Development of a normal mode-based geometric simulation approach for investigating the intrinsic mobility of proteins

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt im Fachbereich Biowissenschaften
der Goethe Universität
in Frankfurt am Main

von
Aqeel Ahmed
aus Phullahdyoon, Pakistan

Frankfurt am Main 2009
(D30)

vom Fachbereich Biowissenschaften
der Goethe Universität als Dissertation angenommen.

Dekan: Prof. Dr. Volker Müller
Gutachter: Prof. Dr. Holger Gohlke
Prof. Dr. Peter Güntert

Datum der Disputation:

Table of contents

1 Introduction and aims ... *1*

2 State of the art ... *7*

2.1 Molecular dynamics (MD) ... **7**

2.2 Normal mode analysis (NMA) ... **9**

2.3 Elastic network model (ENM) ... **11**

2.4 FIRST, ROCK and FRODA ... **14**

2.5 CONCOORD ... **16**

3 Theory and implementation ... *18*

3.1 Rigid Cluster Normal Mode Analysis (RCNMA) approach ... **20**
3.1.1 Elastic Network Model (ENM) ... 20
3.1.2 Coarse-graining in RCNMA ... 21

3.2 Normal Mode Simulation (NMSim) approach ... **22**
3.2.1 Mode extension techniques ... 23
C_α direction ... 24
Random direction ... 24
Distance dependent C_α and random direction ... 24
3.2.2 Mode combination techniques ... 26
Linear combination of modes in freely-evolving NMSim ... 26
Linear combination of modes in target-directed NMSim ... 26
3.2.3 Structure distortion in normal mode directions ... 27
3.2.4 Structure correction module ... 28
General overview ... 28
Constraint types and modeling ... 28
Covalent bonds ... 29
Non-covalent bonds ... 30
Steric clashes ... 31
Phi/psi (φ /ψ) modeling ... 31
Rotamer modeling ... 33
Backbone and side-chain planarity and chirality ... 35
Constraint adjustment ... 37
3.2.5 Pathway selection in ROG-guided NMSim ... 38

3.3 Model testing ... **39**
3.3.1 Testing the φ/ψ model on an Ala-6 system ... 40
3.3.2 Testing the rotamer model on lysozyme ... 43

4 Materials and methods ... *45*

4.1 Comparative study of ENM and ED ... **45**
4.1.1 ED modes and protein data set ... 45
4.1.2 RCNMA and ENM parameters used ... 47
4.1.3 ED and CGNM comparison ... 48
4.1.4 Similarities/dissimilarities in classes/folds: ED and ENM modes ... 50

4.2 NMSim and methodological comparisons ... **51**
4.2.1 Analysis of MD, NMSim, FRODA, CONCOORD and experimental HEWL ensembles 51
4.2.2 Rotamer states and derived measures: rotamericity, heterogeneity, and occupancy ... 53
4.2.3 Structure quality using Procheck ... 54

4.3 NMSim and biological applications 54
4.3.1 The proteins in the dataset 54
4.3.2 The NMSim run: parameters and ensemble generation 56

4.4 NMSim and the pathways of conformational change 57

5 Results and discussions 58

5.1 A large scale comparative study of ENM and ED 58
5.1.1 Influence of the reference structure: Average vs. open 60
5.1.2 Influence of the level of coarse-graining: ENM vs. RCNMA 61
5.1.3 Comparison between ED and ENM modes 62
5.1.4 Similarities/dissimilarities in classes/folds: ED and ENM modes 67

5.2 Comparison of the performance of NMSim to other conformation generation methods 71
5.2.1 Residue fluctuations and correlations 71
5.2.2 Conformational space exploration 75
5.2.3 Essential dynamics 78
5.2.4 Side-chain flexibility and rotamers 80
5.2.5 Structure quality using Procheck 85

5.3 Performance of NMSim in exploring biologically relevant conformational changes 87
5.3.1 Domain motions 88
Comparison of essential dynamics between experimental and NMSim structures 89
Intrinsic fluctuations and conformational changes 92
Ligand bound conformations generated from an unbound one 95
ROG-guided trajectory leads to ligand bound conformation 98
5.3.2 Functionally important loop motions 104
Ligand bound loop conformation computed from unbound 104
Intrinsic fluctuations and conformational changes 109

5.4 NMSim and Conformational change pathways 112
5.4.1 Adenylate kinase: a test case 112
5.4.2 NMSim generated pathways using Close directed and ROG-guided simulations 112

6 Summary 117

Zusammenfassung 122

Outlook 128

Acknowledgements 129

Appendix 131

Bibliography 141

Curriculum vitae 156

1 Introduction and aims

Macromolecules are dynamic, and their motions are critical for their functions.[1] The first evidence of a conformational change was reported in 1938 by Felix Haurowitz.[2] His startling discovery showed that native hemoglobin adapts different conformations during and as part of its functional cycle. Since then, many examples illustrating relationship between molecular motions and functions have been reported. For example, conformational changes are required for the functioning of transport proteins,[3,4] catalytic processes of enzymes,[5,6] molecular mechanism of protein regulations,[7-9] and working of motor proteins.[10,11] Important conformational changes upon ligand binding have also been observed in several proteins, e.g., HIV-1 protease,[12] aldose reductase,[13] adenylate kinase,[14-16] tyrosine phosphatase,[17,18] and calmodulin.[19,20] These conformational changes range from side chain fluctuations to reorientations of domains and partial unfolding and refolding.[21,22]

Several different models have been proposed to explain conformational changes upon ligand binding to a protein. Assuming rigid receptor and shape complementarities of the binding partners, "lock-and-key" was proposed in the nineteenth century by Emil Fischer.[23] Later on, it was found incompatible with the evidences of conformational changes observed in binding partners during binding processes. Consequently, the "induced fit" model was proposed[24] to account for the plasticity in receptor proteins. This model assumes that substrate binding induces a conformational change to a receptor. Thus, a geometric fit is ensured only after the structural rearrangement of the receptor caused by the binding interactions. However, the extent to which the conformational changes are literally induced is questionable. For example, Bosshard[25] has reported that induced fit is possible only if the match between the interacting sites is strong enough to provide the initial complex enough strength and longevity so that induced fit takes place within a reasonable time. In recent years, the "conformational selection/preexisting equilibrium" model[26-29] has emerged as an alternative for induce-fit. Here, it is proposed that proper conformations are "picked" by a ligand from the ensembles of rapidly interconverting conformational species of the unbound molecules. This is supported by experimental evidence for the presence of conformational variability of binding partners prior to their association.[30,31] Furthermore, it explains as to why a single protein can bind multiple unrelated ligands at the same site.[32]

Despite the conceptual differences between “induced fit” and “conformational selection”, it should be noted that both models at least agree with regard to the statement that in every complex, the conformation of both binding partners has to be a specific one for both to fit. It has also been suggested[33-36] that conformational selection and induced fit are not two mutually exclusive processes and that induced fit requires some prior molecular match to provide sufficient affinity,[25] which is likely provided by a conformational selection mechanism. The question is then to assess the extent of each mechanism. A recent study in this direction investigates the interplay between the two mechanisms and concludes that strong and long-range ligand-protein interactions favor induced-fit mechanism whereas weaker and short-range interactions favor a conformational selection mechanism.[37]

The understanding of ligand binding and mechanisms of conformational changes is important in the development of structure-based drug design (SBDD).[38-40] Initially, SBDD approaches relied on the validity of the “lock and key” model,[41] although this assumption leads to clear limitations.[40,42,43] There are considerable efforts nowadays to incorporate the influence of (changes of) protein flexibility and mobility into recent drug design approaches.[38,39,44] These efforts are grounded on the “induced-fit” and “conformational selection” models of ligand binding to proteins. In these lines, incorporating protein mobility information, in terms of multiple structures from X-ray, NMR or MD simulations, has been proven to enhance protein-protein docking,[40,45,46] protein-ligand docking [47-49] and pharmacophore models.[50]

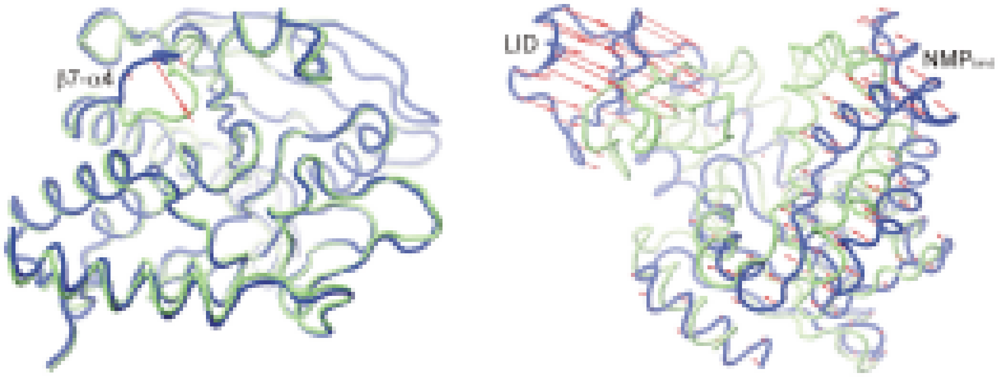
It is important to mention that one needs to distinguish between two different but related concepts, i.e., flexibility and mobility, in order to understand and model conformational changes. Flexibility is a *static* property that only determines the possibility of a motion, whereas nothing actually moves.[51] Mobility in turn describes actual movements in terms of *directions* and *amplitudes*. Flexibility is not necessarily a prerequisite for mobility, as rigid parts of a biomolecule (e.g., domains or helixes) can well move as a whole when connected by hinges. However, mobility provides the origin for receptor plasticity, which enables binding partners to conformationally adapt to each other.

Knowledge about protein mobility can be obtained from different experimental approaches.[52] X-ray crystallography is the major source of structural information; however, it provides the static picture of a single conformation.[53] The underlying protein dynamics can be interpreted using B-factor values or using multiple conformations crystallized in different conformational states. This is, however, restricted to a limited conformational space due to a limited number of available conformations.[54] By contrast, NMR spectroscopy usually provides more direct dynamics information, for example in terms of order parameters and relaxation rates; however, it is restricted to proteins of a limited size.[55]

Different computational approaches targeting the modelling of protein flexibility and plasticity are promising in this context. Molecular dynamics (MD)[56-58] simulation is one of the most widely applied and accurate computational techniques currently being used. However, despite immense increase in computer power, MD simulations are computationally expensive and explore limited conformational space due to slow barrier crossing on the rugged energy landscape of macromolecules.[59,60] Therefore, the MD approach provides only a restricted solution to the challenges posed by protein plasticity in SBDD, for example in generating multiple conformations for flexible docking or high throughput docking approaches.[40,61]

Hence, there have been efforts to develop alternative approaches that are computationally efficient in exploring conformational space. For example, a simple geometry-based approach CONCOORD generates conformations by satisfying distance constraints derived from a stating structure of proteins.[62,63] Another, geometry-based approach FRODA generates conformations by diffusive motions of flexible regions and rigid clusters of proteins.[64] In contrast to MD, these approaches do not provide the time evolution of the molecular movements. However, these approaches are promising due to their efficiency and applicability in SBDD.[65,66] So far, these geometry-based approaches do not use any directional guidance for sampling the biologically *relevant* conformations, which can be helpful, taking into account the complexity of conformational space available to macromolecules.

Coarse-grained normal mode (CGNM) approaches, e.g., elastic network model (ENM) and rigid cluster normal mode analysis (RCNMA), have emerged recently. They provide the directions of intrinsic mobility of biomolecules in terms of harmonic modes (also called normal modes).[67,68] These normal modes can be viewed as possible deformations of proteins and can be sorted by their energetic costs of deformations. More importantly, in agreement with the "conformational selection" model, the conformational changes upon ligand binding of many proteins have been found to occur along a few low-energy modes of unbound proteins calculated using CGNM approaches.[67-71] For example, the directions of conformational changes in tyrosine phosphatase and adenylate kinase upon ligand binding overlap with one of the low-energy modes of the corresponding unbound conformations calculated by the RCNMA approach, as shown in Figure 1.1.[68] Furthermore, the calculations of these modes only take seconds for these proteins and, therefore, can be applied to large macromolecules as well as can be applied iteratively. Realizing the potential of these CGNM approaches, different approaches have utilized these directional information, e.g., in steering MD simulations,[72-74] incorporating receptor flexibility in docking approaches,[75-77] flexible fitting of molecular structures,[78-81] and efficient generation of pathways of conformational changes.[82-84]



a) **b)**

Figure 1.1: Superimposition of open (blue) and closed (green) conformations of tyrosine phosphatase (panel a) and adenylate kinase (panel b). In addition, the amplitudes and directions of motions as predicted by the modes most involved in the conformational changes, respectively, are depicted as red arrows. In both cases, the amplitudes of the motions were scaled for best graphical representation (Figure adopted from Ahmed et al.[68]*).*

Assuming that the low-energy deformation directions of proteins obtained from these CGNM approaches can be helpful in exploring the intrinsic mobility of proteins, the following aims were set for this thesis:

- To validate the directional information obtained from the CGNM approaches on a large dataset of proteins and to study the strengths and limitations of these approaches in capturing the essential motions of proteins.

- To design and develop an efficient geometry-based approach (termed NMSim), utilizing the directional information from a CGNM approach for exploring the intrinsic mobility of proteins.

- To compare and study the usefulness and limitations of different geometry-based approaches, i.e., NMSim, FRODA, and CONCOORD.

- To study the usability of the NMSim approach in exploring the intrinsic mobility of proteins, and in describing ligand induced conformational changes and conformational change pathways.

Keeping these aims in perspective, a large-scale comparative study is performed between principal directions of proteins observed in MD simulations[85,86] and normal modes obtained from CGNM approaches for a large dataset of 335 diverse proteins in section 5.1. A multi-scale approach, termed Normal Mode based Simulation (NMSim), is then developed in this study (chapter 3). The idea behind is to incorporate directional information in a geometry-based simulation technique, in order to sample biologically relevant conformational space, which distinguishes this approach from the previously reported geometry-based simulation approaches CONCOORD[62]and FRODA.[64] In order to analyze the usefulness and the limitations of the different geometry-based approaches, in general, and the NMSim approach, in particular, a methodological comparative study is performed on hen egg white lysozyme in section 5.2. The applicability of the NMSim approach for describing ligand-induced conformational changes is presented in section 5.3. Furthermore, NMSim-generated conformational change pathways from the apo structure to the ligand bound structure of adenylate kinase are compared with previous studies[87-89] and the different crystal structures which lie along the generated pathway are identified in section 5.4.

2 State of the art

The dynamics of biological macromolecules have been of considerable interest since internal motions of proteins were recognized[1] to play an important role in protein function. Different computational approaches targeting the modelling of protein flexibility and mobility are promising in this context. These include force field-based methods, like molecular dynamics;[56] harmonic analysis-based methods like standard normal mode analysis,[90] and elastic network models;[67] and graph-theoretical and geometry-based methods, like FIRST,[91] ROCK,[92] FRODA,[64] and CONCOORD.[62]

2.1 Molecular dynamics (MD)

Molecular dynamics (MD) simulation is one of the most widely applied and accurate computational techniques currently being used in the field of macromolecular computation.[56-58] MD simulation is based on Newtonian dynamics, where instantaneous forces present in the molecular structure are numerically integrated to generate a trajectory through phase space.[60] MD simulations are computationally expensive and limited to the nanosecond or microsecond timescale for most of the systems.[57,58] MD has been applied to a variety of applications, for example, protein folding,[93,94] structure based drug design,[95-97] protein-protein interactions,[98] and protein design.[99]

MD has been utilized successfully for the investigation of receptor plasticity, consequently enhancing structure based drug design (SBDD). For example, a recent MD study[100] of HIV-1 integrase showed an intermittent opening of an unknown favorable binding trench adjacent to the catalytic site, which was experimentally validated later on.[101] Subsequent docking studies of novel ligands with the potential to bind to both regions showed greater selectivity when interacting with the trench.[100] Similarly, dynamic pharmacophore models to compensate for the inherent plasticity of an active site have been developed derived from MD conformations.[50]

Improvements in MD simulation techniques and increased computational power have recently allowed performing MD simulations on unbound protein states that clearly show a potential for generating conformations that mimic bound states. These conformations may well be used subsequently in flexible docking approaches. For example, an artificially low solvent viscosity used in a MD simulation of HIV-1 protease enabled a comprehensive sampling of the conformational space, which shaded light on the flap dynamics of the protein.[102] Although the overall dynamics of the unliganded protease was found to be predominantly populated by semiopen conformations, with closed and fully open structures being a minor component of the overall ensemble, these results strongly support the "conformational selection" model. In another MD study[103] starting from the unbound form of aldose reductase a set of distinct conformational substates that may prove useful as alternative structural templates in virtual screening/docking for new inhibitors was identified. Along these lines, 41 proteins that form protein-protein complexes have been simulated in order to investigate the extent to which conformational fluctuations lead to novel conformational states.[104] Starting again from the unbound structures, it was found that fluctuations take some parts of the molecules into regions of conformational space closer to a bound state, although simulation times of 5 ns were not sufficient in any case to sample the complete bound state.

Atomic MD simulations provide a detailed picture of the dynamics of biomolecules. However, due to the requirement to choose integration time steps on the order of 1 fs, it is computationally expensive and impractical to reach long time scales (> 1 microsecond) for large and complex systems.[105] To deal with this limitation, coarse-grained models have been developed to study large systems, which enables the use of longer time steps (e.g., ~40 fs).[106] For example, a one-bead model (each amino acid is represented as a single particle) has been applied to study the ribosome, revealing the principal direction of motions and the correlations between these motions.[107]

Several efforts have also been made to overcome the problem of restricted sampling in MD due to slow barrier crossing over the rugged energy landscape of biomolecules.[108,109] For example, these includes conformational flooding,[110] replica-exchange molecular dynamics (REMD),[111,112] and targeted molecular dynamics

(TMD).[113,114] However, these methods still lack the required efficiency needed for high throughput approaches.[61]

2.2 Normal mode analysis (NMA)

Normal mode analysis (NMA) alternatively provides an analytical description of a dynamic system. It was first applied to proteins in the early 1980's.[90] NMA is a harmonic analysis that assumes that, over the range of thermal fluctuations, the conformational energy surface can be characterized by a parabolic approximation to a single energy minimum.[115] It starts with creating a harmonic potential well at a local minimum and then finding all possible harmonic modes within this potential well. For molecules, this is usually accomplished by taking an experimentally determined atomic configuration (usually obtained from the Protein Data Bank). The potential energy of the molecule can be calculated for this structure, using well established force fields. Having reached a stationary point after energy minimization, the potential energy surface is then approximated by a parabola, where the shape of the parabola is defined by the Hessian matrix. The elements of this matrix are the second derivatives of the potential energy function with respect to the coordinates of the system. Normal modes are then obtained by diagonalizing the $3N$-dimensional Hessian matrix for the system of N atoms. Each normal mode represents the direction of vibration and the relative displacement of the atoms in that mode. Therefore, it is also termed harmonic or vibrational mode. Each mode is orthogonal to all others, which greatly simplifies the analysis of motion. Every atom in a normal mode vibrates with the same frequency, which represents the energetic cost of displacing the system by one length unit along the eigenvector direction. Hence, normal modes can be viewed as possible deformations of a protein according to their energetic cost, where low-energy deformations correspond to collective or delocalized deformations and high-energy modes are localized deformations. The $3N$ normal modes obtained from the analysis also include the 6 global motions (three translations and three rotations), having no energetic cost. They are usually of no interest and ignored in the analysis. It has also been shown that mostly the lowest-frequency (energetic cost) modes (having frequencies up to 30 cm^{-1}) are responsible for conformational changes and, thus, are considered to be biologically important .[116]

NMA has been successfully applied for the investigation of important conformational changes: For example, to study the hinge-bending motions in human lysozyme[117,118] and citrate synthase,[119] and to study the large-scale conformational changes in allosteric proteins in GroEL chaperonin[120] and aspartate transcarbamylase.[121-123] NMA have also been used as basis vectors for approximate molecular dynamics simulation[124] or refinement of X-ray[125] or NMR data.[126] Furthermore, NMA has also been applied to investigate DNA and RNA dynamics.[127-129] Initially, NMA was applied to only small proteins (approximately up to 500 atoms)[90,130] but advancements in computer hardware and recent efficient approximations to the method now make it possible to analyze large molecular systems as well.[131]

Although the method is straight forward and easy to implement, there are some limitations to it. Despite these limitations, the method seems to work well in describing the conformational changes and predicting internal dynamics.[132] First of all, the method is based on a harmonic approximation of the potential energy surface. However, there are many observations that this approximation breaks down for proteins at physiological temperatures, i.e., by crossing energy barriers of various heights and visiting multiple minima.[133] Even if the energy minimum of a single conformation is considered representative of the motion within all energy minima (as appears to be the case),[134,135] barrier crossing events would be expected to have an even greater influence on the overall motion of the molecule, with no obvious relation to the motion within individual minima. In view of this approximation, the relative success of the normal mode analysis is surprising.[136] The second limitation is that the NMA is performed in *vacuo*, whereas the molecule is usually found in solvent, which has a great influence on its dynamic. Typically, proteins are well known to fold and function in water environment, within a narrow range of pH, temperature, and ionic strength. However, NMA ignores the effect of solvent or any other environment around the system.

Apart from these approximations, there are a few practical limitations to the standard NMA while computationally performing the analysis. As an input, the method takes a minimized structure of a protein, which requires an expensive energy minimization of the structure. This method also has the trivial problem of high memory requirement and being computationally slow. These make it impossible to analyze even medium size (i.e., approx. 5000 atoms) proteins on current desktop computers. However, recent efficient approximations to the method now make it possible to analyze molecular system of the size of the whole ribosome on a single desktop computer, which contains approximately 128,000 atoms.[131] However, this approach only considers translation and rotation degrees of freedom for a residue.

2.3 Elastic network model (ENM)

Standard normal mode analysis (NMA), using an atomic force-field representation of the macromolecules, is computationally expensive, which makes it impossible to analyze large proteins with this method. To overcome this limitation, simplified alternatives in the form of elastic network models (ENM) have emerged recently, triggered by the development of simplified force-fields[137] and coarse-grained models of macromolecules: the Gaussian network model (GNM)[138,139] and the Anisotropic network model (ANM)[67,70,140,141] Here, a protein is modeled as an elastic network; the all-atom representation used in NMA is replaced with a reduced representation by considering, e.g., only C_α atoms between which simplified potentials in terms of Hookean springs of equal strength act (Figure 2.1).[137,138] Hence, the system can be seen as a collection of bodies connected by springs of the same strengths. Further coarse-graining in ENM has also been reported assuming structural rigidity based on secondary structure,[142] rigidity of sequentially consecutive residues,[142-144] or using a rigid cluster decomposition by FIRST.[68] That way, the method can be applied to even macromolecular assemblies.[145-148]

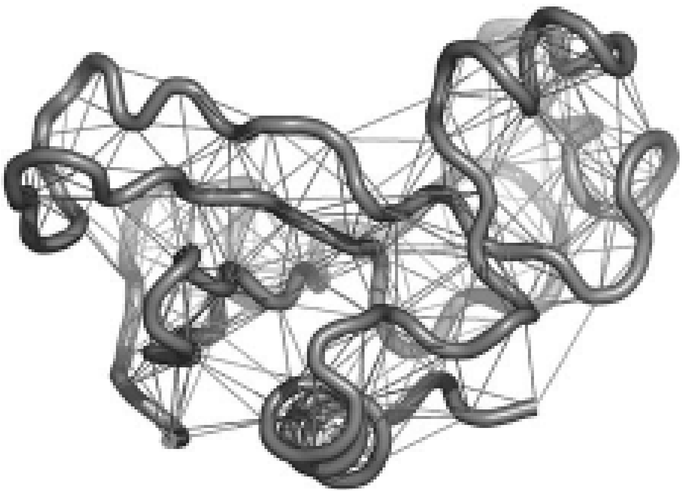


Figure 2.1: ENM representation of barnase. Between C_α atoms (connected by a tube) springs (represented as sticks) of equal strength act (Figure adopted from Gohlke et al.[51]*)*

ENM has been applied to a vast range of problems concerning flexibility/mobility of proteins and other large macro-molecules.[106,136,149,150] In agreement with the "conformational selection model", the conformational change captured by ligands is found for most of the proteins to occur along the lowest energy (frequency) modes calculated by normal mode analysis of the unbound protein. These modes usually involve hinge-bending, large-amplitude, and correlated motions.[70,142] Along these lines ENM has been mostly applied as a *posteriori* analysis in combination with experimental studies, e.g., for examining functional dynamics in *E. coli* adenylate kinase, HIV-1 reverse transcriptase, and influenza virus hemagglutinin,[151-153] cooperative and allosteric dynamics in tryptophan synthase[154] and binding effects in HIV-1 reverse transcriptase.[155] Moreover, several studies showed an efficient conformation and pathway generation by ENM-based techniques, which can be exploited for docking studies.[75,82,83,156]

Apart from a large-scale dynamical analysis, a residue level analysis has also been successfully applied.[157-160] This is surprising, considering the simplicity and coarse-graining of the underlying model. For example, high frequency modes of GNM have been shown to be important for the identification of binding "hot spot" residues,[159] catalytic residues,[158] and protein-binding sites.[157] Catalytic sites were found to be colocalized with global hinge centers predicted by GNM, whereas the ligand binding sites were found to be enjoying flexibility near the catalytic site.[158] In SBDD, these studies can be exploited for efficiently identifying binding hotspot and catalytic residues.

The ENM approach, due to its simplicity and efficiency in predicting large-scale conformational changes, has been found successful when combined with methods that provide atomic detail such as MD. In this respect MD/NMA hybrid methods have been proposed[72-74] that amplify collective motions along normal mode directions in a conventional MD. This method was successfully used for docking in the case of HIV-1 protease.[73]

Loop motions are hard to predict but play an important role in accommodating ligands in binding pockets. Cavasotto et al.[75] introduced a measure of relevance of normal modes to desirable important loop conformational changes upon ligand binding and found that only a few low-frequency modes (< 10 but not usually the first low-frequency modes) are critical and sufficient to represent binding pocket mobility in protein kinases. Using these relevant modes, an ensemble of alternative conformations for holo and apo structures of cAMP-dependent protein kinase, which exhibit backbone rearrangements in two independent loop regions close to the binding pocket, was generated. Considerably improved docking results were observed when docking this ensemble. In my recently work, it was also shown that the coarse-graining of ENM using FIRST can lead to the accurate prediction of loop movements.[68] This can be explained by the fact that the appropriate coarse-graining removes irrelevant modes of the system (without losing the important functional modes), whereas, the modes related to flexible regions became more emphasized.

The success of the ENM approach is based on a reduced protein representation and inherent coarse-graining. This exploits the fact that one is mostly interested in low-frequency modes that are insensitive to atomic level details.[67] Thus, modeling macromolecules at a coarse-grained level instead of an atomic level will still capture the low-frequency motions. This allows predicting surprisingly accurately large conformational changes, which is difficult with force field based methods like MD. However, the ENM approach inherits the same limitations as the standard NMA approach, regarding harmonic approximations and not considering solvent effects.

2.4 FIRST, ROCK and FRODA

Modeling proteins as constraint networks and using graph-theoretical techniques, the flexible regions and rigid clusters in the structures can be identified.[91] This has already been applied for analyzing rigidity in structures of covalent network glasses[161] and engineering structures that consists of struts and joints.[162] For proteins, first, the network corresponding to the protein structure is built such that forces between atoms are transformed into connections between nodes. A fast combinatorial algorithm, the "pebble game", then identifies the flexible (under-constrained), rigid (constrained), and over-rigid (over-constrained) regions by counting bond-rotational degrees of freedom in the network. This algorithm has been implemented into the FIRST (Floppy Inclusion and Rigid Substructure Topology) approach.[91] The outcome of the method is a decomposition of the protein structure into rigid and flexible regions (Figure 2.2). Notably, this approach allows identifying rigid and flexible regions from a single (static) structure in almost no computational time; a FIRST analysis of a molecule of several thousand atoms just takes a few seconds.

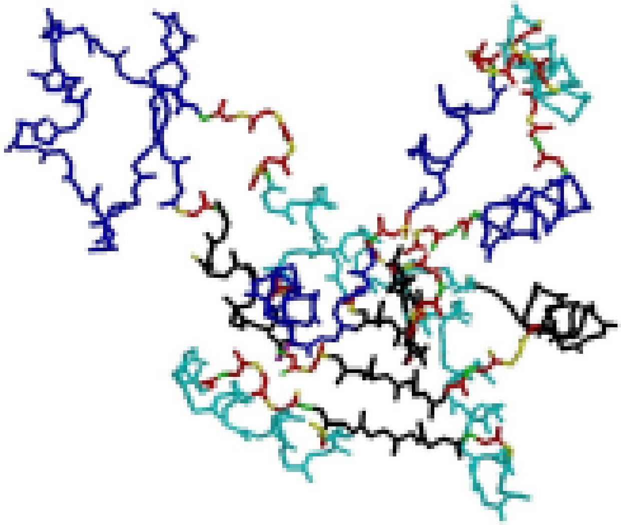


Figure 2.2: Rigid cluster decomposition of adenylate kinase (PDB code: 4ake) obtained from the FIRST approach.[91] Rigid clusters are colored in blue, cyan, black, yellow, red, and green. (Figure adopted from Ahmed et al.[68])

FIRST analyses have been used to accurately identify rigid regions as well as collectively and independently moving regions in a series of proteins.[91,163] An interesting feature of the FIRST analysis is that changes in the flexibility of the binding partners due to complex formation can be investigated in detail. In the case of a protein-protein complex formation,[164] additional interactions across the interface led to a propagation of rigidity through the binding partners. This demonstrates the long range aspect to rigidity percolation. Moreover, the FIRST approach has been applied in combination with MD for investigating the flexibility of prolyl oligopeptidase.[165] Recently, the FIRST approach has also been extended for analyzing flexibility in RNA structures[166] and has been applied to investigate the statics of the ribosomal exit tunnel of large ribosomal subunits.[167]

Flexibility information from FIRST, which leads to a natural coarse-graining of macromolecules based on rigid regions,[51] has been further exploited for simulating protein mobility using constrained geometric simulation.[64,92,168] The ROCK (Rigidity Optimized Conformational Kinetics) approach explores the rigidity-restricted conformational space by satisfying ring closure equations.[92] The FRODA (Framework Rigidity Optimized Dynamic Algorithm) approach makes use of a more efficient algorithm that moves flexible and rigid parts by ghost template rearrangements.[64] FRODA moves flexible parts of a molecule through stereochemically allowed regions of conformational space using random Brownian type (Monte Carlo) dynamics, whereas atoms in rigid clusters are moved collectively.

The ROCK generated structures have been used in flexible docking for the drug targets cyclophilin and estrogen receptor.[169] FRODA has been shown to predict the mobile regions in barnase and qualitatively predict the observed displacements between open and close form in maltodextrin binding protein.[51,64] Docking studies of the multi-subunit protein complex photosystem I, which make use of FRODA conformations and aim at exploring alternative approaching pathways, have also been reported.[66] Furthermore, FRODA has recently been used to flexibly fit an X-ray crystal structure of the bacterial chaperonin GroEL to two different cryo-EM maps.[170]

The ROCK and FRODA approaches completely rely on the flexibility information provided by the FIRST approach in order to explore the conformational space of proteins. In cases where proteins are relatively flexible, these approaches may not be efficient or may not capture the conformational space available to the proteins.

2.5 CONCOORD

CONCOORD (from CONstraint to COORDinates) is another geometry-based method that generates conformations by satisfying constraints.[62] Starting from a random structure, conformational space is captured by fulfilling a set of upper and lower interatomic distance bounds that are derived from the experimental structure of the protein. The differences between upper and lower distance bounds depend on the strengths of interactions, with stronger interactions leading to smaller deviations. Repeating this correction procedure several times leads to an ensemble of structures as a representation of the conformational space, which takes only a few hours of CPU time.

The novel use of CONCOORD generated structures has been to get eigenvectors of essential dynamics; whether it is docking to multiple eigenstructures,[65] analyzing conformational changes in macromolecular assemblies,[171] or exploring different biological mechanisms.[172-175] CONCOORD can generate conformations very efficiently; therefore, it is well suited for larger systems. In the case of hyaluronate lyase,[172] whose size precludes the application of MD to investigate biologically relevant time scales, flexibility (allosteric) information and functional implications were derived from CONCOORD. Two ED modes of motion were identified: the first motion describes an opening and closing of a catalytic cleft, and the second motion demonstrates the mobility of a binding cleft, which may facilitate the binding of the negatively charged hyaluronan to the enzyme. Mustard and Ritchie[65] showed that docking to multiple eigenstructures (obtained by an ED study following a CONCOORD run) generates better docking predictions than docking only to unbound or model built structures.

In a CONCOORD-generated ensemble, each structure is independent from the previous one. On the one hand this implies that no information is obtained about the path along which two conformations are connected and possible energy barriers between them. On the other hand, this approach enables crossing of even high energy barriers and finding other possible conformations. Hence, the CONCOORD approach does not suffer from a sampling problem. However, the sampling completely relies and is sensitive to the inter-atomic distances of the starting structure. Therefore, CONCOORD may not be suitable for large-scale conformational transitions which require change in the distance constraint network (e.g., due to making or breaking of hydrogen bonds). Realizing this limitation, recently, a reimplementation of the original CONCOORD[62] approach has been reported which allows the prediction of conformational transitions as well and therefore has been termed as tCONCOORD.[63] This approach rests on an estimate of the stability of interactions observed in a starting structure, in particular, those interactions that change during a conformational transition.

3 Theory and implementation

Recently, coarse-grained normal mode approaches based on elastic network theory[67,137] have emerged as efficient alternatives for investigating large-scale conformational changes.[136,149,150] Different studies[71,116] have shown that the low frequency modes, which are also found to be involved in functionally important conformational changes of proteins, are robust and insensitive to higher coarse-graining of the elastic network.[143,176,177] Pursing this direction, some high-coarse-graining strategies have been proposed recently.[142-144] RCNMA[68] was proposed to achieve high coarse-graining level, by identifying rigid clusters in protein structures using the FIRST approach[91] and subsequently assuming no internal motion in those rigid clusters, without loosing accuracy.

In this study, a three-step approach for multi-scale modeling of macromolecular conformational changes is developed to further utilize the low frequency modes from RCNMA in order to sample low energy conformational space. The first two steps are based on recent developments in rigidity and elastic network theory.[68] Initially, static properties of the macromolecule are determined by decomposing the macromolecule into rigid clusters using the graph-theoretical approach FIRST[91] at an all-atom representation of the protein. In a second step, dynamical properties of the molecule are revealed by the rotations-translations of blocks approach (RTB)[178] using an elastic network model representation of the coarse-grained protein, i.e., in this step, only rigid body motions are allowed for rigid clusters while links between them are treated as fully flexible.

In the final step, the recently introduced idea of constrained geometric simulations of diffusive motions in proteins[64] is extended. New macromolecule conformers are generated by deforming the structure along low-energy normal mode directions predicted by RCNMA plus random direction components. Here, backbone motions are biased in the low frequency normal mode space, and side-chains have attractive basins derived from a rotamer library.[179] The generated structures are then iteratively corrected regarding steric clashes or constraint violations. This module is termed NMSim. Constraints to be satisfied include torsions of the main and side-chains, distances and angles of covalent and non-covalent interactions such as hydrogen bonds or hydrophobic interactions and the preservation of planar groups. In total, when applied repetitively over all three steps, the procedure generates efficiently series of conformations that lie preferentially in the low energy subspace of normal modes. The pictorial overview of RCNMA/NMSim approach is shown in Figure 3.1.

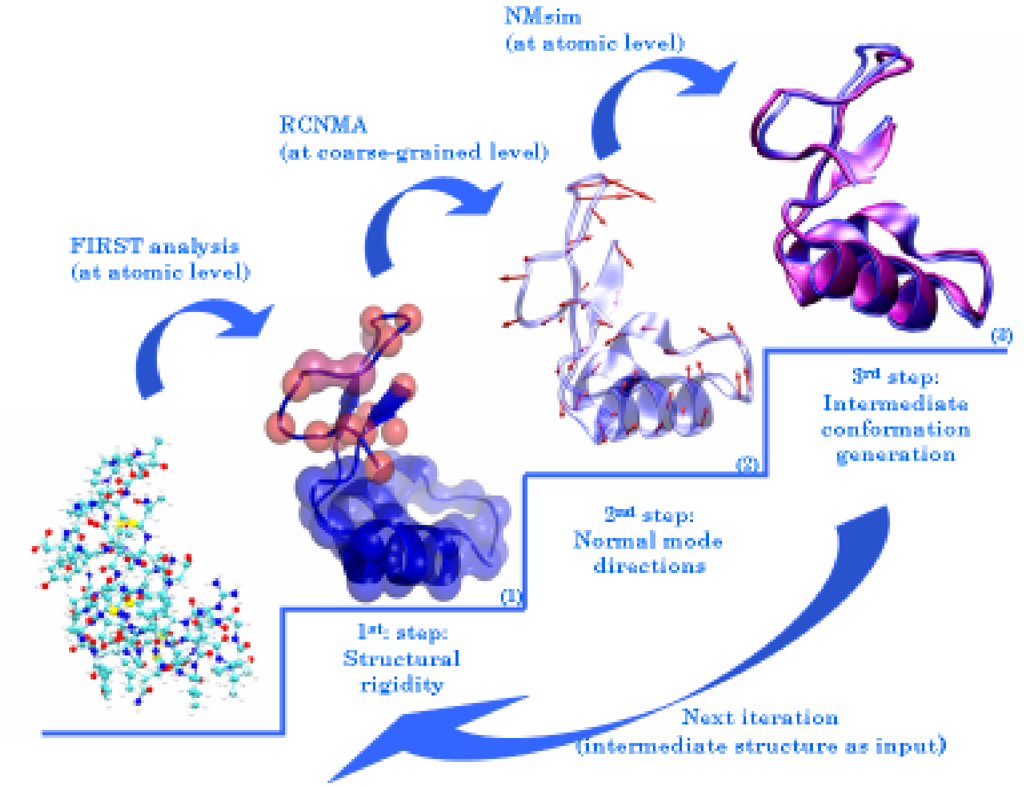


Figure 3.1: Overview of the RCNMA/NMSim approach. In the first step, the FIRST analysis is applied, which provides the rigid cluster decomposition (RCD). In the second step, the RCD is utilized by RCNMA for the calculation of coarse-grained normal mode directions. In the third step, these normal mode directions are then used by the NMSim approach to generate stereochemically allowed conformations. In order to generate an NMSim trajectory, step two and three are repeated using the previously generated structure.

3.1 Rigid Cluster Normal Mode Analysis (RCNMA) approach

The basic idea behind the RCNMA approach[68] is the use of structural flexibility/rigidity information of a molecule prior to the prediction of its dynamic behavior. This is done by a two-step modeling approach. In the first step, a flexibility analysis is performed using a graph theoretical technique, which uses an all atom representation of the protein.[91] In the second step, the information of block formation obtained form the previous step is used to generate a coarse grained model as input for the Block Normal Mode (BNM) approach.[178] A rigid cluster is modeled as a block whereas flexible regions are modeled as fine-grained (one-residue per block). In addition, an elastic network model (ENM) representation is used for the normal mode calculations. An overview of RCNMA is given in Figure 3.1.

3.1.1 Elastic Network Model (ENM)

ENM has been successfully applied to the calculations of coarse-grained normal modes.[136,149,150] Here, based on a simplified representation of the potential energy,[137,138,180] the proteins are described as 3D elastic networks. Each amino acid, i.e., usually the C_α atom, acts as a junction in the network. Interactions between these particles are modeled by Hookean springs based on a harmonic pairwise potential,[137] resulting in a total potential energy of the system given by

$$V = \frac{\gamma}{2}\sum_{i}\sum_{j}\theta\left(r_c - r_{ij}^0\right)\left(r_{ij} - r_{ij}^0\right)^2 , \tag{3.1}$$

where r_c is the cutoff up to which interactions between the C_α atoms are taken into account. r_{ij} and r_{ij}^0 are the instantaneous and equilibrium distances between atoms i and j, respectively. $\theta(x)$ is the Heaviside step function that accounts for the cutoff effect of the interaction; it is 1 if $x > 0$ and 0 otherwise. γ is a phenomenological force constant assumed to be the same for all pairwise interactions.

According to the elastic network model,[67] the elements of a $3N \times 3N$ Hessian matrix H (where N is the number of C_α atoms) are then obtained from the second derivatives of

V with respect to the Cartesian coordinates of atoms i and j. H is then diagonalized to obtain the normal modes.

3.1.2 Coarse-graining in RCNMA

The RCNMA approach [68] adds another level of coarse-graining to ENM by identifying rigid clusters and flexible regions within protein structures using the FIRST approach.[91] The all-atom representation of proteins needed for the FIRST analysis is reduced to a C_α-only representation in the next step. Each rigid cluster, obtained from FIRST approach, forms a block in the subsequent rotations and translations of block (RTB) approach,[142,178] and flexible regions are modeled on a one-residue-per-block basis (in which case only translational motions of the "block" are considered). Interactions between these blocks are modeled as in ENM (Eq. 3.1). The $3N\times3N$ matrix H is therefore reduced to a $6n\times6n$ dimensional matrix H_{sub} by projecting H into the subspace spanned by translation/rotation basis vectors of n blocks according to:

$$H_{sub} = P^t HP, \tag{3.2}$$

with P being an orthogonal $3N\times6n$ projection matrix of the infinitesimal translation/rotation eigenvectors of each block. This leads to a reduction of the memory requirement proportional to $(N/n)^2$ and computational time proportional to $(N/n)^3$, respectively. Diagonalization of the resulting matrix H_{sub} yields the normal modes U_{sub} and eigenvalues Λ:

$$H_{sub}U_{sub} = U_{sub}\Lambda \tag{3.3}$$

Finally, atomic displacements can be obtained by expanding back the eigenvectors U_{sub} from the subspace spanned by translation/rotation basis vectors of the blocks to the Cartesian space (U).

$$U = PU_{sub} \tag{3.4}$$

The $3N\times 6n$ dimensional U matrix, thus obtained, contains $6n$ normal modes $\vec{C}^k$. The k^{th} normal mode direction for j^{th} C_α atom is given by $\vec{C}_j^k = \left[U_{x,k}, U_{y,k}, U_{z,k}\right]$ where $x=3*j-2$, $y=3*j-1$, and $z=3*j$.

3.2 Normal Mode Simulation (NMSim) approach

By combining RCNMA with geometric simulation techniques, a multi-level approach termed Normal Mode based Simulation (NMSim) was developed in this study which was then used for efficient generation of macromolecular conformations. Here, backbone motions are biased in the low frequency normal mode space, and side-chains have attractive basins derived from a rotamer library. An efficient constraint correction approach is applied to generate stereo-chemically allowed conformations. In addition to covalent and non-covalent bonds like hydrogen bonds and hydrophobic interactions, φ/ψ favorable regions are also modeled as constraints.

A schematic diagram of the whole procedure is shown in Figure 3.2. The procedure starts with the structural rigidity analysis of the input protein structure (in PDB format and protonated using the program Reduce[181]) by the FIRST approach,[91] which defines a rigid cluster decomposition (RCD) and a covalent/non-covalent bonded network. The RCNMA module is used to calculate normal modes for the input structure, and the NMSim module is used to generate stereo-chemically allowed conformations based on the input parameter set, the input structure, the calculated normal modes, and the bond network. The NMSim module initially distorts the structure in the low frequency normal mode space, and a stereo-chemically allowed conformation is then generated using that distorted structure (in the structure correction module).

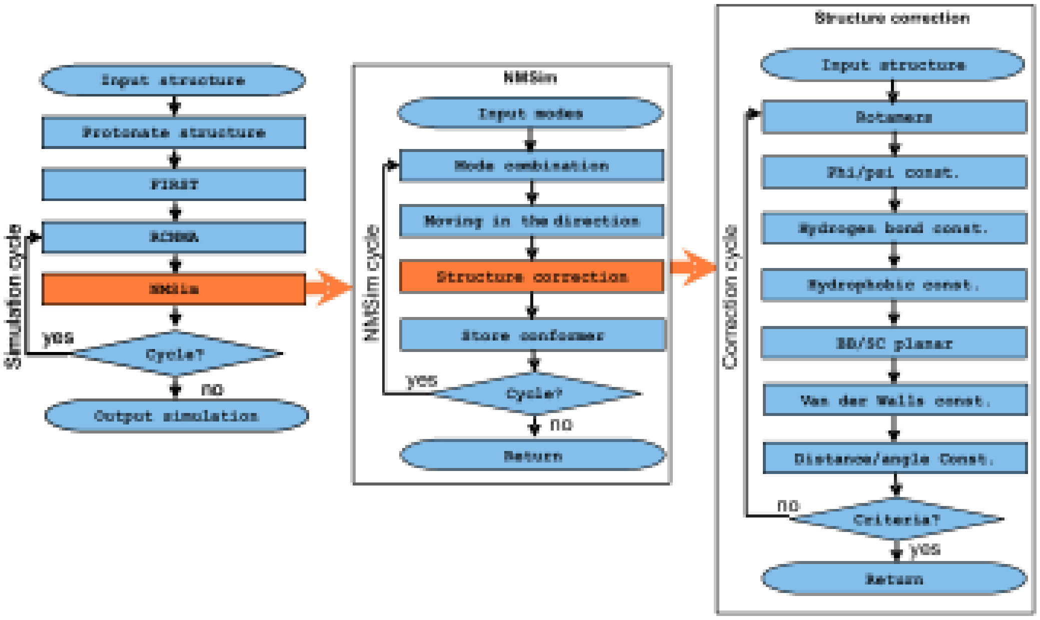


Figure 3.2: A scheme showing the program flow and the different modules of the NMSim approach. The modules in light orange color are further expended on their right. Here, BB stands for backbone, and SC stands for side-chain.

The RCNMA and NMSim modules are alternatively called in simulation cycles. In each RCNMA call, a new set of normal modes are calculated using the previously generated structure in the NMSim module. In each NMSim call, the input set of normal modes is used to generate multiple structures iteratively in NMSim cycles using different linear combinations. In this section, the different components in the NMSim module are explained in detail.

3.2.1 Mode extension techniques

As described above, normal modes from RCNMA/ENM give the direction for C_α atoms only. To move all atoms of the structure, some directions should be given to the remaining non C_α atoms. This extension of the C_α based modes to all-atoms based modes is accomplished by merging two different techniques, called C_α direction and random direction. The idea behind is that the side-chains are allowed to randomly explore the stereo-chemically allowed conformational space whereas the backbone motions are directed in the normal mode space.

C_α direction

Since C_α atom is the representative of whole residue in RCNMA/ENM based modes, it is a good approximation to use the representative C_α direction for all atoms in that residue. However, using this approach limits the side-chain mobility because of the lack of internal movements in the side-chains.

Random direction

The alternative to the C_α direction approach is to use random directions for all non C_α atoms so that side-chains can randomly sample internal motions. However, this would be another extreme, since side-chain positions are also dependent on the backbone motions. Therefore, a combination of C_α direction and random direction would be need for modeling side-chain distortions.

Distance dependent C_α and random direction

The extension of C_α based modes to all-atoms based modes is modeled by merging the above two approaches. As a criteria for mixing the distance of atoms from their representative C_α atoms is used, i.e., the atoms in a side-chain that are closer to their representative C_α atom have large C_α direction components, whereas, the atoms at the tail region of a side-chain have large random components. This distance dependent mixing assures a smooth transition of directionality from C_α to random direction, such that the side-chain conformations can be randomly explored in the stereo-chemically allowed space, and the backbone conformations can be explored in the normal mode space.

To model the above concept, a random unit vector $\vec{R}_i$ for every atom i in 3-dimensional space is generated and scaled by the magnitude of the representative C_α atom direction $\vec{C}_j$ of residue j plus a random component, which is controlled by the parameter RANDSCALING. The resulting random vector $\vec{E}_i$ for every atom i is given by

$$\vec{E}_i = \vec{R}_i * \left((rand * RANDSCALING) + \left| \vec{C}_j \right| \right), \tag{3.5}$$

where, $rand$ is a uniformly distributed random number between 0 and 1, and the default value for RANDSCALING is empirically set to 0.3 Å. Increasing this value causes higher fluctuations in the side-chain regions. The representative C_α atom direction $\vec{C}_j$ of residue j is a normal mode direction as calculated in Eq. 3.5.

A distance dependent weighting factor F_i is used to merge the random direction and the C_α direction of each atom. The F_i is calculated for every atom i of residue j by calculating the distance D_i between atom i and the C_α atom of j, and then normalizing with the maximum distance $D_{\max}$ found in residue j.

$$F_i = \frac{D_i}{D_{\max}} \tag{3.6}$$

Finally, the all-atom normal mode vector $\vec{\mathrm{P}}_i$ for every atom i in residue j is obtained by linearly mixing its representative C_α normal mode direction $\vec{C}_j$ with the random vector direction $\vec{E}_i$ using the distance dependent weighting factor F_i.

$$\vec{P}_i = F_i * \vec{E}_i + (1 - F_i) * \vec{C}_j \tag{3.7}$$

For the representative C_α atom, the weighting factor F_i in Eq. 3.6 is zero and thus no random component is added. For the atom which is farthest away in the residue j from its representative C_α atom, the weighting factor F_i is one in Eq. 3.6 and thus no C_α direction component is added. This procedure is repeated for each mode k and, thus, k all-atom based normal mode vectors $\vec{P}_{i,k}$ are obtained.

3.2.2 Mode combination techniques

Linear combination of modes in freely-evolving NMSim

All-atom based normal mode vectors $\vec{P}_{i,k}$ obtained for each mode k and atom i in 3-dimentional space are linearly combined. The coefficients of the linear combination of $\vec{P}_{i,k}$ vectors are the ratios of a random number O_k to a factor ω_k. The resulting normal mode linear combination vector $\vec{V}_i$ is defined as

$$\vec{V}_i = \sum_{k=7}^{m} \frac{O_k}{\omega_k} \vec{P}_{i,k} \, , \tag{3.8}$$

where O_k is a uniformly distributed random number between -1 and 1, and ω_k is related to eigenvalues Λ_k (as calculated in Eq. 3.3) by $\omega_k = \sqrt{\Lambda_k}$. The low-frequency normal modes are used for the linear combination (default $m = 56$, unless stated explicitly), ignoring first 6 zero-frequency normal modes.

Normal modes are harmonic and can have positive or negative phase (which is not known). Therefore, the sign of a random number O_k assigns the missing phase to a normal mode, whereas the magnitude of a random number emphasizes/de-emphasizes a normal mode randomly in the linear combination. Hence, during the freely-evolving NMSim, each trajectory follows a different path in the low-frequency normal mode space. In addition, the normal modes are emphasized based on their energy of deformation using ω_k, which gives highest weight to the lowest frequency mode and the second highest weight to the second lowest frequency mode and so on.

Linear combination of modes in target-directed NMSim

Since normal modes are harmonic and decoupled, low frequency normal modes are linearly and randomly combined in NMSim to explore the low energy conformational space. This results in a random walk behavior in stereo-chemically allowed low energy space. If the target structure is known, then a pathway leading to the target structure can be traced either by using the best linear combination or by selecting the best overlapping mode with the conformational change direction. It is important to note however that the pathway to the target structure is still restricted in the space spanned by the low frequency normal modes. This type of simulation is termed as target-directed NMSim.

In target-directed NMSim, the conformational change vector $\Delta\vec{r} = \vec{r}_c - \vec{r}_o$ is used to guide the trajectory towards the target structure $\vec{r}_c$ from the starting/intermediate structure $\vec{r}_o$. The vectors $\vec{r}_c$ and $\vec{r}_o$ are the C_α atomic coordinates of the two different conformations. The coefficient O_k for each mode k is calculated by the scalar projection of the conformational change vector $\Delta\vec{r}$ onto the normal mode vector $\vec{C}^k$.

$$O_k = \left(\Delta\vec{r} \cdot \vec{C}^k\right) \tag{3.9}$$

Subsequently, the coefficient O_k of each mode k is either used to select the best overlapping mode or to calculate the target guided linear combination vector $\vec{V}$ in Eq. 3.8.

3.2.3 Structure distortion in normal mode directions

The current structure in each iteration is distorted in low frequency normal mode space using the linear combination vector $\vec{V}$. The magnitude of $\vec{V}$ is adjusted which accounts for the step-size in NMSim. In geometric simulations RMSD can be used as a step-size of a trajectory. The parameter RMSDSTEPSIZE (in Å, see Appendix A) is used for scaling $\vec{V}$ in NMSim. This can be achieved by

$$\vec{Q} = RMSDSTEPSIZE * M^{1/2} * \frac{\vec{V}}{\left|\vec{V}\right|}, \tag{3.10}$$

where, M is the number of atoms in the structure. The current structure when distorted with the displacement vector $\vec{Q}$ causes the distortion of RMSDSTEPSIZE. And thus the distortion in the structure is constant at every NMSim cycle. The default value for RMSDSTEPSIZE is set to 0.5 Å.

3.2.4 Structure correction module

General overview

Studies from ultrahigh resolution crystallography of small molecules have shown strict equilibrium values for bond lengths and angles between constituent atoms of amino acids.[182] The principal degrees of freedom in proteins arise from the dihedral angles, which show a pattern of preferences. For example, φ/ψ dihedral angles show preferences in different regions of the Ramachandran map,[183,184] χ-angles show preferences in terms of different rotamer states,[179] and backbone and side-chain planar groups have strict dihedral angles. Moreover, hydrogen bonds, salt bridges and hydrophobic interactions further restrict the available degrees of freedom in a protein. All these factors need to be considered in a geometry-based structure correction approach.

Constraint types and modeling

Distortions in an intermediate structure, caused by moving atoms in the normal mode directions with random components, are efficiently corrected using the geometry-based constraints correction approach. A network of constraints is built from the protein bonding network where different chemical bonds are modeled as constraints. In addition to covalent and non-covalent bonds φ/ψ favorable regions are also modeled as constraints. For χ-angles, a knowledge-based approach is applied by forcing side-chains into the closest favorable rotamer state during structure correction. Backbone and side-chain chirality and planarity are ensured and steric clashes between atoms are corrected.

Three different types of constraints are used to model the above mentioned chemical bonds and properties: distance, dihedral and planar constraints. Most of the constraints are distance based, which was the preferred type for modeling due to its simplicity and efficiency in correction. All covalent bonds, non-covalent bonds, steric clashes, as well as φ/ψ dihedrals are modeled as one or a combination of distance constraints. These constraints are corrected based on equality, lower limit (as in steric clashes) or upper limit (as in hydrophobic constraints) of the ideal distances. To model the strength of the different interactions or the variability of the different dihedral angles, the model is empirically parameterized for different adjustment factors of the constraints as given in Table 3.1. The adjustment factor is the strength to which constraints are restored during the correction cycles. For rotamer and backbone/side-chain planarity the dihedral and the planar constraint types are used, respectively. A dihedral constraint satisfies a specific dihedral angle by rotating atoms around dihedral bonds. A planar constraint moves all atoms of the disturbed side-chain/backbone planar group towards an imaginary superimposed plane.

Covalent bonds

All covalent bonds (single bond, double bond, or disulphide bridges) in a protein are recognized and modeled as distance constraints between the covalently bonded atoms. Additionally, all possible angles (1-3 connections) in the covalent bond network are recognized and modeled as distance constraints. Ideal distances for distance constraints are taken from the input structure assuming a valid input structure. A covalent bond network of distance constraints for an Ala-3 system is shown in Figure 3.3.

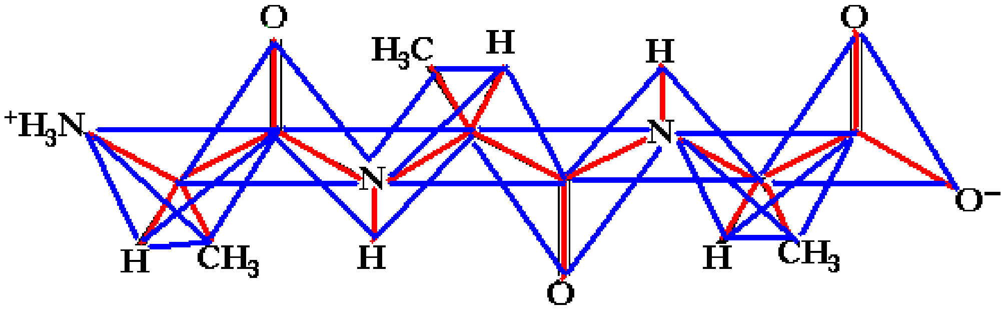


Figure 3.3: A covalent bond network of distance constraints for an Ala-3 system. Covalent bonds (red) and bond angles (blue) are modeled as distance constraints.

Non-covalent bonds

Non-covalent bonds are modeled explicitly and include hydrogen bonds, salt-bridges, and hydrophobic interactions. These are recognized from the input starting structure at the beginning of the program using the FIRST approach[91] and kept throughout the simulation (assuming no breaking or making of bonds during the simulation).

Each hydrogen bond (salt-bridges are modeled similarly) is modeled by three distance constraints: between donor and acceptor, neighboring acceptor and donor, and neighboring donor and acceptor atoms involved in the hydrogen bond (see Figure 3.4). It is important to note that, in general, hydrogen atoms are not considered in the NMSim simulations for efficiency reasons. These constraints ensure that no hydrogen bond breaks or weakens but allows rotations around the D-A constraint.

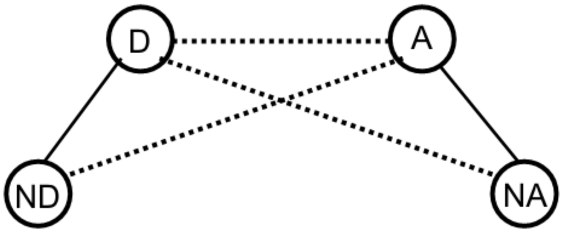


Figure 3.4: A hydrogen bond is modeled using three distance constraints (doted lines) between related atoms. Covalent bonds between donor (D) and neighboring donor (ND) atoms and acceptor (A) and neighboring acceptor (NA) atoms are shown as solid lines.

Hydrophobic interactions are also recognized from the input starting structure using the FIRST approach.[91] Each carbon-carbon, carbon-sulfur, or sulfur-sulfur atoms pair is recognized as a hydrophobic interaction if the atoms in the pair are within a certain cutoff (default cutoff value is 0.35 Å) plus the sum of their van der Waals radii. Each hydrophobic interaction is then modeled as single distance constraint between the interacting atoms. In contrast to the other constraints, a hydrophobic constraint is only restricted by the maximum distance between the two atoms, which allows the atoms to slide with respect to each other yet not pull apart.

Steric clashes

At every structure correction cycle, steric clashes between atoms are checked and corrected. Every atom within a certain cutoff (default value used is 8 Å) of every other atom is connected by a distance constraint, excluding those pairs which are already connected by covalent or non-covalent constraints (except for hydrophobic interactions). Atomic van der Waals (vdW) radii, determined by Tsai *et al.,*[185] are used to assign minimum allowed distance for each vdW constraint. These radii consider the hybridization states of heavy atoms and thus allow implicit hydrogen atom modeling in NMSim.

Each vdW distance constraint is satisfied to assure a minimum distance which is the sum of the vdW radii of the connected atoms. The vdW tolerance values (in fraction of the sum of vdW radii) are parameterized (see Table 3.1) to allow a certain overlap in vdW interactions. Distinctions are made between 1-4 vdW constraints (i.e., atoms pairs that are three covalent bonds apart) and the rest of the vdW constraints. A higher tolerance of 0.2 is set for a 1-4 vdW constraints, which accounts for a higher allowed overlap between these atoms, as compared to 0.07 for the rest of the vdW constraints (see Table 3.1)

Phi/psi (φ/ψ) modeling

Ramachandran *et al.* in 1963 have shown[183] that local steric clashes between atoms restrict the allowed range of φ/ψ angles. An electrostatics effect further contributes to the most-favorable (core) regions in the Ramachandran plot.[186] A study shows that around 82 % of the φ/ψ angles in a dataset of experimentally determined structures lie in core regions, which accounts for only 11 % of the total area in the Ramachandran plot.[187] To model this electrostatic effect, φ/ψ angles were explicitly modeled (see section 3.3 for model testing) using distance constraints.

Three basins of attraction of each core region, i.e.,α_L, α_R, and β (see Figure 3.5) are created using the Ramachandran plot described by Morris *et al.*[187] During the structure correction, the φ/ψ angles that lie in allowed or generously-allowed regions feel attraction towards the center of the core regions. This attraction is in terms of adjusting distance constraints between atoms (see Figure 3.5); such that if these distance constraints are fully satisfied would move φ/ψ angles in the center of a core region. The strength of the attraction is controlled by the adjustment factor of the φ/ψ distance constraints (see Table 3.1).

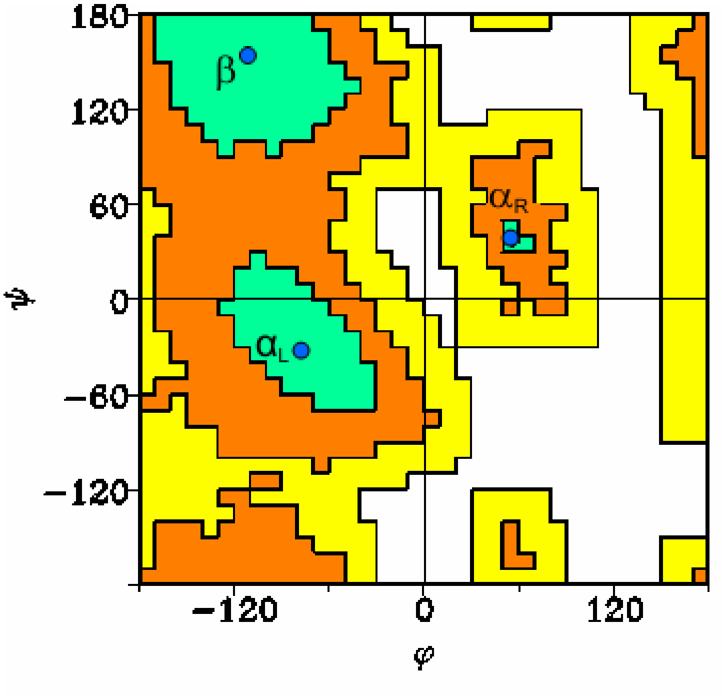


Figure 3.5: The Ramachandran plot with the three basins of attraction for each core region, i.e., α_L, α_R, *and* β. *The coloring on the Ramachandran plot represents the different regions described by Morris et al.,*[187] *i.e., most-favorable or core (light green), allowed (light brown), generously-allowed (yellow) and disallowed (white). The centers of each core regions (blue filed circles) are selected as the basins of attraction.*

Following the above scheme, each φ/ψ angle combination in a protein (excluding Gly residues) is modeled by four distance constraints: for each non-Gly residue r, two distance constraints are used for modeling φ angle, i.e., between C_{r-1} and C_r, and between C_{r-1} and $C\beta_r$ atoms, and the remaining two distance constraints are used for modeling ψ angles, i.e., between N_r and N_{r+1}, and between N_r and $C\beta_r$ atoms (see Figure 3.6). Ideal distances for these constraints are set based on the selected basin of attraction (blue filled circles in Figure 3.5), i.e., the distances of the constraints when the φ/ψ lie at the basin of attraction. It is important to note that, these φ/ψ constraints are used to bias φ/ψ angle towards the core regions. Therefore, these constraints are only slightly adjusted during the structure correction cycles. This is achieved by using a small adjustment factor (see Table 3.1 and Eq. 3.11) of 0.005 (represents the correction of 0.01 times the distance deviation from an ideal distance at every correction cycle). This parameter is set after empirical fitting and testing to ensure a limited biasing.

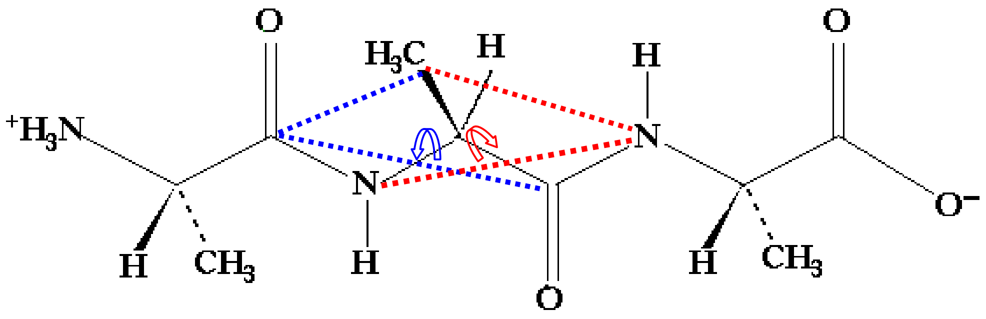


Figure 3.6: The distance constraints used in φ/ψ modeling are shown for an Ala-3 system. Each φ/ψ combination is modeled as four distance constraints i.e., the two distance constraints for modeling the φ angle (blue dotted lines between C_{r-1} and C_r, and between C_{r-1} and $C\beta_r$ atoms) and the remaining two distance constraints for modeling the ψ angle (red dotted lines between N_r and N_{r+1}, and between N_r and $C\beta_r$ atoms). Ideal distances for these constraints are set based on the selected basin of attraction.

Rotamer modeling

Following a similar approach as used for φ/ψ angles, it has been shown that protein side-chain conformations tend to exist in a limited number of conformational states, usually called rotamers.[188,189] Consequently, with the increasing amount of experimental data, many rotamer libraries have been published.[179,190,191] In this study, the Penultimate rotamer library[179] is used, which is based on high resolution crystal structures.[191]

A side-chain in the NMSim approach explores conformational space randomly at the structure distortion/movement step. Subsequently attractive basins derived from the rotamer library are created during the structure correction step. The randomization of the side-chains ensures a proper sampling of conformational space. Furthermore, the biasing during the structure correction step ensures that a side-chain conformation is pushed towards the nearest rotamer state. The state is reached if possible under existing constraints. More rotameric states were sampled during the NMSim simulations than without biasing (see section 3.3). However, the rotamericity of side-chains does not reach 100 % in a protein ensemble, and therefore here it is modeled as such. A study shows that a substantial number of side-chains are under strain: around 5-30 % of the side-chains do not correspond to any rotameric state.[192]

During the structure correction (after initial 50 correction cycles), the nearest rotamer state is selected for each residue r. This is done in two steps:

1) A candidate rotamer list is made for each residue r, i.e., candidates are those rotamers that have all χ-angles within a chi-limit (default CHIDEV_SELLIMIT = ±60°) of the corresponding χ-angles of the residue r.

2) The nearest rotamer is selected from the candidate rotamer list based on the smallest RMSD from residue r.

During the remaining correction iterations (between 50-500 cycles) every χ-angle of every rotamer-assigned residue r is slightly adjusted towards the corresponding selected rotameric χ-angle. This is done by rotating the nearest χ-angle dependent atom around its χ-angle torsion axis. The angle of rotation depends on the χ-angle deviation from the selected rotameric χ-angle and the related adjustment factor (see Table 3.1). A small adjustment factor of 0.001 (representing the correction of 0.001 times the χ-angle deviation from the selected rotameric χ-angle at every correction cycle) is used, after empirical fitting and testing, to ensure a limited biasing and structural stability.

Backbone and side-chain planarity and chirality

Atoms should lie in or near a plane if they are attached to a sp^2 carbon (or equivalent) or in a delocalized aromatic or conjugated system. In the protein backbone, peptide bonds between carboxyl and amino group are planar, i.e., the ω-angle is near 0° for a cis-peptide and near 180° for a trans-peptide. In addition, nine out of 20 natural amino acids (i.e., Arg, Asp, Asn, Glu, Gln, His, Phe, Tyr and Trp residues) also contain a planar group in their side-chain.[193] To achieve planarity in MD simulation a suitable set of improper torsion angles are used. An improper torsion is a rotation around an axis between two atoms that are not bonded to each other. In constraint based correction, all possible 1-4 constraints between atoms of the planar groups can be used to restrict the rotation around any of the torsion angles. However, a small deviation of a 1-4 distance constraint from its ideal value could still result in a large deviation from planarity, which might not be acceptable: according to Procheck criteria, the RMS distance of atoms must be within 0.03 Å and 0.02 Å for rings and others planar groups, respectively.[194]

To acquire better planarity in NMSim, especially for side-chains, a superimposition method was used during the iterative constraint correction procedure. Corrected planar groups are superimposed onto their respective distorted planar groups. Since other distance constraint corrections, as discussed above, would distort the planarity again, an iterative superimposition and constraint correction procedure is applied until convergence, i.e., satisfying both the distance constrains and planarity.

In contrast to a side-chain planar group, a peptide bond shows a higher degree of distortion from ideal planarity. Deviations from planarity can be tolerated with a standard deviation of up to 6° from an ideal angle of 180° for a trans-peptide.[195] However, in some cases, tension in the region might cause an even higher non-planarity (e.g., ω =153.7° was also observed[196]). To model the backbone planarity in NMSim, the same procedure is used as for side-chain planarity, but with a relaxed adjustment factor, i.e., every atom in the distorted planar group is moved only a small fraction (i.e., adjustment factor = 0.02 times the ω-angle deviation from an ideal angle of 180° for a trans-peptide at every correction cycle) towards the superimposed plane (see Table 3.1). This allows variability in the peptide planarity, which depends on the tension level in the molecular environment.

Chirality is another important property. Most amino-acids have an *S* configuration of their chiral centers.[187] During the NMSim simulation, it is assured that the chirality of the C_α atom in the backbone and the C_β atoms in Thr and Ile side-chains does not change. Here hydrogen atoms attached to the chiral centers are also included in the simulation to avoid any chirality change.

Table 3.1: The different constraints used in NMSim modeling with their parameters.

Constraints [a)]	**Adjustment factors** [b)]	**Tolerances** [c)]
Bond/Angle	0.5	0.005 Å
Hydrogen bond	0.2	0.05 Å
Hydrophobic	0.1	0.05 Å
Phi /psi	0.005	0.05 Å
Van der Waals 1-4	0.4	0.20 (fraction of vdW sum)
Van der Waals except 1-4	0.4	0.07 (fraction of vdW sum)
Backbone planarity	0.02	1.0° (from ideal ω -angle)
Side-chain planarity	1.0	0.001 Å (from ideal planarity)
Rotamer	0.001	10° (from each rotameric χ -angle)

a) The different constraints used in NMSim. All constraints are distance-based except backbone/side-chain planarity and rotamer constraints which have planar and angular type respectively. b) An adjustment factor defines the strength of a constraint by which it is restored to its ideal distance/angle in every structure correction cycle. Maximum (full restoration in every structure correction cycle) is achieved at 0.5 for distance based constraints (see Eq. 3.11) and 1 for planar and angular based constraints. c) The tolerance allowed from ideal distances/angles for each constraint.

Constraint adjustment

An iterative approach is applied to satisfy the constraint network, which is built by the above described modeling of the different covalent and non-covalent bonds and the stereo-chemical properties. In every structure correction cycle, every constraint that is unsatisfied is adjusted using respective adjustment factor values (see Table 3.1 for adjustment and tolerance values). A schematic diagram for a distance constraint correction is shown in Figure 3.7. Here, two atoms i and j, connected by a distance constraint having an ideal distance of d_{ij}, are distorted (by moving atoms in the normal mode direction) to new positions ($\vec{a}$ and $\vec{b}$), respectively. Now, the distance is d'_{ij} between them. The constraint is corrected by adding vectors $\vec{G}_{ij}$ and $\vec{G}_{ji}$, respectively, to the current position vectors $\vec{a}$ and $\vec{b}$ to get new coordinate position vectors $\vec{i}\,'$ and $\vec{j}\,'$, respectively.

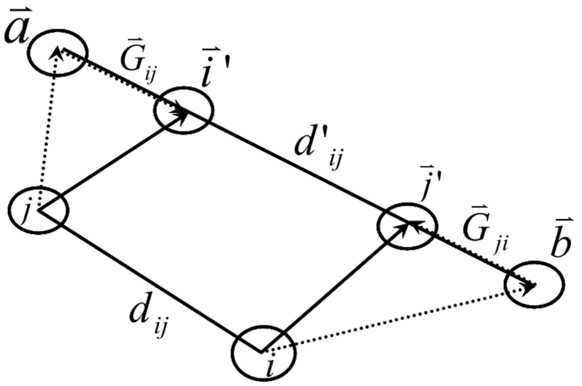


Figure 3.7: A schematic representation of distance constraint correction. Any two atoms i *and* j*, connected by a distance constraint having ideal distance of* d_{ij}*, are distorted to new positions (*$\vec{a}$ *and* $\vec{b}$*), respectively. The constraint is corrected by adding vectors* $\vec{G}_{ij}$ *and* $\vec{G}_{ji}$ *respectively to the current position vectors* $\vec{a}$ *and* $\vec{b}$*, resulting in new coordinate position vectors* $\vec{i}\,'$ *and* $\vec{j}\,'$ *for atoms* i *and* j*.*

The constraint would be adjusted only if the absolute change in ideal and current distances is more than the tolerance value i.e., if $abs\,(\Delta d_{ij}) > Tolerance$ given

$\Delta d_{ij} = d_{ij} - d'_{ij}$ (The criterion for vdW distance constraints is $\Delta d_{ij} > Tolerance$ and for hydrophobic distance constraints is $-\Delta d_{ij} > Tolerance$). The correction vector $\vec{G}_{ij}$ is calculated by

$$\vec{G}_{ij} = \frac{\vec{u}}{|\vec{u}|} * \Delta d_{ij} * AdjustFactor \, , \tag{3.11}$$

where, $\vec{u}$ is defined as $\vec{u} = \vec{a} - \vec{b}$ and the correction vector $\vec{G}_{ji} = -\vec{G}_{ij}$. The $AdjustFactor$ for a distance constraint can have a maximum value of 0.5, which means the constraint would be fully satisfied by moving both connected atoms midway along the line joining the two atoms.

The different types of constraints are satisfied in the sequence shown in the structure correction module in Figure 3.2. The exit criterion for the structure correction cycle is checked every 50th iteration. The criterion is reached when the ratio of the number of unsatisfied covalent distance constraints to the total covalent distance constraints is in the given tolerance value (i.e., by default MISS_SLOPE_TOL=0.01). Additionally, a limit on the maximum number of correction cycle is also considered (i.e., by default SHAKE_ITER=500).

The correction procedure described above was found to be very efficient, e.g., Hen egg white lysozyme structure, which contains 129 residues, needed 5-10 seconds of structure correction time (when distorted with default settings, i.e., step-size = 0.5 Å) on a normal desktop computer. The resulting structure is found to be stereo-chemically valid using Procheck analysis.[194]

3.2.5 Pathway selection in ROG-guided NMSim

A search for a ligand bound conformation of a protein can be drastically improved if some structural properties of the complex are incorporated in order to tailor the trajectory towards those properties. In case of large-scale conformational changes, like domain closures in proteins upon ligand binding, it is well known that the compactness of the protein structure increases upon the ligand binding.[197,198] The radius of gyration (R_g) is often used to describe the compactness of a protein, e.g., during the folding process from a denatured state to the native state. Experimentally, small-angle X-ray scattering has been used to measure the effects of ligand binding on R_g, and the decrease in R_g is used as evidence to domain closure.[197,199] R_g is defined by

$$R_g^2 \cong \frac{1}{n}\sum_{i=1}^{n}(\vec{r}_i - \vec{R}_c)^2 , \tag{3.12}$$

where $\vec{R}_c$ is the center of mass, $\vec{r}_i$ is the atomic position of atom i and n is the number of C_α atoms. Here only C_α atoms are considered.

In a ROG-guided (Radius Of Gyration-guided) simulation, the trajectory can be tailored towards the bound structure by selecting the pathway that leads to a decrease in the R_g, assuming that ligand binding would result in domain or loop closures. It is important to note here that the conformations are still generated by random linear combinations of low frequency normal modes and therefore the pathway still goes though the low energy space. In fact, two or more conformations are generated without any biasing during a simulation cycle (i.e., by calling NMSim module for each conformation). Then the conformation with the lowest R_g is selected for further trajectory exploration in the next simulation cycle. In other words, one of the pathways is selected at every simulation cycle.

3.3 Model testing

The program CONCOORD uses a similar constraint based correction approach as the one described above for NMSim. However, NMSim additionally incorporates explicit modeling of hydrogen bonds, φ/ψ dihedrals, and a rotamer library in a simple constraint based approach. These components are individually tested for their effectiveness and suitability of their relevant parameters. Modeling hydrogen bonds explicitly, instead of rigidifying secondary structures as in CONCOORD, is not only a more natural approach but also a step towards modeling hydrogen bond breaking and forming. The φ/ψ angle and rotamer modeling is a new addition to a geometry-based simulation approach and thus will be discussed below in detail.

3.3.1 Testing the φ/ψ model on an Ala-6 system

Conformational changes in the protein backbone arise mainly due to changes in the φ/ψ dihedral angles. Local steric clashes between atoms restrict the allowed range of φ/ψ dihedrals as shown by Ramachandran *et al.* in 1963.[183] Additionally, φ/ψ dihedrals are restricted due to a dense hydrogen bond network in secondary structures like α-helices and β-sheets. However, φ/ψ dihedrals in loop regions and those forming hinges are critical and need to be modeled correctly in constraint based approaches that lack electrostatic forces.

In NMSim, explicit modeling of φ/ψ was applied as described above and was tested by analyzing the Ramachandran plots of different NMSim generated Ala-6 conformations with and without φ/ψ modeling. Simple alanine systems have been previously used for testing and parameterization, for example, in improving MD force fields.[200] To fully explore the available φ/ψ space, a fully random NMSim simulation was applied for Ala-6, where an Ala-6 structure is randomly distorted and then corrected using the NMSim module. By switching off the steric clashes correction (i.e., by setting VDW_DIST_TOL=1.0 and VDW_ONE4_DIST_TOL =1.0) and the φ/ψ correction in a random NMSim simulation, evenly distributed φ/ψ dihedrals were found (see Figure 3.8-a). The core, allowed, generously-allowed and disallowed regions are occupied to 10, 28, 30, and 32 %, respectively. This is in agreement with the respective area of these regions.[187]

By applying the steric clashes correction in the above simulation, no steric clashes in the generated structures were found by Procheck.[194] The effect of steric clashes correction in NMSim can be seen in Figure 3.8-b, which shows a restriction towards certain φ/ψ regions in Ramachandran map. As expected, the biasing towards a φ/ψ core region was found to be imperfect: core, allowed, generously-allowed and disallowed regions are occupied by 18, 54, 19, and 9 % respectively. This emphasizes the need for explicit φ/ψ modeling. Almost the entire disallowed region and the part of generously-allowed region were restricted due to steric clashes (see Figure 3.8-b). However, a small cluster of φ/ψ angles ($\varphi \cong 50^\circ$ and $\psi \cong -100^\circ$) was found in the disallowed region: further investigation of the conformations in this cluster shows that the distance constraints around these φ/ψ angles are stressed which indicates that this φ/ψ region represents a high energy minimum in NMSim. However, this φ/ψ region is only accessible by crossing high energy barriers: this φ/ψ region was only observed in random NMSim simulations, where each structure is independent of the other and not in default NMSim simulations, where a trajectory follows a low-energy path.

In φ/ψ modeling, a biasing towards the core region was applied as describe above. The Ramachandran plot obtained from NMSim conformations of Ala-6 using φ/ψ modeling is shown in Figure 3.8-c. Here, core, allowed, generously-allowed and disallowed regions are occupied by 64,26,5 and 5 %, respectively. Thus a high φ/ψ distribution shift towards the core region.

A default NMSim simulation (i.e., normal based simulation with default step-size) is also run on an Ala-6 (Figure 3.8-d). Here the core and allowed regions are occupied around 88 % and 12 %, respectively, with no φ/ψ pairs in the generously-allowed and disallowed regions. This is comparable to 82 % and 15 % found in experimental structures.[187] Due to a small biasing value for φ/ψ it is assumed that, in a stressed molecular environment, φ/ψ combination can lie in generously-allowed or disallowed regions.

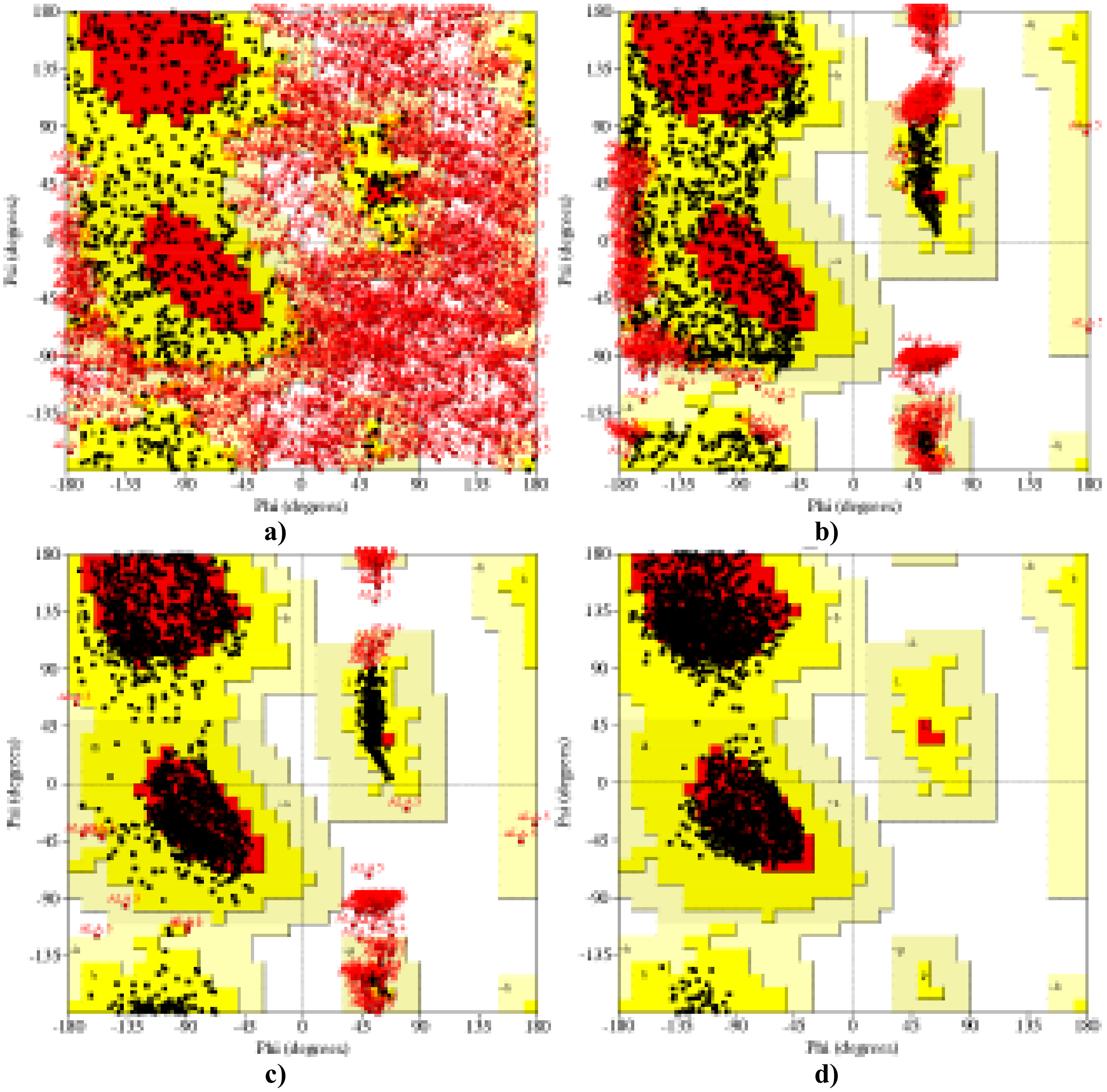


Figure 3.8: The Ramachandran plots for 500 Ala-6 conformations, having 2000 φ/ψ pairs, obtained from different simulations are shown. The conformations generated from random NMSim simulation (i.e., randomly distorted Ala-6 structure and corrected with NMSim correction module) with no steric clashes correction and no φ/ψ modeling (in panel a), with steric clashes correction and no φ/ψ correction (in panel b), with steric clashes correction and φ/ψ correction (in panel c) and conformations generated from a default NMSim simulation (in panel d) are shown. The random NMSim simulation with no steric clashes correction shows evenly distributed φ/ψ pairs (in a) whereas the disallowed regions are restricted due to steric clashes correction (in b). The explicit φ/ψ modeling is applied to bias φ/ψ pairs in the core region in random NMSim simulation (in c) and default NMSim (in d).

3.3.2 Testing the rotamer model on lysozyme

As describe above, a biasing towards the nearest selected rotamer of every residue (excluding Gly, Ala, Pro) is applied during the structure correction. This forces side-chain conformations into the nearest rotameric state. In order to test this model, Hen egg white lysozyme (HEWL) was simulated with and without applying rotamer biasing. The resulting side-chain conformations over the trajectories were analyzed in detail for rotamericity and heterogeneity measures. Here, the rotamericity of a residue in a protein sequence is defined as the ratio of the total number of occurrences of the residue in any of the possible rotamers to the total number of conformers in the ensemble. The heterogeneity measure of a residue in a protein sequence is defined as the ratio of the total number of distinct rotamer states of the residue observed in an ensemble to the total number of available rotamer states for that residue in the rotamer library.[179] Rotamericity measures the quality of side-chain conformations of an ensemble in terms of rotamers. The heterogeneity, in contrast, measures the conformational sampling of a residue in terms of rotamers.

On average, an increase in rotamericity was observed, without trapping in one or few rotameric states, when rotamer biasing was applied in NMSim: in the case of HEWL, the average rotamericity of all residues increases from 0.57 to 0.70. Notably, the biasing applied does not influence the exploration of side-chain conformational space available to each residue: the average value for heterogeneity, which was around 0.46 without biasing doest not change when biasing is applied.

A comparison of these values with different constraint based methods, i.e., CONCOORD and FIRST, and with MD simulations shows that the MD simulation explores rotamer states better than any of the constraint based methods, however, NMSim is the closest to MD among the compared methods (see Table 5.5 and section 5.2.4). CONCOORD does not explore enough rotameric states, i.e., an average heterogeneity value of 0.23 was observed. Higher rotamericity values in CONCOORD are an effect of getting trapped in one or a few of the rotamer states.

The difference in the rotamericity between HEWL trajectory with and without rotamer modeling is shown in Figure 3.9, which gives a qualitative picture of the increase/decrease in rotamericity for each residue. Except for a very few cases, an overall increase in rotamericity can be observed (i.e., positive change in the plot) which is as high as 0.6. Among 103 residues (excluding Gly, Ala, and Pro) nearly half of the residues (i.e., 47 residues) show a considerable increase above 0.1 in their rotamericity values, whereas, only 4 residues show a considerable decrease below 0.1.

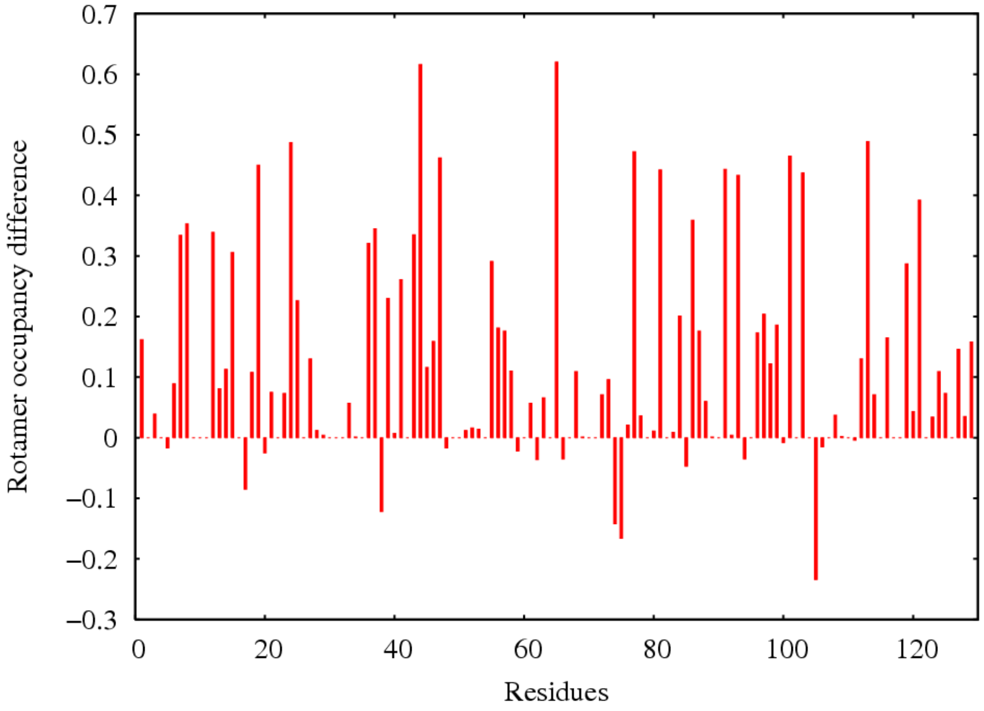


Figure 3.9: Differences in the rotamericity values (defined as the ratio of the total number of occurrences of the residue in any of the possible rotamers to the total number of conformers in the ensemble), obtained from the two ensembles, i.e., the NMSim trajectory of HEWL with and without rotamer modeling, is shown. A positive value represents an increase in rotamericity in the case of HEWL trajectory due to explicit rotamer modeling, whereas a negative value represent a decrease in rotamericity due to the rotamer modeling.

4 Materials and methods

4.1 Comparative study of ENM and ED

The study aims at comparing essential dynamics (ED) modes of proteins observed in MD simulations with normal modes obtained from coarse-grained normal mode methods (CGNM) for a large dataset of 335 diverse proteins. As for MD simulations, the first five ED modes for each protein were obtained from the Molecular Dynamics Extended Library database (MoDEL).[85,86] There, the modes have been extracted from MD trajectories of 10 ns length. Coarse-grained normal modes were calculated using ENM and RCNMA[68] approaches (see section 3.1). The three sets of modes were compared in terms of overlap of directions, correlation of relative magnitudes of motions, and spanning coefficients. The CATH classification[201] of protein structures was used in order to investigate the influence of protein structure similarity/dissimilarity on mode similarities/dissimilarities. For a smaller protein subset, ED, ENM, and RCNMA modes were also compared against experimentally observed conformational changes.

4.1.1 ED modes and protein data set

ED modes were obtained from the MoDEL database (http://mmb.pcb.ub.es/MODEL, version as of May 2006)[85,86] The MoDEL database stores information derived from MD simulations for more than 400 proteins. The MD simulations were performed with the Amber8 suite of programs at 300 K in the NPT ensemble, and the parm99 force field was used together with TIP3P as a water model. The length of each MD trajectory is 10 ns.

The first five available ED modes of 418 proteins were downloaded from the MoDEL database. Here, ED modes are calculated using all atoms; however, for comparison only C_α directions were used. PCA is applied on 5-10 ns trajectories containing snapshot every ps. The corresponding experimental structures were obtained from the RCSB Protein Data Bank.[202] For the sake of compatibility, heavy atoms in the ED modes files were compared with heavy atoms in the PDB files using the PDBParser module of Biopython.[203] Where possible, inconsistencies between the two sets were corrected manually. However, 83 out of 418 cases were removed from the dataset due to deviating numbers of atoms/residues, empty or corrupt ED modes files, C_α-only structures, bad structural quality or inconsistency with the standard amino acid library, or problems in processing by FIRST.[91] Finally, this resulted in a dataset of 335 protein structures. The PDB structures were then protonated using Amber. Disulfide-bridges involving cysteine residues and protonation states of histidines were adopted from the ED mode files. All structures were then aligned to their respective MD average (reference) structure using C_α atoms.

In order to reduce the influence of stereochemical inaccuracies in MD average structure due to the averaging process, minimization was performed. Average MD structure was minimized in the gas phase by using the conjugate-gradient method with a distance-dependent dielectric of 4r (to approximately account for solvation effects, with r being the distance between two atoms) until the root-mean square of the elements of the gradient vector is $< 10^{-4}$ kcal mol^{-1} Å^{-1}.

The dataset of 335 protein structures is diverse with respect to protein size, function, origin, sub-cellular localization, and structure determination method. The proteins contain on average 121 residues, with a minimum of 20 and a maximum of 349 residues. The size distribution of the dataset is shown in Figure 4.1. The distribution is positively skewed with a peak in the range of 60 to 80 residues.

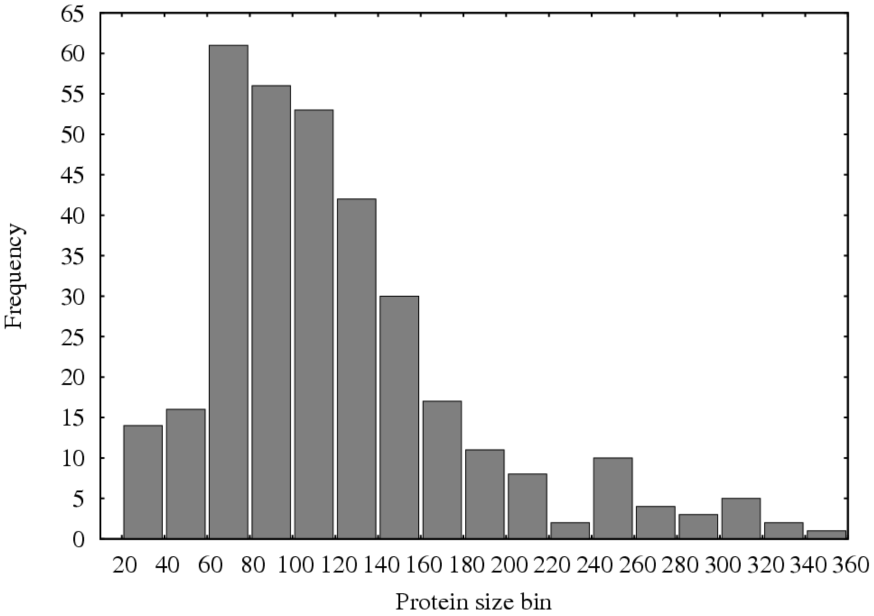


Figure 4.1: Frequency distribution of the protein size, in terms of the residue number, for the dataset of 335 proteins.

4.1.2 RCNMA and ENM parameters used

RCNMA (as described in section 3.1) is performed using the default parameter set which is in accordance with the previous study.[68] Flexible and rigid regions of proteins are identified by FIRST,[91] which identifies and counts the bond-rotational degrees of freedom in a molecular framework of atoms connected by covalent and non-covalent constraints (hydrogen bonds, salt bridges, hydrophobic interactions) based on rigidity theory.[91,161,204] Parameters used for FIRST analysis, i.e., hydrogen bond energy cutoff (i.e. $E_{cut} = -1.0\ \text{kcal mol}^{-1}$) and distance cutoff for hydrophobic interaction (i.e. 0.25Å), are also consistent with a previous study.[68] No profound change in the results was observed by changing these parameters.

The all-atom representation of proteins needed for the FIRST analysis is reduced to a C_α-only representation in RCNMA. Each rigid cluster forms a block in the subsequent rotations and translations of block (RTB)[142,178] approach, and flexible regions are modeled on a one-residue-per-block basis (in which case only translational motion of the "block" is considered). Interactions between these particles are modeled as in

ENM (Eq. 3.1), and the same parameters, for both ENM and RCNMA, are used: interactions cutoff between the C_{α} atoms, i.e., $r_c = 10$ Å and phenomenological force constant, i.e., $\gamma = 1$ kcal mol^{-1} Å^{-2} (see section 3.1).

4.1.3 ED and CGNM comparison

The directions and relative magnitudes of motions described by the first five ED modes were compared with CGNM results. As done previously,[68,70] the overlap of mode directions and the correlation of magnitudes of motions (see Eq. 4.1 and Eq. 4.2) between two sets of modes were calculated for each structure in the protein dataset. Distributions of maximal overlap, maximal correlation, and the mode number involved in maximal overlap between the two sets of modes were analyzed for the dataset. It was further analyzed how well the subspace spanned the first 5 ED modes is described by the 10 %, 25 %, and 50 % lowest frequency CGNM modes by calculating the "spanning coefficient" (see Eq. 4.3). In order to analyze the coarse-grain level achieved by RCNMA based on the rigid cluster decomposition from FIRST, the dimensionality reduction of H (see Eq. 4.4) was calculated.

The overlap I_{in}[119] of the i^{th} CGNM mode $\vec{u}_i$ with the n^{th} ED mode $\vec{v}_n$ ($n = 1, 2, \ldots 5$) was calculated according to:

$$I_{in} = \frac{\left|\vec{u}_i \cdot \vec{v}_n\right|}{\left(\vec{u}_i \cdot \vec{u}_i\right)^{1/2} \cdot \left(\vec{v}_n \cdot \vec{v}_n\right)^{1/2}} \qquad (4.1)$$

An overlap of 1 indicates that the *directions* of the collective atom displacements along the ED mode and the CGNM mode are identical. For each protein structure only the CGNM mode with maximal overlap was considered for further analysis.

Similarly, a correlation coefficient C_{in}[70] of the i^{th} CGNM mode $\vec{u}_i$ with the n^{th} ED mode $\vec{v}_n$ was calculated according to:

$$C_{in} = \frac{\vec{A}_i \cdot \vec{B}_n}{\left(\vec{A}_i \cdot \vec{A}_i\right)^{1/2} \left(\vec{B}_n \cdot \vec{B}_n\right)^{1/2}}\,, \qquad (4.2)$$

where $\vec{A}_i$ and $\vec{B}_n$ are the vectors of mean centered amplitudes of atomic displacements as determined from vectors $\vec{u}_i$ and $\vec{v}_n$. A correlation coefficient of 1 indicates that the *relative magnitudes* of atomic displacements along the ED mode and the CGNM mode are identical.

The "spanning coefficient" S_n^k [205] was computed as the sum of the square of the expansion coefficients:

$$S_n^k = \sum_i^k \left(\vec{u}_i \cdot \vec{v}_n\right)^2 \tag{4.3}$$

Here, the sum over the first *k* CGNM modes was computed in order to determine the lowest percentage of normal modes needed for describing each of the first five ED modes. A spanning coefficient of 1 indicates that the subspace spanned by the ED mode can be completely described by the subspace considered by the *k* CGNM modes.

The dimensionality reduction *D* was calculated based on the reduction of the *H* matrix dimension due to considering rigid blocks in RCNMA:

$$D = 1 - \left(\frac{6n + 3m - 6}{3N - 6}\right), \tag{4.4}$$

where *n* is the number of blocks of size > 2 and *m* is the number of blocks of size 1 (note that for simplicity blocks of size of 2 are not considered *per se* in the H_{sub} matrix and are decomposed into two blocks, each of size one). A dimensionality reduction of 1 indicates that all C_α atoms are in one rigid block, whereas 0 indicates that every block is of size 1. In that case RCNMA becomes equal to ENM.

4.1.4 Similarities/dissimilarities in classes/folds: ED and ENM modes

In order to analyze dynamic similarity within different protein classes or folds, the dataset of proteins was classified according to the CATH classification. Out of 335 proteins, 320 proteins were found in the CATH database.[201] Overlap and correlation results were sorted for these proteins according to different protein classes and folds (Class and Topology levels in CATH), and mean values and standard deviations were calculated accordingly.

Additionally, in order to analyze locality or collectivity of motion within different classes, the collectivity index (Eq. 4.5) was used, which describes the number of atoms that are affected by a mode (or conformational change). The collectivity index proposed by Bruschweiler[206] is calculated according to:

$$\kappa = \frac{1}{N}\exp(-\sum_{i=1}^{N}\Delta\vec{r}_i^{\,2}\,\log\Delta\vec{r}_i^{\,2}), \tag{4.5}$$

where N is the number of atoms, $\Delta\vec{r}_i$ is the relative displacement of the mode or the difference in Cartesian coordinates of atom i if an experimentally determined conformational change of the protein is considered. All values of $\Delta\vec{r}_i$ have been scaled consistently such that $\sum_{i=1}^{N}\Delta\vec{r}_i^{\,2} = 1$. $\kappa = 1$ indicates a mode or conformational change of maximal collectivity, i.e., all $\Delta\vec{r}_i$ are identical. Conversely, if only one atom is affected by the mode or conformational change, κ reaches the minimal value of $1/N$.

4.2 NMSim and methodological comparisons

In order to analyze the usefulness and the limitations of the NMSim approach, it was compared with different counterpart approaches on a test case: the Hen Egg White Lysozyme (HEWL) protein. The HEWL conformations[207] from a state of the art MD[56-58] and different experimental structures are compared with the conformations obtained form the most efficient geometric based methods i.e., FRODA,[64] CONCOORD[62,63] and NMSim.

4.2.1 Analysis of MD, NMSim, FRODA, CONCOORD and experimental HEWL ensembles

The MD trajectory was taken from a recent study by A. Koller *et al.,*[207] where a 100 ns MD simulation of HEWL (PDB code 1hel)[208] was performed with AMBER9 under periodic boundary conditions in the NVT ensemble. The Amber force-field 99SB was used with TIP3P water model at 300 K. This simulation took approximately 4 month on 4 CPUs on a linux cluster. Here, 1,000 equal-spaced conformations were selected from the trajectory, which forms the MD ensemble used in this study.

The NMSim program was applied to the same starting structure with the default parameter set (see Appendix A). In total 10,000 conformations were generated using a simulation cycles of 1,000 and an NMSim cycle of 10. This simulation took 30 hours on a 64-bit desktop computer. Every 10th structure was then selected for the NMSim ensemble.

The FRODA[64] simulation with the latest available version 6.2 was performed using the default parameter set. However, the hydrophobic cutoff –c is set to 0.35Å, because the default cutoff of 0.5Å resulted in a highly rigid protein with no relative motions. For the other parameters the default values were used. In total, 10 million conformations were generated, and every 1000th conformation was saved during the simulations. A total of 10,000 conformations were saved from the simulations. For the analysis, every 10th conformation was selected from the saved conformations, which forms the FRODA ensemble of 1000 conformations. This simulation took 6 days on a 64-bit desktop computer. Here it is important to note that, despite of generating 10 millions of conformations and using approx. 6 days of computational time, the FRODA trajectory was less explorative in terms of RMSD from the starting structure as compared to the NMSim trajectory. The average backbone and heavy atom RMSD of every structure to its previous structure in the FRODA ensemble are 0.25 Å and 0.5 Å, respectively, as compared to 0.4 Å and 0.6 Å in the NMSim ensemble.

The latest available version of the CONCOORD[62] 2.0 program was run with the default parameter set. As recommended on the CONCOORD home page, van der Waals parameters “yamber2” and bonded parameters “Engh-Huber” were used. CONCOORD is a pure conformation generation method with no pathways/trajectories of the simulations. Every conformation is generated using the starting structure distortion and correction procedure, and, hence, does not depend on simulation time; therefore, only 1000 structures were generated for the ensemble. The generation of the 1000 conformation only took 53 minutes on a 64-bit desktop computer.

In order to compare the conformations generated from the different approaches with the experimentally observed conformations of HEWL, an ensemble of experimental structures was made. The experimental structures, which show 100 % sequence similarity with the sequence of the starting structure of HEWL (PDB code 1hel)[208] were downloaded from the RCSB Protein Data Bank.[202] In case of NMR structures, each model was treated separately. The structures that were closest to the starting structure with a C_α RMSD less than 0.5 Å were removed. The experimental ensemble contains 130 different X-ray crystal structures and NMR structures.[209] These structures are listed in Appendix B.

4.2.2 Rotamer states and derived measures: rotamericity, heterogeneity, and occupancy

To compare side-chain conformational sampling in different methods, the Penultimate rotamer library[179] was used in this study. A side-chain conformer was assigned to a rotameric state if every χ-angle of that residue falls within ±30° of the corresponding χ-angle of any of the rotameric states available to that particular residue. Different rotamer derived measures were then used for the analysis.

The rotamericity measure is used to compare the quality of side-chain conformations in different ensembles. The rotamericity of a residue in a protein sequence is defined as the ratio of the total number of occurrences of the residue in any of the possible rotamers to the total number of conformers in the ensemble. It is important to note here that the rotamericity of each residue in the sequence is calculated in this study. This is in contrast to the rotamericity of each amino acid of protein, used by Schrauber *et al.*[192] The rotamericity of amino acids has been used previously, for example to show that a substantial number of side-chains are under strain[192] and that ligand binding induces non-rotamericity.[210]

To analyze the potential of different methods in sampling different rotamer states, different measures are introduced. The heterogeneity measure of a residue in a protein sequence is defined as the ratio of the total number of distinct rotamer states of the residue observed in an ensemble to the total number of available rotamer states for that residue in the rotamer library.[179] This measure defines how well the different methods explore the available side-chain conformational space.

The heterogeneity is normalized with the available rotamer states of a residue. According to the Penultimate rotamer library,[179] some long side-chains like Arg and Lys have 34 and 27 rotamer states respectively, whereas, side-chains like Cys and Ser have only 3 rotamer states. These uneven normalization factors need to be considered for the heterogeneity measure. Therefore, the occupancy measure was also introduced, i.e., the heterogeneity measure without normalization. The occupancy measure of a residue in a protein sequence is defined as the total number of distinct rotamer states of the residue observed in an ensemble. Furthermore, the occupancy vector is introduced, which is simply a vector containing the occupancy value of every residue in a protein. The correlation coefficient between the occupancy vectors is then calculated to compare the patterns of rotamers sampled in the different ensembles.

4.2.3 Structure quality using Procheck

The quality of a subset of the structures obtained from the different types of methods was analyzed using the Procheck[194] program. Here, 100 equal-spaced structures were taken from the ensemble of every method for the analysis. To better judge the structure quality, 100 high resolution crystal structures from Richardson's lab[211] (here named as EXPTOP) were also used for the analysis, in addition to the 130 experimental structures of HEWL. The averages and the standard deviations were calculated for the different properties obtain from Procheck.

The *G*-factor provides a measure of how normal a given stereo-chemical property is. In Procheck, it is computed for dihedrals angles (i.e. $\varphi - \psi$ combination, $\chi_1 - \chi_2$ combination, χ_1 torsion for those residues that do not have a χ_2, combined χ_3 and χ_4 torsion angles, ω torsion angles) and covalent geometry (main-chain bond lengths, main-chain bond angles). The *G*-factor is a log-odd score based on the observed distributions of these stereo-chemical parameters. A low *G*-factor indicates that the property corresponds to a low-probability conformation.

4.3 NMSim and biological applications

4.3.1 The proteins in the dataset

The NMSim approach was applied to a dataset of eight proteins, where important conformational changes have been observed upon ligand binding, and where two crystal structures are available, the unbound (open) and the ligand bound (close) conformations. In order to analyze the usefulness and the limitations of the NMSim approach, the dataset is subdivided into two categories, Domain and Loop, based on the types of the conformational changes observed upon ligand binding. The dataset is listed in Table 4.1 along with the PDB codes and relevant information. The proteins in the dataset have been used previously in different normal mode studies[68,70,75,212] and show both "ligand-induced" and "conformational selection" types of conformational changes.[37,213]

Table 4.1: The protein dataset used in NMSim study.

Proteins a)	**Open structure**b)	**Close structure**b)	**No. of residues**c)	**Interesting regions** d)
Domain:				
Adenylate kinase (ADK)	4ake (A)	1ake	214	Core: 1-28, 80-112, 173-214; ATP: 119-156; NMP: 31-72
Aspartate Aminotransferase (AST)	9aat (A)	1ama	388	Large: 42-322; Small:15-33, 330-356, 362-410
Calmodulin (CLM)	1cfd (A)	1ckk	148	C-term. Domain: 82-146; N-term. Domain: 5-75
Citrate synthase (CTS)	5csc (A,B)	6csc	860	Large: 3-55, 66-272, 330-335, 382-433; Small: 56-63, 284-327, 338-378
LAO binding protein (LAO)	2lao (A)	1lst	238	Lobe I: 1-88 and 195-238 Lobe II: 93-185
Loop:				
Tyrosine phosphatase (TYP)	1ypt (A)	1yts	278	β7-α4 loop: 350-360
Triosephosphate isomerase (TIM)	8tim (B)	1tph	245	Loop 6: 166-176
CAMP-dependent protein kinase (CAPK)	1jlu (E)	1fmo	336	Glycine-rich loop: 50-55

a) The protein dataset is further divided into Domain and Loop dataset. b) The PDB codes for unbound (open) and ligand bound (close) structures. The PDB chain ID used is in brackets. c) 13 residues (SER3-ASP15) in AST were removed, as they were found highly fluctuating. Two and three residues were removed, respectively, from TIM and CAPK to equalize the number of atoms with the close structure. d) The residue numbers for domain[20,213,214] and important loop regions.

4.3.2 The NMSim run: parameters and ensemble generation

In order to explore the extent to which experimentally observed conformational changes can be achieved in NMSim, the method was applied to the open conformation of proteins in the dataset (see Table 4.1). Three different types of simulations were performed using NMSim approach, i.e., freely-evolving, radius of gyration (ROG)-guided and target-directed.

The freely-evolving NMSim is performed with no information of the target conformation, and therefore a random linear combination of normal modes (see section 3.2.2) is used to freely evolve a trajectory in the normal mode space. In general, the freely-evolving NMSim is run with the default parameter set (see Appendix A). As default for freely-evolving NMSim, 5000 conformations are generated in a trajectory (using 500 simulation cycles and 10 NMSim cycles). Every 10th conformation is then selected for analysis. However, in the Domain dataset, the first five normal modes and a smaller side-chain randomization parameter (i.e., RANDSCALING = 0.05 Å) are used in order to explore large-scale backbone conformations. These minor changes in the default parameter set are suited for the large-scale exploration of the conformational space of a protein. Furthermore, 10 different freely-evolving trajectories for each protein in the Domain dataset are run, however, generating 500 conformations in each trajectory (using 500 simulation cycles and one NMSim cycle). This also results in 5000 conformations from 10 different trajectories of each protein in the Domain dataset. For these 10 trajectories, NMSim takes around 2 days of computational time for a normal size adenylate kinase (having 214 residues) on a 64-bit desktop computer.

The ROG-guided NMSim is performed with the assumption that the close structure has smaller radius of gyration (R_g) than the open structure (see section 3.2.5). Consequently, the path which leads to lowering R_g is selected. The 3 different conformations are generated in each simulation. And then the conformation with the lowest R_g is selected (among the 3) for further trajectory exploration in the next simulation cycle. As default settings in ROG-guided NMSim, 1500 conformations are generated in total (using 500 simulation cycles and one NMSim cycle). However, the pathway is represented by the selected 500 conformations.

The target-directed NMSim is performed by using the close conformation information, and hence the best combination of modes (see section 3.2.2) is used at every step of the trajectory. As default settings in target-directed NMSim, 500 conformations are generated (using 500 simulation cycles and one NMSim cycle).

4.4 NMSim and the pathways of conformational change

Pathways of conformational changes from the open to the close structures were generated for ADK using two different types of simulation, i.e., the target-directed NMSim (section 3.2.2) and the ROG-guided NMSim (section 3.2.5). The default parameter set (see Appendix A) for both types of simulations was used. However, each intermediate conformation was generated using the single best mode instead of using a linear combination of modes. For target-directed NMSim, out of 50 modes, the best overlapping mode (i.e., the mode having the best overlap with the conformational change direction) was used at each NMSim cycle (as described in section 3.2.2). For ROG-guided NMSim, 10 structures in either direction of the first 5 modes were generated and the structure with the lowest radius of gyration was selected for further exploration of the pathway, at each NMSim cycle.

In order to analyze the order of the domain closure in ADK, the reaction coordinates described by Whitford *et al.*[89] were used. The reaction coordinates $R^{LID-CORE}$ is defined as the distance between the LID domain and CORE domain centers of mass and $R^{NMP-CORE}$ is defined as the distance between the NMPbind domain and CORE domain centers of mass. In order to further verify the NMSim pathway, the generated intermediate structures were compared with 11 different X-ray crystal structures[87] of ADK in terms of C_α RMSD. The X-ray structures which lie along the pathway from open to close conformation of ADK were identified, by selecting a crystal structure with the lowest RMSD to each intermediate structure along the generated pathway, and compared with a similar study by Maragakis and Karplus.[87]

5 Results and discussions

This chapter is subdivided into four parts: In the first part, a validation of coarse-grained normal mode approaches in describing essential space explored during MD simulations is reported. Furthermore, questions regarding the parameterization of NMSim approach are addressed. In the second part, the usefulness and the limitations of the different geometry-based approaches, in general, and the NMSim approach, in particular, are analyzed on a test case, i.e., the Hen Egg White Lysozyme. The generated conformations from different geometry-based approaches are compared with state of the art molecular dynamics (MD) simulations and experimental conformations. In the third part, the usability of the NMSim approach in exploring the intrinsic dynamics of a protein and in describing conformational changes due to ligand binding is presented. The approach is applied to a dataset of proteins where conformational changes have been observed experimentally either in domain or functionally important loop region. In the last part, NMSim generated pathways of conformational change in adenylate kinase are compared with previous studies.[87-89] Furthermore, the possibility of pathway generation without knowing the target structure is explored.

5.1 A large scale comparative study of ENM and ED

The need for a large-scale comparative study between essential dynamics (ED)[215,216] and elastic network models (ENM)[67] was felt, which investigates the validity and applicability of coarse-grained normal mode approaches. Although ENM has been found successful in describing protein conformational changes,[136,150] a recent study[71] has shown that the success of ENM in describing experimental conformational change (from an unbound to a bound conformation) strongly depends on the collectivity of the conformational change. In order to investigate the successes/limitations of ENM in describing the intrinsic dynamics of a protein, a comparison with ED modes derived from MD simulation is performed here.

Different studies have previously[135,215,217-219] shown the striking similarities between normal modes derived from all-atom force-field potentials[130,220] and ED modes from MD simulations. These early studies have used one or a few proteins with limited simulation size and have focused on the comparisons of frequency spectra.[90,130,135] For example, Hayward *et al.*[135] have used a 200-ps MD trajectory of bovine pancreatic trypsin inhibitor for such studies. In ENM, however, it has been argued[136] that the information provided by the eigenvectors for the directionality of biologically relevant conformational changes has wider applications than the eigenvalues. This is also reflected by the applications and developments of ENM based approaches in recent years.[79,80,82,221,222] In this view, this study focuses on comparing directions of essential protein movements from atomistic MD simulations and coarse-grained normal mode analysis. This analysis was performed on a large dataset of 335 diverse proteins. To our knowledge, a similar study[85] has been reported recently, however, on a relatively small dataset of 30 proteins.

In this section, important questions that are assumed to guide a further development of normal mode-based approaches are addressed. The validity of the coarse-grained normal mode approaches in describing essential space explored during MD simulations is reported. Furthermore, the extents of similarities/dissimilarities between essential directions obtained from the two different methods are presented by comparing overlap of directions, correlation of relative magnitudes of motions, and spanning coefficients between modes. The influence of protein structure similarity/dissimilarity on mode similarities/dissimilarities is analyzed using the CATH [201] classification of protein structures. In view of recent[223-225]evidences regarding evolutionary conservations of vibrational dynamics, modes were compared for proteins within the same fold class, for representative cases, where considerable differences were observed between ENM and ED modes. Here we start with the discussion of the influence of using different reference structures in ENM and the influence of using natural coarse-graining in RCNMA as compared to residue-based ENM.

5.1.1 Influence of the reference structure: Average vs. open

ED modes[215] are based on a reference structure, i.e., the structure obtained by averaging all conformations along the trajectory, which can be conformationally different from the experimental starting (also termed “open”) structure and might have stereochemical inaccuracies. In contrast, CGNM analyses usually use experimental (open) structures.[137] In order to analyze the influence of using different reference structures, i.e., average or open, in CGNM analysis, ED modes were compared with CGNM modes computed from either the average or the open structure. Not unexpectedly, ED modes correlate better with CGNM modes in both directions and amplitudes of motions if the average structure is used for CGNM. For example for ENM, the mean maximal overlap (Eq. 4.1) and mean maximal correlation (Eq. 4.2) values are 0.65 and 0.73, respectively, using the average structure. These values decrease by 0.10 and 0.08 if the open structure is used instead. The lower values are mainly due to those proteins that show large conformational differences between the open and the average structures. For example, those 49 out of 335 protein structures for which the maximal overlap decreases by at least 0.2 have a mean RMSD between open and average structure of 3.11 Å. For comparison, the mean RMSD over all proteins is 2.06 Å.

As a further test, average structures were minimized (see section 4.1.1) to remove stereochemical inaccuracies obtained by the averaging process. The mean maximal overlap and correlation values between ED and CGNM modes were found to be almost unaffected compared to the use of non-minimized average structures, which can be explained[150,176] by the coarse-grained nature of CGNM. Given that in general very similar results are obtained for CGNM from both the open and average structures and for the sake of a fair comparison with ED modes, the average structure will be used as a reference in this study. The average structure has also been used previously[85] in ENM for the sake of comparison.

As for the decomposition of the structure into rigid clusters and flexible regions by FIRST, however, the MD average structures generally result in more flexible decompositions than the open ones. This can be explained by the fact that FIRST requires input at an atomic level, which makes FIRST more sensitive to the accuracy of the input structure.[91] To what extent RCNMA results are influenced by this is discussed in more detail below.

5.1.2 Influence of the level of coarse-graining: ENM vs. RCNMA

FIRST[91] decomposes a protein structure into rigid clusters and flexible regions based on rigidity theory.[91,161,204,226] RCNMA utilizes this information and considers each rigid cluster as a single node with six degrees of freedom in an elastic network representation of the protein. This not only reduces the dimensionality of the problem and, hence, the memory requirements and computational times, but also simplifies and emphasizes important movements of mobile regions.[68] When applying RCNMA, caution is required as an overly rigid representation of a protein might lead to an under-estimation of motion. Average structures obtained from MD trajectories, which are used here as reference structures, generally result in more flexible decompositions than the respective experimental structures: On average the largest rigid cluster comprises 16 % of the residues of the average structure, whereas it comprises 25 % of the residues of the experimental structure. As a more general measure for the level of coarse-graining, the dimensionality reduction (Eq. 4.4) has been introduced. Here, a dimensionality reduction of on average 0.26 for the average structures is found, whereas it is 0.32 for the open structures, in agreement with our previous results.[68] For larger proteins, the dimensionality reduction is even more pronounced. E.g., for proteins with > 200 residues in the dataset, this value amounts to 0.45.

Compared to ED modes, both ENM and RCNMA on average perform similar in terms of the maximal overlap of mode directions and correlation of amplitudes of motions (Table 5.1). There are some differences, however, on the level of individual proteins. Figure 5.1 shows differences in the maximal overlap values between ENM or RCNMA modes and ED modes as a function of the dimensionality reduction. Differences in overlap values occur in both negative and positive directions and are mainly in the range between 0.05 and 0.2, indicating that using a coarse-grained protein representation does not deteriorate the agreement in general. This is also corroborated by the fact that there is no correlation between dimensionality reduction and overlap difference values and that both positive and negative overlap differences are observed even for the highest levels of coarse-graining. Finally, no difference between ENM and RCNMA results were found if the minimized average structures were used instead. For simplicity, we thus present ENM results from here onwards unless stated otherwise.

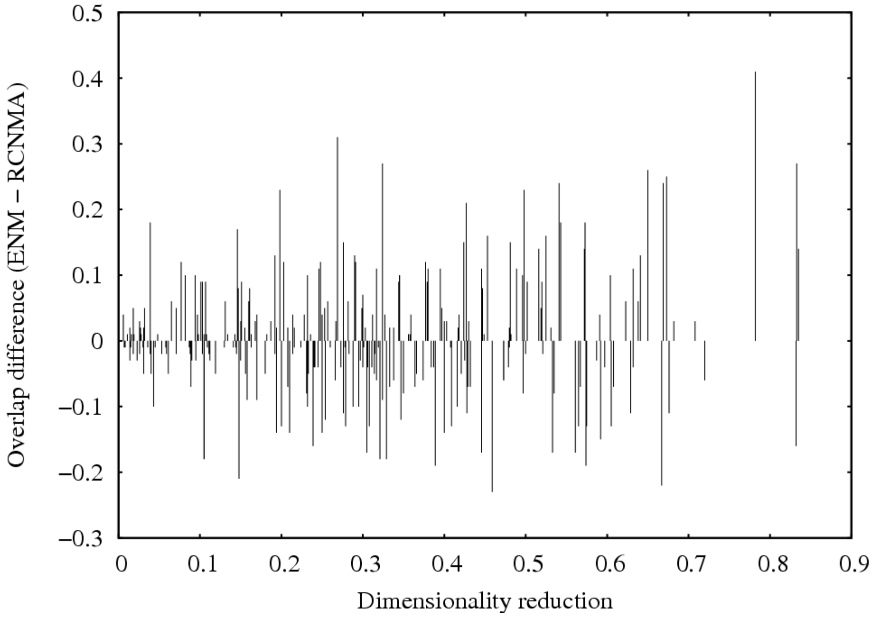


Figure 5.1: The differences between maximum overlap values for modes either obtained from ENM or RCNMA with ED modes for different proteins as a function of the dimensionality reduction (Eq. 4.4) due to coarse-graining the protein in RCNMA.

5.1.3 Comparison between ED and ENM modes

The first five ED modes of each protein of the dataset were compared with ENM modes in terms of overlap, correlation, and spanning coefficient between the two sets of modes. Despite underlying differences between ED and normal mode methods, high maximal overlap and maximal correlation values between the two sets of modes were observed. Table 5.1 shows maximal overlap and maximal correlation values averaged over 335 proteins of ED modes with ENM modes. Only 3 % of the proteins have overlap values < 0.4, indicating an unsatisfactory agreement of mode directions, whereas 83 % of the proteins have maximal overlap values > 0.5 and more than 30 % of the proteins have maximal overlap values > 0.7 (see Figure 5.2). More than one quarter of the proteins have a max overlap value between 0.60 and 0.70. These high overlap values indicate that the essential motions extracted from MD trajectories can likewise be obtained from a coarse-grained normal mode method, albeit at much lower computational expenses. Good overlap values on such a large and diverse dataset support the argument that the ENM approach is successful in describing motions of proteins with different and complex architectures, as long as it describes collective motions.[136,150,227] These collective modes, derived from both ED and ENM, have been shown previously[70,150,218,228-230] to be involved in biologically important conformational changes.

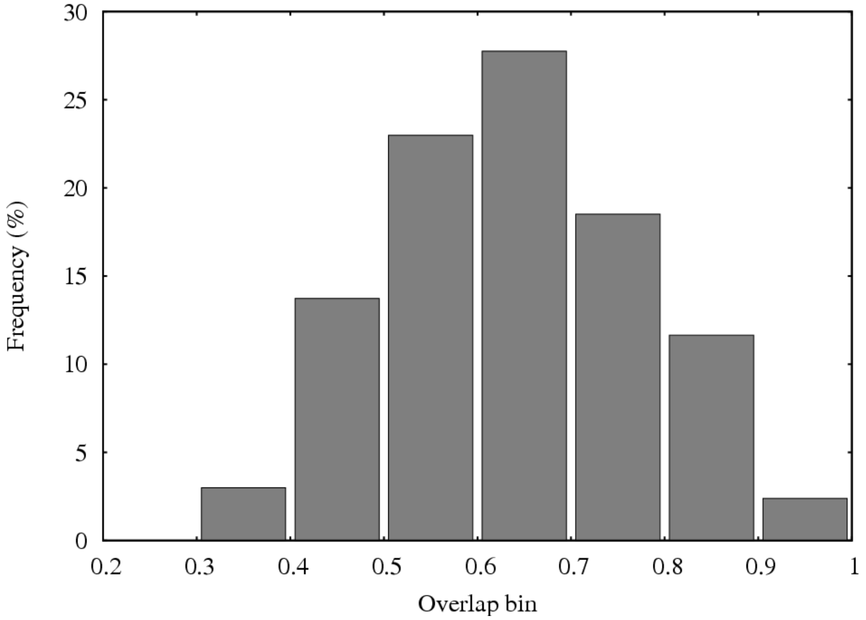


Figure 5.2: Relative frequency distribution of maximal overlap values of ENM modes with ED modes.

Additionally, the frequency distribution of ENM modes involved in the maximal overlap (Figure 5.3) shows that these modes are among the lowest frequency ones. Around 94 % of the overlapping modes are among the first five non-zero modes of ENM. Interestingly, the probability of maximal mode involvement with ED strongly decreases among the first five non-zero ENM modes i.e., the first and fifth non-zero lowest-frequency modes are considered in 45 % and 3 % respectively of all cases.

This result can be helpful in designing normal mode based approaches: it emphasizes that the trend of decreasing importance with increasing frequency of normal modes should be considered when modeling a normal mode-based approach. Interestingly, similar trend is reported[231] for experimental conformational changes on a large dataset of ~4000 proteins. Contrary to ENM, the frequency distribution of the first 5 ED modes involved in the maximal overlap does not show single mode dominance, i.e., the first and fifth non-zero lowest-frequency modes are considered in 21 % and 18 % of all cases, respectively. This is probably an effect of the presence of anharmonic modes[135,217] in ED, which are associated with crossing energy barriers during MD simulation and reside among the first few ED modes. Recently[85] it has also been found that a 1-1 correspondence doesn't exist between overlapping ED and ENM modes.

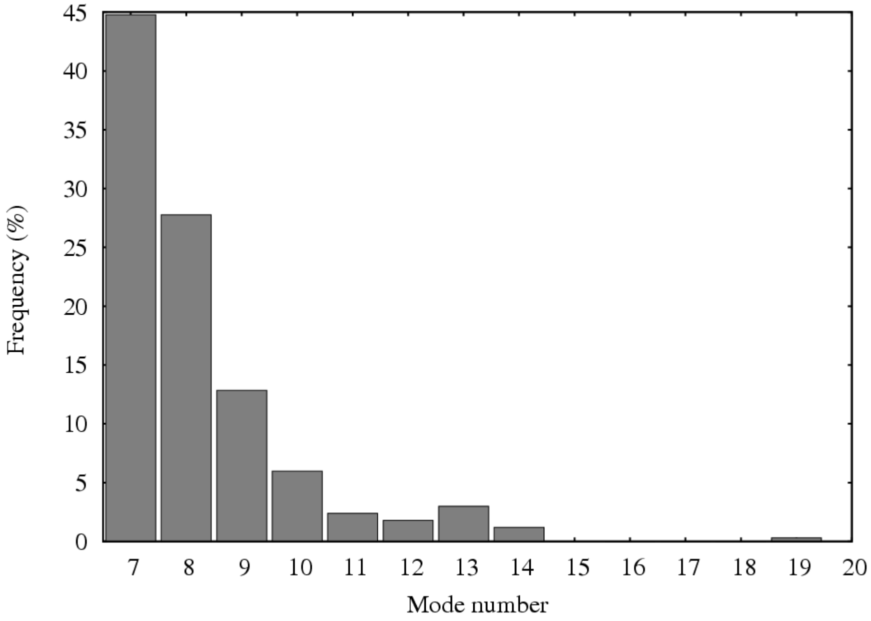


Figure 5.3: Relative frequency distribution of ENM mode numbers involved in the maximal overlap with ED modes. Mode 7 is the first non-zero frequency mode.

Correlations of the *amplitudes of motions* described by ED and ENM modes are even higher than overlap values (Table 5.1) with a mean value around 0.73, more than 94 % of the cases with a correlation value > 0.50, and still more than 40 % of the cases with a correlation value > 0.80 (Figure 5.4). This emphasizes that low frequency modes of ENM do not only well describe directions of motions but also the magnitudes of motions, in comparison to ED modes. ENM has been found[70] to well describe the magnitudes of motions for experimental conformational changes as well, even for non-collective conformational changes.

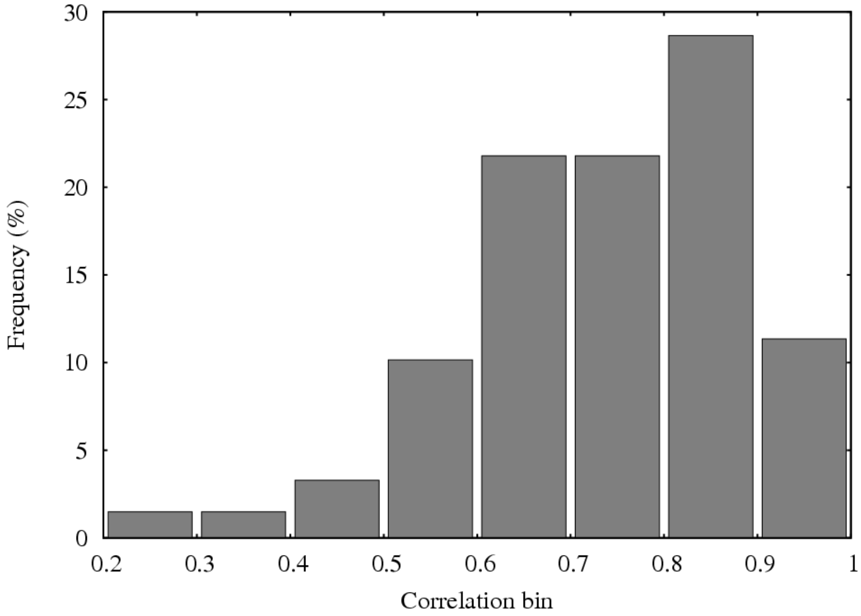


Figure 5.4: Relative frequency distribution of maximal correlation values of ENM modes with ED modes.

To analyze how well each of the five modes of ED can be described by ENM modes collectively and to explore the minimal set of the most contributing ENM modes in the low frequency range, the spanning coefficient (Eq. 4.3) was calculated with a varying mode number. It was found that only a relatively small number of normal modes are needed to describe the space spanned by low-frequency ED modes. The space spanned by the first 10, 25, and 50 % of the ENM modes describes on average around 68 %, 84 % and 92 % of all five modes of ED, respectively (Table 5.1). The spanning coefficient for all five ED modes of all proteins with a varying number of ENM modes (in percentage of the total number of modes) is shown in Figure 5.5. In the case of 10 % (i.e., on average 30) of the modes, a rather broad distribution of points shows that not all of the five ED modes are well represented. On average, the first quarter of ENM modes describes 84 % of the space of ED modes whereas the next quarter of modes describes only 8 % of the space. The last half of the modes describes another 8 % of the space. This emphasizes that the two methods, which completely differ in underlying techniques and coarse-graining levels, not only show high mean maximal overlap (i.e., 0.65) but also good overlap between the two important subspaces (derived first 5 ED modes and 30 (i.e., 10 % of all) or 85 (i.e., 25 % of all) ENM modes). Furthermore, it shows how much dynamic information a single protein structure can provide with almost no computational time. For normal mode based approaches this result can be helpful in deciding the number of modes to be considered in order to explore the essential conformational space. Similar results have been reported recently,[85] however on smaller dataset.

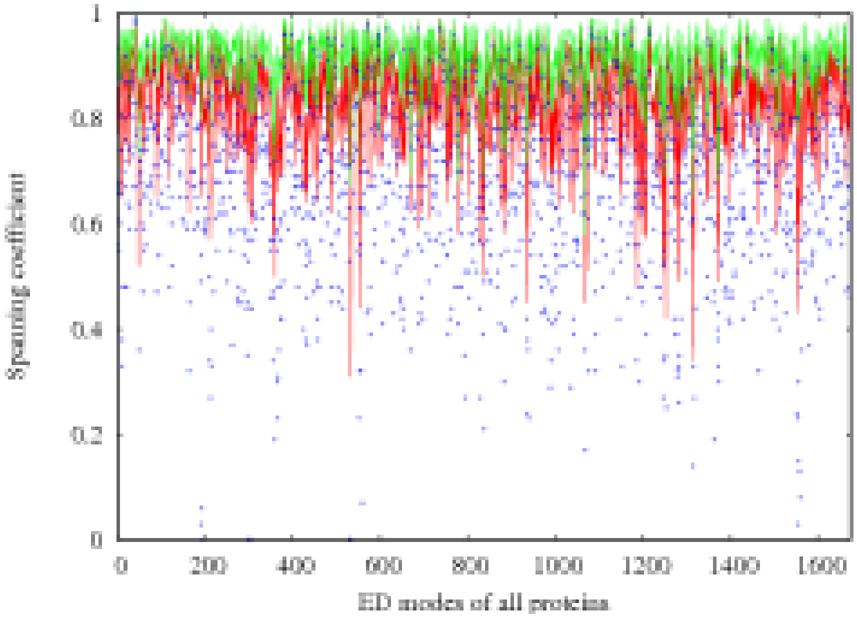


Figure 5.5: The spanning coefficient (Eq. 4.3) of 10 % (blue points), 25 % (red curve), and 50 % (green curve) of all ENM modes as a function of the first five ED modes of all proteins of the dataset. Numbers 1-5 relate to the first five ED modes of the first protein in the dataset, numbers 6-10 of the second protein and so on.

Table 5.1: Comparison of ED modes with ENM and RCNMA results

Methods	**Max overlap** [a)]			**Max correlation** [a)]			**Mean spanning coefficient** [d)]		
	Mean [b)]	**Bad** [c)]	**Good** [c)]	**Mean** [b)]	**Bad** [c)]	**Good** [c)]	**10 %**	**25 %**	**50 %**
ENM	0.65	3 %	83 %	0.73	3 %	94 %	0.68	0.84	0.92
	(0.31, 0.93)			(0.22, 0.98)			(30)	(85)	(176)
RCNMA	0.64	3 %	80 %	0.74	2 %	94 %	0.59	0.78	0.87
	(0.34, 0.95)			(0.26, 0.98)			(20)	(58)	(123)

a) Maximal overlap (Eq. 4.1) or maximal correlation (Eq. 4.2) between ENM/RCNMA and ED modes. b) Average over all 335 proteins in the dataset with lowest and highest values in brackets. c) Percentage of maximal overlap/maximal correlation values < 0.4 (bad) and > 0.5 (good). d) Mean spanning coefficient (Eq. 4.3) over all proteins in the dataset using 10, 25, and 50 % of all available modes. The average number of modes used in each case is given in brackets.

5.1.4 Similarities/dissimilarities in classes/folds: ED and ENM modes

In order to analyze the dynamic similarities/dissimilarities within different classes/folds based on the normal modes and/or essential dynamic modes, the CATH classification was incorporated in our dataset of proteins (as described in methods). Maximal overlap and correlation of amplitudes of motions described by ENM and MD were sorted for these proteins according to different classes/folds, and the mean and the standard deviation values were calculated (see Appendix C). With respect to maximal overlap and correlation in amplitudes between ED modes and ENM, no prominent differences among different classes (i.e., α, β, α+β, and *few secondary structures*) were found (see Table 5.2) considering standard deviations of around 0.1 for all classes. This is in accordance with the recent study on a relatively smaller dataset.[85] This shows that on average ED modes and normal modes do not differentiate on the bases of different classes and therefore normal modes are equally applicable to proteins in different classes.

Additionally, collectivity of the modes involved in maximal overlap between ED and ENM methods were sorted according to CATH classes. A prominent trend of a low collectivity index of ED modes for β and α+β classes as compared to the other two classes was found (see Table 5.2). Although the difference of 0.06 is less significant considering the standard deviation of 0.1, it can not be ignored. It should be noted that collectivity index (Eq. 4.5) does not necessary correspond to correlated motions but motions involving most of the atoms. In this view, relatively high collectivity indices in the α class are probably due to lower packing[232] as compared to β or α+β classes, which provides the required space for collective motions of atoms, whereas β or α+β classes do not show such collectivity due to higher packing. On the contrary, the high collectivity index in the *few secondary structure* class could be attributed to the underlying flexibility of the structures due to lack of secondary structure which probably results in uncorrelated motions (involving most of the atoms).

Table 5.2: Mean results for different protein classes.

Class No. [a)]	**Collectivity index** [b)]		**Max overlap** [c)]	**Max correlation** [c)]
	ED	**CGNM**		
1 (90)	0.41 (±0.11)	0.34 (±0.14)	0.64/0.65	0.73/0.72
2 (103)	0.35 (±0.10)	0.32 (±0.16)	0.64/0.65	0.75/0.73
3 (122)	0.35 (±0.10)	0.30 (±0.13)	0.62/0.63	0.74/0.75
4 (5)	0.42 (±0.13)	0.38 (±0.05)	0.65/0.67	0.75/0.76

a) Protein classes α, β, α+β, and some secondary structures are numbered 1-4, respectively. In brackets are the numbers of domains in the respective class. b) Mean collectivity index (Eq. 4.5) of modes involved in maximal overlaps between ED and ENM modes with standard deviation in brackets. c) Mean of the maximal overlap (Eq. 4.1) or correlation values (Eq. 4.2) of ED and ENM modes.

Furthermore, it was interesting to investigate whether low overlap values observed in some cases as compared to the other members in the same fold and family (Topology and Homologous superfamily levels in CATH) is a limitation of ED or ENM. It is worth noticing that; in general, proteins of the same fold family (topology) show a similar overlap between ED and ENM with a standard deviation of around 0.1 (see Appendix C). However, some of the proteins have an extended N- or C-terminal chain, which results in high overlap between ED and ENM (for example PDB code

1ngl) regardless of their folds. In order to investigate the low maximal overlap in some cases, three pairs of proteins from each of the three main classes of CATH classification were selected such that, despite belonging to the same topology and homologous superfamily, the proteins in each pair highly differ in maximal overlap/correlation values (values are highlighted in bold in Appendix C). These selected pairs are listed in Table 5.3. Assuming that the proteins in the same fold and family should have similar dynamics,[233] the modes derived from either of the method, i.e., ED or ENM of the two selected proteins in each pair were compared in terms of maximal overlap (Eq. 4.1) and maximal correlation (Eq. 4.2). Interestingly, in all three cases the maximal overlaps and maximal correlations in amplitudes of motions obtained from ENM were found to be higher than ED (see Table 5.3). The mean maximal overlap values are 0.31 and 0.56 using ED and ENM respectively, and the mean maximal correlation values are 0.57 and 0.84 using ED and ENM respectively. This illustrates that ENM modes are more robust within a fold than ED modes. Moreover, this might be an indication that in some cases MD simulation time of 10 ns might not be long enough to explore the required conformational space needed to represent the intrinsic motions of a protein.

It is interesting to mention here that functional modes are usually among the most robust modes[177], even to sequence variations.[234] Furthermore, Leo-Macias *et al.*[225] have concluded that, to a significant extent, the structural response of a protein topology to sequence changes takes place by means of collective deformations along combinations of a small number of low-frequency modes. Recently, it has also been argued that dynamics and functional promiscuity are foundation stones of protein evolvability.[235] In this view, results presented here for three selected proteins show that ENM better describe these robust and evolutionary modes than ED and probably MD simulation is restricted in capturing these modes due to slow barrier crossing on the rugged energy landscape.[59,60]

Table 5.3: Comparison of ED and CGNM modes within folds.

PDB codes [a)]	**Max overlap** [b)]		**Max Correlation** [c)]	
	ED	**ENM**	**ED**	**ENM**
1ahq/1cof (43 %)	0.28	0.59 (7)	0.58	0.92
1ccr/1co6 (50 %)	0.26	0.46 (15)	0.41	0.72
1idi/1ntn (66 %)	0.38	0.64 (7)	0.72	0.87

a) PDB codes of three selected protein pairs. Both proteins of a pair are in same fold family and homologous superfamily[201] but highly differ in their maximum overlap values between ED and CGNM. In brackets, the sequence identities of the protein pairs are given. b) Highest overlap (Eq. 4.1) between two sets of modes of each pair of proteins using ED/ENM. The respective ENM mode number is given in brackets. c) Highest correlation (Eq. 4.2) between two sets of modes of each pair of proteins using ED/ENM.

In short, the results in this section validate the directional information obtained from the CGNM approaches, and therefore this information is incorporated in the NMSim approach to guide backbone motions. In the next section, the NMSim approach is validated on hen egg white lysozyme by comparing it to state of the art MD simulation and different experimental structures.

5.2 Comparison of the performance of NMSim to other conformation generation methods

In this section, the advantages and limitations of the different geometry-based approaches are compared to the NMSim approach. The Hen Egg White Lysozyme (HEWL) protein is selected as a test case in this study. HEWL is a well studied[130,220,236-238] protein, comprised of 129 residues, that has also been used for the evaluation of the quality of force-fields.[200,207,239-241] Here, HEWL conformations[207] from a state-of-the-art MD[56-58] simulation and different experimental structures are compared with conformations obtained from the most efficient geometry-based methods FRODA,[64] CONCOORD[62,63] and NMSim (see section 4.2). This section discusses the results from a comparison of the different simulation methods in terms of residue fluctuations, conformational space exploration, essential dynamics,[215,216,242] sampling of side-chain rotamers, and structural quality.

5.2.1 Residue fluctuations and correlations

In order to compare patterns of atomic fluctuation obtained from different methods, the root mean square residue fluctuations (i.e. mass-weighted average of heavy-atom fluctuations for each residue) were calculated for the structural ensembles of MD, NMSim, FRODA, CONCOORD, and experimental structures (see section 4.2). The mobile regions and the magnitudes of fluctuations are well predicted by NMSim taking MD structures and experimental structures as references (see Figure 5.6-a). For example, high fluctuations are observed for residues 45-50 and 68-78, which are associated with β-sheets and turns at the outer edge of the upper lobe of the molecule. This is in accordance with earlier theoretical[130,220,237,241] and experimental[209] studies. Similarly, the regions which are stable, especially the hydrophobic core (i.e., residues 6–15, 25–36 and 89–100) formed by three α-helices (helices A, B and C), show low fluctuations and correlate well with the MD and experimental fluctuations. Differences between the fluctuations are observed for the tail region of the protein, which was found to be highly fluctuating in NMSim, which can be attributed to the "tip-effect" in coarse-grained normal modes.[243] A "tip-effect" results by an imbalance of elastic forces among neighboring harmonic oscillators (C_α atoms) due to lighter packing around “tip” regions. This happens in systems with structural components, the “tips”, protruding out of the main body, e.g., an isolated surface loop or a protruding tail region of N- or C-terminal residues. As a result, relatively higher fluctuations are usually observed for those “tip” regions in coarse-grained normal modes.

The fluctuations obtained from the CONCOORD ensemble also agrees well with MD and experimental fluctuations. However, relatively higher magnitudes of fluctuations occur in the two mobile regions as compared to the MD fluctuations (see Figure 5.6-b). The FRODA simulation underestimates the overall fluctuations of residues resulting in lower magnitudes of fluctuations compared to all other methods. Moreover, even considering relative fluctuations, mobile regions are not well predicted in FRODA. For example, low fluctuations can be seen in the two mobile regions of residues 45-50 and 68-78 as compared to the MD and experimental fluctuations.

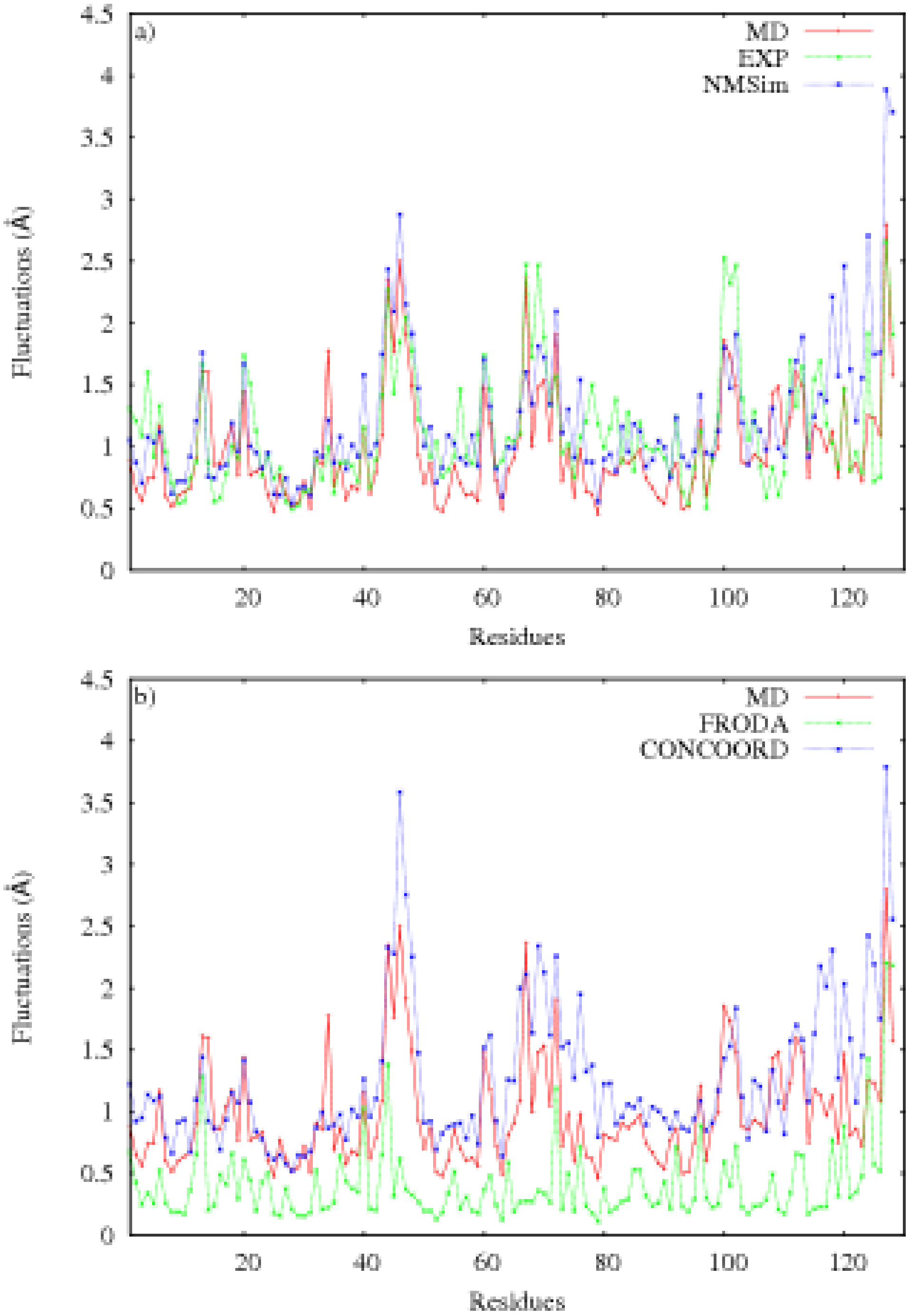


Figure 5.6: The mass-weighted average of heavy-atom fluctuations for each residue in HEWL obtained from the MD (red), experimental (green in panel a), NMSim (blue in panel a), FRODA (green in panel b) and CONCOORD (blue in panel b) structural ensembles. For clarity, fluctuations plots are divided into top and bottom graphs with MD fluctuations (red) as a reference.

The correlation coefficients between the residue fluctuations obtained from the structural ensembles of the different methods and 130 experimental structures are shown in

Table 5.4. Residue fluctuations generated by NMSim and CONCOORD were found to be in good agreement with MD fluctuations with a correlation coefficient of around 0.79 each. Similarly, residue fluctuations obtained from these two methods also highly correlate with fluctuations of the experimental structures (correlation coefficient of around 0.7, respectively). In contrast, the fluctuations obtained from the FRODA ensemble reach only low correlation coefficients of 0.57 and 0.5 with MD and experimental fluctuations. In general, NMSim showed high correlations with every method including FRODA in terms of residue fluctuations.

Table 5.4: The correlation coefficients of residue fluctuations between different methods.

	MD [a)]	**EXP** [b)]	**CONCOORD** [a)]	**FRODA** [a)]
NMSim [a)]	0.792	0.688	0.896	0.792
FRODA	0.568	0.492	0.574	
CONCOORD	0.789	0.702		
EXP	0.730			

a) The mass-weighted average of heavy-atom fluctuations (residue fluctuations) for each residue of HEWL from the structural ensembles of the different methods. b) The residue fluctuations from the structural ensemble of experimental structures of HEWL, which contains 130 structures from both X-ray crystallography and NMR.[209]

5.2.2 Conformational space exploration

In contrast to the MD simulation, progression of a trajectory in a geometry-based simulation is usually measured in terms of RMSD from a reference structure.[64] The plots showing the backbone RMSD between the trajectories/conformers of the different methods and the starting structure are shown in Figure 5.7-a. All methods, except FRODA, show a considerable backbone RMSD from the starting structure, predominantly in the range between 1 to 2 Å. Average backbone RMSD for the different trajectories/conformations from the starting structure were found to be 1.03 Å, 1.40 Å, and 1.26 Å, for MD, NMSim, and CONCOORD, respectively, whereas only 0.37 Å in the FRODA ensemble. This shows that FRODA underestimates the conformational mobility available to a protein structure in terms of backbone RMSD. FRODA has been shown to predict mobile regions in barnase and qualitatively predict observed displacements between open and close form in maltodextrin binding protein.[51,64] However, this study shows that FRODA does not fully explore the backbone conformational space available to HEWL. Interestingly, the FRODA[64] and NMSim approaches share a natural way of coarse-graining,[51] i.e., rigidity analysis using FIRST approach,[91,161,204] at their core levels. However they differ at simulation levels. FRODA uses diffusive motion[64] of rigid regions. Therefore, due to the lack of directions, sampling in FRODA is limited, particularly in the cases where proteins are relatively flexible. In contrast, NMSim uses normal mode directions to guide backbone motions and is therefore less restricted in sampling protein conformational space, at least in this particular case.

CONCOORD explores a conformational space randomly without following any path or trajectory; in this case the minimum and the maximum RMSD with the starting structure was found to be 0.61 Å and 2.34 Å respectively. NMSim explores conformations in a similar range as CONCOORD with relatively higher and more frequent peaks as compared to the MD trajectory. This reflects the coarse-grained nature of the energy landscape which makes it easier to get over barriers. In CONCOORD, however, to get over barriers, each conformation is generated independently of the other, using the starting structure distortion and correction procedure.[62]

The RMSD plots for heavy-atoms (Figure 5.7-b), as expected, show similar patterns as the backbone RMSD plots (Figure 5.7-a). On average, higher heavy-atom RMSD values as compared to the backbone RMSD values for every method were found, which is not surprising, as this is an indication of higher mobility in the side-chains of the protein than its backbone. The average heavy-atom RMSD for the conformations in the structural ensembles of MD, NMSim, FRODA, and CONCOORD with the starting structure were found to be 1.56 Å, 1.86 Å, 1.00 Å and 1.41 Å, respectively. Every method shows an increase in the RMSD values for heavy-atoms compared to backbone-atoms; this increase is 0.67 Å, 0.53 Å, and 0.46 Å in FRODA, MD and NMSim ensembles, respectively, but only 0.15 Å in the CONCOORD ensemble. Although FRODA underestimated the backbone mobility, it extensively explores the side-chain conformational space. In contrast, CONCOORD does not show this high-mobility behavior for side-chain regions, and therefore, it might be restricted in sampling side-chain conformations (as found in section 5.2.4). So far, CONCOORD has been mainly used for generating backbone conformations: The novel use of CONCOORD generated structures has been to get eigenvectors of essential dynamics using backbone atoms; whether it is docking to multiple eigenstructures,[65] analyzing conformational changes in macromolecular assemblies,[171] or exploring different biological mechanisms.[172-175]

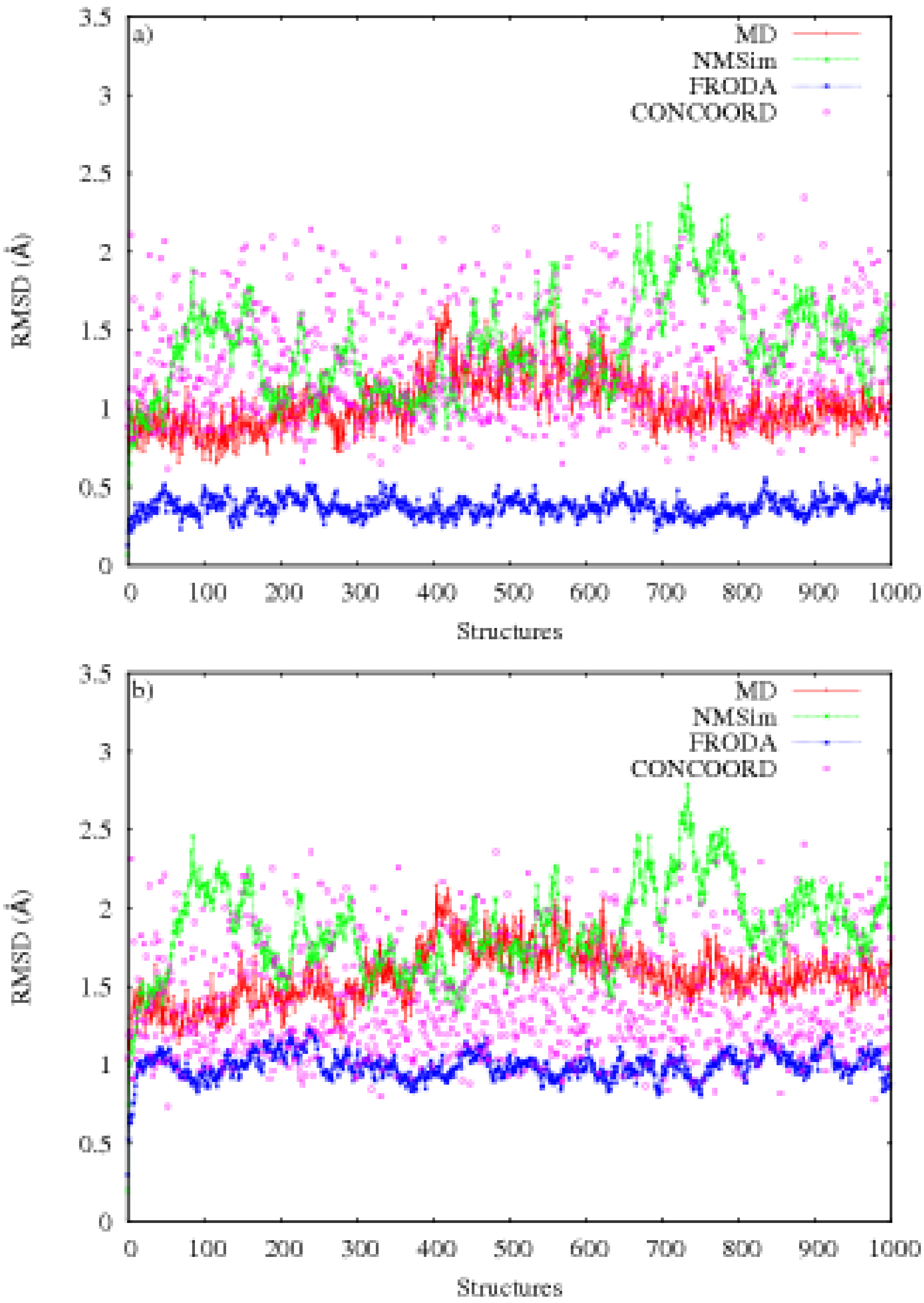


Figure 5.7: Backbone RMSD (a) and heavy-atoms RMSD (b) between the starting structure and the structural ensembles obtained from MD (red), NMSim (green), FRODA (blue), and CONCOORD (magenta). The FRODA trajectories explore only a limited conformational space compared to the other methods.

5.2.3 Essential dynamics

The conformational space exploration in terms of RMSD does not show if two methods explore essentially similar conformational spaces. In order to verify this, essential dynamics (ED)[215,216,242] calculations were performed on the structural ensemble of MD, and the conformations from the different methods were projected onto the plane described by the first two ED modes with the highest Eigen values. ED analysis has previously been extensively applied to extract essential and collective modes,[229] for example from MD trajectories, and has been used not only to investigate protein dynamics[216,218] but also to compare conformational spaces with experimental[244] or generated[62,63] structures. Here, Figure 5.8 shows the 2D projection of conformations from the different methods and the experimental structures onto the plane defined by the first two ED modes derived from the MD ensemble. It describes the maximum diversity of the conformational space that can be captured by different methods in terms of the two essential directions of conformational change explored by MD. As expected, the projection of the MD structural ensemble onto the plane shows an eclipse shape with the major axis aligned with the first principal direction (Figure 5.8-a).

NMSim and CONCOORD conformations were found to be well distributed along the principal directions of MD (see Figure 5.8). The CONCOORD structural ensemble shows the highest diversity onto the plane with the mostly dispersed points. This reflects the uncorrelated nature of the generated conformations. However, it shows that CONCOORD, a simple constraint based method, can efficiently explore the essential conformational space of HEWL as shown previously for other systems.[62,63] NMSim also shows a diverse projection of conformations onto the ED plane, which validates that the conformational exploration in normal mode space are in agreement with the essential dynamics of the MD ensemble. Different previous studies[85,219] showed the striking similarities between normal modes and ED modes. The clustered nature of the points in Figure 5.8-b reflects the typical behavior for trajectories indicating different local minima.

The FRODA conformations capture a very small portion on the ED plane (Figure 5.8-c) compared to the other methods. This confirms the conclusion derived from the

RMSD plots (Figure 5.7): the conformational space explored by FRODA simulation is considerably smaller than compared to the other methods. Furthermore, Figure 5.8-c shows that the FRODA explored conformational space is restricted in exploring the MD principal directions compared to the other methods.

The projections of the 130 experimental structures onto the MD principal directions plane are shown in Figure 5.8-e. The diversity of the points in the MD principal directions plane is relatively less compared to the MD, NMSim, and CONCOORD ensembles but still more than the FRODA ensemble. Interestingly, the two clusters identifiable in Figure 5.8-e can be assigned to a top cluster of 50 NMR[209] structures (PDB code 18el) and a bottom cluster comprising all X-ray crystal structures.

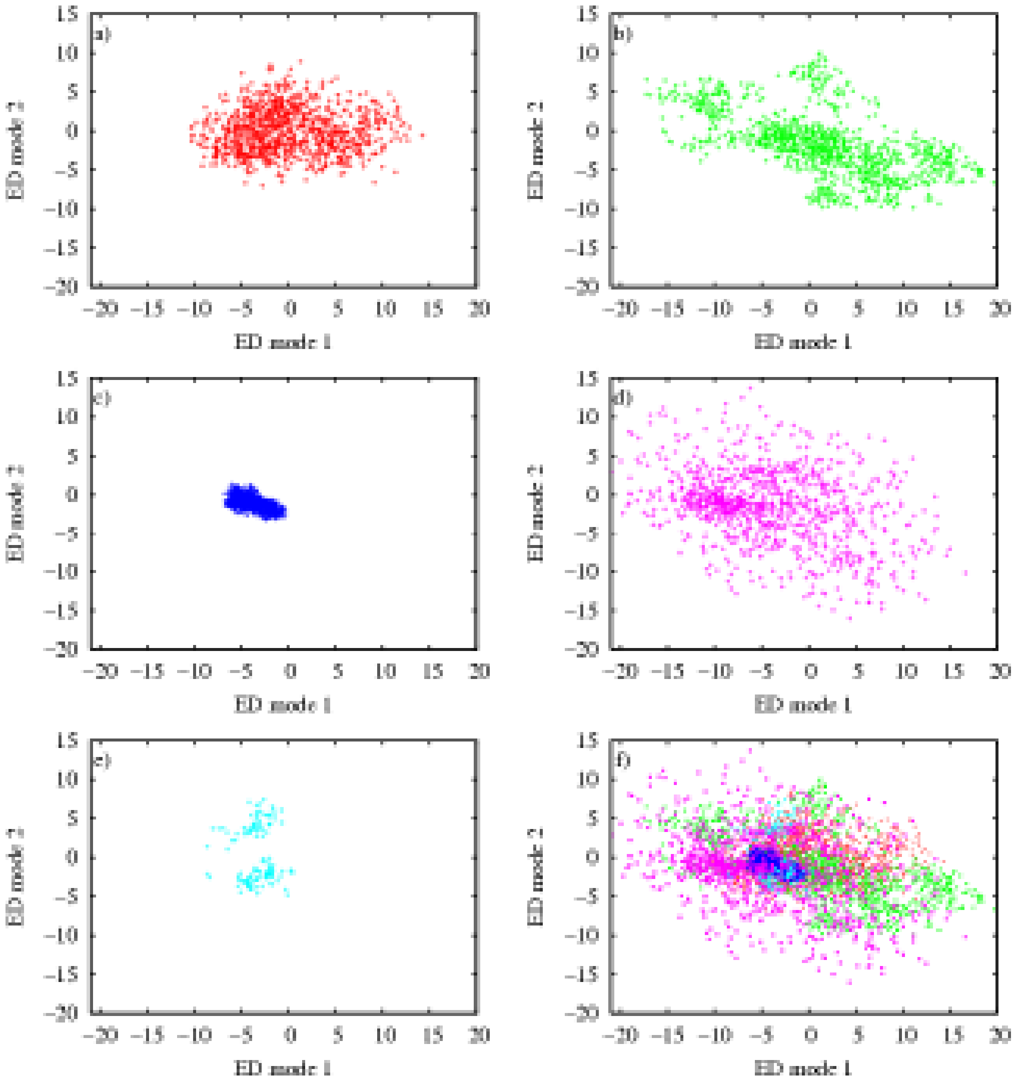


Figure 5.8: The projections of the structural ensembles of 1000 conformations of HEWL obtained from a) MD (red), b) NMSim (green), c) FRODA (blue), d) CONCOORD (magenta), and e) 130 experimental structures (cyan) onto the plane described by the first two ED modes of MD conformations. A collective view is shown in panel (f) by superimposing projections a)-e).

5.2.4 Side-chain flexibility and rotamers

To compare side-chain quality and flexibility in terms of rotamer sampling, the rotamer derived measures (see section 4.2.2), i.e., heterogeneity, occupancy, and rotamericity, were calculated for the structural ensembles of the different methods. Rotamers have been successfully used to account for side-chain flexibility in docking applications.[77,245-247] With the increasing amount of experimental data, many rotamer libraries have been published.[179,190,191] In this study, the Penultimate rotamer library[179] from the Richardson lab has been used. A recent review has regarded the Penultimate rotamer library as the best among the available backbone-independent rotamer libraries.[191]

To analyze how well the different methods sample available rotamer states, a rotamer heterogeneity measure of each HEWL residue was calculated over the structural ensembles of the different methods. The rotamer heterogeneity derived from the 130 experimental structures was taken as reference (see Figure 5.9). Here CONCOORD, which was found to explore good backbone conformation space, poorly explores different rotamer states as compared to the experimentally observed rotamer states (see Figure 5.9-d). None of the residues in the CONCOORD ensemble was found to explore the full range (i.e., heterogeneity = 1) of available rotamer states, whereas the experimental structures show a heterogeneity = 1 for 13 out of 103 residues (i.e., excluding GLY, ALA, and PRO). Furthermore, almost all heterogeneity values observed in the CONCOORD ensemble are lower than the experimentally derived values. This is an interesting observation, since conformations in CONCOORD are generated from randomized atomic positions[62] and thus should be sampling diverse sets of rotamer states. This poor sampling of side-chains should be considered before using CONCOORD structures in side-chain sensitive applications such as ligand docking.

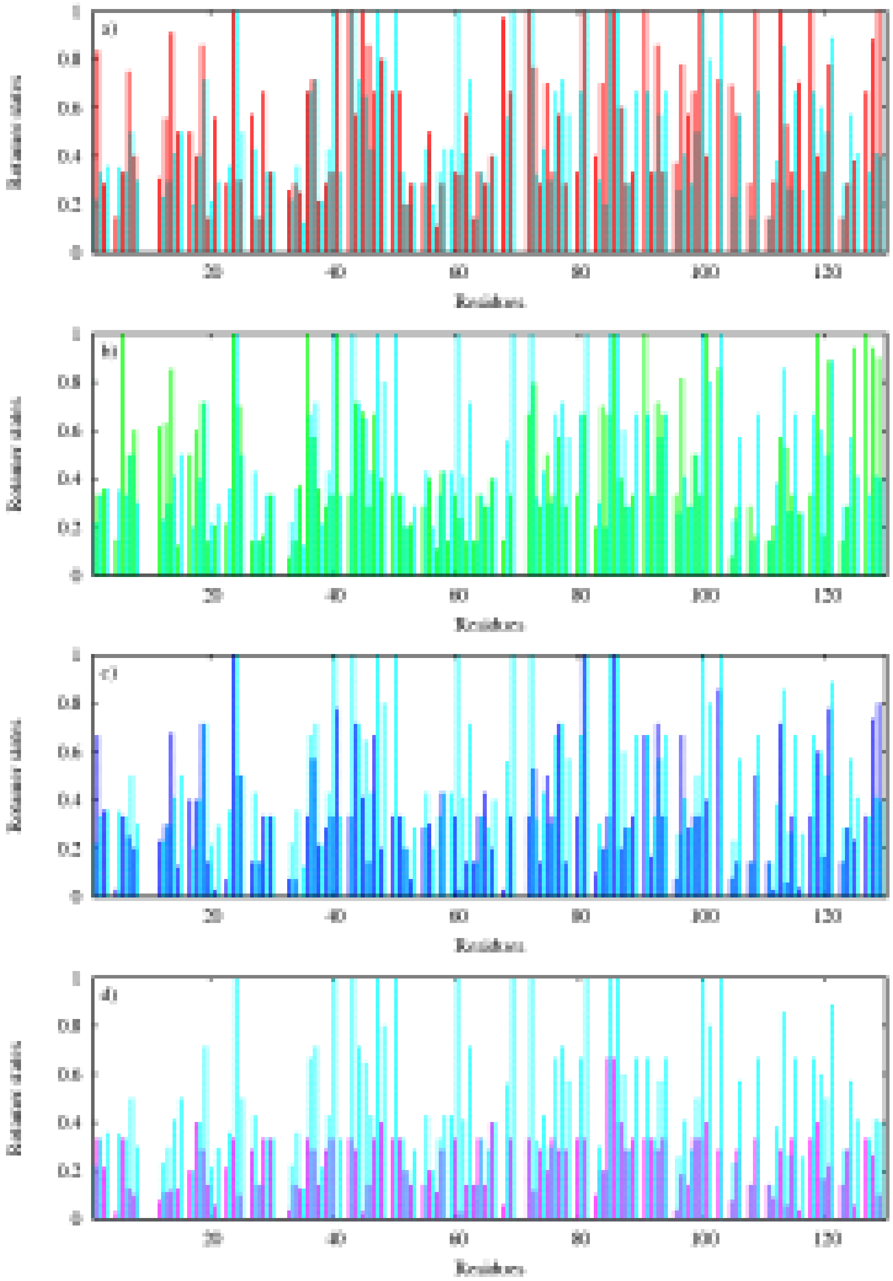


Figure 5.9: The rotamer heterogeneity of HEWL residues in the structural ensembles of 1000 structures obtained from MD (red in panel a), NMSim (green in panel b), FRODA (blue in panel c) and CONCOORD (magenta in panel d). The rotamer heterogeneity values derived from 130 experimental structures (cyan in a-d) are shown as reference in all graphs.

MD, NMSim, and FRODA show a similar pattern of the rotamer heterogeneity, which is also similar to the pattern derived from 130 experimental structures (Figure 5.9 a-c). Small differences occur in the mobile regions of HEWL (i.e., residues 40 to 60), where all methods, especially FRODA, show lower heterogeneity than is experimentally observed. In contrast, in the tail region, all methods show higher heterogeneity than in the experiments. Comparing different methods reveals that MD explores more rotamer states than NMSim, whereas NMSim samples more states than FRODA. This can also been seen by the average heterogeneity values over 103 HEWL residues for different methods (see Table 5.5).

The average of the “rotamer occupancy” measure (see section 4.2.2) can be used to quantify the diversity of the rotamer states captured in an ensemble, and thus reflects the flexibility available to side-chains. It should be noted that the highest possible average rotamer occupancy is ~10 for an HEWL ensemble; i.e., if in a hypothetical case every residue (103 residues, excluding GLY, ALA, and PRO) of HEWL in the ensemble samples all possible rotamer states available in the rotamer library. MD, NMSim, FRODA, and CONCOORD on average sample 5.78, 4.97, 3.14 and 1.63 rotamer states, respectively, out of 10 (see Table 5.5). Here CONCOORD shows around 2.7 times less diversity in rotamer states than the experimentally observed 4.41. This again shows a restricted conformational space available to side-chains in structures generated by CONCOORD. Contrarily, a high average occupancy value for NMSim as compared to FRODA and CONCOORD is observed, which justifies the specific modeling of rotamer states in geometry-based conformational modeling. The correlation coefficient between the occupancy vectors (103-dimensional vector containing occupancy values) is shown in Table 5.5 (see section 4.2.2) which compares the patterns of rotamers sampled in the different ensembles. NMSim was found to have a higher correlation coefficient of 0.71 and 0.80 with the experimental and the MD derived vectors, respectively, as compared to FRODA and CONCOORD.

In order to analyze the probability for any rotamer state to exist for each residue in a protein sequence, the rotamericity measure is calculated over an ensemble of structures (see section 4.2.2). This is related to the quality of side-chains in the ensemble in terms of rotamers. The average rotamericity for 103 residues (Table 5.5) shows a higher value for CONCOORD compared to NMSim and FRODA. This can be expected, if there is a tendency of a method to keep a rotamer state as found in the starting structure over the trajectory/ensemble. However, the average rotamericity measure for NMSim (0.698) and FRODA (0.685) are comparable to the experimentally found value of 0.731, whereas for MD it is even 0.816 (see Table 5.5).

Table 5.5: The rotamer derived measures for different structural ensembles.

Methods	**Average values** [a]			**Occupancy vector** [e]	
	Heterogeneity [b]	**Occupancy** [c]	**Rotamericity** [d]	**EXP**	**MD**
EXP	0.498	4.407	0.731	1.000	0.861
MD	0.537	5.786	0.816	0.861	1.000
NMSim	0.459	4.970	0.698	0.713	0.808
FRODA	0.338	3.145	0.685	0.569	0.733
CONCOORD	0.228	1.631	0.752	0.438	0.520

a) The averages of different measures are calculated over 103 out of 129 residues of HEWL (excluding GLY, ALA, and PRO). b) The heterogeneity measure of a residue in a protein sequence is defined as the ratio of the total number of distinct rotamer states of the residue observed in an ensemble to the total number of available rotamer states for that residue in the rotamer library.[179] c) The occupancy measure of a residue in a protein sequence is defined as the total number of distinct rotamer states of the residue observed in an ensemble. d) The rotamericity of a residue in a protein sequence is defined as the ratio of the total number of occurrences of the residue in any of the possible rotamers to the total number of conformers in the ensemble. e) The correlation coefficients between the different occupancy vectors of different methods. Occupancy vector in HEWL is a 103-dimensional vector containing occupancy values of the residues.

5.2.5 Structure quality using Procheck

The quality of a subset of the structures obtained from the different types of methods was analyzed using the Procheck[194] program (see section 4.2.3). The averages and the standard deviations were calculated for the different properties obtain from Procheck. Table 5.6 summarizes the Procheck results including Ramachandran plot distribution, *G*-factors, and planar groups.

Procheck divides the Ramachandran plot into four areas: core, additionally-allowed, generously-allowed, and disallowed. Every method shows a good Ramachandran plot distribution with almost zero percent of the structures located in disallowed or generously allowed regions and with a highly populated core region. Specific modeling of φ/ψ constraints in NMSim results in the highest core region population on average (i.e., 92 %) as compared to the other methods. Remarkably, this is in agreement with the high resolution experimental structures EXPTOP.

The Procheck *G*-factor provides a measure of how normal a given stereo-chemical property is. This value is computed for dihedrals angles and covalent geometry. A low *G*-factor indicates that the property corresponds to a low-probability conformation; ideally, the *G*-factor value should be above -0.5, whereas structures with values below -1.0 may need investigation. Table 5.6 shows that for every method except for MD the overall *G*-factor value is higher than -0.5. Notably, the covalent *G*-factor (i.e., main-chain bond lengths and main-chain bond angles) for MD is as low as -1.5. NMSim on average achieves 100 % planarity for the planar groups. Considering experimental EXP/EXPTOP, other methods also give acceptable planarity for planar groups except for MD, which gives around 56 % planarity on average. In short, structure quality properties for all methods are within acceptable ranges, as compared to the properties derived from experimental structures, except for main-chain bond lengths and side-chain planarity from MD derived structures. The poor quality of MD structures is understandable as the MD simulation is performed at 300 K, whereas geometry-based methods implicitly minimize each structure during correction cycles.

Table 5.6: The averages and standard deviations for quantities determining structure quality.

Methods	**Ramachandran plot a)**				***G*-factor b)**			**Planar c)**
	Core	**Allow**	**Gen.**	**Disal.**	**Dihe.**	**Cova.**	**Over all**	
MD	84.36	15.05	0.59	0.00	-0.47	-1.51	-0.86	56.08
	±2.73	±2.80	±0.70	±0.00	±0.04	±0.08	±0.04	±5.57
NMSim	92.55	7.42	0.01	0.00	-0.26	-0.36	-0.30	100.00
	±1.61	±1.62	±0.12	±0.00	±0.51	±0.01	±0.31	±0.00
FRODA	88.05	11.90	0.04	0.00	-0.05	-0.36	-0.17	92.02
	±1.30	±1.28	±0.19	±0.00	±0.20	±0.02	±0.12	±2.22
CONCOORD	85.93	13.91	0.14	0.00	-0.09	-0.51	-0.23	92.54
	±2.08	±2.04	±0.33	±0.00	±0.06	±0.12	±0.08	±4.50
EXP	81.31	17.62	0.89	0.18	-0.07	0.24	0.06	98.34
	±7.60	±6.72	±1.22	±0.41	±0.26	±0.91	±0.38	±3.82
EXPTOP	91.26	8.30	0.28	0.14	0.06	-0.28	-0.05	92.82
	±3.80	±3.62	±0.59	±0.33	±0.26	±0.52	±0.25	±11.25

a) Averages/standard deviations of percentages of φ/ψ torsion angles found in different regions (i.e., core, allowed, generously allowed, and disallowed) in the Ramachandran plots of the structural ensembles. b) Averages/standard deviations of Procheck derived *G*-factors for the structural ensembles. A low *G*-factor indicates that the property corresponds to a low-probability conformation. Ideally, *G*-factor should be above -0.5, whereas a value below -1.0 indicates that the structure may need investigation. Procheck calculates *G*-factors for dihedral angles, covalent geometry and overall. c) Averages/standard deviations of percentages of side-chain planarity found in the structural ensembles.

In short, the NMSim approach described in chapter 3 was validated on hen egg white lysozyme in this section. NMSim sufficiently samples both the backbone and the side-chain conformations taking experimental structures and conformations from the state of the art MD simulation as reference. A comparison of different geometry-based simulation approaches shows that FRODA is restricted in sampling the backbone conformational space and CONCOORD is restricted in sampling the side-chain conformational space. NMSim produces structures of a good structural quality. Furthermore, the explicit modeling of rotamer states in NMSim improves the quality of side-chain conformations as compared to without modeling in NMSim and as compared to the other geometry-based approaches. The NMSim approach will be used for exploring biologically relevant motions in the following section.

5.3 Performance of NMSim in exploring biologically relevant conformational changes

Specific functions of biological systems often require conformational transitions of macromolecules. Such changes range from large-scale domain motions to localized loop motions or even single side-chain rearrangements. Understanding the underlying dynamics and knowledge of the different conformational states of macromolecules are important in structure-based drug design (SBDD).[39,44,248,249] In addition to experimental techniques like X-ray crystallography and NMR and theoretical simulation techniques like MD, efficient coarse-grained techniques have also gained importance in describing conformational sub-states and intrinsic motions of macromolecules. Biologically important conformational changes in proteins have been found along low-frequency normal modes.[130,237] Utilizing coarse-grained normal modes, efficient approaches[82,87,88,250] have been developed for conformational pathway and intermediate structure generation between unbound and ligand bound conformations.

The large-scale comparison, shown above in section 5.1 , of essential dynamics (ED) modes from MD simulations and normal modes from coarse-grained approaches further establishes that not only large-scale motions but also intrinsic dynamics from MD essentially follow the directions of low-frequency normal modes. Consequently, the NMSim approach, described above, has been developed, which efficiently exploits structural information available at different levels, i.e., structural rigidity, normal mode directions, rotamer, and stereo-chemical information. In this section, applications of the approach in describing biologically important conformational changes will be described. The usefulness and limitations of the approach will be discussed in detail for important domain and loop motions of different types. It is suggested that a reduction of the radius of gyration (R_g) if used in combination with low-frequency normal modes improves the search for ligand bound conformations.

5.3.1 Domain motions

The domain dataset in Table 4.1 contains diverse proteins in terms of their structures, sizes, and motions. Adenylate kinase (ADK) contains three domains in contrast to two domains for most of the other proteins. The number of residues range from 148 for calmodulin (CLM) to 860 for citrate synthase (CTS). ADK and lysine/argnine/ornithine-binding protein (LAO) show global and hinge-bending motions of domains in contrast to aspartate aminotransferase (AST) and CTS, which show relatively localized motions of small domains and sheer motions.[251] Finally, CLM shows a large-scale bend and twist motion of the two domains.

ADK is a monomeric enzyme that catalyzes the transfer of a phosphoryl group from ATP to AMP. The structure of ADK contains a main domain (CORE), an ATP-binding domain (LID), and a NMP-binding domain (NMPbind)[252]. AST is a homodimeric enzyme that catalyzes a reversible transamination reaction: L-aspartate + 2-oxoglutarate $\rightleftharpoons$ oxaloacetate + L-glutamate.[253] CTS is also a homodimeric enzyme and catalyzes the reaction: acetyl-coenzyme A + oxaloacetate $\rightleftharpoons$ citrate + coenzyme. A study[213] that includes these three domain proteins identifies specific interactions that drive a ligand-induced domain closure. Furthermore, it supports the assumption that each enzyme has a dedicated binding domain, to which the ligand binds first, and a closing domain. CLM is a ubiquitous intracellular protein that plays a critical role in coupling transient Ca^{2+} influx. It consists of two small globular domains separated by a flexible linker, with no stable, direct contacts between the two domains.[20] LAO is a part of bacterial periplasmic transport systems (permeases), which transport a wide variety of substrates. The LAO structure[214] is bi-lobate, and the two lobes (lobes I and II) are held together by two connecting segments.

The NMSim approach is applied to the open conformation of the proteins in the dataset using three different types of simulations: freely-evolving, ROG-guided and target-directed (see section 4.3.2). The conformations obtained over the trajectories are compared with the close conformation in terms of backbone RMSD. The C_{α} RMS fluctuations over the freely-evolving trajectories are compared with the fluctuations derived from respective open and close structures. Adenylate kinase is selected for a detailed analysis, and essential dynamics calculations are applied using eleven experimental structures. Furthermore, the extent to which ROG-guided NMSim can lead to a ligand-bound conformation is discussed in detail.

Comparison of essential dynamics between experimental and NMSim structures

In order to compare the essential dynamics (ED) between the experimental and NMSim structures, ED calculations were performed using eleven crystal structures and NMSim generated structures from ten freely-evolving trajectories, each one starting from the open structure of Adenylate kinase (ADK). ADK is a well studied protein in terms of catalytic mechanism and conformational flexibility and has been used as a test case in different theoretical studies.[87,89,152] Different X-ray crystal structures have been reported[14-16,252] for different conformations of the protein. The eleven crystal structures mainly lie in three groups: structures near the open conformation (4ake_A and 4ake_B; in PDB-code_chain format), intermediate structures in between the open and close conformations (1dvr_B and 1dvr_A; here the LID domain is completely closed and the NMPbind domain is still open), and structures near the close conformation (1e4y_B, 1e4y_A, 1e4v_A, 2eck_A, 1ank_A, 2eck_B and 1ake_A).

The ED calculations were performed on the experimental structures, and the NMSim generated structures were projected onto the plane described by the first two ED modes with highest Eigenvalues (Figure 5.10-a). The projections of the NMSim structures reach very close to both, the intermediate structures (e.g., PDB code 1dvr) and the close structures (e.g., PDB code 1ake). In general, the spread of the projected NMSim structures is broader along the ED mode 1, which in fact represents the movement of the LID domain. This movement has been shown to be a large-scale movement (Figure 5.13-a) and is an important mechanism for ATP binding.[252] It is important to note, however, that the NMSim projected structures show closing as well as further opening of the LID domain, as indicated by projected structures on the right side of 4ake along ED mode 1 in Figure 5.10-a. Furthermore, one NMSim trajectory out of ten shows a closing of the LID domain to an extent seen in the close structure

(PDB code 1ake) along ED mode 1 (dotted line above 1e4y and 1ake in Figure 5.10-a). This suggests that the LID domain is mainly driven by the intrinsic dynamics as argued previously.[89]

Conversely, the ED calculations were also performed on the NMSim generated structures, and the experimental structures were projected onto the plane described by the first two ED modes (See Figure 5.10-b). Here, again, the different close structures were found to be very near to the NMSim structures, whereas an intermediate structure (PDB code 1dvr) was found within one of the clusters of NMSim generated structures on the plane. This shows that the two sets of structures overlap in their essential dynamics. During different trajectories, both the opening and the closing of the LID domain can be seen from the spread of NMSim projected structures along the ED mode 1. However, the overall triangular shape of the NMSim projected structures onto the plane suggests that the ED mode 2 is mostly active upon the LID domain closure.

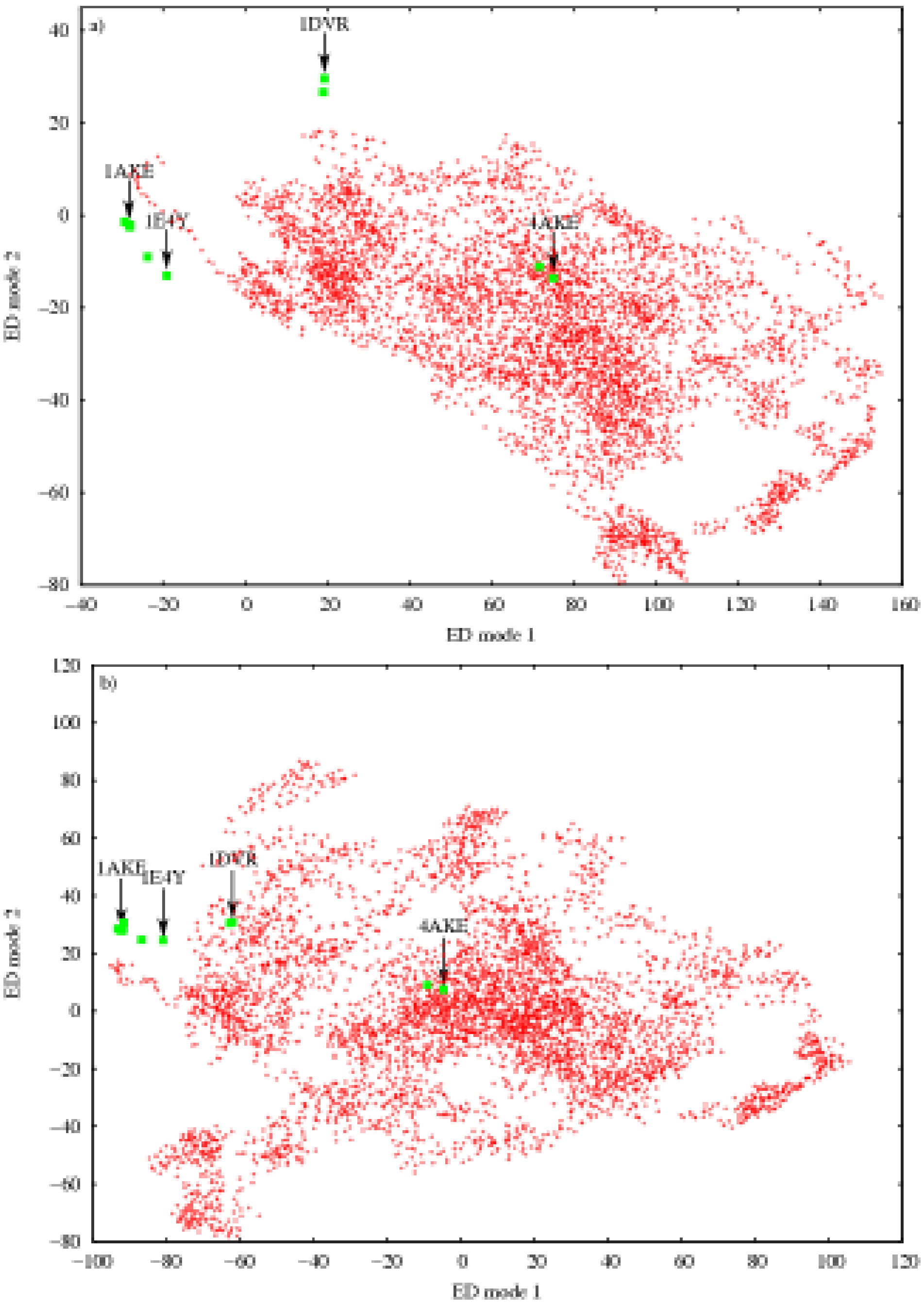


Figure 5.10: The projections of NMSim generated structures (red) using ten freely-evolving trajectories and eleven different experimental structures (green) of adenylate kinase onto the plane described by the first two ED modes derived from eleven different experimental structures (in a) and from NMSim generated structures (in b) are shown. It is shown that the two sets of structures overlap in their essential dynamics.

Intrinsic fluctuations and conformational changes

Intrinsic fluctuations of a protein near its equilibrium state in the open conformation correlate with the conformational change of that protein upon complex formation.[34,254] Theoretically, these fluctuations can be derived from ENM or GNM modes and have been reported to correlate well in different studies.[34,255] In order to verify if this argument holds for NMSim generated structures, C_α RMS fluctuations derived from the freely-evolving NMSim trajectories are compared with the fluctuations derived from their respective open and close conformations (Figure 5.11). Despite considering experimental fluctuations from two extreme conformations in the open and the close forms, good correlations with the fluctuations derived from NMSim generated structures were found (Table 5.7) in 4 out of 5 cases in the domain dataset. This supports the argument, mentioned above, that especially global conformational changes upon complex formation correlate well with the intrinsic motions of proteins in an open form. Furthermore, it shows that the NMSim approach effectively captures the information available in low-frequency normal modes and translates it into structural information in terms of different conformations without disturbing the underlying fluctuation pattern.

Good correlation coefficients above 0.7 (Table 5.7) between the RMS fluctuations derived from NMSim generated structures and the two experimental structures are observed for all cases except CLM. The highest correlation coefficient of 0.92 was observed in ADK between the two fluctuations plots. It is interesting to see that, in contrast to NMSim, the relative fluctuations in the mode best overlapping with the conformational change, as reported previously[68], underestimates the relative motions in the NMPbind domain. This could be explained by the finding[84,89] that the LID domain closure precedes the NMPbind domain movement and, therefore, can not be captured by a single mode in the open conformation. LAO, another protein having hinge bending motion, shows good agreement between the two fluctuations patterns, however, with high fluctuations in some regions as compared to the observed fluctuations between the open and the close structures.

Mobile regions are well recognized in NMSim. For example, in the CTS case, a sheer motion,[251] high fluctuations in small domain comprising residues 284-327 and 338-

378 of chain A and 714-757 and 768-808 of chain B in Figure 5.11 correlate well with the regions of conformational changes upon ligand binding. In the case of CLM, a low correlation coefficient of 0.32 between the two fluctuations is observed. This can be attributed to the local rearrangements (see also open and close structures in Figure 5.13-c) within the two domains of the open structure, which result upon Ca^{2+}-binding.[20] Due to these conformational rearrangements, in both domains of CLM where all four Ca^{2+}-binding sites are occupied, a large hydrophobic surface has been found to become exposed to the solvent.[256] These local rearrangements are not well described in the low-frequency modes,[71] especially in a protein where the intrinsic motion is dominated by the large-scale movement of domains, as in CLM.

In AST, RMS fluctuations derived from NMSim structures are higher than the fluctuations derived from the open and the close structures (Figure 5.11-b). However, a good correlation coefficient of 0.71 between the two is observed. Contrarily, good agreements in the magnitudes of the fluctuations are observed for large-scale motions, for example in ADK (Figure 5.11-a) and CLM (Figure 5.11-c). In general, high fluctuations observed in some proteins, are an indication that the underlying constraint network might be under-constrained in some cases and, therefore, results in a higher mobility of the systems. A similar constraint-based method tCONCOORD[63] also reports high fluctuations as compared to NMR derived structures. In general, therefore, there is a need for improving the underlying constraint network for these methods.

It is important to note here that the reported fluctuations are derived from NMSim generated structures which incorporate low-frequency modes with no prior experimental information. Previously, studies[34,67,68,70,142] have shown good correlation between the fluctuations of the biologically relevant normal mode (which is selected using close structure information) and the observed conformational changes. So, it is almost always true that, in general, the biologically relevant mode is one or several of the low-frequency modes, yet, it is hard to identify that mode without any additional experimental information.[136] For example, LAO and other proteins of the same family have been reported to invoke a single bending low-frequency mode,[69,255] however this information is reported only with the help of experimental structure in its closed form. Recently,[257] it has also been argued that a single mode can be deceiving if used for the purpose of identifying correlated motions in biomolecules. Considering these, it is interesting to see the good correlation values observed in NMSim, which is a normal mode-based method and incorporates a range of low-frequency modes.

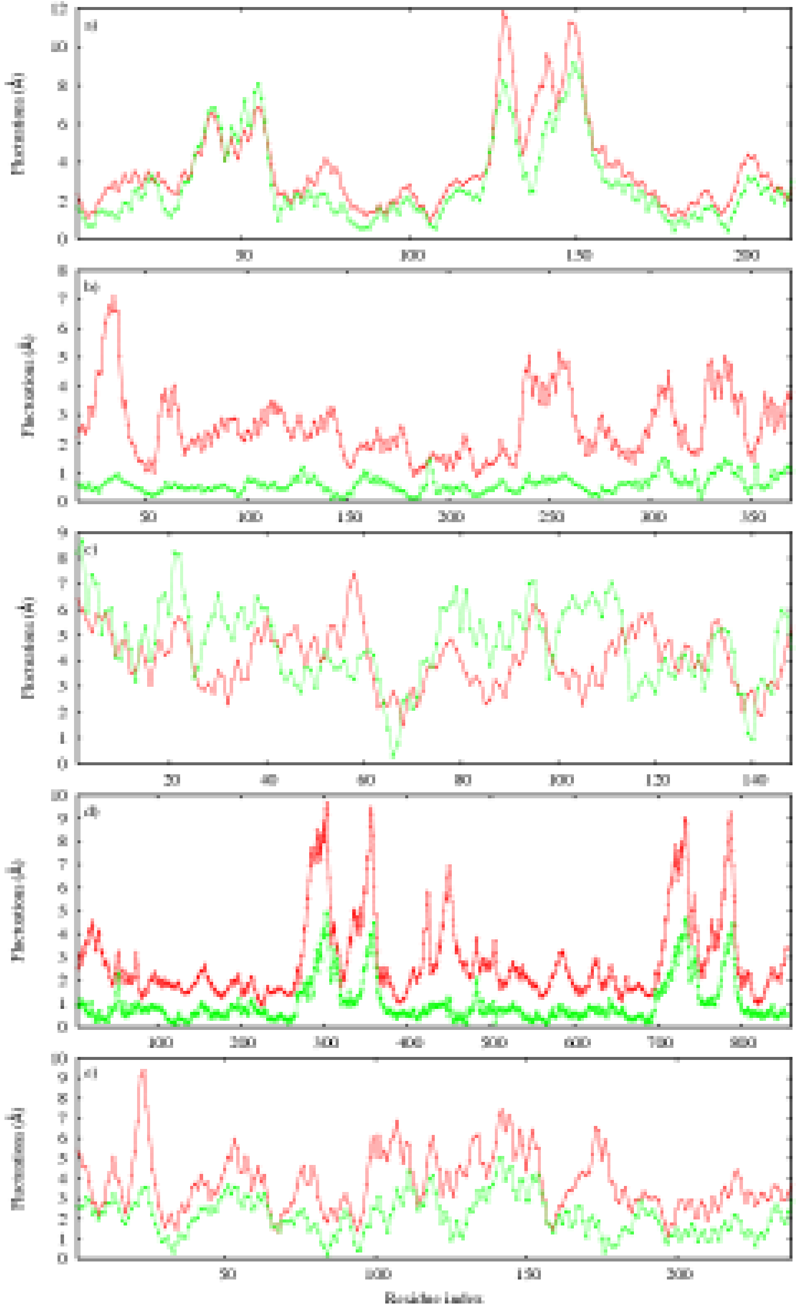


Figure 5.11: The C_α fluctuations of different domain moving proteins: Adenylate kinase (a), Aspartate aminotransferase (b), Calmodulin (c), Citrate synthase (d), LAO binding protein (e) for freely-evolving NMSim trajectory (red) are shown. The C_α fluctuations (green) derived from respective open and close structures are also shown.

Ligand bound conformations generated from an unbound one

In order to observe how close the "close structure" is reached during the NMSim trajectories, freely-evolving NMSim trajectories started from open conformations of different proteins were analyzed in terms of backbone RMSD with their respective close conformation. Figure 5.12 shows the RMSD plots for all 10 different trajectories of every protein in the domain dataset. Each trajectory contains 500 structures and is placed one after the other in the RMSD plot. In general, each trajectory follows a different path and shows different patterns of RMSD distance with the close structures. Hinge bending motions like in ADK and LAO show either an increase or decrease or both in RMSD with the respective close structures in different trajectories, which is an indication of a freely opening and closing of domains. For example in ADK, the first trajectory (structures 1-500) fluctuates around the open conformation, the second trajectory (structures 1-500) shows further opening of domains, the third trajectory (structures 1001-1500) shows a closing of the domains and remains near the close structure, whereas the eighth trajectory (structures 3500-4000) shows an initial opening and then closing of the domains. Sheer motions like in AST and CTS show a more frequent increase in RMSD from their respective close structures. However, interestingly, trajectories do get closer to the respective close structure at the initial stages. It should be noted that, in addition to sheer motions, AST and CTS conformational changes are relatively localized in small domains (see Table 4.1). It has been reported previously[136,258] that for systems involving localized transitions, as in p21^{ras}, normal modes are better suited for initial stages of movements only.

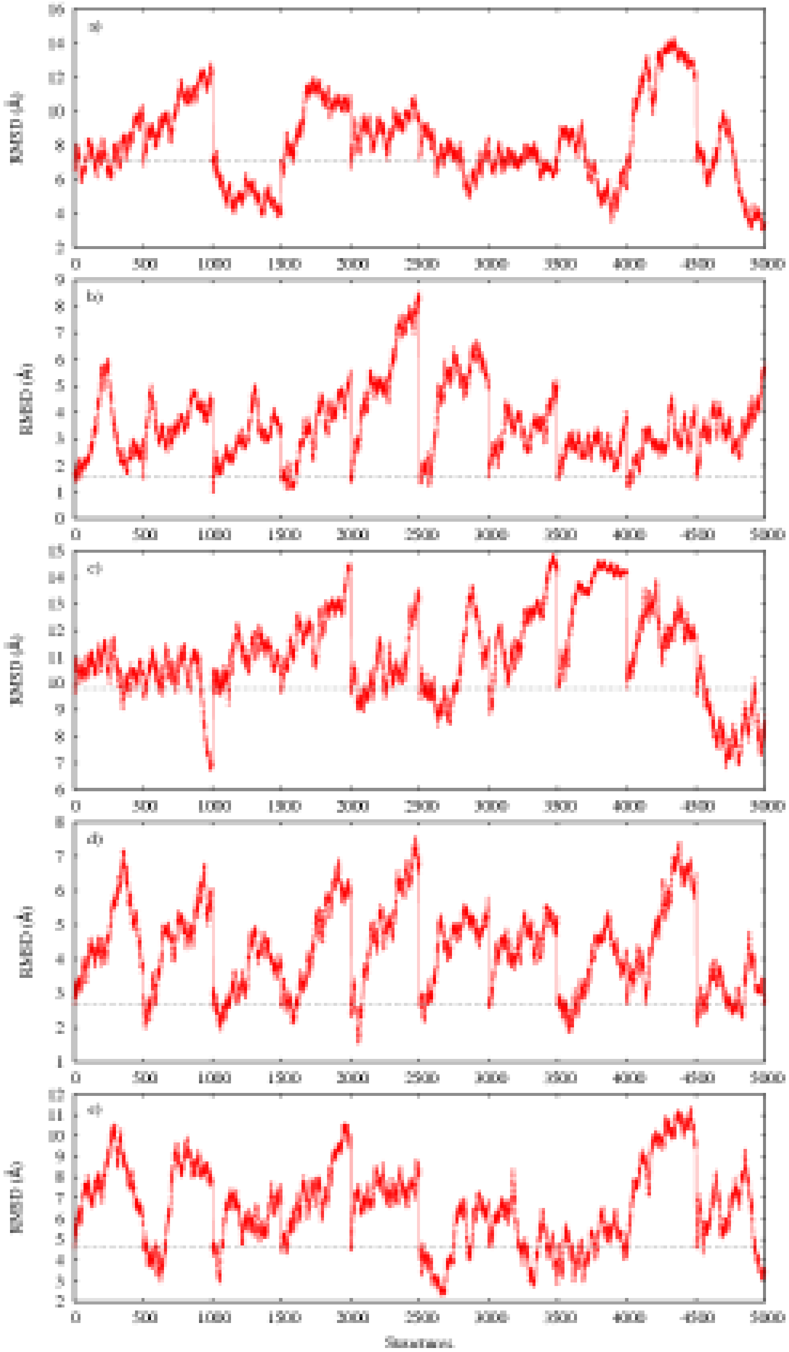


Figure 5.12: The backbone RMSD of the ligand bound (close) structure with the 10 freely-evolving NMSim trajectories (500 structures per trajectory placed in sequence on the x-axis) started from the unbound (open) structure of Adenylate kinase (a), Aspartate aminotransferase (b), Calmodulin (c), Citrate synthase (d) and LAO binding protein (e) are shown. The backbone RMSD between the open and the close structures for each protein (in a-e) is shown as a dotted straight line.

The RMSD between the close structure and the best NMSim generated structure, i.e., the one nearest to the close structure, for each protein is reported in Table 5.7. Considering RMSD between open and close structures, a considerable decrease in RMSD is observed in all cases of the domain dataset. A structure similar to the ADK close structure is achieved with RMSD ~3 Å in NMSim, which is slightly lower than the recently reported[63] RMSD of ~3.3 Å for tCONCOORD for the same structures. In target-directed trajectory, close structure is reached with RMSD ~1 Å using 50 low-frequency modes, however, higher modes would be required to get even closer to the target structures.[230] A similar study,[250] using normal modes but in combination with Monte Carlo simulation for ADK, reports that an RMSD of 2.27 Å is achieved with the close structure using 10 low-frequency modes.

The close structure in LAO is achieved with RMSD as low as ~2.3 Å and ~0.6 Å, respectively, with and without close structure information starting from the open structure, which is ~4.7 Å away from the close structure. This supports the argument in a recent study[37] suggesting a conformation selection mechanism for glutamine-binding protein, which is also a periplasmic binding protein. Proteins having sheer motions, as discussed above, do show initial movements towards the close structure in NMSim trajectories. Considering the large-scale conformational change observed in CLM, the NMSim trajectory does not reach near to the close structure, although it does show a ~3 Å movement towards the close structure. Even a target-directed NMSim trajectory can only reach ~3 Å near to the close structure using the first 50 modes in the CLM case. As discussed above this is due to the local rearrangements within the two domains of the open structure, which results from Ca^{2+}-binding,[20] which are not well described by the low-frequency modes.[230]

Table 5.7: The correlation coefficients and the lowest RMSD achieved by the different types of NMSim simulations. .

Proteins	**RMSD** [a)]				**Correlation** [d)]
	Open [b)]	**Freely-evolving** [c)]	**ROG-guided** [c)]	**Target-directed** [c)]	
Domain:					
Adenylate kinase	7.155	3.059	2.363	0.929	0.919
Aspartate aminotransferase	1.551	0.979	1.214	0.599	0.709
Calmodulin	9.800	6.708	5.319	2.955	0.319
Citrate synthase	2.701	1.551	1.373	0.913	0.860
LAO binding protein	4.675	2.313	1.750	0.593	0.705
Loop:					
Tyrosine phosphatase	3.176	1.862	1.581	0.954	0.427
Triosephosphate isomerase	4.504	2.011	2.236	0.902	0.393
CAMP-dependent protein kinase	1.676	1.141	0.666	0.790	0.279

a) The backbone RMSD with respect to close structures. For loop proteins backbone RMSD only for the loop region is calculated after aligning the rest of the protein. b) The RMSD between open and close structures. c) The lowest RMSD achieved with the respective close structures by different types of simulations, i.e., freely-evolving (see also Figure 5.12), ROG-guided, and target-directed. d) The Correlation coefficient between the two C_α fluctuations (plot shown in Figure 5.11) obtained from conformations generated from the freely-evolving trajectories and obtained from open and close structures.

ROG-guided trajectory leads to ligand bound conformation

Results from freely-evolving and target-directed NMSim trajectories, as discussed above, describe the extent to which the close conformation can be reached without and with prior information of the close conformation using low-frequency modes. Normal modes in combination with different experimental data has been found useful in different applications.[81,145,148,259] It has been shown[260] that a small set of pairwise distance constraints of the end state is helpful in driving one structure into the other using low-frequency modes. However, in the case where experimental information is not known, NMSim can provide an alternative. This is achieved in ROG-guided

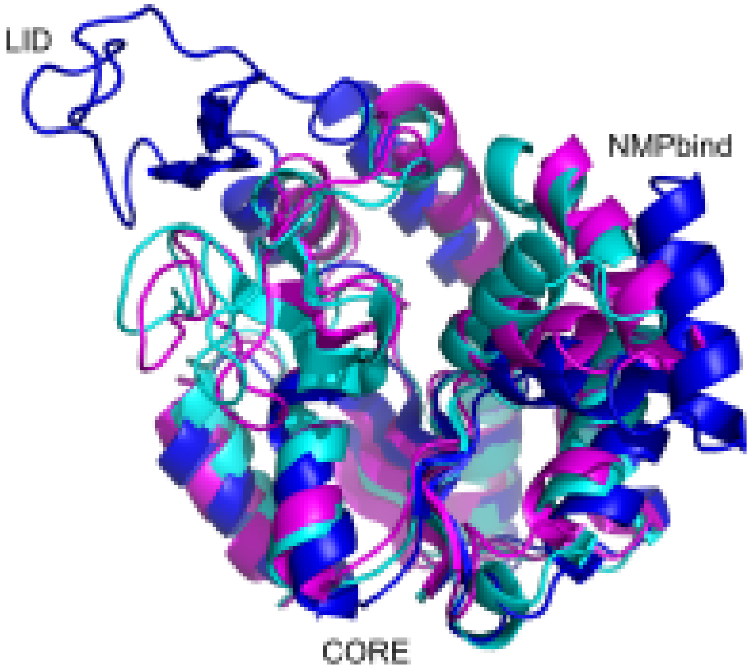
NMSim, which assumes that the ligand binding would result in domain or loop closures. Using normal mode combinations which decrease the radius of gyration (R_g) would then guide to the close conformation. It is important to note here that the conformations are still generated by random linear combinations of low frequency normal modes and, therefore, the pathway still goes though low energy space.

The comparison between the ten freely-evolving NMSim trajectories and the ROG-guided NMSim trajectory for the proteins in the domain dataset shows that the ROG-guided simulations reach nearer to the close structure in 4 out of 5 cases (Table 5.7). This improvement is more obvious for hinge bending motions than sheer motions; this is perhaps because the underlying assumption, that the ligand binding would result in domain closures, is more valid in hinge bending motions. Here, it should be noted that this improvement is achieved with around four times lower computational cost; In contrast to the ten freely-evolving trajectories, a single ROG-guided trajectory (generating 3 structures each step) was run for each protein, because it was found in initial test that different ROG-guided trajectories do not differ significantly.

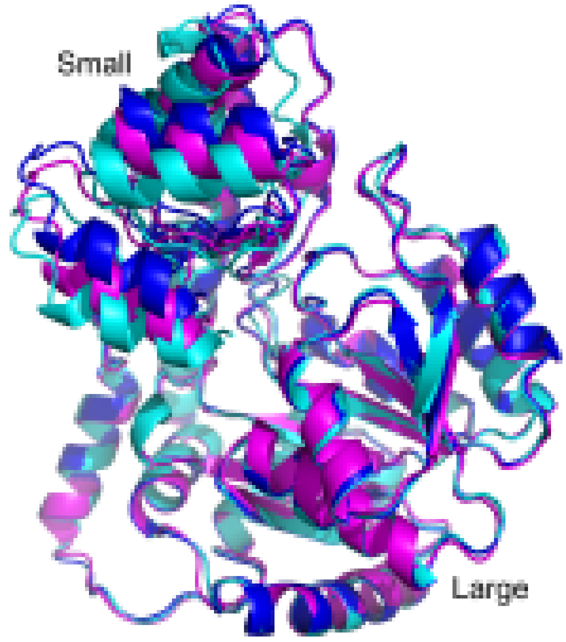
Coarse-grained normal modes usually very well describe functionally important conformational changes,[71,231] however, which mode or combination of modes are involved in a conformational change is not know in advance. This has triggered discussions how to identify functionally relevant mode.[177,230] In this view, the radius of gyration (R_g) can be used as a criterion for selecting normal modes in cases where no experimental information is known.

Figure 5.13 illustrates the extent to which ROG-guided NMSim was successful in reaching the close conformation. The nearest generated structure to the close is shown along with the respective open and close conformations for every protein in the domain dataset. In the ADK case, it is interesting to see that the large-scale conformational change in the LID domain is well reached by ROG-guided NMSim as compared to the close structure with no prior information of the close structure. However NMPbind domain, despite considerable movement, only reaches half-way towards the close conformation. Here, it is important to note that the closing of NMPbind domain has been suggested through ligand-induced mechanism.[89]

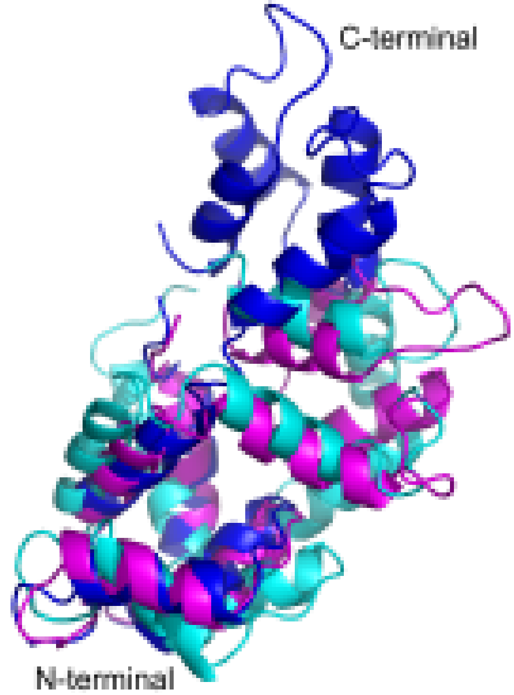
Therefore, probably the full closure of NMPbind domain would only be possible in the presence of a ligand. LAO in Figure 5.13-e again shows a large hinge-bending motion towards the close conformation (~ 3 Å from the starting structure), and the close conformation is almost reached with RMSD 1.7 Å (see Table 5.7) in ROG-guided NMSim. CLM in Figure 5.13-c shows a large scale hinge-bending motion, which can be seen in NMSim generated structure too, however, the local rearrangements within the two domains resulting from Ca^{2+}-binding[20] is not reproduced by NMSim. AST in Figure 5.13-b shows that the sheer type of conformational change is not achieved completely, however, a small movement of 0.3 Å towards the close conformation can be seen. It is interesting to see in Figure 5.13-d that, despite sheer type of motion and localized in the small domain in the case of CTS, NMSim generated structure very well fit to the close structure (with RMSD 1.3 Å). This shows that the underlying assumption in ROG-guided NMSim (i.e., proteins contract upon ligand binding) is justified not only in hinge-bending motions but also in sheer motions. The transition towards the close structure can then be captured using the low-frequency modes without close structure information.



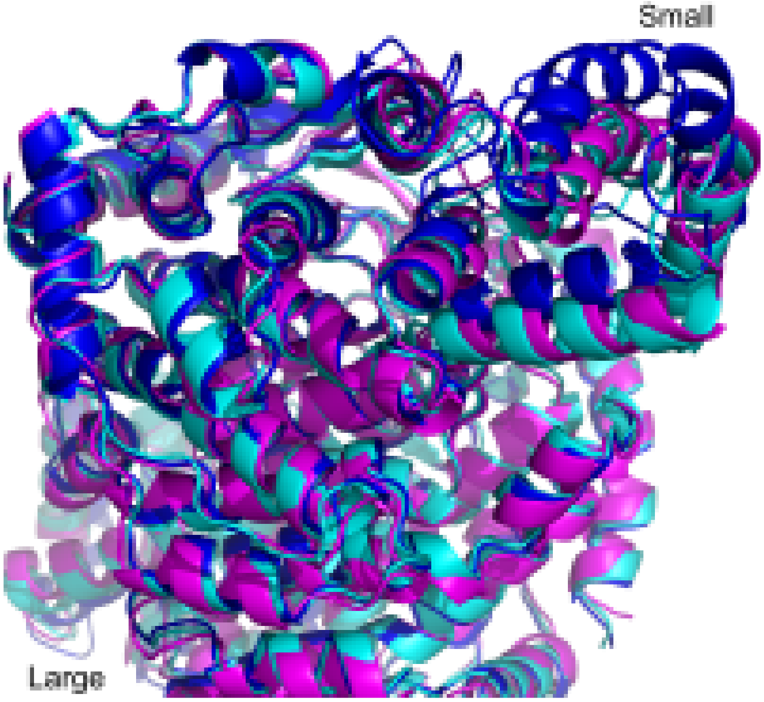
a)



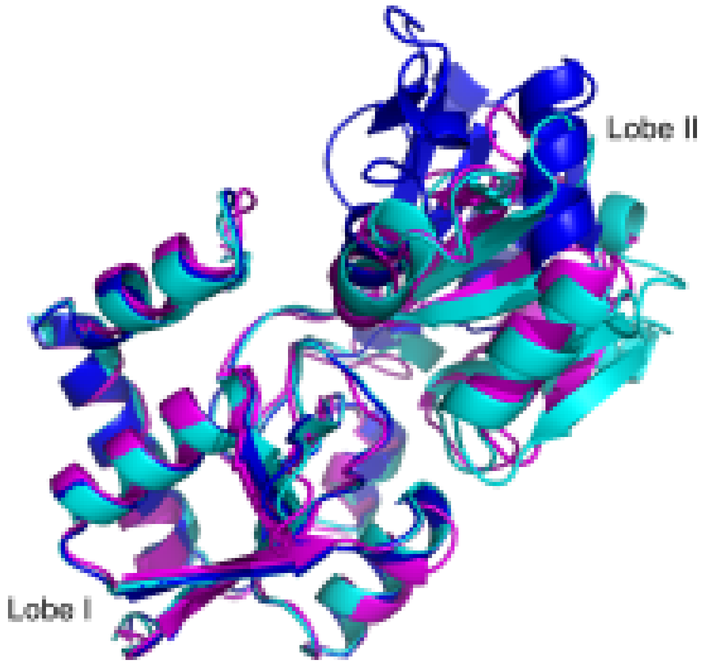
b)



c)



d)



e)

Figure 5.13: The experimental structures i.e., open (blue), close (cyan), and NMSim generated structure nearest to the close (magenta) using ROG-guided trajectories of different domain moving proteins are shown: Adenylate kinase (panel a), Aspartate aminotransferase (panel b), Calmodulin (panel c), Citrate synthase (panel d) and LAO binding protein (panel e).

In order to analyze the effectiveness of using normal mode directions for guiding movements in ROG-guided NMSim, a ROG-guided simulation was also performed using random vectors instead of normal modes. It was found that a random vector-based ROG-guided NMSim simulation hardly moves towards the close structure. For example, in the case of ADK, a random vector-based ROG-guided trajectory moves only ˜0.57 Å towards the close structure in 500 NMSim cycles and reduces R_g by 0.74 Å (the starting structure R_g is 19.46 Å). In contrast, a normal mode-based ROG-guided trajectory moves ~5 Å towards the close structure (Table 5.7) in only 200 NMSim cycles and reduces R_g by 3 Å. This shows that, in the results described above in Table 5.7, the movements towards the close structure in ROG-guided trajectories are due to the collective and functionally relevant modes. And this also shows that, radius of gyration can not be used as a guide for bound conformations in diffusive motion of atoms.

5.3.2 Functionally important loop motions

Three functionally important loop motions used in this study are listed in the loop dataset in Table 4.1. Tyrosine phosphatases (TYP) and kinases coregulate the critical levels of phophorylation necessary for interacellular signaling, cell growth, and differentiation.[261] A ligand-induced conformational change has been observed in TYP, which moves Asp356 on the β7-α4 loop into the active site, where it can function as a general acid.[18] Triosephosphate isomerase (TIM) is an important enzyme in glycolysis, catalyzing the interconversion between dihydroxyacetone phosphate and D-glyceraldehyde-3-phosphate. Low-frequency modes have been shown to be active in the important loop motions in TYP and TIM.[68,212] The catalytic subunit of cAMP-dependent protein kinase (CAPK) catalyzes the phosphorylation of proteins that have several arginines preceding the site of phosphotransfer. A study[75] has suggested that the mid-scale loop rearrangements, like those found in protein kinase, do not involve the first few lowest frequency modes. However, still modes can be selected from the low-frequency range using a relevance measure to describe the loop flexibility in CAPK. This means, normal modes do provide the directions for loop motions as well, however, selecting a mode or combination of modes to predict the conformational changes for loop motions might be complicated than domain motions.

In order to explore the extent to which experimentally observed loop conformational changes can be simulated in NMSim, NMSim was applied to the open conformation of the three proteins in the loop dataset (Table 4.1). For each protein, a single trajectory for each of the three different types of simulation, freely-evolving, ROG-guided and target-directed, is computed (section 4.3.2).

Ligand bound loop conformation computed from unbound

In order to analyze the movements of the selected loop region (Table 5.7), the backbone RMSD of the loop region along the trajectory (after superimposing the rest of the protein) with respect to the close loop conformation is plotted in Figure 5.14 for each protein. In case of TYP and TIM, freely-evolving trajectories show opening and closing movements of the β7-α4 loop and loop 6, respectively. TYP and TIM freely-evolving trajectories reach RMSD 1.8 Å and 2.0 Å with respect to the close loop conformation, respectively, starting from 3.2 Å and 4.5 Å, respectively (Table 5.7). These considerable motions show that even loops have intrinsic motions towards the close conformation, and ligands further stabilize these conformations in the close form. It has been shown, both experimentally[262] and theoretically,[263] that the loop 6 closure in TIM is an intrinsic motion of the protein and not ligand gated, as it can be seen in different unbound crystal structures. In contrast, a ligand induced motion has been proposed[18] for the β7-α4 loop in TYP. However, the considerable movements of the loop in freely-evolving NMSim suggest that there is an intrinsic motion in the β7-α4 loop and probably a ligand influences the receptor conformation at the later stages of the ligand binding. Previously,[68] the lowest-frequency mode of an unbound TYP has also been found to predict the β7-α4 loop movement in TYP, which is also an indication of some intrinsic motions in this loop region. The glycine-rich loop in CAPK does get 0.5 Å (Table 5.7) nearer to close loop conformation in the first 100 structures in the freely-evolving trajectory, however, the loop moves away from the close conformation afterwards in the trajectory (Figure 5.14-c). A study[75] has suggested that, mid-scale loop rearrangements like glycine-rich loop in CAPK, do not involve the first few lowest-frequency modes. Therefore, criteria for mode selection in CAPK have been proposed.[75] It should be noted that, for using the proposed criteria for mode selection, the moving loop region should be known in advance. In contrast, the NMSim approach uses all 50 low-frequency modes and no information of a loop region is provided.

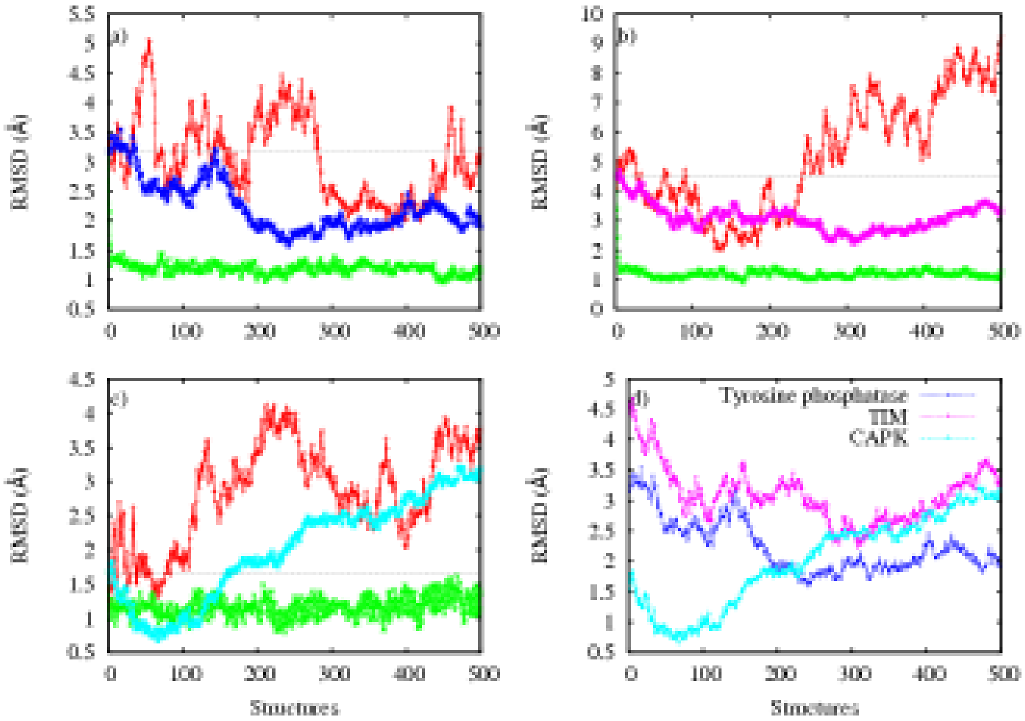


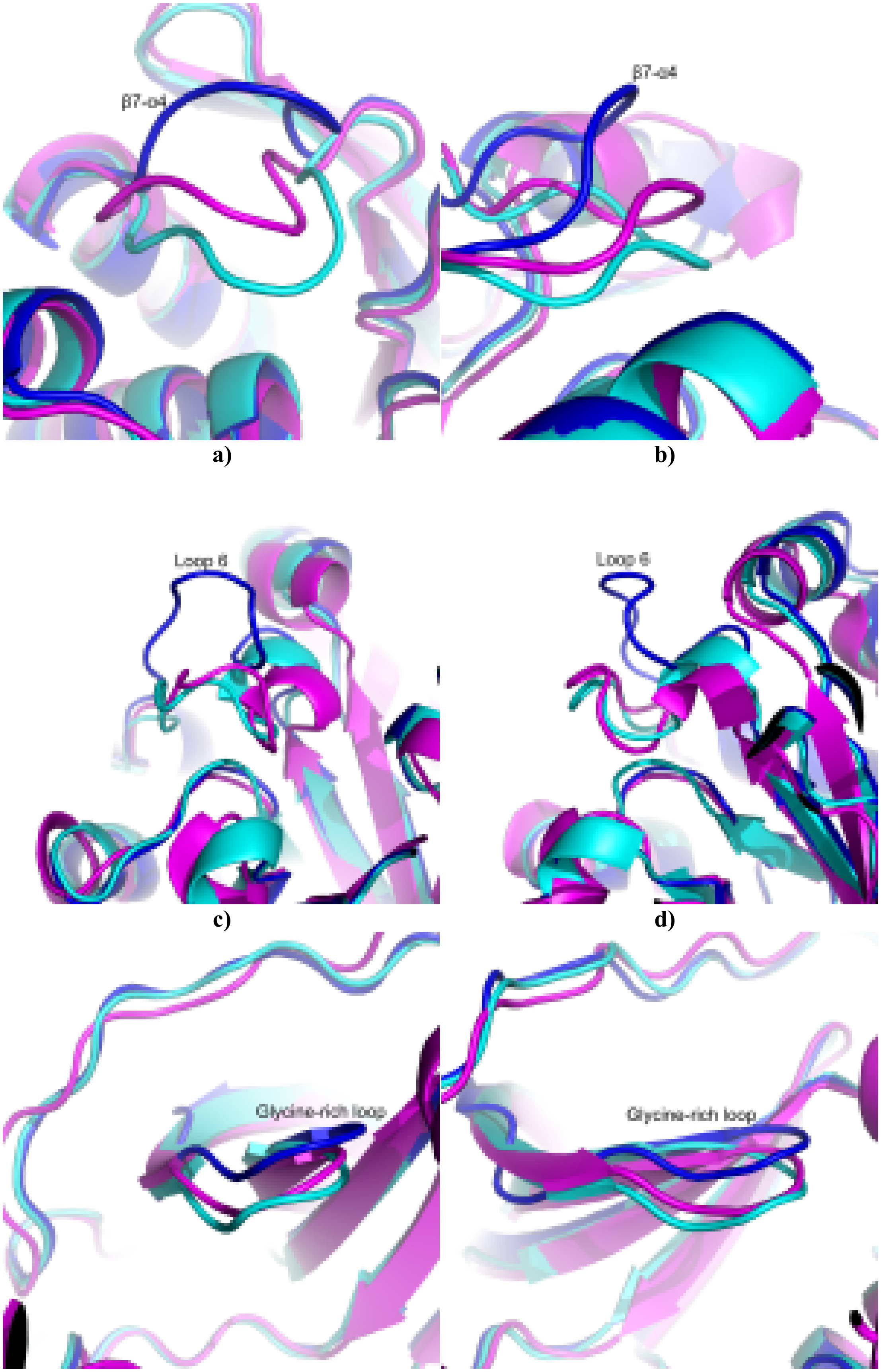
Figure 5.14: The loop region backbone RMSD plots between the ligand bound (close) structure and the NMSim generated structures of Tyrosine phosphatase (a), Triosephosphate isomerase (b), and cAMP-dependent protein kinase (c) are shown. Each plot in a-c shows three different NMSim trajectories, freely-evolving (in red), target-directed (in green), and ROG-guided (in blue, magenta, and cyan). The loop region backbone RMSD between open and close structures for each protein (in a-c) is shown as straight line. For clarity, RMSD plots for ROG-guided trajectories of different proteins are shown in one graph (d).

In ROG-guided trajectories, it is interesting to see that the lowering of R_g in low-energy space does guide the trajectory towards the experimentally observed loop closure in all three cases (Figure 5.14-d) and TYP, TIM, and CAPK ligand-bound loop conformations are reached with RMSD 1.58 Å, 2.23 Å and 0.66 Å, respectively (Table 5.7). This shows that low-frequency modes can be used to predict not only domain closures but also loop closures upon ligand binding with no prior information of close conformation.

Furthermore, in two out of three cases (i.e., in TYP and TIM), the loops in the trajectories fluctuate around to the best achieved loop conformation i.e., nearest to the close structure. This means, further reducing R_g of structures in normal mode space does not influence the loop region. However this is not always the case, as observed in CAPK, where further reducing R_g of structures in normal mode space moves the loop away from its best achieved conformation. This can be expected due to additional compactness of the protein and also probably due to the ligand absence in the environment, which provides the room to over-stretch the loop closure. It is interesting to see that, improvements were observed in two out of three cases with ROG-guided trajectories as compared to freely-evolving trajectories in terms of RMSD achieved nearest to the close conformation (Table 5.7).

The best achieved conformation in ROG-guided trajectory in terms of nearest loop RMSD with the close conformation are shown in Figure 5.15, along with respective open and close conformations. In all three cases, considerable movements towards close conformation can be seen (see Table 5.7). However, further rearrangements are needed to attain completely bound conformation. Probably, a ligand influence is inevitable in those cases. It has been argued[34,255] that conformational selection and induced fit are not two mutually exclusive processes but both can play their part, and the extent to which each mechanism contributes can vary in different proteins.

In the target-directed trajectories, in all three cases loop conformations were obtained that come close to the “close structure” to less than 0.9 Å (see Table 5.7). These values can further be decreased if higher frequency modes are used and if only loop region is considered during mode selection as done previously.[75] As can be seen in Figure 5.14, trajectories immediately move towards close conformations if directions are provided; this again shows that these loop motions are intrinsic and can be captured by low-frequency modes. In the case of CAPK, the ROG-guided trajectory outperforms target-directed trajectory in terms of achieving bound loop conformation. However, values in Table 5.7 shows that this difference is only 0.13 Å. This might happen because; RMSD values are only for loop regions, whereas target-directed NMSim by default uses all C_α atoms in the close conformation as a target.



a) b)

c) d)

e) f)

Figure 5.15: The experimental structures, open (blue) and close (cyan), and the NMSim generated structure closest to the close structure using ROG-guided simulation of different loop moving proteins are shown in front and side views: Tyrosine phosphatase (in panel a and b), Triosephosphate isomerase (in panel c and d), and cAMP-dependent protein kinase (in panel e and f).

Intrinsic fluctuations and conformational changes

In order to see how well the intrinsic fluctuations correlate with the conformational changes observed in the loop dataset, C_α RMS fluctuations derived from NMSim generated structures are compared with the fluctuations derived from their respective open and close conformations and the fluctuations derived from respective B-factor values in open form (see Figure 5.16). In contrast to the domain dataset, low correlation coefficients between the two fluctuations were found (Table 5.7) for the proteins in loop dataset. However, in the case of TYP and TIM high fluctuations can be seen in β7-α4 loop and loop 6, respectively, in NMSim generated structures that match perfectly with the observed conformational changes upon ligand binding (in Figure 5.16-a,b). In lines with the "conformational selection" model, Bahar and co-workers[34] have shown that structural changes involved in protein binding correlate with intrinsic motions of proteins in the open form, and loops possess an intrinsic tendency to move towards the bound conformations. High fluctuations in some parts of the proteins are also seen, for example, residues 335-343 in TYP and residues 65-78 in TIM (Figure 5.16-a,b), which do not correlated with the observed conformational changes. In the case of TYP, these fluctuations do correlate with the B-factor values with a correlation coefficient of 0.64. This shows that, these high fluctuating regions in TYP are probably the regions that have an intrinsic ability to move and the two crystal structures of TYP do not capture these movements. In the case of TIM, B-factor values do not correlate with these high fluctuating regions observed in NMSim. This might be an indication that the underlying constraint network might be under-constrained in some regions and, therefore, results in a higher mobility of those regions.

In contrast to TIM and TYP, the glycine-rich loop in CAPK does not show high fluctuations in NMSim structures. Although this glycine-rich loop has been previously reported[264,265] to be mobile, fluctuations in NMSim are not prominent in this loop region. This is not surprising in view of a recent study[75] that suggested that, the mid-scale loop rearrangements, like glycine-rich loop in CAPK, do not involve the first few lowest-frequency modes. The NMSim approach emphasizes these low-frequency modes and, therefore, does not perform well if the motions of a loop are not guided by low-frequency modes, as found in CAPK case. The region containing F-to-G helix loop and G helix (residues 238-250) is found to be highly mobile in NMSim. This is probably the effect of removing inhibitor near this region for the simulation, as part of the default NMSim setting. In the following section, the usability of NMSim approach to generate conformational change pathways is discussed.

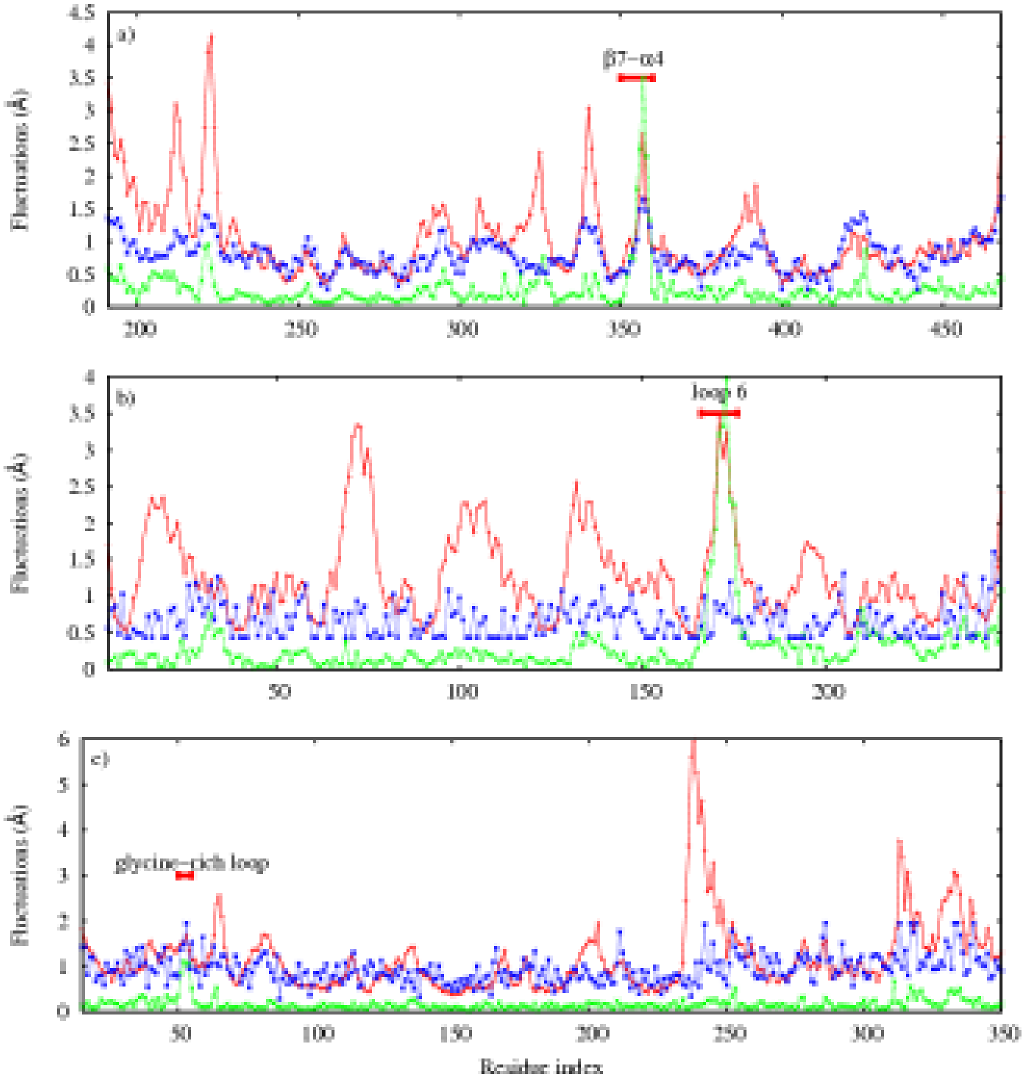


Figure 5.16: The C_α fluctuations of different loop moving proteins, Tyrosine phosphatase (a), Triosephosphate isomerase (b) and cAMP-dependent protein kinase (c) derived from freely-evolving NMSim trajectories (red) are shown. The C_α fluctuations derived from respective open and close structures (green) and derived from B-factor values in open PDB file (blue) are also shown. Residues in the functionally important loop region are marked by a red bar at the top of each plot.

5.4 NMSim and Conformational change pathways

The NMSim approach can be used to generate pathways of conformational change from an apo structure to a ligand-bound structure. In this section, the NMSim generated pathway for Adenylate kinase (ADK) is analyzed and compared with similar studies.[84 89]

5.4.1 Adenylate kinase: a test case

ADK is a monomeric enzyme that catalyzes the transfer of a phosphoryl group from ATP to AMP. The structure of ADK contains a main domain (CORE), an ATP-binding domain (LID), and a NMP-binding domain (NMPbind)[252]. Large conformational changes have been observed in the ADK structure upon ligand binding, where the LID and the NMPbind domains close with respect to the CORE domain. A study[254] has shown that the intrinsic motion of apo ADK occurs preferentially in the direction of the close conformation.

There have been significant efforts to develop theoretical frameworks for describing functional transitions in proteins to fully understand their mechanism. The ADK transition has been studied extensively using different theoretical methods.[87-89,250] A model for landscape hopping between elastic networks has been introduced[83,84] to estimate the barrier of the transition process and to identify regions where local unfolding occurs. A number of studies[87-89,250] have focused on generating intermediate structures and analyzing the pathway between the apo and the bound structures of ADK.

5.4.2 NMSim generated pathways using Close directed and ROG-guided simulations

The NMSim generated pathways (see section 4.4) from the open conformation to the close conformation of ADK using target-directed NMSim trajectory is shown in green in Figure 5.17. In order to analyze the order of the domains closure, the reaction coordinates described by Whitford *et al.*[89] were used (section 4.4). In general, the pathway generated by target-directed NMSim (Figure 5.17) shows that the LID domain closure precedes the NMPbind domain closure. This is in agreement with previous studies.[84,89] It has been suggested that this sequential domain closure is likely evolved to ensure that each conformational rearrangement contribute to the turnover of a substrate by preventing nonproductive substrate binding.[89] Furthermore, the transition seems to be energetically favorable to a large extent: Out of 50 normal modes used in target-directed trajectory, the first 5 lowest frequency modes are active throughout the transition, unless it nearly reaches the close conformation (state e). Interestingly, the initial closing of the LID domain (state a-b) is completely dominated by the first lowest frequency mode. The partial closing of the NMPbind domain (state b-c) mainly originates from the second and third lowest-frequency modes.

A ROG-guided simulation, in contrast to a target-directed simulation, is not biased towards any direction and assumes that the open to close transition would lead to a contraction of the protein, as usually observed for bound structures. The NMSim generated pathway using the ROG-guided trajectory is shown in red in Figure 5.17. Interestingly, this pathway again confirms that the LID domain closure precedes the NMPbind domain, even if no close conformation information is provided. The two pathways, i.e., target-directed and ROG-guided, remarkably resemble each other between state *a* to *e,* however, differ in the last stage. Furthermore, the LID domain in the ROG-guided trajectory closes more as compared to the target-directed trajectory (see Figure 5.17). This could be an effect of the absence of the ligand during the simulation, which provides the required space to contract the protein, whereas in the target-directed simulations this is avoided due to the directional biasing. Despite a small difference in the level of LID domain closure in both trajectories, the same level of NMPbind domain closure is observed in state *b* to *e*. Furthermore, the involvement of higher frequency modes from state *e* to *f* in target-directed trajectory suggests the higher influence of the ligand at this stage. Whitford *et al.*[89] have also suggested that the NMPbind domain closure is an example of a ligand-induced conformational change.

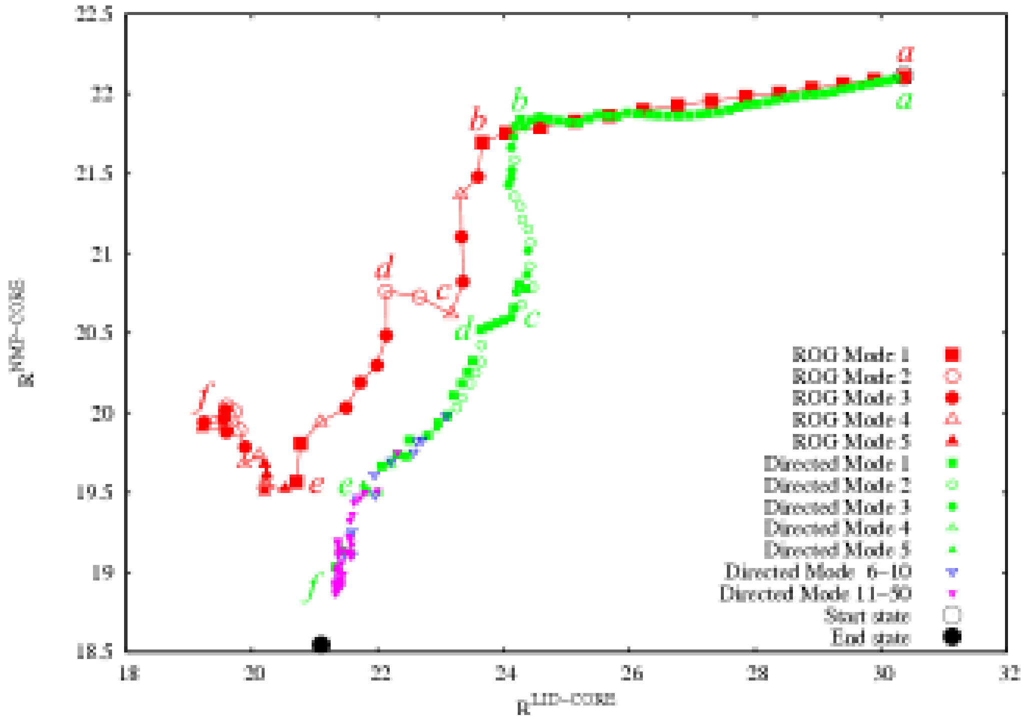


Figure 5.17: The NMSim generated pathway from the open Adenylate kinase conformation is shown for the target-directed trajectory using close conformation information (in green) and the ROG-guided trajectory (in red). On the x-axis is the distance between the LID domain and CORE domain centers of mass, $R^{LID-CORE}$*, and on the y-axis is the distance between the NMPbind domain and CORE domain centers of mass,* $R^{NMP-CORE}$*, over the trajectories. Each point corresponds to an intermediate structure. The different point types represent the modes used for that intermediate conformation generation. The unfilled and the filled black circles mark the starting (PDB code 4ake) and the target (PDB code 1ake) conformations respectively. For discussion, different states are marked from a-f, and higher frequency modes in target-directed trajectory are colored differently (blue and magenta).*

In order to further verify the NMSim pathway, the generated intermediate structures were compared with eleven different X-ray crystal structures of ADK in terms of C_α RMSD. The crystal structures used here can be divided into three groups: structures in the open conformation (4ake_A and 4ake_B: in PDB-code_chain format), structures in between the open and close conformations (1dvr_B and 1dvr_A; here the LID domain is completely closed and the NMPbind domain is still open) and structures near the close conformation (1e4y_B, 1e4y_A, 1e4v_A, 2eck_A, 1ank_A, 2eck_B and 1ake_A). Maragakis and Karplus[87] have identified different crystal structures that lie along the pathway from open to close conformation of ADK by selecting a crystal structure with the lowest RMSD to each intermediate structure along the generated pathway.

The RMSD plots for the target-directed trajectory are shown in Figure 5.18-a. Apart from the minor differences, the structures observed along the generated pathway are in agreement with the previously suggested[87] sequence of structures. Crystal structures 1dvr_A and 1dvr_B have the lowest RMSD < 3 Å in the middle of the transition. Crystal structures 4ake_A, 4ake_B, 1dvr_A, 1e4y_B, 1e4y_A, 1ank_A, 2eck_B and 1ake_A are found along the pathway from the open to the close, when the crystal structures are selected with lowest C_α RMSD along the generated pathway.[87]

The RMSD plots for the ROG-guided trajectory, shown in Figure 5.18-b, do agree with the reported sequence of crystal structures,[87] in the start and the middle of the transition. Crystal structures 4ake_A, 4ake_B, 1dvr_A, 1dvr_B, and 1ake_A are found along the pathway from the open to the close, when the crystal structures are selected with lowest C_α RMSD along the generated pathway. Interestingly, without the information of the close conformation, the crystal structures 1dvr_A and 1dvr_B, which lies in between the open to close transition, is found with a lowest RMSD of ~2.5 Å. However, the end transition is blurred, and does not show any preference for the different close crystal structures. It is interesting to note however that, the close structure is reached ~2.7 Å in the ROG-guided trajectory where no information of close structure is provided (using five low frequency modes, see section 4.4). In the target-directed trajectory the close structure is reached with an RMSD of 1 Å, using the first 50 low frequency modes. However, modes of higher frequency would be required to get even closer to the target structure.[230] A similar study,[250] using normal modes but in combination with Monte Carlo simulation for ADK, reports that the RMSD of 2.27 Å is achieved with the target structure using 10 low frequency modes.

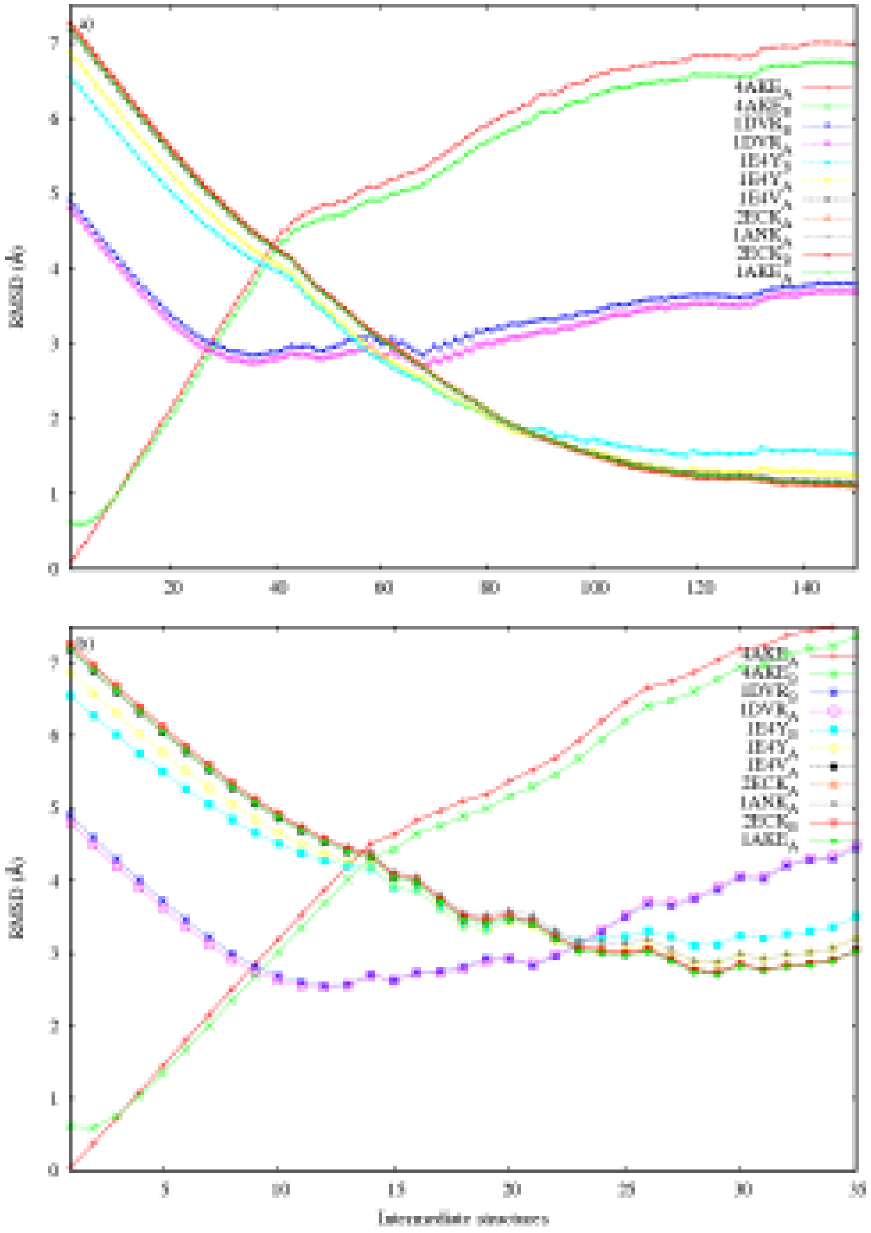


Figure 5.18: The C_α RMSD between intermediate structures, derived from target-directed simulation using close conformation information (in a) and ROG-guided trajectory (in b), and different experimental structures (shown in different colors and point types) of Adenylate kinase are plotted. PDB codes (here subscripts represent chain) for different experimental structures listed in the legend are sorted with the sequence proposed by Maragakis and Karplus,[87] that lie along the pathway from open to close conformational transition of adenylate kinase.

6 Summary

Specific functions of biological systems often require conformational transitions of macromolecules. Thus, being able to describe and predict conformational changes of biological macromolecules is not only important for understanding their impact on biological function, but will also have implications for the modelling of (macro)molecular complex formation and in structure-based drug design approaches. The “conformational selection model” provides the foundation for computational investigations of conformational fluctuations of the unbound protein state. These fluctuations may reveal conformational states adopted by the bound proteins.[33]

Different computational approaches targeting the modelling of protein flexibility and plasticity are promising in this context. Molecular dynamics (MD)[56-58] simulation is one of the most widely applied and accurate computational techniques currently being used. However, despite immense increase in computer power, MD simulations are computationally expensive and explore limited conformational space due to slow barrier crossing on the rugged energy landscape of macromolecules.[59,60] Hence, there have been efforts to develop alternative approaches that are computationally efficient in exploring conformational space. For example, a simple geometry-based approach CONCOORD generates conformations by satisfying distance constraints derived from a starting structure of a protein structure.[62,63] Another geometry-based approach FRODA generates conformations by diffusive motions of flexible regions and rigid clusters of proteins.[64] So far, these geometry-based approaches do not use any directional guidance for sampling the biologically relevant conformational space.

The aim of this work is to incorporate directional information in a geometry-based approach, in order to sample biologically relevant conformational space extensively. Interestingly, coarse-grained normal mode (CGNM) approaches, e.g., the elastic network model (ENM) and rigid cluster normal mode analysis (RCNMA), have emerged recently and provide directions of intrinsic motions in terms of harmonic modes (also called normal modes).[67,68] These normal modes can be viewed as possible deformations of proteins and can be sorted by their energetic costs of deformations. In my previous work[68] and in other studies[67,69-71] it has been shown that conformational changes upon ligand binding occur along a few low-energy modes of unbound proteins and can be efficiently calculated by CGNM approaches.

In order to explore the validity and the applicability of CGNM approaches, a large-scale comparison of essential dynamics (ED) modes from molecular dynamics (MD) simulations and normal modes from CGNM was performed over a dataset of 335 proteins. Despite high coarse-graining, low frequency normal modes from CGNM correlate very well with ED modes in terms of *directions* of motions (average maximal overlap is 0.65) and relative *amplitudes* of motions (average maximal overlap is 0.73). On average, the space spanned by the first quarter of normal modes describes 85 % of the space spanned by the five ED modes. Furthermore, ED and CGNM modes do not differentiate on the basis of protein structural class (Class level in CATH classification). However, for selected cases, it was found that CGNM modes are more robust within the same family (Homologous superfamily levels in CATH) than ED modes. In view of recent[223-225] evidences regarding evolutionary conservation of vibrational dynamics, this suggests that ED modes, in some cases, might not be representative of the underlying dynamics characteristic for a whole family, probably due to insufficient sampling in MD.

The finding that MD essential directions are very well reproduced by CGNM approaches on a large and diverse dataset of proteins illustrates the potential of CGNM approaches in describing the intrinsic motions of proteins. The intrinsic motions of a protein are not only related to its functions according to the "conformational selection model"[26-29] but also to allosteric regulations following a "modern view of allostery"[266,267] and evolvability[225,235] of proteins. Hence, being able to predict the intrinsic motions of proteins with almost no computational cost can be extremely helpful in the development of computational approaches, especially in the field of structural-based drug design (SBDD). In this work, the directional information, provided by the CGNM approach RCNMA, is utilized to sample the biologically relevant conformational space of a protein.

In order to exploit the potential of CGNM approaches, I have developed a three-step approach for efficient exploration of intrinsic motions of proteins. The first two steps are based on recent developments in rigidity and elastic network theory.[68] Initially, *static* properties of the protein are determined by decomposing the protein into rigid clusters using the graph-theoretical approach FIRST[91] at an all-atom representation of the protein. In a second step, *dynamic* properties of the molecule are revealed by the rotations-translations of blocks approach (RTB)[178] using an elastic network model representation of the coarse-grained protein. In the final step, the recently introduced idea of constrained geometric simulations of diffusive motions in proteins[64] is extended for efficient sampling of conformational space. Here, the low-energy (frequency) normal modes provided by the RCNMA approach are used to guide the backbone motions. The side-chains observe diffusive motion biased towards energetically favorable rotamers. This is an iterative approach, which progress in small steps and generates intermediate conformations at every step.

The NMSim approach was validated on hen egg white lysozyme by comparing it to previously mentioned simulation methods in terms of residue fluctuations, conformational space explorations, essential dynamics,[215,216,242] sampling of side-chain rotamers, and structural quality. Residue fluctuations in NMSim generated ensemble is found to be in good agreement with MD fluctuations[207] with a correlation coefficient of around 0.79. A comparison of different geometry-based simulation approaches shows that FRODA is restricted in sampling the backbone conformational space; an average backbone RMSD from the starting structure of 0.37 Å is observed for the FRODA generated ensemble compared to 1.03 Å and 1.40 Å RMSD for MD and NMSim ensembles, respectively. CONCOORD is restricted in sampling the side-chain conformational space; on average, CONCOORD samples 1.63 rotamer states out of 10, in contrast to 5.78 and 4.97 rotamer states sampled in MD and NMSim, respectively. NMSim sufficiently samples both the backbone and the side-chain conformations taking experimental structures and conformations from the state of the art MD simulation as reference. Furthermore, the explicit modeling of rotamer states in NMSim improves the quality of side-chain conformations; the rotamericity increases from 0.57 to 0.70.

It is important to note that the use of directional information differentiates the NMSim approach from the other geometry-based approaches, FRODA and CONCOORD. The FRODA[64] and the NMSim approaches share a natural way of coarse-graining,[51] i.e., rigidity analysis using FIRST approach,[91,161,204] at their core levels. However, they differ at simulation levels. FRODA uses diffusive motion[64] of rigid regions. Therefore, due to the lack of direction, sampling in FRODA is limited, particularly in those cases where proteins are relatively flexible. In contrast, NMSim uses normal mode directions to guide backbone motions, but uses diffusive motions for side-chains. The CONCOORD approach[62] iteratively satisfies inter-atomic distance constraints to generate conformations starting from randomized atomic coordinates. Therefore, the CONCOORD generated structures are sensitive to the inter-atomic distances of the starting structure. In comparison, the NMSim approach relies on the intrinsic mobility information obtained from CGNM approaches of the previously generated structure. This is achieved by moving atomic coordinates of a starting/generated structure, iteratively, in the low-energy normal mode space instead of randomizing atomic coordinates.

The NMSim approach is also applied to a dataset of proteins where conformational changes have been observed experimentally, either in domain or functionally important loop regions. The NMSim simulations starting from the unbound structures are able to reach conformations similar to ligand bound conformations (RMSD < 2.4 Å) in 4 out of 5 cases of domain moving proteins. In these four cases, good correlation coefficients (R > 0.7) between the RMS fluctuations derived from NMSim generated structures and two experimental structures are observed. Furthermore, intrinsic fluctuations in NMSim simulation correlate with the region of loop conformational changes observed upon ligand binding in 2 out of 3 cases. It is suggested in this study that the radius of gyration (R_g), if used in combination with low-frequency normal modes, improves the search for ligand bound conformations in NMSim.

The NMSim generated pathway of conformational change from the unbound structure to the ligand bound structure of adenylate kinase is validated by a comparison to experimental structures reflecting different states of the pathway as proposed by previous studies.[87-89] Different crystal structures that lie along the transition from the unbound structure to the ligand-bound structure are closely sampled in the NMSim generated pathway. Interestingly, the generated pathway confirms that the LID domain closure precedes the closing of the NMPbind domain, even if no target conformation is provided in NMSim.

Hence, the results in this study show that, incorporating directional information in the geometry-based approach NMSim improves the sampling of biologically relevant conformational space and provides a computationally efficient alternative to state of the art MD simulations.

Zusammenfassung

Konformationsänderungen von Proteinen sind häufig eine grundlegende Voraussetzung für deren biologische Funktion. Die genaue Charakterisierung und Vorhersage dieser Konformationsänderungen ist nicht nur für das Verständnis ihres Einflusses auf die Funktion erforderlich, sondern liefert auch hilfreiche Anhaltspunkte für die Modellierung der Protein-Komplexbildung und für das strukturbasierte Wirkstoffdesign (SBDD). Das Konformations-Selektions-Modell liefert die Grundlage für computergestützte Untersuchungen der konformationellen Diversität ungebundener Proteine, welche auch gebundene Konformationen einschließen kann.[33]

In diesem Zusammenhang sind computergestützte Methoden von großem Nutzen, welche die Flexibilität und Plastizität von Proteinen beschreiben. Eines der dafür am häufigsten verwendeten und genauesten computergestützten Verfahren ist die Molekulardynamik-Simulationen[56-58] (MD Simulationen). Trotz der immensen Steigerung der verfügbaren Rechenkapazitäten sind MD Simulationen nach wie vor sehr rechenintensiv und durchmustern den Konformationsraum nur in begrenztem Maße, da die Energiebarrieren in der komplexen Energielandschaft eines Proteins nur langsam überwunden werden können.[59,60] Daher wurden Anstrengungen unternommen, alternative Methoden zu entwickeln, die auf einer reduzierten Darstellung von Proteinen beruhen, dafür aber den biologisch relevanten Konformationsraum rechnerisch viel effizienter durchmustern können. Ein Beispiel ist das geometriebasierte Programm CONCOORD, welches ausgehend von einer Protein-Startstruktur, unter Berücksichtigung von Distanzeinschränkungen, neue Konformationen erzeugt.[62,63] Der alternative geometriebasierte Ansatz FRODA erzeugt Konformationen durch die Diffusionsbewegungen von flexiblen und rigiden Teilbereichen in einer Proteinstruktur.[64] Bisher verwenden diese geometriebasierten Verfahren keine Richtungsinformationen für eine gerichtete Bewegung zur Durchmusterung des biologisch relevanten Konformationsraumes.

Das Ziel dieser Arbeit ist, Richtungsinformationen in einen geometriebasierten Ansatz zu integrieren und so den biologisch relevanten Konformationsraum erschöpfend zu durchmustern. Dies führte kürzlich zur Entwicklung von „coarse-grained normal mode“ (CGNM) Methoden, wie zum Beispiel dem „elastic network model“ (ENM) und der von mir in vorangegangenen Arbeiten entwickelte „rigid cluster normal mode analysis“ (RCNMA). Die beiden Methoden liefern die gewünschte Richtungsinformation der intrinsischen Bewegungen eines Proteins in Form von harmonischen Moden (auch Normalmoden).[67,68] Die Normalmoden entsprechen in diesem Zusammenhang den Deformierungsmöglichkeiten des Proteins und können anhand des Energieaufwandes bei der Deformation sortiert werden. In meinen vorangegangenen Arbeiten[68] und in weiteren Studien[67,69-71] konnte unter Verwendung von CGNM Methoden in Übereinstimmung mit dem Konformations-Selektions-Modell gezeigt werden, dass bei vielen Proteinen die durch die Bindung des Liganden bedingten Konformationsänderung nur entlang weniger, energiearmer Moden des ungebundenen Proteins stattfindet.

Um die Aussagekraft, Robustheit und breite Anwendbarkeit solcher CGNM Verfahren zu untersuchen, wurde im Rahmen dieser Dissertation ein umfangreicher Vergleich zwischen „essential dynamics“ (ED) Moden aus MD Simulationen und Normalmoden aus CGNM Berechnungen durchgeführt. Der zugrundeliegende Datensatz enthielt 335 Proteine. Obwohl die CGNM Verfahren eine stark vereinfachte Darstellung für Proteine verwenden, korrelieren die niederfrequenten Moden dieser Verfahren bezüglich ihrer Bewegungs-Richtung (durchschnittliche maximale Überschneidung: 0,65) und -Amplitude (durchschnittliche maximale Überschneidung: 0,73) sehr gut mit ED Moden. Im Durchschnitt beschreibt das erste Viertel der Normalmoden 85 % des Raumes, der durch die ersten fünf ED Moden aufgespannt wird. In einigen Ausnahmefällen konnte gezeigt werden, dass sich CGNM Moden innerhalb einer Proteinfamilie (homologe Superfamilie in CATH) robuster verhalten als ED Moden. Mit Blick auf neuere Erkenntnisse[223-225] bezüglich der evolutionären Konservierung von Vibrations-Dynamik in Proteinfamilien heißt dies, dass ED Moden die zugrundeliegenden dynamischen Charakteristiken schlechter abbilden. Dies kann möglicherweise durch die ungenügende Durchmusterung des Konformationsraumes durch die MD Simulationen erklärt werden.

Anhand dieses großen und diversen Datensatzes von Proteinen konnte gezeigt werden, dass CGNM essentielle Bewegungsrichtungen äquivalent zu MD

Simulationen abbilden kann und daher über das Potential verfügt, die intrinsische Dynamik von Proteinen zu beschreiben. Die intrinsische Dynamik von Proteinen wiederum steht nicht nur in direktem Zusammenhang mit dem Konformations-Selektions-Modell,[26-29] sondern auch mit allosterischen Regulationswegen in Proteinen im Sinne des „modern view of allostery"[266,267] und der Richtung evolutionärer Strukturveränderungen in Proteinen.[225,235] Die Möglichkeit, intrinsische Dynamik von Biomolekülen mit geringem Rechenaufwand vorherzusagen, ist für die Entwicklung weiterer Computermethoden von Nutzen, insbesondere im Bereich des strukturbasiertem Wirkstoffdesigns. In dieser Arbeit wurde der CGNM Ansatz RCNMA verwendet, um Richtungsinformationen abzuleiten und diese für die Durchmusterung des biologisch relevanten Konformationsraumes zu verwenden.

Um die Leistungsfähigkeit von CGNM Verfahren genauer zu bestimmen, wurde im Rahmen der vorliegenden Studie eine dreistufige Methode zur Untersuchung der intrinsischen Dynamik von Proteinen entwickelt. Die ersten beiden Stufen basieren auf neuen Entwicklungen in der Rigiditäts-Theorie und der Beschreibung von elastischen Netzwerken.[68] Im ersten Schritt werden hierbei statische Eigenschaften des Proteins mit Hilfe des graphentheoretischen Ansatzes FIRST[91] bestimmt, welcher die einzelnen Atome des Proteins in rigide und flexible Teilbereiche zusammenfasst. Im zweiten Schritt wird diese Einteilung in rigide und flexible Teilbereiche verwendet, um die dynamischen Eigenschaften des Proteins durch das sogenannte „rotations-translations of blocks“ (RTB)[178] Verfahren zu beschreiben. Im letzten Schritt wird die kürzlich beschriebene Idee der eingeschränkten, geometrischen Simulation von Diffusionsbewegungen[64] erweitert und zur effizienten Durchmusterung des Konformationsraumes eingesetzt. Dabei werden die Bewegungen des Proteinrückgrates entlang der mittels RCNMA erzeugten niederenergetischen Normalmoden ausgerichtet. Die Seitenkettenkonfomrationen werden dabei durch Diffusionsbewegungen hin zu energetisch günstigen Rotameren erzeugt. Dies ist ein iterativer Prozess, bestehend aus mehreren kleineren Schritten, in denen jeweils intermediäre Konformationen erzeugt werden.

Zur Validierung des NMSim Ansatzes wurde dieser mit den anderen zuvor genannten Simulationsmethoden am Beispiel von Lysozym aus Hühnereiweiß verglichen. Als Bewertungskriterien wurden die Fluktuationswerte der jeweiligen Reste, die

Vollständigkeit der Durchmusterung des Konformationsraumes, die „essential dynamics“[215,216,242] Moden, die Durchmusterung der Seitenkettenrotamere und die Qualität der Struktur verwendet. Die Fluktuationen der Aminosäurereste aus dem mit NMSim erzeugten Ensemble stimmen mit den Fluktuationen aus der MD Simulation[207] gut überein (Korrelationskoeffizient R = 0,79).

Ein Vergleich der unterschiedlichen geometriebasierten Simulationsansätze zeigt, dass bei FRODA die Durchmusterung des Konformationsraumes des Proteinrückrates unzureichend ist. Im Vergleich zu den MD und NMSim erzeugten Ensembles, die jeweils eine durchschnittliche RMS Abweichung zur Startstruktur von 1,03 Å und 1,40 Å erzielen, weist das FRODA generierte Ensemble mit einem durchschnittlichen RMSD von 0,37 Å nur eine geringe Abweichung auf. Bei CONCOORD ist hingegen die Durchmusterung des Konformationsraumes der Seitenketten unzureichend. Verglichen mit durchschnittlich jeweils 5,78 und 4,97 durchmusterten Rotamerzustände von MD und NMSim generierten Ensembles erzeugt CONCOORD durchschnittlich nur 1.63 Rotamerzustände.

NMSim hingegen durchmustert sowohl den Konformationsraum des Proteinrückrates als auch den der Seitenketten angemessen, wenn man die experimentell und mittels MD Simulationen erzeugten Konformationen als Referenz verwendet. Weiterhin führt die explizite Modellierung der Rotamerzustände in NMSim zu einer erhöhten Qualität der Seitenkettenkonformationen: die „rotamericity“ steigt von 0,57 auf 0,70.

Es ist wichtig zu erwähnen, dass sich die NMSim Methode durch die Verwendung richtungsbezogener Information von anderen geometrie-basierten Ansätzen, wie FRODA und CONCOORD, unterscheidet. FRODA und NMSim basieren beide auf einer vereinfachten Darstellung des Proteins,[64] welche beispielsweise mit Hilfe des FIRST Ansatzes[91,161,204] basierend auf der Rigiditätsanalyse erreicht werden kann. Die beiden Methoden unterscheiden sich jedoch auf der Simulationsebene. FRODA verwendet Diffusionsbewegung rigider Bereiche. Durch die fehlende Bewegunsrichtung ist die Durchmusterung in FRODA eingeschränkt, insbesondere bei flexiblen Proteinen. Im Gegensatz dazu verwendet NMSim die Richtung der Normalmoden, um die Bewegungen des Proteinrückrates zu steuern, und Diffusionsbewegungen für die Bewegungen der Seitenketten. Beim CONCOORD

Ansatz werden iterativ interatomare Distanzeinschränkungen (“constraints”) optimiert, um ausgehend von randomisierten Atomkoordinaten sinnvolle Konformationen zu erzeugen. Deshalb sind die mit CONCOORD generierten Strukturen stark abhängig von den interatomaren Distanzen in der Startstruktur. Im Vergleich dazu ist der NMSim Ansatz von der intrinsischen Bewegungsinformation des CGNM Ansatzes abhängig, die aus dessen Anwendung auf die im vorherigen Schritt erzeugte Konformation stammt. Dies wird durch die iterative Veränderung der Atomkoordinaten der vorherigen Konformation im niederenergetischen Normalmodenraum anstatt durch deren Randomisierung erreicht.

Der NMSim Ansatz wurde ebenfalls auf einen Datensatz von Proteinen angewendet, für die Konformationsänderungen in Domänen oder in funktionell wichtigen Schleifenregionen experimentell beobacht wurden. In Übereinstimmung mit dem Konformations-Selektions-Modell ist der NMSim Ansatz bei vier von fünf Proteinen, die eine Domänenbewegung aufweisen, in der Lage, ausgehend von der ungebundenen Struktur neue Konformationen zu erzeugen, die der ligandgebundenen Konformation entsprechen (RMSD < 2,4 Å). In diesen vier erfolgreichen Fällen wurde ein hoher Korrelationskoeffizient (R > 0,7) zwischen der RMS Fluktuation der durch NMSim erzeugten Konformationen und jeweils zwei experimentellen Strukturen erreicht. Hingegen korrelieren die intrinischen Fluktuationen der NMSim Simulation in zwei von drei Fällen mit dem Bereich der ligandinduzierten Konformationsänderung in den Schleifen. In dieser Studie wird gezeigt, dass die Verwendung des Gyrationsradius (Rg) in Kombination mit niederfrequenten Normalmoden in NMSim die Suche nach ligandgebundenen Konformationen verbessert.

Der mit NMSim generierte Pfad für die Konformationsänderungen von der ungebundenen Struktur zur ligandgebundenen Struktur der Adenylat-Kinase wurde durch den Vergleich zu experimentellen Strukturen validiert, die, wie in vorangegangenen Studien gezeigt werden konnte,[87-89] verschiedene Zustände des Pfades widerspiegeln. Die unterschiedlichen Kristallstrukturen, die entlang der Konformationsänderungen von der ungebundenen zur ligandgebundenen Struktur liegen, werden auf dem von NMSim erzeugten Pfad durchmustert. Interessanterweise bestätigt der generierte Pfad, dass die Schließbewegung der LID Domäne derjenigen der NMPbind Domäne vorangeht, sogar wenn keine Zielkonformation für die NMSim Simulation verwendet wurde.

Die Ergebnisse dieser Arbeit zeigen, dass die Einbeziehung richtungsbezogener Information in den geometriebasierten NMSim Ansatz die Durchmusterung des biologisch relevanten Konformationsraumes verbessert und somit eine recheneffiziente Alternative zu den aktuellen MD Simulationen darstellt. Hybride Normalmoden-Ansätze,[72,73,80,81,260] insbesondere in der Kombination mit experimentellen Daten (zum Beispiel Röntgenkristallographie, NMR, Cryo-EM, SAXS), haben sich in verschiedenen Anwendungen als erfolgreich erwiesen. Wie bereits erwähnt, konnte in Analogie dazu in dieser Studie gezeigt werden, dass die Berücksichtigung des Gyrationsradius (Rg) in Kombination mit berechneten Normalmoden in NMSim die Suche nach gebunden Konformationen verbessert. Dies gilt für Scharnierbewegungen („*hinge* bending motions“), Scherbewegung („sheer motions“) und Bewegungen in Schleifenregionen („loop motions“). Eine potentielle Erweiterung für NMSim wäre somit die Einbeziehung experimenteller Daten, wie etwa paarweiser Distanzeinschränkungen oder Gyrationsradien, wodurch sicherlich gebundene Konformationen effizienter vorhergesagt werden könnten.

Die aktuellen Entwicklungen im Bereich der geometriebasierten Simulationsmethoden sind sowohl für die Simulation großer Konformationsänderungen als auch für kombinierte Anwendungen mit molekularem Docking und virtuellen Screening vielversprechend. Offensichtliche Anwendungen liegen hierbei beim Docken in Multiple-Rezeptorkonformationen (MRC) und sogar im Bereich des Hochdurchsatzdockings.[40] Insbesondere bilden solche effizient generierten Konformations-Ensemble die Grundlage für die implizite Berücksichtigung der Rezeptormobilität in Dockinganwendungen. Ein Bespiel hierfür ist eine kürzlich veröffentlichte Studie, die Rezeptormobilität implizit durch eine elastische Netzwerkrepräsentation moduliert.[268]

Outlook

Hybrid normal mode approaches,[72,73,80,81,260] particularly in combination with experimental data (e.g., X-ray, NMR, cryo-EM, SAXS), have been found successful in different applications. Following a similar direction, it was found in this study that the radius of gyration (R_g) if used in combination with normal modes improves the search for ligand bound conformations in NMSim. This is not only true for hinge bending motions but also sheer motions and loop motions. Considering these facts, a potential extension in NMSim would be to incorporate experimental data (for example a small set of pairwise distance constraints or the R_g of the ligand bound conformation) to improve the prediction for the ligand bound conformations.

The recent developments in geometry-based simulation approaches are promising not only in large-scale conformational changes predictions but also in combination with molecular docking and virtual screening approaches. The obvious use of these efficient approaches is in combination with multiple receptor conformations (MRC) docking[40] and even for high throughput docking.[61] Moreover, theses efficiently generated ensembles provide the basis for approaches that implicitly incorporate receptor mobility in docking approaches, for example, as proposed recently,[268] through an elastic representation of a potential grid in the binding pocket region of a receptor.

The NMSim approach can also be extended to nucleic acids. Although normal mode analysis have been applied to investigate DNA and RNA dynamics,[127-129] a large-scale CGNM validation study would be required for nucleic acids too. The NMSim approach can be improved by enhancing the underlying constraint network. For example, by considering the breaking and formation of non-covalent bonds, during the NMSim simulation, based on the atom movements predicted by the normal mode directions. Furthermore, a ligand influence on a receptor can be modeled by biasing modes, and thus the motion of a receptor that influences the binding pocket of a receptor.

Acknowledgements

First of all, a special thanks to my supervisor, Prof. Dr. Holger Gohlke, for his guidance through out the work and a huge amount of support during my stay in Germany, which made it possible to complete this dissertation. His lively, enthusiastic, and energetic personality is inspiration for me. I am particularly very grateful for his scientific advices, discussions and suggestions.

I wish to express my gratitude to the members of Bio- and Chemo-informatics department of Merck Serono, Darmstadt, for a successful collaboration during the course of this work. I would like to convey my deep regards to Dr. Friedrich Rippmann and Dr. Gerhard Barnickel for helpful discussions and for providing feedbacks to the NMSim approach. Every meeting we had provided me the real opportunity to get insight into practical problems in pharmaceutical industry and to guide my research accordingly.

I am deeply grateful to Saskia Villinger for her involvement and cooperation for the large-scale comparison study. I am also grateful to Alrun Koller for providing MD trajectory data for methodological comparisons. Final thesis write-up is never easy; I would like to say thanks to Alrun Koller, Sebastian Radestock and Alexander Metz for their suggestions and discussions during this phase. A special thanks to Sina Kazemi for his interesting feedbacks and valuable suggestions.

I am privileged for having unforgettable friends and colleagues, who have helped me more than I expected and whenever I needed. I would like to mention Domingo González Ruiz, Teresa Jimenez Vaquero, Simone Fulle, Hannes Kopitz, Christopher Pfleger, Dennis Krüger and Johannes Bergs for their support and friendship.

A special gratitude goes to my friends and former colleagues, Benjamin Breu, Elena Schmidt, Christina Wendel, Eva Kestner, Junaid Owasil, and Radhan Ramadass, who have provided great company during all those years of my stay in Frankfurt and supported me in every time of need.

Finally, I would like to thank my parents, Zahoor Ahmed and Jamila Ahmed, for their love, support, and constant prayers, and for simply being there for me in every circumstance.

Appendix

Appendix A: The parameter set used in NMSim.

Parameter names	**Default values**	**Description**
SIM_ITER	500	Number of simulation cycles for calling RCNMA and NMSim alternatively.
MOVE_ITER	10	Number of NMSim cycles.
SHAKE_ITER	500	Maximum number of structure correction cycles.
NMRANGE	7 to 56	Normal modes range used for linear combination.
ECUT	-1.0 kcal/mol	Energy cutoff for hydrogen bonds.
RMSDSTEPSIZE	0.5 Å	Structure distortion (of all atoms) in normal mode directions in an NMSim cycle.
TEMPERATURE	300 K	Temperature for atomic fluctuation calculations from normal modes.
WRITECONFEVERY	1	Frequency of writing out conformations during NMSim cycles.
SELECTCONF	0	Select conformation for next simulation cycle. 1 = lowest ROG, 2 = highest ROG, or 3 = nearest ROG as compared to the previously generated structure.
RANDSCALING	0.3 Å	Scaling factor for random component in side-chain directions.
MISS_SLOPE_TOL	0.01	Exit criteria for structure correction cycle.
VDW_CUT	8.0 Å	Van der Waals cutoff used in structure correction.
CF_DIST_TOL	0.005 Å	Tolerance allowed for covalent distance constraints.
VDW_DIST_TOL	0.07	Tolerance (in fraction of vdW sum) allowed for vdW distance constraints (excluding 1-4 constraints).

VDW_ONE4_DIST_TOL	0.20	Tolerance (in fraction of vdW sum) allowed for vdW 1-4 distance constraints.
HBSB_DIST_TOL	0.05 Å	Tolerance allowed for hydrogen bond distance constraints.
PH_DIST_TOL	0.05 Å	Tolerance allowed for hydrophobic distance constraints.
PHIPSI_DIST_TOL	0.05 Å	Tolerance allowed for φ/ψ distance constraints.
BB_PLANAR_TOL	0.017 Rad. (1°)	Tolerance allowed for backbone planar constraints.
SC_PLANAR_TOL	0.001 Å	Tolerance (from ideal planarity) allowed for side-chain planarity constraints.
ROTAMER_TOL	0.174 Rad. (10°)	Tolerance allowed for χ-angles dihedral constraints from rotamer.
ADJUST_FAC_CF	0.5	Adjustment factor for covalent distance constraints.
ADJUST_FAC_VDW	0.4	Adjustment factor for vdW distance constraints.
ADJUST_FAC_HBSB	0.2	Adjustment factor for hydrogen bond distance constraints.
ADJUST_FAC_PH	0.1	Adjustment factor for hydrophobic distance constraints.
ADJUST_FAC_PHIPSI	0.005	Adjustment factor for φ/ψ distance constraints.
ADJUST_FAC_BB_PLANAR	0.02	Adjustment factor for backbone planar constraints.
ADJUST_FAC_SC_PLANAR	1.0	Adjustment factor for side-chain planar constraints.
ADJUST_FAC_ROTAMER	0.001	Adjustment factor for χ-angle dihedral constraints.
CHIDEV_SELLIMIT	1.047 Rad.(60°)	Chi-limit in making Candidate rotamer list for residues.

Appendix B: The list of 130 experimental structures of Hen Egg White Lysozyme.

1ic5_Y_01
1c08_C_01
2dqd_Y_01
1j1p_Y_01
1ri8_B_01
2dqc_Y_01
2dqg_Y_01
2dqh_Y_01
2dqj_Y_01
2dqf_C_01
2zq3_A_01
1j1o_Y_01
3hfm_Y_01
2dqe_Y_01
1j1x_Y_01
1lys_A_01
1ua6_Y_01
2f4g_A_01
2d4j_A_01
3lyt_A_01
1mlc_E_01
1xgu_C_01
2f30_A_01
2dqi_Y_01
3d9a_C_01
1dqj_C_01
1jto_L_01
2f4a_A_01
1xgp_C_01
2lzt_A_01
1xgt_C_01
1v7s_A_01
4lzt_A_01
2z12_A_01
1lzn_A_01
2f2n_A_01
1lks_A_01
1xgr_C_01
2z19_A_01
2vb1_A_01
1zmy_L_01
1lzt_A_01
1xgq_C_01
1xfp_L_01
2hs7_A_01
3lzt_A_01
1v7t_A_01
1xei_A_01
2hs9_A_01
2hso_A_01
2z18_A_01
1xej_A_01
1sq2_L_01
1jtt_L_01
2yss_C_01
1xek_A_01
2a6u_A_01
1bvk_C_01
1kiq_C_01
1ja6_A_01
1g7j_C_01
1g7i_C_01
1vfb_C_01
1g7m_C_01
1kir_C_01
2fbb_A_01
1g7l_C_01
1kip_C_01
1g7h_C_01
1ja7_A_01
1ja4_A_01
1ja2_A_01
1sfb_A_01
1sf7_A_01
1gxx_A_01
1sf4_A_01
1sfg_A_01
1gxv_2_01
1e8l_A_49
1e8l_A_48
1sf6_A_01
1e8l_A_43
1e8l_A_14
1e8l_A_46
1e8l_A_50
1e8l_A_47
1e8l_A_45
1e8l_A_18
1e8l_A_44
1e8l_A_08
1e8l_A_10
1e8l_A_17
1b2k_A_01
1e8l_A_07
1e8l_A_09
1e8l_A_04
1e8l_A_05
1e8l_A_41
1e8l_A_31
1e8l_A_01
1e8l_A_19
1e8l_A_15
1e8l_A_39
1e8l_A_24
1e8l_A_03
1e8l_A_16
1e8l_A_28
1e8l_A_38
1e8l_A_40
1e8l_A_36
1e8l_A_32
1e8l_A_30
1e8l_A_26
1e8l_A_35
1e8l_A_34
1e8l_A_27
1e8l_A_20
1e8l_A_42
1e8l_A_22
1e8l_A_11
1e8l_A_29
1e8l_A_23
1e8l_A_33
1e8l_A_21
1e8l_A_12
1e8l_A_02
1e8l_A_25
1e8l_A_06
1e8l_A_37
1e8l_A_13

The above list, which is divided into columns for clarity, is sorted with the increasing C_α RMSD to the reference structure of HEWL (PDB code 1hel);[208] The top-left structure has the smallest C_α RMSD of 0.5 Å and the bottom-right has the largest C_α RMSD of 1.8 Å. The experimental structures in the list have the format PDB-code_Chain_Model.

Appendix C: The ED and CGNM mode comparison result along with CATH classifications.

Fold[a]	**A**[b]	**T**[b]	**H**[b]	**PDB**	**Size**[c]	**ENM**[d]		**RCNMA**[e]		**ENM**[d] **Mean ± SD**		**RCNMA**[e] **Mean ± SD**	
						Lap[f]	**Corr**[g]	**Lap**[f]	**Corr**[g]	**Lap**[f]	**Corr**[g]	**Lap**[f]	**Corr**[g]
Class1: mainly-alpha													
Arc Repressor Mutant, subunit A	10	10	10	1bby	69	0.62	0.87	0.63	0.66	0.68 ± 0.11	0.71 ± 0.15	0.69 ± 0.08	0.69 ± 0.11
				1d8k	81	0.74	0.79	0.73	0.80				
				1hks	106	0.73	0.71	0.73	0.70				
				1irf	112	0.65	0.63	0.61	0.47				
				1lea	72	0.68	0.79	0.66	0.71				
			60	1bw6	56	0.69	0.67	0.70	0.68				
				2ezh	65	0.51	0.36	0.59	0.62				
				2ezk	93	0.90	0.83	0.86	0.84				
			250	1fow	76	0.63	0.72	0.68	0.77				
Cytochrome Bc1 Complex; Chain D, domain 2	10	760	10	1c52	131	0.41	0.60	0.46	0.34	0.48 ± 0.10	0.56 ± 0.19	0.52 ± 0.12	0.48 ± 0.24
				1c75	71	0.45	0.29	0.42	0.22				
				1ccr	**111**	**0.58**	**0.60**	**0.67**	**0.68**				
				1co6	**107**	**0.31**	**0.31**	**0.36**	**0.20**				
				1cot	121	0.37	0.53	0.41	0.34				
				1cyj	90	0.46	0.47	0.57	0.44				
				1fi3	82	0.58	0.77	0.59	0.75				
				1gdv	85	0.59	0.88	0.68	0.86				
				3c2c	112	0.55	0.62	0.54	0.52				
Recoverin; domain 1	10	238	10	1b8l	108	0.83	0.84	0.61	0.71	0.62 ± 0.10	0.59 ± 0.26	0.60 ± 0.07	0.60 ± 0.18
				1g33	73	0.67	0.89	0.73	0.79				
				1rk9	110	0.60	0.65	0.58	0.59				
				1rro	108	0.62	0.34	0.65	0.40				
				1sra	151	0.61	0.73	0.59	0.84				
				2bca	75	0.54	0.55	0.56	0.58				
				3pat	109	0.51	0.11	0.51	0.32				
				1c3z	108	0.54	0.57	0.53	0.57				
Lysozyme	10	530	10	1b9o	123	0.68	0.76	0.68	0.76	0.68 ± 0.09	0.68 ± 0.14	0.66 ± 0.10	0.68 ± 0.15
				1gd6	119	0.66	0.48	0.67	0.54				
				1hfx	123	0.69	0.65	0.59	0.53				
				1i56	130	0.50	0.62	0.48	0.56				
				1iiz	120	0.69	0.70	0.68	0.71				
				1jug	125	0.71	0.65	0.73	0.75				
				2eql	129	0.81	0.92	0.81	0.92				
Globins	10	490	10	1a6m	151	0.61	0.50	0.68	0.57	0.64 ± 0.17	0.57 ± 0.26	0.61 ± 0.15	0.59 ± 0.27
				1dlw	116	0.60	0.60	0.47	0.49				
				1hlb	157	0.54	0.41	0.38	0.12				
				2gdm	153	0.57	0.73	0.69	0.88				
				2hbg	147	0.53	0.22	0.67	0.70				
			30	1a87	297	0.98	0.96	0.78	0.76				
Non-ribosomal Peptide Synthetase Peptidyl Carrier Protein; Chain A	10	1200	10	1hqb	80	0.49	0.53	0.48	0.53	0.57 ± 0.07	0.64 ± 0.06	0.52 ± 0.07	0.63 ± 0.16
				1hy8	76	0.58	0.72	0.47	0.46				
				2af8	86	0.51	0.64	0.42	0.49				
			20	1cei	85	0.67	0.62	0.53	0.80				
				1gxg	85	0.62	0.67	0.62	0.67				
				1imq	86	0.57	0.64	0.58	0.83				
Four Helix Bundle (Hemerythrin (Met), subunit A)	20	120	10	1apc	106	0.71	0.67	0.56	0.64	0.69 ± 0.10	0.70 ± 0.17	0.63 ± 0.13	0.71 ± 0.15
			20	1aep	153	0.80	0.67	0.69	0.67				
				1bz4	144	0.56	0.52	0.49	0.58				
			30	1jmw	146	0.68	0.92	0.78	0.93				
5' to 3' exonuclease, C-terminal subdomain	10	150	20	1coo	81	0.85	0.88	0.83	0.80	0.76 ± 0.08	0.82 ± 0.07	0.71 ± 0.10	0.86 ± 0.06
				1doq	69	0.74	0.75	0.66	0.91				
			90	1tam	120	0.69	0.82	0.64	0.86				
Annexin V; domain 1	10	220	10	1ann	315	0.79	0.68	0.79	0.69	0.79 ± 0.05	0.69 ± 0.07	0.79 ± 0.06	0.69 ± 0.06
				1axn	323	0.75	0.63	0.73	0.63				
				1hvf	313	0.84	0.76	0.84	0.74				
434 Repressor (Amino-terminal Domain)	10	260	40	1neq	74	0.69	0.84	0.71	0.87	0.61 ± 0.14	0.62 ± 0.21	0.61 ± 0.13	0.70 ± 0.23
				1pru	56	0.70	0.60	0.66	0.79				
				1r69	63	0.45	0.42	0.46	0.44				
Death Domain, Fas	10	533	10	1ddf	127	0.71	0.91	0.71	0.91	0.66 ± 0.05	0.76 ± 0.21	0.66 ± 0.11	0.73 ± 0.28
				1e3y	104	0.61	0.86	0.73	0.86				
				2ygs	92	0.67	0.52	0.53	0.41				
Phospholipase A2	20	90	10	1lwb	122	0.53	0.44	0.56	0.60	0.60 ±	0.54 ±	0.58 ±	0.46 ±
				1pir	124	0.57	0.52	0.59	0.46				

				1umv	122	0.71	0.66	0.60	0.32	0.09	0.11	0.02	0.14
Growth Hormone; Chain: A;	20	1250	10	1buy	166	0.56	0.58	0.62	0.67	0.59 ± 0.22	0.60 ± 0.32	0.69 ± 0.14	0.73 ± 0.19
				1ijz	113	0.39	0.29	0.60	0.57				
				1irl	133	0.83	0.93	0.86	0.94				
Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat	25	40	20	1bd8	156	0.75	0.61	0.72	0.57	0.67 ± 0.10	0.64 ± 0.05	0.68 ± 0.04	0.64 ± 0.08
				2myo	118	0.71	0.70	0.64	0.64				
			90	1eyh	144	0.56	0.62	0.67	0.72				
Helicase, Ruva Protein; domain 3	10	8	60	1qzm	94	0.56	0.69	0.59	0.74	0.69 ± 0.18	0.80 ± 0.15	0.71 ± 0.16	0.82 ± 0.11
			100	1yub	245	0.82	0.90	0.82	0.90				
Actin-binding Protein, T-fimbrin; domain 1	10	418	10	1aa2	108	0.51	0.56	0.57	0.79	0.69 ± 0.25	0.75 ± 0.26	0.60 ± 0.04	0.83 ± 0.06
				1mb8	243	0.86	0.93	0.62	0.87				
Insulin-like, subunit E	10	100	10	1b9g	57	0.55	0.53	0.63	0.61				
Hydrophobic Seed Protein	10	110	10	1l6h	69	0.68	0.67	0.67	0.64				
Enzyme I; Chain A, domain 2	10	274	10	1eza	259	0.73	0.94	0.84	0.94				
Endonuclease V	10	440	10	2end	137	0.55	0.69	0.63	0.74				
Ribosomal Protein S7	10	455	10	1rss	135	0.92	0.97	0.88	0.93				
Peroxidase; domain 1	10	520	20	1abv	105	0.68	0.94	0.85	0.95				
Major Prion Protein	10	790	10	1ag2	103	0.72	0.72	0.53	0.72				
Cysteine Motif	10	810	10	1hp8	68	0.86	0.81	0.74	0.77				
N-utilizing Substance Protein B Homolog; Chain A	10	940	10	1tzw	142	0.80	0.87	0.79	0.86				
Villin Headpiece Domain; Chain A	10	950	10	1qqv	67	0.71	0.75	0.65	0.79				
Ribosomal Protein S4 Delta 41; Chain A, domain 1	10	1050	10	1c05	159	0.61	0.72	0.65	0.79				
c-terminal domain of poly(a) binding protein	10	1900	10	1i2t	61	0.59	0.55	0.79	0.89				
Pheromone ER-1	20	50	10	2erl	40	0.78	0.65	0.73	0.76				
Acyl-CoA Binding Protein	20	80	10	1mix	206	0.65	0.44	0.60	0.39				
Receptor-associated Protein	20	81	10	1nre	81	0.77	0.60	0.58	0.55				
Glycosyltransferase	50	10	20	1c3d	294	0.58	0.56	0.64	0.61				
Class 2: mainly-beta													
Immunoglobulin-like	60	40	10	1bmg	98	0.70	0.73	0.70	0.73	0.60 ± 0.11	0.63 ± 0.15	0.60 ± 0.12	0.62 ± 0.13
				1cdy	178	0.90	0.86	0.90	0.86				
				1cid	177	0.68	0.79	0.65	0.50				
				1nct	98	0.48	0.63	0.48	0.65				
				1qsz	101	0.57	0.65	0.54	0.64				
				1tit	89	0.68	0.58	0.68	0.58				
				1wit	93	0.58	0.36	0.59	0.38				
			20	1ok0	74	0.69	0.61	0.69	0.62				
			30	1bj8	109	0.55	0.35	0.55	0.35				
				1fna	91	0.61	0.31	0.45	0.42				
				1n6v	212	0.68	0.82	0.68	0.75				
			150	1bci	123	0.46	0.64	0.46	0.49				
			230	1noa	113	0.76	0.78	0.75	0.70				
			290	1e5b	87	0.46	0.63	0.46	0.63				
				1exg	110	0.44	0.65	0.43	0.66				
				1heh	88	0.56	0.67	0.56	0.67				
			420	1a8z	153	0.51	0.62	0.49	0.59				
				1aac	105	0.54	0.60	0.54	0.60				
				1bqk	124	0.62	0.76	0.67	0.84				
				1byp	99	0.47	0.43	0.50	0.45				

				1plb	97	0.54	0.70	0.54	0.70				
				1plc	99	0.70	0.80	0.70	0.80				
				1pmy	123	0.59	0.65	0.59	0.65				
				2cbp	96	0.48	0.65	0.48	0.65				
				2plt	98	0.48	0.33	0.50	0.35				
			550	1ifg	140	0.70	0.65	0.63	0.58				
			740	1amx	150	0.49	0.42	0.48	0.74				
			760	1whp	94	0.56	0.66	0.56	0.66				
			770	1ahk	129	0.68	0.59	0.68	0.59				
			830	1hcz	250	0.78	0.66	0.77	0.66				
			1030	1qts	247	0.74	0.88	0.82	0.76				
			1220	1m42	102	0.60	0.67	0.60	0.67				
OB fold (Dihydrolipoamide Acetyltransferase, E2P)	40	50	40	1b4o	62	0.81	0.87	0.76	0.86	0.73 ± 0.09	0.80 ± 0.10	0.72 ± 0.08	0.80 ± 0.10
				1bo0	76	0.65	0.71	0.73	0.83				
				1dol	71	0.78	0.92	0.77	0.92				
				1hfg	71	0.65	0.85	0.63	0.80				
				1je4	69	0.57	0.69	0.55	0.68				
				1sap	66	0.84	0.86	0.81	0.82				
			100	1dcz	77	0.67	0.81	0.67	0.81				
				1fyc	106	0.88	0.86	0.88	0.86				
				1ghj	79	0.73	0.55	0.71	0.54				
				1iyu	79	0.69	0.84	0.69	0.84				
				1lac	80	0.77	0.78	0.77	0.78				
			140	1ewi	114	0.75	0.84	0.75	0.84				
				1mjc	69	0.66	0.82	0.67	0.82				
Jelly Rolls	60	120	180	1bk1	182	0.71	0.72	0.45	0.74	0.61 ± 0.14	0.66 ± 0.16	0.57 ± 0.15	0.67 ± 0.17
			200	1a3k	137	0.38	0.34	0.38	0.36				
				1gbg	214	0.72	0.67	0.50	0.52				
				2ayh	214	0.61	0.60	0.59	0.57				
			230	1pgs	311	0.52	0.58	0.49	0.68				
			260	1gui	155	0.53	0.61	0.52	0.69				
				1kex	155	0.50	0.86	0.62	0.77				
				1ulo	152	0.78	0.68	0.68	0.74				
			390	1job	162	0.78	0.89	0.87	0.94				
SH3 type barrels	30	30	40	1ark	60	0.74	0.75	0.72	0.83	0.69 ± 0.08	0.80 ± 0.09	0.69 ± 0.08	0.81 ± 0.09
				1awj	77	0.76	0.85	0.76	0.85				
				1hsq	71	0.75	0.90	0.75	0.90				
				1pwt	61	0.70	0.78	0.70	0.78				
				1shg	57	0.54	0.61	0.54	0.61				
				1tuc	61	0.68	0.81	0.68	0.81				
			50	1qp2	70	0.76	0.87	0.75	0.88				
			190	1lpl	95	0.60	0.81	0.60	0.81				
Lipocalin	40	128	20	1bsq	162	0.53	0.54	0.46	0.57	0.53 ± 0.19	0.45 ± 0.34	0.54 ± 0.21	0.60 ± 0.25
				1cbs	137	0.39	-0.03	0.39	0.21				
				1ifc	131	0.36	0.24	0.41	0.62				
				1lpj	133	0.45	0.33	0.37	0.45				
				1ngl	179	0.89	0.93	0.88	0.93				
				1p6p	125	0.54	0.70	0.70	0.80				
Gamma-B Crystallin; domain 1	60	20	10	1ag4	103	0.55	0.69	0.54	0.59	0.64 ± 0.07	0.71 ± 0.07	0.64 ± 0.08	0.68 ± 0.09
				1amm	174	0.72	0.74	0.74	0.72				
			30	1bhu	102	0.68	0.72	0.68	0.72				
				1f53	84	0.58	0.60	0.58	0.60				
				1gh5	87	0.67	0.79	0.67	0.79				
Laminin	10	25	10	1ata	62	0.76	0.70	0.75	0.68	0.72 ± 0.13	0.69 ± 0.21	0.72 ± 0.13	0.70 ± 0.21
				1ip0	50	0.88	0.97	0.88	0.97				
				1k37	46	0.66	0.46	0.66	0.46				
				2tgf	50	0.59	0.61	0.58	0.68				
CD59	10	60	10	1chv	60	0.68	0.62	0.68	0.62	0.67 ± 0.14	0.73 ± 0.16	0.67 ± 0.15	0.73 ± 0.16
				1idi	**74**	**0.49**	**0.74**	**0.46**	**0.74**				
				1ntn	**72**	**0.82**	**0.95**	**0.82**	**0.96**				
				1txa	73	0.69	0.60	0.70	0.61				
Thrombin, subunit H	40	10	10	1arb	263	0.50	0.69	0.54	0.72	0.66 ± 0.16	0.76 ± 0.16	0.66 ± 0.12	0.76 ± 0.16
				1dua	242	0.85	0.95	0.78	0.93				
				1p3c	215	0.56	0.59	0.58	0.56				
				2sfa	191	0.71	0.81	0.73	0.83				
Cyclophilin	40	100	10	1a58	177	0.70	0.88	0.56	0.73	0.63 ± 0.16	0.77 ± 0.19	0.59 ± 0.12	0.80 ± 0.08
				1j2a	166	0.74	0.88	0.73	0.89				
				2cpl	164	0.45	0.55	0.49	0.78				
Trefoil (Acidic Fibroblast Growth Factor, subunit A)	80	10	50	1fmm	132	0.50	0.42	0.56	0.59	0.60 ± 0.14	0.60 ± 0.25	0.63 ± 0.10	0.68 ± 0.13
				1md6	154	0.70	0.77	0.70	0.77				
Cysteine Protease (Bromelain) Inhibitor,	10	69	10	1bbi	71	0.87	0.89	0.87	0.89				

subunit H													
Complement Module; domain 1	10	70	10	1fbr	93	0.82	0.75	0.82	0.75				
Pdz3 Domain	30	42	10	1iu0	91	0.69	0.58	0.69	0.56				
HIV-inactivating Protein, Cyanovirin-n	30	60	10	3ezm	101	0.72	0.84	0.74	0.85				
Heparin-binding Growth Factor, Midkine; Chain A, C-terminal Domain;	30	90	10	1mkc	43	0.90	0.68	0.90	0.68				
Barwin-like endoglucanases	40	40	20	1cz4	185	0.81	0.73	0.81	0.73				
Cathepsin D, subunit A; domain 1	40	70	10	1flh	326	0.60	0.81	0.59	0.60				
Substrate Binding Domain Of DNAk; Chain A, domain 1	60	34	10	1bpr	173	0.53	0.58	0.53	0.58				
Thaumatin	60	110	10	1aun	208	0.76	0.77	0.59	0.77				
Coagulation Factor XIII; Chain A, domain 1	70	50	30	1gdf	145	0.88	0.98	0.87	0.98				
Rieske Iron-sulfur Protein	102	10	10	1rfs	127	0.67	0.50	0.67	0.51				
Pectate Lyase C-like	160	20	10	1ee6	197	0.65	0.64	0.53	0.62				
Calcium-transporting ATPase, cytoplasmic transduction domain A	170	150	10	1h6q	168	0.87	0.94	0.88	0.95				
Class 3: alpha-beta													
Rossman fold	40	50	180	1chd	198	0.67	0.94	0.65	0.83	0.58 ± 0.16	0.62 ± 0.21	0.58 ± 0.14	0.65 ± 0.17
			270	1dg9	157	0.42	0.34	0.48	0.57				
			280	1b1a	137	0.53	0.67	0.53	0.53				
				1be1	137	0.65	0.78	0.67	0.78				
			300	1ak2	220	0.83	0.85	0.82	0.84				
			360	1akq	147	0.55	0.64	0.55	0.64				
			410	1ido	184	0.48	0.52	0.50	0.62				
				1mjn	179	0.49	0.51	0.50	0.61				
			1470	2pth	193	0.80	0.87	0.77	0.85				
			1820	1be0	310	0.35	0.43	0.45	0.57				
				1cex	197	0.46	0.61	0.56	0.66				
			2300	1tmy	118	0.44	0.31	0.49	0.33				
				2fsp	124	0.51	0.47	0.39	0.37				
			10190	1cdz	96	0.71	0.82	0.53	0.70				
				1imo	88	0.85	0.90	0.84	0.91				
Alpha-Beta Plaits	30	70	100	1opz	76	0.60	0.80	0.71	0.80	0.70 ± 0.12	0.76 ± 0.17	0.68 ± 0.13	0.73 ± 0.20
			250	1mla	305	0.90	0.84	0.91	0.86				
			330	1d8z	89	0.80	0.94	0.81	0.94				
				1hd0	75	0.75	0.74	0.67	0.66				
				2mss	75	0.65	0.62	0.63	0.63				
				2sxl	88	0.67	0.88	0.66	0.89				
				2u2f	85	0.69	0.69	0.58	0.43				
			400	1fwp	69	0.48	0.39	0.48	0.41				
			680	1f2h	169	0.63	0.77	0.60	0.74				
			830	1p11	102	0.78	0.95	0.79	0.95				
Ubiquitin-like (Ub-roll)	10	20	10	1pgx	70	0.69	0.83	0.66	0.78	0.68 ± 0.13	0.76 ± 0.11	0.68 ± 0.14	0.78 ± 0.12
			30	1frd	98	0.69	0.73	0.69	0.73				
				2cjn	97	0.56	0.73	0.55	0.75				
				4fxc	98	0.47	0.61	0.48	0.57				
			90	1jru	89	0.68	0.67	0.70	0.89				
				1rrb	76	0.67	0.71	0.67	0.71				
				1ubi	76	0.78	0.89	0.79	0.89				
			240	1ipg	85	0.92	0.94	0.93	0.95				
Defensin A-like	30	30	10	1c56	40	0.76	0.84	0.76	0.84	0.73 ± 0.09	0.78 ± 0.09	0.70 ± 0.16	0.77 ± 0.14
				1chz	64	0.79	0.65	0.83	0.78				
				1jxc	68	0.61	0.79	0.58	0.77				
				1jzb	66	0.60	0.66	0.39	0.46				

				1nra	63	0.81	0.93	0.82	0.93				
				1px9	42	0.68	0.74	0.65	0.70				
				2b3c	64	0.75	0.83	0.75	0.83				
				2sn3	65	0.82	0.81	0.85	0.86				
Glutaredoxin	40	30	10	1a23	189	0.50	0.71	0.46	0.56	0.55 ± 0.12	0.55 ± 0.22	0.53 ± 0.14	0.49 ± 0.22
				1aba	87	0.66	0.65	0.64	0.64				
				1gh2	107	0.48	0.30	0.55	0.31				
				1i5g	144	0.48	0.35	0.38	0.46				
				1o73	144	0.43	0.40	0.42	0.27				
				1thx	108	0.78	0.90	0.78	0.89				
				1trs	105	0.55	0.53	0.46	0.32				
Beta-Lactamase	30	450	20	1ew0	130	0.94	0.98	0.95	0.98	0.61 ± 0.18	0.71 ± 0.16	0.59 ± 0.22	0.69 ± 0.19
			30	1a0k	130	0.44	0.56	0.34	0.53				
				1acf	125	0.53	0.71	0.45	0.59				
				1pne	139	0.49	0.54	0.46	0.50				
			50	1h8m	140	0.66	0.78	0.68	0.82				
			70	1h3q	140	0.60	0.69	0.63	0.72				
SHC Adaptor Protein	30	505	10	1ayd	101	0.65	0.72	0.69	0.87	0.66 ± 0.11	0.73 ± 0.16	0.66 ± 0.11	0.77 ± 0.18
				1bfj	111	0.82	0.89	0.81	0.89				
				1jwo	97	0.57	0.78	0.57	0.79				
				1oo3	111	0.59	0.51	0.58	0.51				
Severin	40	20	10	**1ahq**	**133**	**0.55**	**0.60**	**0.41**	**0.24**	0.63 ± 0.08	0.75 ± 0.11	0.60 ± 0.13	0.65 ± 0.28
				1cof	**135**	**0.71**	**0.74**	**0.70**	**0.73**				
				1svr	94	0.57	0.81	0.60	0.81				
				2vik	126	0.70	0.86	0.67	0.83				
Ricin (A subunit); domain 1	40	420	10	1apa	261	0.68	0.65	0.68	0.66	0.63 ± 0.12	0.68 ± 0.22	0.62 ± 0.12	0.65 ± 0.29
				1d8v	263	0.77	0.93	0.76	0.93				
				1mrg	246	0.54	0.73	0.55	0.75				
				1mrj	247	0.51	0.39	0.50	0.25				
DNA Polymerase III; Chain A, domain 2	10	50	40	1jnt	92	0.48	0.59	0.48	0.60	0.54 ± 0.06	0.67 ± 0.07	0.55 ± 0.06	0.54 ± 0.05
				1rot	118	0.53	0.69	0.58	0.51				
				1yat	113	0.60	0.72	0.60	0.52				
TIM Barrel	20	20	80	1c3f	265	0.56	0.59	0.49	0.46	0.57 ± 0.12	0.60 ± 0.21	0.52 ± 0.03	0.63 ± 0.16
				1jfx	217	0.69	0.81	0.54	0.77				
			140	1vfl	15	0.45	0.40	0.54	0.66				
Double Stranded RNA Binding Domain	30	160	60	2bb8	71	0.70	0.87	0.69	0.87	0.70 ± 0.18	0.80 ± 0.12	0.70 ± 0.19	0.81 ± 0.12
			80	1bbg	40	0.52	0.67	0.52	0.67				
			120	1iqs	88	0.88	0.87	0.89	0.88				
Nucleotidyltransferase; domain 5	30	420	10	1goa	156	0.58	0.77	0.67	0.76	0.63 ± 0.05	0.76 ± 0.08	0.65 ± 0.02	0.74 ± 0.12
				1ril	147	0.67	0.68	0.64	0.61				
			140	1ovq	138	0.64	0.84	0.64	0.84				
Type Iii Antifreeze Protein Isoform Hplc 12	90	1210	10	1hg7	66	0.74	0.85	0.74	0.86	0.66 ± 0.07	0.78 ± 0.10	0.65 ± 0.08	0.76 ± 0.10
				1ops	64	0.60	0.67	0.60	0.66				
				1ucs	64	0.64	0.82	0.62	0.75				
Mannose-Binding Protein A; Chain A	10	100	10	1dv8	128	0.80	0.82	0.78	0.82	0.65 ± 0.22	0.62 ± 0.29	0.63 ± 0.21	0.60 ± 0.32
				1koe	172	0.49	0.41	0.48	0.37				
Ubiquitin Conjugating Enzyme	10	110	10	1a3s	158	0.79	0.87	0.79	0.87	0.81 ± 0.02	0.86 ± 0.02	0.82 ± 0.04	0.87 ± 0.01
				2ucz	164	0.82	0.84	0.85	0.86				
Flavocytochrome B2; Chain A, domain 1	10	120	10	1b5m	84	0.63	0.66	0.55	0.72	0.67 ± 0.06	0.75 ± 0.13	0.73 ± 0.25	0.84 ± 0.17
				1cyo	88	0.71	0.84	0.91	0.96				
Nuclear Transport Factor 2; Chain: A	10	450	10	1cew	108	0.72	0.63	0.80	0.83	0.65 ± 0.10	0.68 ± 0.06	0.69 ± 0.16	0.78 ± 0.08
				1cyv	98	0.58	0.72	0.57	0.72				
60s Ribosomal Protein L30; Chain: A	30	1330	30	1ck2	104	0.46	0.78	0.45	0.79	0.55 ± 0.13	0.84 ± 0.08	0.55 ± 0.14	0.84 ± 0.07
				1go1	102	0.64	0.89	0.65	0.89				
Protein-Tyrosine Phosphatase; Chain A	90	190	10	1jln	297	0.65	0.79	0.64	0.79	0.55 ± 0.14	0.63 ± 0.23	0.54 ± 0.15	0.63 ± 0.23
				1m3g	145	0.45	0.46	0.43	0.47				
P-30 Protein	10	130	10	1a5p	124	0.70	0.72	0.71	0.61				
Mlu1-box Binding Protein; DNA-binding Domain	10	260	10	1bm8	99	0.45	0.60	0.48	0.56				
Trypsin Inhibitor V; Chain A	30	10	10	1mit	69	0.49	0.39	0.48	0.38				
Phenylalanyl-tRNA Synthetase; Chain B, domain 1	30	56	30	1kvv	104	0.64	0.85	0.55	0.78				

Wheat Germ Agglutinin (Isolectin 2); domain 1	30	60	30	1hpt	56	0.80	0.96	0.80	0.96
Viral Topoisomerase I	30	66	10	1vcc	77	0.64	0.80	0.63	0.56
Phosphorylase Kinase; domain 1	30	200	20	1g8a	227	0.62	0.91	0.62	0.91
Rec A Protein; domain 2	30	250	10	1aa3	63	0.60	0.80	0.61	0.87
Barnase; Chain D	30	370	10	1bta	89	0.38	0.16	0.44	0.15
Potassium Channel Kv1.1; Chain A	30	710	10	1cs3	116	0.74	0.82	0.70	0.79
Metal Transport, Frataxin; Chain A	30	920	10	1ew4	106	0.45	0.31	0.59	0.54
Carboxypeptidase Inhibitor; Chain A	30	1040	10	1dtv	67	0.59	0.75	0.59	0.75
Nonspecific Lipid-transfer Protein; Chain A	30	1050	10	1c44	123	0.63	0.80	0.64	0.80
Conserved Hypothetical Protein Mth637; Chain: A	30	1200	10	1jrm	104	0.44	0.52	0.44	0.57
Histidine-containing Protein; Chain: A	30	1340	10	1ptf	87	0.35	-0.02	0.38	0.00
Cell Division Protein Zipa; Chain: A	30	1400	10	1f7w	144	0.61	0.88	0.63	0.88
Lysozyme-like	40	80	10	1j3g	187	0.52	0.74	0.49	0.58
Oxidized Rhodanese; domain 1	40	250	10	1c25	161	0.47	0.52	0.41	0.56
Uracil-DNA Glycosylase, subunit E	40	470	10	1udg	228	0.53	0.20	0.45	0.41
Replication Protein E1; Chain: A	40	1310	20	1l2m	118	0.74	0.81	0.74	0.81
Nuia	40	1460	10	1j57	143	0.71	0.85	0.75	0.85
Hepatocyte Growth Factor	50	4	10	2hgf	97	0.63	0.79	0.64	0.65
GroEL	50	7	10	1srv	145	0.38	0.46	0.36	0.47
Proliferating Cell Nuclear Antigen	70	10	10	1plr	258	0.71	0.71	0.53	0.47
Phenol Hydroxylase P2 Protein	90	56	10	1g10	102	0.53	0.51	0.53	0.53
Phosphatidylethanolamine-binding Protein	90	280	10	1a44	185	0.50	0.69	0.42	0.62
Nucleotide Excision Repair Protein XPA (XPA-MBD); B Chain A	90	530	10	1xpa	113	0.82	0.87	0.70	0.78
Sugar Binding Protein, Amyloid A4 Protein; Chain A	90	570	10	1mwp	96	0.77	0.92	0.65	0.91
Endoglucanase; Chain: A	90	1220	10	1e8p	46	0.64	0.69	0.63	0.78
Class 4: few secondary structure									
Omega-AgatoxinV	10	40	10	1omb	35	0.70	0.64	0.70	0.64

Low-density Lipoprotein Receptor	10	400	10	1j8e	44	0.82	0.83	0.83	0.85
Factor Xa Inhibitor	10	410	10	1dem	60	0.72	0.86	0.73	0.84
H-NS DNA Binding Protein	10	430	10	1hnr	47	0.50	0.49	0.49	0.45
Virus Scaffolding Protein; Chain A	10	810	10	2gp8	40	0.62	0.46	0.50	0.32

a) Fold family (Topology) as classified by CATH.[201] b) CATH protein structure classification,[201] A: architecture, T: topology, H: homologous superfamily. c) Protein size in number of amino acids. d) Elastic Network Model. e) Rigid Cluster Normal Mode Analysis.[68] f) Maximal overlaps as calculated by Eq. 4.1. g) Maximal correlations in magnitudes of modes as calculated by Eq. 4.2. The three pairs of proteins used in Table 5.3 (see section 5.1.4) are highlighted in bold.

Bibliography

1. Karplus M, McCammon JA. Dynamics of proteins: Elements and function. Annu. Rev. Biochem. 1983;52:263-300.
2. Haurowitz F. Das gleichgewicht zwischen hämoglobin und sauerstoff. Z. Physiol. Chem. 1938;254:266-274.
3. Quiocho FA, Ledvina PS. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: Variation of common themes. Mol. Microbiol. 1996;20(1):17-25.
4. Perutz MF, Wilkinson AJ, Paoli M, Dodson GG. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. Annu. Rev. Biophys. Biomol. Struct. 1998;27:1-34.
5. Hammes GG. Multiple conformational changes in enzyme catalysis. Biochemistry (Mosc.) 2002;41(26):8221-8228.
6. Benkovic SJ, Hammes-Schiffer S. A perspective on enzyme catalysis. Science 2003;301(5637):1196-1202.
7. Huse M, Kuriyan J. The conformational plasticity of protein kinases. Cell 2002;2002:275-282.
8. Doyle DA. Structural changes during ion channel gating. Trends Neurosci. 2004;27(6):298-302.
9. Swartz KJ. Towards a structural view of gating in potassium channels. Nat. Rev. Neurosci. 2004;5(12):905-916.
10. Karplus M, Gao YQ. Biomolecular motors: The F1-ATPase paradigm. Curr. Opin. Struct. Biol. 2004;14(2):250-259.
11. Vallee RB, Hook P. Molecular motors: A magnificent machine. Nature 2003;421(6924):701-702.
12. Wlodawer A, Vondrasek J. Inhibitors of hiv-1 protease: A major success of structure-assisted drug design. Annu. Rev. Biophys. Biomol. Struct. 1998;27:249-284.
13. Wilson DK, Tarle I, Petrash JM, Quiocho FA. Refined 1.8 Å structure of human aldose reductase complexed with the potent inhibitor zopolrestat. Proc. Natl. Acad. Sci. U. S. A. 1993;90(21):9847-9851.
14. Schlauderer GJ, Schulz GE. The structure of bovine mitochondrial adenylate kinase: Comparison with isoenzymes in other compartments. Protein Sci. 1996;5(3):434-441.
15. Schlauderer GJ, Proba K, Schulz GE. Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP. J. Mol. Biol. 1996;256(2):223-227.
16. Muller CW, Schlauderer GJ, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding. Structure 1996;4(2):147-156.
17. Stuckey J, Schubert H, Fauman E, Zhang Z-Y, Dixon J, Saper M. Crystal structure of yersinia protein tyrosine phosphatase at 2.5 Å and the complex with tungstate. Nature 1994;370(6490):571-575.

18. Schubert HL, Fauman EB, Stuckey JA, Dixon JE, Saper MA. A ligand-induced conformational change in the yersinia protein-tyrosine-phosphatase. Protein Sci. 1995;4(9):1904-1913.
19. Vandonselaar M, Hickie RA, Quail JW, Delbaere LTJ. Trifluoperazine-induced conformational change in Ca^{2+}-calmodulin. Nat. Struct. Biol. 1994;1(795-801).
20. Kuboniwa H, Tjandra N, Grzesiek S, Ren H, Klee CB, Bax A. Solution structure of calcium-free calmodulin. Nat. Struct. Biol. 1995;2(9):768-776.
21. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein-protein interactions. Curr. Opin. Struct. Biol. 2004;14(1):104-109.
22. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52(1):88-91.
23. Fischer E. Einfluss der konfiguration auf die wirkung der enzyme. Ber. Dtsch. Chem. Ges. 1894;27(3):2985-2993.
24. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. U. S. A. 1958;44:98-104.
25. Bosshard HR. Molecular recognition by induced fit: How fit is the concept? News Physiol Sci 2001;16:171-173.
26. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Protein Eng. 1999;12(9):713-720.
27. Tsai C-J, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Prot Sci 1999;8:1181-1190.
28. Tsai C-J, Ma B, Nussinov R. Folding and binding cascades: Shifts in energy landscapes. Proc Natl Acad Sci USA 1999;96:9970-9972.
29. Berger C, Weber-Bornhauser S, Eggenberger J, Hanes J, Pluckthun A, Bosshard HR. Antigen recognition by conformational selection. FEBS Lett. 1999;450(1-2):149-153.
30. Volkman BF, Lipson D, Wemmer DE, Kern D. Two-state allosteric behavior in a single-domain signaling protein. Science 2001;291(5512):2429-2433.
31. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D. Intrinsic dynamics of an enzyme underlies catalysis. Nature 2005;438(7064):117-121.
32. Ma B, Shatsky M, Wolfson HJ, Nussinov R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. Protein Sci. 2002;11(2):184-197.
33. Bahar I, Chennubhotla C, Tobi D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. Curr. Opin. Struct. Biol. 2007;17(6):633-640.
34. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc. Natl. Acad. Sci. U. S. A. 2005.
35. James LC, Tawfik DS. Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. Proc. Natl. Acad. Sci. U. S. A. 2005;102(36):12730-12735.
36. Hornak V, Okur A, Rizzo RC, Simmerling C. Hiv-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. J. Am. Chem. Soc. 2006;128(9):2812-2813.

37. Okazaki KI, Takada S. Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. Proc. Natl. Acad. Sci. U. S. A. 2008;105(32):11182-11187.
38. Carlson HA. Protein flexibility is an important component of structure-based drug discovery. Curr. Pharm. Des. 2002;8(17):1571-1578.
39. Ahmed A, Kazemi S, Gohlke H. Protein flexibility and mobility in structure-based drug design. Front. Drug Des. Discov. 2007;3:455-476.
40. Totrov M, Abagyan R. Flexible ligand docking to multiple receptor conformations: A practical alternative. Curr. Opin. Struct. Biol. 2008;18(2):178-184.
41. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 1982;161:269-288.
42. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr. Opin. Struct. Biol. 1996;6(3):402-406.
43. Davis AM, Teague SJ. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. Angew. Chem. Int. Ed. Engl. 1999;38(6):736-749.
44. Carlson HA, McCammon JA. Accommodating protein flexibility in computational drug design. Mol. Pharmacol. 2000;57(2):213-218.
45. Bonvin A. Flexible protein-protein docking. Curr. Opin. Struct. Biol. 2006;16(2):194-200.
46. Ehrlich LP, Nilges M, Wade RC. The impact of protein flexibility on protein-protein docking. Proteins 2005;58(1):126-133.
47. Knegtel RM, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. J. Mol. Biol. 1997;266(2):424-440.
48. Broughton HB. A method for including protein flexibility in protein-ligand docking: Improving tools for database mining and virtual screening. J. Mol. Graph. Model. 2000;18(3):247-257, 302-244.
49. Lensink MF, Mendez R. Recognition-induced conformational changes in protein-protein docking. Curr. Pharm. Biotechnol. 2008;9(2):77-86.
50. Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, Lins RD, Briggs JM, McCammon JA. Developing a dynamic pharmacophore model for hiv-1 integrase. J. Med. Chem. 2000;43(11):2100-2114.
51. Gohlke H, Thorpe M. A natural coarse graining for simulating large biomolecular motion. Biophys. J. 2006.
52. Damm KL, Carlson HA. Exploring experimental sources of multiple protein conformations in structure-based drug design. J. Am. Chem. Soc. 2007;129(26):8225-8235.
53. Furnham N, Blundell T, Depristo M, Terwilliger T. Is one solution good enough? Nat. Struct. Mol. Biol.;13(3):184-185.
54. Wall ME, Gallagher SC, Trewhella J. Large-scale shape changes in proteins and macromolecular complexes. Annu. Rev. Phys. Chem. 2000;51:355-380.
55. Kay LE. Protein dynamics from NMR. Nat. Struct. Biol. 1998;5 Suppl:513-517.
56. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature 1977;267(5612):585-590.
57. Hansson T, Oostenbrink C, van Gunsteren W. Molecular dynamics simulations. Curr. Opin. Struct. Biol. 2002;12(2):190-196.

58. Adcock SA, McCammon JA. Molecular dynamics: Survey of methods for simulating the activity of proteins. Chem. Rev. 2006;106(5):1589-1615.
59. Cheatham TE, Young MA. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. Biopolymers 2000;56(4):232-256.
60. Moraitakis G, Purkiss A, Goodfellow J. Simulated dynamics and biological macromolecules. Rep. Prog. Phys. 2003;66(3):383-406.
61. Cavasotto C, Singh N. Docking and high throughput docking: Successes and the challenge of protein flexibility. Curr. Comput.-Aided Drug Des. 2008;4(3):221-234.
62. de Groot BL, van Aalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC. Prediction of protein conformational freedom from distance constraints. Proteins: Struct., Funct., Genet. 1997;29(2):240-251.
63. Seeliger D, Haas J, de Groot B. Geometry-based sampling of conformational transitions in proteins. Structure 2007;15(11):1482-1492.
64. Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. Phys. Biol. 2005;2(4).
65. Mustard D, Ritchie D. Docking essential dynamics eigenstructures. Proteins: Struct., Funct., Bioinf. 2005;60(2):269-274.
66. Jolley CC, Wells SA, Hespenheide BM, Thorpe MF, Fromme P. Docking of photosystem I subunit C using a constrained geometric simulation. J. Am. Chem. Soc. 2006;128(27):8803-8812.
67. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. 2001;80(1):505-515.
68. Ahmed A, Gohlke H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins: Struct., Funct., Genet. 2006;63(4):1038-1051.
69. Trakhanov S, Vyas NK, Luecke H, Kristensen DM, Ma J, Quiocho FA. Ligand-free and -bound structures of the binding protein (livj) of the escherichia coli abc leucine/isoleucine/valine transport system: Trajectory and dynamics of the interdomain rotation and ligand specificity. Biochemistry (Mosc.) 2005;44(17):6597-6608.
70. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng. 2001;14(1):1-6.
71. Yang L, Song G, Jernigan RL. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys. J. 2007.
72. Zhang Z, Shi Y, Liu H. Molecular dynamics simulations of peptides and proteins with amplified collective motions. Biophys. J. 2003;84(6):3583-3593.
73. Tatsumi R, Fukunishi Y, Nakamura H. A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. J. Comput. Chem. 2004;25(16):1995-2005.
74. He J, Zhang Z, Shi Y, Liu H. Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. J. Chem. Phys. 2003;119(7):4005-4017.
75. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. J. Am. Chem. Soc. 2005;127(26):9632-9640.

76. May A, Zacharias M. Protein-protein docking in capri using attract to account for global and local flexibility. Proteins: Struct., Funct., Bioinf. 2007;69(4):774-780.
77. May A, Zacharias M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: Evaluation on kinase inhibitor cross docking. J. Med. Chem. 2008;51(12):3499-3506.
78. Delarue M, Dumas P. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. Proc. Natl. Acad. Sci. U. S. A. 2004;101(18):6957-6962.
79. Hinsen K, Reuter N, Navaza J, Stokes DL, Lacapere JJ. Normal mode-based fitting of atomic structure into electron density maps: Application to sarcoplasmic reticulum Ca-ATPase. Biophys. J. 2005;88(2):818-827.
80. Tama F, Miyashita O, Brooks CL. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J. Mol. Biol. 2004;337(4):985-999.
81. Tama F, Miyashita O, Brooks CL. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-em. J. Struct. Biol. 2004;147(3):315-326.
82. Kim MK, Jernigan RL, Chirikjian GS. Efficient generation of feasible pathways for protein conformational transitions. Biophys. J. 2002;83(3):1620-1630.
83. Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc. Natl. Acad. Sci. U. S. A. 2003;100(22):12570-12575.
84. Miyashita O, Wolynes PG, Onuchic JN. Simple energy landscape model for the kinetics of functional transitions in proteins. J. Phys. Chem. B 2005;109:1959-1969.
85. Rueda M, Chacon P, Orozco M. Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. Structure 2007;15(5):565-575.
86. Rueda M, Ferrer-Costa C, Meyer T, Perez A, Camps J, Hospital A, Gelpi JL, Orozco M. A consensus view of protein dynamics. Proc. Natl. Acad. Sci. U. S. A. 2007;104(3):796-801.
87. Maragakis P, Karplus M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. J. Mol. Biol. 2005;352(4):807-822.
88. Kirillova S, Cortes J, Stefaniu A, Simeon T. An nma-guided path planning approach for computing large-amplitude conformational changes in proteins. Proteins 2008;70(1):131-143.
89. Whitford PC, Gosavi S, Onuchic JN. Conformational transitions in adenylate kinase - allosteric communication reduces misligation. J. Biol. Chem. 2008;283(4):2042-2048.
90. Brooks B, Karplus M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. U. S. A. 1983;80(21):6571-6575.
91. Jacobs D, Rader AJ, Kuhn L, Thorpe MF. Protein flexibility predictions using graph theory. Proteins: Struct., Funct., Genet. 2001;44(2):150-165.
92. Lei M, Zavodszky M, Kuhn L, Thorpe MF. Sampling protein conformations and pathways. J. Comput. Chem. 2004;25(9):1133-1148.

93. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: The energy landscape perspective. Annu. Rev. Phys. Chem. 1997;48:545-600.
94. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253(5494):694-698.
95. Varney MD, Appelt K, Kalish V, Reddy MR, Tatlock J, Palmer CL, Romines WH, Wu BW, Musick L. Crystal-structure-based design and synthesis of novel c-terminal inhibitors of hiv protease. J. Med. Chem. 1994;37(15):2274-2284.
96. Alonso Hn, Bliznyuk AA, Gready JE. Combining docking and molecular dynamic simulations in drug design. Med. Res. Rev. 2006.
97. Perryman A, Lin J-H, McCammon A. Restrained molecular dynamics simulations of hiv-1 protease: The first step in validating a new target for drug design. Biopolymers 2006;82(3):272-284.
98. Elcock AH, Sept D, McCammon JA. Computer simulation of protein-protein interactions. J. Phys. Chem. B 2001;105:1504-1518.
99. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J. Mol. Biol. 1999;293(5):1161-1181.
100. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA. Discovery of a novel binding trench in hiv integrase. J. Med. Chem. 2004;47(8):1879-1881.
101. Hazuda D, Anthony N, Gomez R, Jolly S, Wai J, Zhuang L, Fisher T, Embrey M, Guare J, Egbertson M, Vacca J, Huff J, Felock P, Witmer M, Stillmock K, Danovich R, Grobler J, Miller M, Espeseth A, Jin L, Chen IW, Lin J, Kassahun K, Ellis J, Wong B, Xu W, Pearson P, Schleif W, Cortese R, Emini E, Summa V, Holloway K, Young S. From the cover: A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. PNAS 2004;101(31):11233-11238.
102. Hornak V, Okur A, Rizzo RC, Simmerling C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. Proc. Natl. Acad. Sci. U. S. A. 2006;103(4):915-920.
103. Sotriffer CA, Kramer O, Klebe G. Probing flexibility and "Induced-fit" Phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. Proteins: Struct., Funct., Bioinf. 2004;56(1):52-66.
104. Smith G, Sternberg M, Bates P. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J. Mol. Biol. 2005;347(5):1077-1101.
105. Sherwood P, Brooks B, Sansom M. Multiscale methods for macromolecular simulations. Curr. Opin. Struct. Biol. 2008;18(5):630-640.
106. Tozzini V. Coarse-grained models for proteins. Curr. Opin. Struct. Biol. 2005;15(2):144-150.
107. Trylska J, Tozzini V, McCammon JA. Exploring global motions and correlations in the ribosome. Biophys. J. 2005;89(3):1455-1463.
108. Berne BJ, Straub JE. Novel methods of sampling phase space in the simulation of biological systems. Curr. Opin. Struct. Biol. 1997;7(2):181-189.
109. Lei H, Duan Y. Improved sampling methods for molecular simulation. Curr. Opin. Struct. Biol. 2007;17(2):187-191.

110. Grubmuller H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. Phys. Rev. E 1995;52(3):2893-2906.
111. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 1999;314(1-2):141-151.
112. Zhou R. Replica exchange molecular dynamics method for protein folding simulation. Methods Mol. Biol. 2007;350:205-223.
113. van der Vaart A, Karplus M. Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. J. Chem. Phys. 2005;122(11):114903.
114. Schlitter J, Engels M, Kruger P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. J. Mol. Graph. 1994;12(2):84-89.
115. Case D. Normal mode analysis of protein dynamics. Curr. Opin. Struct. Biol. 1994;4(2):285-290.
116. Zheng W, Doniach S. A comparative study of motor-protein motions by using a simple elastic-network model. Proc. Natl. Acad. Sci. U. S. A. 2003;100(23):13253-13258.
117. Hayward S, Kitao A, Berendsen HJ. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins 1997;27(3):425-437.
118. Gibrat JF, Go N. Normal mode analysis of human lysozyme: Study of the relative motion of the two domains and characterization of the harmonic motion. Proteins 1990;8(3):258-279.
119. Marques O, Sanejouand YH. Hinge-bending motion in citrate synthase arising from normal mode calculations. Proteins 1995;23(4):557-560.
120. Ma J, Karplus M. The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. Proc. Natl. Acad. Sci. U. S. A. 1998;95(15):8502-8507.
121. Thomas A, Field MJ, Mouawad L, Perahia D. Analysis of the low frequency normal modes of the t-state of aspartate transcarbamylase. J. Mol. Biol. 1996;257(5):1070-1087.
122. Thomas A, Field MJ, Perahia D. Analysis of the low-frequency normal modes of the r state of aspartate transcarbamylase and a comparison with the t state modes. J. Mol. Biol. 1996;261(3):490-506.
123. Thomas A, Hinsen K, Field MJ, Perahia D. Tertiary and quaternary conformational changes in aspartate transcarbamylase: A normal mode study. Proteins 1999;34(1):96-112.
124. Zhang G, Schlick T. The Langevin/implicit-Euler/normal-mode scheme for molecular dynamics at large time steps. J. Chem. Phys. 1994;101(6):4995-5012.
125. Kidera A, Inaka K, Matsushima M, Go N. Normal mode refinement: Crystallographic refinement of protein dynamic structure. II. Application to human lysozyme. J. Mol. Biol. 1992;225(2):477-486.
126. Brüschweiler R, Case DA. Collective nmr relaxation model applied to protein dynamics. Phys. Rev. Lett. 1994;72(6):940.
127. Duong TH, Zakrzewska K. Calculation and analysis of low frequency normal modes for DNA. J. Comput. Chem. 1997;18(6):796-811.
128. Matsumoto A, Go N. Dynamic properties of double-stranded DNA by normal mode analysis. J. Chem. Phys. 1999;110(22):11070-11075.

129. Matsumoto A, Tomimoto M, Go N. Dynamical structure of transfer rna studied by normal mode analysis. Eur. Biophys. J. 1999;28(5):369-379.

130. Brooks B, Karplus M. Normal modes for specific motions of macromolecules: Application to the hinge-bending mode of lysozyme. Proc. Natl. Acad. Sci. U. S. A. 1985;82(15):4995-4999.

131. Li G, Cui Q. Analysis of functional motions in brownian molecular machines with an efficient block normal mode approach: Myosin-ii and Ca^{2+}-ATPase. Biophys. J. 2004;86(2):743-763.

132. Case DA. Normal mode analysis of protein dynamics. Curr. Opin. Struct. Biol. 1994;4:285-290.

133. Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, Gunsalus IC. Dynamics of ligand binding to myoglobin. Biochemistry (Mosc.) 1975;14(24):5355-5373.

134. Hayward S, Go N. Collective variable description of native protein dynamics. Annu. Rev. Phys. Chem. 1995;46:223-250.

135. Hayward S, Kitao A, Go N. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. Proteins 1995;23(2):177-186.

136. Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 2005;13(3):373-380.

137. Tirion M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 1996;77(9):1905.

138. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding & Design 1997;2(3):173-181.

139. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. Phys. Rev. Lett. 1997;79(16):3090.

140. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins: Struct., Funct., Genet. 1999;33(3):417-429.

141. Hinsen K, Thomas A, Field MJ. Analysis of domain motions in large proteins. Proteins 1999;34(3):369-382.

142. Tama F, Gadea FX, Marques O, Sanejouand YH. Building-block approach for determining low-frequency normal modes of macromolecules. Proteins 2000;41(1):1-7.

143. Doruker P, Jernigan RL, Bahar I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. J. Comput. Chem. 2002;23(1):119-127.

144. Kurkcuoglu O, Jernigan R, Doruker P. Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions. Polymer 2004;45(2):649-657.

145. Tama F, Feig M, Liu J, Brooks CL, Taylor KA. The requirement for mechanical coupling between head and s2 domains in smooth muscle myosin ATPase regulation and its implications for dimeric motor function. J. Mol. Biol. 2005;345(4):837-854.

146. Tama F, Valle M, Frank J, Brooks CL, 3rd. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. Proc. Natl. Acad. Sci. U. S. A. 2003;100(16):9319-9323.

147. Van Wynsberghe A, Li G, Cui Q. Normal-mode analysis suggests protein flexibility modulation throughout rna polymerase's functional cycle. Biochemistry (Mosc.) 2004;43(41):13083-13096.

148. Zheng W, Brooks B. Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. J. Mol. Biol. 2005;346(3):745-759.
149. Ma J. New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. Curr. Protein Pept. Sci. 2004;5(2):119-123.
150. Bahar I, Rader A. Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol. 2005;15(5):586-592.
151. Bahar I, Erman B, Jernigan R, Atilgan A, Covell D. Collective motions in HIV-1 reverse transcriptase: Examination of flexibility and enzyme function. J. Mol. Biol. 1999;285(3):1023-1037.
152. Temiz NA, Meirovitch E, Bahar I. Escherichia coli adenylate kinase dynamics: Comparison of elastic network model modes with mode-coupling (15)n-nmr relaxation data. Proteins 2004;57(3):468-480.
153. Isin B, Doruker P, Bahar I. Functional motions of influenza virus hemagglutinin: A structure-based analytical approach. Biophys. J. 2002;82(2):569-581.
154. Bahar I, Jernigan RL. Cooperative fluctuations and subunit communication in tryptophan synthase. Biochemistry (Mosc.) 1999;38(12):3478-3490.
155. Temiz NA, Bahar I. Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase. Proteins 2002;49:61-70.
156. Kim MK, Chirikjian GS, Jernigan RL. Elastic models of conformational transitions in macromolecules. J. Mol. Graph. Model. 2002;21(2):151-160.
157. Ertekin A, Nussinov R, Haliloglu T. Association of putative concave protein-binding sites with the fluctuation behavior of residues. Protein Sci. 2006;15(10):2265-2277.
158. Yang LW, Bahar I. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. Structure (Camb) 2005;13(6):893-904.
159. Haliloglu T, Keskin O, Ma B, Nussinov R. How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. Biophys. J. 2005;88(3):1552-1559.
160. Demirel MC, Atilgan AR, Jernigan RL, Erman B, Bahar I. Identification of kinetically hot residues in proteins. Protein Sci. 1998;7(12):2522-2532.
161. Jacobs DJ, Thorpe MF. Generic rigidity percolation: The pebble game. Phys. Rev. Lett. 1995;75:4051-4054.
162. Brejc K, van Dijk WJ, Klaasen RV, Schuurmans M, van der Oost J, Smit AB, Sixma TK. Crystal structure of an ach-binding protein reveals the ligand-binding domain of nicotinic receptors. Nature 2001;411:269-276.
163. Jacobs DJ, Kuhn LA, Thorpe MF. Flexible and rigid regions in proteins. Rigidity theory and applications. New York: Kluwer Academic/Plenum 1999;7:961-967.
164. Gohlke H, Kuhn LA, Case DA. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins 2004;56:322-337.
165. Fuxreiter M, Magyar C, Juhasz T, Szeltner Z, Polgar L, Simon I. Flexibility of prolyl oligopeptidase: Molecular dynamics and molecular framework analysis of the potential substrate pathways. Proteins: Struct., Funct., Bioinf. 2005;60(3):504-512.

166. Fulle S, Gohlke H. Analyzing the flexibility of rna structures by constraint counting. Biophys. J. 2008;94(11):4202-4219.
167. Fulle S, Gohlke H. Statics of the ribosomal exit tunnel: Implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. J. Mol. Biol. 2009;387(2):502-517.
168. Thorpe MF, Lei M, Rader AJ, Jacobs DJ, Kuhn LA. Protein flexibility and dynamics using constraint theory. J. Mol. Graph. Model. 2001;19(1):60-69.
169. Zavodszky M, Lei M, Thorpe MF, Day A, Kuhn L. Modeling correlated main-chain motions in proteins for flexible molecular recognition. Proteins: Struct., Funct., Bioinf. 2004;57(2):243-261.
170. Jolley C, Wells S, Fromme P, Thorpe MF. Fitting low-resolution cryo-em maps of proteins using constrained geometric simulations. Biophys. J. 2008;94(5):1613-1621.
171. de Groot BL, Vriend G, Berendsen HJ. Conformational changes in the chaperonin groel: New insights into the allosteric mechanism. J. Mol. Biol. 1999;286(4):1241-1249.
172. Jedrzejas M, Mello L, de Groot B, Li S. Mechanism of hyaluronan degradation by streptococcus pneumoniae hyaluronate lyase. Structures of complexes with the substrate. J. Biol. Chem. 2002;277(31):28287-28297.
173. Labrou NE, Mello LV, Clonis YD. Functional and structural roles of the glutathione-binding residues in maize (zea mays) glutathione s-transferase i. Biochem. J. 2001;358(Pt 1):101-110.
174. Melo F, Rigden D, Franco Ov, Mello L, Ary M, Maria, Bloch C. Inhibition of trypsin by cowpea thionin: Characterization, molecular modeling, and docking. Proteins: Struct., Funct., Genet. 2002;48(2):311-319.
175. Rigden D, Lamani E, Mello L, Littlejohn J, Jedrzejas M. Insights into the catalytic mechanism of cofactor-independent phosphoglycerate mutase from x-ray crystallography, simulated dynamics and molecular modeling. J. Mol. Biol. 2003;328(4):909-920.
176. Lu M, Ma J. The role of shape in determining molecular motions. Biophys. J. 2005;89(4):2395-2401.
177. Nicolay S, Sanejouand YH. Functional modes of proteins are among the most robust. Phys. Rev. Lett. 2006;96(7).
178. Durand P, Trinquier G, Sanejouand Y-H. A new approach for determining low-frequency normal modes in macromolecules. Biopolymers 1994;34:759-771.
179. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins: Struct., Funct., Genet. 2000;40(3):389-408.
180. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins 1998;33(3):417-429.
181. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 1999;285(4):1735-1747.
182. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst Sect A 1991;47:392-400.
183. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 1963;7(1):95-59.
184. Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide and protein chain conformations .2. Allowed conformations for a pair of peptide units. Biophys. J. 1965;5(6):909-933.

185. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: Standard radii and volumes. J. Mol. Biol. 1999;290(1):253-266.
186. Ho BK, Thomas A, Brasseur R. Revisiting the ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci. 2003;12(11):2508-2522.
187. Morris AL, Macarthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein-structure coordinates. Proteins 1992;12(4):345-364.
188. Bhat TN, Sasisekharan V, Vijayan M. Analysis of side-chain conformation in proteins. Int. J. Pept. Protein Res. 1979;13(2):170-184.
189. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 1987;193(4):775-791.
190. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997;6(8):1661-1681.
191. Dunbrack RL. Rotamer libraries in the 21(st) century. Curr. Opin. Struct. Biol. 2002;12(4):431-440.
192. Schrauber H, Eisenhaber F, Argos P. Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. J. Mol. Biol. 1993;230(2):592-612.
193. Hooft RWW, Sander C, Vriend G. Verification of protein structures: Side-chain planarity. J. Appl. Crystallogr. 1996;29:714-716.
194. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck - a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 1993;26:283-291.
195. MacArthur MW, Thornton JM. Deviations from planarity of the peptide bond in peptides and proteins. J. Mol. Biol. 1996;264(5):1180-1195.
196. Merritt EA, Kuhn P, Sarfaty S, Erbe JL, Holmes RK, Ho WGJ. The 1.25 angstrom resolution refinement of the cholera toxin b-pentamer: Evidence of peptide backbone strain at the receptor-binding site. J. Mol. Biol. 1998;282(5):1043-1059.
197. Olah GA, Trakhanov S, Trewhella J, Quiocho FA. Leucine isoleucine valine-binding protein contracts upon binding of ligand. J. Biol. Chem. 1993;268(22):16241-16247.
198. Egea PF, Rochel N, Birck C, Vachette P, Timmins PA, Moras D. Effects of ligand binding on the association properties and conformation in solution of retinoic acid receptors rxr and rar. J. Mol. Biol. 2001;307(2):557-576.
199. Sinev MA, Razgulyaev OI, Vas M, Timchenko AA, Ptitsyn OB. Correlation between enzyme-activity and hinge-bending domain displacement in 3-phosphoglycerate kinase. Eur. J. Biochem. 1989;180(1):61-66.
200. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins: Struct., Funct., Bioinf. 2006;65(3):712-725.
201. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. Cath - a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.
202. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. Protein data bank - computer-based archival file for macromolecular structures. J. Mol. Biol. 1977;112(3):535-542.

203. Hamelryck T, Manderick B. Pdb file parser and structure class implemented in python. Bioinformatics 2003;19(17):2308-2310.

204. Jacobs DJ, Kuhn LA, Thorpe MF. Flexible and rigid regions in proteins. In: Thorpe MF, Duxbury PM, editors. Rigidity theory and applications. New York: Kluwer Academic/Plenum Publishers; 1999. p 357-384.

205. Li G, Cui Q. A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca^{2+}-ATPase. Biophys. J. 2002;83(5):2457-2474.

206. Bruschweiler R. Collective protein dynamics and nuclear-spin relaxation. J. Chem. Phys. 1995;102:3396-3403.

207. Koller AN, Schwalbe H, Gohlke H. Starting structure dependence of nmr order parameters derived from md simulations: Implications for judging force-field quality. Biophys. J. 2008;95(1):L4-L6.

208. Wilson KP, Malcolm BA, Matthews BW. Structural and thermodynamic analysis of compensating mutations within the core of chicken egg-white lysozyme. J. Biol. Chem. 1992;267(15):10842-10849.

209. Schwalbe H, Grimshaw SB, Spencer A, Buck M, Boyd J, Dobson CM, Redfield C, Smith LJ. A refined solution structure of hen lysozyme determined using residual dipolar coupling data. Protein Sci. 2001;10(4):677-688.

210. Heringa J, Argos P. Strain in protein structures as viewed through nonrotameric side chains: II. Effects upon ligand binding. Proteins 1999;37(1):44-55.

211. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. J. Mol. Biol. 1999;285(4):1711-1733.

212. Kurkcuoglu O, Jernigan RL, Doruker P. Loop motions of triosephosphate isomerase observed with elastic networks. Biochemistry (Mosc.) 2006;45(4):1173-1182.

213. Hayward S. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. J. Mol. Biol. 2004;339(4):1001-1021.

214. Kang CH, Shin WC, Yamagata Y, Gokcen S, Ames GF, Kim SH. Crystal-structure of the lysine-binding, arginine-binding, ornithine-binding protein (lao) from salmonella-typhimurium at 2.7 Å resolution. J. Biol. Chem. 1991;266(35):23893-23899.

215. Amadei A, Linssen AB, Berendsen HJ. Essential dynamics of proteins. Proteins 1993;17(4):412-425.

216. van Aalten DM, Amadei A, Linssen AB, Eijsink VG, Vriend G, Berendsen HJ. The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. Proteins 1995;22(1):45-54.

217. Hayward S, Kitao A, Go N. Harmonic and anharmonic aspects in the dynamics of bpti: A normal mode analysis and principal component analysis. Protein Sci. 1994;3(6):936-943.

218. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. Curr. Opin. Struct. Biol. 1999;9(2):164-169.

219. Dauber-Osguthorpe P. Low frequency motion in proteins comparison of normal mode and molecular dynamics of streptomyces griseus protease A. J. Comput. Phys. 1999;151(1):169-189.

220. McCammon JA, Gelin BR, Karplus M, Wolynes PG. The hinge-bending mode in lysozyme. Nature 1976;262(5566):325-326.

221. Leherte L, Vercauteren DP. Collective motions of rigid fragments in protein structures from smoothed electron density distributions. J. Comput. Chem. 2008;29(9):1472-1489.

222. Song G, Jernigan R. An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins: Struct., Funct., Bioinf. 2006;63(1):197-209.

223. Maguid S, Fernandez-Alberti S, Echave J. Evolutionary conservation of protein vibrational dynamics. Gene 2008;422(1-2):7-13.

224. Maguid S, Fernandez-Alberti S, Ferrelli L, Echave J. Exploring the common dynamics of homologous proteins. Application to the globin family. Biophys. J. 2005;89(1):3-13.

225. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophys. J. 2005;88(2):1291-1299.

226. Tay TS, Whiteley W. Recent advances in generic rigidity of structures. Structural Topology 1984;9:31-38.

227. Chennubhotla C, Rader AJ, Yang LW, Bahar I. Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. Phys. Biol. 2005;2(4):S173-180.

228. Kurkcuoglu O, Robert l, Doruker P. Collective dynamics of large proteins from mixed coarse-grained elastic network model. QSAR Comb. Sci. 2005;24(4):443-448.

229. Berendsen H, Hayward S. Collective protein dynamics in relation to function. Curr. Opin. Struct. Biol. 2000;10(2):165-169.

230. Petrone P, Pande V. Can conformational change be described by only a few normal modes? Biophys. J. 2006;90(5):1583-1593.

231. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins 2002;48(4):682-695.

232. Lobanov M, Bogatyreva N, Galzitskaya O. Radius of gyration as an indicator of protein structure compactness. Mol. Biol. 2008;42(4):623-628.

233. Keskin O, Jernigan RL, Bahar I. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys. J. 2000;78(4):2093-2106.

234. Zheng W, Brooks B, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. PNAS 2006;103(20):7664-7669.

235. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. Science 2009;324(5924):203-207.

236. Blake CCF, Cassels R, Dobson CM, Poulsen FM, Williams RJP, Wilson KS. Structure and binding properties of hen lysozyme modified at tryptophan 62. J. Mol. Biol. 1981;147:73-95.

237. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 1985;181(3):423-447.

238. Bruccoleri R, Karplus M, McCammon A. The hinge-bending mode of a lysozyme-inhibitor complex. Biopolymers 1986;25(9):1767-1802.

239. Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD. Importance of the cmap correction to the charmm22 protein force field: Dynamics of hen lysozyme. Biophys. J. 2006;90(4):L36-L38.

240. Soares TA, Daura X, Oostenbrink C, Smith LJ, van Gunsteren WF. Validation of the gromos force-field parameter set 45a3 against nuclear magnetic resonance data of hen egg lysozyme. J. Biomol. NMR 2004;30(4):407-422.

241. Stocker U, van Gunsteren WF. Molecular dynamics simulation of hen egg white lysozyme: A test of the gromos96 force field against nuclear magnetic resonance data. Proteins: Struct., Funct., Genet. 2000;40(1):145-153.

242. van Aalten DM, Findlay JB, Amadei A, Berendsen HJ. Essential dynamics of the cellular retinol-binding protein evidence for ligand-induced conformational changes. Protein Eng. 1995;8(11):1129-1135.

243. Lu M, Poon B, Ma J. A new method for coarse-grained elastic normal-mode analysis. J. Chem. Theory Comput. 2006;2(3):464-471.

244. Abseher R, Horstink L, Hilbers C, Nilges M. Essential spaces defined by nmr structure ensembles and molecular dynamics simulation show significant overlap. Proteins: Struct., Funct., Genet. 1998;31(4):370-382.

245. Frimurer TM, Peters GH, Iversen LF, Andersen HS, Moller NP, Olsen OH. Ligand-induced conformational changes: Improved predictions of ligand binding conformations and affinities. Biophys. J. 2003;84(4):2273-2281.

246. Kallblad P, Dean PM. Efficient conformational sampling of local side-chain flexibility. J. Mol. Biol. 2003;326(5):1651-1665.

247. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci. 2005;14(5):1328-1339.

248. Carlson HA. Protein flexibility and drug design: How to hit a moving target. Curr. Opin. Chem. Biol. 2002;6:447-452.

249. Wong CF, McCammon JA. Protein flexibility and computer-aided drug design. Annu. Rev. Pharmacol. Toxicol. 2003;43:31-45.

250. Kantarci-Carsibasi N, Haliloglu T, Doruker P. Conformational transition pathways explored by monte carlo simulation integrated with collective modes. Biophys. J. 2008:biophysj.107.128447.

251. Echols N, Milburn D, Gerstein M. Molmovdb: Analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res. 2003;31(1):478-482.

252. Muller CW, Schulz GE. Structure of the complex between adenylate kinase from escherichia-coli and the inhibitor ap5a refined at 1.9 Å resolution - a model for a catalytic transition-state. J. Mol. Biol. 1992;224(1):159-177.

253. Mcphalen CA, Vincent MG, Jansonius JN. X-ray structure refinement and comparison of 3 forms of mitochondrial aspartate-aminotransferase. J. Mol. Biol. 1992;225(2):495-517.

254. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, Hubner CG, Kern D. Intrinsic motions along an enzymatic reaction trajectory. Nature 2007.

255. Keskin O. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: A case study of antibodies. BMC Struct. Biol. 2007;7:31.

256. Tanaka T, Hidaka H. Hydrophobic regions function in calmodulin-enzyme(s) interactions. J. Biol. Chem. 1980;255(23):11078-11080.

257. Van Wynsberghe A, Cui Q. Interpreting correlated motions using normal mode analysis. Structure 2006;14(11):1647-1653.
258. Ma J, Karplus M. Ligand-induced conformational changes in ras p21: A normal mode and energy minimization analysis. J. Mol. Biol. 1997;274(1):114-131.
259. Navizet I, Lavery R, Jernigan R. Myosin flexibility: Structural domains and collective vibrations. Proteins: Struct., Funct., Bioinf. 2004;54(3):384-393.
260. Zheng W, Brooks BR. Normal-modes-based prediction of protein conformational changes guided by distance constraints. Biophys. J. 2005;88(5):3109-3117.
261. Wilkinson DG, Bhatt S, Cook M, Boncinelli E, Krumlauf R. Segmental expression of Hox-2 homeobox-containing genes in the developing mouse hindbrain. Nature 1989;341(6241):405-409.
262. Williams JC, McDermott AE. Dynamics of the flexible loop of triosephosphate isomerase: The loop motion is not ligand gated. Biochemistry (Mosc.) 1995;34(26):8309-8319.
263. Derreumaux P, Schlick T. The loop opening/closing motion of the enzyme triosephosphate isomerase. Biophys. J. 1998;74(1):72-81.
264. Narayana N, Cox S, Shaltiel S, Taylor SS, Xuong N. Crystal structure of a polyhistidine-tagged recombinant catalytic subunit of cAMP-dependent protein kinase complexed with the peptide inhibitor PKI(5-24) and adenosine. Biochemistry (Mosc.) 1997;36(15):4438-4448.
265. Madhusudan, Trafny EA, Xuong NH, Adams JA, Ten Eyck LF, Taylor SS, Sowadski JM. cAMP-dependent protein kinase: Crystallographic insights into substrate recognition and phosphotransfer. Protein Sci. 1994;3(2):176-187.
266. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr. Opin. Struct. Biol. 2003;13(6):748-757.
267. Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically driven protein allostery. Nat. Struct. Mol. Biol. 2006.
268. Kazemi S, Krueger DM, Sirockin F, Gohlke H. Elastic potential grids: Accurate and efficient representation of intermolecular interactions for fully-flexible docking. Chemmedchem 2009;in press.

Curriculum vitae

Aqeel Ahmed

Schloßstraße 100,
60486 Frankfurt am Main (Germany)
Email: Aqeel@bioinformatik.uni-frankfurt.de
Date of Birth: 2nd June, 1979.

Education:

Sep 2006 – till date	**PhD student,** Goethe University, Frankfurt, Germany. *“Development of a normal mode-based geometric simulation approach for investigating the intrinsic mobility of proteins”* Supervisor: Prof. Dr. Holger Gohlke
Sep 2004 - April 2005	**Master thesis,** Goethe University, Frankfurt, Germany. *“RCNMA for multi-scale modeling of macromolecular conformational changes”*
Aug 2003 - Aug 2004	**Master of Science,** Chalmers University of Technology, Sweden. *“Complex Adaptive Systems”*
Jan 1999 - Jan 2003	**Bachelor of Science,** University of Karachi, Pakistan. *“Computer Science*”

Method developments:

NMSim (Normal Mode Simulation)	Normal mode-based geometric simulation for efficient exploration of low-energy conformational space.
RCNMA (Rigid Cluster Normal mode Analysis)	Protein mobility prediction using rigid cluster normal mode analysis. *(Proteins 2006, 63, 1038-1051).*

Publications:

Ahmed, A., Gohlke, H. *Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory.* Proteins 2006, 63, 1038-1051.

Ahmed, A., Kazemi, S., Gohlke, H. *Protein flexibility and mobility in structure-based drug design.* Front. Drug Des. Discov. 2007, 3, 455-476.

Ahmed, A., Villinger, S., Gohlke, H. *Large-scale comparison of protein essential dynamics from molecular dynamics simulations and elastic networks models.* To be submitted.

Ahmed, A., Gohlke, H*. Normal mode-based geometric simulation for efficient exploration of low-energy conformational space of proteins.* To be submitted.

Oral/poster presentations:

Poster presentation at the 4. German Conference on Chemoinformatics, organized by Gesellschaft Deutscher Chemiker (GDCH), Goslar, Germany, in Nov. 2008.

Poster presentation at the MMS07 (Methods of Molecular Simulation), organized by Computational Molecular Biophysics Group, Heidelberg, Germany, in Sept. 2007.

Poster presentation at the 9th International Symposium on “Protein structure function relationship” organized by HEJ Research Institute of Chemistry, Karachi, Pakistan, in Jan. 2007. Poster awarded 3rd prize.

Oral presentation at the 19th Darmstadt Molecular Modelling Workshop, organized by Computer Chemistry Center, Erlangen, Germany, in May 2005.