Native State Protein Dynamics:

A Theoretical Approach

RIJKSUNIVERSITEIT GRONINGEN

Native State Protein Dynamics:

A Theoretical Approach

Proefschrift
ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, Dr. D.F.J. Bosscher,
in het openbaar te verdedigen op
vrijdag 22 januari 1999
om 16.00 uur

door

Berend Lammert de Groot

geboren op 24 augustus 1971
te Emmen

Promotor: Prof. Dr. H.J.C. Berendsen

If this doesn't make you free
It doesn't mean you're tied
If this doesn't take you down
It doesn't mean you're high
If this doesn't make you smile
You don't have to cry
If this isn't making sense
It doesn't make it lies
(Soundgarden)

Beoordelingscommissie:

Prof. dr. B.W. Dijkstra

Prof. dr. G.T. Robillard

Prof. dr. Ir. B. Smit

Contents

1 INTRODUCTION

Protein dynamics

The study of protein dynamics by computer simulations

A large diversity of processes in living organisms critically depend on protein activity. Though in many of these processes the mere structure of a protein dominates its function (e.g. collagen in tissues or α-keratin in hair), protein dynamics is crucial to many others. Virtually all biological processes that involve motion find their origin in protein dynamics. Muscle contraction, for instance, is based on the combined action of actin and myosin. Other examples are the molecular motors kinesin and F1-ATPase. Dynamics also plays an important role in many proteins of which the primary function is not mobility itself. For example, the ability to change conformation is also essential for the function of many transport proteins, proteins involved in signal transduction, proteins in the immune system, and numerous enzymes[1]. In many enzymes, conformational changes serve to enclose the substrate, thereby preventing its release from the protein and ideally positioning it for the protein to perform its function, as in lysozyme. Immunoglobulins are highly flexible in order to be able to deal with a large range of ligands. Another role of protein dynamics is found in G-proteins, binding of a hormone to its receptor triggers the dissociation of the α domain from the rest of the protein after a GTP-mediated conformational change. A special class of conformational transitions are found in so-called allosteric proteins. Substrate binding to one subunit of these multimeric proteins triggers a conformational change that alters the substrate affinity of the other subunits, thereby sharpening the switching response of these proteins.

The conformational changes involved range from very subtle, local changes, as in the case of e.g. myoglobin, to global conformational changes, involving motions of significant amplitude for large parts of a protein (e.g. haemoglobin)[1]. Dynamics plays an important role not only in the functional, native state of many proteins, but also the mechanism by which a protein reaches that native conformation, the protein folding process, is a highly dynamic process.

Although a large part of the current knowledge of conformational flexibility in proteins is derived from experimental data (especially X-ray crystallography and Nuclear Magnetic Resonance (NMR)), there is currently no experimental technique that allows monitoring of protein conformational changes at atomic resolution as a function of time at time-scales of nanoseconds. There are several examples of proteins structurally characterised when trapped in different functional states (for an overview, see ref. 2), and the time resolution of structural studies improves steadily[3]. Nevertheless, details on the pathways between different known conformations often remain obscure. Until now, computer simulation techniques provide the only possibility to obtain dynamic information on proteins at atomic resolution in the picosecond to microsecond time range.

Molecular Dynamics

Out of all possible ways of simulating protein motions, Molecular Dynamics (MD) techniques are among the most popular. In MD, an attempt is made to describe the time evolution of molecular systems as realistically as possible. In a typical simulation, a starting configuration is generated from an experimentally determined structure, and put in an environment that best mimics its natural environment. Obviously, the quality of the obtained dynamic model depends on the quality of the starting model. Once an appropriate starting configuration has been obtained, the actual simulation can be started. In most cases, all particles are treated classically, leaving the problem of solving Newton's equations of motion:

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i \tag{1.1}$$

with $\boldsymbol{F}_i$ the force, m_i the mass and $\boldsymbol{a}_i$ the acceleration of particle i. Atomic positions $\boldsymbol{x}$ are obtained from:

$$\boldsymbol{a} = \frac{d^2\boldsymbol{x}}{dt^2} \tag{1.2}$$

by numerical integration. At every integration step $\boldsymbol{F}$ is evaluated using:

$$\boldsymbol{F} = -\frac{d\boldsymbol{V}}{d\boldsymbol{x}} \tag{1.3}$$

The potential energy $\boldsymbol{V}$ typically includes terms for covalent bond lengths, angles, torsion angles (dihedrals), improper dihedrals (to maintain tetrahedral or planar geometries), and a number of non-bonded terms[4–6]. The non-bonded terms typically consist of a Lennard-Jones term and an electrostatic (Coulomb) contribution, and in some cases an explicit hydrogen-bonding term. Due to the lack of quantum-mechanical terms, specific parameters must be specified for each atom type in each chemical environment. This results in a parameter-set (force-field) that contains many hundreds of parameters. The absence of polarisability in classical force-fields restricts the reliability of MD simulations, especially in systems where polarisability effects are known to play an important role, as for example in ion-binding proteins. Another potential source of artifacts is the calculation of long-range non-bonded forces.

In relatively large molecular systems (tens of thousands of particles) the combinatorial problem of calculating all pairwise interactions makes the force calculations required for MD simulations extremely time-consuming. The next section gives an overview of techniques proposed to alleviate this problem.

Enhanced efficiency methods

Overview

A clear gap exists between time scales that can currently be obtained by computer simulation techniques applied to biological macromolecules and the times required for most biological processes. With current state of the art methods and computers, a typical protein of 1000 amino-acids (100 kD) can be simulated for time-scales of at most nanoseconds[7], whereas most biological processes take place at times ranging between microseconds to seconds (or even minutes). Even if the rate of increase in computer power (an order of magnitude every 5-7 years[7]) continues, simulation of such processes at the required time scales will be beyond those of standard Molecular Dynamics simulation protocols in the next decade. Therefore, several groups have worked on developing techniques to overcome this problem. Conceptually, three categories of techniques can be distinguished [1]: (i) those that aim to mimic biological systems as realistically as possible and focus on sophisticated (mathematical) methods to enhance computational efficiency, affecting the dynamics as little as possible, (ii) those that simplify the molecular models involved, thus gaining computation time by neglecting details and (iii) those that make use of special properties of the simulated system to describe the system in more appropriate, internal coordinates. This division is not exclusive; some methods cannot be assigned to either category whereas others are hybrid methods based on principles from more than one category. A number of examples from each of the categories will be discussed in this section, and in the next section a technique from the third category, the so-called Essential Dynamics technique, will be described in detail since it will play a key role throughout the rest of this thesis.

Methods to speed up Molecular Dynamics with minimal perturbation

Since the first published application of MD to biomolecular systems[9], a little more than 20 years ago, people have devised methods to increase the time scales of Molecular Dynamics simulations. When Newton's equations of motion are integrated, the limiting factor that determines the time step that can be taken is the highest frequency that occurs in the system. In solvated biological macromolecules, the vibrations of bonds involving hydrogen atoms form the highest frequency vibrations. The bond stretching frequency of an O-H bond is typically about 10^{14} Hz, so the average period would be in the order of 10 fs ($10 \cdot 10^{-15}$ s)[10]. This limits the time-step to be taken in MD simulations to about 0.5 fs (a rule of thumb exists that states that for a reasonable sampling of a periodic function, samples should be taken at least twenty times per period). The introduction of a method to constrain these bonds (or, in fact, all covalent bonds) allowed to increase the time step to a

[1] Previously, a subdivision has been suggested according to levels of approximation[8]

typical value of 2 fs[11]. Since these bond vibrations are practically uncoupled from all other vibrations in the system, constraining them does not notably alter the rest of the dynamics of the system. This is not true, however, for bond-angle fluctuations, which form the second-highest frequency vibrations. Constraining bond-angles has a severe effect on many other fluctuations in the system, including even global, collective fluctuations, limiting the use of methods that use bond-angle constraints to only a few specific cases[10].

The notion that a number of discrete classes of frequencies of fluctuations in simulations of biomolecules can be distinguished, however, can be utilised to design more efficient algorithms. Forces that fluctuate rapidly need to be recalculated at a higher frequency than those that fluctuate on a much longer time scale. Although not trivial to implement, a number of successful applications of so-called multiple time-step algorithms have been reported in the literature (for a review, see ref. 10). Speedup factors of 4-5 have been claimed for such methods with respect to unconstrained dynamics, making them only slightly more efficient than simulations with covalent bond-length constraints.

As stated before, the most time-consuming part of Molecular Dynamics simulations is the force evaluation at every time-step. Especially the evaluation of electrostatic forces is notorious since Coulomb terms are inversely proportional to the inter-atomic distances of charged particles. This makes their contribution to the total force non-negligible even at fairly large distances (above 10 Å). Several methods have been proposed to reduce the computational cost to calculate long-range electrostatic forces. The most straightforward of these methods are cut-off methods where interactions beyond a certain radius are simply neglected[12]. This reduces the original order of complexity from N^2 to N (with N the number of particles) but significant artifacts have been reported at the edge of the cut-off radius[13–15]. Ewald methods form the traditional way to calculate electrostatic interactions in a more elegant fashion by calculating infinite lattice sums, but the order of complexity $N^{3/2}$ makes the method unsuitable for simulation of large biomolecular systems. However, approximations like particle-particle particle-mesh (PPPM)[16] and particle-mesh Ewald (PME)[17] that scale with $N{\cdot}\log(N)$, have shown encouraging results[18,19]. Fast Multipole methods (FMM) distribute atomic charges over a hierarchy of clusters and approximate electrostatic interactions by multipolar expansions of the potential generated by the clusters[20]. FMM methods even scale with N but require extra overhead compared with other methods, making them the method of choice only for systems of tens or hundreds of thousands of particles. Combinations of efficient ways to calculate electrostatic interactions with multiple time-step methods have already been described (e.g. FMM together with multiple time step algorithms[21,22]).

Another approach to reach equilibrium conformational properties at an enhanced rate is by performing so called 'mass tensor Molecular Dynamics'[23]. The masses of e.g. hydrogen atoms are increased to slow down the highest-frequency vibrations, allowing for a larger integration step. The dynamics is perturbed in this way, but equilibrium properties are not affected[24]. Another way to get around the problem of high frequency vibrations of hydrogen atoms is by excluding them from the actual integration and regenerating their positions every time step from the positions of the heavy atoms to which they are attached[25]. Although the features of this approach have yet to be explored, initial results have shown that a time-step of 6 to 8 fs is within reach. Another approach has recently been proposed, called "self-guided Molecular Dynamics"[26], that introduces an additional systematic force that is based on earlier parts of the simulation. Enhanced rates of conformational sampling have been claimed for small peptides. Its applicability in the field of protein dynamics still needs to be studied.

Simplified protein models

Before the first all-atom Molecular Dynamics simulation on a protein was performed, simulations of protein folding with a simplified protein model had been reported[27]. This illustrates the limitations of all-atom descriptions of proteins in computer simulations, especially in the presence of explicit solvent.

Simplified protein models have been utilised extensively in the field of protein folding. Employed methodologies include lattice Monte Carlo (MC) models and adapted MD or Langevin Dynamics (LD) models. The Monte Carlo technique is a stochastic method: random displacements are taken at each step, which are only accepted when an energy criterion is fulfilled[28]. Lattice models form perhaps the most simplified models with some resemblance to real proteins[29]. Their advantage is that exhaustive searches of the configuration space can be reached for small proteins (up to about 100 residues) by MC methods[30–33]. However, their applicability is limited due to the lack of detail in the models and the restriction of the search space due to lattice constraints. Continuum models of simplified proteins (bead models) utilising adapted MD or LD algorithms are more promising, in that sense, because of the absence of lattice restrictions. In Langevin Dynamics, compared to eq. 1.1, forces contain an additional friction and noise term to mimic the effect of solvent (which is not treated explicitly)[34]. Although exhaustive searches can usually not be reached by these bead methods, promising results have been reported[8, 35–37]. Another application of simplified protein models for use in protein folding are so called threading techniques[38] (for recent reviews, see refs. 39, 40). The idea is that a discrete number of folds exists to which proteins are restricted. The sequence of a protein with unknown structure is threaded through a set of known protein folds, after which suitable scoring potential (e.g. ref. 41, for a review see ref. 42) reveals which structure is most probable for that sequence.

Monte Carlo calculations using coarse grained protein models similar to those used for threading have shown that native state dynamics of proteins can successfully be simulated at a rate one order of magnitude faster than can be obtained by all-atom models[43, 44]. Also, LD simulations with a multiple time step algorithm showed vast improvements of computational efficiency compared to traditional MD, even when using an all-atom representation[45]. Apart from the advantage of a multiple time-step algorithm, part of the computational efficiency in this model is the result of the absence of explicit solvent molecules. Several methods of solvent treatment by implicit models have been suggested over the years[46–50], but their range of applicability is still a matter of debate[51–54]

Simplification in its most extreme form reduces a protein's conformational space to that of two or more rigid bodies. Domain motions are known to form the basis of the function of several proteins (see e.g. ref. 2) and therefore many properties of the functional mechanism of such proteins may be studied by focusing on the rigid-body motions of the domains involved[55, 56]. Even in single-domain proteins, quasi-rigid parts have been identified (for example secondary structure elements[57–59]). This observation could in principle be used in a simplified protein model, but has so far only been applied in the field of theoretical protein folding[60, 61].

Protein dynamics in internal coordinates

Efficiency of computer simulations can be enhanced by describing the simulated systems in their internal degrees of freedom, as opposed to the usual Cartesian coordinates. The goal, as in the previous section, is to reduce the number of degrees of freedom in the simulated system. The methods described in this section, however, retain the atomic detail of the modeled system. Perhaps the first example of this method was proposed by Ryckaert & Bellemans[62] in their simulation of n-butane, with only one internal degree of freedom (the central torsion angle). For proteins, the use of torsion angles also seems an appropriate choice since dihedral angles are the main degrees of freedom, of which the ϕ and ψ backbone dihedrals play the largest role in large-scale protein motions. Application of torsion angles in the study of protein dynamics has been proposed for MC[63] and MD[64] simulations. The advantage of such techniques is that larger simulation steps (either time-steps in MD or space-steps in MC) can be taken in the simulation. Stable MD simulations with time steps of 13 fs have been described for an Ala_9 peptide[64], whereas time-steps of at most 2 fs can be taken when only bond lengths are constrained. However, a number of problems is encountered when protein dynamics is described in torsion angle space. First, when the equations of motion are solved for these internal coordinates, the inverse of the moments of inertia tensor is required every time step. Since matrix inversion scales with the third power of the number of matrix elements in terms of computation time, application of such methods is limited to small systems. However, a method to get around this problem has been proposed[65], reducing the computational cost to order N instead of N^3. The second problem connected with torsion-angle dynamics is the absence of bond-angle fluctuations. Bond-length fluctuations can safely be neglected, but constraining bond-angles severely restricts dynamics of proteins (see e.g. ref. 10). Due to the altered potential employed in torsion-angle approaches, conformational barriers are overestimated, making the method most useful for simulations at elevated temperatures, used for example in the field of refinement of NMR structures[66].

Torsion angle approaches have also been applied in combination with knowledge-based force-fields. Monte Carlo simulations have been reported claiming enhanced convergence for NMR structure determination[67]. Also in off-lattice simulations MC calculations in torsion-angle space have begun to gain popularity (for a review, see e.g. ref. 68).

Another way to define internal coordinates in proteins is based on the notion that most positional fluctuation occurs along collective degrees of freedom. This was first realised from Normal Mode analyses of a small protein[69–71]. In Normal Mode analyses, the potential energy surface is assumed to be harmonic. Collective variables are obtained by diagonalisation of the Hessian matrix (second derivative of the potential energy) in a local energy minimum. Quasi harmonic analysis[72–75], principal component analysis[76–78] and singular value decomposition[44, 79] of Molecular Dynamics trajectories of proteins have shown that even beyond the harmonic approximation, protein dynamics is dominated by a limited number of collective coordinates. These methods seek those collective degrees of freedom that best approximate the total amount of fluctuation. The subset of largest-amplitude variables form a set of generalised internal coordinates that can be used to effectively describe the dynamics of a protein. As opposed to torsion angles as internal coordinates, these collective internal coordinates are not known beforehand. Unless many experimental structures are available, a simulation is required to obtain a definition of these coordinates. Once an approximation of the collective degrees of freedom has been obtained, simulations in the space spanned by only these coordinates can in principle be initiated. Such a technique has successfully been applied to small molecules[80]. However, coupling of the main modes of collective fluctuation to more constrained coordinates is likely to be responsible for a limited applicability in dynamic simulation of proteins (e.g. ref. 10 and A. Amadei and T. Linssen, personal communication). Methods to bypass the problems of this coupling include biased MD simulations with constraints along collective internal coordinates derived from earlier simulations[81] and form the subject of chapters 3 and 4 of this thesis. The dynamics can also be biased by modifying the potential energy function along such a collective degree of freedom. This is thought to be especially useful for enhancing the rate of conformational transitions in proteins[82].

Essential Dynamics

The Essential Dynamics (ED) technique is a method from the third category of the last section. A brief description will be given here, discussing some important features of the method. For a more rigorous description, see ref. 78. As an analysis technique, ED is based on a principal component analysis of (MD generated) structures. A principal component analysis is a multidimensional linear least squares fit procedure. To understand how this is applicable to protein dynamics, the usual three-dimensional (3D) Cartesian space to represent protein coordinates (which is e.g. used to represent protein conformations in the Brookhaven Protein Data Bank or PDB) needs to be replaced by another, multidimensional space. A molecule of N particles can be represented by N points in 3D space. With 3 coordinates per point, this adds up to 3N coordinates. In a 3N-dimensional space, however, such a structure can be represented by a single point. In this space, this point is characterised by 3N coordinates. This representation is convenient since a collection or trajectory of structures can now be regarded as a cloud of points. Like in the case of a two-dimensional cloud of points, also in more dimensions, always one line exists that best fits all points. As illustrated for a two-dimensional example (Fig. 1.1), if such a line fits the data well, the data can be approximated by only the position along that line, neglecting the position in the other direction. If this line is chosen as coordinate axis, then the position of a point can be represented by a single coordinate. In more dimensions the procedure works similarly, with the only difference that one is not just interested in the line that fits the data best, but also in the line that fits the data second-best, third best, and so on (the principal components). These directions together span a plane, or space, and the subspace responsible for the majority of the fluctuations has been referred to as the 'essential subspace'. Applications of such a multidimensional fit procedure on protein configurations from MD simulations of several proteins has proven that typically the ten to twenty principal components are responsible for 90 % of the fluctuations of a protein[76–78]. These principal components correspond to collective coordinates, containing contributions from every atom of the (protein) molecule. Summarised, a limited number of collective motions is responsible for a large percentage of a protein's conformational fluctuations.

If all atoms in a protein were able to move uncorrelated from each other, an approximation of the total fluctuation by only a few collective coordinates would not be possible. The fact that such an approximation is successful is the result of the presence of a large number of internal constraints and restrictions ('near-constraints') defined by the interactions present in a given protein structure. Atomic interactions, ranging from covalent bonds (the tightest interactions) to weak non-bonded interactions, together with the dense packing of atoms in native-state protein structures form the basis of these restrictions.

In the study of protein dynamics, only internal fluctuations are usually of interest. Therefore, the first step in an Essential Dynamics analysis is to remove overall rotation and translation. This is done by translation of the center of mass of every configuration to the origin after which a least squares rotational fit of the atoms is performed onto to a reference structure. Recently it was suggested that this procedure might lead to a bias in the definition of the internal fluctuations, and that a way to circumvent this bias would be to work in distance space[83]. The actual principal component analysis is based on construction and diagonalisation of the covariance matrix of positional fluctuations. The covariance matrix is constructed from the atomic coordinates according to:

$$\boldsymbol{C}_{ij} = \langle(\boldsymbol{x}_i - \langle\boldsymbol{x}_i\rangle)(\boldsymbol{x}_j - \langle\boldsymbol{x}_j\rangle)\rangle \tag{1.4}$$

where $\boldsymbol{x}$ represents the atomic coordinates and the angle brackets a time or ensemble average. Particles moving in a correlated fashion correspond to positive matrix elements (positive correlation) or negative elements (negative correlation), and those that move independently to small matrix elements. The orthogonal transformation $\boldsymbol{T}$ that diagonalises this (symmetric) matrix contains the eigenvectors or principal components of $\boldsymbol{C}$ as columns and the resulting diagonal matrix $\boldsymbol{\Lambda}$ contains the corresponding eigenvalues:

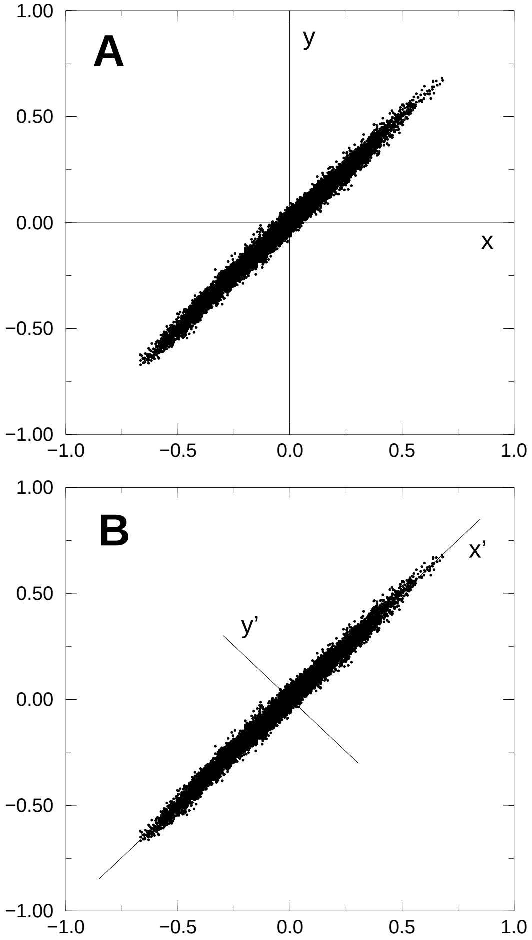

Figure 1.1 Illustration of Essential Dynamics in two dimensions. With a distribution of points as depicted here, two coordinates (x,y) are required to identify a point in the cluster in panel A, whereas one coordinate (x') approximately identifies a point in panel B.

$$\boldsymbol{\Lambda} = \boldsymbol{T}^{\top}\boldsymbol{C}\boldsymbol{T} \tag{1.5}$$

The eigenvalues are a measure of the mean square positional fluctuation along the corresponding eigenvector. When the eigenvectors are sorted to decreasing eigenvalue, the first eigenvectors are those collective motions that best approximate the sum of fluctuations and the last eigenvectors correspond to the most constrained degrees of freedom. The characteristics of these collective fluctuations can be studied by projecting the ensemble of structures onto single eigenvectors and by translation of these projections to 3D space to visualise the atomic displacements connected with that eigenvector. As stated above, analyses of MD trajectories of several proteins have shown that few collective coordinates dominate the dynamics of native proteins (together often referred to as the 'essential subspace'). In a number of cases these main modes of collective fluctuation were shown to be involved in the functional dynamics of the studied proteins[78, 84–86].

ED analyses can be applied to any subset of atoms of the ensemble of structures[78] and are not restricted to ensembles generated by MD simulation. Applications to collections of X-ray structures[86, 87], NMR structures[88] and structures derived from distance constraints[89] have been reported. Since collective (backbone) fluctuations dominate the dynamics of proteins, usually only backbone or C-α coordinates are used to save computation time and to prevent problems with apparent correlation of side chain motions with backbone motions which are merely the result of poor statistics. However, even when the method is applied to only C-α atoms, the diagonalisation of the covariance matrix can still be an enormous computational task. An approximation has been developed to alleviate this problem, allowing analyses of systems with thousands of amino-acids[90].

Although first designed for proteins, the ED method can in principle be applied to any constrained (biomolecular) system. Successful applications to DNA have already been reported[91, 92].

Identification of the dominant modes of collective fluctuation is the first step in the Essential Dynamics technique. As sketched in the previous section, knowledge of the essential subspace can be used in a sampling technique that exploits the limited dimensionality of that space to achieve a more efficient sampling than can be obtained by more conventional techniques.

Outline of this thesis

The second chapter of this thesis is concerned with the convergence of Essential Dynamics results from relatively short MD simulations. In the literature, it had been reported that principal component analysis (Essential Dynamics is a principal components analysis of the atomic fluctuations) of MD simulations of such short time lengths is not suitable for describing long-time scale protein dynamics because this subspace keeps changing throughout the simulations. Apart from the issue of convergence of the essential subspace, the sensitivity of the essential dynamics results to MD parameters is also examined in this chapter. A set of reference simulations is compared to a set in which parameters were modified that were believed to have a potential effect on (protein) dynamical or configurational properties.

The third chapter presents an extension of the Essential Dynamics sampling technique. This ED sampling technique is based on the idea that, since most fluctuations in proteins take place in a hyperspace of limited dimension, a systematic or otherwise enhanced sampling in this subspace will result in an efficient way to explore the configurational space of proteins. A prerequisite for success of this method, of course, is a sufficiently accurate approximation of the subspace. In a first implementation, the method had yielded encouraging results on a small protein which showed that indeed acceptable protein structures were generated which were more widely spread in configuration space than would be obtained by usual MD simulation[81]. This chapter presents the application of a modified sampling algorithm to a peptide hormone. An extensive sampling is performed and the stability of resulting structures is measured by subjecting them to MD simulation without essential dynamics constraints. Based on these results, a model is presented for the free energy surface of this peptide and proteins in general in the space of the major collective conformational coordinates.

Encouraged by the results on the peptide, the ED sampling procedure was applied to a small protein: the Histidine containing Phosphocarrier protein HPr. It was found that some modifications to the algorithm were required because denaturation of the protein was observed when the same criteria were used as with the peptide. In chapter 4, the resulting ensemble of structures is compared to a set of structures collected from unconstrained MD simulations and from simulations with NMR-NOE restraints. Structures extracted from the latter simulations represent the high-resolution NMR structure of HPr[93]. NOE violations from each of the three runs are compared to each other, as well as several geometrical and energetical properties.

Chapter 5 presents a comparison of domain motions in T4 lysozyme calculated from several crystal structures and those obtained from MD simulation. T4 lysozyme is among the best experimentally characterised proteins in terms of conformational properties and therefore is an ideal candidate for a rigorous test how well MD/ED results from simulations in the order of nanoseconds correspond to known, large-scale collective fluctuations in proteins. A newly developed method to characterise domain motions in proteins was employed to compare the experimental and theoretical results in detail. Not only methodological implications, but also functional aspects of the domain fluctuations are described.

The observation that most of a protein's positional fluctuation can be approximated by only a few collective degrees of freedom led to an attempt to derive those degrees of freedom by another, computationally less demanding method. The restriction of a protein's fluctuations to a hyperspace of limited dimension is caused by the presence of a large number of explicit and implicit constraints and restrictions to the configurational freedom of each atom. The idea arose that if the network of interactions responsible for these restrictions could be represented in a simpler way than in e.g. MD simulations, an approximation of the constraint surface, and therefore also of the complementary essential subspace could be obtained. Chapter 6 introduces a technique, named CONCOORD, that generates protein structures within predefined distance bounds. CONCOORD structures of different proteins are compared to structures generated by MD, in terms of Essential Dynamics properties and more conventional techniques.

Chapter 7 presents an application of the CONCOORD method to the molecular chaperonin GroEL. The elucidation of the X-ray structures of GroEL in different conformations together with electron microscopy data had shown that GroEL is a remarkably flexible protein and that allosteric properties play an important role in its function: to assist other proteins to fold to their native conformation. The size of the protein (M$\approx$800kD) makes it unsuitable for other computational techniques that yield protein conformational properties, such as MD, but because of its algorithmic simplicity and efficiency, it proved possible to apply CONCOORD. Essential dynamics analyses were applied to the collection of experimental structures and conformations generated by CONCOORD. Previously unnoticed features of the crystallographic structures are presented, in combination with conformational properties derived from the CONCOORD simulations. Implications for the allosteric mechanism of GroEL are described.

Finally, chapter 8 finishes this thesis with some concluding remarks on theoretical approaches in the field of protein dynamics and an outlook to the future.

2 THE CONSISTENCY OF LARGE CONCERTED MOTIONS IN PROTEINS IN MOLECULAR DYNAMICS SIMULATIONS

B.L. de Groot, D.M.F van Aalten, A. Amadei and H.J.C. Berendsen

Biophys. J. 71: 1554-1566 (1996)

Summary

A detailed investigation is presented into the effect of limited sampling time and small changes in the force field on molecular dynamics simulations of a protein. Thirteen independent simulations of the B1 IgG-binding domain of streptococcal protein G were performed with small changes in the simulation parameters in each simulation. Parameters studied included temperature, bond constraints, cut-off radius for electrostatic interactions and initial placement of hydrogen atoms. The essential dynamics technique was used to reveal dynamic differences between the simulations. Similar essential dynamics properties were found for all simulations, indicating that the large concerted motions found in the simulations are not particularly sensitive to small changes in the forcefield.

A thorough investigation into the stability of the essential dynamics properties as derived from a molecular dynamics simulation of a few hundred picoseconds is provided. Although the definition of the essential modes of motion has not fully converged in these short simulations, the subspace in which these modes are confined is found to be reproducible.

Introduction

Recent studies have provided methods for revealing large concerted motions in proteins from Molecular Dynamics (MD) computer simulations[53,76,78,85,94]. These methods divide the configurational subspace of proteins in a high dimensional subspace in which merely constraint-like motions of high frequency occur (which will from now on be referred to as the near-constraints subspace), and a low dimensional subspace in which all biologically relevant motions occur (the essential subspace). In this paper, we investigate how reproducible these two distinct spaces are in multiple simulations of one protein. The essential dynamics (ED) method, introduced by Amadei *et al.* [78], is used to extract the definition of both subspaces from MD simulations. The sensitivity of the definition of these spaces towards different force field parameters used in MD simulations, as well as the speed of convergence of the description of these subspaces is examined. With this aim, four simulations of a test protein, the B1 IgG-binding domain of streptococcal protein G, were set up, each with one parameter different from seven reference simulations. Apart from these simulations that were performed using explicit solvent, two simulations were run in vacuo. This protein was chosen because it is a small and fairly globular protein, containing both α-helix and β- strand secondary structure elements. Both X-ray[95] and NMR[96] structures are available, as well as NMR relaxation data[97].

Previous work[53,78,84] has suggested that a few hundred picoseconds is usually enough to obtain a rough approximation of the essential subspace of a small protein, although there is still an appreciable amount of noise present in the description of both subspaces after such limited sampling time. Here we use a set of 300 ps simulations as well as a 1 ns simulation to investigate the accuracy of the definition of both subspaces. The influence of a number of simulation parameters is also investigated. All simulations are compared to a set of six solvent simulations of 300 ps. These reference simulations were also compared to each other and to a 1 ns simulation of the same protein to gain insight in the convergence of the ED properties in such short simulations. Apart from comparison of properties derived from ED, conventional structural and geometrical properties were evaluated from the trajectories, to examine the stability and overall structural and dynamic behaviour of the simulations.

Methods and theory

Simulation parameters

Simulations were performed with the GROMOS[4] simulation package. The simulations were started from the crystal structure (Protein Data Bank entry 1PGB[95]). The protein was placed in a truncated octahedral box filled with SPC water[98], except for two simulations that were run in vacuo. The protein consists of 535 (united) atoms. Together with 4 sodium ions which were used to compensate for the net charge of -4 (the ions were placed in the box by replacing water molecules at the lowest electrostatic potential) and approximately 1900 water molecules (the number of water molecules varied from simulation to simulation) the total number of atoms approximated 6500. After energy minimisation, a HEATUP procedure[53] of 25 ps was performed to equilibrate the structure. In short, this involves a slow increase of the temperature, cut-off radius and time step, combined with positional restraining. The simulations were then continued for 275 ps, of which the last 250 ps were used for ED analyses (all other analyses were performed on the full 275 ps trajectories, to include differences in the equilibration period). One simulation was extended to 1 ns. In total, thirteen simulations have been performed, identified below.

1. 275K: This simulation was performed at a constant temperature of 275 K instead of 300 K. All simulations were kept at a constant temperature by coupling to an external temperature bath[99], using a coupling constant of 0.1 ps.;

2. NO_SHAKE: This simulation was performed without SHAKE[11], covalent bond interactions were described by harmonic potentials. In this simulation, a time step of 1 fs was applied. In other simulations SHAKE was used to constrain bond lengths, allowing a time step of 2 fs.;

3. CUT_OFF: This simulation was performed with a twin range cut-off method with radii of 10 and 14 Å instead of 8 and 10 Å for the other simulations. For the short range, the pairlist was updated every time step, for the long range, this list was updated every ten steps.;

4. HPLACE: This simulation was started from a structure in which the positions of the hydrogens were generated using an algorithm which optimises hydrogen bond networks throughout the structure[100]. Other simulations were started with standard GROMOS hydrogen placement, which uses standard hydrogen positioning.;

5. REF_1 till REF_6: 6 reference simulations were performed. These simulations differed in the initial velocities used;

6. REF_7: Identical to the other reference simulations, but this simulation was extended to 1 ns;

7. VAC_1 and VAC_2: Additionally, two simulations in vacuo were performed for comparison. VAC_1 was carried out with reduced charges to mimick the screening effect of the solvent, VAC_2 was performed with full charges.

Comparison techniques

Two types of techniques were used to identify differences between the simulations. First, a number of standard structural analyses were performed to check overall stability. Subsequently, ED analysis was used to compare the dynamic behaviour of the protein in the different simulations. ED analyses were performed on each individual trajectory and compared to the reference simulations. Programs used were those available in the molecular modeling program WHAT IF[101]. Accessible surface calculations and secondary structure evaluations were performed by DSSP[102].

Overlap between eigenvector sets

Overlap between multiple sets of eigenvectors is calculated with two methods. First, the overlap between two essential subspaces is calculated as the sum of all the squared inner products between all pairs of eigenvectors from both essential subspaces, divided by the dimension of that space (see also the next subsection). This definition of the overlap has the disadvantage that it concentrates on similarities between two compared sets, and not on differences. Therefore, we have defined another measure of the overlap between two sets of eigenvectors. It is defined as the product of the square inner product and the difference in eigenvector index (i.e. a difference in relative contribution to the overall fluctuation, eigenvalues and corresponding eigenvectors are sorted to decreasing value), averaged over all pairs of eigenvectors from both sets. This will result in a positive number, being close to zero if the sets are similar. This quantity can be regarded as a penalty function: a high inner product between an eigenvector with a high eigenvector index from one set and an eigenvector with a low eigenvector index from the other set gives a high contribution to this penalty. Significant differences in the dynamic behaviour of two simulations (a motion that is accessible in one simulation but not in another) are therefore immediately evident.

Convergence of trajectories

ED analyses can be used to gain insight in the convergence of MD trajectories, since only a few coordinates are usually required to describe the relevant dynamics of a protein in MD simulations. Here, the overlap between essential subspaces (as defined by the ten eigenvectors with largest eigenvalues) obtained from pairs of simulations is taken as a measure for the similarity of two trajectories in terms of collective motions. This overlap is defined as the cumulative mean square inner product between ten eigenvectors obtained from one set of eigenvectors and ten from another. This results in a number between zero, when there is no overlap, and one, when the two sets are identical. In practice, the lower limit for this value is not zero, even if the actual overlap between the two essential subspaces is negligible, since there will always be some projection of the eigenvectors of one set into the essential subspace of the other.

The overlap of real interest is not the overlap between two sets of eigenvectors obtained from MD, each containing noise, but the overlap of one of such sets with the fully converged set of eigenvectors, as would be obtained after infinite simulation time. This overlap is underestimated by the method described above, since in the comparison of two MD eigenvector sets, both sets contain an appreciable amount of noise, making the overlap smaller than in the case where only one of the sets contains this noise.

Results

Results from structural analyses are summarised in Table 2.1. Apart from the two simulations in vacuo, the observed average properties in the simulations that were performed with different parameters do not differ significantly from those observed in the 300 ps reference simulations. When these properties are plotted as a function of time (data not shown), no significant drift is observed in any of the simulations (apart from those performed in vacuum). Hence, all solvent simulations are stable in terms of these properties in a time window of 300 ps. The simulation performed at lower temperature (275 K) does show a lower total mean square fluctuation than most other simulations, as expected, but one of the reference simulations (REF_1) shows an even lower total fluctuation. This suggests that the spread in the observed fluctuation for the reference simulations of 300 ps covers the difference that might have been caused by the lower temperature.

Compared to other proteins[53, 78, 85, 103], the total sum of fluctuations and the largest eigenvalues are relatively small, indicating a rather rigid molecule. This is in agreement with recent observations[96, 104], where this domain was reported as highly stable.

ED analyses were performed on each individual trajectory. Only α carbon coordinates were used in the covariance analysis. It has been shown that this approach identifies all large scale concerted motions in proteins[78]. It has the advantage over an all-atom analysis (besides saving CPU time in all analyses) that backbone dynamics equilibrates faster than the dynamics in the full coordinate space (apparent correlations between backbone and sidechain motions introduce noise in an ED analysis on all atoms of a simulation of a few hundred picoseconds).

As an illustration of the typical overlap between two eigenvector sets obtained from 300 ps simulations in solvent, an inner products matrix is shown in Fig. 2.1A for two eigenvector sets obtained from REF_1 and REF_2. All high inner products are found close to the diagonal, meaning that directions in configurational space have a similar amount of freedom in both simulations. The same qualitative picture is found for all combinations of the solvent simulations. Fig. 2.1A shows that the eigenvectors spanning the essential subspace (e.g. defined as the first ten eigenvectors) of one set of eigenvectors show the

type	σ^2	RMSD	NRC	HBO	ACC	GYR
275K	0.244	1.19	8.04	44.6	3773	1.018
NO_SHAKE	0.337	1.02	8.95	42.9	3879	1.025
CUT_OFF	0.406	1.15	9.63	44.9	3856	1.015
HPLACE	0.366	1.23	10.61	43.3	3884	1.025
REF_1	0.242	0.93	10.80	41.1	3761	1.016
REF_2	0.351	1.12	8.42	45.8	3840	1.016
REF_3	0.366	1.27	8.40	45.6	3883	1.021
REF_4	0.394	1.11	9.10	43.9	3777	1.018
REF_5	0.365	1.28	8.29	46.0	3849	1.023
REF_6	0.331	1.24	10.70	42.4	3904	1.025
REF_7	0.532	1.45	10.03	44.0	3850	1.024
VAC_1	0.399	1.79	10.47	50.5	3607	0.997
VAC_2	1.261	3.83	18.22	35.3	3600	1.031

Table 2.1 Structural properties. σ^2: total mean square fluctuations (nm^2); RMSD: root mean square deviation from crystal structure (Å); NRC: number of residues adopting random coil conformation; HBO : number of main chain hydrogen bonds; ACC: total solvent accessible surface (Å^2); GYR: radius of gyration (nm). NRC, HBO and ACC were calculated with DSSP[102]

largest inner products with the essential eigenvectors extracted from another simulation, and that the projections outside the essential subspace are mainly concentrated in those eigenvectors which still have a significantly high eigenvalue. The simulations in vacuo show also high inner products further from the diagonal (Fig. 2.1B), indicating that the simulations in vacuo are more different from the solvent simulations than the solvent simulations are from each other.

For all reference simulations, the noise as discussed in the theory section, which causes overlap between eigenvectors from the essential subspace obtained from one simulation and near-constraint eigenvectors from another set, is not homogeneously spread over all near-constraints eigenvectors (Fig. 2.2A). Instead, it is concentrated in the near-constraints which still have an appreciable eigenvalue (eigenvectors 11-50), leaving a negligible overlap with all other eigenvectors. This indicates that the definitions of the essential subspaces from all reference simulations are similar. The overlap of all 300 ps reference simulations (REF_1 through REF_6) with the reference simulation of 1 ns (REF_7) is not significantly higher than the overlap between the 300 ps simulations mutually, although a more converged (i.e. containing less statistical noise) description of the essential subspace was to be expected from this longer simulation. This indicates that the convergence of the definition of the essential subspace is initially fast and does not increase significantly in the time window from 300 ps to 1 ns. This validates the use of relatively short simulations to gain insight in the dynamic properties of such systems.

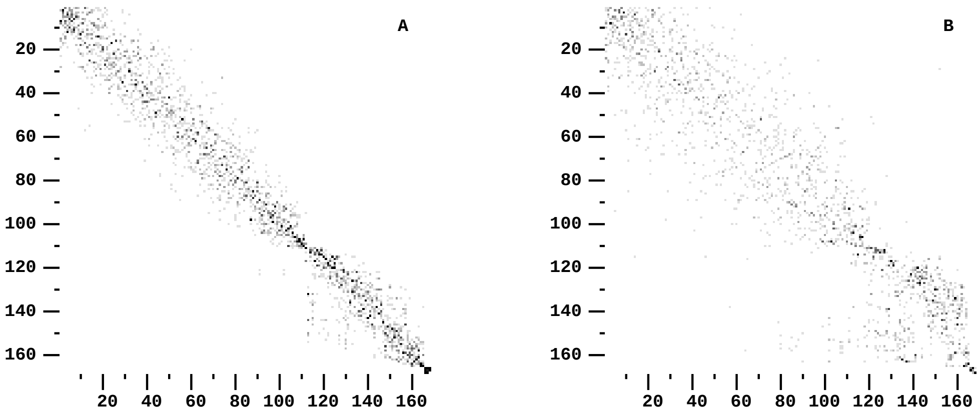

Figure 2.1 Squared inner products matrices. Panel A: Inner products between eigenvectors extracted from REF_1 (y-axis) and REF_2 (x-axis). Panel B: Inner products between REF_1 (y-axis) and VAC_2 (x-axis).

Fig. 2.2B shows the average cumulative square inner products of all eigenvectors of all sets with the first ten eigenvectors from the reference simulations. All curves add up to one because all eigenvectors of one set are always able to rebuild a (subset of) eigenvector(s) of another set (both sets of vectors span the same space). The curves are steep, indicating a high degree of overlap of the first ten eigenvectors of each of the reference simulations in the essential subspaces obtained from the other simulations. The simulations that were run with parameters different from the reference simulations all show an equal amount of overlap with the reference simulations as do the reference simulations among each other. For the simulations that were run in vacuo, the measured overlap is significantly smaller, especially VAC_2. The summed average square inner products (here taken as the overlap of the ten eigenvectors with highest eigenvalues from one set compared to the reference sets) and values for the penalty function as described in the Methods section are summarised in table 2.2. All values are averages over the comparison with the reference simulations. For the reference simulations, values were obtained by comparison of one reference simulation compared to all others, and subsequent averaging over all. The values in the second column of this table are obtained by summation of the square inner products between the first ten eigenvectors of one set with all first ten eigenvectors of another set, and subsequent division by ten. A value of one is obtained when two sets are identical. It should be noted that the first ten eigenvectors span only $10/(56 * 3) = 5.95\%$ of the total space. The fact that the essential subspaces from the different simulations overlap for approximately 50% means that similarities between the essential subspaces of the individual trajectories are significant. The amounts of overlap of the essential subspaces of any combination of (except for the second simulation in vacuo) simulations are similar. The penalty function (Table 2.2), which is more sensitive towards differences between eigenvector sets than the measure of overlap in terms of a cumulative inner product (see theory section), also gives similar values for all solvent simulations. The two simulations in vacuo, however, give significantly higher values for this penalty function, demonstrating dynamic differences caused by the presence of solvent. Based on these data, there are no detectable systematic differences

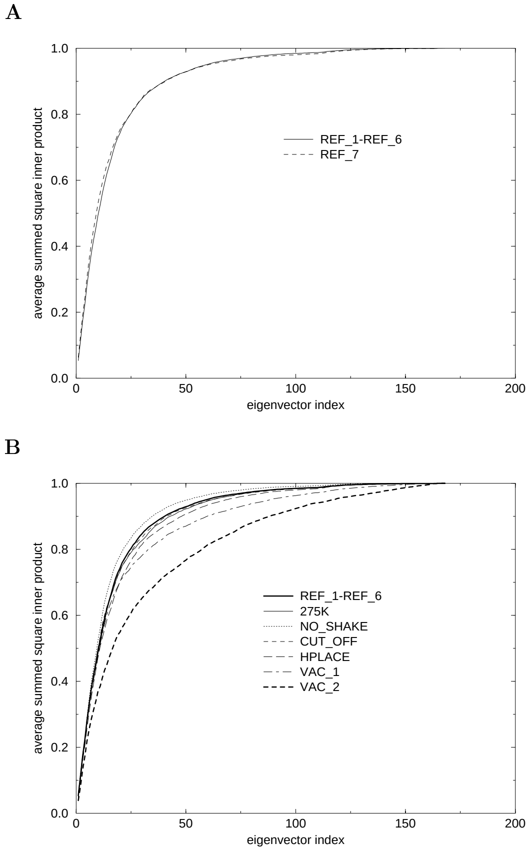

Figure 2.2 Average cumulative square inner products. Panel A: The solid line represents the average summed square inner product of the first ten eigenvectors of one 300 ps reference simulation with all eigenvectors from another 300 ps reference simulation, averaged over all pairs. The dashed line represents the results obtained from the 1 ns reference simulation, compared with the 300 ps simulations. Panel B: For all except the reference simulations, curves were obtained by calculation of the summed squared inner product between all eigenvectors from a single simulation with the first ten eigenvectors of each reference simulation, divided by ten, averaged over all reference simulations. For the reference simulations, the average cumulative squared inner products of the first ten eigenvectors of each set with all eigenvectors from all other reference sets were calculated. The average over all pairs is plotted.

type	MSI		PENAL	
	mean	σ	mean	σ
275K	0.483	0.040	1.172	0.087
NO_SHAKE	0.524	0.037	1.051	0.084
CUT_OFF	0.498	0.017	1.158	0.070
HPLACE	0.472	0.024	1.203	0.103
REFS_1_6	0.494	0.040	1.155	0.087
REF_7	0.530	0.065	1.169	0.132
VAC_1	0.478	0.033	1.517	0.068
VAC_2	0.372	0.038	1.967	0.039

Table 2.2 MSI: Average and root mean square fluctuation of summed square inner product between the first ten C-α eigenvectors from ED analyses of each individual trajectory with each of the first ten eigenvectors from all reference simulations. PENAL: Average and root mean square fluctuation of penalty function (see methods section) between eigenvectors from all simulations and reference simulations.

between the various methods of simulation, apart from the second simulation in vacuo.

Fig. 2.3 shows the kinds of motion that correspond to the most prominent eigenvectors extracted from the reference simulations. For each of the first six eigenvectors, the motion is concentrated in a few specific places that move concertedly.

Conclusions and discussion

The results presented in this paper show, both considering overall structural and dynamic properties, that all solvent simulations that were studied behave essentially similar. Only the simulations that were performed in vacuo showed significantly different behaviour from the reference simulations, although even there the overlap of the essential subspace is still substantial. This is in agreement with previous findings[53, 84]. Of course, all analyses were concentrated on one protein; other proteins may behave differently.

Of the overall quantities, largest differences were found in the total mean square fluctuation and the RMSD from the crystal structure (Table 2.1). As has been noted before[105], the RMSD from a single structure (e.g. the crystal structure) is not necessarily a useful quantity to judge the stability of a simulation when large structural rearrangements are occurring, provided that these rearrangements are reversible. Moreover, backbone RMSD values of the structures in the NMR cluster with respect to the crystal structure are in the same range as the observed values for MD.

From the 300 ps solvent simulations, 275K and CUT_OFF deviate most

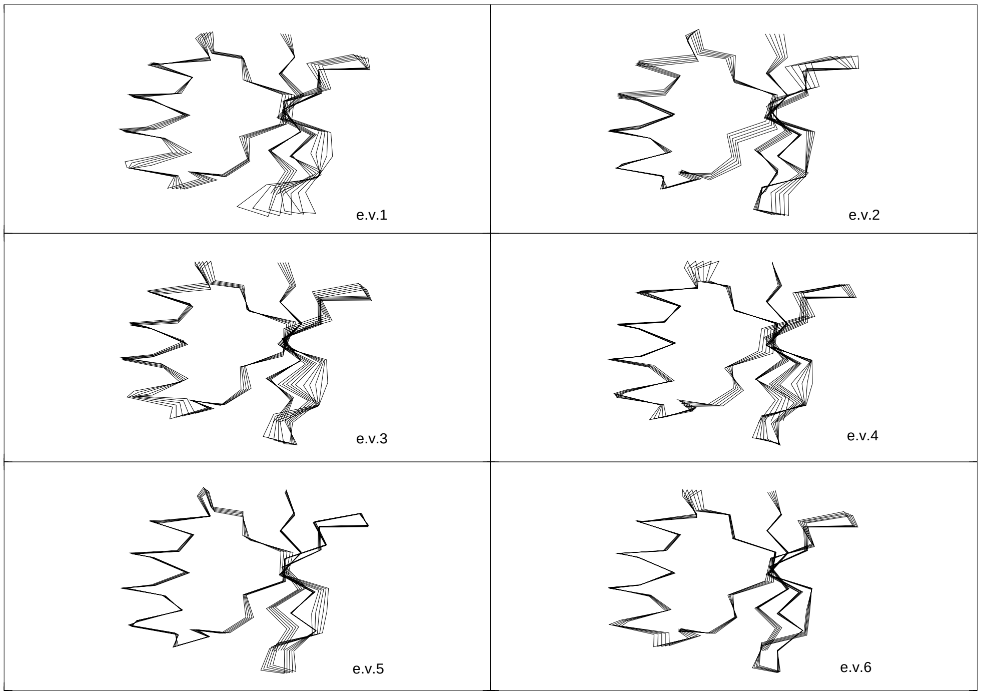

Figure 2.3 Snapshots of single-eigenvector motions. Structures corresponding to the minimum and maximum sampled position along the first six eigenvectors of an ED analysis of solvent trajectories combined are shown, together with three intermediate positions, equally spaced between the minimum and maximum.

from all other simulations. The lower total mean square fluctuation of all protein atoms in the simulation at lower temperature compared to most other simulations can partially be explained by the fact that less thermal motion is present in this simulation. Fast thermal fluctuations can be expected to have a connection to slower, larger fluctuations, which might be reflected in the fact that in the ED analysis, the mean square fluctuations along all essential eigenvectors for this simulation are among the lowest. The CUT_OFF simulation shows highest overall fluctuation (Table 2.1). No obvious explanation can be provided for this observation. Since also REF_1 is quite different with respect to the other reference simulations (it exhibits the lowest total mean square fluctuation of all simulations, Table 2.1) these differences are believed to be based on statistical rather than systematic reasons. In a recent paper[106], where the effects of different protein models on normal modes results were studied, the only significant sensitivity was reported on the description of electrostatic interactions.

For all geometrical properties, the differences between different reference simulations are as high as the differences between the reference simulations and the solvent simulations that were run with adapted parameters. On the basis of these data, therefore, the simulations with different parameters do not differ in a significant way from the reference simulations.

The differences in dynamic behaviour of the ten solvent simulations as revealed by ED do also not appear to be significant. Since there is no systematic connection between the observed dynamic differences and the different simulation parameters (the reference simulations differ as much from each other as from the other solvent simulations), we have the impression that limited sampling time is the most important reason for the presence of these differences. This leads us to the conclusion that the dynamic properties of the simulated protein are not detectably sensitive towards small differences in the force field or in the choice of starting structure in the time span considered here. Only the simulations that were performed in vacuo are significantly different from the reference simulations in solvent. To support these findings, further simulations of other proteins are necessary to be able to draw general conclusions.

As already observed recently[78, 81, 85], single simulations of a few hundred picoseconds of a protein in water seem to yield an acceptable approximation of the essential subspace, although for a fully converged description of this subspace, longer simulations are required. This is supported by the observation that similarities between ED analyses on individual trajectories that were studied here are relatively high (Fig 2.1, table 2.2), considering the fact that in the comparison as presented in table 2.2, always two sets are compared that each contain noise. As explained in the theory section, the overlap (defined as the summed square inner products between vectors from two essential subspaces) between eigenvectors obtained from each of the reference simulations with the fully converged set of eigenvectors can be expected to be larger than the overlap between eigenvectors obtained from two short simulations. The measured overlap of approximately 0.5 between eigenvector sets obtained from multiple simulations (table 2.2) therefore means that for each set, the overlap with the fully converged set of eigenvectors is even higher. This means that in a relatively short simulation, a good approximation of the true essential subspace is reached, within the limitations of the forcefield. The fact that within the essential subspace the individual eigenvectors are not identical in all simulations (although the subspace itself has approximately converged), indicates that in this subspace, the region that has been visited during a single short simulation is only a small fraction of the complete available subspace. This is in agreement with previous findings[81, 107, 108].

Further studies have shown that convergence increases only slowly with simulation time (Fig. 2.2A, table2.2), making predictions about the minimum time required to obtain a fully converged description of the dynamics impossible.

In a recent study[109], two halves of a short MD simulation of myoglobin were compared. It was concluded that a few hundred picoseconds is not sufficient to obtain equilibrated dynamics. In another study[110], two halves of a simulation of 470 ps of G-actin (375 residues) showed significant differences in terms of principal components analysis, analogous to essential dynamics analysis. We have shown here, and before[78, 81, 85] that within a few hundred picoseconds, the definition of both the essential and the near-constraints subspaces are approximately stable, while motions within the essential subspace are still equilibrating. In the present study (as also noted before by us[85] and others[111], the overlap found between the essential subspaces as derived from short simulations is substantial.

The initial description of the essential subspace as derived from a relatively short MD simulation can be used to obtain a more refined definition of this space in an extrapolation method[81]. In such a method, an adapted form of MD is performed, with constraints in the approximated essential subspace. These constraints are chosen such that the system itself determines the regions of the space that it samples. Dynamic coupling between the accessible modes of motion will automatically result in motion in the true essential subspace of that system. Analysis of the cloud of structures thus produced will then yield a more accurate description of that space. The procedure may be repeated until no changes are detectable to obtain a completely converged definition of the modes spanning the essential subspace. We are currently investigating such methods[107, 108].

3 TOWARDS AN EXHAUSTIVE SAMPLING OF THE CONFIGURATIONAL SPACES OF THE TWO FORMS OF THE PEPTIDE HORMONE GUANYLIN

B.L. de Groot, A. Amadei, D.M.F. van Aalten and H.J.C. Berendsen

J. Biomol. Str. Dyn. 13 : 741-751 (1996)

Summary

The recently introduced Essential Dynamics sampling method is extended such that an exhaustive sampling of the available (backbone) configurational space can be achieved. From an initial Molecular Dynamics simulation an approximated definition of the essential subspace is obtained. This subspace is used to direct subsequent simulations by means of constraint forces. The method is applied to the peptide hormone guanylin, solvated in water, of which the structure was determined recently. The peptide exists in two forms and for both forms, an extensive sampling was produced. The sampling algorithm fills the available space (of the essential coordinates used in the procedure) at a rate that is approximately six to seven times larger than that for traditional Molecular Dynamics. The procedure does not cause any significant perturbation, which is indicated by the fact that free Molecular Dynamics simulations started at several places in the space defined by the Essential Dynamics sampling, sample that complete space. Moreover, analyses of the average free Molecular Dynamics step have shown that nowhere except close to the edge of the available space, there are regions where the system shows a drift in a particular direction. This result also shows that in principle, the essential subspace is a constant free energy surface, with well-defined and steep borders, in which the system moves diffusively. In addition, a comparison between two independent essential dynamics sampling runs, of one form of the peptide, shows that the obtained essential subspaces are virtually identical.

Introduction

Recently, the structure of the peptide hormone guanylin was elucidated by NMR[112]. Two distinct conformations were found, denoted A and B, present in equal amounts, which differ in the way two internal disulphide bridges are arranged with respect to the main chain of the peptide. The hormone as it was studied consists of 13 residues and the two conformations can be classified as a right handed spiral (A form) and a left handed spiral (B form)[112]. Interchange between the two forms was not observed experimentally, a finding which was supported by computational methods[112].

Guanylin is an endogenous ligand to the heat stable enterotoxin receptor (STaR), an intestinal guanylyl cyclase[113], causing the production of cyclic GMP when activated. For a review, we refer to[114]. Cyclic GMP plays an important role in fluid regulation in the intestines and overproduction leads to severe diarrhea[115]. Guanylin competes with heat stable enterotoxins (STa) in binding to STaR and is homologous to it[113, 116].

Recently the Essential Dynamics (ED) technique[78] was extended by introduction of a sampling technique that makes use of constraint forces in the essential subspace, where most relevant motions occur[81]. Here, an improved algorithm of this ED sampling technique is presented, which causes less perturbation and performs a rapid filling of the essential subspace. The method is applied to both forms of guanylin, where borders of the allowed region are found in almost every direction, indicating an almost complete sampling. The allowed space found by this method coincides with the space that would be found when a MD simulation would be extended to infinite time. This is shown by starting (free) MD simulations at several places at the border of the essential subspace.

The allowed spaces of the two forms are compared to each other, to the spaces sampled by the initial MD and to the spaces sampled by free MD simulations started at a number of positions in the sampled region. Moreover, we investigated the possible influence of the initial definition of the essential eigenvectors on the results obtained from an extended ED sampling. Careful investigation of the dynamical behaviour of the essential coordinates both during the ED sampling procedure and during MD simulations started at several distinct places in the essential subspace indicate a diffusive behaviour in a constant free energy basin, with well defined borders.

Methods

MD simulations

MD simulations of both the A and the B state were initiated from a corresponding NMR structure (the first structures of the PDB entries 1gna and

1gnb respectively). In both cases, the peptides were surrounded by SPC water molecules[98] (571 and 511 solvent molecules respectively) filling up a truncated octahedron box. The net negative charge was compensated for by a sodium ion which was placed by substitution of the water molecule at lowest potential. Both systems were energy-minimised after which a heatup procedure was used to equilibrate the systems. Subsequently, both systems were simulated for 1 ns, of which the last 750 ps were used for analyses. The temperature was kept constant at 300 K by weak coupling to an external bath[99] (peptide and solvent were coupled separately with a coupling constant of 0.1 ps). The pressure was also kept constant by coupling to a bath with a coupling constant of 0.5 ps. SHAKE[11] was used to constrain bond lengths, allowing a time step of 2 fs. All calculations were performed with the simulation package Gromos[4]. All structure evaluations and visualizations were performed with the program WHAT IF[101].

Initial definition of essential subspace

From the structures of the last 750 ps of both simulations covariance matrices of positional fluctuations (C-α only) were built and diagonalised. Eigenvectors are directions in configurational space and the corresponding eigenvalues indicate the mean square fluctuations along these axes[78]. The procedure corresponds to a linear multidimensional least squares fitting of a trajectory in configurational space[76, 117]. Sorting the eigenvectors by the size of the eigenvalues shows that the configurational space can be divided in a low dimensional (essential) subspace in which most of the positional fluctuations are confined, and a high dimensional (near-constraints) space in which merely small uninteresting vibrations occur.

ED sampling protocol

With ED, all relevant motions, i.e. those with an appreciable amplitude, can be (approximately) described by only a few collective coordinates representing a small fraction of the total number of degrees of freedom. As was shown before[81], this can be used to sample the configurational space more efficiently than by traditional MD. In a previous paper[81] we introduced the concept of constraint dynamics, applying constraints in the essential subspace in the form of an expanding radius (spanned by e.g. three essential coordinates). Here, a modification of that protocol is introduced. Instead of performing an expansion of a radius with a fixed increment in the radius per step, a choice is now made every step between expanding the radius or keeping the radius fixed at the current value, depending on the direction a normal MD step would have taken.

So, at every (usual MD) step an evaluation is made. If the new position in the chosen essential subspace is further from the starting position than the position in the previous step, no correction is applied. If, however, the new position is closer to the starting position, it is moved back, by means of a constraint force on the alpha-carbon eigenvectors, to a position which has the same distance to the starting position. From the new position the velocities are recalculated. Using the principle of least perturbation, the correction is performed in the direction of the radius vector as described in[81]. In this way, the distance from the origin is not forced to increase if it will not spontaneously do so. Instead, it will move on a sphere with fixed radius, until a direction is found in which the system can expand. If the radius does not increase for a certain time (in this case a criterion of 500 subsequent steps was used), indicating that the system approaches a border, a new expansion cycle is started. The last configuration is used as a center of the new expansion sphere.

To avoid oscillation in a particular direction in subsequent expansion cycles, in every cycle an initial linear expansion (of 1000 steps) along one of the eigenvectors used for the radius expansion is performed. The eigenvector used for this linear expansion and the direction are chosen randomly. The principle for such a linear expansion is the same as for the radius expansion: a step is accepted if the distance from the origin increases in an unperturbed MD step. When the distance decreases, it is put back to the original value.

Thus, there are three major differences with respect to the original ED sampling protocol[81]. First, during expansion cycles, a constraint is only applied to prevent the system from going back, not to push it further from the original position. In this way, the system is not forced to move in unfavourable regions, and expansions will stop automatically if a border is reached. Second, the size of the expansion step is not fixed but is determined by the usual MD step, causing least perturbation at the most efficient expansion speed. Third, the initial linear expansion in an arbitrary direction forces the system to move in a direction other than the reverse of the previous cycle, causing a more rapid filling of the allowed space. The software used is an adaptation of the simulation package Gromos[4].

During the ED sampling of the two states, all MD parameters were kept at the same values as during the free simulations. For both states, a three-dimensional ED sampling (using the first three alpha-carbon eigenvectors from each state respectively) of 100 cycles was performed. Because this calculation showed similar behaviour for the two states, it was decided to concentrate further studies on the A state. An additional 100 cycles of the three-dimensional ED sampling were performed for the A state to investigate the completeness of the ED sampling. Also for the A state, several free MD simulations of 100 ps each were started at the borders to investigate if the allowed space as defined by the ED sampling algorithm is consistent with the behaviour of free MD simulation, i.e. to check the stability of the essential subspace.

Finding borders in the essential subspace

As stated above, in the expansion cycles, the expansion stops when there is no spontaneous increase of the distance from the origin anymore. This will cause the procedure to stop when a border of the allowed region in the essential subspace is reached. The constraint used to prevent the system from moving back towards the origin of the expansion makes the system move along a sphere with fixed radius, causing additional sampling of the border region.

To define the location of the borders quantitatively, the average free step of the essential coordinates during the free MD steps (so excluding corrections in the essential positions) was evaluated in every position of the essential space, using a grid. In a position not near a border for a specific coordinate, the average free step vector is expected to be zero (indicating an equal probability to move in each direction). So a non-zero average free step indicates the proximity of a border.

An ED sampling of the A state using only the first two eigenvectors instead of the first three was performed to investigate the average free step in detail on the grid defined by the first two eigenvectors.

Calculation of the configurational volume

To obtain a quantitative measure of the sampled configurational volume, a cubic grid was put over the space spanned by the first three eigenvectors, which were used in the ED sampling protocol. During the ED sampling procedure the number of non-empty grid elements was multiplied by the volume per grid element to give an estimate of the evolution of the sampled configurational volume in these three dimensions. The grid size must be carefully chosen to represent the sampled volume correctly. A fine grid underestimates the volume and makes it proportional to the number of sampled points, while a coarse grid may introduce incorrect connectivity. A suitable compromise was found when for each of the first three eigenvectors 10 intervals were chosen between -2 nm and 2 nm, dividing the 3D space in 1000 grid elements. For this grid size the volume is practically independent of the density of sampled points.

Results

To estimate the efficiency of the ED sampling protocol, three evaluations were done. First, projections of the trajectories produced by the expansion cycles onto the three planes defined by the first three eigenvectors were compared to the projections of the free MD simulations onto these planes (Fig. 3.1). For both conformations, the ED sampling run has not only been able to reproduce the complete region that had been sampled by MD, but has significantly enlarged that region in every direction.

Second, to obtain a more quantitative measure of the efficiency of the ED sampling protocol, the volume of the space sampled in three dimensions as a function of the number of integration steps was compared for the ED sampling runs and the MD simulations (Fig. 3.2). The slope of the curves is a measure of the efficiency of the sampling protocol, and for both conformations, the ED sampling method produces a significantly steeper plot than the MD, indicating a high efficiency of the ED sampling protocol. The ratio of the slopes of the straight lines fitted to the volume curves of the ED sampling technique and MD was approximately 6 to 7 for both states.

The curves of the volume (Fig. 3.2) corresponding to the two ED sampling runs both start to level off after approximately 1 million integration steps (corresponding to 2 ns of simulation with a time step of 2 fs), indicating that the allowed space defined by the first three eigenvectors has been completely sampled.

Third, in Figure 3.3, we compare the eigenvalues of the ED sampling with the eigenvalues of the initial MD runs. This figure shows that the first

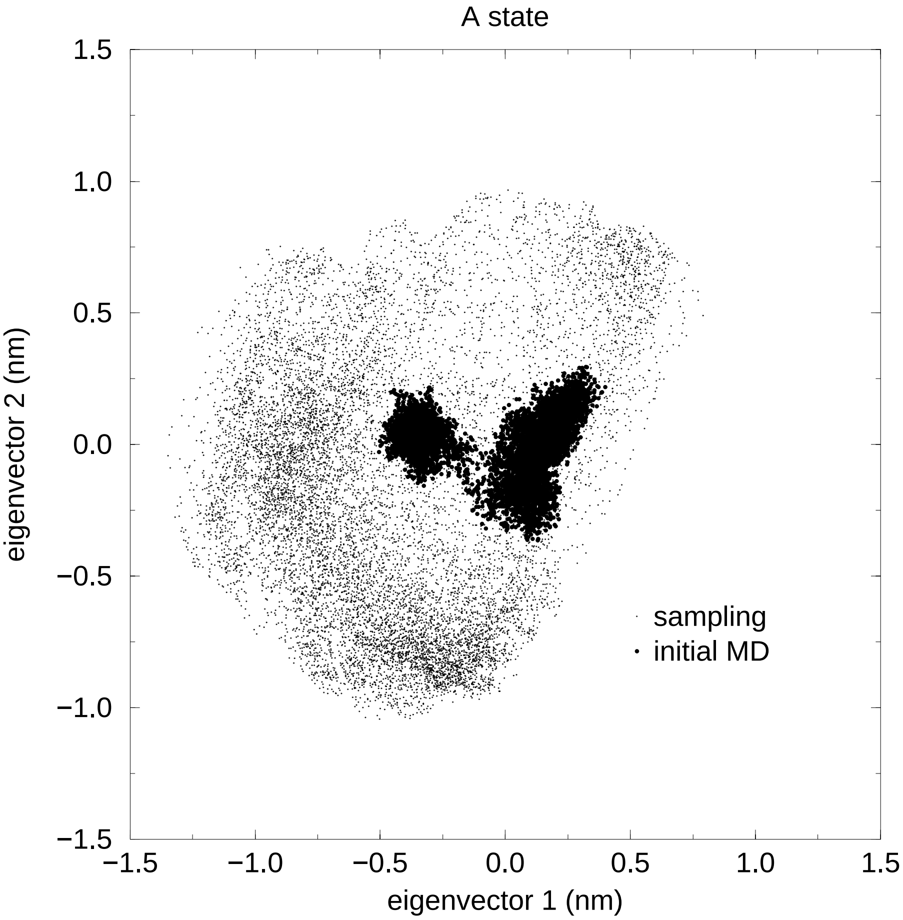

Figure 3.1 A comparison between the region sampled by the initial free MD simulation of 750 ps and by the ED sampling procedure projected in the plane defined by eigenvectors 1 and 2 from the free MD simulation. A state.

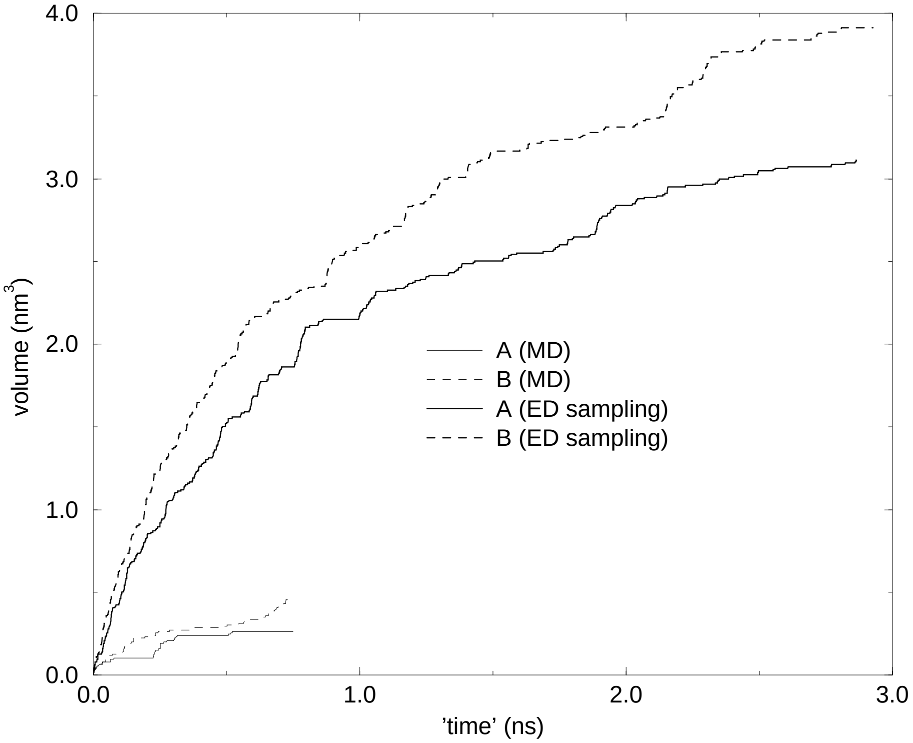

Figure 3.2 The sampled volume calculated over a grid in the space defined by the first three eigenvectors of each state as a function of the number of integration steps, which corresponds to time in a free MD simulation.

ten eigenvalues from the ED sampling are much larger than those from the MD simulations. This again indicates that a much larger essential subspace volume has been covered.

The (2D) ED sampling of the A state in the plane defined by the first two eigenvectors samples the same region in the 1-2 plane as does the 3D ED sampling. The third dimension is somewhat less well sampled because it was not forced to sample the borders, but the sampled 3D volume was very close to that of the three-dimensional ED sampling. Fig. 3.4 shows the average free step in the plane defined by eigenvectors 1 and 2, calculated from the two-dimensional ED sampling of the A state. The arrows indicate the size and direction of the average free step in every point, spread over a square grid. Almost anywhere close to the border of the sampled essential subspace, the average free step is non-zero and points towards the center of the allowed region. This indicates that the sampled space coincides with almost the complete available space of the A state, in this subspace.

To investigate the effect of the ED sampling algorithm on the definition of the borders of the essential subspace, free MD runs were started at several places at the edges of the sampled space of the A state. If the borders are really

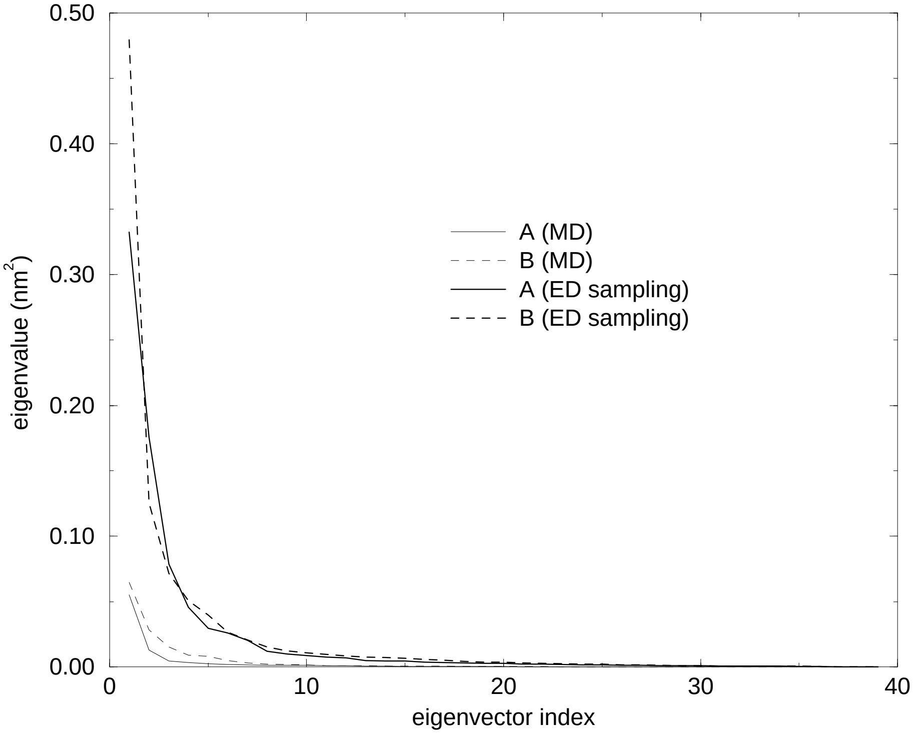

Figure 3.3 The eigenvalues obtained from an initial MD simulation of 1 ns (of which the last 750 ps were used for analyses) for both the A and B state, as well as those calculated from the set of structures built by an extensive ED sampling.

located as indicated by the non-zero average free steps, the region where the average free step is almost zero should also be available in a free simulation. Fig. 3.5 shows the projections of the structures generated by these free runs as well as those from the 2D ED sampling. All free runs move away from the edges for a short time in the direction of the center of the allowed region to fill in the complete allowed region, leaving only the edges unsampled indicating that the whole space produced by the ED sampling is accessible to dynamics.

Energies produced by free runs in regions distinct from the region sampled by the initial free MD simulation of the A state showed no significant differences from the energies produced by the initial free run. The average potential energy of the initial free MD simulation is -25.62 MJ/mol with a standard deviation of 0.17 MJ/mole. For three free simulations in different parts of the essential subspace the averages were -25.66 MJ/mol (with a standard deviation of 0.16 MJ/mole), -25.64 MJ/mol (0.18 MJ/mol) and -25.69 MJ/mol (0.15 MJ/mole) respectively. Also, energies produced during the ED sampling are similar to energies from free MD simulations, in regions not near the borders (average energy: -25.75 MJ/mol, standard deviation: 0.32 MJ/mole). Thus, the fluctuation of the energy is significantly larger during the ED sampling procedure, whereas the average energy is close to the average obtained in a completely free simulation.

To investigate the accuracy of the definition of the essential subspace, i.e. to check if the essential coordinates and near-constraints are consistent in different MD simulations of about 1 ns, a covariance matrix was built and diagonalised for two (uncorrelated) trajectories of the A state of 1 ns each, started in distinct regions of the initial essential subspace. These analyses are compared to an analysis of the structures produced by the 3D ED sampling of the A state.

A way to compare two sets of eigenvectors is to monitor the cumulative square inner product of one eigenvector from one set with all eigenvectors from the other set. This sum will converge to 1.0 because all eigenvectors of one set will always be able to rebuild the other set. Figs. 3.6A and 3.6B show this cumulative square inner product between single eigenvectors of the

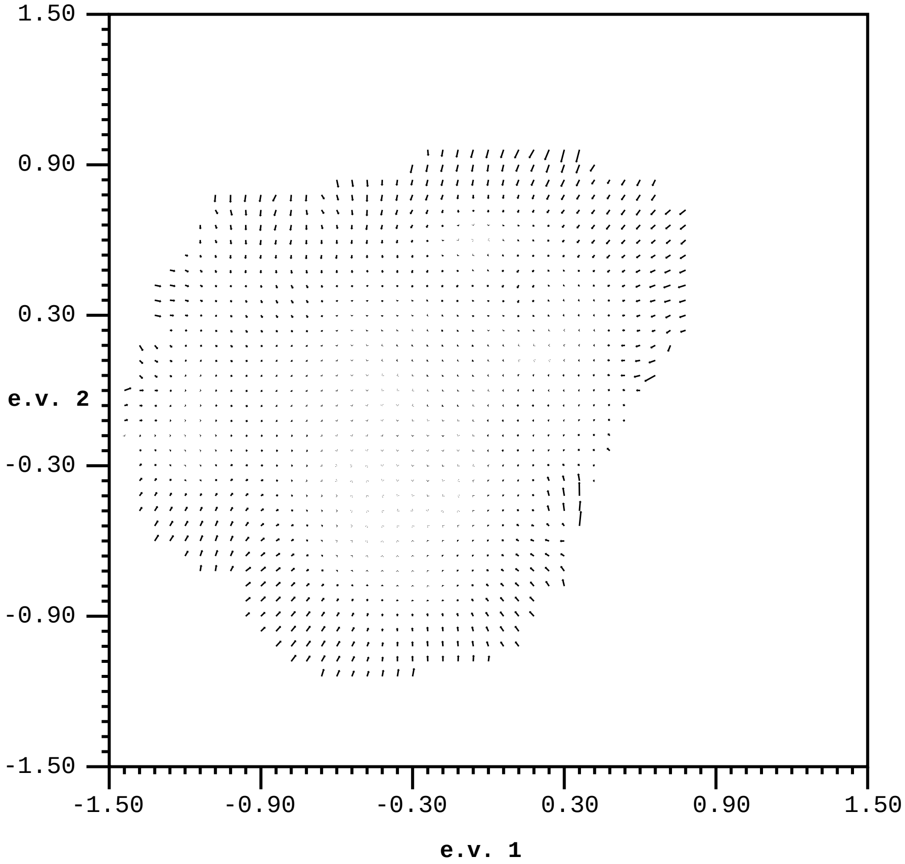

Figure 3.4 The average free steps calculated over a grid projected in the plane defined by the first two eigenvectors. The arrows indicate the direction and size that a free MD step would take on average in each position. A state.

two 1 ns MD runs and those of the whole ED sampling set. Although the analyses of the simulations contain an appreciable amount of noise, they show a considerable similarity of essential subspaces approximately defined by, e.g. , the first five degrees of freedom. It is also important to note that typical near-constraint eigenvectors of the MD runs (like 20 and 30) do not mix at all with the essential eigenvectors of the ED sampling run. Therefore, 1 ns of free simulation is enough to obtain a basic description of the fully converged essential subspace (as was also found previously[81]) and the definition of the essential subspace is consistent in different parts of the configurational space.

Figure 3.6C shows the cumulative square inner product between the eigenvectors from the ED sampling of the A state, and those from another, inde-

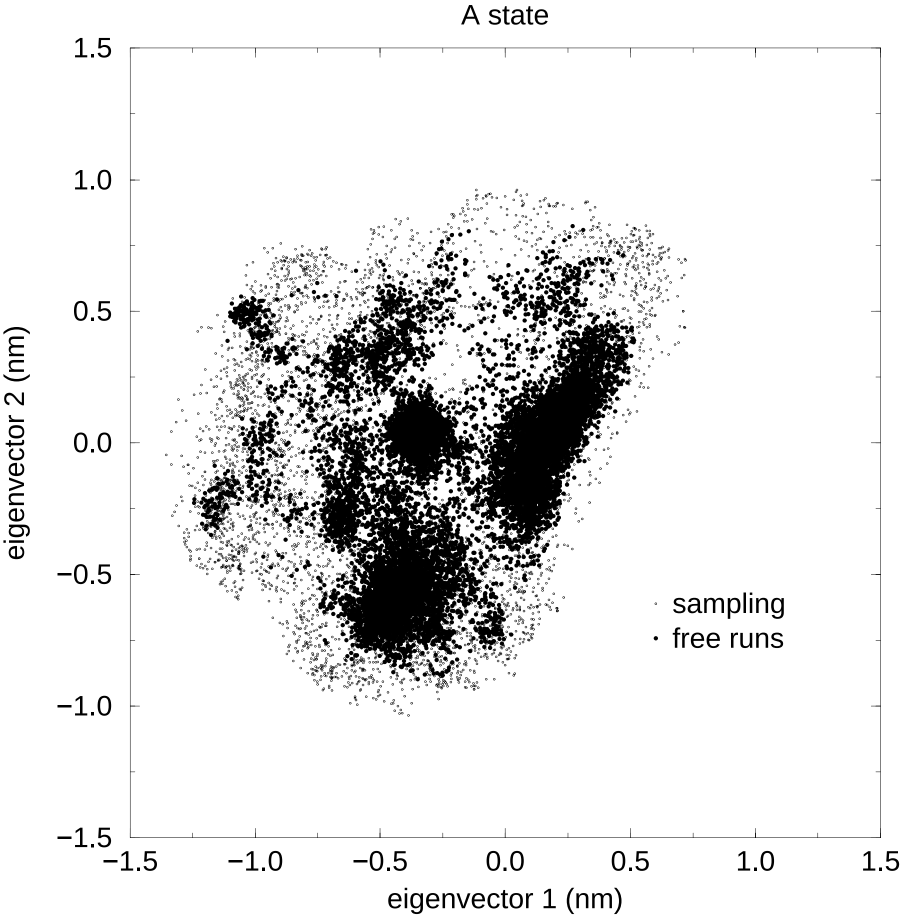

Figure 3.5 A comparison between the region sampled by the ED sampling procedure and by multiple free simulations of 100 ps started at random places in that space. The structures are projected in the plane defined by eigenvectors 1 and 2 from the initial free MD simulation. A state.

pendent, ED sampling of the A state. The latter sampling was produced by constraining the position along eigenvectors constructed from a different initial MD simulation. Compared to Figs. 3.6A and 3.6B, the similarities between the two sets are much higher. This shows that both sets have converged to the same definition of the space. Compared to Figs 3.6A and 3.6B, a much better definition of the essential subspace is obtained because the statistics of the covariance matrix is better when it is based on a complete ED sampling rather than a 1 ns MD run.

Fig. 3.7 shows different structures of the A state produced by the initial MD compared to structures produced by the ED sampling algorithm. The

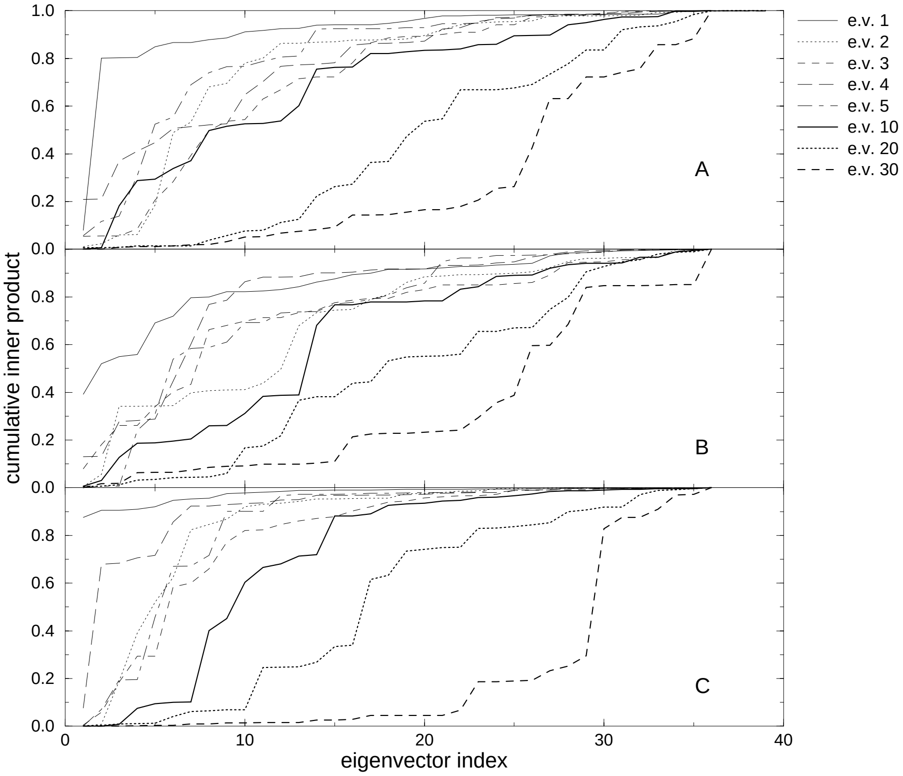

Figure 3.6 Panels A and B: Cumulative square inner products between eigenvectors obtained from two 1 ns free MD simulations and eigenvectors built from the complete collection of structures obtained from ED sampling. Along the X axis are the eigenvector indices of the free MD simulation which are used to rebuild single eigenvectors obtained from sampling. A state. Panel A shows the results from one MD simulation and panel B of another simulation started in a different region of the essential subspace. Panels C shows the cumulative inner products between two sets of eigenvectors, obtained from two independent ED sampling runs. A state.

structures produced by the ED sampling deviate much more from the starting structure than the ones produced by the initial MD, illustrating the fact that a much larger part of the configurational volume has been sampled.

Comparison of the structures produced by the ED sampling procedures of the A with those of the B state showed that there is no overlap between the configurational spaces available to the two forms. Also, direct attempts to drive the system from one state to the other (also using the method of least perturbation[81]), constraining the position along eigenvectors that define the differences between the A and B state, failed to accomplish a transition from one state to the other. This suggests that the free energy barrier to move

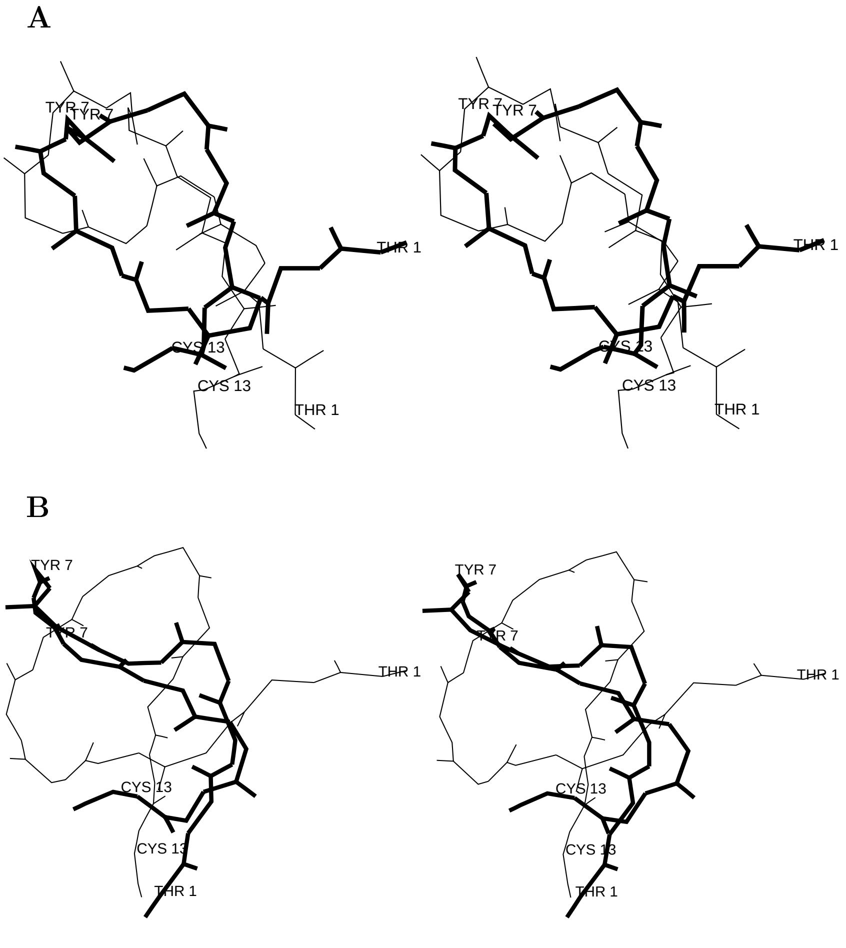

Figure 3.7 Comparison between structures obtained from free MD simulation and ED sampling. A state. panel A: MD, panel B: ED sampling. In both figures, a stereo picture of the structures corresponding to the minimum and maximum sampled position along eigenvector 1 are shown.

from one state to the other is too high to be passed, and that under usual circumstances the two forms each have their own distinct essential subspaces, with no overlap. This is consistent with experimental data[112], which shows distinct species on a time scale of at least seconds.

Conclusions and discussion

The results shown before prove that with the ED sampling technique, improved in this paper, it is possible to approach an almost complete sampling of the essential subspace of a small peptide in water, within a number of integration steps comparable to a simulation time of about 3 ns. The initial slope of the curve of the sampled volume in the subspace defined by the first three eigenvectors which is a measure of the efficiency of a sampling protocol, indicates that the ED sampling algorithm is six to seven times more efficient than usual MD. Therefore, the cluster obtained by ED sampling is much larger than that produced by usual MD of comparable length. This means that for the macroscopical properties that are evaluated by averaging over the ensemble of collected structures, results can be expected to differ between the two clusters. This is especially true for properties that are sensitive to the extreme structures in the ensemble, i.e. properties that depend on the spread in the cluster, rather than on the average.

The fact that the 2D ED sampling reproduces almost all of the 3D space that was obtained by the 3D ED sampling procedure suggests that the use of three essential dimensions is sufficient to obtain a complete sampling of the essential subspace.

The analysis of the average free steps suggests that, apart from the borders, no real free energy gradients are present in the essential subspace. This implies that unconstrained dynamics can be considered as a random walk in the essential subspace, resulting in diffusion like behaviour, as previously observed in the protein HPr[81]. Only those regions where an appreciable average free step is measured (0.0004 nm), remain unsampled by the free MD simulations (Figs. 3.4 and 3.5). This suggests that a non zero average free step is a good indication of a border (or a less favourable region) in the essential subspace and that no significant perturbation is induced in the sampling algorithm.

The ED sampling algorithm searches for possible expansion directions until no progress is found anymore, in every single cycle. Numerous of such cycles have been completed for both forms of the peptide. The results of the average free step calculation (Fig. 3.4) show that, for the A state, in virtually every direction of the essential subspace, a border has been found, indicating that the search in these directions is complete. This suggests that it is improbable that there are paths accessible for the system towards other stable regions in the same essential subspace.

Due to the soft nature of the applied constraints, no unallowed regions have been sampled. This is illustrated by the fact that, for the A state, free MD simulations have filled in the complete allowed space as defined by the ED sampling. Moreover, the free runs span one closed region which indicates that no physical barriers have been passed by the sampling algorithm.

Essential subspaces defined locally from MD simulations in two different parts of the configurational space are similar to the essential subspace defined from the complete ED sampling (Fig. 3.6), which validates the use of a rough initial definition of the essential subspace.

The fact that the two eigenvector sets obtained from two independent sampling runs are so similar (Fig. 3.6) proves that the definition of the space by this ED sampling protocol is consistent in itself, and not dependent on the initial MD simulation from which eigenvectors are extracted that are used to explore the space. Therefore, the approximated essential subspace defined from an initial MD simulation can indeed be used in an extrapolation protocol to sample the complete allowed space and to refine the description of the essential subspace.

Structures in the NMR cluster are much closer to each other for the A state (mean backbone RMSD with respect to the average structure of 0.47 Å) than for the B state (mean RMSD of 1.07 Å), suggesting more configurational freedom for the B state than for the A state[112]. This is partially reflected by the available configurational volume as obtained from the ED sampling runs of both the A and B states (Fig. 3.2),which shows a larger accessible volume for the B state than for the A state. These volumes, however, are much larger than those calculated from the NMR clusters. We think that the refinement protocol used to produce the NMR structures might give a too rigid representation of the molecule, because those structures are selected that simultaneously fulfill most NOE restraints. Thus the cloud of structures produced is close to the average structure, and the differences between the structures are not necessarily a good indication of the main modes of motion for the molecule. This is supported by a recent study[103].

The results presented in this paper show that it is possible to obtain a complete sampling of the essential subspace of a small peptide in water, together with an accurate definition of the location of boundaries. This suggests, together with previous results[81], that similar methods can be used to study the configurational space for larger peptides and proteins. We have evidence that this is indeed the case although the computational effort is considerably larger. The borders of the essential space now contain regions that may involve unfolding pathways.

Acknowledgement

It is a pleasure to thank dr. Ruud Scheek for many stimulating discussions.

4 AN EXTENDED SAMPLING OF THE CONFIGURATIONAL SPACE OF HPR FROM *E. COLI*

B.L. de Groot, A. Amadei, R.M. Scheek, N.A.J. van Nuland and H.J.C. Berendsen

PROTEINS: Struct. Funct. Gen. 26: 314-322 (1996)

Summary

Recently, we developed a method to obtain an extended sampling of the configurational space of proteins, using an adapted form of Molecular Dynamics simulations, based on the Essential Dynamics method. In the present study, this sampling technique is applied to the Histidine containing Phosphocarrier protein HPr from *Escherichia coli*. We find a cluster of conformations that is an order of magnitude larger than that found for a usual MD simulation of comparable length. The structures in this cluster are geometrically and energetically comparable to NMR structures. Moreover, on average, this large cluster satisfies nearly all NMR derived distance restraints.

Introduction

HPr is a component of the phosphoenolpyruvate (PEP)-dependent phosphotransferase system (PTS). The PTS is responsible for the phosphorylation and translocation of sugars from outside the cell to the inside, and the subsequent phosphorylation of these molecules. For a review, see for instance[118]. The role of HPr in the PTS is the transportation of a phosphoryl group from Enzyme I to Enzyme II, the enzyme which performs the actual transport and phosphorylation of the carbohydrates.

The high resolution structure of HPr from *Escherichia coli* was elucidated both from NMR data[93] and X-ray data[119]. *E. coli* HPr consists of 85 residues and shows an open-faced β sandwich fold, with three α-helices on top of a four-stranded β-sheet.

Protein dynamics as derived from MD simulations can be split into two classes of motion[76, 78]. A low-dimensional 'essential' subspace, in which most of the fluctuations are concentrated, is distinguished from a high-dimensional 'near-constraints' subspace in which merely small amplitude, fast equilibrating motions occur. Essential subspaces have been shown[53, 78, 84, 85] to contain functionally relevant information for the simulated proteins. Discussion has been going on about how robust the definition of the essential subspace is[109–111] in view of limited sampling time currently available in MD. We have shown[81, 85, 107] that the description of the two subspaces approximately converges after only a few hundred picoseconds of simulation (although motions within the essential subspace are not adequately equilibrated in such short time, and therefore the definition of the single essential eigenvectors is far from being converged). We have developed a technique[81, 107] that performs an adapted form of MD with constraint forces in the approximated essential subspace. This method yields an enhanced sampling of the configurational space as compared to usual MD, and the extrapolation of the essential subspace, as approximated from a relatively short MD simulation, yields a refined picture of this subspace. When the method was introduced, we presented an application to HPr, where we showed that a cluster of structures generated in this way samples a significantly larger part of the essential subspace, while all structural properties remain intact[81]. Subsequently, we applied a modification of the method to a peptide hormone, for which we obtained an almost complete sampling of the available space[107].

In this chapter, an extended application to HPr is presented in which the quality of the obtained structures, the efficiency of the used protocol compared to usual MD, and the reproducibility of the results is discussed. Physical correctness of the obtained structures is not only measured in terms of geometrical checks, but also in terms of ensemble-averaged atomic distances, which are compared to experimental NMR data. The ED sampling technique was started from two different starting positions, using different initial approximations of the essential subspace, obtained from two MD simulations. A comparison is presented between the two clusters thus obtained, providing insight into the convergence of the essential subspaces.

Methods

An MD simulation in solvent was started from one of the structures in the deposited NMR cluster (pdb entry 1hdn). This simulation of 350 ps is described elsewhere[120]. Over the last 300 ps of this simulation, a covariance matrix was built and diagonalised. Obtained eigenvectors are directions in configurational space and corresponding eigenvalues give the mean square positional fluctuation for each direction[78]. The three eigenvectors with highest eigenvalues (the first three, eigenvectors are ordered to decreasing eigenvalue) were used in a constraint dynamics procedure, where constraint forces are only applied in this essential subspace, as described in[81]. The practical implementation is identical to that described in detail in[107], except for two modifications. Briefly, the algorithm consists of the following steps: a starting position is defined as the set of essential coordinates of the starting conformation; a number of regular MD steps is performed; for each step, a new position is accepted only if it is not closer to the starting position than the previous position, in the subspace defined by the first three eigenvectors (i.e. if the distance from the starting position in this subspace does not decrease). If the new position is closer to the starting position, a correction is applied only in the subspace defined by the first three eigenvectors with least perturbation[81]. The correction is applied along the radius direction such that the position after correction is at the same distance (in the essential subspace) from the starting position as the previous position. In this manner, the system is encouraged to sample new regions in the essential subspace, and prevented from going back to places that have been visited before. When the distance from the initial position does not increase spontaneously anymore, the cycle is finished, and a new cycle is started with the current position being the new starting position. Because of the applied correction the dynamics of the system is altered and therefore, time information is lost, except for local, fast equilibrating motions. Integration steps are therefore merely dynamical sampling steps that cannot directly be interpreted as time steps.

Here, we use the same algorithm, with two modifications. First, backbone N, C-α and carbonyl C atoms were used in the analysis instead of C-α atoms only, to make sure that all backbone motions are included in the analysis. We have shown previously[78, 84] that C-α only analyses usually yield all necessary information, but in the case of HPr, unusual ϕ/ψ combinations have been observed[93, 120–122] and inclusion of the other backbone atoms in the ED analyses is necessary for description of these angles. Second, the criterion for finishing a cycle to start another one was altered. The rate at which the distance from the starting position increases in time is taken as a criterion. This rate was allowed to decrease to zero in the case of the peptide hormone we studied previously[107], but since such a low value resulted in denaturation of HPr, we chose for an average value of $2.5 \cdot 10^{-4}$ nm per step (for comparison, this value is typically $7.0 \cdot 10^{-4}$ nm per step in the initial phase of each cycle). 125 of such cycles were produced, in which an equivalent of approximately 1.8 ns of simulation time was reached. At a position distinct from the starting position of the first MD simulation, a second MD simulation of 350 ps was started. The eigenvectors derived from the covariance matrix built over the last 300 ps of this trajectory were used to direct a second sampling of 75 cycles (consisting of 650.000 steps, equivalent to 1.3 ns). A free MD simulation of 1.0 ns was performed for comparison.

All generated structures were subjected to energy minimization with extra restraining forces (force constants of $4000\ \mathrm{kJ} \cdot \mathrm{mole}^{-1} \cdot \mathrm{nm}^{-2}$) on distances for which experimental NOE data were available[93], until no significant energy change was obtained.

Comparison with time-averaged distance restrained MD is presented. In total, three pieces of 200 ps (which originated from three distinct distance geometry structures) were used in these analyses. The distance restrained simulations are described elsewhere[93].

The software used is an adaptation of the simulation package Gromos[4]. Essential Dynamics analyses, structural checks and visualizations were performed with the molecular modeling program WHAT IF[101]. Secondary structure assignment and solvent accessible surface calculations were calculated by DSSP[102] and dihedral angle evaluations were carried out by PROCHECK[123]. Visualization of secondary structure maps was performed with a gromacs[5] analysis tool.

Results

From the set of structures produced by ED sampling, a covariance matrix was built and diagonalised. The two independent pieces of sampling were combined in this analysis to yield a good definition of both the essential and near-constraints subspaces. Fig. 4.1A shows the structures produced by ED sampling and free simulation, projected onto the plane defined by the first two eigenvectors. The region sampled during free MD is more compact than that obtained from ED sampling. The region obtained from ED sampling includes the area sampled by free MD, and extends it in every direction. The set of NMR structures[93] and the X-ray structure[119] lie close to the center of the region spanned by ED sampling (Fig. 4.1B). The sampled configurational volume[107], expressed in the space spanned by the three most prominent directions in the cloud of structures, is depicted in Fig. 4.2. The slope of the curve obtained from ED sampling is approximately seven times larger than that from free MD and time-averaged distance restrained MD, in the first ns, demonstrating the enhanced efficiency of sampling using the ED approach.

The structures produced by both ED sampling and free MD were validated in several ways. Fig. 4.3 shows the secondary structure as a function of time. Both during free MD (Fig. 4.3A) and ED sampling (Fig. 4.3B), the fold of HPr remains essentially stable. The amount of fluctuation around the mean structure is larger for the ED sampling than for the free MD simulation. All other geometrical properties are summarised in table 4.1. For all the

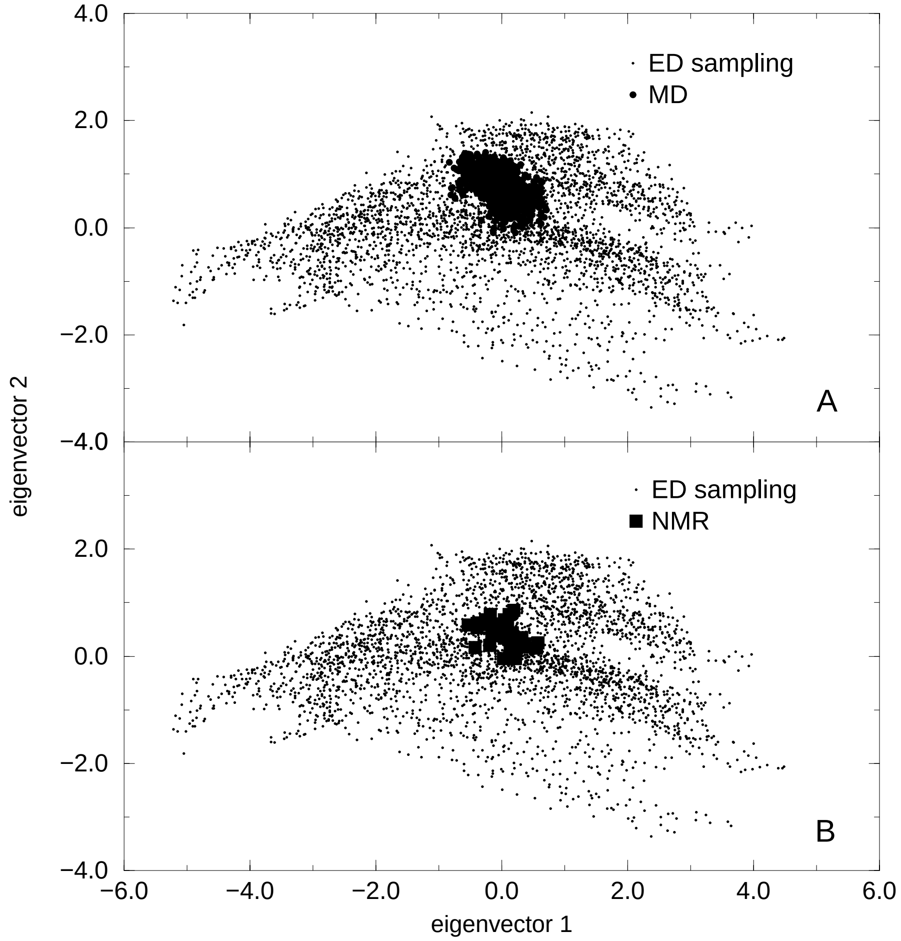

Figure 4.1 Panel A: Projection (in nm) of the collection of structures produced both by free MD (simulation time of 1.0 ns, filled squares) and ED sampling (sampling 'time' corresponding to 3.1 ns, open circles) onto the plane spanned by the two eigenvectors with largest eigenvalues from ED sampling. Panel B: Projection (in nm) of NMR (pdb entry 1hdn, filled squares) and X-ray (1poh, filled circle) structures compared to structures obtained from ED sampling (open circles) onto the plane spanned by the two eigenvectors with largest eigenvalues from ED sampling.

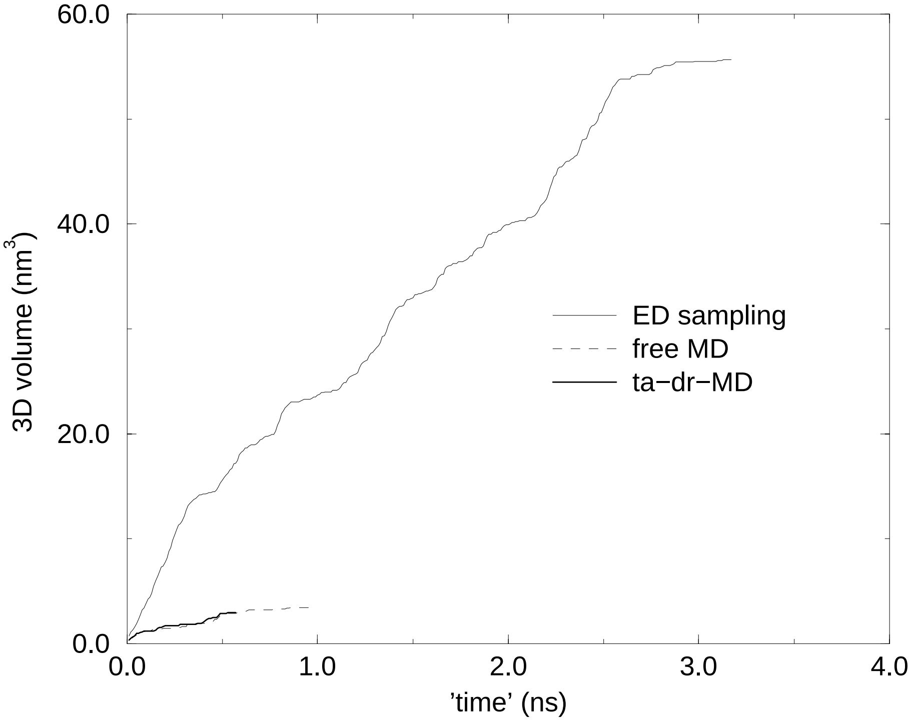

Figure 4.2 Sampled configurational volume (in nm^3) in the space spanned by the first three eigenvectors from ED sampling.

checked properties, averages for the free MD simulation, time-averaged distance restrained MD, and the ED sampling procedure are very similar. Only the RMSD from the mean structure is on average significantly larger in the ED sampling procedure. As observed for the secondary structure (Fig. 4.3), fluctuations of all geometrical properties in table 4.1 are larger during the ED sampling than during free MD. Together with the higher average RMSD, this again demonstrates larger conformational freedom during ED sampling. Energies (Table 4.1) are indistinguishable for the two forms of simulation. Time-averaged distance restrained MD shows even smaller fluctuations, in agreement with previous findings[103].

As a further validation of the produced structures, we monitored correspondence to distances derived from experimentally observed NOE data. Results are summarised in table 4.2. For both the ED sampling and the free MD simulations, violations with respect to the experimentally derived distances were evaluated after averaging over the whole set of structures, that were energy minimised with extra potentials on NOE derived distances. Only a very small fraction of distances (out of 1108 experimentally observed NOE's) is on average larger than the experimental upper limit, both during free MD and ED sampling. The sum of violations is in both cases of the same order of magnitude as calculated from the cluster of structures produced by time-averaged distance restrained MD simulations[93], the last step in the refinement of the NMR structures. It is interesting to note that the running average of the total sum of violations decreases in time (Fig. 4.4). This suggests that a more complete sampling will yield even fewer and smaller violations.

Eigenvectors obtained from two independent ED sampling runs and two independent MD simulations (from which the eigenvectors used to direct the two sampling runs were extracted) were compared to assess whether the definition of the refined essential subspace depends on their initial approximation. Fig. 4.5 shows a comparison between the essential subspaces from MD and ED sampling. For the first MD simulation (Fig. 4.5A), the overlap between the essential subspaces from the ED sampling runs and from MD is approximately equal. This indicates that during sampling, the eigenvectors which are used to direct the sampling do not dominate the produced cloud of structures. For the second MD simulation (Fig. 4.5B), the overlap between the

	mean	σ	mean	σ	mean	σ
	ED	ED	MD	MD	ta-dr-MD	ta-dr-MD
NRC	14.36	2.44	14.88	1.88	12.14	1.93
ACC	5551	184	5490	134	5635	134
DIH	5.08	1.21	4.70	1.11	4.60	0.92
HBO	69.4	4.2	67.6	3.7	71.0	4.0
GYR	1.143	0.011	1.140	0.008	1.165	0.007
RMS	1.715	0.469	1.053	0.150	0.762	0.124
EPOT	-130.48	0.42	-130.56	0.43	N.A.	N.A
EPW	-17404	370	-17391	375	N.A.	N.A.

Table 4.1 Comparison of geometrical and energetical properties between ED sampling, free MD and time-averaged distance restrained MD (ta-dr-MD). The first column contains abbreviations of the studied properties: NRC: number of residues (out of 85) adopting random coil conformation; ACC: solvent accessible surface (Å^2); DIH: number of residues in unfavorable regions of a Ramachandran plot[124]; HBO: number of backbone-backbone hydrogen bonds; GYR: radius of gyration (nm); RMS: root mean square deviation (Å) with respect to mean structure; EPOT: total potential energy (all interactions, including solvent) in GJ/mole; EPW: All energy terms involving the protein in kJ/mole. The other columns show the averages and root mean square fluctuations of these properties for ED sampling, free MD, and ta-dr MD respectively. Energies are not shown for the time-averaged distance restrained MD simulations because in these calculations, an extra term is present in the potential energy function, making direct comparison impossible. It should be noted that the values for the time-averaged distance restrained MD simulations were averaged over three simulations that started from distinct initial NMR structures, making fluctuations larger than what would have been found for a single simulation.

essential subspace of the second sampling is somewhat larger than that for the first sampling. Fig. 4.5C shows that the eigenvectors obtained from the two sampling runs are as similar to one another as are the eigenvectors from two free MD simulations. It should be noted that in all graphs of Fig. 4.5, the overlap between the first 10 eigenvectors of each set is always higher than 40 % and that within the first 50 eigenvectors (out of 765), more than 80 % of the essential subspace of the other set can be rebuilt.

All trajectories were projected onto the eigenvector with highest eigenvalue from the whole cluster of ED sampling structures. Fig. 4.6 shows the structures corresponding to the minimum and maximum position in this direction that were visited both during free MD and ED sampling.

As observed earlier[81], motion inside the essential subspace is of diffusive nature. Fig. 4.7A shows that, for eigenvector 1,2 and 3, during pieces of

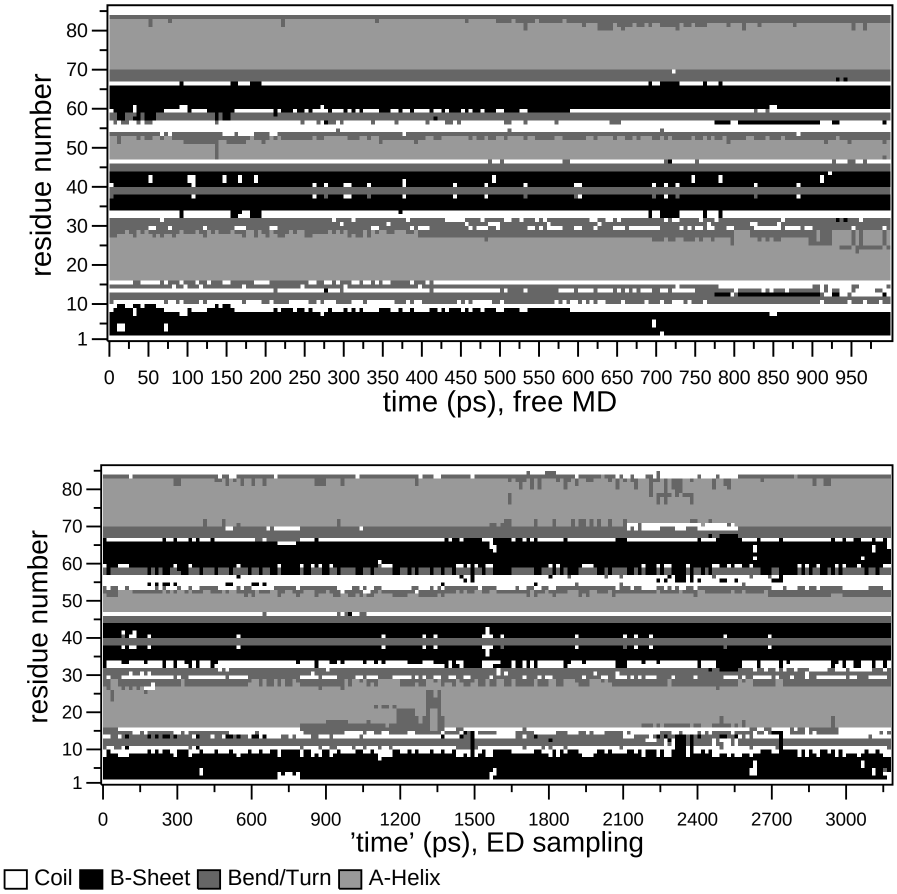

Figure 4.3 Secondary structure as a function of time. panel A: Free MD (1.0 ns). Panel B: ED sampling (3.1 'ns'). Data shown are for the reference simulation of 1 ns, and for the ED sampling, after combination of the two independent pieces.

	ED	MD	ta-dr-MD		
Sum of violations (nm)	1.093	1.300	0.739		
Largest violation (nm)	0.161	0.165	0.140		
Violations larger than 0.1 nm (ED)					
number	residue 1	residue 2	upper limit	average	violation
1	10	85	0.580	0.705	0.125
2	14	57	0.650	0.764	0.114
3	14	80	0.770	0.918	0.148
4	22	55	0.670	0.819	0.149
5	23	55	0.870	1.031	0.161
Violations larger than 0.1 nm (MD)					
1	14	80	0.770	0.929	0.159
2	22	55	0.670	0.809	0.139
3	23	55	0.870	1.035	0.165
4	29	33	0.540	0.643	0.103
5	29	77	0.450	0.592	0.142
6	76	77	0.450	0.556	0.105
7	76	80	0.450	0.592	0.142
Violations larger than 0.1 nm (ta-dr-MD)					
1	26	76	0.650	0.756	0.106
2	73	76	0.450	0.590	0.140

Table 4.2 Violations with respect to NOE derived distance restraints. For comparison, averaged distances were calculated as $< r^{-6} >^{-1/6}$ over clusters generated by ED sampling, free MD, and time-averaged distance restrained MD.

free simulation, the average square displacement increases linearly with time, indicative of diffusive behavior. For comparison, the same evaluation was done for a combination of near-constraints eigenvectors (Fig. 4.7B). In this case, a linear dependence was observed for a few picoseconds, after which the curve levels off, indicating the presence of a significant free energy gradient for these coordinates, as has been postulated[78].

Conclusions and discussion

As we found in an initial study of HPr[81] and for a small peptide (13 residues)[107], with the ED sampling technique it is possible to extend the amount of sampling in the essential subspace of proteins. For the peptide it was even possible to approach a complete sampling within a simulation time corresponding to a few nanoseconds. We showed in the present study that also for HPr, a significant gain in the rate in which the essential subspace is filled during simulation is attained with respect to usual MD.

From a structural and energetical point of view, the clusters produced by

ED sampling show similar properties as those from MD and time-averaged distance restrained MD. This makes the set of structures obtained by ED sampling equally acceptable as the smaller cluster from free MD or an even smaller cluster (e.g. from distance restrained simulations, for which we showed that motions inside the essential subspace may be damped[103]). So far, HPr is the only protein on which the ED sampling technique has been applied, but there is no reason to assume that other proteins will show significantly different dynamic behavior.

During distance restrained energy minimization, positions in the (backbone) essential subspace change only marginally (data not shown). This, together with the observation that violations of the experimental data are small, means that the averages of most of the experimentally derived distances are hardly affected by motions in the essential subspace. In a previous study[103], we showed that distance restraining during MD reduces fluctuations in the essential subspace, due to a few restrained proton pairs, whose

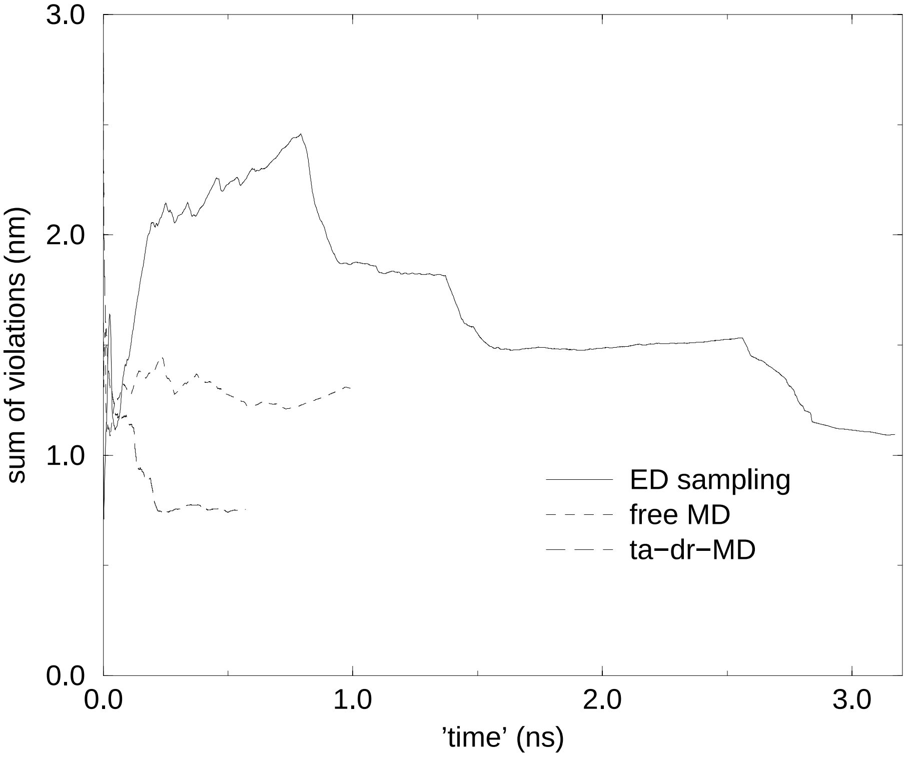

Figure 4.4 Sum of violations of $< r^{-6} >^{-1/6}$ averaged distances with respect to upper limits derived from NOE data as a function of the number of integration steps in the calculation. Data shown are for ED sampling, free MD and time-averaged distance restrained MD. For free MD and time-averaged distance restrained MD, the number of integration steps is proportional to time.

distances were affected by fluctuations in this subspace. Here, we demonstrate that fluctuations within the essential subspace are possible without violating the large majority of experimental restraints. In the graph of the configurational volume (Fig. 4.2), the curve corresponding to free MD is hardly any steeper than that obtained for time-averaged distance restrained MD. This is mainly because the graph consists of a combination of three pieces of simulation, started from different distance geometry structures. In general, a time-averaged distance restrained MD simulation can be expected to sample a smaller fraction of the configurational volume than a free MD simulation of comparable length would. This suggests that the reduced fluctuation in the essential subspace during distance restrained MD is mainly due to the attempt to fulfill all distances simultaneously (or averaged over too short time scales in time-averaged distance-restrained MD). The small number of distances which on average exceed their upper limits by more than 1.0 Å, are

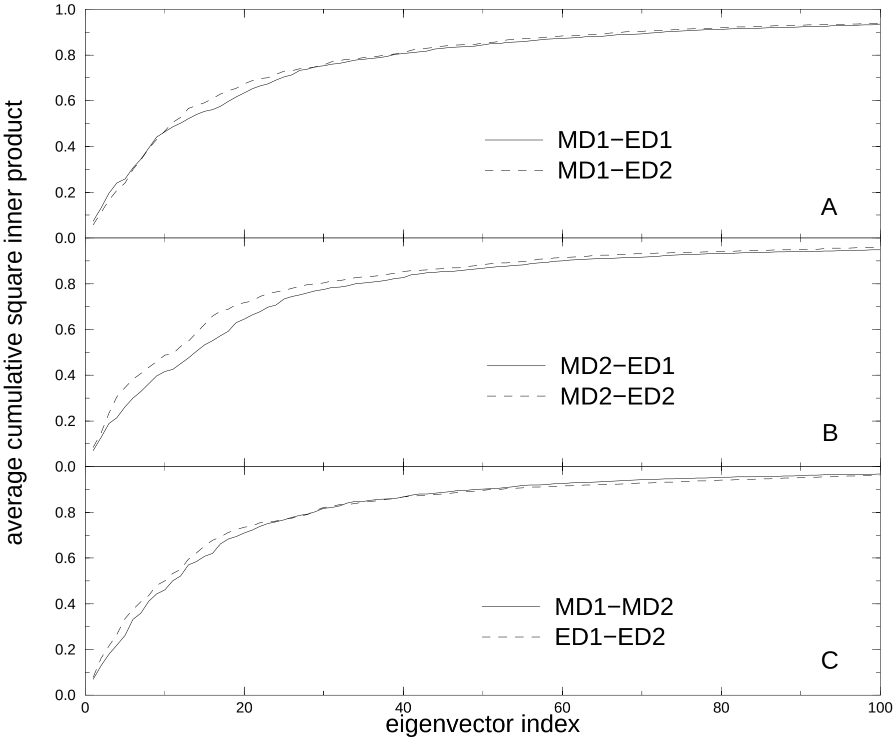

Figure 4.5 Cumulative average square inner product between different sets of eigenvectors. The first 100 eigenvectors (out of 765) of one set (the set which is named first in the legend) are compared to the first 10 eigenvectors of the other set. Panel A compares one MD simulation of 300 ps to two ED sampling runs. Panel B shows the same, now for another simulation of 300 ps. Panel C compares the two MD simulations and the two ED sampling runs.

among the largest distances derived from the set of observed NOE's. These NOE's are the weakest observable peaks in the NOE spectra and are therefore most susceptible to misinterpretation. Therefore we conclude that there is no good reason, based on NMR data, to favor the set of structures obtained from

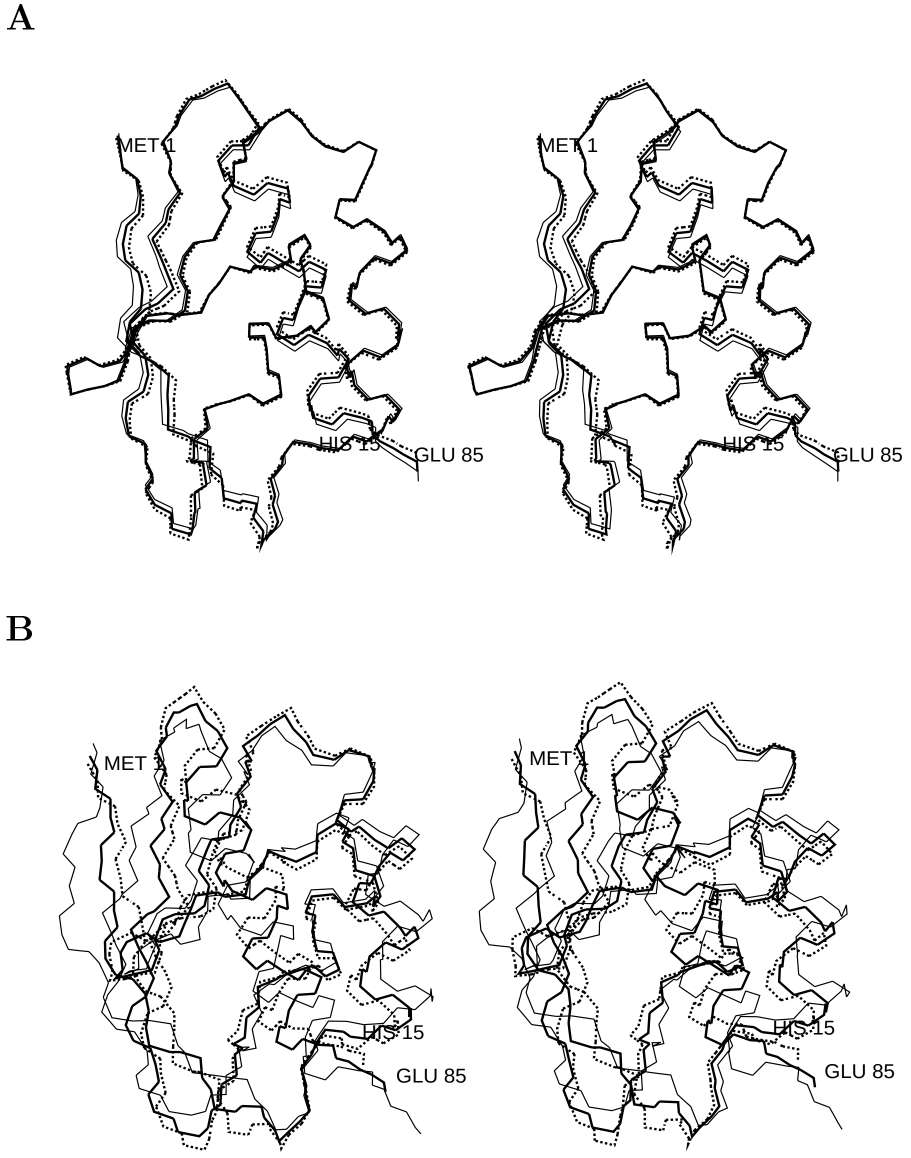

Figure 4.6 Stereo representation of the backbone structures corresponding to minimum (thin line), maximum (dashed line), and average (fat line) position along the first eigenvector from ED sampling. Panel A corresponds to free MD, panel B to ED sampling.

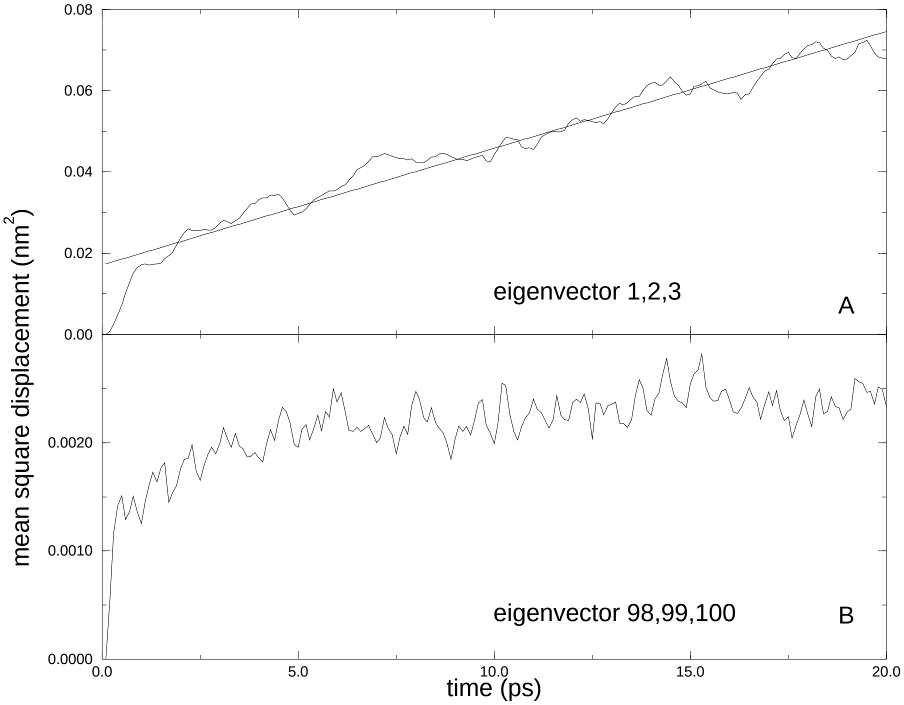

Figure 4.7 Mean square displacement along eigenvectors (in nm^2) averaged over 80 pieces of 20 ps of free simulation. Panel A corresponds to the displacements along the first three eigenvectors, panel B to the displacements along three near-constraints.

time-averaged restrained MD refinement over the present set of structures, as obtained from our ED sampling algorithm. Moreover, the few remaining violations decrease as the ED sampling proceeds (Fig. 4.4), suggesting that averaging over an even larger set of structures could enhance correspondence with the experimental data further.

Other analyses, like the calculation of the configurational volume sampled (Fig. 4.2) and comparison of different sets of eigenvectors (Fig. 4.5) suggest that the sampling is not complete yet. In a peptide of 13 residues, we observed that the curves in the plot of the configurational volume leveled off after a sampling time corresponding to 2 ns of free MD[107], indicating an almost exhaustive sampling. Although the slope of the curve of the volume for ED sampling does decrease with time for HPr, it is not approaching an asymptotic value yet. This suggests that the sampling is still not complete in this subspace and that only a small fraction of all possible configurations has been visited.

The fact that eigenvectors extracted from two independent MD simulations are as similar to each other as are eigenvectors from two independent ED sampling runs (Fig. 4.5), is another indication for this conclusion. The description of the essential subspace has approximately converged, even in a simulation time of a few hundred ps, although individual directions (eigenvectors) spanning this subspace are not fully converged yet. Even in a relatively extensive sampling, complete convergence is not yet reached. The overlap between two sets of eigenvectors, measured as the cumulative mean square inner product between a subset of (ten) eigenvectors of one set with all eigenvectors out of the other set (Fig. 4.5), shows that the largest contribution is always obtained from the first part of the second set. In all graphs, more than 90 % of the overlap is found between the subspace spanned by the first ten eigenvectors of one set and less than 10 % of the (largest eigenvalue) eigenvectors of the other set. It is important to note that in all compared sets, an amount of noise is present. In a pair comparison as presented in Fig. 4.5, this will have a more serious effect than in other forms of comparison.

The dynamic behavior (in the essential subspace) during the ED sampling procedure is affected by the presence of constraint forces. Therefore, the density of sampled configurations cannot be used directly for exact thermodynamic evaluations. However, the dynamic behavior of essential coordinates (Fig. 4.7) appears to be a form of diffusion, indicating that no significant free energy gradients exist in the essential subspace. Hence, a homogeneous equilibrium density distribution in the available essential subspace (as obtained by ED sampling) is to be expected. The sampling of the essential subspace of HPr as presented in this paper is not complete enough to provide insight into the details of the free energy surface in the complete essential subspace. The results suggest however, that the sampled fraction is a rather flat surface with at most many small ($<$ kT) local minima. In a previous paper we showed[107] that for a small peptide the borders of this space are well defined and steep. For HPr, the borders seem less well localised and rather soft in some directions, leaving routes for denaturation.

Acknowledgements

We thank Prof. Dr. G.T. Robillard for many stimulating discussions, and Daan van Aalten for support with a number of analysis programs.

5 DOMAIN MOTIONS IN BACTERIOPHAGE T4 LYSOZYME; A COMPARISON BETWEEN MOLECULAR DYNAMICS AND CRYSTALLOGRAPHIC DATA

B.L. de Groot, S. Hayward, D.M.F van Aalten, A. Amadei and H.J.C. Berendsen

PROTEINS: Struct. Funct. Gen. 31: 116-127 (1998)

Summary

A comparison of a series of extended Molecular Dynamics simulations of bacteriophage T4 lysozyme in solvent with X-ray data is presented. Essential Dynamics analyses were used to derive collective fluctuations from both the simulated trajectories and a distribution of crystallographic conformations. In both cases the main collective fluctuations describe domain motions. The protein consists of an N- and C-terminal domain connected by a long helix. The analysis of the distribution of crystallographic conformations reveals that the N-terminal helix rotates together with either of these two domains: the main domain fluctuation describes a closure mode of the two domains in which the N-terminal helix rotates concertedly with the C-terminal domain, while the domain fluctuation with second largest amplitude corresponds to a twisting mode of the two domains, with the N-terminal helix rotating concertedly with the N-terminal domain. For the closure mode, the difference in hinge-bending angle between the most open and most closed X-ray structure along this mode is 49 degrees. In the MD simulation that shows the largest fluctuation along this mode, a rotation of 45 degrees was observed. Although the twisting mode has much less freedom than the closure mode in the distribution of crystallographic conformations, experimental results suggest that it might be functionally important. Interestingly, the twisting mode is sampled more extensively in all MD simulations than it is in the distribution of X-ray conformations.

Introduction

The notion of domain motions in hen lysozyme, inferred from its X-ray structure[125, 126], is more than twenty years old[127]. Although bacteriophage T4 lysozyme (T4L) has a very different structure, the domain character of the protein is even more pronounced[128]. From the differences between crystallographic structures of various mutants of T4L it has been suggested that a hinge-bending mode of T4L is an intrinsic property of the molecule[129–131]. This hypothesis was recently qualitatively supported by studies of T4L in solution[132]. Also from computer simulations domain motions of the wild-type protein have been observed[52, 133]. The domain fluctuations are predicted to be essential for the function of the enzyme, allowing the substrate to enter and the products to leave the active site. Crystallographic studies of a mutant T4L[134] in which a substrate is covalently bound to the enzyme, suggest that the substrate-bound enzyme is locked in a state in which the two domains have closed around the substrate with respect to the unbound state. The unbound enzyme is expected to display a larger hinge-bending angle on average.

More than 200 T4L structures crystallised in more than 25 different crystal forms are present in the Protein Data Bank[131]. Assuming that each crystal structure represents a possible conformation in solution, this provides a unique experimental view on the conformational flexibility of the protein at atomic resolution. Information on conformational freedom of proteins is usually obtained from only a few experimental structures[135–137] but dynamics of proteins is so complex that these few structures give only an extremely limited view of the dynamics involved. For T4L, the comparatively large number of different experimental conformations should provide us with a more detailed picture of its dynamical behaviour which can then be sensibly compared to an MD simulation[87]. This provides the opportunity to assess the reliability of MD simulations.

T4L is a good system to study, not only for its large number of X-ray conformers but also because it is a rather small domain protein suitable for MD simulation. As domain proteins are usually relatively large, only few MD studies have been published in which domain motions were extensively studied[52, 55, 84, 138–141].

In this study, a detailed comparison is made between the collective (domain) fluctuations in T4L derived from the distribution of X-ray structures and from extensive MD simulations in solvent. Three simulations were conducted, each of one nanosecond, starting from different experimental structures. The Essential Dynamics (ED) analysis[78] was applied both to the distribution of X-ray and MD structures to separate small-amplitude fluctuations from large-amplitude global fluctuations. The largest-amplitude collective fluctuations from the X-ray distribution and from MD were subjected to domain and hinge-bending analyses[55, 56] to monitor domain fluctuations. Collective fluctuations derived from MD can be expected to be affected by limited sampling[109, 110, 142] or imperfections in the inter-atomic interactions or force field. On the other hand, the crystallographic structures may not be representative of solvent-accessible conformations for the wild-type as they may be affected by the different mutations or by crystallisation conditions and/or crystal contacts[131]. Despite these reservations a good correspondence between the MD results and X-ray analysis is obtained. Additionally, the detailed analyses of the domain fluctuations in T4L reveal interesting dynamical aspects that may be important for the function of the protein.

Methods

MD simulations

Three simulations were performed, each of one nanosecond. The first simulation, of the wild-type protein, started from a high-resolution X-ray structure[143] (PDB entry 2LZM). This simulation will from now on be referenced to as WT. The second simulation (M6I) was of the mutant M6I (methionine 6 replaced by isoleucine) and started from the X-ray structure with largest hinge-bending angle of this mutant[129] (PDB entry 150L, hinge-bending angle 31 degrees more open than the WT X-ray structure). The coordinates of the three C-terminal residues not present in this crystal structure were taken from the most closed conformation from the same PDB entry. The third simulation started from the same structure, now mutated back to the wild type (WT*). All simulations were performed in a periodic box filled with SPC[98] water molecules (also crystallographic water molecules were included). Polar and aromatic hydrogens were added to the protein. In each of the simulated systems, 8 Cl^- ions were added to compensate the net positive charge on the protein. These ions were introduced by replacing water molecules with the highest electrostatic potential. This added up to a total of 19195 atoms for the WT simulation and 17101 for the M6I and the WT* simulation. Prior to the simulations, the structures were energy-minimised for 100 steps using a steepest-descents algorithm. Subsequently the structures were simulated for 10 ps with a harmonic positional restraint on all protein atoms (force constant of 1000 kJ mol^{-1} nm^{-2}) for an initial equilibration of the water molecules. Production runs of 1 ns started from the resulting structures. All simulations were run at constant volume. The temperature was kept constant at 300 K by weak coupling to a temperature bath[99] ($\tau = 0.1$ ps). A modification[144] of the GROMOS87[4] force field was used with additional terms for aromatic hydrogens[6] and improved carbon-oxygen interaction parameters[144]. SHAKE[11] was used to constrain bond lengths, allowing a time step of 2 fs. A twin-range cut-off method was used for non-bonded interactions. Lennard-Jones and Coulomb interactions within 1.0 nm were calculated every step, whereas Coulomb interactions between 1.0 and 1.8 nm were calculated every ten steps.

All simulations were performed with the GROMACS simulation package[5].

Analysis techniques

Apart from conventional structural and geometrical analyses to assess the stability of the structures during the simulation, ED[78] analyses were utilised to study large concerted motions. The method yields the directions in configurational space that best describe concerted atomic fluctuations and is related to principal component analysis and quasi-harmonic analyses[73,76,77,94,145]. It consists of diagonalisation of the covariance matrix of atomic fluctuations, after removal of overall translation and rotation. Resulting eigenvectors are directions in configurational space that represent collective motions. Corresponding eigenvalues define the mean square fluctuation of the motion along these vectors. The method can be applied to any (sub)set of atoms using any set of structures[78].

An ED analysis was performed on a cluster of X-ray crystallographic structures. Only structures from different crystal forms were included in the analysis. Zhang *et al.* [131] described 25 different crystal forms. From their list, a set of 21 pdb entries was constructed, including 38 structures. These entries include 149L[146], 152L[146–148], 169L[131], 172L[147], 176L[131], 179L[147], 2LZM[143], 137L[149,150], 150L[129], 167L[131,147], 170L[131], 173L[131], 177L[131], 1L97[130], 151L[131,146], 168L[131], 171L[131], 174L[131], 178L[131,147], 216L[149] and 148L[134].ED analyses were performed on the cartesian coordinates of the main chain N, C-α and C coordinates. Residues 163 and 164 were excluded from the analysis because their coordinates were absent in many of the pdb entries.

The same atoms were used in the ED analyses of the MD simulations. Analyses were performed on each individual MD trajectory (as the potential energies appeared to stabilise in less than 100 ps, the first 100 ps of each trajectory were disregarded) and on a combination of the three simulations. In this combination, the three simulations were not simply concatenated, because the eigenvectors would then be influenced by the differences between the average (starting) structures of each simulation. To remove the bias caused by these static differences, only the fluctuations from the average structure in each simulation were taken into account. This analysis implies the approximation that there are no systematic differences between the individual simulations. This combined analysis will be referenced to as MD_ALL.

ED analyses were carried out using the WHAT IF program[101]. Domains and hinge axes were identified and characterised using the DYNDOM program[55,56]. The method analyses conformational changes in terms of rotational properties. Dynamic domains are identified by clustering each residue's rotation vector in a particular collective mode of motion.

Results

Figure 5.1 shows the root mean square deviation (RMSD) during the three free simulations with respect to the WT X-ray structure and to the most open M6I X-ray structure. Deviations from the respective starting structures are relatively large, suggesting large structural fluctuations. The difference between the two starting structures (0.26 nm) is approximately as large as the drifts from the starting structures in each simulation.

Atomic fluctuations in the set of of X-ray structures were compared to the crystallographic B-factors averaged over the 38 experimental structures and to the atomic fluctuations calculated from the MD simulations (Fig. 5.2). There is poor correspondence between the average B-factors and the atomic fluctuations in the distribution of X-ray structures (correlation coefficient of 0.55) but there is good correspondence between the atomic fluctuations in the X-ray and MD distribution (correlation coefficient of 0.85).

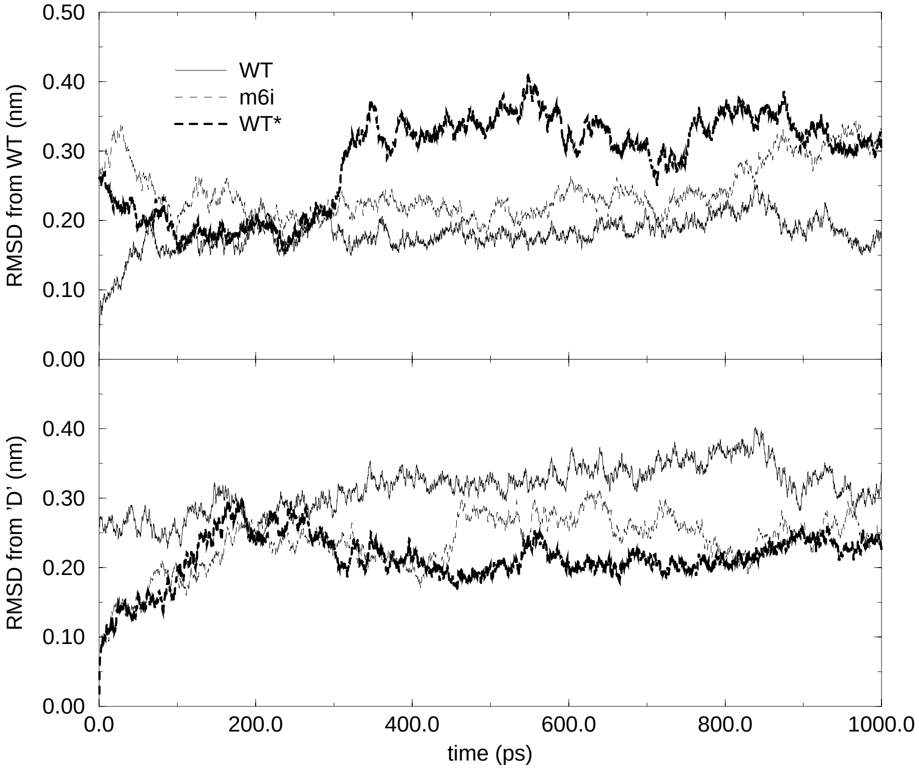

Figure 5.1 Root mean square deviatation of C-α atoms from the WT X-ray structure (upper panel) and from the most open M6I ('D') X-ray structure (lower panel).

Fig. 5.3. shows the eigenvalues of the ED analyses of the set of X-ray structures and of the combination of the three MD simulations (MD_ALL). The eigenvalue curve is very steep in the X-ray analysis with the first eigenvector contributing 86 % to the total mean square fluctuation. For MD_ALL, the eigenvalue curve is less steep and therefore more eigenvectors are required to achieve the same level of approximation of the total mean square fluctuation.

The domain analysis[55, 56] was performed on the motions along single eigenvectors to ascertain whether these main modes of correlated fluctuation correspond to domain motions. Table 5.1 and Fig. 5.4 show that the two most dominant of these modes extracted from the distribution of X-ray structures clearly correspond to the motion of two quasi-rigid bodies with respect to each other[56]. For both modes there are two distinguishable domains. The C-terminal domain is largest and ranges from approximately residue 75 to the C-terminus. The smaller N-terminal domain ranges from approximately

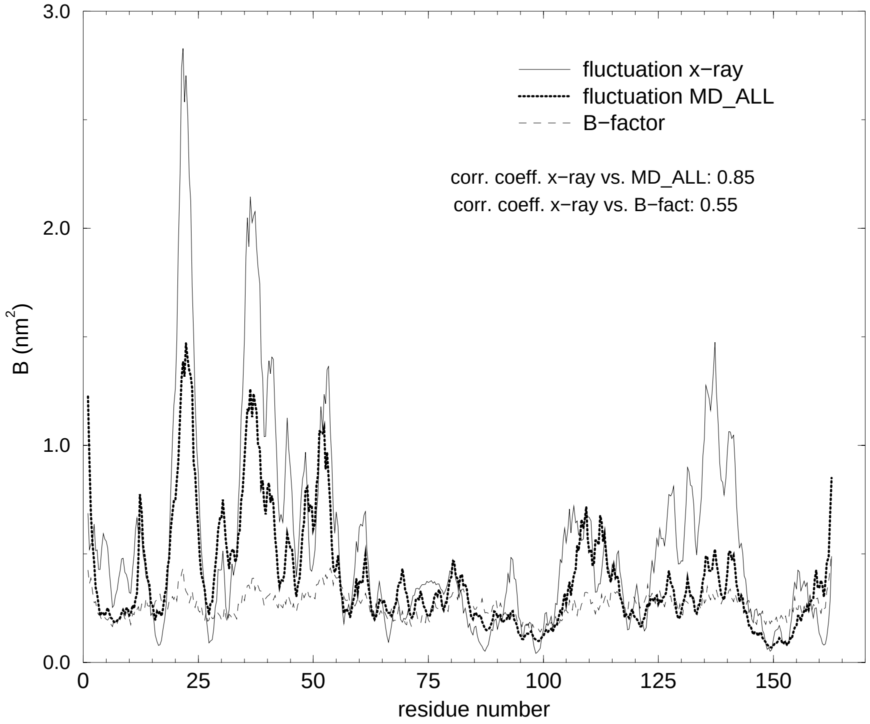

Figure 5.2 Atomic fluctuations (expressed in isotropic B-factors; $B = (\Delta r)^2 \times 8\pi^2/3$, with $(\Delta r)^2$ being the calculated atomic mean square fluctuation) of main chain atoms in the X-ray cluster compared to the B-factors averaged over the 38 crystal structures and to the atomic fluctuations averaged over the three MD simulations.

residue 13 to 65. The first ten N-terminal residues are not statically part of the N- or C-terminal domain, but fluctuate correlated with either of the two domains: with the C-terminal domain in the first eigenvector and with the N-terminal domain in the second. The transition between the N- and C-terminal domains is located between residues 65 and 75, in the middle of the inter-domain helix. The flexible link between the first ten residues and the N-terminal domain consists of residues 11 and 12.

The assignment of residues to the domains given above was used to extract the axes around which the domains rotate with respect to each other. The calculated inter-domain screw-axes are shown as arrows in Fig. 5.4 for the first and second eigenvectors from the ED analysis of the X-ray cluster. Both axes are "effective hinge axes"[56] as they pass near the residues shown to be involved in the inter-domain motion (see table 5.1). The first eigenvector corresponds (mainly) to a closure motion[55] (defined by an effective hinge axis perpendicular to the line connecting the centers of mass of the two domains)

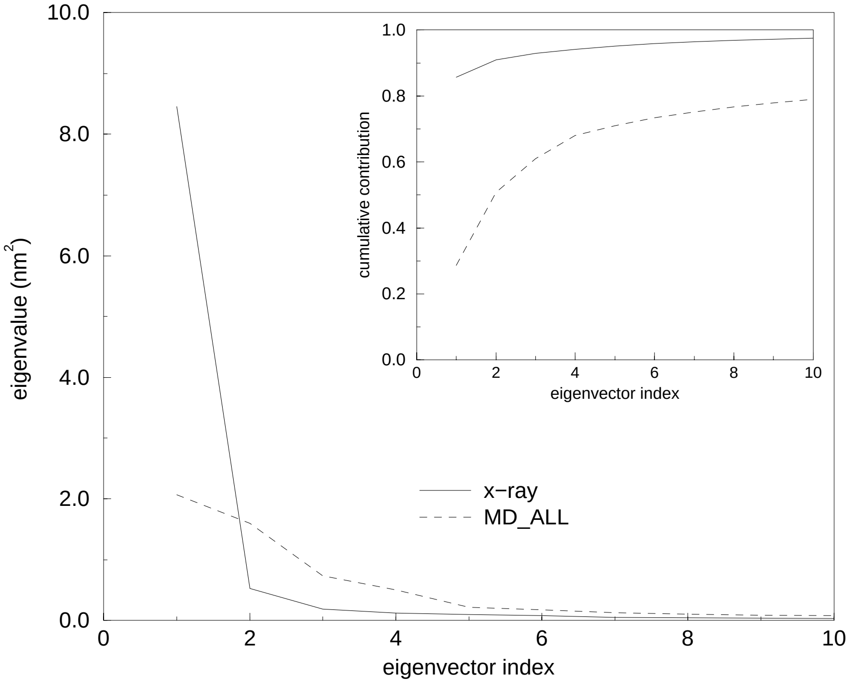

Figure 5.3 Eigenvalues obtained from the Essential Dynamics analyses of the cluster of X-ray structures and of the combination of MD simulations. The inset shows the cumulative contribution of the eigenvectors to the total mean square fluctuation.

	X-ray, e.v. 1	X-ray, e.v. 2	MD_ALL, e.v. 1	MD_ALL, e.v. 2
domain A	14-66	1-65	15-63	12-66
domain B	1-10, 81-162	74- 162	1-12, 75-162	70-162
connecting regions	11-13, 67-80	66-73	13-14, 64-74	1-11, 67-69
angle of rotation	47.1	16.0	39.4	34.5
residues near axis	12,13,29,71-76	6,7,49,50,66,67	13,29,59,102	12,67,69,70
angle ‖	66.5	37.1	29.2	75.3
% closure motion	84.1	36.5	23.9	93.5

Table 5.1 Domain analyses of the two modes with largest amplitudes from X-ray and MD_ALL. Residues were marked near to the effective hinge axis if their $C\alpha$ atoms were found within 3 Å of the axis. Angle || denotes the angle between the effective hinge axis and the line connecting the two centers of mass of the two domains.

(table 5.1, Fig. 5.4A). The angular difference between the most open (PDB entry 178L[147]) and the most closed configuration (PDB entry 152L) is as much as 47 degrees (table 5.1). From the clustering of the endpoints of the rotation vectors in Fig. 5.4A it is visible that the ten N-terminal residues rotate together with the C-terminal domain. The second eigenvector consists (mainly) of a twisting of the two domains, with the effective hinge axis being more parallel to the line connecting the two centers of mass. (table 5.1, Fig. 5.4B). From the clustering of the atoms in Fig. 5.4B it can be seen that the first ten residues now rotate more concertedly with the N-terminal domain.

Domains were identified also from the first two modes of the MD_ALL analysis (table 5.1, Fig. 5.5) and there is good correspondence with the domain demarcation obtained from the X-ray analysis. Again, residues 11-14 and 65-80 form the dynamical links between the two domains. The dynamic behaviour of the N-terminal helix is less pronounced than in the analysis of X-ray structures. It is assigned to the C-terminal domain along the first mode (twist) and is identified as intermediate region along the second mode, which describes a closure mode.

All X-ray and MD structures were projected onto the plane spanned by the two eigenvectors with largest eigenvalue from the distribution of X-ray conformations to compare the kind and extent of fluctuation in the X-ray structures and MD (Fig. 5.6). All MD simulations fluctuate significantly in this plane, indicating that the main modes of collective fluctuation in the X-ray cluster are accessible during MD. This is in agreement with with previous findings[87]. There are differences between the regions sampled in this plane by X-ray and MD, however. The WT simulation shows a 25 degrees opening of the structure along the first X-ray eigenvector with respect to its starting configuration but does not reach any of the most open configurations observed in the X-ray cluster. The M6I simulation starts from a more open configuration and closes 29 degrees, reaching a hinge-bending angle almost

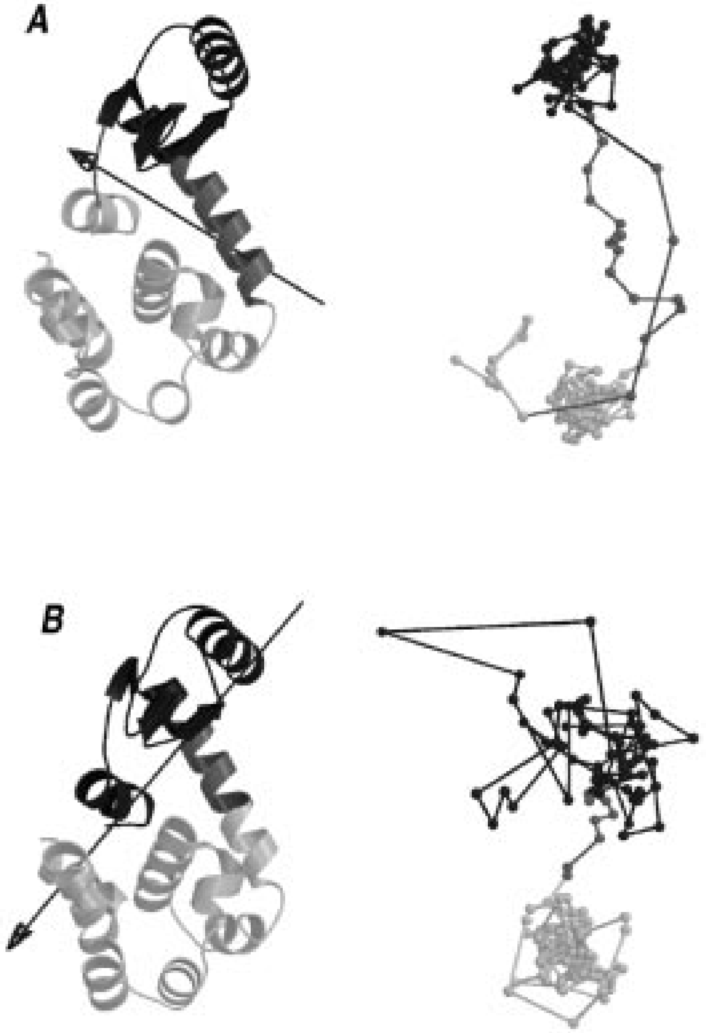

Figure 5.4 Backbone structure (left) and rotation vectors (right) of T4L. End points, depicted as beads, of rotation vectors per residue (the beads are connected as in the amino-acid sequence) were used to identify the domains. The grayscales indicate the different clusters (domains) that were assigned (light and dark residues) and the inter-domain regions (intermediate grayscale) based on the rotation vectors. The arrow indicates the direction of rotation of the light domain relative to the dark domain by the thumb rule of the right hand. Panel A: Eigenvector 1 from the X-ray cluster. Displayed is the most open conformation with the arrow indicating the closure motion. Panel B: Eigenvector 2 from the distribution of X-ray conformations. Domain analyses were performed by DYNDOM[56]. Plots were made with MOLSCRIPT[151] and Raster3D[152,153].

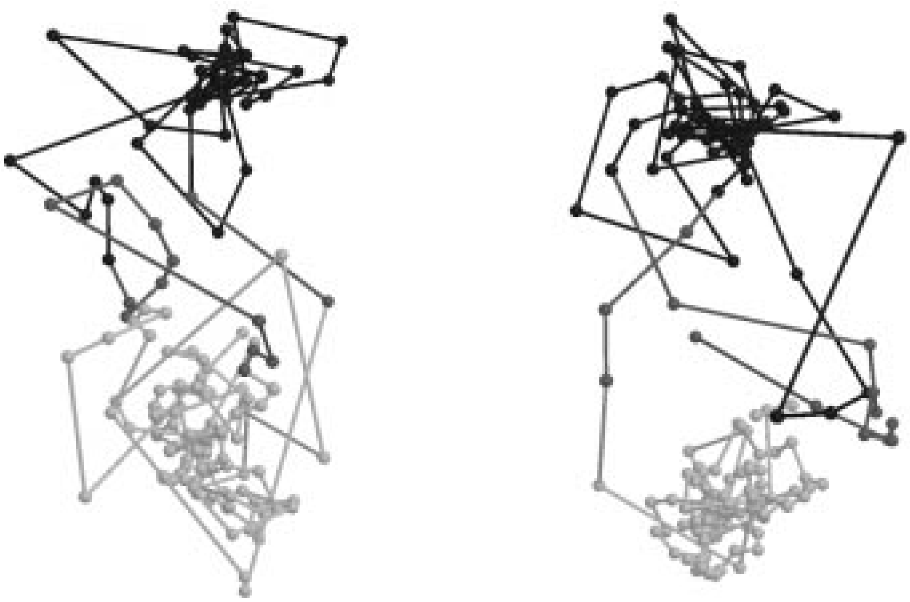

Figure 5.5 Domain identification from the first (left) and second (right) mode from MD_ALL. As in Fig. 5.4, grayscales indicate the different clusters (domains) that were assigned (light,dark) and the inter-domain regions (intermediate gray) from the rotation vectors. Domain analyses were performed by DYNDOM[56]. This plot was made with MOLSCRIPT[151] and Raster3D[152,153].

equal to that of the WT X-ray structure. Both the M6I simulation and the WT simulation spend most of the time at a hinge-bending angle between 7 and 19 degrees more open than the WT X-ray structure. The WT* simulation initially closes and also reaches a conformation similar to that of the WT X-ray structure. After that it opens up again and reaches a conformation with a hinge-bending angle 45 degrees more open than that of the WT X-ray structure, slightly more open than the X-ray structure with largest hinge-bending angle. Along this first eigenvector there seem to be two distinct clouds in the cluster of X-ray structures with only two configurations in between. This is consistent with a two-state mechanism postulated on the basis of these structures[131]. The simulations do not support this hypothesis, however, and indicate that intermediate structures are equally accessible.

The position along the second X-ray eigenvector, which mainly describes a twisting mode, fluctuates uncoupled from the position along the first eigenvector, both in the X-ray cluster as well as in the three MD simulations. The amplitude of the fluctuation in this direction is larger in each of the three simulations than in the cluster of crystal structures.

Table 5.2 lists inner products between eigenvectors obtained from the cluster of X-ray structures and those obtained from MD. This provides a quantitative measure of the overlap in modes of motion derived from the two techniques. Table 5.2a shows that the two eigenvectors with largest eigenvalue from the X-ray analysis are to a large extent present in the space spanned by the first five eigenvectors obtained from each simulation. This means that the modes of domain motion extracted from the differences between the X-ray conformations are also among the most dominant ones in the simulations. It is interesting to note that the overlap between eigenvectors extracted from the

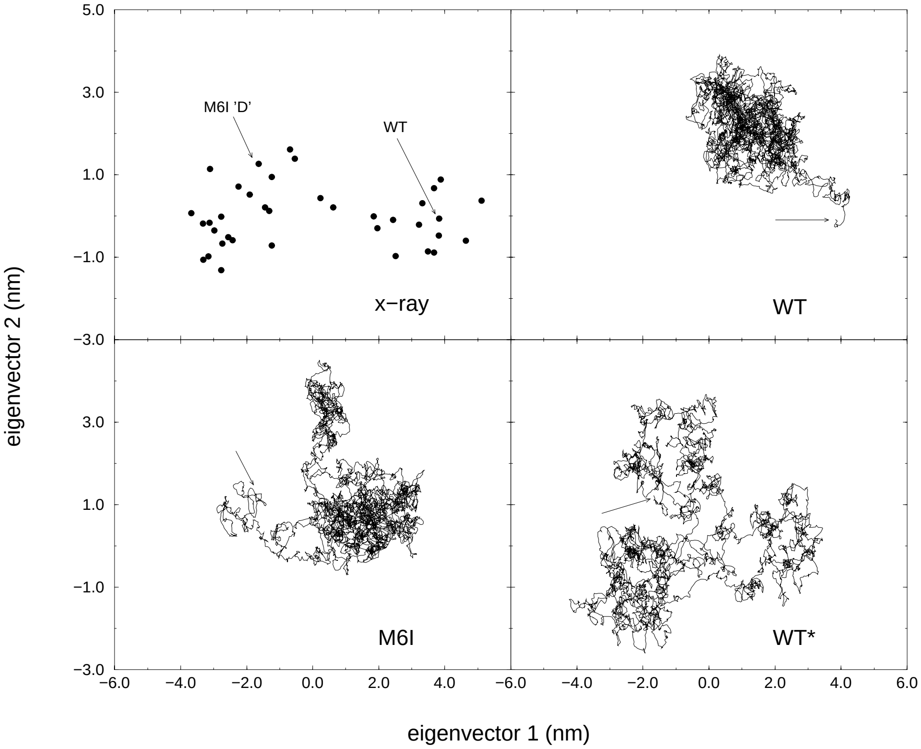

Figure 5.6 Projections (in nm) onto the plane spanned by the two eigenvectors with largest eigenvalues extracted from the cluster of X-ray structures. Upper left panel : X-ray structures; upper right panel : structures from the WT simulation; lower left panel : structures from the M6I simulation; lower right panel : structures from the WT* simulation. The arrows indicate the starting structures of each simulation. In the horizontal direction, structures differ from each other along the closure mode (structures to the left are more open than those on the right); the vertical direction depicts the twisting mode.

combination of the three MD simulations (MD_ALL) and the X-ray eigenvectors is larger than the average of the overlaps between the X-ray eigenvectors and those extracted from each of three simulations individually. When the MD simulations are compared to each other in the same fashion (table 5.2b), the overlap is on average lower than with the X-ray structures (table 5.2a). Therefore, the main modes of motion derived from each of the MD simulations are more similar to the main collective fluctuations derived from the X-ray cluster than to those from the other MD simulations.

In order to compare with the qualitative results of Mchaourab *et al.* [132], fluctuations of the distances between selected pairs of α-carbon atoms were monitored along the two most prominent modes of collective fluctuation derived from the cluster of X-ray structures (Table 5.3). The pairs were selected to study the difference in conformation between the protein free in solution and covalently bound to a substrate. The fluctuations of the distances between pairs 35-137, 22-137, 4-71 and 4-60 are mainly ruled by

a.

	X-ray eigenvector	
e.v. 1-5	1	2
WT	0.72	0.69
M6I	0.80	0.81
WT*	0.91	0.76
MD_ALL	0.92	0.77

b.

	WT eigenvector	
e.v. 1-5	1	2
M6I	0.34	0.40
WT*	0.61	0.45

	M6I eigenvector	
e.v. 1-5	1	2
WT	0.65	0.60
WT*	0.58	0.66

	WT* eigenvector	
e.v. 1-5	1	2
WT	0.64	0.54
M6I	0.85	0.76

Table 5.2 a. & b. Summed squared inner products between one eigenvector of one set and the first five of another. a. MD eigenvectors compared to X-ray eigenvectors. b. MD eigenvectors from different simulations compared to each other.

	Spin labeling	X-ray
Pair	upon substrate binding distance	fluctuation affected by
35-137	decreases	closure
22-109	increases	twist (closure)
22-137	decreases	closure
4-71	decreases	closure
4-60	increases	closure
35-109	-	closure (twist)

Table 5.3 Fluctuation of distances between pairs of α-carbon atoms along the first (closure) and second (twist) collective mode of fluctuation derived from the cluster of X-ray structures (selection of the pairs after Mchaourab *et al.* [132]).

the fluctuation along the eigenvector with largest eigenvalue, describing a closure motion. The observed spin-spin interactions[132] are consistent with a shift along the closure mode (towards closing) upon substrate binding. The distance between residues 35 and 109, however, hardly changes upon 'substrate release' although a fluctuation along the closure mode significantly influences the distance between this pair. The distance between residues 22 and 109 does change upon 'substrate release' but the fluctuation of the distance is much more connected with the twisting mode than with the closure mode, suggesting that substrate binding may also affect the twisting mode.

A web page has been dedicated to the visualisation of the dynamical information presented here:
`http://rugmd0.chem.rug.nl/~degroot/t4l.html`

Discussion and Conclusions

The collective fluctuations in T4L comprise, for the largest part, domain motions. The most dominant modes of fluctuation in the X-ray analysis as well as in all MD analyses correspond to external motions of the domains with respect to each other. Moreover, the main modes of fluctuation obtained from the cluster of X-ray structures are very similar to those obtained from simulation. The amount of overlap between X-ray and MD modes is larger than between modes of two similar MD runs. This is remarkable because it has been observed previously[108–110,142] that the definition of single eigenvectors in an ED analysis has not converged in simulations over time periods in the order of nanoseconds. A possible explanation for this phenomenon lies in the domain character of the protein, which causes two modes of domain motion to dominate over all other fluctuations: the domain fluctuations observed in the X-ray cluster are among the most extensively sampled directions in all MD simulations. The incomplete mutual overlap between MD modes is mainly due to insufficient sampling statistics, suggesting that longer MD simulations will show an even larger amount of overlap with the cluster of X-ray structures. The most important conclusion from the comparison of structural variability in X-ray and MD-generated structures is that MD indeed samples the important, physically relevant space, thus validating the MD method for application to protein dynamics.

The domain fluctuations in the MD simulations indicate that both the wild-type protein and the M6I mutant fluctuate significantly along the domain modes derived from the X-ray cluster. This is consistent with the hypothesis by Zhang *et al.* [131] that domain motions are an intrinsic property of the T4L molecule. The results by Mchaourab *et al.* [132] further support this finding. From the simulated data there is no evidence for the proposed two-state mechanism[131] for the main hinge-bending mode. The WT and M6I simulation do show a preference for intermediate hinge-bending angles for this mode (angles between 7 and 19 degrees more open than the WT X-ray structure) but the WT* simulation indicates that also more open configurations are easily accessible. Since there is no topological difference between the WT and the WT* simulation, a lack of sufficient sampling seems the most probable cause for the apparent difference between these simulations. Since also the differences between the M6I simulation and the WT and WT* simulations are not larger than the difference between the WT and WT* simulation, the conclusion that the hinge bending properties of the M6I mutant are close to those of the WT protein seems justified. This supports our assumption that the combination of the three MD trajectories for ED analysis (MD_ALL) is valid.

In a recent study, Arnold and Ornstein also presented results from MD simulations on native T4L and the M6I mutant[133]. They found that in all their simulations the protein went to a more compact conformation and concluded that a conformation more closed than the WT crystal structure would be the most stable configuration in solution. These findings are not supported by our results. We observe that in all simulations, the large majority of sampled conformations displays a more open conformation than the WT X-ray structure. A possible explanation of this apparent discrepancy is the difference between simulation protocols used. We have used a periodic box filled with a large number of solvent molecules (approximately 5000), whereas Arnold and Ornstein used a shell of solvent containing approximately 2200 water molecules. Protein dynamics in simulations using a shell of solvent molecules might be affected by surface tension effects in such small droplets, resulting in unrealistically compact structures. Since in both cases three simulations have been performed, with consistent results, limited statistics can probably be ruled out as a possible explanation for this observation. Interestingly, Arnold and Ornstein reported that the conformational change towards more compact structures did reveal the domain character of the protein, suggesting once again that domain motions are among the most prominent collective fluctuations of T4L.

The domain modes obtained from MD and the cluster of X-ray structures are essentially similar (Fig. 5.4, Fig. 5.5., Tables 5.1 & 5.2). The protein consists of two domains; an N-terminal domain comprising residues 15 to 65 and a C-terminal domain that ranges from residue 80 to the C-terminus. Residues 70-75, residing in the C-terminal half of the inter-domain helix, form the dynamical bridge between the two domains. The behaviour of the ten N-terminal residues is complex. In the main domain fluctuation derived from the X-ray cluster, mainly a hinge-bending mode describing a closure motion between the two domains, this N-terminal helix rotates concertedly with the C-terminal domain. Along the collective fluctuation with second largest amplitude however, which mainly consists of a twisting of the two domains, this helix appears to be part of the N-terminal domain. The two main modes of collective fluctuation obtained from MD basically form a linear combination of the first two modes from the X-ray cluster. Therefore, the dynamical behaviour of the N-terminal helix is influenced by both the N- and C-terminal domains in these modes and the assignment to either domain is less evident (Fig. 5.4, Table 5.1). Concluding, the N-terminal helix is not a static part of either of the two domains but rather adapts its dynamical behaviour to the kind of domain motion. Upon opening, contacts with residues 93-97 and the C-terminal residues push the N-terminal helix away from its original position. The flexible loop connecting it to the N-terminal domain (the rotation is concentrated around GLU11 and GLY12) allows it to move concertedly with the C-terminal domain along the closure mode. The absence of such a steric effect in the twisting mode causes the helix to move concertedly with the N-terminal domain in this mode.

The large amount of overlap between the domain fluctuations in the cluster of X-ray structures and the MD simulations is the main reason for the close agreement of the atomic fluctuations in both clusters (Fig. 5.2). The much smaller correlation between the fluctuations in the cluster of X-ray structures and the averaged B-factors, together with the significantly lower average level of the B-factors suggests that the main domain motions are significantly suppressed in most of the crystal environments included in this analysis. Although the pattern of thermal factors in some cases (especially those in 176L_A, 176L_B, and to a lesser extent also 2LZM (WT)) does suggest some degree of domain fluctuation[143], we can conclude that, at least for flexible proteins, B-factors may be a less reliable indication of motional freedom in solution than fluctuations derived from MD.

Apart from the similarities between the fluctuations in MD and the X-ray cluster, there are also a few discrepancies. One of the most striking differences is in the shapes of the eigenvalue curves (Fig. 5.3). For the X-ray cluster there is one dominating collective fluctuation (the closure mode) which accounts for 86 % of the total fluctuation, and the first ten eigenvectors together represent 98 % of the fluctuation. In MD, the first mode only contributes 29 % to the total fluctuation and the first ten together represent 79 %. This is not the result of the fact that there far fewer structures present in the X-ray cluster (38) than in the MD cluster (27.000) (when a subset of 38 structures, equally spaced in time, is taken from the MD_ALL cluster, the first eigenvector contributes 32 % and the first ten eigenvectors 85 % to the total fluctuation). This indicates that in the MD, a larger number of collective fluctuations than in X-ray make a significant contribution to the total fluctuation. The difference in sampled regions in the two main directions from the X-ray cluster is illustrated by Fig. 5.6. Both the WT and M6I simulations do not sample the complete range of hinge-bending angles along the main closure mode derived from the X-ray cluster. The WT* simulation, however, indicates that this is the result of limited sampling, since in this simulation almost the complete range that is present in the X-ray cluster is sampled in one nanosecond. For the eigenvector with the second largest eigenvalue derived from the cluster of X-ray structures, the twisting mode, the fluctuation in all three simulations is larger than in the X-ray cluster (Fig. 5.6). Limited sampling in MD cannot be the explanation for this observation since this direction is oversampled with respect to the X-ray cluster. Also, the effect of mutations in the cluster of X-ray structures is not likely to be the reason for this discrepancy since one could expect the mutations to result in a larger fluctuation rather than smaller, with respect to the WT protein. If one assumes that in 25 different crystal forms all conformational freedom has been sampled, then only the effect of crystallisation conditions and/or crystal contacts or the used force field in MD remain as possible explanations for this difference. Further studies (e.g. NMR) will be necessary to distinguish which is the main effect.

The investigation of the fluctuation of distances between selected pairs of α-carbon atoms (Table 5.3) shows that for four out of the six investigated pairs, the experimentally observed changes in distances in solution are in accordance with an opening along the closure mode upon transition from the substrate-bound state to the substrate free state. The fluctuation of the distance between residues 22 and 109, which is found to change upon 'substrate release' is more connected with the twisting mode than with the closure mode, however. This suggests that also the twisting mode is affected by the presence of the substrate. Another distance, between residues 35 and 109, does not seem to change much upon 'substrate release' but is affected substantially by the closure mode. A possible explanation for this observation is a partial compensation by a change along the twisting mode, which also makes a significant contribution to the fluctuation of this distance. This is a further indication that not only the closure motion but also the twisting mode is relevant for the function of this protein and that the two modes are concertedly involved in the dynamics of substrate binding. Interestingly, all MD simulations display a larger extent of fluctuation along the twisting mode than is observed in the cluster of X-ray structures (Fig. 5.6).

In summary, we conclude that T4 lysozyme exhibits a mixture of two hinge-bending modes (a closure and a twist) which are both involved in the dynamic response to substrate binding. Furthermore, we have shown that MD simulations of this protein provide reliable predictions of its functional dynamics.

Acknowledgement

A portion of this work was supported by the Supercomputing Resource for Molecular Biology at the European Molecular Biology Laboratory funded by a European Union Human Capital and Mobility Access to Large Scale Facilities grant (ERBCHGECT940062). S.H. is a recipient of a BIOTECH fellowship from the European Union.

6 PREDICTION OF PROTEIN CONFORMATIONAL FREEDOM FROM DISTANCE CONSTRAINTS

B.L. de Groot, D.M.F. van Aalten, R.M. Scheek, A. Amadei, G. Vriend and H.J.C. Berendsen

PROTEINS: Struct. Funct. Gen. 29: 240-251 (1997)

Summary

A method is presented that generates random protein structures that fulfill a set of upper and lower inter-atomic distance limits. These limits depend on distances measured in experimental structures and the strength of the inter-atomic 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.

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 have 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[154,155]. 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[109,110]. The efficiency of MC calculations is comparable to that of MD due to the presence of internal barriers[156].

Essential Dynamics

Essential Dynamics (ED),[53,78,84,85] equivalent to Principal Component[76,94] 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[84,85,157]. 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[81,108], the definition of individual eigenvectors spanning this subspace from a simulation has not converged in the simulated time[109,110], but the definition of the subspace itself approximately has[90,142].

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[81, 107, 108].

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[158, 159].

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[81, 107, 108].

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 alpha spectrin SH3 domain, HPr from *E. 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 fulfill a set of distance constraints. The first step is to measure all pairwise inter-atomic 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 6.1 and distance limits D are given in table 6.2. 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 6.2 for the distribution of distances over the different classes).

To speed up the search for structures that fulfill all distance criteria, upper and lower bounds are defined for all atom pairs that are not explicitly mentioned in table 6.1. The range of freedom D given to these pairs (0.5 nm) is much larger than for all other pairs (table 6.2) (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[95], 1shg[160], 3cln[161] and 2lzm[143] respectively). For HPr, a snapshot from an equilibrated MD simulation[120] (initiated from the NMR structure with pdb entry 1hdn[93]) was used to extract the distances. All structures were energy minimised using the GROMOS[4] force field before distances were calculated. All non-polar hydrogen atoms were included within united carbon atoms (except for aromatic hydrogens in the case of lysozyme). Polar hydrogens were placed using standard GROMOS/GROMACS 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).

The values given in table 6.2 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

1-3	
Rings	
1-4	pairs of backbone atoms that are part of the same secondary structure element (helix or strand) and are not more than four 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 90^0.
Tight hydrophobic	pairs of atoms between which the inter-atomic 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 inter-atomic distance is smaller than the sum of the Van der Waals radii of the involved atoms plus 1.0 Å.

Table 6.1 Different classes of interacting pairs.

			pgb	SH3	HPr	cal	lys
nr of atoms			535	583	785	1364	1703
nr.	type	D (nm)	nr 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					

Table 6.2 Parameters used in the CONCOORD method. 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.

distances are 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 fulfill 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 inter-atomic 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[162] 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).

For all proteins studied, 500 structures were generated with the CONCOORD method. For the IgG binding domain (56 residues), approximately 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 cut-off radius for inter-atomic 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 minimise 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 diagonalisation of the (mass-weighted) Hessian matrix, as in Normal Modes (NM) analyses[70, 71, 163]. 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[90, 164], although normal modes have the restriction of harmonicity.

Starting from the same distance bounds, diagonalisation 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[165] 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[78] analyses were used for comparison of structural freedom in proteins. The method consists of diagonalisation of the covariance matrix C of atomic fluctuations, after removal of overall translation and rotation:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{6.1}$$

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[78, 85, 90] 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[101] package. ED and all other structural analyses were performed using an interface in the molecular modeling package WHAT IF[101]. Secondary structure analyses and accessible surface calculations were performed with DSSP[102]. Dihedral angle criteria were taken from PROCHECK[123].

Results

All CONCOORD structures were subjected to a number of structural analyses to assess how physically realistic the generated structures are (table 6.3). The same analyses were performed on structures sampled by MD (for simulation details: the IgG-binding domain[142] (1 ns), SH3[53] (1 ns), HPr[120] (300 ps), calmodulin[141] (500 ps) and lysozyme (to be published, 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 Fig. 6.1 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 Fig. 6.2 (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

	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

Table 6.3 Average geometrical properties for structures generated by MD and CONCOORD, compared with the values obtained from experimental structures, for the five proteins studied. Abbreviations used: 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[102]), 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 unfavourable regions in Ramachandran plot[123,124]), QUAL (WHAT IF index indicating the normality of packing[166]). ENE (potential energy after energy minimisation in the GROMOS force field).

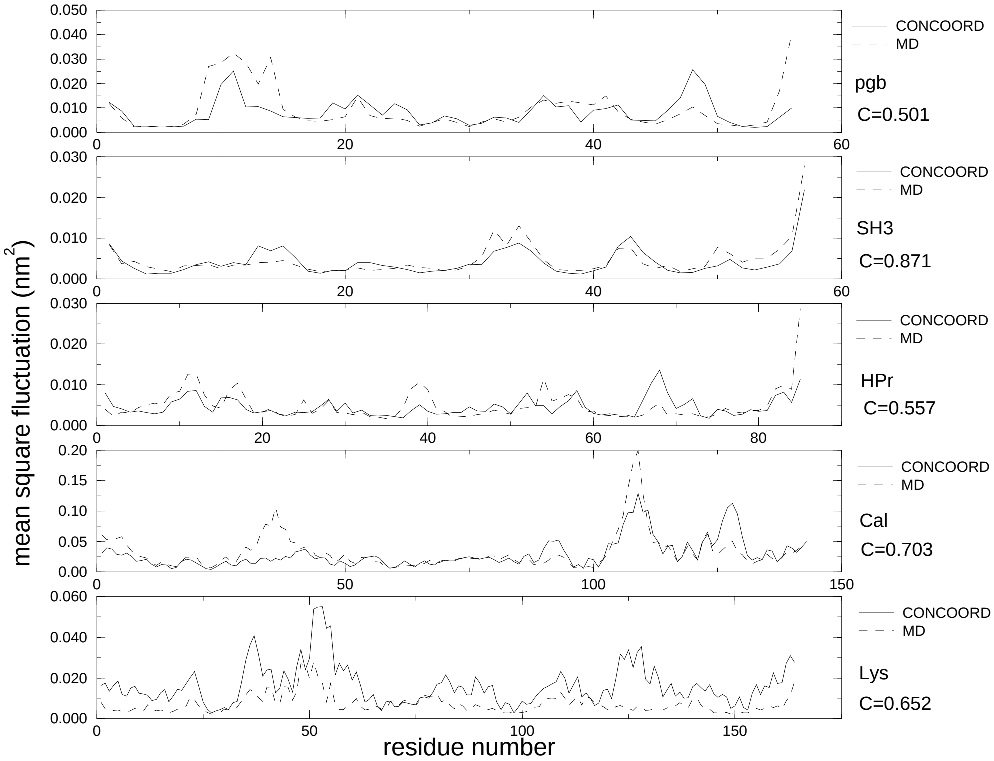

Figure 6.1 Mean square positional fluctuation of C_α atoms. The correlation coefficient between the curves obtained from MD and CONCOORD is shown next to the figures.

in Fig. 6.3A. 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, i.e. directions that show large fluctuations in MD also show relatively large fluctuations in CONCOORD, and vice versa. Fig. 6.3B shows the squared inner products between eigenvectors obtained from two halves of an MD simulation of 1 ns. The overlap between the two eigenvector sets from MD is similar to that between MD and CONCOORD. In Fig. 6.3C 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.

Fig. 6.3 shows that the overlap between MD and CONCOORD is especially high in the essential subspace (defined arbitrarily as the subspace spanned by the ten 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). Fig. 6.4 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 ten 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 ten eigenvectors with largest eigenvalues from MD and CONCOORD are given in table 6.4, for all proteins studied. The overlap between the essential subspaces obtained by MD and CONCOORD is comparable to the overlap obtained from two halves of each MD trajectory. A typical overlap of approximately 0.5 is obtained for

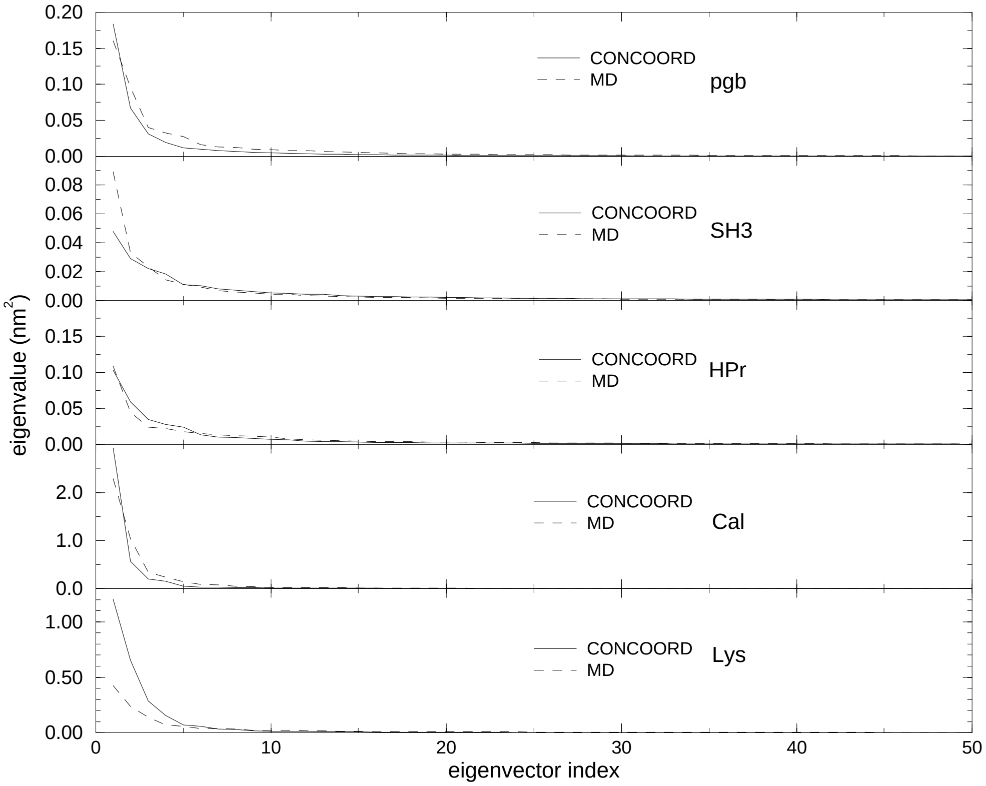

Figure 6.2 Eigenvalues obtained from MD trajectories and ensembles of structures generated by CONCOORD. Only the fifty largest eigenvalues are shown out of 168 (pgb, B1 IgG-binding domain), 171 (SH3), 255 (HPr), 429 (cal) and 492 (lys) respectively.

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 ten CONCOORD eigenvectors with largest eigenvalues with the ten lowest frequency-eigenvectors obtained from diagonalisation 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 analysis). This value is somewhat smaller than the overlap between eigenvectors obtained from two clusters of CONCOORD structures (0.866), indicating that small deviations

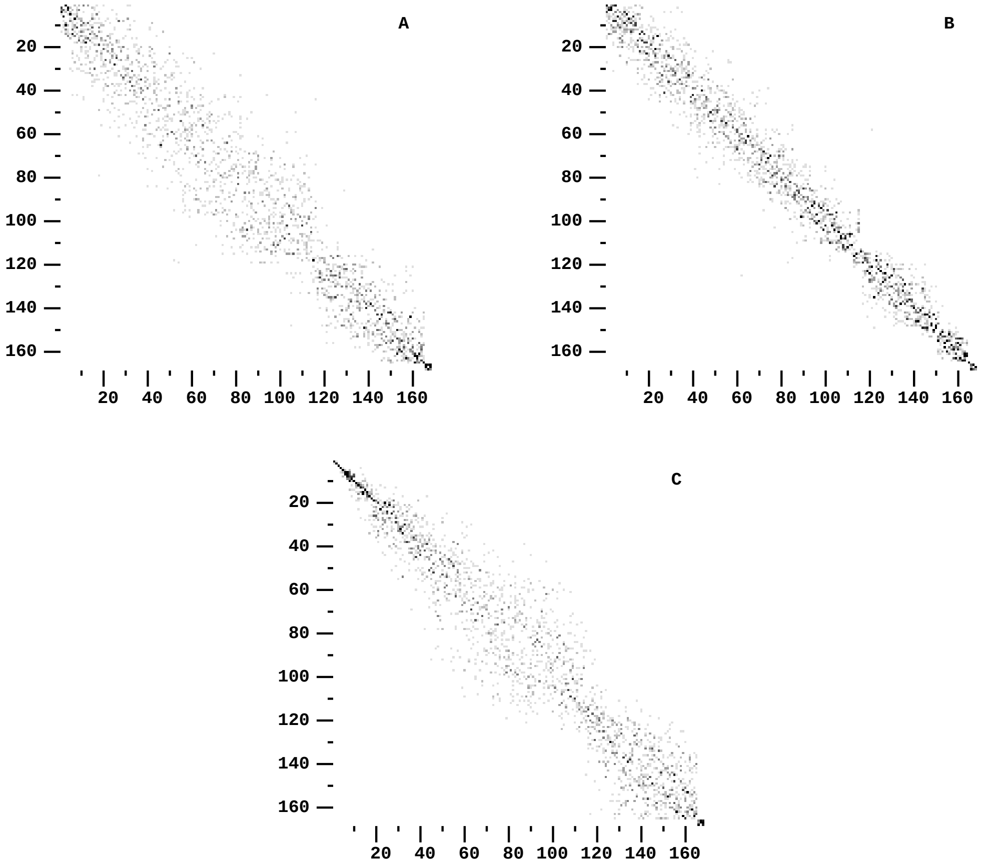

Figure 6.3 Squared inner product matrices for the B1 IgG-binding domain. In panel A eigenvectors from MD (1 ns, y-axis) are compared to those from CONCOORD (500 independent structures, x-axis). In panel B eigenvectors from two halves of the MD run (500 ps each) are compared to each other and the same is done for CONCOORD in Panel C (250 independent structures were used in each analysis).

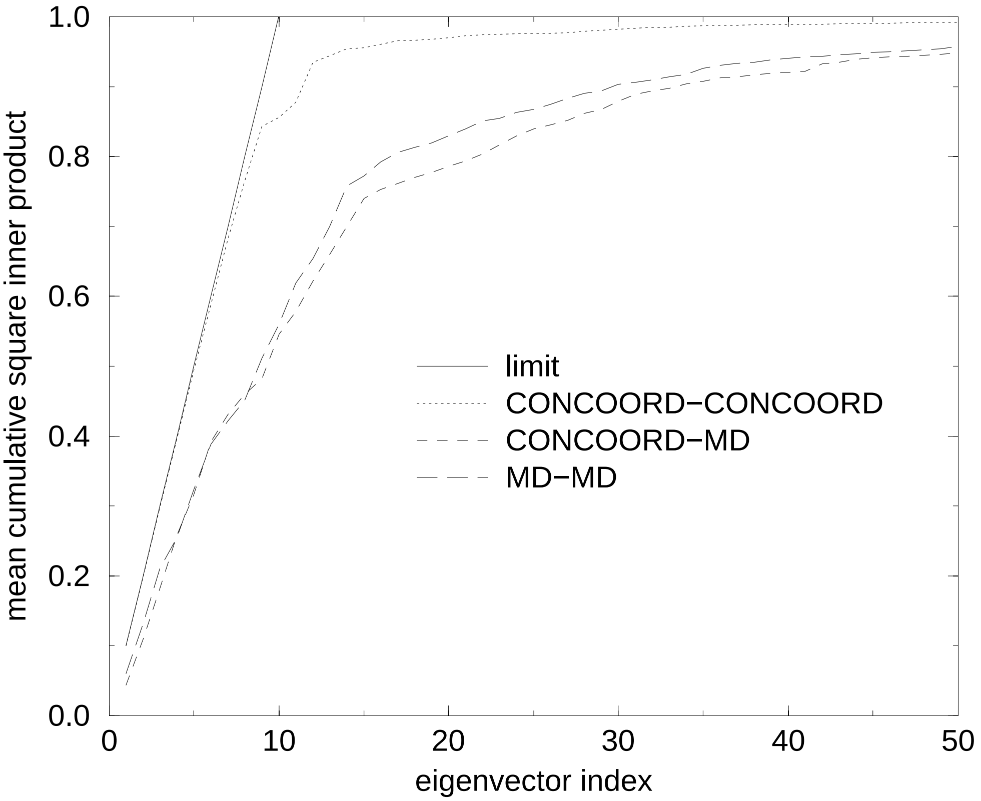

Figure 6.4 Cumulative mean square inner products between the 10 eigenvectors with largest eigenvalues obtained from MD/CONCOORD and all eigenvectors obtained from different MD/CONCOORD runs. After division by ten, all curves converge to 1.0 since every eigenvector from one set is contained in the complete set of vectors from another set. The solid line corresponds to the maximum obtainable overlap. pgb denotes the B1 IgG-binding domain.

protein	mean cumulative square inner product		
	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

Table 6.4 Mean squared inner products between subsets containing the ten eigenvectors with largest eigenvalues. 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.

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 Fig. 6.5. In MD (Fig. 6.5A), a single path is followed that resembles a random walk[81,107,108] whereas in CONCOORD (Fig. 6.5B), 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[81,107,108]. As can be seen from Fig. 6.5C, 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 Fig. 6.6 for calmodulin and lysozyme. The CONCOORD eigenvector with largest eigenvalue for calmodulin corresponds to a combination of a bend and a twist of the inter-domain helix, resulting in a rotation of one domain with respect to the other (Fig. 6.6A). From experiments (hydrogen exchange measurements[167], NMR relaxation data[168] and NMR NOE data[169] from which disorder in the set of NMR structures[170] emerged), the helix is known to break in the middle, which was also observed in MD and Normal Modes analyses[141].

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[131] (Fig. 6.6B). The main domain fluctuation consists of a rotation of the two domains with respect to each other, initiated by a combined twisting and bending of the inter-domain helix. The difference between the most open[131] and the most closed[147] X-ray structure along this rotation axis is as much as 49 degrees. The angular difference between the most open and most closed CONCOORD structure was 33 degrees; for MD this value was 28 degrees. 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 Fig. 6.1, 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 helix 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 fifteen 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 implementation, 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[109, 110, 142]. 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, i.e. 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 one nanosecond. 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 inter-domain 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 diagonalisation 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. Diagonalisation 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 6.2) 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 6.3 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). 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 values in table 6.4. In a recent study, a similar observation was reported[171] for BPTI. The amount of dynamic information that can be derived from NMR/NOE data has been subject of discussion. It has been argued[158, 159] 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 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 inter-atomic 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.

Acknowledgements

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 chapter.

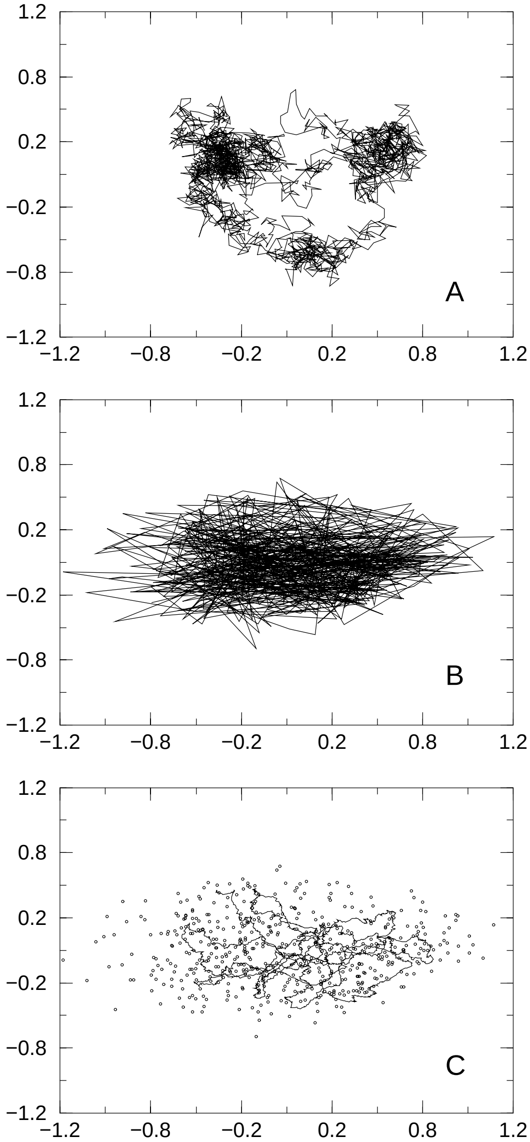

Figure 6.5 Projection of the MD trajectory of the IgG binding domain (panel A) and of the collection of CONCOORD structures (panel B) onto the planes defined by the two eigenvectors with largest eigenvalues from both techniques. The lower panel (panel C) shows a projection of CONCOORD (small circles) and extended MD (continuous line) structures onto the plane defined by the two CONCOORD eigenvectors with largest eigenvalues.

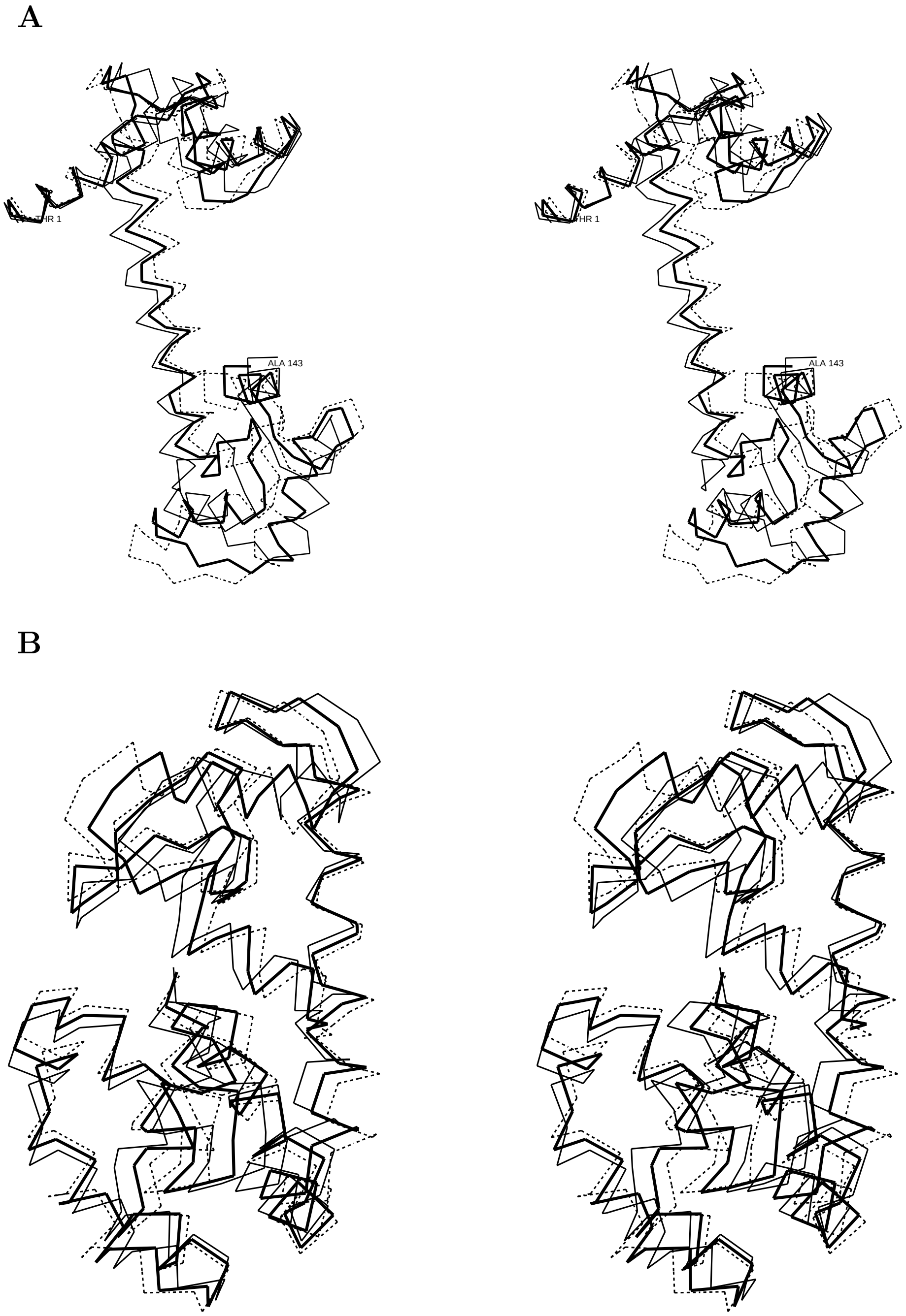

Figure 6.6 Stereo representation of extreme structures (thin line and thin dashed line) along CONCOORD eigenvectors, together with average structures (bold line). Panel A: calmodulin, eigenvector 1. Panel B: lysozyme, eigenvector 2.

7 CONFORMATIONAL CHANGES IN THE CHAPERONIN GROEL: NEW INSIGHTS INTO THE ALLOSTERIC MECHANISM

Bert L. de Groot, Gerrit Vriend and Herman J.C. Berendsen

submitted, September 1998

Summary

Conformational changes are known to play a crucial role in the function of the bacterial GroE chaperonin system. In this study, results are presented from an Essential Dynamics analysis of known experimental structures and from computer simulations of GroEL using the CONCOORD method. The results indicate a possible direct form of inter-ring communication, associated with internal fluctuations in the nucleotide-binding domains upon nucleotide and GroES binding, involved in the allosteric mechanism of GroEL. At the level of conformational transitions in entire GroEL rings, nucleotide-induced structural changes were found to be distinct, and in principle uncoupled from changes occurring upon GroES binding. Nucleotide-induced conformational changes are coupled to GroES-mediated transitions in simulations of GroEL double rings, but not in single rings. This provides another explanation for the fact that GroEL functions as a double ring system.

Introduction

The bacterial chaperonin GroEL and its cofactor GroES are among the best characterised molecular chaperones[172–174]. X-ray studies[175–178] combined with electron microscopy (EM) studies[179–183] have provided insight in the functional cycle of this chaperonin. GroEL is active as a double heptameric ring[184,185] with each ring containing a large central cavity in which substrate protein can be bound[179,186]. The cochaperonin GroES also exists as a heptamer and adopts a dome-like structure[187] that can bind to either GroEL ring to form a cap on the central cavity[179,188,189]. Fig. 7.1 shows the asymmetric crystal structure of GroEL with GroES bound to one GroEL ring[178], showing the packing of the subunits in the assembly, and the topology of each subunit.

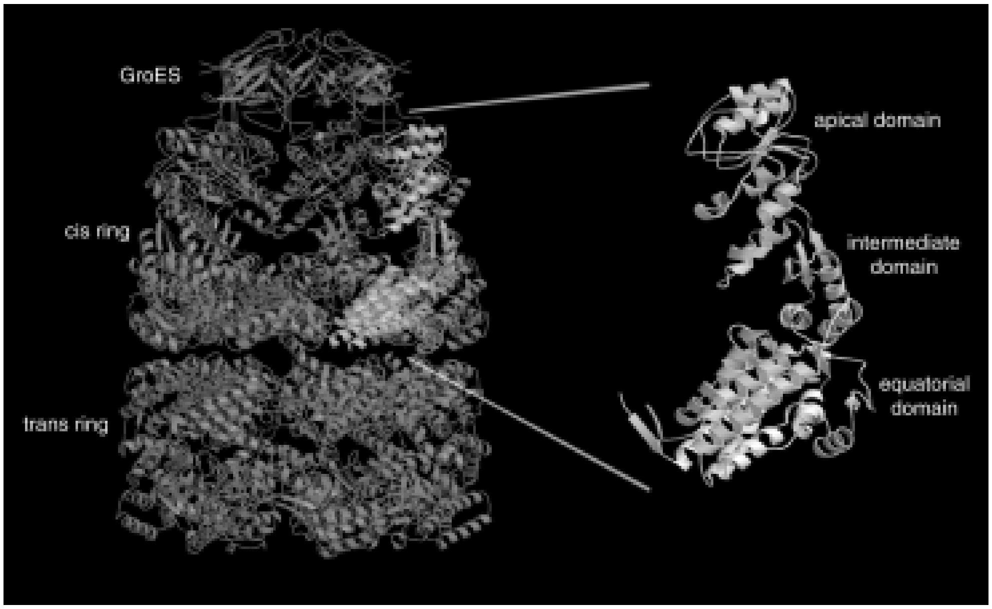

Figure 7.1 Left: the structure of the GroEL/GroES complex[178]. Right: the topology of a single GroEL subunit. Figure generated with Molscript[151,190] and Raster3d[191].

Each subunit of GroEL can be subdivided in three domains[175] (see Fig. 7.1). The equatorial domains form the backbone of the protein and contain an ATP binding site; they are involved in most intra-ring and all inter-ring subunit contacts. The apical domains are involved in interactions with substrate protein and GroES. The third domain, termed intermediate domain, forms the link between the apical and equatorial domains.

The role of GroEL in the substrate folding process is twofold. First, GroEL prevents substrate proteins from aggregating by binding unproductive folding intermediates and forces those to unfold to states more committed towards correct folding[192–196]. Second, it has been proposed that the central cavity works as an Anfinsen cage in which the substrate protein is actively folded[197, 198]. The dramatic conformational changes that are involved in the functional cycle of GroEL are indicative of a highly mobile system and stress the relevance of this flexibility for its biological activity.

GroEL is an allosteric protein. ATP binds cooperatively to the subunits of one ring[199–201], triggering a conformational change that reduces substrate affinity[202, 203] in the ATP bound ring. GroES binding to the ATP bound ring has been reported to complete this conformational change[182]. GroES binding switches the interior surface of the cavity from hydrophobic to hydrophilic, triggering a conformational change in the bound substrate molecule[178]. Negative cooperativity between rings[203–205] also results in a reduced GroES affinity in the ring opposite to the GroES bound ring. ADP binding to one of the rings does not impair ATP or GroES binding to the other ring[206], but ATP binding and hydrolysis in one ring has been proposed to play a role in GroES and substrate release from the other ring[207]. Communication between the two rings must be responsible for this effect, as is supported by the observation that a mutant that impairs dimer formation is defective in GroES release[197, 208, 209], thereby blocking bound substrates to leave the GroEL cavities. On the other hand, under different conditions (higher KCl concentration), productive folding has been observed in this single-ring mutant[210].

Despite the wealth of available experimental information, some aspects of the conformational changes and allosteric mechanism of GroEL remain unresolved. Knowledge of the mechanism underlying these conformational changes would greatly facilitate interpretation of a number of experimental results. Therefore, we have studied the conformational fluctuations in GroEL, with the hope to learn more about the mechanism(s) that govern these fluctuations. The most common method to study conformational fluctuations in proteins is Molecular Dynamics (MD), but with a molecular weight of 800 kDa it would be an impossible task to reach biologically relevant time scales when realistic force-fields are being used. A number of methods exists to speed up the efficiency of conformational sampling in MD[10, 211], and other computational techniques are also avialable. Ma & Karplus recently performed Normal Mode calculations on a minimal subsystem (three subunits) of GroEL that could provide insight into its allosteric mechanism[212]. We have chosen to use CONCOORD[89], a method to generate different protein conformations based on distance restrictions. This method has been shown to yield low frequency collective fluctuations for proteins that are very similar to those that can be extracted from MD simulations, but at a dramatically reduced computational expense[89]. To study the allosteric mechanism of GroEL, CONCOORD simulations have been performed of complete GroEL and GroEL/GroES assemblies based on the different experimentally determined GroEL conformations.

Methods

CONCOORD simulations

Principal component analyses of Molecular Dynamics simulations of proteins have indicated that collective degrees of freedom dominate protein conformational fluctuations[76, 78]. These large-scale collective motions have been shown essential to protein function in a number of cases[84–86]. The notion that internal constraints and other configurational barriers restrict protein dynamics to a limited number of collective degrees of freedom has led to the design of the CONCOORD method to predict these modes without doing explicit MD simulations. The CONCOORD method has been described in detail earlier[89] and will here only be summarised briefly, together with some recent modifications.

The CONCOORD method of prediction of protein conformational freedom generates protein structures within a set of predefined distance bounds. Distance bounds are defined on the basis of interatomic interactions within the starting configuration of the protein and the difference between upper and lower distance bounds depend on the strength of the interaction. A discrete number of categories of interactions has been defined, among which covalent bonds are the least flexible and weakly interacting non-bonded pairs have the largest freedom in distance. Starting from random coordinates, distance and chirality corrections are applied until all distances fulfill their distance bounds. Resulting structures are uncorrelated and hence the technique does not suffer from sampling problems as do techniques like MD in which such correlation is present.

Since the first implementation of CONCOORD, a number of improvements has been made [1]. First, the original algorithm which required all distances to be restricted has been modified to make the method suitable for large systems. Only the distances between atoms involved in pair interactions are now defined. In addition, in order to reach convergence, it appeared necessary to include a fixed number (typically 20N, with N the number of particles) of random pairs with significantly more distance freedom. This way, only up to a few percent of the whole distance matrix needs to be evaluated. Second, categories of distance limits and the difference between upper and lower distance bound for each category were re-evaluated based on crystallographic conformers of T4 lysozyme as well as on distance fluctuations of a number of proteins in MD simulations. The parameters obtained in this way resulted in structures of slightly better quality than those obtained with the previous set. Finally, non-bonded pairs are defined in a different way depending on the number of contacts within a group of residues. Isolated non-bonded interacting atom pairs will have more distance freedom (maximally 4 Å) than pairs which are part of an intensive network of interactions (e.g. pairs contained

[1] The latest version of the CONCOORD program is freely available from the internet: `http://rugmd0.chem.rug.nl/~degroot/concoord.html`

in clusters of more than 50 interactions maximally obtain 1.5 Å of distance freedom.

CONCOORD simulations were performed on each of the three currently available crystallographic double ring structures: the symmetrical (both rings are identical) nucleotide-free structure (pdb entry `1oel`[175,176]), the pseudo-symmetric ATP-γS-bound structure (the inter-ring contact plane is a plane of pseudo-symmetry; pdb entry `1der`[177], and the asymmetric ADP/GroES bound structure (one ring has ADP and GroES bound, the other is empty; pdb entry `1aon`[178]). Additionally, isolated single rings extracted from each of these structures were simulated individually.

Essential Dynamics analysis

Essential Dynamics (ED) analysis is equivalent to a principal component analysis of atomic displacements in an ensemble of structures[76] and is related to the so called 'quasi-harmonic' analysis of protein motions[73]. In practice, ED involves diagonalization of the covariance matrix of positional fluctuations (after removal of the overall rotation and translation). Resulting eigenvectors describe modes of collective fluctuation of which the corresponding eigenvalue is a measure of the mean square fluctuation along that mode[78].

ED analyses were applied to the ensemble of crystallographic structures to assess the main modes of collective fluctuation in GroEL. Ring conformational changes were analysed (inter-subunit fluctuations) by applying ED to the 5 unique ring conformations from the three double ring conformers determined by X-ray crystallography (the two rings of the unliganded GroEL structure `1oel` are symmetry related). The 35 subunit conformations extracted from these structures were subjected to ED analysis to study conformational changes within subunits (intra-subunit fluctuations). CONCOORD structures were projected onto the modes determined from the crystallographic structures to compare the fluctuations predicted by CONCOORD to the differences between crystallographic structures. The way the CONCOORD structures are situated along the collective coordinates derived from the X-ray structures indicate potential dynamic pathways between the experimentally determined conformers.

DYNDOM

Modes of collective fluctuation were analysed for the presence of clear domain motions by the method of Hayward *et al.* [55,56]. This method analyses structural differences in terms of rigid body rotations. The rigid bodies are identified by clustering each residue's rotation vector during a conformational transition.

Results and discussion

Conformational changes in the equatorial domain

An ED analysis of conformations of single subunits extracted from the different experimentally determined structures confirmed the observations of Xu *et al.* [178] that domain motions occur upon GroES binding. Two modes of collective fluctuation were found to dominate the conformational transitions of isolated subunits. The first, most prominent, mode describes differences between subunits extracted from the *cis* and *trans* rings from the asymmetric GroEL/ADP/GroES complex. Apical domains make a rotation of about 90 degrees with respect to the intermediate domains while the equatorial domains are involved in a closure motion of about 30 degrees with respect to the intermediate domains (see Figure 2c of Xu *et al.* [178]). The second mode displays the largest difference between the rings from the ATPγS bound structure and the other structures. Internal fluctuations within the equatorial and apical domains dominate along this second mode.

In contrast to the structural changes of the domains with respect to each other, the internal fluctuations of the equatorial domain are for a large part similar along the first and second mode. Residues involved in nucleotide binding show large displacements along this common mode, suggesting that structural changes necessary to accomodate ATP (or to a lesser extent ADP, or analogues) dominate the internal dynamics of the equatorial domains (Fig. 7.2). Along this common mode, the DYNDOM method[55, 56] identifies two subdomains. The first subdomain consists of residues 12-30, 37-83, 510-521 and the second subdomain of 32-34, 90-137, 411-506. Several residues directly involved in binding nucleotide (Val31-Pro33, Asp87, Thr91)[177, 213] are situated at the interface between the two subdomains (Fig. 7.2). Both groups have two glycine residues in their proximity (32, 35 and 85, 88) that allow for the conformational flexibility needed to adapt to the structural constraints imposed by the bound nucleotide.

Both subdomains of the equatorial domain also exhibit internal fluctuations. The lightest subdomain in Fig. 7.2 contains the two regions forming the most extensive contacts with the other ring (around Ala108 and Ser463). Upon nucleotide and GroES binding, the distance between these inter-ring contact-forming residues changes significantly. In each subunit, the distance between the C-α atoms of residues Ala108 and Ser463 is more than 2 Å smaller in the subunits of the *cis* ring than in those of the *trans* ring in the asymmetric GroEL-GroES complex[178]. These internal fluctuations have a direct effect at the interface and could play a role in the communication between the rings. These observed changes are consistent with the known negative cooperativity between the two GroEL rings, as depicted in Fig. 7.3. A motion of the residues around 108 and 463 towards each other in the equatorial domains of one ring must result in an opposite displacement in the other ring, if the integrity of the interface is to be maintained. The largest displacements of the

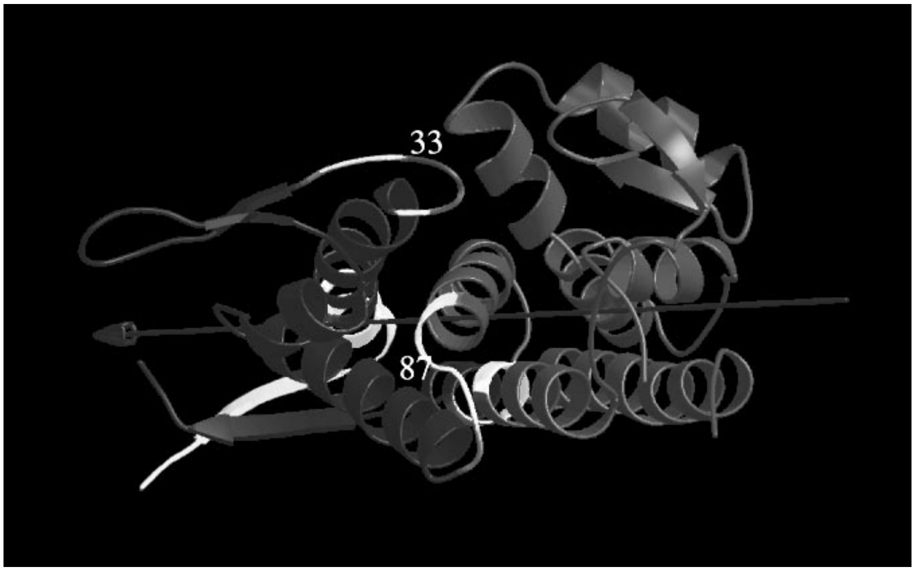

Figure 7.2 Illustration of the main structural changes within the equatorial domains. The results are obtained from DYNDOM[56] based on an ED analysis of X-ray structures of the equatorial domain. The equatorial domain can be considered to consist of two subdomains (in dark and intermedate gray). The residues that form the transition regions between the domains are coloured white to lightgray. The arrow indicates direction of rotation of the lighter domain relative to the darker domain by the thumb rule of the right hand. The loops containing residues 33 and 87, which are known to interact directly with the nucleotide, are situated at the flexible interface between the domains. Figure generated with Molscript[151,190] and Raster3d[191].

residues forming the inter-ring contacts are found to take place in the plane of the rings, but as Ma *et al.* pointed out, fluctuations perpendicular to this plane may also play a role in inter-ring communication[212].

The residues directly involved in inter-ring contacts show displacements both upon GroES binding and upon nucleotide binding (although with a smaller amplitude). The X-ray structures show a conformational change of the stem loop (Lys 34 to Asp 52) only between GroES bound subunits and subunits from GroES-free rings. This stem loop displacement is correlated with the reorientation of the intermediate domain with respect to the equatorial domain. This stem loop displacement also induces a motion of the subunits with respect to each other, resulting in the *en bloc* tilt of the equatorial domains in the cis ring with respect to the trans ring that has been reported by Xu *et al.* [178]. It has also been suggested that the stem loop was involved in the cooperative binding of ATP (and accompanied tertiary structural changes) in one ring from Normal Mode analysis[212]. Our results suggest that, additionally, these residues may be indirectly involved in inter-

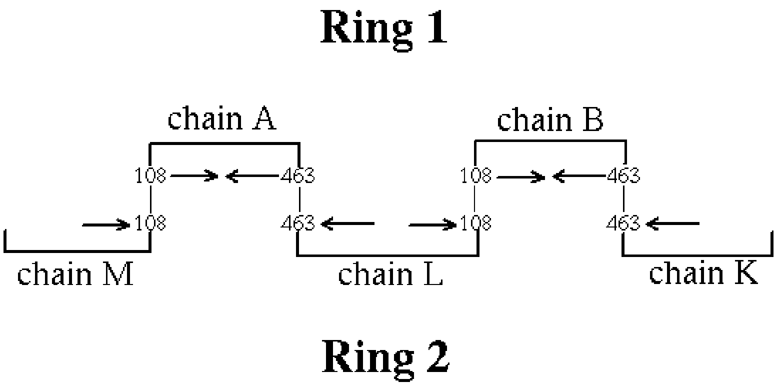

Figure 7.3 Illustration of how the internal fluctuations of the equatorial domains may be involved in the negative cooperativity between the two GroEL rings. A displacement of the two main sites of inter-ring contacts (around residues 108 and 463) in the subunits in one ring has to be compensated by a displacement in the opposite direction in the subunits of the other ring to preserve inter-ring contacts.

ring communication, in which the equatorial domains from one ring directly transmit stuctural changes associated with GroES binding (and to a lesser extent nucleotide binding) to the other ring.

Overall structural changes

Analysis of crystallographic structures reveals dramatic conformational differences between GroES-free rings and GroEL rings bound to the cochaperonin GroES[178]. Previous comparisons between X-ray structures of free GroEL and GroEL bound to ATPγS showed much more modest conformational differences[177]. Figure 7.4 schematically shows the main conformational differences between the different experimentally characterised GroEL rings. The largest difference is observed between the GroES bound *cis* ring and the different GroES-free rings (horizontal direction, first mode; from now on referred to as conformational transition 1 or CT1). The GroES-free rings differ most from each other along the mode with second-largest amplitude (CT2). The largest difference along CT2 is observed between the GroEL rings bound to ATPγS (pdb entry `1der`[177]) and the other ring from the asymmetric GroEL-GroES complex (the *trans* ring of the complex, pdb entry `1aon`). CT2 is likely to be connected with nucleotide binding and/or affinity since it describes the main difference between the rings from the `1oel` and `1der` X-ray structures which only differ from each other by the presence of ATP-γS.

X-ray structures of GroES-free rings have similar positions along CT1, indicating that conformational changes upon nucleotide binding are distinct from those upon GroES binding. The ring *trans* to the GroES bound ring in the asymmetric GroES-bound structure is shifted with respect to the nucleotide-free symmetric GroEL structure along CT2 and not along CT1. GroES binding, therefore, causes a shift along the mode presumably connected with nucleotide binding (in the direction of nucleotide release) in the ring *trans* to GroES.

CONCOORD simulations based on the different experimental structures sample both CT1 and CT2 with a significant amplitude (Fig. 7.4) and are among the largest-amplitude fluctuations in the simulations. Interestingly, there is a clear correlation between the fluctuation along CT1 and CT2 in the different double ring simulations (Fig. 7.4). For GroES-free rings, this correlation links conformational changes in the direction of the change taking place upon GroES binding with changes presumably happening upon nucleotide binding. Therefore, this connection between the two modes of conformational change displays a mechanism by which nucleotide binding in one ring would result in a conformational shift corresponding to a larger GroES affinity in the same ring.

No significant correlation is detected between CT1 and CT2 in the single ring simulations (Fig. 7.4). Apparently, interactions between the rings induce a conformational restriction on both rings which accomplishes the coupling between the two modes. Indeed, when the effect of the CT1 and CT2 on the packing of the equatorial domains is examined in detail, a mechanism emerges which explains the coupling. In the equatorial domains, the major site of contacts with the other ring are formed by residues 461-467. Significant displacements of these residues are observed in both CT1 and CT2 (Fig. 7.5). Looking along the cylindrical axis, the effect of a displacement along CT1 is an inward motion of these residues, whereas displacement along CT2 corresponds to an outward motion. Any steric restrictions that inhibit an overall inward or outward motion would therefore generate a coupling between CT1 and CT2.

To check if the observed coupling is a direct result of extra restrictions of the residues involved in inter-ring contacts in the double rings with respect to the single rings, a CONCOORD simulation was started on a single ring with these residues constrained. As can be seen in Fig. 7.4, CT1 and CT2 are even more strongly coupled than in the case of the double ring simulations. This indicates indeed the existence of a mechanism that correlates CT1 to CT2 (GroES binding to nucleotide binding) in one half of a double ring, induced by restrictions formed by the other ring.

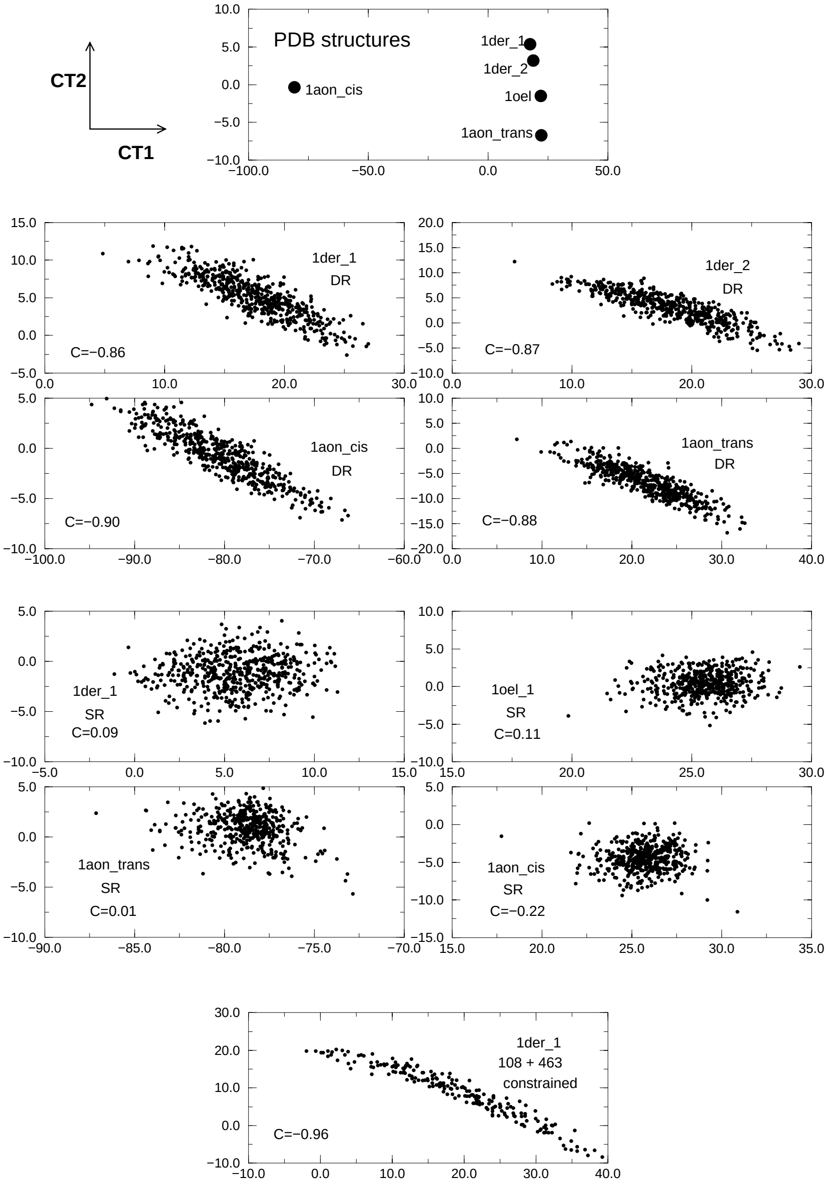

Figure 7.4 Essential dynamics analysis of conformational differences between ring conformations obtained from different experimental (X-ray) structures. Projection of individual rings onto the CT1-CT2 plane. Upper panel: ring conformations from crystallographic structures. Next four panels: CONCOORD-generated double-ring structures. Next four panels: CONCOORD-generated single-ring structures. Lower panel: CONCOORD generated single ring structures with the residues involved in inter-ring contacts (residues 108 and 463 were taken as representative) constrained. The values of C denote the correlation coefficient between the displacements along the two modes. SR: single ring; DR: double ring.

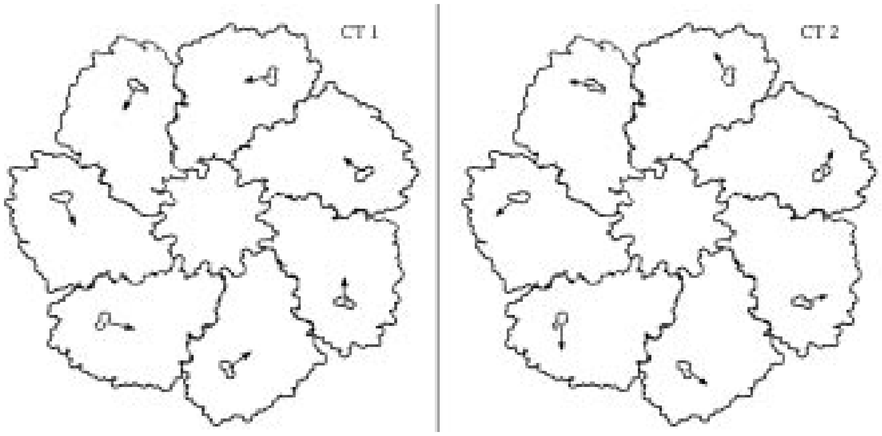

Figure 7.5 Schematic representation of the displacements of the inter-ring contact forming residues (C-α displacements of residues 463 from each subunit were chosen as representative) along the two dominating modes of ring fluctuation. The point of view is along the cylindrical axis formed by the double ring. The arrows indicate the (exaggerated) displacements of residues 463 from each subunit upon GroES release (left: CT1) and nucleotide release (right: CT2).

Conclusions

The results presented here provide new insight into the mechanism underlying the conformational changes of GroEL upon nucleotide and GroES binding. First, an ED analysis of GroEL subunits extracted from X-ray structures shows that within equatorial domains, a direct effect on the inter-ring interface is observable upon both GroES and nucleotide binding which may play a role in the observed negative cooperativity between GroEL rings. This mechanism may enhance (or cooperate with) an earlier observation that nucleotide binding affects the Glu434-Lys105 inter-ring contact[182]. Second, an ED analysis of the crystallographic ring conformers has shown that structural changes that take place upon GroES binding are not an extension (completion) of the changes induced by ATPγS. Rather, these changes are described by two perpendicular modes. Such a completion mechanism has been suggested based on EM data, where larger conformational shifts were observed upon nucleotide binding[182]. The results presented here indicate that structural differences upon GroES and nucleotide binding are described by two perpendicular modes which are not necessarily coupled. However, a coupling between the modes is observed in CONCOORD simulations of the double ring, correlating shifts towards GroES binding to shifts that happen upon nucleotide binding. Since such a coupling is not or hardly present in the simulations of single rings, this leads to the conclusion that the source of this coupling must be provided by the interface between the two rings. This finding is confirmed by the observation that this coupling is present in single ring simulations in which the inter-ring contact-forming residues are constrained. This coupling mechanism may provide an additional explanation for the fact that GroEL acts as a double ring. The double ring has previously been proposed to play a role in substrate release[214, 215], and to provide enhanced efficiency under stress conditions[216–218]. Furthermore, the results show that CONCOORD, despite a few obvious restrictions, is a powerful tool for studying protein conformational freedom for molecular weights and timescales that are currently beyond the scope of explicit dynamic simulation techniques.

Acknowledgements

We thank Dr. Ineke Braakman for critically reading the chapter. Steven Hayward is acknowledged for support with the DYNDOM program, and Frans van Hoesel for help with Figure 7.1.

8 CONCLUDING REMARKS

Current state of the art

As summarised in the introduction of this thesis, recent developments in computer simulations of biological macromolecules have enhanced the range of applicability of these techniques for the study of conformational properties of proteins. The methods described in this thesis form another contribution to this field, and applications to several proteins have yielded interesting results. The Essential Dynamics technique has proven a powerful analysis tool not only for the interpretation of MD simulations of proteins, but also of experimental and CONCOORD generated protein conformations. Enhanced sampling protocols based on the Essential Dynamics technique have shown to reach rates of conformational sampling that are 5-10 times higher than those reached by conventional MD. However, success of the ED sampling technique depends on the accuracy of the initial definition of the collective coordinates along which the positions are constrained in the sampling technique. Comparisons of subspaces spanned by these coordinates from different simulations have shown that the subspace overlap is usually not larger than 60 % (see chapters 2,5,6). Thus, although an approximate convergence of the definition of the principal collective coordinates can be obtained from MD simulations in the nanosecond time range, there is still a significant level of noise in this definition. It has recently been shown that several shorter simulations sample the conformational space of proteins more efficiently than one single, longer simulation[219]. It would be interesting to investigate whether the use of multiple short simulations also yields a faster convergence of the essential subspace.

The CONCOORD method has been described and applied in chapters 6 and 7. It has proven a simple yet powerful technique to study protein conformational freedom. Because it is based on very different principles than for instance Molecular Dynamics, the specific strengths and weaknesses of the CONCOORD method differ from those of the MD technique. Therefore, the two methods partially complement each other, enabling a deeper insight in conformational properties of proteins than can be obtained from either technique individually. Moreover, the ability of the CONCOORD method to yield modes of collective protein fluctuation that are similar to those obtained from MD and experiment proves that protein dynamics is largely governed by restrictions imposed by interactions in the native structure.

Normal Mode analyses form another computational tool to study fluctuations in proteins. Although limited to the fluctuations in a single harmonic well, the method has been shown to sample biologically relevant motions (e.g. ref. 220), and can be applied to relatively large proteins[212,221]. In a recent study, we have shown that combined Normal Modes from multiple (local) minima are more similar to collective modes of fluctuation derived from MD simulations, than are Normal Modes extracted from a single minimum[222].

An efficient, automated procedure to combine Normal Mode results from multiple local minima could form an alternative method to study protein dynamics. Further progress may be obtained by combinations of Molecular Dynamics, CONCOORD and Normal Modes. A method that exploits the specific strengths of each of the three techniques may prove a valuable tool for the study of conformational fluctuations in proteins. It should be noted, however, that both NM and CONCOORD critically depend on the presence of a high-resolution starting structure, whereas in MD, structures can be allowed to equilibrate from a low resolution model or from a structure determined under different conditions (e.g. different solvent).

Limitations

A good illustration of the limitations of the methods described in this thesis is formed by a project in which we studied the coupled tertiary and quarternary structural changes in haemoglobin. After the concept of Essential Dynamics was first conceived, the idea arose that the study of allosteric proteins would be an ideal application of the technique. Allosteric proteins are multi-subunit proteins that are characterised by a cooperative substrate binding. Communication between subunits is responsible for the dependence of the substrate affinity of one subunit on the binding state of the others. The binding affinity can often be further regulated by binding of other molecules at sites distinct from the substrate binding site. Most allosteric proteins exist in two or more conformations that differ in the packing of their subunits (quarternary conformation). One of these states is the preferred conformation in the absence of substrate, and another quarternary conformation is associated with the fully liganded state. In the traditional view[223], the binding of substrate to one of the subunits (slightly) changes the conformation of that subunit, triggering a quarternary conformational change that changes the substrate affinity of the other subunits. Molecules that regulate the activity of such proteins specifically stabilise one of the quarternary conformations. The correlation between the (usually small) tertiary structural changes and the larger global changes, if sampled realistically, would be detected by a covariance analysis like Essential Dynamics, and therefore such an analysis technique could improve our understanding of the mechanisms involved in such conformational changes.

Haemoglobin, probably the best studied allosteric protein, was studied with Molecular Dynamics techniques with the hope to learn about the coupling between the changes that take place in the subunits upon oxygen binding, and overall structural changes. Simulations of 1 nanosecond did not significantly sample the experimentally known quarternary conformational change. This structural change is known to take place in a time scale of microseconds after binding (or removal) of oxygen[224]. The three orders of magnitude time difference between the simulations and the experiment are probably the explanation for the absence of the conformational change in the simulations. However, applications to T4 lysozyme had shown that domain motions were sampled to an appreciable amplitude in simulations of the same length[86]. The critical difference between the two proteins is the presence of a (free) energy barrier between the different conformational states of haemoglobin. In T4 lysozyme the domain motion(s) are not restricted by such an internal barrier and fluctuate diffusively during simulation. CONCOORD simulations of haemoglobin did sample the conformational changes between the different quarternary structures. The CONCOORD method is less sensitive to internal barriers since there is no path dependence between successively generated structures. The conformational changes sampled by CONCOORD, however, did not show a unique mechanism of coupling between tertiary and quarternary structural changes. The specific interaction of oxygen with the haem prosthetic group and the local structural changes are not modeled accurate enough to allow identification of such a coupling mechanism. The CONCOORD results, however, did indicate a direct role of the C-terminal residues of each subunit in the allosteric mechanism.

Outlook

Computational techniques are widely used for the study of conformational properties of biological macromolecules, and their range of application will only grow in the future. From the refinement of experimental structures to the *ab initio* folding of proteins, computer simulation techniques have proven to be valuable tools that can complement insights obtained from experiment. The predictive power of computer simulation techniques applied to proteins is still limited because of the large number of degrees of freedom that need to be treated explicitly. In the introduction of this thesis an overview was given of methods that are currently used to enhance the efficiency of computer simulation techniques to study protein dynamics. Only time can tell which (combination of) techniques will prove most useful in the future.

Based on the resuls presented in this thesis, it follows that a large portion of the configurational freedom is defined by restrictions that are imposed directly by the structure (chapters 6 and 7), Methods that do not use a molecular description on the atomic level will lack features that are directly related to the specific atomic interactions or packing. Another source of artifacts in computer simulation techniques is the representation of atomic interactions. Whereas many aspects of collective protein fluctuations may be correctly described by methods that lack a sophisticated treatment of interactions (e.g. quarternary structural changes in haemoglobin by CONCOORD), other, more subtle, mechanisms may not be correctly represented by such techniques (e.g. the coupling between tertiary and quarternary structural changes in haemoglobin). Therefore, the kind of application defines the method of choice.

Some applications do not require sophisticated all-atom treatment (e.g. determining the hinge-bending mode in lysozyme). Other processes however, depend on subtle interactions and/or take place at time scales that can currently not be simulated realistically (e.g. substrate entry or exit in enzymes or the protein folding process).

In chapters 6 and 7 it was shown that many conformational properties of proteins can be obtained by a much more simple technique (CONCOORD) than Molecular Dynamics. Although the applicability of the CONCOORD method is limited because of the absence of a realistic atomic description, it has the advantage that it does not suffer from sampling problems, at least within predefined limits. Progress with respect to the current implementation can be obtained by releasing some of the constraints imposed by a single conformation. This would allow the generation of conformations more distinct from the starting structure, and a sampling of the paths between those conformations. The difficulty in the design of a method to accomplish this is the prediction of which interactions are to be maintained for each generated conformer, and for which there are alternatives available. One straightforward approach would be to use the method as it is, in a recursive manner. The first step would be the generation of structures based on a single conformation. In next steps, structures generated in the previous steps could be used to define new sets of distance limitations, on the basis of which new structures could be generated. Preferable would be to have multiple experimental structures or reliable MD structures that could be used in the same fashion. Future studies will have to resolve whether meaningful results can be obtained in this way.

Summarising, although biomolecular computer simulations have come of age[225], many interesting processes involving dynamics of biological macromolecules are still beyond the scope of current computational techniques. Simulations employing sophisticated atomic models are limited to short time scales, and more coarse-grained methods lack the atomic detail that is often essential for a full understanding of a dynamical process. However, constant improvement of methodologies, together with a steady increase of (affordable) computer power will allow the study of more complex systems on longer timescales. Backed up by constant thorough experimental validation, methods will be developed in the next decade(s) that will allow detailed simulation of functional dynamics of proteins, the interactions of proteins with other molecules (docking) and the protein folding process[226].

Bibliography

[1] Stryer, L. Biochemistry. 3d Ed. New York: W. H. Freeman and co.. 1988.

[2] Gerstein, M., Lesk, A. M., Chothia, C. Structural mechanisms for domain movements in protein. Biochemistry 33:6739–6749, 1994.

[3] Genick, U. K., Borgstahl, G. E., Ng, K., Ren, Z., Pradervand, C., Burke, P. M., Srajer, V., Teng, T. Y., Schildkamp, W., McRee, D. E., Moffat, K., Getzoff, E. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. Science 275:1471–1475, 1997.

[4] Van Gunsteren, W. F., Berendsen, H. J. C. Gromos manual. BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands 1987.

[5] Van der Spoel, D., Berendsen, H. J. C., Van Buuren, A. R., Apol, E., Meulenhoff, P. J., Sijbers, A. L. T. M., Van Drunen, R. Gromacs User Manual. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: `http://rugmd0.chem.rug.nl/~gmx` 1995.

[6] Van Gunsteren, W., Billeter, S., Eising, A., Hünenberger, P., Krüger, P., Mark, A., Scott, W., Tironi, I. Biomolecular simulation: the GROMOS96 manual and user guide. Biomos b.v., Zürich, Groningen 1996.

[7] Van Gunsteren, W. F., Berendsen, H. J. C. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. Angew. Chem. Int. Ed. Engl. 29:992–1023, 1990.

[8] Grubmüller, H., Tavan, P. Molecular dynamics of conformational substates for a simplified protein model. J. Chem. Phys. 101:5047–5057, 1994.

[9] McCammon, J. A., Gelin, B. R., Karplus, M. Dynamics of folded proteins. Nature 267:585–590, 1977.

[10] Schlick, T., Barth, E., Mandziuk, M. Biomolecular dynamics at long time steps: bridging the timescale gap between simulation and experimentation. Annu. Rev. Biomol. Struct. 26:181–222, 1997.

[11] Ryckaert, J. P., Ciccotti, G., Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints; molecular dynamics of n-alkanes. J. Comp. Phys. 23:327–341, 1977.

[12] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculation. J. Comp. Chem. 4:187–217, 1983.

[13] Brooks III, C. L., Pettitt, B. M., Karplus, M. Structural and energetic effects of truncated long-ranged interactions in ionic and polar fluids. J. Chem. Phys. 83:5897–5908, 1985.

[14] Schreiber, H., Steinhauser, O. Cut-off size does strongly influence molecular dynamics results on solvated polypeptides. Biochemistry 31:5856–5860, 1992.

[15] Perera, L., Essmann, U., Berkowitz, M. L. Effect of the treatment of long-range forces on the dynamics ions in aqueous solutions. J. Chem. Phys. 102:450–456, 1995.

[16] Hockney, R. W., Eastwood, J. W. computer simulation using particles. New York: McGraw-Hill. 1981.

[17] Darden, T., York, D., Pedersen, L. Particle mesh Ewald : an N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98:10089–10092, 1993.

[18] Perera, L., Li, L., Darden, T., Monroe, D. M., Pedersen, L. G. Prediciton of solution structures of the Ca^{2+}-bound gamma-carboxyglutamatic acid domains of protein s and homolog growth arrest specific protein 6: use of the particle mesh Ewald method. Biophys. J. 73:1847–1856, 1997.

[19] Meller, J., Elber, R. Computer simulation of carbon monoxide photodissociation in myoglobin: structural interpretation of the b states. Biophys. J. 74:789–802, 1998.

[20] Greengard, L., Rokhlin, V. A fast algorithm for particle simulations. J. Comp. Phys. 73:325–332, 1987.

[21] Zhou, R., Berne, B. J. A new molecular dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex systems. J. Chem. Phys. 103:9444–9458, 1995.

[22] Eichinger, M., Grubmüller, H., Heller, H., Tavan, P. Famusamm: A new algorithm for rapid evaluation of electrostatic interaction in molecular dynamics simulations. J. Comp. Chem. 18:1729–1749, 1997.

[23] Bennet, C. M. Mass tensor molecular dynamics. J. Comput. Phys. 19:267–279, 1975.

[24] Pomes, R., McCammon, J. A. Mass and step length optimization for the calculation of equilibrium properties by molecular dynamics simulations. Chem. Phys. Lett. 166:425–428, 1990.

[25] Feenstra, K. A., Hess, B., Berendsen, H. J. C. Improving efficiency and accuracy of molecular dynamics simulations of hydrogen-rich systems. in preparation, 1998.

[26] Wu, X., Wang, S. Self-guided Molecular Dynamics simulation for efficient conformational search. J. Phys. Chem. B 102:7238–7250, 1998.

[27] Levitt, M., Warshel, A. Computer simulation of protein folding. Nature 253:694–698, 1975.

[28] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.

[29] Ueda, Y., Taketomi, H., Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulations. II. a three dimensional lattice model of lysozyme. Biopolymers 17:1531–1548, 1978.

[30] Covell, D. G., Jernigan, R. L. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.

[31] Kolinksi, A., Skolnick, J. Monte Carlo simulations of protein folding. I. lattice model and interaction scheme. PROTEINS: Struct. Funct. Gen. 18:338–352, 1994.

[32] Sali, A., Shakhnovich, E. I., Karplus, M. Kinetics of protein folding. a lattice model study of the requirements for folding to the native state. J. Mol. Biol 235:1614–1636, 1994.

[33] Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., Chan, H. S. Principles of protein folding - a perspective from simple exact models. Protein Sci. 4:561–602, 1995.

[34] Chandrasekhar, S. Stochastic problems in physics and astronomy. Rev. Mod. Phys. 15:1–89, 1943.

[35] Honeycutt, J. D., Thirumalai, D. Metastability of the folded states of globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87:3526–3529, 1990.

[36] Srinivasan, R., Rose, G. D. Linus: A hierarchic procedure to predict the fold of a protein. PROTEINS: Struct. Funct. Gen. 22:81–99, 1995.

[37] Berriz, G., Gutin, A., Shakhnovich, E. I. Langevin model for protein folding: cooperativity and stability. J. Chem. Phys. 106:9276–9285, 1997.

[38] Jones, D. T., Taylor, W. R., Thornton, J. M. A new approach to protein fold recognition. Nature 358:86–89, 1992.

[39] Torda, A. E. Perspectives in protein-fold recognition. Curr. Opin. Struct. Biol. 7:200–205, 1997.

[40] Jones, D. T. Progress in protein structure prediction. Curr. Opin. Struct. Biol. 7:377–387, 1997.

[41] Miyazawa, S., Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 256:623–644, 1996.

[42] Jones, D. T., Thornton, J. M. Potential energy functions for threading. Curr. Opin. Struct. Biol. 6:210–216, 1996.

[43] Haliloglu, T., Bahar, I. Coarse-grained simulations of conformational dynamics of proteins: applications to apomyoglobin. PROTEINS: Struct. Funct. Gen. 31:271–281, 1998.

[44] Bahar, I., Erman, B., Haliloglu, T., Jernigan, R. L. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 36:13512–13523, 1997.

[45] Eastman, P., Doniach, S. Multiple time step diffusive langevin dynamics for proteins. PROTEINS: Struct. Funct. Gen. 30:215–227, 1998.

[46] McCammon, J. A., Wolynes, P. G., Karplus, M. Picosecond dynamics of tyrosine side chains in proteins. Biochemistry 18:927–942, 1979.

[47] Wesson, L., Eisenberg, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Prot. Sci. 1:227–235, 1992.

[48] Juffer, A. H. On the modelling of solvent mean force potentials. PhD thesis. Rijksuniversiteit Groningen. 1993.

[49] Stouten, P. F. W., Frömmel, C., Nakumara, H., Sander, C. An effective solvation term based on atomic occupancies for use in protein simulations. Mol. Sim. 10:97–120, 1993.

[50] Fraternali, F., Van Gunsteren, W. F. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. J. Mol. Biol. 256:939–948, 1996.

[51] Fritsch, V., Ravishanker, G., Beveridge, D. L., Westhof, E. Molecular dynamics simulations of poly(dA)-poly(dT): comparisons between implicit and explicit solvent representations. Biopolymers 33:1537–1552, 1993.

[52] Arnold, G. E., Manchester, J. I., Townsend, B. D., Ornstein, R. L. Investigation of domain motions in bacteriophage T4 lysozyme. J. Biomol. Struct. Dyn. 12(2):457–474, 1994.

[53] 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(2):684–692, 1996.

[54] Zeng, J., Treutlein, H. R., Simonson, T. Conformation of the Ras-binding domain of Raf studied by molecular dynamics and free energy simulations. PROTEINS: Struct. Funct. Gen. 31:186–200, 1998.

[55] Hayward, S., Kitao, A., Berendsen, H. J. C. Model free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. PROTEINS: Struct. Funct. Gen. 27:425–437, 1997.

[56] Hayward, S., Berendsen, H. J. C. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. PROTEINS: Struct. Funct. Gen. 30:144–154, 1998.

[57] Post, C. B., Dobson, C. M., Karplus, M. A molecular dynamics analysis of protein structural elements. PROTEINS: Struct. Funct. Gen. 5:337–354, 1989.

[58] Ichiye, T., Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. PROTEINS: Struct. Funct. Gen. 11:205–217, 1991.

[59] Rojewska, D., Elber, R. Molecular dynamics study of secondary structure motions in proteins: application to myohemerythrin. PROTEINS: Struct. Funct. Gen. 7:265–279, 1998.

[60] Karplus, M., Weaver, D. L. Protein-folding dynamics. Nature 260:404–406, 1976.

[61] Rojnuckarin, A., Kim, S., Subramaniam, S. Brownian dynamics simulations of protein folding: access to milliseconds time scale and beyond. Proc. Natl. Acad. Sci. U.S.A. 95:4288–4292, 1998.

[62] Ryckaert, J. P., Bellemans, A. Molecular dynamics of n-butane near its boiling point. Chem. Phys. Lett. 30:123–125, 1975.

[63] Noguti, T., Gō, N. Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. Biopolymers 24:527–546, 1985.

[64] Mazur, A. K., Dorofeev, V. E., Abagyan, R. A. Derivation and testing of explicit equations of motion for polymers described by internal coordinates. J. Comput. Phys. 92:261–272, 1991.

[65] Mathiowetz, A. M., Jain, A., Karasawa, N., Goddard 3rd, W. A. Protein simulations using techniques suitable for very large systems: the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. PROTEINS: Struct. Funct. Gen. 20:227–247, 1994.

[66] Stein, E. G., Rice, L. M., Brünger, A. T. Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. J. Magn. Reson. 124:154–164, 1997.

[67] Abagyan, R., Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983–1002, 1994.

[68] Lee, B., Kurochkina, N., Kang, H. S. Protein folding by a biased Monte Carlo procedure in the dihedral angle space. FASEB J. 10:119–125, 1996.

[69] Levitt, M., Sander, C., Stern, P. S. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. Int. J. Quant. Chem.: Quantum Biology Symposium 10:181–199, 1983.

[70] 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.

[71] 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.

[72] KArplus, M., Kushick, J. N. Methot for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332, 1981.

[73] Levy, R. M., Srinivasan, A. R., Olson, W. K., McCammon, J. A. Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 23:1099–1112, 1984.

[74] Levy, R. M., Karplus, M., Kushick, J., Perahia, J. Evaluation of the configurational entropy for proteins: application to molecular dynamics of an α-helix. Macromolecules 17:1370–1374, 1984.

[75] Teeter, M. M., Case, D. A. Harmonic and quasi harmonic descriptions of crambin. J. Phys. Chem. 94:8091–8097, 1990.

[76] Garcia, A. E. Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699, 1992.

[77] Kitao, A., Hirata, F., Gō, N. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. J. Chem. Phys. 158:447–472, 1991.

[78] Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. Essential dynamics of proteins. PROTEINS: Struct. Funct. Gen. 17:412–425, 1993.

[79] Romo, T. D., Clarage, J. B., Sorensen, D. C., Phillips Jr, G. N. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. PROTEINS: Struct. Funct. Gen. 22:311–321, 1995.

[80] Askar, A., Space, B., Rabitz, H. Subspace method for long time scale molecular dynamics. J. Phys. Chem. 99:7330–7338, 1995.

[81] 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.

[82] Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. Phys. Rev. 52:2893–2906, 1995.

[83] Abseher, R., Nilges, M. Are there non-trivial dynamic cross-correlations in proteins? J. Mol. Biol. 279:911–920, 1998.

[84] 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: Struct. Funct. Gen. 22:45–54, 1995.

[85] 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(11):1129–1136, 1995.

[86] De Groot, B. L., Hayward, S., Van Aalten, D. M. F., Amadei, A., Berendsen, H. J. C. Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. PROTEINS: Struct. Funct. Gen 31:116–127, 1998.

[87] Van Aalten, D. M. F., Conn, D. A., De Groot, B. L., Findlay, J. B. C., Berendsen, H. J. C., Amadei, A. Protein dynamics derived from clusters of crystal structures. Biophys. J. 73:2891–2896, 1997.

[88] Abseher, R., Horstink, L., Hilbers, C. W., Nilges, M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. PROTEINS: Struct. Funct. Gen. 31:370–382, 1998.

[89] De Groot, B. L., Van Aalten, D. M. F., Scheek, R. M., Amadei, A., Vriend, G., Berendsen, H. J. C. Prediction of protein conformational freedom from distance constraints. PROTEINS: Struct. Funct. Gen. 29:240–251, 1997.

[90] 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(2):169–181, 1997.

[91] Garcia, A. E., Soumpasis, D. M., Jovin, T. M. Dynamics and relative stabilities of parallel- and antiparallel-stranded DNA duplexes. Biophys. J. 66:1742–1755, 1994.

[92] Yamaguchi, H., Van Aalten, D. M. F., Pinak, M., Fumkawa, A., Osman, R. Essential dynamics of DNA containing a cis.syn cyclobutane thymine dimer lesion. Nucl. Acids Res. 26:1939–1946, 1998.

[93] 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 *Eschericia coli* determined by restrained molecular dynamics from NMR-NOE data. J. Mol. Biol. 237:544–559, 1994.

[94] Hayward, S., Kitao, A., Hirata, F., Gō, N. Effect of solvent on collective motions in globular proteins. J. Mol. Biol. 234:1207–1217, 1993.

[95] 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.

[96] Gronenborn, A. M., Filipula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T., Clore, G. M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–661, 1991.

[97] Barchi, J. J., Grasberger, B., Gronenborn, A. M., Glore, G. M. Investigation of the backbone dynamics of the igg-binding domain of the streptococcal protein g by heteronuclear two-dimensional ^{1}h-^{15}n nuclear magnetic resonance spectroscopy. Prot. Sci. 3:15–21, 1994.

[98] Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Hermans, J. Interaction models for water in relation to protein hydration. In: Intermolecular Forces. Pullman, B. ed. . D. Reidel Publishing Company Dordrecht 1981 331–342.

[99] Berendsen, H. J. C., Postma, J. P. M., DiNola, A., Haak, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81:3684–3690, 1984.

[100] Hooft, R. W., Sander, C., Vriend, G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. PROTEINS: Struct. Funct. Gen. 26:363–376, 1996.

[101] Vriend, G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8:52–56, 1990.

[102] Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.

[103] Scheek, R. M., Van Nuland, N. A. J., De Groot, B. L., Amadei, A. Structure from NMR and molecular dynamics: distance restraining inhibits motion in the essential subspace. J. Biomol. NMR. 6(1):106–111, 1995.

[104] Kuszewski, J., Clore, G. M., Gronenborn, A. M. Fast folding of a prototypic polypeptide: The immunoglobulin binding domain of streptococcal protein g. Prot. Sci. 3:1945–1952, 1994.

[105] Storch, E. M., Daggett, V. Molecular dynamics simulation of cytochrome b5: Implications for protein-protein recognition. Biochemistry 34(30):9682–9693, 1995.

[106] Janezic, D., Brooks, B. R. Harmonic analysis of large systems. II. comparison of different protein models. J. Comp. Chem. 16(12):1543–1553, 1995.

[107] 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(5):741–751, 1996.

[108] 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: Struct. Funct. Gen. 26:314–322, 1996.

[109] 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(8):3288–3292, 1995.

[110] Balsera, M. A., Wriggers, W., Oono, Y., Schulten, K. Principal component analysis and long time protein dynamics. J. Phys. Chem. 100(7):2567–2572, 1996.

[111] Janezic, D., Venable, M., Brooks, B. R. Harmonic analysis of large systems. III. comparison with molecular dynamics. J. Comp. Chem. 16(12):1707–1713, 1995.

[112] Skelton, N. J., Garcia, K. C., Goeddel, D. V., Quann, C., Burnier, J. P. Determination of the solution structure of the peptide hormone guanylin: Observation of a novel form of topological stereoisomerism. Biochemistry 33(46):13581–13592, 1994.

[113] Currie, M. G., Fok, K. F., Kato, J., Moore, R. J., Hamra, F. K., Duffin, K. L., Smith, C. E. Guanylin: an endogenous activator of intestinal guanylate cyclase. Proc. Natl. Acad. Sci. U.S.A. 89:947–951, 1992.

[114] Forte, L. R., Currie, M. G. Guanylin: a peptide regulator of epithelal transport. The FASEB journal 9(8):643–650, 1995.

[115] Field, M., Graf Jr., L. H., Laird, W. J., Smith, P. L. Heat-stable enterotoxin of *Eschericia coli*: in vitro effects on guanylate cyclase activity, cyclic gmp concentration and ion transport in small intestine. Proc. Natl. Acad. Sci. U.S.A. 75(6):2800–2804, 1978.

[116] De Sauvage, F. J., Keshav, S., Kuang, W.-J., Gillet, N., Henzel, W., Goeddel, D. V. Precursor structure, expression and tissue distributution of human guanylin. Proc. Natl. Acad. Sci. U.S.A. 89:9089–9093, 1992.

[117] Amadei, A., Linssen, A. B. M., De Groot, B. L., Berendsen, H. J. C. Essential degrees of freedom of proteins. In *Modelling of Biomolecular Structures and Mechanisms* (The Netherlands, 1995). et al., A. P., ed. . Kluwer.

[118] Postma, P. W., Lengeler, J. W., Jacobson, G. R. Phosphoenolpyruvate: carbohydrate phosphotransferase systems of bacteria. Microbiol. Rev. 57:543–594, 1993.

[119] Jia, Z., Quail, J. W., Waygood, E. B., Delbaere, L. T. J. The 2.0 Å-resolution structure of *Eschericia coli* histidine-containing phosphocarrier protein HPr. a redetermination. J. Biol. Chem. 268:22490–22501, 1993.

[120] Van Nuland, N. A. J., Boelens, R., Scheek, R. M., Robillard, G. T. High resolution structure of the phosphorylated form of he histidine-containing phosphocarrier protein HPr from *Eschericia coli* determined by restrained molecular dynamics from NMR-NOE data. J. Mol. Biol. 246:180–193, 1995.

[121] Jia, Z., Vandonselaar, M., Quail, J. W., Delbaere, L. T. J. Active-center torsion angle strain revealed in 1.6 Å-resolution structure of histidine-containing phosphocarrier protein. Nature 361:94–97, 1993.

[122] Van Nuland, N. A. J., Wiersma, J. A., Van der Spoel, D., De Groot, B. L., Scheek, R. M., Robillard, G. T. Phosphorylation-induced torsion-angle strain in the active center of HPr, detected by NMR and restrained molecular dynamics refinement. Prot. Sci. 5:442–446, 1996.

[123] Laskowski, R. A., MacArthur, M. ., Moss, D. S., Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. J. Appl Cryst. 26:283–291, 1993.

[124] Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7:95–99, 1963.

[125] Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., Sarma, V. R. Structure of hen egg-white lysozyme, a three dimensional fourier synthesis at 2 angstroms resolution. Nature 206:757–761, 1965.

[126] Diamond, R. Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol. 82:371–391, 1974.

[127] McCammon, J. A., Gelin, B., Karplus, M., Wolynes, P. G. The hinge bending mode in lysozyme. Nature 262:325–326, 1976.

[128] Matthews, B. W., Remington, S. J. The three dimensional structure of the lysozyme from bacteriophage T4. Proc. Natl. Acad. Sci. U.S.A. 71:4178–4182, 1974.

[129] Faber, H. R., Matthews, B. W. A mutant T4 lysozyme displays five different crystal conformations. Nature 348:263–266, 1990.

[130] Dixon, M. M., Nicholson, H., Shewchuk, L., Baase, W. A., Matthews, B. W. Structure of a hinge-bending bacteriophage T4 lysozyme mutant, Ile3 → Pro. J. Mol. Biol. 227:917–933, 1992.

[131] 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.

[132] Mchaourab, H. S., Oh, K. J., Fang, C. J., Hubbell, W. L. Conformation of T4 lysozyme in solution. hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. Biochemistry 36:307–316, 1997.

[133] Arnold, G. E., Ornstein, R. L. Protein hinge bending as seen in molecular dynamics simulations of native and M6I mutant T4 lysozymes. Biopolymers 41:533–544, 1997.

[134] Kuroki, R., Weaver, L. H., Matthews, B. W. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. Science 262:2030–2033, 1993.

[135] Berry, M. B., Meador, B., Bilderback, T., Liang, P., Glaser, M., Phillips Jr., G. N. The closed conformation of a highly flexible protein: The structure of e. coli adenylate kinase with bound amp and amppnp. PROTEINS: Struct. Funct. Gen. 19:183–198, 1994.

[136] Waksman, G., Krishna, T. S. R., Williams Jr., C. H., Kuriyan, J. Crystal structure of *Eschericia coli* thioredoxin reductase refined at 2 Å resolution. Implications for a large conformational change during catalysis. J. Mol. Biol. 236:800–816, 1994.

[137] Sim, K. K., Song, H. K., Shin, D. H., Hwang, K. Y., Suh, S. W. The crystal structure of a triacylglycerol lipase from *Pseudomonas cepacia* reveals a highly open conformation in the absence of a bound inhibitor. Structure 5:173–186, 1997.

[138] Verma, C. S., Caves, L. S. D., Hubbard, R. E., Roberts, G. C. K. Domain motions in dihydrofolate reductase: A molecular dynamics study. J. Mol. Biol. 266:776–796, 1997.

[139] Kasper, P., Sterk, M., Christen, P., Gehring, H. Molecular-dynamics simulation of domain movements in aspartate aminotransferase. Eur. J. Biochem. 240:751–755, 1996.

[140] Komeiji, Y., Uebayasi, M., Yamato, I. Molecular dynamics simulations of trp apo- and holorepressors: Domain structure and ligand-protein interaction. PROTEINS: Struct. Funct. Gen. 20:248–258, 1994.

[141] 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(10):2044–2053, 1996.

[142] 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(4):1554–1566, 1996.

[143] Weaver, L. H., Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. J. Mol. Biol. 193:189–199, 1987.

[144] Van Buuren, A. R., Marrink, S. J., Berendsen, H. J. C. A molecular dynamics study of the decane/water interface. J. Phys. Chem. 97(36):9206–9212, 1993.

[145] Hayward, S., Gō, N. Collective variable description of native protein dynamics. Annu. Rev. Phys. Chem. 46:223–250, 1995.

[146] Zhang, X. J., Matthews, B. W. Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. Prot. Sci. 3:1031–1039, 1994.

[147] Matsumura, M., Signor, G., Matthews, B. W. Substantial increase of protein stability by multiple disulphide bonds. Nature 342:291–293, 1989.

[148] Jacobson, R. H., Matsumura, M., Faber, H. R., Matthews, B. W. Structure of a stabilizing disulfide bridge mutant that closes the active-site cleft of T4 lysozyme. Prot. Sci. 1:46–57, 1992.

[149] Blaber, M., Zhang, X. J., Matthews, B. W. Structural basis of amino acid α-helix propensity. Science 260:1637–1640, 1993.

[150] Blaber, M., Zhang, X. J., Lindstrom, J. D., Pepiot, S. D., Baase, W. A., Matthews, B. W. Determination of α-helix propensity within the context of a folded protein: sites 44 and 131 un bacteriophage T4 lysozyme. J.Mol. Biol. 235:600–624, 1994.

[151] Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24:946–950, 1991.

[152] Bacon, D. J., Anderson, W. F. A fast algorithm for rendering space-filling molecule pictures. J. Mol. Graph. 6:219–220, 1988.

[153] Merritt, E. A., Murphy, M. E. P. Raster3D version 2.0: A program for photorealistic molecular graphics. Act. Cryst. D. 50:869–873, 1994.

[154] Elofsson, A., Nilsson, L. A 1.2 ns molecular dynamics simulation of the ribonuclease t 1)-3'-guanosine monophosphate complex. J. Phys. Chem. 100(7):2480–2488, 1996.

[155] 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: Struct. Funct. Gen. 23:49–62, 1995.

[156] Jorgensen, W. L., Tirado-Rives, J. Monte Carlo vs Molecular Dynamics for conformational sampling. J. Phys. Chem. 100(34):14508–14513, 1996.

[157] Van Aalten, D. M. F., Jones, P. C., De Sousa, M., Findlay, J. B. C. Engineering protein mechanics: Inhibition of concerted motions of the cellular retinol binding protein by site-directed mutagenesis. Prot. Eng. 10:31–38, 1997.

[158] Bonvin, A. M. J. J., Brünger, A. T. Conformational variability of solution nuclear magnetic resonance structures. J. Mol.Biol. 250:80–93, 1995.

[159] 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.

[160] Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., Saraste, M. Crystal structure of a src-homology 3 (sh3) domain. Nature 359:851–854, 1992.

[161] Babu, Y. S., Bugg, C. E., Cook, W. J. Structure of calmodulin refined at 2.2 Å. J. Mol. Biol. 204:191–204, 1988.

[162] Crippen, G. M. A novel approach to calculation of conformation: Distance geometry. J. Comp. Phys. 24:449–452, 1977.

[163] Levitt, M., Sander, C., Stern, P. S. Protein normal-mode dynamics - trypsin-inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.

[164] Hayward, S., Kitao, A., Gō, N. Harmonicity and anharmonicity in protein dynamics: A normal modes and principal component analysis. PROTEINS: Struct. Funct. Gen. 23:177–186, 1995.

[165] Havel, T. F., Kuntz, I. D., Crippen, G. M. The theory and practice of distance geometry. Bull. Math. Biol. 45:665–720, 1983.

[166] Vriend, G., Sander, C. Quality control of protein models: directional atomic contact analysis. J. Appl. Cryst. 26:47–60, 1993.

[167] 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.

[168] 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.

[169] 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.

[170] 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.

[171] 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: Struct. Funct. Gen. 24(3):304–313, 1996.

[172] Fenton, W. A., Horwich, A. L. GroEL-mediated protein folding. Prot. Sci. 6:743–760, 1997.

[173] Martin, J., Hartl, U. Chaperone-assisted protein folding. Curr. Opin. Struct. Biol. 7:41–52, 1997.

[174] Horovitz, A. Structural aspects of GroEL function. Curr. Opin. Struct. Biol. 8:93–100, 1998.

[175] Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D. C., Joachimiak, A., Horwich, A. L., Sigler, P. B. The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. Nature 371:578–586, 1994.

[176] Braig, K., Adams, P. D., Brünger, A. T. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. Nat. Struct. Biol. 2:1083–1094, 1995.

[177] Boisvert, D. C., Wang, J., Otwinowski, Z., Horwich, A. L., Sigler, P. B. The 2.4 Å crystal structure of the bacterial chaperonin GroEL complex with ATPγS. Nat. Struct. Biol. 3:170–177, 1996.

[178] Xu, Z., Horwich, A. L., Sigler, P. B. The crystal structure of the asymmetric GroEL-GroES-(ADP)$_7$ chaperonin complex. Nature 388:741–750, 1997.

[179] Langer, T., Pfeifer, G., Martin, J., Baumeister, W., Hartl, F. U. Chaperonin-mediated proteinfolding: GroES binds to one end of the GroEL cylinder, which accomodates the protein within its central cavity. EMBO J. 11:4757–4765, 1992.

[180] Ishii, N., Taguchi, H., Sumi, M., Yoshida, M. Structure of holo-chaperonin studied with electron-microscopy. FEBS Lett. 299:169–174, 1992.

[181] Chen, S., Roseman, A. M., Hunter, A. S., Wood, S. P., Burston, S. G., Ranson, N. A., Clarke, A. R., Saibil, H. R. Location of a folding protein and shape changes in GroEL-GroES complexes imaged by cryo-electron microscopy. Nature 371:261–264, 1994.

[182] Roseman, A. M., Chen, S., White, H., Braig, K., Saibil, H. R. The chaperonin ATPase cycle: Mechanism of allosteric switching and movements of substrate-binding domains in GroEL. Cell 87:241–251, 1996.

[183] White, H. E., Chen, S., Roseman, A. M., Yifrach, O., Horovitz, A., Saibil, H. R. Structural basis of allosteric changes in the GroEL mutant Arg197 →Ala. Nat. Struct. Biol. 4:690–694, 1997.

[184] Hendrix, R. W. Purification and properties of GroE, a host protein involved in bacteriophage assembly. J. Mol. Biol. 129:375–392, 1979.

[185] Hemmingsen, S. M., Woolford, C., van der Vies, S. M., Tilly, K., Dennis, D. T., Georgopoulos, C. P., Hendrix, R. W., Ellis, R. J. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature 333:330–334, 1988.

[186] Braig, K., Simon, M., Furuya, F., Hainfeld, J. F., Horwich, A. L. A polypeptide bound by the chaperonin GroEL is localized within a central cavity. Proc. Natl. Acad. Sci. U.S.A. 90:3978–3982, 1993.

[187] Hunt, J. F., Weaver, A. J., Landry, S., Gierasch, L., Deisenhofer, J. The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. Nature 379:37–45, 1996.

[188] Chandrasekhar, G. N., Tilly, K., Woolford, C., Hendrix, R., Georgopoulos, C. Purification and properties of the GroES morphogenetic protein of *Eschericia coli.* J. Biol. Chem. 261:12414–12419, 1986.

[189] Saibil, H., Dong, Z., Wood, S., auf der Mauer, A. Binding of chaperonins. Nature 353:25–26, 1991.

[190] Esnouf, R. M. An extensively modified version of molscript that includes greatly enhanced coloring capabilities. J. Mol. Graph. Model. 15:132–134, 112–113, 1997.

[191] Merritt, E. A., Bacon, D. J. Raster3D: photorealistic molecular graphics. Meth. Enzym. 277:505–524, 1997.

[192] Weissman, J. S., Kashi, Y., Fenton, W. A., Horwich, A. L. GroEL-mediated protein folding proceeds by multiple rounds of binding and release of nonnative forms. Cell 78:693–702, 1994.

[193] Ranson, N. A., Dunster, N. J., Burston, S. G., Clark, A. R. Chaperonins can catalyse the reversal of early aggregation steps when a protein misfolds. J. Mol. Biol. 250:581–586, 1995.

[194] Zahn, R., Perrett, S., Stenberg, G., Fersht, A. R. Catalysis of amide proton exchange by the molecular chaperones GroEL and SecB. Science 271:642–645, 1996.

[195] Corrales, F. J., Fersht, A. R. Toward a mechanism for GroEL-GroES chaperone activity: an ATPase-gated and -pulsed folding and annealing cage. Proc. Natl. Acad. Sci. U.S.A. 93:4509–4512, 1996.

[196] Buckle, A. M., Zahn, R., Fersht, A. R. A structural model for GroEL-polypeptide recognition. Proc. Natl. Acad. Sci. U.S.A. 94:3571–3575, 1997.

[197] Weissman, J. S., Rye, H. S., Fenton, W. A., and. A. L. Horwich, J. M. B. Characterization of the active intermediate of a GroEL-GroES-mediate protein folding reaction. Cell 84:481–490, 1996.

[198] Mayhew, M., da Silva, A. C. R., Martin, J., Erdjument-Bromage, H., Tempst, P., Hartl, F. U. Protein folding in the central cavity of the GroEL-GroES chaperonin complex. Nature 379:420–426, 1996.

[199] Fersht, T. E. G. A. R. Cooperativity in ATP hydrolysis by GroEL is increased by GroES. FEBS Lett. 292:254–258, 1991.

[200] Bochkareva, E. S., Lissin, N. M., Flynn, G. C., Rothman, J. E., Girshovich, A. S. Positive cooperativity in the functioning of molecular chaperone GroEL. J. Biol. Chem. 267:6796–6800, 1992.

[201] Jackson, G. S., Staniforth, R. A., Halsall, D. J., Atkinson, T., Holbrook, J. J., Clarke, A. R., Burston, S. G. Binding and hydrolysis of nucleotides in the chaperonin catalytic cycle: implications for the mechanism of assisted protein folding. Biochemistry 32:2554–2263, 1993.

[202] Staniforth, R. A., Burston, S. G., Atkinson, T., Clarke, A. R. Affinity of chaperonin-60 for a protein substrate and its modulation by nucleotides and chaperonin-10. Biochem. J. 300:651–658, 1994.

[203] Yifrach, O., Horovitz, A. Allosteric control by ATP of non-folded protein binding to GroEL. J. Mol. Biol. 255:356–361, 1996.

[204] Yifrach, O., Horovitz, A. Two lines of allosteric communication in the oligomeric chaperoninGroEL are revealed by the single mutation Arg196→Ala. J. Mol. Biol 243:397–401, 1994.

[205] Yifrach, O., Horovitz, A. Nested cooperativity in the ATPase activity of the oligomeric chaperonin GroEL. Biochemistry 34:5303–5308, 1995.

[206] Kad, N. M., Ranson, N. A., Cliff, M. J., Clarke, A. R. Asymmetry, commitment and inhibition in the GroE ATPase cycle impose alternating functions on the two GroEL rings. J. Mol. Biol. 278:267–278, 1998.

[207] Todd, M. J., Viitanen, P. V., Lorimer, G. H. Dynamics of the chaperonin ATPase cycle: implications for facilitated protein folding. Science 265:659–666, 1994.

[208] Weissman, J. S., Hohl, C. M., Kovalenko, O., Kashi, Y., Chen, S., Braig, K., Saibil, H. R., Fenton, W. A., Horwich, A. L. Mechanism of GroEL action: Productuve release of polypeptide from a sequestered position under GroES. Cell 83:577–587, 1995.

[209] Llorca, O., Peréz, J., Carrascosa, J. L., Galán, A., Muga, A. Effects of inter-ring communication in GroEL structural and functional asymetry. J. Biol. Chem. 272:32925–32932, 1997.

[210] Hayer-Hartl, M. K., Weber, F., Hartl, F. U. mechanism of chaperonin action: GroES binding and release can drive GroEL-mediated protein folding in the absence of ATP hydrolysis. EMBO J. 15:6111–6121, 1996.

[211] Berne, B. J., Straub, J. E. Novel methods of sampling phase space in the simulation of biological systems. Curr. Opin. Struct. Biol. 7:181–189, 1997.

[212] Ma, J., Karplus, M. The allosteric mechnism of GroEL: A dynamic analysis. Proc. Natl. Acad. Sci. U.S.A. 95:8502–8507, 1998.

[213] Fenton, W. A., Kasl, Y., Furtak, K., Horwich, A. L. Residues in chaperonin GroEL required for polypeptide binding and release. Nature 371:614–619, 1994.

[214] Inbar, E., Horovitz, A. GroES promotes the T to R transition of the GroEL ring distal to GroES in the GroEL-GroES complex. Biochemistry 36:12276–12281, 1997.

[215] Behlke, J., Ristau, O., Schönfeld, H.-J. Nucleotide-dependent complex formation between the *Eschericia coli* chaperonins GroEL and GroES studied under equilibrium conditions. Biochemistry 36:5149–5156, 1997.

[216] Azem, A., Diamant, S., Kessel, M., Weiss, C., Goloubinoff, P. The protein-folding activity of chaperonins correlates with the symmetric GroEL14(GroES7)2 heterooligomer. Proc. Natl. Acad. Sci. U.S.A. 92:12021–12025, 1995.

[217] Llorca, O., Marco, S., Carrascosa, J., Valpuesta, J. Symmetric GroEL-GroES complexes can contain substrate simultaneously in both GroEL rings. FEBS Lett. 405:195–199, 1997.

[218] Sparrer, H., Rutkat, K., Buchner, J. Catalysis of protein folding by symmetric chaperone complexes. Proc. Natl. Acad. Sci. U.S.A. 94:1096–1100, 1997.

[219] Caves, L., Evanseck, J., Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Prot. Sci. 7:649–666, 1998.

[220] Gibrat, J. F., Gō, N. Normal modes analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. PROTEINS: Struct. Funct. Gen. 8:258–279, 1990.

[221] Mouawad, L., Perahia, D. Motions in hemoglobin studied by normal mode analysis and energy minimization: evidence for the existence of tertiary t-like, quaternary r-like intermediate structures. J. Mol. Biol. 258:393–410, 1996.

[222] Amadei, A., de Groot, B. L., Ceruso, M. A., Paci, M., Nola, A. D., Berendsen, H. J. C. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. submitted.

[223] Monod, J., Wyman, J., Changeux, J.-P. On the nature of allosteric transitions: A plausible model. J. Mol. Biol. 12:88–118, 1965.

[224] Jones, C. M., Ansari, A., Henry, E. R., Christoph, G. W., Hofrichter, J., Eaton, W. A. Speed of intersubunit communication in proteins. Biochemistry 31:6692–6702, 1992.

[225] Berendsen, H. J. C. Bio-molecular dynamics comes of age. Science 271:954–955, 1996.

[226] Berendsen, H. J. C. Protein folding by simulation: A glimpse of the Holy Grail? Science 282:642–643, 1998.

SUMMARY

Protein dynamics plays an important role in the majority of biological processes. The ability of proteins to change conformation is essential for processes as diverse as oxygen transport and immune response. Therefore, a thorough knowledge of the principles governing the dynamics of proteins would greatly facilitate the understanding of these processes. Moreover, it would enhance the possibilities to modify dynamical properties of proteins by mutation for industrial or medical purposes. Much of what is known today about protein structure and dynamics is derived from experimental data. However, no experimental technique is currently available to follow individual protein structures at the nanosecond time scale, the times at which typical relevant protein motions occur. Computer simulation techniques provide the only way to obtain information on conformational properties of proteins at an atomic level at the picosecond to microsecond time scale. The reliability of such computer simulation depends on the accuracy of the starting model of the protein and the sophistication of the simulation procedure. Molecular Dynamics (MD) techniques are among the most popular methods to simulate protein dynamics.

Analyses of Molecular Dynamics simulations of several proteins by the Essential Dynamics (ED) technique have shown that protein dynamics is dominated by a limited number of backbone motions. The ED method is based on a covariance (principal component) analysis of the atomic coordinates and yields collective degrees of freedom that best approximate the full dynamics. The notion that only a few collective coordinates (which together span a hyper-surface, the essential subspace) suffice to approximate the backbone dynamics of a given protein simplifies protein dynamics dramatically. This simplification is not only useful for the interpretation of simulation results but can also be utilised in the design of novel simulation techniques.

Chapter 2 of this thesis is concerned with the convergence of ED results from relatively short MD simulations. A number of groups had reported that principal component analysis of MD simulations of such short time lengths is not suitable for describing long-time scale protein dynamics because the subspace keeps changing throughout the simulations. It is shown in this chapter that even in simulations in the range of hundreds of picoseconds, an approximately converged definition of the essential subspace can be reached for a small protein in an aqueous environment. The individual (eigen)vectors that span this space, however, are not sampled enough in such a short period to allow a fully converged definition. Apart from the issue of convergence of the essential subspace, the sensitivity of the essential dynamics results to MD parameters is also described in this chapter. It was found that essential dynamics results from MD simulations with different parameters deviate as much from those extracted from a set of reference simulations, as do the reference simulations among each other. Only for a simulation in vacuo a significant deviation was observed.

The third chapter describes a technique to perform MD simulations with an enhanced conformational sampling rate, based on ED. An approximation of the essential subspace is obtained from an initial simulation. In subsequent simulations, knowledge of the (approximate) essential subspace is used to impose constraints that encourage sampling of previously unsampled regions of configurational space. An application to the peptide hormone guanylin shows that this ED sampling technique samples configurational space at a rate that is approximately seven times larger than that obtained by conventional MD. Structures generated by the ED sampling technique are not significantly perturbed or strained as is indicated by the fact that when subjected to usual MD, they do not show any drift in a particular direction. Analysis of the sampled configurational space indicates that for both forms of the peptide, an almost exhaustive sampling has been reached. Additionally, except close to the borders of the sampled regions, free energy gradients in the essential subspace are found to be small, indicating a rather flat free energy surface surrounded by steep borders.

Chapter 4 presents an application of the ED sampling technique to a small protein: HPr. A comparable increase in the rate of conformational sampling compared to conventional MD was obtained as with the peptide described in chapter 3. Geometrical properties, like secondary structure and solvent accessibility, as well as energies of structures from the ED sampling ensemble are comparable to those obtained from the much more compact ensembles of MD simulation or NMR refinement. Strikingly, violations of the NMR data are comparable for the ED sampling, usual MD and NMR ensembles, indicating that a much larger cluster equally well satisfies nearly all NMR data.

In chapter 5 a comparison is presented between ED results obtained from MD simulations and those derived from a set of crystallographic structures of bacteriophage T4 lysozyme. T4 lysozyme is probably the best experimentally characterised protein from a structural point of view, with hundreds of experimental structures in the Protein Data Bank. T4 lysozyme consists of two-domains and the domain motions dominate the global fluctuations of this protein. The modes obtained by ED. analysis were characterised in terms of domain motions by the DYNDOM program and results showed that there is significant overlap between the modes derived from the different experimental structures and those obtained from MD simulation. Two modes of domain motion were found in both clusters: the well known hinge-bending mode of lysozyme and a twisting mode. Together with spin-labeling experiments in solution, the results indicate that both the hinge-bending and the twist motion are involved in the catalytic mechanism of T4 lysozyme.

ED analysis of MD simulations of proteins had revealed that protein dynamics is for a large part limited to a small number of collective degrees of freedom. All other degrees of freedom are (virtually) constrained due to internal barriers caused by interactions in compactly folded protein structures. Since both groups of coordinates are complementary, a correct modeling of all interactions in the structure of a given protein should yield the accessible degrees of fluctuations. Chapter 6 describes a method that is based on this idea. The CONCOORD method generates protein structures based on lower and upper distance bounds that are derived from inter-atomic interactions that are present in a starting configuration. Starting from random coordinates, corrections are iteratively applied until all distance bounds are fulfilled. Clusters of protein structures that are obtained in this way were compared to those obtained by MD simulations for several proteins. It was found that for all properties considered, the CONCOORD and MD results were comparable. This suggests that for many purposes, CONCOORD simulations may be used instead of much more CPU-intensive MD simulations.

Chapter 7 presents an application of the CONCOORD method to the molecular chaperonin complex GroEL/GroES. Consisting of 8000 amino-acids, it would be an infeasible task to simulate GroEL by MD even for only one nanosecond. GroEL exists as a double back-to-back ring, and communication between the rings is known to play a role in the allosteric mechanism of the chaperonin. Experimentally it is known that GroEL changes conformation upon both cochaperonin (GroES) and nucleotide (ATP) binding. ED analysis of the different experimental structures showed that conformational changes in single rings upon nucleotide binding are distinct, and in principle uncoupled, from changes upon GroES binding. CONCOORD simulations do show a coupling between the two conformational transitions, but only in simulations of double rings, not in simulations of single rings. This may provide another reason for GroEL to act as a double ring. Internal motions of the nucleotide-binding subunits, which also provide the contact regions between the two rings, were shown to be involved in a possible direct form of inter-ring communication.

The last chapter discusses the current state of the art in computer simulation methods to study protein dynamics. Despite serious limitations, computer simulation methods are essential for a better understanding of protein dynamics and future improvements will allow reliable simulations of systems that are currently beyond the scope of any simulation technique even on the most modern computers.

SAMENVATTING

Eiwitdynamica speelt een belangrijke rol in de meeste biologische processen. De mogelijkheid van eiwitten om hun conformatie te veranderen is essentiëel voor processen zo divers als zuurstoftransport en immuunreacties. Het kennen van de principes die aan eiwitdynamica ten grondslag liggen zou daarom het begrijpen van dit soort processen enorm vergemakkelijken. Bovendien zou het de mogelijkheden vergroten om dynamische eigenschappen van eiwitten door mutaties te veranderen voor industriële of medische doeleinden. Veel van wat er op dit moment bekend is over eiwitstructuren en -dynamica is afgeleid van experimentele data. Echter, tot op heden is er geen experimentele techniek beschikbaar om eiwitstructuren te volgen op de tijdschaal van nanoseconden, de typische tijdschaal waarop relevante eiwitbewegingen plaatsvinden. Computersimulatietechnieken vormen de enige manier om informatie over conformationele eigenschappen van eiwitten te verkrijgen op atomair niveau en tijdschalen van picoseconden tot microseconden. De betrouwbaarheid van zulke computersimulaties hangt af van de nauwkeurigheid van het startmodel van het eiwit en de kwaliteit van de simulatieprocedure. Moleculaire Dynamica (MD) simulaties vormen één van de populairste klassen van methodes om eiwitdynamica te simuleren.

Analyses van MD simulaties van verschillende eiwitten met de Essentiële Dynamica (ED) techniek hebben aangetoond dat eiwitdynamica wordt gedomineerd door een klein aantal bewegingen van de hoofdketen. De ED methode is gebaseerd op een covariantie (principale componenten) analyse van de atomaire verplaatsingen en produceert collectieve vrijheidsgraden die het best de totale fluctuaties benaderen. Het feit dat slechts een paar collectieve coördinaten (die samen een oppervlak in de multidimensionale ruimte spannen, de zogenaamde essentiële subruimte) voldoen om de hoofdketenbewegingen te benaderen, vereenvoudigt eiwitdynamica dramatisch. Deze vereenvoudiging is niet alleen handig voor de interpretatie van simulatieresultaten maar kan ook gebruikt worden bij het ontwerpen van nieuwe simulatietechnieken.

Hoofdstuk 2 van dit proefschrift behandelt de convergentie van ED resultaten verkregen uit relatief korte MD simulaties. Een aantal groepen had gerapporteerd dat principale componenten analyses niet geschikt zijn om langetijdschaal eiwit dynamica mee te beschrijven omdat de essentiële subruimte constant zou veranderen tijdens de simulatie. Het wordt in dit hoofdstuk aangetoond dat zelfs uit simulaties van enkele honderden picoseconden een bij benadering geconvergeerde definitie van de essentiële subruimte kan worden verkregen voor een klein eiwit in een waterige oplossing. De configuratieruimte die in zo'n korte tijd wordt bezocht is echter te klein om een geconvergeerde definitie te verkrijgen van de individuele (eigen)vectoren die deze ruimte opspannen. Naast de kwestie van convergentie van de essentiële subruimte wordt de gevoeligheid van de ED resultaten voor MD parameters beschreven in dit hoofdstuk. Het bleek dat de ED resultaten uit MD simulaties met verschillende parameters even veel verschilden van een set referentiesimulaties als de referentiesimulaties onderling. Alleen voor een simulatie in vacuum werd een significant verschil waargenomen.

Het derde hoofdstuk beschrijft een techniek om MD simulaties uit te voeren die de conformatieruimte sneller afzoekt, gebaseerd op ED. Een benadering van de essentiële subruimte wordt verkregen uit een initiële simulatie. In daaropvolgende simulaties wordt deze benadering gebruikt om het systeem te dwingen gebieden in de configuratieruimte te bezoeken waar het voorheen niet geweest was. Een toepassing op het peptide hormoon guanyline laat zien dat de zoeksnelheid waarmee de configuratieruimte wordt gescand ongeveer een factor zeven groter is dan wordt verkregen met conventionele MD. De structuren die met deze techniek gegenereerd worden zijn niet significant verstoord of gespannen, wat wordt aangetoond door het feit dat wanneer ze worden onderworpen aan normale MD, er geen drift plaatsvindt vanaf de startpositie. Analyse van de bezochte conformatieruimtes laat zien dat voor beide vormen van het peptide bijna de complete ruimte is bezocht. Bovendien bleek dat vrije energie-gradienten in de essentiële subruimte relatief klein zijn, behalve dichtbij de grenzen van de bezochte gebieden. Dit duidt op een vlak vrije energie oppervlak omringd door steile grenzen.

In hoofdstuk 4 wordt de ED zoekmethode toegepast op een klein eiwit: HPr. Een vergelijkbare winst in zoeksnelheid vergeleken met conventionele MD werd verkregen als bij het peptide van hoofdstuk 3. Geometrische eigenschappen als secundaire structuur en water toegankelijkheid als ook energieën van structuren die met de ED zoekmethode zijn gegenereerd zijn vergelijkbaar met die verkregen uit de veel meer compacte clusters die uit MD of NMR verfijning komen. Opvallend is dat overschrijdingen van grenzen die uit NMR data berekend zijn even groot zijn in clusters verkregen met de ED zoekmethode, MD en NMR verfijning, wat aangeeft dat een veel groter cluster even goed aan bijna alle NMR data kan voldoen als een veel kleiner cluster.

Hoofdstuk 5 beschrijft een vergelijking tussen ED resultaten uit MD simulaties en uit een set kristallografische structuren van bacteriofaag T4 lysozym. Vanuit een structureel oogpunt is T4 Lysozym waarschijnlijk het best gekarakteriseerde eiwit, met honderden structuren in de Protein Data Bank. T4 lysozym bestaat uit twee domeinen en de domeinbewegingen domineren de globale fluctuaties van dit eiwit. ED resultaten werden onderzocht op domeinbewegingen met het DYNDOM programma en de resultaten laten zien dat de MD resultaten goed overeenkomen met de data verkregen uit de experimentele structuren. Twee domeinbewegingen werden gevonden in beide clusters van structuren: de welbekende sluitbeweging van lysozym en een twist beweging. Samen met spin-label experimenten in oplossing suggereren deze resultaten dat niet alleen de sluitbeweging, maar ook de twistbeweging een rol speelt in het catalytische mechanisme van T4 lysozym.

ED analyses van MD simulaties hebben keer op keer bevestigd dat eiwitdynamica voor een groot deel beperkt is tot een klein aantal collectieve vrijheidsgraden. Alle andere vrijheidsgraden hebben veel minder bewegingsvrijheid vanwege interne barrières veroorzaakt door interacties in compact opgevouwen eiwitstructuren. Omdat beide groepen vrijheidsgraden complementair zijn zou een correcte modellering van al deze interacties voor een bepaald eiwit de toegankelijke vrijheidsgraden automatisch moeten opleveren. Hoofdstuk 6 beschrijft een methode die uitgaat van dit idee. De CONCOORD methode genereert eiwitstructuren die gebaseerd zijn op een set onder- en bovengrenzen van afstanden die zijn afgeleid van interatomaire interacties die aangetroffen zijn in een startstructuur. Er wordt begonnen met willekeurige coördinaten, waarna iteratief correcties worden toegepast totdat aan alle afstandscriteria voldaan wordt. Clusters van eiwitstructuren die op deze manier zijn verkregen zijn vergeleken met die uit MD simulaties voor een aantal eiwitten. Het bleek dat voor alle eigenschappen die onderzocht zijn, er vergelijkbare resultaten uit MD en CONCOORD kwamen. Dit suggereert dat voor veel doeleinden CONCOORD even goed gebruikt kan worden als MD simulaties, die veel meer computertijd kosten.

In hoofdstuk 7 wordt een toepassing van de CONCOORD methode beschreven op het moleculaire chaperone complex GroEL/GroES. Met zijn 8000 aminozuren zou het onpraktisch zijn dit systeem met MD simulaties te bestuderen, zelfs voor één nanoseconde. GroEL komt voor als twee ringen die met de ruggen tegen elkaar liggen en het is bekend dat communicatie tussen de ringen een rol speelt in het allostere mechanisme van dit eiwit. Experimenteel is vastgesteld dat GroEL conformatieveranderingen ondergaat als het cochaperone (GroES) of nucleotide (ATP) bindt. ED analyse van verschillende experimentele structuren heeft uitgewezen dat structuurveranderingen die optreden in individuele ringen bij het binden van nucleotide verschillen, en in principe ongekoppeld plaatsvinden, van veranderingen die het molecuul ondergaat wanneer het GroES bindt. CONCOORD simulaties laten echter wel een koppeling zien tussen de twee conformatieovergangen, maar alleen in simulaties van dubbele ringen, en niet in enkelring simulaties. Dit is een mogelijke verklaring voor het voorkomen van GroEL als een dubbelring. Interne bewegingen van de nucleotide-bindende subeenheden, die ook de contactregio's tussen de twee ringen vormen, bleken betrokken te zijn bij een mogelijke directe vorm van inter-ring communicatie.

Het laatste hoofdstuk beschrijft de laatste ontwikkelingen in computersimulatiemethoden om eiwitdynamica te bestuderen. Ondanks belangrijke beperkingen zijn computersimulaties essentiëel voor het beter begrijpen van eiwitdynamica en toekomstige ontwikkelingen zullen betrouwbare simulaties mogelijk maken van systemen die nu nog buiten het bereik van elke simulatietechniek vallen, zelfs op de meest moderne computers.

LIST OF PUBLICATIONS

1. A. Amadei, A.B.M. Linssen, B.L. de Groot and H.J.C. Berendsen; ”Essential degrees of freedom of proteins”, in ”Modelling of Biomolecular Structures and Mechanisms” A. Pullmann et. al. (ed.) 85-93 (1995)

2. R.M. Scheek, N.A.J. van Nuland, B.L. de Groot and A. Amadei; ”Structure from NMR and molecular dynamics: distance restraining inhibits motion in the essential subspace”, J. Biomol. NMR. 6: 106-11 (1995)

3. A. Amadei, A.B.M. Linssen, B.L. de Groot, D.M.F. van Aalten and H.J.C. Berendsen; ”An efficient method for sampling the essential subspace of proteins.”, J. Biom. Str. Dyn. 13: 615-626 (1996)

4. B.L. de Groot, A. Amadei, D.M.F. van Aalten and H.J.C. Berendsen; ”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)

5. N.A.J. van Nuland, J.A. Wiersma, D. van der Spoel, B.L. de Groot, R.M. Scheek and G.T. Robillard; ”Phosphorylation-induced torsion-angle strain in the active center of HPr, detected by NMR and restrained molecular dynamics refinement”, Prot. Sci. 5: 442-446 (1996)

6. B.L. de Groot, D.M.F. van Aalten, A. Amadei and H.J.C. Berendsen; ”The consistency of large concerted motions in proteins in Molecular Dynamics simulations” , Biophys. J. 71: 1554-1566 (1996)

7. B.L. de Groot, A.Amadei, R.M. Scheek, N.A.J. van Nuland and H.J.C. Berendsen; ”An extended sampling of the configurational space of HPr from E. coli” , PROTEINS: Struct. Funct. Gen. 26: 314-322 (1996)

8. D. van der Spoel, B.L. de Groot, S. Hayward, H.J.C. Berendsen and H.J. Vogel; ”Bending of the Calmodulin Central Helix: A Theoretical Study” , Prot. Sci. 5: 2044-2053 (1996)

9. D.M.F. van Aalten, B.L. de Groot, H.J.C. Berendsen and J.B.C. Findlay; ” Conformational analysis of retinoids and restriction of their dynamics by retinoid-binding proteins”, The biochemical journal 319(2): 543-550 (1996)

10. D.M.F. van Aalten, B.L. de Groot, H.J.C. Berendsen, J.B.C. Findlay and A. Amadei; ”A Comparison of Techniques for Calculating Protein Essential Dynamics”, J.Comp. Chem. 18: 169-181 (1997)

11. B.L. de Groot, D.M.F. van Aalten, R.M. Scheek, A. Amadei, G. Vriend and H.J.C. Berendsen; ”Prediction of protein conformational freedom from distance constraints”, PROTEINS: Struct. Funct. Gen. 29: 240-251 (1997)

12. D.M.F. van Aalten, D.A. Conn, B.L. de Groot, J.B.C. Findlay, H.J.C. Berendsen and A. Amadei; ”Protein dynamics derived from clusters of crystal structures”, Biophys. J. 73: 2891-2896 (1997)

13. B.L. de Groot, S. Hayward, D.M.F. van Aalten, A. Amadei, H.J.C. Berendsen; ”Domain Motions in Bacteriophage T4 Lysozyme; a Comparison between Molecular Dynamics and Crystallographic Data” PROTEINS: Struct. Funct. Gen. 31: 116-127 (1998).

14. Johannes P.M. Langedijk, Bert L. de Groot, Herman J.C. Berendsen and Jan T. van Oirschot; ”Structural homology of the central conserved region of the attachment protein G of the respiratory syncytial virus with the fourth subdomain of 55 kD tumor necrosis factor receptor”, Virology 243: 293-302 (1998).

15. A. Amadei, B.L. de Groot, M.-A. Ceruso, M. Paci, A. Di Nola and H.J.C. Berendsen; ”A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells”. PROTEINS: Struct. Funct. Gen. In press.

16. Bert L. de Groot, Gerrit Vriend and Herman J.C. Berendsen; ”Conformational changes in the chaperonin GroEL: new insights into the allosteric mechanism”, submitted

NAWOORD

Zo, het zit er op. En hoewel verreweg de meeste inspanningen die nodig waren voor de totstandkoming van dit boekje wel door ondergetekende geleverd zijn, had één en ander toch een stuk langer geduurd als er niet een paar anderen waren geweest die zo nu en dan eens een helpende hand hadden toegestoken. Allereerst wil ik mijn promotor Herman Berendsen bedanken voor de mogelijkheid om in zijn groep te promoveren en de manier waarop. De vrijheid om mijn neus achterna te lopen gecombineerd met een bekwame begeleiding hebben ervoor gezorgd dat ik niet alleen een leuke tijd heb gehad, maar dat ik er ook nog wat van opgestoken heb.

Secondly, I would like to thank Andrea for making me enthousiastic for the world of protein dynamics, and for the invention of Essential Dynamics. Op deze plaats wil ik ook Ton bedanken, voor de steun tijdens mijn afstudeeronderzoek en het begin van mijn promotietijd. Ruud wil ik bedanken voor zijn nooit aflatende interesse in mijn werkzaamheden en voor de dingen die ik in de loop van de tijd over NMR heb geleerd. Ook Nico mag in dit verband niet onvermeld blijven: op het HPr werk kijk ik nog steeds met plezier terug. Gert wil ik bedanken voor de gastvrijheid voor de keren dat ik Heidelberg bezocht en voor de prettige samenwerking waar ik een hoop van geleerd heb. Met Daan heb ik niet alleen een hoop lol gehad maar alle (zeer vermoeiende) tripjes over en weer zijn ook nog eens zeer productief gebleken. “Mijn” student Alex bedank ik voor zijn inspanningen aan myoglobine, waarover wellicht een hoofdstuk in dit boekje zou zijn verschenen als we wat meer mazzel hadden gehad. David, Pieter, Berk en Anton wil ik bedanken voor hun hulp bij mijn Gromacs problemen, Steve voor DYNDOM, Frans voor soepel systeembeheer en hulp bij plaatjes, en mijn kamergenoten Marc, Emile, Danilo en Frank voor antwoord op veel van mijn vragen.

Wellicht ontstaat de indruk dat er op het lab alleen maar leuke dingen gebeuren, maar regelmatig moest er ook hard gewerkt worden. Met name Frans en Alexander wisten vaak niet van ophouden bij een potje tafeltennis, om over de talloze bloedige achtervolgingen in duistere gangen tijdens hectische quake sessies nog maar te zwijgen..

Tot slot wil ik mijn ouders bedanken voor hun onvoorwaardelijke steun.