Chapter 2

Comparison of Experiment and Simulation

Results from molecular dynamics simulations should be compared to experimental observables wherever possible to be able to asses their biological relevance. The methods to calculate observables from X-Ray crystallography, Nuclear Magnetic Resonance spectroscopy and Circular Dichroism spectroscopy from simulation data are reviewed and applications from the literature are discussed.

2.1 Introduction

Molecular dynamics (MD) simulations of biomolecules have become very popular over the last fifteen years. The increase of the number of reported MD studies is due primarily to the concurrent increase in computer power, but also to the increase of available high resolution protein structures, e.g. in the Brookhaven protein databank. Curiously, most of these structures are refined using MD techniques, where experimental restraints are added to the physical force field. Program packages for refinement such as X-Plor [55] or GROMOS [11] are widely used. Although MD techniques perform very well in refinement procedures, the use of unrestrained simulations, is not always without problems [56]. Therefore it is necessary to validate simulations of biomolecules by comparison with experimental data wherever possible. Nowadays, it is a requirement for publication of a biologically oriented paper to include such comparisons. In general, it is preferable to calculate experimental observables directly from the MD trajectory, because interpretation of data from neither experiment nor simulation is necessary this way. To do this, observables from NMR or CD experiments should be generated; some of the equations to do so will be given furtheron. Other features, such as structure factors may be calculated readily from a trajectory, but they are usually known only for pure liquids or liquid mixtures (from neutron or X-ray scattering) or (protein) crystals, but not for biomolecules in solution.

In the following sections, we will address some of the most often used analyses for MD simulations, as well as some less obvious properties. Simple equations for extracting experimental observables from simulation trajectories will be explicitly given, for more complex equations the reader is referred to the appropriate literature. In this review we will limit ourselves to data from X-Ray crystallography, Nuclear Magnetic Resonance (NMR) experiments and Circular Dichroism (CD) experiments.

2.2 X-Ray data

2.2.1 Root Mean Square Deviation

A commonly used criterium for validation of an MD simulation is the Root Mean Square Deviation (RMSD) from the crystal structure of a protein; although proteins are usually simulated in water rather than a crystalline environment, (some exceptions are in [57–61]), this is not a grave restriction. There are some proteins where the crystal structure is known to deviate significantly from the solution structure (such as calmodulin [3, 62, 63]), but it is generally accepted that the differences are small for most proteins. A small and stable RMSD value (typical < 0.2 nm) for protein backbone atoms is a useful quality control for protein simulations. The RMSD can be computed in two ways. The most popular one requires a rotational and translational fit to the reference structure, after which the RMSD for a set of atoms i (e.g. the C_αs) is computed as:

$$RMSD \;=\; \left[\frac{1}{N}\sum_{i}^{N}\left(\boldsymbol{r}_i - \boldsymbol{r}_i^0\right)^2\right]^{\frac{1}{2}} \tag{2.1}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{r}_i^0$ are the atomic position after fitting and the reference position of atom i respectively. Another method that does not require any fitting, employs a distance matrix:

$$RMSD \;=\; \left[\frac{2}{N^2}\sum_{i}^{N}\sum_{j>i}^{N}\left(d_{ij} - d_{ij}^0\right)^2\right]^{\frac{1}{2}} \tag{2.2}$$

where d_{ij} is the distance between atoms i and j and d_{ij}^0 the same distance in the reference structure. eqn. 2.2 yields other RMSD values than eqn. 2.1, usually somewhat higher. Since eqn. 2.2 does not require any fitting, it is the preferred method. However, it is not used very often in simulation studies, in contrast to NMR work. The fit free method was also used as a measure of structural similarity in a study aimed at defining clusters of structures from MD simulations [64]. Other useful properties include secondary structure, which can be determined by the DSSP program [65] and followed in time [47], and tertiary structure, which can be determined using distance matrices (for examples see [47,48]). Of course, it also possible to analyse detailed structural features such as individual hydrogen bonds, ion pairs or side chain packing and compare these to structural data.

Short linear peptides usually do not have stable conformations in aqueous solution [66]. Therefore, the RMSD from the crystal structure is a measure of unfolding rather than a quality criterion [47]. If different peptides are compared, the rate of unfolding determined (a.o.) by the RMSD [44,67] is sometimes used as a measure of stability. The thermodynamic stability, i.e. the free energy difference between unfolded and folded forms can usually not be determined by plain MD since the conformational space is not sampled well enough. However, the ***kinetic stability***, i.e. the resistance against unfolding, can be compared using MD simulations [44].

2.2.2 B-Factors

For another direct comparison to X-Ray crystallography data, the B-factors can be used, which can be obtained for each atom i from the positional fluctuations in a simulation by:

$$B_i \;=\; \frac{8\pi^2}{3}\left\langle\left(\boldsymbol{r}_i - \langle\boldsymbol{r}_i\rangle\right)^2\right\rangle \tag{2.3}$$

where $\boldsymbol{r}_i$ denotes the atomic position and $\langle\rangle$ denotes a time average. eqn. 2.3 is valid only when the positional fluctuations are independent of each other and isotropic, which is not generally the case [68]. Evaluation of B-factors from MD simulations have been performed relatively often for protein simulations [57,69–71]. It is however not easy to reproduce experimental data for a number of reasons, including sources of error in the crystal data such as crystal disorder, and sources of error due to the limited simulation time and possible force field biases. Moreover, proteins are usually simulated in solution rather than a crystalline environment. Thus, B-factors are reproduced qualitatively at best by simulations and no major conclusions can be drawn based on such comparisons. Finally the consistency of crystallographic B-factors between different structures can be questioned, as it was found by Hünenberger *et al.* that the correlation between B-factors for different structures of BPTI (4PTI, 5PTI, 6PTI) varies between 22 and 64% [70].

2.3 NMR Data

2.3.1 Chemical Shifts

The use of chemical shifts for structure determination and the problems associated with acquiring reliable chemical shift data have been reviewed recently by Wishart & Sykes [72]. These authors stress that a wealth of information can be obtained from chemical shift data, provided that accurate measurements are available. It is important to realise that environmental conditions (pH, ionic strength, cosolvents etc.) influence chemical shift values. This issue has been addressed quantitatively by a re-evaluation of ^{1}H random coil shifts at different pH and TFE concentration, by Merutka *et al.* [73]. This new set of random coil shifts is not too different from the old standard of Bundi & Wüthrich [74] but some appreciable differences of up to 0.2 ppm may be important for structural interpretation of chemical shift data. Furthermore a sometimes overlooked phenomenon is the interaction between peptides in the test tube. In contrast to CD experiments, NMR measurements must be performed at high peptide concentrations (up to several mM), leading to an *average* distance between peptide molecules that is less than 10 times the size of the peptide. A well known example that has led to some confusion in the past, is the dimerization of Leu-enkephalin [75]. Even when direct interactions are not important, indirect effects such as dielectric screening by charged peptides may influence the average behaviour. Although NMR experiments may be performed at different concentrations, they require a minimum concentration on the order of 1 mM, which is quite high already.

To be able to compare simulation data to experimental data a method is needed to compute chemical shifts from a structure. Two different approaches can be distinguished:
1. Ab initio methods that use quantum-chemical information to determine chemical shifts [76–79]. In principle these methods work equally well for all nuclei of interest in biological applications (i.e. ^{1}H, ^{13}C, ^{15}N, ^{17}O and ^{19}F), but they are very demanding in computer resources. Some of these methods have been implemented in the Gaussian-94 package [80].
2. Empirical methods. We can make a further subdivision in methods for $^{13}\text{C}_\alpha$ and $^{13}\text{C}_\beta$ chemical shift calculation which can be described in terms of ϕ/ψ angles [81], and methods for $^1\text{H}_\alpha$ chemical shifts. The latter method uses a simple formula to calculate

the deviation δ from random coil shift:

$$\delta \ = \ \delta^{ring} + \delta^{ani} + \delta^{E} \tag{2.4}$$

where δ^{ring} is the contribution of ring currents, δ^{ani} is the contribution due to magnetic anisotropy of peptide bonds and δ^{E} is the contribution due to the local electric field [82, 83]. In eqn. 2.4, the ring contribution δ^{ring} is calculated using either of the well established Haigh-Mallion or Johnson-Bovey theories. The parametrization of these methods was improved recently using density functional chemical shielding calculations [84]. It should be noted that there are methods that employ a combination of empirical and ab initio methods [78], leading to very good correlation between theory and experiment. The entirely empirical methods give good correlation between theory and experiment (up to 85%) for $^{1}\mathrm{H}_{\alpha}$ protons [82], and comparable for $^{13}\mathrm{C}_{\alpha}$ and $^{13}\mathrm{C}_{\beta}$ chemical shifts [81, 85]. Due to the computational cost of ab initio methods, it is not feasible to use them for structure refinement or for analysing protein trajectories from MD simulations. The most important use of quantum chemistry in this context is thus to improve the parametrization of empirical methods [77, 84]. Empirical methods to compute $^{1}\mathrm{H}\alpha$ as well as $^{13}\mathrm{C}\alpha$ and $^{13}\mathrm{C}\beta$ chemical shifts have been used for structure refinement as well [77, 86–89]. For the case of refinement using proton chemical shifts, it was found that the resulting structures did not change much as compared to refinement without chemical shifts, but that the correlation between measured chemical shifts and calculated shifts did improve significantly [88].

Thus, when chemical shifts in an NMR sample show a significant deviation from random coil values, and non-random conformations can be assumed to be present, it makes sense to compare experimental chemical shift data to computed values, that may be averaged over an MD trajectory. This way, the chemical shift may be determined as a time average, which is in the limit of infinite time identical to an ensemble average. Such a comparison has been made using a 2 ns MD simulation of a 25 residue peptide from the coat protein of cowpea chlorotic mottle virus (chapter 5 of this thesis). A fairly good agreement with experimental data was found although it could be determined that part of the amino acids were not equilibrated properly, because a large deviation from the experimental chemical shifts was found. Generally, it is not feasible to use other than the Hα protons for such comparisons, because the correlation between calculated and experimental shifts is not very good for other protons [85]. In special cases, where the ring-current effect of an aromatic group dominates the deviation from random-coil values, other protons than Hαs can be studied as well. In a study of a tetrapeptide from BPTI and some synthetic analogs, a backbone-amide proton was found experimentally to have an upfield shift of 1.5 ppm due to ring current effects [54]; this upfield shift was reproduced by MD simulations (chapter 3 of this thesis). Moreover, an amide chemical shift of one of the synthetic analogs of the BPTI peptide was found to agree very well with experimental data as well [54]. The MD trajectory clearly indicated that the peptide alternated between two different conformations, with different corresponding chemical shift for the amide proton. This example demonstrates the potential of MD simulations in the realm of modeling protein dynamics as was suggested by Wishart & Sykes [72].

2.3.2 J-coupling

Besides NOEs and chemical shifts, coupling constants, measured in NMR experiments, can provide useful structural information about the accessible solution conformations for proteins and peptides. The physical basis of coupling constants J is decribed in textbooks (e.g. [7]) and we will not discuss it here. A recent review by Case [85] describes the NMR experiments necessary to acquire coupling constants. In most analyses, the coupling constants are correlated to the backbone dihedral angles ϕ and ψ, or to the first side chain dihedral angle χ_1. The well-known Karplus equation relates coupling constants J to a torsion angle θ [90]:

$$J(\theta) \;=\; A\cos^2\theta + B\cos\theta + C \tag{2.5}$$

where the values of the constants A,B and C are determined empirically. For $^3\mathrm{J}_{NH\alpha}$, which describes the coupling between the NH proton and the Hα proton, a number of parametrization studies have been performed [91–94]. The results of these are given in Table 2.1; the differences are small, especially when one realises that most of these studies were performed using a single protein. For comparison the coefficients for the Karplus equation correlating $^3\mathrm{J}_{\alpha\beta}$ coupling to rotation about the χ_1 dihedral angle are also given. In the case of Ser and Thr side chains, the net $^3\mathrm{J}_{\alpha\beta}$ should be divided by 1.08 due to the electronegativity of the side chain oxygen [95].

Table 2.1: Empirically determined coefficients for the Karplus equation, for three-bond couplings.

Coupl.	$^3\mathrm{J}_{NH\alpha}$				$^3\mathrm{J}_{\alpha\beta}$	$^3\mathrm{J}_{H\alpha N}$	$^3\mathrm{J}_{H\alpha C'}$	$^3\mathrm{J}_{HNC\beta}$	$^3\mathrm{J}_{Ci-1H\alpha i}$
θ	ϕ-60				χ_1	ψ+60	ϕ	ϕ+60	ϕ+120
Ref.	[91]	[92]	[93]	[94]	[95]	[96]		[97]	
A	6.4	6.0	6.7	6.51	9.5	-0.88	4.0	4.7	4.5
B	-1.4	-1.4	-1.3	-1.76	-1.6	-0.61	1.1	-1.5	-1.3
C	1.9	2.4	1.5	1.60	1.8	-0.27	0.1	-0.2	-1.2

Another three-bond coupling constant in practical use is $^3\mathrm{J}_{N\beta}$ [85]. Furthermore, the one bond coupling $^1\mathrm{J}_{C\alpha H\alpha}$ can be correlated to backbone ϕ and ψ angles as follows [98]:

$$^1J_{C\alpha H\alpha} \;=\; 140.3 + 1.4\sin(\psi + 138) - 4.1\cos 2(\psi + 138) + 2.0\cos 2(\phi + 30) \tag{2.6}$$

Because of the simple relation between J couplings and backbone or side chain dihedral angles, they can be easily calculated from MD trajectories. An interesting example of this is the comparison of GROMOS and CHARMM force fields using simulations of Antamanide [99, 100]. From the simulations, $^3\mathrm{J}_{HH}$ couplings for the four Proline side chains were compared to NMR results, using a modified Karplus equation; both force fields reproduced the $^3\mathrm{J}_{HH}$ couplings from experiments. Another example is presented by Liu and Gierasch [101]. These authors studied a cyclic pentapeptide using both NMR and simulations, and did find good correspondance between calculated and experimentally determined $^3\mathrm{J}_{NH\alpha}$ coupling constants.

Finally, J couplings have been used for structure refinement as well. The most natural form for a penalty function, used first by Kim & Prestegard [102], is:

$$V_J \;=\; \frac{1}{2}K_j\left(J(\theta) - J_0\right)^2 \tag{2.7}$$

To account for mobility and conformational averaging in solution structures, the J(θ) can be taken as a running average during refinement [103].

2.3.3 Relaxation experiments

The use of nuclear magnetic relaxation experiments is of prime importance in the study of proteins and protein folding; observables from these experiments include NOEs and transverse and longitudinal relaxation times. Relaxation effects can arise from dipolar interactions, quadrupolar interaction for nuclei with spin $> 1/2$, and chemical shift anisotropy. The most important of these is the dipolar interaction. The theory behind relaxation phenomena is rather complicated; the standard reference for NMR theory, including relaxation, is the book by Abragam [104]. Valuable contributions to the theory and interpretation of relaxation experiments have been made by Macura & Ernst [105], Tropp [106], Lipari & Szabo [107] and by many others. We will present some useful equations that may be used to compare simulation to experiment, following the "model-free" method of Lipari & Szabo [107], but without fixing the internuclear separation.

The relaxation due to dipole-dipole interactions between two nuclei i and j at positions $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ can be described by a correlation function:

$$C(t) \;=\; \frac{1}{5}\left\langle \frac{P_2(\hat{\boldsymbol{r}}_{ij}^{LF}(0)\cdot\hat{\boldsymbol{r}}_{ij}^{LF}(t))}{r_{ij}(0)^3\, r_{ij}(t)^3}\right\rangle \tag{2.8}$$

Where $\hat{\boldsymbol{r}}_{ij}^{LF}$ is the unit vector connecting i and j in a laboratory frame of reference, P_2 is the second order Legendre polynomial:

$$P_2(x) \;=\; \frac{1}{2}\left(3x^2 - 1\right) \tag{2.9}$$

The spectral density is the Fourier cosine transform of eqn. 2.8:

$$J(\omega) \;=\; 2\int_0^\infty C(t)\cos(\omega t)\mathrm{d}t \tag{2.10}$$

Such a correlation function $C(t)$ contains internal motion as well as diffusive motion. It can be evaluated readily from a MD simulation, and using $C(t)$ the spectral densities $J(\omega)$ for any frequency can be computed. There is a practical problem however. The rotational correlation time of a protein in solution is on the order of 1-10 ns [108]. To obtain a correlation function of this length, a simulation must be performed which has a comparable length, preferably at least twice the rotational correlation time. Since this requirement is usually not met, the tail of the correlation function is not accessible by simulation [109], and we must make some assumptions to continue our analysis.

Order parameters

Usually, it can be assumed that a globular protein molecule undergoes isotropic overall motion which can be separated from internal motions [107]. In this case we can write the total correlation function as:

$$C(t) \;=\; \frac{1}{5}\,\mathrm{e}^{-6D_M t}\,C_1(t) \;=\; \frac{1}{5}\,\mathrm{e}^{-t/\tau_M}\,C_1(t) \tag{2.11}$$

where D_M and τ_M are the rotational diffusion constant and correlation times, respectively. For molecules with a different shape, e.g. a rigid rod, the motion is not isotropic, but it is still possible to describe the motion in similar fashion [106]. Furthermore,

$$C_1(t) \;=\; \left\langle \frac{P_2(\hat{\boldsymbol{r}}_{ij}(0)\cdot\hat{\boldsymbol{r}}_{ij}(t))}{r_{ij}(0)^3\,r_{ij}(t)^3} \right\rangle \tag{2.12}$$

where $\hat{\boldsymbol{r}}_{ij}$ is the unit vector connecting atoms i and j in a reference frame connected to the molecule itself. Note that the overall motion can be removed from a simulation trajectory by an optimal superposition algorithm. As a special case we can see that the correlation at $t = 0$ is given by:

$$C_1(0) \;=\; \left\langle r_{ij}^{-6} \right\rangle \tag{2.13}$$

Lipari & Szabo have defined a *generalized order parameter* $\mathcal{S}^2$:

$$C_1(\infty) \;=\; \mathcal{S}^2 \left\langle r_{ij}^{-6} \right\rangle \tag{2.14}$$

that is a measure for the restriction of motion. If $\mathcal{S}^2 = 0$, the motion of the two particles with respect to each other is not restricted in any way, if $\mathcal{S}^2 = 1$, the internuclear vector $\boldsymbol{r}_{ij}$ is rigidly fixed in the molecular frame. Using the addition theorem for spherical harmonics [110], and the property of correlation functions that

$$\lim_{t\to\infty} \langle A(0)B(t)\rangle \;=\; \langle A\rangle\,\langle B\rangle \tag{2.15}$$

eqn. 2.14 can be written as:

$$\mathcal{S}^2 \left\langle r_{ij}^{-6} \right\rangle \;=\; \frac{4\,\pi}{5} \sum_{m=-2}^{2} \left\langle \frac{Y_{2m}(\theta,\phi)}{r_{ij}^3} \right\rangle^2 \tag{2.16}$$

where θ and ϕ are the angles with respect to a molecular frame and the Y_{2m} are spherical harmonics functions[1]. When the motion is axially symmetric, i.e. independent of ϕ, only

[1] $Y_{20} \;=\; \sqrt{\frac{5}{4\pi}}\;(3\cos^2\theta - 1) \qquad Y_{2\pm1} \;=\; \mp\sqrt{\frac{45}{24\,\pi}}\,e^{\pm i\,\phi}\,\cos\theta\,\sin\theta \qquad Y_{2\pm2} \;=\; \sqrt{\frac{45}{96\,\pi}}\,e^{\pm2\,i\,\phi}\,\sin^2\theta$

the term with $m = 0$ contributes to $\mathcal{S}^2$. When the distance between the two atoms is fixed (e.g. in a covalent bond) and internal rotation is axially symmetric, the generalized order parameter reduces to the usual order parameter:

$$\mathcal{S} \; = \; \langle P_2(\cos\theta)\rangle \; = \; S \tag{2.17}$$

where θ is the angle between a bond vector μ and its symmetry axis. This order parameter can be readily evaluated from a simulation trajectory, and compared to experimental data.

Model free approach

The model-free approach of Lipari & Szabo [107] writes the internal motions in a macromolecule as a series of exponential terms:

$$C_1(t) \; = \; \left\langle r_{ij}^{-6} \right\rangle \sum_{i=0}^{\infty} a_i e^{-t/\tau_i} \tag{2.18}$$

where $\tau_0 = \infty$ and $\tau_{i+1} < \tau_i$. In their original paper, it is demonstrated that it is possible to analyze relaxation experiments when the series is truncated after the second term. From eqn. 2.13 and eqn. 2.14 it then follows that we can write the internal correlation function as:

$$C_1(t) \; = \; \left\langle r_{ij}^{-6} \right\rangle \left(\mathcal{S}^2 + \left(1 - \mathcal{S}^2\right) \, e^{-t/\tau_1} \right) \tag{2.19}$$

Using this approximation we can evaluate the spectral density function (eqn. 2.10) analytically

$$J(\omega) \; = \; \frac{2}{5} \left\langle r_{ij}^{-6} \right\rangle \left[\frac{\mathcal{S}^2 \, \tau_M}{1 + (\omega \, \tau_M)^2} + \frac{\left(1 - \mathcal{S}^2\right) \, \tau_c}{1 + (\omega \, \tau_c)^2} \right] \tag{2.20}$$

using

$$\tau_c^{-1} \; = \; \tau_1^{-1} + \tau_M^{-1} \tag{2.21}$$

In some cases the motions in a macromolecule can not be described by the simple two term approximation. Clore *et al.* have described results for the backbone dynamics of staphylococcal nuclease and interleukin-1β where a third term in the exponential representation was necessary to describe the data [111]. These authors introduce a second order parameter $\mathcal{S}_f^2$, that describes motions on an intermediate time scale, with time constant τ_f.

$\mathbf{T_1}$, $\mathbf{T_2}$ and NOE

The longitudinal and transverse relaxation rates T_1 and T_2 can be described in terms of spectral density functions:

$$T_1^{-1} \; = \; A \, [J(\omega_i - \omega_j) + 3J(\omega_i) + 6J(\omega_i + \omega_j)] \tag{2.22}$$

$$T_2^{-1} \; = \; \frac{1}{2} T_1^{-1} + \frac{A}{2} \, [4J(0) + 6J(\omega_j)] \tag{2.23}$$

with

$$A = \left(\frac{\hbar\gamma_i\gamma_j\mu_0}{8\pi}\right)^2 \tag{2.24}$$

where $\gamma_x = g_x\mu_x/\hbar$ is the gyromagnetic ratio for nucleus x, g_x is the nuclear g-factor, μ_x is the nuclear magneton and μ_0 is the magnetic permeability of free space ($\mu_0/(4\pi) = 10^{-7}$). In this description we have not included chemical shift anisotropy [112]. It should be noted that deviations between computed longitudinal relaxation rates T_1 and measured ones may occur due to spin diffusion [113]. Furthermore, we can write the cross-relaxation σ_{ij}, i.e. the relaxation of nucleus i due to nucleus j, in terms of $J(\omega)$ as well [106]:

$$\sigma_{ij} = (6J(\omega_i + \omega_j) - J(\omega_i - \omega_j)) \tag{2.25}$$

Using the cross-relaxation σ_{ij} and the longitudinal relaxation time T_1 we can write a steady state NOE_{ij} between particles i and j as:

$$NOE_{ij} = 1 + \frac{\gamma_i}{\gamma_j} A\, T_1\, \sigma_{ij} \tag{2.26}$$

Since T_1 occurs in the definition of the NOE, the NOE is also influenced by spin diffusion [113]. The steady state NOE is used primarily in heteronuclear NMR experiments. In homonuclear ^{1}H-NMR experiments, NOEs can be measured by multi-dimensional experiments, and through the NOE intensity I_{ij}, the cross relaxation σ_{ij} can be determined directly [114]:

$$I_{ij} \propto \sigma_{ij} \tag{2.27}$$

A few limiting cases are of particular interest. The first case is that of a slowly tumbling macromolecule with fast internal motions. In this case the cross relaxation is dominated by $J(0)$, and can be written using the model-free approach of Lipari & Szabo (eqn. 2.20):

$$\sigma_{ij} = -\frac{2}{5}\left[\tau_c \left\langle r_{ij}^{-6}\right\rangle + (\tau_M - \tau_c)\frac{4\,\pi}{5}\sum_{m=-2}^{2}\left\langle \frac{Y_{2m}(\theta,\phi)}{r_{ij}^3}\right\rangle^2\right] \tag{2.28}$$

When $\tau_M \gg \tau_c$ this reduces to:

$$\sigma_{ij} = -\frac{8\,\pi\,\tau_M}{25}\sum_{m=-2}^{2}\left\langle \frac{Y_{2m}(\theta,\phi)}{r_{ij}^3}\right\rangle^2 \tag{2.29}$$

As was originally shown by Tropp [106], the distance dependence of σ_{ij} involves r^{-3} rather than r^{-6}, which means that particles at relatively long distances may contribute to cross relaxation. Hence, the NOE intensity depends on the time average of r^{-3}. In the case that the distance is (almost) constant, the distance dependence can be taken out of the averaging, and using the addition theorem for spherical harmonics [110] we can write the cross relaxation σ_{ij} as:

$$\sigma_{ij} = -\frac{2}{5}\tau_M \left\langle r_{ij}^{-6}\right\rangle \left\langle P_2(\hat{\boldsymbol{r}}_{ij}(0)\cdot\hat{\boldsymbol{r}}_{ij}(t))\right\rangle \tag{2.30}$$

using the definitions of eqn. 2.12.

Another important effect is that of multiple protons (e.g. a methyl group) that contribute to cross relaxation at another proton. Some care is required, since in practice one usually works with effective distances r_{ij}^{noe} rather than σ_{ij} (here we omit the angular dependence for the sake of clarity):

$$r_{ij}^{noe} \propto \sigma_{ij}^{-1/6} \tag{2.31}$$

The cross relaxation terms σ_{ij} can be added linearly, but in terms of distances this means that one first has to compute the $\left\langle r^{-3}\right\rangle$ average for each of the particles, and subsequently sum these squared (using eqn. 2.29, i.e. under the assumption that $\tau_M \gg \tau_c$):

$$r_{ij}^{noe} = \left[\sum_{j=1}^{3} \left\langle r_{ij}^{-3}\right\rangle^2\right]^{-1/6} \tag{2.32}$$

Structure refinement based on NOE data

The distance information from NOEs can be used for refinement by MD simulations by introducing a penalty function for *distance restraints* such as:

$$V_{ij} = \begin{cases} 0 & r_{ij} \leq r_{ij}^{noe} \\ \frac{1}{2}k_{dr}\left(r_{ij} - r_{ij}^{noe}\right)^2 & r_{ij} > r_{ij}^{noe} \end{cases} \tag{2.33}$$

where k_{dr} is a force constant, which is usually taken on the order of 1000 kJ mole^{-1} nm^{-2}. The first NMR refinement studies employed eqn. 2.33 directly, by inserting the instantaneous distance r_{ij} between two nuclei in eqn. 2.33 to calculate energy, and forces using the partial derivatives with respect to the nuclear positions. It may be clear from the preceding paragraph however, that internuclear distances derived from NOEs (eqn. 2.31) have to be interpreted as effective average distances. Usually, structure refinement is done for proteins, so that we can use eqn. 2.29 to calculate an effective average distance; the explicit angular dependence has, to our knowledge, never been used. Rather, this term is ignored and an effective distance $\bar{r}_{ij}$ is defined as:

$$\bar{r}_{ij} = \left\langle r_{ij}^{-3}\right\rangle^{-1/3} \tag{2.34}$$

which can then be inserted in eqn. 2.33 instead of r_{ij}. Torda *et al.* introduced an algorithm to take time-averaged distances into account in refinement [115], and in a later paper applied it to refinement of tendamistat [116]. Since this original paper [115] the use of time-averaging has become standard practice in NMR refinement based on NOE restraints. A more extensive treatment of the application of distance restraints in MD simulations can be found in the $GRO\underline{macs}$ user manual [13].

A possible improvement of refinement methodology would be the use of a penalty function based on the σ_{ij} rather than the effective distance (eqn. 2.31). The most natural form for such a function would be (analogous to eqn. 2.7):

$$V_{ij} \;=\; \frac{1}{2}k_{dr}\left(\sigma_{ij} - \sigma_{ij}^{exp}\right)^2 \tag{2.35}$$

where σ_{ij}^{exp} is the experimental value, which is, in contrast to r_{ij}^{noe} (eqn. 2.31), a direct experimental observable. The evaluation of σ_{ij} is somewhat more expensive than that of $\left\langle r^{-3}\right\rangle$, and an optimal superposition is necessary at each timestep of the refinement. Although this makes the refinement as a whole somewhat more expensive in computer time, the advantage of a more rigorous treatment should more than compensate for this. It should be noted here, that the problem of spin diffusion [113] is probably even more important. Attempts at incorporating this in refinement procedures have been published [117].

Relaxation data from MD simulations

A number of interesting studies of NMR relaxation from simulation data are present in recent scientific literature [109, 118–122] as well as some older studies based on short simulations in vcauo [123, 124]. Palmer & Case describe a careful analysis of NMR relaxation in a Zinc-finger peptide [120]. They compute order parameters $\mathcal{S}^2$ using (a.o.) eqn. 2.16, and relaxation time constants τ_c using the definition of Lipari & Szabo:

$$\tau_c \;=\; (1-\mathcal{S}^2)^{-1}\int_0^\infty \left[C_1(t) - \mathcal{S}^2\right] dt \tag{2.36}$$

The authors find almost quantitative correspondence between order parameters from a solvated MD simulation and experiment for C_α-H bonds. Generally, order parameters from MD simulations are well reproduced. However, Smith *et al.* report T_1 and T_2 data from MD simulations of BPTI which do not agree very well at all [109]. The main cause of the discrepancy seems to be the simulation time of 1000 ps which is short compared to the rotational correlation time.

Post has studied the effects of motional averaging on the NOE intensity [118]. She separately studied the effect of angular averaging and radial averaging, by writing the cross relaxation as the product of an order parameter and a radial average:

$$\sigma_{ij} \;=\; -\frac{2}{5}\tau_M \,\left\langle P_2((\hat{\boldsymbol{r}}_{ij}(0)\cdot\hat{\boldsymbol{r}}_{ij}(t))\right\rangle^2 \,\left\langle r_{ij}^{-3}\right\rangle^2 \tag{2.37}$$

Although this simplification is not justified in all cases, because angular and radial motion may be coupled, it gives some insight in the relative importance of the two contributions. In general the angular component works to reduce the NOE intensity to about 90% of the intensity corresponding to a rigid model, while the radial component increases the NOE intensity to about 105 % of the intensity corresponding to a rigid model [118].

2.4 CD Data

Circular dichroism is a spectroscopic method that is well established in the biochemical community. Theoretical aspects as well as applications to proteins, peptides, DNA and RNA have been reviewed thoroughly by Woody recently [125]. Most theoretical CD work is based on the matrix formulation for rotational strength of Bayley *et al.* [126] or older work. The contribution of specific parts of a protein to the CD spectrum can be calculated using molecular orbital calculations, as has been done for aromatic groups [127], peptide groups in poly-Gly helices [128] and of methylated phenols complexed with β-Cyclodextrin [129]. Using the same methodology, the CD spectrum of a cyclic peptide (L-Tyr-LTyr) was calculated as the average of an MD trajectory [130]. All these authors get qualitative agreement with experimental data, but not quantitative. A logical step in refining the methodology for calculating CD spectra seems to be the use of quantum chemistry at a high level of theory, rather than the customary molecular orbital calculations. It must be noted that the relatively new technique of vibrational circular dichroism (VCD) [125] has achieved quite some interest from theoretical chemists, e.g. [131]. However, the physical basis of VCD is very different from conventional CD, and therefore the theoretical work in the VCD field does not apply to CD.

The only direct comparison of an MD simulation with a CD spectrum is at the moment the example of cyclic (L-Tyr-L-Tyr) [130]. However, when we focus our interest on α-helices we can use an empirical method that relates ϕ/ψ angles to CD ellipticity at 222 nm ($[\theta]_{222}$), which is the wavelength characteristic of an α-helix in a CD spectrum. This method, devised by Hirst and Brooks [132], is based on the calculations of Manning and Woody [128]. It defines a residue to be α-helical when:

$$\left[(\phi-\phi_c)^2+(\psi-\psi_c)^2\right]^{\frac{1}{2}} \; < \; 8^{\mathrm{o}} \tag{2.38}$$

where the angles are given in degrees, and the reference angles ϕ_c and ψ_c are taken from a table of 12 ϕ/ψ combinations. Using a relation for the ellipticity $[\theta]_{222}$ based on this definition of α-helicity, the contribution of every residue to the total ellipticity of myoglobin was calculated as the average ellipticity per residue [132]. The study demonstrates the influence of α-helix length and conformation on the ellipticity $[\theta]_{222}$. A further study by the same authors focuses on individual α-helices, and compares their α-helicity to estimate the relative stabilities [67]. Finally, the method was also applied to a peptide from the coat protein of cowpea chlorotic mottle virus (chapter 5 of this thesis). Here, the α-helicity was compared to α-helicity from CD data [133], and a good agreement was found between simulation and experimental data in different environments.

An entirely empirical method of computing a CD spectrum might be feasible when more solution structures become available. However, it would require many parameters to determine what ϕ/ψ angles contribute to each wavelength in the spectrum and therefore a more rigorous method based on high level ab initio calculations is to be preferred.

2.5 Other Experimental Data

Many other experimental methods are available to study (protein) molecules in solution, like Fourier transform infrared spectroscopy, time resolved fluorescence [134], small angle X-ray scattering and neutron scattering. For the sake of brevity we will restrict ourselves to a short description of the latter two.

2.5.1 Small angle X-ray scattering

Small Angle X-Ray Scattering (SAXS) can provide information about the global shape of a molecule through the distance distribution function $P(r)$. The theory of SAXS and its application to macromolecules was desribed by Moore [135]. SAXS experiments are especially interesting in the case of non-globular (protein) molecules, like calmodulin (CaM). Some interesting applications of SAXS to CaM and mutants of CaM have been published [62, 136]. In principle, it is straighforward to compute the $P(r)$ function from a MD trajectory. However, it is currently not feasible to sample the large collective motions that determine the shape of the $P(r)$ function for a protein like CaM in solution.

2.5.2 Neutron scattering

Neutron scattering is a very powerful technique that gives information on the atomic environment of a given solute in another solvent in the form of a structure factor $S(k)$. This structure factor can be converted into a radial distribution function (RDF) by a Fourier transformation [137]. The RDF $g(r)$ is defined such that the quantity $\rho g(r) d\mathbf{r}_0$ is the "probability" of observing a second atom in the spherical shell between $\mathbf{r}$ and $\mathbf{r}$+d$\mathbf{r}$ given that there is an atom at the origin of $\mathbf{r}$, weighted by the nuclear scattering factor, ρ is the particle density. The computation of $g(r)$ is straightforward from a MD trajectory [53], and these function have been used very often to test potential functions, like for water (for a review see [138]). A particular interesting other example in the context of protein folding is that of urea in water. A number of simulation studies have tried to reproduce the experimental $g(r)$ function [139–146].

It should be noted, that it is very well possible to calculate the experimental observable, the structure factor $S(k)$, directly from a simulation trajectory [137]. A comparison of $S(k)$ from simulation to $S(k)$ from experiment has the advantage that it avoids the interpretation of experimental data. Finally, it is also possible to measure frequency dependent structure factors $S(k, \omega)$ by quasi-elastic neutron scattering. From $S(k, \omega)$ one can construct a time-dependent pair correlation function $G(r, t)$. Comparisons of such experiment with simulation data have been performed [137].

2.6 Conclusion

To test the validity of a computer simulation, it is necessary to compare the results to experimental data wherever possible. In doing so, we must distinguish structural properties (e.g. RMSD from a reference structure, NOEs, chemical shifts, J-coupling constants or α-helicity) and dynamical properties (e.g. B-factors or order parameters). The methods of acquiring this information from a MD trajectory are described in this review. For a proper evaluation of these variables it is necessary to average over a long trajectory. Some of the properties can not be computed accurately from simulations because of the limited simulation length. A normal requirement for a simulation trajectory is that it should be long enought to allow molecules to sample all their equilibrium conformations. While this requirement can be met for simulation of pure liquids, where we can average over many liquid molecules, this is impossible for a simulation of a single (macro)molecule in solution (cf. chapter 4 of this thesis). It has been suggested that, for the purpose of sampling, it is better to use a number of short simulations rather than a single long one [69]. However, when one is interested in kinetic effects, such as the unfolding of a protein [37,41–43] or peptide [36,39,44,47,147], processes which take place on the nanosecond timescale, this is not possible.

It should be stressed that the analysis tools given in this review can be used for peptides as well as proteins. However, they should be interpreted with great care because of the fact that a peptide simulation can not give an equilibrium trajectory [148], except for very short peptides (chapter 3 of this thesis). Some analyses, such as chemical shift calculation, or order parameters, can provide information per residue. In this manner it can be determined whether part of a peptide is in local equilibrium (e.g. chapter 5 of this thesis).

It is our hope and belief that further methodological developments, especially for the calculation of chemical shifts and CD spectra, will soon emerge, to the benefit of theoreticians as well as experimentalists. Meanwhile, the currently available tools to analyse MD trajectories should be more than enough to prove whether a simulation is reliable.