PROTEINS: Structure, Function, and Genetics 29:240–251 (1997)

Prediction of Protein Conformational Freedom From Distance Constraints

B.L. de Groot,[1] D.M.F. van Aalten,[2] R.M. Scheek,[1] A. Amadei,[1] G. Vriend,[3] and H.J.C. Berendsen[1]*
[1]*Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, University of Groningen, Groningen, The Netherlands*
[2]*Keck Structural Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York*
[3]*EMBL, Heidelberg, Germany*

ABSTRACT **A method is presented that generates random protein structures that fulfil a set of upper and lower interatomic distance limits. These limits depend on distances measured in experimental structures and the strength of the interatomic interaction. Structural differences between generated structures are similar to those obtained from experiment and from MD simulation. Although detailed aspects of dynamical mechanisms are not covered and the extent of variations are only estimated in a relative sense, applications to an IgG-binding domain, an SH3 binding domain, HPr, calmodulin, and lysozyme are presented which illustrate the use of the method as a fast and simple way to predict structural variability in proteins. The method may be used to support the design of mutants, when structural fluctuations for a large number of mutants are to be screened. The results suggest that motional freedom in proteins is ruled largely by a set of simple geometric constraints. Proteins 29:240–251, 1997.** © 1997 Wiley-Liss, Inc.

Key words: molecular dynamics; essential dynamics; protein dynamics; NMR

INTRODUCTION

Structural studies like X-ray crystallography and NMR spectroscopy often provide insight into the function of a protein. However, detailed questions on many dynamic aspects of enzymatic mechanisms such as regulation or substrate entry, remain unanswered when only static structures are available. Dynamic processes are crucial steps in the functioning of enzymes. Therefore, detailed information on the dynamics of a protein is necessary for a complete understanding of its function.

Simulation techniques can help to obtain dynamic information that cannot be provided by experimental techniques in a straightforward manner. A number of computational techniques has been developed to gain information on protein dynamics and structural fluctuations. Molecular Dynamics (MD) and Monte Carlo (MC) techniques are the most popular ones. The accuracy of these techniques depends on the protocols used (force-field, molecular representation etc.) and on the simulation length. Using the most realistic force fields, at most a few nanoseconds for a small protein in an aqueous environment can be simulated within acceptable computer time.[1,2] This time scale is a few orders of magnitude smaller than that on which most biological processes take place, leaving the MD technique with a significant sampling problem.[3,4] The efficiency of MC calculations is comparable to that of MD due to the presence of internal barriers.[5]

Essential Dynamics

Essential Dynamics (ED),[6–9] equivalent to Principal Component[10,11] analyses of MD trajectories have shown that most (more than 90%) of the simulated atomic fluctuations usually can be described by a few large-scale concerted motions. ED analyses of MD trajectories determine the eigenvectors of the covariance matrix of atomic fluctuations. Diagonalisation of this matrix yields a set of eigenvectors and eigenvalues and the eigenvectors with largest eigenvalues (usually a typical number of ten suffices) describe all large-scale concerted fluctuations. If the eigenvectors are seen as vectors that span a complex space then the few "essential" eigenvectors with largest eigenvalues span a subspace, the essential subspace, and all large concerted motions take place in this subspace. It is assumed that also the true configurational space of most proteins contains a low-dimensional subspace in which most positional fluctuations take place. The essential subspace obtained from simulation is an approximation of that subspace. ED analyses of MD trajectories have been helpful in a number of cases to study functional motions and predict mutants.[7,8,12] As the trajectory of each simulation can be considered as a diffusional path through a part of the available space spanned by the first few eigenvectors,[13,14] the definition of

*Correspondence to: H.J.C. Berendsen, Groningen Biomolecular Sciences and Biotechnology Institute, Department of Biophysical Chemistry, University of Groningen, Nÿenborgh 4, 9747 AG, Groningen, The Netherlands.
E-mail: berends@chem.rug.nl
Received 2 August 1996; Accepted 6 May 1997

individual eigenvectors spanning this subspace from a simulation has not converged in the simulated time,[3,4] but the definition of the subspace itself approximately has.[15,16] This means that the high eigenvalue-eigenvectors constructed from independent (pieces of) simulation(s) are rotated with respect to each other but only in a subspace with limited dimension. The fact that the dynamic behavior of simulated proteins can be captured by only a few directions in configurational space can be used to improve sampling efficiency in MD simulations by driving a second MD run along eigenvectors extracted from an initial MD run.[13,14,17]

Currently the eigenvectors that approximate the essential subspace can only be determined from covariance analyses of long MD runs, requiring considerable computational effort. In the present study, however, an attempt is made to obtain these most prominent collective structure variations in a very simplified way.

Analogy With Structure Determination From NMR Data

Structure solution by NMR is mainly based on the conversion of force-field derived and experimentally determined distances (from NOE data) into a set of three-dimensional coordinates. The available data is often insufficient to reach a unique solution, a problem that is usually circumvented by providing an ensemble of structures. Large local conformational differences between generated structures can represent structural flexibility but are often the result of a lack of experimental data.[18,19]

Here we carry this idea a bit further. If all distances are known, and their upper and lower bounds are set to physically realistic values, then the resulting structures are close to realistic configurations that should, in principle, be reachable (during an MD simulation). An ED analysis of such a set of structures will, if the ensemble of generated structures is large enough, yield directions describing fluctuations that are possible within the selected distance limits. If the distance limits are chosen in a sensible manner, then the observed fluctuations correspond to realistic configurational freedom and the ED results could be used to improve the sampling during MD simulation.[13,14,17]

A technique has been developed to generate random structures, limited by distance criteria. The method has been applied to a number of proteins (the B1 IgG-binding domain of streptococcal protein G, the chicken α-spectrin SH3 domain, HPr from *Escherichia coli*, bacteriophage T4 lysozyme and rat testis calmodulin). These applications indicate that the applied distance restrictions are compatible with acceptable protein structures and that the differences between these structures can be used to extract information on the structural variability of the proteins studied.

METHODS

Distance Bounds

The method in its current implementation is based on a covariance analysis of randomly generated structures that fulfil a set of distance constraints. The first step is to measure all pairwise interatomic distances in the (known) experimental structure of the protein to be studied. The distance limits are now set at this distance plus or minus D nanometers, where D is small for tightly interacting atom pairs and larger for weaker interactions. The different types of interactions that were considered are listed in Table I and distance limits D are given in Table II. For all covalent 1–4 pairs, the upper and lower bounds are corrected such that their distance is always between the distances calculated in the 'cis' resp. 'trans' conformation. There is a special group for atom pairs that are part of the same secondary structure element to make sure secondary structure (helix, strand) is preserved in the generated structures. This way, a total of 4697 distance restrictions (3.3% of the total number of distances) could be defined for the B1 IgG-binding domain. This number was 4197 (2.5%) for SH3, 7333 (2.4%) for HPr, 14388 (1.5%) for calmodulin and 17818 (0.6%) for lysozyme, respectively (see Table II for the distribution of distances over the different classes).

To speed up the search for structures that fulfil all distance criteria, upper and lower bounds are defined for all atom pairs that are not explicitly mentioned in Table I. The range of freedom D given to these pairs (0.5 nm) is much larger than for all other pairs (Table II) (the lower distance limits for these pairs are corrected such that they are not lower than the sum of the van der Waals radii of the atoms involved). If this upper limit is relaxed, the speed of convergence is strongly reduced but the resulting structures are virtually unchanged.

For all studied proteins except HPr, distances were calculated from the experimental (X-ray) structures (pdb entries 1pgb,[20] 1shg,[21] 3cln[22] and 2lzm,[23] respectively). For HPr, a snapshot from an equilibrated MD simulation[24] (initiated from the NMR structure with pdb entry 1hdn[25]) was used to extract the distances. All structures were energy minimised using the GROMOS[26] force field before distances were calculated. All nonpolar hydrogen atoms were included within united carbon atoms (except for aromatic hydrogens in the case of lysozyme). Polar hydrogens were placed using standard GROMOS hydrogen placement. This resulted in 535 atoms for the IgG-binding domain (56 residues), 583 for SH3 (57 residues), 785 for HPr (85 residues), 1364 for calmodulin (143 residues; residues 1–4 and 148 were excluded, since they were not observable in the crystallographic data) and 1703 for lysozyme (164 residues).

TABLE I. Different Classes of Interacting Pairs

1–2	Pairs that are covalently bonded.
1–3	If atom 1 and 2 and atom 2 and 3 are covalently bonded.
Rings	All atom pairs that are part of ring systems.
Side-chain double bonds	ASN, GLN and ARG have one or more (partially delocalized) double bond(s) in the side chain. Torsion angles around these bonds are restricted, making 1–4 pairs (atom 1–2, 2–3 and 3–4 are covalently bonded) around these bonds more restricted than others.
Omega	Distances between C_α atoms from neighboring residues depend on the ω dihedral angle, which is more rigid than the ϕ and ψ torsion angles due to conjugation of the carbonyl bond along the peptide bond, which causes the peptide unit to be rigid and planar (other 1–4 pairs defined by this torsion angle also fall in this category).
Phi/psi	Distances between backbone N atoms depend on ψ dihedral angles, whereas distances between backbone carbonyl C atoms depend on ϕ dihedral angles (other 1–4 pairs defined by ϕ and ψ also fall in this category). ϕ/ψ restricted pairs are subdivided in three groups: —Tight ϕ/ψ: pairs of neighboring residues of which one is a proline and pairs that are part of the same secondary structure element (helix or strand). Backbone dihedrals are relatively more rigid in proline residues because the N and C_α are part of a ring system. Residues in helix and strand conformation have well-defined positions in the Ramachandran plot, from which little deviation is usually observed. —Loose ϕ/ψ: pairs of neighboring residues of which one is a glycine and pairs of residues in loop regions. Glycine residues have relatively much rotational freedom around their ϕ and ψ torsion angles because there is no side chain that induces specific preference for certain ϕ and ψ combinations over others. Loop regions are known to have a relatively poorly defined structure, indicative of conformational flexibility. —Other ϕ/ψ: all other ϕ/ψ restricted pairs.
1–4	Other 1–4 dihedral angle restricted pairs, involving side-chain atoms.
Secondary structure	Pairs of backbone atoms that are part of the same secondary structure element (helix or strand) and are not more than 4 residues apart.
Salt bridge	Oppositely charged groups (all atoms from such a group are restricted) in close proximity ($<$4 Å).
Hydrogen bond	Donor–acceptor distance should not exceed 3.5 Å, the hydrogen–acceptor should not exceed 2.5 Å and the donor–hydrogen–acceptor angle should be minimally 90°.
Tight hydrophobic	Pairs of atoms between which the interatomic distance is smaller than the sum of the van der Waals radii of the involved atoms plus 0.5 Å that do not fall in one of the above categories.
Loose hydrophobic	Identical to tight hydrophobic, but now pairs are included of which the interatomic distance is smaller than the sum of the van der Waals radii of the involved atoms plus 1.0 Å.

The values given in Table II were obtained from an analysis of the distance fluctuations in MD simulations of the B1 IgG-binding domain of streptococcal protein G. The limits were chosen such that the majority of the MD-generated distances is contained within the limits.

Generation of Structures

Having defined distance bounds for all pairs of atoms, the next step is to find structures, other than the reference structure, that fulfil all constraints. We have developed a new, iterative procedure that generates structures fulfilling the requirement that all distances fall between their lower and upper bound. Starting from random coordinates, corrections are applied iteratively to the positions of those atoms that are involved in interatomic distances that violate the upper or lower distance bound. Corrections are applied such that for each violating pair, the distance is put randomly between the upper and lower bound (both atoms involved are displaced by an equal amount). The sum of violations decreases with the number of iterations. The procedure is stopped when the sum of violations is zero. Convergence is usually reached after 100–300 iterations of N steps (N is the number of violations). Occasionally, the algorithm does not converge to a structure satisfying all distance constraints. When the number of iterations exceeds a criterion (typically 500), the algorithm is stopped and restarted with a different set of random starting coordinates. Since no information on chirality is included in the distance bounds, both mirror images are generated. The generated D-amino acid enantiomers are converted into the L form by simply taking the mirror image. The method, called CONCOORD (from CONstraints to COORDinates) resembles a method proposed by Crippen[27] but differs from it in the way the distance corrections are applied.

Since initial coordinates are chosen randomly (from a cube with edges of 2 nm) and distance corrections are applied by choosing distances randomly between their upper and lower bounds, bias in the results is minimal. There is no correlation between any two structures that are generated, and therefore, the accessible space defined by the distance bounds is more efficiently sampled than by procedures in which such correlation is present (like MD).

TABLE II. Parameters Used in the CONCOORD Method*

			pgb	SH3	HPr	cal	lys
	No. of atoms		535	583	785	1364	1703
No.	Type	D (nm)	No. of pairs				
1	1–2	0.002	541	592	792	1376	1723
2	1–3	0.005	780	855	1137	1962	2510
3	Ring	0.01	68	88	34	73	629
4	double bond 1–4	0.01	16	36	40	96	172
5	Omega 1–4	0.01	220	224	336	568	652
6	Tight phi/psi 1–4	0.02	272	190	422	762	893
7	Loose phi/psi 1–4	0.04	120	192	180	265	288
8	Other phi/psi 1–4	0.03	32	56	44	72	76
9	Other 1–4	0.04	254	276	355	624	745
10	Sec. str.	0.05	1556	596	2776	6622	7471
11	Salt bridges	0.075	8	11	1	2	39
12	Hydrogen bonds	0.05	47	60	54	102	86
13	Tight hydrophobic	0.05	278	353	448	741	963
14	Loose hydrophobic	0.1	505	665	714	1132	1571
Total			4697	4194	7333	14388	17818
15	All other pairs	0.5					

*Values indicate the degree of freedom in interatomic distances relative to the experimental structures. The number of distances for all proteins studied in each category are listed.
Abbreviations: pgb, the B1 IgG-binding; cal, calmodulin; lys, lysozyme.

For all proteins studied, 500 structures were generated with the CONCOORD method. For the IgG binding domain (56 residues), ~1 hour of CPU time on a Pentium 100 processor was required (for comparison: a number of weeks would be required for an MD simulation of 1 ns). The speed could be improved by introduction of a cutoff radius for interatomic distances or other methods that reduce the number of pairs that need to be corrected every iteration step. However, the method in its present implementation is fast enough for all practical purposes. Starting from coordinates other than randomly chosen ones may also enhance convergence speed, but since the correction algorithm is particularly efficient in the initial stage and because we want to minimize the amount of bias in the results, we preferred random starting coordinates.

All information on structural variability is stored in the upper and lower distance bounds. Therefore, it should in principle be possible to extract this information directly from the distance bounds, without first generating structures. We have not been able to derive an analytical solution, but an approximation is possible. Given an interaction function, a way to gain insight in the most prominent modes of motion is by diagonalization of the (mass-weighted) Hessian matrix, as in Normal Modes (NM) analyses.[28–30] The matrix elements correspond to second derivatives of the potential energy with respect to the coordinates. The simplest way to implement distance restrictions in such an interaction function is to model all pair interactions by harmonic potentials, with the minimum defined at the distance measured in the experimental structure and the force constant inversely proportional to the difference between upper and lower distance bound (all masses are put to 1.0). Eigenvectors of the Hessian matrix that have the smallest eigenvalues (apart from those that correspond to overall rotation and translation) are directions in configurational space that represent the slowest vibrations of a molecular system. In a detailed force field, these directions have been shown to be similar to the eigenvectors with largest eigenvalues from Principal Component analyses of MD trajectories,[15,31] although normal modes have the restriction of harmonicity.

Starting from the same distance bounds, diagonalization of the Hessian matrix will yield results that are somewhat different from those obtained from diagonalisation of the covariance matrix of positional fluctuations from generated structures for a number of reasons. First, during generation of structures, some distance bounds will never be reached because they are excluded by the presence of other distance limits. Therefore, bound smoothing on the triangulation level[32] had to be performed before calculation of the Hessian matrix. Second, distributions of distances are assumed to be Gaussian in the harmonic approximation, whereas no such assumption is made during the generation of structures in CONCOORD, where the distance distribution may even be asymmetric.

Analysis Techniques

Essential Dynamics[6] analyses were used for comparison of structural freedom in proteins. The method consists of diagonalization of the covariance matrix C of atomic fluctuations, after removal of overall

TABLE III. Average Geometrical Properties for Structures Generated by MD and CONCOORD, Compared With the Values Obtained From Experimental Structures

	RMSD	NRC	HBO	ACC	GYR	DIH	QUAL	ENE
pgb PDB	0.00	8.0	39.0	3391	1.021	1.0	−0.083	−2241
pgb MD	1.43	9.6	44.2	3840	1.023	2.68	−0.662	−2005
pgb CONCOORD	1.04	7.3	42.3	3673	1.023	1.86	−0.337	−2140
SH3 PDB	0.00	14.0	38.0	3665	1.012	3.0	−0.668	−2975
SH3 MD	1.29	14.8	40.0	4051	1.026	2.03	−1.231	−2816
SH3 CONCOORD	0.81	13.3	44.5	3858	1.001	2.94	−0.639	−2811
HPr PDB	0.00	12.0	74.0	4840	1.146	5.0	−0.553	−4237
HPr MD	1.39	14.1	67.9	5031	1.147	5.09	−0.741	−4252
HPr CONCOORD	0.90	12.1	73.2	4892	1.126	4.52	−0.540	−4223
cal PDB	0.00	20.0	110.0	9355	2.095	5.0	−0.160	−7428
cal MD	2.65	21.3	99.4	9851	2.113	10.42	−0.728	−7505
cal CONCOORD	1.93	17.9	108.9	9788	2.091	5.46	−0.509	−7484
lys PDB	0.00	16.0	122.0	8675	1.590	5.0	−0.228	−10869
lys MD	1.81	20.4	122.3	8863	1.562	7.21	−0.953	−10740
lys CONCOORD	1.57	19.7	124.8	8585	1.581	7.91	−0.891	−10493

Abbreviations: pgb, the B1 IgG-binding domain; cal, calmodulin; lys, lysozyme; RMSD, root mean square deviation, expressed in Å; NRC, number of residues in random coil conformation, according to DSSP[34] ; HBO, number of main chain hydrogen bonds (DSSP); ACC, total solvent accessible surface in Å^2 (DSSP); GYR, radius of gyration in nm; DIH, number of residues in unfavorable regions in Ramachandran plot[35,44]; QUAL, WHAT IF index indicating the normality of packing[45]; ENE, potential energy after energy minimization in the GROMOS force field.

translation and rotation:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad (1)$$

where x are cartesian atomic coordinates. Resulting eigenvectors are directions in configurational space of which the corresponding eigenvalues give the mean square fluctuation of the displacement in each direction. ED analyses can be applied to any (sub)set of coordinates of the studied molecular system. Only C_α atoms were included in ED analyses presented here because it has been shown[6,8,15] that this approach best detects the large-scale concerted motions in proteins.

The software for the generation of structures will be available on the WWW (http://rugmd0.chem.rug.nl) and is implemented in the WHAT IF[33] package. ED and all other structural analyses were performed using an interface in the molecular modeling package WHAT IF.[33] Secondary structure analyses and accessible surface calculations were performed with DSSP.[34] Dihedral angle criteria were taken from PROCHECK.[35]

RESULTS

All CONCOORD structures were subjected to a number of structural analyses to assess how physically realistic the generated structures are (Table III). The same analyses were performed on structures sampled by MD (for simulation details: the IgG-binding domain[16] (1 ns), SH3[9] (1 ns), HPr[24] (300 ps), calmodulin[36] (500 ps), and lysozyme (submitted for publication) (1 ns). All MD simulations were performed in explicit solvent at room temperature. Comparison with crystal structures and MD shows that in CONCOORD, with the present set of parameters, structures generally are more similar to their respective experimental structure than in MD. There is good correspondence between the values obtained from MD and CONCOORD for all properties taken into account. Mean square atomic fluctuations of C_α atoms are plotted in Figure 1 for both CONCOORD and MD. There is reasonable qualitative correlation between curves obtained from CONCOORD and MD (correlation coefficients between 0.501 and 0.871).

For all molecules studied the ensembles of conformations generated by MD and CONCOORD were subjected to essential dynamics analyses. In all cases only a few eigenvectors were found with significant eigenvalues. These eigenvalues are shown in Figure 2 (eigenvalues have been sorted by decreasing value). Eigenvalue curves from both techniques are equally steep for all proteins, indicating that also from the CONCOORD results only a few collective fluctuations emerge with appreciable freedom.

Inner products between eigenvectors from MD and CONCOORD were calculated to evaluate whether eigenvectors obtained from both techniques represent similar fluctuations. Squared inner products are shown for every pair of eigenvectors from MD and CONCOORD for the B1 IgG-binding domain in Figure 3a. All high inner products are found close to the diagonal, meaning that for both techniques, directions in configurational space are ordered similarly with respect to the amount of fluctuation, that is, directions that show large fluctuations in MD also show relatively large fluctuations in CONCOORD, and vice versa. Figure 3b shows the squared inner products between eigenvectors obtained from two halves of an MD simulation of 1 ns. The overlap

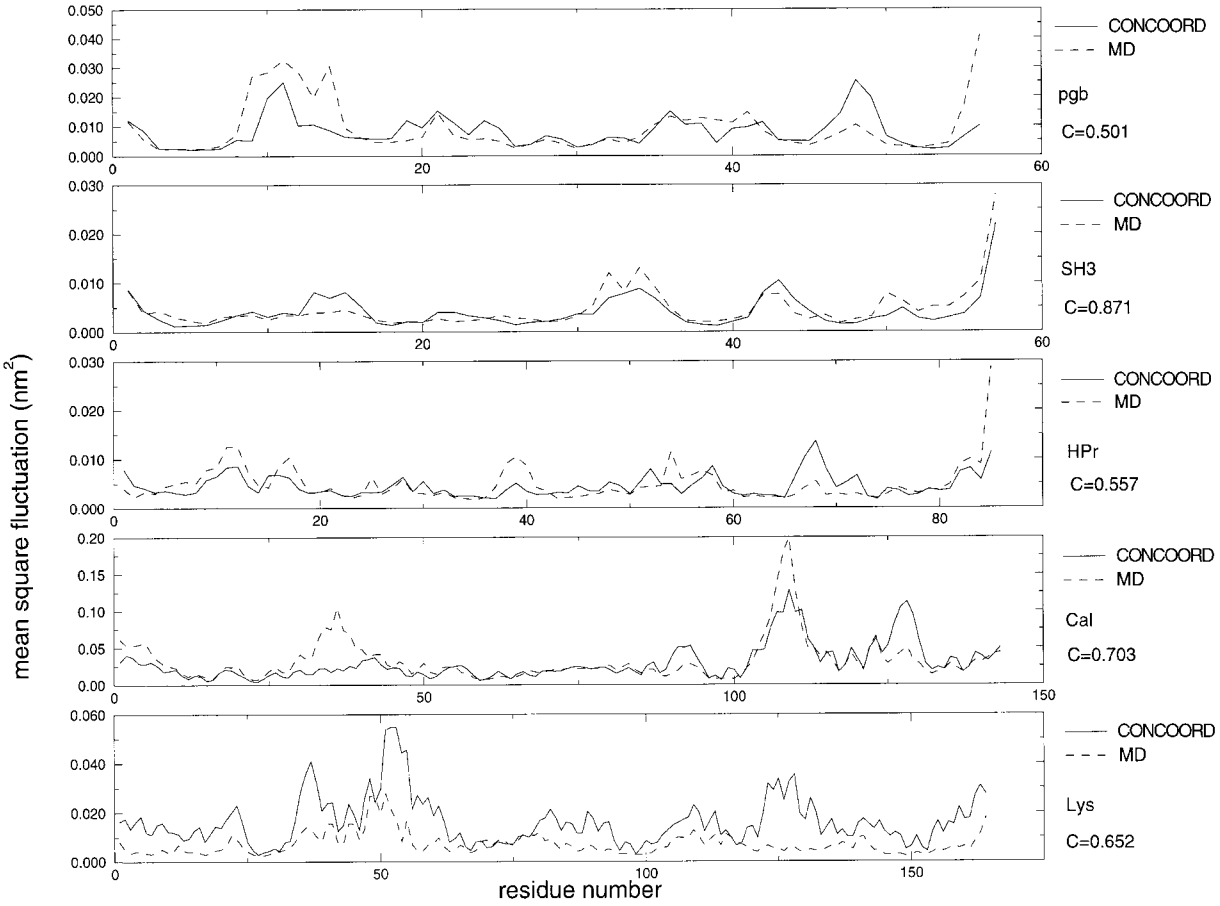


Fig. 1. Mean square positional fluctuation of C_α atoms. The correlation coefficient between the curves obtained from MD and CONCOORD is shown next to the figures.

between the two eigenvector sets from MD is similar to that between MD and CONCOORD. In Figure 3c the same comparison is made for two sets of structures obtained by CONCOORD. Two independent sets of 250 structures were used in the ED analyses.

Figure 3 shows that the overlap between MD and CONCOORD is especially high in the essential subspace (defined arbitrarily as the subspace spanned by the 10 eigenvectors with largest eigenvalues). The overlap of the essential subspaces from MD and CONCOORD has been evaluated in a more quantitative way because the essential subspace is of particular interest (about 80% of the observed structural fluctuation usually occurs in this subspace). Figure 4 shows the mean cumulative squared inner products between eigenvectors (from MD and CONCOORD) spanning this subspace and the first 50 eigenvectors from independent MD/CONCOORD runs, for the IgG binding domain. Overlap is concentrated in the initial part. For example, 80% of overlap with the first 10 CONCOORD eigenvectors is reached within the first 20 MD eigenvectors, indicating that all essential directions found by CONCOORD are also accessible in MD. The overlap between eigenvectors from two independent MD runs is very similar to the overlap between CONCOORD and MD, whereas the overlap between two independent CONCOORD runs is very close to the maximum possible overlap, indicating an almost complete convergence.

The mean squared inner products between the 10 eigenvectors with largest eigenvalues from MD and CONCOORD are given in Table IV, for all proteins studied. The overlap between the essential subspaces obtained by MD and CONCOORD is comparable to the overlap obtained from the two halves of each MD trajectory. A typical overlap of $\sim$0.5 is obtained for all proteins (a value of 1.0 would be obtained if the two sets are identical). Overlap between eigenvectors obtained from two parts of the clusters produced by CONCOORD is significantly larger for all proteins.

Overlap of the 10 CONCOORD eigenvectors with largest eigenvalues with the 10 lowest frequency-eigenvectors obtained from diagonalization of the Hessian matrix was calculated to be 0.678 for the B1 IgG-binding domain (C_α components were extracted from the eigenvectors of the Hessian matrix and the obtained vectors were renormalised before the analy-

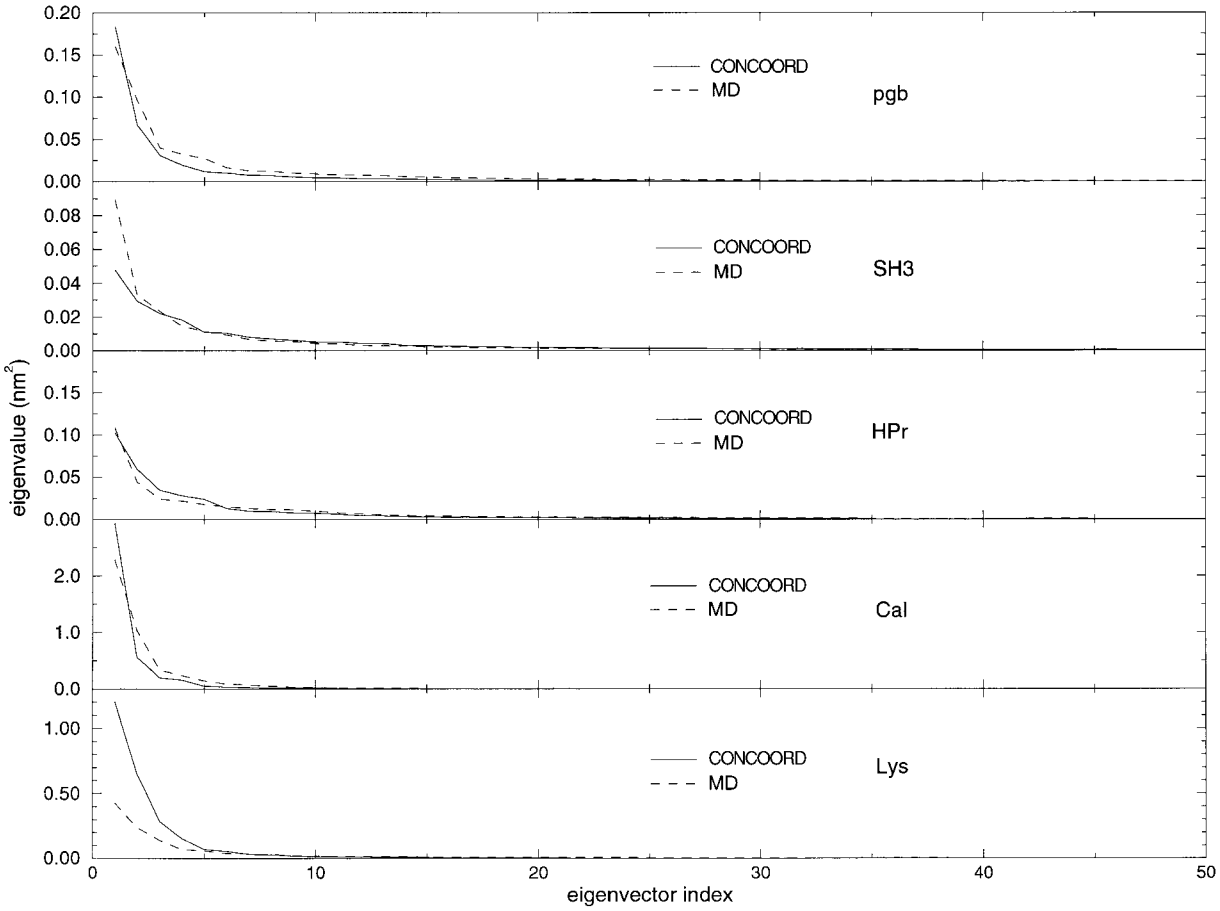


Fig. 2. Eigenvalues obtained from MD trajectories and ensembles of structures generated by CONCOORD. Only the 50 largest eigenvalues are shown out of 168 (pgb, B1 IgG-binding domain), 171 (SH3), 255 (HPr), 429 (cal), and 492 (lys), respectively.

sis). This value is somewhat smaller than the overlap between eigenvectors obtained from two clusters of CONCOORD structures (0.866), indicating that small deviations from the converged CONCOORD results emerge in this approximation. The overlap of the Hessian eigenvectors with MD eigenvectors was calculated to be 0.486. This is slightly lower than the overlap of the eigenvectors obtained from CONCOORD structures with MD eigenvectors (0.532).

The difference in the way the conformational space is sampled in CONCOORD and MD is illustrated in Figure 5. In MD (Fig. 5a), a single path is followed that resembles a random walk,[13,14,17] whereas in CONCOORD (Fig. 5b), a random sampling takes place, with each position independent from the previous one. To investigate in more detail to which extent the modes of motion predicted by CONCOORD are accessible in MD, an extended MD simulation with constraints on the two CONCOORD eigenvectors with largest eigenvalues was performed. The way in which these constraints are applied makes it possible to efficiently assess the portion of the conformational space that is accessible to MD.[13,14,17] As can be seen from Figure 5c, the region sampled by this technique is similar to the region sampled by CONCOORD.

Structures collected along the most important directions defined by CONCOORD are shown in Figure 6 for calmodulin and lysozyme. The CONCOORD eigenvector with largest eigenvalue for calmodulin corresponds to a combination of a bend and a twist of the interdomain helix, resulting in a rotation of one domain with respect to the other (Fig. 6a). From experiments (hydrogen exchange measurements,[37] NMR relaxation data[38] and NMR NOE data[39] from which disorder in the set of NMR structures[40] emerged), the helix is known to break in the middle, which was also observed in MD and Normal Modes analyses.[36]

For lysozyme, the CONCOORD eigenvector with second largest eigenvalue corresponds to a fluctuation that is similar to structural differences that have been observed by crystallography of a number of mutants[41] (Fig. 6b). The main domain fluctuation consists of a rotation of the two domains with respect to each other, initiated by a combined twisting and

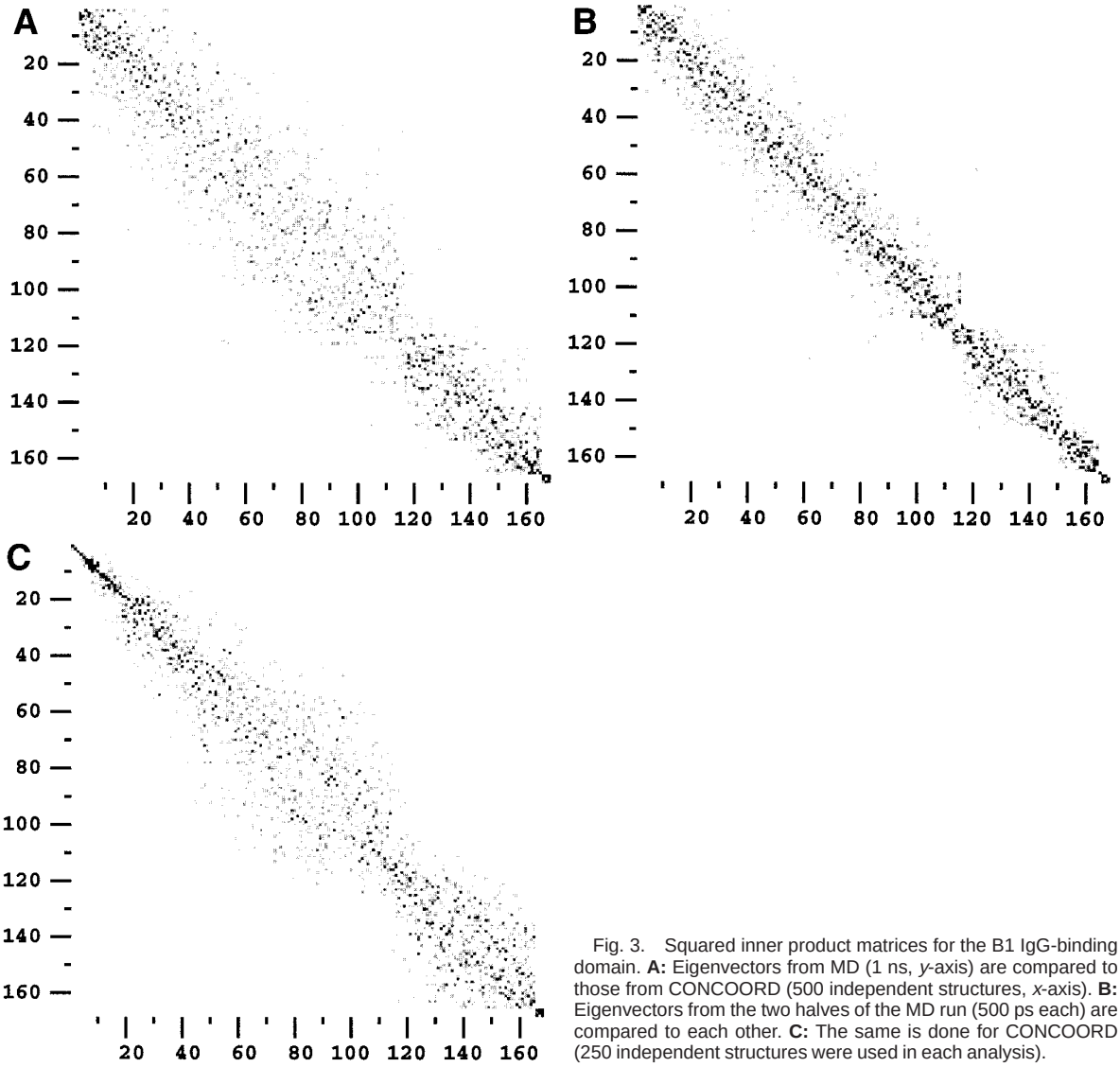


Fig. 3. Squared inner product matrices for the B1 IgG-binding domain. **A:** Eigenvectors from MD (1 ns, *y*-axis) are compared to those from CONCOORD (500 independent structures, *x*-axis). **B:** Eigenvectors from the two halves of the MD run (500 ps each) are compared to each other. **C:** The same is done for CONCOORD (250 independent structures were used in each analysis).

bending of the interdomain helix. The difference between the most open[41] and the most closed[42] X-ray structure along this rotation axis is as much as 49°. The angular difference between the most open and most closed CONCOORD structure was 33°; for MD this value was 28°. Both CONCOORD and MD do not reach the most open experimental configuration.

DISCUSSION

The results show that there are many similarities between MD and CONCOORD. However, there is also a number of apparent discrepancies. In Figure 1, a number of peaks are only observed in the curves obtained from CONCOORD and not from MD, or vice versa. The broad peak near residue 48 (located in the turn connecting β strands 3 and 4) for the B1 IgG-binding domain in CONCOORD that is not present in the curve from MD represents fluctuations that are dominating the CONCOORD eigenvector with largest eigenvalue. This direction is not present within the first two eigenvectors from MD, but is represented 75% by the first six MD eigenvectors, indicating that this motion is also accessible in MD. Likewise, the peak near residue 39 for calmodulin (a surface loop connecting helices 2 and 3) in MD is mostly the result from the motion along the first MD eigenvector. This mode of motion shows little overlap with the first five eigenvectors of CONCOORD but is contained for 75% in the first 15 CONCOORD eigenvectors, indicative of significant fluctuation in the cloud of CONCOORD structures.

The similarity of the MD and CONCOORD results is remarkable, since both techniques differ on several fundamental points. First, the interaction function between particles is much more complex in MD than in CONCOORD, in terms of the number of parameters that determine the amount and kind of fluctuations that are accessible. In the current imple-

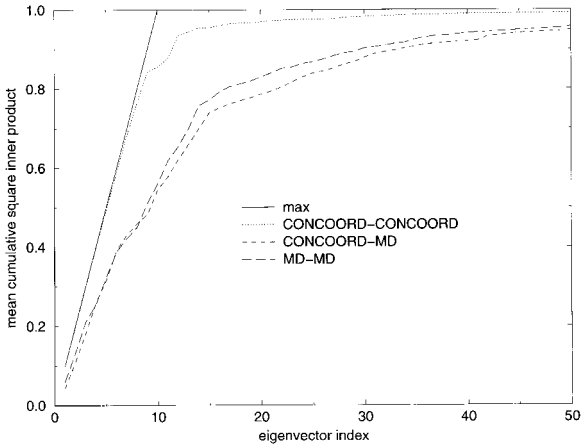


Fig. 4. Cumulative mean square inner products between the 10 eigenvectors with largest eigenvalues obtained from MD/CONCOORD and all eigenvectors obtained from different MD/CONCOORD runs. After division by 10, all curves converge to 1.0, since every eigenvector from one set is contained in the complete set of vectors from another set. The solid line corresponds to the maximum obtainable overlap. pgb denotes the B1 IgG-binding domain.

TABLE IV. Mean Squared Inner Products Between Subsets Containing the 10 Eigenvectors With Largest Eigenvalues

	Mean cumulative square inner product		
Protein	MD–CONCOORD	MD–MD	CONCOORD–CONCOORD
pgb	0.532	0.560	0.866
SH3	0.446	0.494	0.809
HPr	0.416	0.387	0.904
cal	0.440	0.532	0.802
lys	0.454	0.487	0.910

The first column contains a comparison between MD and CONCOORD, the second column compares two halves of each MD trajectory, which is done in the third column for CONCOORD. pgb denotes the B1 IgG-binding domain.

mentation, a total of only 15 parameters is sufficient. Second, in CONCOORD only short-range interactions (roughly smaller than 6 Å) within the protein make a serious contribution, whereas in MD long-range interactions and interactions with solvent are also included. Additionally, all interactions are implemented in the form of distance constraints in CONCOORD. In MD, usually only bond lengths are described this way. Another important difference between MD and CONCOORD is the way in which structures are generated. In MD, the equations of motions are integrated numerically to yield a unique path in configurational space, where each structure is a deterministic result of the previous one. In CONCOORD, structures are generated by a random search method that searches for solutions in a predefined coordinate space. Incomplete sampling is one of the dominating reasons for errors in the definition of an essential subspace from MD simulation.[3,4,16] The fact that the overlap between CONCOORD and MD is similar to the overlap between different parts of MD simulations suggests that these errors are of the same order of magnitude as the errors made in CONCOORD due to a too simple model.

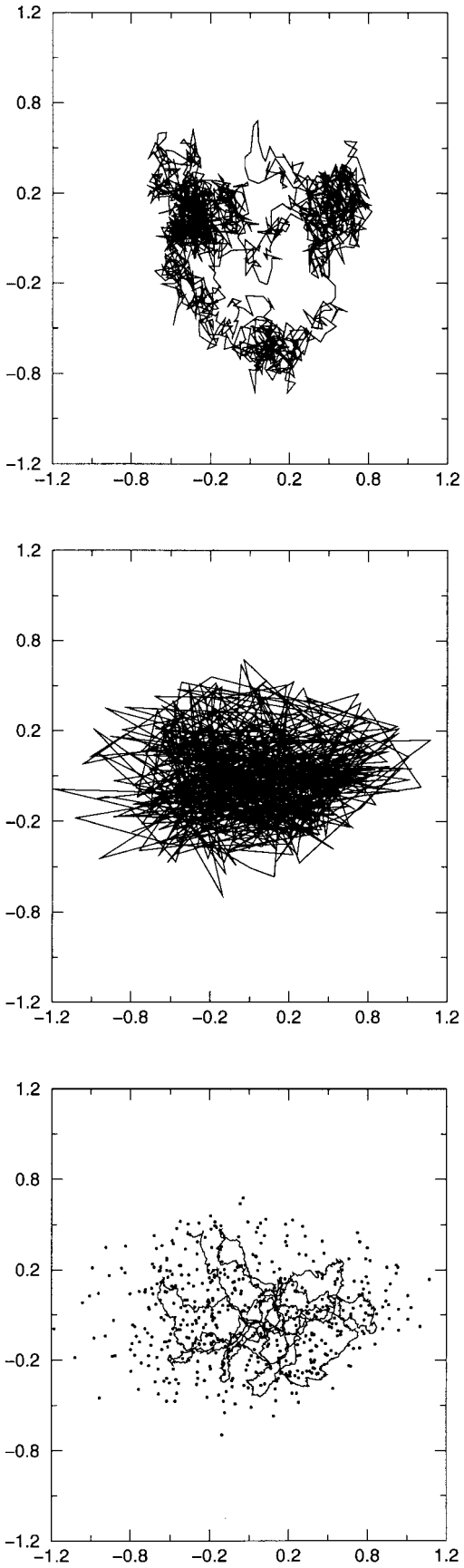
The differences between MD and CONCOORD imply that not all the data that can be obtained by MD can also be obtained by CONCOORD. Dynamic (time-dependent) information, for example, cannot be derived from CONCOORD data. Also, the amplitude of predicted fluctuations can only be derived in a relative sense, that is, the method only predicts certain modes to be more accessible than others. For example, the hinge bending mode in lysozyme was not sampled in the same range as in experiment. However, this also holds for an MD simulation of 1 ns. The local cause of a large overall structure variation cannot be deduced reliably from an analysis of CONCOORD results. The main motion in calmodulin, for example, is known to be the result of the breaking of the interdomain helix. Such a rigorous event is not allowed within the distance bounds as they are defined now. However, it is interesting to note that even in the case of such large conformational changes, the first stage of such changes is already sampled and, in the case of calmodulin, emerges as the fluctuation with largest amplitude.

The comparison of eigenvectors obtained from diagonalization of the Hessian matrix with those from CONCOORD and MD indicates that even without the generation of structures, a rough approximation can be obtained of the subspace in which all significant backbone motions take place. Diagonalization of the Hessian matrix is faster than the generation of a large enough set of structures by CONCOORD for a covariance analysis. In most cases the generation of structures is to be preferred, however, since the produced structures can also be used for other analyses, and the CONCOORD eigenvectors show better overlap with MD.

The parameters used for CONCOORD (Table II) were generated for the B1 IgG-binding domain, but they were applicable without modifications for the other proteins and gave meaningful results. The values in Table III indicate that a set of physically realistic structures has been generated by CONCOORD for all proteins studied.

Structural Variation in Clusters of NMR Structures

A significant level of correlation between essential directions defined from MD and from clusters of NMR structures has been found for a number of proteins (unpublished observations). For the B1 IgG-binding domain of streptococcal protein G for instance, the summed square inner products of the 10 eigenvectors with largest eigenvalues from MD and NMR was found to be 0.35, comparable to the



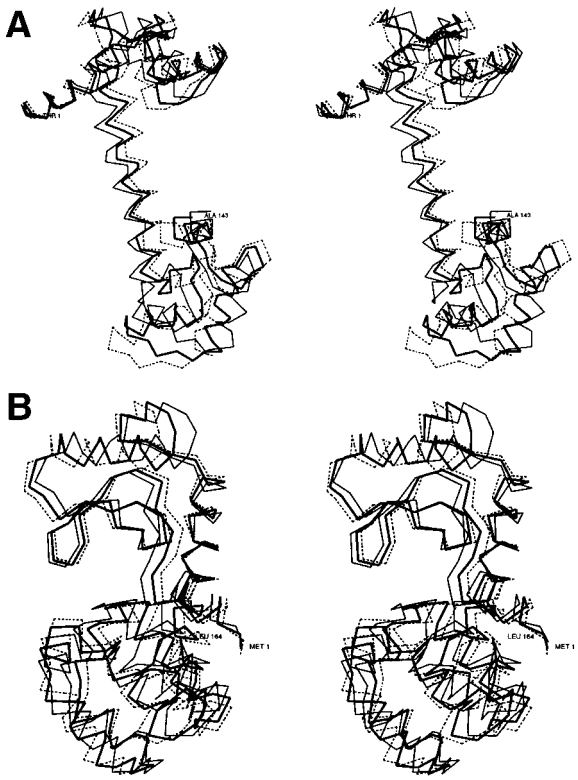


Fig. 6. Stereo representation of extreme structures (thin line and thin dashed line) along CONCOORD eigenvectors, together with average structures (bold line). **A:** Calmodulin, eigenvector 1. **B:** Lysozyme, eigenvector 2.

values in Table IV. In a recent study, a similar observation was reported[43] for BPTI. The amount of dynamic information that can be derived from NMR/NOE data has been subject of discussion. It has been argued[18,19] that the amount of information usually used for structure generation from NMR data is generally too limited to yield information on the conformational flexibility of macromolecules. In line with the results presented in this paper, however, methods that provide a set of protein structures in which all structural constraints are fulfilled can be expected to give insight into the conformational flexibility of these molecular systems. The information derived from a cluster of NMR structures is only partially the result of the experimental data used in the analysis. In NMR structure refinement, not only the experimentally derived (distance) restrictions are used for the analyses, also knowledge of, for

Fig. 5. **a (top):** Projection of the MD trajectory of the IgG binding domain and **b (middle):** the collection of CONCOORD structures onto the planes defined by the two eigenvectors with largest eigenvalues from both techniques. **c (bottom):** Projection of CONCOORD (small circles) and extended MD (continuous line) structures onto the plane defined by the two CONCOORD eigenvectors with largest eigenvalues.

instance, bond lengths and angles is usually included to generate structures. The collection of these restraints restricts the generated configurations to such an extent that meaningful information about (the few) important collective degrees of freedom may be derived from such analyses.

CONCLUSIONS

We have shown that the major fluctuations in protein structures that are predicted by CONCOORD are concentrated in a few directions in configurational space. Apparently, the bounds on interatomic distances, which are on one hand defined by the connectivities in the structure (covalent bonds) and on the other hand by the way the protein is folded (hydrogen bonds, salt bridges, hydrophobic contacts), restrict the conformational freedom of these systems such that only a few collective degrees of freedom fluctuate significantly. Apart from the disadvantages that no time dependent information is obtained and that the extent and structural cause of the fluctuations cannot be determined, an almost converged description of the most important collective degrees of freedom is obtained when only a limited number of structures has been generated. It has been shown that it is not necessary to use sophisticated atomic interaction functions to obtain basic knowledge about the structural fluctuations of proteins in solution. The sum of all interactions in proteins makes fluctuations to be concentrated in a few collective degrees of freedom which can be obtained by a straightforward method. The minimal computational effort involved allows for the screening of fluctuations in many configurations, which could, for example, facilitate the design of mutants, or enhance the capabilities of homology prediction.

ACKNOWLEDGMENTS

We thank David van der Spoel for kindly providing the calmodulin MD trajectory, Nico van Nuland for the HPr MD trajectory, and Michael Nilges (EMBL, Heidelberg) for critically reading the manuscript.

REFERENCES

1. Elofsson, A., Nilsson, L. A 1.2 ns Molecular Dynamics simulation of the ribonuclease T 1)-3′-guanosine monophosphate complex. J. Phys. Chem. 100:2480–2488, 1996.
2. Brunne, R.M., Berndt, K.D., Güntert, P., Wüthrich, K., Van Gunsteren, W.F. Structure and internal dynamics of the bovine pancreatic trypsin inhibitor in aqueous solution from long-time Molecular Dynamics simulations. Proteins. 23:49–62, 1995.
3. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M., Phillips Jr., G.N. A sampling problem in molecular dynamics simulations of macromolecules. Proc. Natl. Acad. Sci. U.S.A. 92:3288–3292, 1995.
4. Balsera, M.A., Wriggers, W., Oono, Y., Schulten, K. Principal component analysis and long time protein dynamics. J. Phys. Chem. 100:2567–2572, 1996.
5. Jorgensen, W.L., Tirado-Rives, J. Monte Carlo vs Molecular Dynamics for conformational sampling. J. Phys. Chem. 100:14508–14513, 1996.
6. Amadei, A., Linssen, A.B.M., Berendsen, H.J.C. Essential dynamics of proteins. Proteins 17:412–425, 1993.
7. Van Aalten, D.M.F., Amadei, A., Vriend, G., Linssen, A.B.M., Venema, G., Berendsen, H.J.C., Eijsink, V.G.H. The essential dynamics of thermolysin: confirmation of hinge-bending motion and comparison of simulations in vacuum and water. Proteins 22:45–54, 1995.
8. Van Aalten, D.M.F., Findlay, J.B.C., Amadei, A., Berendsen, H.J.C. Essential dynamics of the cellular retinol binding protein: evidence for ligand induced conformational changes. Prot. Eng. 8:1129–1136, 1995.
9. Van Aalten, D.M.F., Amadei, A., Bywater, R., Findlay, J.B.C., Berendsen, H.J.C., Sander, C., Stouten, P.F.W. A comparison of structural and dynamic properties of different simulation methods applied to SH3. Biophys. J. 70:684–692, 1996.
10. Garcia, A.E. Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699, 1992.
11. Hayward, S., Kitao, A., Hirata, F., Gō, N. Effect of solvent on collective motions in globular proteins. J. Mol. Biol. 234:1207–1217, 1993.
12. Aalten, D., Jones, P., Sousa, M., Findlay, J. Engineering protein mechanics: inhibition of concerted motions of the cellular retinol binding protein by site-directed mutagenesis. Prot. Eng. 10:31–38, 1997.
13. Amadei, A., Linssen, A.B.M., De Groot, B.L., Van Aalten, D.M.F., Berendsen, H.J.C. An efficient method for sampling the essential subspace of proteins. J. Biom. Str. Dyn. 13(4):615–626, 1996.
14. De Groot, B.L., Amadei, A., Scheek, R.M., Van Nuland, N.A.J., Berendsen, H.J.C. An extended sampling of the configurational space of HPr from *E. coli*. Proteins: 26:314–322, 1996.
15. Van Aalten, D.M.F., De Groot, B.L., Berendsen, H.J.C., Findlay, J.B.C., Amadei, A. A comparison of techniques for calculating protein essential dynamics. J. Comp. Chem. 18:169–181, 1997.
16. De Groot, B.L., Van Aalten, D.M.F., Amadei, A., Berendsen, H.J.C. The consistency of large concerted motions in proteins in Molecular Dynamics simulations. Biophys. J. 71:1554–1566, 1996.
17. De Groot, B.L., Amadei, A., Van Aalten, D.M.F., Berendsen, H.J.C. Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. J. Biomol. Str. Dyn. 13:741–751, 1996.
18. Bonvin, A.M.J.J., Brünger, A.T. Conformational variability of solution nuclear magnetic resonance structures. J. Mol. Biol. 250:80–93, 1995.
19. Bonvin, A.M.J.J., Brünger, A.T. Do NOE distances contain enough information to assess the relative populations of multi-confomer structures? J. Biomol. NMR 7:72–76, 1996.
20. Gallagher, T., Alexander, P., Bryan, P., Gilliland, G.L. Two crystal structures of the B1 Immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 33:4721–4729, 1994.
21. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., Saraste, M. Crystal structure of a Src-homology 3 (SH3) domain. Nature 359:851–854, 1992.
22. Babu, Y.S., Bugg, C.E., Cook, W.J. Structure of calmodulin refined at 2.2 Å. J. Mol. Biol. 204:191–204, 1988.
23. Weaver, L.H., Matthews, B.W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. J. Mol. Biol. 193:189–199, 1987.
24. Van Nuland, N.A.J., Boelens, R., Scheek, R.M., Robillard, G.T. High resolution structure of the phosphorylated form of the histidine-containing phosphocarrier protein HPr from *Escherichia coli* determined by restrained molecular dynamics from NMR-NOE data. J. Mol. Biol. 246:180–193, 1995.
25. Van Nuland, N.A.J., Hangyi, I.W., Van Schaik, R.C., Berendsen, H.J.C., Van Gunsteren, W.F., Scheek, R.M., Robillard, G.T. The high-resolution structure of the histidine-containing phosphocarrier protein HPr from *Escherichia coli* determined by restrained molecular dynamics from NMR-NOE data. J. Mol. Biol. 237:544–559, 1994.

26. Van Gunsteren, W.F., Berendsen, H.J.C. Gromos manual. BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands 1987.
27. Crippen, G.M. A novel approach to calculation of conformation: Distance Geometry. J. Comp. Phys. 24:449–452, 1977.
28. Levitt, M., Sander, C., Stern, P.S. Protein normal-mode dynamics: trypsin-inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.
29. Gō, N., Noguti, T., Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. U.S.A. 80:3696–3700, 1983.
30. Brooks, B.R., Karplus, M. Harmonic dynamics of proteins: Normal Modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. U.S.A. 80:6571–6575, 1983.
31. Hayward, S., Kitao, A., Gō, N. Harmonicity and Anharmonicity in Protein Dynamics: a Normal Modes and Principal Component analysis. Proteins. 23:177–186, 1995.
32. Havel, T.F., Kuntz, I.D., Crippen, G.M. The theory and practice of Distance Geometry. Bull. Math. Biol. 45:665–720, 1983.
33. Vriend, G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8:52–56, 1990.
34. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
35. Laskowski, R.A., MacArthur, M., Moss, D.S., Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26:283–291, 1993.
36. Van der Spoel, D., De Groot, B.L., Hayward, S., Berendsen, H.J.C., Vogel, H.J. Bending of the calmodulin central helix: a theoretical study. Prot. Sci. 5:2044–2053, 1996.
37. Spera, S., Ikura, M., Bax, A. Measurements of the exchange rates of rapidly exchanging amide protons: application to the study of calmodulin and its complex with a myosin light chain kinase fragment. J. Biomol. NMR 1:155–165, 1991.
38. Barbato, G., Ikura, M., Kay, L.E., Pastor, R.W., Bax, A. Backbone dynamics of calmodulin studied by ^{15}N relaxation using inverse detected NMR spectroscopy: the central helix is flexible. Biochemistry 31:5269–5278, 1992.
39. Ikura, M., Spera, S., Barbato, G., Kay, L.E., Krinks, M., Bax, A. Secondary structure and side-chain ^{1}H and ^{13}C resonance assignments of calmodulin in solution by heteronuclear multidimensional NMR spectroscopy. Biochemistry 30:9216–9228, 1991.
40. Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B., Bax, A. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. Science 256: 632–638, 1992.
41. Zhang, X.-J., Wozniak, J.A., Matthews, B.W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. 250:527–552, 1995.
42. Matsumura, M., Signor, G., Matthews, B.W. Substantial increase of protein stability by multiple disulphide bonds. Nature 342:291–293, 1989.
43. Berndt, K.D., Güntert, P., Wüthrich, K. Conformational sampling by NMR solution structures calculated with the program DIANA evaluated by comparison with long-time Molecular Dynamics calculations in explicit water. Proteins. 24:304–313, 1996.
44. Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7:95–99, 1963.
45. Vriend, G., Sander, C. Quality control of protein models: directional atomic contact analysis. J. Appl. Crystallogr. 26:47–60, 1993.