Chapter 18
An Overview on Protein Structure Determination by NMR: Historical and Future Perspectives of the use of Distance Geometry Methods

Fabio C.L. Almeida, Adolfo H. Moraes, and Francisco Gomes-Neto

Abstract Determination of the protein high-resolution structures is essential for the understanding of complex biological mechanisms, for the development of biotechnological methods, and for other applications such as drug discovery. Protein structures solved by nuclear magnetic resonance (NMR) rely on a set of semiquantitative short-range distances and angles information. The exploration of the whole conformational space imposed by the experimental restraints is not a computationally simple problem. The lack of precise distances and angles does not allow to find solutions to this problem by fast geometric algorithms. The main idea is to define an atomic model for the protein structure and to exploit all known geometric angle and distance information along with the semi-quantitative short-range experimental information from NMR. We give an overview of the development of computational methods aimed at solving the problem either by metric matrix distance geometry or using other methods such as simulated annealing. We also discuss future demands and perspectives for structural calculations using NMR data. The need of determining larger and more complex protein structures implies the strong necessity of developing new methods for structural calculation with sparse data.

> *We are like dwarfs sitting on the shoulders of giants. We see more, and things that are more distant, than they did, not because our sight is superior or because we are taller than they, but because they raise us up, and by their great stature add to ours* ... This sentence was written (in Latin) in the logic treatise Metalogicon by John of Salisbury in 1159. Salisbury attributed this sentence to Bernard of Chartres. It was reused later by Isaac Newton to explain the development of western science.

F.C.L. Almeida (✉) • A.H. Moraes • F. Gomes-Neto
National Center of Nuclear Magnetic Resonance, Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
e-mail: falmeida@cnrmn.bioqmed.ufrj.br; amoraes@cnrmn.bioqmed.ufrj.br; fgomes@cnrmn.bioqmed.ufrj.br

A. Mucherino et al. (eds.), *Distance Geometry: Theory, Methods, and Applications*,
DOI 10.1007/978-1-4614-5128-0_18,

18.1 Introduction

The goal of this chapter is to give an overview of protein structure determination by nuclear magnetic resonance(NMR). We will give a historical perspective that illustrates the necessity of solving distance geometry problems in order to determine protein structure. Briefly, the problem consists of exploiting experimental information that is obtained from NMR experiments and that mainly concerns distances between hydrogen atoms, in order to find the three-dimensional structure of a protein. Together with the NMR information, we can also use additional information deduced from the knowledge accumulated during the twentieth century on molecular structures. For this reason, the sentence by John of Salisbury that we quoted perfectly applies to protein structure determination.

This chapter is organized as follows. In Sect. 18.2, we briefly introduce protein structures. In Sect. 18.3, we give a description of the conformational space of protein structures, while we discuss about molecular dynamics in Sect. 18.4. In Sect. 18.5 we briefly describe NMR experiments, and Sect. 18.6 is devoted to the problem of deriving some atomic distance restraints from NMR data. Section 18.7 is devoted to some pseudo-potentials that can be used for modeling the distance restraints, while the distance geometry problem with NMR data is discussed in Sect. 18.8, where the first implemented computational method for protein structural calculation from NMR data is presented. Nowadays, the most used method for solving distance geometry problems with NMR data is the meta-heuristic simulated annealing (SA): we present two variants of this algorithm in Sect. 18.9, one that is based on the Cartesian representation of the protein structures and the other one that is based on the torsion angle representation. We conclude our chapter in Sect. 18.10, where we discuss some future demands for protein structure determination.

18.2 Introduction to Protein Structure

The determination of protein high resolution structures is essential for the understanding of complex biological mechanisms, for the development of biotechnological methods, drug design, and many other applications. Requesting the protein structure to have a high resolution implies that the position of each of its atom is identified precisely (uncertainty smaller than 1 Å).

Proteins are polymeric chains in which the units are the 20 natural L-α-aminoacids that are connected by peptide bonds. Several structures of dipeptides, which have been solved by X-ray crystallography in the early 1930s by the group led by Linus Pauling, demonstrated that the peptide bond can have two configurations: cis and trans [34, 57]. The trans configuration has lower energy and it represents the most abundant configuration in proteins. There are however exceptions, such as cis-prolines which are important for thioredoxin activity [13, 29, 30]. The peptide bond

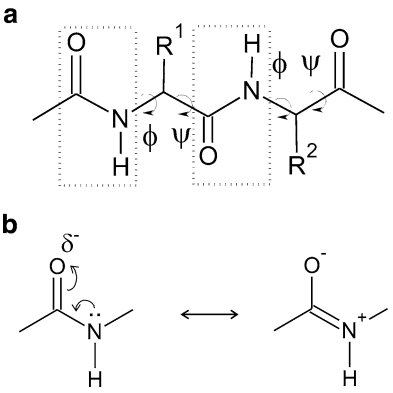

Fig. 18.1 Illustration of a peptide bond planar structure: (**a**) the planar peptide bond (*dotted rectangle*); the dihedral angles that give torsion freedom for the peptide backbone (Φ and Ψ angles) are indicated by *arrows*; the side chains are represented by R^1 and R^2; (**b**) resonance forms of the peptide bond and its double-bond character

is planar because of the resonance effect that gives to it a double-bond character. Figure 18.1 illustrates a polypeptide chain and the planar character of the peptide bond.

Several other geometrical properties of proteins were defined before the first protein structure was solved by Kauzmann in 1964 [34]. Maybe the most important property is given by the presence of secondary structure elements, such as α-helices, that was firstly proposed by Linus Pauling [56–59].

The amino acid sequence is also called primary structure. An amino acid included in a polypeptide chain is called amino acid residue. Secondary structures represent local structural organizations that are stabilized by hydrogen bonds in the main chain. They can be observed in several proteins. The main polypeptide chain, also called protein backbone, is the protein sequence without the radicals of each L-α-amino acids, i.e., without side chains. The backbone chains contain only one hydrogen donor to a hydrogen bond, the amidic hydrogen (N–H), and only one electron pair, which serves as hydrogen acceptor in a hydrogen bond, the free electron pair of the carbonyl (CO). Recall that electron pairs of amide nitrogen on the main chain is "busy" because it is part of the double bond related to one of the resonance forms of the peptide bond (see Fig. 18.1). This means that, in proteins, the only hydrogen bonds stabilizing secondary structures are the ones between the amidic N–H and the carbonyl.

The protein backbone needs to bend in order to stabilize secondary structures, because this is needed for forming hydrogen bonds between amino acids. There are two degrees of freedom that leads to the bending of the main chain. These degrees of freedom are defined by the dihedral angles Φ and Ψ. The dihedral angle Φ among the atoms C_α^{i-1}, N, C_α, and C′ defines the torsion of the bond N–C_α. The dihedral angle Ψ among N, C_α, C′, and $N^{(i+1)}$ defines the torsion of the bond C_α–C′ (where C′ is the carbonyl carbon). It is important to note that the two-dimensional plot of Φ versus Ψ, known as Ramachandran plot, can describe the folding of a protein in the sense that particular pairs (Φ,Ψ) can be identified for each amino acid residue forming the protein. It is also remarkable that the Ramachandran plot defines all the conformational space for the backbone structure of a protein [61]. Note that the

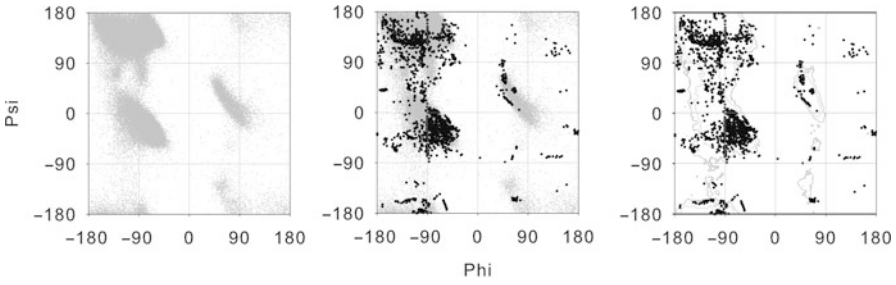

Fig. 18.2 Ramachandran plot. *Leftmost plot*: Φ and Ψ angles for 500 high-resolution crystal structures selected from PDB: the plot shows a general dihedral freedom adopted in proteins [48]. *Plot in the center*: superposition of the angles Φ and Ψ extracted from the 20 lower-energy structures of thioredoxin 1 (PDB id:2I9H) solved by solution NMR [60]. *Rightmost plot*: contour plot showing allowed and generously allowed regions for the angles Φ and Ψ calculated from the initial data set and the superposition of dihedral angles from 2I9H structures

torsion of the peptide bond is not considered as a degree of freedom because of its planarity.

In the 1960s, Ramachandran performed some computational calculations on small peptides and showed that not all combinations of Φ and Ψ are possible in proteins. Moreover, there are high-energetical conformations that can be considered as forbidden [61]. On the other hand, Φ and Ψ combinations that can be observed in secondary structure define the lowest-energy conformations. High energetical states are due to steric effects between large side chains. We can say that, for a given amino acid residue, the larger is the side chain, the smaller are the possible low-energy areas in the Ramachandran plot. Figure 18.2 shows the Ramachandran plot of yeast thioredoxin 1 (PDB id:2I9H, [2]) and the location of the main secondary structures.

The two most frequent secondary structure elements are α-helices and β-sheets (parallel and antiparallel). It is not in the scope of this chapter to describe all the secondary structure elements but to contextualize in regard to the protein structural determination problem. The two secondary structure elements define the lowest-energy regions of the Ramachandran plot, in which the α-helix region is near $\Phi = -60^\circ$ and $\Psi = -30^\circ$, and the β-sheet region is near $\Phi = -120^\circ$ and $\Psi = 135^\circ$ (see Fig. 18.2).

18.3 The Problem of Conformational Space

As we have seen in the previous section, the Ramachandran plot is related to the backbone conformation of a protein. If one knows the dihedral angles Φ and Ψ for all residues, not only the secondary structure is determined, but also the

tertiary structure can be derived from this information. The tertiary structure is given by all three-dimensional coordinates of the atoms forming the protein. The quaternary structure defines the structural organization of proteins in oligomers. The oligomerization of a protein can be homo-oligomerization, where the association involves the same amino acid chain, or hetero-oligomerization, where the association occurs with different chains. There are several levels of oligomerization: dimers, decamers/dodecamers, and virus structures, which may contain thousands of chains.

It is important to discuss the forces that stabilize the tertiary and the quaternary structures of proteins. They are mainly represented by intermolecular non-covalent interactions between atoms belonging to the protein backbone or to the side chains of the amino acids. These interactions are generally called *tertiary contacts*. We give, in the following, some details about the main interaction forces in proteins.

18.3.1 Hydrogen Bonds

Besides the hydrogen bonding between the amidic group N–H and the carbonyl group (CO) of the backbone that stabilizes the secondary structures, there are many others amino acid side chains that can form hydrogen bonds. Amino acid residues serine, threonine, and tyrosine contain a hydroxyl group (–OH) that can be either donor or acceptor of a hydrogen in a hydrogen bond. Moreover, aspartate and glutamate are carboxylic acids that contain a hydroxyl and a carbonyl (donor and acceptor). Asparagine and glutamine contain amide (–NH2, mainly donor) and a carbonyl (acceptor). Finally, lysine (amine) and arginine (guanidinium group) are also good donors and acceptors for hydrogen bond.

In proteins, there is always competition between intramolecular and intermolecular (the protein solvent is water) hydrogen bonding. The more is the residue exposed to water (near the protein surface), the smaller is the contribution of the intramolecular hydrogen bond to the stabilization of tertiary and quaternary structures.

18.3.2 Coulomb Interactions

Several side chains can ionize in water, associating or dissociating protons that become charged. At neutral pH, aspartate, glutamate, and the carboxy terminus are negatively charged, whereas lysine, arginine, histidine, and the amino-terminus are positively charged. The proximity of two opposite charges leads to Coulombic interactions that, when present, strongly contribute to the stabilization of tertiary and quaternary structures.

Charged residues are solvated by water. The dipole of water neutralizes the charge. Coulombic interaction, also known as salt bridge, is restricted to protein

microenvironments where the water access is limited. Similar to the dependence of the strength of hydrogen bonds on water access, the more exposed water is to the charged residue, the weaker is the intramolecular Coulombic interaction.

18.3.3 Van der Waals

van der Waals (vdW) interactions are dipolar–dipolar interactions that occur at very short distances ($r < 5$ Å). Although they are the weakest forces involved in the protein structure stabilization, they are the most important for the tertiary and quaternary structures of proteins because of their high abundance. Every dipole that is close to each other contributes to the protein stabilization.

Water contributes favorably for VdW because the apolar hydrophobic side chains tend to avoid the exposure to the solvent, the so-called hydrophobic effect. In this way, they become part of a hydrophobic core. The exposure of hydrophobic side chains to the bulk water leads to high entropic penalty. The VdW force has two components, one is repulsive, at very short distances ($r < 1.8$ Å), which decays proportionally to r^{-12} and the other one is attractive ($1.8 < r < 5$ Å), which decays proportionally to r^{-6}.

VdW is the "glue" that sticks together the protein structure. Hydrophobic residues are packed in the protein and kept by VdW interactions.

The water contribution is also very important. The exposure of each residue to water determines what kind of interaction is more important for the structure stabilization. Polar residues tend to be found on the surface of proteins and this is the reason why the polar interaction contribution, such as hydrogen bonds and Coulomb interaction, needs to be pondered by the water access.

Water access is also essential for protein dynamics. Polar side chains on the surface of globular proteins have structures that fluctuates among several conformational states. On the other hand, polar side chains, which are packed in the protein core, strongly contributes to the stabilization of the protein structure and are subject to restricted motions and well-defined configurations. The limited access of water increases the interaction energy of intermolecular hydrogen bonds and salt bridges. Apolar side chains in the protein core are packed with restricted motions due to the VdW interactions.

Apolar side chains on the surface of a protein are exposed to water. Any exposure of apolar surface to water leads to entropic penalties due to the super-organization of the water molecules. In order to avoid the entropic penalty, the protein tends to find an alternate organization where the apolar surface is hidden from water. Proteins that contain hydrophobic patches are less soluble in water and/or tend to oligomerize or interact with other proteins.

18.4 Protein Geometry and Introduction to Molecular Dynamics Simulation

It is not our goal to describe all the details of protein geometry and molecular dynamics simulation, but rather emphasize some of its aspects, that are important in the context of protein structural calculation.

Nowadays, we have the possibility of considering the knowledge accumulated over the last century on molecular structures, and particularly the knowledge about protein structure. Force fields are generally based on simplified versions of the classical mechanical equations that can be defined for each geometry element in the molecule and by each interaction force. The creation, and the continuous improvement, of the force fields enables the simulation of the protein geometry, of the intra- and inter-molecular interactions, and of the protein dynamics.

Simulations of molecular dynamics can be performed by solving Newton's equation in discrete time steps (known as integration time). The time step must be small enough to not overcome any polypeptide dynamic event, such as vibrations. Typically, the time step is smaller than 5 femtoseconds (fs, that is 5×10^{-15}s). The mass of each atom, the equilibrium distances, and angles are parameterized in the available force fields [9, 32, 65].

To compute the trajectory at each integration time (dt), the motion equations are obtained using Newton's second law ($\mathbf{F} = m\mathbf{a}$). The resulting external forces can be written as the gradient of the potential energy:

$$\mathbf{F} = -\nabla V. \tag{18.1}$$

The gradient (∇) is a vector operator that, when applied on a function, such as $V(x,y,z)$, results in a vector $\mathbf{F}$:

$$\nabla V(x,y,z) = \frac{\partial V}{\partial x}\mathbf{e_x} + \frac{\partial V}{\partial y}\mathbf{e_y} + \frac{\partial V}{\partial z}\mathbf{e_z}.$$

The combination of the equations above results in a differential equation that is integrated at each time step in order to obtain the trajectory of motion:

$$\nabla V = -m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}. \tag{18.2}$$

Note that the vector force $\mathbf{F}$ is obtained for a potential field $V(x,y,z)$. This is the reason why the set of parameters is also called "force field". The protein structure geometry is defined in the force field by the bond lengths, the bond angles, and by the proper and improper dihedral angles. The nonbonded intramolecular interaction is defined by the nonbonded potential, which mainly considers Coulomb and VdW interactions:

$$V_{\text{total}} = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} + V_{\text{impropers}} + V_{\text{nonbonded}}.$$

The intramolecular interactions with the solvent are also defined by nonbonded terms. The bond and angle potentials are harmonic potentials that model the vibration motion according to Hooke's law:

$$V_{\text{bonds}} = \sum_{\text{bonds}} K_{\text{b}}(r - r_0)^2,$$

$$V_{\text{angles}} = \sum_{\text{angles}} K_{\theta}(\theta - \theta_0)^2,$$

where K_{b} and K_{θ} are spring constants for bonds and angles, respectively. r is the generic bond length, while r_0 is bond length at equilibrium. Similarly, θ is the generic bond angle, whereas θ_0 is the bond angle at equilibrium.

A proper dihedral defines torsion angles which are formed by four atoms joined contiguously through bonds. It defines the geometry of real dihedrals of the protein. Improper dihedrals define the planarity of aromatic rings and peptide bonds, and they avoid stereo centers to interconvert. They also express torsion angles formed by atoms that are not necessarily connected through bonds. Proper dihedrals are usually expressed as periodic potentials:

$$V_{\text{dihedrals}} = \sum_{\text{dihedrals}} K_{\omega}\left[1 + \cos(n\omega - \gamma)\right],$$

$$V_{\text{impropers}} = \sum_{\text{impropers}} \frac{1}{2} K_{\xi}\left[\xi_{ijkl} - \xi_0\right].$$

K_{ω} and K_{ξ} are force constants. ω is the proper dihedral angle and γ is a phase of the periodic potential. ξ_{ijkl} is the generic improper dihedral angle and ξ_0 is the improper dihedral at equilibrium.

The nonbonded potentials are defined as following (the first term represents the Coloumb forces, while the second one represents the VdW forces):

$$V_{\text{nonbonded}} = \sum_{i,jpairs} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{i,jpairs} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right),$$

where q_i is the charge of the atom, ε is the electrical permittivity constant, r_{ij} is the distance between the two atoms i and j, and A_{ij} and B_{ij} are two constants related to the Lennard-Jones potential, modeling the VdW forces. We remark that other potentials, modeling, for example, the hydrogen bonds, can also be defined in force fields.

Tables 18.1 and 18.2 show, as an example, the force field and the topology implemented in the `XPLOR-NIH` and `CNS`.

Note that the topology of each amino acid (we consider the serine in the tables) is defined by the atomic weight, by the charge, and by the covalent connection of each atom. The force field is defined by the parameters (bond, angle, proper and improper dihedrals, and nonbonded interaction) that enables the calculation of all the potentials listed above.

Table 18.1 Selected parts of the topology table used by XPLOR-NIH and CNS. We consider the topology of the serine and report the atom type, charge, bonds description and atoms involved in proper and improper dihedral angle definitions. Note that the improper torsion angles define chirality and stereoisomery of the aminoacid.

atoms			bonds	
atom	type	charge	atom1	atom2
N	NH1	−0.36	N	HN
HN	H	0.26	N	CA
CA	CH1E	0.00	CA	HA
HA	HA	0.10	CA	CB
CB	CH2E	0.08	CB	HB1
HB1	HA	0.10	CB	HB2
HB2	HA	0.10	CB	OG
OG	OH1	−0.68	OG	HG
HG	H	0.40	O	C
C	C	0.48	C	CA
O	O	−0.48		

angles						
improper	HA	N	C	CB	*chirality*	CA
improper	HB1	HB2	CA	OG	*stereo*	CB
dihedral	OG	CB	CA	N	–	–

Table 18.2 Selected parts of the PARALLHDG force field (`parallhdg5.1.param`) [45]

```
BOND C CH1E 1000.000 sd = 0.001 1.525
BOND C CH2E 1000.000 sd = 0.001 1.516
BOND C CH2G 1000.000 sd = 0.001 1.516
...
ANGLe C CH1E CH1E 500.00 sd = 0.031 109.0754
ANGLe C CH1E CH2E 500.00 sd = 0.031 110.1094
ANGLe C CH1E CH3E 500.00 sd = 0.031 110.4838
...
IMPRoper C CH1E HA HA 500.00 sd = 0.031 0 -70.4072
IMPRoper C CH1E N CH1E 500.00 sd = 0.031 0 -179.9829
IMPRoper C CH1E NH1 CH1E 500.00 sd = 0.031 0 180.0000
...
DIHEdral C CH2E CH2E CH1E 5.00 sd = 0.031 3 0.0000
DIHEdral CH1E CH1E CH2E CH3E 5.00 sd = 0.031 3 0.0000
DIHEdral CH1E CH2E CH2E CH2E 5.00 sd = 0.031 3 0.0000
...
NONBonded HA 0.0498 1.4254 0.0450 2.6157 !- charged group.
NONBonded HC 0.0498 1.0691 0.0498 1.0691 ! Reduced vdw radius
NONBonded C 0.1200 3.7418 0.1000 3.3854 ! carbonyl carbon
```

We report the CNS force field description for bonds, angles, and dihedrals, where the atom type, the type of potential, and the spring constant are given [8, 18]

In the next section, we briefly describe conceptual aspects of NMR that help in understanding how to use and convert NMR experimental data in distance restraints. NMR and molecular dynamics simulation, along with other computational methods, can be considered as good partners, in the sense they are complimentary. NMR experiments provide essential structural and dynamical information for parameterization and improvements of the computational methods, while the computational methods provide a unique way to interpret the experimental data.

18.5 Introduction to Nuclear Magnetic Resonance

NMR is a spectroscopy that deals with the nuclear spin and its interaction with magnetic field. Several nuclei are magnetically active, in the sense that they have an associated magnetic moment. Among the magnetically active nuclei, ^{1}H, ^{13}C, and ^{15}N are the most important probes for protein NMR (see Table 18.3). A small protein containing about 100 amino acids approximately contains 2,000 hydrogens, 500 carbons, and 130 nitrogens. Each of these nuclei can be unambiguously assigned, providing precious information. The main physical properties obtained from NMR experiments are chemical shift, scalar coupling, and dipolar interaction (from dipolar coupling).

In practice, proteins prepared for structural determination are enriched with the nuclei presented in Table 18.3. To this purpose, the protein is biosynthesized by a bacterium (among other cells) and grown in an isotope-labeled medium [20].

The magnetism is a consequence of the spin angular momentum. Nuclear magnetism is caused by the nuclear spin. Magnetic active nuclei has a magnetic moment μ, which is associated to the nucleus that is described by the nuclear spin angular momentum $\mathbf{I}$. They are collinear and proportional to each other:

$$\mu = \gamma\hbar\mathbf{I}.$$

The proportionality constant is the magnetogyric ratio γ multiplied by the Planck constant $\hbar = h/2\pi$. See Table 18.3.

The nuclear spin angular momentum, the vector $\mathbf{I}$, has the following magnitude:

$$|\mathbf{I}^2| = \mathbf{I}\cdot\mathbf{I} = \hbar^2\left[I(I+1)\right],$$

where I is the spin angular momentum quantum number.

The spin is a quantum entity without classical analog. Nevertheless, it is useful to use a semiclassical representation based on classical angular momentum to build up a geometric representation of the spin (see Fig. 18.3).

Only one component of the angular momentum $\mathbf{I}$, I_x, I_y, or I_z, can be determined simultaneously with its magnitude $|\mathbf{I}^2|$. By convention, the value of the z component I_z is specified by the equation

$$I_z = \hbar m,$$

Table 18.3 Physical properties of some magnetically active nuclei commonly used in protein NMR

Nucleus	Nuclear spin quantum number (I)	Magnetogyric ratio	Natural abundance
^{1}H	1/2	267.513	100%
^{13}C	1/2	67.262	1%
^{15}N	1/2	27.116	0.377%
^{31}P	1/2	108.291	100%

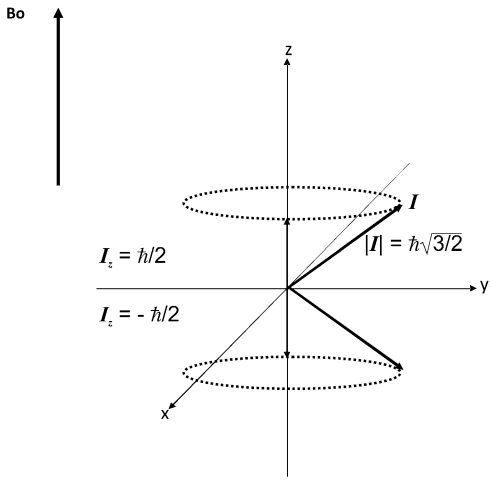

Fig. 18.3 A schematic representation of the angular momentum of nuclei with nuclear spin angular momentum $I = 1/2$. The vector **I**, in *black*, shows the two quantum states, while the vectors in *grey* represent its projection on the z axis. The projection on z can be determined when there is uncertainty in the projection on the xy plane. The uncertainty is represented by the *dotted grey line*. It implies that **I** can be projected in any position of the xy plane

where m is the magnetic quantum number that can have the following values:

$$m \in \{-I, -I+1, -I+2, \ldots, I-2, I-1, I\}.$$

For a nucleus with $I = 1/2$, **I** adopts two orientation. There is certainty in the projection I_z and uncertainty in I_x and I_y. I_z can be either in $+z$ ($m = 1/2$) or in $-z$ ($m = -1/2$). The magnitude of **I** and I_z are

$$|\mathbf{I}| = \frac{\hbar\sqrt{3}}{2}, \quad I_z = \frac{\hbar}{2}, \quad I_z = -\frac{\hbar}{2}.$$

The energy of the interaction of the magnetic moment ($\mu = \gamma\mathbf{I}$) in the presence of an external static magnetic field (**B**) is proportional to the scalar product of μ and **B**:

$$E = -\mu \cdot \mathbf{B}.$$

Both are vectorial quantities and the energy is dependent on the relative orientation of these two vectors. Figure 18.4 shows the energy diagram for a spin $I = 1/2$.

When field **B** is applied along the z direction, the energy becomes

$$E = -\gamma Bo I_z = -m\gamma\hbar Bo,$$

where Bo is the magnitude of the magnetic field **B** along z direction. So, for a spin $I = 1/2$:

- The quantum state $m = 1/2$, which is parallel with Bo, is the minimum energy state ($|\alpha>$ state) with $E = -\gamma\hbar Bo/2$.
- The quantum state $m = -1/2$, which is antiparallel with Bo, is the maximum energy state ($|\beta>$ state) with $E = \gamma\hbar Bo/2$.

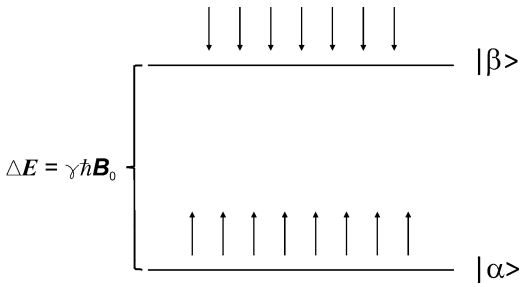

Fig. 18.4 Diagram representing the energy levels of a nuclear spin $I = 1/2$. Note that, at equilibrium, the high-energy level is less populated than the low-energy one. The *arrows* represent the projections on z (*up* and *down*). The *up arrows* indicate spins at the lower-energy state (the z-projection is parallel to the main magnetic field) while the *down arrows* are antiparallel to the static magnetic field (high-energy state)

The difference in energy is

$$\Delta E = \hbar \gamma Bo.$$

Note that the energy difference is proportional to Bo. The energy states are degenerate ($\Delta E = 0$) in absence of the magnetic field.

We have so far discussed about isolated spins only. For an ensemble of spins, we need to consider the vectorial sum of the magnetic moment for each spin in the ensemble. In an ensemble, the x and y components of the magnetic moment are canceled. At thermal equilibrium, the lowest energy state is more important: following a Boltzmann distribution as $\Delta E > 0$ in presence of a static magnetic field. This gives rise to a macroscopic magnetic component along the z axis that is the result of the sum over all spins of the ensemble. This is called magnetization vector **M** (see Fig. 18.5). Note that **M** is zero in absence of an external magnetic field and gets polarized ($|\mathbf{M}| > 0$) in presence of the magnetic field.

The NMR experiment consists of applying a radiofrequency pulse with one quantum of energy ($\Delta E = \hbar\omega = \hbar\gamma Bo$) and consequently of changing the population balance of the energy states. The magnetic component of the radiofrequency pulse $\mathbf{B_1}$ is applied on the xy plane. Figure 18.5 illustrates the magnetic component of the pulse causing the nutation of **M** at the rotating frame. Nutation consists of the evolution of **M** around $\mathbf{B_1}$.

The energy of the radiofrequency pulse is

$$E_{\mathrm{rf}} = \hbar\omega_0.$$

The resonance condition is

$$E_{\mathrm{rf}} = \Delta E \Rightarrow \hbar\omega_0 = \hbar\gamma Bo \Rightarrow \omega_0 = \gamma Bo,$$

where ω_0 is the Larmor frequency.

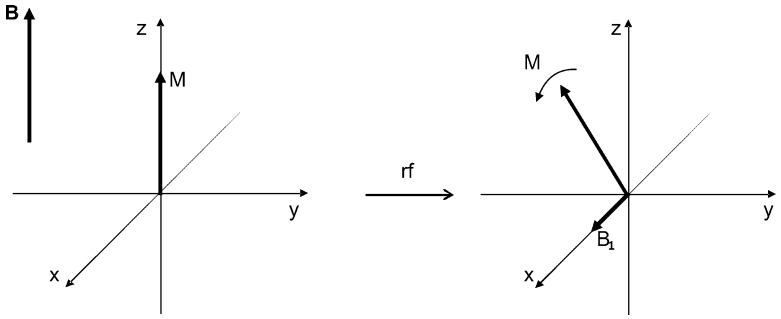

Fig. 18.5 Effect of a radiofrequency pulse represented by its magnetic component $\mathbf{B_1}$ on the magnetization vector **M**. The figure illustrates the nutation of the **M** around $\mathbf{B_1}$ at the rotating frame. Since the pulse is applied in x, the nutation occurs in the zy plane. At the laboratory frame $\mathbf{B_1}$ rotates in the xy plane at the frequency of the applied pulse. The rotating frame is a frame of reference that rotates around the z axis at the same frequency of the applied rf pulse (ω_0). At the rotating frame $\mathbf{B_1}$ is static

The nutation angle of the magnetization is controlled by the rf irradiation time. The spectroscopist calibrates the time necessary for nutating the magnetization at 90° ($M_z = 0, M_{xy} = 1$) or at 180° ($M_z = -1, M_{xy} = 0$), or at any other nutation angle. The calibrated pulse width is then used to set up the pulse sequences necessary for data collection for structure determination.

After excitation with the rf pulse, the transmitter is turned off. The magnetization is free to evolve back to equilibrium, precessing at the Larmor frequency around *Bo*. The frequency of evolution is detected by the receiver, transformed from time to frequency domain by a Fourier transform, which generates the NMR spectrum. Each spin in the ensemble displays in the spectrum. The NMR spectrum contains information of each spin present in the sample (see Fig. 18.6).

The differences in the electronic density in different molecules or parts of those structures cause the magnetic field to vary on a submolecular distance scale. This effect is called chemical shift and is extremely important for the application of NMR spectroscopy to study the molecules. In order to understand this effect, it is important to know how the electronic density of a molecule responds to the application of a static field **B**.

As showed in Fig. 18.7, the mechanism that leads to chemical shift can be simplified in a two-step process:

1. The external magnetic field induces currents in the electron clouds of the molecule.
2. These generated currents induce a magnetic field which can be added vectorially to the static field $\mathbf{B}^{\text{ind}}$:

$$\mathbf{B}^{\mathbf{loc}} = \mathbf{B} + \mathbf{B}^{\text{ind}}.$$

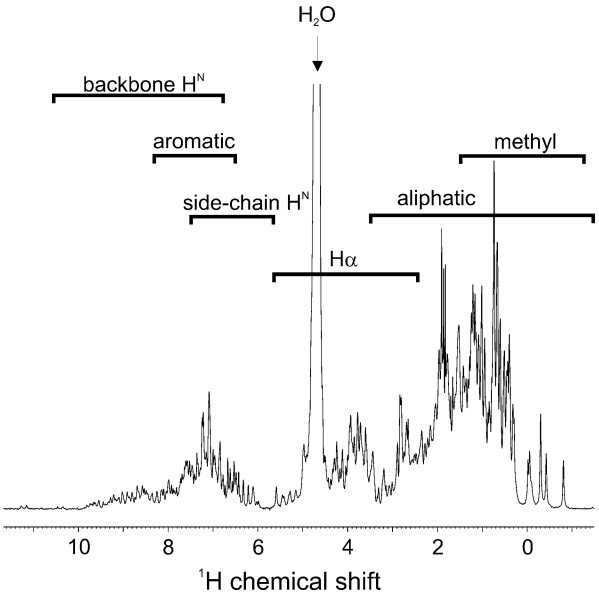

Fig. 18.6 Typical NMR spectrum of a protein. Ranges of chemical shifts expected for the various types of ^{1}H resonances

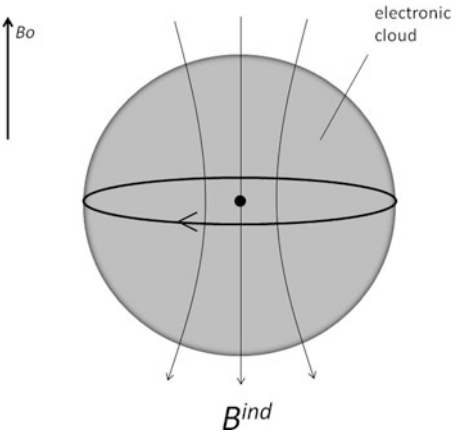

Fig. 18.7 A schematic representation of an atom, which illustrates the nucleus and the effect of the rotation of the electrons inducing a magnetic field B^{ind} which is antiparallel to the static magnetic field

Some important information about $\mathbf{B}^{\text{ind}}$ follows. First, the induced field is approximately linearly dependent on the applied field. Second, the magnitude and direction of some induced magnetic field is dependent on the shape of the molecule and on the location of the nuclear spin in the protein. Assuming these facts, we can write the induced magnetic field as follows:

$$\mathbf{B}^{\text{ind}} = -\boldsymbol{\sigma} \cdot \mathbf{B},$$

where $\boldsymbol{\sigma}$ is called shielding tensor, represented by a 3×3 square matrix. Note that $\boldsymbol{\sigma}$ is not a vector.

18.6 Experimental Restraints Generated by NMR

The main information for protein structural calculation is the nuclear Overhauser effect (NOE). NOE was first observed by Albert Overhauser in 1953 [54]. As previously observed, ensembles of spins get polarized in the presence of an external magnetic field. When two or more spins are near in space, only few angstroms apart, they become coupled (dipolar coupling). Under this condition, they can exchange polarization, affecting the intensities of the resonances of each of the spins. The dipolar coupled spins do not relax independently. The polarization transfer occurs via auto-relaxation but also through cross-relaxation.

Cross-relaxation mix populations between the two spins. The NOE is used to correlate spins through space [36]. The pulse sequence Nuclear Overhauser Effect SpectroscopY (NOESY) is the most important source of restraints [76]. The cross-peaks in a NOESY spectrum provide the distance information between two hydrogens in a protein. The intensity of the NOE cross-peak (I_{NOE}) is proportional to the distance between two hydrogens (the atoms i and j) and depends on the cross-relaxation rate:

$$I_{\mathrm{NOE}} = \alpha \frac{1}{\langle D_{ij} \rangle^6},$$

where α is the proportionality constant and $\langle D_{ij} \rangle$ is the time averaged distance between the two hydrogens. Note that the intensity drops with the sixth power of the distance. Only distances smaller than 6 Å can be therefore measured.

The parameter α contains information on the dynamics of the system ($\alpha = f(\tau)$). τ is the effective correlation time of the nuclei and contains the information about the internal dynamics of each hydrogen, as well as the global dynamics of the protein, such as the overall rotational correlation time. τ cannot be quantitatively treated for each individual hydrogen, and, thus, the NOE information is used in a semiquantitative way. Instead of giving exact distance information, NOEs give ranges of distance, i.e., a lower and upper bound on the actual distance.

There are methods following the local dynamics using a relaxation matrix. These methods provide better-quality distance information, but they still give only time-averaged distances [46].

The step of transforming the NOE intensities into ranges of distances is known as calibration. There are several ways to calibrate NOEs. The most frequent way is to use NOE intensities (or volumes) of hydrogen pairs of known secondary structure elements. The distances of those pairs are indeed well known. One can calculate a certain parameter on the basis of these distances and use the same parameter for all NOEs. This method is the most used for initial protein calculation.

A different NOE calibration method can be used during refinements. At this stage of protein calculations, the structure is already known. Thus, the distances extracted from the structures can be used for NOE calibration.

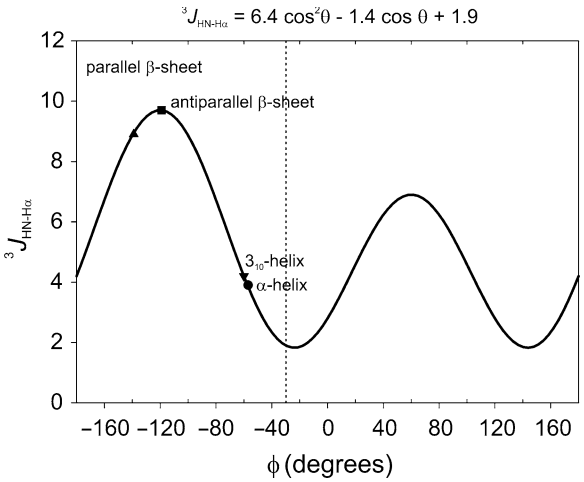

Fig. 18.8 Karplus Plot of $^3J_{H_N-H_\alpha}$ (in Hz) versus the torsion angle Φ. The *grey solid curve* is the best fit of equation parameters (*top of the figure*) where $\theta = |\Phi - 60|$. Values for regular secondary structures are indicated for α-helix (*circle* at $-57°$, 3.9 Hz), 3_{10} helix (*inverted triangle* at $-60°$, 4.2 Hz), antiparallel β-sheet (*square* at $-139°$, 8.9 Hz), and antiparallel β-sheet (*triangle* at $-119°$, 9.7 Hz) [55]. The region on the *left* delimited by the *dotted green line* ($-30°, -180°$) concentrates dihedral angles (Φ) of all amino acids (exception made for the glycines) [75]

18.6.1 Scalar Coupling (J)

The other source of information in the NMR experiments is the scalar couplings (J). Differently from the dipolar coupling that occurs through space, the scalar coupling occurs through bonds. J coupling can be through one, two, or three bonds (^1J, ^2J, ^3J). One-bond J coupling are typically heteronuclear, such as the coupling between amidic nitrogen and hydrogen ($^1J_{^{15}N-^1H}$). Two-bond J coupling occurs between geminal hydrogens, such as CH_2.

Finally, three-bond J coupling are the most important for structural information. Their value gives information about dihedral angles. For instance, the coupling between the amidic hydrogen and alpha hydrogen ($^3J_{H_N-H_\alpha}$) depends on the Φ angle of the Ramachandran plot. Figure 18.8 shows the Karplus relation [33] of the dependence of $^3J_{H_N-H_\alpha}$ with Φ. There are several NMR experiments designed to measure several dihedrals of a protein.

18.6.2 Chemical Shift

As previously shown, chemical shifts are dependent on the microenvironment. They are very sensitive to small changes. A correlation between chemical shifts of

Table 18.4 The correlation between chemical shifts and secondary structures of proteins

Residue	Random coil value of chemical shift (rc, ppm)			
	C′	Ha	CA	CB
	Condition to assign a secondary structure			
	α-helix $->$ rc $+ 0.5$ β-sheet $-<$ rc $- 0.5$	α-helix $->$ rc $- 0.1$ β-sheet $-<$ rc $+ 0.1$	α-helix $->$ rc $+ 0.7$ β-sheet $-<$ rc $- 0.7$	α-helix $->$ rc $+ 0.7$ β-sheet $-<$ rc $- 0.7$
Ala	177.1	4.19	52.5	19
Cys	174.8	4.52	58.3	28.6
Asp	177.2	4.63	54.1	40.8
Glu	176.1	4.24	56.7	29.7
Phe	175.8	4.42	57.9	39.3
Gly	173.6	4.11	45	0
His	175.1	4.59	55.8	32
Ile	176.9	4.09	62.6	37.5
Lys	176.5	4.23	56.7	32.3
Leu	177.1	4.35	55.7	41.9
Met	175.8	4.32	56.6	32.8
Asn	175.1	4.62	53.6	39
Pro	176	4.33	62.9	31.7
Gln	176.3	4.28	56.2	30.1
Arg	176.5	4.32	56.3	30.3
Ser	173.7	4.38	58.3	62.7
Thr	175.2	4.37	63.1	68.1
Val	177.1	4.11	63	31.7
Trp	175.8	4.42	57.8	28.3
Tyr	175.7	4.43	58.6	38.7

The random coil (rc) chemical shift value for each nuclei is presented for each amino acid residue The condition for assigning a secondary structure element on the basis of the chemical shift is given for each nucleus

hydrogen alpha (H_α), carbon alpha ($^{13}C_\alpha$), carbon beta ($^{13}C_\beta$), and the carbonyl ($^{13}C'$) and the secondary structure has been established. It consists in a very important structural information, because after resonance assignments of a protein, it becomes straightforward to determine its secondary structure elements based solely on chemical shifts. Table 18.4 summarizes the correlation between each of the nuclei and the chemical shift.

18.6.3 Residual Dipolar Couplings

As previously observed, the dipolar coupling is responsible for the mechanism of polarization transfer through cross-relaxation, which leads to the NOEs. However, dipolar couplings cannot be measured in the NMR spectra because of the isotropic molecular tumbling.

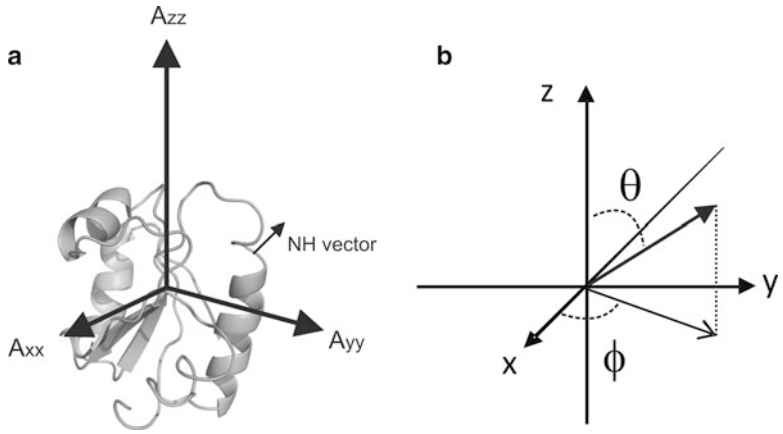

Fig. 18.9 (**a**) A protein structure (yeast thioredoxin, PDB id: 2I9H) showing the calculated molecular alignment tensors A_{xx}, A_{yy}, A_{zz}, as well as the representation of a dipolar vector (the NH vector in this case). By definition, $A_{zz} > A_{yy} > A_{xx}$. The principal molecular alignment tensor is therefore Azz. (**b**) Representation of the dipolar vector (the NH vector) in the molecular orientation frame of reference

In the 1990s, Prestegards and collaborators solubilized proteins in anisotropic media and showed that the residual orientation of the protein was able to recover dipolar coupling information. Anisotropic media consist of colloidal phases, such as bicelles and liquid crystals, or bacteriophages, such as Pf1, which are spontaneously oriented in the magnetic field. They restrict the Brownian motion of proteins in a way that induces a residual orientation due to the intrinsic anisotropic shape of the protein (see Fig. 18.9). Still the proteins keep tumbling fast, maintaining all the good behavior in of sharp lines, necessary for solution NMR.

Still the proteins keep tumbling fast in solution, maintaining all the good-behavior in solution of sharp lines, necessary for solution NMR. The residual orientation induces the reappearance of the dipolar coupling in solution. The residual dipolar coupling constant depends on the degree of orientation of the protein in the anisotropic media. The spectroscopist is able to tune the line shape and the degree of orientation, changing the concentration and other properties of the anisotropic media.

Dipolar coupling depends on the angle between the dipolar vectors with the main static magnetic field. This is true for a static oriented sample. Proteins dissolved in anisotropic media are not static. In this case, the residual dipolar coupling (RDC) does not depend directly on the angle of the dipolar vector with the static magnetic field, but RDCs are the measure of the angle of the dipolar vector with the principal molecular alignment tensor.

The principal molecular alignment tensors can be measured experimentally and also calculated from the molecule shape (Fig. 18.9). Thus, RDCs can be considered as an experimental restraint. This is a good quality restraint because it is a long-range angular restraint. RDCs have been used extensively as a refinement tool and their use allows for improving the geometric quality of the structures [44].

18.7 Experimental Pseudo-potentials

We introduce in this section some experimental pseudo-potentials based on the information obtained by NMR experiments. We describe NOEs as distance restraints, scalar coupling and chemical shifts as short-range angular restraints (proper dihedrals), and RDCs as long-range angular restraints. There are other sources of restraints that we do not discuss here: paramagnetic restraint, which are long-range distance restraints [15], chemical shift anisotropy restraints [42, 43, 74], among others.

The general strategy is to transform the experimental information into pseudo-potentials that can be used in the structural calculations. Next, we describe some pseudo-potential for each information obtained experimentally.

18.7.1 NOEs: Distance Restraints

After NOE calibration, the list of NOEs serves as an input for structural calculation. The NOE assignment list contains the specification of the hydrogen pair and the distance information, determining a lower (L_{ij}) and upper bound distances (U_{ij}). The lower bound is approximately 1.8 Å, which is the shortest possible distance between two hydrogens, accordingly to their atomic VdW radii. The upper bound distance depends on the target distance calculated from NOE calibration. Typically the distance restraints are assigned in classes: weak ($U_{ij} = 6$ Å), medium ($U_{ij} = 3.4$ Å), and strong ($U_{ij} = 2.8$ Å). The interval for each class is somewhat arbitrary and can vary from author to author.

Quadratic Pseudo-potential The pseudo-potential for NOE can be defined as follows. It gives no energy penalty when the distance between the two hydrogens (i and j) is contained in the interval $[L_{ij}, U_{ij}]$. The potential increases quadratically when r does not belong to the given interval:

$$V_{ij} = \begin{cases} C_1(r - L_{ij})^2, & \text{if } r < L_{ij} \\ 0, & \text{if } L_{ij} < r < U_{ij} \\ C_2(r - U_{ij})^2, & \text{if } r > U_{ij}, \end{cases} \tag{18.3}$$

where C_1 and C_2 are force constants that control the steepness of the energy pseudo-potential.

Biharmonic Pseudo-potential The pseudo-potential for NOE can also be defined as a function of a unique target distance D_{ij} that can be calibrated from NOE intensities. In this case, the pseudo-potential is defined as follows:

$$V_{ij} = \begin{cases} C_1(r - D_{ij})^2, & \text{if } r > D_{ij} \\ C_2(r - D_{ij})^2, & \text{if } r < D_{ij}, \end{cases}$$

where C_1 and C_2 are force constants that are weighed by the thermal energy (K_bT) available in the computational system:

$$C_1 = S_1 \frac{K_bT}{2} \quad \text{and} \quad C_2 = S_2 \frac{K_bT}{2}$$

where K_b is the Boltzmann constant and T is the absolute temperature of the system. Note that the potential is not zero when r is within the interval defined by a lower and an upper bound. S_1 and S_2 are scale factors.

18.7.2 Dihedral Restraints

Dihedral restraints can be incorporated in the structural calculation. They are obtained from scalar coupling measurements and chemical shift information. For each dihedral restraint, we have the target dihedral θ_{target} and the permitted variation $\Delta\theta$, which is usually relatively large. This way, it allows the dihedral conformational space to vary freely within the low-energy Ramachandran area.

Pseudo-potential for dihedral angle is defined as follows:

$$V_{dihedral} = \begin{cases} C_1(\theta - \theta_{target})^2, & \text{if } \theta < \theta_{target} - \Delta\theta \\ 0, & \text{if } \theta_{target} - \Delta\theta < \theta < \theta_{target} + \Delta\theta \\ C_2(\theta - \theta_{target})^2, & \text{if } \theta > \theta_{target} + \Delta\theta, \end{cases}$$

where C_1 and C_2 are the two force constants.

18.7.3 Scalar J-Coupling Restraints

The pseudo-potential energy term for scalar coupling makes use of the Karplus relation. This equation uses the dihedral angle θ obtained at each time step of structure calculation to obtain the calculated scalar coupling ($J_{calculated}$).

$$J = A\cos^2(\theta + P) + B\cos(\theta + P) + C,$$

where A, B, and C are the Karplus coefficients and P is a phase. It then uses $J_{calculated}$ to create a pseudo-potential V_J by comparing it to the experimental J coupling ($J_{observed}$). The pseudo-potential is defined as follows:

$$V_J = C(J_{calculated} - J_{observed})^2,$$

where C is the force constant.

18.7.4 Chemical Shift Restraints

^{1}H and ^{13}C chemical shifts correlate with the angles Φ and Ψ and can define secondary structure elements. Several implementations on protein structural calculation include harmonic potentials for chemical shifts. The `X-PLOR-NIH` package for protein structural calculation [64] includes pseudo-potentials for C_α and C_β chemical shifts [37]. It also includes pseudo-potentials for non-exchangeable hydrogens. Chemical shifts are calculated on the basis of semiempirical methods, where random coil values, ring currents, magnetic anisotropy, and electric-field chemical shifts are considered. The experimental chemical shift is compared to the predicted one from the structure, and the pseudo-potential takes care of refining the structure to agree with chemical shifts [37, 38].

The most used strategy to take into account chemical shifts is through the prediction of the Φ and Ψ dihedral angles. The program `TALOS` [66] uses a combination of six chemical shifts information: δ_{H_N}, δ_{H_α}, δ_{C_α}, δ_{C_β}, $\delta_{C'}$, and δ_N. The program is based on a search on a database containing 200 high-resolution protein structures, containing sequence information, Φ and Ψ torsion angles, and chemical shift assignments. It looks for chemical shift similarities between a certain residue and the two adjacent residues (triplets of residues). It always uses triplets of residues to predict backbone torsion angles of a given residue. If there is a consensus of Φ and Ψ angles among the ten best database matches, then `TALOS` uses these database triplet structures to form a prediction for the backbone angles of the target residue.

Based on the matches, `TALOS` calculates a consensus for Φ and Ψ angles (Φ_{target} and Ψ_{target}). The values of Φ_{target} and $\Delta\Phi$ and Ψ_{target} and $\Delta\Psi$ are included as dihedral angle restraints. The accuracy of `TALOS` predictions is about 89%. Most of the errors occur in regions of the Ramachandran that does not define secondary structure elements. `TALOS` prediction can thus be used reliably for secondary structure elements.

18.7.5 Residual Dipolar Coupling Restraints

As observed before, partial orientation of macromolecules in anisotropic media allowed the detection of RDCs. RDCs are good quality restraints because they define angles between a bond vector and the principal molecular alignment tensor (see Fig. 18.9). In order to compute RDCs, it is necessary to use an external orientational axis that is the reference for the angle measurement between the bond vectors. The implementations of RDC pseudo-potentials in the program `Xplor-NIH` can take into account dipolar vectors between atoms that are directly bonded (such as N–H or C–H bonds), or more flexible situations where the dipolar vector is between atoms not directly bonded, such as ^{1}H-^{1}H dipolar couplings. ^{1}H–^{1}H dipolar couplings are more difficult since ^{1}H-^{1}H distances can vary. In this chapter, we describe only the directly bonded RDCs. For more detailed information on other implementations, the reader is referred to [1, 10–12, 49, 63, 70, 71].

A necessary step is the calculation from the structure of the rhombicity and of the amplitude of the molecular alignment tensor. This is accomplished from the shape of the molecule. The molecular alignment tensors from experimental RDC are obtained from the following equation:

$$\mathrm{RDC}(\theta,\Phi) = A_\mathrm{a}\left\{(3\cos^2\theta - 1) + \frac{3}{2}R(\sin^2\theta\cos^2\Phi)\right\},$$

where θ and Φ are the polar angles of the dipolar vector in the molecular frame of reference (see Fig. 18.9), the axial A_a and radial A_r components, and rhombicity R are defined as follows:

$$A_\mathrm{a} = \frac{1}{3}\left\{\frac{A_{zz} - (A_{yy} + A_{xx})}{2}\right\}, \quad A_\mathrm{r} = \frac{A_{xx} - A_{yy}}{3}, \quad R = \frac{A_\mathrm{r}}{A_\mathrm{a}}.$$

The pseudo-potential is defined as a quadratic harmonic potential:

$$V_\mathrm{RDC} = K_\mathrm{RDC}(\mathrm{RDC}_\mathrm{calculated} - \mathrm{RDC}_\mathrm{observed})^2.$$

More frequently, θ, the angle between the internuclear dipolar vector and the reference external vector, which represents A_{zz} in the calculation, is obtained with a good precision. The rhombic component is usually not precise enough to be used in the calculation. Thus, in practice, RDCs are able to define a cone with angle $\pm\theta$ around the principal component of the molecular axis. Of course, the lack of precision in Φ limits the restraining ability of RDCs.

So far, we provided a description of pseudo-potentials which are based on experimental restraints obtained by NMR. In the next sections, we describe some computational solutions for calculating protein structures by using the NMR experimental information.

18.8 Distance Geometry Methods

The most important aspect for protein structure determination by NMR is the exploration of the conformational space imposed by the experimental restraints. X-ray diffraction of a single crystal generates an electron density map, which directly provides structural information. In contraposition, NMR experimental restraints are not able to give structural information, but rather short-range distances and dihedral angles restraints. The result of such a calculation is not a single structure, as for X-ray diffraction, but a set of structures that are all able to satisfy the experimental restraints.

As discussed earlier, NMR experimental restraints consist of semi-quantitative short-range distances and angles information. The structural calculation uses ranges of distances and angles, rather than precise measurements. NMR distance and angle restraints provide upper and lower bounds for both distances and angles.

Ideally, the measurement of precise long-range (in the order of the radius of gyration) distances or angles generates higher-quality restraints. However, this kind of restraints is difficult to measure by NMR. RDCs are better-quality restraints because they give information about long-range angles, but their application is restricted. In fact, only θ angles can be measured with precision. Nevertheless, the inclusion of RDCs in the structure calculation has a dramatic effect on the geometric quality [68]. Recent advances in solid state NMR and paramagnetic relaxation enhancement experiments (PRE) in solution introduced some better-quality long-range distance restraints [21, 31, 40].

What makes structure determination by NMR possible is the fact that the number of short-range distance restraints is generally much larger than the degrees of freedom. There are two degrees of freedom per amino acid residue in the protein backbone (Φ and Ψ dihedral angles), and, typically, good NMR experiments are able to provide more than 15 short-range restraints per amino acid residue.

NMR structure determination is not a computationally simple problem. The lack of precise distances and angles avoid the solution by fast geometric algorithms [3–5]. The computational solution was the inclusion of an all-atom model with all the known protein geometric angle and distances information along with the semiquantitative short-range experimental information. This approach made it possible to obtain the structures of globular proteins.

In the following, we briefly introduce the computational tools that have been particularly conceived in order to tackle with the problem of exploring the whole conformational space imposed by the imprecise experimental restraints.

The most naive way to explore the whole conformational space is to build a systematic grid of potential conformations and exhaustively explore it. However, this method can be applied only to small peptides [67]. Later we consider again this idea in the context of torsion angle simulated annealing.

The problem of finding the structure of a molecule from some distance and angle restraints is known in the scientific literature as the (molecular) *distance geometry* problem. Many methods and algorithms have been developed over the past last years for an efficient solution of this problem. The first method for distance geometry dates back to the 1970s. The basic idea is to define a penalty function which is able to measure the satisfaction of the available restraints, and to optimize this penalty function. One of the advantages is that the minimum value of the penalty function (corresponding to the optimal structure satisfying all restraints) is known a priori, because, when the data are correct, it must be ideally zero. If there is no geometric solution with error near zero, it is a strong evidence of systematic errors in the experimental data [26, 27].

The first method for distance geometry makes use of the metric matrix $\mathbf{G}$, from which it is possible to obtain the Cartesian coordinates of the atoms of the molecule by exploiting the available set of distances between some pairs of atoms. The relation between the elements G_{ij} of the metric matrix $\mathbf{G}$ and the Cartesian coordinates of the two atoms i and j is given by

$$G_{ij} = \mathbf{r_i} \cdot \mathbf{r_j}. \tag{18.4}$$

In the matrix **G**, the diagonal elements are the squares of the Cartesian coordinates of the atom i, whereas the off-diagonal elements represent the projection of $\mathbf{r_i}$ over $\mathbf{r_j}$. The square of the Cartesian coordinates of the atom i can be viewed as an vector, defined by the position of i and the origin $(0,0)$. The diagonal elements can be seen the norm of the vector $\mathbf{r_i}$, which defines the position of each atom in relation to the origin.

As it is well known, the dot product can be written as

$$G_{ij} = |r_i||r_j|\cos\theta,$$

where θ is the angle between the two vectors. Such an angle is 0 for diagonal elements, nonzero for off-diagonal elements.

The metric matrix **G** is built by considering all $N \times N$ possible distances for the set of N atoms. The elements of the metric matrix are obtained through the relations

$$G_{ii} = \frac{1}{N}\sum_{j}^{N} D_{ij}^2 - \frac{1}{2N^2}\sum_{jk}^{N} D_{jk}^2, \quad G_{ij} = \frac{1}{2}\left(G_{ii} + G_{jj} - D_{ij}^2\right),$$

where D_{ij} is the distance between the atoms i and j, and N is the total number of atoms. The metric matrix is positive semi-definite and has rank 3. All eigenvalues are positive or zero and at most three eigenvalues are different from zero.

The general metric matrix decomposition equation is used for the diagonalization, which is necessary to find the coordinates of each atom:

$$G_{ij} = \sum_{\alpha=1}^{n} \lambda_\alpha E_i^\alpha E_j^\alpha. \tag{18.5}$$

E_i^α and E_j^α are the eigenvectors and λ_α is the eigenvalue of the matrix; n is the dimensionality of the system.

The combination of Eqs. (18.4) and (18.5) leads to the following equation, which enable the calculation of the three-dimensional coordinates of the points of the system from the metric matrix elements:

$$r_i^\alpha = \sqrt{\lambda_\alpha} E_i^\alpha.$$

It is implicit in the equations the assumption that every distance is referenced to the origin $(0,0)$. In general, one of the atoms, say the one labeled with 1, is set to the origin.

As discussed before, the distance information is generally given by a list of lower and upper bounds:

$$L_{ij} < D_{ij} < U_{ij}.$$

The basic steps of the first method for distance geometry are [25]:

1. *Bound smoothing*—consists of extrapolating the tightest possible bounds on the incomplete list of interatomic distances

2. *Metrization*—tries to find a matrix of exact values within the lower and upper bound
3. *Embedding*—computes the coordinates of all atoms of the protein
4. *Optimization*—minimizes the penalty function value, i.e., the measure of the violation of both lower and upper bounds on the distances, where some geometric constraints of proteins are also considered

We give the details of these four main steps in the following.

18.8.1 Bound Smoothing

Metric matrix distance geometry algorithms work with exact distances (derived from bond lengths and angles) and NMR experimental data, which are non-exact distances. In the first implementation of algorithms for distance geometry, the distances were chosen independently and randomly within the available lower and upper bounds.

Successively, a bound smoothing was developed for choosing better distances. The technique is based on the fact that interatomic distances always obey triangle inequalities. In fact, the triangle inequality theorem states that any side of a triangle is always shorter than the sum of the two other sides. For a triplet of atoms (i, j, k), it follows that

$$L_{ik} - U_{kj} \leq D_{ij} \leq U_{ik} + U_{kj}.$$

Note that triangle inequality theorem imposes some constraints on D_{ij}. Many algorithms for distance geometry consider these inequalities for all possible triplets (i, j, k) in order to obtain the so-called triangle inequalities bounds.

Another relation that could be used for bound smoothing is given by the tetrangle inequalities. The tetrangle inequality is similar to the triangle inequality, but it considers quadruplets of atoms, not triplets. It is able, in general, to provide tighter bounds on D_{ij}, but it is much more expensive from a computational point of view.

18.8.2 Metrization

The metrization procedure can be used to improve the geometrical consistency of the randomly chosen distances. We suppose that all distances were chosen from bounds previously processed by a bound smoothing technique (based on triangle and/or tetrangle inequalities). The metrization is based on the construction of distance matrices whose elements respect two rules:

1. Their lower and upper bounds satisfy the triangle and the tetrangle inequalities.
2. The chosen distances satisfy the triangle inequality.

The second rule ensures that later interatomic distance choices are consistent with earlier ones. The metrization imposes interdependency between the randomly chosen distances (they are, in fact, not completely independent to each other).

18.8.3 Embedding

The initial distances are chosen as an exact distance contained in the interval defined by the corresponding lower and upper bounds. The metric matrix is calculated, and it frequently results in a non-embeddable matrix in the three-dimensional space. This means that the matrix is not positive semidefinite, i.e., the solution is inconsistent with any conformation in the three-dimensional space.

The main aim is to identify an embeddable metric matrix in three dimensions. Within the bound distances, there is a metric matrix in which the absolute values of the three largest eigenvalues are positive, and their corresponding eigenvectors contain the Cartesian coordinates of the atoms of the molecule. If these values are not positive, the chosen distances are not consistent, and the embedding cannot be performed.

18.8.4 Optimization

This step consists in improving the quality of the protein structure found during the embedding. To this aim, a penalty function (measuring the violations of lower and upper bounds, as well as some geometrical deviations) is defined and optimized. This penalty function must obey to the following rules:

1. Must be nonnegative
2. Must be zero when all the geometric constraints are satisfied
3. Must be twice differentiable in its whole domain

An example of penalty function is

$$F(x) = \sum_{ij} A_{ij}^2(x) + \sum_{ij} B_{ij}^2(x) + \sum_{ijkm} C_{ijkm}^2(x),$$

where:

- $A_{ij}^2(x) = 0$ if and only if the distance between nonbonded pairs of atoms (i, j) is larger than their hard VdW sphere radii.
- $B_{ij}^2(x) = 0$ if and only if the distance between the pair of atoms (i, j) restrained by experimental data lies within the corresponding lower and upper bound.
- $C_{ijkm}^2(x) = 0$ if and only if the angle (i, j, k, l) respects the absolute chirality.

In order to minimize the penalty function, a conjugate gradient minimization method can be used. Different penalty functions have been defined in different distance geometry approaches [25].

18.8.5 Scaling

At the very end, the obtained protein structure can be scaled so that it represents a globular protein. To this purpose, the expected radius of gyration of the structure is calculated. This expected radius can be larger or smaller than the radius of gyration calculated from the embedded coordinates. Therefore, a scaling factor equal to the ratio between expected and actual radius of gyration is computed. The embedded coordinates are then multiplied by this factor, because it makes any successive regularizations easier to perform.

18.9 Simulated Annealing

18.9.1 SA in Cartesian Space

As discussed in the previous section, the first method for distance geometry problems arising in the molecular context makes use of gradient conjugate minimizations of a given penalty function. We remark that such penalty functions do not consider many molecular forces that are instead used in molecular dynamics simulations. As a consequence, structures obtained by this method can produce correct overall folds, but they have poor local geometry. It was realized then that these structure were a very good input for restrained molecular dynamics simulation.

The first approach using restrained molecular dynamics simulation was employed to refine structures calculated from distance matrix distance geometry. The group of Clore and Gronenborn [50–52] used a simulated annealing (SA) algorithm in order to find solutions for multiple variable systems. SA was derived from a metallurgic process where the system is heated at extremely high temperatures and let cooling down slowly. The simulation of this process could allow the atoms of a molecule to assume a low-energy configuration [35].

Standard molecular dynamics simulation force fields are built in order to reproduce the behavior of a molecular system in thermal equilibrium (constant temperatures). High-energy transitions such as cis/trans isomerization and steric hindrance cannot be surpassed using these force fields. For standard molecular dynamics simulation, the calculated structures do not change so much from their initial conformation, or they get stuck at a local minima. In order to partially solve the problem of sampling the conformational space given by the experimental restraints, a set of simplifications was proposed.

The first simplification consists in associating to every atom the same molecular weight (typically 100). This avoids high-frequency bond and angle vibrations, enabling a significant reduction in the number of thermalization steps. If the thermalization is too fast, with a reduced number of integration steps, then high-frequency vibrations, which affect mostly low atomic weight atoms such as hydrogens, can generate strong forces that could break covalent bonds. This simplification is especially important in the SA protocol, where the bath temperature increases up to 2,000 K and the thermalization is essential for the success of the process.

Another simplification is the turning off of attraction nonbonded interaction during the hot phase of SA. The Coulomb term is turned off and the van der Waals potential is replaced by the simplified term (REPEL) [51]:

$$F_{\text{REPEL}} = \begin{cases} 0, & \text{if } r \geq s.r_{\min} \\ k_{\text{rep}}(s^2 r_{\min}^2 - r^2), & \text{if } r < s.r_{\min}, \end{cases}$$

where the values of $r_{\min}$ are the standard values for van der Waals radii (defined in the force fields) [6]. The scale factor s is set to 1.0 in the hot phase and to 0.825 in the cooling phases. In REPEL, only the repulsive term of the Lennard-Jones potential is maintained, reducing in this way the computational cost. This allows for surpassing high-energy barriers, which are due, in many situations, to the attractive forces imposed by Coulomb and VdW interaction, which aid the conformational space sampling.

Additionally, the force field is modified by increasing the penalty for bond and angle geometry violations. Finally, the distance restraint quadratic potential [Eq. (18.3)] is replaced by a simplified linear term, where the penalties increase linearly with the distance restraint violation. It was shown that this modification allows for correcting faster the geometry of the molecule.

During SA, the weight of force field parameters is adjusted to favor the conformational sampling. A typical sequence of events in an SA protocol is showed in Fig. 18.10, where the distance restraint potential (NOE) is weighted high during all phases, while the Coulomb term is turned off.

This new method was included in the structure calculation program `XPLOR` [8], where a hybrid approach to distance geometry was implemented: both target function minimization and simulated annealing in the structure calculation.

The starting structure is calculated using the distance matrix distance geometry algorithm [73]. Successively, target function minimization is performed and finally a series of cycles of simulated annealing calculation are executed. It is common to compute hundreds of structures. However, only the 20 lower-energy structures are selected to represent the protein.

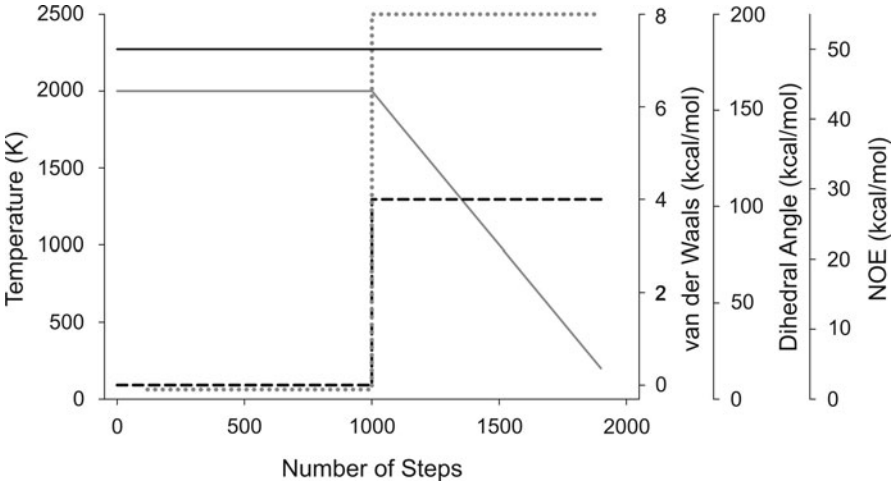

Fig. 18.10 Illustration of a typical Cartesian space SA protocol used for protein structure calculation. Scheduled changes in the parameter values are plotted as a function of the time steps. The bath temperature is represented as the *solid grey line*, the dihedral angle potential as a *dotted grey line*, the distance restraint potentials as *solid black lines*, and the VdW potential by the *dashed black lines* [7, 8]

18.9.2 SA in Torsion Angle Space

Molecular dynamics simulations (as well as SA) in the Cartesian space uses Newton mechanics at discrete time steps in order to describe the protein motion [Eqs. (18.1) and (18.2)]. Newton equations deduce the motion equations of a system from the knowledge of all external forces acting on it.

Another way to approach the molecular mechanics is by solving Lagrange equations. Lagrange mechanics uses scalar equations, which avoid the need to describe all the external forces that act on the system in a vectorial formalism. The Lagrangian function is defined by the difference among kinetic and potential energy:

$$L = T - V,$$

where T is the kinetic energy and V is the potential energy of the system. The motion equations are obtained from the Lagrangian function by the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0, \tag{18.6}$$

where the q_i's represent the coordinates of the system and $\dot{q}_i$ is the time derivative of the system coordinates (velocity). Note that this equation is not vectorial.

In order to illustrate the Lagrangian mechanics, we consider a simple system consisting of a linear spring-mass system on a frictionless table. The Lagrangian function becomes

$$L = T - V = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2,$$

where m is the mass, x is the linear coordinate, and k is the spring constant. The conservative system (18.6) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}(m\dot{x}) + kx = 0 \quad \Rightarrow \quad m\ddot{x} + kx = 0.$$

Note that the differentiation led to the equation of motion of the system in the same form as for the Newtonian formalism of classical mechanics, but without the need of figuring out all external vectorial forces on the system.

The same can be done for simulating the motions of a protein. The great advantage is that we can compute positions and the movements (acceleration) of the atoms by simplifying the coordinate system. The variables are only the torsion angles of a protein. The degree of freedom is decreased about tenfold, because the geometrical parameters, such as bond lengths, bond angles, and improper dihedrals (chirality and planarity), are fixed to their optimal values during the simulation.

As discussed before, what makes the search for conformational space by methods such as SA difficult is the rough energy landscape for a protein. There are many local minima to be avoided by computational methods. The strategies to reach the global minimum and avoid kinetic traps demand high computational time and special algorithms.

In Cartesian SA, much of the computational time is focused on calculations of geometrical parameters that almost do not change. The deviations from optimal geometry of bond lengths, bond angles, chirality, and planarity are small because they are parameterized to be as small as possible. In torsion angle dynamics, instead, these are fixed and so is the number of local minima. This is the main reason why torsion angle dynamics increase the efficiency of the search for conformational space imposed by the NMR experimental restraints.

The force field which is used in Cartesian dynamics considers strong potentials in order to keep the covalent structures. In torsion angle dynamics, the parameters are much simplified. One important aspect is that the time step for numerical integration in Cartesian dynamics must be very small ($<$5 fs), and there is therefore the risk of breaking some covalent structures because of bond and angle with high-frequency vibrations. In torsion angle dynamics, time steps can be three times longer because the covalent structures are fixed and such vibrations are inexistent.

In the implementation of torsion angle dynamics, the protein is described as a tree of rigid bodies connected by single bonds. The only degrees of freedom are rotations around the single bonds. The tree structure starts with a base, typically at the N-terminus and ends with the “leaves” that are the end of the side chains and the

C-terminus. The rigid bodies are labeled from 0 to n. The base is number 0 and each torsion angle is represented as θ_k, where $k \geq 1$. The conformation of the molecule can be uniquely specified by its torsion angle $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$.

The potential energy is defined as

$$V = \begin{cases} 0 & \text{if distances and angles are within the bounds and atoms are not overlapped} \\ V_{\text{target}} & \text{otherwise,} \end{cases}$$

where V_{target} is the target function that is dependent on the upper and lower bounds for the distance and on the angular restraints. ω_0 is a weighting factor. Note that $V > 0$ if the experimental bound are not satisfied or atoms are overlapped. Motion occurs when $V > 0$. The kinetic energy and the inertia tensor are calculated recursively at each time step of numerical integration. For details on the algorithms, see [22, 24].

The Lagrange equation takes the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\theta}_i}\right) - \frac{\partial L}{\partial \theta_i} = 0.$$

The differentiation leads to equation of motions that takes the form:

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) = 0, \tag{18.7}$$

where $M(\theta)$ is the mass matrix and $C(\theta, \dot{\theta})$ is a constant n-dimensional vector. Note that Eq. (18.7) was obtained by using a similar mathematical procedure presented in the simple system of linear spring-mass system in a frictionless table [Eq. (18.6)]. For a detailed description, see [22].

Torsion angle space SA is efficient for searching conformational space because it smoothes the protein energy landscape, avoiding local minima. It also enables the hot phase of SA at very high temperature, such as 50,000 K. However, we have to mention that it is a statistical method and there is no mathematical proof that the global minimum could be actually found.

The introduction of the torsion angle space SA solved the problem of searching the conformational space given by NMR experimental restraints. It is the most frequently used method, and it is implemented in all programs developed for structural determinations, such as `XPLOR-NIH`, `CNS`, and `CYANA`. The algorithm is very efficient and enables the calculation of a protein structure in minutes.

18.10 Future Demands for Protein Structure Determination

This present chapter showed the evolution of computational methods for protein structural determination using NMR experimental data. It is clear that structural determination by NMR does not rely on direct spatial data but on a set of small-range experimental distance and angle restraints that, combined with some structural

geometrical information on proteins, can be exploited for producing structural models. Over the years the NMR structures determined by the methods discussed in this chapter have been accepted by the scientific community as realistic and useful for studying biochemical mechanistic problems.

The torsion angle space SA protocols can be very efficient for searching the conformational space under the constraints given by NMR experiments. All semiautomated methods for structural determination, such as ARIA [62] and UNIO [17, 19, 28, 72], make use of torsion angle space SA. It is also implemented in the software tools for structure determination by NMR, such as XPLOR-NIH [64], CNS [7, 8], and CYANA [23, 24, 28, 47].

Although distance geometry combined with simulated annealing (DGSA) is not the most usual method, it offers many advantages: (1) distance geometry is not a statistical method and can offer mathematical proof that the global minimum has been achieved; (2) DGSA is as fast and efficient in the search of conformational space as it is torsion angle space simulated annealing; (3) since DGSA relies on a geometrical method for the search of conformational space, it can be used for large proteins and complexes. Statistical methods, on the other hand, can become inefficient when the size of the protein is large.

The increase in protein size also imposes a more restricted number of restraints. NMR spectroscopy can nowadays generate structural information for large proteins and protein complexes. However, such a structural information is sparse and new methods for structural calculation with sparse date are becoming increasingly important.

Standley [69] proposed in 1999 a branch-and-bound algorithm for protein refinements with sparse data. They used distance geometry methods to minimize an error function which is based on the experimental restraints, as well as a residue-based protein folding potential. This algorithm is able to identify more compact structures. The protein folding term is based on the idea of using long-range potentials so that the dependence of long-range distance restraints is reduced.

Dong and Wu [16] in 2003 introduced a geometrical method for solving NMR structure with sparse data. In general, NMR spectroscopy generates experimental data that are not complete. They have used geometrical information, in a similar way as bound smoothing and metrization uses triangle and tetrangle inequalities, to build up the "missing" information. The algorithm calculates the coordinates of a given atom on the basis of the coordinates of the previously computed atoms and of the distances between the current and the previous atoms. Some assumptions need to be satisfied in order to use this algorithm. Davis et al. [14] proposed an improved algorithm, called *revised updated geometric build-up algorithm* (RUGB), to build up missing information.

Liberti et al. [41] proposed the use of a discrete search occurring in continuous space for solving protein structure. The main idea is to use distance information between atoms that are contiguous (sequential) in order to discretize the search space (which has the structure of a tree), and to employ a branch-and-prune algorithm for solving the discretized problem. In the branch-and-prune, new candidate atomic positions are generated at each iteration (branching), and their feasibility is verified

immediately so that branches of the tree which do not contain solutions can be removed (pruning). The branch-and-prune can work with both exact or interval data [39] and also in the hypothesis in which only distances between hydrogen atoms are available [53].

In conclusion, structural determination using NMR experimental data needs the use of efficient computational methods. The continuous development of NMR and of computational methods can improve the quality, efficiency, and limits for structural determination by NMR.

Acknowledgments We are grateful to Prof. Antonio Mucherino and Prof. Carlile Lavor for the edition and revision of this chapter.

References

1. Bax, A., Kontaxis, G., Tjandra, N.: Dipolar couplings in macromolecular structure determination. Nucl. Magn. Reson. Biolog. Macromol., Pt B **339**, 127–174 (2001)
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. **28**, 235–242 (2000)
3. Braun, W.: Distance geometry and related methods for protein-structure determination from NMR data. Q. Rev. Biophys. **19**(3–4), 115–157 (1987)
4. Braun, W., Bösch, C., Brown, L.R., Go, N., Wüthrich, K.: Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. Biochimica Et Biophysica Acta **667**(2), 377–396 (1981)
5. Braun, W., Go, N.: Calculation of protein conformations by proton proton distance constraints – A new efficient algorithm. J. Mol. Biol. **186**(3), 611–626 (1985)
6. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M.: CHARMM – a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. **4**(2), 187–217 (1983)
7. Brünger, A.T.: Version 1.2 of the crystallography and NMR system. Nat. Protocol. **2**(11), 2728–2733 (2007)
8. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L.: Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallographica, Section D, Biological Crystallography **54**, 905–921 (1998)
9. Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M. Jr., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J.: The Amber biomolecular simulation programs. J. Comput. Chem. **26**(16), 1668–1688 (2005)
10. Clore, G.M., Gronenborn, A.M., Bax, A.: A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. J. Mag. Reson. **133**(1), 216–221 (1998)
11. Clore, G.M., Gronenborn, A.M., Tjandra, N.: Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. J. Mag. Reson. **131**(1), 159–162 (1998)
12. Clore, G.M., Starich, M.R., Bewley, C.A., Cai, M., Kuszewski, J.: Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. J. Am. Chem. Soc. **121**(27), 6513–6514 (1999)

13. Collet, J.F., Messens, J.: Structure, function, and mechanism of thioredoxin proteins. Antioxidants & Redox Signaling **13**(8), 1205–1216 (2010)
14. Davis, R.T., Ernst, C., Wu, D.: Protein structure determination via an efficient geometric build-up algorithm. BMC Struct. Biol. **10**(1):S7 (2010)
15. Donaldson, L.W., Skrynnikov, N.R., Choy, W.-Y., Muhandiram, D.R., Sarkar, B., Forman-Kay, J.D., Kay, L.E.: Structural characterization of proteins with an attached ATCUN motif by paramagnetic relaxation enhancement NMR spectroscopy. J. Am. Chem. Soc. **123**(40), 9843–9847 (2001)
16. Dong, Q., Wu, Z.: A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. J. Global Optim. **26**, 321–333 (2003)
17. Ellgaard, L., Bettendorff, P., Braun, D., Herrmann, T., Fiorito, F., Jelesarov, I., Guntert, P., Helenius, A., Wüthrich, K.: NMR structures of 36 and 73-residue fragments of the calreticulin P-domain. J. Mol. Biol. **322**(4), 773–784 (2002)
18. Engh, R.A., Huber, R.: Accurate bond and angle parameters for X-ray protein-structure refinement. Acta Crystallographica, Section A **47**, 392–400 (1991)
19. Fiorito, F., Herrmann, T., Damberger, F.F., Wüthrich, K.: Automated amino acid side-chain NMR assignment of proteins using (13)C- and (15)N-resolved 3D (1)H,(1)H -NOESY. J. Biomol. NMR **42**(1), 23–33 (2008)
20. Galvão-Botton, L.M.P., Katsuyama, A.M., Guzzo, C.R., Almeida, F.C., Farah, C.S., Valente, A.P.: High-throughput screening of structural proteomics targets using NMR. FEBS Lett. **552**(2–3), 207–213 (2003)
21. Gillespie, J.R., Shortle, D.: Characterization of long-range structure in the denatured state of staphylococcal nuclease; 2. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. J. Mol. Biol. **268**(1), 170–184 (1997)
22. Guntert, P.: Structure calculation of biological macromolecules from NMR data. Q. Rev. Biophys. **31**(2), 145–237 (1988)
23. Guntert, P., Braun, W., Wüthrich, K.: Efficient computation of 3-dimensional protein structures in solution from Nuclear-Magnetic-Resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J. Mol. Biol. **217**(3), 517–530 (1991)
24. Guntert, P., Mumenthaler, C., Wüthrich, K.: Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. **273**(1), 283–298 (1997)
25. Havel, T.F.: An evaluation of computational strategies for use in the determination of protein-structure from distance constraints obtained by Nuclear-Magnetic-Resonance. Progr. Biophys. Mol. Biol. **56**(1), 43–78 (1991)
26. Havel, T.F., Wüthrich, K.: A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular H-1-H-1 proximities in solution. Bull. Math. Biol. **46**(4), 673–698 (1984)
27. Havel, T.F., Wagner, G., Wüthrich, K.: Spatial structures for BPTI in the crystal and in solution from distance geometry calculations. Experientia **40**(6), 608–608 (1984)
28. Herrmann, T., Guntert, P., Wüthrich, K.: Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J. Biomol. NMR **24**(3), 171–189 (2002)
29. Holmgren, A.: Antioxidant function of thioredoxin and glutaredoxin systems. Antioxidants & Redox Signaling **2**(4), 811–820 (2000)
30. Holmgren, A., Bjornstedt, M.: Thioredoxin and thioredoxin reductase. Methods Enzymol **252**, 199–208 (1995)
31. Iwahara, J., Schwieters, C.D., Clore, G.M.: Ensemble approach for NMR structure refinement against H-1 paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. J. Am. Chem. Soc. **126**(18), 5879–5896 (2004)
32. Jorgensen, W.L., Maxwell, D.S., TiradoRives, J.: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. **118**(45), 11225–11236 (1996)
33. Karplus, M.: Vicinal proton coupling in nuclear magnetic resonance. J. Am. Chem. Soc. **85**(18), 2870–2871 (1963)

34. Kauzmann, W.: Three-dimensional structures of proteins. Biophys. J. **4**, 43–54 (1964)
35. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
36. Kumar, A., Wagner, G., Ernst, R.R., Wuethrich, K.: Buildup rates of the nuclear Overhauser effect measured by two-dimensional proton magnetic-resonance spectroscopy – Implications for studies of protein conformation. J. Am. Chem. Soc. **103**(13), 3654–3658 (1991)
37. Kuszewski, J., Qin, J., Gronenborn, A.M., Clore, G.M.: The impact of direct refinement against C-13(Alpha) and C-13(Beta) chemical-shifts on protein-structure determination by NMR. J. Mag. Reson. Ser. B, **106**(1), 92–96 (1995)
38. Kuszewski, J., Gronenborn, A.M., Clore, G.M.: Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci. **5**(6), 1067–1080 (1996)
39. Lavor, C., Liberti, L., Mucherino, A.: The interval Branch-and-Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances. to appear in J. Global Optim. (2012) DOI: 10.1007/s10898-011-9799-6
40. Liang, B.Y., Bushweller, J.H., Tamm, L.K.: Site-directed parallel spin-labeling and paramagnetic relaxation enhancement in structure determination of membrane proteins by solution NMR spectroscopy. J. Am. Chem. Soc. **128**(13), 4389–4397 (2006)
41. Liberti, L., Lavor, C., Mucherino, A., Maculan, N.: Molecular distance geometry methods: from continuous to discrete. Int. Trans. Oper. Res. **18**, 33–51 (2010)
42. Lipsitz, R.S., Tjandra, N.: Carbonyl CSA restraints from solution NMR for protein structure refinement. J. Am. Chem. Soc. **123**(44), 11065–11066 (2001)
43. Lipsitz, R.S., Tjandra, N.: N-15 chemical shift anisotropy in protein structure refinement and comparison with NH residual dipolar couplings. J. Mag. Reson. **164**(1), 171–176 (2003)
44. Lipsitz, R.S., Tjandra, N.: Residual dipolar couplings in NMR structure analysis. Ann. Rev. Biophys. Biomol. Struct. **33**, 387–413 (2004)
45. Linge, J.P., Nilges, M.: Influence of non-bonded parameters on the quality of NMR structures: A new force field for NMR structure calculation. J. Biomol. NMR **13**(1), 51–59 (1999)
46. Linge, J.P., Habeck, M., Rieping, W., Nilges, M.: Correction of spin diffusion during iterative automated NOE assignment. J. Mag. Reson. **167**(2), 334–342 (2004)
47. Lopez-Mendez, B., Guntert, P.: Automated protein structure determination from NMR spectra. J. Am. Chem. Soc. **128**(40), 13112–13122 (2006)
48. Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C.: Structure validation by C_α geometry: Φ, Ψ and C_β deviation. Protein. Struct. Funct. Genet. **50**(3), 437–450 (2003)
49. Meiler, J., Blomberg, N., Nilges, M., Griesinger, C.: A new approach for applying residual dipolar couplings as restraints in structure elucidation. J. Biomol. NMR **16**(3), 245–252 (2000)
50. Nilges, M., Clore, G.M., Gronenborn, A.M.: Determination of 3-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**(1), 129–136 (1988)
51. Nilges, M., Clore, G.M., Gronenborn, A.M.: Determination of 3-dimensional structures of proteins from interproton distance data by hybrid distance geometry – Dynamical simulated annealing calculations. FEBS Lett. **229**(2), 317–324 (1988)
52. Nilges, M., Gronenborn, A.M., Brünger, A.T., Clore, G.M.: Determination of 3-dimensional structures of proteins by simulated annealing with interproton distance restraints – Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor-2. Protein Eng. **2**(1), 27–38 (1988)
53. Nucci, P., Nogueira, L.T., Lavor, C.: Determining protein backbone from H and H-alpha short interatomic distances, In: Mastorakis, N.E., Mladenov, Demiralp, M, V, Bojkovic Z (eds.) WSEAS Press, Athens, Advances in Biology, Bioengineering and Environment, 43–48 (2010)
54. Overhauser, A.W.: Polarization of nuclei in metals. Phys. Rev.**92**(2), 411–415 (1953)

55. Pardi, A., Billeter, M., Wüthrich, K.: Calibration of the angular-dependence of the amide proton-C-alpha proton coupling-constants, 3JHN-alpha, in a globular protein – Use of 3JHN-alpha for identification of helical secondary structure. J. Mol. Biol. **180**(3), 741–751 (1984)
56. Pauling, L., Corey, R.B.: 2 hydrogen-bonded spiral configurations of the polypeptide chain. J. Am. Chem. Soc. **72**(11), 5349–5349 (1950)
57. Pauling, L., Corey, R.B.: The structure of synthetic polypeptides. Proc. Nat. Acad. Sci. Unit. States Am. **37**(5), 241–250 (1951)
58. Pauling, L., Corey, R.B.: Configuration of polypeptide chains. Nature **168**, 550–551 (1951)
59. Pauling, L., Corey, R.B., Branson, H.R.: The structure of proteins – 2 hydrogen-bonded helical configurations of the polypeptide chain. Proc. Nat. Acad. Sci. Unit. States Am. **37**, 205–211 (1951)
60. Pinheiro, A.S., Amorim, G.C., Netto, L.E., Almeida, F.C., Valente, A.P.: NMR solution structure of the reduced form of thioredoxin 1 from Sacharomyces cerevisiae. Protein. Struct. Funct. Bioinformatics **70**(2), 584–587 (2008)
61. Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V.: Stereochemistry of polypeptide chain configurations. J. Mol. Biol. **7**(1), 95–99 (1963)
62. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., Nilges, M.: ARIA2: Automated NOE assignment and data integration in NMR structure calculation. Bioinformatics **23**(3), 381–382 (2007)
63. Sass, H.J., Musco, G., Stahl, S.J., Wingfield, P.T., Grzesiek, S.: An easy way to include weak alignment constraints into NMR structure calculations. J. Biomol. NMR **21**(3), 275–280 (2001)
64. Schwieters, C.D., Kuszewski, J.J., Clore, G.M.: Using Xplor-NIH for NMR molecular structure determination. Progr. Nucl. Mag. Reson. Spectros. **48**(1), 47–62 (2006)
65. Scott, W.R.P., Hürger, Ph.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., van Gunsteren, P.K.W.F.: The GROMOS biomolecular simulation program package. J. Phys. Chem. A **103**(19), 3596–3607 (1999)
66. Shen, Y., Delaglio, F., Cornilescu, G., Bax, A.: TALOS plus: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR **44**(4), 213–223 (2009)
67. Smith, G.M., Veber, D.F.: Computer-aided, systematic search of peptide conformations constrained by NMR data. Biochem. Biophys. Res. Commun. **134**(2), 907–914 (1986)
68. Spronk, C.A.E.M., Linge, J.P., Hilbers, C.W., Vuister, G.W.: Improving the quality of protein structures derived by NMR spectroscopy. J. Biomol. NMR **22**(3), 281–289 (2002)
69. Standley, D.M., Eyrich, V.A., Felts, A.K., Friesner, R.A., McDermott, A.E.: A branch and bound algorithm for protein structure refinement from sparse NMR data sets. J. Mol. Biol. **285**(4), 1691–1710 (1999)
70. Tjandra, N., Bax, A.: Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science **278**(5340), 1111–1114 (1997)
71. Tjandra, N., Marquardt, J., Clore, G.M.: Direct refinement against proton-proton dipolar couplings in NMR structure determination of macromolecules. J. Mag. Reson. **142**(2), 393–396 (2000)
72. Volk, J., Herrmann, T., Wüthrich, K.: Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J. Biomol. NMR **41**(3), 127–138 (2008)
73. Wagner, G., Braun, W., Havel, H.T., Schaumann, T., Go, N., Wüthrich, K.: Protein structures in solution by Nuclear-Magnetic-Resonance and Distance Geometry – the polypeptide fold of the basic pancreatic trypsin-inhibitor determined using 2 different algorithms, Disgeo and Disman. J. Mol. Biol. **196**(3), 611–639 (1987)
74. Wu, Z.R., Tjandra, N., Bax, A.: P-31 chemical shift anisotropy as an aid in determining nucleic acid structure in liquid crystals. J. Am. Chem. Soc. **123**(15), 3617–3618 (2001)
75. Wüthrich, K.W.T.: NMR of Proteins and Nucleic Acids, vol. 1. Wiley-Interscience, New York (1986)
76. Wüthrich, K., Billeter, M., Braun, W.: Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. J. Mol. Biol. **180**(3), 715–740 (1984)