   2/5 <====> | G*S <====>    | S*V <====>     | V*DNIFY<====> | DNIFY<====>  | DNIFY*PK <====> | R*DNIFY<====>     | DNIFY*S <====>    | -? <====>    | P?T <====>    | T*T <====>    | -*G <====>     | G*EQM <====>  |
| 3 2 (-/-)      | 2 2 (1/1)     | 2 4 (1/1)      | 3 3 (1/1)     | 2 3 (1/1)    | 3 4 (0/3)       | 4 2 (0/0)         | 2 2 (0/0)         | -- (1/1)     | 7 3 (-/-)     | 3 3 (0/2)     | 1 2 (0/0)      | 2 4 (0/1)     |
| V183           | T184          | V185           | I186          | K187         | K188            | D189              | E190              | T191         | F192          | P193          | K194           | N195          |
| X16            | X61           | X32            | X10           | X38          | X53             | X46               | X174              | X353         | X14           | -             | X90            | X251          |
| /   1/1 <====> | V?T <====>    | T*V <====>     | V?I <====>    | I*K <====>   | K*K <====>      | -*DNIFY<====>     | DNIFY*EQM <====>  | EQM*- <====> | T*DNIFY<====> | -? <====>     | P*K <====>     | K*- <====>    |
| 4 1 (0/2)      | 2 3 (1/1)     | 3 4 (0/1)      | 4 6 (0/2)     | 6 7 (1/3)    | 5 3 (0/2)       | 1 2 (0/2)         | 2 3 (1/1)         | 3 1 (0/1)    | 2 3 (0/2)     | -- (0/1)      | 3 5 (1/2)      | 4 3 (0/3)     |
| M196           | T197          | V198           | T199          | Q200         | D201            | D202              | N203              | T204         | S205          | F206          | I207           | F208          |
| X155           | X317          | X86            | X55           | X18          | X7              | X208              | X223              | X340         | X147          | X60           | X574           | X20           |
| /   4/6 <====> | EQM*T <====>  | T?V <====>     | V*T <====>    | T*EQM <====> | EQM?DNIFY<====> | DNIFY*DNIFY<====> | DNIFY*DNIFY<====> | -*V <====>   | V*S <====>    | S?DNIFY<====> | DNIFY*I <====> | I*DNIFY<====> |
| 2 4 (1/1)      | 5 3 (1/1)     | 3 4 (0/3)      | 4 3 (1/7)     | 3 5 (1/7)    | 2 3 (0/3)       | 3 1 (1/1)         | 3 3 (0/0)         | 1 2 (0/2)    | 2 3 (1/1)     | 2 3 (0/2)     | 3 6 (1/7)      | 4 3 (0/8)     |
| N209           | L210          | N211           | S212          | E213         | K214            | X65               | X45               | X54          | X323          | X87           | X92            |               |
| /   2/3 <====> | L*DNIFY<====> | DNIFY*S <====> | S*EQM <====>  | EQM*K <====> | EQM*K <====>    | 2 3 (0/2)         | 2 6 (1/2)         | 4 3 (1/4)    | 2 3 (1/1)     | 3 5 (1/1)     | 3 5 (0/1)      | 5 6 (0/1)     |





|      |        |              |        |        |              |        |        |        |        |              |        |              |        |        |        |       |        |
|------|--------|--------------|--------|--------|--------------|--------|--------|--------|--------|--------------|--------|--------------|--------|--------|--------|-------|--------|
| K170 | Y171   | E172         | M173   | G174   | K175         | Y176   | D177   | I178   | K179   | D130         | V1C:   | G182         |        |        |        |       |        |
| X27  | X5     | X19          | -      | X2     | X189         | X148   | X52    | X230   | X256   | X60          | X52    | X123         |        |        |        |       |        |
| /    | 5/6    | /            | /      | /      | /            | 4/4    | /      | 4/6    | /      | 2/3          | /      | 3/3          | /      | 1/2    |        |       |        |
| -*K  | <====> | R*DNIF<====> | -/?    | <====> | G*K          | <====> | -*DN   | -*I    | 1*     | -*DNFY<====> | DNFY*  | <====>       | V*G    | <====> |        |       |        |
| 1.6  | (1/1)  | 5.3          | (1/1)  | 3.2    | (1/1)        | 4.1    | (1/1)  | 2.5    | (1/1)  | 4.3          | (1/1)  | 3.4          | (0/1)  | 3.2    | (0/0)  |       |        |
| V183 | D134   | N185         | A186   | G187   | A188         | K189   | A190   | G191   | L192   | I193         | F194   | L195         |        |        |        |       |        |
| X303 | X237   | X24E         | -      | X436   | X13          | X276   | X33    | X400   | X118   | X243         | X38    | X153         |        |        |        |       |        |
| /    | 3/4    | /            | /      | /      | /            | 2/5    | /      | 2/2    | /      | 1/2          | /      | 2/3          | /      | 4/7    |        |       |        |
| G*V  | <====> | V*DN         | <====> | A*G    | <====>       | A*K    | <====> | A*G    | <====> | G*H          | <====> | T*DNFY<====> | F*Y*L  | <====> |        |       |        |
| 1.3  | (0/1)  | 3.2          | (0/5)  | 2.3    | (0/5)        | 2.5    | (2/2)  | 2.2    | (0/0)  | 1.4          | (1/1)  | 2.3          | (1/1)  | 2.6    | (0/1)  |       |        |
| V196 | D137   | L19F         | T199   | R200   | N201         | I202   | H203   | M204   | I205   | A206         | D207   | T208         |        |        |        |       |        |
| X216 | X65    | X14E         | X153   | X37    | X206         | X393   | X249   | X310   | X188   | X34          | X353   | X238         |        |        |        |       |        |
| /    | 4/5    | /            | /      | /      | /            | 2/4    | /      | 1/7    | /      | 1/1          | 2/4    | /            | 2/7    | /      | 2/4    |       |        |
| L*V  | <====> | V*DNFY<====> | -*I    | <====> | -*DNFY<====> | -*K    | <====> | -*K    | <====> | -*A          | <====> | A*DNFY<====> | DNFY*  | <====> |        |       |        |
| 4.4  | (1/1)  | 4.3          | (0/1)  | 4.4    | (1/2)        | 2.2    | (0/2)  | 1.3    | (1/2)  | 4.3          | (0/1)  | 2.2          | (0/2)  | 2.3    | (0/0)  |       |        |
| D209 | Y210   | S211         | Z212   | A213   | E214         | A215   | A216   | F217   | N218   | K219         | G220   | E221         |        |        |        |       |        |
| X4   | X222   | X28E         | X40    | X130   | X22C         | X160   | X218   | X163   | -      | X232         | X447   | X245         |        |        |        |       |        |
| /    | 0/0    | /            | /      | /      | /            | 0/0    | /      | 2/2    | /      | /            | 1/1    | 0/0          | /      | -      | -      |       |        |
| T*+  | <====> | DH*+         | <====> | I*H+   | <====>       | -*A    | <====> | A*FY   | <====> | -/?          | <====> | -*G          | <====> | G?+    | <====> |       |        |
| 2.0  | (0/0)  | 3.1          | (0/1)  | 2.2    | (0/0)        | 2.6    | (2/2)  | 4.2    | (1/2)  | 1.1          | (1/1)  | 2.2          | (1/1)  | 1.2    | (2/2)  | 0     | (-)    |
| T222 | A223   | M224         | Z225   | I226   | N227         | G228   | F229   | W230   | A231   | W232         | S233   | N234         |        |        |        |       |        |
| X413 | X15    | X297         | X333   | X82    | X94          | X404   | -      | X369   | X31    | X196         | X305   | -            |        |        |        |       |        |
| /    | 5/3    | /            | /      | /      | 0/1          | 1/2    | /      | 1/2    | /      | 2/2          | /      | -            | /      | -      | /      | -     |        |
| -?T  | <====> | T*H+         | <====> | T*+    | <====>       | -*C    | <====> | -*W*HC | <====> | -*A          | <====> | -*+          | <====> | -*     | <====> | (/)   |        |
| 2.3  | (0/0)  | 3.2          | (0/1)  | 2      | (/)          | 1.1    | (0/1)  | 2.1    | (1/1)  | 1.0          | (/)    | (/)          | 2.1    | (0/0)  | 0.0    | (0/1) | (/)    |
| T235 | D236   | T237         | S238   | K239   | V240         | N241   | Y242   | G243   | V244   | T245         | V246   | L247         |        |        |        |       |        |
| -    | -      | -            | X194   | -      | X137         | X16    | X33    | X463   | X153   | X159         | X54    | X25          |        |        |        |       |        |
| /    | -      | /            | /      | /      | /            | 2/2    | /      | 1/1    | /      | 3/3          | /      | 3/4          | /      | 4/4    | /      | -     | -      |
| -*   | <====> | -*           | <====> | -/?    | <====>       | V*DN   | <====> | DN*    | <====> | V*H          | <====> | T*H*         | <====> | V*H    | <====> | V*H   | <====> |
| --   | (-)    | --           | (-)    | --     | (-)          | 4      | (0/0)  | 2.1    | (0/0)  | 3.3          | (0/1)  | 3.4          | (1/1)  | 4.2    | (0/1)  | 4.2   | (0/1)  |
| F24S | T249   | F25C         | K251   | G252   | Q253         | F254   | S255   | K256   | F257   | F258         | V259   | G260         |        |        |        |       |        |
| -    | X238   | X97          | X8     | X477   | X85          | -      | X206   | X36    | -      | X184         | X100   | X255         |        |        |        |       |        |
| /    | -      | /            | /      | /      | 2/3          | /      | -      | 1/1    | /      | -            | 0/0    | /            | 2/2    | /      | 1/2    | 1/2   | 1/2    |
| -/?  | <====> | -*T          | <====> | -*G    | <====>       | -/?    | <====> | G*+    | <====> | -*           | <====> | -*H          | <====> | -*H    | <====> | V*G   | <====> |
| --   | (-)    | 3.1          | (0/2)  | 1.2    | (0/0)        | 1.2    | (0/0)  | 1.4    | (1/1)  | 1.0          | (0/0)  | 1.3          | (0/0)  | 1.3    | (0/0)  | 3.1   | (0/0)  |

|        |        |        |        |        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| V261   | L262   | S263   | A264   | G265   | I266   | N267   | A268   | A269   | S270   | F271   | N272   | K273   |
| X9     | X481   | X20    | Z51    | X408   | X53    | X26    | X3     | X213   | X320   | -      | X157   | X168   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 1/1    | 1/1    | 1/1    | 0/0    | 1/1    | 2/2    | 1/1    | 0/0    | 2/3    | 2/3    | -      | 1/1    | 1/1    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 3 2    | 3 2    | 5 0    | 1 1    | 2 2    | 2 3    | 5 0    | 1 2    | 2 2    | 2 2    | -      | 3      | 1 2    |
| (0/1)  | (0/1)  | (0/1)  | (0/1)  | (0/0)  | (0/0)  | (0/0)  | (0/0)  | (1/1)  | (0/1)  | (1/1)  | (0/0)  | (0/0)  |
| E274   | L275   | A276   | K277   | E278   | F279   | L280   | E281   | N282   | V283   | L284   | L285   | T286   |
| X192   | X129   | X189   | X199   | X175   | X193   | X147   | X346   | X446   | X244   | X106   | X165   | X448   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 3/4    | 4/6    | 4/6    | 1/1    | -2/2   | 3/3    | 0/0    | 1/1    | 0/2    | 1/1    | 2/3    | 1/1    | 2/3    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 3 3    | 3 6    | 4 2    | 1 3    | -3     | 3 3    | 0 4    | 1 0    | 1 3    | 1 3    | 2 4    | 3 5    | 2 3    |
| (1/1)  | (1/1)  | (1/1)  | (1/1)  | (1/3)  | (0/0)  | (0/0)  | (0/0)  | (0/2)  | (0/0)  | (0/1)  | (1/2)  | (0/1)  |
| D287   | E288   | G289   | L290   | E291   | A292   | V293   | X294   | K295   | D296   | K297   | P298   | L299   |
| X87    | X265   | X335   | X162   | X131   | X154   | X262   | X217   | X253   | X176   | X264   | -      | X65    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 2/6    | 3/4    | 2/2    | 3/4    | 2/6    | 2/5    | 2/3    | 2/3    | 3/3    | 3/4    | 2/4    | -      | 1/3    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 3 2    | 2 3    | 3 2    | 2 6    | 3 4    | 2 2    | 2 4    | 2 3    | 3 4    | 3 2    | 2 6    | -      | 2 4    |
| (1/1)  | (1/1)  | (3/3)  | (0/0)  | (1/2)  | (3/3)  | (1/1)  | (1/1)  | (1/2)  | (0/1)  | (1/4)  | (0/0)  | (1/1)  |
| G300   | A301   | V302   | A303   | L304   | I305   | S306   | V307   | E308   | E309   | E310   | L311   | A312   |
| X475   | X44    | X424   | X227   | X263   | X6     | X365   | X86    | X114   | X266   | X214   | X122   | X202   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 2/3    | 2/2    | 2/2    | 1/3    | 1/3    | 0/2    | 0/3    | 0/3    | 0/1    | 1/1    | 4/5    | 2/5    | 2/2    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2 2    | 2 2    | 2 2    | 4 2    | 1 4    | 2 4    | 0 3    | 0 1    | 0 1    | 2 4    | 4 4    | 2      | 4 2    |
| (-/-)  | (1/1)  | (1/1)  | (0/1)  | (0/0)  | (0/2)  | (0/1)  | (0/0)  | (0/1)  | (1/1)  | (1/1)  | (1/1)  | (0/1)  |
| K315   | I314   | P315   | F316   | I317   | A318   | A319   | F320   | M321   | E322   | N323   | A324   | Q325   |
| X215   | X124   | -      | X229   | X67    | X79    | X178   | X291   | X169   | X105   | X275   | X161   | X242   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 2/5    | 0/0    | -      | 0/0    | 3/4    | -3/3   | 2/2    | 2/2    | 2/2    | 0/2    | 3/5    | 2/4    | 2/2    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2 7    | 2 2    | 2 2    | 1 1    | 0 3    | 3 2    | -2     | 2 3    | 2 2    | 0 3    | 3 2    | 2      | 4      |
| (2/2)  | (2/2)  | (0/0)  | (-/-)  | (1/1)  | (0/0)  | (0/1)  | (1/1)  | (0/0)  | (0/1)  | (1/1)  | (1/1)  | (0/1)  |
| K326   | G327   | E328   | I329   | M330   | F331   | N332   | I333   | F334   | Q335   | M336   | S337   | A338   |
| X208   | X469   | X93    | X10    | X26    | -      | X180   | X450   | -      | X236   | X195   | X318   | X61    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 0/2    | 1/1    | 2/2    | 1/1    | 2/2    | -      | 1/3    | 1/3    | -      | 1/2    | 1/2    | 1/5    | 2/2    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2 2    | 2 2    | 2 3    | 2 6    | 2 2    | -      | 1 1    | 1 5    | -      | 4 3    | 2 1    | 1      | 2      |
| (1/1)  | (1/1)  | (-/-)  | (1/1)  | (1/2)  | (0/1)  | (-/-)  | (1/1)  | (1/2)  | (-/-)  | (0/3)  | (0/1)  | (1/1)  |
| F339   | W340   | Y341   | A342   | V343   | F344   | T345   | A346   | V347   | I348   | N349   | A350   | A351   |
| X173   | X201   | X294   | X101   | X187   | X128   | X339   | X73    | X251   | X190   | X116   | X72    | X126   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| 1/2    | 2/2    | 2/2    | 2/4    | 2/3    | 3/3    | 1/2    | 3/4    | 2/3    | 4/4    | 6/7    | 2/2    | 1/3    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2 3    | 1 2    | 2 2    | 2 2    | 2 4    | 3 2    | 1 3    | 3 2    | 2 4    | 4 6    | 6 2    | 2 2    | 1 2    |
| (1/1)  | (0/0)  | (0/0)  | (1/1)  | (1/1)  | (0/1)  | (0/0)  | (-/-)  | (1/1)  | (1/1)  | (0/3)  | (0/3)  | (1/2)  |

|        |        |        |        |        |        |        |         |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|---------|--------|--------|--------|--------|--------|
| S352   | Q353   | R354   | Q355   | T356   | V357   | D358   | E359    | A360   | L361   | K362   | D363   | A364   |
| X453   | X430   | X122   | X331   | X383   | X104   | X273   | X175    | X107   | X246   | X103   | X143   | X103   |
| /      | /      | /      | /      | /      | /      | /      | /       | /      | /      | /      | /      | /      |
| 2/4    | 1/5    | 4/6    | 4/5    | 4/7    | 3/3    | 4/7    | 2/4     | 2/2    | 4/8    | 3/11   | 3/11   | 3/11   |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====>  | <====> | <====> | <====> | <====> | <====> |
| A*S    | S*C    | G*R    | R*     | -*T    | T*V    | V*DNFY | DNFY*EQ | EQ*A   | A*L    | L?K    | K*DNFY | DNFY*A |
| 2 2    | 2 2    | 1 4    | 4 5    | 4 3    | 3 4    | 4 2    | 2 3     | 3 2    | 2 4    | 4 3    | 4 3    | 3 2    |
| (0/1)  | (1/1)  | (1/1)  | (0/1)  | (0/1)  | (1/1)  | (0/2)  | (1/1)   | (1/1)  | (1/2)  | (1/3)  | (1/2)  | (1/1)  |
| <hr/>  |        |        |        |        |        |        |         |        |        |        |        |        |
| Q365   | T366   | R367   | I368   | T369   | K370   |        |         |        |        |        |        |        |
| X198   | X272   | X150   | X220   | X466   | X12    |        |         |        |        |        |        |        |
| /      | /      | /      | /      | /      | /      |        |         |        |        |        |        |        |
| 1/4    | 3/5    | 2/5    | 4/6    | 4/6    | 3/8    |        |         |        |        |        |        |        |
| <====> | <====> | <====> | <====> | <====> | <====> |        |         |        |        |        |        |        |
| A*-    | -*T    | T*R    | -?I    | I*T    | T*K    |        |         |        |        |        |        |        |
| 2 3    | 1 3    | 3 4    | 2 4    | 4 3    | 3 5    |        |         |        |        |        |        |        |
| (0/2)  | (-/-)  | (1/3)  | (1/1)  | (-/-)  | (2/2)  |        |         |        |        |        |        |        |



Figure 4.9 c Detailed information about spin-systems, clusters, fragments, and connectivities for the  $\alpha$ -chain of HbCO A

|       |       |         |           |       |       |       |           |         |       |         |         |           |       |       |       |       |
|-------|-------|---------|-----------|-------|-------|-------|-----------|---------|-------|---------|---------|-----------|-------|-------|-------|-------|
| V1    | -     | L2      | -         | S3    | -     | P4    | -         | A5      | D6    | K7      | T8      | N9        | V10   | K11   | A12   | A13   |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         | /     | /     | /     | /     |
| -?    | -?    | -?      | -?        | -?    | -?    | -?    | -?        | -?      | -?    | -?      | -?      | -?        | -?    | -?    | -?    | -?    |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       | <=>   | <=>   | <=>   | <=>   |
| (-/-) | (-/-) | (-/-)   | (-/-)     | (-/-) | (-/-) | (-/-) | (-/-)     | (-/-)   | (-/-) | (-/-)   | (-/-)   | (-/-)     | (-/-) | (-/-) | (-/-) | (-/-) |
| W14   | G15   | K16     | V17       | G18   | A19   | H20   | A21       | G22     | E23   | Y24     | G25     | A26       |       |       |       |       |
| X51   | X52   | X101    | X100      | X146  | -     | X116  | X105      | X107    | X60   | X63     | X99     | X125      |       |       |       |       |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         |       |       |       |       |
| 1/5   | 1/6   | 1/6     | 5/10      | 4/7   | -/-   | -/-   | 2/7       | 1/2     | 3/5   | 1/6     | 1/6     | 2/3       |       |       |       |       |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       |       |       |       |       |
| A*    | -*G   | G*K     | R*V       | V*G   | -?    | A*W*H | W*H*G     | -*G     | C*E*Q | E*Q*H   | -*G     | G*H       |       |       |       |       |
| 2.1   | (0/2) | 1.1     | (1/2)     | 1.5   | (1/4) | 4.1   | (0/1)     | 4.1     | (0/1) | 2.2     | (0/1)   | 3.1       |       |       |       |       |
| (1/2) | (1/2) | (1/2)   | (1/4)     | (0/1) | (1/4) | (0/1) | (0/1)     | (0/1)   | (0/1) | (1/2)   | (1/2)   | 1.2       |       |       |       |       |
| (1/1) | (1/2) | (1/2)   | (1/4)     | (0/1) | (1/4) | (0/1) | (0/1)     | (0/1)   | (0/1) | (1/2)   | (1/2)   | 2.2       |       |       |       |       |
| (0/0) | (0/2) | (0/2)   | (0/2)     | (1/3) | 3     | (-/-) | (-/-)     | (-/-)   | (-/-) | (-/-)   | (-/-)   | (0/0)     |       |       |       |       |
| E27   | A28   | L29     | E30       | R31   | M32   | F33   | L34       | S35     | F36   | F37     | T38     | T39       |       |       |       |       |
| X65   | X61   | X95     | X43       | X38   | X39   | X41   | X72       | X24     | X23   | -       | X151    | X147      |       |       |       |       |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         |       |       |       |       |
| 1/8   | 2/4   | 1/6     | 1/6       | 3/5   | 0/3   | 1/3   | 1/7       | 1/4     | 1/4   | -/-     | -/-     | 0/2       |       |       |       |       |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       |       |       |       |       |
| A*    | -*A   | A*      | L*E*Q     | E*Q*H | -?    | -*    | -*L       | -*      | -*    | -?      | -?      | T*H       |       |       |       |       |
| 2.1   | (1/4) | 1.2     | (0/2)     | 2.1   | (0/2) | 4.4   | (1/3)     | 3       | (-/-) | (-/-)   | (-/-)   | 2.3       |       |       |       |       |
| (1/4) | (1/4) | (0/2)   | (0/2)     | (1/3) | (0/2) | (1/3) | (1/3)     | (1/3)   | (1/1) | (0/1)   | (-/-)   | (0/0)     |       |       |       |       |
| (0/1) | (1/4) | (0/2)   | (0/2)     | (1/3) | (0/2) | (1/3) | (1/3)     | (1/3)   | (1/1) | (0/1)   | (-/-)   | (0/0)     |       |       |       |       |
| K40   | T41   | Y42     | F43       | F44   | H45   | F46   | D47       | L48     | S49   | H50     | G51     | S52       |       |       |       |       |
| X56   | X7    | X6      | X45       | -     | X3    | X112  | X113      | X134    | X66   | -       | X97     | X5        |       |       |       |       |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         |       |       |       |       |
| 1/2   | 3/3   | 1/3     | 1/3       | -/-   | -/-   | 1/6   | 1/4       | 0/2     | 0/2   | -/-     | -/-     | 1/2       |       |       |       |       |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       |       |       |       |       |
| -*    | -*T   | T*      | -*D*H*F*Y | -?    | -*    | -?    | -*D*H*F*Y | -?L     | -*S   | -?      | -?G     | G*H       |       |       |       |       |
| 0.1   | (1/1) | 1.3     | (-/-)     | 3.1   | (0/1) | 1.2   | (1/1)     | -       | (0/1) | -       | (0/0)   | 2.3       |       |       |       |       |
| (1/1) | (1/1) | (1/1)   | (-/-)     | (1/1) | (0/1) | (1/1) | (1/1)     | (-/-)   | (0/1) | (-/-)   | (-/-)   | (0/1)     |       |       |       |       |
| A53   | Q54   | V55     | K56       | G57   | H58   | G59   | K60       | K61     | V62   | A63     | D64     | A65       |       |       |       |       |
| -     | X68   | X67     | X53       | X55   | X48   | X50   | X57       | X46     | X53   | X109    | X90     | X86       |       |       |       |       |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         |       |       |       |       |
| 1/9   | 5/9   | 1/3     | 3/4       | 4/7   | 1/2   | 1/5   | 2/4       | 1/6     | 0/1   | 1/7     | 2/6     | 2/7       |       |       |       |       |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       |       |       |       |       |
| -?    | A*E*Q | E*Q*H   | V*H       | K*G   | G*    | -*G   | G*H       | -*      | -?    | V*H     | A*H*F*Y | -*H       |       |       |       |       |
| -     | (-/-) | (1/1)   | (1/4)     | 3.5   | (0/3) | 4.1   | (0/0)     | 1.1     | (0/1) | 1.2     | (0/3)   | 2.2       |       |       |       |       |
| (1/3) | (1/3) | (0/2)   | (0/3)     | (0/3) | (1/2) | (1/2) | (1/1)     | (0/3)   | (0/2) | (1/1)   | (0/2)   | (0/2)     |       |       |       |       |
| (0/0) | (0/0) | (0/0)   | (0/0)     | (0/0) | (0/0) | (0/0) | (0/0)     | (0/0)   | (0/0) | (0/0)   | (0/0)   | (0/0)     |       |       |       |       |
| L66   | T67   | N68     | A69       | V70   | A71   | H72   | V73       | D74     | D75   | M76     | F77     | N78       |       |       |       |       |
| X82   | X75   | X106    | X108      | X111  | X133  | X127  | X4        | X2      | X121  | X27     | -       | X117      |       |       |       |       |
| /     | /     | /       | /         | /     | /     | /     | /         | /       | /     | /       | /       | /         |       |       |       |       |
| 1/8   | 3/7   | 1/6     | 2/6       | 2/5   | 1/5   | 2/3   | 1/8       | 4/6     | 2/7   | 1/7     | -/-     | 2/6       |       |       |       |       |
| <=>   | <=>   | <=>     | <=>       | <=>   | <=>   | <=>   | <=>       | <=>     | <=>   | <=>     | <=>     | <=>       |       |       |       |       |
| A*    | -*T   | T*H*F*Y | -*        | A*V   | V*H   | A*W*H | -*V       | V*H*F*Y | -*    | -*E*Q*H | -?      | -*D*H*F*Y |       |       |       |       |
| 2.1   | (1/3) | 1.3     | (0/2)     | 3.2   | (0/3) | 2.4   | (0/3)     | 2.2     | (0/3) | 2.2     | (0/6)   | 1.2       |       |       |       |       |
| (1/3) | (1/3) | (0/2)   | (0/3)     | (0/3) | (1/2) | (1/2) | (1/1)     | (0/0)   | (1/1) | 1.5     | (-/-)   | (1/3)     |       |       |       |       |
| (0/0) | (0/0) | (0/0)   | (0/0)     | (0/0) | (0/0) | (0/0) | (0/0)     | (0/0)   | (0/0) | (-/-)   | (-/-)   | (1/3)     |       |       |       |       |

|        |        |        |        |        |        |              |        |        |        |              |        |        |
|--------|--------|--------|--------|--------|--------|--------------|--------|--------|--------|--------------|--------|--------|
| A79    | L80    | S81    | A82    | L83    | S84    | D85          | L86    | H87    | A88    | H89          | K90    | L91    |
| X128   | X131   | X25    | -      | X110   | X36    | X35          | X91    | X34    | -      | X83          | X70    | X114   |
| /      | /      | /      | /      | /      | /      | /            | /      | 0/2    | /      | /            | 2/7    | 1/5    |
| 2/8    | 1/7    | 1/7    | -/-    | 3/9    | 2/3    | 1/3          | 1/3    | 0/2    | -/-    | -/-          | 1/5    | 0/2    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====>       | <====> | <====> | <====> | <====>       | <====> | <====> |
| DNF?A  | A*L    | -*     | -?     | A*L    | L*S    | S*DMPY<====> | DMPY?  | -?     | -?     | A*WHC        | WHC?K  | -*     |
| 2 2    | (1/1)  | 2 5    | (-/-)  | 2 4    | (0/2)  | 4 2          | (0/0)  | 2 2    | (0/1)  | 2 2          | (-/-)  | 2 5    |
|        |        |        |        |        |        |              |        |        |        |              |        |        |
| R92    | V93    | D94    | F95    | V96    | N97    | F98          | K99    | L100   | L101   | S102         | H103   | C104   |
| X98    | X12    | -      | -      | X149   | X102   | X21          | X20    | X58    | X62    | X11          | X10    | X47    |
| /      | /      | /      | /      | /      | /      | /            | /      | /      | /      | /            | /      | /      |
| 1/2    | 1/2    | -/-    | -/-    | -/-    | 3/4    | 0/2          | 0/2    | 0/4    | 0/1    | 0/5          | 2/2    | 0/3    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====>       | <====> | <====> | <====> | <====>       | <====> | <====> |
| -*     | -*V    | -?     | -?     | -?V    | V*-    | -*           | -*     | -*     | -*     | L*†          | T*-    | -?     |
| 0 1    | (0/1)  | 1 2    | (-/-)  | - 3    | (1/1)  | 0 1          | (0/1)  | 0 1    | (1/3)  | 5 2          | (-/-)  | - 1    |
|        |        |        |        |        |        |              |        |        |        |              |        |        |
| L105   | L106   | V107   | T108   | L109   | A110   | A111         | H112   | L113   | P114   | A115         | E116   | F117   |
| X71    | -      | X145   | X139   | X74    | X73    | X122         | X34    | X96    | -      | X132         | X104   | X92    |
| /      | /      | /      | /      | /      | /      | /            | /      | /      | /      | /            | /      | /      |
| -/-    | -/-    | -/-    | 4/5    | 1/3    | 1/7    | 2/6          | 2/8    | 1/6    | -/-    | 2/2          | 2/2    | 2/5    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====>       | <====> | <====> | <====> | <====>       | <====> | <====> |
| -*L    | -?     | -?V    | V*†    | -*EQM  | L*A    | A*A          | A*WHC  | -*L    | -?     | -*A          | A*EQM  | -*     |
| 1 5    | (2/4)  | - 4    | (1/2)  | 1 3    | (1/2)  | 2 2          | (0/3)  | 1 3    | (-/-)  | 1 2          | (1/2)  | 2 3    |
|        |        |        |        |        |        |              |        |        |        |              |        |        |
| T118   | P119   | A120   | V121   | H122   | A123   | S124         | L125   | D126   | K127   | F128         | L129   | A130   |
| X47    | -      | X130   | X129   | X87    | X89    | X17          | X15    | X84    | X54    | X28          | X29    | X84    |
| /      | /      | /      | /      | /      | /      | /            | /      | /      | /      | /            | /      | /      |
| -/-    | -/-    | -/-    | 2/4    | 1/6    | 1/7    | 2/7          | 0/3    | 0/6    | 3/6    | 1/5          | 1/5    | 3/5    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====>       | <====> | <====> | <====> | <====>       | <====> | <====> |
| -*†    | -?     | V*A    | A*V    | -*     | -*A    | A*S          | -*L    | -*     | -*†    | -*DMPY<====> | -*L    | L*A    |
| 1 3    | (0/3)  | 1 2    | (1/1)  | 1 1    | (0/3)  | 2 2          | (1/2)  | 0 1    | 1 4    | 3 2          | (0/3)  | 1 4    |
|        |        |        |        |        |        |              |        |        |        |              |        |        |
| S131   | V132   | S133   | T134   | V135   | L136   | T137         | S138   | K139   | Y140   | R141         | X88    | - 4    |
| X22    | X18    | X14    | X13    | X144   | -      | X152         | X44    | -      | -      | -            | -      | -      |
| /      | /      | /      | /      | /      | /      | /            | /      | /      | /      | /            | /      | /      |
| 1/6    | 1/6    | 4/7    | 1/4    | 2/4    | -/-    | -/-          | 2/3    | -/-    | -/-    | -/-          | -/-    | -/-    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====>       | <====> | <====> | <====> | <====>       | <====> | <====> |
| A*-    | -*V    | V*-    | -*†    | T?V    | -?     | -?           | -*S    | -?     | -?     | -?           | -?     | -?     |
| 2 1    | (0/3)  | 1 4    | (0/4)  | 1 3    | (-/-)  | - 2          | (0/0)  | 1 2    | (-/-)  | (-/-)        | (-/-)  | (-/-)  |

**Figure 4.9 d** Detailed information about spin-systems, clusters, fragments, and connectivities for the  $\beta$ -chain of HbCO A

|        |        |        |        |        |        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| V1     | -      | H2     | L3     | T4     | F5     | B6     | E7     | K8     | S9     | A10    | V11    | T12    | A13    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| -      | -      | -      | -      | -      | -      | -      | -      | -      | -      | -      | -      | -      | -      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 1/13   | 1/11   | 1/7    | 1/2    | 1/10   | 1/8    | 4/4    | 1/8    | 0/11   | 0/11   | 2/10   | 0/4    | 1/4    | 1/9    |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2.4    | 2.1    | 2.1    | 1.1    | 1.1    | 3.4    | 4.3    | 0.4    | 0.4    | 0.4    | 3.4    | 1.1    | 2.2    | 1.1    |
| (1/3)  | (0/3)  | (0/3)  | (0/0)  | (1/4)  | 3.4    | (-/-)  | (0/1)  | (0/0)  | (1/5)  | (1/1)  | (1/5)  | (0/4)  | (1/2)  |
| <hr/>  |        |        |        |        |        |        |        |        |        |        |        |        |        |
| L14    | W15    | G16    | K17    | V18    | N19    | M19    | V20    | D21    | E22    | V23    | G24    | G25    | E26    |
| X92    | X67    | X8     | X81    | X19    | X14    | X14    | X19    | X64    | X90    | X77    | X148   | X147   | X31    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 1.2    | 1.5    | 1.2    | 2.1    | 1.2    | 1.0    | 0.6    | 0.4    | 0.3    | 1.1    | 0.4    | 0.1    | 0.1    | 0.1    |
| (0/4)  | (0/5)  | (1/3)  | (1/3)  | (0/3)  | (0/2)  | (0/2)  | (0/3)  | (0/3)  | (0/2)  | (0/0)  | (0/2)  | (0/1)  | (0/1)  |
| <hr/>  |        |        |        |        |        |        |        |        |        |        |        |        |        |
| A27    | L28    | G29    | R30    | L31    | L32    | L32    | V33    | V34    | Y35    | P36    | W37    | T38    | Q39    |
| X53    | X97    | X4     | X91    | X37    | X107   | X107   | X98    | X66    | X83    | -      | X86    | X88    | -      |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 1.2    | 1.4    | 1.6    | 1.4    | 0.3    | 0.3    | 0.3    | 0.6    | 0.6    | 1.6    | 0.6    | 0.3    | 0.3    | 0.3    |
| (0/4)  | (0/5)  | (1/3)  | (1/3)  | (0/3)  | (0/2)  | (0/2)  | (0/3)  | (0/3)  | (0/2)  | (0/2)  | (0/2)  | (0/1)  | (0/1)  |
| <hr/>  |        |        |        |        |        |        |        |        |        |        |        |        |        |
| R40    | F41    | F42    | E43    | S44    | F45    | G46    | G46    | D47    | L48    | S49    | T50    | F51    | D52    |
| X582   | X133   | X99    | X41    | X126   | X30    | X150   | X150   | X28    | X7     | X129   | X142   | -      | X115   |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 1.1    | 1.7    | 1.6    | 1.8    | 0.3    | 1.2    | 1.8    | 1.8    | 2.2    | 1.3    | 1.8    | 2.7    | 0.6    | 0.6    |
| (0/3)  | (1/2)  | (1/3)  | (1/3)  | (0/3)  | (0/4)  | (0/4)  | (0/4)  | (-/-)  | (0/1)  | (0/2)  | (-/-)  | (0/4)  | (-/-)  |
| <hr/>  |        |        |        |        |        |        |        |        |        |        |        |        |        |
| A53    | V54    | M55    | G56    | N57    | P58    | K59    | K59    | V60    | K61    | A62    | H63    | G64    | K65    |
| X22    | X96    | X136   | X6     | X24    | -      | X63    | X63    | X29    | X571   | X46    | X47    | X153   | X44    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 2.2    | 2.5    | 1.5    | 1.6    | 2.3    | 0.6    | 1.1    | 1.1    | 1.1    | 0.9    | 2.7    | 2.5    | 0.6    | 2.5    |
| (2/3)  | (1/2)  | (1/1)  | (1/1)  | (-/-)  | (0/2)  | (-/-)  | (-/-)  | (1/4)  | (1/2)  | (1/2)  | (1/4)  | (0/3)  | (0/3)  |
| <hr/>  |        |        |        |        |        |        |        |        |        |        |        |        |        |
| K66    | V67    | L68    | G69    | A70    | F71    | S72    | S72    | D73    | G74    | L75    | A76    | H77    | L78    |
| X52    | X43    | X58    | X151   | X12    | X79    | X101   | X101   | X21    | X149   | X55    | X71    | X75    | X45    |
| /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      | /      |
| <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> | <====> |
| 5.1    | 1.4    | 3.5    | 5.1    | 1.2    | 2.3    | 3.1    | 3.1    | 3.3    | 3.1    | 0.6    | 4.2    | 2.3    | 1.2    |
| (1/4)  | (1/1)  | (1/1)  | (1/1)  | (0/4)  | (0/2)  | (0/3)  | (0/3)  | (0/1)  | (0/6)  | (0/2)  | (1/5)  | (0/0)  | (0/4)  |

|              |              |                  |             |             |           |           |           |              |              |           |             |             |
|--------------|--------------|------------------|-------------|-------------|-----------|-----------|-----------|--------------|--------------|-----------|-------------|-------------|
| D79          | L80          | L81              | K82         | G83         | T84       | F85       | A86       | T87          | L88          | S89       | E90         | L91         |
| X116         | X68          | -                | X74         | X5          | X78       | X105      | X15       | X117         | X26          | X57       | X51         | X89         |
| /            | /            | /                | /           | /           | /         | /         | /         | /            | /            | /         | /           | /           |
| 2/9          | /            | -/-              | 3/3         | 3/3         | -/        | 3/4       | 1/9       | 2/5          | 3/3          | 2/11      | 0/9         | 3/9         |
| <====>       | <====>       | <====>           | <====>      | <====>      | <====>    | <====>    | <====>    | <====>       | <====>       | <====>    | <====>      | <====>      |
| ->DNPY<====> | ->DNPY<====> | ->               | ->K<====>   | ->K<====>   | ->T<====> | ->T<====> | ->A<====> | A*T<====>    | T*H<====>    | L*S<====> | ->EQM<====> | EQM*-<====> |
| 2.2          | (0/0)        | 2.2              | (0/2)       | 6.1         | (0/0)     | -3        | (1/2)     | 3.1          | (1/3)        | 1.2       | (1/3)       | 3.1         |
|              |              | (0/2)            |             |             |           |           |           |              |              |           |             | (0/6)       |
|              |              |                  |             |             |           |           |           |              |              |           |             |             |
| H92          | C93          | D94              | K95         | L96         | H97       | V98       | D99       | F100         | E101         | N102      | F103        | R104        |
| X134         | X114         | X94              | X110        | X118        | X143      | X60       | X3        | -            | X23          | X132      | X84         | X61         |
| /            | /            | /                | /           | /           | /         | /         | /         | /            | /            | /         | /           | /           |
| -/           | 0/5          | 1/7              | 2/7         | 2/7         | 1/7       | 1/4       | 2/2       | -/           | 2/3          | 2/3       | 0/5         | 2/4         |
| <====>       | <====>       | <====>           | <====>      | <====>      | <====>    | <====>    | <====>    | <====>       | <====>       | <====>    | <====>      | <====>      |
| ->           | ->           | ->DNPY<====>     | ->K<====>   | ->K<====>   | ->K<====> | ->V<====> | ->V<====> | ->           | ->EQM<====>  | R*-<====> | ->A<====>   | ->R<====>   |
| 1.0          | (0/5)        | -1               | (1/2)       | 5.1         | (1/5)     | 1.1       | (1/2)     | 1.4          | (1/3)        | 3.1       | (0/3)       | -5          |
|              |              | (0/3)            |             |             |           |           |           |              |              |           |             | (1/3)       |
|              |              |                  |             |             |           |           |           |              |              |           |             |             |
| L105         | L106         | G107             | N108        | V109        | L110      | V111      | C112      | V113         | L114         | A115      | H116        | H117        |
| X72          | X76          | X2               | X49         | X25         | X39       | X73       | X87       | X50          | X38          | X33       | X93         | X104        |
| /            | /            | /                | /           | /           | /         | /         | /         | /            | /            | /         | /           | /           |
| 0/11         | 0/6          | 0/4              | 1/7         | 1/7         | 0/3       | 0/12      | 0/6       | 1/7          | 1/10         | 1/6       | 1/6         | 0/11        |
| <====>       | <====>       | <====>           | <====>      | <====>      | <====>    | <====>    | <====>    | <====>       | <====>       | <====>    | <====>      | <====>      |
| K*L<====>    | ->           | ->G<====>        | 0*-<====>   | ->*<====>   | ->L<====> | ->V<====> | ->*<====> | ->V<====>    | V*-<====>    | ->A<====> | A*-<====>   | ->WHC<====> |
| 2.3          | (0/4)        | 0.1              | (1/1)       | 1.1         | (1/3)     | 0.1       | (0/5)     | 1.1          | (1/3)        | 1.2       | (0/3)       | 0.3         |
|              |              | (1/1)            |             |             |           |           |           |              |              |           |             | (1/4)       |
|              |              |                  |             |             |           |           |           |              |              |           |             |             |
| F118         | G119         | K120             | E121        | F122        | T123      | P124      | P125      | V126         | Q127         | A128      | A129        | Y130        |
| X112         | X124         | X13              | X95         | X9          | X125      | -         | -         | X42          | X56          | X32       | X54         | X100        |
| /            | /            | /                | /           | /           | /         | /         | /         | /            | /            | /         | /           | /           |
| 1/9          | 1/3          | 4/5              | 2/9         | 2/9         | 0/7       | -/        | -/        | 4/4          | 4/4          | 0/4       | -/          | 1/4         |
| <====>       | <====>       | <====>           | <====>      | <====>      | <====>    | <====>    | <====>    | <====>       | <====>       | <====>    | <====>      | <====>      |
| WHC?-<====>  | ->G<====>    | G*H<====>        | K?EQM<====> | EQM*-<====> | ->*<====> | ->*<====> | ->*<====> | ->V<====>    | V*-<====>    | ->A<====> | ->A<====>   | A*-<====>   |
| 3.1          | (0/5)        | 1.1              | (0/1)       | 4.3         | (1/2)     | 1.1       | (-/-)     | 1.4          | (1/1)        | 0.2       | (0/3)       | 2.1         |
|              |              |                  |             |             |           |           |           |              |              |           |             | (0/7)       |
|              |              |                  |             |             |           |           |           |              |              |           |             |             |
| Q131         | K132         | V133             | V134        | A135        | G136      | V137      | A138      | N139         | A140         | L141      | A142        | H143        |
| X69          | X40          | X65              | X123        | X16         | X152      | X18       | X36       | X113         | X27          | X48       | X71         | X80         |
| /            | /            | /                | /           | /           | /         | /         | /         | /            | /            | /         | /           | /           |
| 0/5          | 1/8          | 4/10             | 4/7         | 4/7         | 2/7       | 1/8       | 1/12      | 2/10         | 1/6          | 3/10      | 1/3         | -/          |
| <====>       | <====>       | <====>           | <====>      | <====>      | <====>    | <====>    | <====>    | <====>       | <====>       | <====>    | <====>      | <====>      |
| ->           | ->           | ->V<====>        | V*V<====>   | V*V<====>   | A*G<====> | G*V<====> | V*V<====> | A?DNPY<====> | DNPY*V<====> | A*L<====> | L?A<====>   | A*WHC<====> |
| 1.1          | (1/2)        | 0.1              | (1/1)       | 4.4         | (0/4)     | 1.3       | (0/7)     | 2.2          | (1/4)        | 2.6       | (0/4)       | 2.3         |
|              |              |                  |             |             |           |           |           |              |              |           |             | (1/3)       |
|              |              |                  |             |             |           |           |           |              |              |           |             |             |
| K144         | Y145         | H146             |             |             |           |           |           |              |              |           |             |             |
| X107         | X82          | X17              |             |             |           |           |           |              |              |           |             |             |
| /            | /            | /                |             |             |           |           |           |              |              |           |             |             |
| 1/6          | 1/12         | 1/1              |             |             |           |           |           |              |              |           |             |             |
| <====>       | <====>       | <====>           |             |             |           |           |           |              |              |           |             |             |
| ->           | ->DNPY<====> | ->DNPY*WHC<====> |             |             |           |           |           |              |              |           |             |             |
| -1           | (0/3)        | 2.1              | (0/4)       | 3.2         |           |           |           |              |              |           |             |             |

## Chapter 5:

# **STARS: software for statistics on inter-atomic distances and torsion angles in protein secondary structures**

5.1 Introduction

5.2 Overview of STARS

5.3 Results and discussion

## Chapter 5:

# STARS: software for statistics on inter-atomic distances and torsion angles in protein secondary structures

## 5.1 Introduction

Structure determination by nuclear magnetic resonance (NMR) and structure validation involve estimation of interatomic distances and dihedral angles. The atom–atom distances are often derived from nuclear Overhauser effects (NOEs), whereas dihedral angles are derived from J-coupling constants and chemical shifts. Assigning each NOE peak in NOE spectroscopy (NOESY) to a specific pair of atoms is a challenging task even for a small protein because of the chemical shift degeneracy of different protons. Knowledge of interatomic distances for atoms located in each type of secondary structure facilitates the assignment of ambiguous NOEs resulting from chemical shift degeneracy on the basis of secondary structures that can be predicted with fair accuracy from chemical shifts or from amino acid sequence with computational techniques alone. However, if some of the NOE assignments are available (e.g. sequential NOEs), the distance knowledge helps in the determination of protein secondary structures too. Similarly, knowledge of dihedral angles for different types of secondary structures is very useful for deriving structural constraints from J-coupling constants. It can also be used to build internal motional models based on experimental J-coupling data.

Sometimes, information about interatomic distance and torsion angle could be very valuable for NMR assignment evaluation (Chapter 4). And besides

applications to NMR, this information may also be used to validate protein structures and compare protein folds. By summarizing different features of some special structure regions within a protein family or between different protein families, one may even analyze or explain some functional differences between several proteins.

Statistics on the distance and dihedral angle are often derived from many known protein structures. It is tedious to obtain the information from a large number of proteins. To the best of our knowledge, there is no tool available for computing the statistics though many tools can calculate distances and dihedral angles for only one given protein structure at one time. In this chapter, we present a software tool for statistics on interatomic distances and dihedral angles in protein secondary structures (STARS). STARS provides highly interactive visualization of statistical results. Its friendly window-based interface makes it extremely easy to use.

## 5.2 Overview of STARS

### 5.2.1 Composition of database

With the aid of CullPDB (Hobohm, Scharf et al. 1992), a non-redundant database of protein crystal structures was generated by extracting structural data from Protein Data Bank (Berman, Westbrook et al. 2000). Hydrogen atoms were added using MOLMOL (Koradi, Billeter et al. 1996). Proteins selected for our database meet the following criteria:

- (1) sequence identity <20%,

(2) resolution  $\leq 1.6\text{\AA}$  and  $R$ -factor  $\leq 0.25$  and

(3) residue number  $> 50$  and without non-standard amino acid and chain break.

The resulting database consisted of 576 protein chains, containing 124037 amino acid residues. Additional structures can be added to the database by users for their own interests.

### 5.2.2 Definition

The definitions and identifiers of amino acids, atoms and torsion angles used in STARS comply with the IUPAC recommendations in 1998 (Markley, Bax et al. 1998).

Secondary structure and chirality were assigned automatically for all proteins in the database using the DSSP method (Wolfgang Kabsch 1983). On the basis of biologist's preference, however,  $\beta$ -sheets were subdivided into three types. Totally, 10 types of secondary structures were defined (Table 5.1), including  $\alpha$ -helix,  $3_{10}$ -helix,  $\pi$ -helix, antiparallel- $\beta$ -sheets, parallel- $\beta$ -sheets and the combination of these two sheets, turn, bend,  $\beta$ -bridge and random coil.

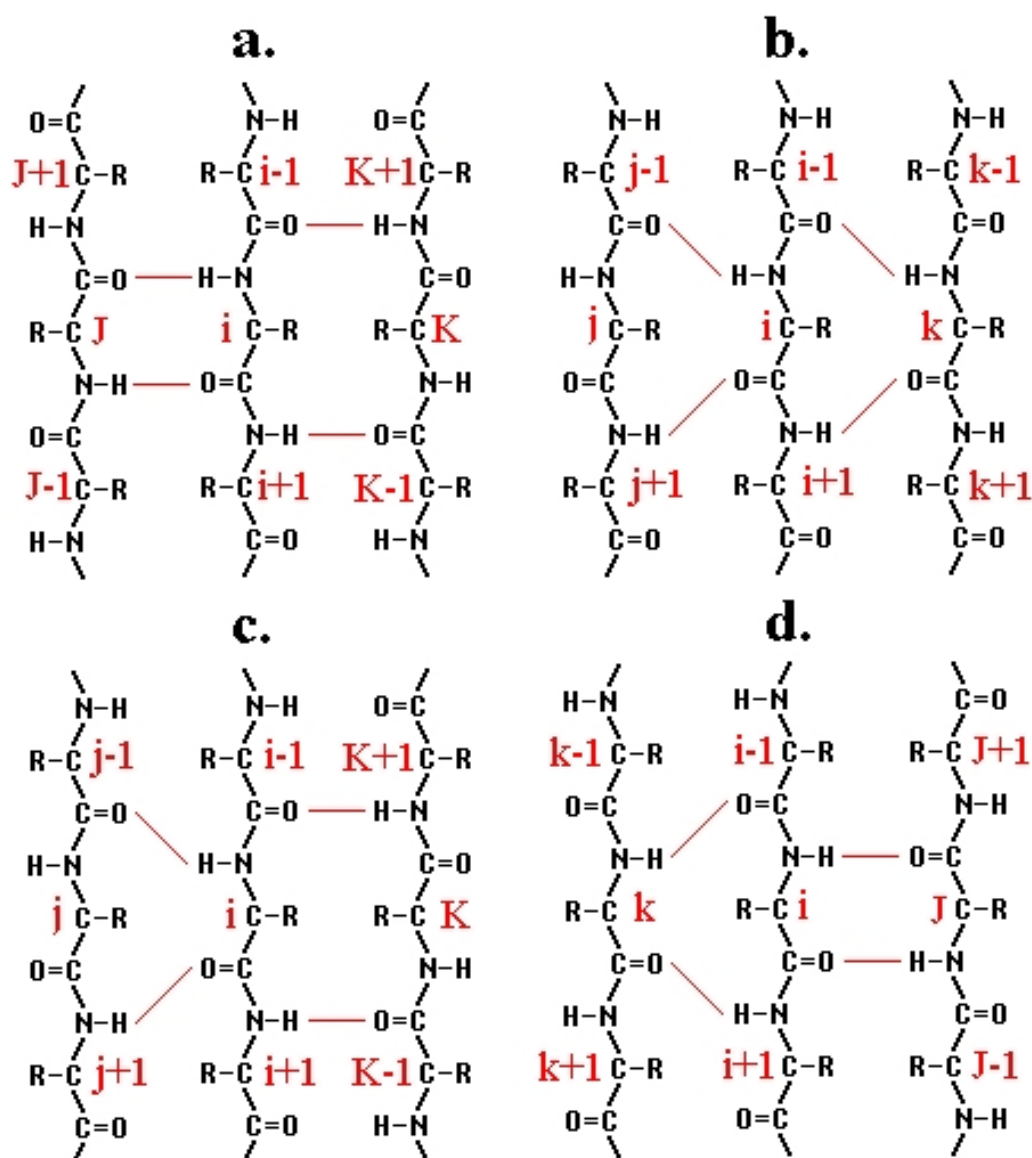
To obtain statistics on atom–atom distances and torsion angles, only relative positions among atoms in a protein chain are required. When the first and second atoms are located at residues  $i$  and  $i+n$ , respectively (where  $i$  is a positive integer while  $n$  is an integer), the relative position of the second atom with respect to the first one is denoted as  $n$ . The definition of residues  $i, J, K, j$  and  $k$  in a  $\beta$ -sheet is shown in Figure 5.1, the relative positions of the second atoms in residues  $J+n$ ,



$K + n, j + n$  and  $k + n$  with respect to the first atom in residue  $i$  are referred to as  $J + n, K + n, j + n$  and  $k + n$ .

**Table 5.1 Ten types of secondary structures defined in STARS and their one-letter symbols.**

| Symbol | Secondary Structure  |
|--------|--|
| A      | all types of secondary structure                                 |
| H      | $\alpha$ -helix (4-helix)  |
| G      | $3_{10}$ -helix (3-helix)  |
| I      | $\pi$ -helix (5-helix)   |
| T      | H-bonded turn (3-turn, 4-turn, 5-turn)                           |
| E      | $\beta$ -strand in antiparallel $\beta$ -sheet                   |
| P      | $\beta$ -strand in parallel $\beta$ -sheet                       |
| W      | $\beta$ -strand between antiparallel and parallel $\beta$ -sheet |
| C      | random coil  |
| S      | bend   |
| B      | isolated $\beta$ -bridge   |



**Figure 5.1** Definition of residues  $i$ ,  $J$ ,  $j$ ,  $K$ ,  $k$  in antiparallel (a), parallel (b) and mixed parallel and antiparallel (c and d)  $\beta$ -sheets.

$i$  is the residue under investigation, at which the first atom is located as shown in the main window (Figure 5.2);  $J$  is the  $\beta$ -bridge partner of residue  $i$  in an antiparallel ladder, with H-bond  $(i, J)$ , where  $i$  and  $J$  are the hydrogen donor and acceptor residues respectively, and H-bond  $(J, i)$ ;  $j$  is the  $\beta$ -bridge partner of residue  $i$  in a parallel ladder, with H-bond  $(i, j-1)$  and H-bond  $(j+1, i)$ ;  $K$  is the  $\beta$ -bridge partner of residue  $i$  in an antiparallel ladder, with H-bond  $(i+1, K-1)$  and H-bond  $(K+1, i-1)$ ;  $k$  is the  $\beta$ -bridge partner of residue  $i$  in a parallel ladder, with H-bond  $(i+1, k)$  and H-bond  $(k, i-1)$ .

The screenshot shows the STARS software interface. At the top, the title bar reads "STARS". Below it is a banner with the text "Statistics on Interatomic Distance & Torsion Angle in Protein Secondary Structures" and the authors "Zheng, Yu & Yang, Daiwen, Dept. of Biological Sciences, National University of Singapore".

The main interface is divided into several sections:

- Selection of Protein Structures:** Includes a "Structure List" button and two radio buttons: "Random" (set to "10 Structures (Very Fast)") and "Specific" (set to "1,3,5-12,345-523"). A note below says "Enter structure IDs and/or ID ranges separated by commas. For example, 1,3,5-12,356-493".
- Navigation:** Tabs for "Distance (Single)", "Distance (Batch)", "Angle (Single)", "Angle (Batch)", and "Result". The "Distance (Single)" tab is active.
- Secondary Structure (SS):** A list of checkboxes for secondary structure elements:
  - A (all)
  - H ( $\alpha$ -helix)
  - G (3<sub>10</sub>-helix)
  - I ( $\Pi$ -helix)
  - T (turn)
  - E ( $\beta$ -strand in antiparallel  $\beta$ -sheet)
  - P ( $\beta$ -strand in parallel  $\beta$ -sheet)
  - W ( $\beta$ -strand between antiparallel  $\beta$ -sheet & parallel  $\beta$ -sheet)
  - C (random coil)
  - S (bend)
  - B ( $\beta$ -bridge)
- Atom Specification:**
  - 1st Atom: Amino Acid (AA1) is "ASN", Atom Name (AN1) is "HN".
  - 2nd Atom: Amino Acid (AA2) is "ILE", Atom Name (AN2) is "MG".
  - Relative Position(s) (RP): "0,j-1~j+1,k".
  - Examples: "0,-1,2~5,j,j,k,k" and "j-1~j+2,k~k+3".
- Chirality:**
  - 1st Atom (CH1): Radio buttons for "ignore (=)", "0 < a < 180 (+)", and "-180 < a < 0 (-)".
  - 2nd Atom (CH2): Radio buttons for "ignore (=)", "0 < a < 180 (+)", and "-180 < a < 0 (-)".

At the bottom, there are three buttons: "Submit" (with a green checkmark), "Clear" (with a red circle and slash), and "Exit" (with a red X).

Figure 5.2 STARS user interface - Main window with the page for interatomic distance statistics in a single mode.

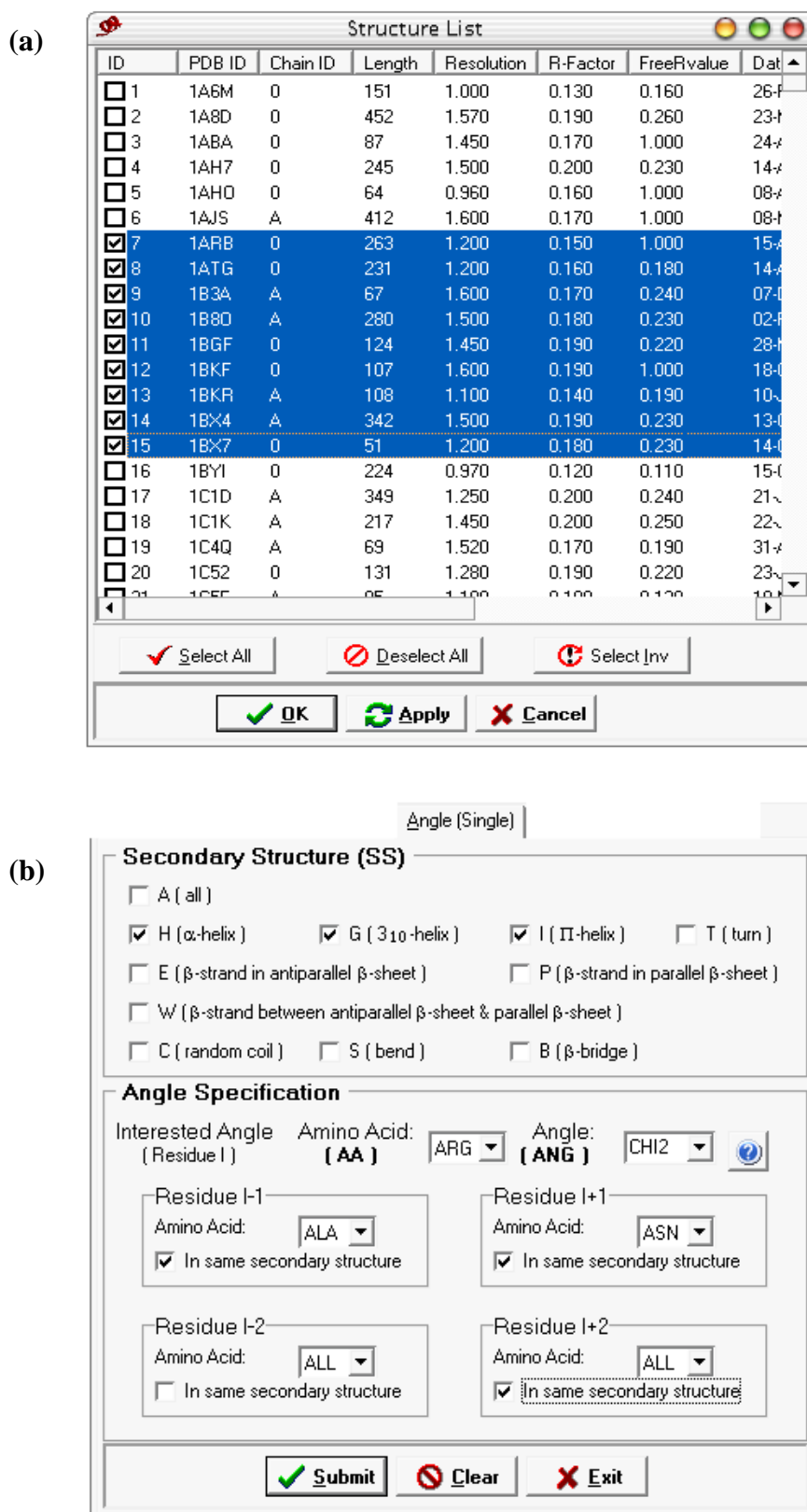


Figure 5.3 STARS user interface – (a) Window for selection of protein structures. (b) Page for torsion angle statistics in a single mode.

(a) Distance (Batch)

**Job Description List**

| SS  | AA1 | AN1 | CH1 | AA2 | AN2  | CH2 | SD  | Output File Name       |
|-----|-----|-----|-----|-----|------|-----|-----|------------------------|
| H   | ASP | HA  | +   | GLU | HB1  | -   | 2   | C:\H ASP HA + GLU ..   |
| P   | TYR | CE2 | =   | VAL | HG23 | -   | -3  | C:\P TYR CE2 = VAL..   |
| B   | LYS | CB  | -   | ALA | O    | +   | 2~4 | C:\B LYS CB - ALA O..  |
| A   | ALL | HN  | =   | ALL | HN   | =   | 3   | C:\A ALL HN = ALL ...  |
| HGI | TRP | HA  | =   | ARG | HA   | =   | 3   | C:\HGI TRP HA = A...   |
| C   | ALL | HA  | +   | ALL | QB   | +   | 0   | C:\C ALL HA + ALL ...  |
| S   | ALL | N   | =   | ALL | N    | =   | 2   | C:\S ALL N = ALL N ... |
| T   | ALA | HA  | -   | ALA | HN   | -   | 0   | C:\T ALA HA - ALA ...  |

(b) Angle (Batch)

**Job Description List**

| SS | AA  | ANG   | I-1 |   | I+1 |   | I-2 |   | I+2 |   | Output File Name |
|----|-----|-------|-----|---|-----|---|-----|---|-----|---|------------------|
| H  | ASN | PHI   | ASN | + | ARG | + | ALL | - | ALL | - | C:\H ASN PHI A   |
| W  | GLU | CHI42 | ARG | + | ALA | + | GLN | + | ARG | + | C:\W GLU CHI4    |
| Iw | TYR | PSI   | ALL | + | ALL | + | ALL | - | ALL | - | C:\W TYR PSI     |
| B  | CYS | OM... | ALA | + | GLN | + | ALL | - | ALL | - | C:\B CYS OMEC    |

**Figure 5.4 STARS user interface – (a) Page for interatomic distance statistics in a batch mode. (b) Page for torsion angle statistics in a batch mode.**



The statistics can be done over all residues, or the residues in one or more specific secondary structures selected by users (Figure 5.2). The relative position(s) of the second atom(s) with respect to the first atom can be specified by a single expression (e.g. 2 or  $J-1$ ), a series of expressions (e.g. -2, 0, 1,  $J-1, K+1$ ), a range of numbers (e.g.  $-2\sim 2$  or  $j-2\sim j+2$ ), or a combination of different expressions (e.g. -3,  $-1-1$ ,  $k-2\sim k+1$ ). If some of the specified atoms are not located in the selected secondary structure(s), the output will not contain distances or angles involved in these atoms.

The statistics can be obtained in a single (Figure 5.2 and 5.3b) or batch mode (Figure 5.4). Since the batch mode uses a parallel process algorithm, it is  $\sim 10$  times faster than the single mode for obtaining the same amount of information. With a job editor, jobs can be created, saved, loaded, edited, sorted, deleted or moved easily in the job list, and submitted at the user's convenience.

The statistic results are displayed in a 3D color-bar-style chart in the result analysis window (Figure 5.5). Almost all features of the chart can be reset by users in terms of color, zoom, mark, label, rotation, range, grid, etc. The software allows the users to view, compare, select, sort, save or load statistics through a result display window. All data files are saved as a common ASCII format which can be read by a normal text editor.

A detailed manual is accessible by clicking the help button in the main window or pressing the F1 key.

### 5.3 Results and discussion

In summary, STARS is a well designed graphics package for performing statistics on interatomic distances and torsion angles in protein secondary structures from a protein crystal structure database. It allows users to obtain both the graphical view and the text format of distributions of the distances and angles for atoms located in 10 types of protein secondary structures. We believe that it will facilitate assignment of ambiguous NOESY peaks, structure determination by nuclear magnetic resonance, structure validation and comparison of protein folds.

All data, documents and execute files are freely downloadable at <http://yangdw.science.nus.edu.sg>. The software works appropriately on Windows system, without any compilation or installation.



## Chapter 6:

# **NMRspy: software package for NMR spectroscopy visualization, analysis and management**

6.1 Introduction

6.2 Feature and advantages of NMRspy

6.3 User's interface

6.4 Results and discussion

## Chapter 6:

# NMRspy: software package for NMR spectroscopy visualization, analysis and management

## 6.1 Introduction

One can use a series of NMR experiments to determine the structure and dynamics of a molecule or protein. The experiments on a reasonably complicated molecule like a protein can generate literally gigabytes of data. Furthermore, to interpret these data requires correlating the information from different experiments. One of the greatest challenges to a modern NMR spectroscopist is to visualize, analyze and manage the great variety of information that is obtainable through NMR techniques.

NMRView (Johnson and Blevins 1994), Sparky (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco), ANSIG (Kraulis, Domaille et al. 1994), and Xeasy (Bartels, Xia et al. 1995) are most widely used software packages for management and analysis of protein NMR datasets. They are capable of working with 2D, 3D and 4D NMR data sets, displaying multiple flexible views on one or more NMR spectra, picking, assigning and integrating peaks in different spectra, and maintaining a flexible database. They also provide a variety of facilities to support the analysis of spectra, such as automatic peak picking, corresponding cursors in different windows tracking each other automatically, and facilitated peak analysis and interactive peak editing. Some even offer sets of flexible commands, data structures and display routines.

With the increase in protein size, 3D, 4D or even higher dimension NMR spectra are often used. With the limitation of sweep width, however, the peaks on these spectra could be folded several times. A lot of troubles and inconveniences appear when one analyzes these folded-spectra. The existing NMR spectroscopy visualization and analysis programs mentioned above lack the strategy to deal with this problem. Although some of them provide facilities to revise the folded peaks into their correct frequency, none of them is capable of visualizing the correlations between folded and unfolded spectra, still less synchronizing them. There is a great demand of a program that can deal with the folded spectra so that users can visually analyze them the same way as they do unfolded spectra.

Inaccurate peak-picking of the 4D spectrum is another major drawback of those existing programs. All existing programs were designed for 2D spectrum initially, whose peak-picking algorithms have been proved to be inaccurate when applying to a complex 4D spectrum, and whose interactive peak-picking facilities are so inconvenient that few would correct the frequencies of a 4D peak. It will be a great relief if a new program can be developed to provide some accurate peak-picking algorithms and convenient interactive peak-picking facilities for high-dimensional NMR spectra.

Here I present a new software package, NMRspy (NMR spectra pinpoint analysis system), for the visualization, analysis and management of NMR spectra. It has been designed to meet all those demands mentioned above and to create a platform which minimizes the number of limits on the user, yet emphasizes on easy-to-use. To facilitate the analysis of complex, crowded and folded high-dimensional spectra, our software package contains a variety of function and

analysis routines that are integrated with the spectral display features. Moreover, the program has intrinsic project management capabilities so that the organization and maintenance of vast quantity of information can be a simple and convenient task.

## 6.2 Feature and advantages of NMRspy

### 6.2.1 Intrinsic capabilities

NMRspy was written in the Java programming language. Java aids one in developing more understandable and reusable code. It puts a lot of emphasis on early checking for possible problems, later dynamic (runtime) checking, and eliminating situations that are error prone. It also has a dynamic automatic-memory-manager that eliminates the possibility of overwriting memory and corrupting data which could happen in C or C++. Benefited from these intrinsic superiority, NMRspy should be more robust (less likely to crash) and users need not worry too much about the computer memory size.

Thanks for the Java's intrinsic support for threading, it becomes possible for NMRspy to perform certain operations in a separate thread of execution, essentially allowing "multitasking" within a single application. On multiprocessor computers each thread can operate on a separate processor and speed up the operation several times. Also thanks for the object oriented nature of Java, the existing NMRspy modules or new modules can simply "plug and play" with the whole package.

The same Java code will run on a variety of systems with a variety of CPU and operating system architectures. With the Java virtual machine installed, which can be freely downloaded from [www.java.com](http://www.java.com), all users can use the same NMRspy distribution without any compilation or installation, even if one is on Windows, another on Macintosh, and still another on Unix. And the design of NMRspy makes it very easy to incorporate Java libraries. This will allow the addition of many new features to NMRspy.

## 6.2.2 Capability of analyzing folded-spectrum

Some spectra, especially 3D and 4D spectra, may be collected with a sweep width narrower than the frequency range of the spectral peaks (for example, Bax et al., 1990, 1991). In these spectra, peaks are folded to a position that differs from their true position by an integer multiple of the sweep width.

### 6.2.2.1 Proper frequency display of aliased peaks

One can differentiate folded and unfolded peaks according to their features (e.g. peak intensity and position) or experimental methods (e.g. comparing correlated spectra).

During the process of analyzing a folded spectrum or determining a folded peak, generally, one would want to know the possible true chemical shifts (or pre-folded chemical shift) of a peak when one or several sweep widths are added to or subtracted from its original chemical shift. And after the folded peaks are determined, one always hopes that these peaks could have their pre-and post-folding chemical shifts recorded so as to find their chemical shift easily in further NOE assignments.

Without software supporting, one usually has to put a lot of effort in tedious jobs such as calculating the possible chemical shifts that may correspond to a particular position or the true position of a folded peak. The manual conversion of folded position into unfolded one seems unremarkable, but in a real process of spectral analysis, facing hundreds of thousands of folded peaks, manual calculation will not only greatly slow down the progress of the analysis, but also tend to be error-prone. When performing the manual chemical shift conversion, one usually takes a rough value with one or none decimal place, this may cause significant risk in chemical shift analysis and peak matching, which will consequently lead to a lot of mistakes that should be avoided.

NMRspy provides a real-time chemical shifts conversion function in the spectral view status bar (Section 6.3.2.3). On the basis of user's preference, when the cursor moves to a certain position on a spectral view, the status bar not only shows the original chemical shifts of that position, but also reveals its possible true chemical shifts. The program could automatically calculate the "true chemical shifts" by adding or subtracting up to 5 integer multiple of the sweep widths from the original chemical shifts, and display them alongside with the original chemical shifts and other information in the status bar. Users could simply move the mouse and read all the accurate chemical shifts they want to know from the status bar after customizing it for their convenience.

Each peak displayed by NMRspy has a set of pre- and post-folding chemical shifts. One or more sweep widths could be easily added to or subtracted from a peak's original position using the Peak Editor dialog (Section 6.3.4.1). With both original and corrected chemical shift sets readily accessible, users

would never need to manually calculate any rough value. As a matter of fact, users don't even need to know the values of the sweep widths at all if they can properly combine and apply the facilities provided by NMRspy.

### 6.2.2.2 Spectra synchronization & cursor correlation

Conventionally, when analyzing a group of correlated spectra, spectroscopists often use a function called "Spectra Synchronization" which synchronizes the axes of 2 or more spectral views. In a group of synchronized views, the positions of the crosshairs in each window are correlated with the positions in every other window in an automatic and intuitive way. The crosshair in different windows automatically tracks each other in an appropriate manner, so that spectroscopists can quickly determine whether two peaks in different spectra match each other in chemical shifts.

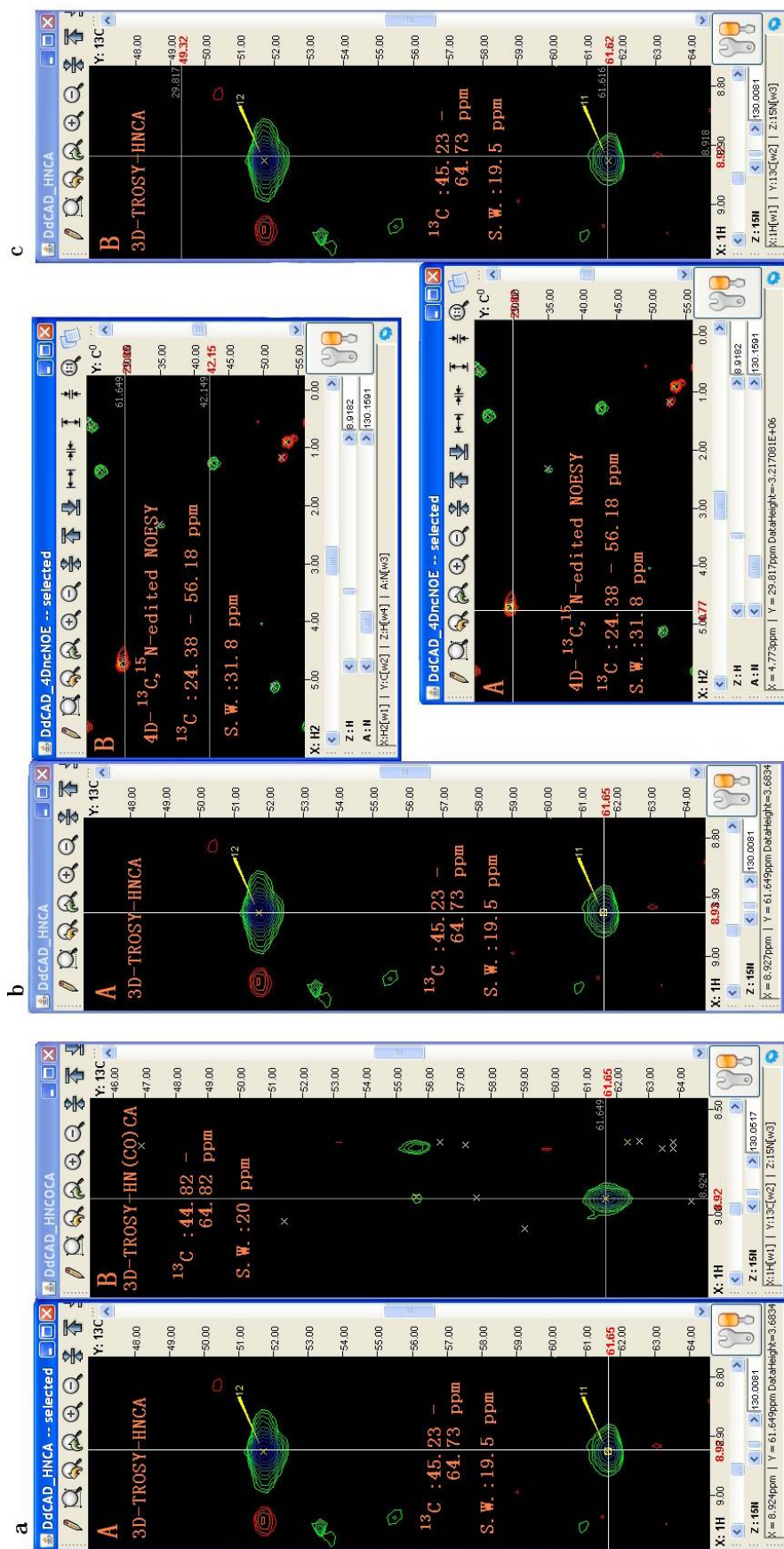
However, with the increased size of protein under NMR study and the limitation of sweep width, spectroscopists often need to correlate and analyze a series of folded multidimensional spectra that have different sweep widths and folding rates. In this case, peaks that should have the same chemical shift may be located differently in different spectra after folding, invalidating the "Spectra Synchronization" function.

NMRspy offers a much more powerful "Spectra Synchronization" function. In addition to synchronizing the unfolded area, it can automatically delineate the possible post-folded location in corresponding spectrum (Figure 6.1). In NMRspy, if the user synchronizes a spectral view A with B, when the cursor

moves to location  $x$  in view A, the program will automatically mark three kinds of corresponding locations in view B:

- 1) If  $x$  falls in the sweep width of spectrum B, the direct corresponding locations in the unfolded area of view B will be marked. (Figure 6.1 a)
- 2) When  $x$  falls out of the sweep width of spectrum B, possible post-folded corresponding locations in view B will be marked. (Figure 6.1 b, c)
- 3) Assume that location  $x$  is already folded, all the possible corresponding locations of  $x$  will be marked in view B. (Figure 6.1 b, c)





**Figure 6.1** Corresponding crosshairs in different windows.

When cursor (marked by white crosshair) moved to position  $^1\text{H}_\text{N}$ :8.924 ppm in spectrum A, a gray crosshair marked the same position in spectrum B. (b) When cursor moved to position  $^{13}\text{C}$ :61.65 ppm in spectrum A, two gray lines marked the possible corresponding positions in spectrum B. One labeled with 61.649, laid in 29.85 ppm marked the possible position that folded once in spectrum B (61.649-31.8 ppm). The other one assumed that the position  $^{13}\text{C}$ :61.65 ppm in spectrum A is already folded. It marked the “true” corresponding position 42.15 ppm, which is the original chemical shift 61.649 ppm subtract the sweep width (19.5 ppm) of spectrum A. (c) When cursor moved to position  $^1\text{H}_\text{N}$ :8.9182 ppm in spectrum A, in spectrum B, one gray line marked the  $^1\text{H}_\text{N}$  position, the second one marked the  $^{13}\text{C}$  position that folded once in spectrum B(29.817+19.5=49.32 ppm), and the third one marked the  $^{13}\text{C}$  position that folded once in spectrum A (29.817+31.8=61.616 ppm).

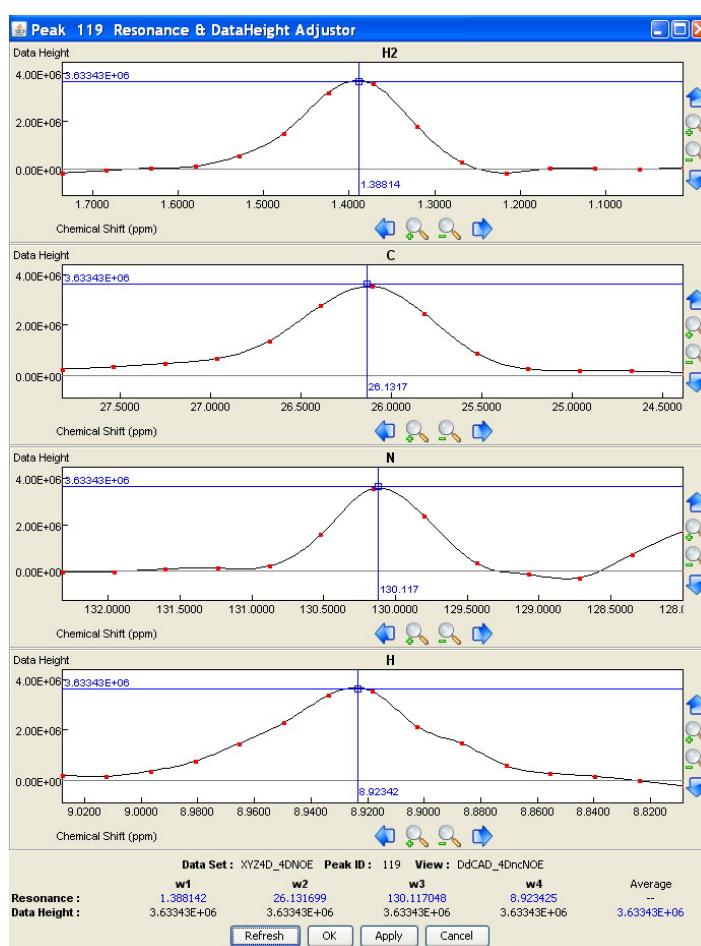
### 6.2.3 Multi-dimension-peakpicking capability

Most of the useful information present in NMR spectra can be extracted in the form of peaks containing information about its positions and intensities. Accordingly, NMRspy provides capabilities for the extraction and analysis of peak data. The first tool required is a peak picker. NMRspy incorporates a simple but fairly robust peak picker similar to that used in many other NMR programs (for example, Kleywegt et al., 1990). Peaks are identified as local maxima in the spectrum. The actual peak position is determined by interpolating to the maximum of a polynomial fit to the five data points nearest to the maximum. The determination of position is repeated along all dimensions of the data set.

However, NMR spectrum peakpicking has never been an easy task for the computer or spectroscopist. Artificial noise, overlapped signal and anomalous peak-shapes present in different dimensions make it a huge challenge to pick peaks in a multidimensional spectrum, especially for those high-dimension low-resolution NMR spectra. Although great efforts have been devoted into developing automatic peakpicking algorithms and programs, none of them can match the accuracy of interactive-peakpicking by a sophisticated spectroscopist. It's not surprising that for most spectroscopists, interactive-peakpicking is still an essential step or at least a necessary reinforcement for the automatic procedure throughout their work.

NMRspy not only allows users to manually add, delete or move peaks in two-dimension plane, as what other programs can do, but also provides a peak Resonance & DataHeight Adjustor (Figure 6.2), which allows users to pinpoint the resonance and intensity of a peak in multi-dimension spectrum. For a

particular peak, the Resonance & DataHeight Adjustor can display the data points and fitting curves around it from all dimensions. Users will find it very convenient to manipulate the interface: display as many data points as they want; enlarge, reduce or move the curves from different dimensions respectively; drag the peak centre along the fitting curves freely with the data height automatically detected. If the peak centre in one dimension has been changed, a flip on the “Refresh” button could quickly synchronize the other displays.



**Figure 6.2 Peak Resonance & DataHeight Adjustor.**

With the aid of this facility, NMRspy users could investigate a peak in great detail and distinguish a lot more peaks that have been overlapped by others.

Moreover, actually looking at the curves shape and distribution of peaks can be very important for discovering artifacts in spectra, which may lead to identification of extra peaks or missing peaks that result from conformational dynamics, multiple species in solution, or ligand binding effects.

#### 6.2.4 Project management capability

A spectroscopist may spend considerable time creating and configuring a group of display windows, importing protein sequence or generating chemical shift lists. To avoid having to re-create the displays, re-import the data or re-generate the lists in the case of a computer malfunction or when starting up a new session of work, it is very useful to save the current state of the window parameters and database.

In order to meet this demand, NMRspy introduced the concept of project management. Users can enclose unlimited amount of spectral view and dataset into a project. A set of protein sequence, atom list and assignment list could also be enclosed into it.

A project contains almost all the parameters describing the NMRspy display environment, including window names, sizes, locations and display attributes, as well as datasets opened in the session. When restarting, NMRspy simply opens the project file and the previous database will be loaded and the display environment will be recreated.

NMRspy features an auto-save function that saves an opened project automatically, helping to reduce the risk or impact of data loss in case of a system crash or freeze. Auto-saving is done in predetermined intervals which

could be customized by users. It is quite a good alternative for users who don't have the good habit of saving files regularly.

With project management, NMRspy users can easily switch between different projects, or continue the unfinished work any time. Those extension program developers can also extend the project manager's functions to save more useful data, making it easier to organize the involved data and interfaces of a complex analysis procedure. (see Chapter 7).

### **6.2.5 Spectral view simplification capability**

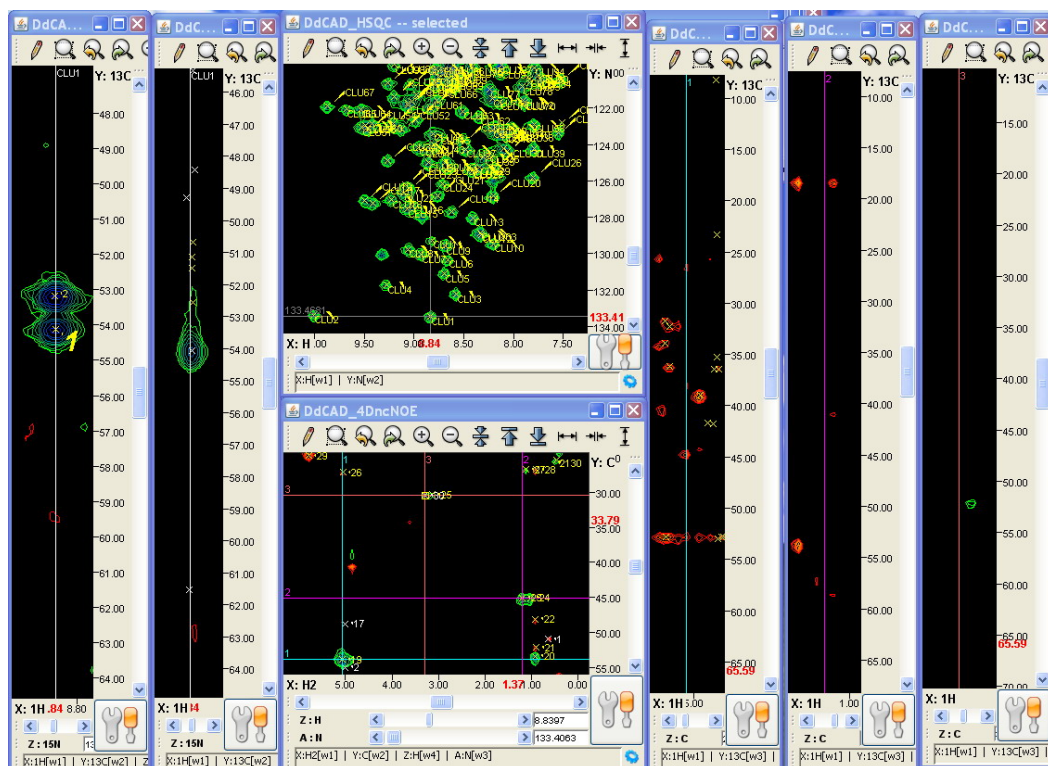
Many NMR spectral analysis strategies, especially those for large protein study, require proper combination of information retrieved from several experiments. This means several NMR spectra may need to be displayed at the same time with dozens of spectral view windows showing different regions of them to facilitate a complex analysis.

With no arbitrary limits on the number of spectral views that can be opened, a major bottleneck of most widely-used NMR spectral visualization software is the space limitation of computer screen. In a traditional (or standard) spectral view window, control and status widgets (e.g. spectrum control bar, chemical shift scales, scroll bar, status bar) usually take up a lot of display space. These "barren" widgets may waste more than 50% of the valuable screen space when numerous spectral view windows are opened, making the information-rich contour plots "compressed" in some small and separate areas (Figure 6.3 a). It is not uncommon that a spectroscopist places several screens side by side to enlarge the visualized spectral area.

NMRspy provides a novel feature -- “Simple Layout” for spectral view window to alleviate the lack of screen space problem. When users press Esc, all “barren” widgets in a spectral view window will be hidden away, with only the contour plot and window’s titles left. Most of the hidden widgets’ functions are replaced by other means, for example, the commands carried by spectrum control bar could be accessed through right-click popup menu or keyboard shortcut, the chemical shift scales replaced by chemical shift labeled crosshairs, and the operations usually by scroll bars performed by the mouse wheel and Cursor Keys instead. The Simple Layout hardly causes any inconvenience to users, and in case a widget is needed (e.g. Status Bar), users could drag it out of the spectral view window before applying the Simple Layout so that it will remain visible and function normally.

Simple Layout raises the utilization of screen space to almost 100% (Figure 6.3 b). It allows for the creation of very complex arrangements of spectra view to further facilitate a complex analysis. The analysis with multiple spectral slices (or spectra) has also been made more intuitive and easier with the gap-free arrangements.

a



b

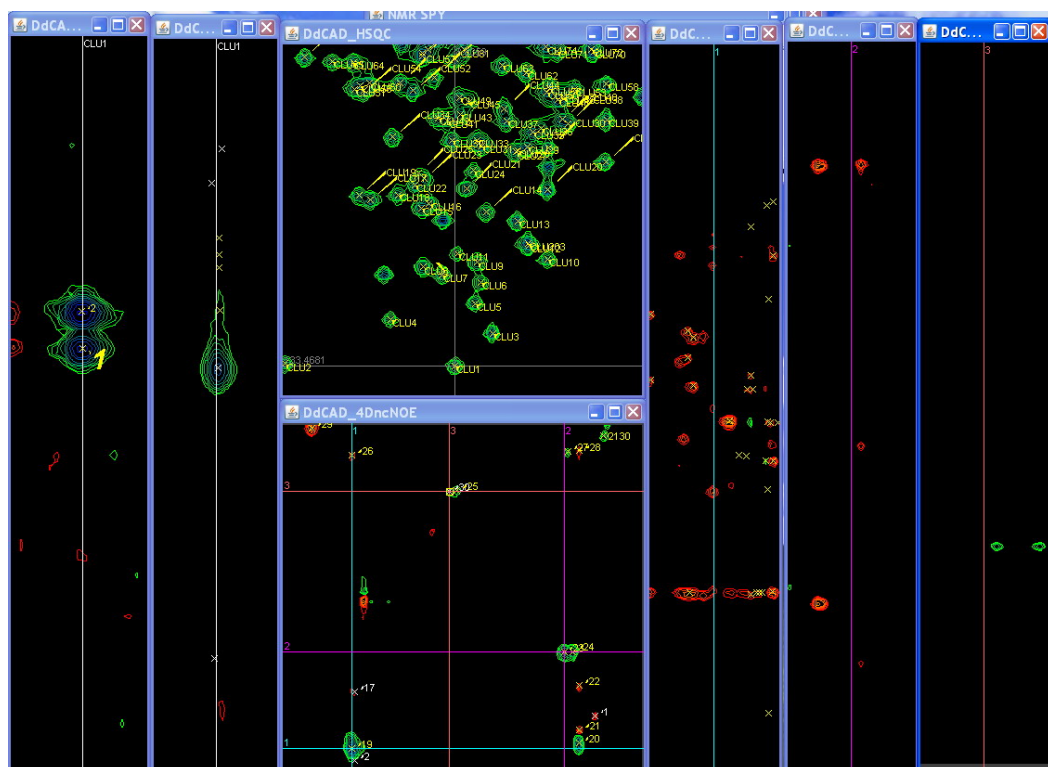
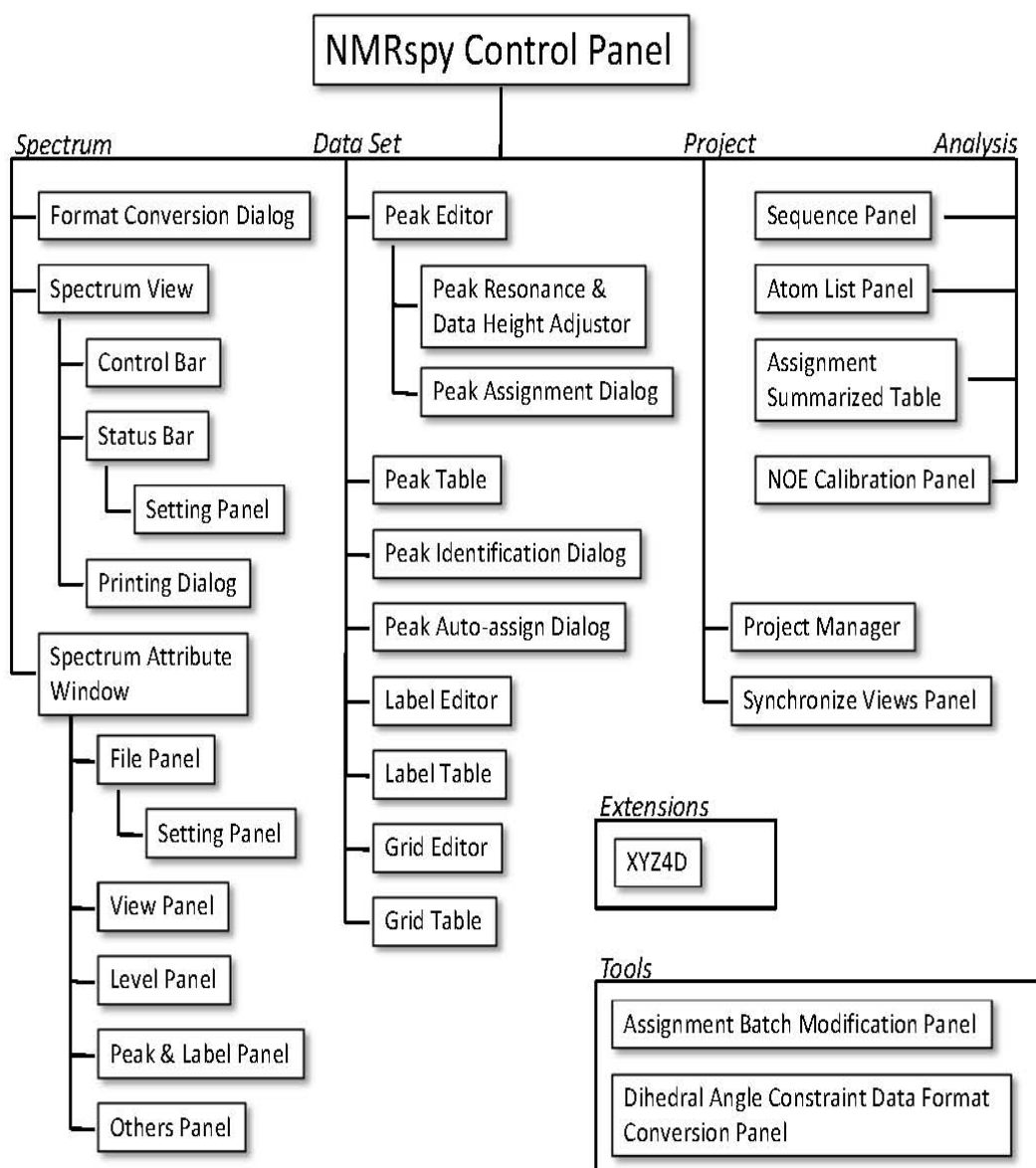


Figure 6.3 Multiple spectral views with standard layout (a) and simple layout (b).

## 6.3 User's interface

NMRspy uses separate windows to perform different functions: overall control, spectrum visualization, attribute configuration, and other operations. (Figure 6.4) These windows are depicted in the following sections of this chapter.



**Figure 6.4 Overall Diagram of interfaces in NMRspy**



### 6.3.1 Control panel

The Control Panel is the first window to be displayed when NMRspy starts up, and consists of a menu bar, a log-display pane and some log-configuration buttons (Figure 6.5).

Logs record the user's important operations. There are two display modes that can be used. Under “Normal” mode, only a summary of operations will be displayed. Under “Detail” mode, the IDs of those objects involved in the operation will also be displayed. With them, NMRspy users can easily examine their operation history and correct the mistake if there is any. For example, if a peak has been mistakenly deleted, the user can track the peak ID by examining the detail-logs and record it with Peak Editor (Section 6.3.4.1) or Peak Table (Section 6.3.4.2). Users can customize the amount of logs that are maintained by NMRspy with Project Manager (Figure 6.6) or save the logs into a text file by using the button located at the down-right corner of the control panel.

The menus available through the menu bar may be used to bring up various other NMRspy windows. The options available through each menu item are summarized as below.

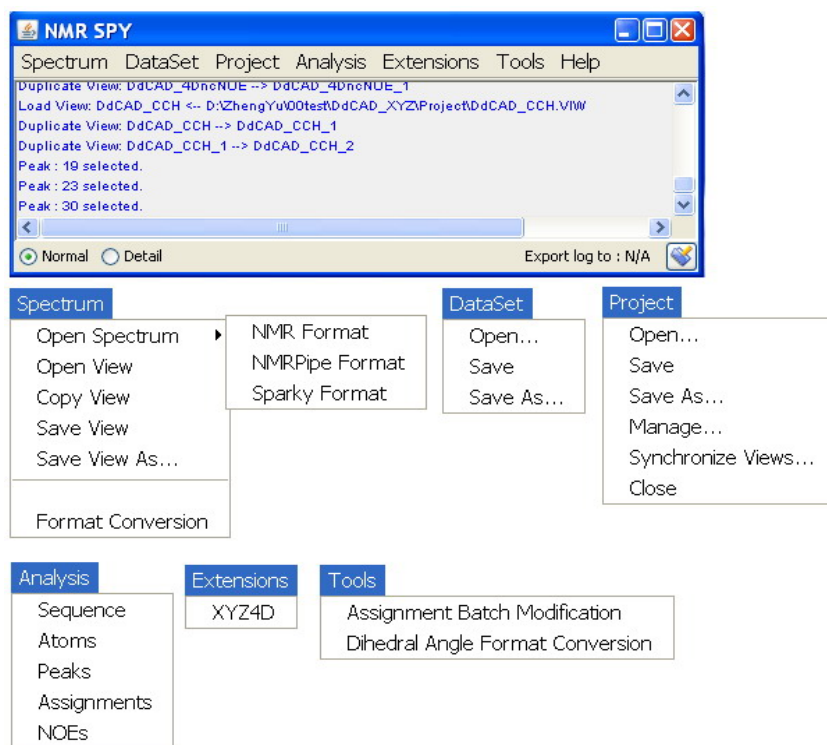


Figure 6.5 NMRspy Control Panel and its menus.



Figure 6.6 Project Manager Window.

### 6.3.1.1 Spectrum menu

- **Data format Conversion**

With the format conversion dialog (Figure 6.7), the user can convert a spectral file in NMRpipe format to the NMR format, change its orientation or provide its title, user and comments at the user's convenience.

Files containing NMR data are commonly found in two different formats: serial format and sub-matrix format. In the serial format (e.g. NMRpipe format), the experimental progression of data points (a complete set of points in one row, followed by another set of points in the next row) corresponds exactly to the order of data points in the actual data file. In the sub-matrix format (e.g. SPARKY, NMRView format) a portion (the block size) of a row is followed by further portions of subsequent rows before the remaining data points of the row are found. Of the two, the sub-matrix format provides a layout of data on the disk that is much more efficient for NMRspy to access. This is particularly true if the data are to be displayed in different orientations or if multiple datasets are to be analyzed simultaneously. Instead of accessing files with a serial format directly, it is generally better to convert them to the more efficient sub-matrix format.

NMR format, a new sub-matrix format defined by us, has been specially optimized for NMRspy. Although NMRspy is capable of reading and displaying other spectral data formats (e.g. NMRpipe, SPARKY) and exerting all NMRspy's powers, users are recommended to convert data files in other formats to the NMR format before using them.

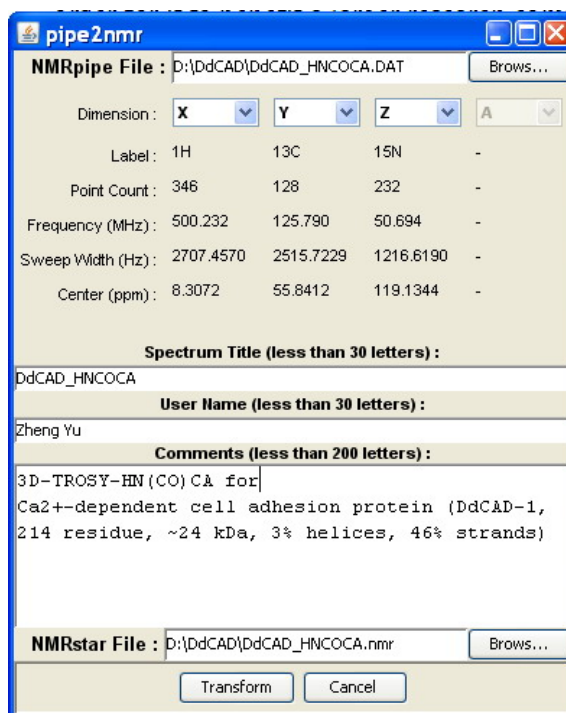


Figure 6.7 Format Conversion Dialog

- **Open Spectrum**

It brings up a file selection panel for reading a spectral data file. NMRspy currently supports three spectral data format: NMR format, NMRpipe format and SPARKY format. The spectrum will be appended to the current project and display in a “Spectral view” window.

- **Open View / Save View / Save View As**

Save the current “Spectral view” window to a “.VIW” file, or load such a window from a “.VIW” file previously saved.

- **Copy View**

Duplicate the current “Spectral view” window.

### 6.3.1.2 DataSet menu

Save the current dataset to a “.DST” file, or load a dataset from a “.DST” file previously saved into the current project.

### 6.3.1.3 Project menu

- **Open / Save / Save As**

Save the currently opened project and its data (e.g. spectral view, dataset, sequence...) to a “.PRO” file and a folder with the same name in the same directory, or close the currently opened project and load a previously saved project.

- **Manage**

Select this to open a Project Manage Panel (Figure 6.6) that can be used to refer to or close spectral views and datasets, set up the log and auto-save parameters.

- **Synchronize Views**

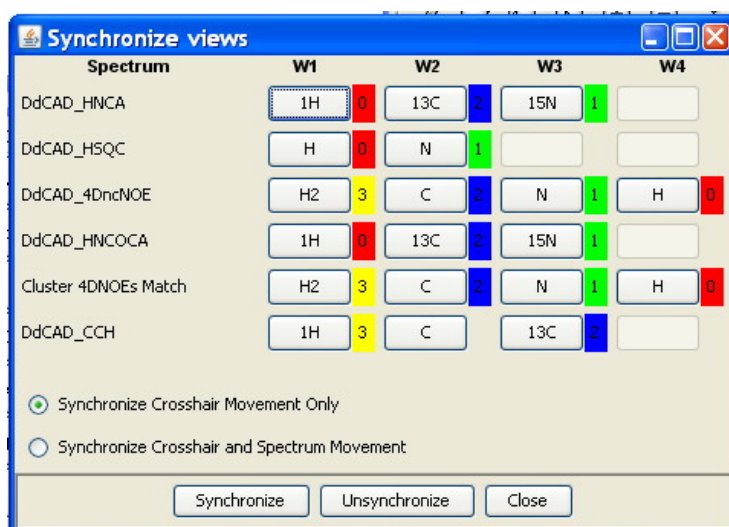
Select this to bring up a “Synchronize Views” panel (Figure 6.8), which could be used to synchronize axes of two or more spectral views so that scrolling one view causes others to scroll.

The Synchronize Views Panel lists the axes for every view in the project. Users may select two or more axes to synchronize and press the synchronize button. Synchronized axes have the number 0, 1, 2, 3, ... displayed next to them, and axes labeled with the same numbers are synchronized. To remove

synchronizations, select one or more axes you no longer want to synchronize and then press the unsynchronize button.

- **Close**

Select this to close currently opened project without saving and open a new one.



**Figure 6.8 Synchronize Views Panel**

### 6.3.1.4 Analysis menu

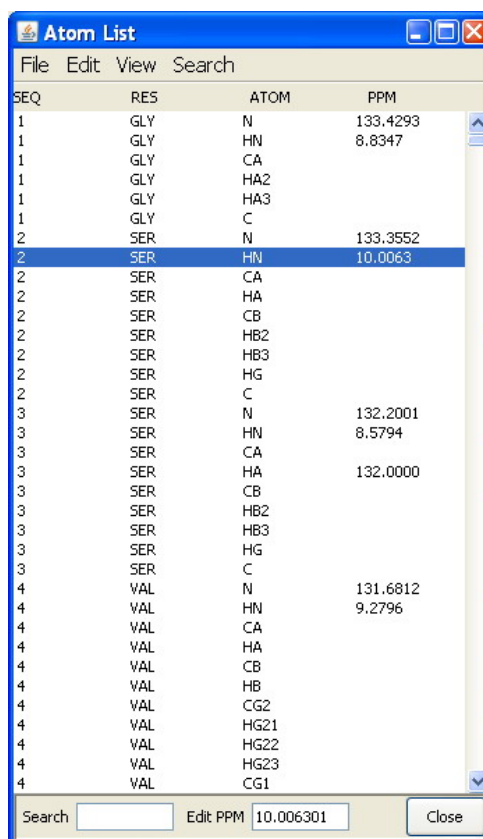
- **Sequence**

Select this to bring up a display panel to display, load or save the of amino acid sequence of a protein.

- **Atoms**

Select this to bring up the Atom List Panel (Figure 6.9) that is used to keep track of chemical shift assignments. The panel allows users to export or import

the atom list, extract chemical shifts from datasets, display the atom list in different patterns or search for a given string in different data fields.



| SEQ | RES | ATOM | PPM      |
|-----|-----|------|----------|
| 1   | GLY | N    | 133.4293 |
| 1   | GLY | HN   | 8.8347   |
| 1   | GLY | CA   |          |
| 1   | GLY | HA2  |          |
| 1   | GLY | HA3  |          |
| 1   | GLY | C    |          |
| 2   | SER | N    | 133.3552 |
| 2   | SER | HN   | 10.0063  |
| 2   | SER | CA   |          |
| 2   | SER | HA   |          |
| 2   | SER | CB   |          |
| 2   | SER | HB2  |          |
| 2   | SER | HB3  |          |
| 2   | SER | HG   |          |
| 2   | SER | C    |          |
| 3   | SER | N    | 132.2001 |
| 3   | SER | HN   | 8.5794   |
| 3   | SER | CA   |          |
| 3   | SER | HA   | 132.0000 |
| 3   | SER | CB   |          |
| 3   | SER | HB2  |          |
| 3   | SER | HB3  |          |
| 3   | SER | HG   |          |
| 3   | SER | C    |          |
| 4   | VAL | N    | 131.6812 |
| 4   | VAL | HN   | 9.2796   |
| 4   | VAL | CA   |          |
| 4   | VAL | HA   |          |
| 4   | VAL | CB   |          |
| 4   | VAL | HB   |          |
| 4   | VAL | CG2  |          |
| 4   | VAL | HG21 |          |
| 4   | VAL | HG22 |          |
| 4   | VAL | HG23 |          |
| 4   | VAL | CG1  |          |

**Figure 6.9 Atom List Panel.**

- **Peaks**

Select this to bring up a Peak Editor Panel which could be used to navigate through peak list and examine or modify the peaks. Find more details in section 6.3.4.1.

- **Assignments**

Select this to display a table of chemical shifts (Figure 6.10). The chemical shift for a given an atom shown in the table is the average obtained from all peaks in all spectra having the same molecule and atom names. To see the

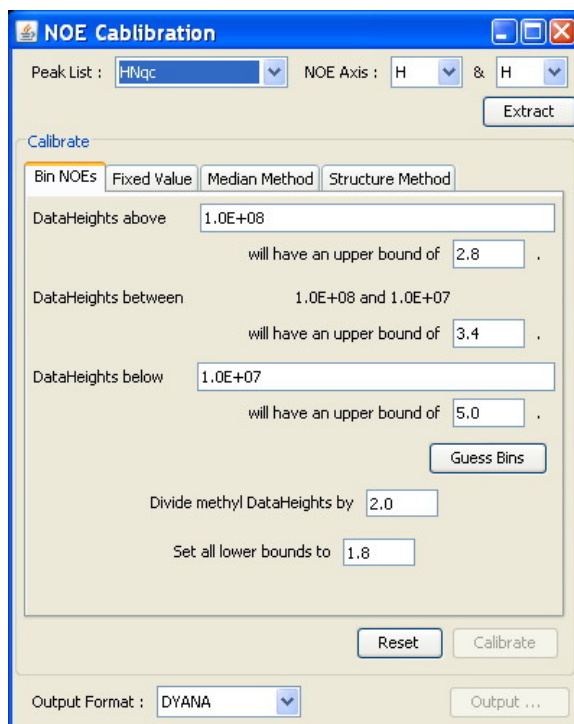
number of assignments contributing to each average chemical shift, click “Show” - “Assignment Counts” switch. You can also have the table entries to include the standard deviation of the peak positions contributing to each chemical shift by clicking on the “Show” - “Standard Deviation” switch. The chemical shifts can also be exported to the atom list if the group labels match with the protein amino acid sequence.

|     |                     |                   | HN                 | N                     |
|-----|---------------------|-------------------|--------------------|-----------------------|
| S2  |                     | 86 (2) [1.543]    | -                  | -                     |
| S3  | 45.539 (4) [11.387] | 3.076 (2) [1.538] | 6.991 (5) [1.398]  | 97.299 (5) [19.460]   |
| V4  | 52.522 (8) [6.569]  | 3.560 (4) [0.890] | 7.809 (25) [0.312] | 115.029 (25) [4.602]  |
| D5  | 48.670 (8) [6.085]  | 3.496 (4) [0.874] | 8.012 (25) [0.320] | 123.905 (25) [4.956]  |
| A6  | 43.331 (4) [10.833] | -                 | 7.502 (11) [0.682] | 112.761 (11) [10.251] |
| N7  | 46.263 (8) [5.784]  | 4.103 (4) [1.026] | 8.062 (21) [0.384] | 106.601 (21) [5.077]  |
| K8  | 48.401 (8) [6.051]  | 4.364 (4) [1.091] | 7.320 (19) [0.385] | 111.909 (19) [5.890]  |
| V9  | 52.539 (8) [6.571]  | 3.708 (4) [0.927] | 7.967 (29) [0.275] | 107.872 (29) [3.720]  |
| K10 | 47.798 (8) [5.976]  | 4.015 (4) [1.004] | 8.343 (27) [0.309] | 120.717 (27) [4.471]  |
| F11 | 49.971 (8) [6.247]  | 4.024 (4) [1.006] | 8.255 (13) [0.635] | 115.394 (13) [8.877]  |
| F12 | 49.157 (8) [6.145]  | 4.479 (4) [1.120] | 7.752 (15) [0.517] | 107.443 (15) [7.163]  |
| F13 | 44.170 (4) [11.043] | 3.737 (2) [1.869] | 8.362 (13) [0.643] | 109.860 (13) [8.451]  |
| G14 | 38.156 (8) [4.772]  | 3.373 (4) [0.928] | 7.761 (17) [0.457] | 100.235 (17) [5.898]  |
| K15 | 50.142 (8) [6.272]  | 3.342 (4) [0.836] | 8.129 (21) [0.387] | 113.143 (21) [5.388]  |
| N16 | 47.409 (8) [5.927]  | 3.616 (4) [0.904] | 9.556 (17) [0.562] | 107.811 (17) [6.343]  |
| C17 | 51.096 (8) [6.391]  | 1.886 (4) [0.471] | 7.749 (13) [0.596] | 101.135 (13) [7.780]  |
| T18 | 52.150 (8) [6.522]  | 3.668 (4) [0.917] | 5.564 (13) [0.428] | 99.090 (13) [7.623]   |
| G19 | 38.538 (8) [4.819]  | 3.344 (4) [0.878] | 7.740 (15) [0.516] | 133.920 (15) [28.386] |
| E20 | 51.083 (8) [6.389]  | 3.180 (4) [0.795] | 8.125 (17) [0.478] | 112.981 (17) [6.646]  |
| S21 | 48.434 (8) [6.055]  | 4.283 (4) [1.071] | 7.105 (19) [0.374] | 107.309 (19) [5.648]  |
| F22 | 48.694 (6) [8.118]  | 3.519 (4) [0.880] | 7.904 (17) [0.465] | 111.386 (17) [6.552]  |
| E23 | 47.827 (8) [5.979]  | 4.450 (4) [1.112] | 7.781 (17) [0.458] | 114.208 (17) [6.718]  |
| V24 | 50.354 (8) [6.298]  | 3.845 (4) [0.961] | 8.254 (19) [0.434] | 112.794 (19) [5.937]  |

**Figure 6.10 Assignment Summarized Table.**

The table summarizes the chemical shifts (p.p.m), assignment counts (in parentheses) and the standard deviations (in square brackets) of each assigned atoms (columns) in different residues (rows).





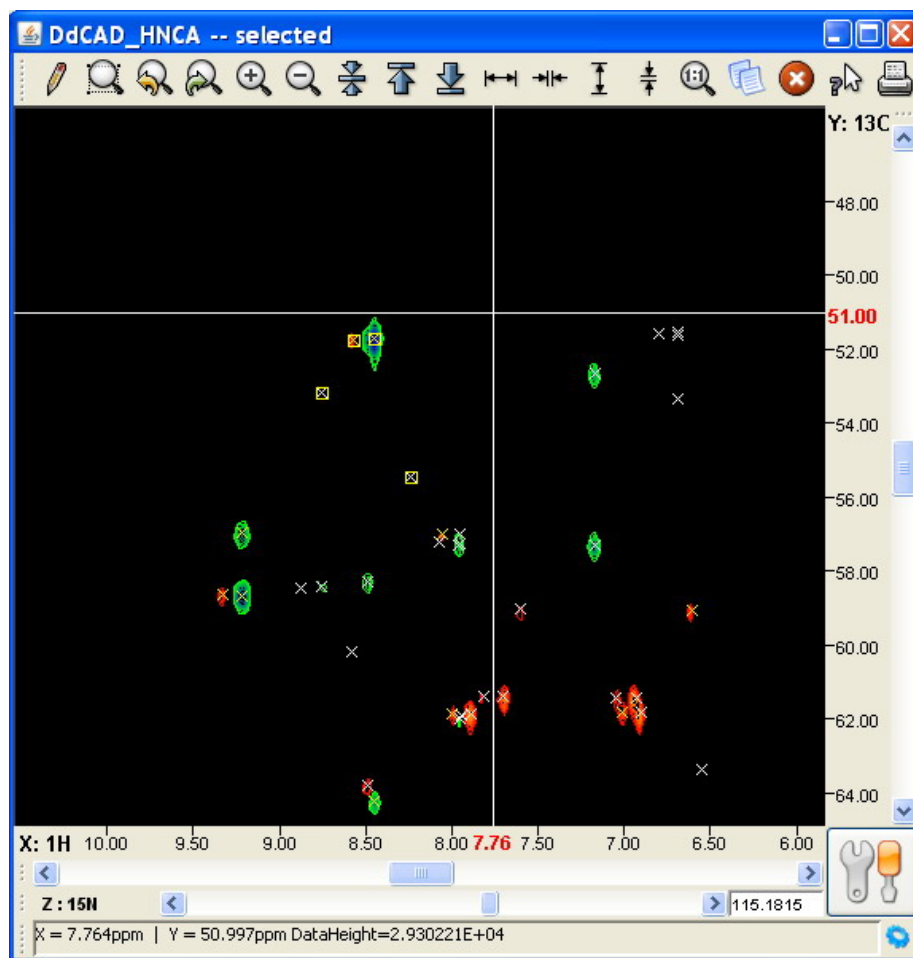
**Figure 6.11 NOE Calibration Panel.**

- **NOEs**

Select this to bring up a panel (Figure 6.11) for creating and editing NOE restraint lists.

### 6.3.1.5 Extensions menu

NMRspy currently has only one extension: “XYZ4D” (Chapter 7) for automatic/semi-automatic backbone assignment using the new strategy developed by our lab (Chapter 4).



**Figure 6.12 Spectral View (Spectral Display Window)**

### 6.3.2 Spectral display windows

Spectral display windows, also called spectral views (Figure 6.12), are the windows in which contour plots of planes of spectra can be displayed.

NMRspy has few limits on the user's ability to visualize NMR spectra. As with the number of datasets that can be opened, there are no limits to the number of spectral views that can be opened at the same time. Likewise, they can be displayed with arbitrary locations and sizes on the desktop.

The spectral view is highly flexible and plastic, almost all colors used in drawing the spectrum (contours, labels, crosshairs, background etc.) can be customized by users. The size and height/width ratio of the spectrum can also be adjusted. Users could even drag a widget out of the spectral view window by pressing and holding the left mouse button with the pointer at the left or top margin of the widget, then keeping the button down and dragging the widget to any arbitrary location.

More spectral views can be displayed in “Simple Layout” as mentioned in Section 6.2.5. While only two-dimensional displays of spectra are possible, the choice of dataset dimensions corresponding to displaying axes is up to the user. Different views, including variations on parameters such as dataset dimensions, plot regions, and contour levels of the same spectrum, can be displayed simultaneously in different spectral views.

### 6.3.2.1 Spectrum control bar

A Control Bar is present across the top of the standard spectral view (Figure 6.12). The icons provide easy access to commands to adjust the spectral view and levels, as well as to print spectra and stop contour drawing that is in progress in a given window.

**Table 6.1 Icons in control bar**



Draw

Draw the spectrum using all currently selected parameters. Use this to refresh the spectrum after changing a parameter which does not result in automatic

redrawing of the spectrum. Keyboard Shortcut: “ D ”



Full

Set the display region of the spectrum to their full extents and draw the spectrum. For 3D and higher dimensional spectra only the dimensions on the x and y axes of the display are set to the full values. Keyboard Shortcut: “ F ”



Previous View

Set the display region to the region displayed before the last spectrum control command issued. Keyboard Shortcut: “ PgUp ”



Next View

Set the display region to the region displayed before the Previous command issued. Keyboard Shortcut: “ PgDn ”



Zoom In

Zoom the display into a region around the center of the currently displayed region. A smaller portion of the spectrum will be displayed, and the displayed peaks will look larger. Keyboard Shortcut: “ + ”



Zoom Out

Zoom the display out from the center of the currently displayed region. A larger portion of the spectrum will be displayed, and the displayed peaks will look smaller. Keyboard Shortcut: “ - ”



Auto Level

Automatically calculate and set the display level to a "reasonable" value. Keyboard Shortcut: “ ? ”



Level Up







Raise the contour threshold of spectrum. Generally, fewer peaks will be displayed, and their displayed footprint will be smaller. Keyboard Shortcut: “ < ”






Level Down

Lower the contour threshold of spectrum. Generally, more peaks will be displayed, and their displayed

footprint will be larger. Keyboard Shortcut: “ > ”

-  **Broaden**      Broaden the display from the center of the currently displayed region. The display region will cover less chemical shift range of the X dimension. Keyboard Shortcut: “ [ ”
-  **Narrow**      Narrow the display from the center of the currently displayed region. The display region will cover more chemical shift range of the X dimension. Keyboard Shortcut: “ ] ”
-  **Heighten**      Enlarge the display from the center of the currently displayed region along the Y-axis. The display region will cover less chemical shift range of the Y dimension. Keyboard Shortcut: “ ; ”
-  **Lower**      Lower the display from the center of the currently displayed region. The display region will cover more chemical shift range of the Y dimension. Keyboard Shortcut: “ ‘ ”
-  **Expand In Proportion**      Expand the spectrum in proportion to the resolution of x and y axes. The height / width ratio of the spectrum will be set as the data point ratio of x and y axes, the longer axes will be set to its full extent in the window. This command affects only the dimensions displayed on the x and y axes of the plot. Keyboard Shortcut: “ \ ”
-  **Duplicate**      Duplicate the spectral view. Create another spectral view using exactly the same parameters with a different title. The new spectral will be displayed at the right side of the original one. Keyboard Shortcut: “ Ins. ”

-  **Stop** Stop drawing the current spectrum. As each spectrum is drawing itself it periodically checks for a "stop" flag. Clicking this button sets the "stop" flag. There may be a short delay between clicking the button and the time at which the spectrum display stops. Keyboard Shortcut: " Break ”
-  **Information** Display some tips and hints for controlling or customizing the spectral view.
-  **Print** Open a dialog for printing the spectrum (Figure 6.13). Use this dialog to set such things as the spectrum title, resolution, labels, grids, tick marks, page size & orientation, output device (including whether to send the output to a file rather than a printer) and previewing of the printing result.

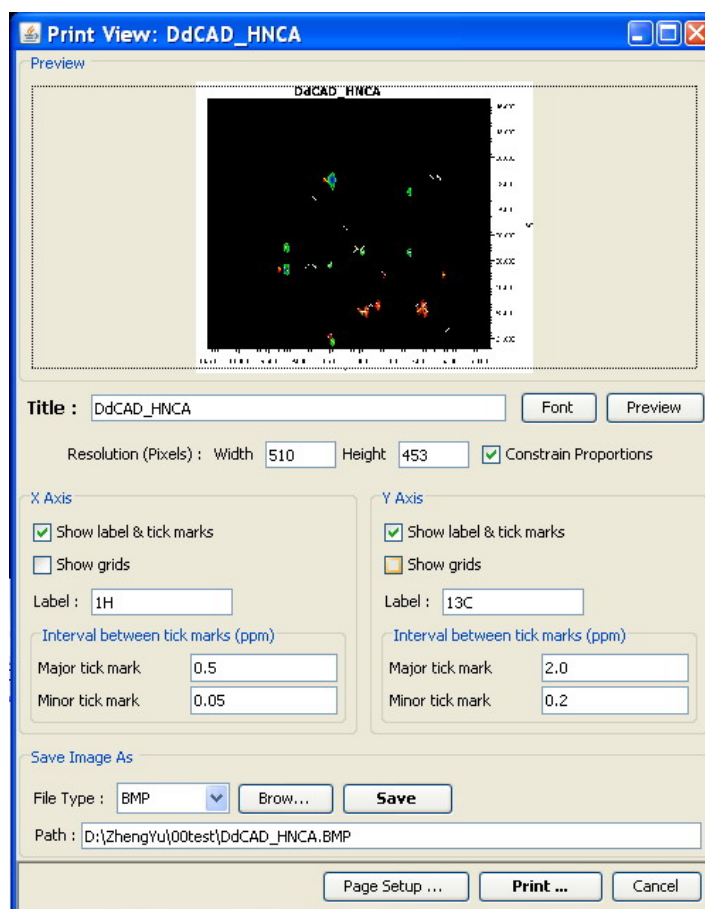


Figure 6.13 Spectrum Printing Dialog.

### 6.3.2.2 Mouse and keypad navigation

You can quickly navigate around a spectrum using buttons or the wheel on the mouse or keys on the auxiliary keypads of the keyboard. To use this feature the active window must have the "focus", that is, it must be the last window that you have clicked.

- **Mouse**

**Left click**                      Select or deselect displayed objects (peaks, labels, grids etc.). Click the left mouse button on an object will select it and deselect other objects. Click the left mouse button on an empty location will deselect any selected object.

**Ctrl + Left click**              Multiple select/deselect object. Hold the Ctrl key and click the left mouse button on an un-selected object will append it into the current selected objects list. Click on a selected object will remove it from the list.

**Right click**                      Bring up a popup-menu which could be used to add/edit/delete displayed object(s), control display region, and open various tables and so on.

**Left drag**                      Select multiple objects in a displayed region. Press and hold the left mouse button with the pointer at the up-left corner of the region. Keep the button down as you drag the

cursor to the down-right corner of the region. A box will appear and follow the cursor. All objects enclosed in the box will be selected after the left mouse button being released.

Move selected object(s). Press and hold the left mouse button with the pointer on a selected object. Keep the button down as you drag the cursor to a new position. The selected object(s) will be moved to the new position.

**Middle drag**                      Expand the display region. Press and hold the middle mouse button (or wheel) as you drag the cursor to draw a box. The display region corresponds to the area currently enclosed in the box will be expand after the middle button being released.

**Wheel up/down**                      Scroll up/down the spectrum.

**Shift + Wheel up/down**                      Scroll left/right the spectrum.

- **Numeric Keypad**

- |   |                            |
|---|----------------------------|
| 1 | Pan down and to the left.  |
| 2 | Pan down.                  |
| 3 | Pan down and to the right. |
| 4 | Pan left.                  |
| 6 | Pan right.                 |



- |   |                          |
|---|--------------------------|
| 7 | Pan up and to the left.  |
| 8 | Pan up.                  |
| 9 | Pan up and to the right. |

- **Cursor Keys**

- |             |                                       |
|-------------|---------------------------------------|
| Down Arrow  | 3D, 4D Spectrum: Move Down a Z plane. |
| Up Arrow    | 3D, 4D Spectrum: Move Up a Z plane.   |
| Left Arrow  | 4D Spectrum: Move Down a A plane.     |
| Right Arrow | 4D Spectrum: Move Up a A plane.       |

- **Other Keys**

- |       |   |
|-------|---|
| Space | Open “Spectral Attribute Window”. (Section 6.3.3)                         |
| A     | Open “Peak Identification Dialog”. (Section 6.3.5.4)                      |
| Enter | Edit the currently selected object(s).                                    |
| Del   | Delete the currently selected object(s).                                  |
| Esc   | Switch between “Standard Layout” and “Simple Layout” (see Section 6.2.5). |

### 6.3.2.3 Status bar

NMRspy provides a very useful status bar (Figure 6.14a). An option dialog (Figure 6.14b) could be brought up by clicking the blue gear icon shown in Fig. 6.13a. It allows users to customize the status bar so that it can display useful information at their convenience.

When the cursor moves into a spectrum, the current position of the crosshair and its folded position(s) (add or subtract up to 5 integer multiple of the

sweep widths) could be displayed in ppm in status bar with its data height. The crosshair position could also be displayed in the unit of pix or/and data point.

When the cursor moves out of a spectrum, spectral information like dimension labels, sweep width, resolutions, data point counts, file name and data set title could be displayed in status bar.

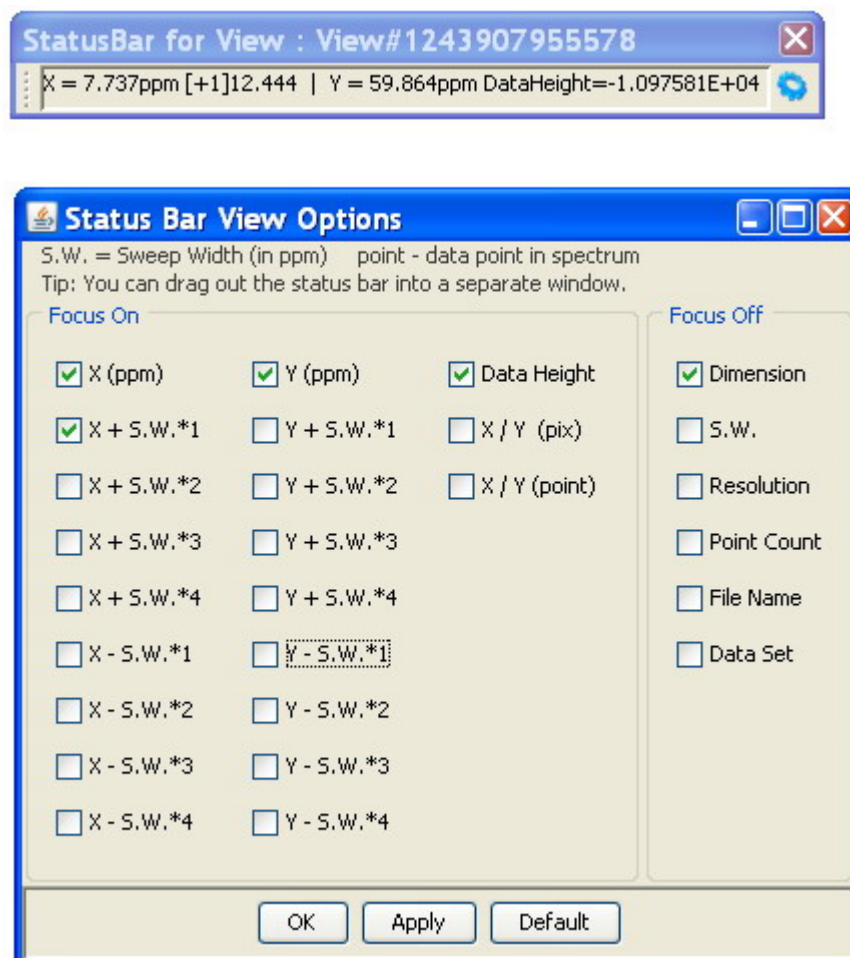


Figure 6.14 Status Bar Setting Dialog.

### 6.3.3 Spectral attribute windows

This window collects a wide variety of controls for interacting with spectral view. Only one view is controlled at a time through this window. The title bar of the window will indicate the name of the view whose attributes are being set. This window is selected when clicking the "Attribute Button" at the

down-right corner of a spectral view, or pressing space key when a spectral view is active.

The Attribute Window is composed of a tabbed panel which forms the majority of the window and four buttons across the bottom. The tabbed panel allows the user to select from a series of panel controlling different aspects of the spectral view.

The various tapped panels that comprise the Attribute Window are described in the following sections.

### **6.3.3.1 File panel**

The File Panel (Figure 6.15) is used to assign individual spectrum to a spectral view, and display/control its references.

NMRspy uses a conservative approach in changing the reference information of a data file. Users could never modify the reference of a non-NMR-format data file. When a NMR-format data file is created (Section 6.3.1.1), an original reference and a copy of it are generated. The original reference is used as a backup and will never be modified; instead, changes to the reference are made persistent by modifying the copy or the so-called active one. A dialog panel (Figure 6.16) is used for adjusting all the reference information and then for modifying the active reference. Both references are saved in the spectral data file.

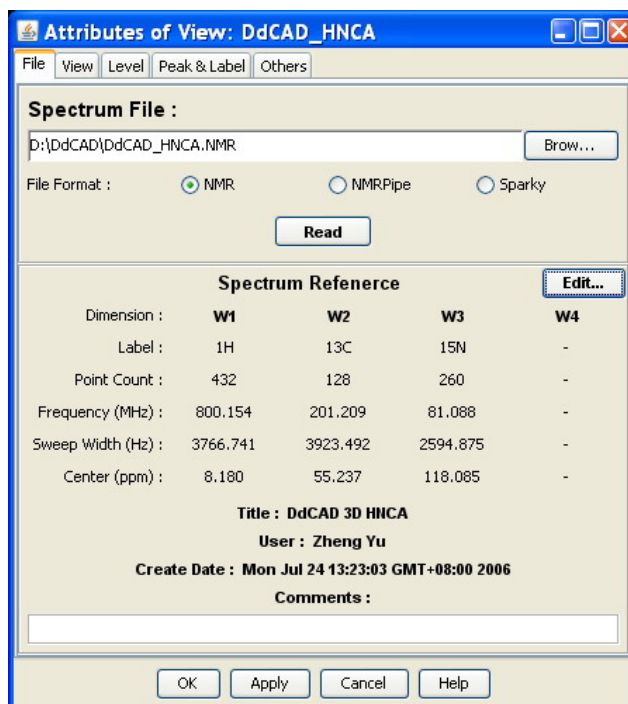


Figure 6.15 Spectrum File Setting Panel.

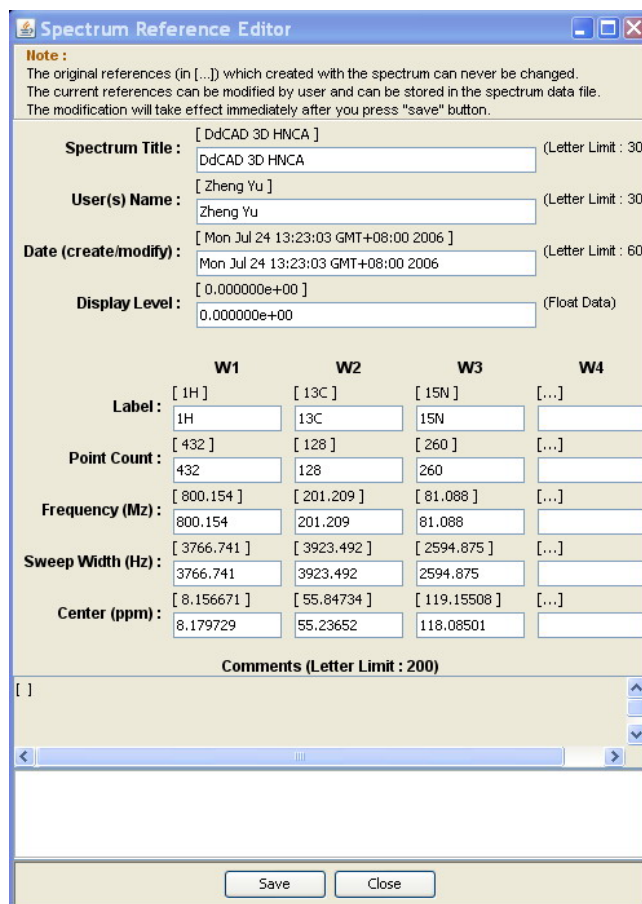


Figure 6.16 Spectrum Reference Editor.

### 6.3.3.2 View panel

The amount of information in an NMR spectrum is generally so great that it is often only informative to look at a portion of the spectrum at one time. In NMRspy, sub-regions may be selected by limiting the region on the x or y axis of the display, and/or choosing a sub-selection of planes in a spectrum with three or more dimensions.

The View Panel (Figure 6.17) provides controls to interactively select the display regions. Press the Left Mouse button over the axis mark (X, Y, Z and A) to pop-up a menu of predefined plot limits, or enter values in the next two text fields to set the plot limits (in ppm). For 3D and 4D spectra, if a range of values for Z and/or A are specified, all planes between the two specified values (inclusive) are drawn with one plane overlaid with another. The file dimensions and the display dimensions can be chosen in any desired manner. The pull down choice box in the third field is used to specify which dataset dimension is specified on the particular axis.

The View Panel also provides controls to overlay spectra. User can overlay the contours of one spectrum on another after opening them in different spectral views. By repeating this process they can overlay as many spectra as they want on one view and control the contour levels, colors and order of each overlaid view separately. The contour levels of all the overlaid spectra could also be changed together. Overlaying spectra is very useful in comparing peaks of spectra collected under different conditions, such as comparing spectra of proteins collected in the presence of different ligands or comparing wild type and mutant forms of a protein.

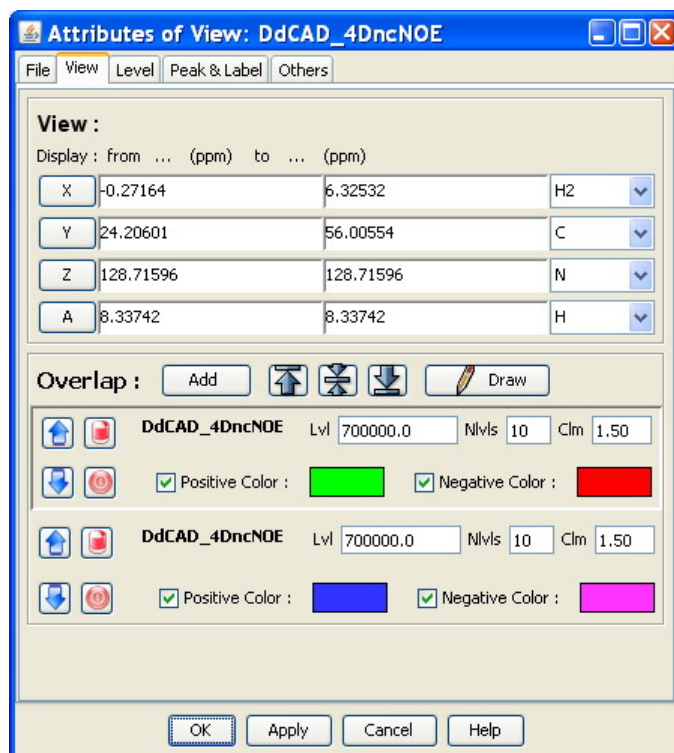


Figure 6.17 Spectral View Setting Panel.

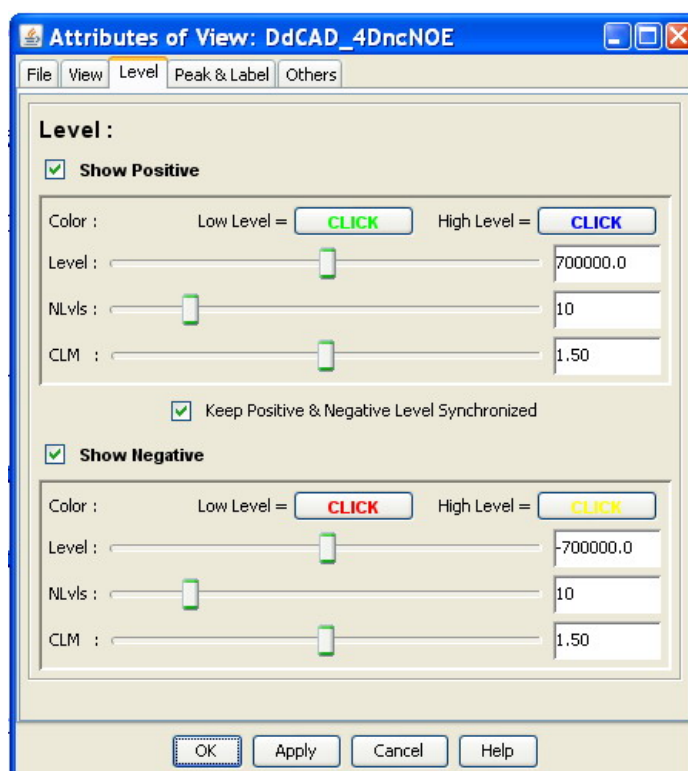


Figure 6.18 Spectral Level Setting Panel.

### 6.3.3.3 Level panel

An NMR spectrum may have a tremendous dynamic range, with some peaks that are orders of magnitude more intense than others. Despite the presence of peaks with great intensity, some of the most important peaks in the spectrum may have intensities only slightly higher than that of the noise. While the NMRspy Auto-Level tool in the Spectral Control Bar can automatically set a level that gives an aesthetically pleasing plot, manual selection of the intensity scale is often required to visually perceive all the information in the spectrum.

The Levels Panel (Figure 6.18) has controls for adjusting the scaling of the NMR spectrum intensity, the levels used for calculating contours, and colors of drawn contours. It is divided into two sub-regions, one for the positive contours and the other for the negative contours of spectrum.

The check-buttons control whether or not the contours will be displayed and synchronized. Select the positive check-button, and deselect the negative check-button to see only positive contours. Select both check-buttons to see both. Select the synchronization check-button to synchronize the level and scaling properties of positive and negative contours.

The color buttons control the color of a drawn spectrum. Clicking the buttons will bring up a Color Selection Dialog. Some of the color choices such as red-blue, red-yellow and green-blue color the contour lines in a range of colors according to their height. In the green -blue scheme the lowest contour level is green and the highest level is blue with intermediate levels having intermediate colors. This can help users see peaks when looking at highly overlapped regions

that are a mass of contours, or when they wish to have low contour levels that show a lot of noise.

The next region of the panel provides three sliders and text boxes to specify the contour level, the ratio between subsequent contours, and the number of contour levels to draw.

**Level** Specify the level at which contours are to be drawn. The slider provides a convenient means to increment or decrement the contour level.

**NLvs** Specify the maximum number of contours to be drawn.

**CLM** Specify the ratio between subsequent contour levels. For example, a value of 1.5 means that each contour will be 1.5 times as high as the previous one.

#### 6.3.3.4 Peak & label panel

While it is possible to analyze NMR data by directly decomposing the raw data into lists of parameters such as frequencies and linewidths, most users still rely on Fourier Transformation of the FID and then identifying peak positions in the transformed data. Accordingly, NMRspy provides many tools for locating and analyzing these spectral peaks. The Peak & Label Panel (Figure 6.19) has controls for generating and displaying peaks in a spectrum.

Peak, label and grid are the only visible objects that could be drawn on the contours of a spectrum. All of them are stored in a dataset. Creating or selecting a dataset of a spectrum is the standard starting point for analyzing it.



To pick peaks in a particular region of a spectrum which is chosen in the spectral view, select the “Current Window” attribute in the region combo box. To pick peaks in the entire spectrum, select the “Whole Spectrum” attribute instead. The peaks that are picked may be appended to the peak list in the dataset, or replace the existing peak list. Click on the Pick button to start the automatic peakpicking or click on the Clear button to clear peaks in the specified region.

Peaks are displayed on the contour plot as small crosses with or without a string labels. The size, color, type and font of the label could be specified by users. With 3D and 4D spectra, only peaks whose Z and/or A dimensions are within a specified number of planes from the currently displayed plane will appear. The range within which planes are displayed is specified with the "Visible Depth" parameter. Peaks whose Z and/or A dimensions are closest to the displayed plane are displayed with the “On Color”. Those peaks that are off the displayed plane but within the specified range appear with the “Off Color”.

The Peak & Label Panel also contains buttons that can access “Peak Auto-assign Dialog” (Section 6.3.4.3) and “Peak (label, grid) Table” (Section 6.3.4.2).

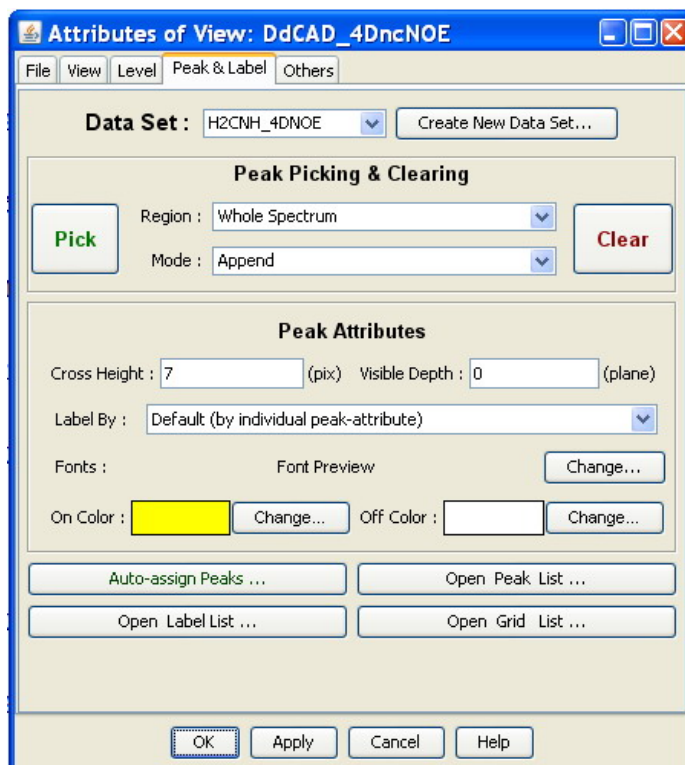


Figure 6.19 Peak & Label Setting Panel.

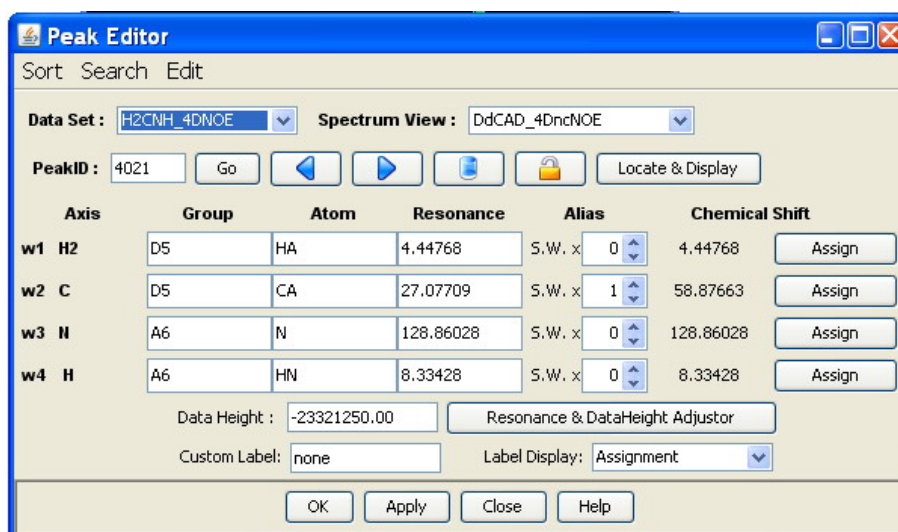
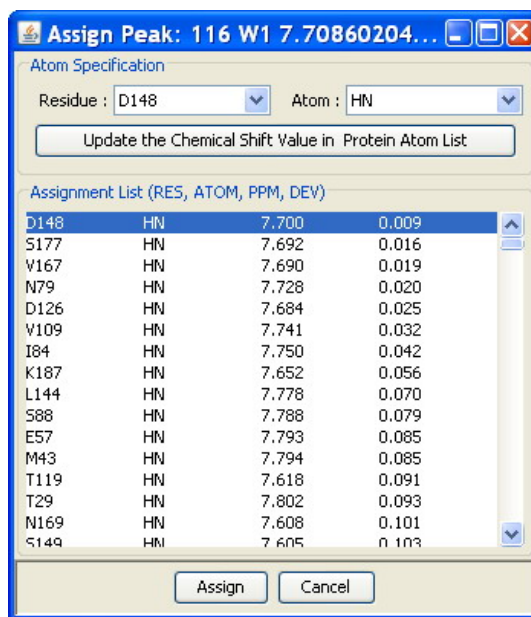


Figure 6.20 Peak Editor Dialog.



**Figure 6.21 Peak Assignment Dialog.**

### 6.3.4 Other dialogs & windows

Besides the major display and configuration windows mentioned above, users may discover many other useful interfaces in NMRspy. We'll brief some important ones.

#### 6.3.4.1 Peak (label, grid) editor

NMRspy provides a very powerful and relatively rapid method for interactively examining peaks. The Peak Editor (Figure 6.20) allows users to rapidly navigate through peak lists, examine, modify or assign peaks with well designed interface.

The Peak Editor is able to sort the whole peak list by various orders (e.g. peak ID, assignment, label, intensity, chemical shift) and navigate through these peaks one by one. It can also search a certain peak that has specified properties (first fully-assigned, next unlock, first deleted etc.). With Peak Editor, Users can send peaks to recycle bin, lock its attributes to avoid farther modification, alias

its chemical shift to the “true position”, adjust its resonance and data-height or customize its display label.

If a spectral view has been selected in the combo box, when clicking the “Locate & Display” button, the display will center on the peak position. If an atom list has been fully or partially assigned, when clicking the “Assign” button at the right side of every dimension’s properties, a peak assignment dialog (Figure 6.21) will pop up and assist users in peak assignment task.

Farther more, the Resonance & DataHeight Adjustor (Figure 6.2) provided by NMRspy allows users to pinpoint the resonance and intensity of a peak in a multi-dimensional spectrum. (Section 6.2.3) Drag the vertical central line (blue line) in every dimension, and the parallel intensity line will move along the fitting curve and extract the accurate data height.

NMRspy also provides Label Editor and Grid Editor. They are similar to Peak Editor but much simpler.

#### **6.3.4.2 Peak (label, grid) table**

A peak Table (Figure 6.22) could be accessed from the Peak & Label Panel (Section 6.3.3.4) or from the right click menu (Section 6.3.2.2). It displays a table of peaks for a specified dataset. Each line could show the properties of a peak (e.g. peak ID, assignment, custom label, data height, chemical shifts, status), and the columns can be displayed or hidden by users.

The peak list can be sorted by every property, and can be saved to a text file. Users can edit or delete an individual peak as well as compress (permanently remove all the peaks that have been sent to recycle bin. Peak State = -1) and/or degap (remove all the gaps in the peak ID numbers) the whole list.

| ID | Assignment | Custom Label | Data Height | H      | N       | Status |
|----|------------|--------------|-------------|--------|---------|--------|
| 1  | ?          | none         | 3775481     | 8.835  | 133.429 | 0      |
| 2  | ?          | none         | 3031795     | 10.006 | 133.355 | 0      |
| 3  | ?          | none         | 6156916     | 8.579  | 132.200 | 0      |
| 4  | ?          | none         | 2446012     | 9.280  | 131.681 | 0      |
| 5  | ?          | none         | 3516011     | 8.698  | 131.088 | 0      |
| 6  | ?          | none         | 2248918     | 8.659  | 130.298 | 0      |
| 7  | ?          | none         | 4766619     | 9.323  | 129.995 | 0      |
| 8  | ?          | none         | 3631917     | 8.926  | 130.020 | 0      |
| 9  | ?          | none         | 3647641     | 9.057  | 129.729 | 0      |
| 10 | ?          | none         | 3700461     | 8.681  | 129.550 | 0      |
| 11 | ?          | none         | 3799574     | 8.197  | 129.439 | 0      |
| 12 | ?          | none         | 1596687     | 8.822  | 129.223 | 0      |
| 13 | ?          | none         | 8320433     | 8.332  | 128.858 | 0      |
| 14 | ?          | none         | 5984429     | 8.409  | 128.018 | 0      |
| 15 | ?          | none         | 6597531     | 8.918  | 127.957 | 0      |
| 16 | ?          | none         | 3302724     | 8.621  | 127.685 | 0      |
| 17 | ?          | none         | 4769349     | 9.068  | 127.493 | 0      |
| 18 | ?          | none         | 4631709     | 9.016  | 127.357 | 0      |
| 19 | ?          | none         | 4140889     | 9.417  | 127.197 | 0      |
| 20 | ?          | none         | 3624648     | 9.499  | 127.024 | 0      |
| 21 | ?          | none         | 3546019     | 9.227  | 127.005 | 0      |
| 22 | ?          | none         | 6746920     | 8.202  | 126.826 | 0      |
| 23 | ?          | none         | 3538471     | 8.756  | 126.777 | 0      |
| 24 | ?          | none         | 3333562     | 9.105  | 126.678 | 0      |
| 25 | ?          | none         | 2952887     | 9.045  | 126.425 | 0      |
| 26 | ?          | none         | 2325816     | 8.708  | 126.122 | 0      |
| 27 | ?          | none         | 5121906     | 8.196  | 125.986 | 0      |
| 28 | ?          | none         | 855626      | 8.116  | 125.807 | 0      |

Figure 6.22 Peak Table.

Auto-Assign dialog box showing settings for two windows, W1 and W2. W1 is set to 'W1: H' with a tolerance of 0.023 ppm. W2 is set to 'W2: N' with a tolerance of 0.185 ppm. Both windows have 'All Residue' and 'All Amino Acid' checked. The dialog also includes checkboxes for Backbone Atoms and Side-Chain Atoms.

Figure 6.23 Peak Auto-assign Dialog.

### 6.3.4.3 Peak auto-assign dialog

Analysis of the peaks in NMR spectra is obviously much more useful if one has the assignment - which atoms in a molecule (or molecules) gives rise to which peaks in the spectrum. Peak Auto-assign Dialog (Figure 6.23) assists users to assign peaks automatically. The peak-list's pattern and tolerance parameters are used when a search is done to find atoms whose assigned chemical shifts are

consistent with those of a given peak. The atom's type, residue ID and amino acid type could be specified by users. If two dimensions are in different groups, their group assignment should not be the same.

#### 6.3.4.4 Peak identification dialog

The Peak Identification Dialog (Figure 6.24) can be brought up from the right click menu or keyboard shortcut "A". It can be used to interactively assign atom identifiers to individual peaks for a molecule with chemical shift assignments. The atoms listed in this dialog are those atoms whose chemical shifts are within the peak-list's tolerance of chemical shift values.

NMRspy users should themselves synthesize the information from the chemical shift deviations, hydrogen pair distances, and spectral display to reach a conclusion about which, if any, of the atom entries are the correct assignment for the peak. Selecting the entry and then clicking the "OK" button will update the assignment labels for the current peak with the names of the atoms in that entry.

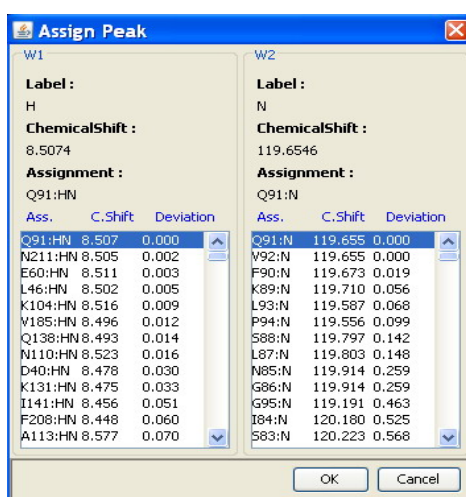


Figure 6.24 Peak Identification Dialog.

## 6.4 Results and discussion

After testing by dozens of users and constantly updating by the developer, NMRspy has become a fairly stable and useful platform for NMR spectral visualization, analysis and management. The features that are specially designed for multi-dimensional spectra, folded peaks and complex tasks have been given remarkable credits by users.

NMRspy tries to provide users with more interactive and self-explanatory user interfaces. The user interfaces are depicted in a way that resembles the general operation system. There is no need to remember any command, because all functions and operations could be accessed by graphic interfaces and most command have more than one access methods (button, menu, shortcut etc.). Users familiar with other spectral visualization software packages can start with NMRspy immediately. Users with basic computer skills but no experience in NMR spectral analysis can handle NMRspy in a short period of time. Once getting familiar with the operation of NMRspy (especially the keyboard shortcuts), a user may work 50% to 200% faster than with other softwares.

With the numerous novel features and functions, NMRspy can perform many operations that are extremely difficult for other software packages. Some operations that may require users to prepare their own script in some well-used software packages can be easily achieved by NMRspy using graphical interface. While some senior spectroscopists may prefer their own script, for most users, especially those who are not familiar with the computer programming, NMRspy's graphical interface could bring them a great release.

However, given the time and practical limitations, NMRspy is far from perfect or complete. Its development and update have never been ceased. At the same time, development of its extensions is still under progress. We have already developed a very useful extension package (Chapter 7).



## Chapter 7:

# XYZ4D: software plug-in for backbone assignment using the new strategy

- 7.1 Introduction
- 7.2 Interface and algorithms
- 7.3 Results and discussion

## Chapter 7:

# XYZ4D: software plug-in for backbone assignment using the new NOESY-based strategy

## 7.1 Introduction

The new strategy introduced in chapter 4 has proven to be an effective and reliable method for backbone assignments of large proteins without the use of deuterium and specific labeling. The manual process, however, is tedious and time-consuming and may require weeks or even months of dedicated work. To overcome the limitation, a software tool has been developed as an NMRspy (Chapter 6) plug-in to facilitate the backbone assignment using this strategy. This chapter elaborates on the implementation of the plug-in, XYZ4D (Xu Yingqi, Yang Daiwen & Zheng Yu's novel strategy for solution structure determination of large proteins without deuteration using 4D NOESY and other 3D NMR spectra.).

Our initial plan was to develop a program that would be capable of managing a set of spectra (2D-TROSY-HSQC, 3D-TROSY-HNCA, 3D-TROSY-HN(CO)CA, 3D-MQ-CCH-TOCSY and 4D-<sup>13</sup>C,<sup>15</sup>N-edited NOESY) used in the strategy, performing calibration and peak picking in them, grouping HC-NH NOE and C<sub>α</sub>-NH (HNCA,HN(CO)CA) correlations that have identical NH chemical shifts into a cluster, identifying spin-systems by separating out intra-residue and sequential HC-NH NOE correlations from other inter-residue NOEs observed in the 4D-NOESY spectrum with the use of HNCA and CCH-TOCSY spectra, then establishing fragments from clusters by matching the

intra-residue spin-system of one cluster with the sequential spin-system of another cluster, and finally mapping the fragments onto the protein sequence. It's expected that the backbone assignments could be obtained with minimal human intervention and a score would be computed for each cluster mapping as an indication of its reliability. The program would also provide graphical interfaces for users to manually check through all the tasks and make corrections should any error occur.

Although a fully automated program with such functionality may appear as a tempting solution to the large protein backbone assignment problem, further analysis has raised the following concerns.

- When searching for NOE peaks in the construction of clusters, the program needs a robust method to filter out the background noise. Since many NOEs, especially those inter-residue NOEs between amide protons and aliphatic protons at the distal end of side-chains, may be weak due to the usually longer distances, it will be hard to find a perfect balance.
- Unlike 3D HNCA, the MQ-CCH-TOCSY spectrum in general contains a lot more noises and the peak pattern is less distinct. As a result, comparison of strip plots cannot be simply based upon the matches of peak positions. Even with a sophisticated algorithm for pattern recognition, the automated spin-system identification result can still be quite futile.
- The scoring scheme for assessing the un-reliability cluster mapping should be sensitive enough to pick up those that can potentially go wrong,

but not so sensitive as to include many correct mappings. Otherwise time saved by automation would be wasted in manual checking. This is again the problem of finding a perfect balance.

- To ensure the efficiency and reliability of a fully automated backbone assignment program, it has to be tested under all conditions with as many data sets and protein samples as possible. This means that many suitable protein samples are required and many 2D, 3D and 4D NMR experiments need to be carried out to generate the test data.

Given the time and practical limitations, the abovementioned concerns have led to the conception of a semi-automated approach with a fully automated option. In such way, users could easily be involved in examining the peaks and resolving the ambiguous assignments or alternatively leave the program to do the whole job. As is well known, a little human intervention to an automatic program could usually lead to a far more satisfactory result, and users who favor the semi-automated approach would soon be aware of the well worthiness of their effort, as differentiation of real peaks from noises and comparing spectral pattern are much easier jobs for human than for computer. In addition, the efficiency and reliability of this program can be readily assessed with just a few NMR data sets.

XZY4D was written in the JAVA programming language as an extension (plug-in) of NMRspy. This not only ensures its maximal compatibility and acceptance with the platform software, but also allows us to tap on many useful functions provided by NMRspy. The whole program was divided into seven modules (Project Preparation, Spectra Calibration, Cluster Identification, CCH & 4DNOE Inspection, Spin-system Identification, Cluster Mapping, Backbone

Assignment), which correspond to the manual assignment steps of the new strategy.

## **7.2 Interface and algorithms**

XYZ4D's user interface consists of eight major components and more than 40 graphic interfaces (Figure 7.1). The following sections present detailed implementation of each of these components.

### **7.2.1 The main application window**

The main application window of XYZ4D is shown in Figure 7.2.

At the left side of the main window lies 7 buttons that can be used to access the 7 modules of XYZ4D. The right side of the main window is a note board which displays a short description of the corresponding module when the cursor moves on to a button. The buttons are arranged from top to bottom in accordance with the routine tasks of the novel backbone assignment strategy, and when clicked, the corresponding module will be activated. To achieve backbone assignment, the 7 modules need to be executed one by one in an appropriate order.

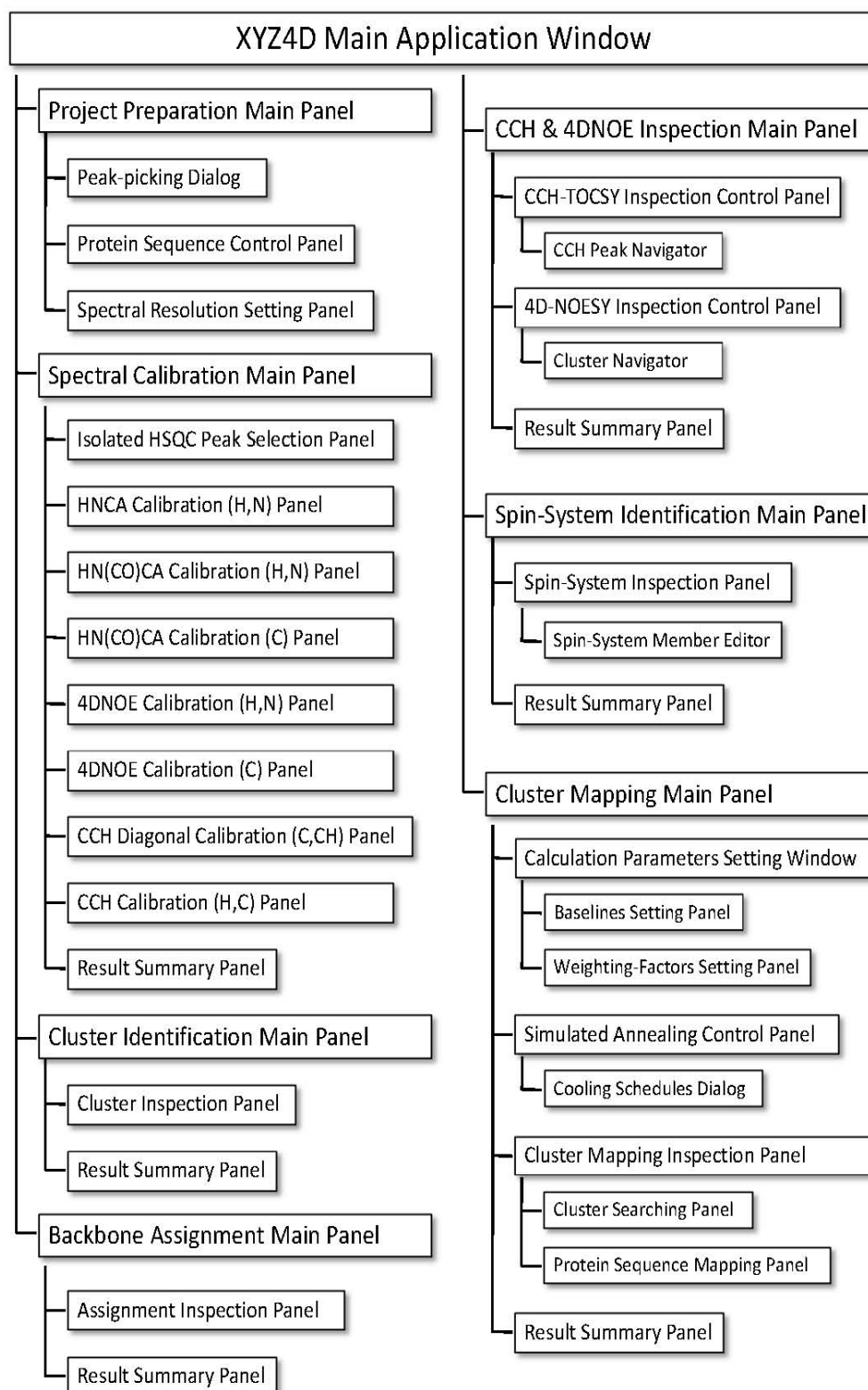
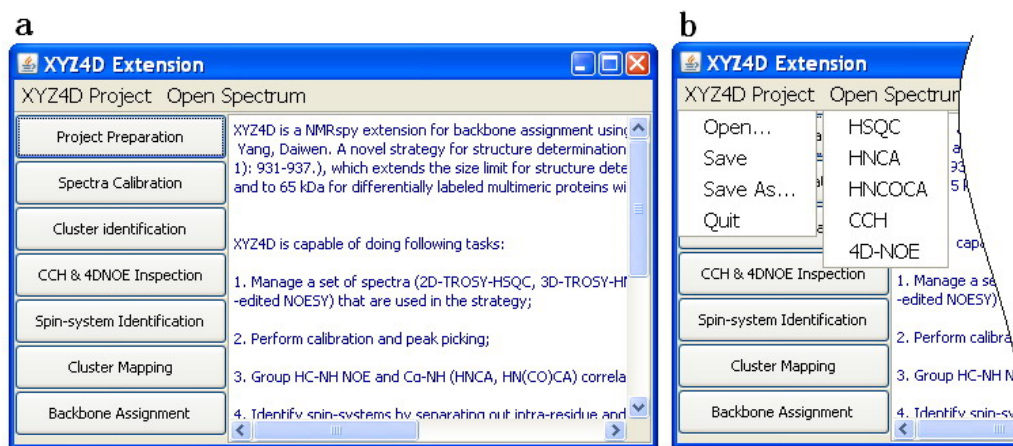


Figure 7.1 Overall Diagram of interfaces in XYZ4D.



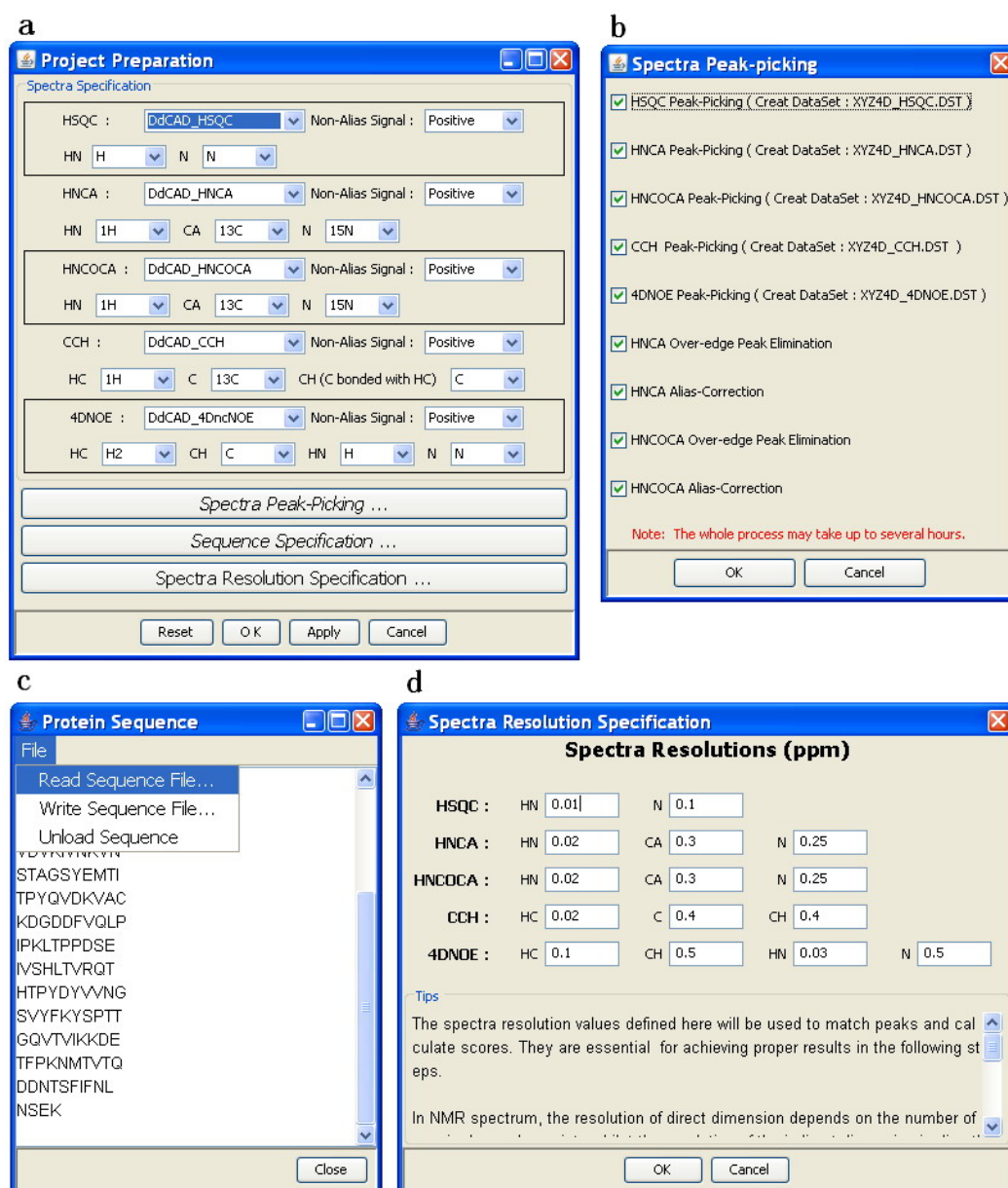
**Figure 7.2** Main application window of XYZ4D (a) and its pull-down menus (b).

On the top of the main window there are two pull-down menus (Figure 7.2 b). The "XYZ4D Project" menu could be used to open or save XYZ4D Project, which is an extension of NMRspy Project. The XYZ4D Project not only has all the features of an NMRspy Project, but also records the parameters, data and results generated by the XYZ4D program. The "Open Spectrum" menu is very useful for quick opening of one of the five spectra (HSQC, HNCA, HN(CO)CA, CCH-TOCSY, 4D-NOESY). With this feature, users could close those temporarily unused spectra to gain a clearer screen and more memory, and quickly reopen them when they are needed.

### 7.2.2 Project preparation module

Project Preparation Module is the first module. It allows XYZ4D to import protein primary sequence, understand the spectra used in the strategy and prepare the spectra for subsequent procedure.

The 5 spectra (2D-TROSY-HSQC, 3D-TROSY-HNCA, 3D-TROSY-HN(CO)CA, 3D-MQ-CCH-TOCSY and 4D- $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY) should be manually opened in NMRspy before users use the pull down menus on the main panel of Project Preparation Module (Figure 7.3 a) to specify the spectral name, dimension labels and the sign of non-alias signals for each spectrum.

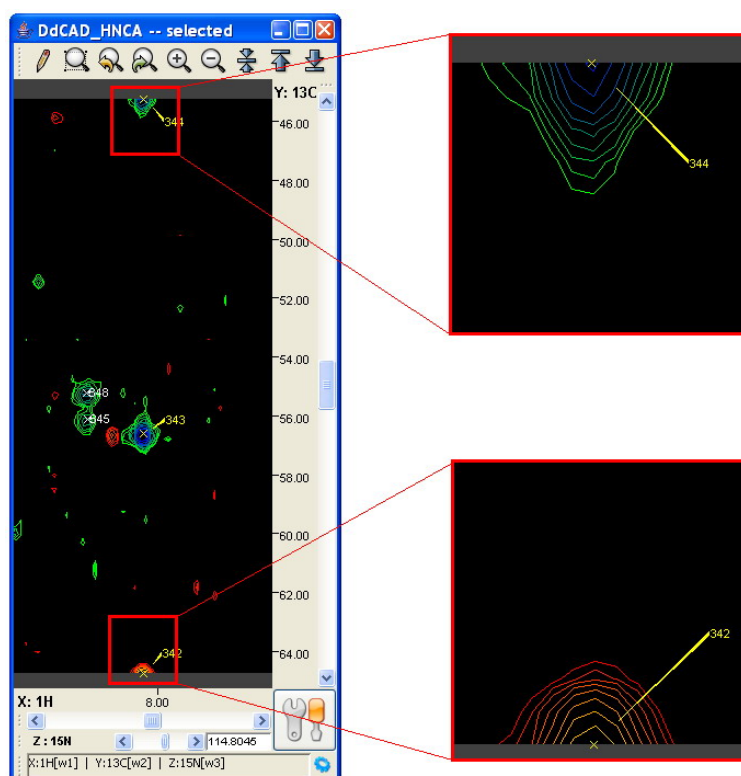


**Figure 7.3** Graphic Interfaces of Project Preparation Module.

(a) The main panel. (b) Peak-picking dialog. (c) Protein sequence control panel. (d) Spectral resolution setting panel.



After careful adjustment of the level for each spectrum, users may click "Peak-Picking" button to bring up a dialog (Figure 7.3 b) that can be used to automatically peak-pick the 5 spectra and correct the over-edge and folded peaks in HNCA and HN(CO)CA spectra. These tasks, especially 4DNOE peak-picking, may take a very long time. Users may execute them separately at their convenience. They can also redo certain tasks at any time when it's needed.



**Figure 7.4 Over-edge peak.**

A peak locates at the upper edge of  $^{13}\text{C}$  dimension on HNCA spectrum. Its upper half folds to the lower edge of the spectrum and appears as a negative peak.

The “Over-edge peak elimination” will automatically detect those peaks that locate within 5 points from the up/low edge of the HNCA or HN(CO)CA spectra. If two peaks have common  $^1\text{H}_\text{N}$ ,  $^{15}\text{N}$  chemical shifts and are located at

the upper and lower edges of the spectrum respectively, XYZ4D will eliminate the weaker one. (Figure 7.4) The "Alias-Correction" will automatically detect those folded peaks in HNCA and HN(CO)CA spectra, and will try to fold them into the normal  $C_{\alpha}$  chemical shift region, i.e. 40-70ppm.

If users skip or forget one of the peakpicking, over-edge peak elimination or Alias-Correction operation, it will be carried out automatically by XYZ4D when its result is required by the subsequence step, in order to avoid any unexpected error caused by human mistakes.

Users need to manually import the protein primary sequence (Figure 7.3 c) and define the resolutions of every dimension for each spectrum (Figure 7.3 d). The spectral resolution values will be used to match peaks and calculate scores in the later steps. They are essential for achieving proper results in the following steps.

After users pressing "OK", XYZ4D will save all the parameters, peak lists and protein sequence as a XYZ4D Project. Unless specifically stated otherwise, all the files generated during the subsequence procedure will be saved in a sub-folder which has the same name as the project and under the same directory. Therefore, when necessary, users could easily find the relevant files.

### 7.2.3 Spectral calibration module

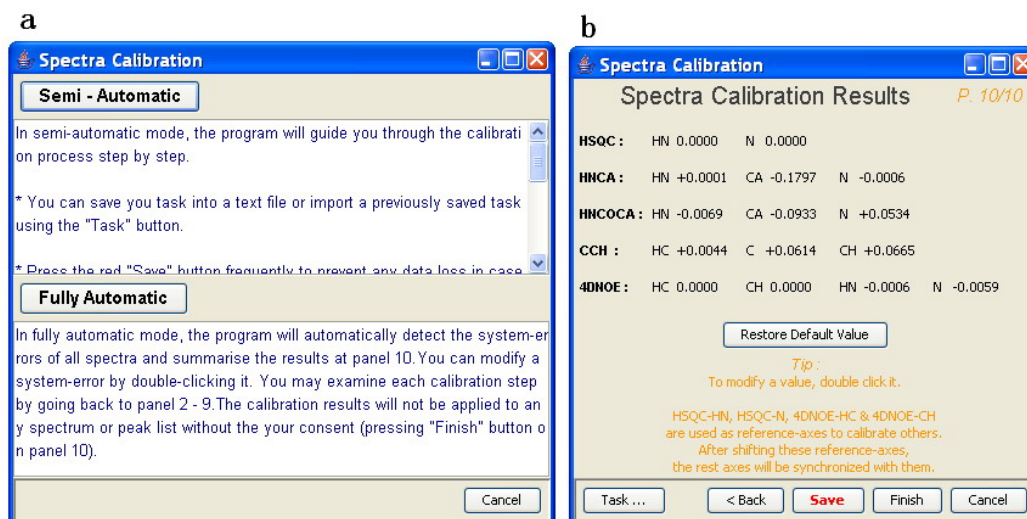
Although spectral calibration is not a mandatory step, it is strongly recommended to do so for any spectrum that is used in the backbone assignment

approach. Our extension tool will much more likely produce satisfactory results if the calibration could be as accurate as possible.

The Spectral Calibration Module uses  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shifts of isolated HSQC peaks as references to calibrate the  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  dimensions of HNCA, HN(CO)CA and 4D  $^{13}\text{C},^{15}\text{N}$ -edited NOESY spectra. It also uses  $^{13}\text{C}$  chemical shifts of strong 4D-NOESY peaks as references to calibrate the  $^{13}\text{C}$  dimensions of HNCA, HN(CO)CA and CCH-TOCSY spectra. The  $^1\text{H}$  dimension of CCH-TOCSY is calibrated from  $^1\text{H}_\text{C}$  chemical shifts of strong 4D-NOESY peaks. This module consists of 10 graphic interfaces.

### 7.2.3.1 Main panel

The main panel of the Spectral Calibration Module is shown in Figure 7.5 a.



**Figure 7.5 Main panel (a) and result summary panel (b) of the Spectral Calibration Module.**

Like all the subsequent modules, Spectral Calibration Module provides both semi-automatic and fully automatic options.

If users choose the semi-automatic mode, the module will guide them through the calibration process step by step using the interface described in section 7.2.3.2-7.2.3.9; if users choose fully automatic mode, the module will directly jump to the summary interface described in section 7.2.3.10, which shows only the final results (Figure 7.5 b). No matter which mode is selected, the calibration results will not be applied to any spectrum or peak list without the user's consent.

In the main panel, below either the "fully automatic" or "semi-automatic" button, a brief guide and some tips are provided to users for a better understanding of the operation and quick assistant.

During the Spectral Calibration procedure, users can save their task into a text file, or import a previously saved task. This text file can be opened or modified by any text editor, allows users to quickly check the relevant data, and prevents any data loss in case of a computer crash or freeze. Note that an unfinished task is not a part of the XYZ4D project, and saving a XYZ4D project will not save the currently working task.

### **7.2.3.2 Selection of isolated HSQC peaks**

Similar to manual calibration, the program attempts to choose isolated, strong HSQC peaks as reference peaks. In practice, 20 most isolated peaks which have the longest distances from other peaks will be picked out from HSQC spectrum (peaks in the side-chain region, i.e.  $^{15}\text{N} = 106.5\sim 119.2$  ppm,

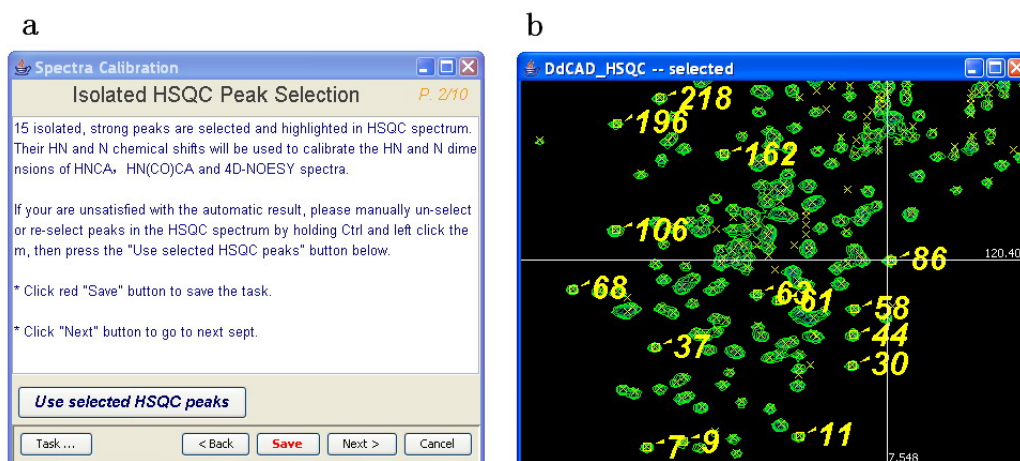
$^1\text{H}_\text{N}$ =5.77~8.64 ppm, will be ignored), and the 15 strongest peaks among them will be selected as the “referential HSQC peak” and highlighted in HSQC spectrum. (Figure 7.6 b)

If users are unsatisfied with the automatically generated result, they can manually un-select or re-select peaks in the HSQC spectrum by holding Ctrl and left clicking them, and then press the "Use selected HSQC peaks" button provided by the interface (Figure 7.6 a) to select them as referential HSQC peaks.

### 7.2.3.3 HNCA calibration (H, N)

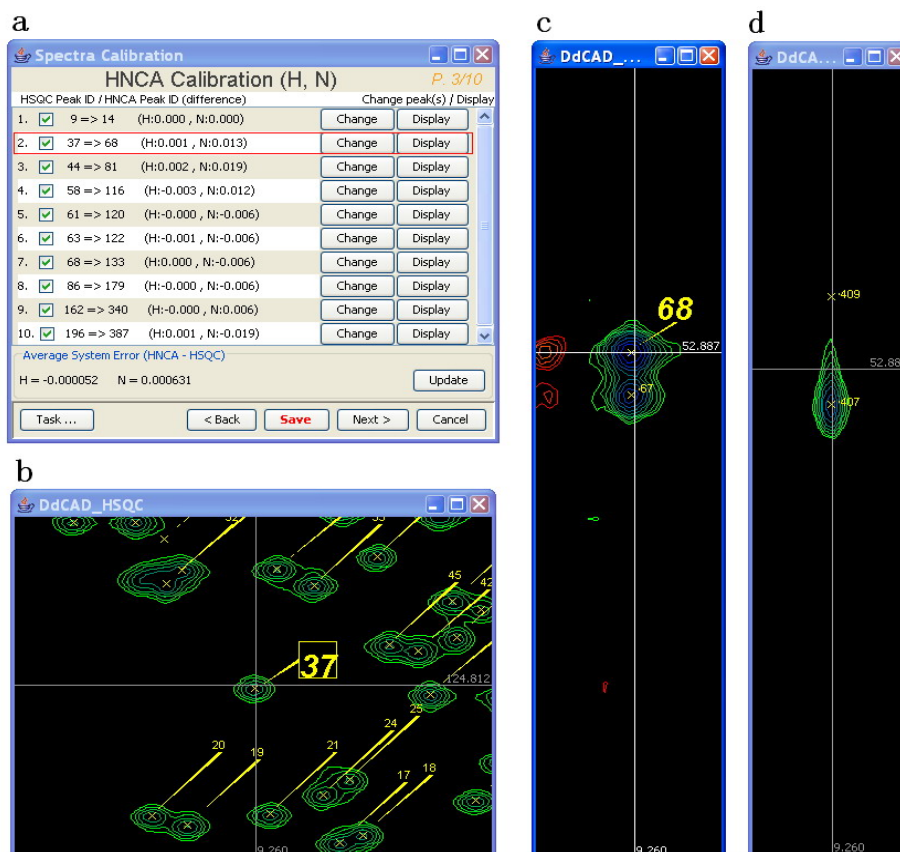
For each referential HSQC peak, the program will automatically detect a strongest correlated peak in HNCA spectrum based on their common  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shift values. Ten pairs of the peaks (each pair consists of one HSQC and HNCA peak) that have the most similar  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  systematic errors will be selected and displayed on the “HNCA Calibration (H, N) Panel”.(Figure 7.7 a)

The peak IDs and systematic errors of each peak-pair are clearly visible on the interface. A "Display" button could be used to display and highlight the corresponding HSQC peak, HNCA strip plot, and HN(CO)CA strip plot (Figure 7.7 b - d). The "Change" button at the right top corner (Figure 7.7 a) could be used to replace the HSQC and/or HNCA peak(s) in the peak-pair. Un-tick the check-box in front of a peak-pair, and then the pair of peaks will be ignored when calculating the average systematic error.



**Figure 7.6** Isolated HSQC peak selection panel (a) and its correlated HSHC spectrum (b).

In HSQC spectrum, the 15 referential HSQC peaks are automatically selected and highlighted by large labels.



**Figure 7.7** Graphic interfaces for HNCA Calibration (H, N).

(a) Main panel, (b) HSQC spectrum, (c) HNCA spectrum and (d) HN(CO)CA spectrum. Spectra (b-d) are centre on  $^1\text{H} = 9.260$  ppm and  $^{15}\text{N} = 124.812$  ppm.

If there is any artificial or overlapped peak among the referential HSQC peaks, in this step, such a peak can be easily identified using the HN(CO)CA spectrum, as a single HSQC peak (corresponding to one residue) should be correlated with only one HN(CO)CA peaks (Figure 7.7 d). Users may go back to the last step and re-select the isolated HSQC peak.

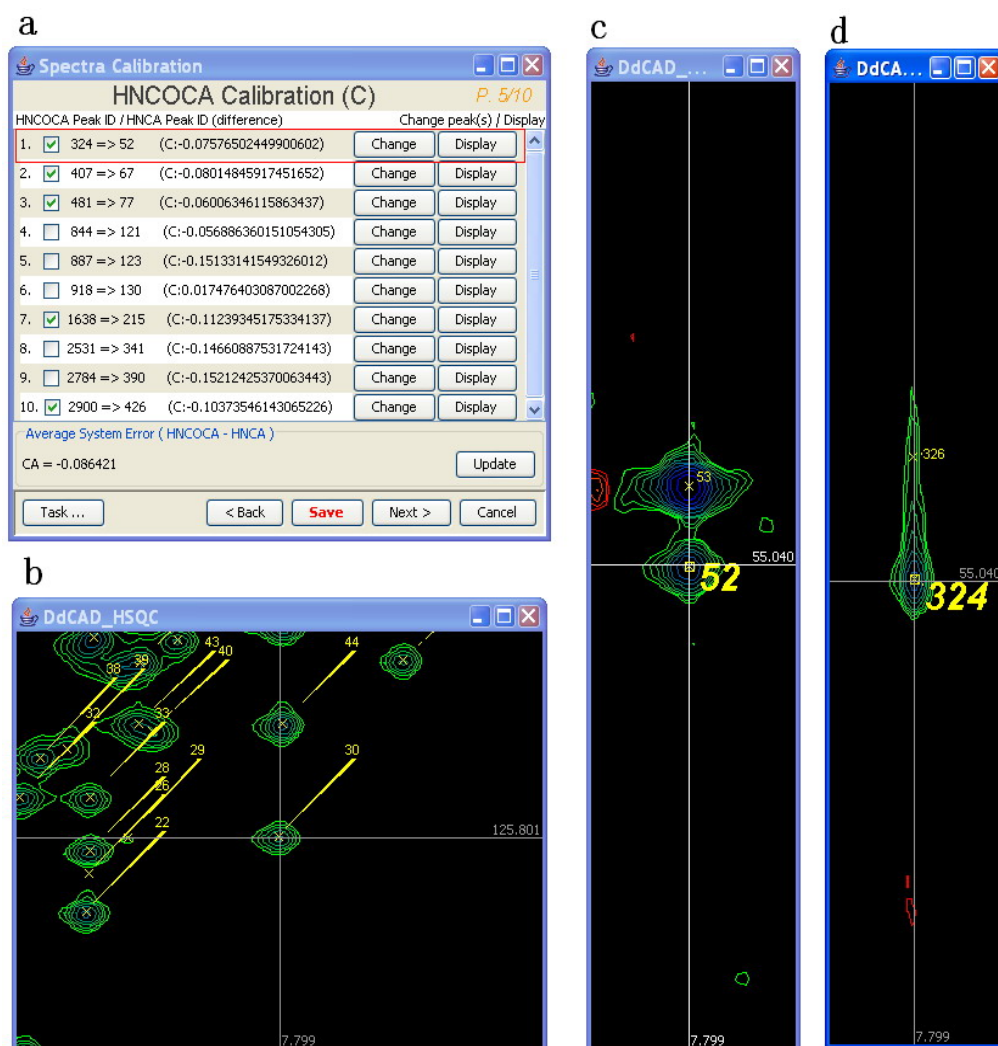
#### **7.2.3.4 HN(CO)CA calibration (H, N)**

This step is very similar to the previous step.

#### **7.2.3.5 HN(CO)CA calibration (C)**

For each HNCA peak that has been confirmed in HNCA calibration (H,N) (7.2.3.3), the program will automatically detect its correlated peak from HN(CO)CA spectrum based on its common  $^1\text{H}_\text{N}$ ,  $^{15}\text{N}$  and  $^{13}\text{C}$  chemical shifts. All the peak pairs will be displayed on the “HN(CO)CA Calibration (C) Panel” (Figure 7.8 a), but only five pairs that have the most similar  $^{13}\text{C}$  systematic errors will be selected.

The rest of this step is similar to the previous steps.



**Figure 7.8** Graphic interfaces for HN(CO)CA Calibration (C).

(a) Main panel, (b) HSQC spectrum, (c) HNCA spectrum and (d) HN(CO)CA spectrum. Spectra (b-d) are centre on  $^1\text{H}=7.799$  ppm and  $^{15}\text{N}=125.801$  ppm.

### 7.2.3.6 4DNOE calibration (H, N)

For each referential HSQC peak (there should be 15), a strongest correlated peak from 4D-NOESY spectrum will be automatically detected by the program based on its common  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shifts. Since the resolution of 4D-NOESY spectrum is normally not very high and the systematic error between the

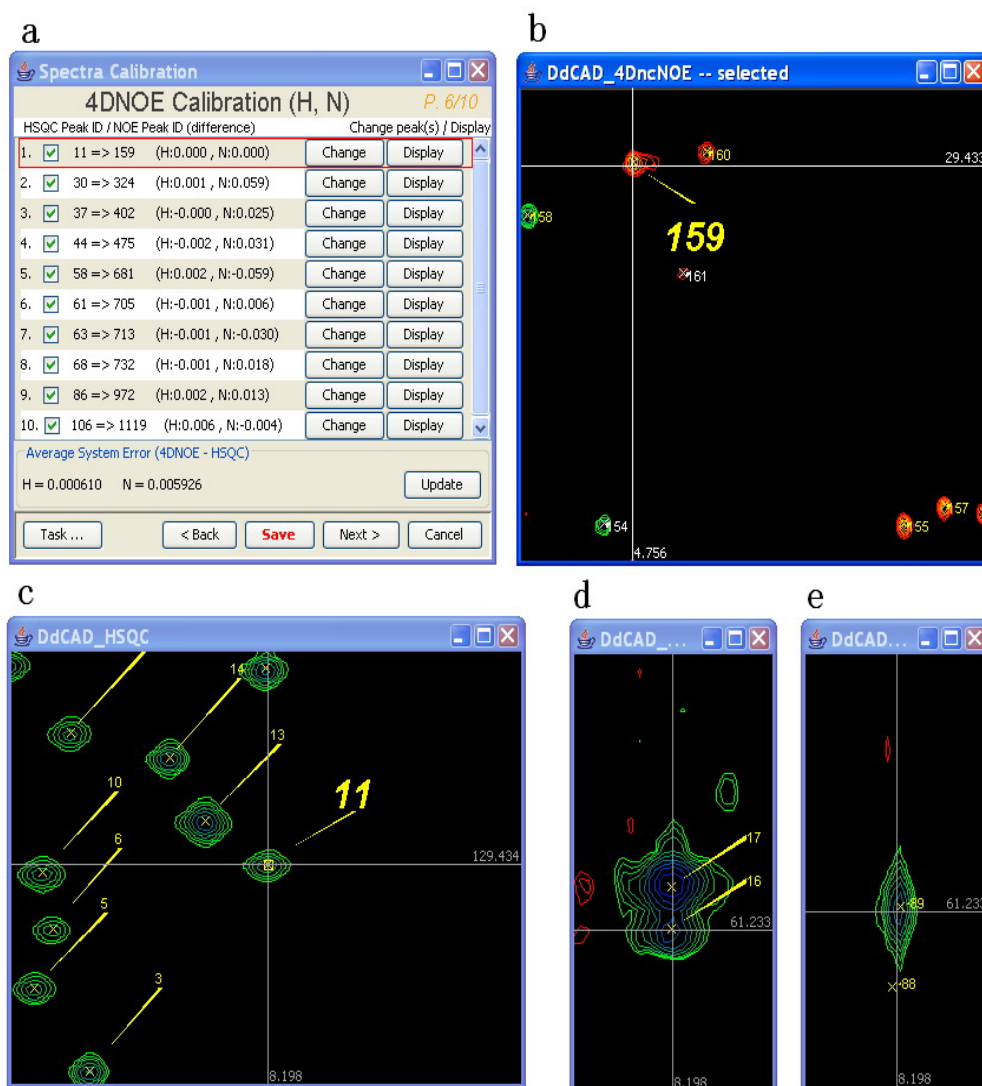


NOESY and HSQC spectra could be quite large, all the NOE peaks within a region of the referential HSQC peak (which is defined as a circle with the center at the position defined by the HSQC and with a radius equal to half of the distance between the HSQC peak and its nearest peak) will be considered when attempting to find the strongest one. Ten pairs of peaks that have the most similar  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  systematic errors will be selected and displayed on the “4DNOE Calibration (H, N) Panel”.(Figure 7.9 a)

Similar to previous steps, the HSQC peak and 4DNOE slice of each peak-pair could be easily displayed with the chosen peaks highlighted by an extreme large label. Any peak or peak-pair could be manually modified or ignored.

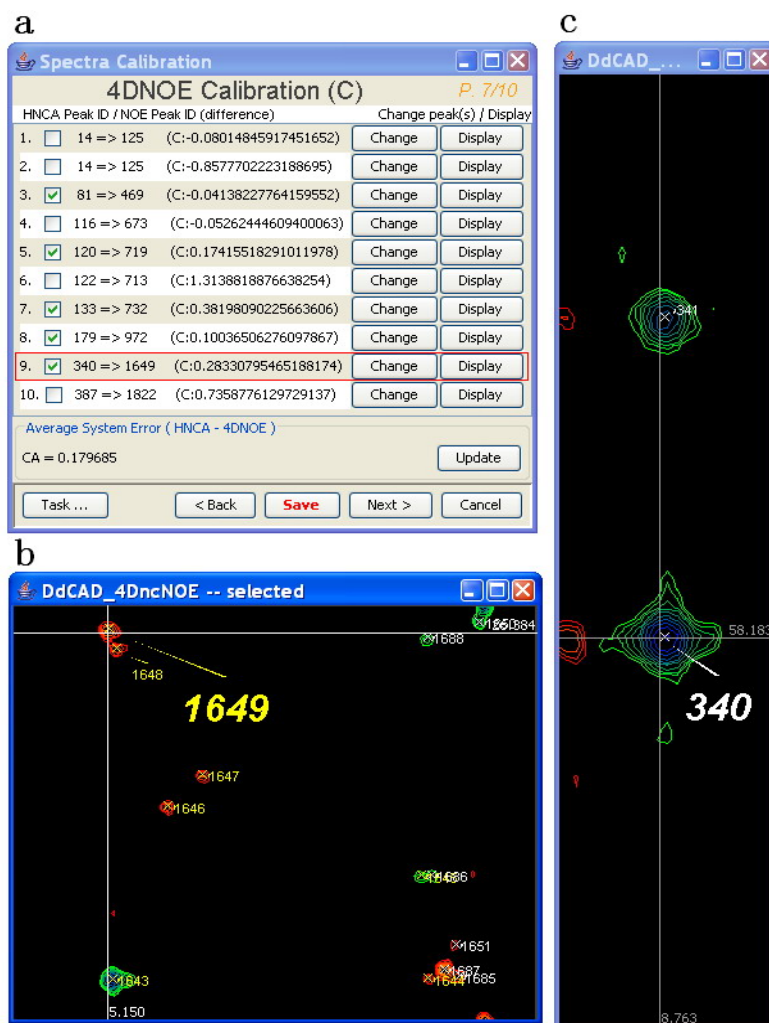
#### 7.2.3.7 4DNOE calibration (C)

The idea of this step is to calibrate the  $^{13}\text{C}$  systematic error between HNCA and  $\text{H}_\text{N}\text{-CH}\alpha$  NOE peaks. The program uses the confirmed HNCA peaks as reference to search for the  $\text{H}_\text{N}\text{-CH}\alpha$  NOE peaks. The searching is limited within the  $\text{H}\alpha$  region with a  $^1\text{H}_\text{C}$  chemical shift range from 3.0 to 6.0 ppm. All the resulted peak-pairs will be displayed on the “4DNOE Calibration (C) Panel” (Figure 7.10 a), but only the best five of them will be selected.



**Figure 7.9** Graphic interfaces for 4DNOE Calibration (H,N).

(a) Main panel, (b) 4D-NOESY spectrum, (c) HSQC spectrum, (d) HNCA spectrum and (e) HN(CO)CA spectrum. Spectra (b-e) are centre on  $^1\text{H}=8.198$  ppm and  $^{15}\text{N}=129.434$  ppm.



**Figure 7.10** Graphic interfaces for 4DNOE Calibration (C).

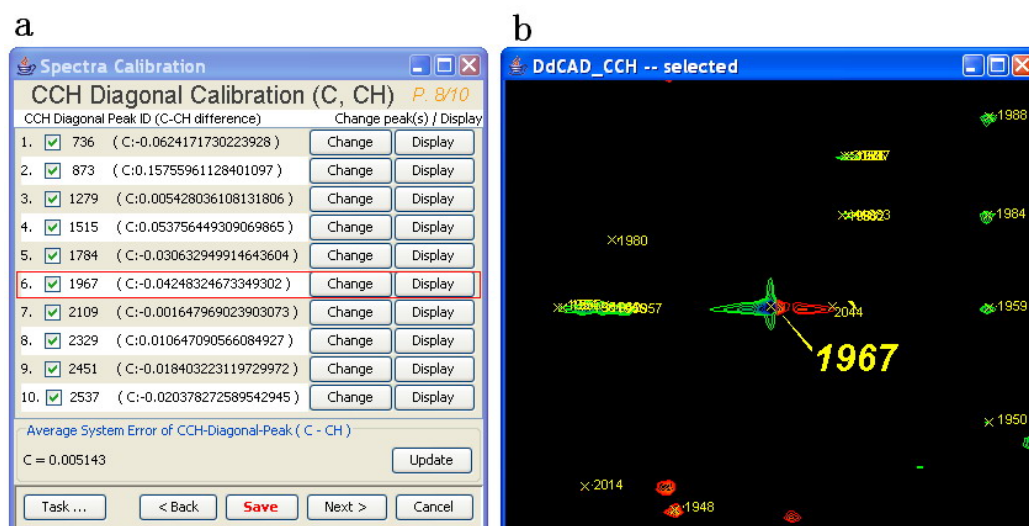
(a) Main panel, (b) 4D-NOESY spectrum and (c) HNCA spectrum. Spectra (b, c) are centre on  $^1\text{H}=8.763$  ppm and  $^{15}\text{N}=114.928$  ppm.

### 7.2.3.8 CCH diagonal calibration (C, CH)

Theoretically, the diagonal peaks in MQ-CCH-TOCSY spectrum should have the same chemical shift values in the two  $^{13}\text{C}$  dimensions. The systematic error is always the main cause for inconsistency. To calibrate the MQ-CCH-TOCSY spectrum, users should, first of all, pick out the diagonal peaks and make  $^{13}\text{C}$  shifts of each peak in the two dimensions consistent.

Normally, the diagonal peaks are also the strongest peaks in CCH-TOCSY spectrum. The program will pick out 30 strongest peaks within the diagonal regions ( $\pm 3$  ppm in  $^{13}\text{C}$  dimension), and then choose 10 peaks that have similar systematic errors or differences between the two  $^{13}\text{C}$  dimensions and show them in the “CCH Diagonal Calibration (C, CH) Panel”. (Figure 7.11 a)

The selected peaks are not necessarily unfolded peaks. The program will automatically test the possible alias of a given peak in one  $^{13}\text{C}$  dimension while searching for the diagonal peaks.



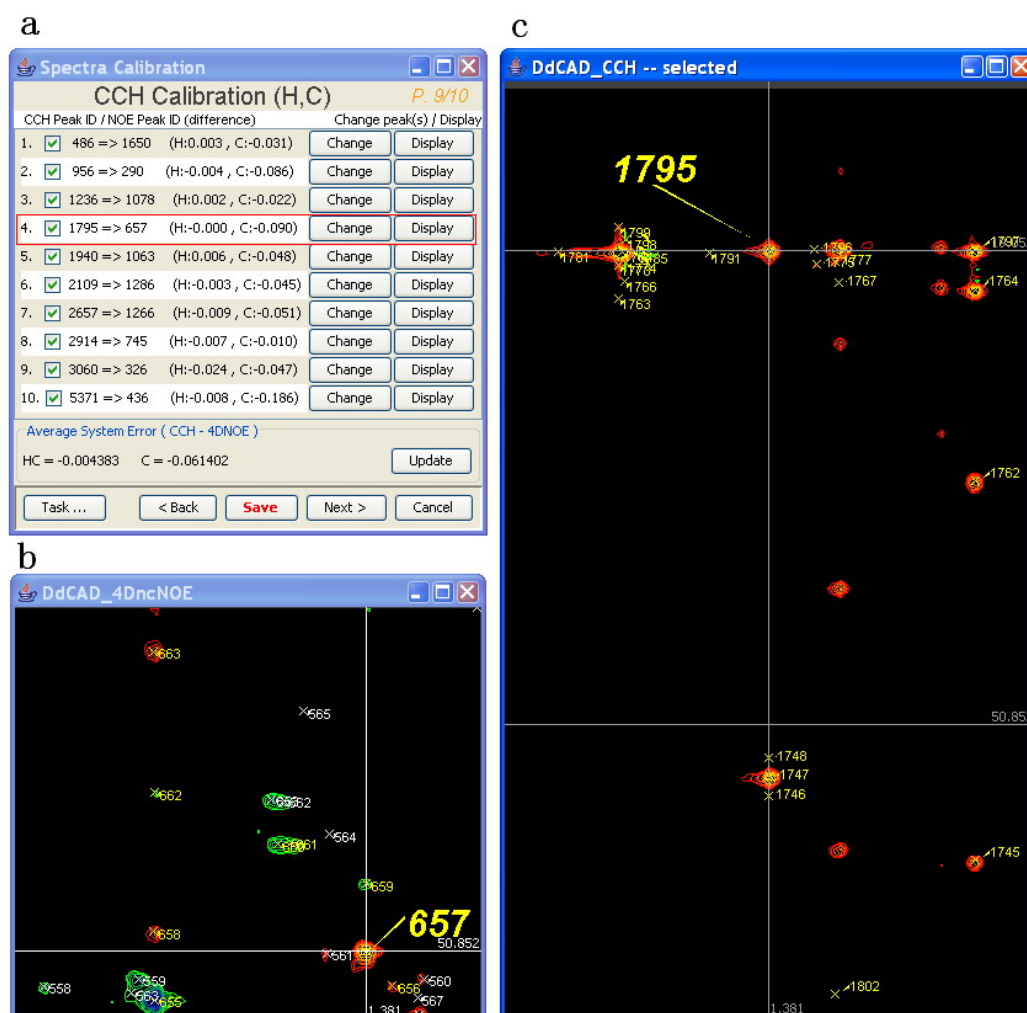
**Figure 7.11** Graphic interfaces for CCH Diagonal Calibration (C, CH).

(a) Main panel, and (b) CCH-TOCSY spectrum. Chemical shifts of peak 1967:  $^1\text{H}=3.896$  ppm,  $^{13}\text{C}_\text{H}=42.848$  ppm,  $^{13}\text{C}_\text{H}=42.806$  ppm.

### 7.2.3.9 CCH calibration (H,C)

XYZ4D will select 30 strongest and most isolated CCH diagonal peaks. For each of them, the program will search through the entire 4D-NOESY spectrum for the strongest NOE peak that may be correlated to them. Ten

4DNOE-CCH correlated peak-pairs that have similar systematic errors will be subsequently used as references and displayed on the “CCH Calibration (H,C) Panel”. (Figure 7.12 a)



**Figure 7.12** Graphic interfaces for CCH Calibration (H, C).

(a) Main panel, (b) 4D-NOESY spectrum, and (c) CCH-TOCSY spectrum.

The systematic errors or differences showed in Figure 7.12a are between the  $^{13}\text{C}$  dimension of the TOCSY which has the larger sweep width and the  $^{13}\text{C}$  dimension of the 4D NOESY. After users correct this dimension, the program will automatically correct the other  $^{13}\text{C}$  dimension of the TOCSY spectrum

according to the calibration result for the two  $^{13}\text{C}$  dimensions of the TOCSY spectrum obtained from the last step (Section 7.2.3.8).

The theory behind the automatic detecting process is that strong peaks in CCH-TOCSY and 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectra mainly involve methyl groups since the number of protons of a  $\text{CH}_3$  group is three times as large as that of a CH group and the proton and  $^{13}\text{C}$  spins of the  $\text{CH}_3$  group have longer relaxation times than CH and  $\text{CH}_2$  groups due to the free rotation along the C3 axis (Figure 7.12 b). Moreover, since the methyl protons always have the lowest  $^1\text{H}$  and  $^{13}\text{C}$  chemical shift values, it's most unlikely their NOE peaks will mess up with other peaks, which will make the automatic detecting process and manual inspection easier and more reliable. Otherwise, it will be very difficult to tell the relationship between NOE peaks and CCH peaks. Users may also consider using some idiosyncratic protons (which have extreme high or low or strange  $^1\text{H}/^{13}\text{C}$  chemical shift values) or those protons from  $\text{CH}_2$  groups (which may appear by pairs on the 4D NOESY spectrum and have identical peak patterns in the CCH-TOCSY spectrum) to identify reliable 4DNOE-CCH peak-pairs.

#### 7.2.3.10 Results panel

The “Spectral Calibration Results Panel” (Figure 7.5 b) summarizes all the systematic errors between the five essential spectra using the results of the 7 steps (7.2.3.3 – 7.2.3.9) described above. Users could modify the systematic error values by double-clicking them or going backwards to any of the previous steps, reprocess it or examine the results without affecting other steps.

Before the “Finish” button being clicked, all the calibration process will “remain only on paper”, with no actual operation carried out on any spectrum. Users can play around with the interfaces and files without affecting any spectrum or peak. But when “Finish” button is pressed, the program will apply systematic error corrections on all the spectra and their dataset, which means the center of each dimension for each spectrum will be changed according to the calibration results and each peak’s position will also been synchronized with its spectrum. This is a critical step because if there is any analysis job to be done on the spectrum or dataset before this systematic error correction, its result may be inaccurate. This step is irreversible, and although users can manually modify the dimension center back to its original value (see Section 6.3.3.1, Chapter 6), the shifted peaks will remain in the post-correction-positions.

#### **7.2.4 Cluster identification module**

Clusters are formed by grouping HSQC, HNCA, HN(CO)CA and 4D-NOESY correlated peaks according to their common NH chemical shifts, where NH denotes amide spins  $^{15}\text{N}$  and  $^1\text{H}$ . Each non-overlapped cluster contains all the signals collected from a single amide proton and can be considered as a representation of a residue in our backbone assignment approach. Building clusters is one of the most essential steps for achieving good backbone assignment with our strategy. The Cluster Identification Module is designed to help users quickly and precisely create clusters.

### 7.2.4.1 Method

The Cluster Identification Module creates clusters by dividing HNCA peaks into different groups according to their common  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shifts, and then searching through HSQC and HN(CO)CA spectra for their associated peaks.

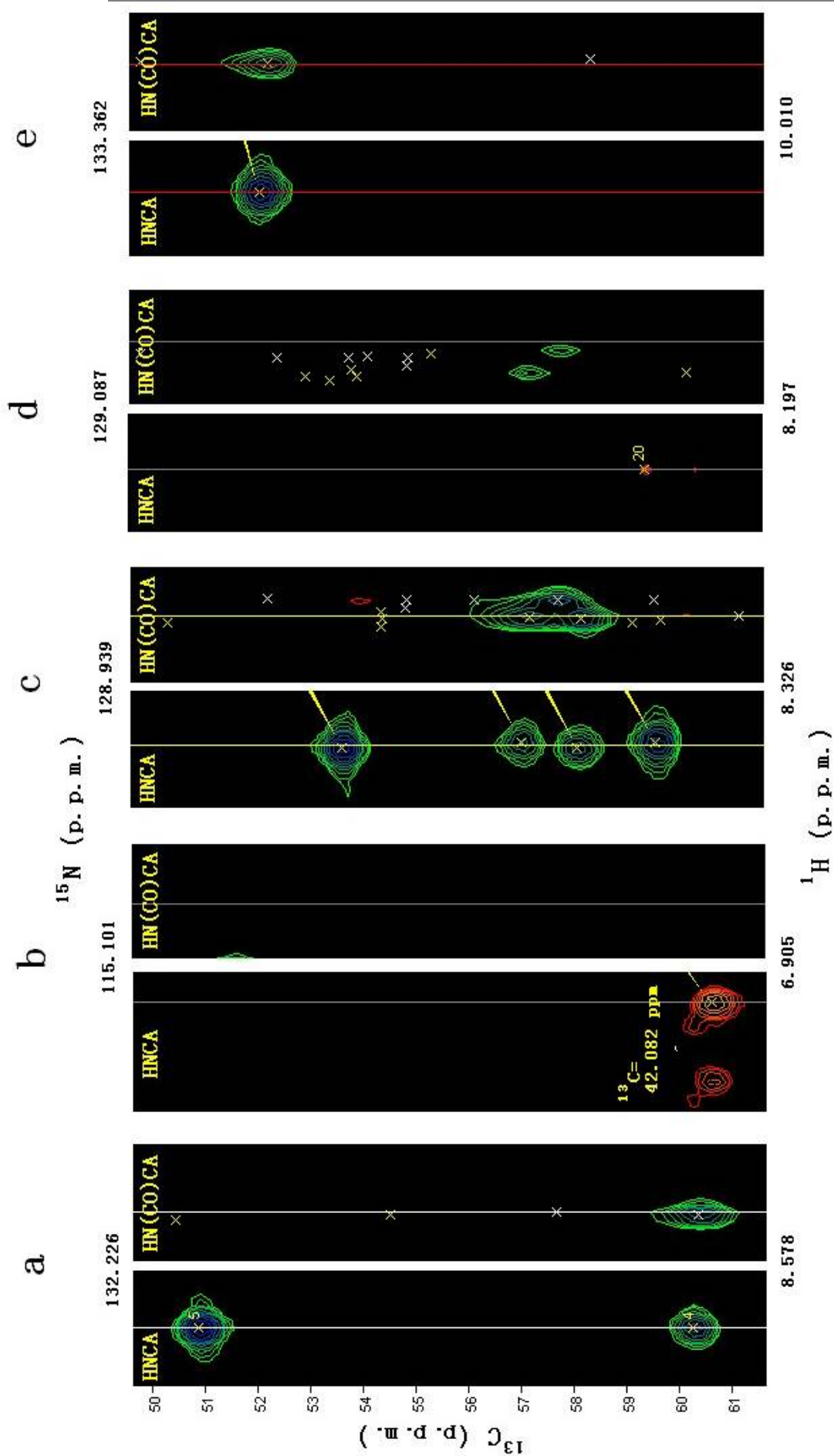
According to the following conditions, the program classifies all the newly-created clusters into 5 categories: corrupted, normal, overlapped, suspected, and side-chain.

1. If a cluster has two HNCA peaks and one HN(CO)CA peak, it will be classified as “normal” (Figure 7.13 a). Meanwhile, based on the chemical shift of the correlated HN(CO)CA peak, the program will automatically identify the sequential HNCA peak that comes from the preceding residue (i-1).
2. If a cluster has only one HNCA peak,
  - a. its  $^{15}\text{N}$  and  $^1\text{H}_\text{N}$  chemical shifts are located within the side-chain region (  $^{15}\text{N}= 106.5\sim 119.2$  ppm,  $^1\text{H}_\text{N} =5.77\sim 8.64$  ppm) ,
  - b. its  $^{13}\text{C}$  chemical shift value is less than 50ppm, and
  - c. there is another peak that has the same  $^{15}\text{N}$  and  $^{13}\text{C}$  chemical shifts in the HNCA spectrum,

this cluster will be classified as a side-chain cluster (Figure 7.13 b).



3. If a cluster has more than two strong HNCA peaks or more than one possible HN(CO)CA peak, it will be defined as “overlapped” (Figure 7.13 c), which means it comprises signals from two or more amides that have degenerated NH chemical shifts. Based on the fact that one amide gives rise to only one HN(CO)CA peak and no more than two HNCA peaks, the program will automatically determine the “number of amides” of a cluster according to the number of its peaks. A “strong HNCA peak” should be 30% larger in intensity than the possible HNCA peaks. The “possible HNCA peaks” are the first  $n$  strongest peaks in a HNCA spectrum, where  $n = (\text{total residue number in the protein} - \text{proline residue number}) * 2 - 2$ . The “possible HN(CO)CA peaks” are the first  $m$  strongest peaks in a HN(CO)CA spectrum, where  $m = n/2$ .
4. If a cluster has only weak HNCA peaks that do not match any HN(CO)CA peak in (H, N C) shifts and HSQC peak in (H, N) shifts, it will be classified as “corrupted” (Figure 7.13 d).
5. The rest clusters, such as those that have only one HNCA peak or those that do not match any HSQC or HN(CO)CA peak, will be defined as suspected clusters (Figure 7.13 e).



**Figure 7.13** Examples of cluster classification.

Examples of (a) Normal cluster, (b) side-chain cluster, (c) overlapped cluster, (d) corrupted cluster, and (e) suspected cluster.

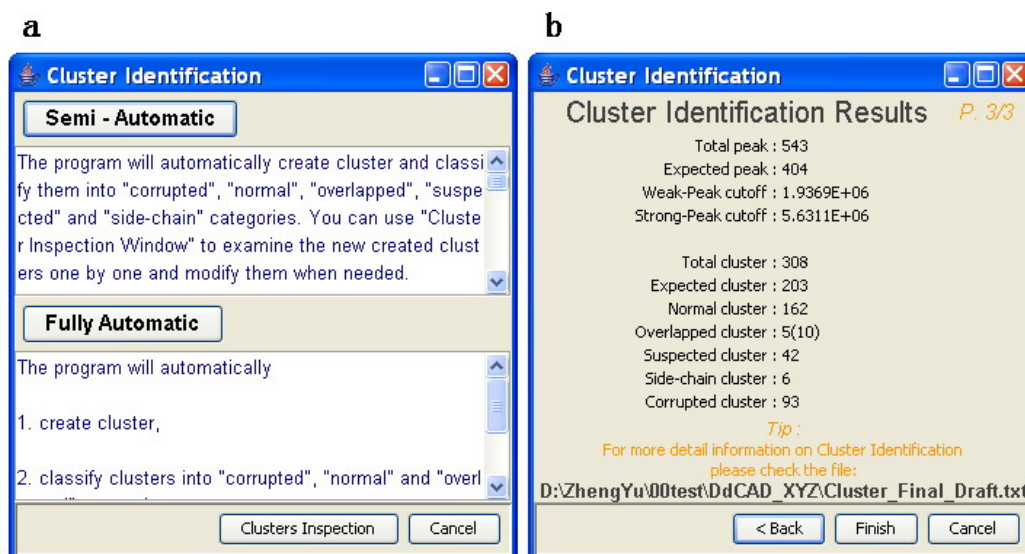
The result of this automatic classification procedure is normally quite reliable. From our experience, more than 99% of the “corrupted”, “normal” and “side-chain” clusters will be correctly classified. Those “overlapped” clusters may need some manual inspection, for experienced spectroscopists may be able to spot some overlapped HNCA peaks or divide an overlapped cluster into several single clusters.

In a fully automatic mode, the program will score the suspected clusters based on the perfectibility and intensity of their peaks, and then re-classify them as “corrupted” or “normal”. In semi-automatic mode, users could carefully inspect the suspected clusters one by one with the interface provided by the Cluster Identification Module, and manually re-classify them as “corrupted” or “normal”.

At the end of the cluster identification process, corrupted and side-chain clusters will be eliminated and all the remained clusters will be reorganised and given identity numbers. The results will be saved in the file: ClusterList.txt .

#### **7.2.4.2 Main panel**

Figure 7.14a shows the main panel of the Cluster Identification Module.



**Figure 7.14** Main window (a) and result summary window (b) of Cluster Identification Module.

As mentioned above, if users choose fully automatic mode, the program will automatically create clusters and classify them into “corrupted”, “normal” and “overlapped” categories. After that, the program will directly jump to “Cluster Identification Results Panel” (Figure 7.14 b). If users choose semi-automatic mode, the clusters will be classified into 5 categories, and a “Cluster Inspection Panel” will be active (Figure 7.15 a). Users may notice that there is a disabled “Cluster Inspection” button located in the bottom of the main panel. This button will be enabled after users finish the whole cluster identification procedure. It can be used to modify the cluster identification results later on in the case of any improper-identification discovered during the subsequent study.

### 7.2.4.3 Cluster inspection panel

The “Cluster Inspection Panel” (Figure 7.15 a) allows users to examine the created and classified clusters one by one. A temporary ID number has been

given to each of these clusters which can be used to navigate through them smoothly.

When switched to a particular cluster, the HSQC spectrum window will immediately centre on its correlated HSQC peak, with the pinpoint position of the cluster marked out and the average chemical shift values labelled on both dimensions (Figure 7.15 b). The HNCA and HN(CO)CA strip plots of the cluster will be painted in their window and a vertical grid with the cluster's ID labelled on its top will denote the centre of the cluster in the  $^1\text{H}$  dimension and its status (Figure 7.15 c, d). For a "normal" cluster the grid and label will be in white colour, while for a "corrupted" or "side-chain" cluster they will be in dark grey. An "overlapped" cluster will be labelled with a bright yellow grid and a "suspected" one with a red grid (Figure 7.13). In the 4D-NOESY spectrum, the corresponding NOE slice will also be displayed (Figure 7.15 e), and with all the spectra synchronized, when users move the cursor into one spectrum, the correlated crosshairs will appear in other spectra so that user could easily compare the centre of several peaks in different spectra at the same time.

The cluster's status, number of amides (if available), average  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shift values and the HNCA peak(s) contained in the cluster will be displayed on the control panel (Figure 7.15 a). The status and number of amides (if available) could be changed using the pull down menus. The average  $^1\text{H}_\text{N}$  and  $^{15}\text{N}$  chemical shift values will be generated automatically by the program, while users may need to update them after adding or removing peak(s) from the cluster.

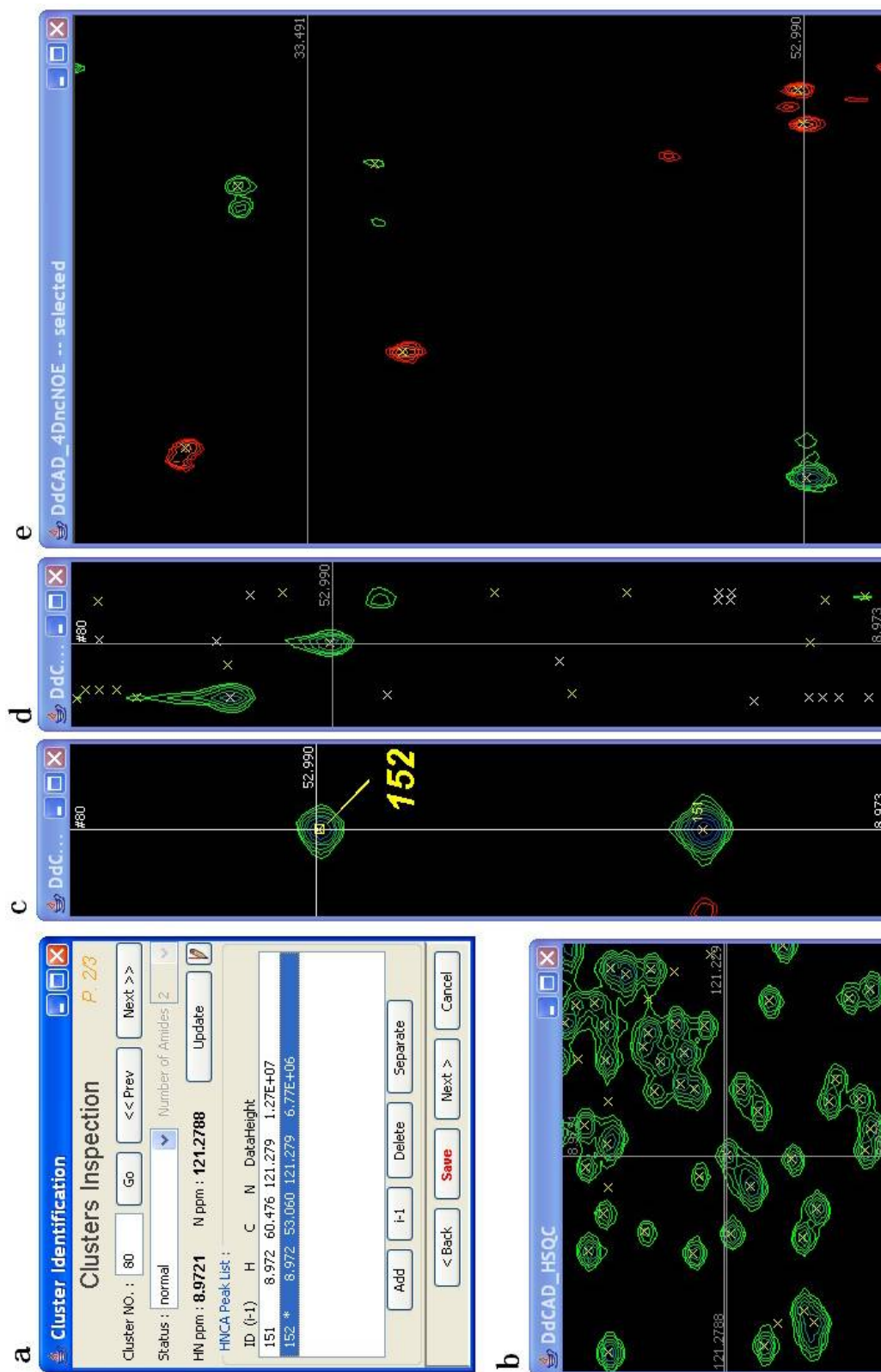


Figure 7.15 Cluster inspection interface.

(a) Control panel and (b) HSQC, (c) HNCA, (d) HN(CO)CA, (e) 4D-NOESY spectra.

The “HNCA Peak List” panel shows all the HNCA peak(s) composed of the cluster with their ID (a “\*” will appear on the right side if the peak comes from the previous residue, i.e. “i-1” residue), chemical shifts and Intensity. Users may add or remove peak(s) from it, define an “i-1” HNCA peak or separate one or more peaks into a new cluster. Peak(s) selected in the HNCA peak list will be highlighted by an extreme large label in the HNCA strip plot.

The small button located at the right side of the “Update” button with a “pencil-drawing-a-line” icon could be used to change the display mode of the grids that mark out the clusters in HNCA and HN(CO)CA spectra. There are three modes of showing those grids: no grid, only the grid for currently focused cluster and all grids. By keep pressing the button, the display mode could be changed within a blink.

The red “Save” button at the bottom of the panel could be used to save the cluster list at anytime if anything about the cluster has been changed. It’s recommended that users should press it once it pops up in order to reduce the risk of data loss in case of a computer crash or freeze. The operation of saving clusters is totally separated from the project saving procedure, and although the files are saved in the same directory, users should not expect that by saving the project, the cluster list and current operation state will be restored next time when they open the software.

#### **7.2.4.4 Results panel**

After “overlapped” and “suspected” clusters been closely inspected (all the “suspected” should be altered to other status, otherwise a warning message will

pop out), users may proceed to the “Cluster Identification Results Panel” ( Figure 7.14 b).

By clicking the “Finish” button, those remained “suspected” clusters will be automatically changed to “corrupted” or “normal” according to their scores. Then, those “corrupted” and “side-chain” clusters will be eliminated, leaving the rest “normal” and “overlapped” clusters for reorganization.

### 7.2.5 CCH & 4DNOE inspection module

The quality of MQ-CCH-TOCSY and 4D- $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectra is usually quite bad compared to the 3D HNCA and HN(CO)CA spectra. The CCH-TOCSY spectrum is always severely overlapped due to the overwhelming number of resonances (Figure 7.17 a). With the intervention of artificial signals and intense  $\text{H}_2\text{O}$  resonances (at  $\sim 4.7\text{ppm}$ ), it's extreme difficult for computer to recognize any peak patterns in the CCH-TCSY. Yet at the 4D-NOESY spectrum, the low signal-noise ratio combined with the low spectrum resolution brings a lot of inconvenience in distinguishing the NOE signals and matching their peak positions.

In order to minimize the difficulty of following steps and increase the accuracy of spin-system identification, it's necessary to spend some time on inspecting the two spectra and eliminating those useless or artificial peaks before the 4D-NOESY peaks could be assigned to clusters and further divided into spin-systems with the HNCA and CCH-TOCSY spectra. This is the reason behind the development of CCH & 4DNOE Inspection Module.



### 7.2.5.1 Interface

The interface of CCH & 4DNOE Inspection Module is relatively simple compared to other modules. Buttons that access to 5 relatively independent tasks are placed in two separate panels. One panel is specifically for CCH-TOCSY spectrum containing 2 buttons: "Eliminate Water-peaks" and "Eliminate Artificial-peaks" (Figure 7.16 a). The other one is only for the 4D-NOESY spectrum containing 3 buttons: "Collect NOE-peaks for Clusters", "Eliminate Over-edge NOE Peaks" and "Correct Alias of NOE-peaks" (Figure 7.16 b). Meanwhile, the module offers two navigators: "CCH Peak Navigator" (Figure 7.17 a) and "Cluster Navigator" (Figure 7.17 b), allowing users to quickly locate a CCH peak or cluster in the spectra. These interactive navigators could be very valuable since actually observing the shape and distribution of peaks can be very important for discovering artefacts in spectra.

The working principle of "Eliminate Over-edge NOE Peaks" is very similar to the one described in the previous section (Section 7.2.2). The principle of the other four tasks will be elaborated in the following sections.

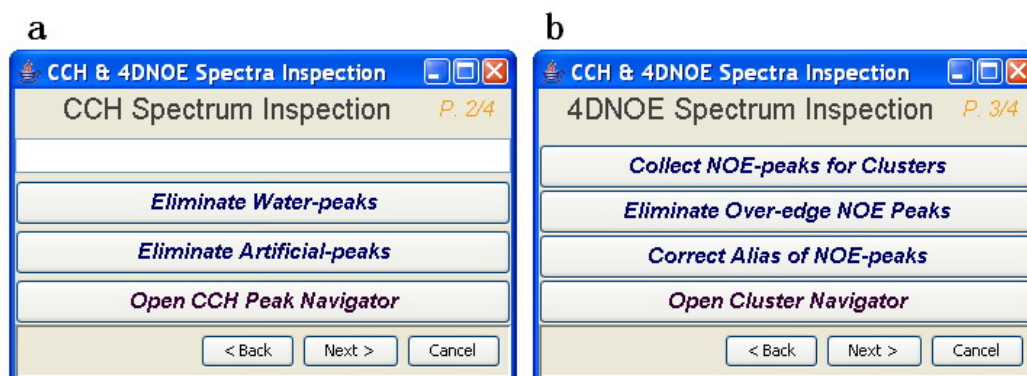


Figure 7.16 Control panels of (a) CCH-TOCSY and (b) 4D-NOESY Inspection.

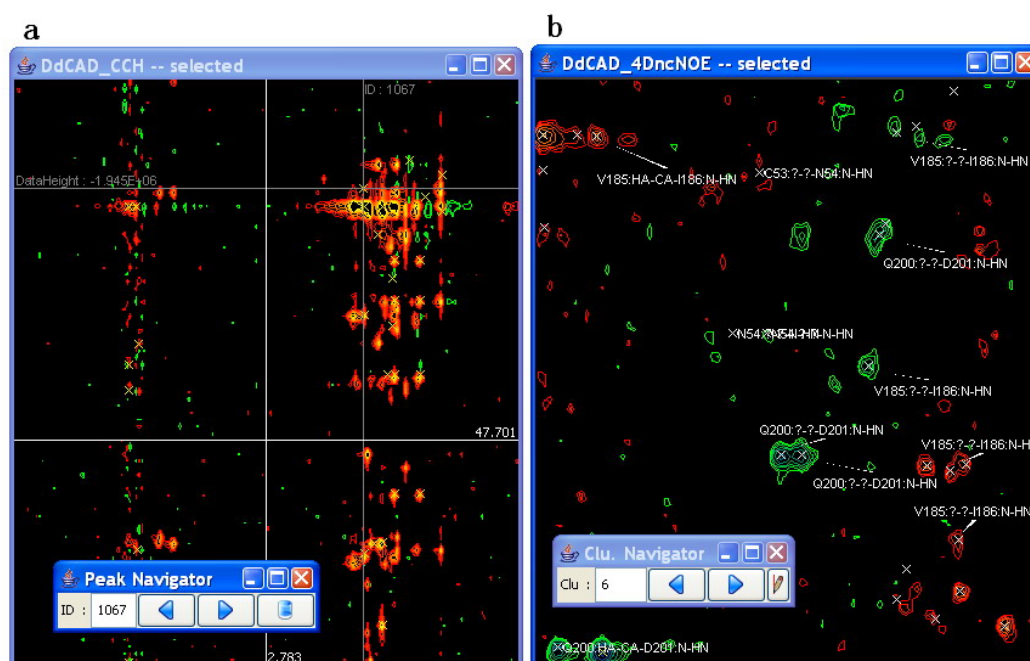


Figure 7.17 Interfaces of (a) CCH Peak Navigator and (b) Cluster Navigator.

### 7.2.5.2 CCH water-peak elimination

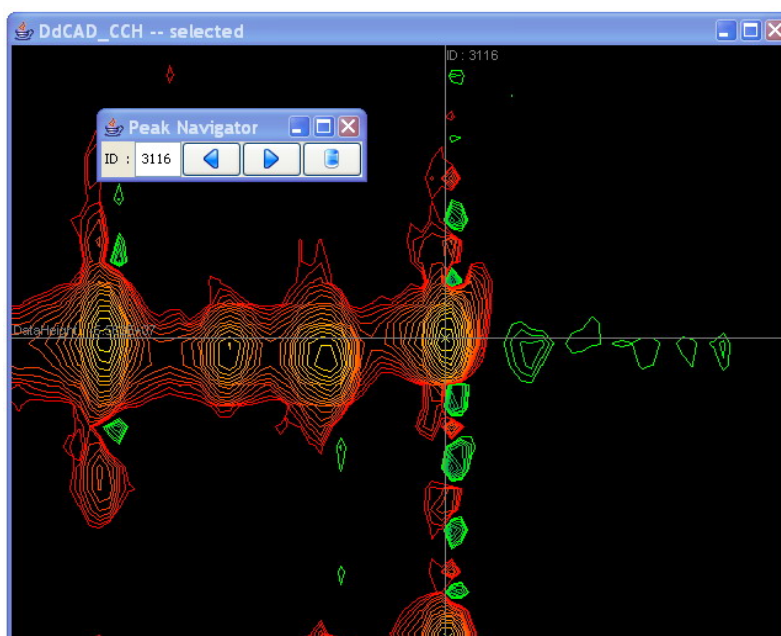
There are a huge number of peaks between 4.5 ~ 5.0 ppm in  $^1\text{H}$  dimension, which result from the strong water signal. These peaks, whether positive or negative, are much stronger than normal noise. They hide a majority of weak  $\text{H}\alpha$  signal and interfere with the others, making it extreme difficult to recognize any useful peaks around this region.

In order to eliminate these water-peaks, the program will carefully examine each CCH slice and scans the  $^1\text{H}$  dimension from left to right at a scale of 0.01 ppm. If there are more than 10 peaks crowded in a small column (width = two times of the spectral resolution) around a certain  $^1\text{H}$  chemical shift value and the ratio between positive and negative peaks are about the same (40% ~ 60%), this  $^1\text{H}$  chemical shift will be defined as the start / end line of the water-peak-region in this particular slice, and the strongest peak on this line will be defined as a “referential water-peak”. When a water-peak-region has been detected in more than 70% of the slices with very similar start/end lines (difference < 0.2ppm), the average of 30% of most similar start/end lines will be used to define the water-peak-region for the whole spectrum. In this region, if a peak is weaker than 50% of referential water-peaks, it will be eliminated.

Using such an approach, more than 80% of the water peak can be successfully removed; while those real  $\text{H}\alpha$  signals will not be affected even they are only slightly stronger than the water peaks. Certainly, most of the weak  $\text{H}\alpha$  signals may be mistakenly deleted, but even the most experienced spectroscopists is unable to distinguish them from the water peak, let alone a computer.

### 7.2.5.3 CCH artificial -peak elimination

In CCH-TOCSY spectrum, one may find wiggling artificial-peaks (Figure 7.18) surrounding a very strong peak. These artificial-peaks are quite common when linear prediction is applied to process the raw data. In this case, positive and negative peaks appear in a stagger form, with a constant distance between peaks, and the peak intensity decreases with the increasing distance from the strong peak.



**Figure 7.18** An example of artificial-peaks that surround strong peaks along the Y-axis in CCH-TOCSY spectrum.

A strong peak is denoted by crosshair. The wiggling artificial-peaks above and below it in the  $^{13}\text{C}$  dimension are clearly visible.

The program will scan the surrounding area of strong peaks (the intensity is 10 times larger than the weakest peak) in the CCH-TOCSY spectrum, and if there are more than 5 peaks appearing in an artificial-peak-pattern, they will be automatically deleted.

As the artificial-peaks are often interfered by real peaks or other noise, the program can only identify about 50% of them. It's recommended that users perform a slice-by-slice inspection, and manually delete those obvious artificial peaks.

#### 7.2.5.4 NOE-peak collection

A pair of unique tolerance (threshold) values, one for  $^1\text{H}_\text{N}$  and another for  $^{15}\text{N}$  are assigned to every cluster when the cluster is generated by the Cluster Identification Module (Section 7.2.4). The tolerance value of a given cluster for  $^1\text{H}_\text{N}$  ( $^{15}\text{N}$ ) is equal to half of the distance between the  $^1\text{H}_\text{N}$  ( $^{15}\text{N}$ ) of this cluster and that of the nearest cluster, but not smaller than the spectral resolution of the 4D - NOESY spectrum in the  $^1\text{HN}$  ( $^{15}\text{N}$ ) dimension. The program will automatically assign a NOE peak to a cluster if the peak is located within the tolerance range of that cluster. A NOE peak could be assigned to more than one cluster. Those that could not be assigned will be permanently removed to save computer memory and waiting time for reading/writing peak list.

#### 7.2.5.5 NOE-peak alias correction

With “Selected Chemical Shift Statistics” provided by Biological Magnetic Resonance Data Bank (Ulrich, Akutsu et al. 2008), we studied the chemical shift distribution of C-H groups of proteins and calculated the  $^{13}\text{C}$ - $^1\text{H}$  correlation regions (Table 7.1) in which the (C, H) chemical shift of >99% HC-NH NOE peaks should be located .

Table 7.1 Statistic  $^{13}\text{C}$ - $^1\text{H}$  chemical shift region.

| $^1\text{H}$ region<br>(ppm) | Min. $^{13}\text{C}$<br>(ppm) | Max. $^{13}\text{C}$<br>(ppm) | $^1\text{H}$ region<br>(ppm) | Min. $^{13}\text{C}$<br>(ppm) | Max. $^{13}\text{C}$<br>(ppm) |
|------------------------------|-------------------------------|-------------------------------|------------------------------|-------------------------------|-------------------------------|
| 0.12 ~ 0.18                  | 10.16                         | 16.78                         | 0.18 ~ 0.22                  | 10.16                         | 27.41                         |
| 0.22 ~ 0.41                  | 10.16                         | 27.86                         | 0.41 ~ 0.84                  | 10.16                         | 31.17                         |
| 0.84 ~ 1.25                  | 10.16                         | 45.96                         | 1.25 ~ 2.29                  | 13.35                         | 45.96                         |
| 2.29 ~ 2.83                  | 13.35                         | 45.12                         | 2.83 ~ 2.88                  | 20.54                         | 45.12                         |
| 2.88 ~ 3.05                  | 20.54                         | 52.28                         | 3.05 ~ 3.53                  | 20.54                         | 68.13                         |
| 3.53 ~ 3.84                  | 20.54                         | 73.01                         | 3.84 ~ 3.88                  | 26.04                         | 73.01                         |
| 3.88 ~ 4.72                  | 42.85                         | 73.01                         | 4.72 ~ 4.82                  | 49.29                         | 73.01                         |
| 4.82 ~ 5.13                  | 49.29                         | 68.13                         | 5.13 ~ 5.32                  | 49.84                         | 68.13                         |
| 5.32 ~ 5.40                  | 49.84                         | 67.37                         | 5.40 ~ 5.41                  | 51.56                         | 67.37                         |
| 5.41 ~ 5.78                  | 51.56                         | 64.78                         | 6.09 ~ 6.15                  | 118.34                        | 124.46                        |
| 6.15 ~ 6.18                  | 113.71                        | 127.19                        | 6.18 ~ 6.36                  | 113.71                        | 132.25                        |
| 6.36 ~ 6.67                  | 113.71                        | 135.60                        | 6.67 ~ 6.99                  | 111.27                        | 135.60                        |
| 6.99 ~ 7.92                  | 111.27                        | 142.28                        | 7.92 ~ 8.12                  | 117.16                        | 142.28                        |
| 8.12 ~ 8.98                  | 132.80                        | 142.28                        |                              |                               |                               |

In 4D-NOESY spectrum, if a HC-NH NOE peak does not fall in this region, it's most likely that this peak is folded in  $^{13}\text{C}$  dimension. The program will try to “unfold” it into the statistic region, which means one or more sweep-width(s) may be automatically added to or subtracted from its  $^{13}\text{C}$  chemical shift.

The program is unable to correct those peaks folded in  $^1\text{H}_\text{C}$  dimension. Fortunately, this kind of peak is extremely rare.

In the step of inspecting CCH-TOCSY and 4D NOESY, the Z axis of the CCH-TOCSY should be set as the  $^{13}\text{C}$  dimension with a smaller sweep width, while the Z axis and A axis of the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectrum (or sub-spectrum) should be set as the  $^{15}\text{N}$  and direct  $^1\text{H}$  dimensions, respectively. This setting should be kept the same in the following steps.

## 7.2.6 Spin-system identification module

With the use of HNCA and MQ-CCH-TOCSY spectra, intra-residue and sequential HC-NH NOE peaks of each cluster can be separated out from other

inter-residue HC-NH NOE peaks observed in the 4D  $^{13}\text{C}$ ,  $^{15}\text{N}$ -edited NOESY spectrum or sub-spectrum. Meanwhile, by grouping the intra-residue and sequential HC-NH NOE peaks into two separate spin-systems, an intra-residue spin-system and a sequential spin-system can be established for each cluster. The amino acid type of both spin-systems can subsequently be determined based on the  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts.

The “Spin-system Identification Module” is designed for these tough jobs.

### 7.2.6.1 Methods

Even the most experienced spectroscopists may be unable to unambiguously identify two spin-systems for every cluster or classify them into intra-residue or sequential spin-system straightforward. For many clusters which simply don't have enough NOEs or have some disturbing inter-residue NOEs, spectroscopists need to combine all the information retrieved from subsequent studies to unravel which peak actually belongs to which spin-system. That's why the Spin-system Identification Module tries to create as many spin-systems as possible and collect as much information as it can for each cluster.

For each cluster generated from the cluster identification procedure (chapter 7.2.4), the program will first extract all the possible  $\text{H}_\alpha\text{C}_\alpha\text{-NH}$  NOEs from its NOE peak list by matching every NOE peak with its HNCA peaks with a tolerance equal to the spectral resolution. A spin-system will be created for each of the possible  $\text{H}_\alpha\text{C}_\alpha\text{-NH}$  NOEs which contains information about both HNCA peak and possible  $\text{H}_\alpha\text{C}_\alpha\text{-NH}$  NOE peak. If possible  $\text{H}_\alpha\text{C}_\alpha\text{-NH}$  NOE peak

could not be detected for a HNCA peak, an empty spin-system that only contains information about the HNCA peak will be created.

From CCH-TOCSY slices defined by the CH spin-pairs of individual  $H_C-N_H$  NOEs, the CCH peak pattern of each NOE peak could be extracted. By comparing them, those NOE peaks that give rise to similar CCH peak patterns in which one of the  $^{13}C$  match the  $C_\alpha$  of identified  $H_\alpha C_\alpha$ -NH NOE or HNCA peak (mentioned above for a given cluster) will be grouped into the spin-system defined by the HNCA peak. A confident level (from 0 to 10) will be given to each of these NOE peaks based on how well their CCH peak pattern matches that of the possible  $H_\alpha C_\alpha$ -NH NOE peak. Other than that, the confident level of a NOE peak may also be affected by its fellow-peaks in the same spin-system, and if several peaks have very similar CCH peak patterns, they will enhance each other's confident level, and the better their peak patterns match with each other, the more their confident level will increase.

The program could easily and unambiguously recognize most Glycine spin-systems. For other spin-systems, instead of typing them as a single or a group of amino acids, a score will be given for each possible amino acid, which means there are twenty scores for each spin-system delineating its possibility of being one of the twenty types. The scores are calculated mainly based on the  $^1H$  and  $^{13}C$  chemical shifts of the identified spin-system members (or NOEs), but the amount, intensity and pattern of its NOE peaks are also considered. For example, an Alanine spin-system should only have two peaks located around H:4.26ppm, C:53.17ppm ( $H_\alpha C_\alpha$ ) and H:1.36ppm, C:18.94ppm ( $H_\beta C_\beta$ ), and if a spin-system



has three NOE peaks or has a strong peak at an odd position of  $^1\text{H}= 3.12$ ,  $^{13}\text{C}=43.14$ , its score for being an Alanine spin-system will be very low.

In the end, all the information of the candidate-spin-systems will be recorded in the file: SpinSystemList.txt. The cluster list file (ClusterList.txt) will also be updated.

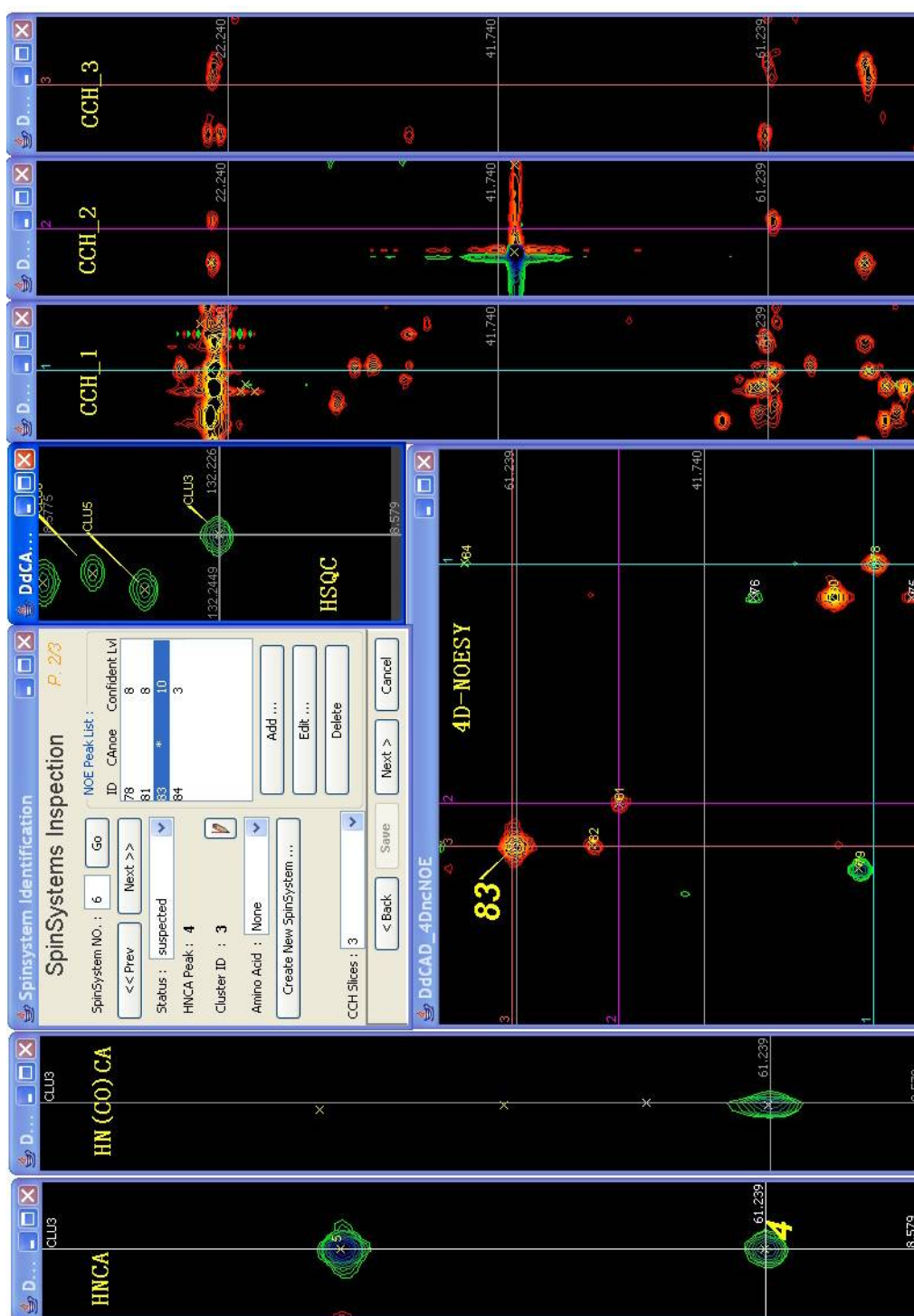
### 7.2.6.2 Interface

The main panel and the results panel of the Spin-system Identification Module are very similar to those of Cluster Identification Module (Chapter 7.2.4).

The “Spin-Systems Inspection Panel” (Figure 7.19) allows users to examine the newly created spin-systems one by one. The HSQC, HNCA, HN(CO)CA and 4D-NOESY spectra will be automatically synchronized and highlighted when users switch from one spin-system to another. Users may need to compare the CCH pattern of every NOE peak in a cluster. However, due to the limited size of the computer screen, it's unfeasible to display all the CCH slices. XYZ4D displays up to eight CCH windows, and every window is associated with a keyboard shortcut: one of the 1-8 number-keys. If users want to check the CCH pattern of a NOE peak, they could simply select the NOE peak and press one number-key, the associated window will immediately display the CCH slice of the selected peak and mark out the slice centre using a unique colour. Furthermore, users could show the CCH slice of a certain position in the NOESY spectrum by moving the crosshair to the position and pressing a shortcut without selecting any NOE peak. This feature could be very handy when some NOE peaks are overlapped or too weak to be picked. If users want to check a folded

position, holding Ctrl or Alt key when pressing the shortcut will display the CCH slice that is associated with the given position with one sweep-width in the  $^{13}\text{C}$  dimension added or subtracted.

At the beginning, all of the newly-created spin-systems are defined as “Suspected”, in order to encourage users to examine the spin-system one by one, and manually change their status to either “Confirmed” or “Corrupted”. Upon completion of this module, all corrupted spin-system will be removed, while suspected spin-systems are treated the same way as confirmed ones. Thus, unless there is a reminder that there are suspected spin-systems left, unconfirmed spin-system will not affect future work.



**Figure 7.19** The graphic interface of spin-system identification.

The HSQC, HNCA, HN(CO)CA and 4D-NOESY spectra are automatically synchronized by control panel. The 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> CCH-TOCSY window display the CCH pattern of HC-NH NOE peak 78, 81 and 83 respectively. NOE peak 83 is selected in the control panel, which cause it been highlighted in 4D-NOESY spectrum.

### 7.2.7 Cluster mapping module

With fairly complete clusters and spin-system information, it's time to assemble them into fragments and map the fragments onto protein sequence. The most intuitive approach is to compare every 2 clusters, generate dipeptide segment if they are consecutive, assemble the dipeptides into fragments and then uniquely map them to protein primary sequence.

Two consecutive clusters, A and B, should fulfill the following conditions:

1. The HNCA peak (i) of A and the HNCA peak (i-1) of B should have common  $^{13}\text{C}$  chemical shift.
2. The  $\text{C}_\alpha\text{H}_\alpha$ -NOE (i) of A and the  $\text{C}_\alpha\text{H}_\alpha$ -NOE (i-1) of B should have common  $^{13}\text{C}$  and  $^1\text{H}_\text{C}$  chemical shift.
3. The intra-residue spin-system of A and the sequential spin-system of B should match each other.
4. A and B probably have some common inter-residue NOEs.

If they are supposed to be mapped onto *a* and *b* position of the protein sequence, they need to fulfill the following conditions:

1. For cluster A, the amino acid type of its intra-residue spin-system should be consistent with position *a*.

2. For cluster B, the amino acid type of its sequential spin-system should be consistent with position *a*, and the amino acid type of its intra-residue spin-system should be consistent with position *b*.
3. The order of A and B should be consistent with the predicted order established based on HN(CO)CA spectrum.

XYZ4D calculates 7 scores ( $S_{c1}$ ,  $S_{c2}$ ,  $S_{c3}$ ,  $S_{c4}$ ,  $S_{m1}$ ,  $S_{m2}$  and  $S_o$ ) for each mapped cluster based on the above mentioned 7 conditions. An overall energy (E) of the cluster will subsequently be generated based on the scores. These scores and energy reflect both the signature information of cluster mapping and the adjacency information between clusters.

Under the semiautomatic mode, these scores and energy will appear as rainbow stripes making it visible to check which fragment are badly assembled or wrongly mapped onto the sequence. Under fully-automatic mode, energy will be used by a simulated annealing-Monte Carlo approach to achieve the best mapping result.

#### 7.2.7.1 Methods

If cluster A and B are consecutive and they have been mapped to positions *a* and *b* on the protein sequence, the scores and energy of cluster B are calculated as follows:

$S_{c1}$  (Connecting Score No.1 / HNCA matching score) indicates how well the HNCA peaks of cluster B and cluster A agree with each other. It's calculated by the following equation (7.2.7.1.1):

$$S_{c1} = B_{c1} + (100 - B_{c1}) \times \exp \left[ - \left( \frac{\omega_C^A - \omega_C^B}{2 \times \Delta^{13}C} \right)^2 \right] \quad (7.2.7.1.1)$$

where  $B_{c1}$  is a user-defined constant to differentiate “confirmed no-match” and “potential-match”,  $\omega_C^A$  is the  $^{13}C$  chemical shift value of HNCA peak of cluster A in an intra-residue spin-system,  $\omega_C^B$  is the  $^{13}C$  chemical shift value of a HNCA peak of cluster B in a sequential spin-system,  $\Delta^{13}C$  is the tolerance that equal to  $3 \times$  resolution of the HNCA spectrum on the  $^{13}C$  dimension.

$S_{c1} = 0$  if their HNCA peaks do not match.  $S_{c1} = B_{c1}$  if one or both HNCA peaks are missing.  $S_{c1}$  has the value 100 for a perfect match ( $\omega^A = \omega^B$ ).

$S_{c2}$  (Connecting Score No.2 /  $C_\alpha H_\alpha$ -NOE matching score) indicates how well the  $C_\alpha H_\alpha$ -NOE peaks of cluster B and cluster A agree with each other. It's calculated by equation 7.2.7.1.2:

$$S_{c2} = B_{c2} + (100 - B_{c2}) \times \exp \left[ - \left( \frac{|\omega_C^A - \omega_C^B| + |\omega_H^A - \omega_H^B|}{2 \times \Delta^{13}C \times \Delta^1H_C} \right)^2 \right] \quad (7.2.7.1.2)$$

where  $B_{c2}$  is a user-defined constant,  $\omega_C^A$  and  $\omega_H^A$  are the  $^{13}C$  and  $^1H_C$  chemical shift values of a  $C_\alpha H_\alpha$ -NOE peak of cluster A in the intra-residue spin-system,  $\omega_C^B$  and  $\omega_H^B$  are the  $^{13}C$  and  $^1H_C$  chemical shift values of a  $C_\alpha H_\alpha$ -NOE peak of cluster B in the sequential spin-system,  $\Delta^{13}C$  and  $\Delta^1H_C$  are tolerance values, which were set to the spectral resolutions of the 4D-NOESY spectrum in the  $^{13}C$  and indirect  $^1H$  dimensions, respectively.

$S_{c2} = 0$  if their  $C_{\alpha}H_{\alpha}$ -NOE peaks do not match.  $S_{c2}=B_{c2}$  if one or both  $C_{\alpha}H_{\alpha}$ -NOE peaks are missing.  $S_{c2}$  has the value 100 for a perfect match ( $\omega_{C}^B=\omega_{C}^A$ ,  $\omega_{H}^B=\omega_{H}^A$ ).

$S_{c3}$  (Connecting Score No.3 / spin-system matching score) indicates how well the sequential spin-system of cluster B matches the intra-residue spin-system of cluster A.

For each matched NOE peak-pair P, a contribution  $V_p$  (0~1) will be defined similarly to the fore-mentioned step and differentiated by considering the confidential level L of both peaks. Assume that there are  $n$  matched peak-pairs among the two spin-systems, the sum of their contributions divided by anticipated matched-peak-number N will be used to calculate the score  $S_{c3}$ . Anticipated matched-peak-number varies with different amino acid types, and a pair of Ala spin-systems should have at most two matched peak-pairs while a pair of Lys spin-systems could have nine. A user-defined constant  $B_{c3}$  will be given to  $S_{c3}$  if one or both spin-systems are “Empty Spin-system” caused by missed or overlapped  $C_{\alpha}H_{\alpha}$ -NOE peak.

$$V_p = \exp \left[ - \left( \frac{|\omega_C^{Ap} - \omega_C^{Bp}| + |\omega_H^{Ap} - \omega_H^{Bp}|}{2 \times {}^{13}C \times {}^1H_C} \right)^2 \right] \times \frac{L^{Ap} + L^{Bp}}{2 \times 10} \quad (7.2.7.1.3)$$

$$S_{c3} = B_{c3} + (100 - B_{c3}) \times \frac{\sum_{i=1}^n V_p^i}{N} \quad (7.2.7.1.4)$$

$S_{c4}$  (Connecting Score No.4 / cluster matching score) is complementary to  $S_{c3}$ . Besides those sequential and intra-residue NOEs that have been taken into account in  $S_{c3}$ , the matching peak-pairs of cluster A and B may come from their common inter-residue NOEs, which would contribute to  $S_{c4}$ .  $S_{c4}$  is positively related to the amount of matched peak-pairs between two clusters, until reaching the maximum value of 100.

$$S_{c4} = 20 \times \sum_{i=1}^n \exp \left[ - \left( \frac{|\omega_C^{Ai} - \omega_C^{Bi}| + |\omega_H^{Ai} - \omega_H^{Bi}|}{2 \times {}^{13}\text{C} \times {}^1\text{H}_C} \right)^2 \right] \quad (7.2.7.1.5)$$

$S_{m1}$  (Mapping Score No. 1 / sequential spin-system amino acid typing score) is calculated mainly based on how well the  ${}^1\text{H}$  and  ${}^{13}\text{C}$  chemical shifts of spin-system match the amino acid type of residue  $\mathbf{a}$ . The method has been briefed in section 7.2.6.1. A user-defined constant  $B_m$  will be given to  $S_{m1}$  if the spin-system is a “Empty Spin-system”.

$S_{m2}$  (Mapping Score No. 1 / intra-residue spin-system amino acid typing score) is similar to  $S_{m1}$ , but it indicates how well the intra-residue spin-system matches the amino acid type of residue  $\mathbf{b}$ .

$S_o$  (Ordering Score) indicates whether the mapping is consistent with the predicted order. A perfect cluster should have two HNCA peaks: HNCA peak (i) from current residue and HNCA peak (i-1) from the preceding residue. Using HN(CO)CA spectrum (Section 7.2.4) one can identify the sequential HNCA peak (i-1). After mapping a cluster onto the sequence, its HNCA peak (i-1) should be correlated with the sequential spin-system and should have a common  ${}^{13}\text{C}$  chemical shift with the HNCA peak (i) of the preceding cluster. In this case,  $S_o$



will be set to 100. A user-defined constant  $B_0$  will be given to  $S_0$  if the cluster only has one HNCA peak (cause by lacking or overlapping peaks). If the cluster has more than one HNCA peak and HNCA peak (i-1) hasn't been identified (missing HN(CO)CA peak),  $S_0$  will be calculated by the following equation:

$$S_0 = \begin{cases} 0 & (0 < |I_i|/|I_{i-1}| \leq 0.5) \\ B_0 \times (|I_i|/|I_{i-1}| - 0.5) \times 2 & (0.5 < |I_i|/|I_{i-1}| \leq 1) \\ B_0 + (100 - B_0) \times (|I_i|/|I_{i-1}| - 1) & (1 < |I_i|/|I_{i-1}| < 2) \\ 100 & (|I_i|/|I_{i-1}| \geq 2) \end{cases} \quad (7.2.7.1.6)$$

where  $I_i$  and  $I_{i-1}$  are the intensities of HNCA peak (i) and HNCA peak (i-1) respectively.

$E$  is the overall energy of a cluster. It is negatively related to the 7 scores mentioned above and ranges from 0 to 10000. The user-defined constants e.g.  $G_{c1}$ ,  $G_{m1}$ ,  $G_0$ ... are weight factors that control the contribution of each score to the cluster energy. After mapping as many clusters as possible onto the protein sequence, the total energy of the mapping scheme is the sum of energies of individual residues, while an empty residue that hasn't been assigned to any cluster has the maxim energy value of 10000.

$$E = 100 \times \frac{\sum G \times (100 - S)}{\sum G} \quad (G = G_{c1}, G_{c2}, G_{c3}, G_{c4}, G_{m1}, G_{m2}, G_0) \quad (S = S_{c1}, S_{c2}, S_{c3}, S_{c4}, S_{m1}, S_{m2}, S_0) \quad (7.2.7.1.7)$$

An automated simulated annealing-Monte Carlo approach tailored for obtaining the best mapping scheme with the above energy equation will be carried out if users choose fully-automatic mode for cluster mapping. Monte Carlo methods are particularly powerful in this particular application because they explore the landscape of possible solutions during the mapping process. Consequently, they are able to report both the most favorable set of mapping as well as an ensemble of mapping schemes that are closely related to the best one. This ensemble of mapping schemes can be inspected to detect possible errors in the previous steps or to identify weak links or wrong assignments.

However, a limitation of the simulated annealing-Monte Carlo approach is the slow convergence of the algorithm. In larger proteins, the solution space grows rapidly with the number of residues and cannot be searched extensively in practical time scales, unless additional constraints are used to reduce the search space.

XYZ4D provides users with an option to perform a “Best-first approach” before the simulated annealing-Monte Carlo approach. It will search the best matched clusters from cluster pool and generate dipeptide segments from them. As more dipeptide segments are generated, one or more reliable fragments will eventually be established and uniquely mapped to protein sequence. Fix these assignments before the simulated annealing-Monte Carlo approach could significantly speed up the procedure and provide better solutions.

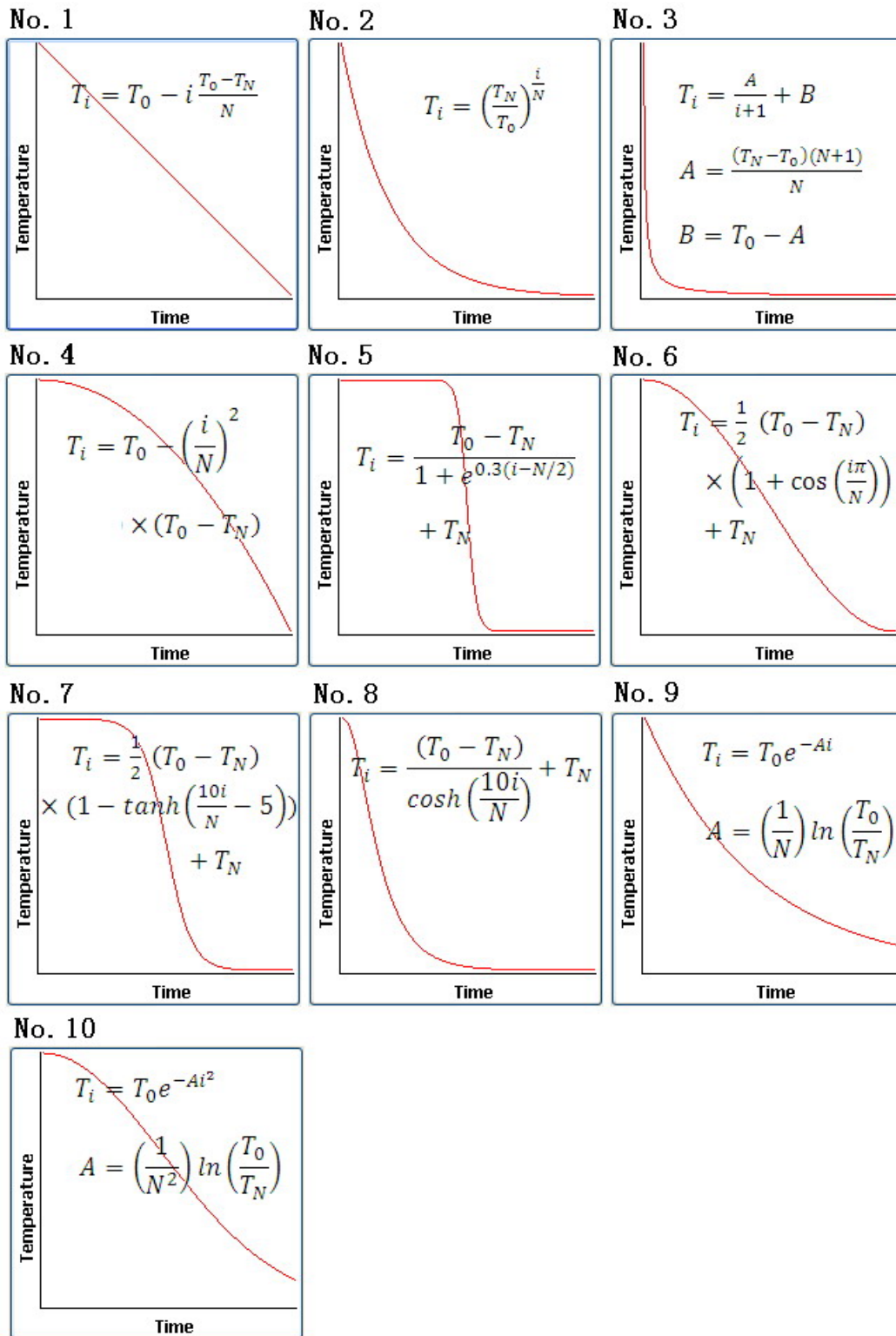


Figure 7.20 Ten simulated annealing cooling schedules provide by XYZ4D.

$T_i$  is the temperature for step  $i$ , where  $i$  increases from 0 to  $N$ .  $T_0$  is the initial temperature and  $T_N$  is the final temperature.

With or without the Best-first approach, the un-assigned clusters will be assigned randomly to the protein sequence. After calculating the energy of the initial random assignments, a cooling schedule (Figure 7.20) will be used to perform the simulated annealing-Monte Carlo approach. Each cooling schedule consists of a beginning temperature, a final temperature and a number of cooling steps.

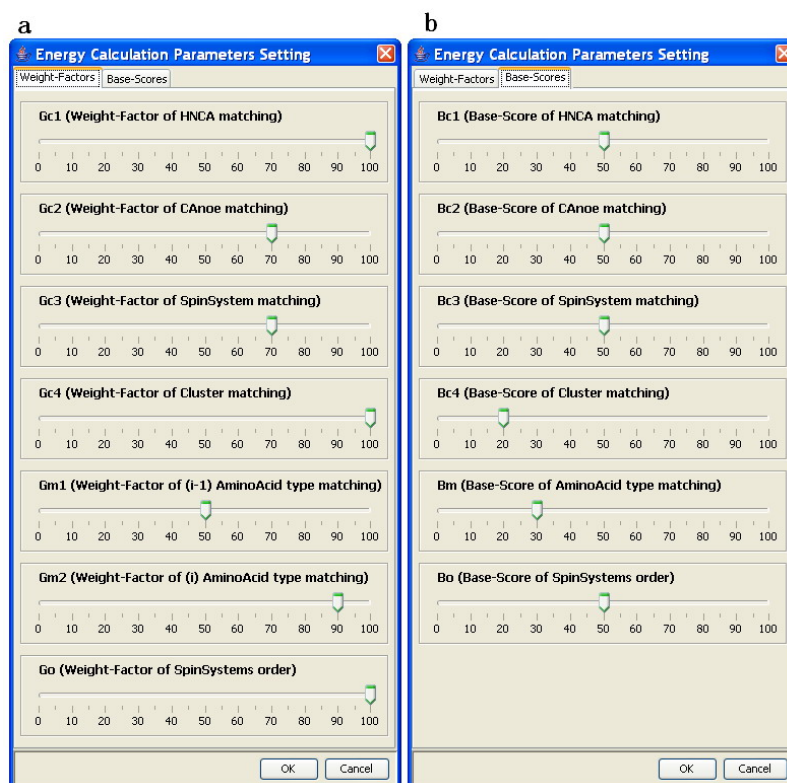
The program optimizes the assignment by exchanging, or swapping, one or more consecutive clusters from within the primary sequence with an identically sized collection of clusters from either the primary sequence or from the cache. This operation is called “perturbation”, and the segments are selected randomly from within the primary sequence and are of random length, with the maximum size of 5. The energy of this new mapping result is then calculated. If the energy is lowered, then the new mapping is retained. If the score rose, then a decision is made to either keep or discard the new mapping. This decision is based on the ratio of the increase in energy to the current temperature of the system. If the increase in the energy is equal to the current temperature then, on average,  $1/e$  ( $e$  is Euler–Mascheroni constant) of the solutions are retained. If the increase in energy is smaller than the current temperature then the probability of retaining the new mappings is larger than  $1/e$ . An increase in the energy that exceeds the current temperature causes the probability to be less than  $1/e$ . The larger the increase the lower the probability.

The temperature is initially set sufficiently high that most proposed perturbations are accepted. The temperature is gradually lowered during the run; consequently it becomes increasingly less likely to accept swaps that increase the

overall energy. To insure that the system remains in equilibrium during the annealing process it is necessary to use a large number of perturbations at each cooling step. In practice, the number of perturbations within each cooling step is defined by the user or equal to the square of the “empty residue” number, which is bigger.

### 7.2.7.2 Interface

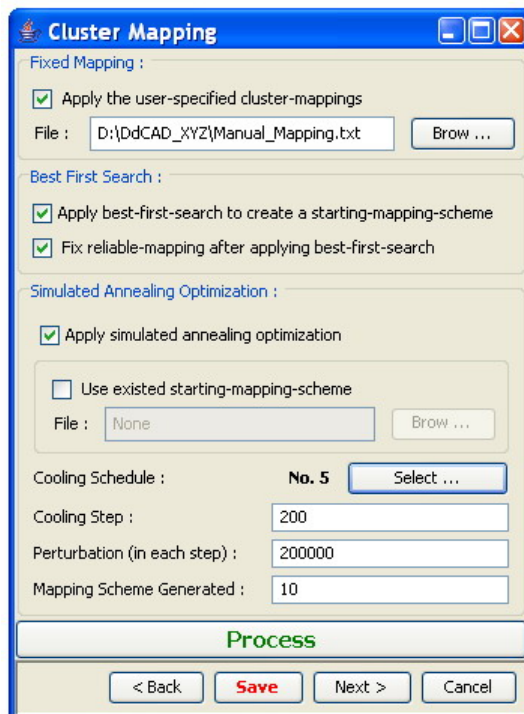
The main panel of the Cluster Mapping Module is slightly different from other modules’ with an additional “Parameters Setting” button which allows users to define the 6 baselines ( $B_{c1}$ ,  $B_{c2}$ ,  $B_{c3}$ ,  $B_{c4}$ ,  $B_m$ ,  $B_o$ ) and 7 weighting factors ( $G_{c1}$ ,  $G_{c2}$ ,  $G_{c3}$ ,  $G_{c4}$ ,  $G_{m1}$ ,  $G_{m2}$ ,  $G_o$ ) of the scores. (Figure 7.21)



**Figure 7.21** Setting Panels of Energy Calculation Parameters.

(a) Baselines setting panel. (b) Weighting factors setting panel.

If the fully-automatic mode is chosen, the program will ask the user to define the cooling schedule, cooling steps, number of perturbations within each cooling step and how many mapping schemes that would be generated (Figure 7.22). The annealing process will run concurrently on multiprocessor computer. The results and a statistics-analysis report will be saved as a text file in a user defined folder. The final mapping schemes could be exported into the semi-automatic interface for manual analysis.



**Figure 7.22 Control panel of Simulated Annealing-Monte Carlo approach.**

In the semi-automatic mode, XYZ4D provides an efficient and flexible interface to help users quickly and intuitively assemble the cluster (Figure 7.23). Starting from any cluster, users can search its following or preceding cluster, and XYZ4D lists all the candidate-clusters and their corresponding spin-systems ordered by how well they match with the given one, using yellow colour to

indicate a cluster that has been connected to other clusters or red colour to indicate a connected and locked cluster.

Choosing a candidate-cluster, the program will display HSQC peaks, HNCA strips, HN(CO)CA strips and 4DNOE slices of the two clusters side by side with an additional double-layer NOE slice shows the contour plot of NOE signals from both clusters in different colours and allows users to inspect the matching of spin-system intuitively. By moving the crosshair in the spectra with the mouse, users could easily check the matching of peaks. By selecting a peak ID in any of the panels, the peak will immediately be highlighted in the corresponding spectrum.

Double-clicking the candidate-cluster will connect it to the current one; users could then use the “Lock” button to lock this connection so that no other cluster could be connected to it until the connection is unlocked manually. Using the “Jump” button located beside the “Lock” and “Search” button, the interface will jump to the following or preceding cluster, allowing users to repeat the above routine until the fragment cannot be extended any longer. Then, users can apply the routine on another cluster forward or backward to assemble another fragment.

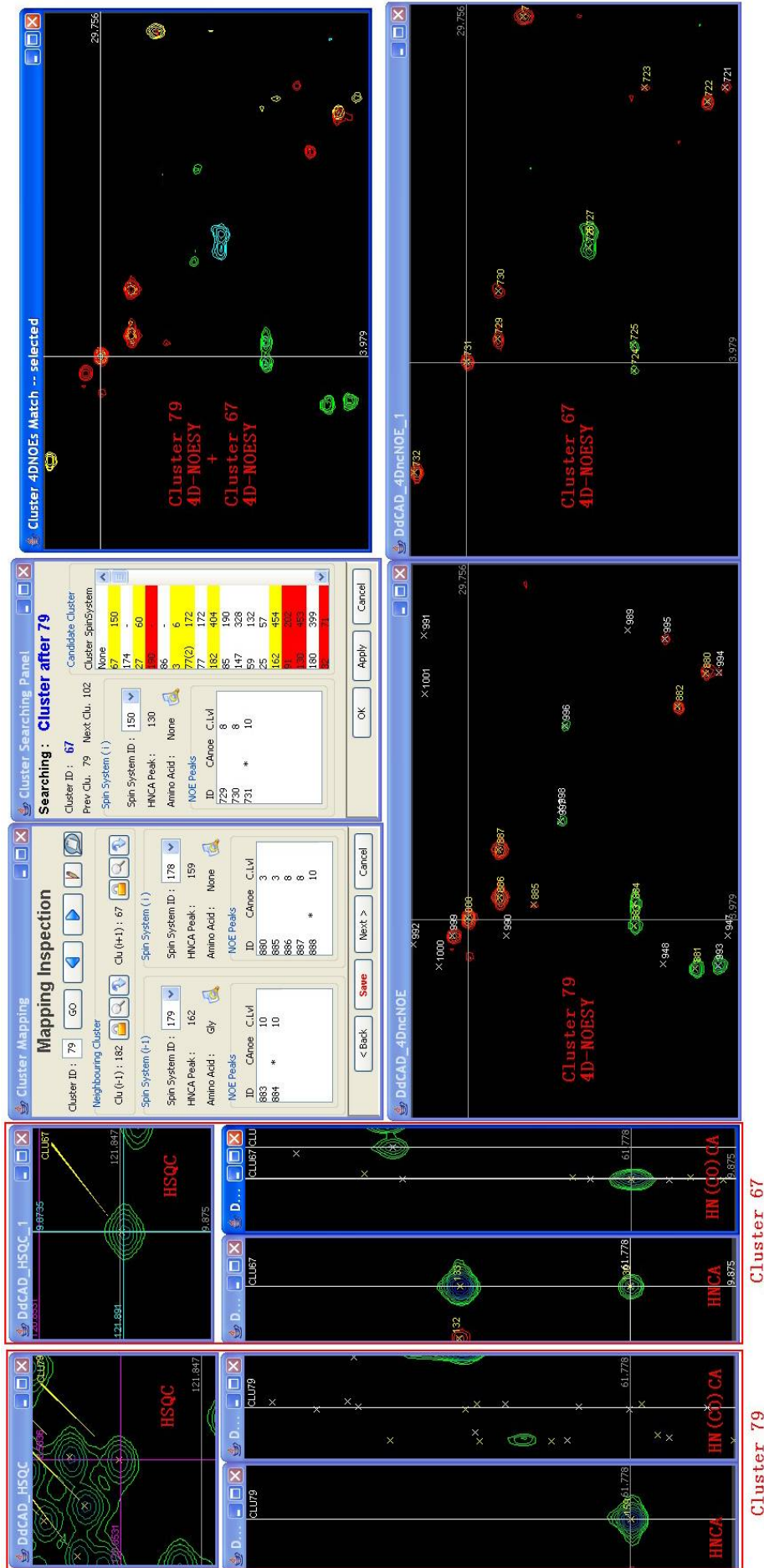
The fragments are delineated vividly as a bunch of “Cluster Cards” on the “Protein Sequence Map” (Figure 7.24). A normal cluster is correlated with a single cluster card while an overlapped cluster is correlated several. The card labelled with the ID of the cluster and its sequential spin-system (low left) and intra-residue spin-system (low right). If the amino acid type of a spin-system is characterized, it will appear right below the spin-system’s ID. Middle click a

cluster card can display the cluster in the main panel (Figure 7.23), while middle click a spin-system will pop up a table shows its 20 amino acid typing score. Cluster cards and fragments could be dragged to anywhere within the map, or placed into sequence. By placing two cards or fragments side by side, a connection will be generated automatically and indicated by a yellow light at the edge of the cards. Middle clicking the yellow light can break the connection, but if the light turns into red which means the connection is locked, it will not be broken by middle clicking.

If the spin system types are well characterized, then a fragment of four or five clusters usually is sufficient to achieve sequence-specific assignment. The ambiguity in the assignment process can be reduced by weak connections between fragments and the identification of other medium-range NOEs. The assignments encompass all spin-systems and clusters, and self-consistency is the best measure of the validity of the results.

After connecting clusters together or placing them into the sequence map, scores and a mapping energy that are described in section 7.2.7.1 will be calculated immediately and displayed as rainbow stripe above and below the cluster card, respectively. A high score means the connection or mapping is good, and the vertical rainbow stripes above the cluster card are shorter. A better mapping gives rise to lower energy, with the horizontal rainbow stripe below the sequence slot shows less red colour. The accurate score and energy value could be easily accessed by right click menu. Users can use the rainbow stripes to tell which fragment is badly assembled or wrongly mapped onto the sequence.





**Figure 7.23** Graphic interfaces for cluster mapping.

Cluster 79 is chosen in the main control panel. Its HSQC peaks, HNCA strips, HN(CO)CA strips and 4DNOE slices are automatically displayed one set of windows. The program searched Cluster 79's following cluster and listed the candidate-clusters in another panel. By selecting cluster 67, its HSQC peaks, HNCA strips, HN(CO)CA strips and 4DNOE slices are automatically displayed in the second set of windows. A double-layer NOE slice shows the contour plot of NOE signals from both clusters.

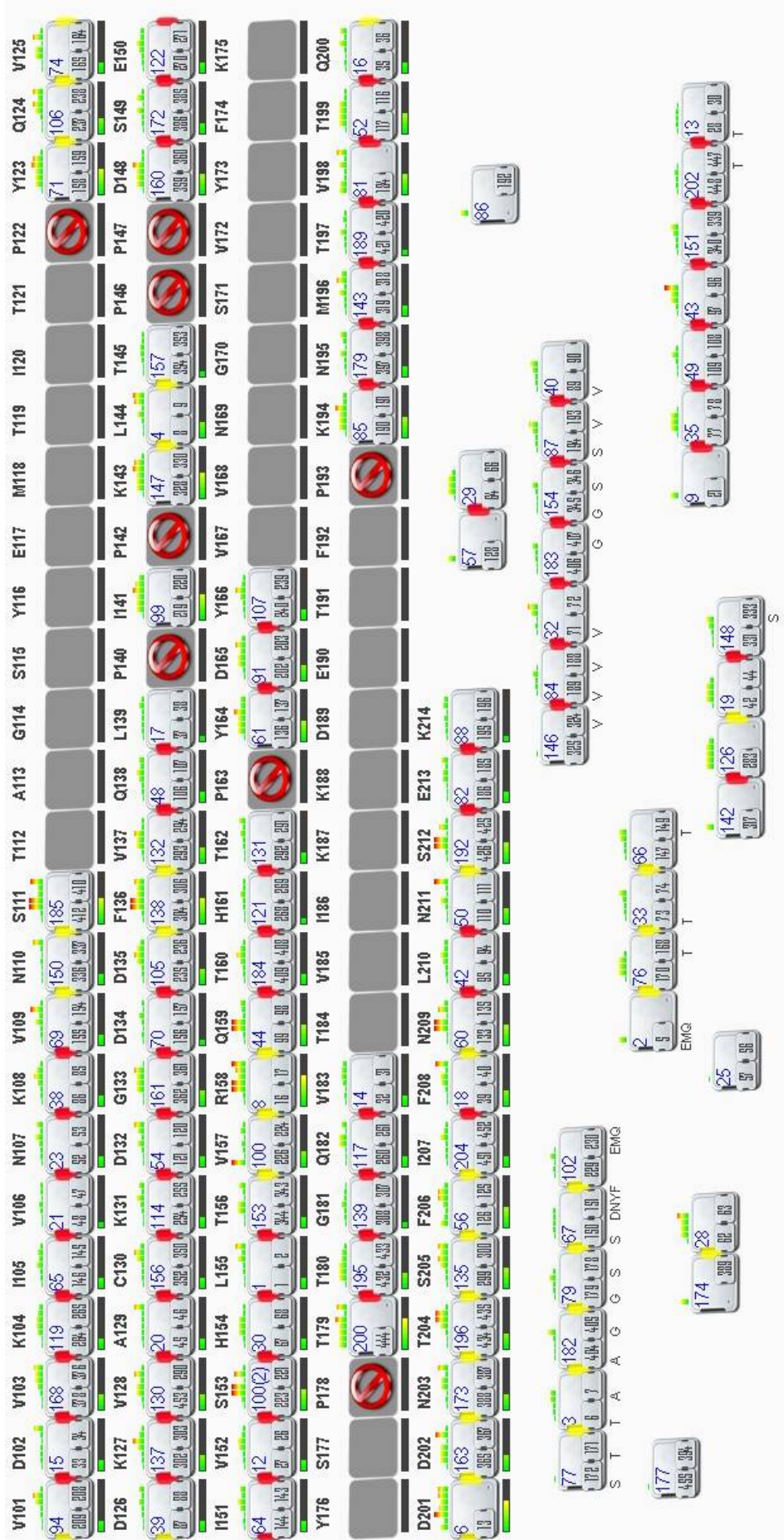
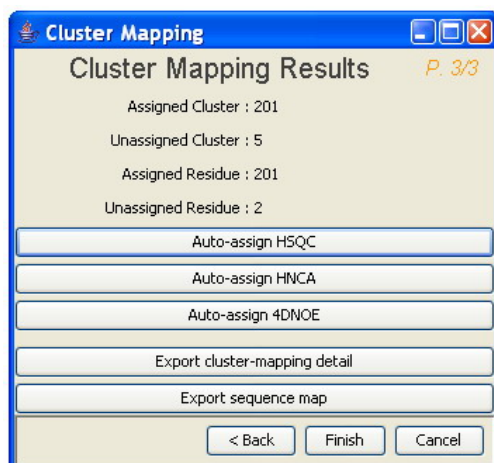


Figure 7.24 Protein Sequence Mapping.

The “Protein Sequence Map” can be saved as image or text file using right click menu at any time. The final mapping scheme could be used to assign the HSQC, HNCA and 4DNOE spectra. (Figure 7.25)



**Figure 7.25** The panel of cluster mapping module.

## 7.2.8 Backbone assignment module

After achieving a well optimized mapping scheme, the chemical shift values of most backbone atoms (HN, N, C $\alpha$ , H $\alpha$ ) could be easily established and filled into the atom list provided by NMRspy. Only based on this atom list, users could continue working on side chain assignment, NOE assignment and structure calculation.

XYZ4D establishes the  $^1\text{H}_\text{N}$ ,  $^{15}\text{N}$  chemical shifts of a residue by averaging the corresponding chemical shift values of the HSQC, HNCA and NOE peaks in the assigned cluster. The C $\alpha$  chemical shift of a residue is established using the HNCA peak (i) as reference, and H $\alpha$  chemical shift using the C $\alpha$ H $\alpha$  NOE peak within the intra-residue spin-system as reference.

A graphic interface (Figure 7.26) allows users to display the related spectrum region and highlight a peak (by selecting a cluster or peak in the panel) or change the chemical shift reference (by double clicking a cluster or peak in the panel). The references of different atoms are highlighted with yellow colour.

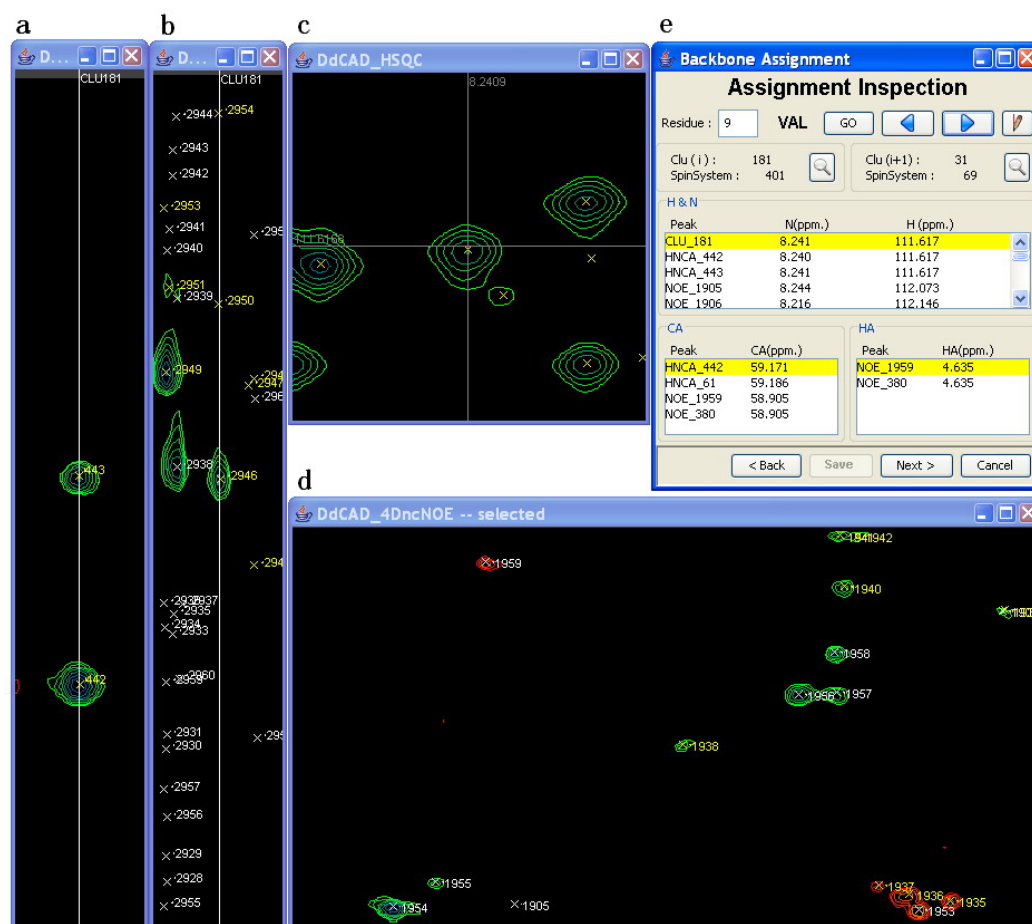


Figure 7.26 Graphic interface of Backbone Assignment Module.

### 7.3 Results and discussion

In this study, we have developed an NMRspy extension, XYZ4D, to facilitate the backbone assignment of large proteins without deuterium and specific labeling proteins by adopting a robust assignment strategy which makes

use of 2D-TROSY-HSQC, 3D-TROSY-HNCA, 3D-TROSY-HN(CO)CA, 3D-MQ-CCH-TOCSY and 4D- $^{13}\text{C}$ , and  $^{15}\text{N}$ -edited NOESY spectra.

The benefits of using XYZ4D are twofold. Firstly, the whole assignment process is greatly accelerated and alleviated due to computer automation. Secondly, the user is freed from the tedious routine calculation and spectra handling, and will focus only on resolving errors and ambiguities. This, coupled with the handy features of multiple spectrum view, uncluttered but powerful control panels and effective sequence mapping, will improve the accuracy of the assignments.

## References

- Barrick, D., Lukin, J. A., Simplaceanu, V., and Ho, C., Nuclear magnetic resonance spectroscopy in the study of hemoglobin cooperativity. *Methods Enzymol* **379**, 28 (2004).
- Bartels, Christian et al., The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR* **6** (1), 1 (1995).
- Bax, A., Clore, G.M., and Gronenborn, A.M., H-1-H-1 CORRELATION VIA ISOTROPIC MIXING OF C-13 MAGNETIZATION, A NEW 3-DIMENSIONAL APPROACH FOR ASSIGNING H-1 AND C-13 SPECTRA OF C-13-ENRICHED PROTEINS. *J. Magn. Reson.* **88** (2), 425 (1990).
- Berman, H. M. et al., The Protein Data Bank. *Nucleic Acids Res* **28** (1), 235 (2000).
- Billeter, M., Braun, W., and Wuthrich, K., Sequential resonance assignments in protein <sup>1</sup>H nuclear magnetic resonance spectra. Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *J Mol Biol* **155** (3), 321 (1982).
- Boomershine, W. P. et al., Structure of Mth11/Mth Rpp29, an essential protein subunit of archaeal and eukaryotic RNase P. *Proc Natl Acad Sci U S A* **100** (26), 15398 (2003).
- Brunger, A. T. et al., Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54** (Pt 5), 905 (1998).
- Choy, W. Y., Shortle, D., and Kay, L. E., Side chain dynamics in unfolded protein states: an NMR based <sup>2</sup>H spin relaxation study of delta131delta. *J Am Chem Soc* **125** (7), 1748 (2003).
- Clore, G. M. and Bewley, C. A., Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings. *J Magn Reson* **154** (2), 329 (2002).
- Clore, G. M., Kay, L. E., Bax, A., and Gronenborn, A. M., Four-dimensional <sup>13</sup>C/<sup>13</sup>C-edited nuclear Overhauser enhancement spectroscopy of a protein in solution: application to interleukin 1 beta. *Biochemistry* **30** (1), 12 (1991).

- Coggins, B. E., Venters, R. A., and Zhou, P., Filtered backprojection for the reconstruction of a high-resolution (4,2)D CH<sub>3</sub>-NH NOESY spectrum on a 29 kDa protein. *J Am Chem Soc* **127** (33), 11562 (2005).
- Cordier, F., Rogowski, M., Grzesiek, S., and Bax, A., Observation of through-hydrogen-bond 2hJHC' in a perdeuterated protein. *J Magn Reson* **140** (2), 510 (1999).
- Cornilescu, G., Delaglio, F., and Bax, A., Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* **13** (3), 289 (1999).
- Cowburn, D. et al., Segmental isotopic labeling for structural biological applications of NMR. *Methods Mol Biol* **278**, 47 (2004).
- Delaglio, F. et al., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6** (3), 277 (1995).
- Delaglio, F., Kontaxis, G., and Bax, A., Protein Structure Determination Using Molecular Fragment Replacement and NMR Dipolar Couplings. *J. Am. Chem. Soc.* **122** (9), 2142 (2000).
- Dickerson, Richard Earl and Geis, Irving, *Hemoglobin : structure, function, evolution, and pathology*. (Benjamin/Cummings Pub. Co, Menlo Park, Calif. ; London, 1983).
- Dunker, A. K. and Obradovic, Z., The protein trinity--linking function and disorder. *Nat Biotechnol* **19** (9), 805 (2001).
- Eghbalnia, Hamid R. et al., Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. *Journal of Biomolecular NMR* **32** (1), 71 (2005).
- Eletsky, A., Moreira, O., Kovacs, H., and Pervushin, K., A novel strategy for the assignment of side-chain resonances in completely deuterated large proteins using <sup>13</sup>C spectroscopy. *J Biomol NMR* **26** (2), 167 (2003).
- Fan, D., Zheng, Y., Yang, D., and Wang, J., NMR solution structure and dynamics of an exchangeable apolipoprotein, *Locusta migratoria* apolipoprotein III. *J Biol Chem* **278** (23), 21212 (2003).
- Fernandez, C. and Wider, G., TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr Opin Struct Biol* **13** (5), 570 (2003).
- Fesik, Stephen W. et al., 2D and 3D NMR spectroscopy employing carbon-13/carbon-13 magnetization transfer by isotropic mixing. Spin system identification in large proteins. *J. Am. Chem. Soc.* **112** (2), 886 (1990).

- Fischer, M. W. F., Zeng, L., and Zuiderweg, E. R. P., Use of  $^{13}\text{C}$ - $^{13}\text{C}$  NOE for the Assignment of NMR Lines of Larger Labeled Proteins at Larger Magnetic Fields. *J. Am. Chem. Soc.* **118** (49), 12457 (1996).
- Foster, M. P., McElroy, C. A., and Amero, C. D., Solution NMR of large molecules and assemblies. *Biochemistry* **46** (2), 331 (2007).
- Gardner, K. H. and Kay, L. E., The use of  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct* **27**, 357 (1998).
- Gardner, Kevin H., Konrat, Robert, Rosen, Michael K., and Kay, Lewis E., An (H)C(CO)NH-TOCSY pulse scheme for sequential assignment of protonated methyl groups in otherwise deuterated  $^{15}\text{N}$ ,  $^{13}\text{C}$ -labeled proteins. *Journal of Biomolecular NMR* **8** (3), 351 (1996).
- Gardner, K. H., Rosen, M. K., and Kay, L. E., Global Folds of Highly Deuterated, Methyl-Protonated Proteins by Multidimensional NMR. *Biochemistry* **36** (6), 1389 (1997).
- Gardner, K. H., Zhang, X., Gehring, K., and Kay, L. E., Solution NMR Studies of a 42 KDa Escherichia Coli Maltose Binding Protein-Cyclodextrin Complex: Chemical Shift Assignments and Analysis. *J. Am. Chem. Soc.* **120** (45), 11738 (1998).
- Giesen, A. W., Homans, S. W., and Brown, J. M., Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. *J Biomol NMR* **25** (1), 63 (2003).
- Gross, J. D., Gelev, V. M., and Wagner, G., A sensitive and robust method for obtaining intermolecular NOEs between side chains in large protein complexes. *J Biomol NMR* **25** (3), 235 (2003).
- Grzesiek, Stephan, Anglister, Jacob, Ren, Hao, and Bax, Ad, Carbon-13 line narrowing by deuterium decoupling in deuterium/carbon-13/nitrogen-15 enriched proteins. Application to triple resonance 4D J connectivity of sequential amides. *J. Am. Chem. Soc.* **115** (10), 4369 (1993).
- Grzesiek, Stephan, Kuboniwa, Hitoshi, Hinck, Andrew P., and Bax, Ad, Multiple-Quantum Line Narrowing for Measurement of H.alpha.-H.beta. J Couplings in Isotopically Enriched Proteins. *J. Am. Chem. Soc.* **117** (19), 5312 (1995).
- Grzesiek, Stephan et al., Four-Dimensional  $^{15}\text{N}$ -Separated NOESY of Slowly Tumbling Perdeuterated  $^{15}\text{N}$ -Enriched Proteins. Application to HIV-1 Nef. *J. Am. Chem. Soc.* **117** (37), 9594 (1995).
- Gschwind, Ruth M., Gemmecker, Gerd, and Kessler, Horst, A Spin System Labeled and Highly Resolved ed-H(CCO)NH-TOCSY Experiment for



- the Facilitated Assignment of Proton Side Chains in Partially Deuterated Samples. *Journal of Biomolecular NMR* **11** (2), 191 (1998).
- Guntert, P., Automated NMR structure calculation with CYANA. *Methods Mol Biol* **278**, 353 (2004).
- Habeck, M., Rieping, W., Linge, J. P., and Nilges, M., NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol Biol* **278**, 379 (2004).
- Hajduk, P. J. et al., NMR-Based Screening of Proteins Containing  $^{13}\text{C}$ -Labeled Methyl Groups. *J. Am. Chem. Soc.* **122** (33), 7898 (2000).
- Henry, G. D., Weiner, J. H., and Sykes, B. D., Backbone dynamics of a model membrane protein:  $^{13}\text{C}$  NMR spectroscopy of alanine methyl groups in detergent-solubilized M13 coat protein. *Biochemistry* **25** (3), 590 (1986).
- Herrmann, T., Guntert, P., and Wuthrich, K., Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* **319** (1), 209 (2002).
- Herrmann, T., Guntert, P., and Wuthrich, K., Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* **24** (3), 171 (2002).
- Hilty, Christian, Fernández, César, Wider, Gerhard, and Wüthrich, Kurt, Side chain NMR assignments in the membrane protein OmpX reconstituted in DHPC micelles. *Journal of Biomolecular NMR* **23** (4), 289 (2002).
- Hobohm, U., Scharf, M., Schneider, R., and Sander, C., Selection of representative protein data sets. *Protein Sci* **1** (3), 409 (1992).
- Hua, Q. et al., A thermophilic mini-chaperonin contains a conserved polypeptide-binding surface: combined crystallographic and NMR studies of the GroEL apical domain with implications for substrate interactions. *J Mol Biol* **306** (3), 513 (2001).
- Hung, Ling-Hong and Samudrala, Ram, Accurate and automated classification of protein secondary structure with PsiCSI. *Protein Sci* **12** (2), 288 (2003).
- Ikura, M., Kay, L. E., and Bax, A., A novel approach for sequential assignment of  $^1\text{H}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$  spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **29** (19), 4659 (1990).
- Ishima, R., Louis, J. M., and Torchia, D. A., Transverse  $^{13}\text{C}$  Relaxation of CHD2 Methyl Isotopomers To Detect Slow Conformational Changes of Protein Side Chains. *J. Am. Chem. Soc.* **121** (49), 11589 (1999).

- Iwadate, Mitsuo, Asakura, Tetsuo, and Williamson, Michael P.,  $C_{\alpha}$  and  $C_{\beta}$  Carbon-13 Chemical Shifts in Proteins From an Empirical Database. *Journal of Biomolecular NMR* **13** (3), 199 (1999).
- Janin, J., Miller, S., and Chothia, C., Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* **204** (1), 155 (1988).
- Johnson, Bruce A. and Blevins, Richard A., NMR View: A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR* **4** (5), 603 (1994).
- Kainosho, M., Isotope labelling of macromolecules for structural determinations. *Nat Struct Biol* **4 Suppl**, 858 (1997).
- Kainosho, M. et al., Optimal isotope labelling for NMR protein structure determinations. *Nature* **440** (7080), 52 (2006).
- Kay, L. E., Torchia, D. A., and Bax, A., Backbone dynamics of proteins as studied by  $^{15}\text{N}$  inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28** (23), 8972 (1989).
- Koradi, R., Billeter, M., and Wuthrich, K., MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14** (1), 51 (1996).
- Kraulis, P. J. et al., Solution structure and dynamics of ras p21.GDP determined by heteronuclear three- and four-dimensional NMR spectroscopy. *Biochemistry* **33** (12), 3515 (1994).
- Laskowski, R. A. et al., AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8** (4), 477 (1996).
- Le, Hongbiao and Oldfield, Eric, Correlation between  $^{15}\text{N}$  NMR chemical shifts in proteins and secondary structure. *Journal of Biomolecular NMR* **4** (3), 341 (1994).
- Lee, A. L. and Wand, A. J., Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature* **411** (6836), 501 (2001).
- LeMaster, D. M., Deuterium labelling in NMR structural analysis of larger proteins. *Q Rev Biophys* **23** (2), 133 (1990).
- LeMaster, D. M. and Richards, F. M., NMR sequential assignment of Escherichia coli thioredoxin utilizing random fractional deuteration. *Biochemistry* **27** (1), 142 (1988).
- Lin, Y. et al., Solution structure of the catalytic domain of GCN5 histone acetyltransferase bound to coenzyme A. *Nature* **400** (6739), 86 (1999).

- Lin, Z., Huang, H., Siu, C. H., and Yang, D.,  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$  resonance assignments of  $\text{Ca}^{2+}$ -free DdCAD-1: a  $\text{Ca}^{2+}$ -dependent cell-cell adhesion molecule. *J Biomol NMR* **30** (3), 375 (2004).
- Lin, Z., Xu, Y., Yang, S., and Yang, D., Sequence-specific assignment of aromatic resonances of uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled proteins by using  $^{13}\text{C}$ - and  $^{15}\text{N}$ -edited NOESY spectra. *Angew Chem Int Ed Engl* **45** (12), 1960 (2006).
- Linge, J. P., O'Donoghue, S. I., and Nilges, M., Automated assignment of ambiguous nuclear overhauser effects with ARIA. *Methods Enzymol* **339**, 71 (2001).
- Liu, Dingjiang et al., Letter to the Editor: Backbone  $^1\text{H}$ ,  $^{15}\text{N}$  and  $^{13}\text{C}$  resonance assignments of the Staphylococcus aureus acyl carrier protein synthase (AcpS). *Journal of Biomolecular NMR* **24** (3), 273 (2002).
- Liu, H. L. and Hsu, J. P., Recent developments in structural proteomics for protein structure determination. *Proteomics* **5** (8), 2056 (2005).
- Liu, W., Zheng, Y., Cistola, D. P., and Yang, D., Measurement of methyl  $^{13}\text{C}$ - $^1\text{H}$  cross-correlation in uniformly  $^{13}\text{C}$ -,  $^{15}\text{N}$ -, labeled proteins. *J Biomol NMR* **27** (4), 351 (2003).
- Logan, T. M., Olejniczak, E. T., Xu, R. X., and Fesik, S. W., A general method for assigning NMR spectra of denatured proteins using 3D HC(CO)NH-TOCSY triple resonance experiments. *J Biomol NMR* **3** (2), 225 (1993).
- Lopez-Mendez, B. and Guntert, P., Automated Protein Structure Determination from NMR Spectra. *J. Am. Chem. Soc.* **128** (40), 13112 (2006).
- Luginbuhl, Peter, Szyperski, Thomas, and Wuthrich, Kurt, Statistical Basis for the Use of  $^{13}\text{C}$ [ $\alpha$ ]Chemical Shifts in Protein Structure Determination. *Journal of Magnetic Resonance, Series B* **109** (2), 229 (1995).
- Lukin, J. A. and Ho, C., The structure--function relationship of hemoglobin in solution at atomic resolution. *Chem Rev* **104** (3), 1219 (2004).
- Lukin, J. A. et al., Quaternary structure of hemoglobin in solution. *Proc Natl Acad Sci U S A* **100** (2), 517 (2003).
- Lukin, J. A. et al., Backbone resonance assignments of human adult hemoglobin in the carbonmonoxy form. *J Biomol NMR* **28** (2), 203 (2004).
- Markley, J. L. et al., Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. *J Biomol NMR* **12** (1), 1 (1998).

- McElroy, C. et al., TROSY-NMR studies of the 91kDa TRAP protein reveal allosteric control of a gene regulatory protein by ligand-altered flexibility. *J Mol Biol* **323** (3), 463 (2002).
- Metzler, W. J. et al., The three-dimensional solution structure of the SH2 domain from p55blk kinase. *Biochemistry* **35** (20), 6201 (1996).
- Mittermaier, A. and Kay, L. E., New tools provide new insights in NMR studies of protein dynamics. *Science* **312** (5771), 224 (2006).
- Montelione, Gaetano T., Lyons, Barbara A., Emerson, S. Donald, and Tashiro, Mitsuru, An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. *J. Am. Chem. Soc.* **114** (27), 10974 (1992).
- Mueller, G. A. et al., Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *J Mol Biol* **300** (1), 197 (2000).
- Mueser, T. C., Rogers, P. H., and Arnone, A., Interface sliding as illustrated by the multiple quaternary structures of liganded hemoglobin. *Biochemistry* **39** (50), 15353 (2000).
- Muhandiram, D. R., Yamazaki, Toshio, Sykes, Brian D., and Kay, Lewis E., Measurement of  $2H$  T1 and T1. $\rho$ . Relaxation Times in Uniformly  $^{13}C$ -Labeled and Fractionally  $2H$ -Labeled Proteins in Solution. *J. Am. Chem. Soc.* **117** (46), 11536 (1995).
- Mulder, F. A. et al., Studying excited states of proteins by NMR spectroscopy. *Nat Struct Biol* **8** (11), 932 (2001).
- Nicholson, L. K. et al., Dynamics of methyl groups in proteins as studied by proton-detected  $^{13}C$  NMR spectroscopy. Application to the leucine residues of staphylococcal nuclease. *Biochemistry* **31** (23), 5253 (1992).
- Nietlispach, D. et al., An Approach to the Structure Determination of Larger Proteins Using Triple Resonance NMR Experiments in Conjunction with Random Fractional Deuteration. *J. Am. Chem. Soc.* **118** (2), 407 (1996).
- Palmer, A. G., 3rd, Kroenke, C. D., and Loria, J. P., Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol* **339**, 204 (2001).
- Pastore, A. and Saudek, V., The relationship between chemical shift and secondary structure in proteins. *J Magn Reson* **90**, 165 (1990).
- Permi, P., Tossavainen, H., and Hellman, M., Efficient assignment of methyl resonances: enhanced sensitivity by gradient selection in a DE-MQ-(H)CC(m)Ht (m)-TOCSY experiment. *J Biomol NMR* **30** (3), 275 (2004).

- Pervushin, K., Riek, R., Wider, G., and 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* **94** (23), 12366 (1997).
- Post, C. B., Exchange-transferred NOE spectroscopy and bound ligand structure determination. *Curr Opin Struct Biol* **13** (5), 581 (2003).
- Rosen, M. K. et al., Selective methyl group protonation of perdeuterated proteins. *J Mol Biol* **263** (5), 627 (1996).
- Salzmann, M. et al., NMR Assignment and Secondary Structure Determination of an Octameric 110 kDa Protein Using TROSY in Triple Resonance Experiments. *J. Am. Chem. Soc.* **122** (31), 7543 (2000).
- Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M., The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **160** (1), 65 (2003).
- Selenko, P. and Wagner, G., Looking into live cells with in-cell NMR spectroscopy. *J Struct Biol* **158** (2), 244 (2007).
- Serber, Z. et al., Methyl groups as probes for proteins and complexes in in-cell NMR experiments. *J Am Chem Soc* **126** (22), 7119 (2004).
- Shang, Z., Swapna, G. V. T., Rios, C. B., and Montelione, G. T., Sensitivity Enhancement of Triple-Resonance Protein NMR Spectra by Proton Evolution of Multiple-Quantum Coherences Using a Simultaneous  $^1\text{H}$  and  $^{13}\text{C}$  Constant-Time Evolution Period. *J. Am. Chem. Soc.* **119** (39), 9274 (1997).
- Sharff, A. J., Rodseth, L. E., and Quioco, F. A., Refined 1.8-Å structure reveals the mode of binding of beta-cyclodextrin to the maltodextrin binding protein. *Biochemistry* **32** (40), 10553 (1993).
- Silva, M. M., Rogers, P. H., and Arnone, A., A third quaternary structure of human hemoglobin A at 1.7-Å resolution. *J Biol Chem* **267** (24), 17248 (1992).
- Simplaceanu, V. et al., Chain-selective isotopic labeling for NMR studies of large multimeric proteins: application to hemoglobin. *Biophys J* **79** (2), 1146 (2000).
- Skrynnikov, N. R. et al., Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin. *J Mol Biol* **295** (5), 1265 (2000).

- Skrynnikov, N. R. et al., Probing Slow Time Scale Dynamics at Methyl-Containing Side Chains in Proteins by Relaxation Dispersion NMR Measurements: Application to Methionine Residues in a Cavity Mutant of T4 Lysozyme. *J. Am. Chem. Soc.* **123** (19), 4556 (2001).
- Smith, Brian O. et al., An approach to global fold determination using limited NMR data from larger proteins selectively protonated at specific residue types. *Journal of Biomolecular NMR* **8** (3), 360 (1996).
- Spera, Silvia and Bax, Ad, Empirical correlation between protein backbone conformation and C.alpha. and C.beta. <sup>13</sup>C nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society* **113** (14), 5490 (1991).
- Szilagyi, L. and Jardetzky, O.,  $\alpha$ -Proton chemical shifts and secondary structure in proteins. *J Magn Reson* **83**, 441 (1989).
- Takeuchi, K. and Wagner, G., NMR studies of protein interactions. *Curr Opin Struct Biol* **16** (1), 109 (2006).
- Teng, Q., Iqbal, M., and Cross, T. A., Determination of the carbon-13 chemical shift and nitrogen-14 electric field gradient tensor orientations with respect to the molecular frame in a polypeptide. *J. Am. Chem. Soc.* **114** (13), 5312 (1992).
- Tjandra, N. and Bax, A., Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278** (5340), 1111 (1997).
- Tugarinov, V., Choy, W. Y., Orekhov, V. Y., and Kay, L. E., Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci U S A* **102** (3), 622 (2005).
- Tugarinov, V. and 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.* **125** (45), 13868 (2003).
- Tugarinov, V. and Kay, L. E., Side Chain Assignments of Ile 1 Methyl Groups in High Molecular Weight Proteins: an Application to a 46 ns Tumbling Molecule. *J. Am. Chem. Soc.* **125** (19), 5701 (2003).
- Tugarinov, V., Kay, L. E., Ibraghimov, I., and Orekhov, V. Y., High-resolution four-dimensional <sup>1</sup>H-<sup>13</sup>C NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. *J Am Chem Soc* **127** (8), 2767 (2005).
- Tugarinov, V., Muhandiram, R., Ayed, A., and Kay, L. E., Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase g. *J Am Chem Soc* **124** (34), 10025 (2002).

- Uhrin, D. et al., 3D HCCH<sub>3</sub>-TOCSY for resonance assignment of methyl-containing side chains in <sup>13</sup>C-labeled proteins. *J Magn Reson* **142** (2), 288 (2000).
- Ulrich, E. L. et al., BioMagResBank. *Nucleic Acids Res* **36** (Database issue), D402 (2008).
- Venters, R. A., Farmer, B. T., 2nd, Fierke, C. A., and Spicer, L. D., Characterizing the use of perdeuteration in NMR studies of large proteins: <sup>13</sup>C, <sup>15</sup>N and <sup>1</sup>H assignments of human carbonic anhydrase II. *J Mol Biol* **264** (5), 1101 (1996).
- Wagner, G. and Wuthrich, K., Sequential resonance assignments in protein <sup>1</sup>H nuclear magnetic resonance spectra. Basic pancreatic trypsin inhibitor. *J Mol Biol* **155** (3), 347 (1982).
- Wand, A. J., Ehrhardt, M. R., and Flynn, P. F., High-resolution NMR of encapsulated proteins dissolved in low-viscosity fluids. *Proc Natl Acad Sci U S A* **95** (26), 15299 (1998).
- Wang, C. C., Chen, J. H., Lai, W. C., and Chuang, W. J., 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. *J Biomol NMR* **38** (1), 57 (2007).
- Williamson, Michael P., Havel, Timothy F., and Wuthrich, Kurt, Solution conformation of proteinase inhibitor IIA from bull seminal plasma by <sup>1</sup>H nuclear magnetic resonance and distance geometry. *Journal of Molecular Biology* **182** (2), 295 (1985).
- Wishart, D. S. and Nip, A. M., Protein chemical shift analysis: a practical guide. *Biochem Cell Biol* **76** (2-3), 153 (1998).
- Wishart, D. S., Sykes, B. D., and Richards, F. M., The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31** (6), 1647 (1992).
- Wolfgang Kabsch, Christian Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** (12), 2577 (1983).
- Wüthrich, Kurt, *NMR of proteins and nucleic acids*. (Wiley, New York ; Chichester, 1986).
- Xu, Y., Lin, Z., Ho, C., and Yang, D., A general strategy for the assignment of aliphatic side-chain resonances of uniformly <sup>13</sup>C, <sup>15</sup>N-labeled large proteins. *J Am Chem Soc* **127** (34), 11920 (2005).

- Xu, Y., Zheng, Y., Fan, J. S., and Yang, D., A new strategy for structure determination of large proteins in solution without deuteration. *Nat Methods* **3** (11), 931 (2006).
- Yamazaki, Toshio et al., A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of  $^{15}\text{N}$ ,  $^{13}\text{C}$ ,  $^2\text{H}$  Labeled Proteins with High Sensitivity. *J. Am. Chem. Soc.* **116** (26), 11655 (1994).
- Yang, D. and Kay, L. E., TROSY Triple-Resonance Four-Dimensional NMR Spectroscopy of a 46 ns Tumbling Protein. *J. Am. Chem. Soc.* **121** (11), 2571 (1999).
- Yang, D., Zheng, Y., Liu, D., and Wyss, D. F., Sequence-specific assignments of methyl groups in high-molecular weight proteins. *J Am Chem Soc* **126** (12), 3710 (2004).
- Zheng, Deyou et al., Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci* **12** (6), 1232 (2003).
- Zheng, Y. et al., Side-chain assignments of methyl-containing residues in a uniformly  $^{13}\text{C}$ -labeled hemoglobin in the carbonmonoxy form. *J Biomol NMR* **30** (4), 423 (2004).
- Zheng, Y. and Yang, D., Measurement of dipolar cross-correlation in methylene groups in uniformly  $^{13}\text{C}$ -,  $^{15}\text{N}$ -labeled proteins. *J Biomol NMR* **28** (2), 103 (2004).
- Zheng, Y. and Yang, D., STARS: statistics on inter-atomic distances and torsion angles in protein secondary structures. *Bioinformatics* **21** (12), 2925 (2005).
- Zuiderweg, E. R., Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **41** (1), 1 (2002).
- Zwahlen, C. et al., Methods for Measurement of Intermolecular NOEs by Multinuclear NMR Spectroscopy: Application to a Bacteriophage; N-Peptide/boxB RNA Complex. *J. Am. Chem. Soc.* **119** (29), 6711 (1997).



## Publications

- Xu, Yingqi\*; Zheng, Yu\*; Fan, Jing-song; Yang, Daiwen. **A novel strategy for structure determination of large proteins without deuteration.** *Nature Methods*(2006), 3(11): 931-937. (\* Equal contribution.)
- Zheng, Yu; Yang, Daiwen. **STARS: statistics on inter-atomic distances and torsion angles in protein secondary structures.** *Bioinformatics* (2005), 21(12):2925-2926.
- Zheng, Yu; Giovannelli, Janel L.; Ho, Nancy T.; Ho, Chien; Yang, Daiwen. **Side-chain assignments of methyl-containing residues in a uniformly  $^{13}\text{C}$ -labeled hemoglobin in the carbonmonoxy form.** *Journal of Biomolecular NMR* (2004), 30(4), 423-429.
- Yang, Daiwen; Zheng, Yu; Liu, Dingjiang; Wyss, Daniel F. **Sequence-specific assignments of methyl groups in high-molecular weight proteins.** *Journal of the American Chemical Society* (2004), 126(12), 3710-3711.
- Zheng, Yu; Yang, Daiwen. **Measurement of Dipolar Cross-Correlation in Methylene Groups in Uniformly  $^{13}\text{C}$ -,  $^{15}\text{N}$ -Labeled Proteins.** *Journal of Biomolecular NMR* (2004), 28(2), 103-116.
- Liu, Weidong; Zheng, Yu; Cistola, David P.; Yang, Daiwen. **Measurement of methyl  $^{13}\text{C}$ - $^1\text{H}$  cross-correlation in uniformly  $^{13}\text{C}$ -,  $^{15}\text{N}$ -, labeled proteins.** *Journal of Biomolecular NMR* (2003), 27(4), 351-364.
- Fan, Daping; Zheng, Yu; Yang, Daiwen; Wang, Jianjun. **NMR Solution Structure and Dynamics of an Exchangeable Apolipoprotein, *Locusta migratoria* Apolipoprotein III.** *Journal of Biological Chemistry* (2003), 278(23), 21212-21220.