Journal of Biomolecular NMR, **19:** 305–320, 2001.
KLUWER/ESCOM

Assessing the effect of conformational averaging on the measured values of observables

Roland Bürgi, Jed Pitera & Wilfred F. van Gunsteren*
Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, CH-8092 Zürich, Switzerland

Received 1 September 2000; Accepted 22 December 2000

Key words: averaging, molecular dynamics, NOE, ^3J-coupling constant, saddle-point approximation

Abstract

Experiment and computer simulation are two complementary tools to understand the dynamics and behavior of biopolymers in solution. One particular area of interest is the ensemble of conformations populated by a particular molecule in solution. For example, what fraction of a protein sample exists in its folded conformation? How often does a particular peptide form an alpha helix versus a beta hairpin? To address these questions, it is important to determine the sensitivity of a particular experiment to changes in the distribution of molecular conformations. Consequently, a general analytic formalism is proposed to determine the sensitivity of a spectroscopic observable to the underlying distribution of conformations. A particular strength of the approach is that it provides an expression for a weighted average across conformational substates that is independent of the averaging function used. The formalism is described and applied to experimental and simulated nuclear Overhauser enhancement (NOE) and ^3J-coupling data on peptides in solution.

Introduction

Statistical mechanics shows that any simple observable of a macroscopic system (a solution of protein molecules) can be described as an ensemble average over microscopic states (individual protein molecules in that solution). When considering an observable at a particular time, this ensemble average is carried out linearly over all of the N molecules in the observed volume:

$$\langle A \rangle = \frac{1}{N} \sum_{i=1}^{N} A_i. \tag{1}$$

Since the value of observable A is usually some function of the molecular conformation r, $A(r)$, the sum over molecules above is typically re-cast as an integral over possible conformations r weighted by the probability of each conformation $p(r)$:

$$\langle A \rangle = \int A(r)\, p(r)\, \mathrm{d}r, \tag{2}$$

where

$$\int p(r)\, \mathrm{d}r = 1. \tag{3}$$

Equation 2 shows how to calculate the value of an ensemble average once the distribution $p(r)$ of microscopic states in a sample is known. Unfortunately, it is often straightforward to determine $\langle A \rangle$ from the experiment, while the underlying distribution $p(r)$ is experimentally inaccessible since the experiment is both a time- and ensemble-average over molecular conformations. In contrast, molecular dynamics and Monte Carlo simulations provide a direct sampling of microscopic states r with probability $p(r)$ but often have difficulty reaching the time scales necessary to yield converged values of $\langle A \rangle$.

For bulk materials and macroscopic properties of condensed matter systems, we are not particularly interested in the distribution of microscopic states. However, this is not the case for proteins and other biomolecules. For these systems, conformation is intimately linked to function – for example, it can modulate ligand affinity, as in the T and R states of hemoglobin (Moffat et al., 1979) or affect the access

*To whom correspondence should be addressed. E-mail: wfvgn@igc.phys.chem.ethz.ch

of substrates to an enzyme active site, like the open and closed forms of hexokinase (Bennett and Steitz, 1980) or the gated states of acetylcholinesterase (Zhou et al., 1998). To understand the biological function it is essential to understand the populations and dynamics of these conformational substates.

Towards this end, theoretical and experimental studies of biomolecules are rapidly converging. While the first computer simulation of protein dynamics was 3 picoseconds in simulated duration (McCammon et al., 1977), today routine protein simulations span several nanoseconds (Stocker et al., 2000) and typical peptide simulations reach 50 to 100 nanoseconds (Daura et al., 1999b). In parallel, experimental methods have increased enormously in both time resolution and molecular resolution. Single molecule fluorescence spectroscopy is providing information on the behaviour (and variations) of individual enzymes and proteins (Nie et al., 1994).

Recent long-timescale simulations of proteins and peptides have underscored the fact that it is possible for an ensemble of populations (or a trajectory of a single molecule) to yield ensemble averages compatible with a particular microscopic state even though the simulated ensemble contains only fractional populations of the particular state in question. For example, Daura et al. (1999a) have shown that molecular dynamics trajectories of a β-heptapeptide populated a particular 'folded' conformation only 50% of the time at 340 K, yet yielded no NOE distance bound violation greater than 0.06 nm.

This problem has actually been well known in the spectroscopic community for some time. While it is extensively discussed in the NMR literature (Jardetzky and Roberts, 1981), attempts to connect NMR observables to the underlying conformational ensemble have concentrated on simplified models of the molecular geometry. For example, Braun et al. (1981) showed that a uniform distribution of interatomic distances would yield an NOE signal if at least 10% of the distribution falls within a threshold distance. For the relationship between ^3J-coupling constants and torsion angles, Jardetzky considered the influence of averaging over several discrete conformations on the apparent value of several NMR observables (Jardetzky, 1980), and showed that a particular value of an observable is often compatible with a range of distributions over several different conformations. It is well known that certain ranges of ^3J-coupling constants provide limited structural information due to the degeneracy or multiple-valuedness of the Karplus curve relating ^3J-coupling constants and the corresponding torsion angles (Syberts et al., 1987).

However, these analyses have been based on either uniform distributions or sampling between a small number of discrete model conformations. Considering more realistic examples, Bonvin and Brunger (1996) have explored how well a collection of NOEs describes a mix of several realistic conformations of a protein loop. Similar analyses have been performed using conformations from a molecular dynamics trajectory by Daura et al. (1999a). In both cases it was shown that the available NOE distance information was not able to precisely define the conformations populated by the molecule in question. This is a significant issue given the use of time- and ensemble-averaging schemes in modern NMR refinement protocols (Bonvin et al., 1994). More detailed comparisons have been performed between experimental and simulated cross-relaxation rates and order parameters (Bruschweiler et al., 1992; Philippopoulos and Lim, 1994; Beutler et al., 1996), but primarily with the intent of reproducing the experimental observations.

If more observations are available than the number of relevant degrees of freedom, it becomes possible to use statistical techniques of sensitivity analysis or component analysis to infer the underlying probability distribution or potential (Ho and Rabitz, 1993; Lazarides et al., 1994; Utz, 1998). In the single observable case we consider, such techniques are not applicable since they require more data than parameters. As a result we make use of a simple parametric sensitivity analysis in this paper.

Although the problem of connecting the value of a spectroscopic observable with its underlying ensemble has probably been discussed the most in the context of NMR, it is a general issue for any observation of a molecule that undergoes some sort of averaging (van Gunsteren et al., 1994, 1999). It is particularly crucial for thermodynamic interpretations that have been inferred from spectroscopic data, or for attempts to describe the structure of highly flexible systems such as short peptides and protein loops.

The formalism described in this paper permits the analytic use of realistic probability distributions to estimate the sensitivity of a particular spectroscopic observable to the composition of the underlying ensemble. We must note, however, that our analysis is restricted to situations where the kinetics of conformational exchange do not themselves influence the observable measured for each molecule. For NMR, this corresponds to the case where exchange between

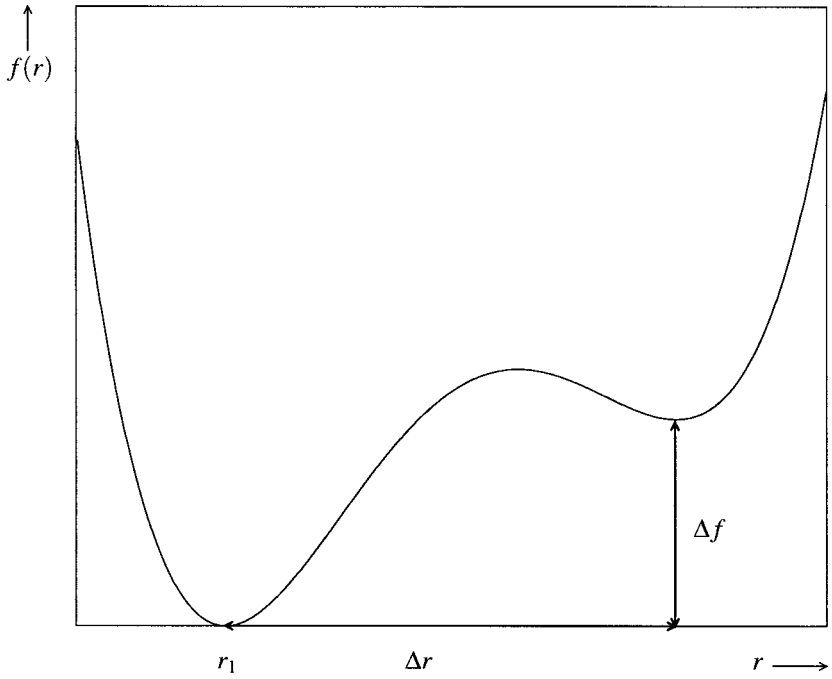

Figure 1. Double-well potential $f(r)$ (Equation 5) described by the parameters r_1, Δr, and Δf.

conformations is slow with respect to overall rotation but fast with respect to the macroscopic relaxation rate (T_2).

While we describe the averaging problem in terms of single observables, a more complex case often occurs where several observables are measured for a particular sample. For example, cyclic structures are commonly determined by considering all measured ^3J-coupling values around the cycle and selecting the conformation that best fits this set of data. Though we do not deal with this case explicitly, it is straightforward to extend our single-observable formalism to the realm of two or more observables. If the observables all correspond to the same degree of freedom, r, then when taken as a whole they can significantly constrain the underlying probability distribution $p(r)$. Finding the bounds on $p(r)$ becomes a question of solving a system of N equations, one for each observable but all with common parameters describing the underlying distribution. When the observables apply to different degrees of freedom (e.g., $\langle A_1 \rangle$ and $\langle A_2 \rangle$ with independent distributions $p(r_1)$ and $p(r_2)$), far less information can be derived. In the case of nonlinear averaging, the formalism described herein can be used to estimate the minimum fractional population necessary to generate a particular ensemble average. If the sum of these minimum fractions is greater than 1, it implies that the underlying $p(r)$ probability distributions must be somewhat correlated. If the sum is less than unity, however, all the individual ensemble averages can be satisfied even if the underlying probability distributions are uncorrelated.

Sensitivity analysis

This section has the purpose of estimating analytically the effect of different distributions and averaging methods on observables. Using the saddle-point approximation (see Appendix A), we can calculate a weighted average for a distribution together with a general averaging function, where the weights do not depend on the averaging function itself. This provides us with a function that can be readily used in the sensitivity analysis of a given observable by analysing the derivatives of the weighted average with respect to the parameters describing the distribution.

In this work, we focus on the analysis of a bimodal distribution. For many questions asked in a sensitivity analysis, considering a bimodal distribution is sufficient – the sensitivity of an averaging function towards shifting two maxima of a distribution with respect to each other can be analysed. If more complex distrib-

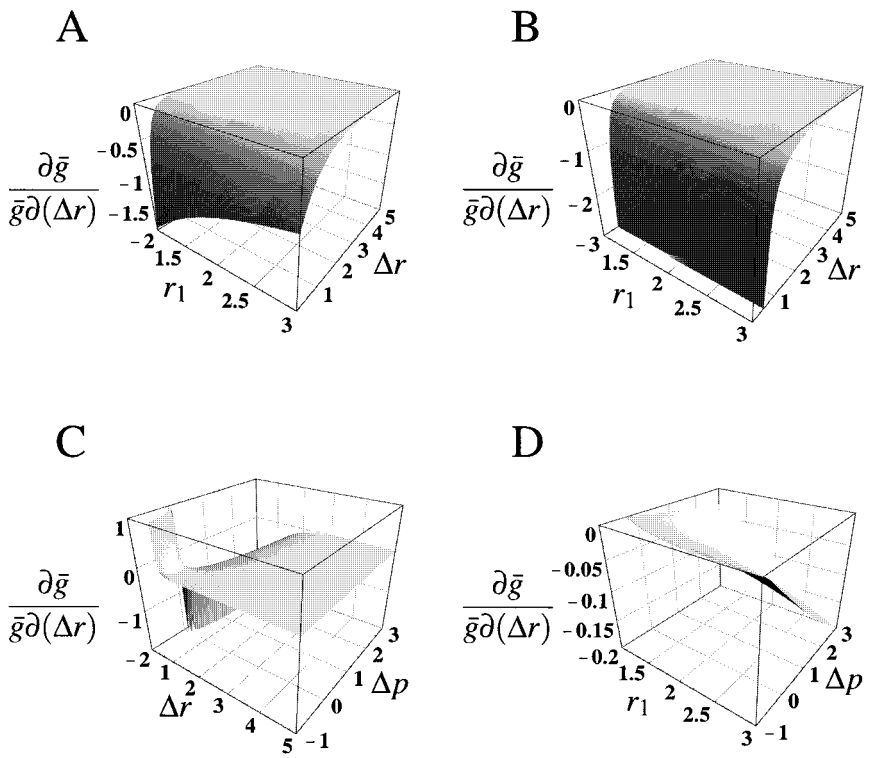

Figure 2. $(\partial\bar{g}/\partial(\Delta r))/\bar{g}$ for the bimodal distribution (Equations 4 and 5) and averaging Equation 13 with $\beta = 100$. (A) shows the derivative as a function of r_1 and Δr at $\Delta p = 0$; in (B), the value of Δp is 1. (C) shows the derivative as a function of Δr and Δp at $r_1 = 2$, and (D) as a function of r_1 and Δp at $\Delta r = 2$.

utions have to be considered, the formalism is easily expandable to general distributions.

Weighted average in a bimodal distribution

For simplicity, we will denote the degree of freedom over which is averaged by r. The averaging function defining the observable will be called $g(r)$. It is easiest to analyse a bimodal distribution $p(r)$ of the conformational degree of freedom r, as the number of parameters describing such a distribution is sufficiently small. The bimodal distribution $p(r)$ can be described by a double-well effective potential $f(r)$ (see Figure 1) through

$$f(r) = -\frac{1}{\beta}\log p(r), \tag{4}$$

where β is one of the parameters describing the distribution $p(r)$. The potential $f(r)$ is described by the function

$$f(r) = \frac{1}{4\Delta r^3}\left\{(r-r_1)^2\left[\Delta r^3(r-r_1-\Delta r)^2 + 4\Delta f(3\Delta r - 2r + 2r_1)\right]\right\} + a_0, \tag{5}$$

for $0 \le r \le \infty$. Its functional form (Equation 5) has been chosen such that $f'(r_1) = f'(r_1+\Delta r) = 0$, $f(r_1) = a_0$, and $f(r_1+\Delta r) = a_0 + \Delta f$. In order to obtain a proper probability distribution $p(r)$ through Equation 4, it is required that $p(r)$ satisfies Equation 3. In other words, a_0 has to be chosen such that

$$\int_0^\infty e^{-\beta f(r)}\,\mathrm{d}r = 1. \tag{6}$$

This condition yields, using the saddle-point approximation Equation 17 considering both minima of the potential,

$$a_0 = \frac{1}{\beta}\log\left[2\sqrt{\frac{\pi}{\beta}}\Delta r\,(A+B)\right], \tag{7}$$

with

$$A = \frac{1}{\sqrt{\Delta r^4 + 12\Delta f}} \tag{8}$$

$$B = \frac{e^{-\beta\Delta f}}{\sqrt{\Delta r^4 - 12\Delta f}} \tag{9}$$

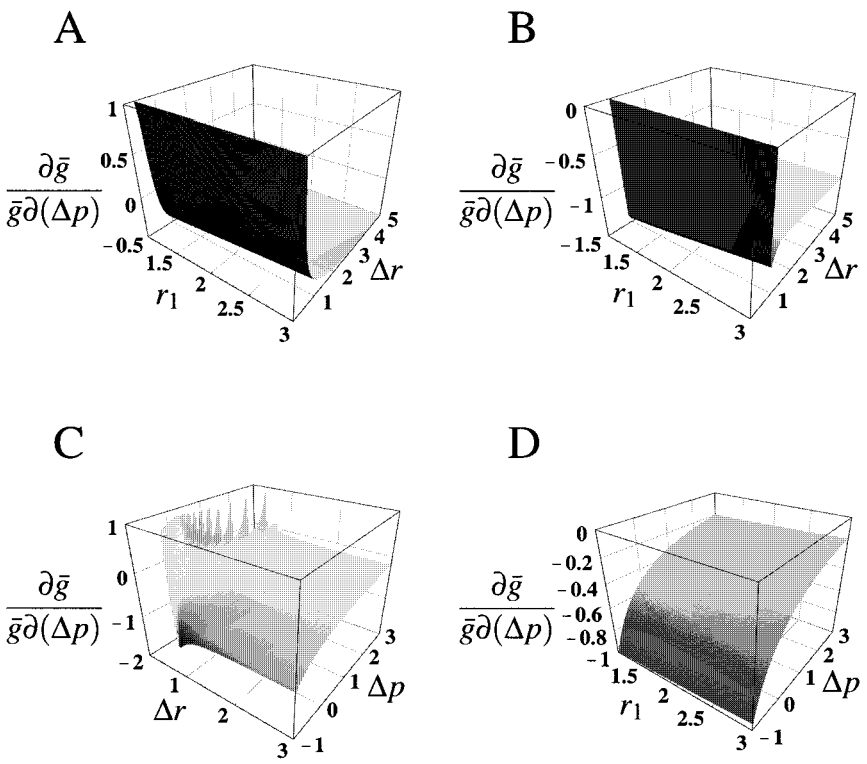

Figure 3. $(\partial\bar{g}/\partial(\Delta p))/\bar{g}$ for the bimodal distribution (Equations 4 and 5) and averaging Equation 13 with $\beta = 100$. (A) shows the derivative as a function of r_1 and Δr at $\Delta p = 0$; in (B), the value of Δp is -0.98. (C) shows the derivative as a function of Δr and Δp at $r_1 = 2$, and (D) as a function of r_1 and Δp at $\Delta r = 2$.

Note that the condition

$$|\Delta f| < \frac{\Delta r^4}{12} \tag{10}$$

must be fulfilled. Thus we have four parameters describing the bimodal distribution: r_1, Δr, Δf, and β.

The average quantity $\bar{g}$ can then be calculated, using the saddle-point approximation, as

$$\bar{g} = \frac{Ag(r_1) + Bg(r_1 + \Delta r)}{A + B}. \tag{11}$$

In the special case of $\Delta f = 0$, this leads to

$$\bar{g} = \tfrac{1}{2}\left(g(r_1) + g(r_1 + \Delta r)\right). \tag{12}$$

Extension to general distributions

The formalism described above can easily be extended to general distributions. For each new peak, two additional parameters have to be introduced in a similar way as for the bimodal distribution, describing the vertical and horizontal distances of the maxima with respect to each other. Due to the saddle point approximation, each peak will contribute as a summand to the average. Needless to say that the formalism cannot be applied to a distribution without a single peak.

Application of the formalism to NOE distances

In this section, the derived formulae are used to investigate the effect of the r^{-6} averaging when obtaining a value for an observable, i.e., for the case

$$g(r) = r^{-6}. \tag{13}$$

Analysis of the r^{-6} average

To analyse the influence of the r^{-6} averaging in a double-well potential, it is of interest to look at the behaviour of the derivatives $\partial\bar{g}/\partial(\Delta r)$ and $\partial\bar{g}/\partial(\Delta f)$ for several cases. However, as Δf is not very easy to interpret directly, we will introduce the parameter

$$\begin{aligned}\Delta p &= \frac{p(r_1 + \Delta r) - p(r_1)}{p(r_1)} \\ &= e^{-\beta\Delta f} - 1,\end{aligned} \tag{14}$$

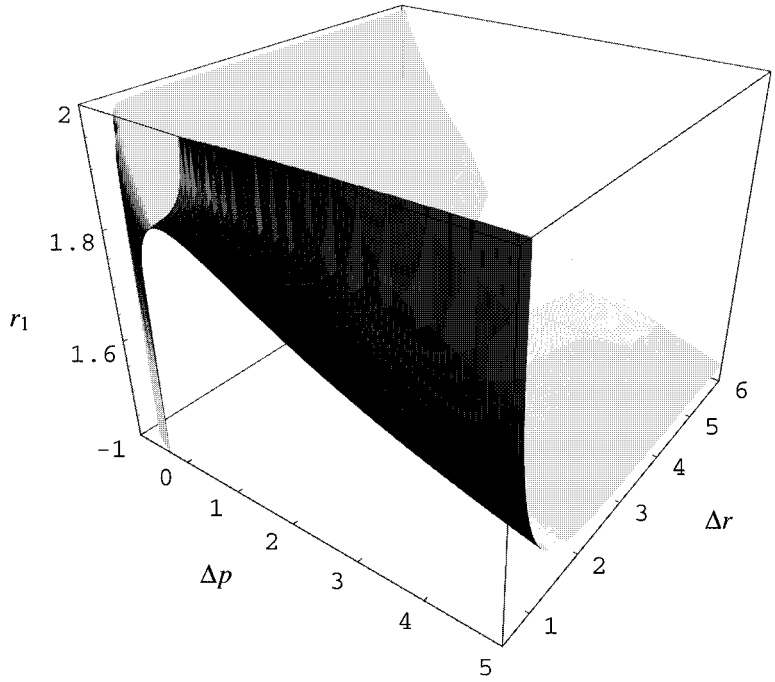

Figure 4. Parameter values defining bimodal distance distributions. All points in the surface $(r_1, \Delta r, \Delta p)$ yield the average $\bar{g} = 2^{-6}$ as determined by Equations 4–11 and 14.

Table 1. Four example interatomic distance distributions taken from two MD simulations (Bürgi et al., 2001; Daura et al., 1998, 1999b). Indicated are the parameters r_1, Δr, Δf, and β (Equations 4 and 5) of a bimodal distribution fitted to the simulated distributions. The last five lines show the r^{-6} weighted average distance $\bar{r}$ (MD) calculated over the simulated distribution, the average distance $\bar{r}$ (SPA) calculated using the saddle-point approximation (SPA) to the bimodal distribution, the upper bound distance $\bar{r}$ (exp.) derived from NOE experiments, and the value of the derivative of the average $\bar{g} = \left\langle r^{-6} \right\rangle$ with respect to Δr and Δp of the bimodal distribution

Atom pair	Octapeptide in DMSO 150 ns		Heptapeptide in MeOH 200 ns	
	1CB1-2HN	2CB2-5HN	2HN-5HCB	3HN-5HCB
r_1 (nm)	0.318	0.455	0.314	0.364
Δr (nm)	0.117	0.125	0.440	0.405
Δf (nm^4)	-1.12×10^{-6}	-1.23×10^{-6}	2.16×10^{-4}	8.54×10^{-5}
β (nm^{-4})	8.72×10^{5}	9.25×10^{4}	3.62×10^{3}	4.50×10^{3}
$\bar{r}$ (MD) (nm)	0.335	0.487	0.357	0.402
$\bar{r}$ (SPA) (nm)	0.371	0.495	0.336	0.397
$\bar{r}$ (exp.) (nm)	0.404	0.471	0.330	0.340
$\frac{\partial \bar{g}}{\bar{g}\partial(\Delta r)}$ (nm^{-1})	−4.82	−2.65	0.186	0.091
$\frac{\partial \bar{g}}{\bar{g}\partial(\Delta p)}$	−0.154	−0.132	−0.650	−0.538

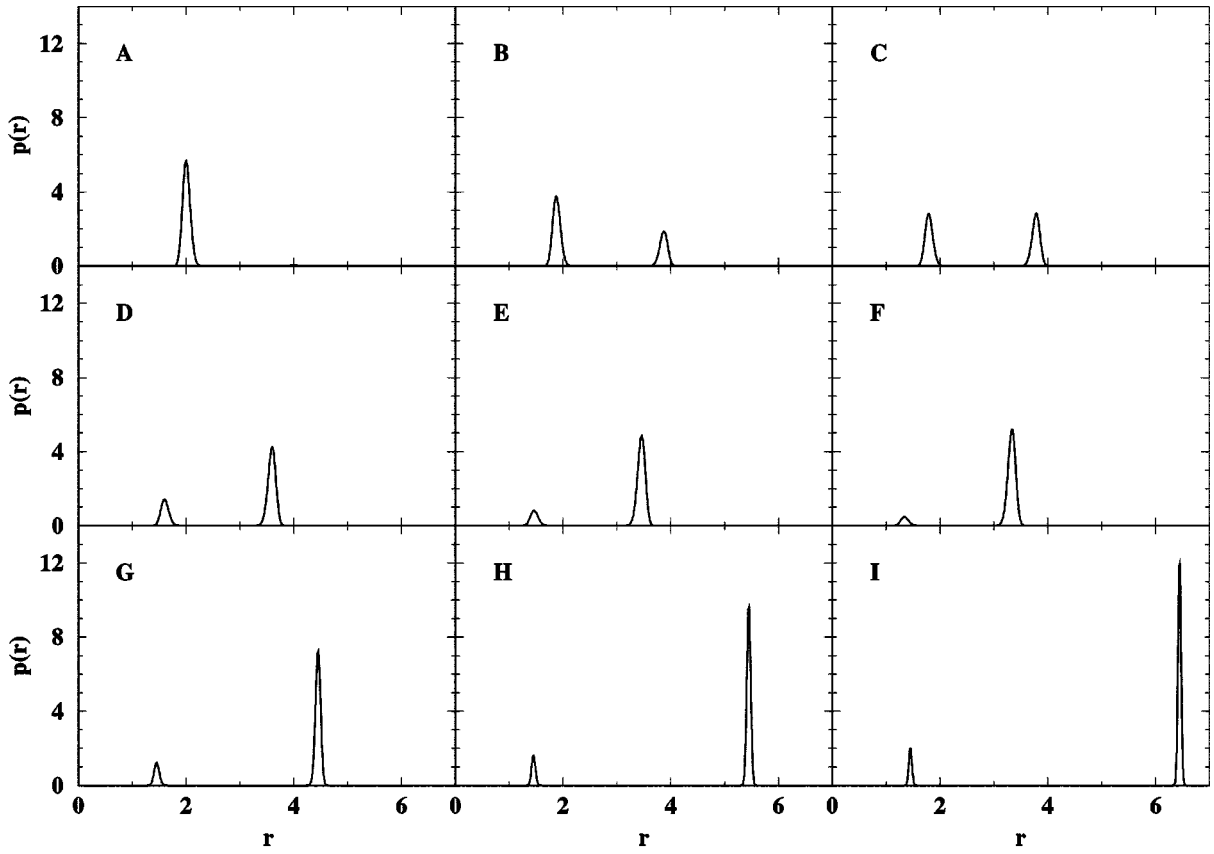

Figure 5. Distance distributions $p(r)$ that yield an average $\bar{g} = \left\langle r^{-6} \right\rangle = 2^{-6}$. The parameter values are $\beta = 100$ for (A–I), $\Delta r = 2$ for (A–F), $\Delta p = 5$ for (G–I), and $r_1 = 2.0$, $\Delta p = -0.99$ (A), $r_1 = 1.87$, $\Delta p = -0.5$ (B), $r_1 = 1.79$, $\Delta p = 0$ (C), $r_1 = 1.59$, $\Delta p = 2$ (D), $r_1 = 1.46$, $\Delta p = 5$ (E), $r_1 = 1.34$, $\Delta p = 10$ (F), $r_1 = 1.45$, $\Delta r = 3$ (G), $r_1 = 1.45$, $\Delta r = 4$ (H), $r_1 = 1.45$, $\Delta r = 5$ (I).

where $p(r)$ denotes the distribution of the degree of freedom r, and therefore Δp denotes the relative height difference of the two peaks of the bimodal distribution.

The derivative $(\partial \bar{g}/\partial(\Delta r))/\bar{g}$ for $\beta = 100$ is shown in Figure 2.

According to Equation 10, $|\Delta f| < \Delta r^4/12$ must be fulfilled. At the border surface $\Delta f = -\log(1 + \Delta p)/\beta = \Delta r^4/12$, the value of the derivative of $\bar{g}$ with respect to Δr goes to infinity for negative Δp and to minus infinity for positive Δp. Qualitatively, all the derivatives are negative. That means, when changing to a bigger Δr, the average $\left\langle r^{-6} \right\rangle$ gets smaller, i.e. the average distance $\left\langle r^{-6} \right\rangle^{-1/6}$ gets bigger. From Figure 2A, $\Delta p = 0$, we can see that changing Δr has hardly any influence on the average unless Δr is very small. The average is not defined at $\Delta r = 0$, and therefore, the derivative goes to minus infinity. If $\Delta p = 1$ (Figure 2B), the average is not defined for $\Delta r \leq 0.537$. Otherwise, the surface looks the same as the one in Figure 2A. The same situation is observed in Figure 2C ($r_1 = 2$): the derivative is nearly zero until it approaches the disallowed area for Δr. Figure 2D ($\Delta r = 2$) shows clearly what we have seen already in Figure 2A–C, namely that the value of the derivative only slightly depends on the choice of r_1 and Δp as long as $\Delta r \gg (12|\Delta f|)^{1/4}$.

Furthermore, if we want to extend this analysis to the observable $\left\langle r^{-6} \right\rangle^{-1/6}$, we must consider the Taylor expansion

$$\left(\bar{g} + \frac{\partial \bar{g}}{\partial(\Delta r)}\Delta(\Delta r)\right)^{-1/6} = \bar{g}^{-1/6} - \bar{g}^{-7/6}\frac{1}{6}\frac{\partial \bar{g}}{\partial(\Delta r)}\Delta(\Delta r) + \cdots \tag{15}$$

Therefore, the influence of a slight change in Δr on the average distance is even reduced more.

The dependence of $(\partial \bar{g}/\partial(\Delta r))/\bar{g}$ on r_1 and Δr for negative Δp is not separately shown because, as

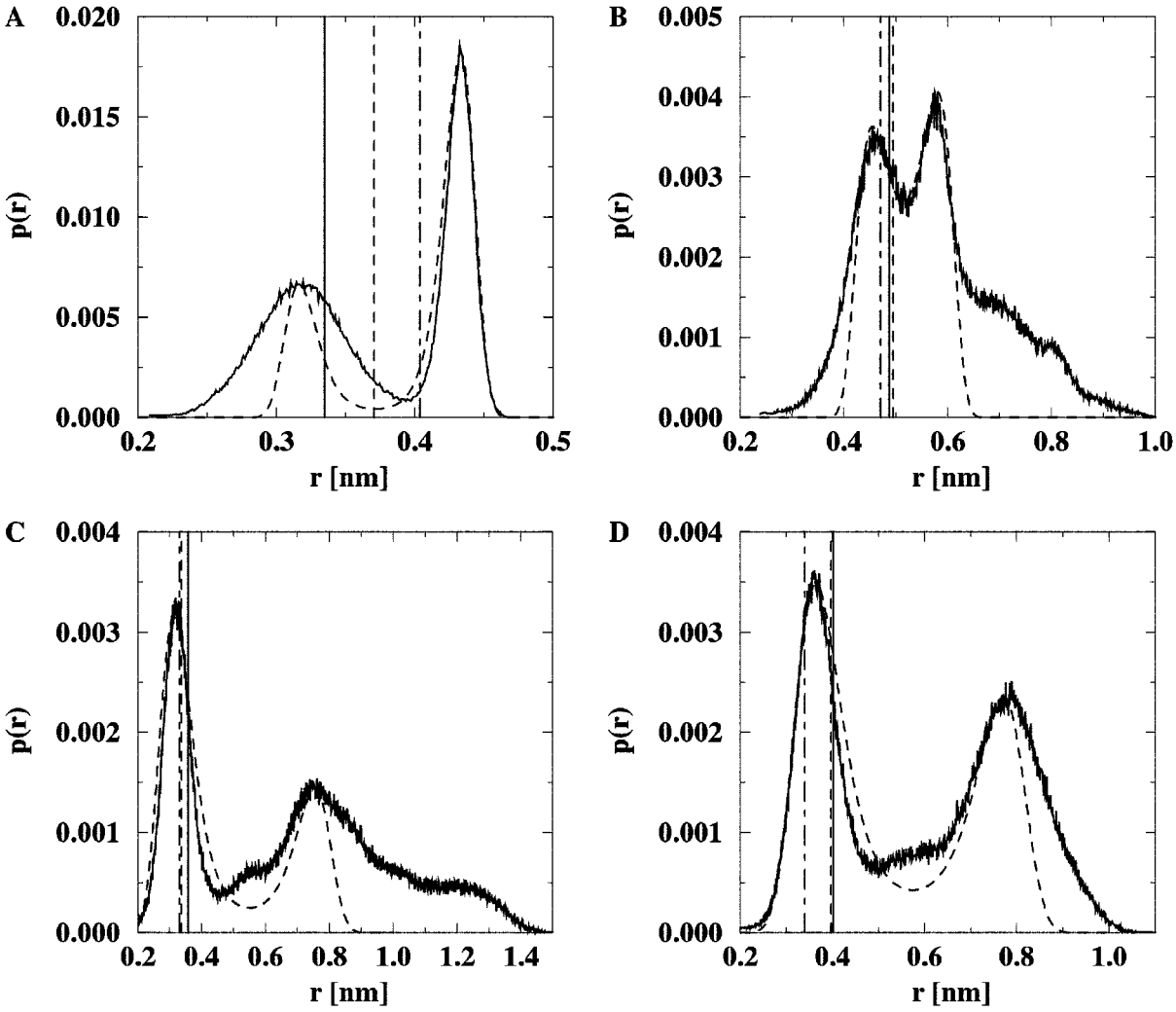

Figure 6. Four examples of atom-atom distance distributions taken from two MD simulations. (A) shows the distribution of the distance 1CB1-2HN between residues 1 and 2 and (B) the distribution of the distance 2CB2-5HN between residues 2 and 5 of an octapeptide in DMSO (150 ns at 298 K) (Bürgi et al., 2001), (C) the distribution of the distance HN to H-CB between residues 2 and 5 and (D) the distribution of the corresponding atoms of residues 3 and 5 of a β-heptapeptide in methanol (200 ns at 340 K) (Daura et al., 1998, 1999b). The solid curves are the distribution of the distances as taken from the MD simulations, the dashed curves are obtained by fitting a bimodal distribution to the simulated distributions. The solid line indicates the value of the r^{-6} average over the simulated distribution, the dashed line the average as calculated using the saddle-point approximation, and the dashed-dotted line the upper bound derived from NOE measurements.

can be inferred from Figures 2C and 2D ($r_1 = 2$ and $\Delta r = 2$ respectively), the influence of changing Δr is always about zero for negative Δp.

The derivative $(\partial \bar{g}/\partial(\Delta p))/\bar{g}$ for $\beta = 100$ is shown in Figure 3. As in the previous case, the values of the derivatives hardly depend on r_1 and Δr, as long as $\Delta r \gg (12|\Delta f|)^{1/4}$. The value of the derivative is around -1 for $\Delta p = -0.98$ (Figure 3B) and around -0.5 for $\Delta p = 0$ (Figure 3A). It vanishes for larger Δp (Figures 3C and 3D), as then the first maximum in the distribution dominates the average completely. We can conclude that the r^{-6} average is even more insensitive to changes in Δp than to changes in Δr.

Analysing the space of parameter values that yield the same average is even more interesting than investigating the derivatives of $\bar{g}$. The space that yields the average $\bar{g} = 2^{-6}$ is shown for $\beta = 100$ in Figure 4. For each pair Δr and Δp, a value for r_1 can be found that yields the average $\bar{g} = 2^{-6}$, as long as Equation 10 is fulfilled. As Δp approaches -1, r_1 approaches the value 2. This is clear, as in this case, there would be only one maximum in the distribution. It is also

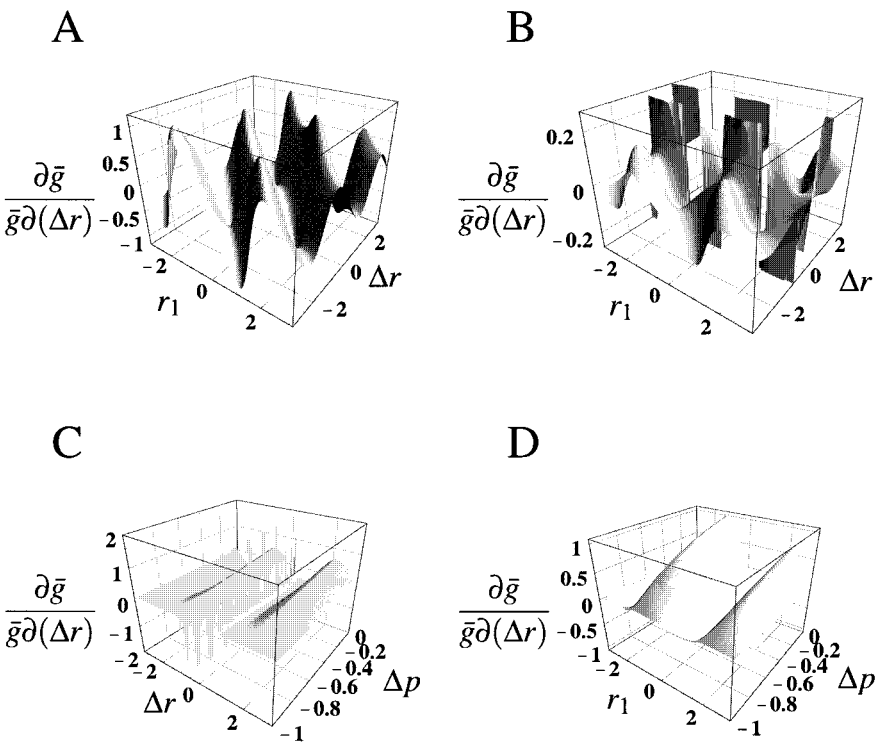

Figure 7. $(\partial \bar{g}/\partial(\Delta r))/\bar{g}$ for the bimodal distribution (Equations 4 and 5) and averaging Equation 16 with $\beta = 100$. (A) shows the derivative as a function of r_1 and Δr at $\Delta p = 0$, in (B), the value of Δp is -0.95. (C) shows the derivative as a function of Δr and Δp at $r_1 = 0$, and (D) as a function of r_1 and Δp at $\Delta r = -\pi$.

Table 2. Four example torsion-angle distributions taken from two MD simulations (Bürgi et al., 2001; Daura et al., 1998, 1999b). Indicated are the parameters r_1, Δr, Δf, and β (Equations 4 and 5) of a bimodal distribution obtained by fitting such a distribution to the simulated distributions. The last five lines show the average ^3J-coupling constant $\bar{r}$ (MD) calculated over the simulated distribution, the average ^3J-coupling constant $\bar{r}$ (SPA) calculated using the saddle-point approximation (SPA) to the bimodal distribution, the experimentally determined ^3J-coupling constant $\bar{r}$ (exp.), and the value of the derivative of the average $\bar{g} = \left\langle a\cos^2(r) + b\cos(r) + c \right\rangle$ with respect to Δr and Δf of the bimodal distribution

	Octapeptide in DMSO 150 ns		Heptapeptide in MeOH 200 ns	
Residue	2Aib	6Leu	2Ala	6Ala
Torsion angle	H-N-CA-CB	H-N-CA-HA	C-N-CB-CA	N-CB-CA-C
r_1 (rad)	-1.94	-2.165	-3.05	1.18
Δr (rad)	1.60	-0.963	3.18	1.71
Δf (rad^4)	2.66×10^{-2}	2.08×10^{-3}	1.09	0.115
β (rad^{-4})	14.4	65.2	3.96	16.4
$\bar{r}$ (MD) (Hz)	1.59	6.99	8.93	4.11
$\bar{r}$ (SPA) (Hz)	1.62	7.06	9.58	4.04
$\bar{r}$ (exp.) (Hz)	–	6.8	9.2	3.9
$\frac{\partial \bar{g}}{\bar{g}\partial(\Delta r)}$ (rad^{-1})	0.60	8.0×10^{-3}	-2.1×10^{-3}	6.9×10^{-2}
$\frac{\partial \bar{g}}{\bar{g}\partial(\Delta p)}$	0.31	6.3×10^{-2}	-8.2×10^{-2}	1.9

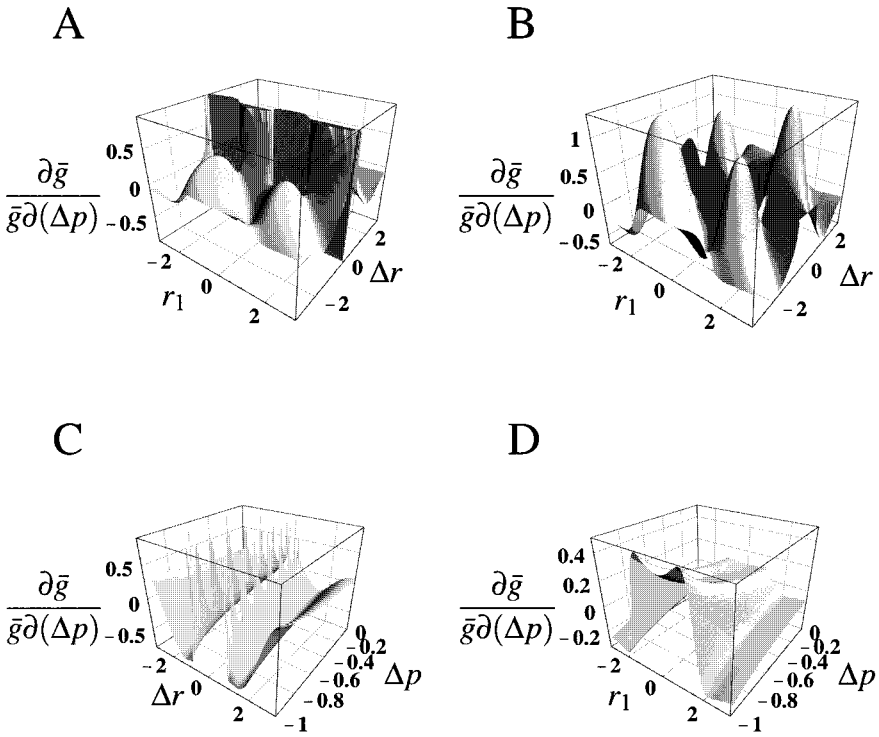

Figure 8. $(\partial\bar{g}/\partial(\Delta p))/\bar{g}$ for the bimodal distribution (Equations 4 and 5) and averaging Equation 16 with $\beta = 100$. (A) shows the derivative as a function of r_1 and Δr at $\Delta p = 0$, in (B), the value of Δp is -0.1. (C) shows the derivative as a function of Δr and Δp at $r_1 = 0$, and (D) as a function of r_1 and Δp at $\Delta r = -\pi$.

obvious that the larger Δp is, the smaller r_1 has to be to compensate for the larger (second) maximum in the distribution. What is rather surprising, however, is that even for $\Delta p = 5$, the value of r_1 is hardly dependent on Δr. This is another effect of the dominant contribution of low distances within the r^{-6} averaging.

As the average $\bar{g}$ is hardly dependent on Δr, we can generate very different distributions that each yield the average $\bar{g} = 2^{-6}$. A selection of such distributions is shown in Figure 5. Figure 5A shows the situation as it is expected to yield a single r^{-6} weighted average distance of 2. However, the distributions can be very different as Figures 5B–I illustrate. Looking at Figures 5F–I, for instance, it is intuitively hard to believe that the average distance is $\langle r^{-6}\rangle^{-1/6} = 2$. All examples in Figure 5 are quite moderate – it is easy to pick more extreme examples by taking large values for Δp and Δr.

Examples from molecular dynamics simulations

To illustrate the analysis of the previous section and emphasise its relevance, it is of interest to show the distributions of distances obtained through molecular dynamics (MD) simulations. Figure 6 shows four distributions of distances obtained from two simulations (Daura et al., 1998, 1999b; Bürgi et al., 2001).

The location of the two maxima of the MD distribution and thus the parameters r_1, Δr, and Δf of the bimodal distribution (Equations 4 and 5) was determined by a polynomial fit. The parameter β was obtained by a nonlinear fit to the MD distribution. The data for the fitted bimodal distribution curves as well as the calculated averages of Equation 13, the experimentally determined NOE distance bounds and the value of the derivatives of the averaging function with respect to the parameters Δr and Δp of the fitted bimodal distribution are given in Table 1.

We see that β is large enough for the saddle-point approximation to be valid. Furthermore, it is interesting to note the difference between the two peptides: Whereas the octapeptide has in most distributions a positive Δp, the heptapeptide has for the crucial distributions mostly a negative Δp. This might be one of the reasons why the NOE distances for the latter as calculated from the MD trajectories match the experi-

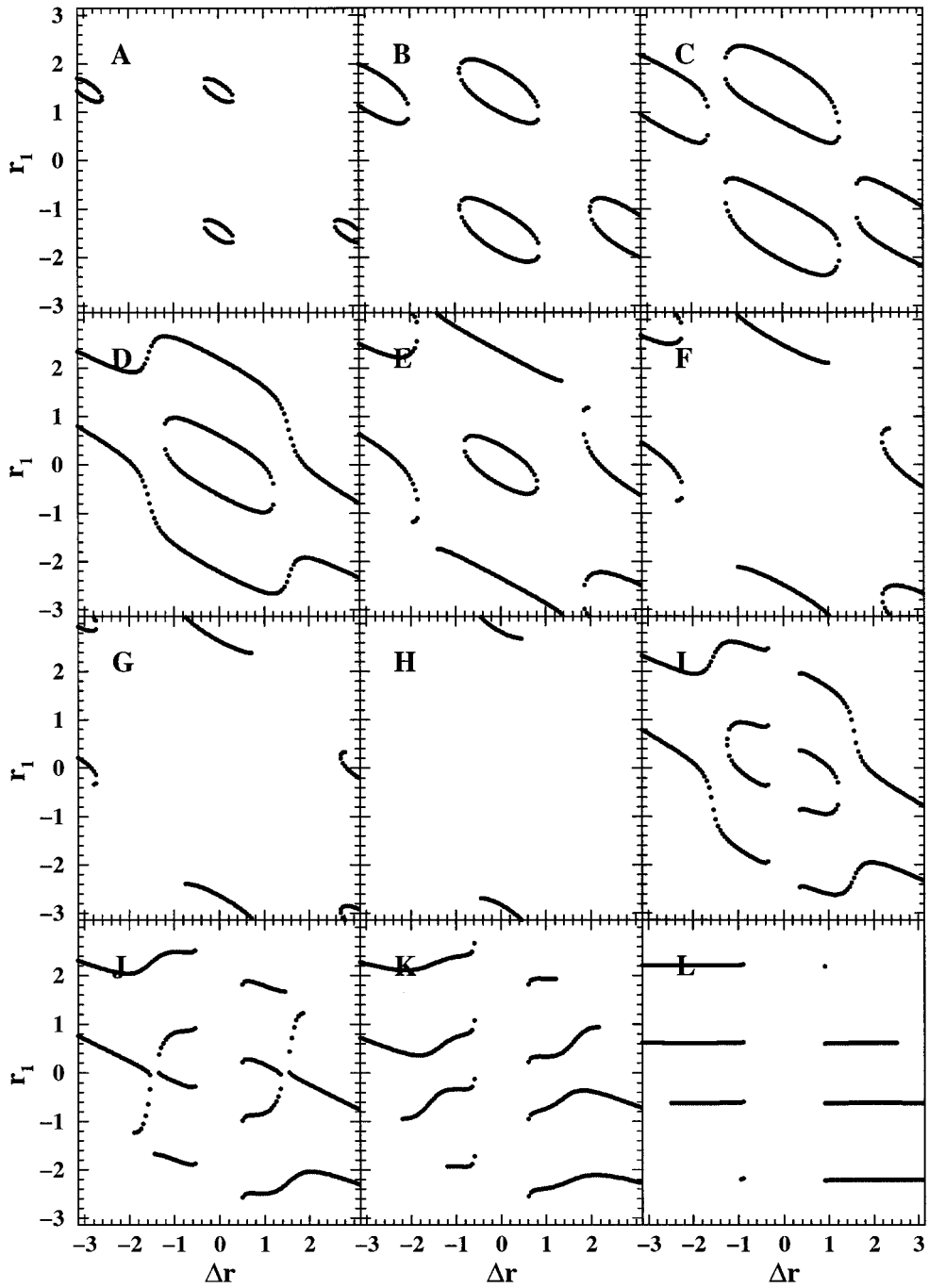

Figure 9. Parameter values defining bimodal torsion-angle distributions with $\beta = 100$ and $\Delta p = 0$ that yield the averages $\bar{g} = 2$ (A), $\bar{g} = 3$ (B), $\bar{g} = 4$ (C), $\bar{g} = 5$ (D), $\bar{g} = 6$ (E), $\bar{g} = 7$ (F), $\bar{g} = 8$ (G), $\bar{g} = 9$ (H), as well as of the bimodal distributions that yield the average $\bar{g} = 5$ for $\beta = 100$ and $\Delta p = -0.1$ (I), $\Delta p = -0.39$ (J), $\Delta p = -0.63$ (K), and $\Delta p = -0.99$ (L).

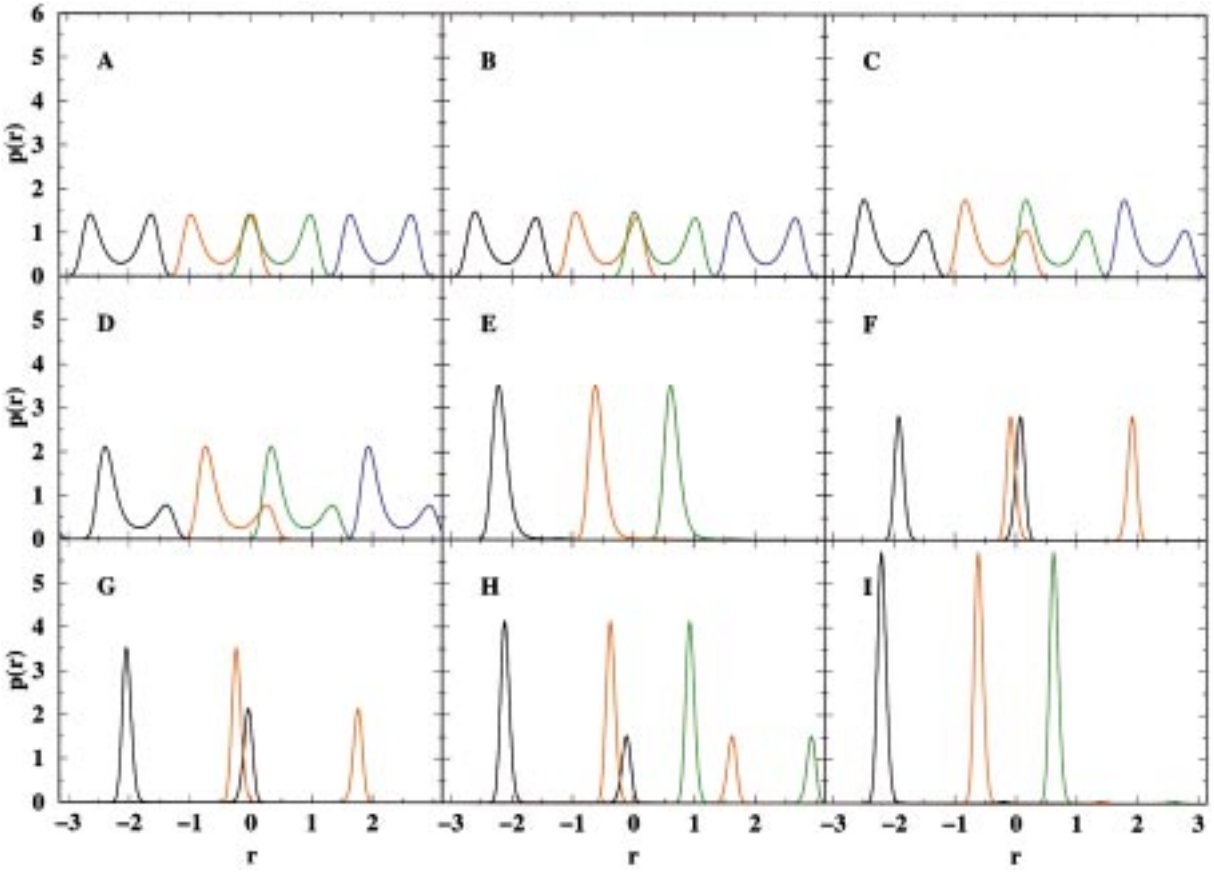

Figure 10. Torsion-angle distributions $p(r)$ that yield the average $\bar{g} = \left\langle a\cos^2 r + b\cos r + c \right\rangle = 5$. The parameter values are $\beta = 100$ for (A–I), $\Delta r = 1$ for (A–E), $\Delta r = 2$ for (F–I) and $\Delta p = 0$ (A), $\Delta p = -0.1$ (B), $\Delta p = -0.39$ (C), $\Delta p = -0.63$ (D), $\Delta p = -0.99$ (E), $\Delta p = 0$ (F), $\Delta p = -0.39$ (G), $\Delta p = -0.63$ (H), $\Delta p = -0.99$ (I).

mental bounds better; the averages are less sensitive to changes in the distributions, which is also indicated by the smaller values of the derivatives of $\bar{g}$ with respect to Δr and Δp for the heptapeptide.

Application of the formalism to ^3J-coupling constants

In this section the derived formalism is used to analyse the effect onto the ^3J-coupling constants using the averaging function

$$g(r) = a\ \cos^2 r + b\ \cos\ r + c. \tag{16}$$

For consistency with the previous sections we are using the same notation: r will in the case of ^3J-coupling constants denote the corresponding torsion angle. As the analysis yields only small qualitative differences for different Karplus parameters a, b, and c, all numerical examples are based on the following Karplus parameters for a H-N-C_α-H_α torsion angle: $a = 6.4$ Hz, $b = -1.4$ Hz, $c = 1.9$ Hz. Furthermore, it is sufficient to discuss cases for $\Delta p < 0$ ($\Delta f > 0$), since the absolute maximum of the distribution can always be denoted as r_1 with Δr being chosen accordingly, due to the periodicity of the cosine function.

As in the previous section, we shall first analyse the derivatives of the averages $\bar{g}$ with respect to Δr and Δp. Figure 7 shows the derivative $(\partial\bar{g}/\partial(\Delta r))/\bar{g}$, and Figure 8 shows the derivative $(\partial\bar{g}/\partial(\Delta p))/\bar{g}$, both at $\beta = 100$.

Unlike in the case of NOEs, in the case of ^3J-coupling constants both derivatives can be either positive or negative. The sign and magnitude of the derivatives are dependent on the choice of r_1 and Δr. Both derivatives depend hardly on Δp, $(\partial\bar{g}/\partial(\Delta r))/\bar{g}$ approaches 0 for $\Delta p = -1$ and has its greatest values for $\Delta p = 0$, whereas $(\partial\bar{g}/\partial(\Delta p))/\bar{g}$ has its greatest

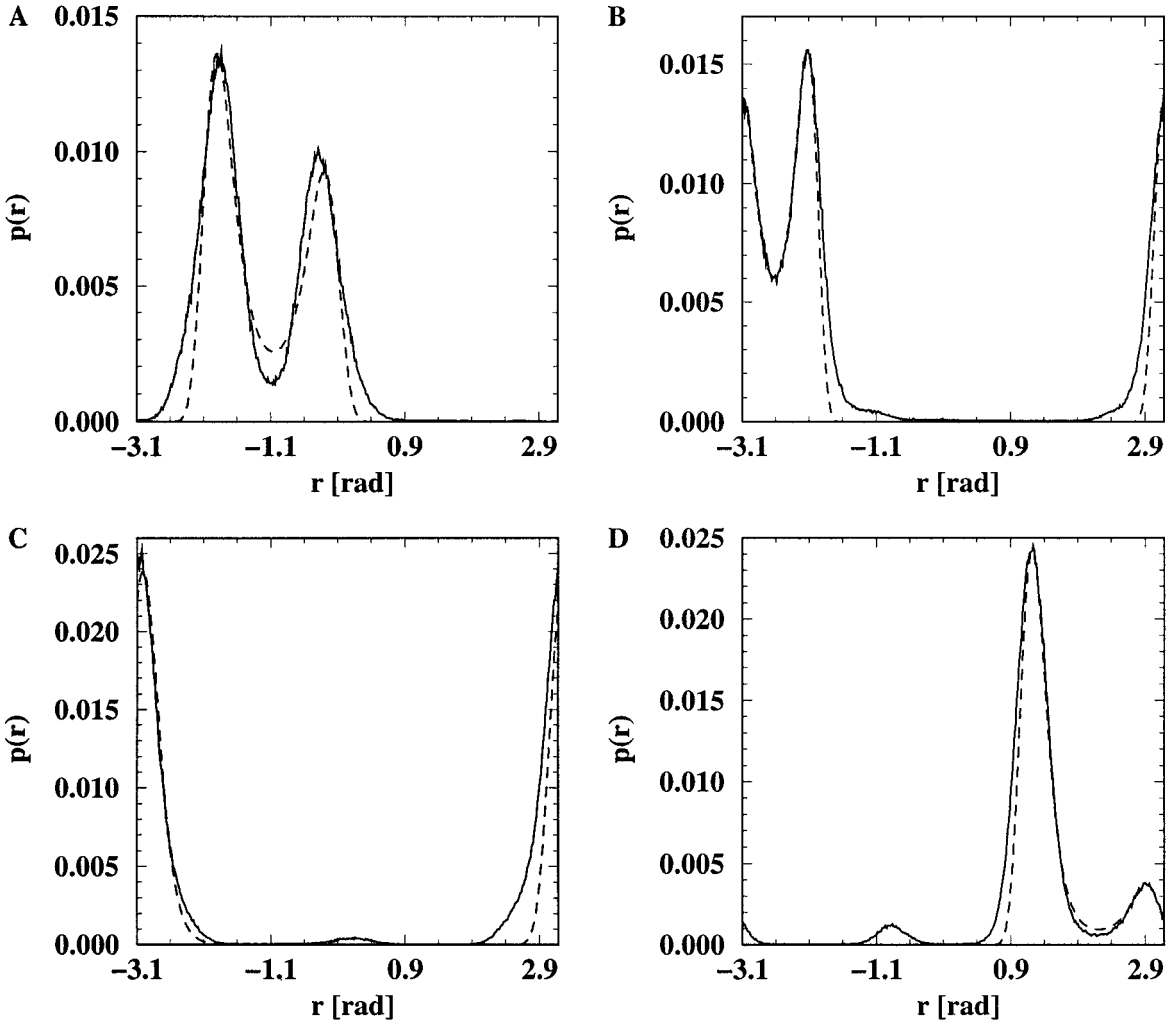

Figure 11. Four examples of torsion-angle distributions taken from two MD simulations. (A) shows the distribution of the torsion angle H-N-CA-CB of residue 2Aib and (B) the distribution of the torsion angle H-N-CA-HA of residue 6Leu of the octapeptide in DMSO (150 ns at 298 K) (Bürgi et al., 2001), (C) the distribution of the torsion angle C-N-CB-CA of residue 2Ala and (D) the distribution of the torsion angle N-CB-CA-C of residue 6Ala of the β-heptapeptide in methanol (200 ns at 340 K) (Daura et al., 1998, 1999b). The solid curves are the distribution of the torsion angles as taken from the MD simulations, the dashed curves are obtained by fitting bimodal distributions to the simulated distributions.

absolute values for $\Delta p = -1$ (Figures 8C and 8D). Both derivatives go to plus or minus infinity when Δr approaches the disallowed region defined by Equation 10 (Figures 8B and 8C). $(\partial \bar{g}/\partial(\Delta p))/\bar{g}$ is not defined at $\Delta r = 0$. However, apart from the disallowed regions, we can conclude the same as for the r^{-6} averaging: The magnitude of both derivatives is rather small, namely between -1 and 1. Therefore, the influence of slightly changing the torsion-angle distributions will generally be small.

As for the NOEs, we have calculated a set of torsion-angle distributions for several average ^3J-coupling constant values (see Figure 9). As we can see, there are several choices for torsion angle distributions that yield the same average ^3J-coupling constant. A selection of torsion angle distributions that yield the average $\bar{g} = 5$ is shown in Figure 10. It gives an impression of how diverse these distributions can be and yet have the same average ^3J-coupling constant value.

Examples from molecular dynamics simulations

As for the r^{-6} averaging, we illustrate the analysis of the previous section with torsion-angle distributions obtained from the same two simulations (Bürgi et al., 2001; Daura et al., 1998, 1999b). Figure 11 shows four examples of torsion-angle distributions taken from these simulations. To obtain the bimodal distributions, the same fitting procedure as for the distance distributions was applied. The data for the fitted bimodal distribution curves as well as the calculated averages of Equation 16, the experimentally determined ^3J-coupling constants and the value of the derivatives of the averaging function with respect to the parameters Δr and Δp of the bimodal distribution are shown in Table 2.

Even though β is much smaller for all example torsion-angle distributions than it was for the NOE distances, the saddle-point approximation still seems to be quite accurate. All four averages are very insensitive to changes of the distributions. All derivatives of the average with respect to the distribution parameters are of the order of 10^{-3} to 10^{1}. For small changes in the distribution, the average will not change significantly.

Conclusions

We have presented a formalism to analyse a general averaging function $g(r)$, which represents an observable, in terms of sensitivity of the average of the observable to small changes in the distribution of the degree of freedom r. The formalism is based on the saddle-point approximation, which yields as the average over a bimodal distribution a weighted average, where the weights do not depend on the averaging function. Therefore, it is straightforward to calculate derivatives of the weighted average with respect to the parameters of the bimodal distribution.

For the two examples of averaging functions given here, $g(r) = r^{-6}$ (NOEs) and $g(r) = a\ \cos^2(r) + b\ \cos(r) + c$ (^3J-coupling constants), it was shown that the averages are not very sensitive to a variety of changes in the distribution of r. Furthermore, we have calculated the range of parameters that yield the same average value and shown the diversity of distributions that yield the same average value. For the cases we have studied, this implies that an experimentally averaged value does not contain much information on the underlying distribution of molecular conformations.

It is our expectation that the approach outlined in this paper will be generally useful as a quantitative way both to assess experimental data or simulation results and as a way to deepen and make more precise the connection between computer simulations and experimental data.

Acknowledgements

Financial support was obtained from the Schweizerischer Nationalfonds, project number 21-57069.99, which is gratefully acknowledged. R.B. thanks Dr Xavier Daura for providing his data on the heptapeptide and Prof. Herman Berendsen for useful discussions on the subject.

Appendix A: Saddle-point approximation

The formulation of the theorem as well as the proof follow the ideas given in (Jänich, 1983).

Theorem. *Let $g(r)$ and $f(r)$ be functions that are sufficiently many times differentiable on an interval (a, b). The cases $a = -\infty, b = +\infty$ are also allowed. The function $f(r)$ should be real and have a non-degenerate absolute minimum at $r_0 \in (a, b)$. Furthermore, there should exist $\delta > 0$ and $\epsilon > 0$ so that $f(r)$ is decreasing monotonically on the interval $(r_0 - \epsilon, r_0)$ and increasing monotonically on the interval $(r_0, r_0 + \epsilon)$ and that $f(r) \geq f(r_0) + \delta$ for all r outside the interval $(r_0 - \epsilon, r_0 + \epsilon)$. $g(r)$ should be chosen such that $g(r_0) \neq 0$ and that $\int_a^b |g(r)|e^{-\beta f(r)}\,\mathrm{d}r$ exists for a $\beta = \beta_0$ (and therefore for all $\beta \geq \beta_0$). The following expression is then valid:*

$$\int_a^b g(r)e^{-\beta f(r)}\,\mathrm{d}r = g(r_0)e^{-\beta f(r_0)}\left[\sqrt{\frac{2\pi}{f''(r_0)\cdot\beta}} + \mathcal{O}\left(\frac{1}{\beta^{3/2}}\right)\right] \qquad (17)$$

Proof. All contributions of the integral outside the interval $(r_0 - \epsilon, r_0 + \epsilon)$ can be neglected, as their absolute value is $\leq e^{-\delta\beta}e^{\delta\beta_0} \cdot \int_a^b |g(r)|e^{-\beta_0 f(r)}dr$. These contributions are therefore absorbed in the error term $\mathcal{O}(\beta^{-3/2})$. $f(r)$ can be expanded around r_0: $f(r) = f(r_0) + \frac{1}{2}f''(r_0)(r - r_0)^2 +$ higher-order terms. Let us call $f''(r_0) = c^2$, since $f(r_0)$ is a non-degenerate minimum. If we also expand $g(r) =$

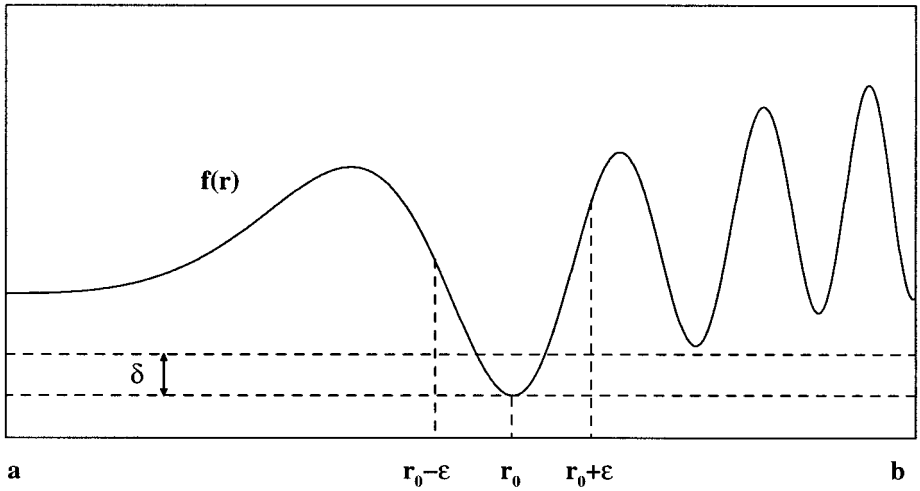

Figure 12. Conditions for $f(r)$ are that $f(r) \geq f(r_0) + \delta$ for all r outside the interval $(r_0 - \epsilon, r_0 + \epsilon)$.

$g(r_0) + c_1(r - r_0) + (r - r_0)^2\psi(r)$ around r_0, all that remains to be solved is the integral

$$\begin{aligned}
&\int_{r_0-\epsilon}^{r_0+\epsilon} g(r)e^{-\beta f(r_0)-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r = \\
&g(r_0)\int_{r_0-\epsilon}^{r_0+\epsilon} e^{-\beta f(r_0)-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r \\
&+c_1\int_{r_0-\epsilon}^{r_0+\epsilon}(r-r_0)e^{-\beta f(r_0)-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r \\
&+\int_{r_0-\epsilon}^{r_0+\epsilon}(r-r_0)^2\psi(r)e^{-\beta f(r_0)-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r.
\end{aligned} \tag{18}$$

The first term is approximated by

$$\begin{aligned}
&g(r_0)e^{-\beta f(r_0)}\int_{r_0-\epsilon}^{r_0+\epsilon} e^{-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r \simeq g(r_0)e^{-\beta f(r_0)} \\
&\int_{-\infty}^{\infty} e^{-\frac{1}{2}\beta c^2 r^2}\,\mathrm{d}r = g(r_0)e^{-\beta f(r_0)}\sqrt{\frac{2\pi}{\beta c^2}}.
\end{aligned} \tag{19}$$

The second term vanishes, as the integrand is anti-symmetric around r_0. Since $\psi(r)$ is bounded in the interval $(r_0 - \epsilon, r_0 + \epsilon)$, i.e. $\psi(r) \leq c_2$ with $c_2 > 0$, the third term is approximated by

$$\begin{aligned}
&e^{-\beta f(r_0)}\int_{r_0-\epsilon}^{r_0+\epsilon}(r-r_0)^2\psi(r)e^{-\frac{1}{2}\beta c^2(r-r_0)^2}\,\mathrm{d}r \\
&\simeq c_2 e^{-\beta f(r_0)}\int_{-\infty}^{\infty} r^2 e^{-\frac{1}{2}\beta c^2 r^2}\,\mathrm{d}r = \frac{c_2}{\beta c^2}\sqrt{\frac{2\pi}{\beta c^2}}.
\end{aligned} \tag{20}$$

Therefore, the third term is also included in the error term $\mathcal{O}(\beta^{-3/2})$.

References

Bennett, W.S. and Steitz, T.A. (1980) *J. Mol. Biol.*, **140**, 183–230.

Beutler, T.C., Bremi, T., Ernst, R.R. and van Gunsteren, W.F. (1996) *J. Phys. Chem.*, **100**, 2637–2645.

Bonvin, A.M.J.J. and Brunger, A.T. (1996) *J. Biomol. NMR*, **7**, 72–76.

Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1994) *J. Biomol. NMR*, **4**, 143–149.

Braun, W., Boesch, C., Brown, L.R., Go, N. and Wüthrich, K. (1981) *Biochim. Biophys. Acta*, **667**, 377–396.

Bruschweiler, R., Roux, B., Blackledge, M., Griesinger, C., Karplus, M. and Ernst, R.R. (1992) *J. Am. Chem. Soc.*, **114**, 2289–2302.

Bürgi, R., Daura, X., Mark, A., Bellanda, M., Mammi, S., Peggion, E. and van Gunsteren, W.F. (2001) *J. Pept. Res.*, **57**, 107.

Daura, X., Antes, I., van Gunsteren, W.F., Thiel, W. and Mark, A. (1999a) *Proteins*, **36**, 542–555.

Daura, X., Jaun, B., Seebach, D., van Gunsteren, W.F. and Mark, A.E. (1998) *J. Mol. Biol.*, **280**, 925–932.

Daura, X., van Gunsteren, W.F. and Mark, A.E. (1999b) *Proteins*, **34**, 269–280.

Ho, T.S. and Rabitz, H. (1993) *J. Phys. Chem.*, **97**, 13447–13456.

Jänich, K. (1983) *Analysis für Physiker und Ingenieure: Funktionentheorie, Differentialgleichungen, spezielle Funktionen*. Springer-Verlag, Berlin.

Jardetzky, O. (1980) *Biochim. Biophys. Acta*, **621**, 227–232.

Jardetzky, O. and Roberts, G.C.K. (1981) *NMR in Molecular Biology*, Academic Press, New York, NY, Chapter 4.

Lazarides, A.A., Rabitz, H. and McCourt, F.R.W. (1994) *J. Chem. Phys.*, **101**, 4735–4749.

McCammon, J.A., Gelin, B.R. and Karplus, M. (1977) *Nature*, **267**, 585–590.

Moffat, K., Deatherage, J.F. and Seybert, D.W. (1979) *Science*, **206**, 1035–1042.

Nie, S.M., Chiu, D.T. and Zare, R.N. (1994) *Science*, **266**, 1018–1021.

Philippopoulos, M. and Lim, C. (1994) *J. Phys. Chem.*, **98**, 8264–8273.

Stocker, U., Spiegel, K. and van Gunsteren, W.F. (2000) *J. Biomol. NMR*, **18**, 1–12.

Syberts, S.G., Maerki, W. and Wagner, G. (1987) *Eur. J. Biochem.*, **164**, 625–635.

Utz, M. (1998) *J. Chem. Phys.*, **109**, 6110–6124.

van Gunsteren, W.F., Bonvin, A.M.J.J., Daura, X. and Smith, L.J. (1999) In *Structure Computation and Dynamics in Protein NMR* (Eds, Krishna, N.R. and Berliner, L.J.), Vol. 17 of *Biol. Magnetic Resonance*, Plenum Publishers, New York, NY, pp. 3–35.

van Gunsteren, W.F., Brunne, R.M., Gros, P., van Schaik, R.C., Schiffer, C.A. and Torda, A.E. (1994) *Methods Enzymol.*, **239**, 619–654.

Zhou, H.X., Wlodek, S.T. and McCammon, J.A. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 9280–9283.