Coarse-grained Modeling of Protein Dynamics using Elastic Network Models

Dissertation

zur Erlangung der Doktorwürde

der Fakultät für Biologie, Chemie und Geowissenschaften der

Universität Bayreuth

vorgelegt von

Silke Andrea Wieninger

2012

Die vorliegende Arbeit wurde in der Zeit von Januar 2008 bis Oktober 2012 in Bayreuth am Lehrstuhl Biopolymere unter Betreuung von Herrn Prof. Dr. G. Matthias Ullmann angefertigt.

Vollständiger Abdruck der von der Fakultät Biologie, Chemie und Geowissenschaften der Universität Bayreuth genehmigten Dissertation zur Erlangung des akademischen Grades Doktor der Naturwissenschaften (Dr. rer. nat.).

Dissertation eingereicht am: 07.11.2012
Zulassung durch die Prüfungskommision: 14.11.2012
Wissenschaftliches Kolloquium: 10.05.2013

Amtierende Dekanin:
Prof. Dr. Beate Lohnert

Prüfungsausschuss:
Prof. Dr. G. Matthias Ullmann (Erstgutachter)
Prof. Dr. Rainer Böckmann (Zweitgutachter)
Prof. Dr. Stephan Förster (Vorsitz)
Prof. Dr. Franz X. Schmid

Danksagungen

Mein herzlicher Dank gilt

Prof. Dr. Matthias Ullmann für die ausgezeichnete Betreuung, die Freiheit, eigenen Gedanken und Ideen nachzugehen, und die Einführung in die Welt der Wissenschaft,

Prof. Dr. Engin Serpersu für die spannende Zusammenarbeit an APH und seinen Besuch in unserer Arbeitsgruppe,

Prof. Dr. Heinrich Sticht und Dr. Pia Rücker für die gute und interessante Zusammenarbeit an VSV-G,

Lars Müller und Dr. Timm Essigke für jegliche Unterstützung bei Computerproblemen,

allen aktuellen, früheren und kurzzeitigen Mitgliedern der Arbeitsgruppe,

Lars Ackermann und Sebastian Buchholz für ihren Programmiercode,

BIGSS für finanzielle Unterstützung und interessante Treffen mit Doktoranden anderer Disziplinen, sowie Dr. Stephan Schwarzinger und Violaine Zigan für die gute Organisation,

Dr. Eva-Maria Krammer und Thomas Ullmann für das Korrekturlesen und viele hilfreiche Kommentare und

Dr. Wolfgang Löhr für stete mathematische sowie nicht-mathematische Unterstützung und das Korrekturlesen dieser Arbeit.

Abstract

Dynamics is crucial for the functioning of biological macromolecules. Because of severe limitations in studying protein dynamics experimentally or with all-atom simulations, coarse-grained methods, especially elastic network models (ENMs), are frequently employed. In the last years, studies on various proteins showed that ENMs reliably reproduce experimental data, despite the simplified protein representation and the purely harmonic potential function. This work on two proteins with very different dynamical properties highlights the remarkable success of ENMs and shows which kind of questions can be answered using coarse-grained methods.

The allosteric enzyme aminoglycoside phosphotransferase(3′)-IIIa (APH), which confers resistance against a broad range of aminoglycoside antibiotics to pathogenic bacteria, drastically changes its flexibility upon binding of substrates, but without changing its average conformation. In contrast, the homotrimeric vesicular stomatitis virus glycoprotein G (VSV-G), which triggers the pH-dependent fusion of viral and host membrane, undergoes a large structural rearrangement. A striking difference between the two proteins is their shape. VSV-G contains weakly constrained protein segments, the fusion loops, which can undergo large-scale motions at low energetic cost, whereas APH is not obviously arranged into different protein segments. Nevertheless, ENM calculations show that also APH consists of independently moving segments with correlated internal motion, so-called dynamic domains. The concept of dynamic domains can explain the differential effects of ligand binding on the dynamics of APH.

The first chapter of this thesis describes how experimental evidence for the importance of dynamics successively replaced the former idea of static proteins, and explains the dynamic basis of ligand binding, allostery and conformational changes. In the second chapter, theoretical methods for the analysis of protein dynamics are introduced, with emphasis on the ENM-based methods used in my studies. The studies are summarized in the third chapter. In the study on APH, I employ the Gaussian network model to analyze the ligand-dependent dynamics, the broad substrate specificity and the perturbation-sensitivity of the ligand binding sites. In a second study, ENM-based as well as all-atom molecular dynamics simulations are used to analyze the conformational change of VSV-G. Both approaches detect the fusion loops

of VSV-G as most flexible parts of the protein, and thus as most likely starting point for the structural rearrangement, but only the all-atom model can generate deviations from the average structure at low pH. The last study describes the implementation and application of a dynamic domain assignment method, called CovarDom, which is based on covariances of residue fluctuations. Calculation of dynamic domains for a large protein set demonstrates the general applicability of CovarDom.

Zusammenfassung

Dynamik ist entscheidend für die Funktion von biologischen Makromolekülen. Aufgrund von starken Beschränkungen in Experimenten oder vollatomaren Simulationen werden häufig grob auflösende Methoden, insbesondere elastische Netzwerkmodelle (ENMs), verwendet. In den letzten Jahren zeigten Untersuchungen an verschiedenen Proteinen, dass ENMs experimentelle Daten zuverlässig reproduzieren, trotz der vereinfachenden Proteinbeschreibung und der rein harmonischen Energiefunktion. Diese Arbeit an zwei Proteinen mit sehr unterschiedlichen dynamischen Eigenschaften zeigt den bemerkenswerten Erfolg von ENMs auf und schildert, welche Fragenstellungen mit Hilfe von grob auflösenden, so genannten coarse-grained Methoden beantwortet werden können.

Das allosterische Enzym Aminoglykosid-Phosphotransferase(3′)-IIIa (APH), welches pathogenen Bakterien Resistenz gegen eine breite Palette an Aminoglykosid-Antibiotika verleiht, verändert seine Flexibilität beim Binden von Substraten, allerdings ohne seine mittlere Konformation zu ändern. Im Gegensatz dazu erfährt das homotrimere Vesikuläre-Stomatitis-Virus Glykoprotein G (VSV-G), welches die pH-abhängige Fusion von viraler Membran und Wirtsmembran auslöst, eine große Strukturumlagerung. Ein auffälliger Unterschied zwischen den beiden Proteinen liegt in ihrer Gestalt. VSV-G enthält nur wenigen Beschränkungen unterliegende Proteinsegmente, die Fusionsloops, welche große Bewegungen bei niedrigem Energieaufwand ausführen können, während APH nicht offensichtlich aus verschiedenen Proteinsegmenten aufgebaut ist. Dennoch zeigen Berechnungen mit ENM, dass auch APH aus sich unabhängig bewegenden Segmenten mit korrelierter interner Bewegung, sogenannten dynamischen Domänen, besteht. Das Konzept dynamischer Domänen kann die unterschiedlichen Effekte von Ligandenbindung auf die Dynamik von APH erklären.

Das erste Kapitel dieser Arbeit beschreibt, wie experimentelle Belege für die Bedeutung von Dynamik nach und nach das zuvor verbreitete Bild von statischen Proteinen verdrängte, und erläutert die dynamische Grundlage von Ligandenbindung, Allosterie und Konformationsänderungen. Im zweiten Kapitel werden theoretische Methoden zur Untersuchung von Proteindynamik eingeführt, mit den ENM-basierten Methoden, welche ich in meinen Studien

verwendet habe, als Schwerpunkt. Die Studien sind im dritten Kapitel zusammengefasst. In der Studie an APH verwende ich ein Gaußsches Netzwerkmodell, um die ligandenabhängige Dynamik, die breite Substratspezifität sowie die Sensitivität der Ligandenbindungsstellen gegenüber Störeinflüssen zu untersuchen. In einer zweiten Studie werden ENM-basierte sowie vollatomare Molekulardynamik-Simulationen eingesetzt, um die Konformationsänderung von VSV-G zu untersuchen. Beide Verfahren ermitteln die Fusionsloops von VSV-G als flexibelste Proteinsegmente, und daher als wahrscheinlichsten Startpunkt für die Strukturumlagerung, doch nur das vollatomare Modell kann Abweichungen von der mittleren Struktur bei niedrigem pH-Wert generieren. Die letzte Arbeit beschreibt die Implementierung und Anwendung einer Methode zur Zuordnung von dynamischen Domänen, CovarDom, welche auf Kovarianzen der Fluktuationen von Resten beruht. Berechnung der dynamischen Domänen für ein große Auswahl an Proteinen demonstriert die allgemeine Anwendbarkeit von CovarDom.

Contents

Chapter 1

Functional Importance of Protein Dynamics

Protein dynamics is crucial for all protein functions, including interaction with other molecules, allostery, cell signaling and catalysis. It is defined as the time-dependent change in atomic coordinates, including both equilibrium and non-equilibrium motions. The first section gives a short historical overview over the findings that led to our present-day picture of protein dynamics. Thereby developed principles of protein dynamics are used in the second section to illustrate the importance of dynamics for ligand binding and allosteric regulation of proteins. The third section describes the influence of protein architecture on the range of possible conformational changes.

1.1 From Static Structures to Dynamical Systems

In 1965, Monod *et al.* published a model explaining the allosteric behavior of hemoglobin.[1] The authors postulated the existence of two alternative hemoglobin structures, the T and the R quaternary states, which are characterized by low and high oxygen affinity, respectively. Shortly afterward, an X-ray study of oxygen-free methemoglobin strengthened the assumption that different hemoglobin states must exist, because it showed side chains blocking the entrance to the heme pocket, which must swing away for ligand binding or release.[2] Such ligand-dependent conformational differences between oxy- and deoxyhemoglobin were confirmed by X-ray structures in 1970.[3] In the follow-

ing years, rapid structural fluctuations of proteins in the nanosecond range were inferred from several techniques, including hydrogen exchange,[4] fluorescence quenching[5] and NMR.[6] The first experimental support for the existence of many different protein conformers came in 1974 from carbonmonoxide rebinding measurements to myoglobin after photodissociation.[7] At low temperature, i.e. below 210 K, rebinding is concentration-independent and nonexponential. The nonexponential time dependence results from a range of activation energies of the binding process instead of a single activation energy, which indicates that myoglobin is frozen in closely related structures, the so-called conformational substates. Above the transition temperature, the substates readily interconvert and the rebinding kinetics becomes exponential. The concentration-independence of CO and O_2 rebinding below the transition temperature or if myoglobin is embedded in a solid shows that the ligand remains in one of the specific docking sites in the distal heme pocket after photodissociation, and rebinds from there.[8] Only above the transition temperature, relaxation processes take place and the ligand can dissociate into the solvent. As predicted a decade earlier for hemoglobin,[2] also in myoglobin conformational changes are necessary to open a transient channel to the ligand binding site.

Since then, myoglobin served as model system of protein dynamics in numerous studies, which gradually refined the concept of protein energy landscapes and motions thereon. X-ray crystallography,[9] which yields average positions and mean-square displacements (MSD) of all non-hydrogen atoms, and Mössbauer spectroscopy,[10–12] which yields spatial and temporal information about ^{57}Fe, showed a transition from linear to non-linear temperature dependence of the MSD. Only above the transition temperature, the MSD differs strongly between the atoms and has a contribution additional to the thermal vibrations, which arises from fluctuations between the substates. It was demonstrated that a simultaneous description of the temperature dependence of the MSD determined by X-ray crystallography and by Mössbauer spectroscopy is only possible assuming a complex energy landscape with deep traps formed by the conformational substates and shallow basins in the transition states, which cause friction.[13] Additionally, the study of nonequilibrium motions after photodissociation, leading from carbonmonoxymyoglobin to ligand-free myoglobin, indicated a hierarchical organization of the energy landscape with

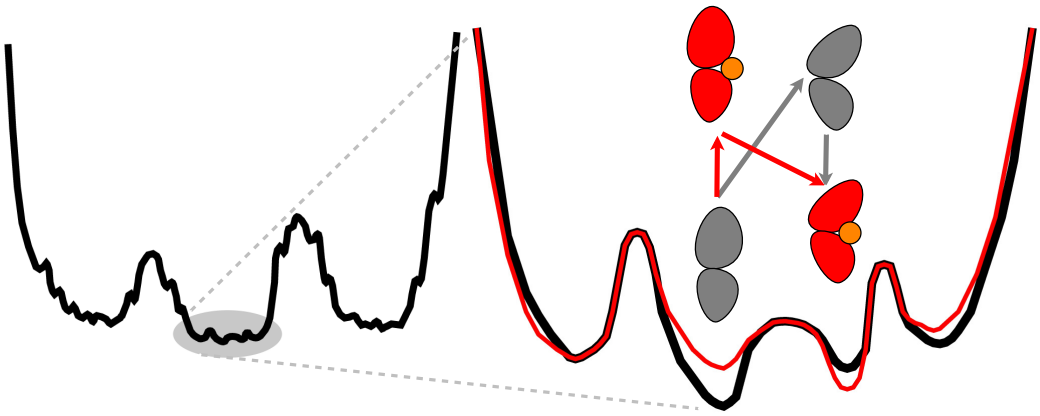

Figure 1.1. Schematic view of the protein energy landscape. Magnification of the shaded region shows the dynamic energy landscape upon ligand binding. The open conformation on the left side is more favorable in the unbound form (gray), while the closed conformation on the right side is more favorable when the ligand is bound (red). Gray arrows indicate the conformational selection pathway, red arrows indicate the induced-fit pathway.

several tiers of decreasing free energy barriers[14] (see Figure 1.1). The conformations in the top tier are called taxonomic substates, because they can be fully characterized. Infrared absorption spectra of the CO stretch bands of carbonmonoxymyoglobin showed that the top tier of myoglobin contains three taxonomic substates.[15] The taxonomic substates of myoglobin fulfill different functions.[16] The taxonomic substate prevailing at high pH stores dioxygen, the one prevailing at low pH is involved in NO enzymatics. Each taxonomic substate assumes a large number of statistical substates.

Over the years it became clear that besides temperature, the MSD also depends critically on the hydration level of the protein. The inability of dry myoglobin to exchange CO with the solvent[8] was a first hint at the essential role of solvent in controlling functionally important protein fluctuations, and was confirmed by the absence of non-vibrational fluctuations in the nuclear gamma resonance of dry myoglobin.[12] The dielectric relaxation of hydration water, which consists of about two layers of water that surround the protein, and of the bulk solvent was measured[17,18] and compared to the temperature-dependent rate coefficients of different myoglobin processes.[19,20] The kinetics of many processes were already investigated before, including covalent CO

binding to the heme iron,[8] CO exit into the solvent,[8,21] fluctuations between taxonomic substates,[22,23] fast fluctuations observed by vibrational echo experiments,[24] slower fluctuations observed after spectral hole burning,[25] and relaxations after pressure release.[15,22] The data showed that large-scale, collective fluctuations follow the dielectric fluctuations in the bulk solvent. They are nonexponential in time, do not follow the Arrhenius law and are absent in rigid environments and dehydrated proteins. These fluctuations govern, for instance, the entrance and exit of ligands in myoglobin. In contrast, local fluctuations are coupled to the fluctuations in the hydration shell, but are essentially independent of the fluctuations of the bulk solvent. They obey the Arrhenius law and are absent in dehydrated proteins, but not in rigid environments. These fluctuations permit the passage of ligands inside myoglobin. Both types of fluctuations are slaved, meaning that the rates are proportional to the fluctuation rate of the surrounding water, but smaller. A third type of motion observed in proteins are vibrational fluctuations. These are nonslaved processes which are independent of the solvent and the hydration shell.

A large number of studies applying different techniques on various proteins shed light on the relation between protein dynamics and function. Electron transfer rates between cytochrome c and cytochrome c peroxidase go to zero at approximately 200 K.[26] Neutron scattering of lysozyme showed that the dependence of anharmonic motions on hydration and temperature correlates well with catalytic activity.[27] Studies on the light-activated enzyme protochlorophyllide oxidoreductase showed that the formation of the first reaction intermediate can occur below the glass transition temperature of the solvent, while the second intermediate is only build above the transition temperature.[28] Internal, non-slaved protein motions drive the first step of the reaction cycle, whereas solvent-slaved motions control the second step. Based on the experimental findings, one can formulate a picture of protein dynamics and the energy landscape underlying it. An instantaneous structure is one point in conformation space and is characterized by the positions of all atoms in the protein and in the surrounding solvent. Conformational substates are the minima and transition states the saddle points in the energy landscape, which is tied to an individual set of temperature, pressure and solvent conditions. Protein structures can adopt a very large number of nearly isoenergetic conformational substates. Protein motions are transitions among the points in

the conformation space and cover time scales from femtoseconds to seconds and corresponding distance scales of fractions of an Ångström to nanometers. Dynamics on a slow timescale of microseconds at physiological temperature occur between substates that are separated by energy barriers of several $k_{\mathrm{B}}T$. Typically, these are large-amplitude collective motions. Fast timescale dynamics occur between states that are separated by energy barriers of less than one $k_{\mathrm{B}}T$ and result in local, small-amplitude picosecond to nanosecond fluctuations, as for example loop motions. Even more local atomic fluctuations as side chain rotations occur on the picosecond timescale, while bond vibrations are motions on the femtosecond timescale. The overall atomic fluctuations can be described as local oscillations superposed on motions with a more collective character.

1.2 Ligand Binding and Allostery

Whether two molecules bind is determined by the associated change in free energy, composed of both enthalpic and entropic terms. In the classical view, ligand binding is enthalpy-driven, and counteracted by unfavorable entropic effects. While close packing of hydrophobic residues and the formation of hydrogen bonds and salt bridges leads to a favorable enthalpy change, it also increases the rigidity of the binding residues, which corresponds to a decrease in entropy. Further significant entropy loss originates from the total number of translational and rotational degrees of freedom, which is reduced from twelve to six upon association. Nevertheless, the entropy change due to ligand binding is not necessarily strongly unfavorable. The missing external degrees of freedom are transformed into six additional internal degrees of freedom of the complex, which recover a large amount of entropy. Furthermore, desolvation of the ligand and the protein binding pocket can release water molecules into the bulk solvent, resulting in a favorable entropy change. The desolvation effect can even lead to entropy-driven ligand binding, when the buried hydrophobic surface is very large, as in inhibitor binding to HIV-1 protease.[29] Calorimetric methods enable the determination of the overall entropic contribution to the free energy of association.[30] For a deeper understanding of the impact of residue flexibility on binding, site-specific entropy changes can be estimated from NMR relaxation.[31] With this method, relaxation parame-

ters for backbone or side chain atoms are determined, which depend on the amplitude of fast time scale motions. Although the resulting order parameters are no quantitative measure of conformational entropy, binding-induced entropy changes can be reasonably deduced from a comparison between free and complexed proteins. NMR relaxation measurements on mouse major urinary binding protein (MUP)[32] disproved the prevalent assumption that ligand binding always leads to motional restriction. Pheromone binding to MUP results in a small increase in backbone motion for nearly all residues, which adds up to a significant increase in backbone conformational entropy, suggesting a dominant role in the stabilization of the complex. Another binding strategy of several proteins, called entropy-entropy compensation, was revealed in relaxation experiments on calcium-loaded calmodulin[33] and a PDZ2 domain from tyrosine phosphatase,[34] which counterbalance the loss of dynamics of binding site residues by increased entropy of side chains distal to the binding site.

The influence of ligand binding on the free energy is described by the concept of dynamic energy landscapes.[35,36] Ligand binding shifts energy landscapes, leading to altered funnel shapes and a redistribution of the populations of conformational substates. Two models of ligand binding, induced-fit[37,38] and conformational selection,[39] describe extreme cases of the coupling mechanism between ligand binding and conformational change. According to the more than fifty years old induced-fit model, the ligand binds to the protein and triggers the conformational change. This model was supported by the growing number of proteins with known crystal structure of a ligand-free, open form and a ligand-bound, closed form. In contrast, according to the conformational selection view, the protein already samples binding competent conformations in the ligand-free state. The ligand selects complementary protein conformations from this native ensemble, depleting the binding conformer from the solution and shifting the equilibrium in favor of the closed form. This model is supported by the finding that some proteins undergo transient motion toward the closed conformation also in their ligand-free state, for example Ca-free calmodulin.[40] A beneficial feature of the conformational selection model is that it can easily explain the binding promiscuity of very flexible proteins. In a simplified view, the two models of ligand binding can be described as transition between four different states[41] (see Figure 1.1). In the energy landscape of the ligand-free protein, the open conformation is lowest

in energy and therefore most populated, whereas the closed conformation is less populated. In the shifted energy landscape of the ligand-bound protein, the closed conformation is more favored than the open conformation. There are multiple possibilities to traverse the dynamic energy landscape from the ligand-free open to the ligand-bound closed state. According to the model of conformational selection, the protein first makes a transition to the closed state and then binds the ligand. Alternatively, the protein can first bind the ligand and then go to the closed state, following the induced-fit model. In reality, the ligand most likely chooses one of many possible intermediate ways, binds to a conformation which has certain characteristics of the closed state, and induces further structural changes. Thus, induced-fit and conformational selection are not mutually exclusive, but different protein-ligand systems may tend more toward one mechanism or the other.

The shift of energy landscapes due to ligand binding allows for allosteric regulation of proteins. An allosteric effector causes a redistribution of conformations to states with increased or decreased substrate affinity. Allosteric effectors that bind to a second, equivalent binding site on a distinct subunit of oligomeric proteins are called homotropic. Heterotropic effectors bind to a different site on the same or a distinct subunit. Typical examples of enzymes underlying allosteric control catalyze the first irreversible step in a metabolic pathway, and are activated by substrate and inhibited by the end product. A key goal of the investigation of allosteric proteins is the understanding of the molecular pathways communicating signals between the allosteric and the active site. Long-range communication is mediated not only by changes in the mean conformation, giving the enthalpic contribution, but also by changes in the dynamic fluctuations about the mean conformation, giving the entropic contribution.[42] Some proteins even show purely dynamics-driven allostery, an effect referred to as dynamic allostery.[43,44] One example is catabolite activator protein, a homodimeric transcriptional activator which binds two molecules cAMP. Although binding of the first cAMP molecule to one subunit has only minimal effects on the conformation of the other subunit, binding of the second molecule cAMP is clearly disfavored.[45] Binding of the first cAMP molecule activates fluctuations between an ensemble of alternate conformations on the μs to ms timescale. This favorable entropy contribution to the binding free energy is missing in the second binding step, leading to negative cooperativity

of cAMP binding. Nowadays, allosteric sites have become the target of a class of drugs, called allosteric drugs.[46] The existence of a second site, distant from the active site, increases the possibilities to affect protein function. In contrast to competitive inhibitors, non-competitive ligands need no chemical similarity to the substrate and can not only decrease, but also increase the enzyme activity. For the prediction of physiological and non-physiological allosteric sites, a deeper understanding of the influence of ligand binding on protein conformation and flexibility is needed.

1.3 Conformational Changes and Protein Domains

For many proteins, the structures of alternative conformations are known and can be used to explore the repertoire of protein conformational changes. The architecture of a protein determines its range of possible low-energy motions.[47] When only few interactions are present between protein segments, they can move away from each other at low energetic cost. Few constraints compared to the interactions in globular proteins can occur between the subunits of oligomeric proteins or in proteins consisting of clearly separated segments, so-called structural domains. If the domains are only linked by a flexible hinge region, a few large changes in main-chain torsion angles of the hinge suffice for large opening and closing movements. A typical example is the so-called hinge bending motion of the two lobes forming the active site cleft of lysozyme[48,49] (see Figure 1.2A). In contrast, closed packed segments are constrained and can only undergo small shear movements, which maintain the interface contacts. Such proteins often have layered architectures with one layer sliding over another, such that a number of small shear motions combines to give a large effect. A typical example is the homodimeric enzyme citrate synthase[50] (see Figure 1.2B). In each monomer, the active site lies between a large domain of fifteen α-helices and a small domain of five α-helices, that closes over the large one. Extensive interactions between the two domains take place in both the closed and the open state. Many protein conformational changes can be described by a combination of hinge and shear motions. Besides movement of quasi-rigid domains relative to each other, also motions of smaller fragments, like surface loops and secondary structure elements, can accomplish the conformational change. A classification of low-energy conforma-

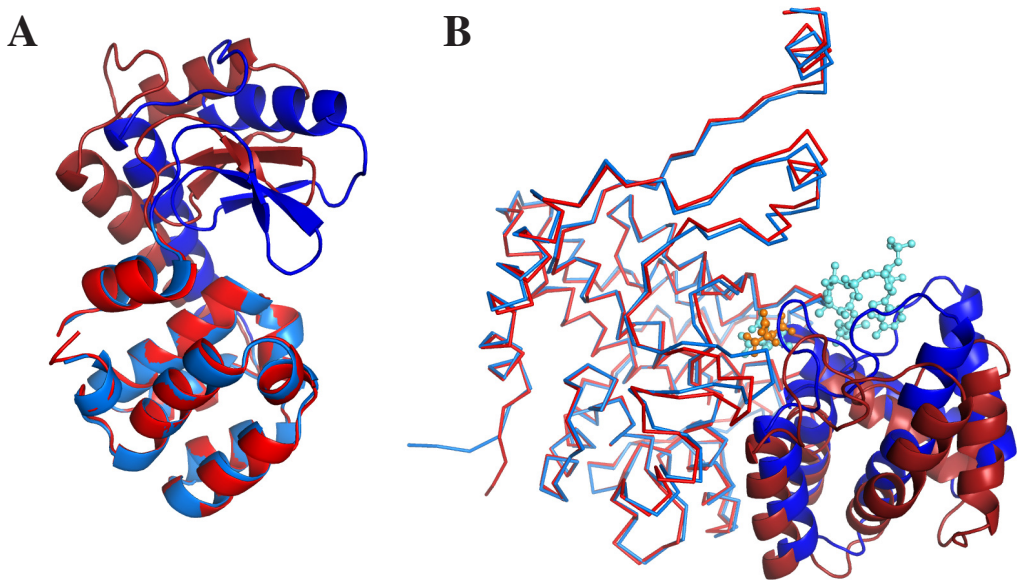

Figure 1.2. Possible low-energy motions of proteins. Hinge motions are closure movements which create new interactions, whereas shear motions occur between protein segments which interact in both conformations. A) Hinge motion in I3P mutant of T4 lysozyme.[51] The open conformation (PDB code 1l97) is colored in red, the closed conformation (PDB code 1.96) in blue. B) Shear motions in the homodimeric citrate synthase, depicted for one subunit.[50] The open conformation (PDB code 1cts) is colored in red, the closed conformation (PDB code 2cts) in blue. Citric acid, which is bound to both forms, is shown as orange and light-blue ball and stick model for the open and the closed conformation, respectively. Binding of coenzyme A, shown as light-blue ball and stick model, results in the closed conformation by shear movements of five α-helices, shown in cartoon representation. The images were produced using PyMOL.[52]

tional changes of proteins can be found at the Molecular Movements Database www.molmovdb.org.[53] Conformational changes which occur on a much slower timescale can also involve breaking and rebuilding of many interresidue contacts and lead to a larger structural rearrangement. One example is the pH-dependent structural rearrangement of Vesicular Stomatitis Virus glycoprotein G triggering fusion between virus and host cell membrane.[54]

Chapter 2

Modeling of Protein Dynamics

Theoretical models of protein dynamics can be used to investigate the amplitude, time dependence and spatial correlation of fluctuations. Examples from literature, which demonstrate how theoretical methods can complement and explain experimental data and deliver information not accessible to experiment are given in the following sections. The description of all-atom models in the first section allows for an understanding of the approximations made in the elastic network models, which are employed in my studies and explained in detail in the last section.

2.1 All-Atom Models

The dynamics of proteins is too complex to be computed by quantum mechanical approaches. Instead, molecular mechanics is employed, a force field method which describes the potential energy of the system as a function of nuclear positions only. The electronic motions can be ignored according to the Born-Oppenheimer approximation, because they are fast enough to equilibrate in the time needed for nuclei motions. There are several empirical force fields which describe the energy landscape of proteins.[55] They differ in parametrization and the exact form of the potential function, but are all composed of a sum of different energy terms. Deviations of bond-lengths, angles and dihedral angles from equilibrium values are penalized by bonded terms, while nonbonded terms account for van der Waals interactions, Pauli repulsion and electrostatic interactions. Molecular dynamics (MD) simulations are based on

molecular mechanics force fields and explore the time-dependent behavior of proteins, providing a detailed picture of the way in which a system passes from one conformation into another. Successive configurations of the system are generated by applying Newton's equations of motion. First protein MD simulations published in 1977 were applied to bovine pancreatic trypsin inhibitor (BPTI) and were carried out in vacuum and without explicitly considering hydrogen atoms.[56] They showed a high flexibility of the termini, the loop region and exposed side chains, in contrast to α-helices and β-sheets. Following MD simulations of BPTI in explicit solvent showed protein dynamics as a superposition of local, high-frequency oscillations and collective, low-frequency fluctuations.[57] MD simulations of myoglobin confirmed the assumed complexity of the energy surface, which is characterized by a large number of thermally accessible minima in the neighborhood of the native structure, and illuminated the structural differences between nearly isoenergetic minima.[58] It was estimated that twenty to thirty percent of the root-mean-square (RMS) fluctuations of main chain atoms are contributed by harmonic oscillations within a well and the rest arises from anharmonic transitions among wells. MD simulations on myoglobin at different temperatures could affirm the transition from linear to non-linear temperature-dependence of dynamics of hydrated proteins around 210 K.[59] At low temperature there is purely vibrational motion, while above the transition temperature the atomic fluctuations exhibit both harmonic and anharmonic behavior. MD simulations also proved to be successful in the prediction of NMR order parameters, suggesting that MD can be useful for the determination of entropy changes.[60]

To analyze the trajectories generated by MD, one can cluster conformations to detect highly sampled regions in conformation space.[61] Alternatively, one can employ Principal Component Analysis (PCA)[62,63] to extract large-scale motions present in a MD trajectory. It allows to reduce the complicated dynamics to a lower-dimensional description of the functional motions by a change of orthonormal basis. First the overall translational and rotational motion must be eliminated from the snapshot structures. Then the symmetric covariance matrix **C** is constructed, which gives the mass-weighted atomic displacements in configuration space, defined by the $3N$ Cartesian coordinates. The covariance between atom i and j is given by

$$C_{i,j} = \sqrt{m_i m_j}\langle(\vec{r}_i(t) - \vec{r}_i^{\,av})(\vec{r}_j(t) - \vec{r}_j^{\,av})\rangle. \tag{2.1}$$

$\vec{r}_i^{\,av}$ is the mean position of atom i, averaged over all snapshot positions $\vec{r}_i(t)$, and m_i is its mass. The diagonal elements of **C** give the variances, which measure the average amplitude of motion along one coordinate, while the off-diagonal elements give the covariances, which measure the degree of linear relationship between motions. The goal of PCA is to find uncorrelated directions along which large-amplitude fluctuations take place. Expressing the protein motions as linear combination of vectors along such directions diagonalizes the covariance matrix. Computationally, the diagonalization is achieved by solving the eigenvalue problem of the covariance matrix and using the eigenvectors of **C**, called principal components, as new orthogonal basis. The principal components are sorted by their associated eigenvalues, which give the mean-square fluctuations, with mode 1 being the largest-amplitude motion. PCA assumes that the probability distributions are fully characterized by the mean and the variance. This assumption is true for Gaussian probability distributions, but not in general. It is fulfilled by harmonic motions and approximately also by many anharmonic motions, but not by modes traversing multiple minima. Most protein fluctuations can be described by a subspace spanned by the first principal components, called essential subspace, as for example the conformational change in lysozyme.[64]

Another method used for the identification of large-scale protein motions is Normal Mode Analysis (NMA). NMA was originally employed for the assignment of high-frequency bands in vibrational spectra of infrared, Raman or inelastic neutron scattering spectroscopy[65] and later established as computational tool for analysis of harmonic protein motions.[66,67] Instead of numerically solving Newton's equations of motion, NMA yields a unique analytical solution of collective modes by expansion of the potential function in a Taylor series. If $\vec{r}_0$ is the coordinate vector of a reference structure and $\vec{r} = \vec{r}_0 + \Delta\vec{r}$ is the coordinate vector of a structure diplaced by a small amount $\Delta\vec{r}$, the Taylor series is

$$V(\vec{r}) = V(\vec{r}_0) + \vec{g}^{\,\mathrm{T}}\Delta\vec{r} + \frac{1}{2}\Delta\vec{r}^{\,\mathrm{T}}\mathbf{H}\Delta\vec{r} + ... \tag{2.2}$$

where the first-derivative vector of the energy, $\vec{g}$, and the second-derivative matrix, **H**, are determined at the reference structure, $\vec{r}_0$. The reference structure must be properly energy-minimized, such that the gradient vanishes and the Hessian matrix is positive-semidefinite, that is all of its eigenvalues are non-negative. Terms after second order are neglected. This harmonic approxi-

mation to the potential function is only valid for dynamics in a single potential well. We can use Newton's law $\vec{f} = \mathbf{M}\vec{a}$ to describe the motion of the atoms in the system, with $f_i = -\frac{\partial V}{\partial r_i}$, $a_i = \frac{\mathrm{d}^2 r_i}{\mathrm{d}t^2}$ and mass matrix $\mathbf{M}$, and obtain

$$-\mathbf{H}\Delta\vec{r} = \mathbf{M}\frac{\mathrm{d}^2\Delta\vec{r}}{\mathrm{d}t^2}. \tag{2.3}$$

The solutions of this second-order differential equation are of the form

$$\Delta\vec{r}_i = \vec{u}_i \cos(\omega_i t + \phi_i). \tag{2.4}$$

Substitution into the differential equation and usage of mass-weighted Cartesian coordinates yields

$$\mathbf{H}'\vec{u}_i' = \lambda_i \vec{u}_i', \tag{2.5}$$

with $\mathbf{H}' = \mathbf{M}^{-\frac{1}{2}}\mathbf{H}\mathbf{M}^{-\frac{1}{2}}$, $\vec{u}_i' = \mathbf{M}^{\frac{1}{2}}\vec{u}_i$, $\lambda_i = \omega_i^2$. The eigenvectors $\vec{u}_i'$ of the Hessian matrix are called normal modes. The associated eigenvalues λ_i give the frequency of the harmonic motion, which depends on the curvature of the potential along the normal mode directions. The first six normal modes have an eigenvalue of zero and describe translational and rotational rigid-body movements.

The frequency spectra of different proteins are very similar, because most of the modes describe motions that are common to all proteins, ranging from hydrogen vibrations to vibrations of secondary-structure elements.[68] In contrast to principal components, the large-amplitude normal modes describing specific motions of a protein have small eigenvalues. These low-frequency modes generally involve nonlocalized motions of the molecule and contribute dominantly to the mean-square displacements of C_α atoms. Early NMA studies showed that RMS backbone fluctuations calculated from normal modes correlate well with RMS fluctuations obtained from MD simulations[69] and with experimentally observed X-ray temperature factors.[70] Also conformational changes of proteins known from different crystal structures were predicted sucessfully by low-frequency normal modes, like the hinge bending motions of lysozyme,[71,72] citrate synthase[73] and the E.coli ABC Leu/Ile/Val transport system.[74] The time-dependent displacement $\Delta\vec{r}$ of a protein along the first n non-zero collective modes can be expressed as linear combination of the orthogonal normal modes weighted by the inverse of their frequencies ω_i,

$$\Delta\vec{r}(t) = \mathbf{M}^{-\frac{1}{2}} \sum_{i=7}^{n} \frac{\sqrt{2k_\mathrm{B}T}}{\omega_i} \vec{u}_i' \cos(\omega_i t + \phi_i), \tag{2.6}$$

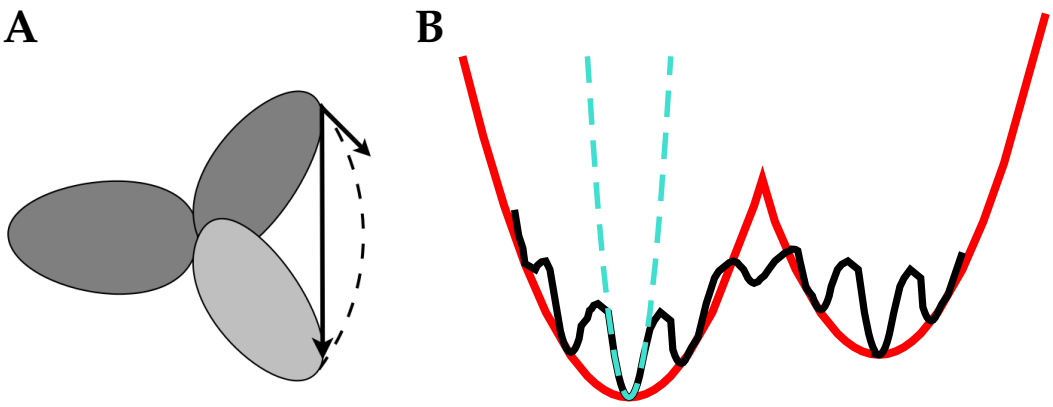

Figure 2.1. Description of conformational changes by normal modes. A) Schematic drawing of a rotating domain. For rotations, the direction of infinitesimal motion of a normal mode, depicted as small arrow, deviates from the finite motion of the conformational change, depicted as long arrow. B) Schematic representation of the harmonic approximation of NMA (cyan dotted curve) and ENM (red) to the potential energy surface (black curve). The smoother landscape of the ENM allows for sampling of nearby local minima. But a transition from the left potential well to the potential well of another conformation, shown as second red parabola on the right side, is not possible.

with Boltzmann factor k_B, temperature T and phases ϕ_i. Comparing the finite motions between two protein conformations and the infinitesimal motion directions given by NMA, one must bear in mind that they are different for rotational motion[75] (see Figure 2.1A). For a better comparison between conformational change and normal mode directions, one can replace the finite conformational change by an infinitesimal rigid-body motion of separate domains.[76] Determination of such axes from PCA of a MD simulation of solvated lysozyme and NMA of lysozyme in vacuum yielded similar hinge axes, showing good agreement despite the neglect of anharmonic and solvent effects in NMA.[77] Anharmonic behavior is present if higher-order terms of the Taylor expansion contribute significantly to the dynamics. Due to the harmonic approximation, NMA can in principle only describe the first steps of a conformational transition, because transitions between energy minima would require barrier crossing. But the energy barrier for conformational transitions caused by ligand binding can be significantly lowered by the continuosly developing protein-ligand interactions in the process of binding, reducing the contribution of anharmonic motion.[78] As a further limitation, the calculation of normal

modes in vacuum neglects the slowing down of large-amplitude motions by solvent damping. But the directions of low-energy motion are determined by the potential surface and thus hardly affected by solvent. Accordingly, comparison of covariance matrices from MD in vacuum and in solvent and from NMA on BPTI showed close agreement.[79] However, the assignment of time scales and amplitudes of motion would require a detailed model that incorporates anharmonic and solvent effects.[68] The observation that the subspace spanned by the lowest frequency modes is robust, meaning that it does not depend very sensitively on the energy function, was also made for normal modes in dihedral angle in comparison to cartesian coordinate space.[80] But single low-frequency modes can be arranged or combined differently, especially if the modes are nearly equal in energy. Thus, one should always analyze the essential subspace instead of single normal modes.

NMA can be used to test the accuracy of force fields.[81,82] Another application of NMA is the determination of the vibrational entropy of a system and the increase of vibrational entropy of proteins due to ligand binding or protein association. Dimerization of insulin showed that binding does not only add six vibrational modes, but also alters the overall density of states,[83] resulting in lower frequency modes of the dimer in comparison to the monomer. It was not possible to identify a small number of specific modes of the complex that give rise to the vibrational entropy increase. Instead, small alterations in the frequencies of many modes were found to contribute.

Projection is a valuable tool for comparing data from MD, PCA and NMA. Projection methods can be used to determine the contribution of a mode to the motion under consideration, for example of a secondary structure element, to analyze the effect of changed conditions on MD trajectories or to investigate the harmonic and anharmonic contributions to a trajectory. Projecting MD trajectories of solvated and unsolvated lysozyme onto the normal modes of the protein showed that solvent effects are important for the slowest motions with frequencies below 1 ps^{-1}, but negligible for faster motions.[84] In vacuum, there are no conformational transitions, and the motion is restricted to the surroundings of a single stable conformation. Only the slowest modes change the shape of the protein and thereby its surface, making interactions with the surrounding water molecules more important. Projections of MD trajectories onto the normal modes allow to determine the deviation from harmonicity and to de-

tect large conformational changes in MD simulations, because a transition to a new minimum changes the contribution of single modes to the overal motion.[80] The harmonic approximation to the original energetic minimum is not a good approximation to the new minimum anymore. Instead of directly projecting the MD trajectory onto the normal modes, one can also use principal components. PCA in combination with NMA showed that the transition from linear to nonlinear temperature-dependence of the root-mean-square displacement (RMSD) of hydrated myoglobin arises from collective motions along a few anharmonic principal components.[85]

2.2 Coarse-graining and Multiscale Modeling

Despite an immense increase in computational power, there is a trend to use coarse-grained models for the simulation of the dynamics of macromolecules. All-atom descriptions in explicit aqueous environment are in general still limited to a time scale of nanoseconds at a spacial scale of nanometers. In contrast, many relevant dynamics and interactions of proteins occur on a timescale of microseconds to milliseconds and involve large macromolecular aggregates. Therefore, coarse-grained methods are often applied to huge complexes like the ribosome.[86] Generation of structural ensembles is crucial to reliably predict free energy changes,[87] for example upon ligand binding or protein-protein association. Coarse-graining accelerates the dynamics not only by reducing the number of degrees of freedom, it also reduces the ruggedness of the potential energy surface, allowing for a larger time step in MD simulations. By uniting groups of atoms into single interacting centers or pseudoatoms one gets rid of the irrelevant degrees of freedom. C_α-C_α pseudo bond stretching, which is the fastest vibration in C_α based models, has a lower frequency than the O-H and N-H bond vibrations of atomistic models. The longer time scales accessible to coarse-grained simulations allow for a direct comparison between simulation and experiment. In NMA, coarse-graining allows for sampling of nearby conformations which would be inaccessible in classical NMA, because the coarse-grained description smooths out local energy barriers in the potential surface (see Figure 2.1B).[88] But enhanced sampling is by far not the only goal of coarse-grained approaches. Coarse-grained models can be used to describe systems for which no high-resolution structures are known. Moreover,

the identification of the simplest models that are able to capture the essential features determining protein motions helps in understanding the properties underlying dynamics. Reduced models can be justified by the observation that time scales present in macromolecular systems are separated into slowly and rapidly evolving degrees of freedom.[89] A set of slow degrees of freedom regulates the behavior of the system over long time scales, while the remaining, much faster degrees of freedom easily equilibrate around each given point in the space spanned by the slow degrees of freedom. With the same reasoning as the neglect of electronic degrees of freedom in Molecular Mechanics, one can neglect the fast degrees of motion of certain nuclei, and integrate the effect of the rapidly changing variables into the definition of effective interactions between the slower variables.

The degree of coarse-graining varies from a few atoms to entire domains or macromolecules. The least reduced is the united atom model that eliminates only some hydrogens. In four-bead models,[90] the side chain is represented by a single bead, whereas the coordinates of the three heavy atoms of the backbone are represented explicitly, allowing an explicit description of the hydrogen bonds. One-bead models represent each residue by one bead and reduce the number of interacting particles by an order of magnitude. The parametrization of protein models can be structure-independent and transferable, like molecular mechanics, or rely on a certain protein structure, like elastic network models[91,92] and Gō models.[93] The most difficult aspect of protein-independent models is parametrization. The smaller the number of beads representing an amino acid, the harder it is to build a parametrization transferable to other proteins.[86] A variety of coarse-grained models have been introduced in MD. The MARTINI force field for MD of proteins and lipids was implemented into Gromacs.[94] Further freely available programs allowing for coarse-grained MD are CafeMol,[95] ESPResSo[96] and YUP.[97] Multiscale techniques combine the efficiency of coarse-grained simulations with the detail of all-atom simulations. One can use different resolutions in different regions of the molecule during a single simulation, for example represent only the active site in detail,[98] or coarsen lipid and water molecules in a membrane-bound ion channel, while using an all-atom representation for the ion channel itself.[99] Also mixed levels of coarse-graining are applied to analyze different parts of the structure with different detail, from atomistic to dozens

of residues as one coarse-grained site.[98,100,101] In contrast, the resolution exchange method[102] switches between different levels of structural detail during the simulation in order to cross energy barriers. Another strategy applies simplified models of the whole system to generate alternative, all-atom protein structures. It assumes that it is possible to reliably and efficiently move between coarse-grained and all-atom models, and that the coarse-grained model is physically realistic so that the protein structures being sampled represent relevant conformations of the protein.[103] Normal mode directions obtained from ENM can be used to iteratively deform structures[104,105] or to steer MD simulations.[106] The obtained structural ensembles are useful as templates for homology modeling and for generating putative transition pathways or incorporating receptor flexibility in docking approaches.

2.3 Elastic Network Models

In 1996, Tirion proposed a model which eliminates the time-consuming and inaccurate energy minimization prior to NMA.[107] The simplification is achieved by assuming that the input conformation corresponds to a local minimum. The molecular mechanics force field is replaced by a single-parameter potential. Atom pairs are connected with Hookean springs with a uniform force constant γ, and the equilibrium distances r_{ij}° are given by the atom distances in the experimentally determined structure. The total energy of a molecule consisting of N atoms is

$$E_{\text{Tirion}} = \sum_{i,j=1}^{N} \frac{\gamma}{2}(r_{ij} - r_{ij}^{\circ})^2 H(r_{\text{cut}} - r_{ij}^{\circ}). \tag{2.7}$$

The Heaviside step function $H(x)$ equals one if $x \geq 0$ and zero otherwise, ensuring that only atom pairs with a separation closer than a cutoff distance r_{cut} are connected. In the following years, several modifications of Tirion's model were described. The anisotropic network model (ANM)[92] also employs the potential function of Eq. 2.7, but replaces the atomic description by a one-bead model (see Figure 2.2A). Each amino acid is represented by a node located at the position of the C_α atom. For nucleic acids, phosphate atom positions are used. A few years earlier, Hinsen had already proposed an ANM with spring constants which exponentially decay with the atom pair separation, eliminating the need for a cutoff distance.[68] Another widely applied elastic network

model (ENM), the Gaussian network model (GNM),[91] is deduced from polymer science[108] and based on a different potential function. Assuming that the fluctuations are Gaussian and isotropic, the resulting harmonic potential can be written in terms of the coordinate changes $\Delta x_i = x_i - x_i^\circ$, $\Delta y_i = y_i - y_i^\circ$ and $\Delta z_i = z_i - z_i^\circ$ as

$$E_{\mathrm{GNM}} = \sum_{i,j=1}^{N} \frac{\gamma}{2} \left[(\Delta x_i - \Delta x_j)^2 + (\Delta y_i - \Delta y_j)^2 + (\Delta z_i - \Delta z_j)^2 \right] H(r_{\mathrm{cut}} - r_{ij}^\circ). \tag{2.8}$$

GNM penalizes not only changes in internode distances, but also any change in the direction of the internode vector (see Figure 2.2B). The isotropy leads to a threefold degeneration of the $3N \times 3N$-dimensional Hessian matrix, which can thus be reduced to the $N \times N$-dimensional Kirchhoff matrix $\mathbf{\Gamma}$, defined by

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } r_{ij} \leq r_{\mathrm{cut}} \\ 0, & \text{if } i \neq j \text{ and } r_{ij} > r_{\mathrm{cut}} \\ -\sum_{k,k\neq i} \Gamma_{ik}, & \text{if } i = j. \end{cases} \tag{2.9}$$

To consider interactions from residues of the first coordination shell,[109] the cutoff distance r_{cut} is usually set to a value around 7Å.

GNM allows for the calculation of variances $\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_i \rangle$ of residue fluctuations and covariances$\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_j \rangle$ of residue fluctuations, which are evaluated from the diagonal and off-diagonal elements of the inverse Kirchhoff matrix, respectively, using

$$\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_i \rangle = \frac{3k_BT}{\gamma} (\mathbf{\Gamma}^{-1})_{ii}, \quad \langle \Delta\vec{r}_i \cdot \Delta\vec{r}_j \rangle = \frac{3k_BT}{\gamma} (\mathbf{\Gamma}^{-1})_{ij}. \tag{2.10}$$

Solving the eigenvalue problem of the Kirchhoff matrix delivers $N-1$ nonzero eigenvalues λ_i and corresponding eigenvectors $\vec{u}_i$, which are used to determine the pseudo-inverse of the Kirchhoff matrix $\tilde{\mathbf{\Gamma}}^{-1}$ as

$$\tilde{\mathbf{\Gamma}}^{-1} = \sum_{i=2}^{N} \frac{1}{\lambda_i} \vec{u}_i \vec{u}_i^{\mathrm{T}}. \tag{2.11}$$

In the anisotropic models, the expectation values are accordingly calculated from the trace of the 3×3-dimensional submatrices $\mathbf{H}_{ii}$ of the pseudo-inverse Hessian matrix. The theoretically determined fluctuations can be compared to mean-square displacements in X-ray diffraction data, which are related to the

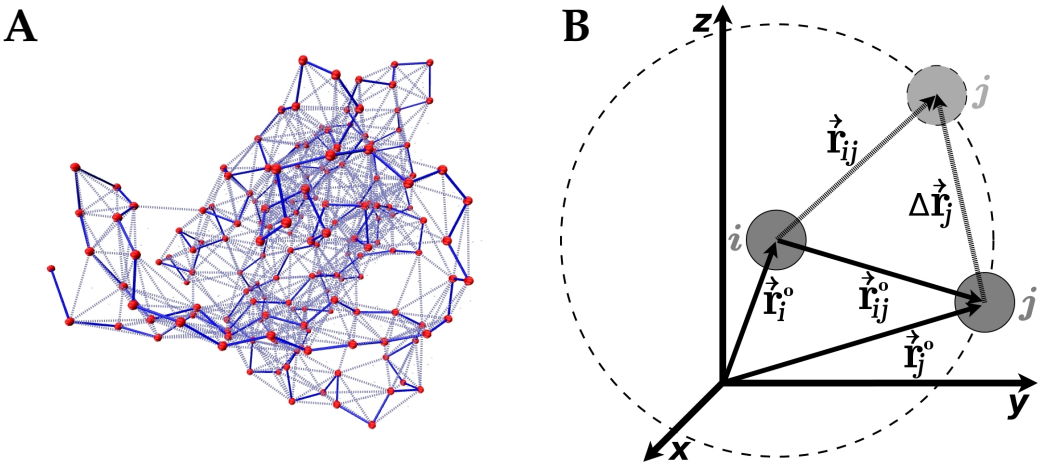

Figure 2.2. Protein representation of the elastic network model. A) One-bead ENM of M-Ras[110] constructed using a cutoff distance of 8 Å. Nodes located at the coordinates of C_α atoms are shown in red, bonds between nodes representing sequential residues are indicated by solid blue lines, and bonds between nodes representing non-sequential residues are indicated as blue dotted lines. The image was produced using VMD.[111] B) Potential function difference between GNM and ANM. In ANM, the displacement of node j costs no energy, because $|\vec{r}^{\circ}_{ij}| = |\vec{r}_{ij}|$. In GNM, also the change in direction of the internode vector $\vec{r}_{ij}$, given by $\Delta\vec{r}_{ij} = \vec{r}_{ij} - \vec{r}^{\circ}_{ij}$, is penalized.

crystallographic B-factors by

$$B_i = \frac{8\pi^2}{3}\langle \Delta\vec{r}_i \cdot \Delta\vec{r}_i \rangle. \tag{2.12}$$

The experimental mean-square displacements originate from both static disorder due to the ensemble of substates trapped in the crystal, and dynamic disorder due to fluctuations that occur in the crystal. While the static differences between conformations and the largest contributions to thermal atomic vibrations can be described by collective modes, the rigid-body motions of the entire molecule are not considered in ENM. Also crystal contacts are usually neglected, although they reduce the flexibilities of exposed atoms, as was shown by comparison of B-factors of proteins known in different crystal forms.[112] Nevertheless, calculations on 1250 non-homologous proteins showed reasonable agreement between crystallographic B-factors and B-factors computed by GNM over a broad range of cutoff distances from 7 to 15Å.[113] Considering crystal contacts by inclusion of neighboring molecules[114] or by periodic boundary conditions[115] and including the influence of lattice vibrations[116] further improves the prediction of crystallographic B-factors. The theoretical fluctuations can also be compared to data from NMR experiments, like order

parameters,[117] hydrogen-deuterium exchange times[118] or the size of RMSDs of NMR ensembles.[119] The latter study showed that excluding the slowest mode from the calculation of B-factors reduces the correlation to NMR data, but hardly affects the correlation to X-ray data, demonstrating that large-scale motions are restricted in the crystal environment.

Although B-factors computed by ANM were reported to correlate less well with experimental data than those computed by GNM,[120] they are beneficial if anisotropic displacement parameters (ADPs) are available,[121,122] which occurs commonly for X-ray structures with a resolution higher than 1.2 Å. Anisotropic models are also needed for the prediction of functional protein motions, which requires directional information. Application to various large macromolecular complexes, for example DNA-dependent polymerases,[123] the ribosome[124,125] and hemoglobin,[126] and to an extensive set of proteins known in different conformations[127,128] showed that, just as in all-atom NMA, a few low frequency normal modes are usually sufficient to explain the conformational change. It is preferable to use the open conformation as reference structure, because the closure motions are usually easily accessible from the open state. In contrast, additional contacts in the closed form hinder a low-energy transition into the open conformation, as described for several proteins.[75,117,126,129,130] Figure 2.3 shows the open and closed conformation of adenylate kinase.[131,132] Comparison between ANM and MD showed that the ENM successfully reproduces the essential subspace of proteins.[133,134] Furthermore, in a study employing classical NMA, Tirion's full-atom EN and two different one-bead ENMs,[135] it was shown that normal modes concordantly obtained in all models are often involved in functional protein motions. Several studies confirmed the insensitivity of slow collective motions to details of the protein model and the energy function.[68,88] The RTB (rotations-translations of blocks) method,[136,137] which divides the protein into a number of blocks being made of a few consecutive residues, was shown to predict the slowest motions sufficiently, provided the shape of the protein is properly captured. Even in an ENM study with much lower resolution, that is one node representing ten to forty residues, the global motions were only sligthly affected.[138] The robustness of low-freqency modes was further explored by representing the protein structure on a cubic lattice[139] and by randomly changing the non-zero Hessian matrix elements,[140] showing that the absolute values of stiffness and directionality of local in-

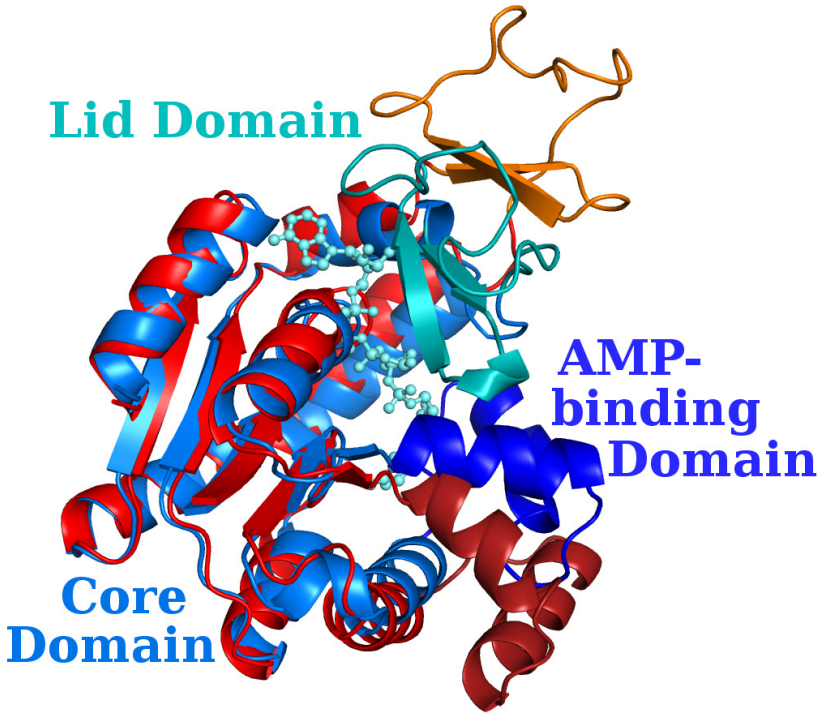

Figure 2.3. Hinge motions in adenylate kinase. The open conformation (PDB code 4ake[132]) is colored in red, the closed conformation (PDB code 1ake[131]) in blue. Lid domain and AMP-binding domain close over the inhibitor bis(adenosine)-5′-pentaphosphate, which is shown as light-blue ball and stick model. The additional interactions formed in the closed conformation hinder a low-energetic motion towards the open conformation, whereas the transition from open to closed conformation is successfully described by the first normal modes obtained from ANM. The image was produced using PyMOL.[52]

teractions hardly influence the low-frequency motions. The low-frequency subspace of eigenvectors is predominantly determined by the shape of the molecule, strengthening the foundation of coarse-grained ENMs with single force constant.

Despite these observations, many alternative EN models were proposed. In the β Gaussian model,[133] C_β centroids are rigidly tethered to the C_α nodes. Often, a more complex assignment of force constants than a single value for all interactions is suggested. Usage of additional force constant parameters was proposed for covalently bound residues,[122] for interactions within α-helices,[141] for intradomain contacts[76,142] and for different amino acid types.[143] Force constants can be assigned by comparison of computed fluctuations to crystal B-factors[144] or to fluctuations from a all-atom MD simulation.[145–147] The chemical network model (CNM) evaluates atomic contacts to determine residue interactions.[148] Various types of ENM calculations can be performed on the web servers elNémo,[149] oGNM,[113] MAVEN[150] and ProDy,[151] and by the programs MMTK[152] and RedMD.[153] The collective motions calculated by ENM can be used to deduce further protein properties. For the assignment of protein domains, the absence of local deformations in low-frequency normal modes[68] or covariance patterns of residue fluctuations are exploited.[154] ENM was also used to assign allosterically coupled sites, i.e. sites where binding can cause a change in ligand-affinity at another site, by determining which binding sites are simultaneously affected by the same motion.[155] Various ENM based methods were proposed for generating transition pathways between equilibrium conformations, for example elastic network interpolation,[156] the double-well network model,[157] the plastic network model (PNM),[130] mixed ENM[158] or interpolated ENM.[159] Another often described application of ENMs is the analysis and refinement of low-resolution data from X-ray crystallography,[160] cryo-electron microscopy[161] and small-angle X-ray scattering.[162] Alternative conformational substates are detected by fitting a high-resolution X-ray structure into low-resolution data of a different conformational state using normal modes.[163]

Chapter 3

Manuscript Overview

3.1 Motivation and Synopsis

Elastic network models reliably reproduce experimental data, can be applied to large biomolecular complexes and highlight the properties governing protein dynamics. In my work, I used different ENM-based methods. The studies are presented in the following chapters and elucidate the relationship between protein structure and dynamics, but also investigate the applicability and limitations of ENMs. Manuscript A describes a GNM study of the ligand-dependent dynamics of the bacterial enzyme aminoglycoside phosphotransferase(3′)-IIIa (APH), which confers resistance against a broad range of aminoglycoside antibiotics. In manuscript B, the large structural rearrangement of the homotrimeric 65-kDa protein Vesicular Stomatitis Virus Glycoprotein G (VSV-G), which triggers the pH-dependent fusion of the viral membrane with the host membrane, is simulated by coarse-grained MD. Different processes perturb the dynamics of the two proteins. In APH, binding of nucleotide and binding of various aminoglycosides have very different effects on the dynamics.[164,165] The binding of ligands can be simulated by adding a few nodes, which represent the ligand, to the elastic network of the protein. In VSV-G, the structural rearrangement is caused by protonation changes of residues. I performed electrostatic calculations on the prefusion conformation to determine the protonation states of all titratable residues at pH 5 and 7. Based on the titration curves, differentially protonated histidine residues could be detected, which represent promising triggers for the structural change. They are posi-

tioned at functionally important interfaces between domain IV, which contains the fusion loop, and the protein core, and are conserved in homologs, as shown in conservation studies carried out by Pia Rücker. Based on these results, Pia Rücker performed two all-atom MD simulations, 50 ns each, with protonation states representing the two pH values. In the coarse-grained MD simulation, integer charges corresponding to the protonation states of the residues were assigned to the nodes. The coarse-grained MD simulation was computed using the program RedMD[153] and an ANM force field combined with Coulomb interactions.

VSV-G contains weakly constrained protein segments, the so-called fusion loops, which can undergo large-scale motions at low energetic cost. The coarse-grained MD simulation confirmed the assumed high flexibility of the fusion loops. Because the ENM is based on topological constraints, it is expected to correctly predict the high flexibility of quasi-independently moving protein regions. But is ENM also applicable to proteins which are not obviously arranged into different domains, like for example APH? The existence of a domain structure of APH could neither be deduced from visual inspection nor from comparison between different conformations, because only small structural differences can be seen in the X-ray structures of APH in the apo form and different substrate-bound forms. Nevertheless, ENM calculations suggested that APH consists of quasi-independent segments with correlated internal motion. Such segments are called dynamic domains, and are characterized by the similarity of the dynamic properties of their residues. The assignment of three dynamic domains to APH demonstrated that dynamic domains are a valuable concept for understanding the differential effects of ligand binding on APH dynamics. I could show that perturbation-sensitive sites of ligand binding, which may be interesting for mutation studies and drug design, lie between the anticorrelated dynamic domains, just as the natural ligand binding sites. Manuscript A describes the computational method used to assign dynamic domains to APH, which is based on covariances of residue fluctuations. Manuscript C generalizes the domain assignment method and compares the dynamic domains of a large set of proteins to manual domain assignments.

Why are ENMs successful in describing protein flexibilities and collective motions, also for quite compact proteins as APH? One can approach this question by looking at other simple models, which were proposed to reproduce

residue flexibilities. The translation libration screw model (TLS) describes a crystalline protein as internally rigid body undergoing motion along translation, libration and screw axes.[114] It works comparably well relative to GNM on highly spherical structures. CONCOORD (CONstraints to COORDinates) generates random protein structures that fulfill a set of upper and lower interatomic distance limits.[166] The authors concluded that motional freedom in proteins is ruled largely by a set of simple geometric constraints. Halle[167] proposed a direct inverse proportionality between crystallographic B-factors and the local packing density, that is the number of noncovalent nonhydrogen neighbor atoms within the first coordination shell. It performs well because most types of interactions are manifested in the local packing density, for example secondary structures are not only extensively hydrogen-bonded but are also densely packed. This implicit encoding of interactions in the structure is definitely an important factor explaining the success of ENMs. But is the number of atomic neighbors really enough to understand residue flexibilities, or does the overall architecture of the protein influence the dynamics, too? My study on APH dynamics gave valuable insights into this question. A stabilization of APH upon antibiotic binding, seen in H/D exchange experiments,[165] was confirmed by the calculations, which predict reduced node flexibilities for the binding residues when nodes representing the antibiotic are added to the elastic network. In manuscript A, I developed an approach which allows to determine the contribution of connectivity to the flexibility change upon ligand binding. It turned out that the connectivity is indeed a decisive factor influencing the flexibility of nodes within the cutoff radius of the ligand. But also protein nodes which are too far away from the ligand binding site to be connected to the ligand change their flexibilities upon binding, which demonstrates that also the overall architecture of the protein influences the node flexibilities.

Apparently, few constraints lead to high flexibility. What happens when more and more constraints are added? A surprising destabilization of β-sheet residues of APH upon nucleotide binding, again seen in H/D exchange experiments,[165] contradicts the idea of reduced flexibility due to increased connectivity. The destabilization can be understood by considering the location of the nucleotide binding site between anticorrelated dynamic domains. Ligand binding adds further constraints to the stable β-sheet region, which can result in frustration, tilting of the β-sheet, and disrupted hydrogen bonds. The exis-

tence of overconstrained protein regions was also postulated by the authors of the graph-theoretical approach FIRST (Floppy Inclusion and Rigid Substructure Topology), which uses an all-atom representation to decompose the molecule into rigid clusters.[168] It distinguishes between underconstrained or flexible regions, constrained regions and overconstrained regions, with more crosslinking bonds than needed to rigidify the region. Can we draw conclusions about the effect of additional constraints on the enthalpy and entropy changes of proteins? An isothermal titration calorimetry (ITC) study[164] showed that binding of antibiotic to apo-APH is driven by a much more favorable enthalpy change than antibiotic binding to the APH-nucleotide complex. This finding can be explained by the overconstrained nature of the dynamic domain participating in both nucleotide and antibiotic binding. The smaller enthalpy change upon binding to the binary complex is compensated by a smaller entropic penalty. Calculations on the X-ray structure of the binary APH-AMPPNP complex confirm the higher flexibilities of residues near the antibiotic binding site when nucleotide is bound, resulting from an open conformation of the antibiotic binding loop. This feature could enable the broad substrate selectivity of the allosteric enzyme APH, because binding of nucleotide leads to enhanced flexibility of antibiotic binding residues.

ENMs were applicable to all issues presented so far. Are there also questions which cannot be answered by coarse-grained models, and necessitate a calculation with atomic resolution and complex energy function? The study on VSV-G serves as an example. All-atom and coarse-grained simulations concordantly demonstrated the high flexibility of the fusion loops of VSV-G, and suggested that the motion of the fusion loops is the initial step of the conformational change of VSV-G. But only the all-atom MD simulation could detect a directional deviation of the fusion loops from the equilibrium structure, whereas in the coarse-grained simulation, the fusion loops merely fluctuate around the equilibrium state. Thus, the electrostatic attraction or repulsion between charges is not sufficient to overcome the harmonic constraints of the elastic network. One obvious limitation of ENMs is that the bonds between nodes cannot break. Another limitation, when charge differences are the source of perturbation, is the neglect of the solvation effect. As I could show by electrostatic calculations, the solvation effect contributes critically to the motion of the fusion loops of VSV-G detected in all-atom MD. Inclusion of the solvation

effect and usage of a non-harmonic potential function are important for the proper simulation of the pH-dependent structural rearrangement of VSV-G. But the differential effects of antibiotic and nucleotide binding to APH can be explained by GNM, which showed that the rigidity and thus the architecture of dynamic protein domains allows for substrate-adjustable protein dynamics. The assignment of dynamic protein domains helps to understand protein dynamics, but could also serve as criterion to decide which bonds are allowed to break during the simulation of conformational changes of proteins. If changes of protein dynamics rely on topological properties, coarse-grained methods can successfully be applied to study the interplay between protein stability, flexibility and conformational changes, which is often hard to examine experimentally.

3.2 Contributions to the Joint Publications

Manuscript A

I performed all calculations presented in the manuscript. The results were interpreted by me together with Engin H. Serpersu and G. Matthias Ullmann. Most parts of the manuscript were written by me, supported by G. Matthias Ullmann. Engin H. Serpersu wrote the manuscript parts relating the theoretical results to experimental data.

Manuscript B

The electrostatic calculations and the coarse-grained MD simulation were performed by me. The conservation analysis and the all-atom MD simulation were performed by Pia Rücker. The results of my calculations were analyzed by me and G. Matthias Ullmann, and interpreted with regard to the all-atom simulations by Pia Rücker, Heinrich Sticht, G. Matthias Ullmann and me. I wrote the manuscript parts concerning the electrostatic calculations and the coarse-grained simulation.

Manuscript C

I performed and interpreted all calculations presented in the manuscript. The manuscript was written by me with the help of G. Matthias Ullmann.

Bibliography

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

[2] M F Perutz and F S Mathews. An x-ray study of azide methaemoglobin. *J. Mol. Biol.*, 21:199–202, 1966.

[3] M F Perutz. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228:726–739, 1970.

[4] S W Englander, N W Downer, and H Teitelbaum. Hydrogen exchange. *Annu. Rev. Biochem.*, 41:903–924, 1972.

[5] J R Lakowicz and G Weber. Quenching of protein fluorescence by oxygen. Detection of structural fluctuations in proteins on the nanosecond time scale. *Biochemistry*, 12:4171–4179, 1973.

[6] W C Jones, T M Rothgeb, and F R Gurd. Nuclear magnetic resonance studies of sperm whale myoglobin specifically enriched with 13C in the methionine methyl groups. *J. Biol. Chem.*, 251:7452–7460, 1976.

[7] R H Austin, K W Beeson, L Eisenstein, H Frauenfelder, I C Gunsalus, and V P Marshall. Activation energy spectrum of a biomolecule: photodissociation of carbonmonoxy myoglobin at low temperatures. *Phys. Rev. Lett*, 32:403–405, 1974.

[8] R H Austin, K W Beeson, L Eisenstein, H Frauenfelder, and I C Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14:5355–5373, 1975.

[9] H Frauenfelder, G A Petsko, and D Tsernoglou. Temperature-dependent x-ray diffraction as a probe of protein structural dynamics. *Nature*, 280:558–563, 1979.

[10] F Parak and H Formanek. Untersuchung des Schwingungsanteils und des Kristallgitterfehleranteils des Temperaturfaktors in Myoglobin durch Vergleich von Mössbauerabsorptionsmessungen mit Röntgenstrukturdaten. *Acta Cryst. A*, 27:573–578, 1971.

[11] H Keller and P G Debrunner. Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys. Rev. Lett*, 45:68–71, 1980.

[12] F Parak, E N Frolov, R L Mössbauer, and V I Goldanskii. Dynamics of metmyoglobin crystals investigated by nuclear gamma resonance absorption. *J. Mol. Biol.*, 145:825–833, 1981.

[13] F Parak and E W Knapp. A consistent picture of protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 81:7088–7092, 1984.

[14] A Ansari, J Berendzen, S F Bowne, H Frauenfelder, I E Iben, T B Sauke, E Shyamsunder, and R D Young. Protein states and proteinquakes. *Proc. Natl. Acad. Sci. U.S.A.*, 82:5000–5004, 1985.

[15] I E Iben, D Braunstein, W Doster, H Frauenfelder, M K Hong, J B Johnson, S Luck, P Ormos, A Schulte, P J Steinbach, A H Xie, and R D Young. Glassy behavior of a protein. *Phys. Rev. Lett*, 62:1916–1919, 1989.

[16] H Frauenfelder, B H McMahon, R H Austin, K Chu, and J T Groves. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc. Natl. Acad. Sci. U.S.A.*, 98:2370–2374, 2001.

[17] G P Singh, F Parak, S Hunklinger, and K Dransfeld. Role of adsorbed water in the dynamics of metmyoglobin. *Phys. Rev. Lett*, 47:685–688, 1981.

[18] J R Huck, G A Noyel, and L J Jorat. Dielectric properties of supercooled glycerol-water solutions. *IEEE Trans. Electr. Insul.*, 23:627–638, 1988.

[19] P W Fenimore, H Frauenfelder, B H McMahon, and F G Parak. Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc. Natl. Acad. Sci. U.S.A.*, 99:16047–16051, 2002.

[20] P W Fenimore, H Frauenfelder, B H McMahon, and R D Young. Bulk-solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and functions. *Proc. Natl. Acad. Sci. U.S.A.*, 101:14408–14413, 2004.

[21] T Kleinert, W Doster, H Leyser, W Petry, V Schwarz, and M Settles. Solvent composition and viscosity effects on the kinetics of CO binding to horse myoglobin. *Biochemistry*, 37:717–733, 1998.

[22] H Frauenfelder, N A Alberding, A Ansari, D Braunstein, B R Cowen, M K Hong, I E T Iben, J B Johnson, S Luck, M C Marden, J R Mourant, P Ormos, L Reinisch, R Scholl, A Schulte, E Shyamsunder, L B Sorensen, P J Steinbach, A Xie, R D Young, and K T Yue. Proteins and pressure. *J. Phys. Chem.*, 94:1024–1037, 1990.

[23] J B Johnson, D C Lamb, H Frauenfelder, J D Müller, B McMahon, G U Nienhaus, and R D Young. Ligand binding to heme proteins. VI. Interconversion of taxonomic substates in carbonmonoxymyoglobin. *Biophys. J.*, 71:1563–1573, 1996.

[24] K D Rector, J Jianwen, M A Berg, and M D Fayer. Effects of solvent viscosity on protein dynamics: infrared vibrational echo experiments and theory. *J. Phys. Chem. B*, 105:1081–1092, 2001.

[25] Y Shibata, A Kurita, and T Kushida. Solvent effects on conformational dynamics of Zn-substituted myoglobin observed by time-resolved hole-burning spectroscopy. *Biochemistry*, 38:1789–1801, 1999.

[26] J M Nocek, J S Zhou, S De Forest, S Priyadarshy, D N Beratan, J N Onuchic, and B M Hoffman. Theory and practice of electron transfer within protein-protein complexes: Application to the multidomain binding of cytochrome c by cytochrome c peroxidase. *Chem. Rev.*, 96:2459–2490, 1996.

[27] J H Roh, V N Novikov, R B Gregory, J E Curtis, Z Chowdhuri, and A P Sokolov. Onsets of anharmonicity in protein dynamics. *Phys. Rev. Lett*, 95:038101, 2005.

[28] G Durin, A Delaunay, C Darnault, D J Heyes, A Royant, X Vernede, C N Hunter, M Weik, and D Bourgeois. Simultaneous measurements of solvent dynamics and functional kinetics in a light-activated enzyme. *Biophys. J.*, 96:1902–1910, 2009.

[29] H Ohtaka and E Freire. Adaptive inhibitors of the HIV-1 protease. *Prog. Biophys. Mol. Biol.*, 88:193–208, 2005.

[30] J E Ladbury and B Z Chowdhry. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chem. Biol.*, 3:791–801, 1996.

[31] M J Stone. NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Acc. Chem. Res.*, 34:379–388, 2001.

[32] L Zídek, M V Novotny, and M J Stone. Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nat. Struct. Biol.*, 6:1118–1121, 1999.

[33] A L Lee, S A Kinnear, and A J Wand. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat. Struct. Biol.*, 7:72–77, 2000.

[34] S Gianni, T Walma, A Arcovito, N Calosci, A Bellelli, A Engström, C Travaglini-Allocatelli, M Brunori, P Jemth, and G W Vuister. Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure*, 14:1801–1809, 2006.

[35] C J Tsai, B Ma, and R Nussinov. Folding and binding cascades: shifts in energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 96:9970–9972, 1999.

[36] S Kumar, B Ma, C J Tsai, N Sinha, and R Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.*, 9:10–19, 2000.

[37] D E Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, 44:98–104, 1958.

[38] S Hayward. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *J. Mol. Biol.*, 339:1001–1021, 2004.

[39] B Ma, M Shatsky, H J Wolfson, and R Nussinov. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.*, 11:184–197, 2002.

[40] J Evenäs, S Forsén, A Malmendal, and M Akke. Backbone dynamics and energetics of a calmodulin domain mutant exchanging between closed and open conformations. *J. Mol. Biol.*, 289:603–617, 1999.

[41] K Okazaki and S Takada. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, 105:11182–11187, 2008.

[42] A J Wand. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nat. Struct. Biol.*, 8:926–931, 2001.

[43] A Cooper and D T Dryden. Allostery without conformational change. A plausible model. *Eur. Biophys. J.*, 11:103–109, 1984.

[44] C J Tsai, A del Sol, and R Nussinov. Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.*, 378:1–11, 2008.

[45] N Popovych, S Sun, R H Ebright, and C G Kalodimos. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.*, 13:831–838, 2006.

[46] A Panjkovich and X Daura. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct. Biol.*, 10:9–9, 2010.

[47] M Gerstein, A M Lesk, and C Chothia. Structural mechanisms for domain movements in proteins. *Biochemistry*, 33:6739–6749, 1994.

[48] D C Phillips. The hen egg-white lysozyme molecule. *Proc. Natl. Acad. Sci. U.S.A.*, 57:484–495, 1967.

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

[50] S Remington, G Wiegand, and R Huber. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. *J. Mol. Biol.*, 158:111–152, 1982.

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

[52] W L DeLano. The PyMOL Molecular Graphics System. *DeLano Scientific LLC, Palo Alto, CA, USA. http://www.pymol.org*, 2008.

[53] M Gerstein and W Krebs. A database of macromolecular motions. *Nucleic Acids Res.*, 26:4280–4290, 1998.

[54] S Roche, A A Albertini, J Lepault, S Bressanelli, and Y Gaudin. Structures of vesicular stomatitis virus glycoprotein: membrane fusion revisited. *Cell Mol. Life Sci.*, 65:1716–1728, 2008.

[55] J W Ponder and D A Case. Force fields for protein simulations. *Adv. Protein Chem.*, 66:27–85, 2003.

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

[57] S Swaminathan, T Ichiye, W van Gunsteren, and M Karplus. Time dependence of atomic fluctuations in proteins: analysis of local and collective motions in bovine pancreatic trypsin inhibitor. *Biochemistry*, 21:5230–5241, 1982.

[58] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235:318–321, 1987.

[59] R J Loncharich and B R Brooks. Temperature dependence of dynamics of hydrated myoglobin. Comparison of force field calculations with neutron scattering data. *J. Mol. Biol.*, 215:439–455, 1990.

[60] M Philippopoulos, A M Mandel, A G Palmer, and C Lim. Accuracy and precision of NMR relaxation experiments and MD simulations for characterizing protein dynamics. *Proteins*, 28:481–493, 1997.

[61] J Hartigan. *Clustering algorithms*. Wiley, New York, 1975.

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

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

[64] A Amadei, A B Linssen, and H J Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[65] E B Wilson, J C Decius, and P C Cross. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. McGraw-Hill, New York, 1955.

[66] M Tasumi, H Takeuchi, S Ataka, A M Dwivedi, and S Krimm. Normal vibrations of proteins: glucagon. *Biopolymers*, 21:711–714, 1982.

[67] T Noguti and N Go. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature*, 296:776–778, 1982.

[68] K Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

[69] B Brooks and M Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U.S.A.*, 80:6571–6575, 1983.

[70] N Go, T Noguti, and T Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. U.S.A.*, 80:3696–3700, 1983.

[71] B Brooks and M Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. U.S.A.*, 82:4995–4999, 1985.

[72] J F Gibrat and N Go. Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. *Proteins*, 8:258–279, 1990.

[73] O Marques and Y H Sanejouand. Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, 23:557–560, 1995.

[74] S Trakhanov, N K Vyas, H Luecke, D M Kristensen, J Ma, and F A Quiocho. 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*, 44:6597–6608, 2005.

[75] K Hinsen, A Thomas, and M J Field. Analysis of domain motions in large proteins. *Proteins*, 34:369–382, 1999.

[76] G Song and R L Jernigan. An enhanced elastic network model to represent the motions of domain-swapped proteins. *Proteins*, 63:197–209, 2006.

[77] S Hayward, A Kitao, and H J Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins*, 27:425–437, 1997.

[78] J Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373–380, 2005.

[79] T Ichiye and M Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11:205–217, 1991.

[80] T Horiuchi and N Go. Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins*, 10:106–116, 1991.

[81] B Melchers, E W Knapp, F Parak, L Cordone, A Cupane, and M Leone. Structural fluctuations of myoglobin from normal-modes, Mössbauer, Raman, and absorption spectroscopy. *Biophys. J.*, 70:2092–2099, 1996.

[82] A V Goupil-Lamy, J C Smith, J Yunoki, S F Parker, and M Kataoka. High-resolution vibrational inelastic neutron scattering: a new spectrocopic tool for globular proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 119:9268–9273, 1997.

[83] B Tidor and M Karplus. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J. Mol. Biol.*, 238:405–414, 1994.

[84] K Hinsen and G R Kneller. Solvent effects in the slow dynamics of proteins. *Proteins*, 70:1235–1242, 2008.

[85] A L Tournier and J C Smith. Principal components of the protein dynamical transition. *Phys. Rev. Lett*, 91:208106, 2003.

[86] J Trylska. Coarse-grained models to study dynamics of nanoscale biomolecules and their applications to the ribosome. *J. Phys. Condens. Matter*, 22:453101, 2010.

[87] A Benedix, C M Becker, B L de Groot, A Caflisch, and R A Böckmann. Predicting free energy changes using structural ensembles. *Nat. Methods*, 6:3–4, 2009.

[88] P Doruker, A R Atilgan, and I Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, 40:512–524, 2000.

[89] C Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.*, 17:1–6, 2007.

[90] A V Smith and C K Hall. Assembly of a tetrameric alpha-helical bundle: computer simulations on an intermediate-resolution protein model. *Proteins*, 44:376–391, Aug 2001.

[91] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.

[92] A R Atilgan, S R Durell, R L Jernigan, M C Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[93] N Go. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, 12:183–210, 1983.

[94] H J C Berendsen, D van der Spoel, and R van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91:43–56, 1995.

[95] H Kenzaki, N Koga, N Hori, R Kanada, W Li, K Okazaki, X Q Yao, and S Takada. CafeMol: a coarse-grained biomoleculer simulator for simulating proteins at work. *J. Chem. Theory Comput.*, 7:1979–1989, 2011.

[96] H J Limbach, A Arnold, B A Mann, and C Holm. ESPResSo – an extensible simulation package for research on soft matter systems. *Comput. Phys. Commun.*, 174:704–727, 2006.

[97] R K Tan, A S Petrov, and S C Harvey. Yup: A molecular simulation program for coarse-grained and multi-scaled models. *J. Chem. Theory Comput.*, 2:529–540, 2006.

[98] M Neri, C Anselmi, M Cascella, A Maritan, and P Carloni. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett*, 95:218102, 2005.

[99] Q Shi, S Izvekov, and G A Voth. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane-bound ion channel. *J. Phys. Chem. B*, 110:15045–15048, Aug 2006.

[100] O Kurkcuoglu, R L Jernigan, and P Doruker. Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions. *Polymer*, 45:649–657, 2004.

[101] O Kurkcuoglu, O T Turgut, S Cansu, R L Jernigan, and P Doruker. Focused functional dynamics of supramolecules by use of a mixed-resolution elastic network model. *Biophys. J.*, 97:1178–1187, 2009.

[102] E Lyman, F M Ytreberg, and D M Zuckerman. Resolution exchange simulation. *Phys. Rev. Lett*, 96:028105–028105, 2006.

[103] A P Heath, L E Kavraki, and C Clementi. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins*, 68:646–661, 2007.

[104] Q Yang and K A Sharp. Building alternate protein structures using the elastic network model. *Proteins*, 74:682–700, 2009.

[105] A Ahmed and H Gohlke. *Multi-scale modeling of macromolecular conformational changes*. 1st international conference on mathematical and computational biomedical engineering, CMBE2009, 2009.

[106] Z Zhang, Y Shi, and H Liu. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J.*, 84:3583–3593, 2003.

[107] M M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett*, 77:1905–1908, 1996.

[108] P. J. Flory. Statistical thermodynamics of random networks. *Proc. R. Soc. A*, 351:351–378, 1976.

[109] I Bahar and R L Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266:195–214, 1997.

[110] M Ye, F Shima, S Muraoka, J Liao, H Okamoto, M Yamamoto, A Tamura, N Yagi, T Ueki, and T Kataoka. Crystal structure of M-ras reveals a GTP-bound "off" state conformation of ras family small GTPases. *J. Biol. Chem.*, 280:31267–31275, 2005.

[111] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[112] E Eyal, S Gerzon, V Potapov, M Edelman, and V Sobolev. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.*, 351:431–442, 2005.

[113] L W Yang, A J Rader, X Liu, C J Jursa, S C Chen, H A Karimi, and I Bahar. oGNM: online computation of structural dynamics using the gaussian network model. *Nucleic Acids Res.*, 34:24–31, 2006.

[114] S Kundu, J S Melton, D C Sorensen, and G N Phillips. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.*, 83:723–732, 2002.

[115] K Hinsen. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24:521–528, 2008.

[116] G Song and R L Jernigan. vGNM: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.*, 369:880–893, 2007.

[117] N A Temiz, E Meirovitch, and I Bahar. Escherichia coli adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)N-NMR relaxation data. *Proteins*, 57:468–480, 2004.

[118] I Bahar, A Wallqvist, D G Covell, and R L Jernigan. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry*, 37:1067–1075, 1998.

[119] L W Yang, E Eyal, C Chennubhotla, J Jee, A M Gronenborn, and I Bahar. Insights into equilibrium dynamics of proteins from comparison of NMR and x-ray data with computational predictions. *Structure*, 15:741–749, 2007.

[120] L W Yang and C P Chng. Coarse-grained models reveal functional dynamics–I. Elastic network models–theories, comparisons and perspectives. *Bioinform. Biol. Insights*, 2:25–45, 2008.

[121] E Eyal, C Chennubhotla, L W Yang, and I Bahar. Anisotropic fluctuations of amino acids in protein structures: insights from x-ray crystallography and elastic network models. *Bioinformatics*, 23:175–184, 2007.

[122] D A Kondrashov, A W Van Wynsberghe, R M Bannen, Q Cui, and G N Phillips. Protein structural variation in computational models and crystallographic data. *Structure*, 15:169–177, 2007.

[123] M Delarue and Y H Sanejouand. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.*, 320:1011–1024, 2002.

[124] F Tama, M Valle, J Frank, and C L Brooks. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl. Acad. Sci. U.S.A.*, 100:9319–9323, 2003.

[125] Y Wang, A J Rader, I Bahar, and R L Jernigan. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.*, 147:302–314, 2004.

[126] C Xu, D Tobi, and I Bahar. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T↔R2 transition. *J. Mol. Biol.*, 333:153–168, 2003.

[127] W G Krebs, V Alexandrov, C A Wilson, N Echols, H Yu, and M Gerstein. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, 48:682–695, 2002.

[128] V Alexandrov, U Lehnert, N Echols, D Milburn, D Engelman, and M Gerstein. Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. *Protein Sci.*, 14:633–643, 2005.

[129] F Tama and Y H Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, 14:1–6, 2001.

[130] P Maragakis and M Karplus. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, 352:807–822, 2005.

[131] C W Müller and G E Schulz. Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 A resolution. A model for a catalytic transition state. *J. Mol. Biol.*, 224:159–177, Mar 1992.

[132] C W Müller, G J Schlauderer, J Reinstein, and G E Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4:147–156, Feb 1996.

[133] C Micheletti, P Carloni, and A Maritan. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins*, 55:635–645, 2004.

[134] M Rueda, P Chacón, and M Orozco. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, 15:565–575, 2007.

[135] S Nicolay and Y H Sanejouand. Functional modes of proteins are among the most robust. *Phys. Rev. Lett*, 96:078104, 2006.

[136] P Durand, G Trinquier, and YH Sanejouand. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, 34:759–771, 1994.

[137] F Tama, F X Gadea, O Marques, and Y H Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41:1–7, 2000.

[138] P Doruker, R L Jernigan, and I Bahar. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.*, 23:119–127, 2002.

[139] P Doruker and R L Jernigan. Functional motions can be extracted from on-lattice construction of protein structures. *Proteins*, 53:174–181, 2003.

[140] M Lu and J Ma. The role of shape in determining molecular motions. *Biophys. J.*, 89:2395–2401, 2005.

[141] D Ming and M E Wall. Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett*, 95:198103–198103, 2005.

[142] L Yang, G Song, and R L Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophys. J.*, 93:920–929, 2007.

[143] K Hamacher and J A McCammon. Computing the amino acid specificity of fluctuations in biomolecular systems. *J. Chem. Theory Comp.*, 2:873–878, 2006.

[144] B Erman. The gaussian network model: precise prediction of residue fluctuations and application to binding problems. *Biophys. J.*, 91:3589–3599, 2006.

[145] J W Chu and G A Voth. Coarse-grained modeling of the actin filament derived from atomistic-scale simulations. *Biophys. J.*, 90:1572–1582, 2006.

[146] K Moritsugu and J C Smith. Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance hessian. *Biophys. J.*, 93:3460–3469, 2007.

[147] E Lyman, J Pfaendtner, and G A Voth. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys. J.*, 95:4183–4192, 2008.

[148] D A Kondrashov, Q Cui, and G N Phillips. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys. J.*, 91:2760–2767, 2006.

[149] K Suhre and Y H Sanejouand. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, 32:610–614, 2004.

[150] M T Zimmermann, A Kloczkowski, and R L Jernigan. MAVENs: motion analysis and visualization of elastic networks and structural ensembles. *BMC Bioinformatics*, 12:264–264, 2011.

[151] A Bakan, L M Meireles, and I Bahar. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*, 27:1575–1577, 2011.

[152] K Hinsen. The molecular modeling toolkit: a new approach to molecular simulations. *J. Comput. Chem.*, 21:79–85, 2000.

[153] A Górecki, M Szypowski, M Dlugosz, and J Trylska. RedMD–reduced molecular dynamics package. *J. Comput. Chem.*, 30:2364–2373, 2009.

[154] S O Yesylevskyy, V N Kharkyanen, and A P Demchenko. Hierarchical clustering of the correlation patterns: new method of domain identification in proteins. *Biophys. Chem.*, 119:84–93, 2006.

[155] S Mitternacht and I N Berezovsky. Coherent conformational degrees of freedom as a structural basis for allosteric communication. *PLoS Comput. Biol.*, 7, 2011.

[156] M K Kim, R L Jernigan, and G S Chirikjian. Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.*, 83:1620–1630, 2002.

[157] J W Chu and G A Voth. Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophys. J.*, 93:3860–3871, 2007.

[158] W Zheng, B R Brooks, and G Hummer. Protein conformational transitions explored by mixed elastic network models. *Proteins*, 69:43–57, 2007.

[159] M Tekpinar and W Zheng. Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins*, 78:2469–2481, 2010.

[160] M Delarue and P Dumas. On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc. Natl. Acad. Sci. U.S.A.*, 101:6957–6962, 2004.

[161] K Hinsen, N Reuter, J Navaza, D L Stokes, and J J Lacapère. Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys. J.*, 88:818–827, 2005.

[162] C Gorba and F Tama. Normal mode flexible fitting of high-resolution structures of biological molecules toward SAXS data. *Bioinform. Biol. Insights*, 4:43–54, 2010.

[163] P Gniewek, A Kolinski, R L Jernigan, and A Kloczkowski. Elastic network normal modes provide a basis for protein structure refinement. *J. Chem. Phys.*, 136:195101–195101, 2012.

[164] C Özen and E H Serpersu. Thermodynamics of aminoglycoside binding to aminoglycoside-3'-phosphotransferase IIIa studied by isothermal titration calorimetry. *Biochemistry*, 43:14667–14675, 2004.

[165] A L Norris and E H Serpersu. NMR detected hydrogen-deuterium exchange reveals differential dynamics of antibiotic- and nucleotide-bound aminoglycoside phosphotransferase 3'-IIIa. *J. Am. Chem. Soc.*, 131:8587–8594, 2009.

[166] B L de Groot, D M van Aalten, R M Scheek, A Amadei, G Vriend, and H J Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29:240–251, 1997.

[167] B Halle. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 99:1274–1279, 2002.

[168] D J Jacobs, A J Rader, L A Kuhn, and M F Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44:150–165, 2001.

Chapter 4

Manuscript A

ATP Binding Enables Broad Antibiotic Selectivity of Aminoglycoside Phosphotransferase(3′)-IIIa: An Elastic Network Analysis

Silke A. Wieninger, Engin H. Serpersu and G. Matthias Ullmann
J. Mol. Biol. 2011, 409, 450-465
doi: 10.1016/j.jmb.2011.03.061

Chapter 5

Manuscript B

pH-dependent Molecular Dynamics of Vesicular Stomatitis Virus Glycoprotein G

Pia Rücker, Silke A. Wieninger, G. Matthias Ullmann
and Heinrich Sticht
Proteins 2012, 80, 2601-2613
doi: 10.1002/prot.24145

Chapter 6

Manuscript C

CovarDom: Identifying Dynamic Protein Domains based on Covariance Matrices of Motion

Silke A. Wieninger and G. Matthias Ullmann

to be submitted

CovarDom: Identifying Dynamic Protein Domains based on Covariance Matrices of Motion

Silke A. Wieninger and G. Matthias Ullmann

to be submitted

Abstract

Complex protein structures are frequently classified into separate domains to facilitate the study of protein folding, dynamics and function. Still, domain assignments are often based on subjective criteria and not unique. We describe the program package CovarDom, which assigns protein domains automatically based on the dynamical behavior of the protein residues. The dynamic input data in form of covariances of residue fluctuations is calculated by a Gaussian network model. A program called DomainClusterer determines the domain boundaries, while a second program, DomainTester, decides the usually more difficult question if the protein or part of the protein actually consists of several dynamic domains. Comparison of the dynamic domains to structural domains assigned by the authors of protein structures for a large set of proteins demonstrates analogies and differences between the two approaches. Dynamic and structural domains coincide if proteins consist of clearly separated parts. But in contrast to structural domains, dynamic domains are often discontinuous in sequence, and small groups of residues can belong to another dynamic domain than their sequential neighbors. Application of CovarDom to the enzymes 4-hydroxyphenylacetate decarboxylase and acetylene hydratase shows the importance of these properties of dynamic protein domains for the functionality of the enzymes.

Introduction

Many large proteins are composed of several domains. Even if there is no unique definition of protein domains, they are usually considered as quasi-independent compact structural units. Depending on whether protein evolution, stability or function is investigated, the domain assignment may differ. Evolutionary defined domains are considered as functional building blocks, which can recombine on the genetic level to proteins with different functions.[1] Accordingly, evolutionary defined domains are only formed by one continuous sequence segment of the protein chain. They are expected to fold independently and often carry out a special function, such as DNA binding or phosporylation. In contrast, domains assigned on the basis of protein structures, called structural domains, can be sequentially discontinuous and rely on criterions like a compact structural appearance or the presence of a hydrophobic core. The partitioning of proteins into weakly interacting compact units allows for conformational changes at low energetic cost.[2] Therefore, structural domain assignments are frequently used to predict functional motions of proteins, which are important for biological processes like signal transduction and ligand binding. The assignments are often performed manually, but there is also a wide range of automatic methods, which use the generally accepted principle that intradomain interactions are stronger than interdomain interactions. The programs evaluate the number of domain contacts,[3–7] the similarity of contact environments,[8] the distribution of hydrophobic cores[9] and secondary structure elements[10] or van der Waals energy profiles[11] to determine the boundaries of structural domains.

Instead of inferring protein dynamics from structural domains, one can directly define dynamic domains based on concerted motions of amino acid residues.[12] Such dynamic domains can deviate from structural domains if the protein structure is not clearly divided into separate parts. Then the interplay between the interactions of different protein parts and the compactness and size of the protein parts themselves is too complex to predict protein dynamics just by viewing. Dynamic domains can be used to analyze potential large-scale protein motions or the effect of ligand binding and oligomerization on protein dynamics. In a previous study on the enzyme aminoglycoside phosphotransferase 3′-IIIa,[13] we showed that binding of substrates between different dy-

namic domains leads to either stabilization or destabilization, depending on the architecture of the involved dynamic domains. Besides, dynamic domains can help to identify perturbation-sensitive sites of proteins, where addition or removal of a few interactions leads to large changes of protein dynamics. Input data for the identification of dynamic domains can originate from principal components deduced from a molecular dynamics simulation[14] or from normal mode analysis. The large-amplitude principal components or normal modes describe global protein movements and allow to identify residues belonging to the same quasi-rigid domain based on the directions of motion. Different methods exist which identify rigid protein parts analyzing one[15] or several low-frequency normal modes[16,17] calculated by an elastic network model (ENM).[18,19] The ENM uses purely topological constraints deduced from the protein structure to determine single-residue fluctuations and collective protein motions. To consider the contributions of all normal modes, one can use covariances of motion as input data, as described by Yesylevskyy et al.[20,21]

The here described method CovarDom also clusters covariances of residue motion to predict dynamic domains. In contrast to the work of Yesylevskyy et al., where the number of domains is determined based on the largest correlation difference between two clustering steps, CovarDom implements a separate method, which checks whether a protein or protein part actually consists of several domains. As input data, CovarDom only depends on the connectivity of the residues and on the covariance matrix, calculated for one protein conformation by an optional simulation method. In this work, we calculate the covariance matrices by a Gaussian network model (GNM),[18] one variant of the ENM. Other than most domain assignment methods, CovarDom does not use any postprocessing steps to alter unexpected domain classifications after the actual assignment procedure. The dynamic domains are allowed to be discontinuous and to include small fragments. Secondary structure elements and the strands of β-sheets can be spread over several dynamic domains.

In the following, we describe the algorithms used by the programs DomainTester and DomainClusterer, as well as the overall workflow of CovarDom. Besides CovarDom, a slightly different approach, CovarZeroDom, is introduced, which employs an alternative stopping criterion of the clustering algorithm. We compare our predictions to manual domain assignments for a dataset of 135 proteins and investigate analogies and discrepancies between

the approaches. We investigate the influence of GNM parameters and parameters of DomainTester on the domain assignments and compare the domain assignments of CovarDom and CovarZeroDom. Finally, we show on the examples of 4-hydroxyphenylacetate decarboxylase[22] and acetylene hydratase[23] how the dynamic domains can help to understand protein functionality.

Theory

The domain identification is based on dynamical information in form of covariance matrices, which we determine using a Gaussian Network Model (GNM), as described in the Methods section. But the covariance matrices could as well be obtained from the anisotropic network model,[19] an all-atom normal mode analysis or a principal component analysis of molecular dynamics simulations. The covariance matrix is a symmetric $N \times N$ matrix, and the sum over all entries of the covariance matrix equals zero, because translational and rotational motions are described by the eigenvectors with zero eigenvalues, which are excluded.[24] The sum over all correlations, which are normalized covariances, is not zero anymore. Thus we use covariances instead of correlations as similarity measure in the agglomerative clustering procedure. Because the clustering program DomainClusterer is not able to distinguish between 1-domain and multidomain proteins, the program DomainTester is needed to check if the structure can be partitioned. If DomainTester detects several domains, DomainClusterer performs an agglomerative clustering of the residues into domains. In the following description of the algorithms, the term domain is only used for the final residue partition. The term cluster is used for groups of residues which have to be combined or split to become domains.

DomainTester: Differentiation between 1-Domain and Multidomain Proteins

Distinguishing 1-domain proteins from multidomain proteins is a crucial part of the domain identification procedure. Figure 1 shows the obvious differences which exist between covariance matrices of 1-domain and multidomain proteins. We need to find rules that describe these differences and allow for a computational evaluation. Covariance matrices of multidomain proteins have

large sequential areas of positive values, in contrast to covariance matrices of 1-domain proteins. Therefore we search for regions of at least $s_{\min}$ nodes with positive covariance cov_{ij} between all pairs of nodes, that is

$$cov_{ij} > 0 \text{ for all } i,j \in [k,l] \quad \text{and} \quad l-k > s_{\min}. \tag{1}$$

We call these regions positive-covariance segments. The segments are usually overlapping, meaning that one node is part of several segments. $\langle x \rangle_{\text{seg}}$ is the fraction of nodes that belong to at least one positive-covariance segment. If $\langle x \rangle_{\text{seg}} < \langle x \rangle_{\text{seg}}^{\min}$, meaning that less than a fraction of $\langle x \rangle_{\text{seg}}^{\min}$ of the nodes can be grouped into positive-covariance segments, the protein is considered as 1-domain protein. Additionally, we request that the number of non-overlapping positive-covariance segments, n_{seg}, is at least two for a multidomain protein. In summary, for being a multidomain protein, the following two criteria must be met:

$$n_{\text{seg}} \geq 2 \quad \text{and} \quad \langle x \rangle_{\text{seg}} \geq \langle x \rangle_{\text{seg}}^{\min}. \tag{2}$$

DomainClusterer: Agglomerative Clustering of Covariances

First, every node builds one cluster. Then the two clusters with highest positive covariance to each other are merged into one cluster. We denote the two clusters with highest intercluster covariance as a and b. They comprise N_a and N_b nodes, and their covariance is denoted as cov_{ab}. The covariance of cluster a to any other cluster i is cov_{ai}. Let us denote the merged cluster consisting of nodes a and b as a'. The covariance of the new cluster a' to any other cluster i is calculated by averaging over the covariances of cluster a and b to cluster i:

$$\text{cov}_{a'i} = \frac{\text{cov}_{ai} \cdot N_a + \text{cov}_{bi} \cdot N_b}{N_a + N_b} \tag{3}$$

The new cluster a' consists of $N_a + N_b$ nodes, and the total number of clusters is reduced by one. The intracluster covariance of the merged cluster is $\text{cov}_{a'a'}$. It is given by the average over all covariances within the cluster:

$$\text{cov}_{a'a'} = \frac{\text{cov}_{aa} \cdot N_a^2 + \text{cov}_{bb} \cdot N_b^2 + \text{cov}_{ab} \cdot 2N_aN_b}{N_a^2 + N_b^2 + 2N_aN_b} \tag{4}$$

The stepwise execution of Eqs. 3 and 4 has the same effect as clustering all nodes at once and calculating the inter- and intracluster covariances by the

arithmetic mean over the covariances of all cluster nodes. As we do not know in advance which nodes belong to which cluster, we have to assign the covariances in this stepwise manner. An example of the clustering algorithm can be found in Figure 2A. With n giving the number of clusters, the relation

$$\sum_{a=1}^{n} \sum_{b=1}^{n} \mathrm{cov}_{ab} \cdot N_a N_b = 0 \tag{5}$$

is true after each step, because the sum over all entries of the covariance matrix equals zero. The program stops either when a certain number of clusters is reached, or when the highest intercluster covariance is smaller than a given cutoff value, which we typically set to zero.

CovarDom and CovarZeroDom

The two approaches CovarDom and CovarZeroDom both use the programs DomainTester and DomainClusterer to predict dynamic domains, but differ in the stopping criterion of the clustering procedure. Figures 2C shows the interplay of the different programs in CovarDom and CovarZeroDom. In CovarDom, the program DomainClusterer merges residues until they are divided in two clusters. After each splitting, the program DomainTester is used again to check if the two clusters are in turn composed of several clusters, and if so, new covariance matrices are calculated by GNM. This approach assumes that the motions of the dynamic domains are independent, and can be calculated separately for each cluster. The final dynamic domains are arranged into different hierarchical levels, and one can easily see from the domain numbering which domains are split first and which domains are split in a later step, meaning that they are less anticorrelated to each other (Figure 2B). In contrast, CovarZeroDom uses DomainTester only in the very beginning to check if the protein consists of several domains, and calculates the covariance matrix only once. The program DomainClusterer stops merging residues when the largest intercluster covariance is negative. The final domain number corresponds to the number of remaining clusters. The advantage of CovarZeroDom is that no covariance matrices of splitted protein structures must be calculated, which might be impossible using for example Molecular Dynamics for the generation of dynamical data.

One should keep in mind that the positive-covariance segments of Domain-

Tester do not necessarily coincide with parts of the final protein domains. The final domains are allowed to be smaller than $\langle x \rangle_{\mathrm{seg}}^{\mathrm{min}}$, the minimal fraction of nodes which belong to positive-covariance segments, and can be discontinuous. While DomainTester makes use of the sequential information, in DomainClusterer the sequential information is neglected.

Methods

Gaussian Network Model

The Gaussian Network Model (GNM) is a coarse-grained method which uses the atom coordinates of the protein to build a network consisting of one or several nodes per residue.[13,18] The nodes are connected covalently if they represent sequential residues. Nodes representing non-sequential residues are only connected if their distance to each other is smaller than a certain cutoff radius r_{cut}. The potential energy V of the network is given by

$$V = \sum_{i,j=i}^{N} \Theta(r_{\mathrm{cut}} - r_{ij}^{\circ}) \frac{\gamma_{ij}}{2} \left((\Delta x_i - \Delta x_j)^2 + (\Delta y_i - \Delta y_j)^2 + (\Delta z_i - \Delta z_j)^2 \right) \quad (6)$$

where γ_{ij} is the covalent or non-covalent force constant, depending on the connection between nodes i and j. $\Theta(x)$ is the Heaviside step function and is 1 if nodes i and j are connected, and 0 otherwise. The energy function penalizes distortions from the equilibrium coordinates r_{ij}° of the experimental structure by summation over pairwise energy terms. GNM allows calculation of variances $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_i \rangle$ and covariances $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle$ of residue fluctuations, which are evaluated from the diagonal and off-diagonal elements of the pseudo-inverse Kirchhoff matrix $\tilde{\Gamma}^{-1}$, respectively.

$$\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle = \frac{3k_B T}{\gamma_{ij}} (\tilde{\Gamma}^{-1})_{ij} \quad (7)$$

The number of nodes per residue, the ratio between covalent and non-covalent force constants and the cutoff radius are parameters of the GNM which can be optimized by comparison of the theoretical atom flexibilities to experimental B-factors from x-ray crystallography.[25] Covariance matrices of residue motion indicate which residues tend to move simultaneously into the same direction (positive covariance values), and which residues are anticorrelated to each

other (negative covariance values). As protein domains have only few connections to each other, they can move apart at low energetic cost, while residues of the same domain stick together. Therefore, covariance matrices can be used to identify the number and boundaries of domains present in a protein.

Computational Details

In the GNM, amino acids are represented by one node at the C_α position. If cofactors are present in the crystal, they influence the dynamics. Hence, they should also be included in the elastic network. The number of nodes representing cofactors is set depending on the size of the molecule. The GNM cutoff radius is varied from 6 to 11 Å. The force constant for non-covalent interactions adopts values of 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 kcal (mol Å^2)$^{-1}$, while the covalent force constant is fixed to 10 kcal (mol Å^2)$^{-1}$. The ratio between covalent and non-covalent force constant influences the covariances of residue fluctuations, while the absolute values of the force constants only influence the absolute scale of the fluctuations. As default parameters for DomainTester calculations, we use a minimal size for positive covariance segments of $s_{\min} = 40$, and a minimal fraction of nodes which must belong to positive-covariance segments of $\langle x \rangle_{\text{seg}}^{\min} = 0.5$. The covariance plots, protein images and charts were produced using GMT,[26] PyMOL[27] and gnuplot, respectively.

Protein Dataset and Evaluation

The domain identification algorithm is applied to proteins of the benchmark dataset 3 from the pDomains website,[28] which offers domain information about 135 proteins. Table 1 of the Supporting Information gives the PDB codes and the number of assigned structural domains for all proteins of the dataset. The dataset is constructed based on domain assignments from methods which are not or at least not fully automatic. SCOP[29] relies on structural and evolutionary relationships between proteins, CATH[30] classifies proteins according to their structure by a combination of automatic and manual procedures, and AUTHORS collects assignments of the authors of protein structures. In the following, we refer to these methods as manual assignments, and call the resulting domains structural domains. For proteins of the dataset, the three manual methods agree about the number of domains and at least to 90% about the

domain boundaries. For comparing dynamic domains to structural domains, we use the domain boundaries assigned by the authors of the structures. As measure for the similarity between domain boundaries assigned by different methods, the percentage of domain overlap[31] is used. For each pair of domains identified by two different methods, the number of common residues is determined. The best combination of domains from the different methods corresponds to the combination with the highest sum of matching residues. If the domain numbers assigned by the two methods differ, the spare domains remain unpaired. The number of matching residues is divided by the total number of protein residues assigned to domains. This total residue number can differ between CovarDom and the manual assignments, because more than one quarter of the multidomain proteins of the dataset has missing C_α coordinates within the protein chain, which are not assigned to domains by CovarDom, in contrast to some manual assignments. On the other hand, CovarDom assigns each residue present in the ENM to a domain, while expert methods sometimes leave out residues. Cofactors are represented by several nodes in the ENM, which can be assigned to different domains by CovarDom. In manual assignments, they are not considered at all. Therefore, they do not count to the number of residues assigned to domains.

Results and Discussion

Comparison between Structural and Dynamic Protein Domains

We determine the dynamic domains of the 135 proteins of the pDomains dataset.[28] The covariance matrices are calculated using the Gaussian network model (GNM) with a cutoff radius of 7 and force constants of 10 kcal $(\mathrm{mol}^2)^{-1}$ for covalent interactions and 5 kcal $(\mathrm{mol}^2)^{-1}$ for non-covalent interactions. In total, 106 of the 135 proteins are split into the same number of domains manually and by CovarDom, including the assignments as 1-domain protein. Because five of the eighty proteins assigned as multidomain manually are assigned as 1-domain protein by CovarDom, seventy-five multidomain proteins remain for calculating the average percentage of domain overlap. 64% of these proteins have a percentage of domain overlap higher than 90%. If we only consider

the fifty-six multidomain proteins with the same number of structural and dynamic domains, 85.7% have an average percentage of domain overlap higher than 90%. These data show that for most proteins which are classified into the same number of structural and dynamic domains, also the domain boundaries are alike. Although the assignments of structural and dynamic domains are based on different criteria, the agreement in domain number and boundaries for many proteins indicates that the underlying principal ideas are the same. A small number of interdomain contacts is reflected in the dynamics of multidomain proteins. Thus, structural and dynamic domains coincide if the domains are clearly separated from each other. Figure 3 depicts proteins with corresponding domain assignments by manual methods and CovarDom, meaning that the domain number is equal and the domain overlap is higher than 90%. Still, there are small differences. α-Helices and β-strands can be spread over two dynamic domains, as in 5′-nucleotidase[32] (Figure 3B) and CryIA(a) toxin from *Bacillus thuringiensis*[33] (Figure 3C). In manual domain assignments, such secondary structure elements lying between two domains like the α-helix of 5′-nucleotidase usually remain unassigned. Another discrepancy between the methods is that CovarDom occasionally clusters small groups of residues to another dynamic domain than their neighbors. Often, these are loop residues, as in aminopeptidase P[34] (Figure 3A) and neuraminidase from *Vibrio cholerae*[35] (Figure 3D), but also small α-helices can be dynamically coupled to another domain, as in CryIA(a). Other automatic domain assignment methods usually change the assignment of such residues in a postprocessing step. But the location of such residues can give information about the interactions between dynamic domains and possible hinge regions.

In the following, we analyze which structural properties can lead to larger differences between dynamic and structural domains. Figure 4 shows the dynamic domains and covariance matrices of the protein Rab geranylgeranyltransferase.[36] It is partitioned into three domains by the authors of the structure and into six dynamic domains by CovarDom. The dynamic domains belong to different hierarchies and demonstrate how CovarDom creates dynamic domains through iterative splitting of the structure and recalculation of covariance matrices. The first CovarDom splitting step already cuts in the middle of the large structural domain assigned by the authors (Figure 4C). Although this structural domain can be classified as evolutionary domain, be-

cause its helical fold is also found in other proteins,[36] from a dynamical view its residues clearly belong to at least two different dynamic domains, as one can recognize in the covariance matrix of the whole protein (Figure 4A top). Further splitting leads to classification of this structural domain into four dynamic domains. The manual assignment of more than three hundred residues to one huge structural domain could, besides the evolutionary aspect, also be induced by the similar arrangement of the α-helices forming the structural domain. Human prediction tends to identify conspicuous protein folds as structural domains. Figure 5 shows further examples of such protein architectures. The iron-sulfur protein of carbon monoxide dehydrogenase[37] (Figure 5A) is assigned as 1-domain protein by CovarDom, but as 2-domain protein manually. In contrast, neuraminidase N9 of influenza virus[38] (Figure 5B) and nitrous oxide reductase[39] (Figure 5C) consist of more dynamic domains than assigned manually. The recurrence of folding patterns, as in iron-sulfur protein, and the similar arrangement of secondary structures, as in Rab geranylgeranyltransferase, neuraminidase and nitrous oxide reductase, seem to lead to under- and overestimation, respectively, of the contacts between the residues. The interrelation between the number of connections between two protein parts and the size and compactness of the protein parts themselves is too complex to be predicted just by visual inspection. On the one hand, protein parts with correlated movement are not always visible as clearly separated domains. On the other hand, connections between compact regions of different structural domains can impede the independent movement. At sensitive sites, even small changes in the elastic network connections can have large effects on the covariances.[13] For example, the binding of cofactors can change the dynamic properties of a protein, as in flavohemoglobin[40] (Figure 5D). It is assigned as 2-domain protein by CovarDom if FAD and heme are bound to it. Manually, three domains are assigned. Interestingly, also CovarDom assigns three dynamic domains if heme and FAD are neglected. But they are part of the functional enzyme and influence the dynamic behavior, thus they should be included in the calculation. This view is confirmed by a study on high-resolution X-ray structures, which showed that adding ligands and cofactors to a GNM improves the correlation between theoretical and experimental B-factors.[41]

An example of a protein for which the manual domain assignment seems more plausible than the one resulting from CovarDom is cytochrome f, which

is assigned as 1-domain protein by CovarDom, but consists of two structural domains according to the authors.[42] By visual inspection of the structure and the covariance matrix (Figure 6), one would agree that cytochrome f consists of two dynamic domains, because two separated, clearly anticorrelated protein parts exist. The residues of the smaller structural domain are highly positively correlated, and show a strong anticorrelation to most of the residues of the larger structural domain. Also the residues of the larger structural domain are dynamically coupled, but the sequence of residues building the central β-sheet is disrupted by long loops, short α-helices and the residues of the small domain. Therefore, the corresponding positive-covariance segment includes only forty-three residues. In total, the positive-covariance segments comprise 43% of the residues, which leads to the classification as 1-domain protein if default values are used in CovarDom. By lowering the required fraction of nodes in positive-covariance segments from 0.5 to 0.4, also CovarDom assigns two domains to cytochrome f.

Influence of GNM Parameters on Dynamic Protein Domains

The identification of dynamic protein domains is not only influenced by the parameters of DomainTester, but also by the GNM parameters chosen for the calculation of the covariance matrix. To investigate the influence of GNM parameters, we vary the cutoff radius from 6 to 11 and the force constant for non-covalent interactions from 0.1 to 10 $\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$. The covalent force constant is fixed to 10 $\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$. First, we examine the influence of GNM parameters on the classification as 1-domain or multidomain protein by the program DomainTester. A larger cutoff radius and a higher non-covalent force constant lead to a higher ratio of proteins assigned as 1-domain protein (Figure 7). The dependence of the 1-domain prediction on GNM parameters can be understood by looking at covariance matrices of phophatidylinositol transfer protein,[43] calculated for different parameter pairs (Figure 8). Depending on the parameters, one to three dynamic domains are assigned. For a cutoff radius of 7 and a non-covalent force constant of 5 $\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$, two dynamic domains are assigned by CovarDom which agree with the domains assigned manually. For smaller non-covalent force constants, the covariance matrices are less scattered and there is a broad zone of positive covariance along the di-

Table 1. Combinations of cutoff radii r_{cut} and non-covalent force constants k_{ncov} of parameter set 13. These GNM parameter pairs lead to an accordance of at least 90% in the classification as 1-domain or multidomain protein by manual methods and DomainTester (see Figure 9A).

r_{cut} []	7						8				9		10
k_{ncov} [kcal $(\mathrm{mol}^2)^{-1}$]	5	6	7	8	9	10	1	2	3	4	0.5	1	0.5

agonal, which allows for the detection of many positive-covariance segments, whereas for high cutoff radii, the covariance matrices are very fragmented, such that only few positive-covariance segments remain.

Figure 9A shows the percentage of 1-domain or multidomain proteins, according to manual predictions, which are assigned correspondingly by CovarDom. Obviously, at a higher ratio of 1-domain proteins, more proteins which are assigned as 1-domain manually are also assigned as 1-domain by CovarDom. The opposite is true for multidomain proteins. Although it is not our primary goal to reproduce the assignments of structural domains, we use the pDomains dataset to adjust our program parameters, because this standardization helps to figure out the essential differences between structural and dynamic domains. For good agreement between manual methods and CovarDom, a compromise between the contrary trends for 1-domain and multidomain proteins must be found. Thirteen GNM parameter pairs which lead to an accordance of at least 90% for both 1-domain and multidomain proteins lie at the intersection between the two curves in Figure 9A. The cutoff radii and non-covalent force constants of this parameter set, which we call set 13, are given in Table 1. In contrast, set 72 denotes the full test set with all seventy-two possible combinations of cutoff radii and non-covalent force constants.

For five proteins of the pDomains dataset, the domain numbers assigned manually and by CovarDom differ for all seventy-two parameter pairs. One example is the assignment of three structural domains to Rab geranylgeranyltransferase (Figure 4), whereas CovarDom assigns five dynamic domains for 56% of the parameter pairs, six domains for 31% of the parameter pairs, and seven or even eight dynamic domains for the remaining parameter pairs. For twelve multidomain proteins of the pDomains dataset, the domain number agrees for all GNM parameter pairs of set 72 (PDB codes 1a8y, 1au7, 1b24, 1cun, 1eif, 1grj, 1lck, 1prt, 1tbr, 1urk, 1vol, 2cgp). But for many proteins, the assigned

domain numbers agree only for some GNM parameter pairs, as shown previously for phophatidylinositol transfer protein (Figure 8). In some cases, the number of domains assigned by CovarDom agrees with the manually defined domains when small cutoff radii are used, while for others, agreement occurs preferentially using large cutoff radii. Is there a way to determine suitable parameters separately for each protein? A possibility to distinguish 1-domain from multidomain proteins is to choose the assignment which occurs most often for the different GNM parameter pairs. We refer to this procedure as frequency approach. Another approach, which additionally allows to determine domain numbers and boundaries, is to compare theoretical to crystallographic B-factors[44–46] using the linear correlation coefficient.[25] Table 1 of the Supporting Information gives the GNM parameter pairs determined by comparison to experimental B-factors out of set 72 and set 13 and the corresponding number of dynamic domains assigned by CovarDom. Figure 9B compares the consensus between the manual and DomainTester predictions in the classification as 1-domain or multidomain protein for the different approaches (frequency vs. B-factor) and parameter sets (13 vs. 72). Additionally, Figure 9B shows the consensus for each parameter pair, which corresponds to the average over the two curves of Figure 9A, but is calculated for a total protein number of 122 instead of 135, because only proteins with known crystallographic B-factors could be used for the comparison. Again, the curve shows that the agreement between manual assignments and CovarDom is quite low for a combination of small cutoff radii with small non-covalent force constants and of large cutoff radii with large non-covalent force constants. The frequency and the B-factor approach lead to higher agreement than most fixed parameter pairs of set 72, but it is more favorable to select GNM parameter pairs only from set 13 than from set 72. Using set 72 in the frequency approach, the parameter pairs leading to low agreement are just more numerous than the ones leading to high agreement. Using set 72 in the B-factor approach often selects GNM parameters which lead to low agreement between manual methods and CovarDom. The parameter pair selected most often out of set 72 has a cutoff radius of 11 and a non-covalent force constant of $0.1\,\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$, selected 16 times out of 122 (see Table S1). It is followed by the parameter pair $r_{\mathrm{cut}} = 11$ and $k_{\mathrm{ncov}} = 0.5\,\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$, selected twelve times. Thus, for high agreement it is better to select ENM parameters only from a smaller set which is appro-

priate for most proteins. The frequent choice of unsuitable GNM parameters could amongst others result from the high number of proteins with low linear correlation coefficient between crystallographic and theoretical B-factors. Only 56% of the 122 proteins have a linear correlation coefficient of at least 0.6 for the best choice from set 72 (see Table S1). One possible reason is the high fraction of proteins in the pDomains dataset which were crystallized as larger complex. 51% of the multimeric proteins and 63% of the proteins crystallized as monomers have a linear correlation coefficient of at least 0.6. Besides, the correlation between B-factors is usually higher if the theoretical B-factors are calculated considering the crystal environment of the protein.[47,48]

Next, we study the influence of GNM parameters on the number of dynamic domains. Only proteins with available crystallographic B-factors which were assigned as multidomain both manually and by DomainTester for all parameter pairs used are considered. The sixty proteins fulfilling this condition are highlighted in Table S1 of the Supporting Information. Figure 10 shows the agreement in domain numbers between the manual predictions and CovarDom for different parameter pairs and the B-factor approach. Apart from parameter combinations of small non-covalent force constants and small cutoff radii, for which CovarDom tends to assign too many domains, the agreement between manual assignments and CovarDom is quite insensitive to the different GNM parameters. There is no advantage of determining GNM parameters for each protein separately, as done for set 13 and set 72, over simply using a non-covalent force constant of $5\,\mathrm{kcal\,(mol\ \AA^2)^{-1}}$ and a cutoff parameter of 7 Å, like employed in the first results section. In contrast to the differentiation between 1-domain and multidomain proteins, where large cutoff radii in combination with large non-covalent force constants lead to the classification of too many proteins as 1-domain protein, an according parameter selection does not affect the number of domains assigned to multidomain proteins. However we should stick to parameter pairs with cutoff radii of 7 or 8 Å and a ratio of covalent to non-covalent force constants smaller than 100 to ensure high agreement for the whole domain assignment process. A cutoff radius of 7 Å corresponds to the typical value chosen in GNM to include the interactions in the first shell of neighbors.[49] Several studies proposed the usage of stronger force constants for covalent than for non-covalent interactions[41,46] or distance-dependent force constants.[16] As our analysis shows, the nonbonded interac-

tions should also not be underestimated, because non-covalent force constants which are a hundred times weaker than the covalent force constant lead to the assignment of too many dynamic domains.

CovarZeroDom: Negative Covariance as Stopping Criterion

CovarZeroDom uses an intercluster covariance cutoff value as stopping criterion in the clustering procedure, in contrast to CovarDom, which recalculates the covariance matrices and uses DomainTester after each splitting in two clusters to decide whether they can be further divided. CovarZeroDom is applicable if covariance matrices of the splitted protein cannot be determined, for example if molecular dynamics is used instead of an ENM. The intercluster covariance is usually positive until a small number of cluster is reached. Thus, we can choose a covariance of zero as stopping criterium, which is physically meaningful, because all residues with nonnegative covariance are clustered into one dynamic domain. But the program as well allows to choose another value than zero as final intercluster covariance. In comparison to CovarDom, CovarZeroDom assigns less often two and more often three dynamic domains to proteins (see Figure S1 in Supporting Information). The additional domains assigned by CovarZeroDom are often small fragments lying between larger domains, as for example in endonuclease I-Dmol (Figure 11A) and TAFII250 (Figure11B). If these small domains connect larger domains, they can possibly act as hinges in conformational changes of the protein. The small domains have a negative intercluster covariance to all other domains and are therefore not merged by CovarZeroDom. In CovarDom, the precursor cluster containing the small domain and a larger domain are classified as one domain by DomainTester, because mostly not even one positive-covariance segment is found in the corresponding covariance matrix. Thus, the small fragments are not separated from the large domain using CovarDom.

But the use of CovarZeroDom can also lead to the assignment of less domains than by CovarDom, if two domains are strongly anticorrelated, but consist of further dynamic domains themselves. Figure 11C shows the covariance matrices of elongation factor Tu,[50] calculated using a non-covalent force constant of $5\,\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$ and a cutoff radius of 7. In the last step, DomainClusterer merges domains 2_1 and 2_2 with a positive intercluster covariance of

0.0028. Thus, if a negative intercluster covariance is chosen as stopping criterion, the final domain number of elongation factor Tu is two, although from the covariance matrix, it is obvious that three dynamic domains are present. CovarDom allows for the assignment of three domains, because the removal of the highly anticorrelated domain 1 from the EN shifts the intercluster covariance between domains 2_1 and 2_2 to negative values, as the sum over all covariances is always zero.[24]

Another possibility besides CovarZeroDom to avoid the recalculation of covariance matrices after each splitting is a renormalization of the parts of the original matrix which belong to one cluster, such that the sum over all cluster elements is again zero. To re-normalize, the average covariance is subtracted from all matrix elements. The corresponding covariance matrix of cluster 2 of elongation factor Tu is shown in Figure 11C. The difference of this approach to CovarDom is that the influence of the residues of the other dynamic domains are still present in the covariances, whereas in CovarDom, the dynamic domains are considered as being independent from each other. Thus, the difference between the two differently calculated covariance matrices gives the deviation from independent behavior. With values between -0.06 to +0.02, the differences are small. Nevertheless, using re-normalized covariances results in the assignment of two instead of three dynamic domains to elongation factor Tu, because DomainTester detects only one positive-covariance segment in cluster 2 comprising the residues 7-97 of domain 2_1. Domain 2_2 has more interactions with Domain 1, which leads to less independent movement of its residues. Employing the renormalization strategy to all multidomain proteins of the pDomains benchmark shows that it leads to less overall agreement with manual domain assignments.

Substrate Channels between Dynamic Domains: Acetylene Hydratase

Acetylene hydratase catalyzes the hydration of acetylene to acetaldehyde.[23,51] Catalysis occurs by a water molecule bound to a bis-molybdopterin guanine dinucleotide-ligated tungsten atom. The water molecule is activated by an aspartate residue, Asp13, whose deprotonation is shifted to unusually high pH values by interaction with a nearby [4Fe-4S] cluster. CovarDom assigns two

dynamic domains to acetylene hydratase (see Figure 12A). For calculating the covariance matrix, the tungsten atom and the [4Fe-4S] cluster are each represented by one node in the elastic network, while each molybdopterin guanine dinucleotide molecule is represented by five nodes. No substrate nodes are included in the calculation. Choosing the best GNM parameters based on comparison of B-factors out of set 13, we use a force constant of $0.5\,\mathrm{kcal}\,(\mathrm{mol}^2)^{-1}$ and a cutoff radius of 10.

Asp13, the [4Fe-4S] cluster, the tungsten atom and one molybdopterin molecule belong to domain 1, the second molybdopterin molecule belongs to domain 2. The substrate channel lies between the two dynamic domains (Figure 12B). This location between two large dynamic domains could lead to an easier entry of the substrate acetylene. Interestingly, in all other known enzymes of the DMSO reductase family of molybdenum and tungsten enzymes,[52,53] a different position is found for the channel to the active site, which is sealed in acetylene hydratase by the residues 328 to 393 (Figure 12C). Residues 331 to 367 and 385 to 393 belong to dynamic domain 1, while the residues 328 to 330 and 368 to 384 are located in domain 2. Thus, the lid over the original substrate channel is not a flexible structure which could easily move away, but is connected to the dynamic domains.

Dynamic Domains of a Multimer: 4-Hydroxyphenylacetate Decarboxylase

All proteins of the pDomains dataset are single subunits, although several of them are part of a larger protein complex in their active form. The splitting reflects the assumption that domains do not spread over several subunits. But from a dynamical point of view, residues of different subunits can belong to the same dynamic domain, which may be of functional importance in protein-protein interactions. In the following, the domain assignment method is demonstrated on the multimer 4-hydroxyphenylacetate decarboxylase (HPD). HPD is a glycyl radical enzyme which catalyzes the chemically difficult decarboxylation of 4-hydroxyphenylacetate to *p*-cresol.[22,54,55] The $(\beta\gamma)_4$ tetramer consists of heterodimers built of a catalytic β-subunit harboring a glycyl/thiyl dyad (Gly873, Cys503) and a small γ-subunit with two [4Fe-4S] clusters. We investigate the dynamic domains of the tetramer comprising eight subunits.

The [4Fe-4S] clusters are represented by one node each, lying in the center of the cluster. The substrate 4-hydroxyphenylacetate is not included in the calculation. Choosing the best GNM parameters out of set 13, with a correlation between crystallographic and theoretical B-factors of 0.74, we use a force constant of 7 kcal $(\mathrm{mol}^2)^{-1}$ and a cutoff radius of 7.

The protein is split into 28 dynamic domains. In the first CovarDom splitting step, two $(\beta\gamma)_4$ dimers are separated from each other (see Figure 13A). Because the resulting clusters, labeled 1_* and 2_*, are identical, we can only analyze the further splitting of cluster 2_*. The next splitting step separates the two heterodimers from each other, but the separation is not complete, because residues 252 to 262 of each heterodimer are grouped into the cluster of the other heterodimer. The further splitting of the two heterodimers is nearly identical. Figure 13B shows cluster 2_1_*. For better readibility, we omit these first two digits in the domain denomination of the dynamic domains originating from cluster 2_1_* in the following discussion. The β- and the γ-subunit are built from different dynamic domains, like expected. The β-subunit of HPD consists of five dynamic domains, the γ-subunit of two dynamic domains. But the splitting between β- and the γ-subunit does not occur in the next step of the domain assignment procedure. First, cluster 2_1_* is split in the middle of the β-subunit, while the γ-subunit still lies in cluster 2_1_2_* together with a part of the β-subunit (see Figure 13A), meaning that the interactions between this part of the β-subunit and the γ-subunit are stronger than the interactions within the β-subunit. Another evidence for the strong interaction between β- and γ-subunit is that the separation between β- and γ-subunit is not complete. Residue 14 of the the γ-subunit belongs to domain 2_1_2, while residue 87 of the β-subunit belongs to domain 2_1_1_2. The γ-subunit is not present in all glycyl radical enzymes and is proposed to be involved in regulation of the oligomeric state and catalytic activity of HPD.[56]

In the large β-subunit, the two interacting residues of the radical dyad belong to domain 1_1. The channel to the active site is flexible due to its location between two dynamic domains, 1_2_1 and 2_1_2 (Figure 13B). Residues interacting with the substrate 4-hydroxyphenylacetate belong to the domains 1_1 (Arg223, Ser344, Gly345, Phe405, Glu505 and Ile750), 1_2_1 (Phe214, Ile219, His536, Phe537 and Glu637), 1_2_2 (Val752) and 2_1_2 (Val399 and Leu400). Their distribution onto several domains with uncorrelated motion allows to

arrange the active site residues in a manner to prebuild the transition state of the reaction. The radical on Gly873 is obtained by interaction of HPD with an activating enzyme (AE). Reductive cleavage of S-adenosylmethionine in the AE generates a transient 5′-deoxyadenosyl radical which then generates the Gly873 radical. The radical domain is flanked by two peptide sequences that are weakly structured in x-ray crystallography. The flexibility of these sequences is proposed to play a role in the postulated opening of the radical domain upon complex formation with the AE.[22] The stretch Gln121 to Lys167 belongs mostly to domain 2_1_2 (Figure 13C). Only residues 144 to 149 which build an extended loop belong to domain 1_1. The residues Asn672 to Glu700 fully belong to domain 1_2_2. Thus, the flexible sequences are mostly dynamically decoupled from the radical domain and can move away, which could lead to conformational changes in the radical domain.

Conclusions

Dynamic domains have direct functional relevance, because the functionality of a protein is tightly connected to its dynamics. For example, ligand binding sites often lie between dynamic domains. The uncorrelated motion of different domains can allow for an easier entry of the substrates and a perfect arrangement of the active site residues. Besides, sites lying between anticorrelated domains are often perturbation-sensitive, such that ligand binding has a large effect on the dynamics of the protein, which can for example lead to allostery. In contrast to structural domains, dynamic domains are often sequentially discontinuous, highlighting the residues which mediate interdomain relations. Additional information about the strength of interactions between dynamic domains is given by their hierarchical organisation created by CovarDom. For multimeric proteins, all subunits can be included in the calculation. Even though the different subunits are most likely assigned as different dynamic domains, small segments assigned against expectation and the order of the splitting events can highlight important structural properties of the complex. The assignment of dynamic protein domains by CovarDom is not influenced by human conception, but purely based on previously calculated dynamical data. Still, the automatic domain assignment is influenced by the choice of both GNM and CovarDom parameters, and should only be used

as guideline which is followed by manual inspection. Using a cutoff radius of 7, a standard value for GNM calculations, and a non-covalent force constant which is smaller than the covalent force constant, but not by several orders of magnitude, seems to work well for most proteins. Instead, the elastic network parameters can be chosen by comparison of theoretical to crystallographic B-factors from a reasonable set of cutoff radii and ratios between covalent and non-covalent force constants.

Bibliography

[1] R F Doolittle. The multiplicity of domains in proteins. *Annu Rev Biochem*, 64:287–314, 1995.

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

[3] S A Islam, J Luo, and M J Sternberg. Identification and analysis of domains in proteins. *Protein Eng*, 8:513–525, 1995.

[4] L Wernisch, M Hunting, and S J Wodak. Identification of structural domains in proteins by a graph heuristic. *Proteins*, 35:338–352, 1999.

[5] N Alexandrov and I Shindyalov. PDP: protein domain parser. *Bioinformatics*, 19:429–430, 2003.

[6] J T Guo, D Xu, D Kim, and Y Xu. Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res*, 31:944–952, 2003.

[7] H Zhou, B Xue, and Y Zhou. DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. *Protein Sci*, 16:947–955, 2007.

[8] Z Y Xuan, L J Ling, and R S Chen. A new method for protein domain recognition. *Eur Biophys J*, 29:7–16, 2000.

[9] M B Swindells. A procedure for detecting structural domains in proteins. *Protein Sci*, 4:103–112, 1995.

[10] R Sowdhamini and T L Blundell. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci*, 4:506–520, 1995.

[11] I N Berezovsky. Discrete structure of van der Waals domains in globular proteins. *Protein Eng*, 16:161–167, 2003.

[12] S Hayward, A Kitao, and H J Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins*, 27:425–437, 1997.

[13] S A Wieninger, E H Serpersu, and G M Ullmann. ATP binding enables broad antibiotic selectivity of aminoglycoside phosphotransferase(3')-IIIa: an elastic network analysis. *J Mol Biol*, 409:450–465, 2011.

[14] S Bernhard and F Noé. Optimal identification of semi-rigid domains in macromolecules from molecular dynamics simulation. *PLoS One*, 5, 2010.

[15] S Kundu, D C Sorensen, and G N Phillips. Automatic domain decomposition of proteins by a gaussian network model. *Proteins*, 57:725–733, 2004.

[16] K Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

[17] K Hinsen, A Thomas, and M J Field. Analysis of domain motions in large proteins. *Proteins*, 34:369–382, 1999.

[18] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2:173–181, 1997.

[19] A R Atilgan, S R Durell, R L Jernigan, M C Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[20] S O Yesylevskyy, V N Kharkyanen, and A P Demchenko. Hierarchical clustering of the correlation patterns: new method of domain identification in proteins. *Biophys. Chem.*, 119:84–93, 2006.

[21] S O Yesylevskyy, V N Kharkyanen, and A P Demchenko. Dynamic protein domains: identification, interdependence, and stability. *Biophys. J.*, 91:670–685, 2006.

[22] B M Martins, M Blaser, M Feliks, G M Ullmann, W Buckel, and T Selmer. Structural basis for a Kolbe-type decarboxylation catalyzed by a glycyl radical enzyme. *J Am Chem Soc*, 133:14666–14674, 2011.

[23] G B Seiffert, G M Ullmann, A Messerschmidt, B Schink, P M Kroneck, and O Einsle. Structure of the non-redox-active tungsten/[4Fe:4S] enzyme acetylene hydratase. *Proc Natl Acad Sci U S A*, 104:3073–3077, 2007.

[24] S O Yesylevskyy, V N Kharkyanen, and A P Demchenko. The change of protein intradomain mobility on ligand binding: is it a commonly observed phenomenon? *Biophys. J.*, 91:3002–3013, 2006.

[25] S Kundu, J S Melton, D C Sorensen, and G N Phillips. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.*, 83:723–732, 2002.

[26] P Wessel and W H F Smith. New, improved version of Generic Mapping Tools released. *EOS Trans. Amer. Geophys. U.*, 79:579, 1998.

[27] W L DeLano. The PyMOL Molecular Graphics System. *DeLano Scientific LLC, Palo Alto, CA, USA. http://www.pymol.org*, 2008.

[28] T A Holland, S Veretnik, I N Shindyalov, and P E Bourne. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol*, 361:562–590, 2006.

[29] A G Murzin, S E Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, 1995.

[30] C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, and J M Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[31] S Veretnik, P E Bourne, N N Alexandrov, and I N Shindyalov. Toward consistent assignment of structural domains in proteins. *J Mol Biol*, 339:647–678, 2004.

[32] T Knöfel and N Sträter. X-ray structure of the escherichia coli periplasmic 5'-nucleotidase containing a dimetal catalytic site. *Nat Struct Biol*, 6(5):448–453, May 1999.

[33] P Grochulski, L Masson, S Borisova, M Pusztai-Carey, J L Schwartz, R Brousseau, and M Cygler. Bacillus thuringiensis cryia(a) insecticidal toxin: crystal structure and channel formation. *J Mol Biol*, 254:447–464, Dec 1995.

[34] M C Wilce, C S Bond, N E Dixon, H C Freeman, J M Guss, P E Lilley, and J A Wilce. Structure and mechanism of a proline-specific aminopeptidase from escherichia coli. *Proc Natl Acad Sci U S A*, 95(7):3472–3477, Mar 1998.

[35] S Crennell, E Garman, G Laver, E Vimr, and G Taylor. Crystal structure of vibrio cholerae neuraminidase reveals dual lectin-like domains in addition to the catalytic domain. *Structure*, 2:535–544, Jun 1994.

[36] H Zhang, M C Seabra, and J Deisenhofer. Crystal structure of rab geranylgeranyltransferase at 2.0 A resolution. *Structure*, 8:241–251, 2000.

[37] P Hänzelmann, H Dobbek, L Gremer, R Huber, and O Meyer. The effect of intracellular molybdenum in Hydrogenophaga pseudoflava on the crystallographic structure of the seleno-molybdo-iron-sulfur flavoenzyme carbon monoxide dehydrogenase. *J Mol Biol*, 301:1221–1235, 2000.

[38] W R Tulip, J N Varghese, A T Baker, A van Donkelaar, W G Laver, R G Webster, and P M Colman. Refined atomic structures of N9 subtype influenza virus neuraminidase and escape mutants. *J Mol Biol*, 221:487–497, 1991.

[39] K Brown, M Tegoni, M Prudêncio, A S Pereira, S Besson, J J Moura, I Moura, and C Cambillau. A novel type of catalytic copper cluster in nitrous oxide reductase. *Nat Struct Biol*, 7:191–195, 2000.

[40] U Ermler, R A Siddiqui, R Cramm, and B Friedrich. Crystal structure of the flavohemoglobin from Alcaligenes eutrophus at 1.75 A resolution. *EMBO J*, 14(24):6067–6077, Dec 1995.

[41] D A Kondrashov, Q Cui, and G N Phillips. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys. J.*, 91:2760–2767, 2006.

[42] G Sainz, C J Carrell, M V Ponamarev, G M Soriano, W A Cramer, and J L Smith. Interruption of the internal water chain of cytochrome f impairs photosynthetic function. *Biochemistry*, 39:9164–9173, 2000.

[43] B Sha, S E Phillips, V A Bankaitis, and M Luo. Crystal structure of the Saccharomyces cerevisiae phosphatidylinositol-transfer protein. *Nature*, 391:506–510, 1998.

[44] E Eyal, L W Yang, and I Bahar. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, 22:2619–2627, 2006.

[45] W Zheng, B R Brooks, and D Thirumalai. Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations. *Biophys J*, 93:2289–2299, 2007.

[46] Q Yang and K A Sharp. Building alternate protein structures using the elastic network model. *Proteins*, 74:682–700, 2009.

[47] G Song and R L Jernigan. vGNM: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.*, 369:880–893, 2007.

[48] K Hinsen. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24:521–528, 2008.

[49] I Bahar and R L Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266:195–214, 1997.

[50] G Polekhina, S Thirup, M Kjeldgaard, P Nissen, C Lippmann, and J Nyborg. Helix unwinding in the effector region of elongation factor EF-Tu-GDP. *Structure*, 4:1141–1151, 1996.

[51] B Schink. Fermentation of acetylene by an obligate anaerobe Pelobacter acetylenicus sp. nov. *Arch Microbiol*, 142:295–301, 1985.

[52] C Kisker, H Schindelin, and D C Rees. Molybdenum-cofactor-containing enzymes: structure and mechanism. *Annu Rev Biochem*, 66:233–267, 1997.

[53] H Dobbek and R Huber. The molybdenum and tungsten cofactors: a crystallographic view. *Met Ions Biol Syst*, 39:227–263, 2002.

[54] L D'Ari and H A Barker. p-Cresol formation by cell-free extracts of Clostridium difficile. *Arch Microbiol*, 143:311–312, 1985.

[55] P I Andrei, A J Pierik, S Zauner, L C Andrei-Selmer, and T Selmer. Subunit composition of the glycyl radical enzyme p-hydroxyphenylacetate decarboxylase. a small subunit, HpdC, is essential for catalytic activity. *Eur J Biochem*, 271:2225–2230, 2004.

[56] L Yu, M Blaser, P I Andrei, A J Pierik, and T Selmer. 4-hydroxyphenylacetate decarboxylases: properties of a novel subclass of glycyl radical enzyme systems. *Biochemistry*, 45:9584–9592, 2006.

[57] T S Peat, J Newman, G S Waldo, J Berendzen, and T C Terwilliger. Structure of translation initiation factor 5A from Pyrobaculum aerophilum at 1.75 A resolution. *Structure*, 6:1207–1214, 1998.

[58] T Keitel, O Simon, R Borriss, and U Heinemann. Molecular and active-site structure of a Bacillus 1,3-1,4-beta-glucanase. *Proc Natl Acad Sci U S A*, 90:5287–5291, 1993.

[59] G H Silva, J Z Dalgaard, M Belfort, and P Van Roey. Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. *J Mol Biol*, 286:1123–1136, 1999.

[60] R H Jacobson, A G Ladurner, D S King, and R Tjian. Structure and function of a human TAFII250 double bromodomain module. *Science*, 288:1422–1425, 2000.

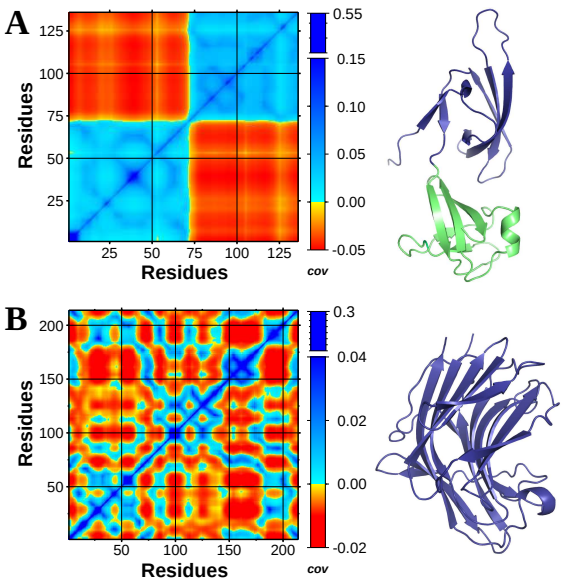

Figure 1. Covariance matrix and structure of a two-domain protein (A) and a 1-domain protein (B). A) The covariance matrix of translation initiation factor 5A (PDB code 1bkb[57]) has two separate positive-covariance areas. B) The covariance matrix of beta-glucanase (PDB code 1byh[58]) shows no large area of only positive covariances.

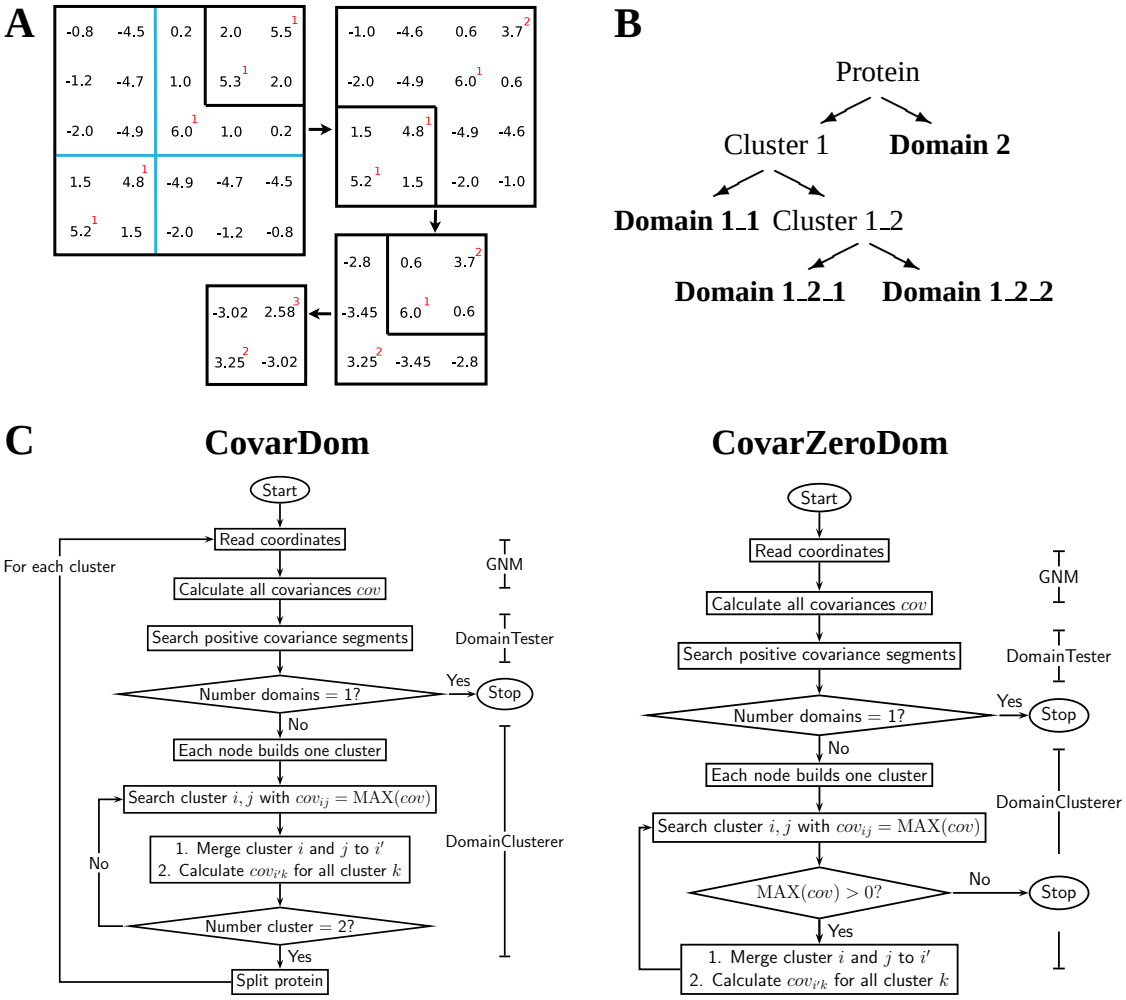

Figure 2. Calculation of dynamic protein domains. A) Example calculation of inter- and intracluster covariances after merging of two cluster. The red digits indicate the number of nodes within the cluster. The two cluster in the black box are merged next. The blue boxes show the final classification of the nodes into two cluster. Averaging over all covariances within one blue box gives the final intracluster covariance value, while averaging over all covariances between the two blue boxes results in the final intercluster covariance. B) Example for domain numbering in CovarDom. The different hierarchical levels are obvious from the names. Domains which only differ in the last digit of the name can be combined into the precursor cluster. C) Differences in the overall workflow of CovarDom and CovarZeroDom. CovarZeroDom uses a negative intercluster covariance as stopping criterion in the clustering procedure, while CovarDom always clusters all nodes into two cluster and then employs DomainTester to test if the cluster can be further divided.

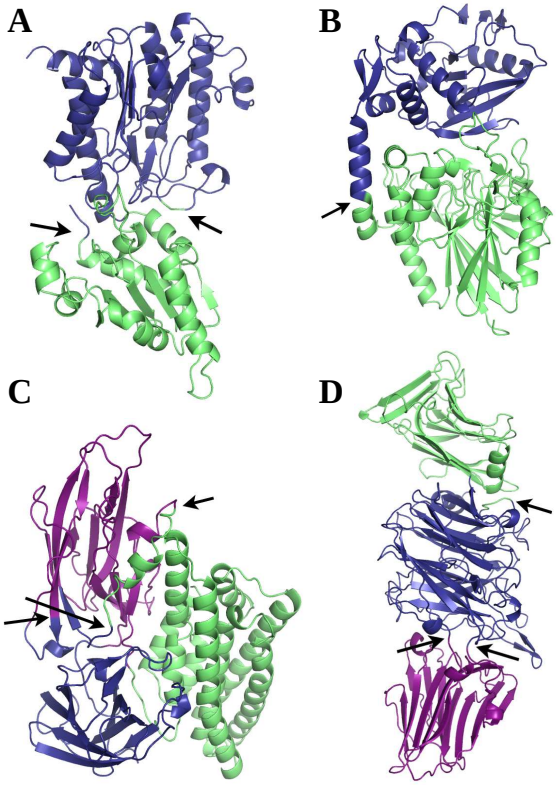

Figure 3. Dynamic domains which are very similar to the manually assigned domains. Arrows indicate splitting of secondary structure elements or clustering of a few residues to another dynamic domain by CovarDom. If three dynamic domains exist, first the green domain is split from the other two domains. A) Aminopeptidase P (PDB code 1jaw[34]) consists of two domains. The dynamic domains are continuous, except for a few loop residues which are clustered to another dynamic domain than their sequential neighbors. B) 5′-nucleotidase (PDB code 1ush[32]) consists of two continuous domains. An α-helix is split between the two dynamic domains. Manually the residues of this helix were not assigned to any domain. C) CryIA(a) toxin from *Bacillus thuringiensis* (PDB code 1ciy[33]) consists of three domains. Two β-strands are split between the dynamic domains shown in blue and magenta. A small α-helix belongs to the dynamic domain shown in green, although according to the peptide sequence it would belong to the magenta domain. D) Neuraminidase from *Vibrio cholerae* (PDB code 1kit[35]) consists of three domains. The blue domain is discontinuous, be cause the residues of the magenta domain are inserted into its sequence. Additionally, a few loop residues neighboring the blue domain are clustered to the green domain.

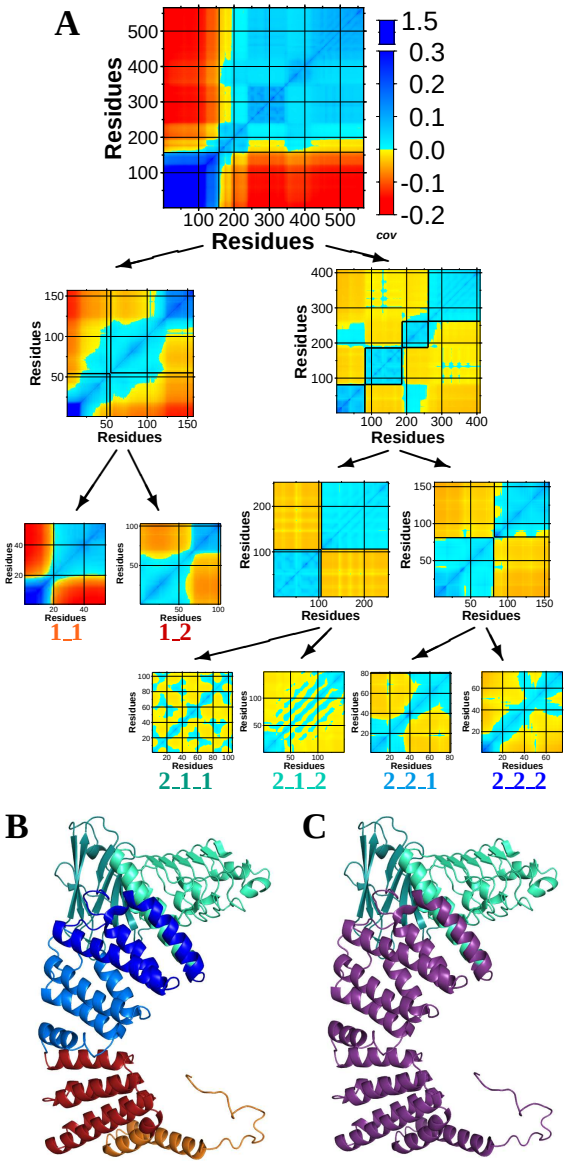

Figure 4. Clustering of Rab geranylgeranyltransferase (PDB code 1dce[36]), using a cutoff radius of 7 Å and a non-covalent force constant of 5 kcal (mol Å^2)$^{-1}$. A) Hierarchical clustering of the covariance matrix. First, the precursor of the red and orange domain are split from the rest. Then, the precursor of the blue and iceblue domain are split from the precursor of the lightblue and cyan domain. When only five dynamic domains are assigned by CovarDom, the splitting of the precursor domain of domains 2_2_1 and 2_2_2, does not occur. B) Six dynamic domains assigned by CovarDom. C) Three domains assigned manually.

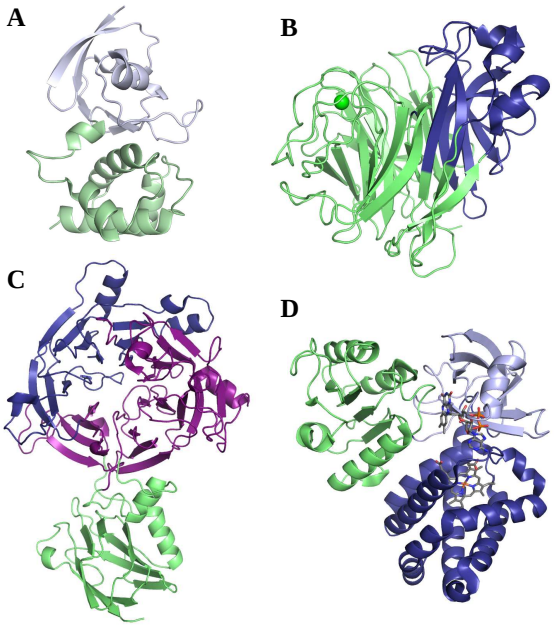

Figure 5. Proteins with dynamic domains differing from the manual domain assignments. A) Iron-sulfur protein of carbon monoxide dehydrogenase (PDB code 1ffu[37]) is assigned as 1-domain protein by CovarDom and as 2-domain protein manually. The two domains assigned by the authors of the structures are colored in pale blue and pale green. B) One subunit of neuraminidase N9 of influenza virus (PDB code 5nn9[38]) is assigned as 1-domain protein manually, but as 2-domain protein by CovarDom. The two domains assigned by CovarDom are colored in blue and green. A bound calcium ion is shown as green sphere. C) Nitrous oxide reductase (PDB code 1qni[39]) is divided into two domains manually, but into three domains shown in green, blue and purple by CovarDom. The blue and the magenta domain build one structural domain. D) Flavohemoglobin from *Alcaligenes eutrophus* (PDB code 1cqx[40]) is assigned as 2-domain protein by CovarDom if FAD and heme are bound to it (shown in sticks representation). Manually, three domains are assigned. The blue and the pale blue domain build one dynamic domain. Neglecting the ligands, CovarDom also assigns three dynamic domains to the flavohemoglobin.

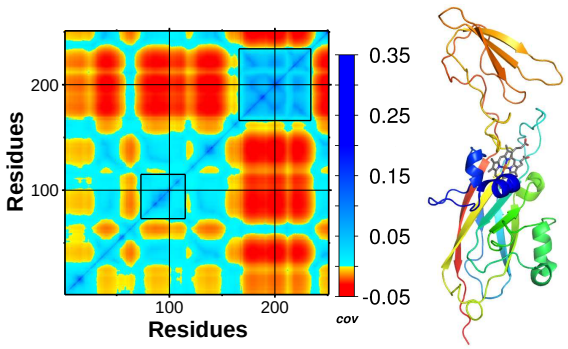

Figure 6. Structure and covariance matrix of cytochrome f (PDB code 1e2v[42]). DomainTester detects two positive-covariance segments, which comprise 43% of the residues. These residues are encircled in the covariance plot. If the required fraction of nodes in positive-covariance segments is lowered to 0.4, two dynamic domains are assigned by DomainClusterer. The structure of cytochrome f in cartoon representation is colored by sequence to show that the smaller domain is inserted between two β-strands belonging to the larger domain.

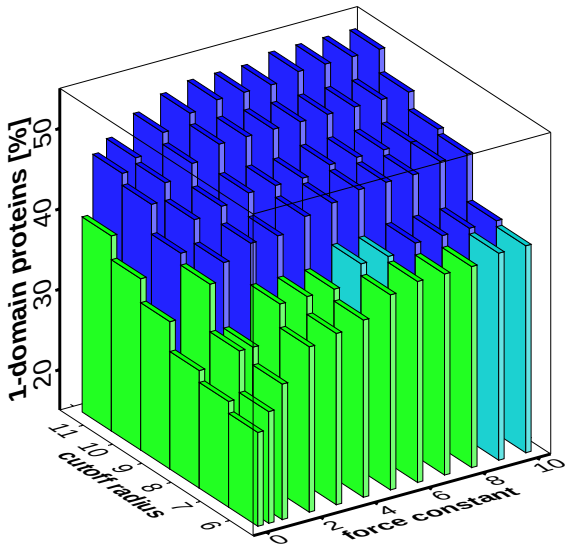

Figure 7. Influence of the cutoff radius and the non-covalent force constant of the GNM on the classification as 1-domain or multidomain protein by DomainTester. The height of the bars indicates the percentage of proteins assigned as 1-domain protein. All 135 proteins of the pDomains dataset were used for the calculation. Cyan bars represent GNM parameter pairs which result in a percentage of 1-domain proteins of 40.7, which corresponds to the percentage of 1-domain proteins assigned manually. Green bars indicate a lower percentage and blue bars indicate a higher percentage of 1-domain proteins. A combination of small cutoff radii with small non-covalent force constants leads to the assignment of many multidomain proteins, while calculations using large cutoff radii and large non-covalent force constants result in the assignment of many 1-domain proteins.

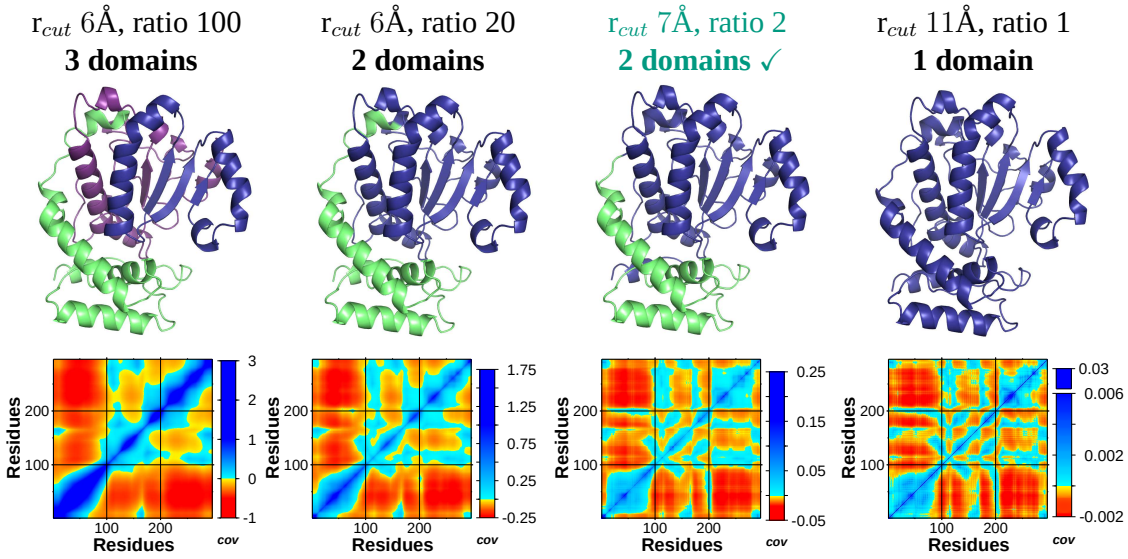

Figure 8. Covariance matrices of phophatidylinositol transfer protein (PDB code 1aua[43]) calculated using different GNM parameter pairs. Ratio gives the ratio of the covalent to the non-covalent force constant, with the covalent force constant being fixed to 10 kcal (mol Å^2)$^{-1}$. Small cutoff radii in combination with low non-covalent force constants underestimate the nonbonded interactions and lead to a broad zone of positive covariance along the diagonal, which results in the assignment of more domains by CovarDom. The highlighted clustering using a cutoff radius of 7Å and a non-covalent force constant of 5 kcal (mol Å^2)$^{-1}$ agrees with the manual assignment in domain number and boundaries.

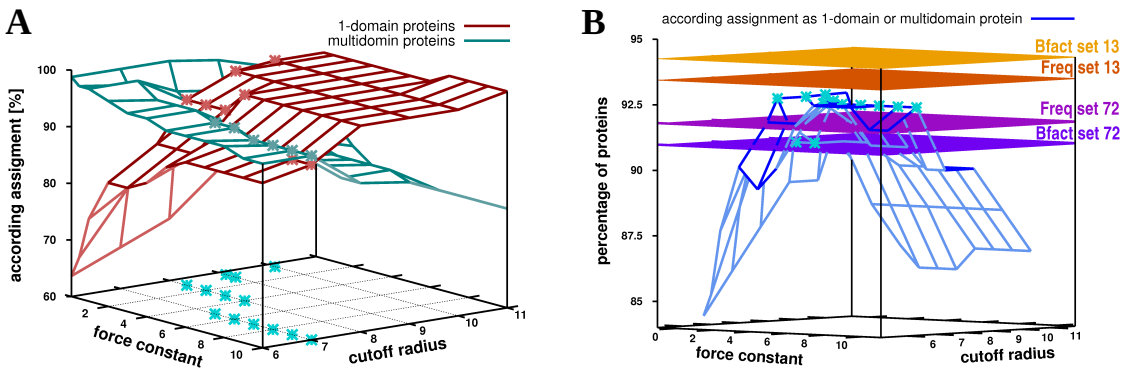

Figure 9. Influence of the cutoff radius and the non-covalent force constant on the classification as 1-domain or multidomain protein. A) DomainTester was applied to the eighty proteins of the pDomains dataset which are assigned as multidomain proteins and to the fifty-five proteins which are assigned as 1-domain protein. The curves give the percentage of proteins which are classified accordingly by DomainTester. Thirteen GNM parameter pairs which lead to an accordance of at least 90% between DomainTester and manual predictions for both protein sets are highlighted by stars. B) DomainTester was applied to all proteins of the pDomains dataset with available experimental B-factors. The curve gives the percentage of proteins which are assigned accordingly by DomainTester and by manual methods as 1-domain or multidomain protein. The planes situated at 91.0%, 91.8%, 93.4% and 94.3% give the accordance if the GNM parameter pairs are chosen separately for each protein by comparison to experimental B-factors or by the frequency approach from parameter set 72 or 13. The parameter pairs of set 13 are highlighted by cyan stars.

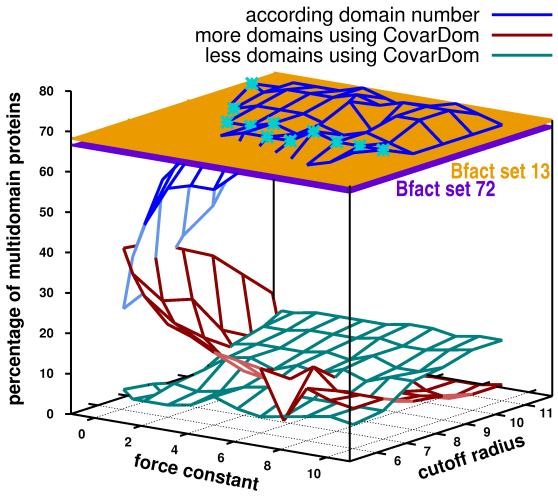

Figure 10. Influence of the cutoff radius and the non-covalent force constant on the domain number of multidomain proteins. The calculations are performed on sixty proteins which are assigned as multidomain in the pDomains dataset and by CovarDom for all GNM parameter pairs, and for which experimental B-factors are available. These proteins are specified in Table 1 of the Supporting Information. Cyan stars indicate the position of GNM parameter pairs of set 13. Percentage of multidomain proteins for which CovarDom assigns the same, a smaller or a larger number of domains than given by manual predictions. The planes situated at 66.7% and 68.3% give the percentage of proteins with according domain number if the GNM parameter pairs are chosen separately for each protein by comparison to experimental B-factors out of set 72 and set 13, respectively.

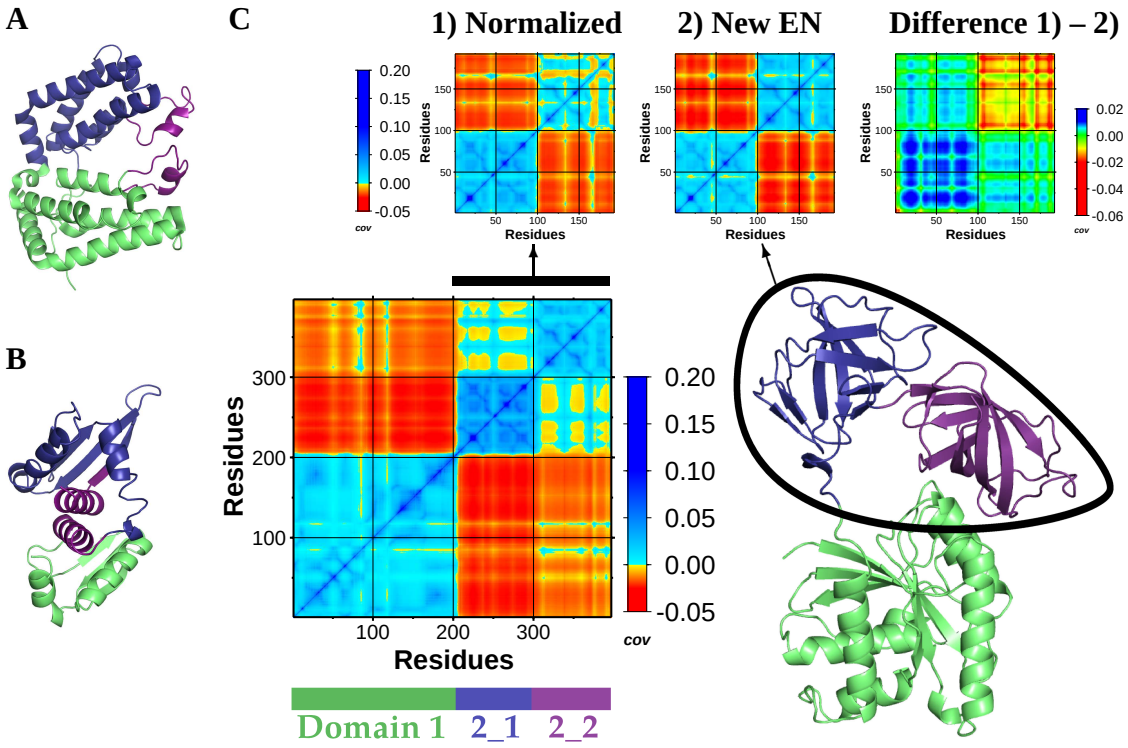

Figure 11. Comparison between domain assignments using CovarDom and CovarZeroDom. A,B) The two proteins endonuclease I-Dmol (A, PDB code 1b24[59]) and TAFII250 (B, PDB code 1eqf[60]) are assigned as 2-domain proteins by CovarDom and as 3-domain proteins by CovarZeroDom. The colors indicate the dynamic domains assigned by CovarZeroDom. Blue and magenta domains build one CovarDom domain. C) Elongation factor Tu (PDB code 1tui[50]) is assigned as 3-domain protein by CovarDom, but as 2-domain protein by CovarZeroDom. Using CovarZeroDom, the domains 2_1 and 2_2 are merged into one dynamic domain, because their intercluster covariance is positive. The covariance matrix of the residues of cluster 2 of elongation factor Tu is once calculated for the whole protein and once only for the residues of cluster 2. The difference between the two covariance matrices indicates the deviation from independent behavior of domain 1 and cluster 2, that is it shows the influence of domain 1 on cluster 2. The difference matrix is normalized such that the sum over all matrix elements equals zero.

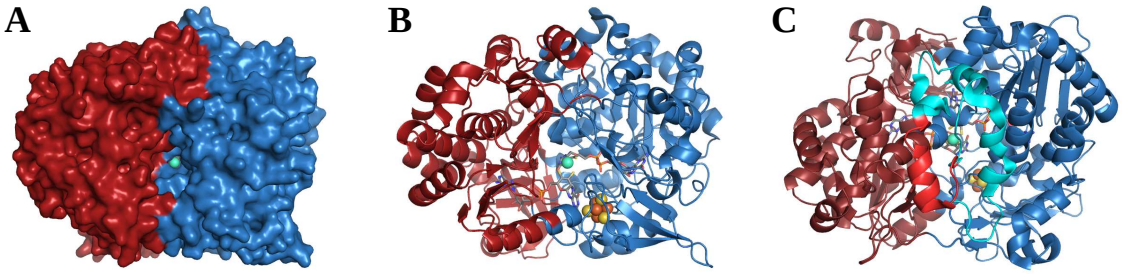

Figure 12. Dynamic domains 1 (blue) and 2 (red) of acetylene hydratase. The [4Fe-4S] cluster and the tungsten atom are shown in VdW representation. Tungsten is shown in green. The two molybdopterin guanine dinucleotide molecules are represented as bonds. A,B) View on the substrate channel. C) View on the alternative substrate channel found in other molybdenum and tungsten enzymes. Residues 328 to 393 seal this substrate funnel. Residues 328 to 367 and 385 to 393, shown in cyan, belong to domain 1, residues 368 to 384, shown in light red, belong to domain 2. The orientation is rotated in comparison to A) and B).

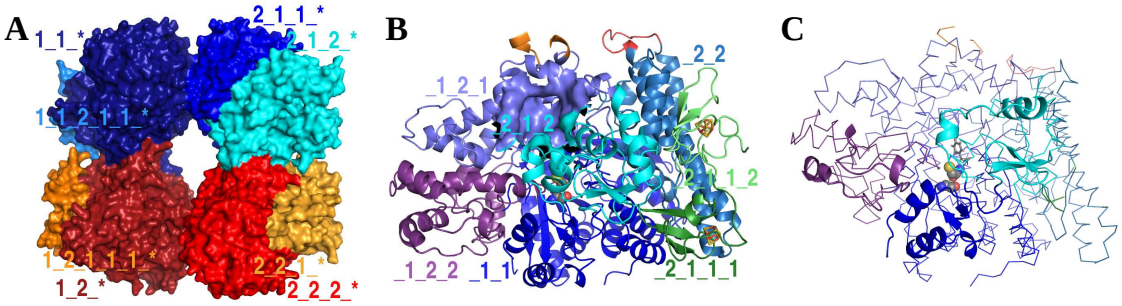

Figure 13. Dynamic domains of 4-hydroxyphenylacetate decarboxylase. A) The left half shows subunit β in dark blue and red and the γ subunits in light blue and red. The right half depicts the first splitting step of the heterodimers by CovarDom. B) The five dynamic domains of the β-subunit are pictured in shades of blue and the two dynamic domains of the γ-subunit are pictured in light and dark green. The glycyl/thiyl radical dyad is shown in VdW representation. The channel to the active site is shown in surface representation. The two [4Fe-4S] cluster of the γ-subunit are shown as sticks. Residues shown in red are part of this heterodimer, but belong to the dynamic domain of the neighboring heterodimer. C) Radical segment (dark blue) and flexible segments (cyan, purple) of subunit β are shown in cartoon representation, colored according to their domain affiliation. The rest of the subunit is shown in ribbon representation. The glycyl/thiyl radical dyad is shown in VdW representation, and the substrate 4-hydroxyphenylacetate as sticks.

CovarDom: Identifying Dynamic Protein Domains based on Covariance Matrices of Motion

Silke A. Wieninger and G. Matthias Ullmann

to be submitted

Supporting Information

Table 1. Number of domains assigned to the proteins of the pDomains dataset. Manual predictions are compared to CovarDom predictions using different GNM parameter. For parameter sets 13 and 72, the parameter pair used in the calculation is chosen based on comparison between experimental and theoretical B-factors. k gives the force constant for non-covalent interactions, r the cutoff radius and *corr* the linear correlation coefficient between the B-factors. Only proteins with bold name are used in the calculation of the domain overlap. They are assigned as multidomain both manually and by CovarDom for all GNM parameter pairs, and experimental B-factors are available for these proteins.

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1a8y		3	0.5	10	0.52	3	0.5	10	0.52	3	3
1aba		1	10	7	0.61	1	10	6	0.64	1	1
1alc		1	1	9	0.51	1	1	9	0.51	1	1
1aog	A	3	0.5	10	0.71	3	0.5	11	0.71	3	3
1aps		1									1
1au7	A	2	10	7	-0.05	2	10	10	0.11	2	2
1aua		2	5	7	0.65	2	0.1	7	0.72	2	2
1b24	A	2	10	7	0.69	2	0.1	11	0.70	2	2
1bbw	A	2	0.5	10	0.50	2	1	11	0.53	2	2
1bc5	A	2	5	7	0.60	3	0.1	7	0.66	4	3
1bds		1	0.5	10	0.81	1	1	11	0.85	1	1
1bhg	A	3	0.5	10	0.34	3	0.5	10	0.34	3	3
1bi3	A	3	10	7	0.71	3	10	7	0.71	3	3
1bkb		2	5	7	0.72	2	0.1	11	0.73	2	2
1bpm		2	0.5	9	0.65	2	0.5	9	0.65	2	2
1btc		1	0.5	10	0.60	1	0.5	11	0.66	1	2
1byh		1	4	8	0.54	1	4	8	0.54	1	1

Table 1 – continued

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1c0m	A	2	0.5	10	0.51	2	0.5	11	0.52	2	2
1c5a		1	10	7	0.36	1	10	9	0.41	1	1
1cad		1	4	8	0.72	1	10	10	0.73	1	1
1caj		1	0.5	10	0.83	1	0.5	10	0.83	1	1
1cgi	I	1	0.5	9	0.70	1	0.1	6	0.78	1	1
1chc		1									1
1chu	A	3	9	7	0.85	2	9	7	0.85	2	2
1ciy		3	1	9	0.73	3	0.5	11	0.74	3	3
1cne		2									2
1cqx	A	3	6	7	0.54	2	1	10	0.55	2	2
1crx	A	2	4	8	0.09	2	10	11	0.15	2	3
1cs6	A	4	0.5	9	0.49	3	0.1	8	0.63	4	4
1csg	A	1									1
1ctn		3	10	7	0.33	3	10	11	0.40	2	3
1cun	A	2	4	8	0.62	2	1	11	0.63	2	2
1cwv	A	5	10	7	0.22	4	10	11	0.28	4	4
1d0g	T	3	4	8	0.47	2	10	9	0.48	2	2
1dce	A	3	0.5	10	0.55	6	10	10	0.55	5	6
1dfq	A	2	0.5	10	0.61	2	0.5	11	0.63	2	2
1dg3	A	2	4	8	0.71	3	10	11	0.79	3	4
1djz	A	3	1	9	0.70	2	3	11	0.73	2	3
1dnk	A	1	0.5	10	0.79	1	0.5	11	0.83	1	1

Table 1 – continued

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1duy	A	2	0.5	10	0.63	2	1	11	0.66	2	2
1dvp	A	2	10	7	0.02	3	10	11	0.13	2	3
1e1l	A	2	0.5	10	0.62	2	0.5	10	0.62	2	2
1e2v	A	2	0.5	9	0.79	1	0.5	7	0.83	2	1
1e39	A	3	0.5	10	0.70	3	0.5	11	0.71	2	3
1eaf		1	0.5	10	0.40	1	0.1	7	0.44	3	1
1ega	A	2	0.5	9	0.28	2	0.1	6	0.35	3	2
1eif		2	5	7	0.75	2	10	6	0.79	2	2
1eqf	A	2	0.5	9	0.58	2	0.1	11	0.59	2	2
1fba	A	1	1	8	0.37	1	0.1	6	0.46	4	1
1fbl		2	0.5	10	0.47	2	0.5	11	0.49	2	2
1fc2	C	1	1	8	0.53	1	1	8	0.53	1	1
1fdn		1	5	7	0.69	1	3	6	0.71	1	1
1ffh		2	0.5	10	0.52	2	0.5	11	0.55	2	2
1ffu	A	2	0.5	9	0.73	1	0.1	9	0.77	2	1
1fha		1	10	7	0.43	2	10	7	0.43	2	2
1fmt	A	2	1	9	0.71	2	1	11	0.75	2	2
1fnm	A	5	4	8	0.33	5	2	6	0.46	6	5
1fxi	A	1	4	8	0.40	1	10	8	0.40	1	1
1gdc		1									1
1gdd		2	0.5	10	0.46	2	10	11	0.54	1	2
1gps		1									1

Table 1 – continued

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1grj		2	10	7	0.37	2	9	6	0.53	2	2
1hcc		1									1
1hiv	A	1	0.5	10	0.11	1	10	6	0.17	1	1
1hjp		3	10	7	0.68	2	10	7	0.68	2	2
1hum	A	1	4	8	0.86	1	3	11	0.90	1	1
1ira	Y	3	10	7	0.45	3	0.5	7	0.47	3	3
1jaw		2	0.5	9	0.67	2	0.1	10	0.67	2	2
1kit		3	0.5	9	0.27	3	0.1	11	0.29	4	3
1ksi	A	3	0.5	9	0.59	5	0.1	9	0.62	5	3
1lck	A	2	10	7	0.41 2	10	10	0.48	2	2	
1mda	A	1	0.5	10	0.19	1	0.1	6	0.30	1	1
1mla		2	0.5	9	0.69	2	0.5	9	0.69	2	2
1mrr	A	1	10	7	0.39	1	10	10	0.40	1	1
1mup		1	10	7	0.69	1	10	7	0.69	1	1
1myt		1	0.5	9	0.69	1	0.1	9	0.72	1	1
1ovb		1	5	7	0.65	1	2	7	0.66	1	1
1pdz		2	5	7	0.58	3	0.1	11	0.60	3	3
1pky	A	3	0.5	10	0.70	3	0.1	11	0.71	4	3
1prs		2									2
1prt	B	2	4	8	0.47	2	1	11	0.49	2	2
1qba		4	0.5	10	0.72	5	1	10	0.72	4	4
1qd1	A	2	4	8	0.56	2	10	8	0.56	2	2

Table 1 – continued

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1qdn	A	2	0.5	10	0.67	1	0.1	11	0.72	1	1
1qlf	A	2	1	9	0.67	2	0.5	11	0.70	2	2
1qmd	A	2	1	9	0.38	2	2	11	0.41	2	2
1qmh	A	2	6	7	0.46	3	6	7	0.46	3	3
1qni	A	2	0.5	10	0.32	3	5	10	0.33	2	3
1qnt	A	2	0.5	10	0.46	1	0.5	11	0.49	1	2
1rcb		1	10	7	0.62	1	10	6	0.64	1	1
1rve	A	1	10	7	0.51	1	10	7	0.51	1	1
1sky	E	3	1	8	0.75	2	0.5	8	0.77	3	2
1tbr	S	2	4	8	0.18	2	10	11	0.28	2	2
1tnr	A	1	1	9	0.83	1	1	9	0.83	1	1
1ttb	A	1	0.5	10	0.48	1	0.1	11	0.64	1	1
1tui	A	3	0.5	9	0.42	3	8	6	0.48	3	3
1ula		1	10	7	0.57	1	0.1	10	0.59	1	1
1urk		2									2
1ush		2	1	9	0.79	2	1	9	0.79	2	2
1utg		1	0.5	9	0.58	1	0.1	11	0.62	1	1
1vdc		2	0.5	10	0.64	2	1	10	0.64	2	2
1vfa	B	1	0.5	10	0.56	1	0.1	11	0.57	1	1
1vol	A	2	5	7	0.69	2	1	7	0.76	2	2
1vvc		2	4	8	0.61	2	8	8	0.62	1	1
1wgt	A	4	0.5	10	0.65	3	1	10	0.66	3	4

Table 1 – continued

		manual	parameter set 13				parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
1whe		2									1
1wrp	R	1	5	7	0.73	2	0.1	7	0.75	2	2
1xim	A	1	10	7	0.20	4	10	6	0.29	4	4
1ytf	D	2	0.5	9	0.61	2	0.1	11	0.73	2	2
2bs2	B	2	0.5	9	0.62	2	0.1	10	0.64	2	2
2cbl	A	3	1	8	0.59	3	0.1	11	0.60	3	2
2cgp	A	2	1	9	0.66	2	1	10	0.66	2	2
2ech		1									1
2fcr		1									1
2gb0	A	2	0.5	10	0.74	2	0.1	11	0.77	2	2
2gli	A	5	10	7	0.26	2	10	11	0.36	2	2
2mad	L	1	0.5	10	0.64	1	0.1	11	0.67	1	1
2msb	A	1	1	9	0.65	1	5	6	0.68	1	1
2pcd	M	1	5	7	0.44	1	0.5	6	0.53	1	1
2pf2		1	0.5	10	0.35	1	8	11	0.45	1	1
2por		1	4	8	0.27	1	10	10	0.32	1	1
2shp	A	3	0.5	10	0.69	3	0.5	10	0.69	3	4
2xsc	A	1	1	9	0.23	1	0.5	11	0.26	1	1
3chy		1	4	8	0.79	1	4	8	0.79	1	1
3hdh	A	2	5	7	0.60	2	0.1	6	0.68	4	2
3tec	I	1	0.5	9	0.08	1	0.1	6	0.39	1	1
3tf4	A	2	1	8	0.51	4	10	11	0.53	2	3

Table 1 – continued

		manual	parameter set 13			parameter set 72				$k5, r7$
PDB	chain	#dom	k	r	*corr*	#dom	k	r	*corr*	#dom	#dom
4cp4		1	0.5	10	0.76	1	0.1	11	0.78	2	1
4htc	I	1									1
4icb		1	1	8	0.69	1	1	7	0.72	1	1
4pti		1	5	7	0.76	1	1	7	0.81	1	1
5eau		2	2	8	0.70	3	2	8	0.70	3	4
5nn9		1	0.5	10	0.54	1	0.1	11	0.55	3	2
5p21		1	2	8	0.52	1	2	8	0.52	1	1
6ebx	A	1	2	8	0.78	1	2	8	0.78	1	1

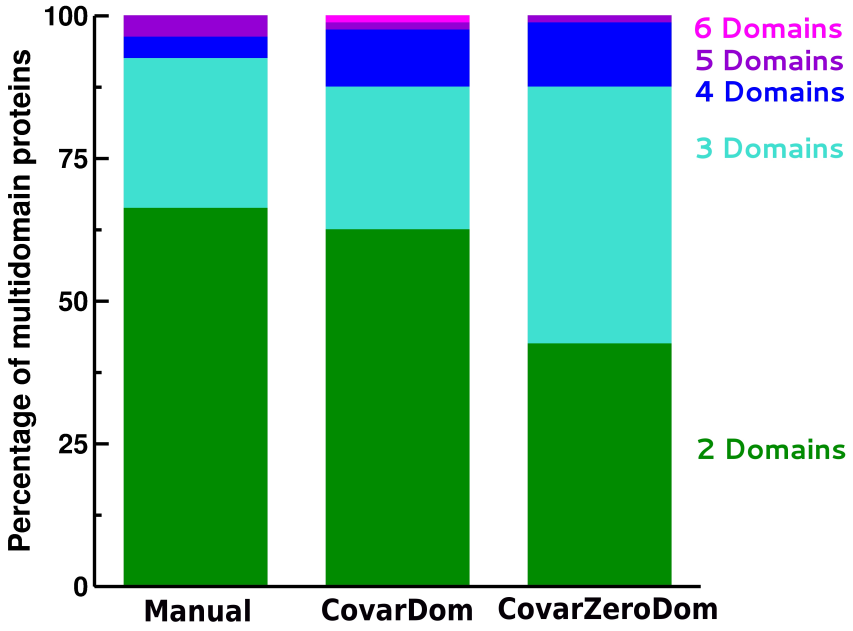

Figure 1. Comparison of the domain numbers assigned to multidomain proteins manually, by CovarDom and by CovarZeroDom. For all three, the total number of proteins assigned as multidomain is eighty, but the protein sets are not the same, because CovarDom assigns five proteins assigned as multidomain manually as 1-domain proteins and five proteins assigned as 1-domain manually as multidomain proteins. The bars indicate the percentage of the multidomain proteins with the specified domain number. CovarDom employs DomainTester after each splitting into two cluster, whereas CovarZeroDom uses negative intercluster covariance as stopping criterion in the clustering procedure.

List of Abbreviations

ADP	anisotropic displacement parameter
ANM	anisotropic network model
APH	aminoglycoside phosphotransferase(3′)-IIIa
BPTI	bovine pancreatic trypsin inhibitor
ENM	elastic network model
GNM	Gaussian network model
H/D	hydrogen/deuterium
MD	molecular dynamics
MSD	mean-square displacement
MUP	mouse major urinary binding protein
NMA	normal mode analysis
NMR	nuclear magnetic resonance
PDB	Brookhaven Protein Data Bank (www.rcsb.org/pdb)
RMS	root-mean-square
RMSD	root-mean-square displacement
VSV-G	vesicular stomatitis virus glycoprotein G

Originalitätserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Ferner erkläre ich, dass ich nicht bereits anderweitig mit oder ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

Bayreuth, den 2. August 2013,

Silke A. Wieninger