



Sampling of conformation space in torsion angle dynamics calculations

Peter Güntert*, Kurt Wüthrich

Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule Hönggerberg, CH-8093 Zürich, Switzerland

Received 16 January 2001; accepted 22 February 2001

Abstract

Torsion angle dynamics (TAD) performs molecular dynamics simulation using torsion angles instead of Cartesian coordinates as degrees of freedom. TAD algorithms used in conjunction with simulated annealing are one of the common methods for the calculation of three-dimensional protein structures from NMR data. For this application of TAD, unbiased sampling of conformation space is essential. This paper presents a systematic study of the sampling of conformation space in protein structure calculations with the TAD algorithm implemented in the program DYANA, and compares the results with those obtained with a different TAD algorithm in the program CNS. Examples used are unconstrained poly-alanine peptides of length 20 to 100 residues, and the globular protein *Antennapedia*(C39S) homeodomain, which comprises unstructured polypeptide segments at the two chain termini and was calculated from a high-quality experimental NMR data set. The results show that the different implementations of TAD all have good sampling properties for calculating protein structures that are well-constrained by experimental NMR data. However, if TAD is used for studies of long unconstrained polypeptides, the results obtained in this paper show that the molecule needs to reorient freely in space, and that the total angular and linear momenta of the system are conserved and periodically reset to zero. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Conformational sampling; Conformation space; Torsion angle dynamics; Angular momentum; NMR structure calculation; Polypeptide; Protein structure; Molecular dynamics simulation

Abbreviations: Ala₂₀, Ala₄₀, Ala₆₀ and Ala₁₀₀, polypeptide chains consisting of 20, 40, 60 and 100 L-alanyl residues, respectively; *Antp*(C39S), *Antennapedia* homeodomain with cysteine 39 replaced by serine; *Antp*(C39S) [-10–77] and *Antp*(C39S)[-20–87], *Antp*(C39S) polypeptides elongated by Ala₁₀ or Ala₂₀ segments at both chain ends; NOE, nuclear Overhauser effect; REDAC, use of redundant dihedral angle constraints; TAD, torsion angle dynamics; TF, target function; VTF, variable target function; three software packages for NMR structure calculation: CNS, crystallography and NMR system; DIANA, distance geometry algorithm for NMR applications; DYANA, dynamics algorithm for NMR applications.

* Corresponding author.

E-mail address: guentert@mol.biol.ethz.ch (P. Güntert).

1. Introduction

Molecular dynamics simulations are an important computational tool for the study of biological macromolecules [e.g., [1]]. Generating trajectories by numerical integration of the equations of motion of classical mechanics is instrumental in “standard” molecular dynamics simulations, which try to simulate the behavior of a protein molecule in its natural environment, as well as in structure calculation techniques based on simulated annealing. The latter are commonly used as efficient minimization methods in *ab initio* structure calculations from experimental X-ray or NMR data [2,3].

Molecular dynamics simulations have a high demand of computing power [e.g., [4,5]], since trajectories have to be generated in sufficiently short time-steps to follow the fastest motions in the system. Torsion angle dynamics (TAD) allows to increase the length of the time steps by up to one order of magnitude when compared to Cartesian space molecular dynamics, since the restriction to the torsion angles as the only degrees of freedom eliminates the highest frequency motions, i.e. oscillations of covalent bond lengths and bond angles [6,7]. This is particularly advantageous when TAD is used as a NMR structure calculation method, since in these applications the trajectory is driven dominantly by the constraints derived from experimental NMR measurements and, hence, possible biasing effects from freezing of the high-frequency motions are negligibly small.

The NMR solution structure of a protein is commonly represented by a bundle of conformers that have been calculated using the same experimental input data and the same computational protocol but starting from different, randomly chosen initial conditions [8]. The rationale for this representation of solution structures is that the precision with which the experimental data define the three-dimensional structure can be estimated from the spread among the individual conformers. Obviously, this approach gives a realistic picture only if the structure calculation algorithm performs uniform sampling of the conformation space that is accessible to the molecule with the given set of experimental constraints. Therefore, considerable effort has been directed at investigations of conformational sampling by different computational techniques used for NMR structure calculations. In this paper we extend these earlier studies, which focused on metric matrix distance geometry [9–12], the variable target function method [13, 14], simulated annealing by molecular dynamics in Cartesian space [e.g., [2]], and the TAD method, which has been implemented in the programs DYANA [15], XPLOR [16] and CNS [17], and currently appears to be the most efficient way to calculate NMR structures of biological macromolecules [3].

The program DYANA has been written for efficient calculation of protein structures from comprehensive sets of conformational constraints derived from NMR experiments, which define a unique molecular architecture with limited uncertainties in the atom positions. We previously showed that for such applications the sampling of the allowed conformation space by the TAD algorithm implemented in DYANA compares well with that provided by other algorithms in use for NMR structure calculation of proteins [15]. The present systematic study of conformational sampling by TAD was motivated by observations with atypical proteins, in which one or several long, contiguous segments of the polypeptide chain are only marginally constrained by NMR data. A typical example are mammalian prion proteins, which contain an unconstrained “tail” of about 100 residues [18–21]. The sampling of conformation space for this tail by the original TAD implementation in DYANA [15] was markedly non-uniform, leading to elongated structures. Considering the practical applications of DYANA and other programs with TAD implementations [16,17] for calculations with well-constrained systems, the observation of limited sampling for this extreme situation may appear to be of little practical interest. However, since it could not be excluded *a priori* that the cause of the non-uniform sampling for unconstrained chains might influence, in subtle and as yet unnoticed ways, also the outcome of structure calculations for well-constrained systems, there was a fundamental interest in a detailed investigation.

2. Methods

2.1. TAD implementation in DYANA that enables free global reorientation of the molecule

In this section a modification of the original implementation of the TAD algorithm in DYANA [15] is described that allows for free global reorientation of the molecule (see the *Results and Discussion* section below). The molecule is represented as a tree structure consisting of $n + 1$ “clusters”, which are connected by n rotatable bonds [3,15]. Each cluster is a rigid body made up of one or several mass points (atoms) with fixed relative positions. The only degrees of freedom are rotations about single bonds, and parameters that define the position and orientation of the molecule in space. The clusters are numbered from 0 to n . The base cluster has the number $k = 0$ and is located at one end of the polymer chain, for example, in proteins typically at the N-terminus. Each of the other clusters, with $k \geq 1$, has a single nearest neighbor in the direction toward the base, which has a number $p(k) < k$. The following quantities are defined for each cluster k (see Fig. 14 in [3]): the “reference point”, \mathbf{r}_k , which is the position vector of the end point of the bond between the clusters $p(k)$ and k ; $\mathbf{v}_k = \dot{\mathbf{r}}_k$, the velocity of the reference point; ω_k , the angular velocity of the cluster; \mathbf{Y}_k , the vector from the reference point to the center of mass of the cluster; m_k , the mass of the cluster k ; I_k , the inertia tensor of the cluster k with respect to the reference point. $I_k = \sum_{\alpha} m_{\alpha} I(\mathbf{y}_{\alpha})$, where the sum runs over all atoms in the cluster k , m_{α} is the mass of the atom α , \mathbf{y}_{α} is the vector from the reference point of cluster k to the atom α , and $I(\mathbf{y}_{\alpha})$ is the symmetric 3×3 matrix defined by the relation $I(\mathbf{y})\mathbf{x} = \mathbf{y} \wedge (\mathbf{x} \wedge \mathbf{y})$ for all three-dimensional vectors \mathbf{x} . All position vectors are in an inertial frame of reference that is fixed in space.

The position and orientation of the molecule in the inertial frame are specified by the three Cartesian coordinates of the reference point of the base cluster, \mathbf{r}_0 , and by 4 quaternion parameters $q = (q_0, q_1, q_2, q_3)$, which are subject to the normalization condition $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$. The rotation matrix that describes the orientation of the base cluster with respect to the inertial frame is given in terms of the quaternion parameters by [22]

$$R(q) = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(-q_0q_3 + q_1q_2) & 2(q_0q_2 + q_1q_3) \\ 2(q_0q_3 + q_1q_2) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(-q_0q_1 + q_2q_3) \\ 2(-q_0q_2 + q_1q_3) & 2(q_0q_1 + q_2q_3) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}. \quad (1)$$

The following two relations hold between the angular velocity of the base cluster, ω_0 , and the time-derivative of the quaternion parameters, $\dot{q} = dq/dt$ [22]:

$$\omega_0 = 2A(q)\dot{q} \quad (2)$$

and

$$\dot{q} = \frac{1}{2}A(q)^T\omega_0, \quad (3)$$

where the superscript T denotes the transpose and

$$A(q) = \begin{bmatrix} -q_1 & q_0 & -q_3 & q_2 \\ -q_2 & q_3 & q_0 & -q_1 \\ -q_3 & -q_2 & q_1 & q_0 \end{bmatrix}. \quad (4)$$

The additional degrees of freedom that enable free global reorientation of the molecule are incorporated into the implementation of the Jain TAD algorithm [7] in DYANA (Eqs. (5)–(21) of [15]) by extending the loops over the clusters $1, \dots, n$ so as to include the base cluster (cluster 0). Instead of using $\omega_0 = 0$ and $\mathbf{v}_0 = 0$ in Eqs. (5) and (6) of [15], the angular and linear velocities of the base cluster are now given by Eq. (2), and by $\mathbf{v}_0 = \dot{\mathbf{r}}_0$, respectively, and the value of α_0 in the recursion of Eq. (21) in [15] is the solution of the six-dimensional linear system of equations $P_0\alpha_0 = -z_0$, in which P_0 and z_0 are calculated as described in [15]. The second time derivatives of the quaternion parameters are obtained by differentiation of Eq. (3). With this implementation of the algorithm the total linear and angular momenta of the system are preserved.

2.2. Compensation of angular and linear momentum

The total angular momentum of the molecule with respect to the origin of the inertial frame is

$$\mathbf{L} = \sum_{k=0}^n m_k [(\mathbf{r}_k + \mathbf{Y}_k) \wedge \mathbf{v}_k + \mathbf{r}_k \wedge (\omega_k \wedge \mathbf{Y}_k)] + I_k \omega_k. \quad (5)$$

Similarly, the total linear momentum of the molecule is

$$\mathbf{P} = \sum_{k=0}^n m_k (\mathbf{v}_k + \omega_k \wedge \mathbf{Y}_k). \quad (6)$$

A change in the angular and linear velocity of the base cluster, $\omega_0 \rightarrow \omega_0 + \Delta\omega_0$ and $\mathbf{v}_0 \rightarrow \mathbf{v}_0 + \Delta\mathbf{v}_0$, leads to concomitant changes in the velocities of the other clusters to the extent of $\Delta\omega_k = \Delta\omega_0$ and $\Delta\mathbf{v}_k = \Delta\mathbf{v}_0 + \Delta\omega_0 \wedge (\mathbf{r}_k - \mathbf{r}_0)$. Hence, the changes induced in the total angular and linear momenta of the molecule are

$$\Delta\mathbf{L} = M(\mathbf{C} \wedge \Delta\mathbf{v}_0 - \Delta\omega_0 \wedge \mathbf{r}_0) + \sum_{k=0}^n [m_k(I(\mathbf{r}_k + \mathbf{Y}_k) - I(\mathbf{Y}_k)) + I_k] \Delta\omega_0, \quad (7)$$

and

$$\Delta\mathbf{P} = M[\Delta\mathbf{v}_0 + \Delta\omega_0 \wedge (\mathbf{C} - \mathbf{r}_0)], \quad (8)$$

where $M = \sum_{k=0}^n m_k$ is the total mass of the molecule, and $\mathbf{C} = \frac{1}{M} \sum_{k=0}^n m_k (\mathbf{r}_k + \mathbf{Y}_k)$ the position of its center of mass. The angular and linear momentum changes are linear functions of the base cluster velocity changes:

$$\begin{bmatrix} \Delta\mathbf{L} \\ \Delta\mathbf{P} \end{bmatrix} = B \begin{bmatrix} \Delta\omega_0 \\ \Delta\mathbf{v}_0 \end{bmatrix}. \quad (9)$$

B is a 6×6 matrix with elements given by Eqs. (7) and (8). Solving Eq. (9) for the case $\Delta\mathbf{L} = -\mathbf{L}$ and $\Delta\mathbf{P} = -\mathbf{P}$ provides the values $\Delta\omega_0$ and $\Delta\mathbf{v}_0$ by which the velocities of the base cluster have to be changed in order to set the angular and linear momenta of the molecule from given values of \mathbf{L} and \mathbf{P} to zero.

Using this procedure, the angular and linear momenta were reset to zero after each sequence of 10 TAD steps for all momentum-compensated DYANA calculations in this paper. The additional computational effort for the momentum compensation is minimal and scales linearly with the size of the molecule. Thus, the high efficiency of the TAD algorithm in DYANA is preserved with the newly implemented algorithm [15].

2.3. Molecular systems and structure calculations

Structure calculations for poly-alanine chains of 20, 40, 60 and 100 residues were performed using the covalent geometry of the ECEPP/2 force field [23] as implemented in the standard library of the program DYANA. In the structure calculations that included steric repulsion, the repulsion was modeled by semi-harmonic lower distance limits given by the sums of the repulsive core radii of 0.95 Å for amide hydrogens, 1.00 Å for other hydrogens, 1.35 Å for aromatic carbons, 1.40 Å for other carbons, 1.30 Å for nitrogens, 1.20 Å for oxygens, and 1.60 Å for sulphurs [13]. For each polypeptide chain, the following eight sets of 50 conformers each were generated:

- (i) Random structures, obtained by choosing the individual torsion angle values from independent uniform distributions using a random number generator [24].
- (ii) Conformers obtained using TAD and simulated annealing with the program DYANA, with 4000 TAD steps per conformer [15], where the standard protocol was modified as follows: The steric repulsion was turned off, and the N-terminus of the polypeptide chain was kept fixed in space, so that no compensation of the total angular and linear momenta could be applied.

- (iii) Same as (ii), except that free global reorientation of the molecule was allowed, and the total angular and linear momenta were periodically reset to zero (see next two sections).
- (iv) Conformers obtained using TAD in the standard annealing protocol of the program CNS, version 0.9 [17]. The macros *generate_seq.inp*, *generate_extended.inp*, and *anneal.inp* were used with standard parameters, except that all potential energy terms were set to zero. The TAD algorithm [25] implemented in CNS uses compensation of the total angular and linear momenta.
- (v) Conformers obtained using the variable target function (VTF) algorithm [13,14] as implemented in DYANA [15], with inclusion of steric repulsion.
- (vi)–(viii) Same as (ii), (iii) and (iv), except that steric repulsion with the aforementioned atomic radii was activated.

The *Antp*(C39S) homeodomain [26] contains chain-terminal segments of residues 0–6 and 60–67 that are devoid of structurally relevant NOE distance constraints and are thus expected to be “unstructured” in solution. This protein has previously been thoroughly investigated with the program DIANA [14,26], using conjugate gradient minimization with the variable target function method and the REDAC strategy [27], and is therefore a suitable system to investigate conformational sampling by novel procedures. The NMR data set for the *Antp*(C39S) homeodomain consists of 844 NOE upper distance limits, 50 upper and 50 lower distance limits for 25 hydrogen bonds, and 171 ϕ , ψ , and χ^1 torsion angle constraints obtained by systematic analysis of the local conformation with the program HABAS [28]. Structures were calculated using the standard DYANA simulated annealing protocol with 4000 TAD steps. Two sets of 50 conformers each were calculated without and with compensation of angular and linear momenta (see the next two sections). The 20 conformers with the lowest final target function values were retained for analysis. These structures were compared to a bundle of 20 conformers calculated with the VTF algorithm in the program DYANA, using the REDAC strategy as described in [27].

Additional structure calculations were performed for two hypothetical *Antp*(C39S) homeodomain polypeptides that had been extended at both chain ends by either 10 or 20 alanyl residues, *Antp*(C39S)[-10–77] and *Antp*(C39S)[-20–87]. For the longest construct, structures were also calculated by a “hybrid procedure” that combines the efficiency of simulated TAD annealing for the globular domain with re-randomization and minimization for the unstructured regions. To this end, the backbone torsion angles ϕ and ψ of the residues outside of the structured part of the protein (residues 7–59) were randomized in the range -180° to $+180^\circ$ at the end of the cooling phase of the standard DYANA TAD simulated annealing protocol, and 1000 VTF minimization steps each at the target function level 3 and at the maximal target level [14] were performed before the standard protocol was continued by including the steric repulsion of the hydrogen atoms [15].

Structure calculations with the program DYANA were performed on Compaq Alpha computer systems using 4 to 8 processors in parallel. Structure calculations with the program CNS were done on a Compaq Alpha computer.

2.4. Structure comparison and criteria used to characterize the global shape of molecules

The measure commonly used to quantify the spread in a bundle of NMR conformers is the global root mean square distance (RMSD) of atomic positions with respect to the mean coordinates. In a well-defined structure, the average of the RMSD values thus calculated for the individual conformers then has the intuitive meaning of representing the radius of a cylindrical conformation space spanned by the structure bundle, and thus is a measure of the precision of the structure determination [3,8]. For poorly defined polypeptide segments, however, the mean coordinates are no longer meaningful because the averaging may lead to a strongly distorted, collapsed set of coordinates. In this paper we therefore report RMSD values that are calculated as the average over all the $N(N-1)/2$ pairwise RMSD values among the N individual conformers of a structure bundle.

Since the RMSD value is known to be quite insensitive to certain deviations from randomness in predominantly extended chains [11], the extension of the polypeptide chains was assessed independently by calculating the end-to-end distance, i.e. the distance between the backbone nitrogen atom of the N-terminal residue and the backbone carboxyl carbon atom of the C-terminal residue [29]. The overall shape of the molecule has further

been characterized by the inertia tensor, with principle axes $a \geq b \geq c$. The ratio between the largest and smallest principle axes, a/c , is then used as measure for the relative elongation of the structure, i.e. its deviation from spherical shape, which is independent of the absolute size of the polypeptide and is therefore suitable for comparison of results obtained for different chain lengths [30].

Superpositions of structure bundles and calculation of the RMSD values and inertia tensors were performed with the program MOLMOL [31], which was also used for the preparation of the figures.

3. Results and discussion

In the original implementation of TAD in the program DYANA [15], the N-terminus of the protein was fixed in space. This was not a prerequisite of the TAD algorithm [7], but it appeared conceptually more appealing since the system then had only one type of degrees of freedom, i.e. the torsion angles, whereas for the description of the global position and orientation of a molecule that can freely reorient in space, one has to introduce six additional degrees of freedom of a different type. As a consequence of fixing the N-terminus of the polypeptide chain, the total angular and linear momenta of the system cannot be conserved, since the conservation laws for these quantities result from the fact that a mechanical system is invariant under global rotation and translation [30]. Therefore, if a part of the molecule is fixed in space, it is in general not possible for the total angular momentum to be zero throughout the calculation. Global rotations associated with a non-vanishing total angular momentum may lead to centrifugal forces that can influence the sampling of conformation space for long, flexible molecules. In the following we investigate whether and to which extent this effect is noticeable in structure calculations of polypeptides and proteins.

3.1. Unconstrained polypeptide chains

Structure calculations with vanishing target function or potential energy have commonly been used as sensitive tests of conformational sampling [10,11]. Since in such systems each combination of torsion angle values should occur with the same probability in a Boltzmann ensemble, optimal sampling would lead to an ensemble of conformers for which all torsion angles are independent, uniformly distributed random variables. To test whether TAD yields such bundles of random conformers, we performed structure calculations with unconstrained polypeptide chains of 20, 40, 60, and 100 alanyl residues (see Methods). Literature references are available, since Ala₂₀, Ala₄₀ and Ala₆₀ chains have previously been used for sampling tests with different techniques of structure calculation [11].

The results (Table 1) show close agreement of the RMSD values measured for a bundle of random conformers with those obtained by TAD with and without momentum compensation. The end-to-end distances, D_{ee} , and the ratios of the inertia tensor axes, a/c , of the random structures are also very close to those for the structures calculated with DYANA or CNS with momentum compensation. In fact, for Ala₂₀, Ala₄₀ and Ala₆₀ the three structure bundles cannot be distinguished on the basis of the numbers in Table 1, and the differences seen for Ala₁₀₀ are hardly significant. The reference a/c value measured for random structures is approximately 3.5, showing that random structures are, on the average, not spherical but markedly oblong (Fig. 1). The structures computed without momentum compensation are significantly more elongated than either the random structures or those obtained with momentum compensation (Fig. 1), as indicated by larger values of D_{ee} and a/c , which also indicate a tendency that elongation increases for the longer chains. Further tests showed that release of the N-terminus and compensation of the momenta is needed to achieve random sampling.

The distribution of the backbone torsion angles ϕ and ψ for Ala₆₀ in Ramachandran plots is uniformly random for the randomized structure (Fig. 2(a)) as well as for those obtained using TAD with momentum compensation (Fig. 2(b)), whereas the structure from TAD without momentum compensation shows a slight clustering of torsion

Table 1
Structure calculations for unconstrained poly-Ala chains^a

Chain length and computation technique	RMSD (Å) ^b	D_{ee} (Å) ^c	a/c ^d
Ala ₂₀			
random torsion angles	6.7 ± 1.6	21.8 ± 8.3	3.5 ± 1.3
TAD (DYANA, uncompensated) ^e	5.8 ± 1.3	32.6 ± 8.2	4.9 ± 1.4
TAD (DYANA, compensated) ^f	6.8 ± 1.4	19.0 ± 7.9	3.1 ± 1.1
TAD (CNS) ^g	6.7 ± 1.4	17.1 ± 6.9	2.8 ± 1.0
Ala ₄₀			
random torsion angles	10.4 ± 2.1	32.0 ± 12.3	3.4 ± 1.3
TAD (DYANA, uncompensated)	10.5 ± 3.3	62.6 ± 18.0	6.8 ± 2.5
TAD (DYANA, compensated)	10.6 ± 2.3	29.1 ± 12.4	3.4 ± 1.2
TAD (CNS)	10.5 ± 2.1	24.8 ± 11.5	3.1 ± 1.1
Ala ₆₀			
random torsion angles	13.5 ± 3.0	37.4 ± 15.3	3.6 ± 1.2
TAD (DYANA, uncompensated)	17.9 ± 6.9	86.3 ± 35.6	8.0 ± 4.0
TAD (DYANA, compensated)	14.9 ± 4.6	34.7 ± 22.2	3.5 ± 1.6
TAD (CNS)	14.0 ± 3.4	33.5 ± 17.5	3.4 ± 1.2
Ala ₁₀₀			
random torsion angles	17.2 ± 3.3	48.1 ± 20.4	3.3 ± 1.1
TAD (DYANA, uncompensated)	23.6 ± 10.4	226.0 ± 46.0	15.5 ± 6.4
TAD (DYANA, compensated)	26.3 ± 9.8	65.5 ± 38.8	4.2 ± 2.4
TAD (CNS)	22.4 ± 8.0	87.8 ± 37.8	6.4 ± 2.7

^a For each entry the structure bundle consists of 50 conformers (see Methods).

^b Average value and standard deviation of the pairwise RMSDs for the backbone atoms N, C^α and C' of the entire polypeptide chain.

^c Average value and standard deviation of the end-to-end distances, i.e. the distance between the backbone nitrogen atom of the N-terminal residue and the carboxyl carbon atom of the C-terminal residue.

^d Average value and standard deviation for the ratio between the largest and smallest half-axes of the inertia tensor.

^e Original implementation of TAD in DYANA [15], with the N-terminus fixed in space so that there is no momentum compensation.

^f Modified implementation of TAD in DYANA (see Methods), where the N-terminus is not fixed in space and momentum compensation is applied.

^g Use of the program CNS as described in Section 2.3.

angle values in the corners of the Ramachandran plot (Fig. 2(c)). Nonetheless, Fig. 2(c) also shows that even in the absence of momentum compensation the local randomness of the structure is not seriously impaired.

3.2. Polypeptide chains with steric repulsion

The same calculations as in Section 3.1 were repeated for poly-Ala chains with steric repulsion (Table 2). As a reference we use structures calculated using a variable target function method based on conjugate gradient minimization, which has previously been shown to exhibit good sampling properties [9,11,13]. There is no

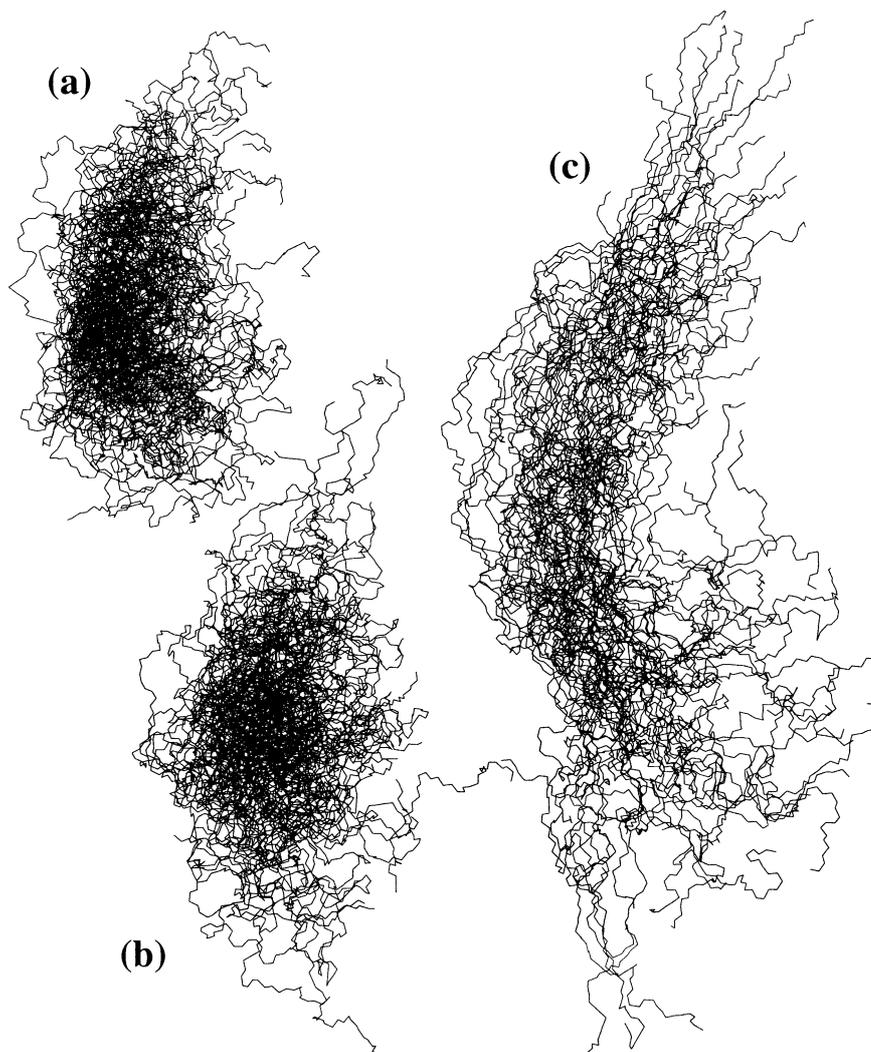


Fig. 1. Structures of unconstrained Ala_{60} polypeptide chains. (a) Generated by randomizing the torsion angles. (b) Calculated using TAD in DYANA with momentum compensation. (c) Calculated using TAD in DYANA without momentum compensation. Each structure bundle consists of 50 conformers that were superimposed for pairwise minimal RMSD (see Methods).

significant difference between these reference structures and those calculated using TAD with momentum compensation in either DYANA or CNS. Again, without compensation of the momenta, the TAD structures are more elongated. When compared with the results for unconstrained chains, the values of D_{ee} are consistently larger in the structures for which steric repulsion was taken into account, whereas the overall shape of the molecules is similar, as evidenced by the closely similar a/c values. This indicates, not unexpectedly, that the excluded volume effect of the steric repulsion results in globally enlarged structures but has at most a minor influence on the global shape when compared to the corresponding fully unconstrained chains. Overall, the results obtained for unconstrained polypeptide chains with “realistic” steric features of the covalent structure showed that good sampling of conformation space by TAD can be achieved when using momenta compensation.

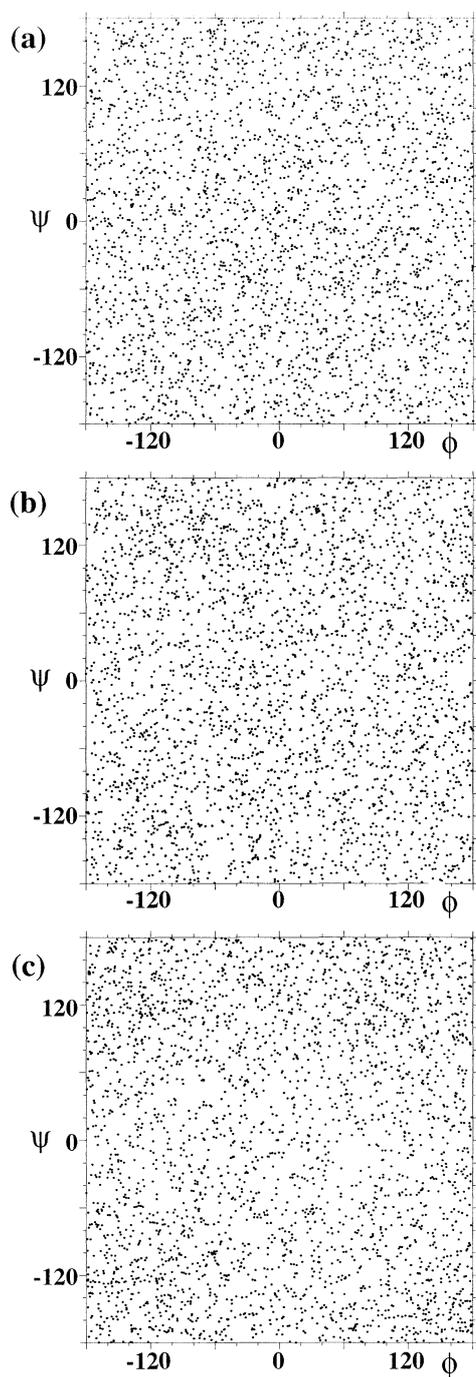


Fig. 2. Ramachandran ϕ , ψ -plots for 50 conformers of an unconstrained Ala₆₀ polypeptide. (a), (b) and (c) represent the data for the structure bundles in Fig. 1, (a), (b) and (c), respectively.

Table 2
Structure calculations for poly-Ala chains with steric repulsion^a

Chain length and computation technique	RMSD (Å) ^a	D_{ee} (Å) ^a	a/c^a
Ala ₂₀			
VTF ^b	6.3 ± 1.4	26.0 ± 8.7	3.9 ± 1.3
TAD (DYANA, uncompensated) ^a	6.0 ± 1.7	35.7 ± 9.5	5.2 ± 1.6
TAD (DYANA, compensated) ^a	6.2 ± 1.4	28.0 ± 9.1	4.0 ± 1.1
TAD (CNS) ^a	6.3 ± 1.3	27.3 ± 7.5	3.8 ± 1.2
Ala ₄₀			
VTF	10.9 ± 2.2	37.9 ± 13.8	3.9 ± 1.2
TAD (DYANA, uncompensated)	9.8 ± 3.2	75.9 ± 17.3	8.3 ± 3.2
TAD (DYANA, compensated)	10.7 ± 2.3	42.8 ± 12.7	3.7 ± 1.2
TAD (CNS)	10.9 ± 2.3	36.6 ± 14.9	3.7 ± 1.4
Ala ₆₀			
VTF	14.5 ± 2.7	45.3 ± 17.3	3.5 ± 1.1
TAD (DYANA, uncompensated)	13.9 ± 5.0	116.1 ± 23.9	9.5 ± 4.0
TAD (DYANA, compensated)	14.4 ± 2.6	46.2 ± 15.4	3.3 ± 0.9
TAD (CNS)	15.0 ± 3.6	47.9 ± 20.5	3.8 ± 1.2
Ala ₁₀₀			
VTF	20.0 ± 3.6	61.3 ± 20.9	3.4 ± 1.2
TAD (DYANA, uncompensated)	31.5 ± 13.8	176.5 ± 62.3	11.7 ± 6.5
TAD (DYANA, compensated)	23.5 ± 6.9	93.7 ± 32.6	5.6 ± 2.5
TAD (CNS)	21.9 ± 1.4	77.3 ± 31.5	4.4 ± 1.8

^a See footnotes to Table 1.

^b Structures calculated using the variable target function method [13] implemented in DYANA [15].

3.3. Structures calculated using experimental NMR data sets

In NMR structure calculations with experimental constraints and steric repulsion, regions of the structure that are not well defined by the experimental data are most susceptible to uneven sampling of the accessible conformation space. These may include chain ends, exposed loops, and long amino acid side-chains on the protein surface. We therefore investigated the sampling of conformation space by the TAD algorithm of DYANA for a globular protein with extensive unstructured parts, i.e. the *Antp*(C39S) homeodomain [26]. This protein comprises N- and C-terminal segments of the residues 0–6 and 60–67, respectively, for which complete lack of medium-range and long-range NOE distance constraints indicates that they are “unstructured” in solution.

The results of the structure calculations with the *Antp*(C39S) homeodomain polypeptides are summarized in Table 3. The TAD calculations find structures with lower residual target function values than the calculations based on variable target function conjugate gradient minimization, which manifests the superior performance of TAD for the calculation of the well-constrained region of residues 7–59 (Fig. 3). A comparison of the sampling properties for the unstructured tail regions in the “natural” polypeptide chain of residues 0–67, *Antp*(C39S)[0–67], shows

Table 3
Structure calculations for *Antp*(C39S) homeodomain polypeptides^a

Protein construct and computation technique ^b	TF (Å ²) ^c	RMSD (Å) ^d	D_{ee} (Å) ^e	a/c ^e
<i>Antp</i> (C39S)[0–67]				
VTF/REDAC ^f	1.21 ± 0.16	5.4 ± 1.2	35.9 ± 11.2	1.77 ± 0.16
TAD (DYANA, uncompensated) ^e	0.82 ± 0.11	4.9 ± 1.3	34.4 ± 8.9	1.75 ± 0.16
TAD (DYANA, compensated) ^e	0.81 ± 0.10	4.6 ± 1.1	34.3 ± 9.4	1.70 ± 0.13
<i>Antp</i> (C39S)[-10–77]				
VTF/REDAC	1.22 ± 0.14	11.2 ± 2.3	41.9 ± 11.9	1.90 ± 0.39
TAD (DYANA, uncompensated)	0.77 ± 0.12	14.0 ± 4.1	61.5 ± 23.5	2.30 ± 0.34
TAD (DYANA, compensated)	0.86 ± 0.17	13.4 ± 3.3	46.9 ± 16.2	1.91 ± 0.34
<i>Antp</i> (C39S)[-20–87]				
VTF/REDAC	1.14 ± 0.13	14.1 ± 2.5	46.2 ± 9.4	2.27 ± 0.48
TAD (DYANA, uncompensated)	0.88 ± 0.15	17.3 ± 5.3	114.4 ± 28.0	3.79 ± 0.84
TAD (DYANA, compensated)	0.83 ± 0.09	19.4 ± 5.0	96.1 ± 32.8	3.03 ± 0.69
Hybrid TAD/VTF ^g	0.95 ± 0.14	13.1 ± 3.2	48.0 ± 22.1	2.43 ± 0.32

^a For each entry the structure bundle consists of the 20 conformers with the lowest residual target function values selected out of all conformers that were calculated (see Methods).

^b *Antp*(C39S) is the previously described construct [26] of a variant *Antennapedia* homeodomain, with flexibly disordered chain ends of residues 0–6 and 60–67. In the other two constructs the polypeptide chain has been elongated at both ends by Ala₁₀ segments, or Ala₂₀ segments, respectively.

^c Average value and standard deviation of the residual DYANA target function value for the entire polypeptide chain.

^d Average value and standard deviation of the pairwise RMSDs for the backbone atoms N, C^α and C^β of the entire polypeptide chain.

^e See the Methods section and footnotes to Table 1.

^f Structure calculated with the variable target function method implemented in DYANA, using the REDAC strategy [27].

^g Structure calculated with DYANA using a hybrid method that combines TAD calculations and variable target function minimization as described in the Methods section.

that all three calculations of Table 3 produced structures that are equivalent with respect to the global measures of sampling, elongation and shape. The DYANA TAD algorithm with or without momentum compensation thus provides comparably good sampling of conformation space as the less efficient variable target function (VTF) conjugate gradient minimization method, which has previously been shown to yield good sampling [9,14,27] and is therefore used as a reference. In the *Antp*(C39S) homeodomain polypeptides that have been extended by 10 or 20 Ala residues at both chain ends, the number of unstructured chain-terminal residues increases from 15 (22%) in the “natural” protein to 35 (40%) in *Antp*(C39S)[-20–77] and 55 (51%) in *Antp*(C39S)[-20–87]. The extension by 10 residues at both chain ends has no marked effect on the sampling properties of the TAD algorithm with momentum compensation, when compared with VTF minimization, whereas the structures obtained using TAD without momentum compensation show a significant trend towards more pronounced elongation (values for D_{ee} and a/c in Table 3). For the longest polypeptide, *Antp*(C39S)[-20–87], there is a trend to increased elongation in the TAD structures obtained with and without momentum compensation. The artifactual elongation could be suppressed with a hybrid approach, where structures were first calculated with TAD, which was followed by randomization and VTF conjugate gradient minimization of the unstructured regions. Since the TAD result for the well-structured parts of the protein was preserved during the additional treatment of the “tails”, this approach

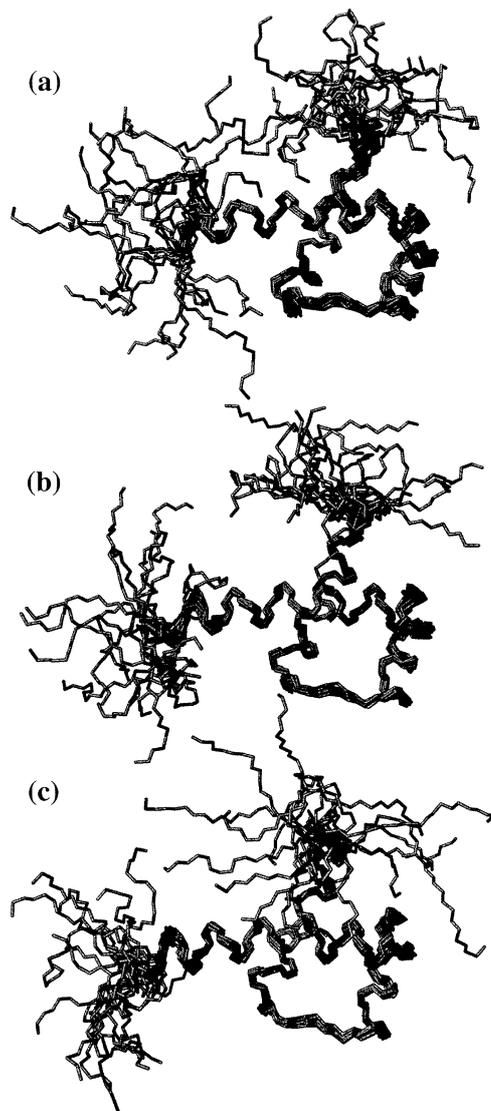


Fig. 3. Structures of the *Antp*(C39S) homeodomain calculated using the program DYANA with the following techniques (see also Table 3): (a) Variable target function method with the REDAC strategy. (b) TAD calculation with momentum compensation. (c) TAD calculation without momentum compensation. Each structure bundle consists of 20 conformers that were superimposed for pairwise minimal RMSD of the backbone atoms N, C $^{\alpha}$ and C' of the residues 7–59.

combines the high efficiency of TAD to escape from local minima with the virtues of VTF conjugate gradient minimization for uniform sampling of conformation space with long unstructured polypeptide chains.

For a meaningful comparison of the sampling of conformation space for the well-confined residues 7–59 of the *Antp*(C39S) polypeptides, the differences in the quality of the results obtained with the different methods (as manifested by the final target function values) have to be taken into account. It is well known that the results of calculations in which larger residual constraint violations are tolerated for a given set of experimental constraints will sample larger regions of conformation space than “more precise” structures with lower residual target function values [3]. The average RMSD to the mean atom coordinates for the backbone atoms N, C $^{\alpha}$ and C' of residues

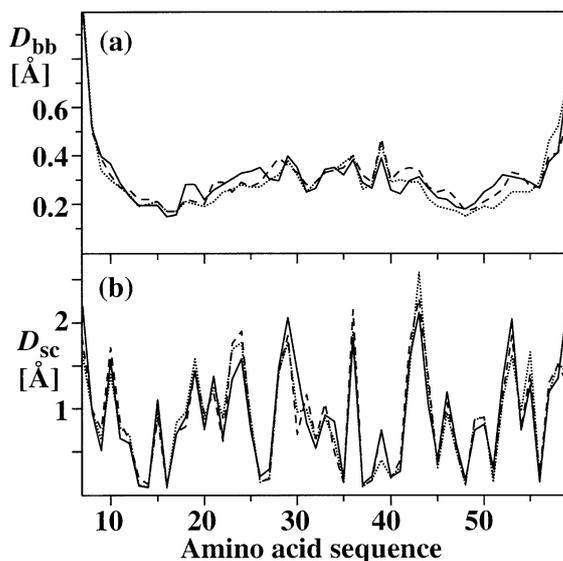


Fig. 4. (a) Global backbone displacements, D_{bb} , and (b) local side-chain displacements, D_{sc} , calculated for the well confined residues 7–59 in the structures of the *Antp*(C39S) homeodomain that were calculated with different techniques, using the program DYANA. The average of the displacement values [35] for the individual conformers relative to the corresponding mean coordinates are plotted as follows: Solid lines for the structure calculated using the variable target function (VTF) method with the REDAC strategy. Broken lines for the structure calculated using TAD without momentum compensation. Dotted lines for the structure calculated using TAD with momentum compensation (see the Methods section and Table 3 for details). To compensate for the fact that the structure obtained with the VTF method has larger residual constraint violations than those obtained with TAD, the displacements for the VTF structure have been scaled down as described in the text. The global backbone displacement of residue i is obtained by evaluating the RMSD for the backbone atoms N, C^α and C' of residue i after global superposition computed for the same backbone atoms of residues 7–59. The local side-chain displacement for residue i is the RMSD for the side-chain heavy atoms of residue i evaluated after superposition of the backbone atoms N, C^α and C' of the residues $i - 1$, i and $i + 1$.

7–59 is 0.46 \AA for the *Antp*(C39S) structures calculated with the VTF/REDAC method, which was cut off with an average target function value of 1.21 \AA^2 , and 0.36 \AA for those obtained using TAD with momentum compensation, where the average final target function value was 0.81 \AA^2 (for TAD structure calculations with a cut-off at a residual TF-value of 1.22 \AA^2 , which was achieved by reducing the number of TAD steps per conformer from 4000 to 1500, an RMSD value of 0.43 \AA was obtained, which coincides with the VTF/REDAC method). In Fig. 4 we accounted for the correlation between residual target function values and RMSD by downscaling the global backbone displacements, D_{bb} , and the local side chain displacements, D_{sc} , for the VTF/REDAC structures by the ratios of the relevant global RMSD. For the residues 7–59 these ratios are $0.36/0.46 = 0.78$ for the backbone atoms, and $0.98/1.17 = 0.84$ for all heavy atoms. Fig. 4 then shows that the sampling of conformation space by VTF/REDAC and by TAD is comparable for the backbone (Fig. 4(a)) as well as for the amino acid side-chains (Fig. 4(b)), and that no significant differences are apparent between the TAD structures obtained with and without momentum compensation.

4. Conclusions

The NMR structure calculations from experimental conformational constraints (Table 3) clearly demonstrate the superior performance of TAD algorithms when compared to the variable target function approach [13] in the most advanced implementations [15,27]. Superior performance of TAD relative to molecular dynamics calculations in Cartesian space has previously been documented [15,16], and the efficiencies of the TAD implementations in

the programs DYANA and CNS have also been compared [15]. For molecular structures that are, overall, well constrained by experimental data, the TAD implementations with or without compensation of momenta in the DYANA program [15] and with momenta compensation in the CNS program [16] all combine high efficiency with good sampling of conformation space. Although this positive evaluation applies for proteins with unconstrained “tails” of up to about 10 residues, and for individual unconstrained long amino acid side chains (Table 3), good sampling of conformation space for longer unconstrained polypeptide segments must be based on TAD implementations that enable compensation of momenta during the structure calculation (Tables 1–3). The results collected in Figs. 1–4 and the Tables 1–3 provided the foundation for a revised implementation of TAD in DYANA (Sections 2.1 and 2.2; a new release of the program, DYANA 1.6, can be obtained from the authors: www.mol.biol.ethz/dyana), and may serve as a guideline for possible new TAD implementations in future software packages.

For the critical situation encountered with macromolecules containing unconstrained chain segments corresponding in length to 10 or more amino acid residues, comparably good sampling of conformation space is achieved using the implementations of TAD with momentum compensation in both commonly used programs, DYANA [15] and CNS [16]. Since the use of the TAD algorithm in DYANA with compensation of momenta has no disadvantages in computational efficiency when compared to the TAD algorithm without momentum compensation (see Sections 2.1 and 2.2), it should be used for all structure calculations with DYANA. In future, novel TAD implementations for structure calculation it should be ascertained that the molecule can freely reorient in space, so that the total momenta will be conserved and can periodically be reset to zero. The compensation of momenta should be achievable at minimal reduction of the efficiency of the TAD algorithm, similar to the situation with DYANA.

Finally, the present study indicates that very long unconstrained polypeptides tend to adopt a slightly elongated shape and do not span a spherical area of conformation space. This effect appears not to be the result of details of the TAD implementation, since it is apparent in structures calculated with DYANA as well as with the program CNS, which employs a different TAD algorithm. The elongation may be due to metric tensor effects inherent in TAD [32]. Although there exist efficient methods to compensate for metric tensor effects [33,34], their implementation in programs for NMR structure calculation, such as DYANA and CNS, seems hardly warranted, since structure calculations of long unstructured polypeptides are of limited practical interest. For the rare situations of globular proteins comprising long unstructured polypeptide segments, we recommend a simple hybrid approach that combines TAD calculations for the structured part with re-randomization and variable target function minimization for the flexible tails. The program DYANA includes all the routines needed for this approach, as exemplified in this paper for the elongated homeodomain polypeptide *Anip*(C39S)[-20–87] (Table 3).

Acknowledgements

The use of computing resources of the Competence Center for Computational Chemistry of ETH Zürich is gratefully acknowledged. We thank the Schweizerischer Nationalfonds for financial support (project 31.49047.96), and Mrs. M. Geier for the careful processing of the manuscript.

References

- [1] M.P. Allen, D.J. Tildesley, *Computer Simulation of Liquids*, Clarendon, Oxford, 1989.
- [2] A.T. Brünger, M. Nilges, *Q. Rev. Biophys.* 26 (1993) 49.
- [3] P. Güntert, *Q. Rev. Biophys.* 31 (1998) 145.
- [4] A.M.J.J. Bonvin, A.E. Mark, W.F. van Gunsteren, *Comput. Phys. Commun.* 128 (2000) 550.
- [5] G.S. Heffelfinger, *Comput. Phys. Commun.* 128 (2000) 219.
- [6] A.K. Mazur, V.E. Dorofeev, R.A. Abagyan, *J. Comput. Phys.* 92 (1991) 261.
- [7] A. Jain, N. Vaidehi, G. Rodriguez, *J. Comput. Phys.* 106 (1993) 258.

- [8] K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, New York, 1986.
- [9] G. Wagner, W. Braun, T.F. Havel, T. Schaumann, N. Go, K. Wüthrich, *J. Mol. Biol.* 196 (1987) 611.
- [10] W.J. Metzler, D.R. Hare, A. Pardi, *Biochemistry* 28 (1989) 7045.
- [11] T.F. Havel, *Biopolymers* 29 (1990) 1565.
- [12] J. Kuszewski, M. Nilges, A.T. Brünger, *J. Biomol. NMR* 2 (1992) 33.
- [13] W. Braun, N. Go, *J. Mol. Biol.* 186 (1985) 611.
- [14] P. Güntert, W. Braun, K. Wüthrich, *J. Mol. Biol.* 217 (1991) 517.
- [15] P. Güntert, C. Mumenthaler, K. Wüthrich, *J. Mol. Biol.* 273 (1997) 283.
- [16] E.G. Stein, L.M. Rice, A.T. Brünger, *J. Magn. Reson.* 124 (1997) 154.
- [17] A.T. Brünger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren, *Acta Cryst. D* 54 (1998) 905.
- [18] R. Riek, S. Hornemann, G. Wider, R. Glockshuber, K. Wüthrich, *FEBS Lett.* 413 (1997) 282.
- [19] D.G. Donne, J.H. Viles, D. Groth, I. Mehlhorn, T.L. James, F.E. Cohen, S.B. Prusiner, P.E. Wright, H.J. Dyson, *Proc. Nat. Acad. Sci. USA* 94 (1997) 13 452.
- [20] R. Zahn, A. Liu, T. Lührs, R. Riek, C. von Schrötter, F. López-García, M. Billeter, L. Calzolari, G. Wider, K. Wüthrich, *Proc. Natl. Acad. Sci. USA* 97 (2000) 145.
- [21] F. López-García, R. Zahn, R. Riek, K. Wüthrich, *Proc. Natl. Acad. Sci. USA* (1997) 8334.
- [22] G.R. Kneller, K. Hinsen, *Phys. Rev. E* 50 (1994) 1559.
- [23] F.A. Momany, R.F. McGuire, A.W. Burgess, H.A. Scheraga, *J. Phys. Chem.* 79 (1975) 2361.
- [24] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1986.
- [25] D.S. Bae, E.J. Haug, *Mech. Struct. Mech.* 15 (1987) 359.
- [26] P. Güntert, Y.Q. Qian, G. Otting, M. Müller, W.J. Gehring, K. Wüthrich, *J. Mol. Biol.* 217 (1991) 531.
- [27] P. Güntert, K. Wüthrich, *J. Biomol. NMR* 1 (1991) 446.
- [28] P. Güntert, W. Braun, M. Billeter, K. Wüthrich, *J. Am. Chem. Soc.* 111 (1989) 3997.
- [29] C. Tanford, *Physical Chemistry of Macromolecules*, Wiley, New York, 1961, p. 150.
- [30] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer, New York, 1978.
- [31] R. Koradi, M. Billeter, K. Wüthrich, *J. Mol. Graph.* 14 (1996) 51.
- [32] M. Fixman, *Proc. Nat. Acad. Sci. USA* 71 (1974) 3050.
- [33] A. Jain, *J. Comput. Phys.* 136 (1997) 289.
- [34] T. Mülders, W. Swegat, *Mol. Phys.* 94 (1998) 395.
- [35] M. Billeter, A.D. Kline, W. Braun, R. Huber, K. Wüthrich, *J. Mol. Biol.* 206 (1989) 677.