Structure calculation of proteins from solution and solid-state NMR data : Application to monomers and symmetric aggregates

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of the Freie Universität Berlin

by

Benjamin Bardiaux
from Nantes, France

January, 2009

1st Reviewer: Prof. Dr. Michael Nilges (Institut Pasteur, Paris)
2nd Reviewer: Prof. Dr. Hartmut Oschkinat (FMP, Berlin)

date of defence: 8 Sept. 2009

This work was carried out from October 2004 to September 2008 under the joint supervision of Prof. Dr. Michael Nilges (Institut Pasteur, Paris), Dr. Thérèse E. Malliavin (Institut Pasteur, Paris) and Prof. Dr. Hartmut Oschkinat (Leibnizinstitut für Molekulare Pharmakologie, Berlin) at the Unité de Bioinformatique Structurale of the Institut Pasteur in Paris.

Acknowledgements

I would like to thank Michael Nilges for giving me the opportunity to work in an excellent scientific environment and entrusting me with all ARIA related projects during the last four years. I am deeply grateful to Thérèse Malliavin for introducing me to the world of NMR and structure calculations. Her everyday support, inventiveness and cheerfulness was of great help during my thesis. I wish to thank Hartmut Oschkinat for his effective co-supervision of my thesis from Berlin.

I would also like to express my gratitude to the persons across Europe without whom this work would not have been possible : Michael Habeck and Wolfgang Rieping who initiated me to the mysteries of the ARIA program; Anja Böckmann for her constant enthusiasm for structure calculations; Antoine Loquet, Carole Gardiennet, Manuel Etzkorn and Marc Baldus for fruitful and pleasant discussions; Richard Lavery for welcoming me in his lab and Alexey Mazur for his contribution to the work on ICMD; Ernest Laue, Tim Stevens, Wim Vranken and all the CCPN developer team; Michael Sattler and Bernd Simon for providing data on the B2 dimer; Christian Wasmer and Beat Meier for discussions on HET-s fibrils; Michele Fossi for sharing his experience on SOLARIA; Iñaki Guijarro and all ARIA users and contributors.

My visits to Berlin were always a great pleasure, thanks to the availability and kindness of Hartmut Oschkinat, Frank Eisenmenger and Victoria Highman. Particular thanks to Johanna Becker and Barth-Jan van Rossum for their close collaboration on CA150.WW2 fibrils and SH3 projects.

My everyday work at the Institut Pasteur has been greatly facilitated by Tru Huyn and Renée Communal. I would like to thank both of them for their crucial help with computer and office matters. I am thankful to all the BIS people for the pleasant atmosphere in the lab and more particularly Olivier, Giacommo, Laura and Elodie. Special thanks to Aymeric and David who were always present for work or fun.

I would like to thank the French “Ministère de la Recherche“ for financial support through the “Action Concerté Incitative” ICMD_RMN and the ANR “Regul lipids+”. This work was also supported by the EU projects SPINE and Extend-NMR.

Je remercie également mes parents, Marie et Jean-Claude, qui m’ont toujours soutenu durant mes études, ainsi que mon frère et Catherine. Un grand merci à Annie et toute la famille de Morgane.

Of course, I could not have done this thesis without the support of Morgane, my beloved little fay, who is always beside me.

CONTENTS

Abbreviations viii

1 General Introduction 1

1.1 Concepts in NMR spectroscopy . . . 1
  1.1.1 Chemical shift . . . 2
  1.1.2 Scalar and dipolar couplings . . . 2
  1.1.3 Magical angle spinning solid-state NMR spectroscopy . . . 3
1.2 NMR data for protein structure determination . . . 3
  1.2.1 Nuclear Overhauser Effects . . . 4
  1.2.2 J couplings . . . 5
  1.2.3 Chemical shifts . . . 6
  1.2.4 Residual dipolar couplings . . . 6
  1.2.5 Structural restraints from solid-state MAS NMR . . . 7
1.3 Structure calculation algorithms . . . 8
  1.3.1 Molecular dynamics simulated annealing . . . 8
  1.3.2 NMR data energy target functions . . . 9
  1.3.3 Torsion angle dynamics . . . 11
  1.3.4 Internal Coordinate Molecular Dynamics approach . . . 12
  1.3.5 Inferential Structure Determination . . . 12
1.4 Automated NOE assignment and structure calculation . . . 13
  1.4.1 Current state of the field . . . 13
  1.4.2 ARIA . . . 14
1.5 Validation of NMR structures . . . 15
  1.5.1 Experimental data validation . . . 15
  1.5.2 Geometric quality evaluation . . . 16
1.6 NMR structure determination of symmetric oligomers . . . 17
1.7 Objectives of the thesis . . . 19

2 Programming contribution to the ARIA software 21

2.1 Description of the ARIA program . . . 21
  2.1.1 Detailed protocol steps . . . 21
  2.1.2 Bookkeeping in ARIA . . . 27
  2.1.3 General ARIA implementation . . . 27
2.2 Network anchoring in ARIA . . . 28
2.3 Symmetric homo–dimer calculation . . . 29
2.4 Adaptation to solid-state MAS NMR data . . . 31
2.5 Graphical analysis of NMR structural quality . . . 32
  2.5.1 Interactive analysis of peaks assignments . . . 32
  2.5.2 Per-residue structural quality . . . 34

3 Influence of automated NOE assignment and NMR structure calculation protocols on structural quality 35

3.1 Effect of network anchoring on the structure of the HRDC domain 36
  3.1.1 Calculation schemes 36
  3.1.2 Results 36
3.2 Comparison of different torsion angle approaches for NMR structure determination. 40
  3.2.1 Calculation schemes 40
  3.2.2 Analysis of the ICMD structures with floating restraints 44
  3.2.3 Ambiguous and floating form of restraints 46
  3.2.4 Comparison between ICMD and RECOORD 47
3.3 Conclusion 52

4 Automated structure determination of symmetric homo-dimers with ARIA. 55

4.1 Symmetric homo-dimers from solution NMR 55
  4.1.1 Calculation schemes 55
  4.1.2 Efficiency of the chain assignment 57
  4.1.3 Application of the network anchoring with ambiguous data 60
  4.1.4 Impact of the spin diffusion correction with different levels of ambiguity 62
4.2 Structure determination of the *Crh* homo-dimer from solid-state NMR 64
  4.2.1 Experimental context 64
  4.2.2 Automatic assignments of highly ambiguous restraints. 65
  4.2.3 Calculation of the Crh 3D structure from the ARIA unambiguous restraints 67
  4.2.4 Comparison with other protein structures obtained from ssNMR data 69
4.3 Conclusion 69

5 Analysis of the conformational landscape of Crh from solution and solid-state NMR data 71

5.1 Introduction 71
5.2 Materials and Methods 73
  5.2.1 Input files of the conformers calculations 73
  5.2.2 ARIA-CNS calculation 73
  5.2.3 Structure analysis 74
  5.2.4 Clustering algorithms 75
  5.2.5 Molecular dynamics simulations 75
  5.2.6 Ensemble crystallographic refinement 75
5.3 Results 76
  5.3.1 Filtering of the liquid-state NMR restraints 76
  5.3.2 Presentation of data inputs and dimeric calculations 77
  5.3.3 Convergence of the calculation and fit to the NMR restraints 78
  5.3.4 Conformers quality and accuracy 82
  5.3.5 Relative monomer orientation in the dimer 83
  5.3.6 Dimer and tetramer architecture 86
5.4 Discussion 89

6 A general method for NMR structure calculation of symmetric aggregates **91**
6.1 Introduction 91
6.2 Material and Methods 92
6.2.1 Symmetric systems and data sets used 92
6.2.2 Structure calculation with strict symmetry 93
6.3 Results and Discussion 95
6.3.1 Phospholamban homo-pentamer 95
6.3.2 Modelling of amyloid fibrils from solid-state NMR data 99
6.3.3 Refinement of SH3 domain microcrystalline structure with crystal symmetries 104
6.4 Conclusion 105

7 General conclusions and Perspectives **107**
7.1 Concluding remarks 107
7.2 Perspectives 109
7.3 Publications 110

References **111**

Summary **123**

Zusammenfassung **124**

Résumé **125**

C.V. **127**

Appendix **129**

ABBREVIATIONS

2D	two-dimensional
3D	three-dimensional
ADR	ambiguous distance restraint
ARIA	ambiguous restraints for iterative assignments
CPU	Central Processing Unit
Crh	Catabolite repressor HPr-like protein
HRDC	helicase and RNaseD C-terminal
ICMD	Internal Coordinate Molecular Dynamics
ISPA	isolated spin-pair approximation
MAS	magic angle spinning
MD	molecular dynamics
MDSA	molecular dynamics based simulated annealing
NCS	non-crystallographic symmetry
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
PDB	Protein Data Bank
PDSD	proton driven spin-diffusion
PLN	phospholamban
RMS	root mean square
RMSD	root mean square deviation
RMSF	root mean square fluctuation
RDC	residual dipolar coupling
SA	simulated annealing
ssNMR	solid-state NMR
SH3	Src homology 3
TAD	torsion angle dynamics
WW2	Trp-Trp domain 2

1

General Introduction

1.1 Concepts in NMR spectroscopy

The aim of this part is to present general concepts in NMR spectroscopy that will appear in the thesis. *Nuclear Magnetic Resonance* (NMR) spectroscopy relies on the quantum effects induced by an external magnetic field to the magnetic momentum of an atomic nucleus, the magnetic momentum μ being defined by:

$$\mu = \gamma \mathbf{I};\, \mu_z = \gamma I_z \tag{1.1}$$

where γ is the gyromagnetic ratio of the nucleus and $\mathbf{I}$ the nuclear spin angular momentum; I_z and μ_z are the z components of $\mathbf{I}$ and μ, respectively.

The nuclear spin angular momentum number I is a multiple of $\frac{1}{2}$, and the number of quantum states is $2I + 1$. High resolution NMR mainly focuses on $I = \frac{1}{2}$ nuclei (mainly ^{1}H,^{13}C and ^{15}N), that have only two spin states. The use of isotopes enriched (^{13}C and ^{15}N) samples alleviates the problem of the natural abundance these nuclei.

In the absence of external field, there is no energy difference between the two spins states. When applying a external static magnetic field ($\mathbf{B}$) along the z-axis, the interaction of the magnetic field with the nuclear spin produces a Zeeman effect. The spin nuclei present in the sample are divided in two populations: slightly more than half of the nuclei exist in the lower energy state called α than in the higher energy state called β. In the lower energy level, the spin magnetic moment does not oppose the applied field, whereas it opposes it in the higher energy level. The transition energy between the two states is given by:

$$\Delta E = \gamma \hbar B_0 \tag{1.2}$$

In a "classical" view of the behaviour of the nucleus - that is, the behaviour of a charged particle in a magnetic field, the axis of spin rotation precesses around the magnetic field B_0. The frequency of precession ν_0 is the Larmor frequency, and is identical to the transition frequency corresponding to the transition energy described above:

$$\Delta E = \hbar\nu_0 = \hbar\gamma B_0 \tag{1.3}$$

The Larmor frequency is the resonance frequency of the NMR phenomenon.

1.1.1 Chemical shift

Chemically different hydrogens in a molecule do not experience the same magnetic field. Indeed, electrons shield the nucleus thereby reducing the effective magnetic field and requiring energy of a lower frequency to cause resonance. On the other hand, when electrons are withdrawn from a nucleus, the nucleus is deshielded and feels a stronger magnetic field requiring more energy (higher frequency) to cause resonance.

As the Larmor frequency described above (Eq. 1.3) is related to the magnetic field experienced by the observed nucleus, nuclei in different electronic environments resonate at different Larmor frequencies. This frequency shift is called the *chemical shift* of a nuclei and is calculated as the difference between the Larmor frequency ν of the nucleus and a standard ν_{ref}, relative to the standard. This quantity is reported in ppm (part per million):

$$\delta = \frac{\nu - \nu_{ref}}{\nu_{ref}} 10^6 \tag{1.4}$$

The intimate relationship between the chemical shift and the chemical environment of a nucleus has given the basis for the NMR development in chemistry and biology.

1.1.2 Scalar and dipolar couplings

Nuclear spins within macromolecular structures may interact with each other. The interaction of nuclei mediated by electrons through chemical bonds is named *scalar couplings* (or *J coupling*). Its strength is measured by the scalar coupling constant ${}^{n}J_{ab}$ for two nuclei a and b separated by n covalent bonds.

Moreover, since spins behave as magnetic moments, they interact also with each other through space. This *dipolar coupling* is orientation-dependant, but, in solution, it is averaged to zero due to molecular tumbling. Nevertheless, it still influences spin relaxation rates. The transfer of magnetisation between spins coupled by the dipole–dipole interaction in a molecule that undergoes Brownian motion in a liquid is called *Nuclear Overhauser Effect* (NOE). The intensity of an NOE is related to the distance between the two interacting spins. Magnetisation can also be transferred from one spin to another not only directly but also by *spin diffusion*, i.e. indirectly via other spins in the vicinity [Solomon, 1955; Macura and Ernst, 1980; Neuhaus and Williamson, 1989]. The application of NOE spectroscopy for structure determination will be discussed in the next section (1.2).

1.1.3 Magical angle spinning solid-state NMR spectroscopy

Solid-state NMR spectroscopy (*ssNMR*) is used for immobilised proteins (micro-crystalline powders of soluble proteins, membrane proteins in lipid bilayers or fibrillar proteins) on samples with very slow or inexistent reorientation. ssNMR offers much lower spectral resolution compared to solution NMR, but is less size-limited. Indeed, band-narrowing mechanism do not appear due to the absence of molecular tumbling.

However, anisotropic interactions often contribute to a line-broadening effect in NMR spectra. To simulate motional averaging, the sample is subjected to a rapid spinning at the magic angle (θ = 54.7˚relative to the magnetic field B_0). Under this *Magic Angle Spinning* (MAS), anisotropic interactions vanish and the normally broad lines become narrower, increasing the resolution. Since the spinning rate has to be larger than the interactions strength, the strong dipolar coupling involving protons prevents the detection of ^{1}H in MAS NMR. Therefore, the nuclei observed are generally low-γ nuclear spins, i.e. ^{13}C and ^{15}N.

1.2 NMR data for protein structure determination

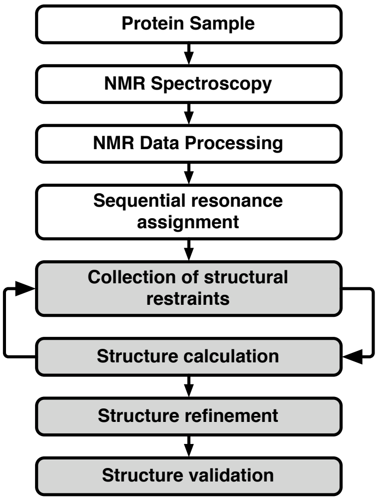


Figure 1.1: Outline of the procedure for protein structure determination by NMR.

The basic process of structure determination by NMR consists of different steps of experimental data acquisition, processing and resonance assignment. Then, structural restraints are derived and used for structure calculation. The general procedure of protein structure determination is illustrated in Figure 1.1. The most important experimental parameter for the determination of biomolecular structure in solution is the NOE [Wüthrich, 1986]. A series of different data can be directly used for the determination of NMR structures in addition to the NOEs, such as J-couplings, chemical shifts and dipolar couplings (Figure 1.2)

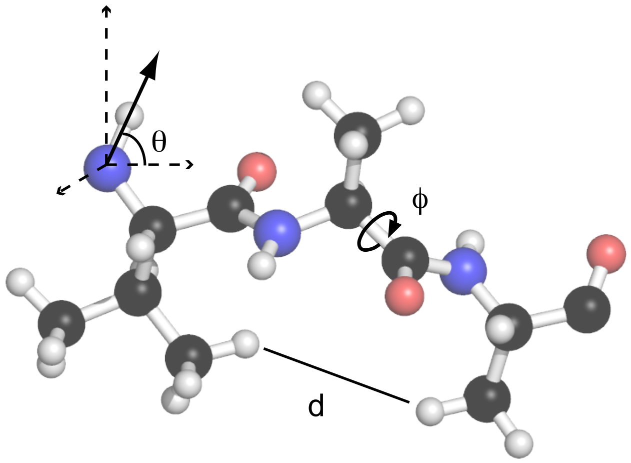


Figure 1.2: Structural data that can be measured by NMR : inter-proton distance (**d**) derived from a NOE, torsion angle (ϕ) obtained from a coupling constant (J-coupling) or chemical shifts and angle (θ) between a covalent bond and an external coordinate system from residual dipolar couplings (RDC).

1.2.1 Nuclear Overhauser Effects

NOEs have the important property of depending on inter-nuclear distances, providing a way to measure them. NOESY [Jeener et al., 1979] is the 2D solution NMR experiment, which directly exploits this effect to correlate nuclei which are close in space (distance smaller than 5Å). Using a first-order approximation, the volume V of each NOE may be expressed as a function of the inter-nuclear distance r:

$$V = \left\langle r^{-6} \right\rangle f(\tau_c) \tag{1.5}$$

The remaining dependence of the magnetisation transfer on the motion is expressed as a function of the rotational correlation time (τ_c), which includes effects of global and internal motions of the molecule. Distances can be then derived from the measurement of NOE cross-peak intensities V and used for structure calculation by restraining the distance between the involved protons.

The approximation of isolated spin pairs is valid only for very short mixing times in the NOESY experiment. For longer mixing times, the spin-diffusion phenomenon must be considered in the estimation of the distance (Fig. 1.3). The Solomon equation [Solomon, 1955] provides a semi-classical description of two interacting spins. The cross-relaxation rates R_{ij} predicted by the Salomon's equation is a simple function of the distance r_{ij} and the correlation time τ_c:

$$R_{ij} = \frac{1}{10}\gamma^4 \hbar \frac{1}{r_{ij}^6}\left(-\tau_c + \frac{6\tau_c}{1+(2\omega_0\tau_c)^2}\right) \tag{1.6}$$

In matrix form, the theoretical NOE intensities A_{ij} in a 2D NOE spectrum recorded at a mixing time τ_m are given by:

$$\mathbf{A} = e^{-\mathbf{R}\tau_m} \tag{1.7}$$

where $\mathbf{R}$ is the matrix of all R_{ij}. Different methods exist to determine the values of the relaxation rates and/or distances from the NOE intensity by the inversion of Eq. 1.7 [Boelens et al., 1989; James et al., 1990].

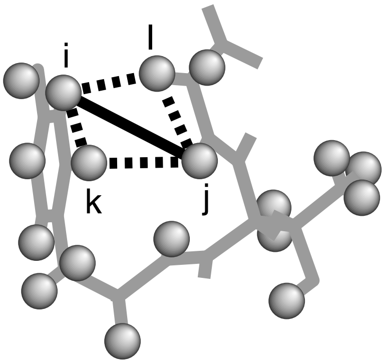


Figure 1.3: Illustration of spin-diffusion in a protein. Indirect magnetisation transfer via spins k and and l may influence the NOE between i and j.

1.2.2 J couplings

Dihedral angle restraints are the second group of important restraints that can be derived from NMR spectra. Figure 1.4 gives an illustration of dihedral angles in protein backbone. Dihedral angles influence the 3-bonds scalar ^3J couplings [Karplus, 1963]:

$$^3J(\theta) = A\cos^2(\theta) + B\cos(\theta) + C \tag{1.8}$$

Constants A,B and C have been determined for the most common dihedral angles θ (ϕ, ψ or χ). While NOEs provide information on short and long-range interactions, scalar coupling constants give insights only on the local conformation of a polypeptide chain. These restraints are therefore important in order to accurately define the local conformation.

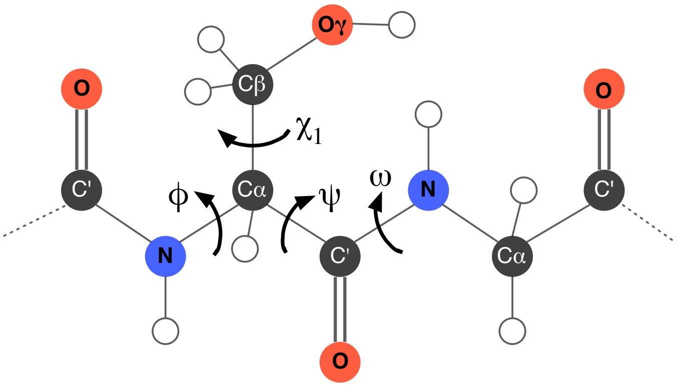


Figure 1.4: Overview of important dihedral angles (ϕ, ψ, ω and χ_1) in proteins.

1.2.3 Chemical shifts

Chemical shifts of backbone protons and heteronuclei in proteins are good indicators of the secondary structure elements of the system under investigation. It has been shown that a relationship exists between the “secondary chemical shifts” and the secondary structure [Dalgarno et al., 1983; Wüthrich, 1986]. The secondary chemical shift corresponds to the part of the chemical shift that is induced by the 3D structure, i.e. the difference between the measured chemical shift and the chemical shifts in a random coil peptide.

The Chemical Shifts Index (CSI) [Wishart et al., 1992] methodology predicts qualitatively the secondary structure of the amino-acids of a protein from the backbone chemical shifts (mainly Hα and Cα shifts). Such information can be included in a structure calculation in the form of canonical torsion angle restraints or, for the case of α-helices, of hydrogen bonds restraints. Also, more quantitative dihedral angle restraints can be predicted automatically with the program TALOS [Cornilescu et al., 1999] from a database of tripeptide backbone secondary chemical shifts of high resolution structures.

More recently, it has been shown that the structural information contained in the chemical shifts is sufficient for *de novo* structure determination of proteins from chemical shifts only. Two comparable approaches, CHESHIRE [Cavalli et al., 2007] and CS-ROSETTA [Shen et al., 2008], are based on a selection of molecular fragments consistent with the sequence and chemical shifts from a database of high- resolution structures, followed by Monte-Carlo simulations including a chemical shift agreement energy term.

1.2.4 Residual dipolar couplings

It is possible to partially restore dipolar couplings in solution by adding agents in the sample to orient, to a certain extent, the proteins along a particular direction [Tjandra, 1999]. In that context, *residual dipolar couplings* (RDC) provide information on the orientation of bond vectors with respect to the direction of the external magnetic field B_0. When considering a rigid molecule, the residual dipolar coupling between two spins i and j with respect to the external axis system (A) can be expressed as:

$$D^{ij}(\phi,\psi) = D_a\left[(3\cos^2(\theta)-1)+\frac{3}{2}R\sin^2(\theta)\cos(2\phi)\right] \tag{1.9}$$

The two parameters D_a and R are the magnitude and the rhombicity of the RDC tensor, with $R = A_r/A_a$ (A_a and A_r being the axial and rhombic components of the alignment tensor). D_a is equal to

$$D_a = -\left(\frac{\mu_0 h}{16\pi^3}\right)\gamma_i\gamma_j\left\langle r_{ij}^{-3}\right\rangle A_a \tag{1.10}$$

and depends on r_{ij}, the inter-nuclear distance. γ_i,γ_j are the gyromagnetic ratio of spin i and j respectively. Figure 1.5 illustrates the relationship between the RDC tensor and the relative orientation of bond vectors in the tensor.

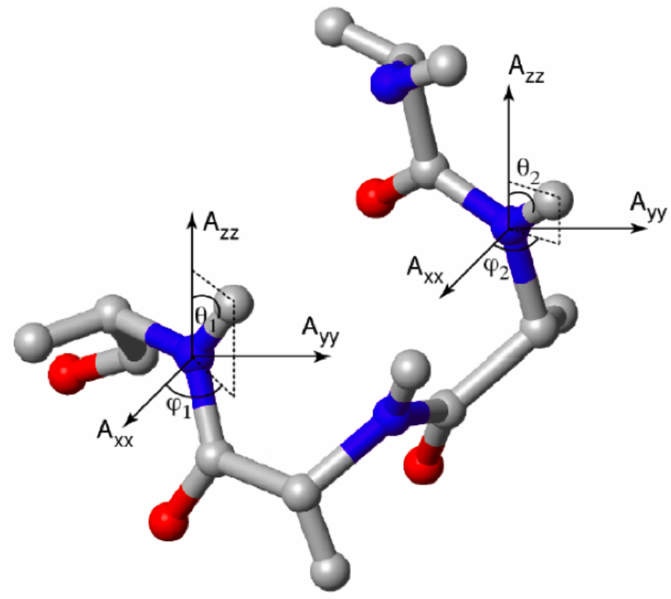


Figure 1.5: Orientation of two dipolar coupling vectors in a protein segment. The vectors connect the amide ^{1}H and ^{15}N atoms and they coincide with the chemical bond. The axis system of the alignment tensor is designated as A_{xx}, A_{yy}, A_{zz}. The angles θ_1, ϕ_1, and θ_2, ϕ_2 define the orientation of both dipolar vectors with respect to the alignment tensor.

1.2.5 Structural restraints from solid-state MAS NMR

For the determination of protein 3D structure, it is required to collect data on long-range interactions between ^{13}C,^{15}N and/or ^{1}H spins. In solid-state NMR, homonuclear and heteronuclear (^{13}C,^{15}N) correlations are mostly mediated by dipolar interactions. 2D or 3D Proton-driven spin-diffusion (*PDSD*) experiments explore indirect (^{13}C,^{13}C) couplings by transferring magnetisation through the protons in a spin-diffusion process. However, polarisation transfer is dominated by the strong dipolar couplings between covalently bonded carbon spins, and the measurement of weak couplings (through space) is attenuated. To alleviate this problem, strategic labelling schemes have been developed. Proteins are expressed on a medium containing [1,3-^{13}C]glycerol or [2-^{13}C]glycerol as carbon source [Castellani et al., 2002, 2003]. In the resulting sample, approximately one carbon out of two is labelled and only a few amino acid types have ^{13}C labels in adjacent positions. This method permits to decongest PDSD spectra by limiting dipolar truncation and consequently access long-range CC correlations.

Meanwhile, proton-proton distances could provide precious information for structure determination, like in solution NMR. A series of experiments have been designed [Lange et al., 2002, 2003; Heise et al., 2005] to indirectly probe ^{1}H-^{1}H interactions by a detection in 2D homonuclear or heteronuclear (CHHC/NHHC) correlations spectra. These experiments do not require selectively-labelled proteins and can be recorded on uniformly labelled ^{13}C,^{15}N samples. In contrast to carbon correlations, since no protons are chemically bonded to each other, the majority of the correlations observed with this strategy yields valuable structural restraints.

The derivation of distance restraints from these two types of ssNMR experiments is hampered by two major problems. First, it is more difficult to infer an adequate relation between the observed signal intensity and the distance. In both types of experiments, the transfer rate is proportional to the inverse sixth-power of the distance, but numerous effects affect the relationship (partial mobility, transfer efficiency, spin-diffusion). Nevertheless, a semi-quantitative evaluation of the distance can be achieved from “build-up curves” analysis [Castellani et al., 2002; Gardiennet et al., 2008] or with a calibration protocol [Manolikas et al., 2008]. Second, detection on rare spins in solid-state NMR produces larger line-width compared to solution NMR (0.5 to 1 ppm). This results in a higher number of overlapping cross-peaks, which complicates the assignment process. Nevertheless, there are several examples of protein structures determined by MAS ss-NMR [Castellani et al., 2002; Zech et al., 2005; Seidel et al., 2005; Lange et al., 2006; Korukottu et al., 2008; Loquet et al., 2008].

1.3 Structure calculation algorithms

Calculation of NMR structurs is usually performed using a molecular dynamics based simulated annealing protocol to minimise an energy function. Simulated annealing is a non-linear optimisation technique that allows to overcome local energy barriers due to the kinetic energy of the system (Fig. 1.6).

1.3.1 Molecular dynamics simulated annealing

In Cartesian coordinates, molecular dynamics based simulated annealing (MDSA) consists in the numerical solution of Newton’s equation of motions:

$$\frac{d^2\vec{r_i}}{dt^2} = -\frac{1}{m_i}\frac{\partial E_{hybrid}}{\partial \vec{r_i}} \tag{1.11}$$

where m_i and $\vec{r_i}$ are the mass and position of a particle i.

The aim of the MDSA protocol is to find a global energy minimum of the potential energy function E_{hybrid}. This energy function contains terms relative to the chemical structure of the system and also from the NMR experimental data :

$$\begin{aligned} E_{hybrid} &= E_{chem} + E_{data} \\ &= \sum_i w_i Ei \\ &= w_{bond}E_{bond} + w_{angle}E_{angle} + w_{dihed}E_{dihed} + w_{imp}E_{imp} \\ &\quad + w_{non-bonded}E_{non-bonded} \\ &\quad + w_{NOE}E_{NOE} + w_{torsion}E_{torsion} + w_{RDC}E_{RDC} + \ldots \end{aligned} \tag{1.12}$$

The chemical term E_{chem} encompasses the contributions of geometrical properties (like bond length, angles, dihedral and improper angles) as well as non-bonded interactions (electrostatics and van der Waals forces), and is based on a molecular modelling *force-field*. After a high-

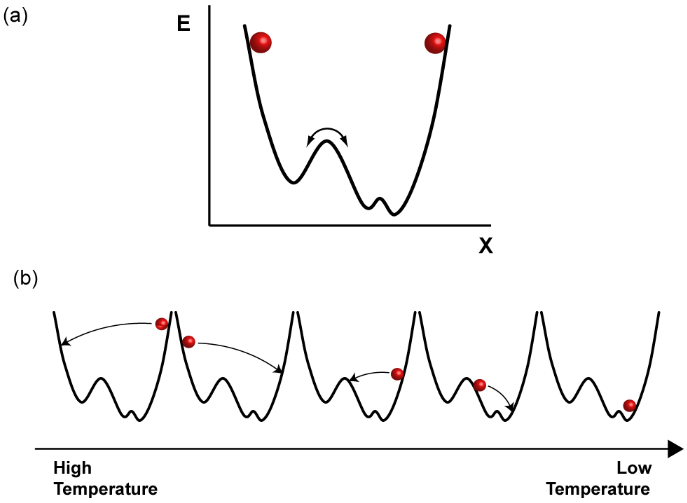


Figure 1.6: Illustration of the benefits of simulated annealing as a minimisation techniques. (a) Minimisation by molecular dynamics, (b) Temperature control: simulated-annealing

temperature stage search where the conformational landscape of the system is largely explored, the temperature is progressively reduced to converge toward the global minimum of E_{hybrid}. The temperature is controlled by coupling the system to an external heat bath that is fixed at the desired temperature [Berendsen et al., 1984]. The bath acts as a source of thermal energy, supplying or removing heat from the system as appropriate, and allowing the system to fluctuate about a desired temperature. In standard NMR structure calculation protocols, a simple repulsive potential [Nilges et al., 1988] is generally used to replace the Lennard-Jones and Coulombic interactions of the full empirical energy function. This short-range repulsive function can be calculated much faster and significantly facilitates large-scale conformational changes that are required during the folding process by lowering energy barriers.

1.3.2 NMR data energy target functions

Models to derive distances from the NOE cross-peak intensities are only approximate. Thus, it is a common practice to introduce *lower* and *upper* distance bounds to restrain the distance to an interval rather than to a unique value. For that reason, the energy term for NOE-based distance restraints is usually an harmonic potential with a “flat-bottom” (zero-energy between the lower (L) and upper (U) limits) (Figure 1.7 : *square-well*):

$$E_{noe} = \begin{cases} (L-d)^2 & \text{if d} < \text{L} \\ 0 & \text{if L} \leq \text{d} \leq \text{U} \\ (d-U)^2 & \text{if d} > \text{U} \end{cases} \tag{1.13}$$

in which d is the distance measured from the atomic coordinates. Similarly, dihedral angles are restrained with the same form of energy potential. In the context of automated NOE assignment (§ 1.4), large distance violations may occur during structure calculation. The shape of the poten-

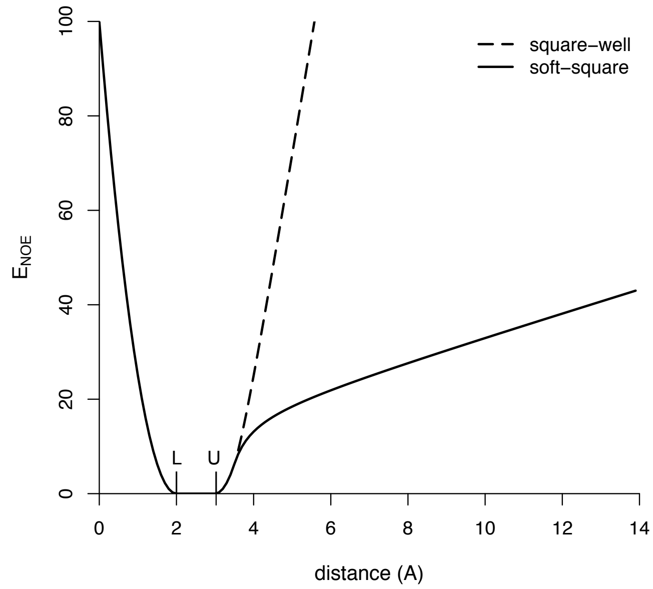


Figure 1.7: Flat-bottom NOE distance restraints potentials : “square-well” with harmonic walls. “soft-square” with smooth switching and linear asymptote for large distance.

tial is adapted by replacing the harmonic wall with a linear behaviour for large distance violations [Nilges, 1995] (Figure 1.7 : *soft-square*) :

$$E_{noe} = \begin{cases} (L-d)^2 & \text{if d} < \text{L} \\ 0 & \text{if L} \leq \text{d} \leq \text{U} \\ (d-U)^2 & \text{if U} < \text{d} < \text{U+}s \\ a + \frac{b}{(d-U)} + c(d-U) & \text{if d} > \text{U+}s \end{cases} \tag{1.14}$$

where *s* is the value where the potential switches from harmonic to asymptotic shape, c being the slope of the asymptote. The coefficients a and b are determined such that the potential and its first derivative are continuous. Recently, a new potential has been developed from the observation of the distribution of distances and NOE intensities [Rieping et al., 2005b]. The advantage of this *log-harmonic* potential is twofold: it has a single minimum and it is more tolerant for large violation [Nilges et al., 2008].

RDCs can be also incorporated in the structure calculation with a harmonic potential, like in the *SANI* term in the CNS program [Brünger et al., 1998]. With this method, a pseudo-molecule that is free to rotate represents the reference axes. During the calculation, it reorients itself to yield the best fit between experimental and calculated RDCs (from the angle of the bond-vector in the reference axis). The parameters of the RDC tensor (magnitude and rhombicity) need to be known. Several methods exist to predict these parameters, from the distribution of the RDC values [Clore et al., 1998] or from the shape of the molecule [Zweckstetter and Bax, 2000].

Alternatively, in the SCULPTOR approach [Hus et al., 2000], both the orientation and parameters of the alignment tensor are left free to float in a large flat-bottomed potential well and optimised throughout the course of structure calculation.

1.3.3 Torsion angle dynamics

In torsion angle dynamics (*TAD*), a protein is represented as a tree structure consisting of rigid bodies that are connected by rotatable bonds (Figure 1.8). The tree structure starts from the N-terminus of the chain and the coordinates of all the successive bodies are only defined by the preceding torsion angles. This concept greatly reduces the number of variables compared to Cartesian dynamics (by a factor $\sim$9). In the context of TAD, equations of motion are expressed for the torsion angles θ, and the torsional accelerations are defined by:

$$\ddot{\theta} = \mathbf{M}(\theta)^{-1}\mathbf{C}(\theta, \dot{\theta}) \tag{1.15}$$

where $\mathbf{M}$ is the $n \times n$ mass matrix and $\mathbf{C}$ a n-vector containing all inertial contributions (n is the total number of degrees of freedom). Both $\mathbf{M}$ and $\mathbf{C}$ can be calculated explicitly. However, the mass matrix $\mathbf{M}$ is non-diagonal and non-constant, i.e. depends on θ. Thus, a particular recursive algorithm is used to compute torsional accelerations [Jain et al., 1993]. The fixed standard geometry of the polypetidic chain allows for a significant increase of the time-step and of the temperature in MDSA protocols, because of the absence of harmonic potential for bond lengths or angles. Moreover, TAD was shown to improve the structure convergence as well as the structure quality [Güntert et al., 1997; Stein et al., 1997; Schwieters and Clore, 2001]. Torsion angle molecular dynamics is now implemented in numerous molecular dynamics programs (ICMD [Mazur, 1997, 1999], DYANA [Güntert et al., 1997], CNS [Brünger et al., 1998], Xplor-NIH [Schwieters et al., 2003]).

TAD has some similarities with the *variable target function* algorithm [Braun and Go, 1985]. The variable target function algorithm is based on a conjugate gradient minimisation of a target function in torsion angle space and where the NMR restraints are introduced in a certain order

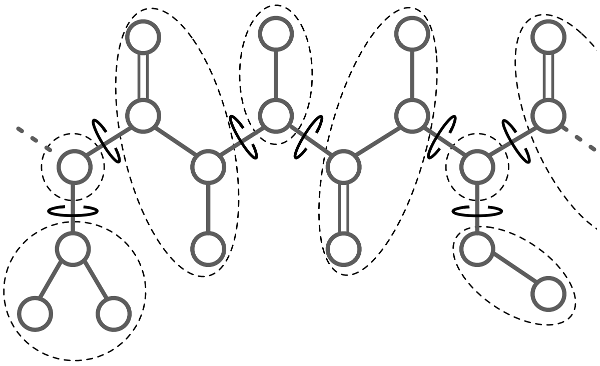


Figure 1.8: Tree structure of in torsion angles dynamics. Dotted shapes represent rigid units. Rotatable torsion angles between the units are indicated by solid circles.

(from intraresidual to long-range) that allows the refinement to proceed via valleys of potential energy landscapes.

1.3.4 Internal Coordinate Molecular Dynamics approach

The Internal Coordinate Modelling (ICM) approach was one of the first approaches [Mazur and Abagyan, 1989; Abagyan and Mazur, 1989] for modeling polymers with partially fixed spatial structure. In the last ten years, this approach was significantly enhanced by making possible propagation of MD trajectories in the space of canonical variables with new symplectic numerical integrators and the recursive mass matrix inversion algorithm [Jain et al., 1993; Mazur, 1997] with the Internal Coordinate Molecular Dynamics (ICMD) approach [Mazur, 1997, 1999]. ICMD is not limited to dynamics with a single choice of variables corresponding to the torsion angle space. In contrast, it can vary the degree of molecular flexibility by fixing and/or unfixing torsions, bond angles, and bond lengths. These technique was adapted for all-atom simulations of proteins and nucleic acids in explicit aqueous environments [Mazur, 1998b,a, 2002]. All these advances are potentially useful for NMR-based refinement applications since they greatly improve the quality and stability of MD trajectories and also allow modelling of multimolecular complexes. A new algorithm for NOE-based determination of biomolecular structures has been recently introduced in ICMD that has been successfully applied in studies of peptides and peptide-ion complexes [Kozin et al., 2001; Zirah et al., 2006]. This algorithm applies the idea of the variable target function [Braun and Go, 1985] in the context of torsion angle MD and also involves a simplified treatment of floating and ambiguous restraints. The application of the ICMD methodology for protein structure calculation from NMR data with a complete force-field will be presented in section 3.2.

1.3.5 Inferential Structure Determination

In contrast to the standard MDSA approach presented above, the Inferential Structure Determination (ISD) [Rieping et al., 2005a, 2008] method considers structure calculation as an “inference problem”. The method consists in exploring the conformational space (by a sampling strategy [Habeck et al., 2005]) to obtain a probability distribution of the structure ensemble. Each structure is ranked with a “posterior probability” computed with Bayesian probabilistic calculus. This probability combines two terms: a “prior probability” and a “likelihood function”. By analogy with equation 1.12, the prior knowledge is represented by the E_{chem} energy. The likelihood function encompasses a forward model, to calculate the data from the structure, and an error distribution. In addition to the determination of the structure uncertainty, one advantage of the method is that all unknown parameters (such as the distance restraint weight) are estimated in the course of the calculation. In theory, any kind of experimental data can be included provided that a suitable forward model exists. For example, RDC data can be incorporated in the structure calculation [Habeck et al., 2008] and all the unknown RDC tensor elements estimated during the calculation.

1.4 Automated NOE assignment and structure calculation

Manually analysing NOESY spectra is a tedious task. Unambiguously assigning NOE cross-peaks is sometimes very difficult due to spectral resolution and potentially overlapped cross-peaks. Consequently, several automated procedures for NOESY interpretation and structure calculation have been developed. Automated approaches generally follow the global strategy presented earlier (§ 1.2 and fig. 1.1) and aim at limiting manual interventions. Moreover, automation of data analysis in structure determination by NMR does not only lead to a significant speed up but also offers possibilities to document the process to a degree that is difficult to reach with manual approaches. Most methods require a complete or nearly complete list of assigned chemical shifts. This sequential resonance assignment is achieved by means of numerous "through-bond" NMR experiments, either manually or with the help of computational approaches (see [Moseley and Montelione, 1999; Baran et al., 2004] for reviews)

1.4.1 Current state of the field

In the last decade, several semi-automated or fully automated programs were published for NOE assignment and NMR structure calculation (see reviews in [Güntert, 2003; Altieri and Byrd, 2004]). The majority rely on an iterative process with several rounds of structure calculation and cross-peak assignment. While some approaches are semi-automated, like SANE [Duggan et al., 2001], most of the programs exhibit a high-level of automatisation, with direct interfaces to structure calculation engines. For example, in the program NOAH [Oezguen et al., 2002], structures are calculated with a variable target function method while restraints are examined by a violation analysis to automatically assign NOEs. ARIA[1] [Linge et al., 2003a; Rieping et al., 2007] and CANDID/CYANA [Herrmann et al., 2002; Güntert, 2004] both apply the concept of *Ambiguous Distance Restraint* (ADR) [Nilges, 1995] to handle multiple assignment ambiguities and rely on an MDSA protocol for structure calculation (§ 2.1.1 for a detailed description). ARIA includes all assignment possibilities from the beginning of the protocol and assumes that incorrectly assigned restraints are correlated and not compatible with the 3D structure. However, CANDID attempts to assign NOEs by searching for data corroborating each possible assignment, using the fact that the distance restraints should be mutually consistent. Thus, CANDID implements a *network-anchoring* procedure, which weights each assignment possibility with respect to the density of the network of all other assignment possibilities and with the help of the covalent structure (see § 2.2 for details). AutoStructure [Huang et al., 2006] is an expert system with a topology-constrained bottom-up approach. Another approach (PASD [Kuszewski et al., 2004]) uses a probabilistic method to estimate the validity of all NOEs and assignment possibilities during the course of the structure calculation, presuming that a cross-peak originates from a single interaction. While all strategies presented above require sequential assignment, the CLOUDS [Grishaev and Llinás, 2005] method constructs a "cloud" of atoms from unassigned chemical

---

[1] Ambiguous Restraints for Iterative Assignment

shifts and NOEs that is iteratively regularised to a 3D protein structure. Completely automated methods for NMR data analysis are also emerging in the context of structural genomics projects [Grishaev et al., 2005; Kobayashi et al., 2007].

1.4.2 ARIA

The ARIA procedure [Nilges et al., 1997; Linge et al., 2003a; Rieping et al., 2007] is widely used for structure determination from unassigned NOESY spectra. ARIA is based on an iterative protocol to derive distance constraints from NOE cross-peaks that will be applied to calculate an ensemble of structure. The newly created bundle of molecular conformations then serves to evaluate the consistency of the assignments, and the ambiguity present in the initial NOE assignments is progressively reduced (Figure 1.9).

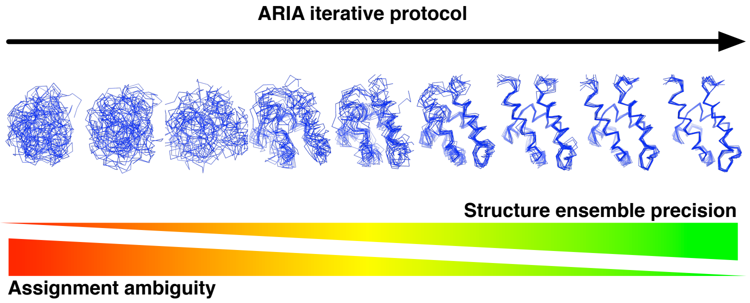


Figure 1.9: Illustration of the convergence of ARIA protocol.

The most important idea that underlies the ARIA methodology is the concept of *Ambiguous Distance Restraints* (ADR) [Nilges, 1995]. In the frame of the ADR, each NOESY cross-peak is treated as the superposition of the signals from each of its multiple assignments possibilities: the NOE intensity depends on the sum of the inverse sixth power of all the individual proton-proton distances that contribute to the signal. An effective distance $\bar{D}$ is thus derived as:

$$\bar{D} = \left(\sum_{c=1}^{N_c} d_c^{-6} \right)^{-\frac{1}{6}} \tag{1.16}$$

where c runs through all N_c assignment possibilities and d_c is the inter-atomic distance between the two protons corresponding to the c-th contributions. As a consequence of the r^{-6}-sum, $\bar{D}$ is always smaller than any individual d_c. During structure calculation, in a similar fashion as for unambiguous distance constraints, the distance $\bar{D}$ in the molecular coordinates is restrained through the distance target energy function (§ 1.3.2). The iterative ARIA protocol will be presented in detail in chapter 2.

1.5 Validation of NMR structures

One of the key steps in the determination of protein structure by NMR is the evaluation of their reliability and accuracy. Due to the sparse nature of the NMR data and due to the methods carried out to interpret them, NMR structures are generally more prone to errors compared to structures determined by X-ray crystallography. To validate the quality of protein structures determined by NMR, several criteria are commonly assessed, and can be separated into two groups: validation of experimental data and validation of the overall geometric quality based on comparison to high-resolution X-ray structures (see [Spronk et al., 2004; Nabuurs et al., 2004] for a thorough presentation).

In contrast to structure determined X-ray crystallography, an NMR structure does not correspond to a single set of atomic coordinates but to an ensemble of conformers calculated in the same way with same data. Thus, two important notions need to be defined, as they contribute to the overall validation process (Figure 1.10):

- ***Precision*** represents the positional uncertainty of the molecular coordinates in the structure ensemble. It is normally calculated as the average *root mean square deviation* (RMSD) of all the conformers from the average structure coordinates.

- ***Accuracy*** designates the closeness of the structure ensemble to a reference "true" structure. For NMR structures, accuracy is often estimated from a relevant X-ray structure and expressed as a RMSD. In the next chapters, the accuracy of an ensemble of NMR conformers will be occasionally computed from another average NMR structure.

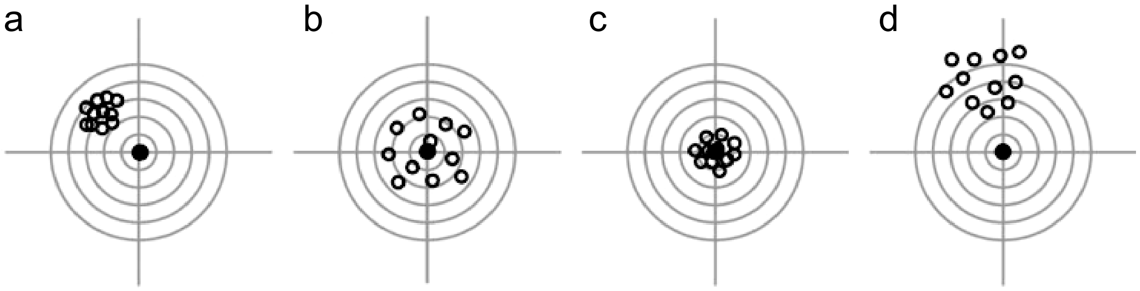


Figure 1.10: Illustration of precision and accuracy. The true x,y-coordinate of a given atom is indicated by a filled circle, sampled values of the coordinates are indicated by open circles. (a) Precise but inaccurate; (b) accurate but imprecise; (c) accurate and precise; (d) inaccurate and imprecise. From Spronk et al. [2004]

1.5.1 Experimental data validation

There are various measures to describe the accordance of NMR structures with the experimental restraints. For instance, the *number of violated restraints* as well as the *RMS of deviations* from the distance bounds in the structure ensemble are basic indicators. Violated restraints can be enumerated according to different violation cut-offs; the number of restraints violated in more than 50% of the conformers is referred to as the *number of consistent violations*. In addition, the QUEEN method [Nabuurs et al., 2003] tries to quantify the information content of the restraint data set (e.g. *restraints completeness, redundancy* and *uniqueness*). Some other measurements judge the agreement of a subset of data, excluded from the structure calculation, with the final structures. This independent check is particularly appropriate to validate a structure against RDC (*Q-factor* [Cornilescu et al., 1998]).

1.5.2 Geometric quality evaluation

The applications of NMR restraints for structure calculation may induce distortions in the geometry of the molecular structure. For this purpose, three major programs (PROCHECK [Laskowski et al., 1993], WHAT IF [Hooft et al., 1996] and MOLPROBITY [Davis et al., 2004]) aim at detecting outliers and abnormalities in macromolecular structure by comparing several characteristic geometric properties to a database of small molecules and/or high-resolution X-ray structures. A basic, but non trivial, inspection of bond length, bond angle, tetrahedral geometry and side chain planarity is performed to flag irregularities in the structures under analysis.

Another important concern is the backbone conformation. PROCHECK evaluates the distribution of the ϕ and ψ dihedral angles by classifying residues in different regions of the Ramachandran plot. The Ramachandran plot delineates the most favourable areas of (ϕ,ψ) couples that have been determined from steric considerations and statistics on high-resolution X-ray structures. WHAT IF proposes a more precise approach to quantify the similarity of the NMR structure to high-resolution X-ray structures, with the introduction of *Z-scores*. A Z-score represents the number of standard deviations a value is away from the average of a normal distribution. In the case of WHAT IF, the distribution is determined from the reference database of X-ray structures. The structures under analysis are then compared to the database distribution. By definition, the closer is a Z-score to zero, the more “ideal” is the conformation (here, the ideal value is the most represented conformation in the database). Positive and negative Z-scores are equally likely to occur and the generally acceptable range is (-4,+4), but a positive Z-score represents structures that are “too ideal”. In that way, WHAT IF offers several checks for the Ramchandran plot appearance, backbone normality and $\chi1/\chi2$ rotamer distribution. Furthermore, non-bonded interactions are also investigated through the quality of the hydrogen bonding network and of the residue packing. Steric clashes may also occur, but one must note that the three programs employ different methods and parameters to detect *inter-atomic bumps*, and the net influence of the atom radii used for structure calculation must not be neglected.

Yet, it has been shown that the choice of the force-field has a significant impact on the overall quality of the structures determined from NMR data [Spronk et al., 2002; Linge and Nilges, 1999; Linge et al., 2003b]. More recently, some studies stressed that global structural indicators are not sufficient to detect errors in structures and also encourage to look at parameters on a per-residue basis [Nabuurs et al., 2005, 2006]. Another recent study shows that some geometric indicators could be more informative than others in the detection of flaws [Saccenti and Rosato, 2008].

1.6 NMR structure determination of symmetric oligomers

Dealing with symmetric oligomeric structures is an important issue in the context of NMR for structural proteomics. Indeed, it is estimated that about 60% of the proteins in every genome are homo-oligomers, and half of all homo-oligomers structures in the Protein Data Bank (PDB) are homo-dimers [Goodsell and Olson, 2000; Levy et al., 2006]. Over the last ten years, the size range of proteins that can be explored by solution NMR spectroscopy has extended [Foster et al., 2007], and the number of structures of protein symmetric oligomers determined by NMR is continually increasing (Fig. 1.11). The problem with resolving ambiguity in NOE data is particularly severe for symmetric aggregates, due to the fact that the nuclei in the monomers are surrounded by exactly the same chemical environment and are therefore indistinguishable in NMR. Consequently, structure determination of symmetric oligomers is severely hampered by the difficulty to differentiate inter-monomeric and intra-monomeric correlations.

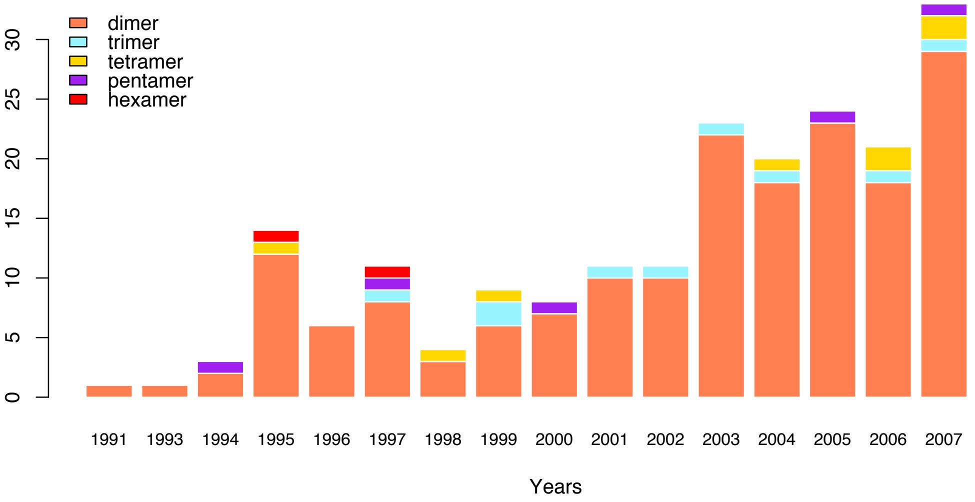


Figure 1.11: Symmetric oligomers : number of NMR structures deposited per year in the Protein Data Bank (PDB).

Symmetric aggregates are usually energetically more favourable than asymmetric aggregates [Blundell and Srinivasan, 1996; Wolynes, 1996]. *In vivo*, most symmetric aggregates have *point group symmetry*, which forms symmetric oligomers. A point group is defined by one or more symmetry axes and the rotation operators that relate the monomers arranged around each axis. The majority of homo-oligomers adopt either cyclic (C_n) or dihedral (D_n) symmetry [Levy et al., 2008], as illustrated in figure 1.12. In $\mathrm{C_n}$ symmetry, n monomers are arranged around a single symmetry axis with rotations of $360°/n$. The dihedral group $\mathrm{D_n}$ contain an axis of rotational symmetry and a perpendicular axis of two-fold symmetry. Oligomers with dihedral symmetry have different types of interface, and are often involved in allosteric controls. Another class of symmetry, *linear symmetry*, is observed in protein fibres, for instance. Here, the monomers are repeated along the fibre axis, by translational and rotational operators. A particularly interesting example of protein fibres are *amyloid fibrils* that play a role in various neurodegenerative diseases. The progress of MAS solid-state NMR spectroscopy shows that this technique is a powerful tool to study the structural organisation of amyloid fibrils [Tycko, 2006; Heise, 2008]

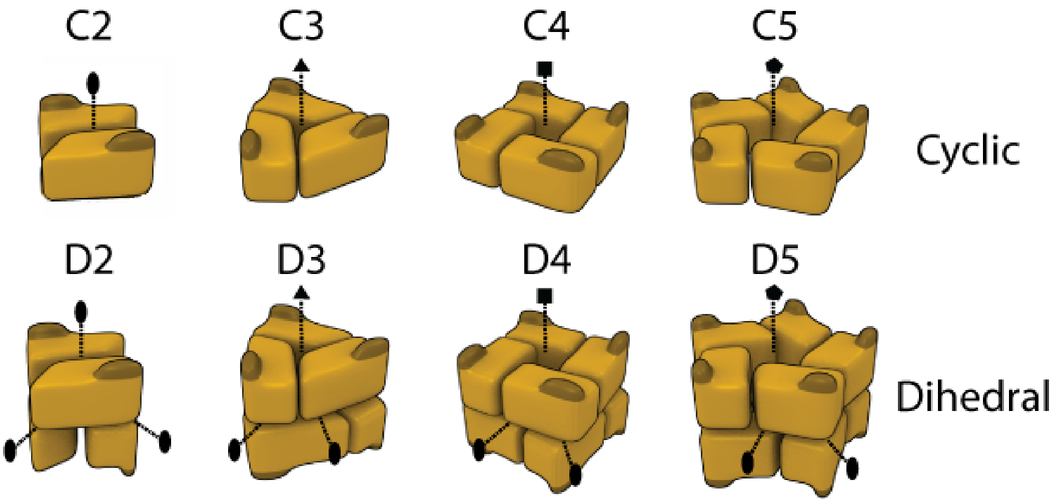


Figure 1.12: Illustration of crystallographic point group symmetries occurring in symmetric oligomers. The symmetry axes are represented with dotted lines. Adapted from www.3Dcomplex.org

Experimental approaches can resolve the ambiguity for dimers, but are not always straightforward to interpret. Most of the methods rely on the creation of asymmetrically labelled samples, where complexes are formed from mixtures of labelled and unlabelled or differentially labelled (e.g. ^{15}N, ^{12}C and ^{14}N, ^{13}C) protein monomers [Weiss, 1990; Folkers et al., 1993; Breeze, 2000]. Recently, the same principle have been adapted to solid-state MAS NMR [Etzkorn et al., 2004]. Marginally, paramagnetic probes can be introduced to disrupt the symmetry in the NMR spectra of homo-dimers [Gaponenko et al., 2002]. While these methods are usually sufficient to unambiguously assign intra- and inter-monomeric NOEs and to determine the structure of homo-dimers, the data remain highly ambiguous for higher-order oligomers.

Algorithmic support for the assignment seems therefore particularly important for symmetric oligomers. Different approaches using branch-and-bound algorithms have been proposed [Wang et al., 1998; Potluri et al., 2006, 2007]. Similarly, structures of symmetric homo-oligomers can be determined with a *docking* strategy in a symmetry restricted search space [Pierce et al., 2005; Berchanski et al., 2005] or by taking advantage of the information provided by RDCs to determine the relative orientation of the monomers [Bewley and Clore, 2000; van Dijk et al., 2005; Wang et al., 2008]. Also, structures of symmetrical protein assemblies can be predicted by Monte-Carlo simulated annealing algorithm [Andre et al., 2007].

These methods can explore systematically the entire space of the orientations of the monomers for a given symmetry, but they rely on the exact knowledge of the monomer structure, which is unrealistic during a *de novo* structure determination. An approach based on ambiguous distance restraints [Nilges, 1993] has been used successfully applied in a number of *bona fide* structure determinations [Junius et al., 1996; O'Donoghue et al., 2000; Kovacs et al., 2002]. It does not require the positions, and orientations of the symmetry axes and the symmetry is imposed by distance restraints (see § 2.3 for details). Still, this strategy has limited convergence properties and is somewhat difficult to apply for higher symmetry.

Likewise, solid-state MAS NMR faces the problem of discriminating between inter-molecular and intra-molecular correlations. In solid micro-crystalline samples of proteins, inter-molecular interactions with crystallographic neighbours are expected to cause additional cross-peaks in the spectra. This requires the preparation of asymmetrically labelled samples (*diluted samples*) to identify intermolecular contacts [Castellani et al., 2002]. In the case of amyloid fibrils, the ideal approach would be to consider a set of neighbouring monomers instead of a single one. Ferguson et al. [2006] illustrates the calculation of CA150.WW2 protofilament with six consecutive monomer units. Some rare examples of amyloid fibrils structure determined with additional "diluted samples" [Iwata et al., 2006; Wasmer et al., 2008] or from mutagenesis experiments [Lührs et al., 2005] demonstrate the difficulty of the task.

1.7 Objectives of the thesis

As we have seen in this chapter, NMR is a formidable technique to investigate the structure of proteins and protein aggregates. It brings also new computational challenges in terms of automation, data integration, structure calculation and validation. In this work, several aspects of the computational effort for NMR structure determination are covered.

One aim of this thesis is to develop and integrate methods for NMR restraint analysis and structure calculation and also to evaluate their relative impact on the resulting structural quality. These methods will be implemented in the framework of the ARIA (**Chapter 2**) and ICMD (**Chapter 3**) programs. The goal is to improve the standard protocols for NOE assignment and structure calculation in terms of structural quality and assignment speed. The influence of the network-anchoring will be assessed for the determination of protein structures from ambiguous data from solution-state NMR with ARIA and a new procedure for NMR structure calculation, based on the Internal Coordinate Molecular Dynamics (ICMD) approach will be compared to state-of-the-art procedures (**Chapter 3**).

The second main goal of this work is to tackle the problem of structure determination of symmetric protein assemblies from both solution and solid-state NMR data. For this purpose, a protocol based on symmetry distance restraints will be implemented in the ARIA methodology to determine the structure of symmetric homo-dimers. This method and the impact of network-anchoring will be evaluated with different levels of assignment ambiguity (**Chapter 4**). The investigation of symmetric protein structures by solid-state NMR is also to be analysed. Thus, the ARIA methodology will be applied to determine the structure of the dimeric protein *Crh* from unassigned solid-state NMR cross-peaks (**Chapter 4**) and its conformational landscape during oligomerisation and crystallisation will be studied with a simultaneous use of solution-state NMR, solid-state NMR and crystallographic restraints (**Chapter 5**). Despite the existence of different approaches for structure calculation of symmetric protein aggregates from NMR data, a general procedure is to be designed. A new methodology for structure calculation and refinement of symmetric protein assemblies from ambiguous NMR restraints will be developed, with a strict treatment of symmetries. The efficiency of the method will be evaluated on characteristic modern challenges in NMR, such as symmetric membrane proteins studied by solution-state NMR or amyloid fibrils investigated by solid-state NMR (**Chapter 6**).

2

Programming contribution to the ARIA software

Several methods for structure calculation, automated NOE assignment and validation that are evaluated in this work have been included in the ARIA software package and made available to the users community[1]. This chapter begins with a detailed presentation of the ARIA protocol followed by a description of the functionality implemented in ARIA during this thesis: *network anchoring*, support for *symmetric homo–dimers* and *solid-state NMR data* . Finally, extensions of the ARIA graphical interface are described. These extensions aim at providing new ways to detect potentially erroneous NOE assignments.

2.1 Description of the ARIA program

The general workflow of the ARIA methodology is presented in figure 2.1. After an initial chemical-shift based cross-peak assignment and a calibration step, ambiguous distance restraints are derived from the NOEs. From these restraints, an ensemble of conformer is calculated. On the basis of these conformers, noise peaks are detected with a violation analysis and unlikely assignment possibilities are discarded. This process is iterated several times (nine by default) with optimised parameters for each iteration. In the next paragraphs, each step of the protocol will be described in details.

2.1.1 Detailed protocol steps

Initial NOE assignment ❶

For every NOE cross-peak, ARIA uses the chemical shift lists from the sequential resonance assignment to derive possible assignments. As illustrated in figure 2.2, the peak position is defined by its frequency coordinates (c_1, c_2) in each dimension of the spectrum. In order to account for the limited precision in chemical shifts measurement, for the uncertainty of the NOE cross-peak coordinates and for systematic experimental errors, chemical shift tolerances (δ_1,δ_2)

---

[1] http://aria.pasteur.fr

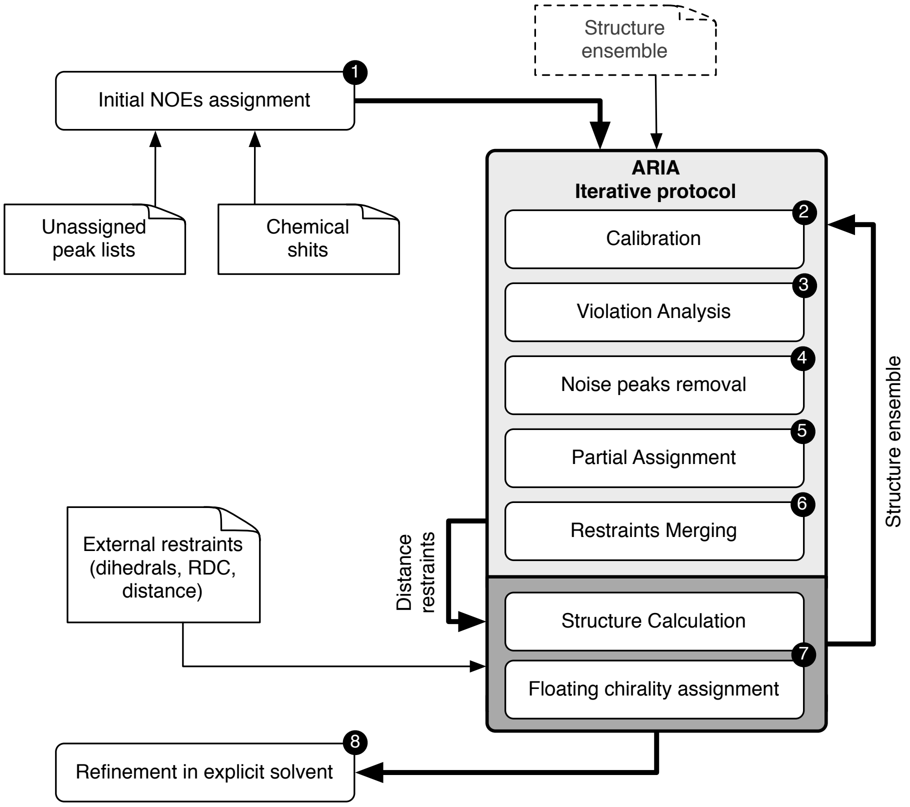


Figure 2.1: Description of the ARIA protocol workflow (see text for a detailed explanation of the different steps)

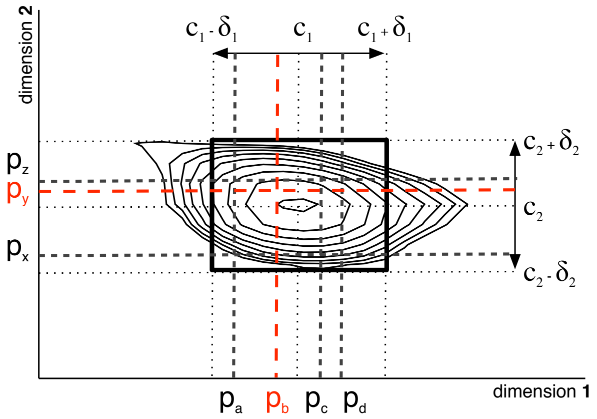


Figure 2.2: Illustration of the assignment of cross-peak. c_1, c_2 denote the peak coordinates in the frequency space. The assignment frequency windows is materialised by the solid black square, defined from the chemical shifts tolerances δ_1 and δ_2. The coordinates of the only correct assignment is represented by the red dashed lines (p_b, p_y). Multiple resonances laying in the tolerance windows (p_a, p_b, p_c,p_d in dimension 1 and p_x, p_y, p_z in the other dimension) give rise to 12 assignment possibilities.

are applied around the peak position. The tolerances should be chosen sufficiently large to obtain frequency windows that could compensate for all sources of inconsistencies between the list of resonance assignments and the cross-peak lists. Then, for each peak dimension, all protons whose chemical shifts fall in the peak frequency windows are collected. In the case of 3D or 4D heteronuclear spectra, the hetero atom attached to the proton must also match the corresponding chemical shift window. The list of all the assignment possibilities (or *contributions*) for a cross-peak is generated from the combination of the resonances assignment (Figure 2.2).

The sizes of the frequency windows play an important role in the initial NOE assignment step: too narrow windows induce potentially incomplete assignments, while large window sizes lead to highly ambiguous initial assignments, which is often the source of severe convergence issues during the ARIA protocol. Therefore, window size must be chosen carefully; the ideal situation is reached when the windows size is sufficiently large to contain the correct assignments, but without increasing unduly the number of assignment possibilities. Typical window size values for NOESY spectra are 0.02 ppm and 0.04 ppm for the direct and indirect protons dimension, respectively, and 0.5 ppm for the heteronuclear dimensions. The maximum number of assignment possibilities ($\mathrm{n_{max}}$) also affects the quality of the initial assignment, as some peaks that could correctly be assigned are rejected due to an excessively large number of assignment possibilities. Fossi et al. [2005b] have developed a strategy, based on a pre-calculation analysis, for choosing optimal values for δ and and $\mathrm{n_{max}}$ for a particular data set.

Distance restraints calibration ❷

As shown in paragraph 1.2.1 the simplest model to derive distance from NOE signal intensity is to consider an isolated-spin pairs (Isolated Spin pair Approximation). For short mixing times, ISPA provides a good approximation to relate an NOE volume (V_{ij}) to the distance d_{ij} of two interacting spins i and j:

$$V_{ij} = C d_{ij}^{-6} \tag{2.1}$$

The scale factor C (also named *calibration factor*) cannot be measured directly since it depends of the system under investigation and the experimental setup. The *calibration factor* is estimated for all NOEs from the ratio of the average of the experimental volume to the average of the theoretical volume:

$$C = \frac{\sum_i V_{exp}}{\sum_i \hat{d_i}^{-6}} \tag{2.2}$$

where $\hat{d_i}$ is the effective distance for NOE i in the conformer ensemble. In the case of multiple assignment possibilities, $\hat{d}$ is calculated according to equation of ADR (Eq. 1.16). Finally, the calibrated distance is obtained by:

$$d = (C^{-1} V_{exp})^{-\frac{1}{6}} \tag{2.3}$$

From this calibrated distance (or *target distance*), upper (U) and lower (L) bounds are derived:

$$\begin{aligned} L &= d - \Delta^- \\ U &= d + \Delta^+ \end{aligned} \tag{2.4}$$

where $\Delta^- = \Delta^+ = 0.125d^2$

Alternatively, ARIA offers another way to model the NOE signal by correcting the distance for the spin-diffusion phenomenon [Linge et al., 2004]. This is achieved by calculating the relaxation matrix ($\mathbf{R}$) (after applying a distance cut-off) and a fast integration scheme for the NOE matrix $\mathbf{A}$:

$$\mathbf{A}(\tau_m) = (\mathbf{I} - \mathbf{R}\Delta t)^N \mathbf{A}(0) \tag{2.5}$$

in which $N\Delta t = \tau_m$. The theoretical NOE matrix at mixing time τ_m is obtained by simplifying the calculation to N successive matrix multiplications ($\log_2 N$ matrix squaring). The resulting NOE back-calculated volumes, that take into account the bias induced by spin-diffusion, will then serve for the determination of the calibration factor C, leading to corrected target distances d:

$$d = \hat{d}\left(C^{-1}\frac{V_{exp}}{V_{th}}\right)^{-\frac{1}{6}} \tag{2.6}$$

When using spin-diffusion corrected distances, the computation of distance bounds from the theoretical volume may also be of interest for the structure calculation [Linge et al., 2004].

Violation analysis and noise peaks removal ❸+❹

To identify wrong assignments and noise peaks, the calibrated restraints are treated with a violation analysis, following the structural consistency hypothesis [Mumenthaler and Braun, 1995; Nilges and O'Donoghue, 1998]: incorrectly assigned peaks or noise peaks are not consistent with the 3D structure determined with all experimental data. To assess whether a particular restraint follows the "general trends" imposed on the structures by the entire data set, the obtained distance bounds are compared to the corresponding distances found in the conformer ensemble. A restraint is considered as *violated* if the distance found in the structure lies outside the bounds by more than a user-defined *violation tolerance*, t. To identify systematically violated restraints, each conformer in the ensemble is analysed. The fraction, f_i, of conformers violating restraint i is calculated according to:

$$f_i = \frac{1}{S}\sum_{j=1}^{S} \max\left(\theta(L_i - t - d_i^{(j)}), \theta(d_i^{(j)} - U_i - t)\right) \tag{2.7}$$

where L_i and U_i denote the lower and upper bounds of the i-th restraint, $d_i^{(j)}$ designates the distance found in the j-th conformer; θ is the Heaviside function and S is the total number of conformers analysed. A restraint is classified as violated if f_i exceeds a user-defined *violation threshold*, typically 50 %. The corresponding NOE cross-peak is thus removed of the list of active peaks for the next iteration. During the course of the protocol, the violation tolerance t is reduced from iteration to iteration to ensure that most of the inconsistent peaks are removed.

Partial assignment ❺

The assignment of cross-peaks is made in an indirect fashion by progressively eliminating unlikely assignment possibilities. Due to the r^{-6} dependence, assignments with large distances contribute only little to the NOE intensity. Thus, for a particular cross-peak, each assignment possibility is weighted by its normalised partial volume, w_c, calculated as follows :

$$w_c \propto d_c^{-6} \tag{2.8}$$

$$\sum_{c=1}^{N_c} w_c = 1 \tag{2.9}$$

where d_c is the effective distance of the contribution c in the structure ensemble and N_c, the number of contributions for the cross-peak. To reduce the number of assignment possibilities, only the m largest contributions satisfying the following condition are kept:

$$\sum_{1}^{m} w_c \geq p \tag{2.10}$$

where p designates a user-defined *ambiguity cut-off* that is set to 1.0 in the first iteration. This cut-off is gradually decreased to 0.8 to obtain almost unambiguous assignments at the last iteration.

Restraints merging ❻

Symmetry peaks or duplicate peaks from different experiments lead to equivalent restraints (restraints involving the same set of atoms). To avoid overrepresentation of certain distance data, non-violated restraints with equivalent atom content are detected. The restraint with the smallest distance is kept, while the others are discarded for the rest of the protocol.

Structure calculation ❼

On the basis of the merged restraints list, a new structure ensemble is calculated with the program CNS [Brünger et al., 1998] through a molecular dynamics simulated annealing (MDSA) protocol. ARIA provides two forms of molecular dynamics : in *Cartesian* or *torsion angle* space. As explained in section 1.3, torsion angle molecular dynamics (*TAD*) reduces the calculation time and allows for higher MDSA temperature, while generally increasing the convergence radius. The molecular structures obtained with TAD also present better local geometries. As shown in figure 2.3, the MDSA protocol is divided in two phases : i) an initial high temperature search phase and ii) a cooling phase where the temperature slowly decreases. The second part of the cooling stage is performed in Cartesian coordinates. The length of the cooling stages determines the slope of the bath temperature cooling function. It has been shown that this parameter plays an important role in the convergence properties of the ARIA calculation for highly ambiguous data [Fossi et al., 2005c]. The MDSA protocols implemented in ARIA [Nilges and O'Donoghue, 1998] are optimised for the application of ambiguous distance restraints and for the violation analysis method. The minimisation protocols are mostly based on separate scaling of different energy terms with relatively low force constants (see Fig. 2.3) and error tolerant restraint potential (*soft-square*, § 1.3.2). Any other structural constraints available may also be used during the structure calculation (hydrogen bonds restraints, dihedral angles and RDCs).

The treatment of unassigned prochiral group is realised with a floating chirality assignment approach [Folmer et al., 1997]. The two substituents of a prochiral center (methylene protons

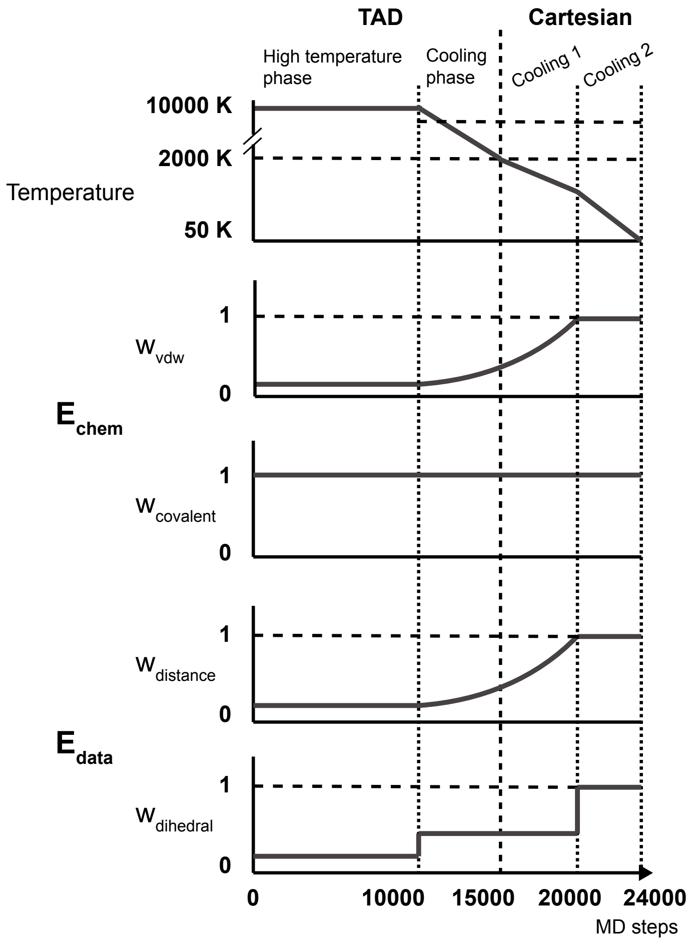


Figure 2.3: Temperature and energy constants in the molecular dynamics simulated annealing protocol in ARIA. A first SA in torsion angle space is performed, followed by a low-temperature cooling stage in Cartesian coordinates. The weights w_{vdw} and $w_{covalent}$ designates the energy scale for the non-bonded and bonded interactions. The weights for distance and dihedral angle restraints are $w_{distance}$ and $w_{dihedral}$, respectively. The Cartesian cooling is separated in two stages (cooling 1 and cooling 2) with different cooling rates. As a result of the longer time-step employed in TAD (27 fs), the effective length of the TAD period is actually 9-time shorter. For the sake of consistency, the lengths of the different phases are here expressed in number of Cartesian MD steps.

or methyl protons of isopropyl groups) are often difficult to assign stereo-specifically, in terms of chemical shifts. In each proton dimension, a resonance matching one of the chemical shifts may potentially involve either of the two prochiral substituents. In ARIA, the two assignment alternatives are tested in the course of the structure calculation and the most energetically favourable possibility is used.

The number of calculated conformers is an important parameter of the structure calculation protocol. Each structure is calculated independently, starting from an elongated polypeptide chain with randomised backbone torsion angles. Among all calculated conformers, only the *n*-lowest energy ones (usually *n*=30%) will be used in the next ARIA iteration to re-calibrate and re-assign NOEs.

Refinement in solution ❽

The simplified force field parameters for non-bonded contacts (cf. § 1.3) applied in structure calculation in vacuum often produces structures that contain artefacts (unrealistic side-chain packing and unsatisfied hydrogen bond donors or acceptors). Therefore, the final structures of the last ARIA iteration are automatically refined in a shell of explicit solvent (water or DMSO molecules). This refinement consists in a short MD with a complete force field, which includes electrostatics and Lennard-Jones potential. The parameters applied in the refinement [Linge and Nilges, 1999] are consistent with the force field used for structure calculation and validation, avoiding systematic differences that could influence validation results. It has been shown that the refinement in solution significantly improves structure quality [Linge and Nilges, 1999; Linge et al., 2003b].

2.1.2 Bookkeeping in ARIA

To facilitate data validation and integration, ARIA uses a data format based on the extensible markup language (XML) to describe molecular systems, chemical shifts and NOE cross-peaks. Since most NMR software packages use proprietary formats for data storage, the interconversion step required to transfer data with other applications like ARIA can lead to a loss of information. The Collaborative Computing Project for NMR (CCPN) has developed a data model [Fogh et al., 2005] to store, in a common framework, all information emerging in a structure determination project. ARIA integrates an interface to communicate with the CCPN data model. In that way, NMR data, such as chemical shifts, NOE peak-lists and assignments or restraints lists, can be imported directly from a CCPN project into ARIA. Moreover, assigned peak lists, restraint lists, and the structure ensembles calculated by ARIA are also automatically exported to the CCPN data model (Fig. 2.4).

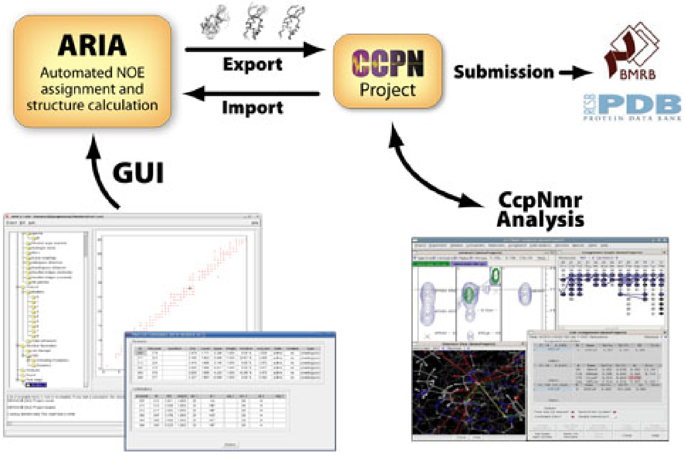


Figure 2.4: Integration of ARIA with the CCPN data model : an ARIA project can be prepared directly through the GUI with NMR data from a CCPN project. ARIA results are exported back to CCPN that offers automated routines to facilitate the data submission.

2.1.3 General ARIA implementation

ARIA (version 2) is written in the programming language Python[2]. The modular and highly object-oriented design of the program facilitates the addition of new features, such as the ones presented in the next sections. For computationally intensive matrix operations, ARIA takes advantage of the Python extension Numpy[3], which employs optimised C and Fortran libraries.

---

[2] http://www.python.org/

[3] http://numpy.scipy.org/

For setting-up a project, ARIA offers a *graphical user interface* (GUI) (Fig. 2.4) written in Python and based on the Tcl/Tk[4] and Tix graphics libraries.

2.2 Network anchoring in ARIA

The network anchoring approach was proposed [Herrmann et al., 2002] to reduce the number of possibilities of NOE peak assignment during the iterative NMR structure refinement with CANDID and CYANA [Güntert, 2004]. The approach is based on the ranking of each NOE assignment, using the information about the assignments of neighbouring nuclei in 3D space, and is efficient because the true assignments form a self-consistent subset of the network of all possible assignments (Fig. 2.5). In this work, the network anchoring approach was implemented in ARIA to improve the standard NOE assignment protocol.

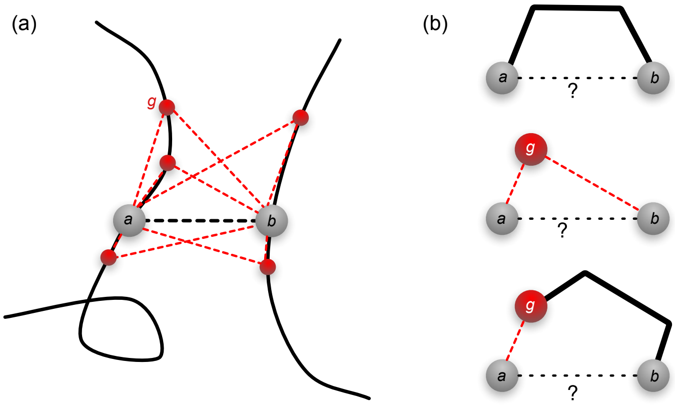


Figure 2.5: Illustration of the network anchoring procedure: (a) for a given NOE assignment possibility (between protons *a* and *b*), all the protons *g* in the same or neighbouring residues of either *a* or *b* and that are simultaneously connected to *a* and *b* are searched. (b) the connections *a*-*g* or *b*-*g* can be either another assignment possibility (red dashed lines) or an expected interaction with respect to the covalent structure (solid black lines). Adapted from Güntert [2004]

The initial network of all possible assignments is first extended with information from the covalent structure of the molecule, coming from the NOEs that are observable independently of the 3D structure. An XML (eXtensible Markup Language) file contains the list of all possible spin pairs separated by at most two dihedral angles in proteins, and whose distance is smaller than 5.5 Å whatever the dihedral angle value. The core of the list was built from Wüthrich et al. [1983]; in case of pseudo-atoms, the distances were re-evaluated from an exhaustive sampling of the conformations of the amino-acids and amino-acid pairs, using the CNS [Brünger et al., 1998] topology definition and only allowing dihedral angle rotations. This list can be easily supplemented by user-defined spin-pairs, like short proton-proton distances observed in regular secondary structure elements. Moreover, distance restraints imported from a CCPN [Fogh et al., 2006] project can be added to the network of possible assignments. The CCPN restraints may be also filtered out by the network anchoring procedure.

The usual contribution weight of the ARIA protocol, which serves as a criterion to eliminate the least likely possible assignments, is then modified as:

$$w = w_d . w_c \tag{2.11}$$

---

[4] http://www.tcl.tk/

where w_d is the original contribution weight derived from the distances observed in the structure ensemble, and w_c is the network anchoring score of the contribution c, calculated as follows.

Since pseudo-atoms are not used in ARIA, each assignment possibility (contribution) of a particular cross-peak may contain multiple pairs of spins. For each pair of spins a, b of a contribution, all the atoms (g) in the neighbourhood of either atom a or b and connected by the network to both a and b atoms are collected. The spin-pair network anchoring score w_{ab} is then computed from the weights of the connections a-g and g-b:

$$w_{ab} = \sum_g \sqrt{\nu_{ag}\nu_{gb}} \tag{2.12}$$

The definition of the weight ν_{ag} for the connection a-g is identical to the one proposed in CANDID [Herrmann et al., 2002].

The network anchoring score w_c of a contribution c is then calculated as the average of the network anchoring scores obtained for each spin-pair a, b, belonging to the contribution:

$$w_c = \left(\sum_{a,b} w_{ab} \right) / n_{ab} \tag{2.13}$$

where n_{ab} is the number of spin pairs in the contribution.

Residue-wise (S_{res}) and atom-wise (S_{atom}) network anchoring scores, calculated in the same way as in CANDID, are used to detect possible erroneous cross-peaks. A peak is conserved if one of the following rules is verified:

$$S_{res} \geq N_{res}^{high} \tag{2.14}$$

$$S_{res} \geq N_{res}^{min} \ \& \ S_{atom} \geq N_{atom}^{min} \tag{2.15}$$

where N_{res}^{high}, N_{res}^{min} and N_{atom}^{min} are user-defined thresholds. N_{res}^{high} corresponds to the minimal network anchoring score for a peak to be conserved. N_{atom}^{min} and N_{res}^{min} can be seen as combined thresholds to determine whether the network anchoring scores per atom and per residue, respectively, are sufficient to consider a peak as reliable.

The efficiency and impact of the network anchoring protocol in ARIA has been evaluated for automated structure determination of a monomeric protein (§ 3.1.2) and symmetric homo–dimers as well (§ 4.1).

2.3 Symmetric homo–dimer calculation

As described earlier, the symmetry degeneracy in NMR spectra usually impedes automated NOE assignment and structure calculation of symmetric oligomers. In order to overcome these difficulties, a method based on Ambiguous Distance Restraints was proposed few years ago [Nilges, 1993]. In this *symmetry-ADR* strategy, the problem of assigning inter- and intra-monomer NOEs is circumvented by adapting the distance target function to reflect the NOE signals arising in a symmetric homo-oligomers, where both intra- and intermonomer effects contribute to the signal.

If we considered the simple case of a symmetric homo-dimer, for each peak observed in the spectrum at the chemical shifts corresponding to the spins i and j, the following ambiguous inter- and intra-monomer restraints will be applied [Nilges, 1993]:

$$\frac{1}{{d_{ij}}^6} = \frac{1}{(d_{ij}^{intra})^6} + \frac{1}{(d_{ij}^{inter})^6} \tag{2.16}$$

where d_{ij}^{intra} is the intra-monomer distance between the spins i and j, and d_{ij}^{inter} is the inter-monomer distance between the spins i and j.

Nevertheless, this ADR formalism is not sufficient for the structure calculation of symmetric homo-oligomers. The symmetry properties of the molecule are also taken into account by adding a *symmetry target function* to the total energy E_{hybrid} :

$$E_{hybrid} = E_{chem} + E_{data} + E_{sym} + E_{NCS} \tag{2.17}$$

where E_{sym} ensures the symmetry of the molecule (C_2 for dimer) and E_{NCS} minimises the RMSD between the chain coordinates. To avoid the problem of positioning the symmetry axes, the given symmetry of the molecule is imposed through *distance symmetry restraints* instead of constraining the symmetry [Nilges, 1993]. The minimisation of the RMSD between the monomers is achieved by *non-crystallographic symmetry* (NCS) restraints. Additionally, a packing restraint is applied between the centres of mass of each chain to keep the chains in the vicinity of each other. This restraint is progressively reduced from 10 kcal/mol to zero during the simulated annealing protocol.

The data model in ARIA2 was adapted to allow for the definition of a symmetric molecule, and to provide a complete implementation of the symmetry-ADR method for NMR structure calculation of symmetric homo–dimers. In order to take advantage of filtered experiments, ARIA offers an option to specify the content of each NOESY spectrum (intra-monomer peaks only, inter-monomer only or chain ambiguous). The chain ambiguity corresponds to known spin assignment for all peaks, only the distinction between intra- and inter-molecular assignment being ambiguous. Moreover, this standard input for a symmetric homo–dimer calculation can be supplemented by information about unambiguously assigned inter–monomer peaks. In the implementation developed in this work, ARIA uses simple rules [Duggan et al., 2001] to assign some NOEs as inter–monomer before the structure calculation, using the secondary structure predicted by the Chemical Shift Index [Wishart et al., 1992]. If two secondary structure elements are facing at the interface, NOEs observed within the same element between residues separated by more than five residues in sequence cannot arise from intra-molecular contacts, and are thus unambiguously classified as inter-molecular.

2.4 Adaptation to solid-state MAS NMR data

One of the biggest challenges in both rare-spin and proton-mediated correlation experiments remains the unambiguous assignment of the peaks to distance restraints between two spins. In solid-state NMR, detection on rare spins (^{13}C or ^{15}N) yields line widths of about 0.5 to 1 ppm. Compared to solution NMR, the number of isolated cross-peaks is greatly reduced, and the size of the chemical shift tolerance window must be increased. As a consequence, the majority of cross-peaks in ^{13}C-^{13}C or ^{15}N-^{13}C correlation spectra remain ambiguously assigned. The use of Ambiguous Distance Restraints in the context of solid-state NMR data has been previously explored for the assignment of selectively labelled proteins using a dedicated algorithm (SOLARIA) [Fossi et al., 2005a]. To complete this work, we have adapted the program ARIA for the automated assignment and structure calculation of protein from proton-mediated homonuclear (CHHC/NHHN) and heteronuclear (NHHC) correlation spectra on uniformly [^{13}C,^{15}N] labelled samples.

As in these experiments proton-proton interactions are indirectly detected on rare spins, the peak assignment and restraint generation routines of ARIA were modified to allow the use of ^{13}C and ^{15}N chemical shifts to create ^{1}H-^{1}H distance restraints. The rest of the general program workflow of ARIA is unchanged. In this type of solid state experiments, cross-peaks intensities may be affected by more parameters compared to solution NMR (§ 1.2.5). Moreover, medium- and long-range contacts may be more subject to relay mechanisms than intra-residue or sequential contacts, as already observed in solution NMR. Consequently, the model of an isolated spin pair is not valid anymore, preventing a detailed comparison of the magnitude of the observed cross signals. In this context, the application of a uniform calibration, as it is done for NOESY data (§ 2.1.1), to convert peak volumes into distance constraints might lead to a misinterpretation of the experimental data. The standard calibration process of ARIA was thus reprogrammed to allow the specification of fixed restraints bounds (lower and upper) for each spectrum. This functionality permits the utilisation of multiple peak list corresponding to different distance classes (generally estimated from build-up curves). Nevertheless, the function used in ARIA to weight each assignment possibility from the measured distance (§ 2.1.1) still exploits the inverse sixth power model. It has been shown, in fact, that for short proton mixing times in N/CHHC experiments, the signal intensity correlates well with the inter-proton distance with an $1/r^{-6}$ proportionality [Lange et al., 2003] .

As a consequence, the modified software is also able to handle C-C correlation spectra (like PDSD and DARR [Takegoshi et al., 2001]) recorded on uniformly labelled samples. A general purpose version of ARIA for solid-state NMR is currently being developed with the incorporation of the special functionalities of the program SOLARIA [Fossi et al., 2005a] (initially developed on the previous version of ARIA). For ssNMR experiments on site-directed ^{13}C-enriched samples, the management of labelling schemes would be entrusted to the CCPN data model [Vranken et al., 2005].

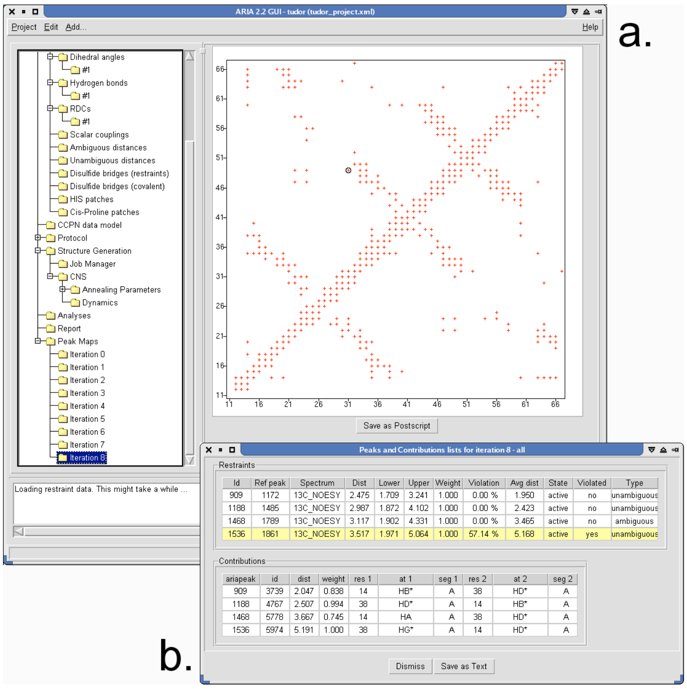
2.5 Graphical analysis of NMR structural quality

The current state of the ARIA protocol including ambiguous assignments and distance violations is summarised in several report files located in each iteration directory. Analysing such text files is difficult since they contain a large number of data. ARIA was thus extended to allow the generation of an *interactive contact map*, which provides a detailed analysis of the restraints and restraint contributions. Moreover, analysing the quality of NMR structure is a key step into the validation of an ARIA calculation. In that respect, it was recently shown [Nabuurs et al., 2006] that profiles of quality scores calculated on individual residues along the biomolecular structure can be essential to detect possible sources of error in the spectral assignment. Several extensions of ARIA were therefore implemented in order to generate postscript files describing the structural quality and the restraint violations at the residue level.

2.5.1 Interactive analysis of peaks assignments

In each iteration, the current assignments are stored in the form of a binary file that can be analysed afterwards. An additional section in the GUI provides a way to read back the assignments and display them as a clickable contact map. This map is defined as a Tk canvas widget and each pixel is clickable to present additional information about this particular contact. A pop-up window displays the corresponding assignments in tables that can be exported as text files. The peak map can be saved in Postscript format. For each ARIA iteration, the interactive peak map displays the pairs of residues involved in one or more assignment possibilities. Such maps can be generated from the current state of the assignment using three classes of restraints: (i) all restraints, (ii) ambiguous restraints and (iii) unambiguous restraints. Clicking on a pixel located at the position (i, j) on the map (Figure 2.6a) opens a pop-up window (Figure 2.6b) that shows a list of ARIA restraints involving atoms from residues i and j. It also gives information about each assignment possibility (contribution) of these restraints. Multiple pixel selection is possible. The restraints lists are displayed in tables, indicating different parameters such as the target distance, the percentage of structures in which a restraint was violated or the average distance found in the structure ensemble. A colour code indicates whether a restraint is globally violated. For each assignment possibility, the table indicates the relative weight, the effective distance as well as the description of the pairs of atoms involved.

This interactive tool allows the user to get a detailed analysis of the peak assignment procedure at each step of the ARIA protocol. Since the results are presented as a two dimensional map, this tool significantly extends the information content with respect to the standard ARIA reports. Moreover, the dynamic and graphical nature of the map may allow a rapid detection of possible errors in the assignment process, or of potential inconsistencies in the data.



Loading restraint data. This might take a while ...

Restraints

Id	Ref peak	Spectrum	Dist	Lower	Upper	Weight	Violation	Avg dist	State	Violated	Type
909	1172	13C_NOESY	2.475	1.709	3.241	1.000	0.00 %	1.950	active	no	unambiguous
1188	1485	13C_NOESY	2.987	1.872	4.102	1.000	0.00 %	2.423	active	no	unambiguous
1468	1789	13C_NOESY	3.117	1.902	4.331	1.000	0.00 %	3.465	active	no	ambiguous
1536	1861	13C_NOESY	3.517	1.971	5.064	1.000	57.14 %	5.168	active	yes	unambiguous

Contributions

ariapeak	id	dist	weight	res 1	at 1	seg 1	res 2	at 2	seg 2
909	3739	2.047	0.838	14	HB*	A	38	HD*	A
1188	4767	2.507	0.994	38	HD*	A	14	HB*	A
1468	5778	3.667	0.745	14	HA	A	38	HD*	A
1536	5974	5.191	1.000	38	HG*	A	14	HD*	A

Figure 2.6: Interactive peak map (a) Right panel of the ARIA 2.2 GUI showing the interactive peak map at the iteration 8 of an ARIA run. Each pixel of the map located between residues i and j is clickable and an assignment report (b) can be opened, containing the list of ARIA peaks existing between the residues i and j, along with their contributions.

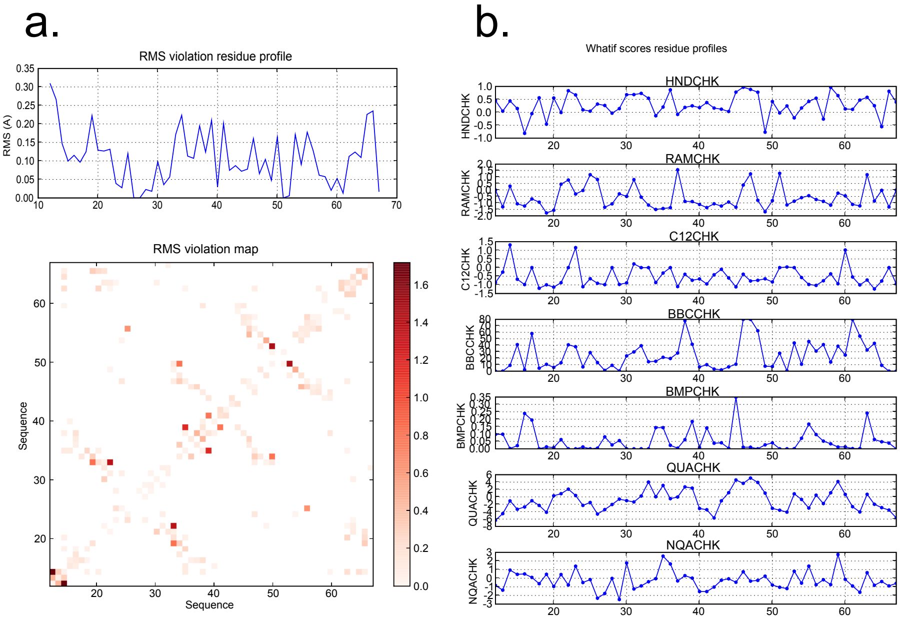


Figure 2.7: Per-residue quality plots (a) Contact map displaying the sums of RMS deviations and profile of the RMS deviations (b) WHAT IF score profiles along the sequence. The RMS deviations are plotted with a colour scale.

2.5.2 Per-residue structural quality

Postscript files describing (i) the restraints, through the RMS of deviations from the distance bounds, and (ii) the structure quality, through WHAT IF [Hooft et al., 1996] scores are automatically created at the end of each iteration or after the final structure analysis. These parameters are displayed at the residue level, in the form of a profile along the protein sequence, or as a contact map for the RMS deviations per residue pair. The graphics are plotted with the matplotlib[5] plotting library interfaced with Python. Quality and RMS profiles data are also stored in formatted text files for further use. The contact map displays the sum of the RMS deviations (Figure 2.7a) per residue pair. In the profiles, the sum of the RMS of violations per residues and the mean values over the conformers of the WHAT IF scores are plotted along the protein sequence (Figure 2.7b). The most informative WHA TIF scores are plotted, such as the packing quality Z-score (1st and 2nd generation), inter-atomic bumps as well as the backbone conformation Z-score.

An essential part in the validation of an ARIA calculation is the analysis of the quality of the NMR structures. Classically, the overall number of violations or the RMS deviations in addition to global WHAT IF scores of the whole molecule are used to assess the quality of a structure. In the light of recent investigations [Nabuurs et al., 2006], it is clear that these global parameters may not suffice to readily detect errors in the local or global fold of a protein. The analysis of quality scores of each residue along the molecular sequence is essential to precisely detect possible sources of error in the spectral assignment. The automated generation of per-residue profiles for RMS deviations and for WHAT IF scores provides a highly integrated tool to rapidly identify regions of the structure that exhibit abnormal quality factors, and where restraints and assignments should be more thoroughly investigated.

---

[5] http://matplotlib.sourceforge.net/

3

Influence of automated NOE assignment and NMR structure calculation protocols on structural quality

The purpose of this chapter is twofold : (i) assess the efficiency of the network anchoring to automatically assign NOEs inside the ARIA approach (§ 2.2), (ii) compare different torsion angle structure calculation protocols with assigned NOE-derived distance restraints. Each method has been evaluated by analysing the quality of the final obtained structure ensemble.

Network anchoring was designed to speed up the structure determination process by reducing the NOE data ambiguity in the early stages of the iterative protocol. This is achieved by searching for data corroborating each possible assignment, using the fact that the distance restraints should be mutually consistent. It was therefore interesting to investigate how this method influences the NOE assignment process in ARIA and its impact on the structural quality.

When restraints have been correctly assigned, structures calculated with the standard NMR protocol with simplified force-field often exhibit defects. Further improvement of the structures is generally achieved by additional in-water MD simulations under normal temperature with standard force-fields [Linge and Nilges, 1999; Linge et al., 2003b]. The molecular dynamics algorithm ICMD (§ 1.3.4), with the variable target function and a fast TAD algorithm allows to use a general purpose force-field at all stages of structural refinement. It also permits to obtain, in gas phase, structures of similar quality than the ones refined in water. A simulated annealing protocol was designed for the classical problem of NOE-based determination of protein structures. We also performed a thorough comparative test on a representative group of proteins earlier studied by the state-of-the-art methods. For such comparison, a set of eight proteins was selected from the database RECOORD [Nederveen et al., 2005]. This database resulted from an effort of protein structure re-calculation, with methods corresponding to the state-of-the-art of the NMR structure determination.

3.1 Effect of network anchoring on the structure of the HRDC domain

3.1.1 Calculation schemes

ARIA calculations were performed on the HRDC domain (PDB id: 1D8B [Liu et al., 1999]), with different setups: (i) partially assigned restraints without network anchoring (hrdc$_{ori}$), (ii) totally ambiguous restraints without network anchoring (hrdc$_{nonet}$), and (iii) totally ambiguous restraints with network anchoring (hrdc$_{net}$, hrdc$_{net+}$). The difference between hrdc$_{net}$ and hrdc$_{net+}$ concerns the network anchoring parameters (N_{res}^{high}, N_{res}^{min} and N_{atom}^{min}: Table 3.2), which were more stringent in hrdc$_{net+}$ to increase the assignment efficiency. Network anchoring was applied only during iterations 0 to 3, since the expected gain of convergence is less significant at the next iterations. In each case, the number of calculated conformers is 80 for the first iteration and 60 for the next 8 iterations. The 10 lowest total energy structures of the last iteration were refined in a shell of water [Linge et al., 2003b].

Run	Original assignments	Network Anchoring
hrdc$_{ori}$	✓	-
hrdc$_{nonet}$	-	-
hrdc$_{net}$	-	✓
hrdc$_{net+}$	-	✓

Table 3.1: Summary of the different calculation schemes

	hrdc$_{net}$			hrdc$_{net+}$		
iteration	N_{res}^{high}	N_{res}^{min}	N_{atom}^{min}	N_{res}^{high}	N_{res}^{min}	N_{atom}^{min}
0	6.0	1.0	0.5	6.0	1.0	0.75
1	4.0	1.0	0.25	6.0	1.0	0.6
2	4.0	1.0	0.25	4.0	1.0	0.25

Table 3.2: Network anchoring parameters for the HRDC calculations

3.1.2 Results

A comparison of the superimposed conformers of the HRDC domain for the different runs (Fig. 3.1) illustrates the acceleration of the convergence due to network anchoring, as already shown by Herrmann et al. [2002]. This acceleration was also analysed by calculating the RMSD values determined at each iteration between the atomic coordinates of all calculated conformers (Fig. 3.2(a,b) “Precision”) or between the conformers and the deposited PDB structure (Fig. 3.2(c,d) “Accuracy”). The RMSD values are smaller in presence (hrdc$_{net}$, hrdc$_{net+}$) than in absence (hrdc$_{nonet}$) of network anchoring for all iterations; the largest difference being in the 2-3 Å range

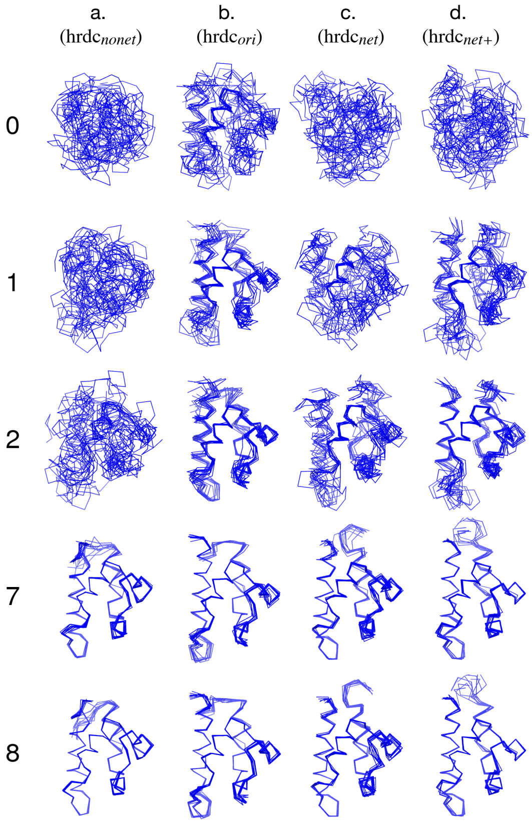


Figure 3.1: Efficiency of network anchoring in ARIA 2.2. The example shown here is run on the domain HRDC (PDB id: 1D8B [Liu et al., 1999]) for the following conditions: no prior NOE assignment (hrdc$_{nonet}$), partial NOE assignments (hrdc$_{ori}$), network anchoring without prior NOE assignment (hrdc$_{net}$ and hrdc$_{net+}$). The HRDC backbone of the 10 best conformers obtained at iteration 1,2,3,7 and 8 of the runs were superimposed.

for the iterations 1 to 3. Nevertheless, the use of partial manual assignments (hrdc$_{ori}$) gives the smallest RMSD values for the iterations 1 to 6. The final RMSD of the ensemble in the presence of network anchoring is slightly smaller than the one observed in the absence of network anchoring, even with manual assignments. The comparison of the RMSD between the mean calculated structures and the reference PDB structures in the presence and in the absence of network anchoring shows a decrease of about 0.4 Å if network anchoring is applied (Fig. 3.2(c,d)). The computational cost of the network anchoring analysis in the first iteration (30.1 sec) is counterbalanced by a substantial reduction of the conformer generation time (278.5 sec/structure vs. 524.9 sec/structure without network anchoring on a 64 bits CPU at 2.8 GHz).

The quality of the final sets of conformers were compared using PROCHECK [Laskowski et al., 1993] v3.5.4 and WHAT IF [Hooft et al., 1996] v6.0, as well as the statistics on restraint violations (Table 3.3). The PROCHECK core percentage increases if network anchoring is applied with more stringent parameters (hrdc$_{net+}$), whereas the other PROCHECK percentages decrease. However, the largest PROCHECK core and the smallest PROCHECK allowed percentages are obtained for the run with partial manual assignment (hrdc$_{ori}$). Similar numbers of bad contacts are determined by PROCHECK in all cases.

Concerning the WHAT IF scores, the best values are generally obtained for the run with partial manual assignments (hrdc$_{ori}$). Otherwise, similar WHAT IF Z-scores values are reported in all runs, for the packing quality (NQACHK) and the Ramachandran plot appearance (RAMCHK) whereas the backbone conformation (BBCCHK) Z-scores display worse values in the case of

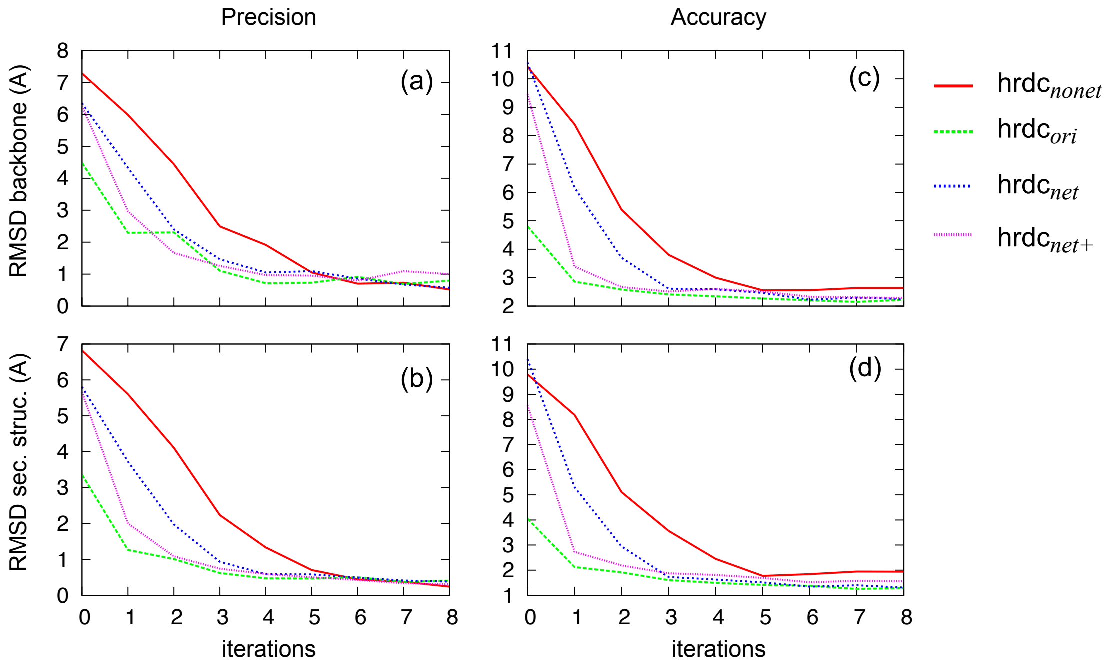


Figure 3.2: Precision and Accuracy per iteration. RMSD from the mean structure of the final ensemble over backbone (a) and over the secondary structure elements (b). RMSD from the mean PDB structure over backbone (c) and over the secondary structure elements (d) The curves are shown for the following conditions, described in Table 3.1: hrdc$_{nonet}$ (red solid), hrdc$_{ori}$ (green dashed), hrdc$_{net}$ (blue dotted), hrdc$_{net+}$ (pink dotted).

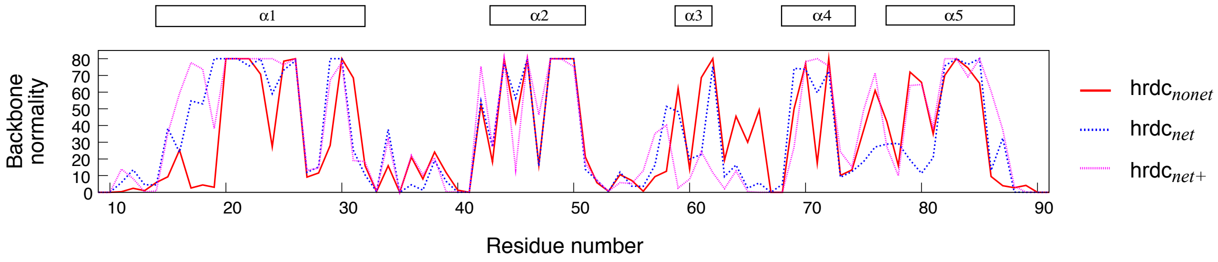


Figure 3.3: WHAT IF backbone normality score (BBCCHK) along the sequence for the HRDC calculation: hrdc$_{nonet}$ (red solid), hrdc$_{net}$ (blue dashed), hrdc$_{net+}$ (pink dotted). Secondary structures are plotted on top.

fully automatic assignment (hrdc$_{nonet}$, hrdc$_{net}$, hrdc$_{net+}$). In all calculations, steric clashes (reported by BMPCHK) are observed in the same regions: between helix $\alpha 1$ and $\alpha 2$ and around residue 60 (belonging to the $\alpha 3$ helix). The number of NOE violations as well as the RMS of NOE violations increase in the case of fully automatic assignment. It is also interesting to note that the analysis of the NQACHK and BBCCHK scores profiles along the sequence reveals differences in the residue range 15-20 (Fig. 3.3), which arise from a shorter $\alpha 1$ helix observed in absence of network anchoring. In fact, in absence of network anchoring or manual assignments, the N–terminal part of helix $\alpha 1$ is folded as a 3_{10} helix instead of the regular α helix (Fig. 3.4).

To sum-up, the use of the network anchoring without any prior NOE cross-peak assignment allows to obtain structure ensembles with a level of structural quality similar to the structures obtained in absence of initial assignment and of network anchoring. ARIA calculations without network anchoring nor initial NOE assignment converge to a globally correct fold, but display local errors in regions with low NMR information content. These errors are removed by the utilization of network anchoring. One should also notice that the NOE fitting is better if the initial assignment information is used.

	Partial NOE assignments	No prior NOE assignment	No prior NOE assignment with network anchoring	
	hrdc$_{ori}$	hrdc$_{nonet}$	hrdc$_{net}$	hrdc$_{net+}$
PROCHECK[a]				
Most favored	79.6 ± 3.5	75.0 ± 4.2	74.0 ± 4.3	77.8 ± 4.0
Allowed	15.7 ± 3.4	21.8 ± 3.6	19.6 ± 3.7	17.7 ± 3.1
Gener. allow.	1.8 ± 1.4	2.6 ± 1.0	5.3 ± 1.9	3.6 ± 2.2
Disallowed	3.0 ± 1.7	0.7 ± 1.0	1.1 ± 1.1	0.8 ± 0.7
Bad contacts	0.7 ± 0.5	0 ± 0	1.5 ± 0.7	0.2 ± 0.4
WHAT IF scores[b]				
NQACHK	-1.4 ± 0.4	-2.4 ± 0.2	-2.1 ± 0.3	-1.2 ± 0.3
RAMCHK	-4.8 ± 0.2	-5.6 ± 0.4	-5.7 ± 0.4	-5.2 ± 0.5
C12CHK	-2.7 ± 0.3	-3.3 ± 0.2	-3.9 ± 0.4	-3.7 ± 0.4
BBCCHK	-3.8 ± 0.7	-6.8 ± 1.0	-6.5 ± 0.7	-6.1 ± 1.0
BMPCHK	25.7 ± 3.7	48.9 ± 2.6	46.6 ± 4.2	45.9 ± 3.3
NMR ensemble precision[c]				
Backbone RMSD (Å)	0.79 ± 0.23	0.53 ± 0.10	0.61 ± 0.09	0.84 ± 0.18
Heavy RMSD (Å)	1.19 ± 0.22	0.98 ± 0.18	1.05 ± 0.12	1.26 ± 0.20
NOE Violations				
# viol. $\geq$ 0.3 Å	7.0 ± 2.2	56.5 ± 4.9	61.0 ± 5.5	59.5 ± 4.0
RMS violations (Å)	0.07 ± 0.03	0.1 ± 0.02	0.1 ± 0.01	0.1 ± 5.7E-03

Table 3.3: Quality parameters for the monomeric calculations on HRDC domain with or without network anchoring

[a] PROCHECK results: percentage of residues in most favoured region, allowed region, generously allowed region and disallowed region.

[b] The WHAT IF scores are the following: 2nd generation packing quality Z-score (NQACHK), Ramachandran plot appearance Z-score (RAMCHK), $\chi1/\chi2$ rotamer normality Z-score (C12CHK), Backbone conformation Z-score (BBCCHK) and the number of inter-atomic bumps (BMPCHK).

[c] RMSD between the final conformers calculated for the backbone and heavy atoms. Residues 11-91 where used for superposition and RMSD calculation.

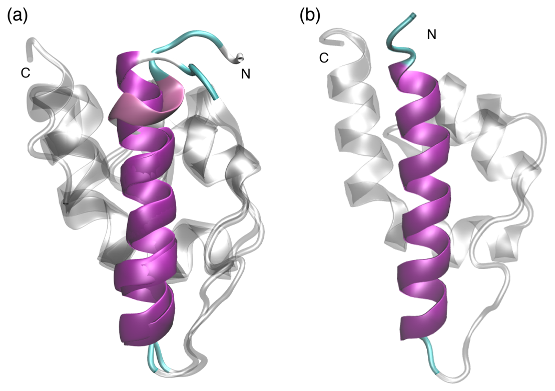


Figure 3.4: (a) Superposition of the mean structures from the runs hrdc$_{nonet}$ and hrdc$_{net}$. In hrdc$_{nonet}$, the N terminal helix ends with a 3_{10} helix (pink) instead of a regular α helix (purple) as in the hrdc$_{net}$ structure. (b) Mean structure of the PDB deposited 1D8B structure [Liu et al., 1999].

3.2 Comparison of different torsion angle approaches for NMR structure determination.

3.2.1 Calculation schemes

The set of proteins used

The set of eight proteins used to analyse the results obtained with ICMD and to compare the ICMD conformations with those stored in the RECOORD database are ribosomal proteins with sizes in the range of 60-104 residues (Table 3.4). They are thus medium-size proteins with respect to the sizes which are now attainable in NMR structural studies [Miclet et al., 2003]. Most of the proteins are α-β proteins (L25, L30, L11, L23 and S19), and there are one α protein (L20) and two β proteins (S28e and S27e). The NOE completeness[1] is in the 0.31-0.64 range, but half of the proteins (L30, S19, S27e and L20) display a NOE completeness smaller than 0.45, and thus belong to the low completeness region [Doreleijers et al., 1999]. Proteins L11, L23, S28e and S19 contain long (about or more than 10 residues) loops and/or N or C terminal tails. These regions may display larger internal flexibility. The structures of the majority of these proteins, except L25 and L30, were determined only by NMR. X-ray [Lu and Steitz, 2000] and NMR [Stoldt et al., 1999] structures of the L25-RNA complex were determined. The structure of the protein L30 isolated [Chen et al., 2003] was determined by X-ray crystallography, and the structure of the L30-RNA complex was determined from a joint X-ray and NMR refinement [Chao and Williamson, 2004]. Backbone dihedral angle restraints are available only for the

Name	Size[a]	α(%)[b]	β(%)[b]	NOE_c[c]	PDB id[d]	N_{float}[e]	N_{adr}[f]	RMSD[g]	NOE_{res}[h]
L25	94	9.0	43.5	0.54	1b75	106	99	8-76	16.4
L30	104	37.6	19.5	0.31	1ck2	101	111	10-101	14.2
L11	76	45.9	5.9	0.49	1fow	110	100	8-76	12.0
L20	60	57.1	0.0	0.42	1gyz	112	109	1-60	13.1
L23	96	20.6	27.1	0.50	1n88	104	100	7-60,75-96	17.7
S28e	82	0.5	33.8	0.64	1ny4	114	109	7-56	21.1
S19	92	11.3	9.4	0.36	1qkh	107	137	3-67	15.1
S27e	66	0.2	30.0	0.37	1qxf	134	139	4-53	11.5

Table 3.4: Set of the proteins used for the comparison.

[a] The size of the protein is expressed as the number of residues.

[b] The percentage of α and β secondary structures calculated by DSSP [Kabsch and Sander, 1983].

[c] The NOE completeness taken from the database RECOORD [Nederveen et al., 2005].

[d] References : 1b75 [Stoldt et al., 1998], 1ck2 [Mao and Willamson, 1999], 1fow [Markus et al., 1997], 1gyz [Raibaud et al., 2002], 1n88 [Ohman et al., 2003], 1ny4 [Aramini et al., 2003], 1qkh [Helgstrand et al., 1999], 1qxf [Herve du Penhoat et al., 2004]

[e] Number of conformations calculated using the floating restraints.

[f] Number of conformations calculated using the ambiguous restraints.

[g] Residue ranges used in the conformation superposition for the RMSD calculation.

[h] Number of NOE restraints by residue.

---

[1] Ratio between the number of observed and the number expected restraints.

proteins L30, L11, L23 and S28e. Sidechain χ_1 angle restraints were also measured on L30. The dihedral angles given for L23 were automatically determined using TALOS [Cornilescu et al., 1999]. For each protein, about one hundred conformations were calculated with the SA protocol described in §3.2.1. The exact numbers of conformations are given in Table 3.4.

RECOORD [Nederveen et al., 2005] is an effort of protein structure recalculation, based on the most recent protocols and software available (CYANA [Herrmann et al., 2002] and CNS [Brünger et al., 1998] using ARIA simulated annealing scripts) and further refined in explicit water [Linge et al., 2003b]. For each entry, the RECOORD database is constituted of four structure ensemble sets:

CNS : models recalculated in CNS (with a simplified force-field)
CYA : models recalculated in CYANA (with a simplified force-field)
CNW : **CNS** models water-refined in CNS
CYW : **CYA** models water-refined in CNS

NMR restraints

In the case of ambiguous distance restraints (more than one atom in one of the two atom groups involved in the restraints), the ICMD approach provides two different procedures to restrain the distance. In the first procedure (*floating restraint*), all distances between all possible atom pairs are calculated, and the restraint is applied to the smallest distance. This *swapping* is repeated each time the restraint list and/or forces are changed (see below). This simple approach is considered as standard in the ICMD variable target function protocol. The second procedure (*adr*) makes use of the concept of Ambiguous Distance Restraint [Nilges, 1993]) to determine the effective distance (see Eq. 1.16, § 1.4.2)

Distance restraints are applied using the variable target function algorithm [Braun and Go, 1985], modified as follows. The *torsion angle separation* (TAS) is defined as the number of torsion angles connecting a given pair of atoms. The corresponding TAS values are assigned to all atom pairs potentially involved in distance restraints. The variable target function calculations starts from TAS=1 and the following TAS levels are added consecutively. In the case of *floating* restraints, the swapping procedure is applied before every cycle of dynamics or minimisation. All NOE restraints are checked and the atom pair corresponding to the smallest distance chosen; the TAS of this atom pair is next used to decide if the corresponding restraint should be applied at the current stage of the protocol. In the case of *adr* restraints, the contributions to each restraint is introduced successively, according to their corresponding TAS value.

Molecular dynamics simulations

Both NOE distance and dihedral angle restraints (when available) were applied using a square-well potential (§ 1.3.2). Two forms of restraints (*floating* and *adr*) were tested. This second method (*adr*) was introduced for better comparison because the reference structures from the

RECOORD database were obtained with the ambiguous form of restraints. We note that the *floating* restraint can be considerably more stringent than the *adr* restraint.

The NMR-based molecular dynamics simulations were performed in the gas phase. The force-field parameters were taken from the AMBER all-atom parameter set [Cornell et al., 1995], and the covalent geometry of amino-acids was derived from the CNS paramallhdg5.2 parameters [Linge and Nilges, 1999]. The non-bonded van der Waals energy was modelled through the Lennard-Jones potential with simple spherical cutoff. The Coulomb electrostatic energy was computed with a force-shift truncation method [Levitt et al., 1997]. The cutoff distance was 6 Å for all non-bonded interactions.

Simulated annealing protocol

The numerical integration of ICMD equations of motion employed an implicit leapfrog integrator in torsion angle space, which makes possible the use of 10fs a time-step [Mazur, 1997]. This time-step is optimal for torsion MD of biopolymers with standard atom parameters [Mazur, 1998b]. The temperature coupling is performed using the Berendsen algorithm [Berendsen et al., 1984]. The parameters of the simulated annealing protocol for NOE-based structure calculation are given in Figure 3.5. The starting protein conformation is a random pseudo-extended polypeptide chain. The protocol starts with a *high temperature unfolding phase* at 7000K. The second step is the *variable target function annealing phase* (3000 K). Here, for each level of torsion angle separation (TAS), the energy constant of the NOE restraints is increased up to 10 kcal mol^{-1} Å^{-2}. The following condition *C* is then checked: the value R of the restraint energy per restraint is larger than 1.2, and the TAS level was not processed more than three times. The larger the ratio R, the larger is the number of unsatisfied restraints. If the condition *C* is true, the variable target annealing phase is again performed with the same TAS level. If the condition is false, the next TAS level is processed. After the last TAS levels has been processed, all restraints are applied simultaneously during the *NOE regrouping phase*, where NOE and dihedral restraint force constants are increased up to 50 kcal mol^{-1} Å^{-2} and 50 kcal.mol^{-1} rad^{-2}, respectively. The *cooling phase* is then taking place: the temperature is reduced by stages from 600K to 0K while the energy constants are kept constant at respectively of 5 kcal mol^{-1} Å^{-2} and 50 kcal.mol^{-1} rad^{-2} for the NOE and dihedral restraints. Finally, two stages of energy minimisation are carried out, each one being preceded by the distance restraint *swapping* for floating restraints.

Structure analysis

The comparison of different calculation protocols was performed by using parameter values averaged over the corresponding ensembles of conformations. The first set of parameters describes the convergence of calculations and the fitting of the NMR restraints, namely, the RMSD between the calculated conformers on selected protein regions (Table 3.4), and the number of

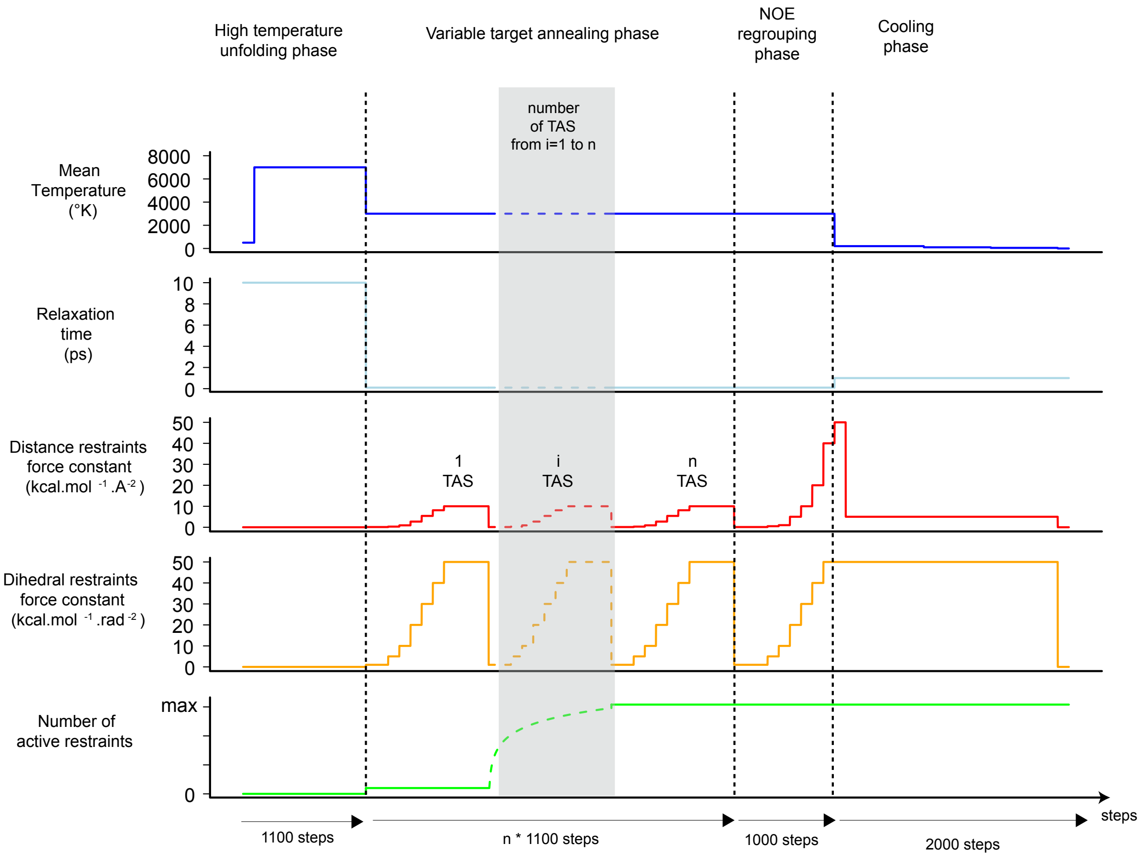


Figure 3.5: Parameters of the simulated annealing protocol used in the ICMD methodology. The grey zone represents the part of the variable target annealing phase repeated for each TAS (Torsion Angle Separation) levels

consistent violations. The second set of parameters describes the quality of the structures obtained, namely, the percentage of residues in the core region of the PROCHECK Ramachandran diagram, three WHAT IF Z-scores, the Ramachandran score, the 2nd-generation packing score and the backbone conformation score, and the number of bumps detected by WHAT IF.

The precision (RMSD$_c$) of the structures ensemble was calculated as the average of the RMSD of each conformation with respect to the mean coordinates of all conformations. The average CNW structure were compared to other RECOORD and ICMD average structures by calculating the RMSD values (RMSD$_d$) between them. The RMSD values analysed here were calculated on the backbone atoms.

The violations of the distances restraints were calculated in two different ways according to the type of restraints used in calculations (*adr* or *floating*). For *floating* restraints, the violations were calculated by taking the distances of all possible pairs and checking if they were larger than U+0.5 Å, where U is the upper bound value of the restraint. If all distances were larger than U+0.5 Å, then the restraint was considered as violated. For ambiguous restraints (*adr*), the distance d (Eq. 1.16) was calculated from the set of distances d_i between the two groups of atoms involved. A restraint was considered as violated if d was larger than U+0.5 Å. The consistent violations of the analysed conformations [Nederveen et al., 2005] were included in the analysis, as well as the RMS of violations, calculated on all violations larger than U.

3.2.2 Analysis of the ICMD structures with floating restraints

The following analysis involves the complete set of ICMD conformations as well as the 25 lowest energy conformations. The total energy is the sum of the ICMD potential energy, and of the restraint energies.

Structure ensemble precision

For all proteins analysed, the RMSD$_c$ values, calculated on the set of the 25 best conformations, are smaller than the corresponding values calculated on the total set of conformations (Fig. 3.6a). When selecting the 25 best conformations, the RMSD changes is smaller than 0.5 Åfor all proteins (Fig. 3.6a) except for S19 (3 Å). This difference for the S19 structure may come from the presence of the long C terminal tail (residues 63-80) which exhibits a non-canonical secondary structure, formed by consecutive loops. Two structure clusters, exhibiting different energies, were observed in the total set of conformations. All proteins, except L11 and S19, have RMSD values over the complete set of conformations smaller than 1.7 Å, indicating that the protocol used by ICMD converged to a well-defined 3D structure. The protein L11 displays a large conformational variability in the loop 19-33: this variability increases the corresponding RMSD.

Distance restraints violations

The number of consistent violations (Fig. 3.6b) varies significantly between different proteins. L25 and L20 show the largest number of violations whereas for L30, L11, S28e and S27e the number of violations is low. The presence and the quality of dihedral angle restraints influence strongly the number of consistent violations. In fact, no dihedral angle restraints were available for the proteins exhibiting the largest number of violations (L25, L20 and S19). Moreover, the dihedral angle restraints for L23 were automatically derived from the TALOS protocol, and may contain incorrect predictions [Cornilescu et al., 1999]. Thus, it seems that the application of reliable restraints on dihedral angles reduces the number of violations. This tendency is probably intensified by the use of larger energy constants for dihedral angles than for distances (Fig. 3.5). The number of consistent violations does not change much over all generated conformations or over the best 25 conformations. This is in agreement with the small improvement of convergence obtained by the selection. The majority of violated restraints involve methyl or methylene protons. In general, one-half of them concerns long-range NOEs. The majority of violations in S19 concerns the C-terminal β strand (residues 56-62) and the C-terminal tail (residues 63-80) of the protein. The majority of other proteins (L25, L30, L23) exhibits violations involving residues located in β strands. Protein L11 shows violations for long-range NOEs between α-helices.

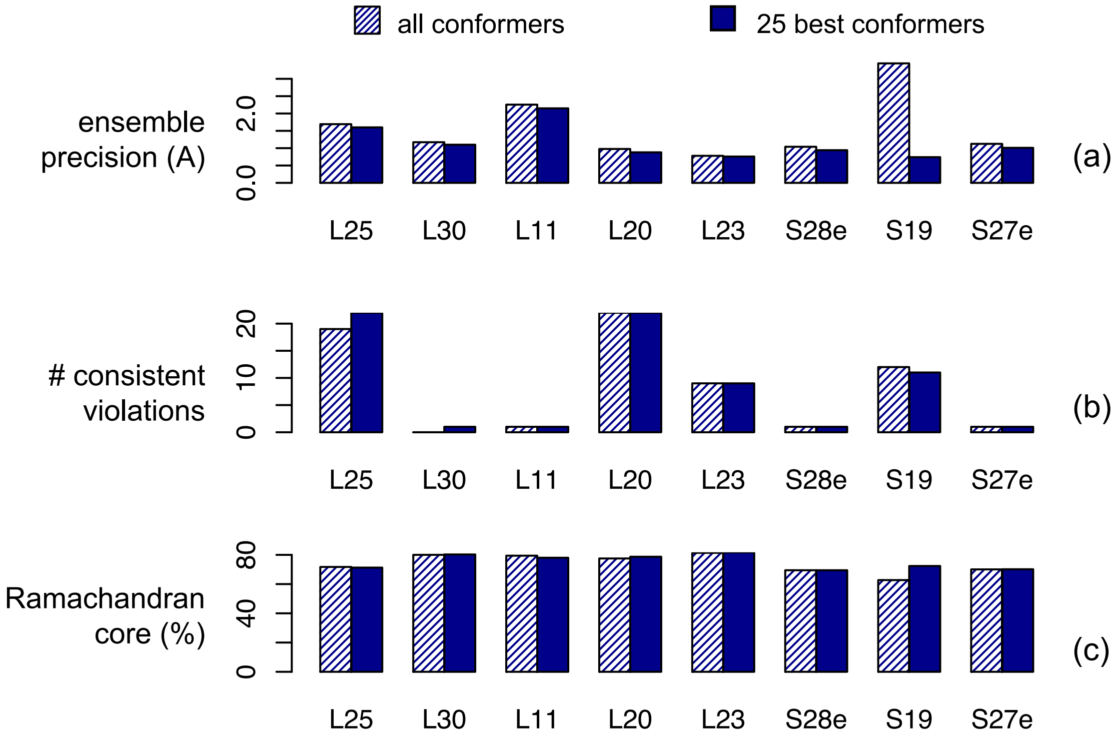


Figure 3.6: ICMD structures with floating restraints : comparison of the ensemble precision (Å) (a), of the number of consistent violations (b) and of the percentage of residues in the core Ramachandran regions (c), for calculations run with ICMD on the set of proteins from Table 3.4 with floating restraints. The results shown are those obtained using all conformations (shaded dark blue) and the 25 best conformations (dark blue). The compared values are mean values calculated on the set of the analysed conformations.

Structural quality

The percentage of residues in the core Ramachandran regions, determined by PROCHECK, is displayed in Fig. 3.6c. All percentage values for the best 25 conformations are larger than 69%. The variation of percentage between all conformations and the 25 best ones is smaller than 2 %, except for S19, and is parallel to the variation of RMSD. Proteins S28e, S19 and S27e display a percentage smaller than 75%. The residues located by PROCHECK in the non-allowed Ramachandran regions were analysed in details for the best 25 conformations. These residues are mainly located in loops, or in the C terminal tails of S19 and S28e.

The conformations obtained with ICMD were also analysed by calculating WHAT IF quality Z-scores (Fig. 3.7). For the majority of proteins analysed here, the Z-scores are larger than -4 and the conformations generated by ICMD are thus of good quality. Also, the Z-scores increase when the 25 best conformations are selected. Z-scores smaller than -4 are observed for L25 (Ramachandran and backbone conformation scores), for L20 (Ramachandran score), for S28e (backbone conformation score) and for S19 (Ramachandran, 2nd-generation packing and backbone conformation scores). For S19, L20 and S28e, the results of WHAT IF are in agreement with the results previously described for PROCHECK. The WHAT IF results were analysed in details for the 25 best conformations, to detect which residues are responsible for lowering the

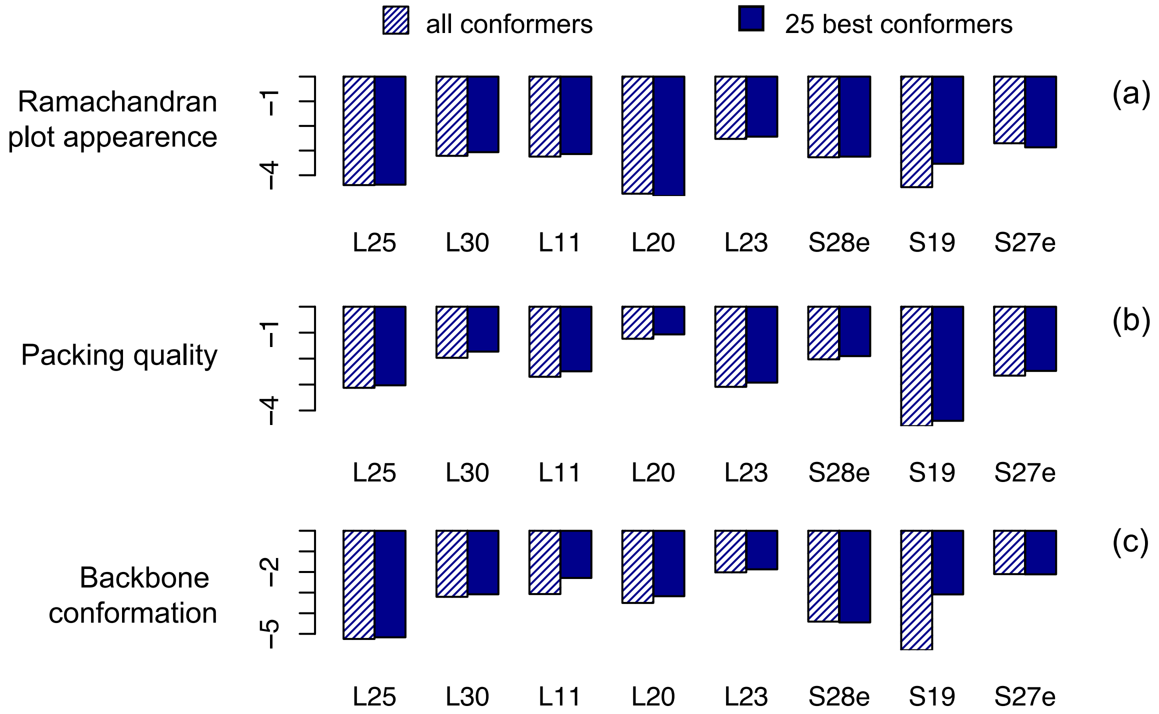


Figure 3.7: WHAT IF Z-scores of ICMD structures with floating restraints: comparison of WHAT IF Ramachandran Z-score (a), of the 2nd-generation packing Z-score (b), and of the backbone conformation Z-score (c) for calculations run on the set of proteins from Table 3.4 with floating restraints. The results shown are those obtained using all conformations (shaded dark blue), and the 25 best conformations (dark blue). The compared values are mean values calculated on the set of the analysed conformations.

backbone conformation Z-score. Similarly to PROCHECK, these residues are mainly located in loops and in the N or C terminal tails. Almost the same set of residues was detected whatever the restraint type (floating or ambiguous) was used.

The selection of the 25 best conformations, based on their total energy, improves the convergence of the calculation, as the RMSD between conformations decreases, and also improves the structural quality, as well as the restraint fit. The convergence and the quality of the obtained conformations are generally good before the selection, except for S19. A larger number of violated NOE restraints are observed for proteins exhibiting no dihedral angle restraints, or automatically predicted dihedral angle restraints.

3.2.3 Ambiguous and floating form of restraints

Comparison of these two different treatments of the restraints presents a non-negligible methodological interest. It should be clear that they were used separately, in different refinement series. With the floating restraints, one always assigns a restraint to a definite pair of protons. This procedure is simple and transparent and it can be employed only in the framework of the variable target function protocol where re-attribution of restraints is made frequently anyway. With ambiguous restraints, many atoms contribute to the restraint simultaneously, which in some cases is equivalent the first approach, and in some cases not. The latter method is standard in the

ARIA-CNS refinement algorithm used for producing the reference structures from the RECOORD database, and we should check if any large difference in the final structures can arise just from the different type of NOE restraints.

The data assembled in Table 3.5 demonstrate that the results obtained with ICMD using the two types of restraints are not significantly different. This concerns the quality scores, the radii of gyration and the RMSD$_c$ values. The number of consistent violations is sometimes different, but one should keep in mind that this number is evaluated differently for ambiguous and floating restraints (see § 3.2.1). Using a single universal definition would be misleading because structures obtained with ambiguous restraints exhibit an exaggerated apparent number of floating violations and *vice versa*. We will show below that the apparently different number of NOE violations in structures produced with floating and ambiguous restraints disappears when the corresponding conformations are re-minimised in identical conditions.

3.2.4 Comparison between ICMD and RECOORD

In order to compare the ICMD methodology with the state-of-the-art methods in NMR structure determination, we compared the 25 lowest energy conformations obtained with ICMD with the conformations of the same proteins, stored in the RECOORD database. The comparison presented here was performed with the conformations in the CNS and CNW sets. Both comparisons are important because the ICMD approach can be considered as an intermediate between the CNS and CNW protocols. A comparison with the RECOORD conformations calculated with CYANA gives results similar to those obtained in the ICMD/CNS comparison (not shown).

Structure ensemble precision and resemblance

Within each set of conformations (CNS, CNW, and ICMD), the RMSD value RMSD$_c$ between the best 25 conformations are shown in Fig. 3.8a and in Table 3.5 (f,g). The variation of the RMSD$_c$ values between the different sets is small (0.1-0.6 Å): there is no large difference in convergence, whatever the method used. One can notice that larger RMSD$_c$ values are generally obtained within the ICMD conformations than within RECOORD conformations. The largest differences of RMSD$_c$ values are observed for proteins L23 and L11. To compare the structures obtained in each set of calculation, the RMSD values RMSD$_d$ between the average CNW structure and the three average structures of the sets CNS and ICMD, are shown in Table 3.5. The RMSD$_d$ values are smaller than the RMSD$_c$ values, except for L23, S28e and S27e. For the majority of proteins, the difference between the structures is within the conformational variability of each structure, and the structures are thus very similar. The large RMSD$_d$ observed for L23, S28e and S27e arise from local conformational variability in loops and in helix 15-23 of L23.

For the majority of cases studied here, the largest RMSD$_c$ values are obtained in the ICMD structures. The use of the ICMD protocol thus produces a slightly better exploration of the conformational space than the other methods. In RECOORD, the refinement of the protein structure by a short molecular dynamics simulation in water (CNW set) increases the RMSD

Name	set	short[a] H pairs	PROCHECK[b] % core	RMS of[c] violations	WHAT IF[d] bumps	sidechains[e] bumps (%)	RMSD$_c$[f]	RMSD$_d$[g]	R$_{gyr}$[h]
L25	cns	8	48.8	0.03 (0.07)	20.7	56.4	1.5	0.8	13.3 ± 0.1
L25	cnw	7	66.1	0.03 (0.07)	6.6	28.7	1.6	-	12.9 ± 0.1
L25	float	0	71.3	0.13 (0.12)	16.2	79.2	1.6	1.2	13.1 ± 0.1
L25	adr	1	75.8	0.09	15.4	82.7	1.6	1.5	13.1 ± 0.1
L30	cns	4	78.1	0.01 (0.02)	4.8	69.5	0.9	0.7	13.5 ± 0.1
L30	cnw	0	80.4	0.01(0.02)	2.5	44.7	1.1	-	13.0 ± 0.1
L30	float	1	80.3	0.05 (0.07)	9.0	89.3	1.1	0.7	13.1 ± 0.2
L30	adr	0	81.5	0.05	9.2	92.1	1.3	0.6	13.1 ± 0.1
L11	cns	1	83.5	0.01(0.03)	6.7	90.7	1.8	0.9	14.1 ± 0.3
L11	cnw	0	78.8	0.01 (0.03)	3.6	36.7	1.8	-	13.3 ± 0.4
L11	float	1	78.0	0.06 (0.06)	8.4	85.8	2.1	1.1	13.2 ± 0.5
L11	adr	1	81.1	0.05	8.8	85.6	2.4	1.1	13.2 ± 0.6
L20	cns	24	71.6	0.04 (0.08)	25.3	44.1	0.9	0.5	10.7 ± 0.1
L20	cnw	1	77.3	0.04 (0.09)	7.6	41.1	0.8	-	10.6 ± 0.1
L20	float	1	78.7	0.19 (0.17)	10.9	68.9	0.9	0.6	10.7 ± 0.1
L20	adr	0	78.8	0.12	7.8	75.2	0.9	0.5	10.8 ± 0.1
L23	cns	20	73.6	0.06 (0.04)	23.2	58.0	0.5	1.0	14.6 ± 0.2
L23	cnw	11	81.6	0.06 (0.03)	10.8	48.0	0.6	-	14.1 ± 0.3
L23	float	1	81.8	0.09 (0.09)	10.3	85.0	0.8	1.8	14.6 ± 0.5
L23	adr	0	81.5	0.05	8.2	85.8	0.8	1.0	14.7 ± 0.5
S28e	cns	6	70.0	0.01 (0.04)	7.1	61.2	0.9	1.8	14.1 ± 0.6
S28e	cnw	5	72.7	0.01 (0.04)	5.0	24.2	1.1	-	13.0 ± 0.6
S28e	float	1	69.5	0.06 (0.05)	9.6	79.8	0.9	1.6	13.2 ± 0.7
S28e	adr	3	69.4	0.05	9.4	78.8	1.0	2.0	13.4 ± 0.7
S19	cns	4	64.4	0.02 (0.06)	16.0	62.9	0.7	0.6	12.2 ± 0.2
S19	cnw	0	71.2	0.03 (0.07)	7.9	52.2	0.8	-	11.8 ± 0.2
S19	float	4	72.4	0.13 (0.11)	12.5	87.5	0.7	0.7	11.9 ± 0.3
S19	adr	0	71.7	0.12	11.8	86.6	0.9	1.0	12.2 ± 0.3
S27e	cns	15	70.8	0.05 (0.04)	4.7	75.7	1.0	0.8	11.5 ± 0.1
S27e	cnw	1	77.0	0.04 (0.04)	3.4	62.0	1.2	-	11.0 ± 0.1
S27e	float	1	70.2	0.05 (0.04)	6.8	89.6	1.0	1.2	11.1 ± 0.2
S27e	adr	1	71.2	0.01	7.6	86.9	1.2	1.4	11.3 ± 0.2

Table 3.5: Comparison of the RECOORD conformations from the CNS and CNW sets with the 25 best ICMD conformations calculated using floating (float) and ambiguous restraints (adr). The calculations were performed on the set of proteins described in Table 3.4.

[a] Total number of hydrogen pairs closer than 1.5 Å

[b] Mean % of residues in the PROCHECK Ramachandran core

[c] RMS of NOE violations (Å). Numbers in parentheses are calculated on the minimised conformations described in Fig. 3.10

[d] Mean number of WHAT IF bumps

[e] Percentage of WHAT IF bumps involving one or two sidechain atoms.

[f] RMSD (Å) value calculated between the conformations.

[g] RMSD (Å) value of the average structures (cns, float, adr) with respect to the cnw average structure.

[h] Mean values and standard deviations (Å) of the protein radius of gyration.

values with respect to the CNS set. The use of ICMD has thus an effect similar to those produced by a molecular dynamics simulation in water: this may come from the general purpose force-field used for the ICMD calculations.

Structural quality

The largest percentages of residues in the core Ramachandran region are observed (Fig. 3.8b and Table 3.5(b)) for the ICMD structures of L25, L30, L20, L23 and S19. The WHAT IF quality

Z-scores (Fig. 3.9) display different trends when the ICMD and the RECOORD structures are compared. The worst Ramachandran (Fig. 3.9a) and 2nd-generation packing (Fig. 3.9b) Z-scores are observed for the CNS structures. The best Ramachandran Z-scores (Fig. 3.9a) are generally observed in the ICMD structures, whereas the best 2nd-generation packing Z-scores (Fig. 3.9b) are generally obtained in the CNW structures. Exceptions are observed for the proteins L30, L11 and L20, for which the CNW structure exhibit the best Ramachandran Z-scores, and for protein S19, for which ICMD structure displays the best 2nd-generation packing Z-score. Concerning the backbone conformation Z-score (Fig. 3.9c), the worst values are mainly exhibited by the RECOORD structures, except for the proteins L30 and L11. For all calculations, the residues showing most of the poor ϕ, ψ values are mainly located in loops, C terminal and N terminal tails.

The number of hydrogen pairs closer than 1.5 Å, are smaller in ICMD than in RECOORD structures (Table 3.5(a)). On the other hand, the mean numbers of bumps detected by WHAT IF (Table 3.5(d)) are in the 2.5-25.3 range for the RECOORD structures, and in the 7.6-16.2 range for the ICMD structures. A closer inspection reveals that the bumps in ICMD and in RECOORD conformations are somewhat different on average. In RECOORD structures, for all proteins except L11 (CNW set) the percentage of bumps where one or two sidechain atoms are involved lies between 24.2 and 75.7% (Table 3.5(e)). For the ICMD structures, the fraction of such bumps is significantly higher (68.9-92.1%). The decrease of the percentage of main chain atoms involved in the bumps is also in agreement with the improvement of the Ramachandran Z-scores in ICMD structures. Furthermore, visual inspection of the residues involved in the bumps, shows that their sidechains are mainly directed towards the protein exterior.

The mean radii of gyration of the ICMD structures were calculated and compared to those of the CNS and CNW sets (Table 3.5(h)). For proteins S28e and S27e the polypeptide chain con-

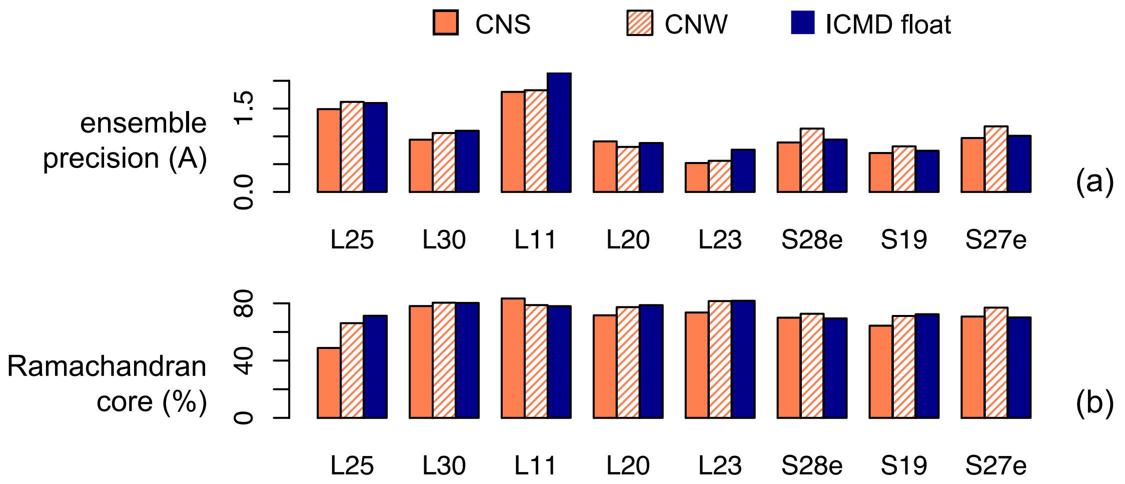


Figure 3.8: Comparison of the CNS, CNW and ICMD structures : comparison of the ensemble precision (Å) (a), and of the percentage of residues in the core Ramachandran regions (b), for calculations run on the set of proteins from Table 3.4. The results shown are those obtained using CNS (orange), CNW (shaded orange) and ICMD with floating restraints (dark blue). The compared values are mean values calculated on the set of the analyzed conformations.

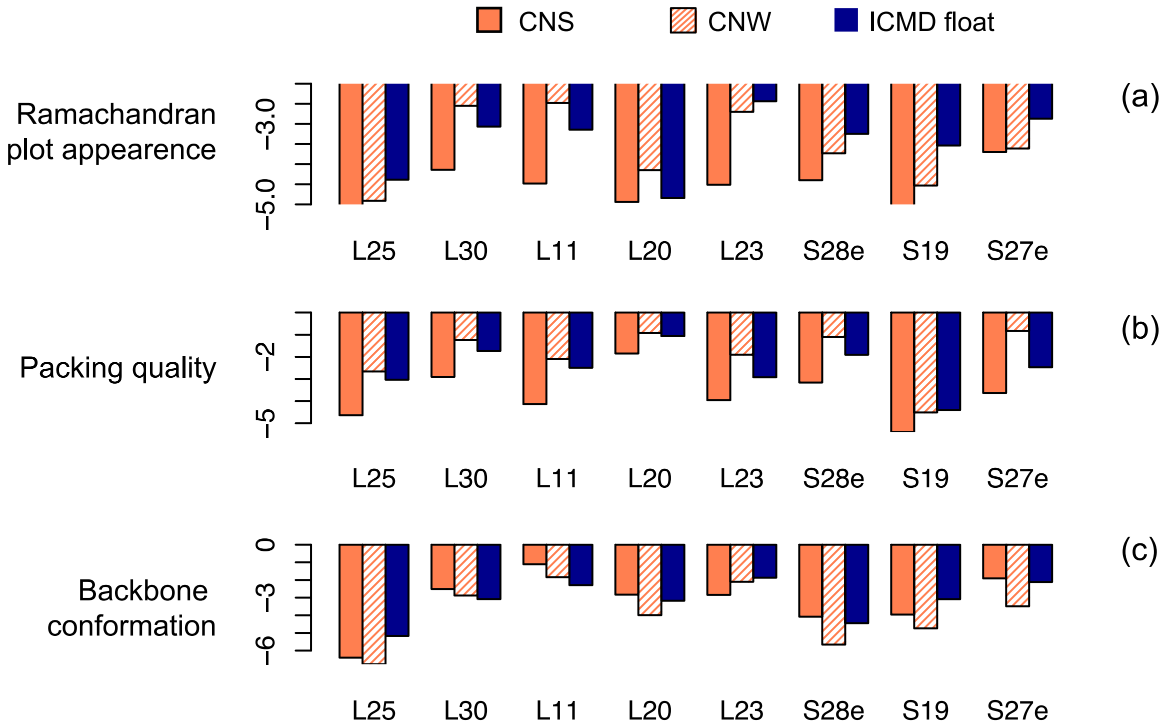


Figure 3.9: Comparison of WHAT IF Z-scores of CNS, CNW and ICMD structures : comparison of the WHAT IF Ramachandran Z-score (a), of the 2nd-generation packing Z-score (b), and of the backbone conformation Z-score (c) for calculations run on the set of proteins from Table 3.4. The results shown are those obtained using CNS (orange), CNW (shaded orange), ICMD with floating restraints (dark blue). The compared values are mean values calculated on the set of the analysed conformations.

sidered in the ICMD calculation is larger than the ones in the PDB and RECOORD databases. This is because some residues are unrestrained, and their conformational variability increases significantly the radius of gyration. To remove this bias, the radii of gyration of S28e and S27e were calculated by selecting only the sub-sequence present in the RECOORD structures. The mean radii of gyration are smaller in ICMD structures than in CNW one, for protein L11. For the other proteins, the mean radii of gyration are larger in ICMD structures than in the CNW structures, but these differences are smaller than or equal to the sum of the corresponding standard deviations. One can conclude, therefore, that the use of the all atom force-field in the gas phase calculations in ICMD introduces no undesirable general bias towards more compact or more extended protein structures.

Distance restraints violations, floating restraints and energy minimisation

The quality of NMR structures depends upon the relative weight attributed to violations of experimental data with respect to the conformational energy. By changing this balance, the number of violated restraints can always be reduced at the expense of degraded chemical geometry, and *vice versa*. In order to compare equivalent objects, we decided to energy-minimise the CNS, CNW and ICMD floating conformations in identical conditions. To this end, the internal biopolymer geometry was set free and all ICMD floating conformations were re-minimised. Si-

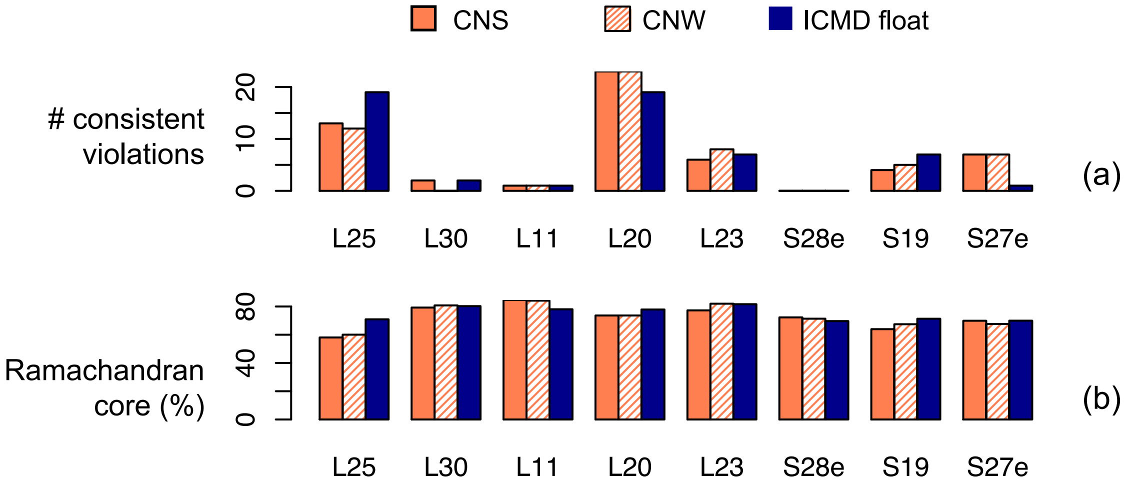


Figure 3.10: Comparison of re-minimised CNS, CNW and ICMD float structures : Comparison of the number of consistent violations (a) and of the percentage of residues in the core Ramachandran region (b) for calculations run on the set of proteins from Table 3.4. The results shown are those obtained on minimised ICMD conformations obtained with floating restraints (dark blue), obtained on CNS conformations minimised with floating restraints (orange) and obtained on CNW conformations minimised with floating restraints (shaded orange). All minimisations were performed with ICMD. The compared values are mean values calculated on the set of the analysed conformations.

multaneously, the corresponding structures were imported from RECOORD database to ICMD and energy minimised in identical conditions. The results of these calculations are compared in Fig. 3.10. Similar numbers of consistent violations are obtained for the three sets of conformations (Fig. 3.10a). The quality of minimised conformations was evaluated through the percentage of residues in the core Ramachandran regions (Fig. 3.10b). The comparison of the percentage of residues in the core Ramachandran diagram shows that these percentages were not sensibly modified with respect to Fig. 3.8b. The conformation quality is thus not significantly altered by the minimisation.

Calculation times

The CPU time required to produce one conformation with ICMD was between 1 and 3 hours on a processor operating at 2.4 GHz for the proteins analysed here. On the other hand, the CPU time needed for the production of one conformation using the CNS (respectively CNW) protocol is 15 minutes (respectively $\sim$10 minutes). The ICMD protocol is thus more time-consuming than the CNS and CNW protocols, which is not surprising since the number of integration steps is 20-fold larger. Indeed, the number of steps in the CNS protocol is 20 000, whereas the number of steps for ICMD is 3100+(100*n_{max}), where n_{max} is the number of TAS levels contained in the set of restraints. The CPU time used for running the ICMD protocol varies much within the set of proteins studied, as the variable target function algorithm [Braun and Go, 1985] may behave in a very different way for α and β secondary structures, and depends on the protein size. The protocol is also certainly sensitive to the consistency of the restraints, as too many inconsistent restraints produce repeated cycles of variable target phase (Fig. 3.5) and thus delay the conformation calculation.

In summary, the comparison between the RECOORD and the ICMD structures shows that the ICMD methodology exhibits an efficiency similar to those of the state-of-the-art methods, as far as the convergence of the calculation and the quality of the obtained structures are concerned. The quality of the ϕ, ψ distributions is slightly better in the ICMD than in the RECOORD structures. The bumps detected by WHAT IF in the ICMD structures concern mainly sidechain atoms, in contrast to the observations made in the RECOORD structures. The radii of gyration display similar values in ICMD and in RECOORD structures. The RMSD$_c$ values are slightly larger, which is the sign of a better exploration of the conformational space. Thus, the ICMD approach permits a slight improvement of the quality of the obtained structures as well as of the exploration of the conformational space, justifying the more important computational effort.

3.3 Conclusion

In this chapter, the application of two original approaches for efficient automated NOE assignment and NMR structure calculation were described. First, the implementation of the network anchoring algorithm in the ARIA protocol proved particularly efficient in accelerating the NOE assignment process with completely unassigned data, yielding, from the third iteration, backbone precision and accuracy similar to those observed when the peaks are already assigned. The global structural quality of the structures obtained with the inclusion of the network anchoring stays close to the one observed with initial partial NOE assignments. Moreover, it also managed to improve on the standard ARIA protocol in avoiding local assignment and structure calculation errors. Detection of such errors has been made easier by the analysis of quality score profiles along the sequence, and this supports a generalised use of the graphical tools presented in § 2.5.

Second, with the ICMD methodology, structures of a quality similar to that observed in RECOORD are obtained. For small proteins, ICMD is thus competing well with the current NMR refinement methods. The calculations performed here also indicate that it may be important to use a general purpose force-field, including the Coulombic and Lennard-Jones non-bonded interactions, from the beginning of the refinement. Indeed, ICMD, which uses such a force-field, produces conformations with generally better Ramachandran Z-scores scores than the RECOORD structures. The use of a relatively short (6 Å) cut-off value for the non-bonded interactions reduces the influence of long-range interactions during the high temperature phase of the refinement. The use of a slower cooling protocol in ICMD can be another reason for obtaining better Ramachandran Z-scores in the structures. This tendency is in agreement with the observation recently made [Fossi et al., 2005c] that a slow cooling in structure determination is more productive. The ICMD conformations produced with the floating restraints often exhibit a larger number of consistently violated restraints, but this apparent difference results from the definition of restraints and is removed by minimisation of the conformations in identical conditions. The residues lowering the structure quality are mainly located in the less-defined parts of the molecules, in long loops and tails. However, the methods to check structure quality rely on the knowledge of the X-ray crystallographic structures, and the flexible parts (loops and tails) are usually invisible in the electronic densities [Kwasigroch et al., 1997]. Thus, the poor quality Z-scores due to residues located in such protein regions may arise from the protein structure itself as well as from the lack of database knowledge from X-ray structures.

4

Automated structure determination of symmetric homo-dimers with ARIA.

This chapter describes the application of the *symmetry-ADR* strategy implemented in ARIA (§ 2.3) for automated structure determination of symmetric homo-dimers for both solution and solid-state NMR data. The first section (§ 4.1) reports the results obtained on three dimeric structures previously determined by solution NMR. Different levels of initial assignment ambiguity have been compared. We have also analysed the ability of the network anchoring to cope with the highly ambiguous present in symmetric assemblies. The spin-diffusion correction proposed in ARIA intends to describe more quantitatively the observed NOE intensities. It was therefore interesting to investigate the interplay between these two methods and their relative influence on NOE assignment of homo-dimeric proteins. The second part (§ 4.2) addresses the problem of *de novo* structure determination of protein from solid-state NMR data on uniformly labelled sample, using the concept of ambiguous distance restraints (ADR) with the ARIA protocol. In collaboration with Dr. Anja Böckmann (IBCP, Lyon, France), we used here as a model the 2x10.4 kDa Crh protein, whose structure, determined by X-ray diffraction, shows a domain-swapped dimer [Juy et al., 2003]. For this protein, the structure determination problem consists in obtaining the structure of the monomer, as well as the relative orientation of the two monomers to form the dimer. Compared to solution NMR, the number of isolated cross-peaks is greatly reduced, and the size of the chemical shift tolerance window must be increased due to the lower spectral resolution. As a consequence, the majority of cross-peaks in ^{13}C-^{13}C or ^{15}N-^{13}C correlation spectra remain ambiguously assigned, in addition to the intrinsic symmetry ambiguity.

4.1 Symmetric homo-dimers from solution NMR

4.1.1 Calculation schemes

The analysis is conducted on a set of symmetric homo–dimeric structures, representing three different interfaces: anti-parallel β sheet (PDB id: 1NEI [Yee et al., 2002]), intertwined bundle of four α-helices (PDB id: 1PZQ [Broadhurst et al., 2003]) and α-helical (PDB id: 2B9Z [Lingel et al., 2005]). Both 1PZQ and 1NEI deposited structures were originally calculated with manually detected inter-monomeric NOEs. Manually assigned NOEs with remaining chain ambiguity were used in the first place to determine the 2B9Z structure. Several ARIA calculations were performed with different levels of initial ambiguity and with or without the application of the network anchoring and spin diffusion correction (Table 4.1).

Run	Original assignments	Network Anchoring	Spin diff. correction	Chain ambiguity	Full ambiguity
2b9z$_{ori}$	✓[a]	-	-	-	-
2b9z$^{sd}_{ori}$	✓	-	✓	-	-
2b9z$_{amb}$	-	-	-	✓	-
2b9z$^{sd}_{amb}$	-	-	✓	✓	-
2b9z$_{amb\star}$	-	✓	-	✓	✓
2b9z$^{sd}_{amb\star}$	-	✓	✓	✓	✓
2b9z$^{nonet}_{amb\star}$	-	-	-	✓	✓
1nei$_{ori}$	✓	-	-	-	-
1nei$_{amb}$	-	-	-	✓	-
1nei$_{amb\star}$	-	✓	-	✓	✓
1nei$^{nonet}_{amb\star}$	-	-	-	✓	✓
1pzq$_{ori}$	✓	-	-	-	-
1pzq$_{amb}$	-	-	-	✓	-
1pzq$_{amb\star}$	-	✓	-	✓	✓
1pzq$^{nonet}_{amb\star}$	-	-	-	✓	✓

Table 4.1: Summary of the different calculation schemes

---

[a] Legend: ✓ : present, - : not present

Chain ambiguous restraints

For each dimer, two calculations were performed, with or without chain ambiguity, producing the sets (2b9z$_{ori}$, 1nei$_{ori}$, 1pzq$_{ori}$) and (2b9z$_{amb}$, 1nei$_{amb}$, 1pzq$_{amb}$). In the first case (ori), cross-peaks are assigned unambiguously, while in the second (amb), the assignments as intra– or inter–monomer were ambiguous for all restraints, except a few ones determined automatically using the secondary structure elements, as described in § 2.3. For all ARIA runs, 40 conformers were calculated for each iteration 0 to 7, 100 conformers were generated at the iteration 8, and the 10 conformers having the best total energy underwent a refinement step in water [Linge et al., 2003b]. The number of MD steps for the two Cartesian cooling stages of the ARIA simulated annealing protocol was three times larger that the default values, leading to a total of 27000 steps.

Fully ambiguous data and network anchoring

Two structure calculations were carried out on the structures 2B9Z, 1NEI and 1PZQ with totally unassigned cross-peaks (no proton and no chain assignments), with or without the application of the network anchoring procedure. For a symmetric dimer, the use of network anchoring biases the assignment towards intra-monomer NOEs, as they are better supported by the rest of the network. In that way, the dimer structure is disrupted and each monomer folds separately. To avoid this artefact, only one monomer was considered when applying network anchoring. Then, the chain ambiguity is reintroduced for each restraint, except for the unambiguous inter-monomeric restraints determined using the secondary structure elements.

The structure calculations with full ambiguity and with network anchoring are named 2b9z$_{amb\star}$, 1nei$_{amb\star}$ and 1pzq$_{amb\star}$, whereas the calculations without network anchoring are 2b9z$^{nonet}_{amb\star}$, 1nei$^{nonet}_{amb\star}$ and 1pzq$^{nonet}_{amb\star}$ (Table 4.1). Except otherwise stated, network anchoring was used during the first three iterations. The initial data sets consisted in: 2156 ambiguous restraints (29.1restraints/residue) for 2B9Z, 1477 ambiguous restraints (24.6 restraints/residue) for 1NEI and 1562 ambiguous restraints (26.0 restraints/residue) for 1PZQ. The length of the SA cooling phase as well as the numbers of generated and analyzed conformers are identical to the ones presented in previous section.

For the calculations performed in § 4.1.3, as no experimental peak list was available, synthetic cross-peak lists were simulated from the 1NEI and 1PZQ structures according to the following protocol. Each deposited restraint was converted into an artificial heteronuclear 3D NOESY cross-peak, the peak volume being equal to the inverse sixth power of the corresponding target distance. The two sets of protons involved in the restraint represent the two proton dimensions, and one of the attached Carbon or Nitrogen atoms, the third dimension. For each dimension, the corresponding chemical shifts were retrieved from the assignments deposited in the BioMagResBank [Ulrich et al., 2008] (1NEI: 5621, 2PZQ: 5885). For ambiguous restraints, the average chemical shift value of the possible assignments was used.

Spin-diffusion correction

In the case of 2B9Z, the availability of NOE peak lists and mixing times made it possible to study the effect of spin diffusion. The data set consisted in three 3D-NOESY peak lists derived from spectra recorded with proton mixing time of 90 and 100 ms. The calculations are named as follows (Table 4.1): unambiguous distance restraints (2b9z$_{ori}$, 2b9z$^{sd}_{ori}$), chain ambiguity (2b9z$_{amb}$, 2b9z$^{sd}_{amb}$), full ambiguity with network anchoring (2b9z$_{amb\star}$, 2b9z$^{sd}_{amb\star}$). In case of spin-diffusion correction (sd), the error bounds estimation was based on calibrated distances (§ 2.1.1).

4.1.2 Efficiency of the chain assignment

In this section, we will focus on the influence of chain ambiguity for the structure determination. The contact maps derived from the restraints at the first iteration are compared to contact maps obtained at the end of each run (Fig. 4.1). Only few (less than 4) contacts were known as inter-molecular at the iteration 0, and the resulting contact maps are thus almost completely ambiguous (Fig. 4.1(a,c,e)), whereas all inter- and intra-molecular contacts were correctly assigned at the iteration 8 (Fig. 4.1(b,d,f)). The final superimposed conformers are shown in Figure 4.2. The WHAT IF and PROCHECK quality parameters obtained for the dimeric structures (Table 4.2) are in the admitted range (-4/4) and are similar for the unambiguous calculations (ori) as well as for the calculations with chain ambiguity (amb). The worse quality observed for 1PZQ may be related to the structural context of the dimer. This protein was constructed by fusing a pair of unstable interface regions from Polyketide Synthase [Broadhurst et al., 2003]. The accuracy of the calculation was determined as the coordinate RMSD from a known reference structure. Comparable accuracies were obtained for the calculations with initial ambiguous restraints or with unambiguous restraints (Table 4.2). Nevertheless, the RMSD from the deposited

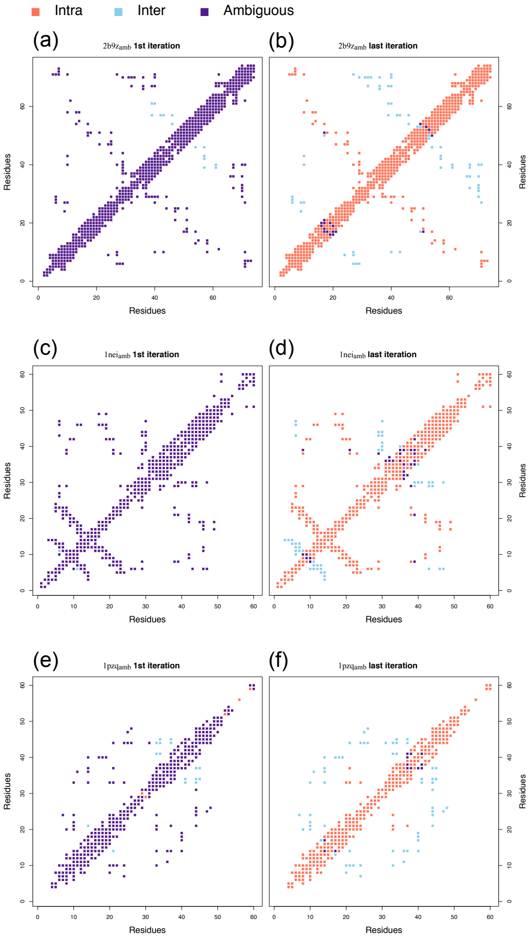


Figure 4.1: Contact maps observed at the first (a,c,e) and last iteration (b,d,f) of the ARIA runs with chain ambiguity on symmetric homo–dimers: (a,b) 2b9z$_{amb}$, (c,d) 1nei$_{amb}$ and (e,f) 1pzq$_{amb}$. The inter-molecular contacts are shown in blue and the intra-molecular ones in red, while the ambiguous contacts are in purple.

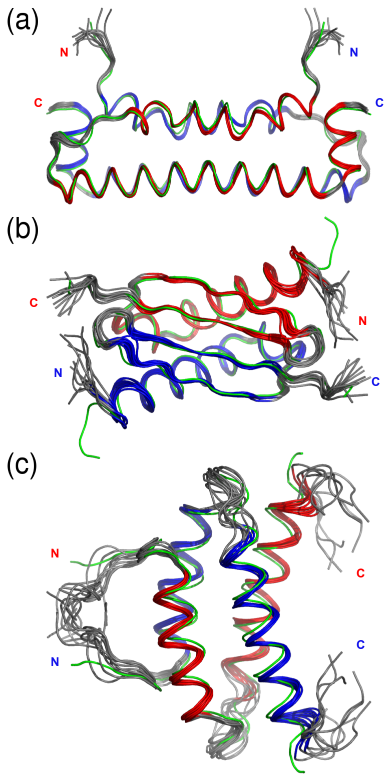


Figure 4.2: Superpositions of the final conformers obtained for the symmetric homo–dimers : (a) 2b9z$_{amb}$ (b) 1nei$_{amb}$ and (c) 1pzq$_{amb}$. The chains are drawn in tube, the A chain in red and the B one in blue. Only the secondary structure elements are coloured. The mean structure of the corresponding reference PDB structures is show in green. Figures have been rendered with PyMOL (http//www.pymol.org)

	2B9Z		1NEI		1PZQ	
	2b9z$_{ori}$	2b9z$_{amb}$	1nei$_{ori}$	1nei$_{amb}$	1pzq$_{ori}$	1pzq$_{amb}$
PROCHECK[a]						
Most favored	98.0 $\pm$ 0.7	97.4 $\pm$ 1.1	92.4 $\pm$ 2.2	90.0 $\pm$ 2.9	69.5 $\pm$ 3.0	71.3 $\pm$ 3.5
Allowed	1.9 $\pm$ 0.6	2.0 $\pm$ 0.8	7.2 $\pm$ 2.4	7.9 $\pm$ 4.3	24.2 $\pm$ 3.3	24.7 $\pm$ 2.6
Gener. allow.	0.0 $\pm$ 0.0	0.6 $\pm$ 0.8	0.1 $\pm$ 0.3	0.9 $\pm$ 1.0	2.6 $\pm$ 0.7	0.9 $\pm$ 1.0
Disallowed	0.1 $\pm$ 0.2	0.0 $\pm$ 0.0	0.3 $\pm$ 0.6	1.1 $\pm$ 1.3	3.8 $\pm$ 1.4	3.0 $\pm$ 1.3
Bad contacts	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.1 $\pm$ 0.3	0.1 $\pm$ 0.3
WHAT IF scores[b]						
NQACHK	0.1 $\pm$ 0.2	-0.2 $\pm$ 0.2	-1.3 $\pm$ 0.2	-1.4 $\pm$ 0.2	-2.0 $\pm$ 0.3	-2.3 $\pm$ 0.5
RAMCHK	-1.2 $\pm$ 0.2	-0.2 $\pm$ 0.4	-0.8 $\pm$ 0.3	-1.1 $\pm$ 0.4	-3.9 $\pm$ 0.4	-3.5 $\pm$ 0.5
C12CHK	-2.6 $\pm$ 0.2	-2.9 $\pm$ 0.2	-2.0 $\pm$ 0.5	-1.7 $\pm$ 0.7	-2.5 $\pm$ 0.3	-2.3 $\pm$ 0.7
BBCCHK	0.3 $\pm$ 0.2	0.5 $\pm$ 0.3	0.3 $\pm$ 0.4	-0.6 $\pm$ 0.6	-4.8 $\pm$ 1.1	-4.4 $\pm$ 0.8
BMPCHK	18.8 $\pm$ 3.8	28.4 $\pm$ 7.3	3.9 $\pm$ 2.8	6.8 $\pm$ 5.3	20.9 $\pm$ 5.3	14.8 $\pm$ 2.7
NMR ensemble precision[c]						
Backbone RMSD (Å)	0.43 $\pm$ 0.09	0.49 $\pm$ 0.16	0.64 $\pm$ 0.10	0.63 $\pm$ 0.11	0.76 $\pm$ 0.10	0.71 $\pm$ 0.11
Heavy RMSD (Å)	0.66 $\pm$ 0.05	0.64 $\pm$ 0.08	0.99 $\pm$ 0.10	0.91 $\pm$ 0.13	1.14 $\pm$ 0.12	1.18 $\pm$ 0.11
Backbone accuracy[d]						
NMR ref. RMSD (Å)	0.46 $\pm$ 0.05	0.65 $\pm$ 0.04	1.17 $\pm$ 0.13	1.21 $\pm$ 0.14	1.22 $\pm$ 0.10	1.33 $\pm$ 0.09
X-ray ref. RMSD (Å)	1.39 $\pm$ 0.06	1.39 $\pm$ 0.06	n/a	n/a	n/a	n/a
NOE Violations						
# viol. $\geq$ 0.3 Å	0.2 $\pm$ 0.4	2.9 $\pm$ 1.1	0.8 $\pm$ 0.6	0.0 $\pm$ 0.0	7.3 $\pm$ 1.5	0.6 $\pm$ 0.9
RMS violations (Å)	0.02 $\pm$ 0.00	0.03 $\pm$ 0.00	0.02 $\pm$ 0.00	0.02 $\pm$ 0.00	0.05 $\pm$ 0.00	0.02 $\pm$ 0.00

Table 4.2: Quality parameters for the dimer calculations with or without chain assignment.

---

[a] PROCHECK results: percentage of residues in most favoured region, allowed region, generously allowed region and disallowed region.

[b] The WHAT IF scores are the following: 2nd generation packing quality Z-score (NQACHK), Ramachandran plot appearance Z-score (RAMCHK), $\chi1/\chi2$ rotamer normality Z-score (C12CHK), Backbone conformation Z-score (BBCCHK) and the number of inter-atomic bumps (BMPCHK).

[c] RMSD between the final conformers calculated for the backbone and heavy atoms. The residues ranges used for the RMSD calculation is 1-74 for 2B9Z, 8-50 for 1NEI and 1-50 for 1PZQ.

[d] RMSD between the final conformers and the average coordinates of the PDB reference structures calculated for the backbone atoms, with the same residues ranges as for the precision, except for 2B9Z (residues 4-73). The X-ray reference structure for the 2B9Z calculations is PDB id : 2AZ0 [Chao et al., 2005]

NMR structure is slightly larger in 2b9z$_{amb}$ than in 2b9z$_{ori}$, whereas the structure remains similar to the corresponding X-ray structure (PDB id: 2AZ0 [Chao et al., 2005]) for both calculations.

These results confirm the efficiency of the automated procedure presented here, to dissipate the chain ambiguity and to determine the structure of symmetric homo-dimers. The resulting structure ensembles exhibit a satisfactory level of structural quality for three types of dimer interface.

4.1.3 Application of the network anchoring with ambiguous data

The focus of this section is on evaluating the efficiency of the network anchoring approach for the structure determination of symmetric homo–dimeric structures from NOE data with chain and chemical shift ambiguity. The ARIA calculations for the three proteins without any prior assignments converged to the correct fold, if network anchoring was applied (Table 4.3). Without network anchoring, the 2b9z$_{amb\star}^{nonet}$ calculation converged, but inaccurately and the 1nei$_{amb\star}^{nonet}$ did not converge at all. With the standard assignment protocol, only 1pzq$_{amb\star}^{nonet}$ succeeded in folding correctly. This may be due to the fact that most of the long-range restraints in 1PZQ are inter-molecular (Fig. 4.1(f)), making network anchoring less crucial for filtering the assignment possibilities. The calculations with full ambiguity and network anchoring (Table 4.3) were then compared to the calculations previously performed (Table 4.2) with chain ambiguity and to the results obtained on the monomeric protein HRDC (Chap. 3, Table 3.3). The use of network anchoring slightly decreases the PROCHECK Ramachandran quality for 2B9Z and 1PZQ, which are two α helical structures. This observation is similar to the one made on the HRDC domain (Table 3.3). However, no large variations are observed for the WHAT IF Z scores between $amb\star$

	2B9Z		1NEI		1PZQ	
	2b9z$_{amb\star}$	2b9z$_{amb\star}^{nonet}$	1nei$_{amb\star}$	1nei$_{amb\star}^{nonet}$	1pzq$_{amb\star}$	1pzq$_{amb\star}^{nonet}$
PROCHECK[a]						
Most favored	96.1 ± 1.3	92.5 ± 1.1	91.1 ± 1.6	86.2 ± 2.0	66.6 ± 3.3	69.6 ± 3.5
Allowed	2.0 ± 1.5	2.3 ± 1.3	8.2 ± 2.2	13.0 ± 1.4	27.9 ± 4.1	26.0 ± 3.1
Gener. allow.	0.6 ± 1.1	4.1 ± 1.2	0.8 ± 1.0	0.8 ± 1.3	3.6 ± 1.7	2.3 ± 1.5
Disallowed	1.2 ± 1.0	1.1 ± 1.0	0.0 ± 0.0	0 ±0	1.9 ± 1.3	2.1 ± 1.4
Bad contacts	0.6 ± 1.0	4.6 ± 1.0	0.4 ± 0.8	0.1 ± 0.3	0.8 ± 1.0	1.0 ± 1.1
WHAT IF scores[b]						
NQACHK	0.1 ± 0.3	-3.4 ± 0.1	-1.9 ± 0.2	-4.4 ± 0.3	-2.1 ± 0.5	-2.0 ± 0.2
RAMCHK	-0.3 ± 0.2	-4.2 ± 0.3	-0.8 ± 0.6	-4.1 ± 0.3	-3.5 ± 0.3	-3.6 ± 0.4
C12CHK	-2.9 ± 0.4	-4.1 ± 0.3	-1.7 ± 0.5	-2.1 ± 0.6	-2.3 ± 0.5	-2.1 ± 0.5
BBCCHK	0.6 ± 0.3	-6.6 ± 0.5	-0.3 ± 0.3	-3.8 ± 0.6	-3.3 ± 0.6	-3.3 ± 1.0
BMPCHK	35.6 ± 7.9	78.8 ± 4.9	3.4 ± 3.3	37.1 ± 4.8	16.7 ± 3.6	18.8 ± 4.7
NMR ensemble precision[c]						
Backbone RMSD (Å)	0.48 ± 0.26	0.37 ± 0.08	0.55 ± 0.19	7.58 ± 1.36	0.75 ± 0.15	0.84 ± 0.17
Heavy RMSD (Å)	0.70 ± 0.13	0.60 ± 0.06	0.80 ± 0.21	7.91 ± 1.35	1.17 ± 0.16	1.28 ± 0.18
Backbone accuracy[d]						
NMR ref. RMSD (Å)	1.04 ± 0.08	15.82 ± 0.02	1.43 ± 0.15	13.47 ± 1.80	1.64 ± 0.14	1.67 ± 0.23
X-ray ref. RMSD (Å)	1.58 ± 0.12	16.44 ± 0.02	n/a	n/a	n/a	n/a
NOE Violations						
# viol. $\geq$ 0.3 Å	4.3 ± 2.1	6.00 ± 2.19	0.0 ± 0.0	0.40 ± 0.80	0.2 ± 0.6	1.00 ± 1.34
RMS violations (Å)[e]	0.03 ± 0.00	0.04 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00

Table 4.3: Quality parameters for the fully ambiguous dimer calculations with or without network anchoring

---

[a] see Table 4.2 for the legends.

calculations (Table 4.3) and all calculations described in Table 4.2. In contrast to the calculation performed on the HRDC domain (§ 3.1), the number of NOE violations larger than 0.3 Å is small for all runs. The comparison of amb runs (Table 4.2) with $amb\star$ runs (Table 4.3) shows that the network anchoring moderately decreases the backbone RMSD between NMR conformers or leaves it unchanged (for 1PZQ). Increasing the ambiguity reduces the structure accuracy expressed as the coordinate RMSD to a known NMR or X-ray reference structure.

Two different applications of network anchoring were used for the 1nei$_{amb\star}$ calculations: in the first iteration only (1nei$_{amb\star}$) or during iterations 0 to 2 (1nei$_{amb\star 2}$). When superimposing the average final structure obtained with the two setups (Fig. 4.3), it is striking that the helices $\alpha 2$ of the two monomers are swapped in 1nei$_{amb\star 2}$; the rest of the fold remains identical. The comparison of the corresponding WHAT IF score profiles reveals large differences, especially for the number of inter-atomic bumps and the backbone conformation Z-score. (Fig. 4.4). The major differences are located in the region of the residues 32 to 36, which corresponds to the N-terminal part of the helix $\alpha 2$ and coincides with the hinge region of the observed swapping. To a lesser extent, differences are observed for the $\beta 1$ strand, mainly arising from wrongly assigned $\beta 1$-$\alpha 2$ correlations. In this specific case, an application of network anchoring during more ARIA iterations biased the assignment process toward a wrong structure. The situation of symmetric dimers is particularly difficult since the corresponding inter or intra-monomer distances are often equivalent near the dimer interface. Nonetheless, the observation of certain structural quality

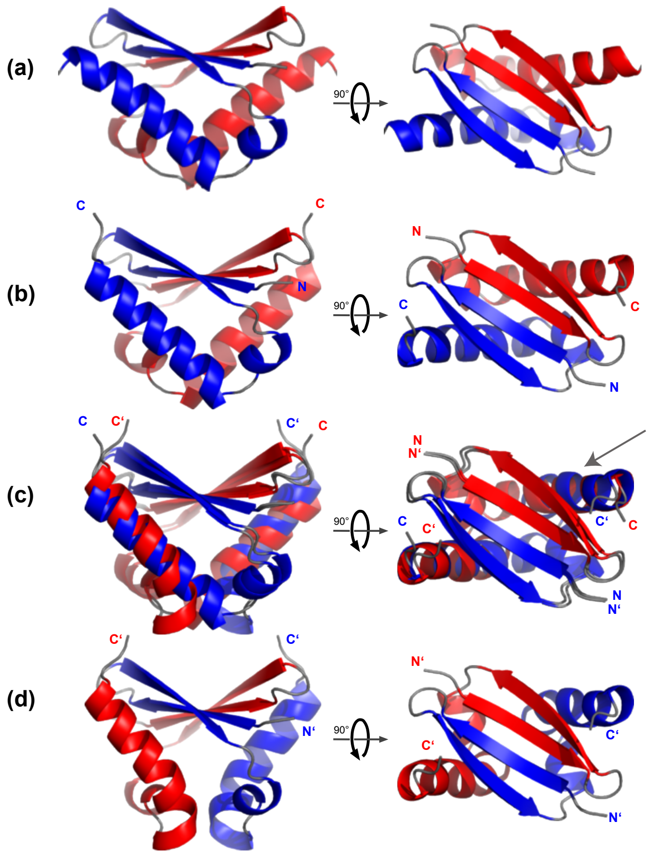


Figure 4.3: Average structure of the 1NEI ensembles: deposited in the PDB (a), average structure of the final ensembles of the 1nei$_{amb\star}$ (b) and 1nei$_{amb\star 2}$ (d) calculations. Superposition of the 1nei$_{amb\star}$ and 1nei$_{amb\star 2}$ (c) average structures. The swapped helix in 1nei$_{amb\star 2}$ is indicated by the grey arrow. The right part illustrates a rotation of 90° around the x axis. The A chain is coloured in red and the B one in blue. Only the secondary structure elements are colored. Figures have been rendered with PyMOL (http//www.pymol.org)

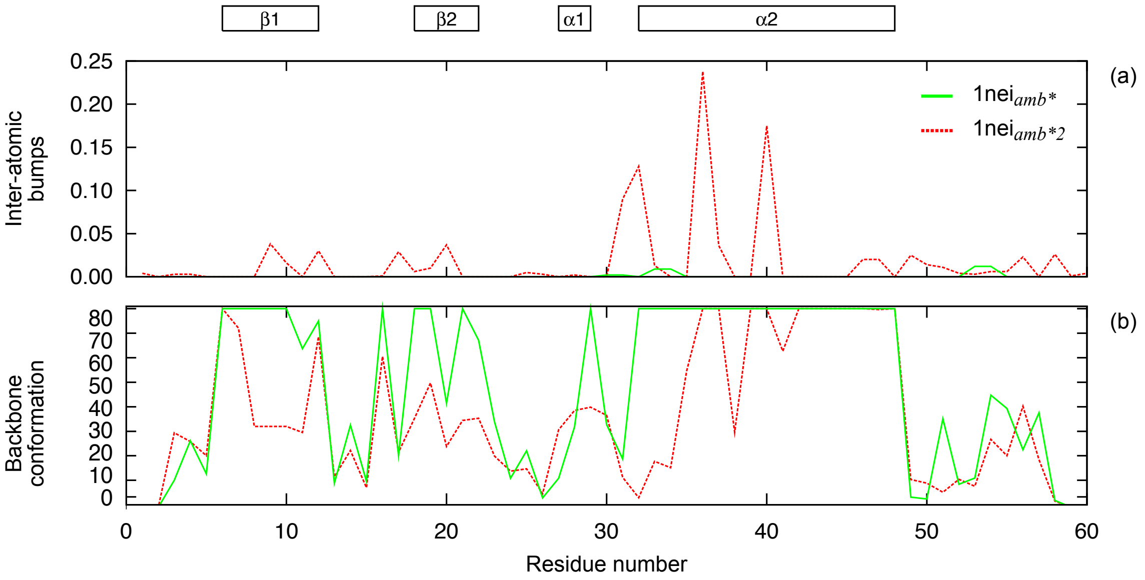


Figure 4.4: Comparison of the WHAT IF score profiles: scores obtained in the calculation 1nei$_{amb\star}$ (red solid) and 1nei$_{amb\star2}$ (green dashed) on the structure 1NEI. (a) Number of inter-atomic bumps and (b) Backbone conformation Z-score. Secondary structure elements are shown on top.

indicators along the protein sequence allows to identify this kind of local errors. Besides, the structure analysis performed here supports the idea that performing multiple structure calculations with network anchoring present in different iterations and comparing the outputs will help to detect possible errors.

The adaptation of the network anchoring to the difficult case of the symmetric homo–dimers is efficient to correctly assign this type of highly ambiguous NOE spectra in an automated way. The quality parameters observed on conformations obtained in presence and in absence of network anchoring have similar values, and similar trends are observed with monomer calculations previously performed (§ 3.1). This proves that same criteria can be used to check all types of calculations.

4.1.4 Impact of the spin diffusion correction with different levels of ambiguity

The comparison of calculations realised in similar conditions of ambiguity and network anchoring, but in presence and in absence of spin diffusion correction reveals that the correction consistently reduces the number of WHAT IF inter-atomic bumps (Fig. 4.5). This observation is in agreement with the previous calculation of the PH spectrin domain [Linge et al., 2004]. Indeed, as the target distances determined in absence of spin diffusion correction are shorter than the corresponding distances in the X-ray structure, too short distances induce steric clashes and consequently large BMPCHK values.

Whatever the level of restraint ambiguity, the WHAT IF scores are better for the backbone conformation (Fig. 4.5(d)) and the Ramachandran plot (Fig. 4.5(e)) when spin diffusion correction is applied. The less tight restraints observed when spin-diffusion correction is used could explain the degradation of the packing quality (Fig. 4.5(c)). The spin diffusion correction also increases the coordinate RMSD between NMR conformers (data not shown), which is in agreement with the derivation of looser restraints. Interestingly, the spin diffusion correction increases also the RMSD of the calculated structures to the reference NMR structure (Fig. 4.5(a)), similarly to the observation on PH spectrin domain [Linge et al., 2004], but decreases the RMSD to the X-ray structure 2AZ0 [Chao et al., 2005] (Fig. 4.5(b)). As a consequence, the structures 2b9z$^{sd}_{ori}$, 2b9z$^{sd}_{amb}$, and 2b9z$^{sd}_{amb\star}$ are less precise, but closer to the crystallographic structure.

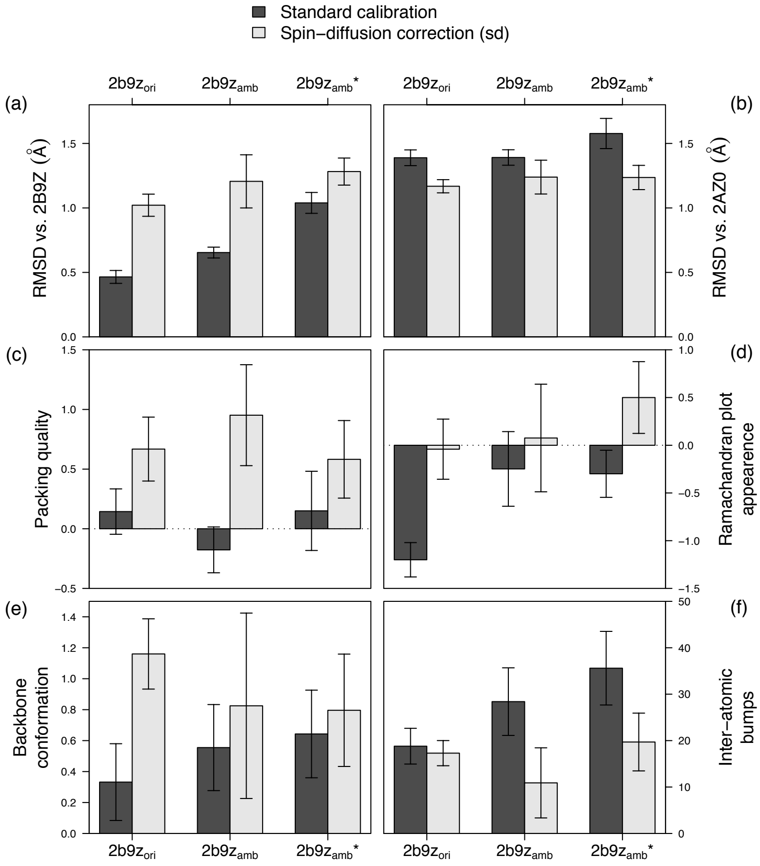


Figure 4.5: Effect of the spin diffusion correction on 2b9z$_{ori}$, 2b9z$_{amb}$ and 2b9z$_{amb\star}$ calculations. Backbone RMSD between 2b9z NMR ensembles and deposited PDB structures 2B9Z (a) and 2AZ0 (b) and comparison of the WHAT IF score profiles (c-f) obtained in absence (dark grey) and in presence (light grey) of spin diffusion correction.

To conclude, by correcting the distance restraints for the spin-diffusion effect, the resulting dimeric structure ensembles are less precise, but are of better quality and become more similar to the crystallographic structure. This more quantitative evaluation of the NOE signals is still effective for the high level of ambiguity in NMR data encountered in symmetric homo–dimers.

4.2 Structure determination of the *Crh* homo-dimer from solid-state NMR

4.2.1 Experimental context

***B*acillus subtilis Crh**

Crh is a 85 residue protein, with a tertiary structure formed by a four-stranded anti-parallel β-sheet (β1-4), and three α-helices α1, α2 and α3 (Chap. 5, Fig. 5.1). A conformational exchange between a monomeric and a dimeric form was observed in solution and its monomeric structure was determined in solution by liquid-state NMR [Favier et al., 2002]. The dimeric structure of Crh was then determined in the crystalline form by X-ray crystallography (PDB id: 1mu4, 1mo1 [Juy et al., 2003]). The transition from the monomeric to the dimeric form is produced by the 3D domain swapping of strand β1. In the hinge region, an intermolecular short β1a-sheet is formed. Solid-state NMR sequential assignments and structural analysis of a microcrystalline form of the protein [Böckmann et al., 2003] have revealed that the dimeric domain-swapped form is present in this preparation as well.

Data-set used and solid-state NMR spectroscopy

Sample preparation and NMR data acquisition on Crh micro-cystals have been performed in Lyon and is detailed elsewhere [Böckmann et al., 2003; Loquet et al., 2008]. Peak lists were generated from solid-state CHHC and NHHC spectra on U-[^{13}C,^{15}N] labelled microcrystalline Crh, which results in a total number of 1002 cross-peaks. Additionally, a set of 25 inter-monomer distance restraints, previously identified in a uniformly but heterogeneously [^{13}C:^{15}N] labelled protein Crh [Etzkorn et al., 2004], was used in the ARIA calculations. The TALOS [Cornilescu et al., 1999] software was applied to predict torsion angles from N, Cα, Cβ and C´ chemical shifts [Böckmann et al., 2003]. Dihedral angle predictions for 56 out of 85 residues were considered as “good” by TALOS and converted to dihedral angle restraints with the error margins given by the program.

ARIA protocol

In order to assign cross-peaks from ssNMR peak lists without the use of a homology model, we used the solid-state version of the program ARIA for C/NHHC spectra on uniformly labelled, with symmetry-ADR method, presented in § 2.4. The chemical shifts and the peak lists generated from the CHHC and NHHC spectra provided the input for the distance restraint assignments using ARIA. A detailed analysis of the ratio of the influence of the chemical shift tolerance window on the number of ambiguities (described in [Loquet et al., 2008]) suggested a chemical shift tolerance window of 0.25 ppm and a maximum number of assignment possibilities per peak of 25. Considering the high degree of ambiguity of the NMR-derived restraints, slow cooling stages were used[Fossi et al., 2005c]: 80000/64000 steps for the first run, and 60000/48000 steps for the second run. ^{1}H-^{1}H restraints were defined by a common lower bound of 1.8 Å, a target distance of 3.5 Å for proton-proton distances identified in the NHHC spectrum, and 3.8 Å for distances identified in the CHHC spectrum, with a common upper bound of 5.0 Å. No distance classes were established, as suggested by a previous analysis of the correlation between polarization transfer build-up curves and distances measured on the Crh X-ray structure [Gardiennet et al., 2008].

Structure calculations from unambiguous restraints determined by ARIA.

In the ARIA protocol, the structures were calculated with both unambiguous and ambiguous restraints. Subsequently, all unambiguous restraints assigned in the last iteration of the ARIA run were used as input for a final 3D structure calculation round using XPLOR-NIH [Schwieters et al., 2003]. Here, a symmetry-ADR protocol, similar to the one depicted in § 4.1 was applied [Loquet et al., 2008]. 200 dimer structures were generated and the 10 lowest-energy conformers, as well as the NMR restraint data file, were deposited to the Protein Data Bank (PDB id: 2RLZ). The program PROCHECK [Laskowski et al., 1993] was used to analyze the quality of the obtained conformers. The number of violations larger than 0.5 and 0.1 Å, as well as the RMS of violations were also determined. The convergence was assessed by calculating the RMSD between the conformers superimposed either on a hypothetical monomer (residues 2-12 of chain A, and 13-80 of chain B) or the complete dimer. The distance between the centres of mass of the monomers, as well as the angles between the two monomers, were compared to the corresponding values in the crystallographic structure.

4.2.2 Automatic assignments of highly ambiguous restraints.

Due to the size of the tolerance window, the total number of assignment ambiguities was 32600, with an average number of 2x16.3 ambiguities per peak, the factor of two taking into account the dimeric nature of the protein. The left panel of Figure 4.6 shows the 10 lowest energy conformers obtained after iterations 0, 3, 5 and 7 from the first run of ARIA. The quality of the structures calculated for each iteration was evaluated with respect to the average number of ambiguities per cross-peak (Figure 4.7a), the number of unambiguous and ambiguous restraints (Figure 4.7b), as well as the precision and accuracy of the structures (Figure 4.7c). These Figures reveal the good correlation between the convergence of the conformers and the progress of the automated assignment.

The automated assignments performed by the ARIA program highly reduce the average number of ambiguities per peak in each iteration (Figure 4.7a), from 16.3 in iteration 0 to 2.5 in the last iteration. After the last iteration, 397 cross-peaks represent more than one possible assignment, The automated assignment procedure achieved unambiguous assignments for 593 ^{1}H-^{1}H contacts after the last iteration of the first run (Figure 4.7b), including 115 long-range restraints. Iteration 7 was the first iteration where the number of unambiguous restraints is

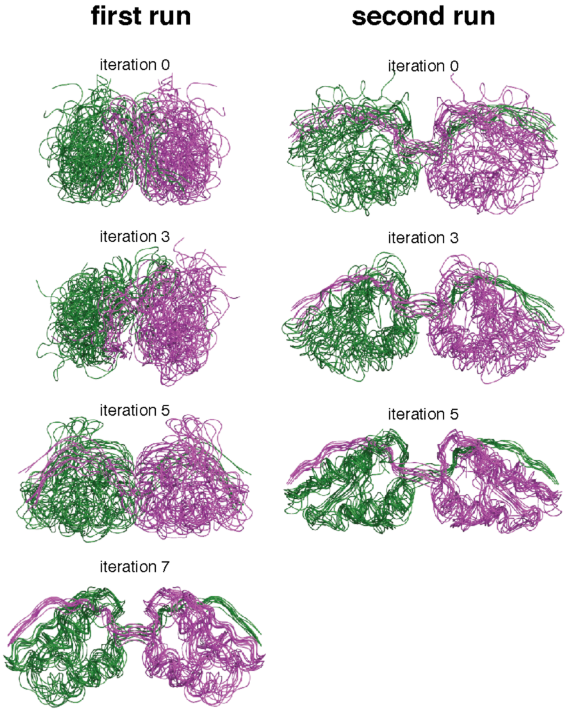


Figure 4.6: Comparison of conformers calculated in different iterations of ARIA runs : iterations 0, 3, 5 and 7 for the first run (left panel), iterations 0, 3, 5 for the second run (right panel). 28 conformers are calculated for each iteration. The 10 lowest-energy conformers superimposed on the backbone atoms are shown

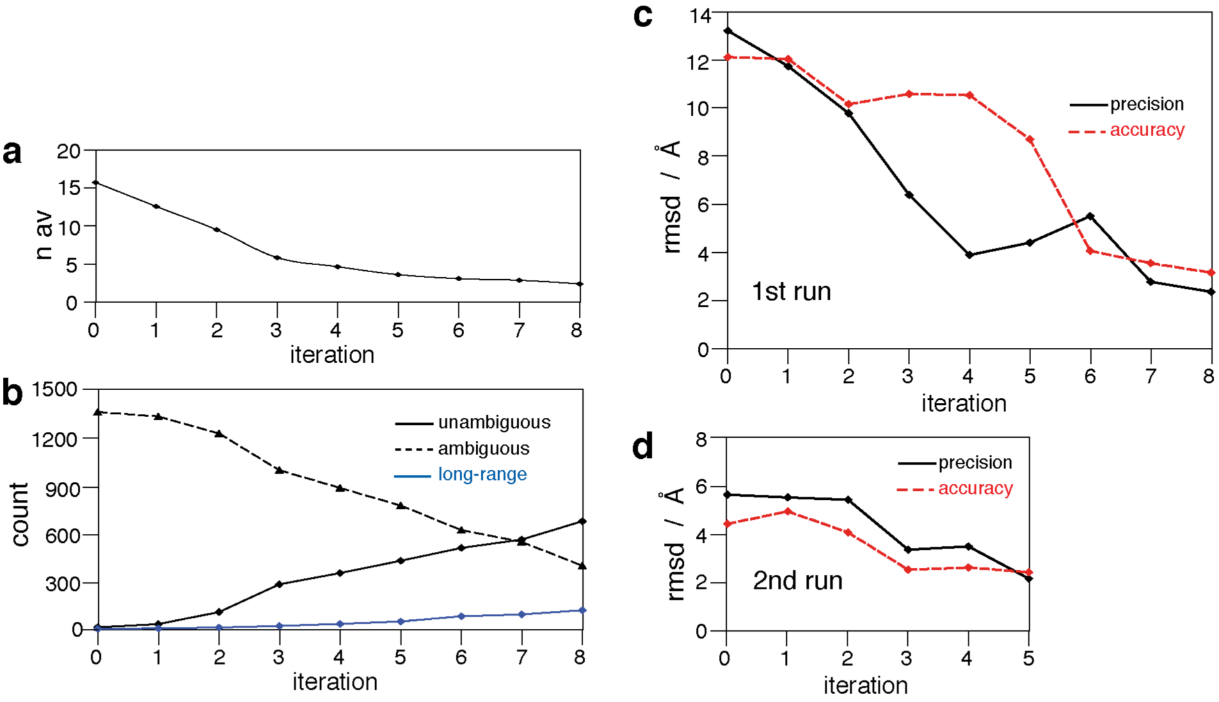


Figure 4.7: Analysis of the two runs performed using ARIA for the structure calculations of the Crh dimer. Evolution as a function of the iteration number: (a) average number of possible assignments per cross-peak for the first run, (b) number of unambiguous, ambiguous and long-range distance restraints for the first run, (c) backbone RMSD of the 10 lowest-energy conformers (precision) and backbone RMSD between the 10 lowest-energy conformers and the crystal structure (accuracy) for the first run and (d) RMSD of the 10 lowest-energy conformers (precision) and RMSD between the 10 lowest-energy conformers and the crystal structure (accuracy) for the second run. RMSD values were calculated for residues 2-81. In each iteration, 28 conformers were calculated, and the RMSD values is indicated for the 10 lowest-energy conformers.

higher than the number of ambiguous restraints. The precision of the structures increased with each iteration (Figure 4.7c), to reach a value of 2.8 Å for iteration 8, and the RMSD decreased steeply during the five first iterations down to 4 Å.

In order to enhance the quality of the assignment, a second round of ARIA iterations was performed. The 7 lowest-energy conformers obtained after iteration 8 of the first ARIA run were used as template structures to filter the possible assignments in the second ARIA run. The same solid-state NMR data was used for this second round (chemical shift and peak lists, TALOS dihedral angle restraints and inter-monomer restraints), but the input of the template structures allowed to decrease the number of possible assignments already during the first iteration. We set the parameter of violation tolerance to 2 Å instead of the standard value of 1000 Å used for the first iteration, and reduced the number of iterations to 6. The right panel in Figure 4.6 shows the 10 lowest-energy conformers for iterations 0, 3, 5 of the second ARIA run. The addition of the template structures highly increased the convergence of the first iteration resulting in a precision of 5.6 Å (Figure 4.7d) compared to 13.2 Å at iteration 0 of the first run. As shown in figure 4.6, the global fold of the Crh protein is defined after the first iteration of the second run, and the calculations converge to reach a precision of 2.3 Å. The automated assignment procedure achieved unambiguous assignments for 643 ^{1}H-^{1}H contacts in the last iteration of the second run, including 131 long-range restraints. The iterative assignment resulted in unambiguously assigned peaks for about 65 % of the initial 1002 peak entries.

4.2.3 Calculation of the Crh 3D structure from the unambiguous restraints determined using ARIA.

In the final structure calculation realised using XPLOR-NIH, we used only the 643 unambiguous ^{1}H-^{1}H restraints assigned by ARIA, as well as the TALOS dihedral restraints, to calculate a total of 200 conformers. A set of conformers was chosen on the basis of the absence of distance violations $>$ 0.5 Å. These 10 structures show good covalent geometry (Table 4.4); 74.2 % of the residues have backbone conformations in favourable regions, and 20.6 % in allowed regions of the Ramachandran plot. The selected 10 conformers show a precision of 1.33 Å when superimposing the backbone atoms for the dimer, and of 1.20 Å when superimposing the heavy atoms in regular secondary structures elements. The higher precision, with respect to the second ARIA run, may be due to the fact that only unambiguous distance restraints were considered in this calculation round, as well as to the higher number of conformers calculated. If we consider a hypothetic monomer (residues 2-12 from chain A, and residues 13-80 from chain B) , the RMSD decreases to 0.84 Å, and to 0.76 Å if only regular secondary structure elements are taken into account. The backbone fold of the solid-state NMR structure shows an accuracy of 2.89 Å with respect to the single crystal structure of the dimeric Crh protein. The geometry of the obtained structure was also compared to the geometry of the crystallographic dimer using the mean distance of the monomer centres-of-mass, and the orientation between the monomers, calculated as the angle formed by the two vectors going from the centre-of-mass of the interface region (residues 10 to 15) to the centre-of-mass of each monomer. Despite the fact that the orientation of the two monomers displays close mean values to the crystallographic structures (Table 4.4), the inter-monomer distance is strikingly larger in the ssNMR structure ensemble (see Chap. 5 for a comprehensive analysis of the inherent orientation question of the Crh monomers). The accuracy improves to 1.62 Å when considering the hypothetic monomer. Besides showing that the accuracy obtained for Crh is comparable to the one achieved for other proteins using solid-state NMR methods (see below), this also underlines that multimeric protein structure determination by NMR remains a challenge.

Unambiguous distance restraints	
total	643
sequential ($\mid i-j \mid = 1$)	181
medium-range ($1 < \mid i-j \mid < 5$)	85
long-range ($\mid i-j \mid > 4$)	131
Backbone dihedral angle restraints	58
Distance restraints violations	
>0.50 Å	0
>0.10 Å	2.20 ± 1.30
RMS (Å)	0.012 ± 0.005
NMR ensemble precision[a] (RMSD, Å)	
monomer, backbone	0.84 ± 0.12
monomer , SSE[b]	0.76 ± 0.09
dimer, backbone	1.33 ± 0.23
dimer, SSE	1.20 ± 0.24
dimer, SSE all heavy atoms	1.74 ± 0.21
Accuracy[c] (RMSD, Å)	
monomer	1.62
dimer	2.89
Ramachandran data[d]	
residues in most favored region (%)	74.2
residues in allowed regions (%)	20.6
residues in generously allowed regions (%)	4.0
residues in disallowed regions (%)	1.2
angle between the monomers (°)	
final set of ssNMR structure	164.3 ± 4.0
X-ray structure (1MU4)	164.8
distance between centres of mass (Å)	
final set of ssNMR structure	22.1 ± 0.7
X-ray structure (1MU4)	20.7

Table 4.4: Structural statistics for the 10 lowest-energy conformers of the Crh dimer protein calculated with XPLOR-NIH using the unambiguous distance restraints assigned after the second ARIA run

---

[a] calculated for residues 2-12 from chain A, and residues 13-80 from chain B for the monomer, and residues 2-80 from both chains for the dimer.

[b] SSE: secondary structure elements.

[c] from the X-ray structure (PDB entry 1MU4 [Juy et al., 2003]; calculated on secondary structure elements

[d] from PROCHECK [Laskowski et al., 1993]

4.2.4 Comparison of the convergence of the Crh structure with that of other protein structures obtained from ssNMR data.

The convergence of the Crh dimer structure calculated here was compared to the convergence of protein structures previously calculated from ssNMR data by calculating the backbone RMSD obtained on the NMR conformers. Two structures had been determined from extensively labelled protein samples: SH3 (62 residues) and Ubiquitin (76 residues). The SH3 structure shows a precision of 0.7 Å (calculated on the Cα atoms in the secondary structure elements), and an accuracy of 1.3 Å with respect to the crystal structure [Castellani et al., 2002, 2003]. Ubiquitin shows a precision on the backbone atoms of about 1.0 Å; no accuracy is given [Zech et al., 2005]. Two structures were obtained for small fully labelled proteins by means of simple manual assignments: Kaliotoxin (38 residues) [Lange et al., 2005] and GB1 (56 residues) [Zhou et al., 2007; Peng et al., 2008]. They show backbone RMSDs of 0.81 Å and 0.82 Å respectively, and an accuracy of 1.9 Å. More recently, the incorporation of relative backbone orientation restraints, derived from dipolar line shapes, allowed to reach an backbone accuracy of 1.43 Å for protein GB1 [Franks et al., 2008]. The backbone RMSD of 1.33 Å (1.20 Å on secondary structure elements) and the accuracy of 2.89 Å obtained for the fully labelled Crh dimer (170 residues) is slightly higher than the values obtained for the above proteins; this increase can mainly be explained by the larger size of the protein, combined with the additional difficulty of the calculation of an elongated dimeric molecule. This is supported by the RMSD calculated for a hypothetic monomer, where the backbone RMSD for regular structure elements decreases to 0.76 Å, and which shows an accuracy of 1.62 Å, values which compare favourably to those of previously determined ssNMR structures.

4.3 Conclusion

The symmetry-ADR method integrated in the program ARIA found to be an effective way to correctly assign ambiguous cross-peaks in symmetric homo-dimers from both solution and solid-state NMR. Moreover, the relative effects of network anchoring and spin-diffusion correction were assessed on a set of symmetric homo-dimers corresponding to different structure interfaces in solution. Although producing less precise structures, the spin-diffusion correction appears to improve the structural quality, especially the number of bumps, as well as the number of NOE violations. This observation should encourage a quantitative analysis of NOE intensity during NMR calculation as well as the use of quantitative target distances [Rieping et al., 2005b]. The comparison of homo-dimers calculations with those conducted in similar conditions on monomeric proteins shows that the effects of these methods are similar in monomers or in homo-dimers. Furthermore, similar levels of structure quality are obtained, for each structure studied, in the presence or absence of network anchoring and spin-diffusion correction. The adaptation of an efficient automated assignment tool such as the network anchoring in the ARIA software allows now to address the difficult problem of assigning NOE spectra of symmetric oligomers, without degrading the overall quality of the structures. For the second time, we also confirm here that the observation of local quality parameters contributes significantly to the detection of possible errors during the automated assignment process (cf. § 3.1).

We also could show how highly ambiguous proton-mediated, rare-spin detected solid-state NMR data sets of a fully labelled protein sample can be used for structure calculation through automated iterative assignments. The input of over 1000 cross-signals as ambiguous restraints, 25 inter-monomer restraints and chemical shifts derived dihedral angles yields a total of 643 unambiguously assigned distance restraints, including 131 long range restraints. The Crh dimer structure calculated using this data shows a precision of 1.33 Å, and an accuracy with respect to the crystal structure of 2.89 Å. Crh is the largest protein structure which has been determined so far from solid-state NMR data; this work shows that even complex structural features, like the dimeric and elongated nature of Crh, are not an obstacle to high-resolution structure determination by solid-state NMR.

5

Analysis of the conformational landscape of Crh from solution and solid-state NMR data

5.1 Introduction

Solid-state NMR is becoming a tool for structural biology, aiming notably at structure elucidation of insoluble proteins, e.g. membrane proteins, cytoskeletal proteins and protein fibrils. Simple oligomerisation of monomeric proteins, or more complex interactions including soluble and insoluble protein fragments are at the origin of the formation of multimeric or polymeric assemblies. Monomeric proteins are often soluble, or at least fragments of them. Indeed, several membrane proteins are formed by multimerisation of small subunits, as viroporins [Park et al., 2003], light-harvesting complexes [Roszak et al., 2003], phospholamban [Andronesi et al., 2005]. Similarly, prion proteins can be often divided [Wider et al., 1998; Wasmer et al., 2008; Bousset et al., 2004] into a soluble globular part and a prion-forming domain, as this is also the case for Het-s [Wasmer et al., 2008], the human prion protein [Helmus et al., 2008] and Ure2p [Bousset et al., 2004]. Cytoskeleton proteins are also often made up of soluble protein subunits. All these proteins have in common that the structural study of the soluble part can be carried out by solution NMR, in aqueous solution or in detergents. The possibility of a combined use of restraints from NMR/X-ray, and information, only accessible from solid-state NMR, about the interface in the native polymer or the membrane protein in lipid bilayers, would be of great use to analyse the protein conformational landscape during the transition between the monomeric and multimeric state.

We use as a model the *Bacillus subtilis* Crh protein (Catabolite repression HPr), a phosphocarrier protein of the phosphoenolpyruvate:carbohydrate phosphotransferase system (Fig. 5.1), which displays a large conformational variability. Its monomeric structure was determined in solution by liquid-state NMR (PDB id: 1k1c, Favier et al. [2002]) and is close to the HPr structure [Azuaga et al., 2005; Kalbitzer and Hengstenberg, 1993; Hahmann et al., 1998]. A conformational exchange between a monomeric and a dimeric structure was observed in solution, but selective precipitation of the dimeric protein hampered dimer structure determination in solution [Penin et al., 2001]. The dimeric structure was then determined in the crystalline form by X-ray crystallography (PDB ids: 1mo1, 1mu4 Juy et al. [2003]), as well as from micro-crystals by ssNMR (see § 4.2, PDB id: 2rlz, Loquet et al. [2008]). The transition from the monomeric to the dimeric form consists in the 3D domain swapping of strand β1 (Fig. 5.1b,c). In the crystal, the dimers are interacting two-by-two to form a tetramer. The wealth of data available for this protein, and its conformational variability make Crh a good example to evaluate the feasibility of structure determination for multimeric, insoluble proteins. As the ssNMR was only recently applied to proteins of this size, Crh occupies a quite unique position according to the availability of data from solution NMR, ssNMR and X-ray crystallography.

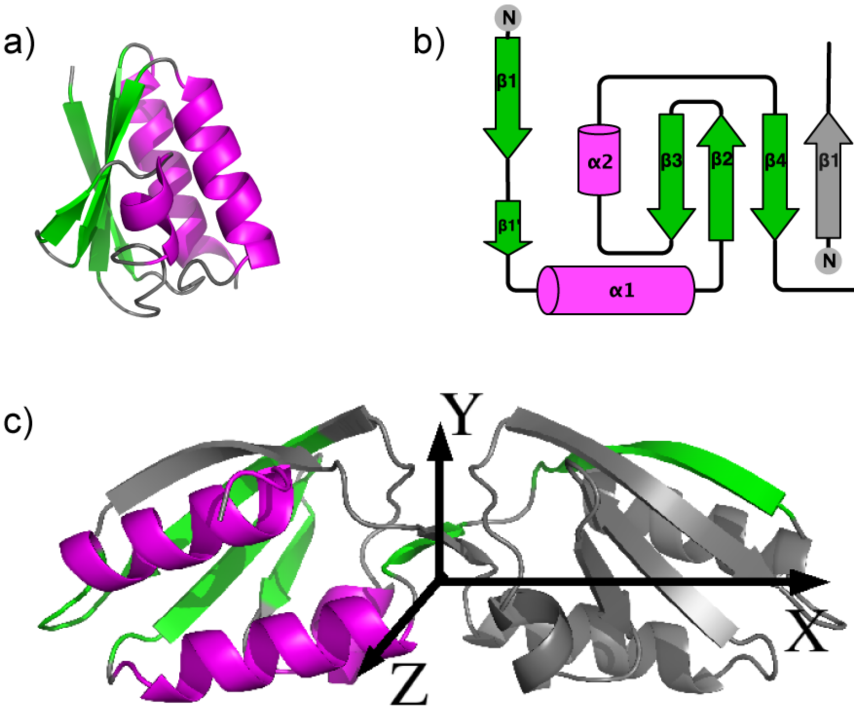


Figure 5.1: Structures of the Crh protein. (a) The monomer structure determined by liquid-state NMR (PDB id: 1k1c) contains 3 α helices (α1: residues 17-28, α2: residues 47-50 and α3: residues 70-83) as well as a β sheet formed from four β strands (β1: residues 3-9, β2: residues 31-37, β3: residues 40-42 and β4: residues 60-67). (b) Topology of a Crh monomer, inside the dimer, (c) Dimer structure determined by X-ray crystallography (PDB id: 1mo1). The β strands are in green and the α helices in magenta. In the dimer (b,c), one monomer is coloured and the other one displayed in grey. In c, the axes X, Y, Z allowing to define the angles Ψ, Θ and Φ are drawn.

The objectives of this chapter are to study the possibility to use NMR restraints from different origins (solution and solid-state NMR) as well as X-ray crystallography to analyse the interface and the relative orientation of the monomers during the transition from solution to crystal. Getting structural information about such a transition is rare and could be of great value to understand the crystallisation process of proteins. The development of ssNMR provides the opportunity to obtain information about states which are the starting points of crystallogenesis [Schmidt et al., 2007].

Three series of NMR conformer generation were performed using a version of ARIA dedicated to ssNMR (§ 2.4). The first series was performed using precise distance restraints determined on the X-ray crystallographic structure, in order to evaluate the convergence attainable with the procedure. The two last series explore the conformational landscape of Crh dimers during the oligomerisation and the crystal formation. The sets of ARIA conformers are compared to MD simulations starting from crystallographic structures of the Crh dimers and tetramers. We observe that the conformational heterogeneity of the conformers is concentrated in the relative orientation of the Crh monomers. This heterogeneity is conserved even if non-crystallographic symmetry within the crystal tetramers is taken into account. The variability of monomers orientation inside the dimer is similar among ARIA conformers, during the MD simulations, and in a crystallographic ensemble refinement along structure factors. A balance between protein-water interactions and protein-protein interactions plays a crucial role in the stabilisation of the monomer orientation observed in the crystal.

5.2 Materials and Methods

5.2.1 Input files of the conformers calculations

The NMR assignments of Crh in solution [Favier et al., 2002] and in the crystal [Böckmann et al., 2003] were obtained from the BMRB [Ulrich et al., 2008] (ids: 4972 and 5757). The inter-monomer assigned cross-peaks measured on the NHHC spectrum [Etzkorn et al., 2004] were also used. The two crystallographic dimeric structures [Juy et al., 2003] (PDB ids: 1mo1, 1mu4) were used to analyse the accuracy of the conformers. The monomeric structure (PDB id: 1k1c) [Favier et al., 2002] and the corresponding restraint file provided a synthetic NMR peak list for the monomer. The ψ and ϕ dihedral angle restraints were determined from TALOS [Cornilescu et al., 1999], using the ssNMR chemicals shifts (BMRB id: 5757). This prediction also yields 26 hydrogen bonds in α-helices.

5.2.2 ARIA-CNS calculation

Conformers were generated using the version of ARIA presented in § 2.3 and the version 1.1 of CNS [Brünger et al., 1998]. An iterative ARIA calculation was used to filter the monomer NMR restraints using the ssNMR chemical shifts: eight iterations in vacuum were performed starting from the synthetic monomer peak list described above and the ϕ, ψ dihedral restraints determined from the ssNMR chemical shifts. The other ARIA calculations (Fig. 5.2) did not perform an iterative assignment of restraints, and required only a single iteration in vacuum, generating 360 conformations. The main differences with an usual ARIA structure refinement is a larger number of steps (50.000 to 100.000 steps) during the second cooling stage, as already proposed by Fossi et al. [Fossi et al., 2005c]. The 50 lowest-energy structure were then submitted to a refinement step with non-crystallographic symmetry (NCS) restraints to calculate tetrameric conformers, and analysed afterward by two clustering methods. A water refinement with NCS restraints was then performed on one or two resulting clusters.

Non-crystallographic symmetry (NCS) restraints were applied using a strict NCS relationship [Brunger, 2007] reproducing the transformation (rotation and translation) between the two dimers forming the tetramer structure in 1mo1. Only coordinates for one dimer were calculated, the coordinates of the other dimer being generated using the NCS. The interaction between dimers was calculated using a simple repulsive term, and electrostatic Coulombic interactions were added during the water refinement step.

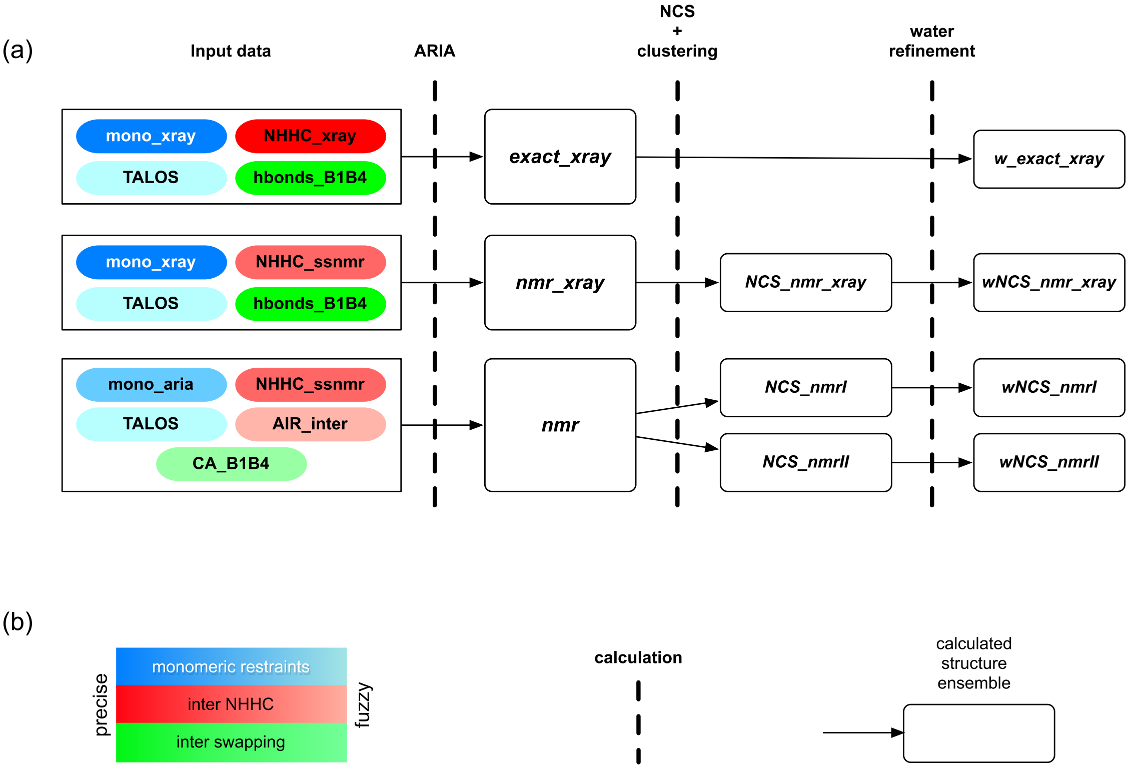


Figure 5.2: ARIA calculations scheme. (a) The coloured boxes on the left describe the input data, the blue denoting intra-monomer restraints, the red denoting inter-monomer restraints coming from the interpretation of the NHHC spectrum, the green denoting restraints of the swapping topology. The dark colours stand for precise restraints, whereas light colours stand for fuzzy restraints. The dashed vertical lines correspond to the calculation or to the processing of conformers, and the sets of conformers are given inside the white rectangles. (b) Visual legend of (a).

5.2.3 Structure analysis

The RMSD between the backbone atoms was calculated by iterative fitting [Bardiaux et al., 2008b]. Residues 1-85 were superimposed for the dimers, and residues 15-85 for the monomers. The fit of the structure to the experimental restraints is evaluated by determining the number of violated distance restraints and the RMS of violations. A restraint is considered as violated if the distance is larger than U+0.5 Å or smaller than L-0.5 Å, where U and L are the restraint upper and lower bounds. The quality of the structures is described by the PROCHECK percentages of residues in the core and allowed regions of the Ramachandran diagram, and several WHAT IF Z scores: the 2nd-generation packing score (NQACHK), the Ramachandran score (RAMCHK), the $\chi1/\chi2$ rotamer normality (C12CHK), the backbone conformation score (BBCCHK), the inside/outside distribution of hydrophobic and hydrophilic residues (INOCHK), and the number of WHAT IF inter-atomic bumps (BMPCHK).

The relative monomer orientation is described through the distance between the monomer centres-of-mass and through the Euler angles Ψ, Θ and Φ of the rotations around the axes X, Y and Z, describing the transformation from one monomer to the other, the axes X, Y, Z being aligned along the principal inertia axes of the structure 1mu4 (Fig. 5.1c). The variation of Crh tertiary structure is monitored by calculating the minimum distances between axes of secondary structure elements. In the α-helices the points belonging to the axes are determined as the middles of the backbone atom segments (N(i), N(i+2)), (Cα(i), Cα(i+2)) and (C'(i), C'(i+2)), where the i is the residue number. In the β strands, axes are defined from the positions of backbone heavy atoms.

5.2.4 Clustering algorithms

Two clustering algorithms were used to analyse the conformers. The first clustering algorithm *clustering-I*, similar to the one used in HADDOCK [Dominguez et al., 2003], is based on the iterative processing of the pairwise coordinates RMSD matrix, using a RMSD cutoff and a minimal size of a cluster. The conformers are sorted into clusters, the cluster having the largest size is removed from the conformer pool, and the algorithm is run again on the remaining conformations. The distance cutoff varies from 2 to 4 Å, by steps of 0.1 Å, and the minimal cluster size varies from 5 to 10 by steps of 1, a given conformation belonging generally to several clusters. The second clustering algorithm (*clustering-II*) groups conformers in two dimensions (clusters I versus conformers), by two successive hierarchical clusterings (command hclust of R [Gordon, 1999]), applied first on the conformer axis, and then on the cluster axis.

5.2.5 Molecular dynamics simulations

The X-ray crystallographic structure 1mo1 [Juy et al., 2003] containing a Crh tetramer (chains A, B, C, D) was used as a starting point for the molecular dynamics (MD) simulations (see Appendix). The simulations were performed with periodic boundary conditions: *sol_dimer* on the chains A and B, and *sol_tetra*, *cryst_tetra* on the chains A, B, C and D. The simulations *sol_dimer* and *sol_tetra* intend to describe the behaviour of the Crh dimer and tetramer in solution, whereas the simulation *cryst_tetra* models the Crh tetramer in a more restricted environment, including qualitatively crystal packing. A similar approach was used recently [Walser et al., 2002] to simulate proteins in the crystalline state.

5.2.6 Ensemble crystallographic refinement

The tetrameric Crh structure was refined along the structure factors measured on 1mo1 (file: 1mo1-sf.cif). The protocol for ensemble crystallographic refinement, recently proposed [Levin et al., 2007], was used to generate 16 conformations of the tetramers. The calculation was performed using CNS 1.2 [Brunger, 2007]. The starting conformation was the one observed in the PDB structure 1mo1. The water molecules and the cofactors (sulfate ions and glycerol) were not duplicated. Ten percents of the structure factors were used for data cross-validation. A R factor of 0.14 and a free R factor of 0.18 were observed on the set of conformers.

5.3 Results

5.3.1 Filtering of the liquid-state NMR restraints

The comparison of the dimer crystallographic structure [Juy et al., 2003] and of the monomer NMR structure [Favier et al., 2002] supports the hypothesis that the intra-monomer architecture is partly conserved during the oligomerisation and crystallisation processes. However, the ss-NMR [Böckmann et al., 2003] and solution NMR [Favier et al., 2002] chemical shifts are different in the N-terminal $\beta 1$ strand, as well as in the region around loop comprising residues 10-15 [Böckmann et al., 2003]. All monomer restraints are thus not valid in the dimer, and one needs an objective filtering method. The use of ssNMR chemical shifts is a way to filter these restraints, while avoiding the use of the crystallographic structure. As the same spectrum is observed in the micro-crystalline and precipitate states of Crh [Etzkorn et al., 2007], the restraints filtered using ssNMR chemical shifts can give information on the conformational landscape of Crh in the precipitate as well.

In order to guide the building of the dimer architecture during the calculation, the interactions between strands $\beta 1$ and $\beta 4$ were imposed through distance restraints corresponding to hydrogen bonds (1.2-2.2 Å: HN-O, 2.2-3.2 Å: N-O) between the β strands. This information is directly extracted from the X-ray crystallographic structure of Crh. As a possible interaction between $\beta 1$ and $\beta 4$ strands has indeed been predicted for Crh from solution NMR chemical shifts [Favier et al., 2002], one can assume that this interaction is formed very early during the protein oligomerisation and crystallisation. For other molecules, equivalent information about interacting segments could be accessed in by mutational studies [Fay et al., 2005], electron microscopy [Ranson et al., 2006] or molecular dynamics simulations [Paci et al., 2004; Cecchini et al., 2006].

An iterative ARIA calculation was performed to filter among the monomer NMR distance restraints those still verified in the dimer structure. The inputs were: (i) the dihedral and intra-monomer hydrogen bond restraints deduced from the TALOS analysis of the ssNMR chemical shifts, (ii) the synthetic peak list corresponding to the NMR distance restraints of 1k1c concerning the residues 16-85 for which the chemical shift does not change between solution and ssNMR assignment. In this way, the restraints involving the region 1-15 were excluded from the filtering process. The synthetic peak list was submitted to ARIA along with the assignments already performed during the structure calculation of 1k1c. These assignments were used to limit the number of possible contributions to each peak and consequently reduce the combinatorial analysis for the generation of ambiguous distance restraints. The restraints resulting from the

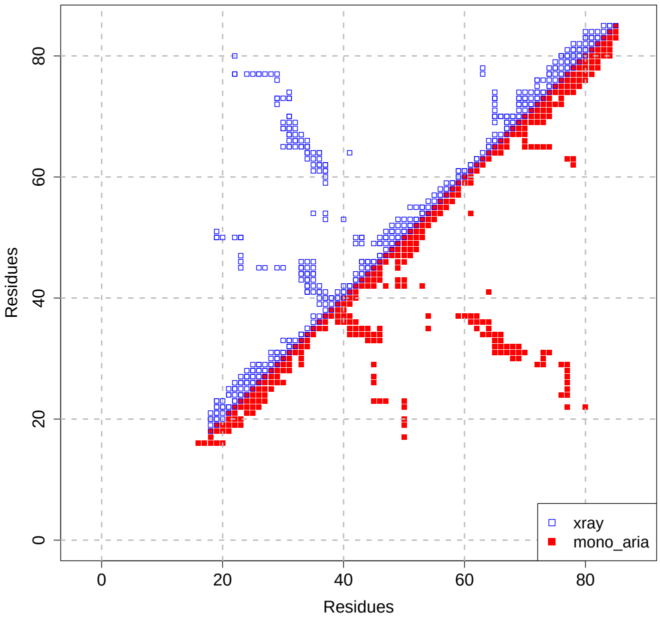


Figure 5.3: Comparison of the intra-monomer restraints obtained by filtering on the X-ray structure 1mo1 (upper triangle, blue empty square) and of the intra-monomer restraints obtained by filtering monomer peak list using ssNMR chemical shifts (lower triangle, red filled squares). The restraints are plotted along the residue numbers.

automatic assignment protocol of ARIA, run on a symmetric homo–dimer, finally provided the mono_aria set of restraints, used as the intra-monomer restraints during the ARIA calculations described below.

The contact map obtained from restraints mono_aria (Fig. 5.3, upper triangle) is quite similar to the contact map obtained by filtering the 1k1c restraints directly on the structure 1mo1 (Fig. 5.3, lower triangle). The filtering of the liquid-state NMR restraints by the ssNMR chemical shifts produces thus restraints similar to the ones observed in the X-ray crystallographic structures. As the solution NMR restraints involving the region 1-15 were excluded from the filtering process, the restraints of the same region filtered on the structure 1mo1 are not displayed on the contact map.

5.3.2 Presentation of data inputs and dimeric calculations

Data inputs for the calculations include the mono_aria restraint set and the TALOS restraints sets, described above, as well as the following sets of inter-monomer restraints (Fig. 5.2). Exact distances, measured on the structure 1mo1, for the inter-monomer peaks assigned on the NHHC spectrum [Etzkorn et al., 2004], produce the restraints NHHC_xray. The NMR restraints built from the NHHC peaks, using invariant bounds 2.5-4.5 Å, correspond to the restraints NHHC_ssnmr. Additional inter-monomer restraints between the strands $\beta 1$ and $\beta 4$ were applied using hydrogen bond restraints (hbonds_B1B4) or using restraints between Cα (CA_B1B4), determined from 1mo1. The restraint bounds are, for hbonds_B1B4 (1.2-2.2 Å: HN-O, 2.2-3.2 Å: N-O) and for

CA_B1B4 (4.5-5.5 Å: Cα-Cα). Ambiguous inter-monomer restraints (AIR_inter) were applied between all residues assigned to the dimer interface in solution by chemical shift perturbation [Favier et al., 2002]. This ambiguous restraint application is similar to the one used in HADDOCK [Dominguez et al., 2003].

In the ARIA runs, no iterative peak list assignment was used, and the protein conformers were obtained after one iteration followed by a water refinement step. Three sets of conformers were generated using the ARIA protocol (Fig. 5.2). The first conformer set (*exact_xray*) was calculated using the restraints: mono_xray, NHHC_xray, TALOS and hbonds_B1B4. These calculations intended to evaluate the precision attainable using distance restraints. This calculation was followed by a water refinement step (*w_exact_xray*), and was compared to an ensemble refinement along the crystallographic structure factors [Levin et al., 2007]. The second conformer set (*nmr_xray*) was obtained using the restraints: mono_xray, NHHC_ssnmr, TALOS and hbonds_B1B4, in order to explore the conformational landscape in presence of restraints close to those observed in the crystal. Finally, the third conformer set (*nmr*), based on the restraints: mono_aria, NHHC_ssnmr, TALOS, CA_B1B4 and AIR_inter, was devoted to the description of the conformational landscape of the Crh dimer and tetramer during oligomerisation and crystal formation. The conformations obtained from the set *nmr_xray* and *nmr* were further refined using additional non-crystallographic symmetry (NCS) restraints to produce the sets *NCS_nmr_xray* and *NCS_nmr*. An additional water refinement step was finally performed on a cluster extracted from *NCS_nmr_xray* and on two clusters extracted from *NCS_nmr*, to produce the sets *wNCS_nmr_xray*, *wNCS_nmrI* and *wNCS_nmrII*. The sets *wNCS_nmrI* and *wNCS_nmrII* will be described more precisely in the § 5.3.3.

The application of the NCS restraints intends to apply a long-range order to the ARIA conformers similar to the one observed in Crh crystals or precipitate. As the crystallisation conditions are different for the samples used in ssNMR [Böckmann et al., 2003] and in crystallographic [Juy et al., 2003] studies, the NCS restraints represent only a qualitative modelling of the long-range crystal or precipitate order.

5.3.3 Convergence of the calculation and fit to the NMR restraints

Clustering-I, described in § 5.2.4 was performed on the 50 lowest-energy conformers. The clustering was analysed (Table 5.1) through the number of clusters, the cluster sizes, the backbone precision of the clusters and the backbone accuracy with respect to the X-ray structure 1mu4. A similar analysis was performed for the 50 conformers of the sets *NCS_nmr_xray* and *NCS_nmr*, refined in gas-phase with non-crystallographic symmetry (NCS) restraints. The small number of clusters detected for *exact_xray* and their large sizes prove the convergence of the conformations. For *nmr_xray*, the slight decrease clusters precision and accuracy, and the appearance of 5-members clusters, reveal a diminution of the convergence. Nevertheless, more than 9 clusters are larger than 30 members, and the lower bounds of the RMSD are similar to those observed for *exact_xray*. The refinement of the *nmr_xray* conformers, performed in vacuum with the appli-

Set	Number of clusters	Clusters size	Accuracy (Å)	Precision (Å)
exact_xray	6	38-45	2.2-2.3	1.5-1.8
nmr_xray	19	5-44	2.4-3.3	1.3-2.1
NCS_nmr_xray	4	41-45	2.5-2.6	1.3-1.5
nmr	33	5-21	3.7-6.8	1.5-2.4
NCS_nmr	35	5-26	3.0-7.3	1.4-2.3

Table 5.1: Results of the clustering-I performed on the 50 best-energy conformers generated in vacuum (exact_xray, nmr_xray, nmr), and performed on the 50 conformers obtained after a refinement in presence of NCS restraints (NCS_nmr_xray, NCS_nmr). The number of clusters obtained is given along with the range of cluster sizes, the range of the coordinate RMSD to the crystallographic structure 1mu4, and the range of the coordinate RMSD between the conformers inside each cluster. The RMSD values were calculated by superimposing the backbone heavy atoms.

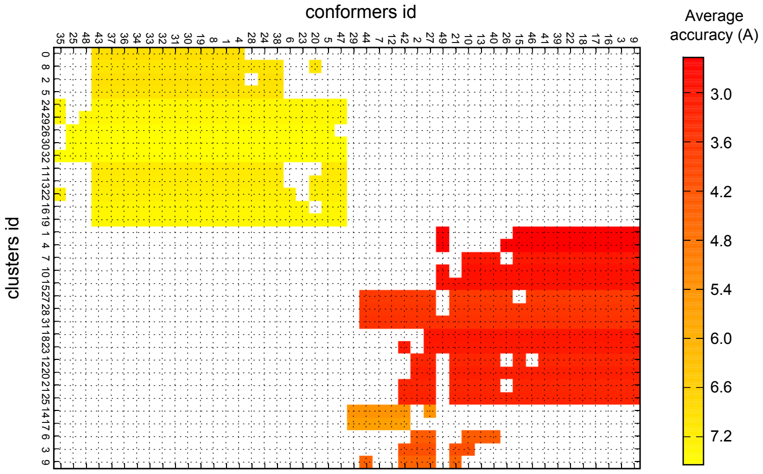


Figure 5.4: Results of the clustering-II algorithm applied on the ARIA conformers obtained in the run NCS_nmr. The x-axis describes the conformers id, whereas the y-axis describes the clusters id. The average RMSD of each cluster with respect to the X-ray structure 1mo1 is plotted according to the color scale given.

cation of NCS restraints (*NCS_nmr_xray*), displays convergence and accuracy close to the ones observed for *exact_xray*. In the set *nmr*, the convergence towards 1mu4 is not obtained, as the RMSD distance to 1mu4 is doubled with respect to other runs. Also, the number of clusters is doubled, and the maximum cluster size divided by two with respect to *exact_xray*. The application of NCS restraints (*NCS_nmr*) does not modify much this situation, which may be due to the relatively local symmetry applied.

The clusters obtained by the clustering-I method on *nmr_xray*, *NCS_nmr_xray* and *NCS_nmr*, were then analysed through a hierarchical clustering method (clustering-II) described in § 5.2.4. The method clustering-II detected one group of conformers in *nmr_xray* and *NCS_nmr_xray*, and two groups in *NCS_nmr* (Fig. 5.4). This detection of two groups was performed independently of the knowledge of the crystallographic structure, producing a first group displaying an RMSD to 1mo1 around 3 Å, and a second one around 7 Å. A cluster closest to the structure 1mu4 was extracted from the first group (*NCS_nmrI*) and was supplemented by the largest cluster extracted from the second group (*NCS_nmrII*). Clusters extracted from the sets *exact_xray*, *NCS_nmr_xray* and *NCS_nmr* were submitted to a refinement step in water and in presence of NCS restraints

(Fig. 5.2). For *exact_xray* and *NCS_nmr_xray*, the 38 best-energy conformers and the cluster of 41 conformers (clustering-I) closest to the structure 1mu4 respectively, were selected. In *NCS_nmr*, the 12-members cluster (*NCS_nmr_I*) and the 22-members cluster (*NCS_nmr_II*), described above, were selected. These four clusters were then refined in water and in presence of NCS restraints, to provide the sets *w_exact_xray*, *wNCS_nmr_xray*, *wNCS_nmrI* and *wNCS_nmrII*, which will be analyzed below, along with the sets *exact_xray*, *nmr_xray* and *nmr* obtained.

The clustering simplifies the description of the Crh conformational landscape. The use of NCS restraints improves the structure convergence (Table 5.2), and along with clustering techniques, allows to detect sets of conformers (*NCS_nmr_xray*, *wNCS_nmrI*) displaying accuracy smaller than 2.6 Å to 1mu4. The dimer precisions obtained for *w_exact_xray*, *wNCS_nmr_xray* and *wNCS_nmrI* are similar to the precision (1.3 Å) of the ssNMR structure of Crh determine previously with ARIA (§ 4.2, Loquet et al. [2008]). The conformer local RMSD along the sequence (Fig. 5.5b) qualitatively resembles to the fluctuations by residues in MD simulations (Fig. 5.5a) and to the B factors in 1mo1 (Fig. 5.5c), with local maxima located in the same protein regions (residues 27, 40, 57, 60, 67). The monomer convergence is good for all sets, with coordinate RMSD values in the 0.6-0.9 Å range, close to the value of 0.8 Å obtained on the ssNMR structure (§ 4.2). The convergence only slightly decreases when looser restraints are applied. On the other hand, the RMSD calculated on the whole dimer, is larger: the lack of convergence of the Crh dimer comes mainly from the lack of variability in the relative orientation of the two monomers.

The small numbers restraint violations in all conformers sets along with violation RMS in the 0.11-0.16 Å range prove the good fit of the conformations to the restraints. The introduction of non-crystallographic symmetries and of water molecules induces a better fitting of the conformation to the restraints with respect to the corresponding calculations in vacuum.

To summarise, the lack of convergence in the Crh dimer mainly arises from a lack of definition in the monomer orientation. The conformers convergence decreases as fuzzier restraints are applied. On the other hand, the use of a long-range order, reminiscent of the crystal configuration, improves the dimer convergence. A relatively limited long-range order, involving only the Crh tetramer, permits to obtain most of the convergence toward the crystallographic structures. This is an argument in favour of the appearance of the tetramer interaction early in transition path from solution to crystal.

	exact_xray	*nmr_xray*	*nmr*	*w_exact_xray*	*wNCS_nmr_xray*	*wNCS_nmrI*	*wNCS_nmrII*
# conformers	50	50	50	38	41	12	22
PROCHECK (%) core	86.6 ± 2.2	87.8 ± 2.5	86.3 ± 3.5	92.8 ± 2.3	92.1 ± 2.2	91.6 ± 3.4	91.0 ± 2.7
PROCHECK (%) allowed	12.9 ± 2.3	12.0 ± 2.5	13.1 ± 3.2	6.8 ± 2.5	7.5 ± 2.2	7.7 ± 3.2	8.4 ± 2.9
NQACHK	-1.5 ± 0.4	-2.0 ± 0.4	-3.1 ± 0.5	-0.2 ± 0.4	-1.0 ± 0.5	-1.9 ± 0.5	-2.5 ± 0.5
RAMCHK	-3.5 ± 0.4	-3.7 ± 0.4	-2.5 ± 0.6	-3.0 ± 0.5	-3.1 ± 0.5	-1.2 ± 0.7	-2.1 ± 0.7
C12CHK	-1.5 ± 0.4	-1.4 ± 0.4	-0.2 ± 0.6	-2.2 ± 0.5	-1.9 ± 0.6	-0.7 ± 0.7	-2.1 ± 0.5
BBCCHK	0.3 ± 0.5	0.3 ± 0.5	0.1 ± 0.7	0.6 ± 0.3	0.4 ± 0.3	0.7 ± 0.7	0.1 ± 0.6
INOCHK	0.5 ± 0.5	0.5 ± 0.5	0.5 ± 0.5	1.0 ± 0.02	1.0 ± 0.02	0.5 ± 0.5	1.0 ± 0.03
BMPCHK	36.4 ± 6.2	35.4 ± 6.1	19.6 ± 5.5	27 ± 6.6	26.2 ± 7.6	19.3 ± 4.4	46.2 ± 13.8
# viol. $\geq$ 0.5 Å	58.4 ± 8.4	56.2 ± 8.7	33.9 ± 7.3	52.0 ± 6.4	46.8 ± 6.6	19.3 ± 5.3	21.6 ± 5.5
Violation RMS (Å)	0.16 ± 0.01	0.16 ± 0.01	0.14 ± 0.02	0.15 ± 0.01	0.14 ± 0.01	0.12 ± 0.02	0.11 ± 0.02
Precision Monomer (Å)	0.6 ± 0.2	0.8 ± 0.1	0.9 ± 0.2	0.6 ± 0.2	0.7 ± 0.2	0.7 ± 0.1	0.9 ± 0.2
Precision Dimer (Å)	2.5 ± 2.5	2.8 ± 2.1	5.7 ± 2.5	1.2 ± 0.4	1.1 ± 0.3	1.2 ± 0.4	1.8 ± 0.6
Accuracy Monomer (Å)	1.0 ± 0.2	1.3 ± 0.2	1.8 ± 0.2	1.0 ± 0.2	1.2 ± 0.2	1.6 ± 0.1	1.9 ± 0.2
Accuracy Dimer (Å)	3.1 ± 2.3	3.3 ± 2.0	6.7 ± 2.3	2.0 ± 0.3	2.3 ± 0.3	2.4 ± 0.4	7.6 ± 1.0
Ψ (X) (°)	-66.7 ± 26.2	-69.5 ± 32.4	-110.7 ± 39.8	-61.6 ± 15.2	-68.4 ± 11.0	-62.9 ± 15.2	-124.3 ± 34.9
Θ (Y) (°)	10.3 ± 2.7	10.7 ± 3.1	11.0 ± 5.3	11.3 ± 2.2	12.4 ± 1.2	13.2 ± 2.2	9.7 ± 3.8
Φ (Z) (°)	-7.6 ± 3.8	-9.0 ± 5.2	-19.9 ± 24.2	-7.5 ± 3.1	-8.9 ± 2.3	-8.0 ± 3.4	-17.8 ± 5.3
$D_{A\text{-}B}$ (Å)	20.6 ± 0.7	20.7 ± 1.0	20.1 ± 2.1	20.0 ± 0.6	19.9 ± 0.5	19.4 ± 0.9	21.5 ± 0.9

Table 5.2: Quality, convergence and restraint fitting of the ARIA conformers calculated in gas phase (*exact_xray*, *nmr_xray*, *nmr*) and refined in presence of NCS restraints and water (*w_exact_xray*, *wNCS_nmr_xray*, *wNCS_nmrI*, *wNCS_nmrII*). Relative position of the monomers inside the dimer in the ARIA conformers, described through the distance between the monomer centers of mass, and through the angles Ψ, Θ and Φ (see § 5.2.3).

	sol_dimer	*sol_tetra*	*cryst_tetra*	*exact_xray*	*nmr_xray*	*nmr*	*w_exact_xray*	*wNCS_nmr_xray*	*wNCS_nmrI*	*wNCS_nmrII*
intra-monomer										
α2-β1a	3.3 ± 0.5	4.0 ± 0.2	4.4 ± 0.2	5.0 ± 0.2	5.1 ± 0.2	7.6 ± 2.4	4.8 ± 0.2	5.0 ± 0.2	6.8 ± 1.8	7.2 ± 2.5
α2-β3	5.5 ± 0.5	5.8 ± 0.2	5.6 ± 0.2	5.9 ± 0.1	5.9 ± 0.1	5.0 ± 0.4	5.7 ± 0.1	5.8 ± 0.2	5.8 ± 0.3	5.2 ± 0.4
α3-α1	9.0 ± 0.3	9.5 ± 0.2	9.4 ± 0.3	9.1 ± 0.2	9.2 ± 0.2	9.8 ± 0.3	9.1 ± 0.2	9.2 ± 0.2	9.3 ± 0.3	9.5 ± 0.3
α3-β4	6.0 ± 0.2	6.3 ± 0.2	6.3 ± 0.2	6.1 ± 0.1	6.1 ± 0.2	6.3 ± 0.1	6.0 ± 0.2	6.1 ± 0.3	6.4 ± 0.1	6.3 ± 0.2
β2-β3	3.0 ± 0.1	3.2 ± 0.2	3.4 ± 0.1	3.1 ± 0.1	3.1 ± 0.1	3.1 ± 0.1	3.1 ± 0.1	3.2 ± 0.1	3.2 ± 0.1	3.2 ± 0.1
β4-β2	3.1 ± 0.3	3.9 ± 0.1	3.9 ± 0.1	3.7 ± 0.03	3.7 ± 0.1	3.9 ± 0.1	3.6 ± 0.1	3.6 ± 0.1	3.9 ± 0.1	3.8 ± 0.1
inter-monomer										
β1a-β1a	3.4 ± 0.5	4.0 ± 0.1	4.0 ± 0.1	3.7 ± 0.3	3.8 ± 0.4	3.4 ± 0.4	3.9 ± 0.1	4.0 ± 0.2	3.6 ± 0.2	3.9 ± 0.3
β1-β4	3.4 ± 0.5	4.0 ± 0.1	4.0 ± 0.1	3.9 ± 0.2	3.7 ± 0.1	3.8 ± 0.2	3.7 ± 0.1	3.7 ± 0.1	3.7 ± 0.1	3.7 ± 0.1

Table 5.3: Distances (Å) between secondary structure elements in the ARIA conformers and during the MD simulations. In the simulations, the mean values were calculated over the 2-10 ns interval and over the monomers.

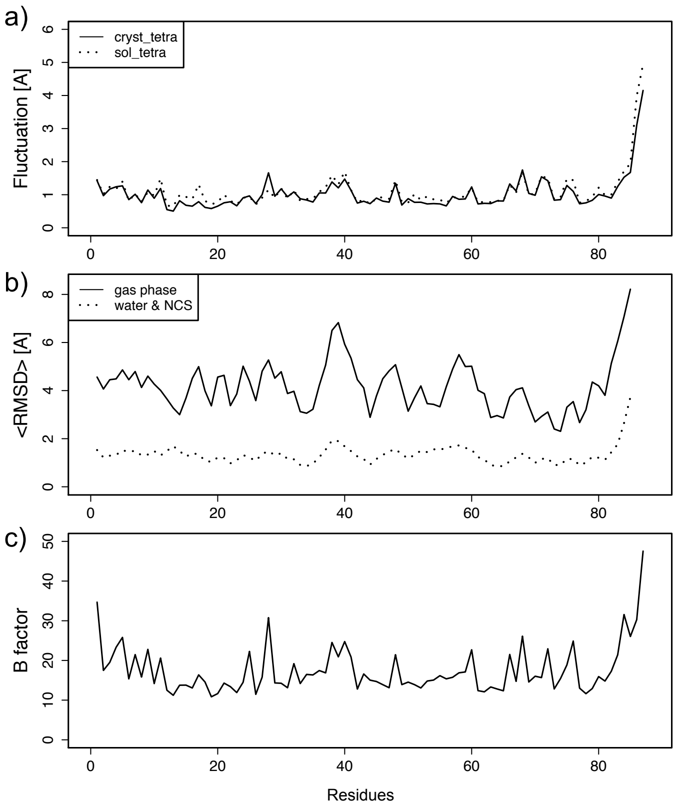


Figure 5.5: Comparison of the (a) fluctuations by residues (*cryst_tetra*: solid, *sol_tetra*: dotted), of (b) the mean local RMSD among ARIA conformers (gas phase: solid, water and NCS refinements: dotted) plotted along the sequence, and of (c) the mean B factors in the four chains of 1mo1.

5.3.4 Conformers quality and accuracy

The structure quality parameters, determined from PROCHECK and WHA TIF analyses (Table 5.2), correspond to good-quality solution NMR structures. Indeed, for all runs, more than 86% of the residues are located in the core PROCHECK Ramachandran diagram Similarly, the WHAT IF parameters are in the -4/4 range, the worse values being observed for Ramachandran plot Z-scores. These RAMCHK values are produced by the individual scores of residues 34, 37, 62, 66, located in β strands. The mean number of inter-atomic bumps (BPMCHK) observed by WHAT IF is around 36 for *exact_xray* and *nmr_xray*, and around 20 for *nmr*. The bumps are mainly observed for residues 27, 29, 47, 48, 69, 70, all located in α helices and on the same side of the Crh structure. All quality parameters are constant in the three runs in vacuum, which means that the application of looser restraints does not degrade the physical relevance of the generated conformations in the set *nmr*. The quality of the conformers calculated in presence of water (Table 5.2) shows an overall tendency to improvement with respect to vacuum, in agreement with the observations of the literature [Linge et al., 2003b]. The run *wNCS_nmrII* displays the highest number of bumps, arising from residues mainly located in the N terminal region 1-30, which is the sign of a badly defined dimer interface.

The application of long-range order (NCS restraints) improves the accuracy of about 1.0 Å for all sets except for *nmr*. In *w_exact_xray*, *wNCS_nmr_xray* and *wNCS_nmrI*, similar accuracy on the dimer is observed, whereas in *wNCS_nmrII*, the RMSD to 1mu4 displays a threefold increase. The accuracy in all sets, except for *nmr* and *wNCS_nmrII*, compares well to the results obtained on the Crh dimer structure calculated from ssNMR restraints (§ 4.2) for which the accuracy on the monomer is 1.6 Å and 2.9 Å for the dimer . The conformation quality is overall good, the *wNCS_nmrII* conformers displaying the worst parameter values. The accuracy of the ARIA conformers compares well to the accuracy observed for the ssNMR structure [Loquet et al., 2008], except if the conformations are calculated using the most loose restraints. Thus, the sets extracted from *nmr* may correspond to a range of the conformational transition from solution to crystal.

5.3.5 Relative monomer orientation in the dimer

Among the ARIA conformers, the relative orientation of the monomers into the dimer was monitored (Table 5.2) through the calculation of the Euler angles Ψ, Θ and Φ. These angles define the rotations around the axes X, Y and Z (Fig. 5.1c) describing the transformation from one monomer to another. The standard deviations of the angles are larger in *nmr* and *wNCS_nmrII* compared to the other sets calculated in vacuum and in presence of NCS restraints. The largest standard deviation is always observed for Ψ, which corresponds to a largest variability around the X axis (Fig. 5.1c). This is in agreement with the relative orientation of the monomers in the ssNMR structure of Crh 2rlz and in the crystallographic structures 1mo1 and 1mu4 (Table 5.4). In a similar way, the set of 16 tetrameric conformers (ens_XR) obtained from a crystallographic ensemble refinement [Levin et al., 2007] displays Ψ as the most variable angle (Table 5.4), and thus show the same variability features as the ARIA conformers and the ssNMR structure 2rlz. The mean values of Euler angles are more or less constant for *exact_xray*, *nmr_xray*, *w_exact_xray*, *w_nmr_xray* and *wNCS_nmrI* (Table 5.2), and are of the same order of value than in the structures 1mo1, 1mu4, 2rlz and in the set of conformers ens_XR (Table 5.4). The bias observed in *nmr* and *wNCS_nmrII* arises mainly from different relative orientations of the monomers around the axis X (Fig. 5.1c). The lowest-energy conformers of the runs *w_exact_xray* (Fig. 5.6b), *wNCS_nmr_xray* (Fig. 5.6c), and *wNCS_nmrI* (Fig. 5.6d) are close to the structure 1mu4 (Fig. 5.6a), whereas the lowest energy conformer of *wNCS_nmrII* displays clearly (Fig. 5.6e) a difference in the relative orientation of the monomers. Except for *wNCS_nmrII*, the distances between the monomer centres-of-mass are slightly smaller after the refinement step in water than in gas-phase. They also become more different than the one observed in crystallographic structures (1mo1, 1mu4) and in ens_XR.

During the molecular dynamics (MD) simulations, larger conformational drifts, estimated from the coordinate RMSD from the starting point, are observed (Fig. 5.7a, b, c) in the simulation *sol_dimer* than in the two tetramer simulations. In all simulations, the monomer RMSD is smaller than the dimer RMSD. These two features agree well with the variability of monomer relative

	MD simulations			Dimeric structures		
	sol_dimer	*sol_tetra*	*cryst_tetra*	2rlz	1mo1/1mu4	ens_XR
Distances (Å)						
A-B	19.2 ± 0.4	20.7 ± 0.5	20.0 ± 0.1	22.1 ± 0.7	21.3 ± 0.2	20.8 ± 0.1
C-D	-	20.3 ± 0.4	20.1 ± 0.2	-	-	-
Angle (axis) (°)						
Ψ A-B (X)	-77.8 ± 7.1	-80.3 ± 6.1	-79.5 ± 5.6	-67 ± 11.6	-74.5 ± 1.8	-77.0 ± 1.6
Ψ C-D (X)	-	-75.5 ± 4.8	-76.0 ± 5.0	-	-	-
Θ A-B (Y)	15.3 ± 0.8	14.2 ± 1.0	15.8 ± 1.0	12.8 ± 1.1	15.7 ± 0.6	16.0 ± 0.2
Θ C-D (Y)	-	14.8 ± 0.8	15.6 ± 0.9	-	-	-
Φ A-B (Z)	-11.1 ± 1.8	-11.5 ± 1.5	-11.9 ± 1.9	-7.9 ± 2.4	-11.2 ± 1.0	-12.3 ± 0.5
Φ C-D (Z)	-	-10.3 ± 1.3	-11.2 ± 1.5	-	-	-

Table 5.4: Analysis of the relative position of the monomers inside the dimer during the MD simulations (*sol_dimer*, *sol_tetra*, *cryst_tetra*) in the Crh structures (2rlz, Loquet et al. [2008], 1mo1 and 1mu4, Juy et al. [2003]) and in the sets of 16 tetrameric conformers (ens_XR) obtained from the crystallographic ensemble refinement [Levin et al., 2007]. The relative position of the monomers are described through the distance between the monomer centres of mass, and through the angles Ψ, Θ and Φ (see § 5.2.3).

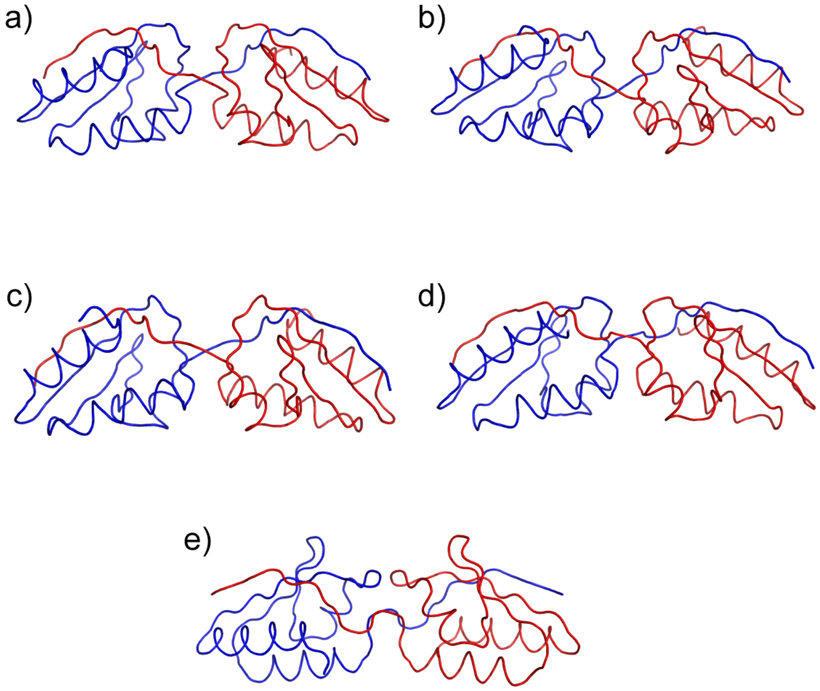


Figure 5.6: Crh dimer structure : 1mu4 (a), lowest-energy conformers from *w_xray_exact* (b), *wNCS_xray_nmr* (c), *wNCS_nmrI* (d) and *wNCS_nmrII* (e). The Crh chains are coloured in blue and red.

orientation, observed among the ARIA conformers, and with the drift from the X-ray crystallographic structure observed in *nmr*.

In MD simulations, the Euler angles stay close to the values observed in crystallographic structures (Table 5.4), but again Ψ displays the largest standard deviations. The Ψ values observed in MD simulations lie between the Ψ values observed for *w_exact_xray*, *wNCS_nmr_xray* and *wNCS_nmrI* from one side, and the Ψ values observed for *wNCS_nmrII* from the other side. This variation is in agreement with the observations made for ARIA conformers and suggests a larger variability of the relative orientation of the monomers around the axis X (Fig. 5.1c). The distance between monomer centres of mass is generally smaller than the value observed in the crystallographic structures. This distance decreases along the whole *sol_dimer* simulation (Fig. 5.7e, f) from 21 Å down to 19.5 Å, whereas it stays constant for simulations *sol_tetra* (Fig. 5.7e) and *cryst_tetra* (Fig. 5.7f).

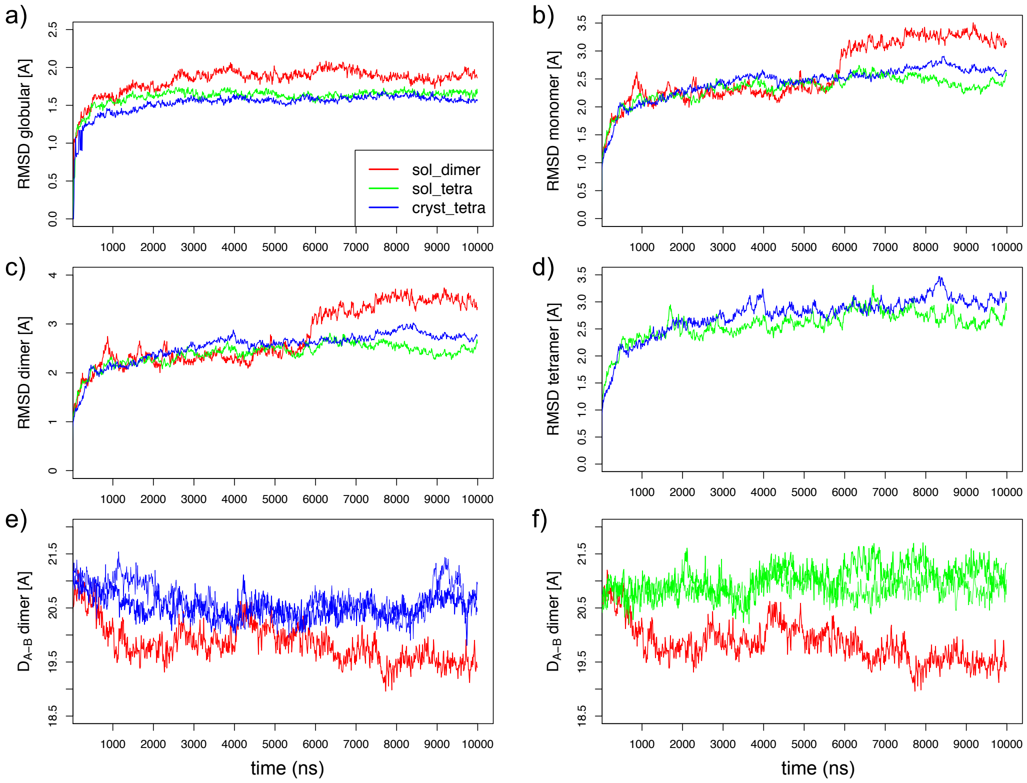


Figure 5.7: Conformational drifts of the MD simulations : (a) region 15-80 of Crh, (b) monomers, (c) dimers, (d) tetramers for the simulations *sol_dimer* (red), sol_tetra (green) and *cryst_tetra* (blue). If several similar regions are present in the simulation, the mean value of RMSD is displayed. (e, f) Variation of the distances between the monomer centres-of-mass. (e) *sol_dimer* (red), *cryst_tetra* (blue lines). (f) *sol_dimer* (red), *sol_tetra* (green lines).

To sump up, the relative orientation of the monomers displays the largest variability for the rotations around the longitudinal axis X of the dimer. This feature is observed for several independently calculated conformations, i.e. the sets of ARIA conformers calculated here, the ssNMR structure 2rlz, or the conformers obtained from a crystallographic ensemble refinement [Levin et al., 2007].

5.3.6 Dimer and tetramer architecture

The Crh structure was analysed by calculating: the distances between secondary structure elements of the Crh dimer in order to have a more local view of the variability and deformations observed in the tertiary structure of the protein (Table 5.3) . The hydrogen bonds in the secondary structure elements (Fig. 5.8) as well as the water bridges (Fig. 5.9) were monitored.

In vacuum, small variations are observed between the sets *exact_xray* and *nmr_xray*, but the distances vary more in *nmr* than among the previous sets and most of them increase. The increase of the β4-β2, β4-α3 and α3-α1 distances is compatible with the disorder observed in the Crh precipitate for these secondary structure elements [Etzkorn et al., 2007]. On the other hand, α2 seems to go apart from β1a and get closer to the monomer core, given that the α2-β1a distance increases and the α2-β3 distance decreases: this is in agreement with the variability in the relative orientation of monomers.

A variation of the intra-monomer distance between α2 and β1a is also observed among the conformers calculated using NCS restraints and water refinement. Larger mean distance and standard deviations are present in *wNCS_nmrII* and in *wNCS_nmrI* and are the sign of the different relative orientations of the monomers. In the MD simulations, the distances (Table 5.3) generally increase from *sol_dimer* up to *sol_tetra* and *cryst_tetra*, but they are smaller than the distances observed in ARIA conformers. This probably comes from the pressure induced on the protein by the water box. The slight decrease of the distance standard deviations from *sol_dimer* down to *sol_tetra* agrees with a more rigid tetrameric structure (Fig. 5.7d). The largest intra-monomer distance expansion ($>$0.5 Å) are observed for α2-β1a and β4-β2.

Contrary to the other distances between secondary structure elements, the inter-monomeric distance β1a-β1a shows a large variability and a tendency to decrease in ARIA conformers generated in gas phase when the restraints become looser (*nmr*) or in the MD simulation *sol_dimer*. However, the distance β1-β4 stays more or less constant in ARIA conformers whatever the level of precision of the restraints. But, the inter-monomer distance β1-β4 is smaller in *sol_dimer* than in the tetramer MD simulations. In a dimeric structure, the β strands involved in inter-monomer interaction thus display the tendency to get closer to each other, whereas the application of restraints describing the long-range crystal order forces these strands to go apart.

The lifetime of an hydrogen bond in the secondary structures was monitored as the percentage of simulation time or of ARIA conformers for which the distance is smaller than 2.2 Å. These lifetimes are color-coded according to their values and displayed on contact maps (Fig. 5.8). The comparison of the contact maps obtained for *wNCS_nmrI* (Fig. 5.8d), *wNCS_nmrII* (Fig. 5.8e) and *wNCS_nmr_xray* (Fig. 5.8f) shows that the number of lifetimes larger than 50 % is more important in *wNCS_nmr_xray* than in *wNCS_nmrI* and *wNCS_nmrII*. The largest decrease is observed inside the helices α1, α2 and between the strands β2 and β4. Within each contact map, the hydrogen bonds in the helices are the least formed in α2, which is not surprising as this helix was shown to be labile in MD simulations of HPr [Canalia et al., 2004]. Similar behaviours are observed for MD simulations: in *sol_dimer* , the helices α2 and α3 are less stable than α1 (Fig.

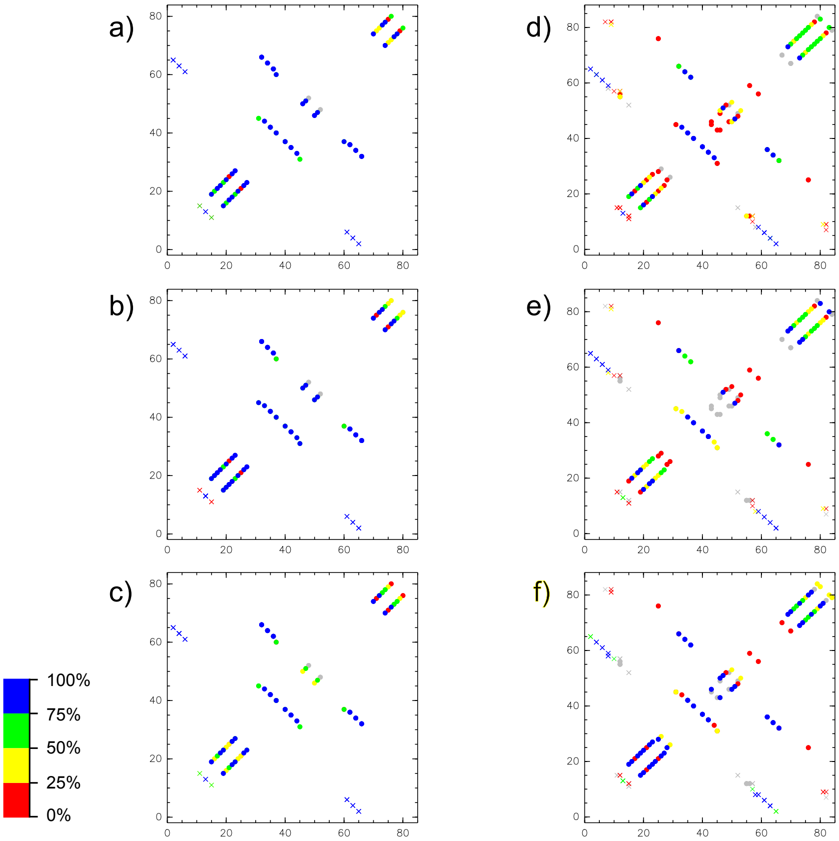


Figure 5.8: Contact maps describing the percentage of hydrogen bond formation among the MD durations or the ARIA conformers. (a) *cryst_tetra* (chains A, B), (b) *sol_tetra* (chains A, B), (c) *sol_dimer*, (d) *wNCS_nmrI*, (e) *wNCS_nmrII*, (f) *wNCS_nmr_xray*. The percentage of formation is described through color-coding: 5-25%: red, 25-50%: yellow, 50-75%: green, 75-100%: blue. Inter-monomer hydrogen bonds are marked with crosses and intra-monomer hydrogen bonds with bullets.

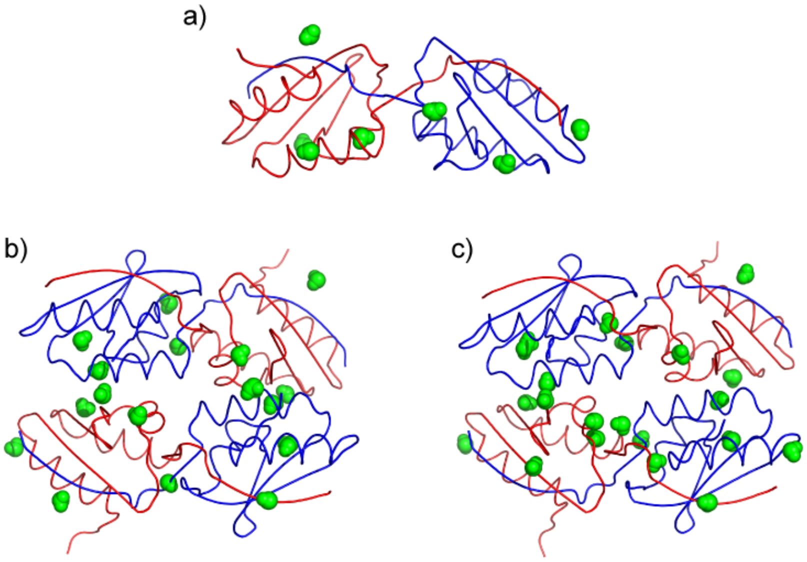


Figure 5.9: Water bridges observed in MD simulations: (a) *sol_dimer*, (b) *sol_tetra*, (c) *cryst_tetra*, The cutoff distance for the detection of hydrogen bonds between water atoms and acceptor/donor groups was 2.5 Å. The Crh chains are coloured in blue and red, and the water molecules are drawn in green CPK. This figure was realised with pymol 0.98 DeLano [2002].

5.8c), but improve their stability in *sol_tetra* (Fig. 5.8b) and *cryst_tetra* (Fig. 5.8a). Otherwise, the mean hydrogen bond lifetime involving residues from the strands β1a is 26% in *wNCS_nmrl*, whereas it is smaller than 20% in other sets of conformers, in particular for *w_exact_xray*. As *w_exact_xray* converges best to the crystallographic structure, the lack of hydrogen bonds is not due to a lack of restraints, but reveals a feature also encountered in crystallographic structures. In a way similar to the observations made on ARIA conformers, the inter-monomer hydrogen bonds between the strands β1a are more stable in *sol_dimer* than in *cryst_tetra*. The values of hydrogen bond lifetimes agree to the variations of the β1a-β1a distance, described above.

The water bridges were detected in MD simulations as water molecules for which at least two atoms display a distance smaller than 2.5 Å to a protein acceptor or donor groups. Seven bridges are present in more than 25 % of the molecular dynamics (MD) simulation *sol_dimer*, but this number is multiplied by 3 in the tetramer simulations, where the size of the solute is multiplied by 2. The water bridges thus appear in presence of a more rigid structure and may induce this rigidity. In *sol_dimer* (Fig. 5.9a), two water bridges are observed at the dimer interface, between O Met-51 (chain A)/H Arg-17 (chain B) and between the strands β4 and $\beta 1$: H Glu-7 (chain A)/Oϵ1 Gln-82 (chain B). During the tetrameric MD simulations (*sol_tetra*, *cryst_tetra*), more water bridges are located (Figs. 5.9b,c) between the dimers and at the monomer interfaces, and inter-monomer bridges appear between: H Ala-54/Oϵ1 Gln-24, O Lys-11/H Gln-15, Leu-21/Hζ3 Lys-40, H Lys-41/O Gln-24, O Met-51/H Ala-16. The appearance of water bridges between monomers and dimers is observed for MD simulations, for which the strands β1a tend to separate from each other. Consequently, these water molecules may be thought to compensate for the induced structure destabilisation.

The inter-molecular crystallographic bridge between Thr-57 and Thr-12, which corresponds to a bridge conserved in the crystallographic structure [Lesage et al., 2006], is not observed in any of the MD simulations. This means that, in *cryst_tetra*, the corresponding crystallographic water molecules are diffusing away from the molecule, and may be the consequence of not modelling the exact environment and packing of the crystal. On the other hand, the disappearance of the Thr-57/Thr-12 bridge is correlated with a larger variability of the monomer orientation than in the crystallographic ensemble refinement, and supports the importance of this bridge for stabilising the dimeric structure, in agreement with the observation of Lesage et al. [Lesage et al., 2006]. Furthermore, the position of this bridge is appropriate to block the rotation around the axis X, which is responsible for the largest part of the variability in monomers orientation inside the dimer.

To summarise, the architecture of the Crh dimer displays variations mainly located in the interface or in secondary structure elements close to the interface, like the α2 helix. The most striking effect is the paradoxical separation of the β1a strands if the conformation gets closer to the X-ray crystallographic structure. This separation is accompanied by the apparition of a larger number of water bridges. Furthermore, the absence in MD simulations of water bridges present in the crystal is in agreement with the variability of the relative orientation of the monomers.

5.4 Discussion

Several structure calculation were performed using ARIA with sets of distance restraints based on the solution-state and solid-state NMR experiments. These structure calculations allowed to sample the Crh conformational landscape in solution and during oligomerisation and crystal nucleation. Molecular dynamics simulations were performed to complement the structural information. In ARIA conformers, a good convergence was observed for monomers. Furthermore, the clustering of solutions and the use of environment restraints (NCS, water), allow to generate sets of conformers exhibiting good precision and accuracy with respect to the crystallographic structure.

The approach proposed here makes full profit of the data recorded in solution state, which are easier accessible than ssNMR data, an on ssNMR chemical shifts which can be nowadays obtained even for proteins of up to 100 amino acids [Pintacuda et al., 2006]. We can thus concentrate on the definition of monomer orientations within a multimer, guided by intermolecular restraints from differentially labelled protein multimers. The information about intermolecular restraints were here partly obtained from the X-ray crystallographic structure, but, as methodology for the measurement of restraints is advancing at a fast pace [Lewandowski et al., 2007], one can expect in the future to rely more on the ssNMR information.

The present study allows an exploration of the protein conformation during the transition from solution to crystal. The short range of NMR distance restraints make them to agree with the local order observed during the first steps of oligomerisation and crystallisation. Then, modelling a long-range order (non-crystallographic symmetries) and the environment (water molecules), allows to improve the similarity to the X-ray crystallographic structure. The inclusion of specific water molecules at the interfaces was already shown to be important in the prediction of complex structures [van Dijk and Bonvin, 2006]. Similarly, the presence of water molecules allows the apparition of water bridges stabilising the inter-molecular interactions. An independent test of the orientation variability of the Crh monomers inside the dimer is provided by the crystallographic conformers calculated by ensemble refinement [Levin et al., 2007]. The comparable orientations variability supports a funnel shape of the conformational space during the transition from solution to crystal.

The exploration of the conformational space, during the solution-crystal transition, can be exploited to detect alternative orientations of the monomers inside the dimer, as in *wNCS_nmrII*. These orientations, transiently populated during the transition, provide a model of the Crh crystallogenesis and explain the importance of specific water molecules in the crystal.

6

A general method for NMR structure calculation of symmetric aggregates

6.1 Introduction

As shown in previous chapters, structure determination of symmetrical protein aggregates by NMR is a delicate task, which entails the development of specific assignment approaches. It is not only hampered by the inherent ambiguity of the NMR data due to signal superposition but it also represents a twofold problem. Actually, on top of solving the structure of the monomeric subunit, the arrangement of monomers in space needs to be determined. For systems exhibiting high-order symmetry, NMR experiments on asymmetrically labelled molecule turn out to be not always as helpful as expected and somehow difficult to interpret. Distinction of inter-molecular from intra-molecular correlations impedes structure determination of symmetric aggregates by NMR (§ 1.6). Computational approaches provide a substantial support for structure elucidation of symmetric assemblies from NMR data. As suggested by the observations made in the previous chapter, it is necessary to consider the symmetric system as a whole, which incidentally increases the number of atoms to be simulated. Also, using symmetry information directly in the calculation allows to reduce the conformational space to explore. Apart from the symmetry-ADR strategy (§ 2.3), existing methods generally consist in docking-like protocols in a symmetry restricted search space (§ 1.6). However, these methods usually require a rigid or semi-rigid structure of the monomeric unit and do not permit to model all types of symmetry.

In this chapter, a new general method for structure calculation of symmetric aggregates from NMR restraints will be presented. Since structure of symmetric assemblies is entirely defined by the structure of the monomeric unit and the symmetry group of the system, this method implements strict symmetry operators to virtually simulate subunits as images of a single, explicitly modelled monomer. To assess the efficiency of the method, we have applied this protocol to three types of symmetric assemblies illustrating the current challenges in both solution and solid-state NMR: the pentameric form of the membrane phospholamban protein studied by solution NMR [Oxenoid and Chou, 2005], amyloid fibrils of the second WW domain of the human CA150 protein [Ferguson et al., 2006] and of the prion domain of the HET-s protein from *Po-*

dospora anserina [Wasmer et al., 2008] and finally the SH3 domain of the chicken α-spectrin [Castellani et al., 2002] analysed by MAS solid-state NMR in microcrystals.

6.2 Material and Methods

6.2.1 Symmetric systems and data sets used

For our calculations, we have used different data sets representing three types of symmetrical protein assemblies studied by NMR: the pentameric form of the phospholamban protein (PLN, Oxenoid and Chou [2005]), amyloid fibrils of the WW2 domain of the human CA150 protein (CA150.WW2, Ferguson et al. [2006]) and of the prion domain (residues 218 to 289) of the HET-s protein from *Podospora anserina* (HET-s, Wasmer et al. [2008]). Additionally, the method was applied to refine a low resolution structure of the microcrystalline form of SH3 domain of the chicken α-spectrin (SH3, Castellani et al. [2002]; Fossi et al. [2005a]). Table 6.1 gives an overview of the symmetrical properties of the calculated multimeric systems. The restraints sets used here are summarised in Table 6.2. For PLN and HET-s calculations, the restraints deposited at the PDB were used. Distance restraints and TALOS predictions for the CA1500.WW2 fibrils were obtained from Johanna Becker (FMP, Berlin). PDSD cross-peaks recorded on SH3 microcrystals were provided by Barth-Jan van Rossum (FMP, Berlin).

Molecular system	PLN	HET-s	CA150.WW2	SH3
Symmetry group	C_5	helicoïdal	linear	P_{212121}
Number of images	4	6	4	107
Unitary symmetry operator	rotation 72°	translation 9.46 Å + rotation 5.7 Å	translation 4.73 Å	-
Number of residues/monomer	52	79	30	55

Table 6.1: Symmetry definition of the calculated assemblies : group of symmetry, number of virtual images of the monomeric molecule (n), unitary symmetry operator (repeated n times for the whole simulated system) and number of residues per monomer.

		PLN	HET-s	CA150.WW2	SH3	
					assigned	SOLARIA
Distance restraints	intra	193	35	-	1024	-
	inter	9	-	-	23	-
	ambiguous	-	103	23	62	1163
Hydrogen bonds	intra	35	7	-	-	-
	inter	-	7	18	-	-
	ambiguous	-	9	-	-	-
Dihedral angles	ϕ,ψ	84	74	42	70	70
	$\chi1$,$\chi2$	32	-	-	-	-
RDC	NH	43	-	-	-	-
	CCα	31	-	-	-	-
	NC	41	-	-	-	-
PDB id		1ZLL	2RNM	2NNT	1U06	
Experimental method		solution NMR	MAS ssNMR	MAS ssNMR	MAS ssNMR, X-ray	

Table 6.2: Number and type of restraints available for the calculated systems. All values reported here are for a single monomer.

6.2.2 Structure calculation with strict symmetry

Strict non-crystallographic symmetry

For systems exhibiting non-crystallographic symmetry (like C_n oligomers or fibrils), structure calculation is performed by a MDSA protocol under NMR restraints with the program CNS [Brünger et al., 1998]. As in the standard protocol presented earlier, the initial structure consisted in a pseudo-extended polypeptidic chain with randomised ϕ, ψ angles. Here, the key concept is that only one monomeric unit is explicitly modelled and the symmetry related partners simulated as virtual images (Fig. 6.1). According to the symmetry group of the system, the corresponding symmetry operators are specified to impose the given *strict* symmetry during the MDSA protocol. All symmetry operators required to reconstruct the entire system are entered with the `NCSstrict` statement in CNS. The expression of the symmetry operators is chosen such that at least one of the intrinsic symmetry axes of the molecule corresponds to one axis of the internal CNS system coordinates. Consequently, there is absolutely no degree of freedom between the monomers and its images, and the perfect symmetry of the system is thus maintained throughout the whole simulation.

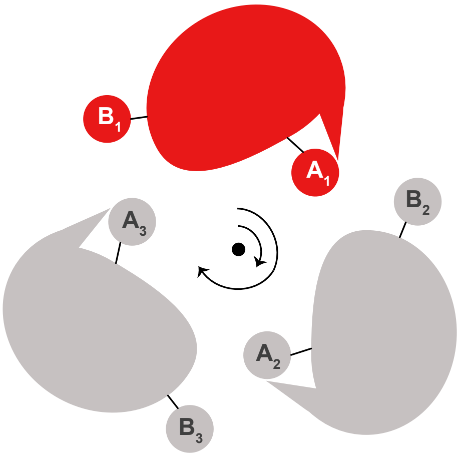


Figure 6.1: Illustration of the strict symmetry method : for a trimer with C_3 symmetry, the symmetry relationships to produce the two images (grey) of the reference monomer (red) consist in two rotations around the symmetry axis (e.g. z-axis) of 120° and 240° (black circle arrows). Atomic coordinates of the images are calculated "on-the-fly" by applying the given rotation operators to the coordinates of the modelled reference monomer.

During MD simulations, the monomeric molecule is free to rotate and translate in all directions. At each energy evaluation, the atomic coordinates of the images are computed "on the fly" by applying the symmetry operators to the coordinates of the modelled monomer (Fig. 6.1). Non-bonded interactions between the monomer and its symmetry related images are included in the hybrid energy function (§ 1.3.1) through the E_{nb_sym} term:

$$E_{hybrid} = E_{chem} + E_{data} + E_{nb\_sym} \tag{6.1}$$

Intermolecular non-bonded interactions are described in the same way as intramolecular ones by a single repulsion term (§ 1.3.1). Distance restraints derived from NOEs are applied with a classical soft-square energy potential (§ 1.3.2) but the corresponding distance computation procedure has been adapted to consider the inter-monomeric contribution between the modelled monomer and the rest of the molecule (see § 6.2.2). Dihedral angle and RDC restraints can also be included in an similar way as in the standard MDSA protocol but need to be defined for one monomer only.

Distance evaluation for ambiguous distance restraints

Standard CNS routines were modified in order to allow the evaluation of the inter-atomic distances between the modelled monomer and its symmetry images. Distances restraints are divided into three classes: *intra-monomeric*, *inter-monomeric* and *ambiguous*. Intra-monomeric restraints are restraints that only involve atoms of one monomer. For inter-monomeric restraints, all inter-atomic distances between the reference monomer and its images are included in the evaluation of the restrained distance. In this formalism, inter-monomeric correlations are implicitly ambivalent in the sense that they are not specific to a particular pair of distinct monomers. Finally, ambiguous restraints are applied according to the ADR description (§ 1.4.2) by including both intra- and inter-monomeric contributions. The effective distance between atoms A and B for an ambiguous restraint (Fig. 6.1) is thus calculated with the following equation:

$$d_{AB} = \left(\sum_{n}^{N} d_{AB_n}^{-6} \right)^{-\frac{1}{6}} \tag{6.2}$$

where n is the monomer number, with a total number of N monomers.

Structure refinement with crystallographic symmetry

In the case of systems defined by *crystallographic symmetry*, unit cell parameters (length and angles) and symmetry operators for a given space group are specified with the CNS `xray` statement. The symmetry operators are expressed with the Jones-Faithfull notation, e.g. (x, y, z) and will be internally applied to the fractionalized coordinates[1] of the modelled monomeric molecule for the in-memory reconstruction of the whole crystal. Since a unit cell is surrounded by 26 other cells in the crystal lattice, the total number of potential neighbouring images is $27n - 1$, where n is the number of molecule per unit cell.

For a given space group, crystallographic symmetry operators are only valid for a particular position and orientation of the main coordinates in the unit cell. In that regard, structure calculation is performed with a different algorithm from the one presented above, and illustrated in Fig. 6.2. In the first stage of the protocol, the position and orientation of the molecule are chosen at random. A rigid body minimisation is then performed under distance restraints. The second stage consists in a cooling stage, constituted of successive short MDs and rigid-body minimisations and where the bath temperature is regularly decreased from 2000 K to 100 K. This protocol is iteratively repeated and the final monomer atomic coordinates of one iteration constitute the starting structure of the next iteration. Non-bonded interactions between the molecule and its images are still modelled with a single repulsion energetic term with a cutoff of 6.5 Å.

---

[1] as fractions of the a, b and c unit cell vectors

Due to the high number of images that the system can be constituted of, inter-molecular distance restraints are evaluated differently whether they involve images belonging to the same or to a different cell than the modelled molecule. In the first situation, distances are computed according to Eq. 6.2, where the sum if performed over all symmetry related molecules in the primary cell. However, distances between molecules in the primary cell and neighbouring cells are determined according to a “minimum image convention”[2]. We have also replaced the flat-bottom distance restraints potential by a log-harmonic potential [Nilges et al., 2008]. This potential has a single minimum, and, in contrast to standard soft-square potential, is more tolerant to large violations (the asymptotic value of the slope is zero). The choice of this form of restraint potential is justified by the fact that incorrect assignments of inter- and intra-molecular correlations may lead to relatively larger distance violations. In that regard, we used a uniform target distance for all restraints instead of individual distance bounds. A basic calibration, that estimates the target distance for which the global energy is minimum, is regularly performed during the course of the calculation.

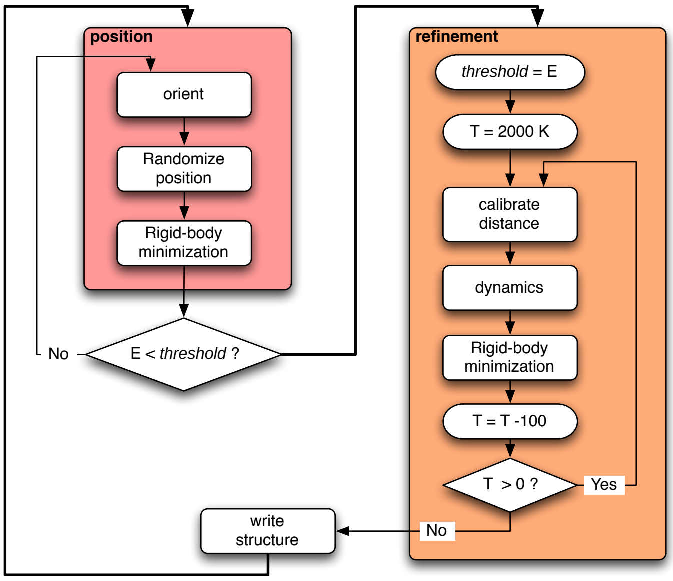


Figure 6.2: Schematic representation of the algorithm for structure refinement with crystallographic symmetry. Coloured boxes outline the two stages of the protocol (positioning and refinement)

6.3 Results and Discussion

6.3.1 Phospholamban homo-pentamer

Phospholamban (PLN) is a symmetric homopentameric membrane protein that regulates the calcium levels between cytoplasm and sarcoplasmic reticulum by interacting with the SERCA protein. The structure of pentameric PLN in DPC (dodecylphosphocholine) micelles has been

---

[2] distance from the closest image

determined by solution NMR (PDB id 1ZLL, Oxenoid and Chou [2005]) and presents C_5 symmetry. Each monomer is constituted of two alpha helices (AP, cytoplasmic and TM, transmembrane) separated by a linker region.

We have calculated an ensemble of 100 conformers with the strict symmetry method using the deposited set of restraints (Table 6.2) and by defining the symmetry operators corresponding to C_5 symmetry. The precision of the 20 lowest energy conformers ensemble (PLN_{sym}) is given in Table 6.2. When considering only the transmembrane domain of the pentamer, the calculation PLN_{sym} converged well and the average structure is similar to the original PDB structure (1ZZL, [Oxenoid and Chou, 2005]), with an RMSD of 1.09 Å. However, in the PLN_{sym} structure, the orientation of the AP helix is slightly different compared to the 1ZZL structure, yielding a larger RMSD between the two structures for the full-length pentamer (Table 6.3 and Fig. 6.3a,b).

	1ZLL		**PLN_{sym}**		**PLN_{dock}**	
	TM	All	TM	All	TM	All
1ZLL	0.80 ± 0.07	1.00 ± 0.11	1.09	5.17	1.53	5.38
PLN_{sym}	-	-	0.97 ± 0.16	1.99 ± 0.42	1.16	2.26
PLN_{dock}	-	-	-	-	0.99 ± 0.08	1.29 ± 0.13

Table 6.3: Precision and similarity of pentameric PLN ensembles for the 1ZLL, PLN_{sym} and PLN_{dock} structures. The values in the diagonal corresponds to the ensemble precision and the non-diagonal elements are the RMSD between average structures. Structures have been superimposed on all heavy atoms of the five chains for residues 2-51 (All) or 23-51 (TM). The same atom sets were used for the atomic RMSD calculation. All values are given in Å.

In the light of this observation, a second ensemble of pentamer conformers was calculated (PLN_{dock}) following the protocol described by Oxenoid and Chou [2005]. Here, the structure of an isolated monomer is first calculated from intra-molecular distance, dihedral angle and RDC restraints with a standard simulated annealing protocol (PLN_{mono}). In a second step, five copies of the lowest energy PLN_{mono} conformer are together refined with a low-temperature simulated annealing protocol in presence of inter-molecular distance restraints (PLN_{dock}). In this docking-like step, backbone atoms of the monomeric units are treated as rigid-bodies. As shown in table 6.3, the PLN_{dock} structures are also very different from the 1ZZL structures (RMSD of 5.4 Å). As for the PLN_{sym} calculation, the structural difference comes from the relative orientation of the AP helix, which is more directed towards the interior of the molecule and more orthogonal to the membrane plane in the 1ZLL structure.

Interestingly, the analysis of the RMS of distance restraints deviations per residue revealed that, in the 1ZLL structure, restraints are less satisfied for residues belonging to the AP helix and to the linker region than in the PLN_{mono}, PLN_{dock} and PLN_{sym} (Fig. 6.4). From our point-of-view, this observation could explain the two different AP helix orientations reported in the 1ZLL and PLN_{dock}/PLN_{sym} structures. Indeed, regardless the general philosophy of the calculation methods, variations in the simulation conditions may still exists (force-fieds, restraints weights) that would account for differences in the resulting structures. Moreover, the pentameric structure calculated in presumed similar conditions (1ZLL, PLN_{dock}) share a common violation profile in the region of the TM helices involved in inter-molecular NOEs (Fig. 6.5) with larger deviations

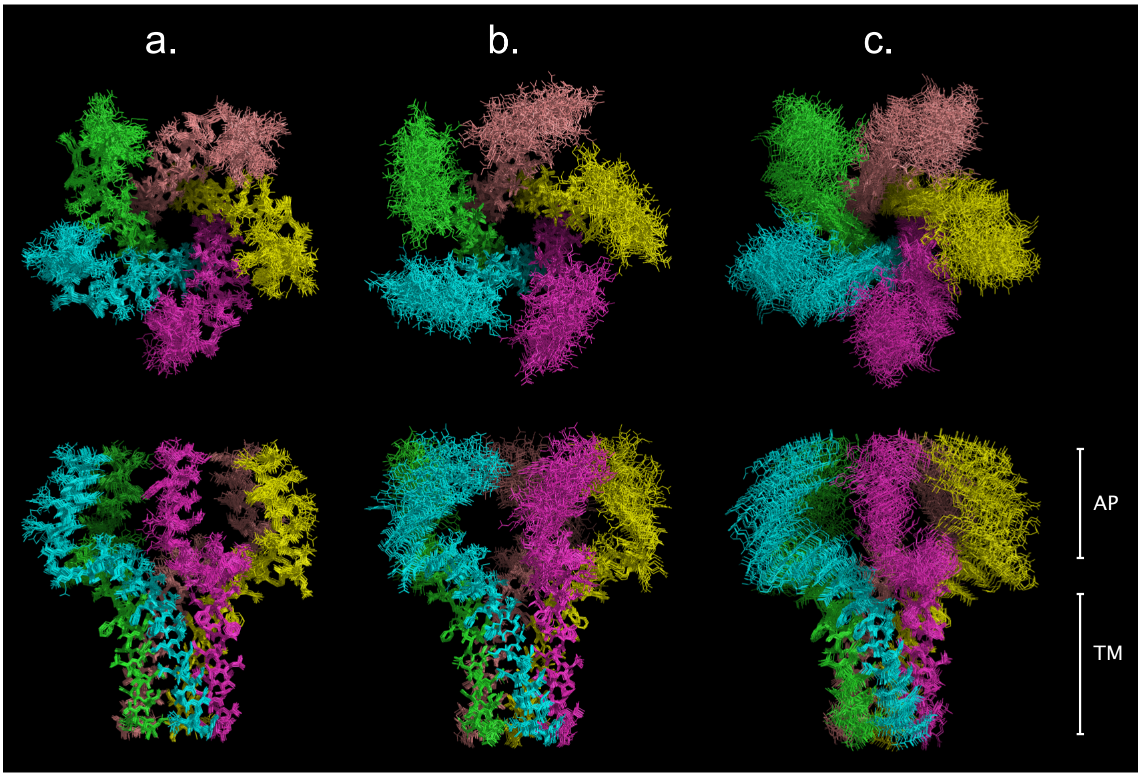


Figure 6.3: Pentameric PLN structure ensembles: 1ZLL (a), PLN$_{sym}$ (b) and 2HYN (c). Each chain is represented by different colour. In the top panel, structures are oriented so that the C_5 symmetry axis is orthogonal to the viewing plane. The bottom panel corresponds to a 90° rotation around the horizontal axis perpendicular to the C_5 symmetry axis. The white marks on the bottom right give the position of the AP and TM helices.

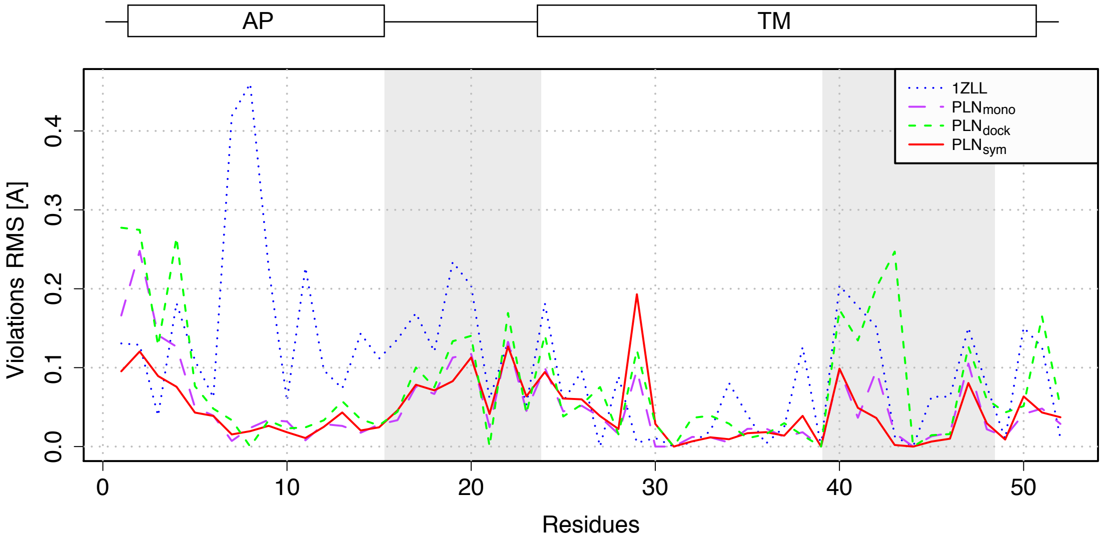


Figure 6.4: RMS of distance restraint violations along the PLN monomer sequence for the 1ZLL (blue-dotted line), PLN$_{mono}$ (solid-yellow), PLN$_{dock}$ (green-dashed) and PLN$_{sym}$ (solid-red) structures. Secondary structures are plotted on top (cytoplasmic (AP) and transmembrane (TM) helices). The first grey zone between the 2 helices represents the linker region and the second one, the region involved in inter-molecular NOEs.

compared to the strict symmetry protocol (PLN$_{sym}$).

NMR relaxation analysis [Oxenoid and Chou, 2005; Oxenoid et al., 2007] on the pentameric phospholamban reveal that the AP helix must be partially mobile relative to the TM helix. The comparison of the atomic RMS fluctuation per residue show that position of the AP helix fluctuates more in the PLN$_{sym}$ ensemble than in 1ZLL. The maximum difference in fluctuation between TM and AP helices is less than 1 Å in 1ZLL whereas it exceeds 4 Å in the structures calculated with the strict symmetry method (PLN$_{sym}$). Potluri et al. [2006] recently proposed a new structure ensemble of the pentameric phospholamban (PDB id 2HYN). It has been determined with a branch-and-bound algorithm that explores the possible arrangements of semi-rigid 1ZLL monomers with optimised van der Waals packing and inter-molecular restraints. As shown in Fig. 6.4, the atomic RMS fluctuation in the twenty 2HYN lowest energy structures are consistently larger than in the PLN$_{sym}$ ensemble. Unfortunately, it is difficult to infer the lability of AP helix position from the 2HYN ensemble since the monomers were considered as rigid bodies. On the other hand, it largely reflects the impossibility of inter-molecular restraints to determine a single possible arrangement of the five monomers.

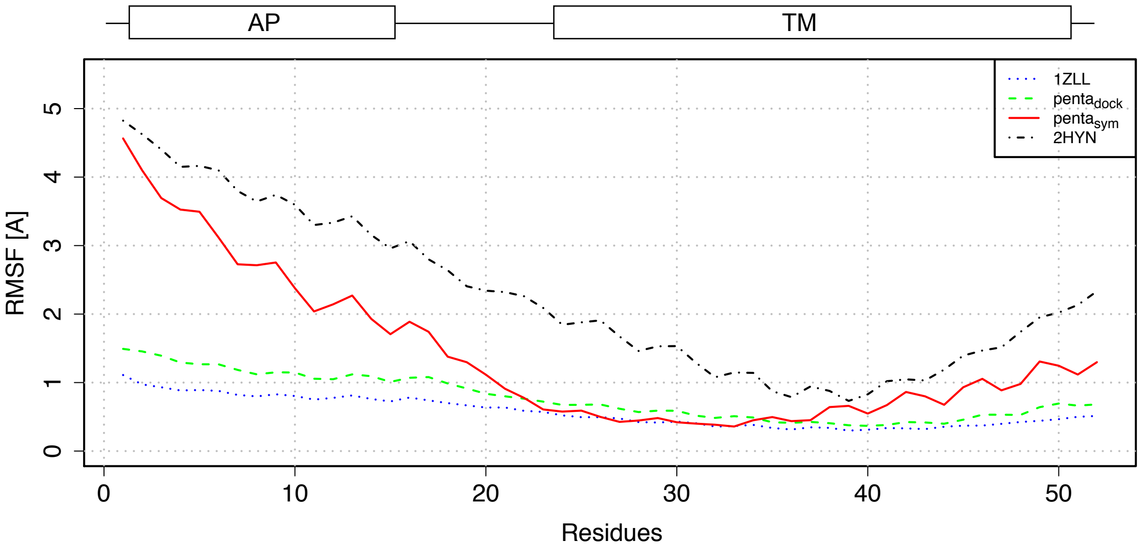


Figure 6.5: RMS fluctuations of the atomic coordinates within a structures ensemble along the PLN monomer sequence for the 1ZLL (blue-dotted line), PLN$_{dock}$ (green-dashed), PLN$_{sym}$ (solid-red) and 2HYN (black dash-dotted) structures. Secondary structures are plotted on top (cytoplasmic (AP) and transmembrane (TM) helices). Structures have been superimposed on the backbone atoms of the five chains of the TM domain (residue 23 to 51). RMSF is calculated as the mean backbone RMSD per residue from the average structure for all conformers in the ensemble.

The question of the orientation and mobility of the AP helix in the pentameric phospholamnban is largely discussed in the literature. Two different models have been proposed for the pentamer PLN : the “bellflower” [Oxenoid and Chou, 2005] (PDB id 1ZLL) and “pinwheel” [Robia et al., 2005] (PDB id 1XNU) models. The latter model suggests that AP helices lay on the reticulum membrane surface. Several studies have reported experimental evidences of contacts between the N-terminal part of AP helix with the membrane [Abu-Baker et al., 2007; Traaseth et al., 2008; Kelly et al., 2008]. Nevertheless, it has been also suggested that the “bellflower” and "pinwheel" structures are not necessarily exclusive alternatives, and may represent sub-states of a dynamic quaternary conformation [Kelly et al., 2008]. Moreover, Nesmelov et al. [2007] have put forward a possible two-state equilibrium for the AP helix position in the monomeric form of phospholamban. In this model, the AP domain may reside in either a R state (not in contact with the membrane) or T state (parallel to the membrane). The T state is predominant and more stable, whereas in the R state, the AP helix moves rapidly with large movements. EPR (Electron Paramagnetic Resonance) experiments also showed that the cytoplasmic domain of PLN is just as dynamic in the pentamer as in the monomer [Traaseth et al., 2007, 2008].

To sum up, the strict symmetry method allowed to recalculate the structure of the phospholamban pentamer without determining the structure of an isolated monomer beforehand. The structure of the transmembrane domain is accurately defined but the position of the cytoplasmic helices remain imprecise while satisfying the NMR restraints. This behaviour supports the highly dynamical and flexible nature of this domain, as suggested in the literature.

6.3.2 Modelling of amyloid fibrils from solid-state NMR data

CA.150 WW2 domain

The second WW domain of the human transcriptional activator CA150, named CA150.WW2, can form amyloid fibrils under physiological conditions. A structural model of these fibrils was constructed based on alanine scanning, electron microscopy and MAS NMR spectroscopy derived constraints (PDB id 2NNT, Ferguson et al. [2006]). In this study, 25 long-range Carbon-Carbon distance constraints, TALOS predicted dihedral angles restraints, as well as inter–molecular hydrogen-bonds restraints were used in a standard simulated-annealing protocol including six copies of a monomeric CA150.WW2 peptide. According to the 2NNT model, the fibril consists of two β-strands separated by a loop incorporated in parallel β-sheets, where hydrogen-bonds are oriented parallel to the fibril axis. The distance restraints were derived from ssNMR spectra that were recorded on an homogeneous sample, where all monomers were identically labelled. To generate the model, distance restraints were however treated as purely intra-monomeric. It was therefore interesting to apply the strict symmetry calculation protocol with the same restraints, but considering them as ambiguous: intra- or inter-monomeric.

An ensemble of 1000 conformers of the CA1500.WW2 fibril (residue 1 to 30) was calculated from the C-C distance restraints described above. These restraints were considered as ambiguous and separated in two classes with different upper bounds (5.5 Å and 7.5 Å, respectively). The neighbouring peptides were simulated by four images (2 upstream and 2 downstream) separated by 4.73 Å, the translation being along the fibril axis. In addition, 42 dihedral angle restraints and 18 inter-molecular hydrogen bond restraints [Ferguson et al., 2006], corresponding to parallel in-register β-sheets, were included in the calculation.

The final pentameric conformer ensemble exhibits an unexpected poor convergence pattern as illustrated in Figure 6.6. A significant number of conformers have a very high total energy (Fig. 6.6a), whereas, in the low energy area (Fig. 6.6b), there is no clear convergence towards

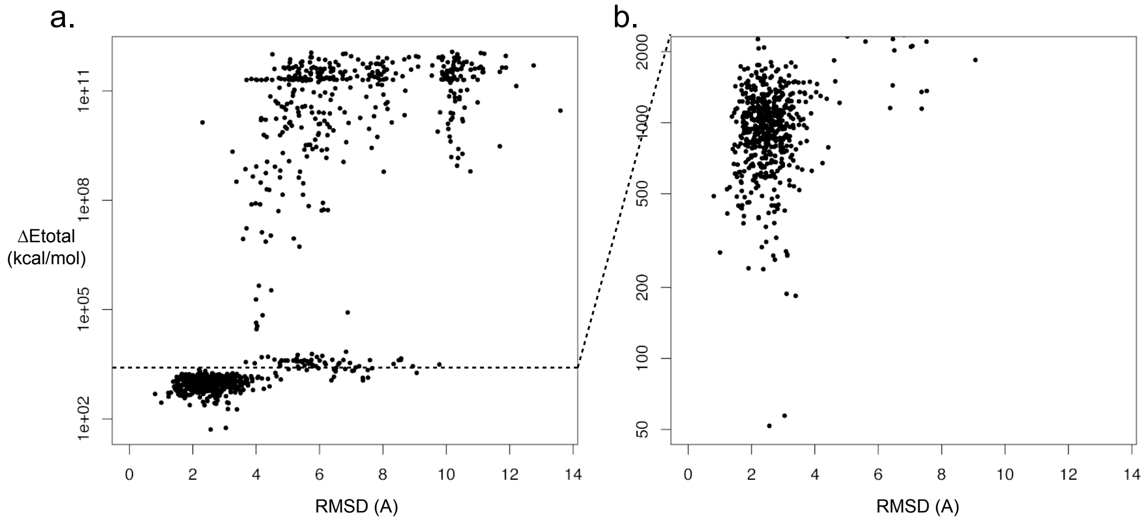


Figure 6.6: Convergence plot of CA150.WW2 conformers calculated with strict symmetry (a) 1000 calculated conformers (y-axis is in log-scale) (b) Magnification of the region where $\Delta E_{\mathrm{total}}$ is below 2000 kcal/mol. The total energy and N,C,Cα atoms RMSD values of each structure are determined from the lowest energy conformer on the five monomers.

a unique fibril conformation. In fact, high energy conformers present unrealistic conformations with severe steric clashes between simulated images. This arises from the fact that probably not enough symmetry related images were used in the calculation, so that the primary image can occupy space that would be occupied by the 3rd or 4th image (only two were calculated for the sake of computational efficiency). To circumvent these limitations, a clustering step was performed on a subset of 542 conformers. The subset comprises conformers with a total energy $<$ 2000 kcal/mol and presenting a MOLPROBITY clashscore [Davis et al., 2004] smaller than 50. Structures with similar backbone conformations were clustered together with a hierarchical clustering approach [Gordon, 1999] from the matrix of all pairwise backbone RMSD values.

Two interesting clusters (named **a** and **b**) containing respectively 42 and 83 conformers were further analysed. Both clusters correspond to relatively well defined ensembles (residue 2 to 28 backbone RMSD of 1.20 Å for **a** and 1.32 Å for **b**) but their global folds differ slightly. In fact, when superimposing backbone atoms of β-strand 1 (residue 2 to 13), one can observe dissimilar arrangements of the strands forming the second β-sheet (Fig. 6.7). In cluster **a**, the two strands are globally in the same plane (orthogonal to the fibril axis), whereas, in cluster **b**, β-stands 2 explore a continuous range of displacement (or *shift*) along the fibrils axis, directed toward the following monomer. In the most shifted position observed in cluster **b**, β-strand 2 faces the inter-strand region, between β-strands 1 of the monomer it belongs to and the following monomer. This observation is in agreement with a recent related study of CA150.WW2 fibrils [Becker, 2008], where analogous calculations provide evidences for two favourable β-sheets arrangement alternatives (in-plane and positive shift), but without excluding a possible heterogeneous architecture throughout the fibril sample.

Nevertheless, the diversity of conformations and β-sheets arrangements observed in our calculation may have several causes. First, this behaviour might illustrate the fact that the position of

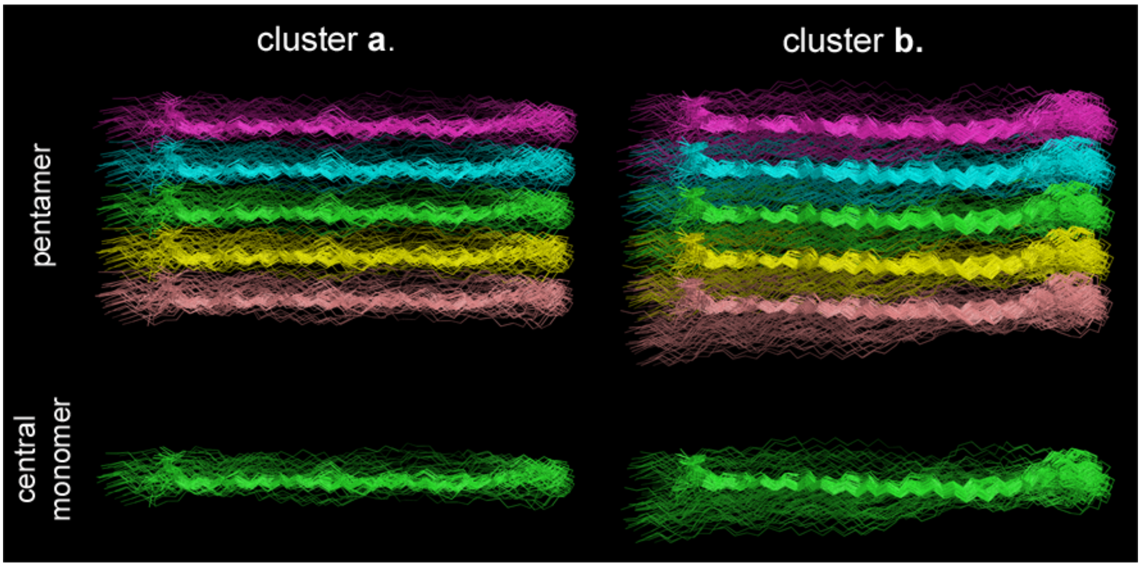


Figure 6.7: Backbone conformation of CA150.WW2 clusters *a* and *b*: Conformers are superimposed on the backbone atoms of β-strand 1(residue 2 to 13). Top panel: pentamer structures where each peptide is represented by a different colour. Bottom panel : central monomer. β-strand 1 is in the foreground.

the second strand is intrinsically not well defined in the fibrils. However, it can also reflect a lack of structural data (less than one distance restraints per residue) and the inability of the method to properly converge for symmetric system like fibrils. Also, it should be noticed that, in the strict symmetry method, there is no rotational degree of freedom between the monomers around the fibril axis. Thus, by prohibiting the twist inherent to β-sheets assemblies [Kajava and Steven, 2006], the strict symmetry calculation method could introduce a bias in the determination of the correct monomeric fold and the overall arrangement of the fibril.

HET-s(218-289) prion

The atomic structure of the HET-s amyloid fibril was determined from MAS ssNMR experiments (PDB id 2RNM, Wasmer et al. [2008]). Fibrils arrange in a left-handed β-solenoid structure. The core of the fibril is defined by three β-strands per winding (six β-strands per molecule) that form continuous in-register parallel β sheets (Fig 6.8**B**). The structure of the fibril was originally calculated for a set of seven monomers, based on numerous distance restraints (Table 6.2) from C-C correlations or C/NHHC experiments on both uniformly labelled and diluted samples. The latter allowed the detection of purely intra-molecular distance and hydrogen bonds restraints. Additionally, dihedral angle restraints predicted by TALOS were included. The symmetry of the system was indirectly enforced by more than 200 distance restraints.

To assess the ability of the strict symmetry method to recalculate the structure of HET-s fibrils, two distinct calculations were performed. For each round, 200 conformers were calculated with the data specified in Table 6.2. Six images of a monomer were simulated and the average helicoïdal twist observed in the 2RNM [Wasmer et al., 2008] structure ensemble was entered through of a rotation of 0.6°/Å around the fibril axis. In the first calculation (HET-s_{ori}),

	2RNM	**HET-s$_{ori}$**	**HET-s$_{ambi}$**
2RNM	0.71 ± 0.22	0.52	1.03
HET-s$_{ori}$	-	0.31 ± 0.06	0.77
HET-s$_{ambi}$	-	-	0.45 ± 0.11

Table 6.4: Precision and similarity of HET-s ensembles for the 2RNM, HET-s$_{ori}$ and HET-s$_{ambi}$ structures. The values in the diagonal correspond to the ensemble precision and the non-diagonal elements are the RMSD between average structures. The twenty lowest energy conformers have been superimposed on the backbone atoms the core region (226-242,262-278) of the five central monomers. The same atom sets were used for the atomic RMSD calculation. All values are given in Å.

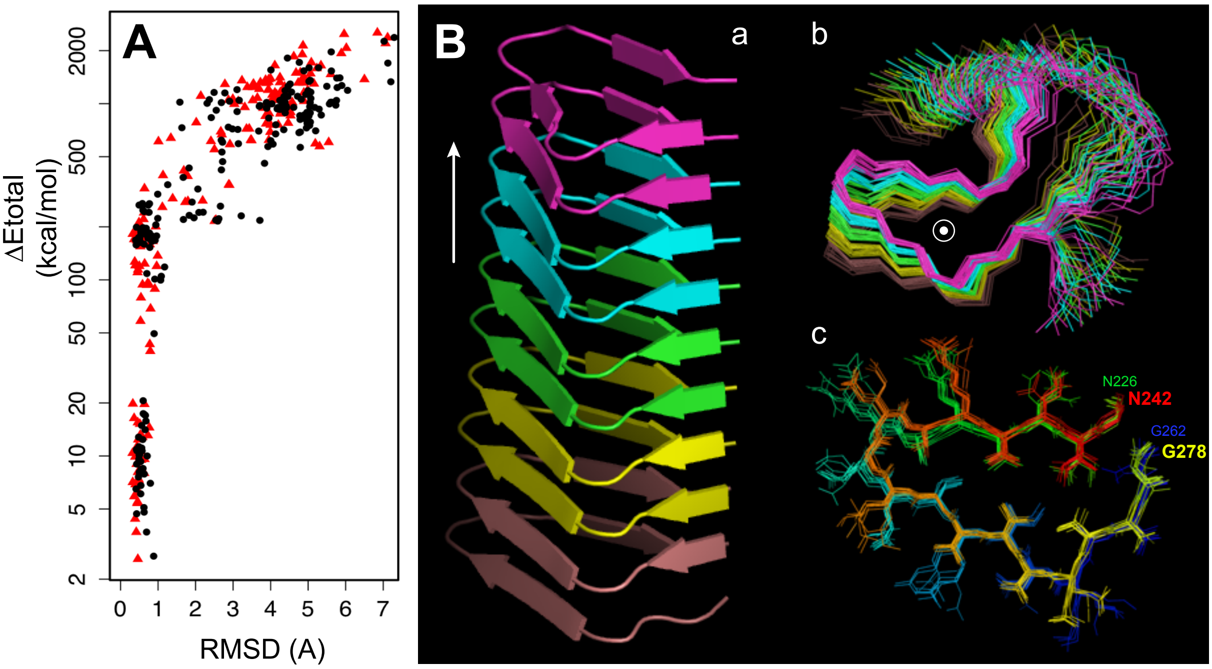


Figure 6.8: Pentameric HET-s structure ensembles: **A** Convergence plot of the HET-s$_{ori}$ (red triangle) and HET-s$_{ambi}$ calculations (black point). Cα RMSD (core region, five central monomers) and ΔE_{total} are referenced from the lowest energy conformer. **B** (a) Side view of the average pentameric HET-s$_{ambi}$ structure (core region) . Each chain is represented by a different colour and the fibril axis is materialised by a white arrow. (b) Top view of the 10 lowest energy structure. (c) Top view of the central monomer core region of the 10 lowest energy structures. The two stacked β strand layers (residue 226 to 242 and residue 262 to 278) are coloured according to the residue numbering.

distance restraints were used with their original level of ambiguity, while in the other calculation (HET-s$_{ambi}$), they were all treated as ambiguous (intra- or inter-molecular), except for intramolecular hydrogen-bond restraints.

As illustrated in Figure 6.8**A** and Table 6.4, both HET-s$_{ori}$ and HET-s$_{ambi}$ calculations precisely converged with a high accuracy with respect to the 2RNM ensemble. The precision of the HET-s$_{ori}$ and HET-s$_{ambi}$ structure ensembles are similar and slightly higher than the deposited 2RNM ensemble. In the structure ensemble determined from ambiguous restraints (HET-s$_{ambi}$, Figure 6.8**B**), the backbone conformation and arrangement of the five central monomers differ moderately from the structure observed in the 2RNM and HET-s$_{ori}$ ensemble. However, this difference corresponds to a maximum RMSD of 1.03 Å which is very satisfactory given the size

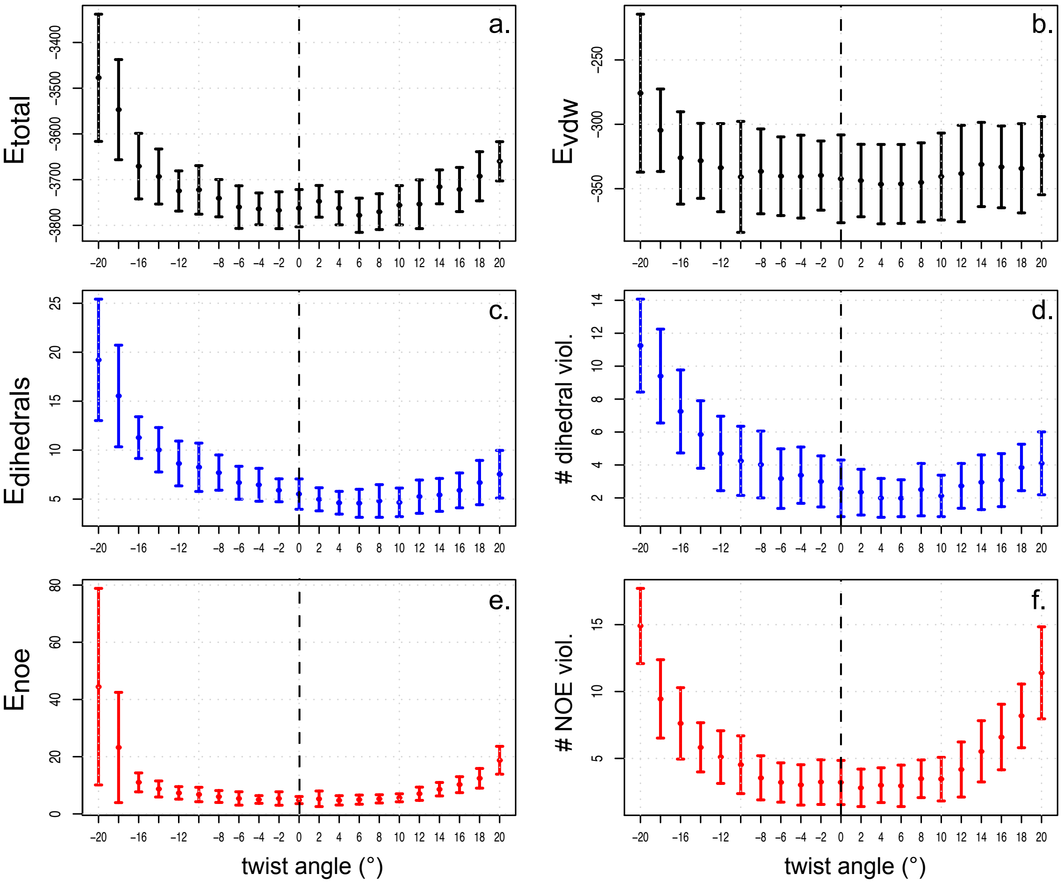


Figure 6.9: Energies and restraints violations dependence on the rotational HET-s twist angle. Total energy (a), van der Waals energy (b), dihedral angles (c) and distance (e) restraints energy, number of dihedral angle (d) and distance restraint violations (f) as a function of the imposed twist angle, from -20˚ to +20˚. The thresholds to consider a restraint as violated are 3˚ for dihedral angles and 0.1 Å for distances. The dotted lines show the position of a zero-degree twist.

of the system and the fact that only few restraints were considered unambiguously as intramolecular (Table 6.4).

As previously stated, the calculation approach used here does not allow for rotations between monomers other than the one imposed in the strict symmetry. Still, by performing the same calculations with different fixed twist angles, we were able to evaluate the contribution of both distance and dihedral angle restraints in the final fibril conformations for a given angle (Figure 6.9). For angles between -6˚ to +8˚, both total and van der Waals energies seem insensitive to the twist angle since no clear minima become apparent. More strikingly, satisfaction of distance restraints also appears independent of the twist for small angles. The only component showing a favourable tendency toward positive angles around +6˚ are dihedral angle restraints. Indirectly, this observation confirm that HET-s fibrils have a left-hand-twist (positive angle), based on available structural data. Nevertheless, from our calculations, it was impossible to infer a precise value for the twist, as observed in the 2RNM ensemble (between +3.4˚ and +8.0˚). In a similar manner of the β-sheets arrangement in CA150.WW2 fibrils, the helicoïdal twist may not be constant all through the HET-s fibril protofilament.

The strict symmetry approach performed well in recalculating the structure of the HET-s prion fibril, even without prior assignment of distance restraints as intra- or inter-molecular. This behaviour contrasts with the results obtained on the CA1500.WW2 fibril, but the amount of available structural data and the rigidity of the two-layered core region of HET-s fibrils certainly help to attain this high level of precision. Moreover, inclusion of a predefined twist angle facilitated the convergence of the calculation.

6.3.3 Refinement of SH3 domain microcrystalline structure with crystal symmetries

We have applied the strict symmetry calculation approach to refine the structure of an approximate conformation of the chicken α-spectrin SH3 domain determined from ambiguous ssNMR distance restraints. The initial atomic positions of the model to be refined correspond to a monomeric structure calculated with a variant of the ARIA protocol dedicated to ssNMR data recorded on site-directed ^{13}C-enriched samples (SOLARIA, Fossi et al. [2005b]), without any particular optimisation regarding the inter- or intra-molecular assignment ambiguity. The initial ARIA peak-lists contained more than 1000 totally unassigned C-C cross-peaks representing intra-molecular correlations or correlations between neighbouring molecules in the crystal lattice. The final monomeric structure presents several distortions when compared to the X-ray structure (PDB id 1U06, Chevelkov et al. [2005]), with a secondary structure backbone RMSD of 3.4 Å. These errors are mainly due to inter-molecular cross-peaks assigned incorrectly by the ARIA methodology.

To first validate the protocol described in section 6.2.2, two series of calculations were carried out with a set of distance restraints derived from the peak-lists described above, but that were correctly assigned with the help of the complete reconstructed crystal (1U06) (Table 6.2). We here assumed that the crystal lattice should be identical in both X-ray and ssNMR samples, with the same space group ($P2_12_12_1$) and cell dimensions. In the first calculation, restraints were applied with their correct assignments, whereas in the second calculation, they were all considered ambiguously (no knowledge of the inter- or intra-molecular state). The two calculations converged very precisely with a remarkable accuracy, the RMSD from the reference X-ray structure being smaller than 0.65 Å (Table 6.5a). More surprisingly, the initial approximate structure is improved towards an highly accurate fold even if information about the inter- or intra-molecular nature of restraints are not considered. As a consequence, this improvement of the SH3 structure can be largely attributed to the inclusion of the crystal lattice configuration in the simulation and the capacity of the method to cope with such ambiguous restraints. Somehow, the incorporation of the log-harmonic restraint potential may also contribute to a faultless interpretation of ambiguous restraints. Nevertheless, in the context of *de novo* structure determination of protein from MAS ssNMR experiments on microcrystals or powder, the restraint list used so far depicts a relatively ideal situation.

To further analyse the efficiency of our method, the same refinement approach was con-

		Precision	**Accuracy**
a) Assigned restraints	correct	0.37 ± 0.07	0.63
	ambiguous (inter/intra)	0.45 ± 0.15	0.59
b) SOLARIA restraints	all	0.99 ± 0.27	2.62
	without noise	0.77 ± 0.18	1.21
	without noise, $\max_n$ = 8	0.64 ± 0.14	0.86

Table 6.5: Precision and accuracy of SH3 conformers refined with different restraint ambiguity conditions Coordinates were superimposed on backbone atoms of residue 8-17 and 23-58 and RMSDs calculated on the same regions. Accuracies were determined from the 1U06 X-ray structure. All values are given in Å.

ducted, but with totally ambiguous restraints, corresponding to the pool of restraints generated at the first SOLARIA iteration. The same approximate structure (3.4 Å accuracy) was used as the starting conformation. Here, both the ambiguity attributable to chemical shifts degeneracy and to the existence of possible inter-molecular correlations are present. In these conditions, accuracy of the initial approximate SH3 model is increased by only 0.8 Å (Table 6.5b), which represents only a slight improvement when compared to the results obtained with the correctly assigned restraints. This lack of success of the refinement protocol might be imputed to the high ambiguity of the restraints and also to the presence of potential "noise" peaks, i.e. peaks for which no plausible assignment could be found in the structure of the complete crystal. Indeed, elimination of "noise" peaks from the restraints allowed to enhance this improvement with a final accuracy of 1.2 Å (Table 6.5b). Moreover, when considering only restraints with less than eight assignment possibilities, accuracy of the refined structure dropped down to 0.86 Å, a value which is very close to what has been observed with ideally assigned restraints. As a matter of fact, by selecting reliable peaks, it was possible to accurately determine the structure of SH3 domain from fully ambiguous restraints, by taking into account the symmetric arrangement of the crystal lattice.

6.4 Conclusion

We have developed a general method to determine the structure of symmetrical protein aggregates from NMR restraints. The method uses molecular dynamics simulated annealing where only one monomeric unit is explicitly modelled to reduce the computational complexity. For a given symmetric assembly, atomic positions of the symmetrically related partners are simulated as virtual images of the modelled monomer using knowledge of the symmetry type.

The method has been applied to systems with cyclic and helicoïdal symmetry but also to crystalline structure. The approach is not limited to these examples and can be extend to model all types of symmetric systems where all monomers are structurally equivalent. The results show that highly accurate models can be produced with this approach. Moreover, the protocol was found to be efficient in unravelling the ambiguity present in NMR data owing to the symmetric nature of the systems. In contrast with existing methods, prior knowledge of the monomeric structure is not required to determine the structure of large oligomeric aggregates, from both solution and solid-state NMR data. As a consequence, the method produces ensembles of structures that may also reflect the flexible and dynamical aspects of the aggregates under investigations. The main disadvantage of our method is the absence of degree of freedom between the subunits potentially necessary for helical symmetry. Nevertheless, it is still conceivable to partially release some degrees during the simulation.

The promising results obtained on amyloid fibrils and microscrystalline protein structure studied by MAS solid-state NMR suggest that the strict symmetry method may be a powerful supplement to experimental techniques in determining the structures of such systems by solid-state NMR. Furthermore, this type of approach can obviously be integrated in an automated iterative assignment protocol, like ARIA, to limit the necessity to resort to complex labelling strategies.

7

General conclusions and Perspectives

7.1 Concluding remarks

NMR is a powerful technique to investigate the structure and dynamics of proteins and protein complexes. The constant progress in solution and solid-state NMR experiments in terms of spectral resolution and labelling patterns opens the door to detailed analyses of large macromolecular structures, and notably, symmetric protein aggregates. In this context, the determination of high resolution 3D structures by NMR spectroscopy critically relies on efficient and reliable automated assignment strategies. Moreover, structural proteomics projects are now coming of age. These projects, which aim at generating accurate three-dimensional models for all folded, globular proteins and domains in the protein universe, often include an NMR component. It also results in new computational challenges for automation, data integration, structure calculation and validation.

In this work, several proficient methods were developed for automated assignment and structure calculation from NMR data in the frame of the ARIA protocol. This methods tackle the majors concerns of structure determination by modern NMR spectroscopy : (i) obtaining and validating structures of high quality, (ii) resolving structures of symmetric protein assemblies and (iii) exploiting solid-state NMR data for high resolution 3D structure elucidation.

First, the inclusion of a network anchoring approach in the ARIA methodology was shown to considerably speed-up the NOE assignment process while preserving a high structural quality of the obtained models. It also manages to improve on the standard ARIA protocol by reducing flaws in local assignment and structure calculation processes from ambiguous data. We have also demonstrated, with a new protocol based on the ICMD methodology, the possibility of a structure calculation entirely performed with a complete general purpose force-field, including the Coulombic and Lennard-Jones interactions. We believe that such an approach can modify the further analysis of structures, as well as their use in molecular modelling studies. In a certain sense, the ICMD approach can be considered as intermediate between structure calculation with simple force-field and refinement in a shell of water molecules. Compared to the latter two methods, the precision of the ICMD structure ensembles are similar and the percentage of residues in the core Ramachandran diagram as well as the WHAT IF Z-scores for ICMD structures are most often slightly superior. On top of that, our observations on local quality parameters confirmed the idea that this kind of analysis are essential to detect possible sources of error in the spectral assignment [Nabuurs et al., 2006]. In this view, graphical analysis tools, such as WHAT IF quality profiles, were integrated within the widely used ARIA program.

In a second stage, we have considered the delicate problem of *de novo* structural determination of symmetric protein aggregates from NMR data. For the simplest and most recurrent case, i.e. symmetric homo–dimers, a specific approach was developed, based on symmetry ambiguous distance restraints and integrated in the ARIA protocol. On a set of three dimers representing different types of interface, this method was shown to perform well when the ambiguity of the restraints is limited to an ambiguity of chain. Interestingly, for some homo–dimeric folds, network anchoring has been shown essential for unravelling both chain and proton assignment ambiguities and for calculating high quality dimeric structures from completely unassigned NOE cross-peaks. Furthermore, a new general method, based on strict symmetry definitions, was conceived for structure calculation of any type of symmetric assemblies from both solution and solid-state NMR data. The successful application of this method to the calculation of a pentameric membrane protein structure and of amyloid fibril protofilaments from ambiguous ssNMR distance restraints outlines the possibilities of this approach for structural investigations of large symmetric systems by NMR.

In addition, an analysis of the conformational landscape of the Crh protein during oligomerisation and crystallisation was conducted with a simultaneous use of solution-state, solid-state NMR and X-ray crystallography data. In summary, the Crh structure displays two features typical for the transfer of a biomolecule from solution to solid state: on the one hand a conformational change between solution and solid-state structure, and on the other hand extended intermolecular interactions. We explored here an approach based on a combination of monomer solution NMR restraints, easier to record, a minimal set of ssNMR restraints, and information from the X-ray crystallographic structure. The generation of NMR and crystallographic conformers, as well as molecular dynamics simulations allowed us to describe the impact of the long-range solid-state order on the convergence. The variability of monomer orientation is concentrated in rotations around the dimer longitudinal axis, and this is just amplified but not created by the use of distance restraints, as this feature is observed also in a set of crystallographic conformers obtained from ensemble refinement along the structure factors. In this regard, a combined analysis by solution and solid-state NMR seems particularly well-suited to an investigation of structural polymorphism of protein oligomers. A recent study on αb-crystalline oligomers with solid and solution state NMR [Jehle et al., 2008] also supports this idea.

Finally, we have addressed the problem of high resolution structure determination from MAS solid-state NMR, which is often hampered by two crucial complications : the high level of ambiguity present in the spectra and the detection of inter-molecular correlations. These limitations generally require preparation of complex labelling strategies. By adapting the ARIA methodology to the specific case of proton-mediated, rare-spin detected solid-state NMR spectra recorded on uniformly labelled sample, we were able to determine the structure of the dimeric Crh protein without manual cross-peak assignment. This result is a further step towards the structure determination of insoluble proteins by the more general approach using fully labelled protein samples, and paves the way for the study of larger molecules. This automated assignment and structure calculation strategy is complementary to the SOLARIA approach [Fossi et al., 2005a]. Moreover, in the context of our strict symmetry approach, a specific protocol was designed to refine the structure of micro-crystalline proteins incorrectly resolved from MAS solid-NMR spectra, due to the presence of inter-molecular correlations. By taking into account the overall configuration of the crystal lattice in the calculation, it was possible to drastically increase the accuracy of the SH3 domain structure from a subset of reliable unassigned cross-peaks. This method relies on the knowledge of the crystal lattice parameters (cell dimensions and symmetry space group). Recent progresses in powder diffraction data analysis [Margiolaki et al., 2007; Margiolaki and Wright, 2008] reveal that such information could be determined on protein microcrystalline samples as well, without first resolving the atomic structure.

7.2 Perspectives

The determination of high-resolution 3D structures of protein by solid-state NMR is to some extent still complicated by the presence of large spectral overlaps. Clearly, the fast convergence quality of the network anchoring and its ability to cope with highly ambiguous restraints would be of great use to obtain even higher resolution structures from solid-state NMR data. Introduction of local backbone relative angular restraints [Franks et al., 2008] in the ARIA calculation may also improve convergence for protein regions lacking long-range structural information. Obviously, a natural continuation of this work will be the distribution to the NMR community of a solid-state NMR dedicated variant of the software ARIA2 for automated assignment of 2D/3D proton *and* carbon spectra, and that would handle specific labelling schemes.

Despite the encouraging results of the strict symmetry approach, some obstacles remain. The absence of degree of freedom between the monomers may be circumvented with an hybrid protocol, mixing symmetry-ADR and strict symmetry strategies. For the time being, it is not possible to use the strict symmetry approach to directly assign ambiguous distance restraints, but a future implementation within the iterative ARIA procedure would lead to a powerful tool to detect inter-molecular correlations and calculate the structure of high-order symmetric assemblies from ambiguous NMR spectra.

One of the greatest challenges in modern structural biology is the characterisation of the topology of multi-domain macromolecular complexes that govern a major part of important cellular functions. As demonstrated by a recent approach which allowed to determine the molecular architecture of the yeast's nuclear pore complex [Alber et al., 2007, 2008], this can be achieved by integrating data from diverse experimental methods. In combination with other structural methods, NMR spectroscopy could have a key role in accessing structures of large macromolecular complexes. Such a structural determination approach requires the development of new integrated computational strategies that can use heterogeneous spatial information from various experimental methods, and the work presented here falls within this scope. In this regard, medium or low-resolution data from electron microscopy (EM) emerge as an ideal complement of solid-state NMR to explore supramolecular structures, like association of amyloid fibrils protofilaments [Luca et al., 2007; Paravastu et al., 2008; Vilar et al., 2008]. High-resolution cryoelectron microscopy provides intermediate resolution density map that can be used to refine structure of macromolecular assemblies [Topf et al., 2008]. It is conceivable to incorporate shape information from EM in the NMR structure determination process of protein assemblies, in similar fashion to small angle scattering data [Mareuil et al., 2007; Gabel et al., 2008; Förster et al., 2008].

7.3 Publications

The work described here has so far resulted in four publications:

Bardiaux et al. [2008b] describes the analysis of the network-anchoring and symmetry-ADR approaches for structure calculation of symmetric homo-dimers with the ARIA protocol.

Bardiaux et al. [2008a] details the graphical tools for structure analysis implemented in ARIA.

Loquet et al. [2008] presents the structure determination of the dimeric *Crh* protein by solid-state NMR. In this publication, both the experimental work done in Lyon and the application of the ARIA methodology are described.

Bardiaux et al. [2006] reports the application of the ICMD procedure for NMR structure calculation and the comparison with the RECOORD structures.

REFERENCES

Abagyan, R. and Mazur, A. (1989). New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles. *J Biomol Struct Dyn.*, 6:833–845.

Abu-Baker, S., Lu, J.-X., Chu, S., Shetty, K. K., Gor'kov, P. L., and Lorigan, G. A. (2007). The structural topology of wild-type phospholamban in oriented lipid bilayers using 15N solid-state NMR spectroscopy. *Protein Sci*, 16(11):2345–9.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2007). Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–94.

Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem*, 77:443–77.

Altieri, A. and Byrd, R. (2004). Automation of NMR structure determination of proteins. *Curr Opin Struct Biol*, 14(5):547–53.

Andre, I., Bradley, P., Wang, C., and Baker, D. (2007). Prediction of the structure of symmetrical protein assemblies. *P Natl Acad Sci Usa*, 104(45):17656–61.

Andronesi, O., Becker, S., Seidel, K., Heise, H., Young, H., and Baldus, M. (2005). Determination of membrane protein structure and dynamics by magic-angle-spinning solid-state NMR spectroscopy. *J Am Chem Soc*, 127:12965–12974.

Aramini, J., Huang, Y., Cort, J., Goldsmith-Fischman, S., Xiao, R., Shih, L., Ho, C., Liu, J., Rost, B., Honig, B., Kennedy, M., Acton, T., and Montelione, G. (2003). Solution NMR structure of the 30S ribosomal protein S28E from Pyrococcus horikoshii. *Protein Sci.*, 12:2823–2830.

Azuaga, A., Neira, J., and van Nuland, N. (2005). HPr as a model protein in structure, interaction, folding and stability studies. *Protein Pept Lett.*, 12:123–137.

Baran, M., Huang, Y., Moseley, H. B., and Montelione, G. (2004). Automated analysis of protein NMR assignments and structures. *Chem Rev*, 104(8):3541–56.

Bardiaux, B., Bernard, A., Rieping, W., Habeck, M., Malliavin, T., and Nilges, M. (2008a). Graphical analysis of NMR structural quality and interactive contact map of NOE assignments in ARIA. *BMC Struct Biol*, 8(1):30.

Bardiaux, B., Bernard, A., Rieping, W., Habeck, M., Malliavin, T. E., and Nilges, M. (2008b). Influence of different assignment conditions on the determination of symmetric homodimeric structures with ARIA. *Proteins*, in press, doi:10.1002/prot.22268.

Bardiaux, B., Malliavin, T., Nilges, M., and Mazur, A. (2006). Comparison of different torsion angle approaches for NMR structure determination. *J. Biomol. NMR*, 34:153–166.

Becker, J. (2008). *Structural investigation of CA150.WW2 amyloid fibrils by MAS NMR spectroscopy*. PhD thesis, Freie Universität Berlin.

Berchanski, A., Segal, D., and Eisenstein, M. (2005). Modeling oligomers with Cn or Dn symmetry: application to CAPRI target 10. *Proteins*, 60(2):202–6.

Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A., and Haak, J. (1984). Molecular dynamics with coupling to an external bath. *J Chem Phys*, 81:3684–3690.

Bewley, C. and Clore, G. M. (2000). Determination of the relative orientation of the two halves of the domain-swapped dimer of cyanovirin-N in solution using dipolar couplings and rigid body minimization. *J Am Chem Soc*, 122(25):6009–6016.

Blundell, T. L. and Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multi-component protein systems. *Proc Natl Acad Sci USA*, 93(25):14243–8.

Böckmann, A., Lange, A., Galinier, A., Luca, S., Giraud, N., Juy, M., Heise, H., Montserret, R., Penin, F., and Baldus, M. (2003). Solid state NMR sequential resonance assignments and conformational analysis of the 2x10.4 kDa dimeric form of the Bacillus subtilis protein Crh. *J Biomol NMR*, 27(4):323–39.

Boelens, R., Koning, T., Vandermarel, G., VanBoom, J., and Kaptein, R. (1989). Iterative procedure for structure determination from proton proton NOEs using a full relaxation matrix approach : Application to a DNA octamer. *J Magn Reson*, 82:290–308.

Bousset, L., Redeker, V., Decottignies, P., Dubois, S., Maréchal, P. L., and Melki, R. (2004). Structural characterization of the fibrillar form of the yeast Saccharomyces cerevisiae prion Ure2p. *Biochemistry*, 43:5022–5032.

Braun, W. and Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J Mol Biol.*, 186:611–626.

Breeze, A. (2000). Isotope-filtered NMR methods for the study of biomolecular structure and interactions. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 36(4):323–372.

Broadhurst, R., Nietlispach, D., Wheatcroft, M., Leadlay, P., and Weissman, K. (2003). The structure of docking domains in modular polyketide synthases. *Chem Biol*, 10(8):723–731.

Brunger, A. (2007). Version 1.2 of the Crystallography and NMR system. *Nat Protoc*, 2:2728–2733.

Brünger, A., Adams, P., Clore, G., DeLano, W., Gros, P., Grosse-Kunstleve, R., Jiang, J., Kuszewski, J., Nilges, M., Pannu, N., Read, R., Rice, L., Simonson, T., and Warren, G. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, 54(Pt 5):905–921.

Canalia, M., Malliavin, T., Kremer, W., and Kalbitzer, H. (2004). Molecular dynamics simulations of HPr under hydrostatic pressure. *Biopolymers*, 74:377–388.

Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., Duke, R., Merz, K., Wang, B., Pearlman, D., Crowley, M., Brozell, S., Tsui, V., Gohlke, H., Mongan, J., Hornak, V., Cui, G., Beroza, P., Schafmeister, Cand Caldwell, J., Ross, W., and Kollman, P. (2004). *AMBER 9*.

Castellani, F., van Rossum, B., Diehl, A., Rehbein, K., and Oschkinat, H. (2003). Determination of solid-state NMR structures of proteins by means of three-dimensional 15N-13C-13C dipolar correlation spectroscopy and chemical shift analysis. *Biochemistry*, 42(39):11476–83.

Castellani, F., van Rossum, B., Diehl, A., Schubert, M., Rehbein, K., and Oschkinat, H. (2002). Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature*, 420(6911):98–102.

Cavalli, A., Salvatella, X., Dobson, C. M., and Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. *P Natl Acad Sci Usa*, 104(23):9615–20.

Cecchini, M., Curcio, R., Pappalardo, M., Melki, R., and Caflisch, A. (2006). A molecular dynamics approach to the structural characterization of amyloid aggregation. *J Mol Biol*, 357:1306–1321.

Chao, J., Lee, J., Chapados, B., Debler, E., Schneemann, A., and Williamson, J. (2005). Dual modes of RNA-silencing suppression by Flock House virus protein B2. *Nat Struct Mol Biol*, 12(11):952–957.

Chao, J. and Williamson, J. (2004). Joint X-ray and NMR refinement of the yeast L30e-mRNA complex. *Structure*, 12:1165–1176.

Chen, Y., Bycroft, M., and Wong, K. (2003). Crystal structure of ribosomal protein L30e from the extreme thermophile Thermococcus celer: thermal stability and RNA binding. *Biochemistry*, 42:2857–2865.

Chevelkov, V., Faelber, K., Diehl, A., Heinemann, U., Oschkinat, H., and Reif, B. (2005). Detection of dynamic water molecules in a microcrystalline sample of the SH3 domain of alpha-spectrin by MAS solid-state NMR. *J Biomol NMR*, 31(4):295–310.

Clore, G., Gronenborn, A., and Bax, A. (1998). A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J Magn Reson*, 133(1):216–21.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*, 13(3):289–302.

Cornilescu, G., Marquardt, J., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc*, 120(27):6836–6837.

Dalgarno, D., Levine, B., and Williams, R. (1983). Structural information from NMR secondary chemical shifts of peptide alpha C-H protons in proteins. *Biosci Rep*, 3(5):443–52.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *J Chem Phys*, 98:10089–10092.

Davis, I., Murray, L., Richardson, J., and Richardson, D. (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res*, 32(Web Server issue):W615–9.

DeLano, W. (2002). The PyMOL Molecular Graphics System. *http://www.pymol.org*.

Dominguez, C., Boelens, R., and Bonvin, A. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125:1731–1737.

Doreleijers, J., Raves, M., Rullmann, T., and Kaptein, R. (1999). Completeness of NOEs in protein structures: A statistical analysis of NMR data. *J Biomol NMR.*, 14:123–132.

Duggan, B., Legge, G., Dyson, H., and Wright, P. (2001). SANE (Structure Assisted NOE Evaluation): an automated model-based approach for NOE assignment. *J Biomol NMR*, 19(4):321–329.

Etzkorn, M., Böckmann, A., Lange, A., and Baldus, M. (2004). Probing molecular interfaces using 2D magic-angle-spinning NMR on protein mixtures with different uniform labeling. *J Am Chem Soc*, 126(45):14746–51.

Etzkorn, M., Böckmann, A., Penin, F., Riedel, D., and Baldus, M. (2007). Characterization of folding intermediates of a domain-swapped protein by solid-state NMR spectroscopy. *J Am Chem Soc*, 129:169–175.

Favier, A., Brutscher, B., Blackledge, M., Galinier, A., Deutscher, J., Penin, F., and Marion, D. (2002). Solution structure and dynamics of Crh, the Bacillus subtilis catabolite repression HPr. *J Mol Biol*, 317(1):131–44.

Fay, N., Redeker, V., Savistchenko, J., Dubois, S., Bousset, L., and Melki, R. (2005). Structure of the prion Ure2p in protein fibrils assembled in vitro. *J Biol Chem*, 280:37149–37158.

Ferguson, N., Becker, J., Tidow, H., Tremmel, S., Sharpe, T. D., Krause, G., Flinders, J., Petrovich, M., Berriman, J., Oschkinat, H., and Fersht, A. R. (2006). General structural motifs of amyloid protofilaments. *Proc Natl Acad Sci USA*, 103(44):16248–53.

Fogh, R., Boucher, W., Vranken, W., Pajon, A., Stevens, T., Bhat, T., Westbrook, J., Ionides, J., and Laue, E. D. (2005). A framework for scientific data modeling and automated software development. *Bioinformatics*, 21(8):1678–84.

Fogh, R., Vranken, W., Boucher, W., Stevens, T., and Laue, E. (2006). A nomenclature and data model to describe NMR experiments. *J Biomol NMR*, 36(3):147–155.

Folkers, P., Folmer, R., Konings, R., and Hilbers, C. (1993). Overcoming the ambiguity problem encountered in the analysis of Nuclear Overhauser Magnetic-Resonance spectra of symmetrical dimer proteins. *J Am Chem Soc*, 115:3798–3799.

Folmer, R., Hilbers, C., Konings, R., and Nilges, M. (1997). Floating stereospecific assignment revisited: Application to an 18 kDa protein and comparison with J-coupling data. *J Biomol NMR.*, 9:245–258.

Förster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A., and Sali, A. (2008). Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. *J Mol Biol*, 382(4):1089–106.

Fossi, M., Castellani, F., Nilges, M., Oschkinat, H., and van Rossum, B. (2005a). SOLARIA: a protocol for automated cross-peak assignment and structure calculation for solid-state magic-angle spinning NMR spectroscopy. *Angew Chem Int Ed Engl*, 44(38):6151–4.

Fossi, M., Linge, J., Labudde, D., Leitner, D., Nilges, M., and Oschkinat, H. (2005b). Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning NOE data. *J Biomol NMR*, 31(1):21–34.

Fossi, M., Oschkinat, F., Nilges, M., and Ball, L. (2005c). Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data. *J Magn Reson*, 175(1):92–102.

Foster, M. P., McElroy, C. A., and Amero, C. D. (2007). Solution NMR of large molecules and assemblies. *Biochemistry*, 46(2):331–40.

Franks, W. T., Wylie, B. J., Schmidt, H. L. F., Nieuwkoop, A. J., Mayrhofer, R.-M., Shah, G. J., Graesser, D. T., and Rienstra, C. M. (2008). Dipole tensor-based atomic-resolution structure determination of a nanocrystalline protein by solid-state NMR. *P Natl Acad Sci Usa*, 105(12):4621–6.

Gabel, F., Simon, B., Nilges, M., Petoukhov, M., Svergun, D., and Sattler, M. (2008). A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints. *J Biomol NMR*, 41(4):199–208.

Gaponenko, V., Altieri, A. S., Li, J., and Byrd, R. A. (2002). Breaking symmetry in the structure determination of (large) symmetric protein dimers. *J Biomol NMR*, 24(2):143–8.

Gardiennet, C., Loquet, A., Etzkorn, M., Heise, H., Baldus, M., and Böckmann, A. (2008). Structural constraints for the Crh protein from solid-state NMR experiments. *J Biomol NMR*, 40:239–250.

Goodsell, D. and Olson, A. (2000). Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure*, 29:105–53.

Gordon, A. (1999). *Classification*. London Chapman and Hall / CRC.

Grishaev, A. and Llinás, M. (2005). Protein structure elucidation from minimal NMR data: the CLOUDS approach. *Meth Enzymol*, 394:261–95.

Grishaev, A., Steren, C., Wu, B., Pineda-Lucena, A., Arrowsmith, C., and Llináss, M. (2005). ABACUS, a direct method for protein NMR structure computation via assembly of fragments. *Proteins*, 61(1):36–43.

Güntert, P. (2003). Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43(3-4):105–125.

Güntert, P. (2004). Automated NMR structure calculation with CYANA. *Methods Mol Biol*, 278:353–378.

Güntert, P., Mumenthaler, C., and Wütrich, K. (1997). Torsion Angle Dynamics for NMR Strucutre Calculation with the New Program DYANA. *J. Mol. Biol.*, 273:283–298.

Habeck, M., Nilges, M., and Rieping, W. (2005). Replica-exchange Monte Carlo scheme for bayesian data analysis. *Phys Rev Lett*, 94(1):018105.

Habeck, M., Nilges, M., and Rieping, W. (2008). A unifying probabilistic framework for analyzing residual dipolar couplings. *J Biomol NMR*, 40(2):135–44.

Hahmann, M., Maurer, T., Lorenz, M., Hengstenberg, W., Glaser, S., and Kalbitzer, H. (1998). Structural studies of histidine-containing phosphocarrier protein from Enterococcus faecalis. *Eur J Biochem*, 252:51–58.

Heise, H. (2008). Solid-state NMR spectroscopy of amyloid proteins. *Chembiochem*, 9(2):179–89.

Heise, H., Seidel, K., Etzkorn, M., Becker, S., and Baldus, M. (2005). 3D NMR spectroscopy for resonance assignment and structure elucidation of proteins under MAS: novel pulse schemes and sensitivity considerations. *J Magn Reson*, 173(1):64–74.

Helgstrand, M., Rak, A., Allard, P., Davydova, N., Garber, M., and Hard, T. (1999). Solution structure of the ribosomal protein S19 from Thermus thermophilus. *J Mol Biol.*, 292:1071–1081.

Helmus, J., Surewicz, K., Nadaud, P., Surewicz, W., and Jaroniec, C. (2008). Molecular conformation and dynamics of the Y145Stop variant of human prion protein in amyloid fibrils. *Proc Natl Acad Sci U S A*, 105:6284–6289.

Herrmann, T., Güntert, P., and Wüthrich, K. (2002). 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):209–227.

Herve du Penhoat, C., Atreya, H., Shen, Y., Liu, G., Acton, T., Xiao, R., Li, Z., Murray, D., Montelione, G., and Szyperski, T. (2004). The NMR solution structure of the 30S ribosomal protein S27e encoded in gene RS27_ARCFU of Archaeoglobus fulgidis reveals a novel protein fold. *Prot Sci.*, 13:1407–1416.

Hooft, R., Vriend, G., Sander, C., and Abola, E. (1996). Errors in protein structures. *Nature*, 381(6580):272.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–725.

Huang, Y., Tejero, R., Powers, R., and Montelione, G. (2006). A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins*, 62(3):587–603.

Hus, J. C., Marion, D., and Blackledge, M. (2000). De novo determination of protein structure by NMR using orientational and long-range order restraints. *Journal of Molecular Biology*, 298(5):927–36.

Iwata, K., Fujiwara, T., Matsuki, Y., Akutsu, H., Takahashi, S., Naiki, H., and Goto, Y. (2006). 3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR. *P Natl Acad Sci Usa*, 103(48):18119–24.

Jain, A., Vaidehi, N., and Rodriguez, G. (1993). A Fast Recursive Algorithm for Molecular Dynamics Simulations. *J. Comput. Phys.*, 106:258–268.

James, T., Borgias, B., Bianucci, A., and Zhou, N. (1990). Determination of DNA and protein structures in solution via complete relaxation matrix analysis of 2D NOE spectra. *Basic Life Sci*, 56:135–154.

Jeener, J., Meier, B., Bachmann, P., and Ernst, R. (1979). Investigation of exchange processes by 2-dimensional NMR-spectroscopy. *J Chem Phys*, 71(11):4546–4553.

Jehle, S., van Rossum, B., Stout, J., Noguchi, S., Falber, K., Rehbein, K., Oschkinat, H., Klevit, R., and Rajagopal, P. (2008). αb-crystallin: A hybrid solid-state/solution-state NMR investigation reveals structural aspects of the heterogeneous oligomer. *J Mol Biol*, in press.

Junius, F. K., O'Donoghue, S. I., Nilges, M., Weiss, A. S., and King, G. F. (1996). High resolution NMR solution structure of the leucine zipper domain of the c-Jun homodimer. *J Biol Chem*, 271(23):13663–7.

Juy, M., Penin, F., Favier, A., Galinier, A., Montserret, R., Haser, R., Deutscher, J., and Böckmann, A. (2003). Dimerization of Crh by reversible 3D domain swapping induces structural adjustments to its monomeric homologue Hpr. *J Mol Biol*, 332:767–776.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

Kajava, A. V. and Steven, A. C. (2006). Beta-rolls, beta-helices, and other beta-solenoid proteins. *Adv Protein Chem*, 73:55–96.

Kalbitzer, H. and Hengstenberg, W. (1993). The solution structure of the histidine-containing protein (HPr) from Staphylococcus aureus as determined by two-dimensional 1H-NMR spectroscopy. *Eur J Biochem*, 216:205–214.

Karplus, M. (1963). Vicinal proton coupling in Nuclear Magnetic Resonance. *J Am Chem Soc*.

Kelly, E. M., Hou, Z., Bossuyt, J., Bers, D., and Robia, S. L. (2008). Phospholamban oligomerization, quaternary structure, and SERCA-binding measured by FRET in living cells. *J Biol Chem*.

Kobayashi, N., Iwahara, J., Koshiba, S., Tomizawa, T., Tochio, N., Güntert, P., Kigawa, T., and Yokoyama, S. (2007). KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. *J Biomol NMR*, 39(1):31–52.

Korukottu, J., Schneider, R., Vijayan, V., Lange, A., Pongs, O., Becker, S., Baldus, M., and Zweckstetter, M. (2008). High-resolution 3D structure determination of kaliotoxin by solid-state NMR spectroscopy. *PLoS ONE*, 3(6):e2359.

Kovacs, H., O'Donoghue, S., Hoppe, H., Comfort, D., Reid, K., Campbell, I., and Nilges, M. (2002). Solution structure of the coiled-coil trimerization domain from lung surfactant protein D. *J Biomol NMR*, 24(2):89–102.

Kozin, S. A., Bertho, G., Mazur, A. K., Rabesona, H., Girault, J. P., Haertle, T., Takahashi, M., Debey, P., and Hoa, G. H. (2001). Sheep prion protein synthetic peptide spanning helix 1 and beta-strand 2 (residues 142-166) shows beta-hairpin structure in solution. *J. Biol. Chem.*, 276:46364–46370.

Kuszewski, J., Schwieters, C., Garrett, D., Byrd, R., Tjandra, N., and Clore, G. (2004). Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J Am Chem Soc*, 126(20):6258–6273.

Kwasigroch, J., Chomilier, J., and Mornon, J. (1997). A global taxonomy of loops in globular proteins. *J Mol Biol*, 259:855–872.

Lange, A., Becker, S., Seidel, K., Giller, K., Pongs, O., and Baldus, M. (2005). A concept for rapid protein-structure determination by solid-state NMR spectroscopy. *Angew Chem Int Ed Engl*, 44(14):2089–92.

Lange, A., Giller, K., Hornig, S., Martin-Eauclaire, M., Pongs, O., Becker, S., and Baldus, M. (2006). Toxin-induced conformational changes in a potassium channel revealed by solid-state NMR. *Nature*, 440(7086):959–62.

Lange, A., Luca, S., and Baldus, M. (2002). Structural constraints from proton-mediated rare-spin correlation spectroscopy in rotating solids. *J Am Chem Soc*, 124(33):9704–5.

Lange, A., Seidel, K., Verdier, L., Luca, S., and Baldus, M. (2003). Analysis of proton-proton transfer dynamics in rotating solids and their use for 3D structure determination. *J Am Chem Soc*, 125(41):12640–8.

Laskowski, R., Macarthur, M., Moss, D., and Thornton, J. (1993). PROCHECK : a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*, 26:283–291.

Lesage, A., Emsley, L., Penin, F., and Böckmann, A. (2006). Investigation of dipolar-mediated water-protein interactions in microcrystallune Crh by solid-state NMR spectroscopy. *J Am Chem Soc*, 128:8246–8255.

Levin, E., Kondrashov, D., Wesenberg, G., and Phillips, G. (2007). Ensemble refinement of protein crystal structures: validation and application. *Structure*, 15:1040–1052.

Levitt, M., Hirshberg, M., Sharon, R., Laidig, K., and Daggett, V. (1997). Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J Phys Chem*, 101:5051–5061.

Levy, E., Pereira-Leal, J., Chothia, C., and Teichmann, S. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput Biol*, 2(11):e155.

Levy, E. D., Erba, E. B., Robinson, C. V., and Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, 453(7199):1262–5.

Lewandowski, J., Paepe, G. D., and Griffin, R. (2007). Proton assisted insensitive nuclei cross polarization. *J Am Chem Soc*, 129:728–729.

Linge, J., Habeck, M., Rieping, W., and Nilges, M. (2003a). ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, 19(2):315–6.

Linge, J., Habeck, M., Rieping, W., and Nilges, M. (2004). Correction of spin diffusion during iterative automated NOE assignment. *J Magn Reson*, 167(2):334–342.

Linge, J. and Nilges, M. (1999). Influence of non-bonded parameters on the quality of NMR structures: A new force field for NMR structure calculation. *J Biomol NMR.*, 13:51–59.

Linge, J., Williams, M., Spronk, C., Bonvin, A., and Nilges, M. (2003b). Refinement of protein structures in explicit solvent. *Proteins*, 50(3):496–506.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2005). The structure of the flock house virus B2 protein, a viral suppressor of RNA interference, shows a novel mode of double-stranded RNA recognition. *EMBO Rep*, 6(12):1149–1155.

Liu, Z., Macias, M., Bottomley, M., Stier, G., Linge, J., Nilges, M., Bork, P., and Sattler, M. (1999). The three-dimensional structure of the HRDC domain and implications for the Werner and Bloom syndrome proteins. *Structure*, 7(12):1557–1566.

Loncharich, R., Brooks, B., and Pastor, R. (1992). Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, 32:523–535.

Loquet, A., Bardiaux, B., Gardiennet, C., Blanchet, C., Baldus, M., Nilges, M., Malliavin, T., and Böckmann, A. (2008). 3D Structure Determination of the Crh Protein from Highly Ambiguous Solid-State NMR Restraints. *J Am Chem Soc*, 130:3579–3589.

Lu, M. and Steitz, T. (2000). Structure of Escherichia coli ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-A resolution. *Proc Natl Acad Sci U S A*, 97:2023–2028.

Luca, S., Yau, W.-M., Leapman, R., and Tycko, R. (2007). Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR. *Biochemistry*, 46(47):13505–22.

Lührs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Döbeli, H., Schubert, D., and Riek, R. (2005). 3D structure of Alzheimer's amyloid-beta(1-42) fibrils. *P Natl Acad Sci Usa*, 102(48):17342–7.

Macura, S. and Ernst, R. (1980). Elucidation of cross relaxation in liquids by two-dimensional NMR-spectroscopy. *Mol Phys*, 41(1):95–117.

Manolikas, T., Herrmann, T., and Meier, B. (2008). Protein structure determination from (13)C spin-diffusion solid-state NMR spectroscopy. *J Am Chem Soc*.

Mao, H. and Willamson, J. (1999). Local folding coupled to RNA binding in the yeast ribosomal protein L30. *J Mol Biol*, 292:345–359.

Mareuil, F., Sizun, C., Perez, J., Schoenauer, M., Lallemand, J.-Y., and Bontems, F. (2007). A simple genetic algorithm for the optimization of multidomain protein homology models driven by NMR residual dipolar coupling and small angle X-ray scattering data. *Eur Biophys J*, 37(1):95–104.

Margiolaki, I. and Wright, J. P. (2008). Powder crystallography on macromolecules. *Acta Crystallogr A Found Crystallogr*, 64(Pt 1):169–80.

Margiolaki, I., Wright, J. P., Wilmanns, M., Fitch, A. N., and Pinotsis, N. (2007). Second SH3 domain of ponsin solved from powder diffraction. *J Am Chem Soc*, 129(38):11865–71.

Markus, M., Hinck, A., Huang, S., Draper, D., and Torchia, D. (1997). High resolution solution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding to RNA. *Nat Struct Biol.*, 4:70–77.

Mazur, A. (1997). Quasi-Hamiltonian equations of motion for internal coordinate molecular dynamics of polymers. *J. Comput. Chem.*, 18:1354–1364.

Mazur, A. (1998a). Accurate DNA dynamics without accurate long range electrostatics. *J. Am. Chem. Soc.*, 120:10928–10937.

Mazur, A. (1998b). Hierarchy of fast motions in protein dynamics. *J. Phys. Chem. B*, 102:473–479.

Mazur, A. (1999). Symplectic integration of closed chain rigid body dynamics with internal coordinate equations of motion. *J. Chem. Phys.*, 111:1407–1414.

Mazur, A. (2002). DNA dynamics in a water drop without counterions. *J. Am. Chem. Soc.*, 124:14707–14715.

Mazur, A. and Abagyan, R. (1989). New methodology for computer-aided modelling of biomolecular structure and dynamics. 1. Non-cyclic structures. *J Biomol Struct Dyn.*, 6:815–832.

Miclet, E., Duffieux, F., Lallemand, J., and Stoven, V. (2003). Backbone H(N), N, C(alpha), C', and C(beta) assignment of the 6-phosphogluconolactonase, a 266-residue enzyme of the pentose-phosphate pathway from human parasite Trypanosoma brucei. *J Biomol NMR*, 25:249–250.

Moseley, H. and Montelione, G. M. (1999). Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol*, 9(5):635–42.

Mumenthaler, C. and Braun, W. (1995). Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *Journal of Molecular Biology*, 254(3):465–80.

Nabuurs, S., Krieger, E., Spronk, C., Nederveen, A., Vriend, G., and Vuister, G. (2005). Definition of a new information-based per-residue quality parameter. *J Biomol NMR*, 33(2):123–34.

Nabuurs, S., Spronk, C., Krieger, E., Maassen, H., Vriend, G., and Vuister, G. (2003). Quantitative evaluation of experimental NMR restraints. *J Am Chem Soc*, 125(39):12026–12034.

Nabuurs, S., Spronk, C., Vriend, G., and Vuister, G. W. (2004). Concepts and tools for NMR restraint analysis and validation. *Concepts in Magnetic Resonance*.

Nabuurs, S., Spronk, C., Vuister, G., and Vriend, G. (2006). Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comput Biol*, 2(2):e9.

Nederveen, A., Doreleijers, J., Vranken, W., Miller, Z., Spronk, C., Nabuurs, S., Guntert, P., Livny, M., Markley, J., Nilges, M., Ulrich, E., Kaptein, R., and Bonvin, A. M. (2005). RECOORD: a REcalculated COORdinates Database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, 59:662–672.

Nesmelov, Y. E., Karim, C. B., Song, L., Fajer, P. G., and Thomas, D. D. (2007). Rotational dynamics of phospholamban determined by multifrequency electron paramagnetic resonance. *Biophys J*, 93(8):2805–12.

Neuhaus, D. and Williamson, M. (1989). *The Nuclear Overhauser Effect in Structural and Conformational Analysis*. VCH Publishers, New York.

Nilges, M. (1993). A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins*, 17(3):297–309.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol*, 245(5):645–60.

Nilges, M., Bernard, A., Bardiaux, B., Malliavin, T., Habeck, M., and Rieping, W. (2008). Accurate NMR structures through minimisation of an extended hybrid energy. *Structure*, 16(9):1305–12.

Nilges, M., Clore, G., and Gronenborn, A. (1988). Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Lett*, 239:129–136.

Nilges, M., Macias, M., O'Donoghue, S., and Oschkinat, H. (1997). Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J Mol Biol*, 269(3):408–422.

Nilges, M. and O'Donoghue, S. (1998). Ambiguous NOEs and automated NOE assignment. *Prog Nucl Mag Res Sp*, 32:107–139.

O'Donoghue, S., Chang, X., Abseher, R., Nilges, M., and Led, J. (2000). Unraveling the symmetry ambiguity in a hexamer: calculation of the R6 human insulin structure. *J Biomol NMR*, 16(2):93–108.

Oezguen, N., Adamian, L., Xu, Y., Rajarathnam, K., and Braun, W. (2002). Automated assignment and 3D structure calculations using combinations of 2D homonuclear and 3D heteronuclear NOESY spectra. *J Biomol NMR*, 22(3):249–63.

Ohman, A., Rak, A., Dontsova, M., Garber, M., and Hard, T. (2003). NMR structure of the ribosomal protein L23 from Thermus thermophilus. *J Biomol NMR.*, 26:131–137.

Oxenoid, K. and Chou, J. J. (2005). The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *P Natl Acad Sci Usa*, 102(31):10870–5.

Oxenoid, K., Rice, A. J., and Chou, J. J. (2007). Comparing the structure and dynamics of phospholamban pentamer in its unphosphorylated and pseudo-phosphorylated states. *Protein Sci*, 16(9):1977–83.

Paci, E., Gsponer, J., Salvetella, X., and Vendruscolo, M. (2004). Molecular dynamics studies of the process of amyloid aggregation of peptide fragments of transthyretin. *J Mol Biol*, 340:555–569.

Paravastu, A. K., Leapman, R. D., Yau, W.-M., and Tycko, R. (2008). Molecular structural basis for polymorphism in Alzheimer's β-amyloid fibrils. *P Natl Acad Sci Usa*, 105(47):18349–54.

Park, S., Mrse, A., Nevzorov, A., Mesleh, M., Oblatt-Montal, M., Montal, M., and Opella, S. (2003). Three-dimensional structure of the channel-forming trans-membrane domain of virus protein ”u” (Vpu) from HIV-1. *J Mol Biol*, 332:409–424.

Peng, X., Libich, D., Janik, R., Harauz, G., and Ladizhansky, V. (2008). Dipolar chemical shift correlation spectroscopy for homonuclear carbon distance measurements in proteins in the solid state: application to structure determination and refinement. *J Am Chem Soc*, 130(1):359–69.

Penin, F., Favier, A., Montserret, R., Brutscher, B., Deutscher, J., Marion, D., and Galinier, A. (2001). Evidence for a dimerisation state of the Bacillus subtilis catabolite repression HPr-like protein, Crh. *J Mol Microbiol Biotechnol*, 3(3):429–32.

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*, 21(8):1472–8.

Pintacuda, G., Giraud, N., Pierattelli, R., Böckmann, A., Bertini, I., and Emsley, L. (2006). Solid-State NMR of a Paramagnetic Protein: Assignment and Study of the Human Dimeric Oxidized Zn(II)-Cu(II) Superoxide Dismutase (SOD). *Angew. Chem. Int. Ed. Engl*, 46:1079–1082.

Potluri, S., Yan, A., Chou, J., Donald, B., and Bailey-Kellogg, C. (2006). Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der Waals packing. *Proteins*, 65(1):203–219.

Potluri, S., Yan, A., Donald, B., and Bailey-Kellogg, C. (2007). A complete algorithm to resolve ambiguity for intersubunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Sci*, 16(1):69–81.

Raibaud, S., Lebars, I., Guillier, M., Chiaruttini, C., Bontems, F., Rak, A., Garber, M., Allemand, F., Springer, M., and Dardel, F. (2002). NMR structure of bacterial ribosomal protein l20: implications for ribosome assembly and translational control. *J Mol Biol.*, 323:143–151.

Ranson, N., Stromer, T., Bousset, L., Melki, R., and Serpell, L. (2006). Insights into the architecture of the Ure2p yeast protein assemblies from helical twisted fibrils. *Protein Sci*, 15:2481–2487.

Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T., and Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382.

Rieping, W., Habeck, M., and Nilges, M. (2005a). Inferential Structure Determination. *Science*, 309(5732):303–6.

Rieping, W., Habeck, M., and Nilges, M. (2005b). Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. *J Am Chem Soc*, 127(46):16026–16027.

Rieping, W., Nilges, M., and Habeck, M. (2008). ISD: A software package for Bayesian NMR structure calculation. *Bioinformatics*, 24(8):1104–5.

Robia, S. L., Flohr, N. C., and Thomas, D. D. (2005). Phospholamban pentamer quaternary conformation determined by in-gel fluorescence anisotropy. *Biochemistry*, 44(11):4302–11.

Roszak, A., Howard, T., Southall, J., Gardiner, A., Law, C., Isaacs, N., and Cogdell, R. (2003). Crystal structure of the RC-LH1 core complex from Rhodopseudomonas palustris. *Science*, 302:1969–1972.

Ryckaert, J., Ciccotti, G., and Berendsen, H. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*, 23:327–341.

Saccenti, E. and Rosato, A. (2008). The war of tools: how can NMR spectroscopists detect errors in their structures? *J Biomol NMR*, 40(4):251–61.

Schmidt, H. L. F., Sperling, L. J., Gao, Y. G., Wylie, B. J., Boettcher, J. M., Wilson, S. R., and Rien stra, C. M. (2007). Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. *The journal of physical chemistry B*, 111(51):14362–9.

Schwieters, C. and Clore, G. (2001). Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *Journal of Magnetic Resonance*, 152:288–302.

Schwieters, C., Kuszewski, J., Tjandra, N., and Clore, G. (2003). The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*, 160(1):65–73.

Seidel, K., Etzkorn, M., Heise, H., Becker, S., and Baldus, M. (2005). High-resolution solid-state NMR studies on uniformly [13C,15N]-labeled ubiquitin. *Chembiochem*, 6(9):1638–47.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax, A. (2008). Consistent blind protein structure generation from NMR chemical shift data. *P Natl Acad Sci Usa*, 105(12):4685–90.

Solomon, I. (1955). Relaxation processes in a system of two spins. *Phys. Rev.*, 99(2):559–565.

Spronk, C., Linge, J., Hilbers, C., and Vuister, G. (2002). Improving the quality of protein structures derived by NMR spectroscopy. *J Biomol NMR*, 22(3):281–9.

Spronk, C., Nabuurs, S., Krieger, E., and Vriend, G. (2004). Validation of protein structures derived by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*.

Stein, E. G., Rice, L. M., and Brünger, A. T. (1997). Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J Magn Reson*, 124(1):154–64.

Stoldt, M., Woehnert, J., Goerlach, M., and Brown, L. (1998). The NMR Structure of Escherichia Coli Ribosomal Protein L25 Shows Homology to General Stress Proteins and Glutaminyl-tRNA Synthetases. *EMBO J.*, 17:6377–6384.

Stoldt, M., Wohnert, J., Ohlenschlager, O., Gorlach, M., and Brown, L. (1999). The NMR structure of the 5S rRNA E-domain-protein L25 complex shows preformed and induced recognition. *EMBO J*, 18:6508–6521.

Takegoshi, K., Nakamura, S., and Terao, T. (2001). C-13-H-1 dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem Phys Lett*, 344(5-6):631–637.

Tjandra, N. (1999). Establishing a degree of order: obtaining high-resolution NMR structures from molecular alignment. *Structure*, 7(9):R205–11.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure*, 16(2):295–307.

Traaseth, N. J., Ha, K. N., Verardi, R., Shi, L., Buffy, J. J., Masterson, L. R., and Veglia, G. (2008). Structural and dynamic basis of phospholamban and sarcolipin inhibition of Ca(2+)-ATPase. *Biochemistry*, 47(1):3–13.

Traaseth, N. J., Verardi, R., Torgersen, K. D., Karim, C. B., Thomas, D. D., and Veglia, G. (2007). Spectroscopic validation of the pentameric structure of phospholamban. *P Natl Acad Sci Usa*, 104(37):14676–81.

Tycko, R. (2006). Molecular structure of amyloid fibrils: insights from solid-state NMR. *Q Rev Biophys*, 39(1):1–55.

Ulrich, E., Akutsu, H., Doreleijers, J., Harano, Y., Ioannidis, Y., JLin, Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C., Tolmie, D., Wenger, R., Yao, H., and Markley, J. (2008). BioMagResBank. *Nucleic Acids Res*, 36:D402–D408.

van Dijk, A. and Bonvin, A. (2006). Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics*, 22:2340–2347.

van Dijk, A. D. J., Fushman, D., and Bonvin, A. M. J. J. (2005). Various strategies of using residual dipolar couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against 15N-relaxation data. *Proteins*, 60(3):367–81.

Vilar, M., Chou, H.-T., Lührs, T., Maji, S. K., Riek-Loher, D., Verel, R., Manning, G., Stahlberg, H., and Riek, R. (2008). The fold of alpha-synuclein fibrils. *P Natl Acad Sci Usa*, 105(25):8637–42.

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, 59(4):687–696.

Walser, R., Hunenberger, P., and van Gunsteren, W. (2002). Molecular dynamics simulations of a double unit cell in a protein crystal: volume relaxation at constant pressure and correlation of motions between the two unit cells. *Proteins*, 48:327–340.

Wang, C. S., Lozano-Pérez, T., and Tidor, B. (1998). AmbiPack: a systematic algorithm for packing of macromolecular structures with ambiguous distance constraints. *Proteins*, 32(1):26–42.

Wang, X., Bansal, S., Jiang, M., and Prestegard, J. H. (2008). RDC-assisted modeling of symmetric protein homo-oligomers. *Protein Sci*, 17(5):899–907.

Wasmer, C., Lange, A., Melckebeke, H. V., Siemer, A., Riek, R., and Meier, B. (2008). Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. *Science*, 319:1523–1526.

Weiss, M. (1990). Distinguishing symmetry-related intramolecular and intermolecular Nuclear Overhauser Effects in a protein by asymmetric isotopic labeling. *J Magn Reson*, 86:626–632.

Wider, R., G, R., Billeter, M., Hornemann, S., Glockshuber, R., and Wüthrich, K. (1998). Prion protein NMR structure and familial human spongiform encephalopathies. *Proc Natl Acad Sci U S A*, 95:11667–11672.

Wishart, D., Sykes, B., and Richards, F. (1992). The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651.

Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc Natl Acad Sci USA*, 93(25):14249–55.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. Wiley.

Wüthrich, K., Billeter, M., and Braun, W. (1983). Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with Nuclear Magnetic Resonance. *J Mol Biol*, 169(4):949–961.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C., Cort, J., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L., Gehring, K., Kennedy, M., Edwards, A., and Arrowsmith, C. (2002). An NMR approach to structural proteomics. *Proc Natl Acad Sci USA*, 99(4):1825–1830.

Zech, S., Wand, A., and McDermott, A. (2005). Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. *J Am Chem Soc*, 127(24):8618–26.

Zhou, D., Shea, J., Nieuwkoop, A., Franks, W., Wylie, B., Mullen, C., Sandoz, D., and Rienstra, C. (2007). Solid-state protein-structure determination with proton-detected triple-resonance 3D magic-angle-spinning NMR spectroscopy. *Angew Chem Int Ed Engl*, 46(44):8380–3.

Zirah, S., Kozin, S., Mazur, A., Blond, A., Cheminant, M., Ségalas-Milazzo, I., Debey, P., and Rebuffat, S. (2006). Structural changes of the 1-16 region of the Alzheimer's disease amyloid beta -peptide upon zinc binding and in vitro aging. *J. Biol. Chem.*, 281(4):2151–61.

Zweckstetter, M. and Bax, A. (2000). Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *J Am Chem Soc*, 122(15):3791–3792.

Summary

Nuclear Magnetic Resonance (NMR) spectroscopy has become an important tool to investigate the structure and dynamics of proteins and protein complexes. The recent progress in solution and solid-state NMR opens the door to detailed analyses of large macromolecular structures, and notably, symmetric protein aggregates. In this context, the determination of high resolution three-dimensional structures of proteins by NMR critically relies on efficient and reliable automated assignment strategies. In this thesis, several methods were developed for automated assignment and structure calculation from NMR data in the framework of the Ambiguous Restraints for Automated Assignment (ARIA) protocol. These methods tackle the majors issues of structure determination by modern NMR spectroscopy. First, the incorporation of a network anchoring approach in the ARIA methodology was shown to considerably speed-up the NOE assignment process while preserving a high structural quality of the obtained models. With a new protocol based on the Internal Coordinates Molecular Dynamics (ICMD) methodology, it was also demonstrated that the structures of small proteins can be calculated through a molecular dynamics based simulated annealing protocol entirely performed with a complete general purpose force-field. With regard to the structural quality of the models, this approach can be considered as intermediate between structure calculation with a simple force-field and refinement in a shell of water molecules.

In a second stage, the difficult problem of *de novo* structure determination of symmetric protein assemblies from NMR data was considered. For symmetric homo–dimers, a specific approach based on ambiguous distance restraints was shown to perform well when the ambiguity of the restraints is limited to an ambiguity of chain. For some homo–dimeric folds, network anchoring was found to be essential for unravelling both chain and proton assignment ambiguities and for calculating high quality dimeric structures from completely unassigned NOE cross-peaks. Furthermore, a new general method based on strict symmetry definitions was conceived for structure calculation of any type of symmetric assemblies from both solution and solid-state NMR data. This method was successfully applied to the calculation of a pentameric membrane protein structure and of amyloid fibrils from ambiguous ssNMR distance restraints. In addition, an analysis of the conformational landscape of the Crh protein during oligomerisation and crystallisation was conducted with a simultaneous use of solution-state, solid-state NMR and X-ray crystallography data. The generation of NMR and crystallographic conformers, as well as molecular dynamics simulations allowed us to describe the impact of the long-range solid-state order on the convergence. The variability of Crh monomer orientation is concentrated in rotations around the dimer longitudinal axis, which is amplified but not created by the use of distance restraints.

Finally, we have addressed the problem of high-resolution structure determination from Magic Angle Spinning (MAS) solid-state NMR data, which is often hampered by two crucial complications : the high level of ambiguity present in the spectra and the detection of inter-molecular correlations. By adapting the ARIA methodology to the specific case of CHHC/NHHC ssNMR spectra recorded on a uniformly labelled sample, it was possible to determine the structure of the dimeric Crh protein without manual cross-peak assignment. Moreover, a specific protocol was designed to refine the structure of micro-crystalline proteins incorrectly resolved from MAS ssNMR spectra, due to the presence of inter-molecular correlations. By taking into account the overall configuration of the crystal lattice in the calculation, it was possible to drastically increase the accuracy of the SH3 domain structure from a subset of reliable unassigned cross-peaks.

Zusammenfassung

Die Kernspinresonanzspektroskopie (NMR) ist ein wichtiges Werkzeug, um die Struktur und Dynamik von Proteinen und Proteinkomplexen zu untersuchen. Neue Fortschritte in der Lösungsmittel- und der Festkörper-NMR (ssNMR) ermöglichen detailierte Analysen grosser makromolekularer Strukturen, insbesondere symmetrischer Proteinaggregate. In diesem Kontext hängt die Bestimmung von hochaufgelösten dreidimensionalen Proteinstrukturen mittels NMR kritisch von effizienten und verlässlichen automatisierten Assignmentstrategien ab. In der vorliegenden Arbeit wurden mehrere Methoden für die automatische Zuordnung und Strukturrechnung aus NMR-Daten im Rahmen des Protokolls “Ambiguous Restraints for Automated Assignment” (ARIA) entwickelt. Diese Methoden nehmen die wichtigsten Schwierigkeiten der Strukturbestimmung mittels moderner NMR-Spektroskopie in Angriff. Im ersten Teil der Arbeit wurde gezeigt, dass die Verwendung einer Netzwerkverankerung im ARIA-Protokoll die NOE-Zuordnung erheblich beschleunigt und zugleich die hohe Qualität der Strukturen erhält. Desweiteren wurde gezeigt, dass ein neues Protokoll, das auf der Molekulardynamik (MD) in internen Koordinaten (ICMD) beruht, die Berechnung der Strukturen kleiner Proteine mittels MD-basiertem Simulated Annealing in einem vollen MD-Kraftfeld ermöglicht. Bezüglich der Qualität der Strukturen kann dieser Zugang als ein Zwischenschritt in der Strukturberechnung mit vereinfachtem Kraftfeld und einer Verfeinerung in einer Box von Wassermolekülen angesehen werden.

Im zweiten Teil dieser Arbeit wurde das schwierige Problem der De-novo-Strukturbestimmung symmetrischer Proteinkomplexe behandelt. Für symmetrische Homodimere wurde gezeigt, dass ein Zugang basierend auf mehrdeutigen Distanzeinschränkungen zufriedenstellend abschneidet, wenn die Mehrdeutigkeit in den Distanzeinschränkungen auf die Mehrdeutigkeit in der Zuweisung der Polypeptidkette beschränkt ist. Für einige Faltungen von Homodimeren stellte sich die Netzwerkverankerung als essentiell heraus, um Mehrdeutigkeiten sowohl in der Ketten- als auch der Protonenzuordnung aufzulösen, und um qualitativ hochwertige Dimerstrukturen aus komplett nicht-zugeordneten NOEs zu berechnen. Desweiteren wurde eine neue Methode entwickelt, die auf strikten Symmetrydefinitionen basiert und die Strukturrechnung von Komplexen, die eine beliebige Symmetrie aufweisen, von Lösungmittel- und Festkörper-NMR-Daten ermöglicht. Diese Methode wurde erfolgreich in der Berechnung einer Membranproteinstruktur mit fünffacher Symmetrie sowie von Amyloidfibrillen aus mehrdeutigen ssNMR-Distanzeinschränkungen eingesetzt. Zusätzlich wurde eine Analyse der Konformationslandschaft des Crh Proteins während der Oligomerisierung und Kristallisation durchgeführt unter gleichzeitiger Verwendung von Lösungmittel-NMR, Festkörper-NMR und röntgenkristallographischen Daten. Die Generierung von NMR- und kristallographischen Konformern sowie MD-Simulationen erlaubten uns, den Einfluss der langreichweitigen Festkörperordnung auf die Konvergenz zu beschreiben. Die Variabilität der Crh-Monomer-Orientierung konzentriert sich in Rotationen um die Längsachse des Dimers, diese wird durch den Gebrauch von Distanzeinschränkungen verstärkt, aber nicht erzeugt.

Schliesslich wurde das Problem der hochaufgelösten Strukturbestimmung aus Magic Angle Spinning (MAS) Festkörper-NMR-Daten bearbeitet; hier treten zwei Schwierigkeiten auf: der hohe Grad an Mehrdeutigkeit in den Spektren und das Aufspüren von intermolekularen Korrelationen. Durch Anpassung der ARIA-Methode für den spezifischen Fall der CHHC/NHHC-ssNMR-Spektren, welche für eine uniform markierte Probe gemessen wurden, konnten wir die Struktur eines Crh-Dimers berechnen, ohne zuvor die Kreuzsignale manuell zuzuordnen. Ausserdem wurde ein Protokoll entwickelt, um die Struktur mikrokristalliner Proteine zu verfeinern, die aufgrund inter-molekularer Korrelationen falsch aus MAS-ssNMR-Spektren aufgelöst wurden. Durch Berücksichtigung der gobalen Konfiguration des Kristallgitters in der Strukturrechnung war es möglich, die Genauigkeit der Struktur einer SH3-Domäne aus einer Untermenge verlässlicher, nicht-zugeordneter Kreuzsignale dramatisch zu verbessern.

Résumé

La spectroscopie par Résonance Magnétique Nucléaire (RMN) est devenue un outil important pour étudier la structure et la dynamique des protéines et des complexes protéiques. Les progrès récents de la RMN en solution et à l'état solide ouvrent la voie à l'étude détaillée de grandes structures macromoléculaires et, notamment, des agrégats symétriques de protéines. Dans ce contexte, la détermination de structure tridimensionnelle de haute résolution de protéines par RMN repose de manière critique sur des des stratégies d'attribution automatisée efficaces et fiables. Dans cette thèse, plusieurs méthodes ont été développées pour l'attribution automatique et le calcul de structure à partir de données RMN dans le cadre du protocole ARIA. Ces méthodes s'attaquent aux problèmes majeurs de la détermination de structure par spectroscopie RMN. Tout d'abord, il a été montré que l'inclusion d'un réseau d'ancrage dans la méthodologie ARIA accélère considérablement le processus d'attribution des NOEs, tout en conservant une haute qualité structurale des modèles obtenus. Avec un nouveau protocole basé sur la méthode ICMD (Dynamique moléculaire en coordonnées internes), il a été aussi démontré que des structures de petites protéines peuvent être calculées via un recuit simulé en dynamique moléculaire entièrement réalisé avec un champ de force complet et général. Au regard de la qualité structurale des modèles cette approche peut être considérée comme intermédiaire entre le calcul de structure avec champ de force simplifié et un raffinement avec des molécules d'eau.

Dans un deuxième temps, le délicat problème de la détermination *de novo* de la structure d'assemblage symétriques de protéines à partir de données RMN a été étudié. Pour les

homo-dimères symétriques, une approche spécifique, basée sur des contraintes ambiguës de distance et intégrée au protocole ARIA, a montré de bons résultats lorsque l'ambiguïté des contraintes est limitée à une ambiguïté de chaîne.Pour certains repliements homo-dimerique, le réseau d'ancrage s'est avéré indispensable pour élucider à la fois l'ambiguïté de chaîne et de déplacements chimiques, et pour calculer des structures de dimères de haute qualité à partir de pics NOEs non attribués. En outre, une nouvelle méthode générale, basée sur des symétries strictes, a été conçue pour le calcul de structures symétriques à partir de données de RMN en solution et à l'état solide. Cette méthode a été appliquée avec succès au calcul de la structure d'une protéine membranaire pentamerique et de fibrilles amyloïdes à parir de contraintes de distance ambiguës de RMN du solide. Par ailleurs, une analyse du paysage conformationnel de la protéine Crh au cours de l'oligomérisation et de la cristallisation a été réalisée par une utilisation simultanée de données de RMN en solution, à l'état solide et de cristallographie aux rayons X. La génération de conformations RMN et cristallographiques, ainsi que des simulations de dynamique moléculaire ont permis de décrire l'influence d'un ordre à longue distance dans l'état solide sur la convergence. La variabilité de l'orientation des monomères de Crh est concentrée dans une rotation autour de l'axe longitudinal du dimère, qui n'est qu'amplifiée et non créée par l'emploi de contraintes de distance.

Enfin, nous avons abordé le problème de la détermination de structure de haute-résolution par spectroscopie RMN à l'état solide à l'angle magique, qui est souvent entravée par deux complications cruciales : le haut niveau l'ambiguïté présent dans les spectres et la détection des corrélations inter-moléculaires. En adaptant la méthodologie ARIA au cas spécifique des spectres CHHC/NHHC de RMN à l'état solide enregistrés sur des échantillons uniformément marqués, il a été possible de déterminer la structure de la protéine dimérique Crh sans attributions manuelles des pics. De plus, un protocole spécifique a été conçu pour affiner la structure micro-cristalline de protéines incorrectement résolue par des spectres RMN à l'état solide, en raison de la présence de corrélations inter-moléculaires. En tenant compte de l'ensemble de la configuration du réseau cristallin dans le calcul, nous avons pu améliorer considérablement la justesse de la structure du domaine SH3 à partir d'un sous-ensemble fiable de pics non-attribués.

CURRICULUM VITAE

Bardiaux Benjamin
Nationality: French
Age: 30

Education

2001 Bachelor degree in Biology (Louis Pasteur University, Strasbourg, France)

2002 1st year of Master in Molecular and Cell Biology (Louis Pasteur University, Strasbourg, France)

2003 Masters degree in Bioinformatics and Applied Genomics (European School of the Higher Rhine Universities and Louis Pasteur University, Strasbourg, France).
"*In silico* analysis of disulfide bonds data from biological databanks using data-mining methods."

Additional courses:

2005 Protein NMR. Recording, structure calculation and evaluation, Summer Course, University of Copenhagen, Denmark

2007 Structure determination of biological macromolecules by solution NMR, EMBO Practical Course, Biozentrum Basel, Switzerland

Publications

Bardiaux B, Favier A, Etzkorn M, Baldus M, Böckmann A, Nilges M and Malliavin TE. (2008) *Simultaneous use of liquid and solid-state NMR to study the conformational landscape of the Crh protein during oligomerisation and crystallization.* Submitted to Proteins.

Nilges M, Bernard A, **Bardiaux B**, Malliavin TE, Habeck M and Rieping W. (2008) *Accurate NMR structures through minimisation of an extended hybrid energy.* Structure, 16 (9), 1305-12

Bardiaux B, Bernard A, Rieping W, Habeck M, Malliavin TE and Nilges M. (2008) *Influence of different assignment conditions on the determination of symmetric homo-dimeric structures with ARIA*. Proteins (2008) vol. 75 (3) pp. 569-585

Bardiaux B, Bernard A, Rieping W, Habeck M, Malliavin TE and Nilges M. (2008) *Graphical analysis of NMR structural quality and interactive contact map of NOE assignments in ARIA*. BMC Struct Biol. 8 (1), 30

Loquet A, **Bardiaux B**, Gardiennet C, Blanchet C, Baldus M, Nilges M, Malliavin T and Böckmann A. (2008) *3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints*. J Am Chem Soc. 130 (11), 3579-89

Rieping W, Habeck M, **Bardiaux B**, Bernard A, Malliavin T and Nilges M. (2007) *ARIA 2: Automated NOE assignment and data integration in NMR structure calculation.* Bioinformatics. 23, 381-382

Bardiaux B, Malliavin T, Nilges M and Mazur AK. (2006) *Comparison of different torsion angle approaches for NMR structure determination.* J. Bio. NMR. Mar. 34 (3), 153-66

Tessier D, **Bardiaux B**, Larré C, Popineau Y. (2004) *Data mining techniques to study the disulfide-bonding state in proteins: signal peptide is a strong descriptor.* Bioinformatics. 20, 2509-2512.

Oral communications and Posters

Bardiaux B. *Automated NOE assignment and structure determination of symmetric dimers.* French-Benelux Meeting on Magnetic Resonance, Blankenberge, Belgium, March 2006 (talk)

Bardiaux B. *ARIA 2.2: outils et exemples pour le calcul de structures RMN de complexes protéiques.* SFBBM, RMN structurale et fonctionnelle : de la molécule aux systèmes integrés Dourdan, France, May 2007 (talk)

Bardiaux B, Malliavin TE and Nilges M. *NMR structure calculation of symmetric aggregates*, Joint NMR-life and Extend-NMR Meeting, Berlin, Germany, October 2008 (poster)

Bardiaux B, Bernard A, Malliavin TE, Rieping W, Habeck M and Nilges M. *ARIA 2.2: Automated structure determination from NOESY*, 22nd International Conference on Magnetic Resonance in Biological Systems, Göttingen, Germany, August 2006 (poster)

Bardiaux B, Favier A, Böckman A, Nilges M and Malliavin TE. *Simultaneous use of liquid and solid-state NMR restraints for NMR structure determination.* 22nd International Conference on Magnetic Resonance in Biological Systems, Göttingen, Germany, August 2006 (poster)

ACI ICMD_RMN, *Le calcul en coordonnées internes pour la détermination des contraintes RMN inter- et intra-moléculaires.*, Réunion ACI Informatique, Mathématique, Physique en Biologie Moléculaire, Lyon, France, July 2005 (poster)

APPENDIX

Crh molecular dynamics simulations

The simulations *sol_dimer* and *sol_tetra* were performed at constant pressure, and the simulation *cryst_tetra* was performed at constant volume. A cutoff distance of 10 Å was used to determine the water box size in *sol_dimer* and *sol_tetra*, whereas a cutoff distance of 2 Å was used in *cryst_tetra*, in order to model the effect of long-range order observed in the solid state. In *cryst_tetra*, the 10 sulfate molecules and the 9 glycerol molecules observed in the crystal were kept in the simulation box.

The simulations of 10 ns were performed using the package AMBER 9.0 [Case et al., 2004], and the ff99SB force field [Hornak et al., 2006]. A cutoff of 10 Å was used for Lennard-Jones interactions, and long-range electrostatic interactions were calculated with the Particle Mesh Ewald (PME) protocol [Darden et al., 1993]. The systems total charge was neutralised using sodium counterions ions. Pressure was regulated with isotropic position scaling and a relaxation time of 1 ps, and temperature using a Langevin thermostat [Loncharich et al., 1992] with a collision frequency of 2 ps^{-1}. For *sol_dimer* and *sol_tetra*, the simulations were performed at temperature 298 K and pressure 1 atm, whereas for *cryst_tetra*, the temperature was 283 K, in order to keep conditions close to those of the ssNMR experiments [Böckmann et al., 2003]. The SHAKE algorithm [Ryckaert et al., 1977] was used to keep rigid all covalent bonds involving hydrogens, enabling a time step of 2 fs. Atom coordinates were saved every ps. The MD analysis was mainly performed using ptraj [Case et al., 2004].

Simulations were initiated by some rounds of semi-restrained and then unrestrained minimisation of the entire system. Heating of the system up to 300 K was realised during 20 ps at constant volume, while restraining the positions of the atoms of the Crh dimer or tetramer with a force constant of 25 kcal/(mol.Å^2). The equilibration process was then performed: 1 MD round of 5 ps at constant volume and 4 MD rounds of 2.5 ps were run while reducing the position restraints from 25 kcal/(mol.Å^2) down to 5 kcal/(mol.Å^2); eventually a last MD round of 70 ps was performed with a restraint of 2.5 kcal/(mol.Å^2).

	sol_dimer	*sol_tetra*	*cryst_tetra*
Number of counterions	6	12	32
Water box dimensions (Å)	74.1 x 89.7 x 58.3	72.5 x 91.7 x 81.1	58.9 x 79.2 x 68.9
Number of water molecules	9396	12560	6266
Total number of atoms	30918	43140	24454

Table 1: Preparation details of the MD simulations.