INAUGURAL – DISERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen
Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

Vorgelegt von
Dipl.-Ing. Sidonia E. Mesentean
aus Musatesti/Arges, Romania

Tag der mündlichen Prüfung: 20.07.2007

Analysis of Large-Scale Structural Changes in Proteins with focus on the Recovery Stroke Mechanism of Myosin II

Gutachter: Prof. Dr. Joachim Spatz
Prof. Dr. Jeremy C. Smith

Summary

The mechanisms through which proteins achieve their functional three-dimensional structure starting from a string of amino acids, as well as the manner in which the interactions between different structural elements are orchestrated to mediate function are largely unknown, despite the large amount of data accumulating from theoretical and experimental studies. One clear view emerging from all these studies is that function is a result of the intrinsic protein dynamics and flexibility, namely the motions of its well-defined structural elements and their ability to change their position and shape in space to allow large conformational transitions necessary for the function.

Simulation techniques have been increasingly used over the past years in the endeavour to solve the structure-function puzzle as they have proven to be powerful tools to investigate the dynamics of proteins. However, extracting useful dynamical information from trajectories thus generated in order to draw functionally relevant conclusions is not always straight forward, especially when the protein function involves concerted movements of entire protein domains. This is due to the high dimensionality of the energy surface the proteins can explore. Therefore, a decrease in complexity is to be desired and can be achieved in principle by reducing the number of dimensions to the ones capturing only the dominant motions of the protein.

To this purpose, in this thesis two different dimensionality reducing techniques, namely Principal Component Analysis and Sammon Mapping are applied and compared on four proteins that undergo conformational changes with different amplitudes and mechanisms. In particular, the present thesis tackles the large conformational change occurring during the recovery stroke of myosin, using these methods and rigidity analysis algorithms in the attempt to elucidate in atomic detail the structural mechanism underlying the function of this protein that couples ATP hydrolysis to the mechanical force needed to achieve muscle contraction.

The results presented in this thesis show the successful applicability of certain dimensionality reducing methods to large conformational changes and their suitability in analyzing and dissecting dynamical transitions in computationally generated trajectories. The findings regarding the recovery stroke step in the myosin cycle are consistent with experimental data coming from mutational studies and confirm the

previously postulated communication mechanism between the active sites of the protein, thus representing a major contribution to the field of molecular motors and a strong evidence of the importance of theoretical studies in complementing the experimental investigations.

Zusammenfassung

Der Mechanismus durch welchen Proteine, ausgehend von einer Folge von Aminosäuren, ihre funktionsfähige dreidimensionale Struktur erlangen, sowie auch die Art und Weise, in der die Wechselwirkungen zwischen verschiedenen Strukturelementen orchestriert sind, um Funktion zu vermitteln, ist, trotz der großen Datenmengen, die sich aus theoretischen und experimentellen Studien angesammelt haben, größtenteils unbekannt. Eine Anschauung, die sich aus all diesen Untersuchungen herausbildet, ist, dass Funktion ein Resultat der intrinsischen Proteindynamik und -flexibilität ist, und zwar der Bewegungen ihrer wohldefinierten Strukturelemente und deren Fähigkeit, ihre Position und Form zu verändern, um große, für die Funktion notwendige, Konformationsänderungen zu ermöglichen.

In den letzten Jahren sind vermehrt Simulationstechniken in dem Bestreben eingesetzt worden, das Struktur-Funktions-Puzzle zu lösen, da sie sich als mächtige Werkzeuge zur Erforschung der Dynamik von Proteinen erwiesen haben. Nützliche dynamische Informationen aus den so erzeugten Trajektorien zu extrahieren, um daraus funktionsrelevante Schlüsse zu ziehen, ist allerdings nicht immer einfach, besonders wenn die Arbeitsweise des Proteins mit gemeinschaftlichen Bewegungen ganzer Domänen einhergeht. Dies liegt an der hohen Dimensionalität der Energiefläche, die Proteine ablaufen können. Daher ist eine Verringerung der Komplexität erwünscht und kann im Prinzip durch Reduktion der Dimensionen auf jene, welche die dominanten Bewegungen des Proteins erfassen, erreicht werden.

Zu diesem Zweck werden in dieser Arbeit zwei dimensionsreduzierende Techniken, nämlich Hauptkomponentenanalyse (principal component analysis) und Sammon Abbildung (Sammon mapping) auf vier Proteine angewendet und verglichen, die Konformationsänderungen verschiedenen Umfangs und verschiedener Mechanismen durchlaufen. Vornehmlich befasst sich die vorliegende Arbeit mit der großen Konformationsänderung während des “recovery stroke” von Myosin. Die genannten Methoden werden zusammen mit Rigiditätsanalysen benutzt, um den strukturellen Mechanismus aufzuklären, welcher der Funktion dieses Proteins, das ATP-Hydrolyse an die mechanische Kraft koppelt, welche zur Muskelkontraktion benötigt wird, zugrunde liegt.

Die in dieser Arbeit dargestellten Ergebnisse zeigen die erfolgreiche Anwendbarkeit bestimmter dimensionsreduzierender Methoden auf große Konformationsänderungen und deren Eignung für die Analyse und Aufgliederung dynamischer Übergänge in computergenerierten Trajektorien. Die Erkenntnisse hinsichtlich des “recovery strokes” im Myosin-Zyklus sind im Einklang mit experimentellen Daten aus Mutationsstudien und bestätigen den zuvor postulierten Kommunikationsmechanismus zwischen den aktiven Zentren des Proteins, und stellen daher einen bedeutenden Beitrag auf dem Gebiet der molekularen Motoren und einen deutlichen Beweis für die Wichtigkeit theoretischer Studien in Ergänzung zu experimentellen Untersuchungen dar.

Acknowledgments

The content of these acknowledgment pages are at least as important for me as the obtained results presented in the following pages. Without the persons mentioned here none of the published work would have been possible.

First of all, I want to thank to Prof. Jeremy C. Smith for giving me the chance to work with him. His guidance along my PhD work was like a road, always showing the direction along which I have to go in order to achieve my purpose. His charm and geniality was present not only during the working hours but also in the social life of his group. Another mentor to which I want to thank is Dr. Stefan Fischer. He always knew how to bring into the light the best part of my work. He was walking with me side by side through the problems arising from the tasks I had to complete.

To my husband I am deeply indebted for his permanent support, for his unique way of bringing me back on the track when I was lost, for listening my complains and worries and not the last for his love. Together with my family, they found the perfect words for cheering me up and the best mood for celebrating.

A special thank to my parents for their love and for believing in me till the end. Also for fighting against the social and economical hindrances present along my way, they deserve a special place in my mind and heart. My sister also played an important role due to the fact that she was my shadow in all this years and followed my spirit all through the way.

I miss the words when I have to express my gratitude for the help received from dear friends like Crina, Andreea, Durba or Bogdan. They always surrounded me with hope and optimism, independent if it was work, celebrations or just simple discussions.

I would like to thank Ms. Ellen Vogel for her kindness and patience with which she solved most of my administrative matters and for always finding a few minutes to hear my complains. Lots of thanks to all the members of the Computational Molecular Biophysics and Computation Biochemistry groups, for helpful input and discussions regarding my research but also for making my stay in Heidelberg unforgettable.

Last but not least I would like to acknowledge the Bundesministerium für Bildung und Forschung and the Deutsche Forschungsgemeinschaft which financed most of this project.

The following list comprises the publications that resulted from the work presented in this thesis:

[1] Mesentean S., Smith J. C. and Fischer S – "Analyzing Large-Scale Structural Change in Proteins: Comparison of Principal Component Projection and Sammon Mapping", *Proteins: Structure, Function and Bioinformatics*, **64**, 210-218, 2006 (included here as Chapter 3).

[2] Mesentean S., Koppole S, Smith J. C. and Fischer S. - "The principal Motions Involved in the Coupling Mechanism of the Recovery Stroke of the Myosin Motor", *Journal of Molecular Biology*, **367 (2)**, 591-602, 2007 (included here as Chapter 5).

[3] Noe F., Mesentean S., Smith J. C. and Fischer S – "Formation and Dissociation of Rigid Domains in Myosin's Functional Cycle", manuscript in preparation, 2007 (included here as Chapter 6).

Contents

Contents

Contents

Chapter 1

INTRODUCTION

1.1 Structure of proteins

Proteins are miniatural molecular machines that constitute major components of any living organism, being part of the cell nucleus, cytoplasm and membrane. They are made of monomeric building blocks called aminoacids that, according to the interactions they establish with each other and the relative positions they adopt in space, give rise to four hierarchical levels of structure (see Fig. 1.1. below).

Proteins are synthesized as unorganized, linear aminoacid strings which define their primary structure (see Fig. 1.1). When these aminoacids rearrange in space and interact through specific hydrogen bonds, they yield elements of secondary structure such as alpha helices, loops and beta-sheets (Fig. 1.1). These elements of secondary structure can interact through non-covalent bonds such as hydrophobic or ionic interactions and adopt a unique arrangement in space known as the three dimensional structure of proteins, which is the functionally competent one. In some cases, proteins consist of multiple homo- or hetero- subunits (i.e., independent polypeptidic chains with already attained three dimensional structure) that assembly non-covalently, thus giving rise to quaternary structure as exemplified in Fig. 1.1, which is also related to the function of the protein as a whole.

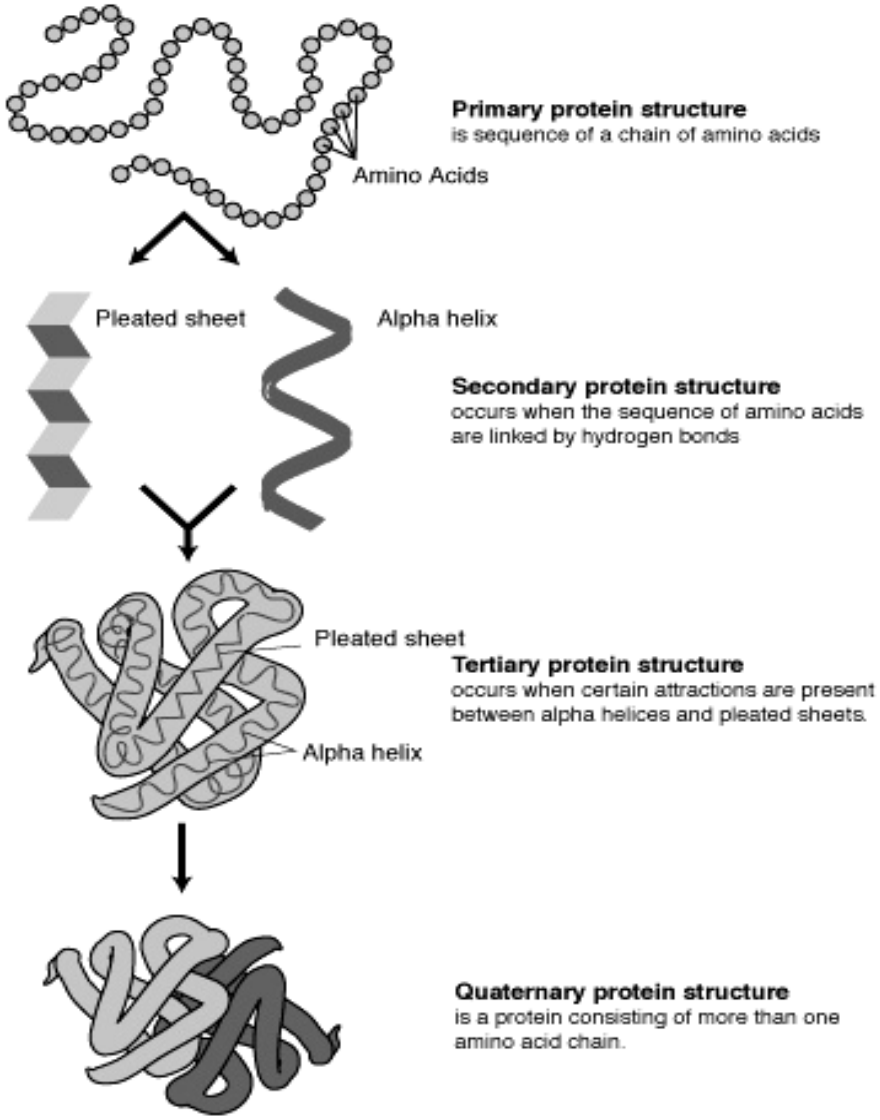


Fig. 1.1 Schematic drawing of the four different levels of protein structure. Picture was taken from http://www.accessexcellence.org/RC/VL/GG/protein.html.

1.2 Protein conformation: the energy landscape concept

The building blocks of any protein, namely the aminoacids contain an amino and carboxy end, as well as one characteristic side-chain for each of the 20 essential aminoacid that form the proteins. To make a protein, the aminoacids form a peptydic

bond via their amnio and carboxy termini, thus yielding the main chain (or the backbone) of the polypeptidic chain (see Fig. 1.2).

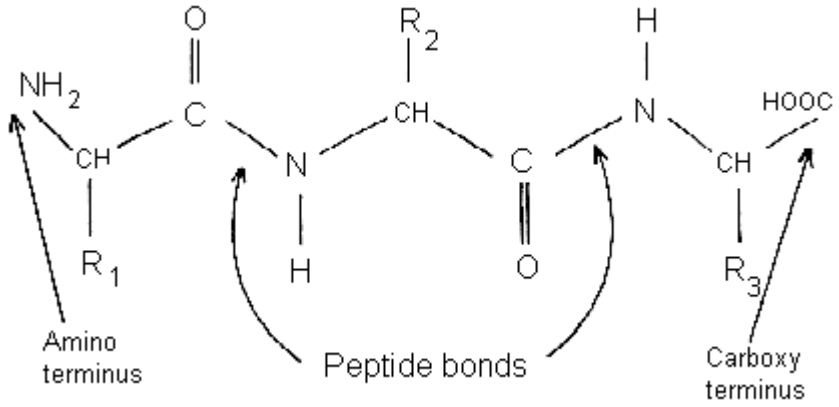


Fig. 1.2 Schematic drawing of a tri-peptide. R1, R2 and R3 denote different or identical aminoacid sidechains.

Both the protein backbone and the individual sidechains of the aminoacids display an intrinsic flexibility due to free rotation possibility around the alpha carbon of the backbone (-CH- atom in Fig.1.2). The rotation occurs around two angles denoted phi and psi as represented in Fig. 1.3.

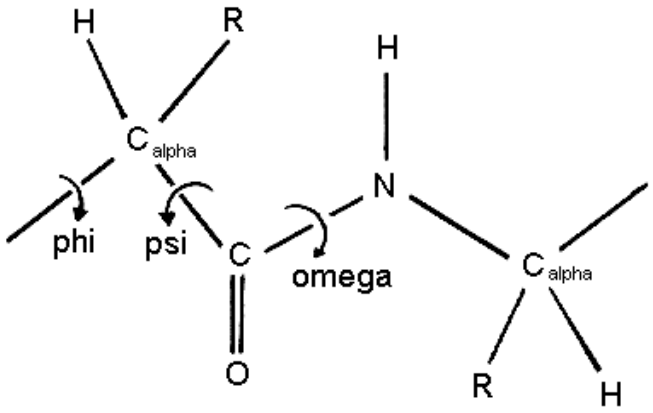


Fig. 1.3 Schematic drawing of a peptidic bond: rotation angles psi and phi around the carbon alpha atom, as well as the rotation angle around the peptidic bond are shown.

This free rotation around the alpha carbon atom provides the protein chain with a huge number of possible combinations of psi and phi angles it can adopt. Each of these combinations for each of the individual aminoacid in the protein defines one possible configuration or conformation. The total number of possible configurations theoretically available to a polypeptidic chain defines the **configurational landscape** or the **conformational landscape** of that protein. However, due to physical clashes between atoms for certain values of the psi or phi angles, not all configurations are explored or physically possible, even though in theory they exist. Of all possible conformations, some exhibit more stability than others. For example, the elements of secondary structure are characterized by energetically and geometrically stable specific combinations of psi and phi angles.

The configurational space of a protein is a function of the free energy of each configuration or arbitrarily chosen configuration parameters such as psi and phi angles. Therefore, one can refer to the configurational space as the **(free) energy landscape** of the protein. Considering the many degrees of freedom of each residue in a protein, the flexibility of the sidechains and main backbone, the energy landscape for one protein is populated by high energy and low energy states corresponding to favorable and unfavorable states. In Fig. 1.4 the free energy landscape for a small helical protein, namely protein G is presented as it resulted from computer simulations. As it can be seen in Fig. 1.4, the free energy landscape contains regions of high energy (from light blue to green) which are not very likely to be populated for long time or at all and regions of low energy (dark blue) which correspond to conformations that are likely to be visited by the protein in its native state. However, one question that arises from such a representation is the following: what are the populated conformations by a protein and how do these conformations look like? The answer to this question is still not given, in spite of the big amount of data accumulating from both experimental and theoretical studies. However, the protein folding theory is able to give parts of the answer as it will explained in the following.

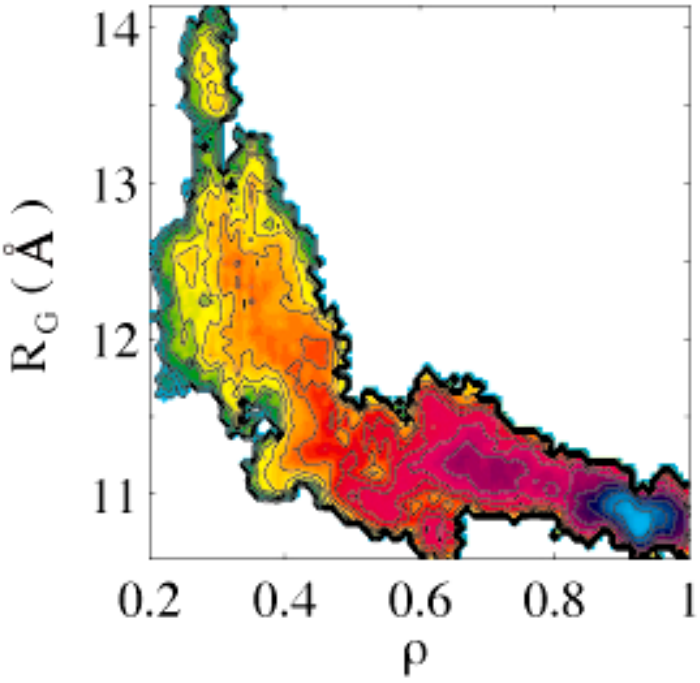


Fig. 1.4 Free energy landscape of protein G as a function of the radius of gyration R_G (i.e., how far it is spreading from its center) and the fraction of native contacts ρ as it results from computational studies performed by Sheineman and Brooks. Colour contours show free energy from the lowest (deep blue) of the native state to the others (blue to pink to red to orange to yellow to green) of less favorable free energy. Picture was taken from http://www.psc.edu/science/Brooks98/.

1.3 Protein folding

One of the major challenges in modern molecular biology is to understand how the functionally competent structure of a protein is attained. From the unfolded primary structure proteins are designed to rapidly adopt an organized three dimensional structure by undergoing a sequential set of spatial rearrangements of their aminoacid main- and side-chains. This process is referred to as **protein folding**. The process is very complex, as proteins exhibit many degrees of freedom. The larger the

number of atoms, the more complicated the protein and its folding becomes. In the past 20 years, one picture has emerged that seems to explain the protein folding in thermodynamical and kinetics terms and that is the **folding funnel** theory as introduced by Onuchic et al[1]. Briefly, according to this theory, a polypeptidic chain exists in a large ensemble of random conformations characterized by high energy which rapidly converge downhill towards the native structure which corresponds to the minimum of free energy and is characterized by very low energy as shown in Fig. 1.5. This state corresponds to the functionally competent state of the protein.

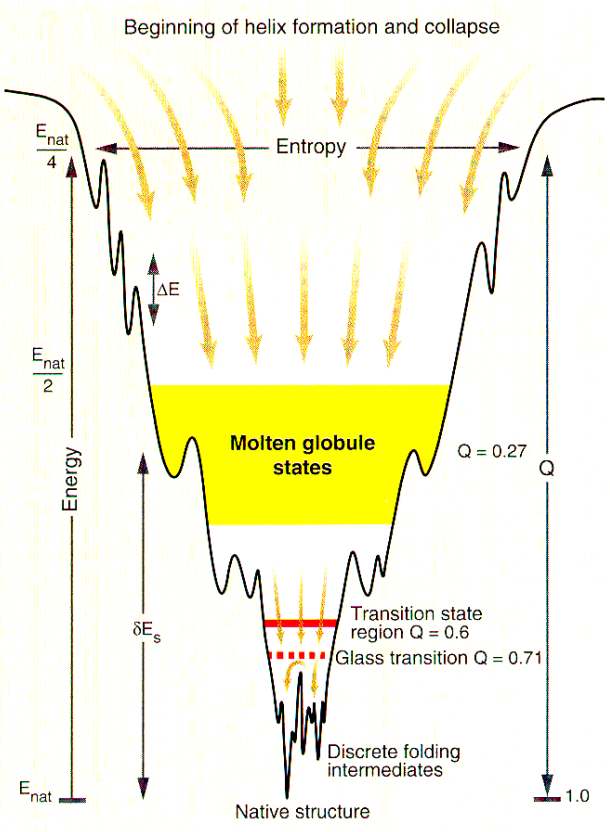


Fig. 1.5. Schematic representation of the folding funnel of a 60-residue helical protein. The width represents the entropy and the depth the enthalpy. The number of native contacts (Q) correctly established is indicated for each collection of the states. Arrows indicate the flow of the molecule from the unfolded to the native structure. The picture was taken from Ref. 1.

According to this figure, the energy landscape of the folding process is characterized by hills and valleys corresponding to high energy and low energy conformations, respectively. The folding funnel explains that, although many conformations are accessible, only the ones that establish correctly the native contacts are allowed to happen. Any misfolded conformations remain trapped in valleys that are situated at high energy levels and, even though they may exhibit a certain stability, they are usually degraded by the cell machinery. The functionally relevant conformation is located in the lowest energy minimum (the native structure in Fig. 1.5).

The folding funnel concept, besides giving useful insight into the folding process, promotes the idea that proteins are dynamical structures and they exist, even in their native state, in many possible conformations at room temperature. This concept came to contradict the idea advanced by X-ray crystallography which provided only static three dimensional structures of one protein conformation. But it is widely accepted and known nowadays that proteins are permanently switching from one conformation to another. Then the question that arises from here is: how is this possible? While folding, the protein sidechain and mainchain undergo a lot of changes in their structure in order to adopt the spatial arrangement with the most favorable energy. This leads to constant rearrangements of structure elements and motions thereof that yield in the end the native state. Since the subject of the present thesis is not concerned with conformational changes during protein folding, which represents the most dramatical conformational change a protein can undergo, only conformational changes associated with protein function, i.e. those that happen in the native state are discussed. The nature and types of these changes are described hereafter.

1.4 Conformational changes: types of motions

In their native state, proteins undergo constant thermal fluctuations that allow them to perform minor (or major) small rearrangements of their backbone or sidechains, thus leading to conformational changes. These changes may occur with only relatively small expenditure of energy. At the molecular structural level, conformational changes in single polypeptides are the result of changes in mainchain torsional angles and side chain orientations. The overall effect of such changes may be localised with reorientations of a few residues and small torsional changes in the regional mainchain. On the other hand torsional changes localised at very few critically placed residues may lead to large changes in tertiary structure. The later type of conformational changes is described as domain motions.

Large proteins are often composed either of domains, i.e., elements of secondary structure that fold independently and interact in later steps of the folding process through non-covalent interactions, and/or of multiple homo or hetero- subunits that lead to quaternary structures. At room temperature, in solution, due to thermal fluctuations, the different domains or subunits can undergo motions relative to each other, thus leading to a different conformation that may be functionally relevant. For example, ligand binding (be that ligand another protein, ATP, small molecules such oxygen) often induces conformational changes in the one protein. These can be domain motions that can be either hinge-like motions or shear-like motions or they can be allosteric motions defined by concerted movement of the different subunits of a protein.

Some very well known and characterized examples are the following. The cooperative rearrangement of the hemoglobin subunits between the oxygen-bound and –unbound states allows an efficient uptake of oxygen in the lungs and its transport to the muscles.[2,3] Another example of complex processes is RasP21 protein which caries a binary state of information that is communicated to other proteins involved in the

process. The transition between the active and inactive forms of Ras, called molecular switch, is an essential control for cell metabolism and is strongly related to the occurrence human tumors[4,5]. A third example is the contraction of the muscle, based on the relative sliding motion of filaments consisting of actin and myosin proteins. This motion is caused by a large conformational change in myosin, called power-stroke, which rotates the myosin head relative to the lever arm[6,7].

Such conformational changes are ubiquitous processes, critical for the biological function. A detailed understanding of complex transitions is interesting from the theoretical point of view and has a high potential impact on medical and biotechnological applications. Conformational changes can be nowadays nicely characterized based on the energy landscape of the protein as explained hereafter.

1.5 Transitions in proteins: types, and timescale

One interesting question, considering the types of motion a protein may undergo is: what kind of transitions are associated with these conformational changes? How does the energy landscape of a protein look like in its native state? For a simpler representation, the energy landscape in this case can be considered a construct in $3N$ dimensions, where N is the number of atoms in a protein. The energy landscape contains valleys and saddle points between valleys[8] (a two dimensional representation can be seen in Fig. 1.6). The roughness of the surface corresponds to local energy minima arising from the many possible stable conformations accessible to the protein. A valley (shown in blue in Fig. 1.6) describes a low energy structure, whereas hills (shown in red-magenta in Fig. 1.6) correspond to high energy structures. This energy landscape can be perceived as a relatively large basin situated at the bottom of the folding funnel. In this large basin, the protein can fall into more than one stable, low energy minima each corresponding to a stable three dimensional and even functional structure. The differences between these structures are in their tertiary and quaternary structures due to conformational changes.

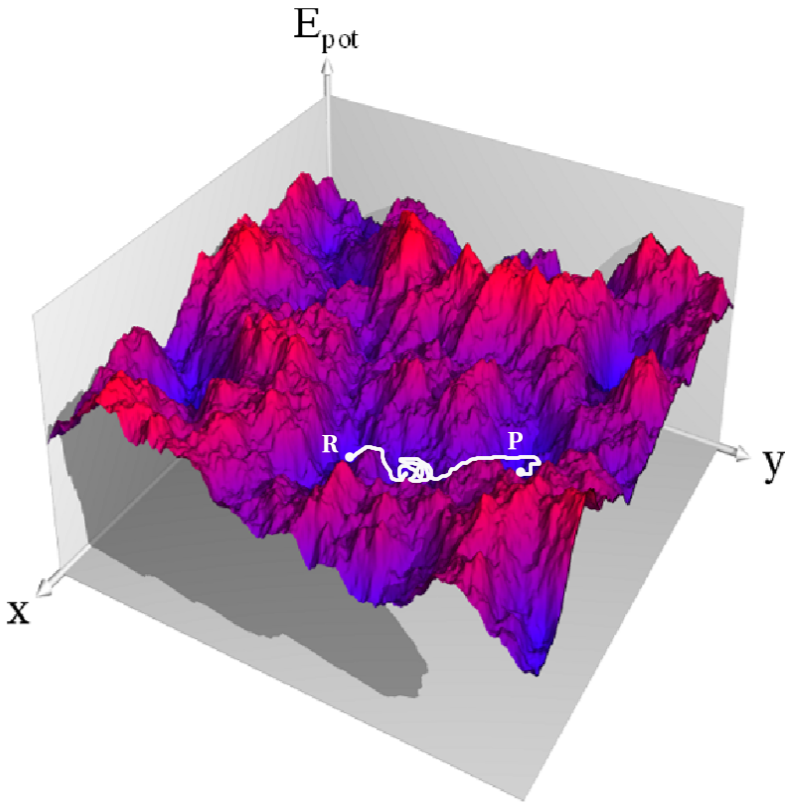


Fig. 1.6. A hypothetical conformational landscape containing a vast number of valleys (blue color) and saddle-points (purple color). Two end-states of a reaction corresponding to two stable conformations, namely the Reactant (R) and the Product (P) states are shown, as well as a possible path connecting them. The landscape is represented as the potential energy as a function of two arbitrarily chosen conformational variables, X and Y.

Due to the fact that proteins have a high number of degrees of freedom (even a medium sized one), the exploration of their conformational space is relative difficult. Transitions between valleys correspond to conformational changes made as sequential events along a path (one example is shown in white in Fig. 1.6). Unfortunately, for complex protein the energy landscape is not known and therefore is difficult to understand the transitions or know the low energy pathways the transition will occur on.

One connection of the energy landscape to function is obvious: Transitions among different conformations are defining protein motions, and protein motions are essential for protein function. Even if this connection in reality is much deeper, finding connections between the structure and the function of a protein remains a challenge.

But how does the protein switch from many possible sub-states? Usually, proteins that are found in one or more conformations are energetically confined to one energy basin whose roughness depends on the complexity of the protein itself. In these basins conformational transitions are a consequence of thermal fluctuations. The exploration of a single energy basin is generally associated with minor conformational changes due to sidechains rotations and small rearrangements. Such small conformational changes are often not responsible for protein function; they occur on timescales from picoseconds to microseconds and are associated with crossing low energy barriers on the energy landscape. In contrast, large conformational transitions involving domain motions for instance are thermally activated and they represent the conformational jump from one energy basin to another. This involves that the end-states of the transition (the *reactant* and the *product* structures, R and P in Fig. 1.6) are both metastable, ***i.e.*** if the protein is in one of these states, it remains there on a relatively long timescale (usually up to microseconds or milliseconds), until a sufficiently strong thermal activation carries it out of that state. This allows designing experiments using X-ray crystallography and nuclear magnetic resonance spectroscopy that provide atomic-detail structures of the end-states, and sometimes long-lived intermediates, of conformational transitions[9]. Such large conformational transitions are associated with protein function; they occur on timescales of microseconds to seconds and involve the activated crossing of high-energy barriers on the energy landscape. The probability to find a protein in a particular conformation is related to the energy of that conformation *via* the Boltzmann distribution [$\exp(-E/k_BT)$], where k_B is the Boltzmann constant and T is the temperature. A thermally activated process is the departure of the system out of such an energy basin by overcoming an energy barrier through random thermal excitation (see Fig. 1.7). At constant temperature, the

probability of such an event per unit time decreases exponentially with increasing the barrier height.

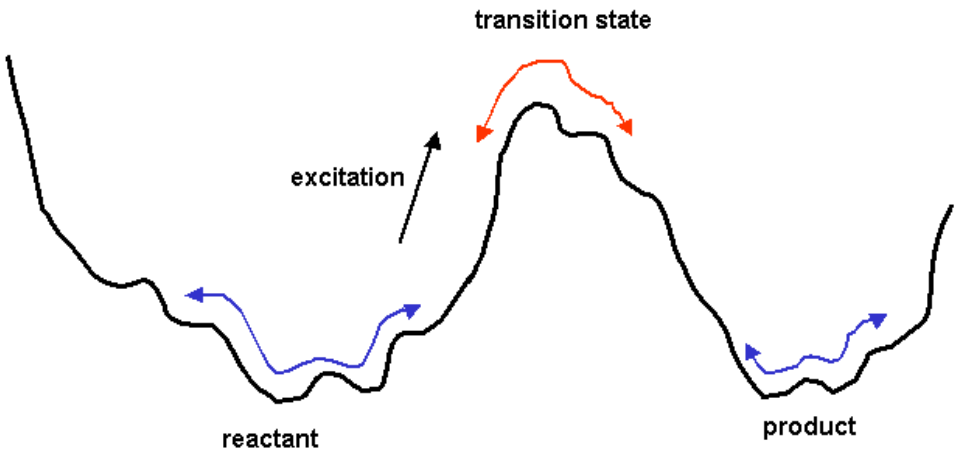


Fig. 1.7. Diagram of the energy barrier corresponding to a transition. Biomolecules diffuse and vibrate in their stable end-states (reactant and product) in a nanosecond time scale (blue color). Occasionally (microsecond time scale – red color), sufficient thermal energy is accumulated to overcome the transition barrier (the transition state).

1.6 Importance of modelling techniques

Large conformational changes in proteins typically involve barriers such that one transition event is expected to happen on a timescale of microseconds or longer (two different energy basins are explored). The activated transition states involved in such a transition, however, are very short-lived as the system quickly relaxes towards lower-energy regions. Because of this transition-state instability, the structural mechanism of transitions can not be experimentally resolved. Complex or large-scale conformational transitions pose a particular challenge because their mechanism (*i.e.,* the order and nature of their sub-transitions) is difficult to predict and may, in principle, occur *via* various pathways. Computer simulation can help to gain insight into these mechanisms.

Molecular biology studies have shown that biological function is without a doubt a consequence of intramolecular motions. The activity and functional properties of the biological molecules are closely associated with their three-dimensional structure and flexibility. Molecules contain atoms that are in continuous motion. Thus, it is necessary to have a representation of atomic motions of the molecules in order to understand the behavior of the whole system. In other words, it is important to have a clear picture of the microscopic phenomena in order to understand the macroscopic world. Molecular modeling techniques have proven to be powerful tools in understanding the dynamical aspects of protein function.

Among the most commonly used molecular modeling techniques employed to gain insight into the dynamics of proteins, Molecular Dynamics (MD) simulations and Molecular Kinematics (MK) methods have been the most successful ones. With the help of MD simulations, only the exploration of one energy basin in atomic detail is possible, due to the limited time of the simulated trajectory (up to microseconds, depending on protein size). MK methods offer the possibility to investigate in atomic detail transitions involving large conformational changes associated with overcoming of high energy barrier, as well as the computation of multiple pathways connected to the transition in order to distinguish the most probable one from a pool of many possible paths.

However, the trajectories generated with computational techniques are very often difficult to analyze due to the complexity of the system. Moreover, motions essential to the protein function may be very difficult to isolate, extract and analyze from such trajectories due to the background noise of the thermal fluctuations. One main challenge when analyzing the dynamical behavior of a protein consists in finding ways to simplify the data, to extract only the significant motions without losing important information and to correlate the motion with function. This may be possible in many cases with the help of dimensionality reducing methods as is explained hereafter.

1.7 Dimensionality reducing methods

Experimental studies involving a large diversity of advanced techniques such as infrared spectroscopy, fluorescence spectroscopy, nuclear magnetic resonance, neutron scattering or x-ray diffraction carried out on a huge number of biological systems have been unable to provide a complete description of the dynamics of a molecule at atomic level. Therefore, molecular modeling methods have been proven useful in bringing the missing details in the picture of the structure/function relationship. In order to avoid problems (like time limitation of the simulations or convergence of the obtained data) arising when trajectories generated with such molecular modeling methods are analyzed, projection methods offered a simple and useful solution by reducing the total information content to the essential one (a simplified example is shown in Fig. 1.8). The main goal when projecting is to represent the input data in a lower-dimensional space in such a way that certain properties of the structure corresponding to the analysed data set are preserved as faithfully as possible. The projection can be used to visualize the data set if a sufficiently small output dimensionality is chosen.

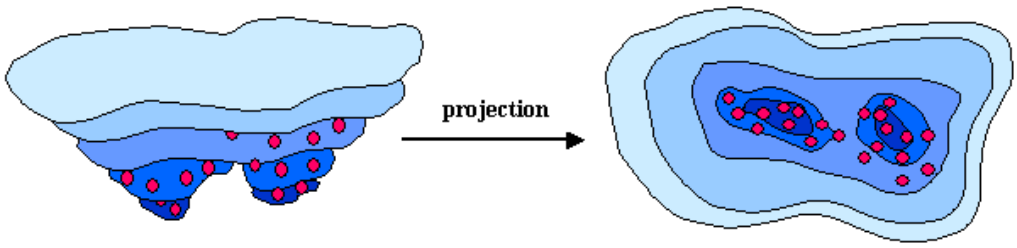


Fig. 1.8. Schematic representation for the projection of a 3D conformational space (left) to a 2D one (right). Color coding corresponds to energy levels: dark blue to low energy and light blue to high energy. Red dots represent possible conformations explored during a simulation trajectory.

There are two main classes of projection methods namely linear and nonlinear. Principal component analysis[10] (PCA) is one of the methods used to display the data as a linear projection on such a subspace of the original data space that best preserves the variance in the data. Another linear projection method is the exploratory projection pursuit[11,12], where as much of the non-normally distributed structure of the data set as possible is sought. This is mainly done by assigning a numerical "interestingness" index (how much the projected data deviates from normally distributed data in the main body of its distribution) to each possible projection, and by maximizing this index. PCA cannot take into account nonlinear structures, structures consisting of arbitrarily shaped clusters or curved manifolds, since it describes the data in terms of a linear subspace. Projection pursuit on the other side tries to express some nonlinearities, but if the data set is high-dimensional and highly nonlinear it may be difficult to visualize it with linear projections onto a low-dimensional display. Therefore, several approaches have been proposed for reproducing nonlinear higher-dimensional structures on a lower-dimensional display. The most common methods allocate a representation for each data point in the lower-dimensional space and try to optimize these representations so that the distances between them would be as similar as possible to the original distances of the corresponding data items. The nonlinear methods differ basically in how the different distances are weighted and how the representations are optimized. One widely used method is the Multidimensional scaling (MDS). The starting point of MDS is a matrix consisting of the pairwise dissimilarities of the entities. The key idea of the method is to approximate the original set of distances with distances corresponding to a configuration of points in a Euclidean space which can be used for constructing a nonlinear projection[13,14]. A problem with the nonlinear MDS is that it is computationally very intensive for large data sets. Another nonlinear projection method closely related to MDS is Sammon Mapping[15]. The main difference between Sammon mapping and the MDS is that the errors in distance preservation are normalized with the distance in the original space and therefore, the preservation of small distances will be this way emphasized.

All this methods have a main common issue namely the fact that they reduce the high dimensionality of the data. This way the complexity of modeling flexibility in proteins can be significantly decreased by reducing the number of necessary dimensions. If linear or nonlinear, these methods are able, to a certain extent, to provide a 2D conformational space where analyzing important macromolecular motions is much simple.

In this thesis the effectiveness of one linear (PCA) and one nonlinear (Sammon Mapping) projection methods is investigated. PCA is a standard method in data analysis; it is well understood and it is one of the most used methods in the analysis of molecular dynamics trajectories of proteins[16-19]. Another advantage is the fact that effective algorithms for computing a PCA projection exist. As a nonlinear alternative Sammon Mapping was chosen for its ability to normalize the errors in the hope that it may highlight details not captured by the complementary PCA. Both methods are applied on pathways of varying complexity corresponding to four different proteins and the corresponding results together with a detailed description of the advantages and disadvantages present on both sides are present in Chapter 3. For a better understanding of the importance and relevance of the theoretical studies in general, focus on a particularly complex protein is made. Since the discovery of myosin (a particular molecular motor that enables skeletal muscle contraction), many experimental and theoretical studies have attempted to elucidate its mechanism at atomic detail. Filtering out and understanding protein motions with a functional role, is a major challenge.

1.8 Aims of the thesis

As explained before, the motions involving large parts of proteins occur in nature mostly upon ligand binding (whatever the nature of the ligand may be) and are associated with the protein activation from one state that is not active to one that can perform a certain function. Therefore, identifying, cahracterizing and understanding these motions is essential to bridge the structure to function.

Large scale changes in proteins pose a challenge to computational approaches and therefore the main aims of this thesis are to find better ways of employing existing methods to identify, extract and analyze the essential protein motions responsible for the protein function during complex transitions in simulated trajectories. In this context, two main questions were addressed during this work:

- How can we find a simplified representation of the conformational space explored by protein dynamics?

- Can the motion behind a large conformational transition be captured in a low dimensional space?

In order to answer these questions, Sammon Mapping and PCA algorithms were compared in a first series of studies included in this thesis on four well-studied proteins whose architecture, function and dynamics are known to be very different from one another. The corresponding trajectories were created using two different simulation methods, i.e., Molecular Dynamics and Conjugate Peak Refinement. In a second series of analysis included in the present work, more focus is being made towards understanding the coupling between chemical and mechanical energy in myosin by analysing the essential motions during the myosin II recovery stroke captured by the dominant Involvement Coefficients. In this context, more specific questions were asked:

- Being known that the recovery stroke is coupled to the activation of myosin's ATPase by a mechanism that is essential for an efficient motor cycle; can the principal motions of myosin during this step be extracted out of equilibrium molecular dynamics simulations?
- Is there a sequence of events present behind the large conformational change of myosin during the recovery stroke? If yes, is there an easy way of filtering this out?

- Is this mechanism governed by the motion of rigid domains? Are they formed with a specific purpose?

- Are these motions correlated?

The results presented in this thesis should provide insight into the usefulness and applicability of Sammon Mapping and PCA algorithms in the analysis of complex transitions. In the same time, biologically relevant questions should be answered by extracting essential motions of myosin in the endeavour to elucidate the mechanism of this protein.

References

1. Onuchic J.N., Wolynes P.G., Luthey-Schulten Z. & Socci ND.- Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci USA* **92(8)**, 3626-3630, 1995.
2. Olsen K., Fischer S. & Karplus M. – A continuous path fort he T R allosteric transition of hemoglobin, *Biophys. J.*, **78**, 394A, 2000.
3. Perutz M.F., Wilkinson A.J., Paoli M. & Dodson G.G. – The stereochemical mechanism of the cooperative effects in hemoglobin revised. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 1-34, 1999.
4. Coleman M.L., Marschall C.J. & Olson M.F. – Ras and Rho GTPases in G1-Phase Cell-Cycle Regulation, *Nat. Rev. Mol. Cell Bio.*, **62**, 851-891, 1993.
5. Vojtek A.B. & Der C.J. – increasing complexity of the Ras signaling pathway, J. Biol. Chem., **62**, 851-891, 1993.
6. Fischer S., Windshügel B., Horak D., Holmes K. C. & Smith J. C. – Structural mechanism of the recovery stroke in the Myosin molecular motor, *Proc. Natl. Acad. Sci*. **102**, 6873-6878, 2005.
7. Geeves M. A. & Holmes K. C. - Structural mechanism of muscle contraction, *Ann. Rev. Biochem.* **68**, 687-728, 1999.
8. Frauenfelder H. – Proteins: paradigms of complexity, PNAS, 99, 2479-2480, 2002.
9. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N & Bourne P. E. – The protein data bank, *Nucl. Acids Res.* **28**, 235-242, 2000.
10. Hotelling H. - Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417-441,498-520, 1933.
11. Friedman J. H. - Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249-266, 1987.
12. Friedman J. H. & Tukey J. W. - A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881-890, 1974.

13. Kruskal J. B. - Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1-27, 1964.
14. Shepard R. N. - The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125-140; 219-246, 1962.
15. Sammon Jr. J. W. - A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401-409, 1969.
16. Hayward S., Kitao A. & Go N. - Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and Principal Component Analysis, *Protein Sci*. **3**, 936-943, 1993.
17. Amadei A., Linssen A. B. M. & Berendsen H. J. C. - Essential dynamics of proteins, *Proteins: Structure, Function, and Genetics* **17(3)**, 412-425, 1993.
18. Hayward S., Kitao A., Hirata F. & Go N. - Effect of solvent of collective motions in globular protein, *J. Mol. Biol*. **234**, 1204-1217, 1993.
19. Tournier A. & Smith J. C. - Principal Components of the Protein Dynamical Transition, *Physical Review Letters* **91(20)**, 208106-4, 2003.

Chapter 2

METHODS

In this chapter are presented the simulation techniques used to generate trajectories of complex transitions in hemoglobin, SNase, Ras p21 and myosin, as well as the methods employed to analyze these trajectories by reducing the space dimensionality. For better understanding and coherence of the analysis, a brief description of the simulation techniques is given in the beginning, followed by a thorough and more detailed presentation of the methods used to analyze the simulated trajectories of the conformational changes in the chosen protein systems.

2.1 Force field description

As mentioned in the general Introduction, proteins explore different regions of the free energy landscape in order to achieve their function. To understand the mechanism of the transitions between different conformations on the conformational space at atomic detail, simulation techniques were used to generate trajectories starting from different initial configurations of the selected protein systems. As it is very difficult to estimate the free energy in simulation techniques due to the unknown entropic contribution, the potential energy is used. This is represented as a function depending on different empirically-derived parameters as will be described hereafter and describes the set of interactions between the atoms in the system.

As it will be later on presented, the energy function contains many parameters, determined by fitting simulation data obtained by using the energy function against data from experiments or quantum-mechanical *ab initio* calculations. Molecular simulation packages include a definition of the energy functional terms and also deliver parameters for the atom types that typically occur in biomolecules. By specifying the topology of a protein (*i.e.* which atoms are contained in the protein and how are they bonded), the energy function for this particular protein is defined and can be used as a computational model.

The underlying potential energy surface (PES) is essential in describing the native state and folding dynamics of real proteins. Therefore it is important to visualise and analyse the features of the energy landscape[1]. For minimalist systems of even moderate size (tens of beads) the PES has very high dimensionality, and these dimensions cannot all be represented on a single plot. A simple representation of its roughness containing a vast number of minima and saddle points is shown in Fig. 2.1.

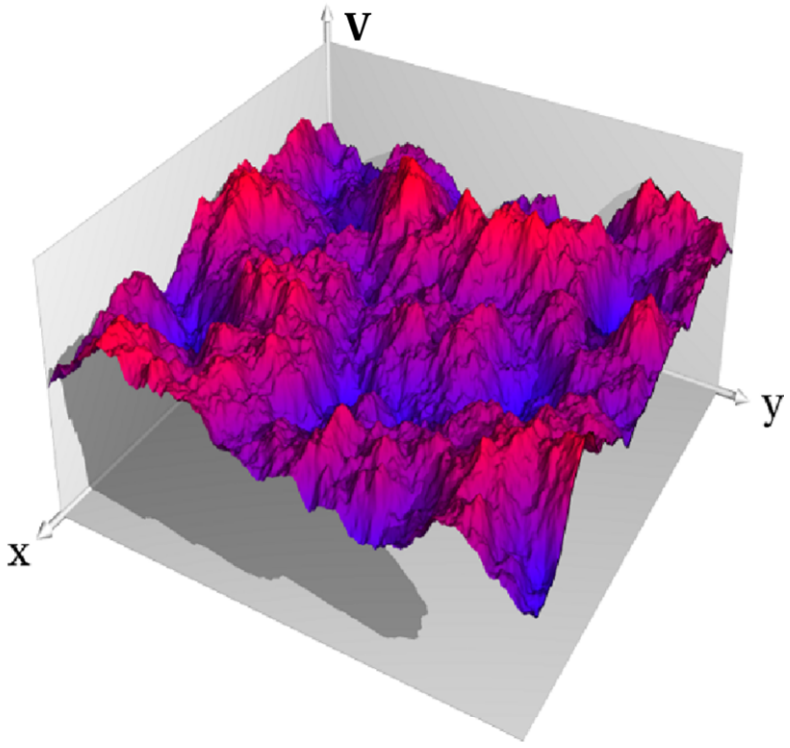


Fig. 2.1 A very rough potential energy surface defined in a 3*N*-dimensional space and containing a vast number of minima (blue color) and saddle-points (purple color).

The energy of large biomolecules is usually calculated by empirical potential energy functions *i.e.,* by molecular mechanics force fields[2]. Among the most commonly used potential energy functions are the AMBER[3], GROMOS[4], and CHARMM[5] force fields. Current generation of force fields provide a reasonably good compromise between accuracy and computational efficiency. Their ability to reproduce experimentally measured physical properties has been extensively tested. The trajectories studied in this thesis were calculated using CHARMM force field, where the description of the interatomic forces is split into two categories: the bonded terms (V_{bonded}) and the non-bonded terms ($V_{nonbonded}$). The bonded terms regroup simple covalent binding as well as the more complex hybridization and π-orbital effects and include bonds, angles and dihedrals terms. The non-bonded terms describe the van der Waals forces and the electrostatic interactions between the atoms. The total energy, V is given by:

$$V(\vec{R}) = V_{bonded}(\vec{R}) + V_{nonbonded}(\vec{R})$$
$$= \underbrace{V_{bonds} + V_{angles} + V_{impr} + V_{dihedrals}}_{V_{bonded}} + \underbrace{V_{vdW} + V_{elec}}_{V_{nonbonded}} \quad (2.1)$$

where, V_{bonds} is the bond stretching energy term, V_{angles} is the angle bending energy term, $V_{dihedrals}$ accounts for rotation along a bond, V_{impr} is the distortion energy term, V_{vdW} the van der Waals energy and V_{elec} the electrostatic energy. These terms are schematically drawn in Fig. 2.2 and are further presented in more detail.

Bond stretching. The bond stretching term describes the forces acting between two covalently bonded atoms. The potential is assumed to be almost harmonic:

$$V_{bonds} = \sum_{bonds} k_b (l - l_0)^2 \quad (2.2)$$

where l is the distance between the two atoms and l_0 is the ideal bond length. The force constant, k_b, determines the strength of the bond.

Angle bending. The angle bending terms describes the force obtained from the deformation of the valence angles between three covalently bonded atoms. The angle bending term is also described using a harmonic potential:

$$V_{angles} = \sum_{angles} k_\theta (\theta - \theta_0)^2 \quad (2.3)$$

where θ is the angle between three atoms. Two parameters characterize each angle in the system: the reference angle θ_0 and a force constant k_θ.

Torsional terms. The torsional terms are weaker than the bond stretching and angle bending terms. They describe the barriers to rotation existing between four bonded atoms. There are two types of torsional terms: proper and improper dihedrals. Proper torsional potentials are described by a cosine function:

$$V_{dihedrals} = \sum_{1,4\,pairs} k_\phi (1 + \cos(n\phi - \gamma)) \quad (2.4)$$

where ϕ is the angle between the planes formed by the first and the last three of the four atoms. Three parameters characterize this interaction: γ sets the minimum energy angle, k_ϕ is a force constant and n is the periodicity.

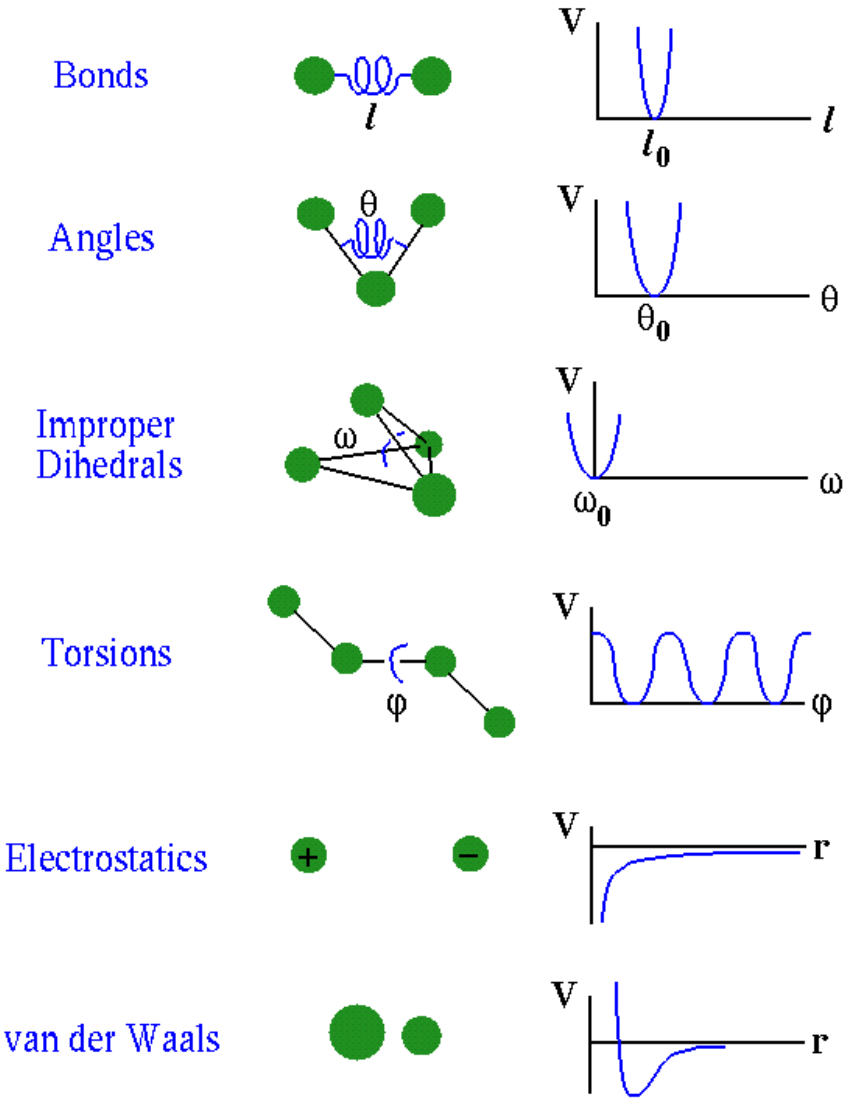


Fig. 2.2 Schematic representation of the bonded and nonbonded interaction terms contributing to the force field: bond stretching, angle bending, torsional terms, electrostatic and van der Waals interactions.

The improper dihedral term is designed both to maintain chirality about a tetrahedral heavy atom and to maintain planarity about certain atoms. The potential is described by a harmonic function:

$$V_{impr} = \sum k_w (\omega - \omega_{eq})^2 \qquad (2.5)$$

where ω is the angle between the plane formed by the central atom and two peripheral atoms and the plane formed by the peripheral atoms (see Fig. 2.2).

Van der Waals interations. The van der Waals force acts on atoms in close proximity but which are not covalently bonded together. It is strongly repulsive at short range and weakly attractive at medium range. The interaction is described by a Lennard-Jones potential:

$$V_{vdW} = \sum_{i,k} A\left(\frac{\sigma_{ik}}{r_{ik}^{12}} - \frac{\sigma_{ik}}{r_{ik}^{6}}\right) \tag{2.6}$$

where r is the distance between two atoms. It is parameterized by σ, the collision parameter (the separation for which the energy is zero) and A, the depth of the potential well.

Electrostatic interactions. Finally, the long distance electrostatic interaction between two atoms is described by Coulomb's law:

$$V_{elec} = \sum_{i,k} \frac{q_i q_k}{\varepsilon r_{ik}} \tag{2.7}$$

where q_i and q_k are the charges of both atoms and r_{ik} the distance between them. ε is the effective dielectric function for the medium.

So finally, the equation for the potential energy describing the force field can be written:

$$\begin{aligned} V(\vec{R}) = & \sum_{bonds} k_b (l - l_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \\ & + \sum_{dihedral} k_\phi (1 + \cos(n\phi - \gamma)) + \sum_{impr} k_\omega (\omega - \omega_0)^2 + \\ & + \sum_{i,k} A\left(\frac{\sigma_{ik}}{r_{ik}^{12}} - \frac{\sigma_{ik}}{r_{ik}^{6}}\right) + \sum_{i,k} \frac{q_i q_k}{\varepsilon r_{ik}} \end{aligned} \tag{2.8}$$

The energy function assigns a value V to each molecular configuration R, *i.e.* to each atom position vector. For an N-atomic system each configuration is defined by $3N$ coordinates. Therefore V defines an energy surface on a $3N$-dimensional hyperspace.

2.2 Molecular Dynamics Simulations

One of the principal tools in the theoretical study of biological molecules is the method of molecular dynamics simulations (MD). Molecular dynamics simulations permit the study of complex, dynamic processes that occur in biological systems. These include, for example: protein stability, conformational changes, protein folding, molecular recognition or ion transport in biological systems. This computational method calculates the time dependent behaviour of a molecular system consisting in the integration of Newtonian equations of motions for a system of interacting atoms under a particular set of forces. MD simulations provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids. The outcome of the simulation consists in the trajectory of all the atoms during the time covered by the simulation. This trajectory can then be analyzed to reach a new understanding of the system based on the atomistic description offered by MD. Many quantities that characterize the dynamical behavior of the system can be computed from the trajectory (for example radius of gyration, root mean square fluctuations, percentage of secondary structure) and directly compared with experiments. Therefore, MD simulations are not performed to replace experiments, but rather to complement and help interpreting the experimental studies that lack atomic level description of the biological processes.

Initial conditions are defined by setting the atomic coordinates according to one of the experimentally known transition end-states of the protein, and the corresponding velocities according to a distribution that yields a desired overall temperature. Then a system of differential equations is integrated (*e.g.* Newtonian dynamics or Langevin dynamics) and the trajectory of the system through phase-space is followed. The limitation of this approach is that the presently accessible simulation time is in the range of 10 to 100 nanoseconds, given the complexity of the calculations and the allowable length for an integration time step. Therefore, MD simulations were usually applied for the study of small conformational transitions. Nowadays this

approach seems to be appropriate also when studying larger conformational changes and one such case will be presented later on in this thesis.

Basic steps in performing MD

MD simulations yield successive configurations of the system that represent its natural evolution in real time. The essence of the molecular dynamics technique is the numerical integration of Newton's second law relating the mass and acceleration of an atom in the system to the gradient of the potential energy field. The atomic positions and velocities of every atom *i* of the system are then derived by integrating the equation:

$$\vec{F}_i = m_i \vec{a}_i = m_i \frac{\delta^2 r_i}{\delta t^2} \tag{2.9}$$

At every step of the simulation this equation is solved for every atom *i* characterized by a mass m_i and subjected to a force F_i and acceleration a_i. The force F_i is the first derivative of the potential function *V* (given in Equation 2.8) with respect to the cartesian coordinates of every atom *i*:

$$\vec{F}_i = \frac{\delta V(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)}{\delta \vec{r}_i} \tag{2.10}$$

while the velocity of atom *i* is given by: $\vec{v}_i = \frac{\delta \vec{r}_i}{\delta t}$

The acceleration a_i is introduced in a Taylor expansion that gives the position of atom *i* at time $t+\delta t$. The high-order terms of the Taylor expansion are neglected to reduce the computation time:

$$\begin{aligned} \vec{r}_i(t+\delta t) &= \vec{r}_i(t) + \frac{\delta \vec{r}_i(t)}{\delta t} dt + \frac{1}{2} \frac{\delta^2 \vec{r}_i(t)}{\delta t} dt^2 + ... \\ &= \vec{r}_i(t) + \vec{v}_i(t) dt + \frac{1}{2} \vec{a}_i(t) dt^2 + ... \end{aligned} \tag{2.11}$$

which represents the integration equation used by the Verlet algorithm implemented in CHARMM. Leap frog and velocity Verlet variants are also available. For detailed

discussion of these methods and other numerical methods for integrating differential equations, see Ref. 6 and 7.

A typical molecular dynamics run involves the following basic steps:

1. *Preliminary preparation* - A molecular structure with all Cartesian coordinates defined is required for a dynamics simulation. After determining the internal coordinates values of the molecule, total energy as a function of the Cartesian coordinates is computed by evaluating the individual terms of the energy equation.

2. *Minimization* - All dynamics simulations begin with an initial structure that may be derived from experimental data. Energy minimization is performed on structures prior to dynamics to relax the conformation and remove steric overlap that produces bad contacts. In the absence of an experimental structure, a minimized ideal geometry can be used as a starting point.

3. *Heating* - A minimized structure represents the molecule at a temperature close to absolute zero. Heating is accomplished by initially assigning random velocities according to a Gaussian distribution appropriate for that low temperature and then running dynamics. The temperature is gradually increased by assigning greater random velocities to each atom at predetermined time intervals.

4. *Equilibration* - Equilibration is achieved by allowing the system to evolve spontaneously for a period of time and integrating the equations of motion until the average temperature and structure remain stable. This is facilitated by periodically reassigning velocities appropriate to the desired temperature. Generally, the procedure is continued until various statistical properties of the system become independent of time.

5. *Production* - In the final molecular dynamics simulation, CHARMM takes the equilibrated structure as its starting point. In a typical simulation, the trajectory traces the motions of the molecule through a period of at least 10 picoseconds. Just as with energy minimization, provision is made to update the nonbonded and hydrogen bonded lists periodically. Additional options are available, making the dynamics facility quite flexible.

A molecular dynamics run generates a dynamics trajectory consisting of frames (set of coordinates) and velocities that represent the trajectory of the atoms over time. Using trajectory data, you can compute the average structure and analyze fluctuations of geometric parameters, thermodynamics properties, and time-dependent processes of the molecule. Preliminary analysis is possible using commands provided in the coordinate manipulation facility. More detailed perturbations can be monitored using correlation functions. Because molecular dynamics runs often require considerable amounts of computer time, a restart facility is available that allows you to suspend the simulation and resume the calculation.

2.3 Conjugate Peak Refinement

MD simulations allow the proteins to explore limited regions in the conformational space. This is due to the fact that the timescales accessible to MD simulations are on the order of tens of nanoseconds (10^{-9} s). In this time range molecules cannot easily undergo large conformational transitions because they are confined to a limited area on the energy surface where they can only jump between energy minima separated by low barriers. Consequently, only low-energy conformations are being sampled with such a technique.

Existing methods (like Molecular Kinematics or Conjugate Peak Refinement) meant to identify pathways between reactant and product, are available in order to investigate activated processes. The definition of the reactant and the product should be understood in terms of two structures located in two minima on the energy surface corresponding to stable molecular conformations. A reaction pathway describes a transition from the reactant to the product meant to characterize the essential motions that occur during this process. The use of such techniques implies that the structures of the reactant and product are known. The path starting from the reactant follows the energy gradient up to a maximum point called “saddle point” and then down to the minima in which the product is situated. Therefore, finding the reaction pathway

connecting two energy minima implies finding the saddle points along that pathway (see Fig. 2.3).

The description of the reaction pathway between a given reactant and a product is not always a trivial issue (due to the roughness of the energy landscape). There are many possible pathways connecting the reactant and the product and most of them are not explored by the system in nature because they involve crossing of very high barriers meaning very slow transitions. Due to the fact that in nature such transitions are optimized (the lowest barrier corresponding to the energy minimum path should be overcome), searching for a pathway connecting the reactant and the product is reduced to the search for the adiabatic reaction path (also called minimum energy path).

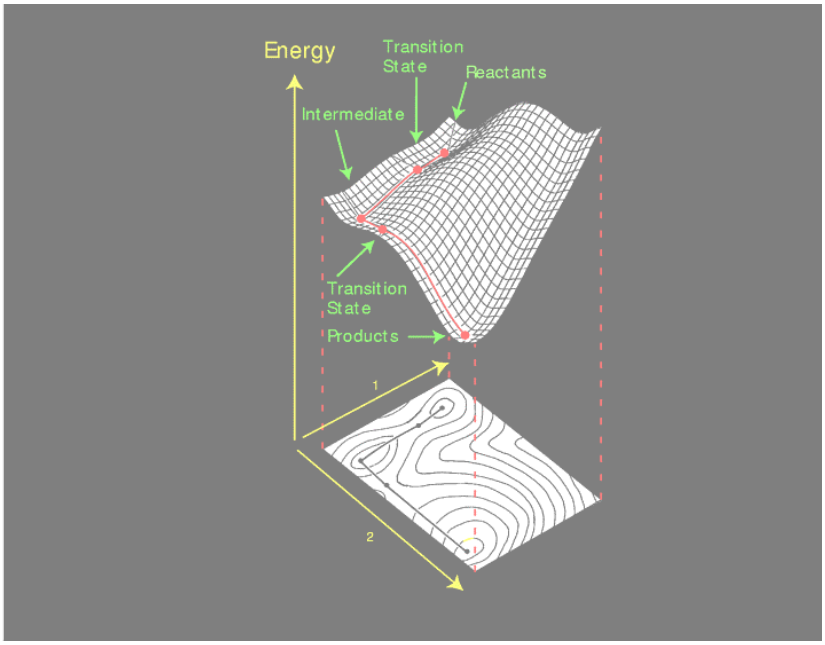


Fig. 2.3 Two-dimensional energy landscape showing a pathway (red line, upper scheme) connecting the reactant and product states through an intermediate state via two saddle points. Reactant and product states are situated in the valleys of the energy surface, while transitions states are characterized by high energies (hills) along the pathways. The two reaction coordinates 1 and 2 are arbitrarily chosen.

2.3.1 Method description

Conjugate Peak Refinement (CPR) belongs to the category of global methods used to find reaction pathways and it is a reliable tool to accurately determine the first-order saddle points along a path connecting the reactant and the product states[8]. Finding the exact saddle point enables the correct activation barrier height to be computed. The method requires a continuous energy function and the first derivative of this function.

The CPR method is a very robust procedure that can be applied to complicated reactions that involve multiple saddle points. The algorithm is capable of identifying all the saddle points along a path and therefore there are no hidden energy barriers within path segments. The parameters of the algorithm are designed to be applicable to any system regardless of the complexity of the reaction and of the size of the system. For this reason, CPR is well-suited for the study of large-scale conformational changes like the one myosin undergoes during the recovery stroke. It has also been successfully applied in elucidating the catalytic mechanism of cis-trans prolyl isomerase, in describing the entry pathway of ligands into binding sites, in explaining the quaternary transitions in hemoglobin, in identifying phases along the recovery stroke transition of myosin or in revealing the mechanism of sugar transport across membranes[9-13].

One of the main characteristics and advantages of the CPR method is that the reaction is not steered along a pre-defined reaction coordinate. On the contrary, rather than defining a reaction coordinate as a function of one of the degrees of freedom in the system, in the CPR method no constraints are applied and all degrees of freedom are allowed to contribute to the reaction. This way the reaction coordinate is determined by the intrinsic reaction coordinate (λ, describing the progress of a conformer in the configurational space) which during such a pathway is defined as a sum of root mean square (RMS) distances along path segments:

$$\lambda_n = \sum_{i=1}^{n} d_i \tag{2.12}$$

where d_i represents the RMS distance between two consecutive path points.

The main focus of the CPR method is to accurately find the structures corresponding to saddle points along a given path. The transition states represent the key towards understanding the nature of the reaction mechanism and therefore their accurate identification is crucial when computing reaction pathways. CPR optimizes the path connecting the reactant and the product as a chain of conformers finding all the saddle points along that path. The continuity of the path is given by the path points, which are different from the saddle points. These intermediate path points are optimized so that the energy decreases monotonically from the saddle points towards the neighboring minima. The number of the path points is not fixed as in the case of other global methods. It can increase or decrease during the refinement, thereby providing flexibility in accommodating any degree of complexity along the path. However, the path points are not uniformly distributed along the trajectory. They are more abundant in the region of the saddle point because that region is thoroughly refined by the algorithm.

The result of a CPR calculation is a continuous trajectory (a kind of "movie" corresponding to the studied reaction) that is time- and temperature-independent. The time parameter is described by:

$$k = \frac{k_B T}{h} \cdot e^{-\frac{\Delta G^{\ddagger}}{RT}} \tag{2.13}$$

where k is the rate constant, k_B is the Boltzmann constant, T is the temperature, h is the Planck constant and $\Delta G^{\ddagger}$ is the free energy of activation (also called the barrier height) and R is the gas constant.

2.3.2 Algorithm

The theoretical basis of the method is given in detail elsewhere[8] but a brief introduction into the algorithm features is given in this section.

The input for the CPR algorithm, as previously mentioned, is represented by the two structures of the reactant and the product. These are available from X-ray or

NMR experiments as discussed before and they have to be minimized before using them for a CPR calculation so that they lie in local minima on the energy surface.

The starting point for a CPR calculation is represented by a first guess of the path connecting two energy minima. The first guess of the path can be a linear interpolation between the reactant and the product structures. This initial path is than optimized. However, many possible paths connecting the reactant and product exist and they need to be explored in order to identify the one with the lowest activation barrier. Therefore, in search for the minimum energy pathway, many paths have to be generated. Alternative paths are created by making different initial guesses. This is achieved by inserting intermediate structures into the initial path. These intermediate structures can be also guesses of the possible structures connecting the reactant and the product (usually constructed manually). The different initial guesses are also optimized and upon refinement the intermediate structures are removed from the path.

For the purpose of speeding up the computational costs when optimizing an initial path, some of the atoms that are not expected to contribute to the transition from reactant to product can be fixed. Since these fixed atoms still contribute to the shape of the reaction path through the non-bonded and bonded interactions with the moving atoms, their coordinates have to be the same in the reactant, intermediate and product structures.

The procedure for finding saddle points is based on energy minimization and maximization. Line maximization is employed to find the energy peak on a path segment, from which line minimization is performed in a direction conjugated to the path to find an energy minimum (see Fig. 2.4 for an example). This energy minimum is added as a path point along the path and the algorithm starts looking for energy peaks on other path segments. At each step of the algorithm, energy peaks from all path segments are compared and the highest one is optimized first. The algorithm is heuristic in the sense that this procedure for finding peaks of high energy along the path is iterative, path points being dynamically added, removed or improved in **CPR cycles**. The path is considered fully refined when the only remaining peaks along the path are the exact saddle-points.

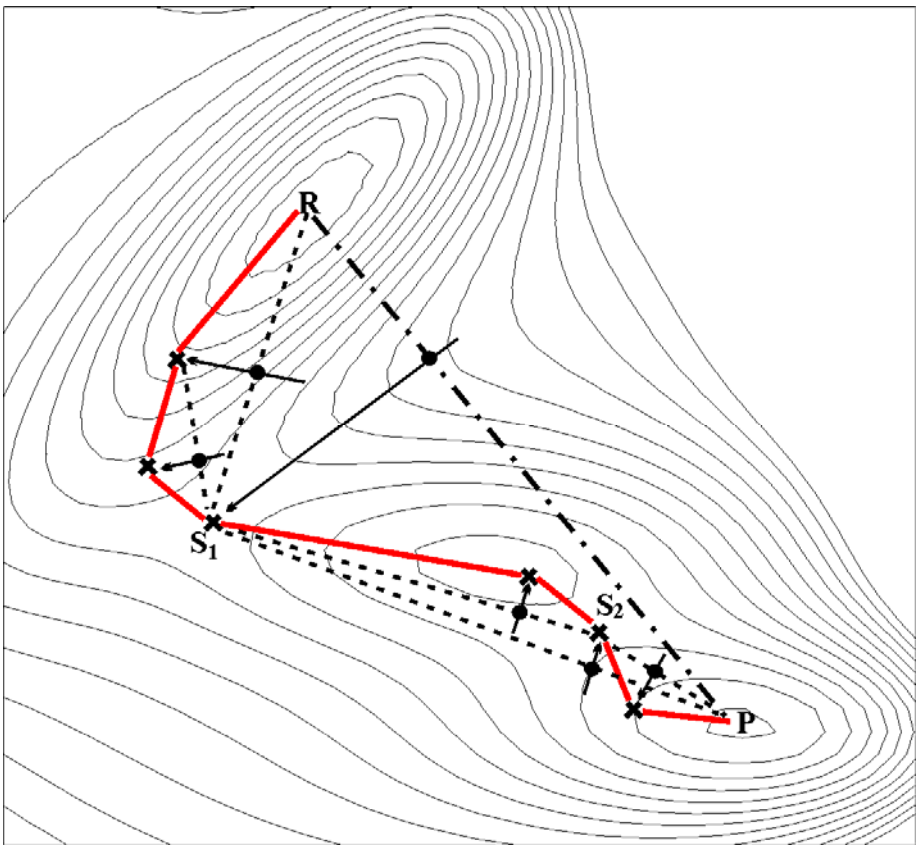


Fig. 2.4 Exemplification of the CPR algorithm on a 2D energy surface. Line maximizations find peaks of high energy (black dots) along path segments (dotted lines), while line minimizations of the peaks in a direction conjugated to the direction of the path lead to path points on the paths (black crosses). The procedure is applied iteratively for each path segment until the only high energy peaks on the path are represented by first-order saddle-points (S_1 and S_2 in this case). The algorithm yields a continuous path (red line) connecting the reactant R and product P (both lying in two energy minima on the energy surface). Picture was adapted from Ref. 8.

2.3.3 Comparison with experimental data

Exploration of multiple pathways connecting two given end-states leads to accumulation of paths characterized by different transition states and barrier heights. In order to decide which path is the most likely one, comparison with the experimental data available for the studied process is vital. However, experimental activation

barriers ($\Delta G^{\ddagger}$) cannot be directly compared to the computed barrier heights because the entropic effects that contribute to the experimental barriers cannot be accounted for in these force fields. Therefore, the comparison must be done between experimentally-derived and computationally-determined activation enthalpies ($\Delta H^{\ddagger}$). Another way to validate the reaction path computed with CPR is to compare the experimental energy difference (ΔE) between the product and the reactant (derived from calorimetric and kinetic studies) with the calculated ones. These two quantities ($\Delta H^{\ddagger}$ and ΔE) represent reliable possibilities to directly relate the CPR simulations to experimental studies.

2.4 Analysis of simulated molecular transitions

Proteins show evidence of moving between multiple conformations at the room temperature. These conformational substates can exist on many levels, from large domain or hinge motions to small rearrangements of side chains[14]. Using computational techniques we can begin to characterize the nature of these substates.

Previously, a few computational methods used to produce the trajectories analyzed in this thesis were presented. Analysis of molecular conformation spaces, over which the potential energy surfaces is defined, are used for locating stable structures in order to analyze molecular flexibility. For small systems, with only few minima, it is possible to use a direct approach and describe most of the PES, while for systems with many degrees of freedom (a molecule with N atoms has 3N degrees of freedom, and its corresponding conformation space is 3N-6 dimensional) and a very large number of minima, a direct approach to the PES becomes very difficult[15].

The conformation of a protein refers to the high dimensional arrangement of its constituent atoms. Since the expression of the biological activity of a protein depends on its conformation, it is clear that full characterization of a protein involves an understanding of its high dimensional structure. Therefore, a related question is that of how can we visualize a multidimensional PES. In general, creating useful representation of molecular conformation spaces is complicated by the high

dimensionality of these spaces. As a result, even relatively small molecules have very large conformation spaces. Thus, methods that project multidimensional data on low-dimensional subspaces are very suitable for representing and visualizing trajectories that traverse these spaces. In the following Sections techniques used for dimension reduction are presented, namely Principal Components Analysis (Section 2.4.1) and Sammon Mapping (Section 2.4.3), followed in Section 2.4.2, 2.4.4 and 2.4.5 by methods like Involvement Coefficients, DynDom or Rigidity analysis, used to further analyze the projected motions describing large conformational changes of different biological systems.

2.4.1 Principal Components Analysis

Principal component analysis (PCA) or essential dynamics is a computational tool that reduces the effective dimensionality of molecular conformational spaces while retaining an accurate representation of the interconformational distances. This task is accomplished by projecting the original multidimensional data onto an optimal low-dimensional subspace, allowing visual inspection of conformational spaces and of dynamic trajectories that traverse these spaces. A simplified representation of the way PCA determines the most important directions of motion within a data set can be seen in Fig. 2.5.

The PCA method allows identification of the essential degrees of freedom in a protein from an MD trajectory. These essential degrees of freedom are large concerted atomic motions. First, all transitional and rotational motion present in a protein MD trajectory are removed. Subsequently, a covariance matrix is constructed from the resulting trajectory. Normally for big proteins, C_α atoms trajectory is used to construct the covariance matrix. C_α atoms have been shown to contain all the information for a reasonable description of the protein large concerted motions[16]. Then the average is taken over the whole trajectory. Upon diagonalization of the covariance matrix, a set of eigenvectors/eigenvalues is obtained. These eigenvectors represent a direction in a multidimensional space along which a concerted motion of atoms takes place. The amplitude of each motion is indicated by the corresponding eigenvalue. The central

hypothesis of the essential dynamics method is that *only the motions along the eigenvectors with large eigenvalues are important for describing the functionally significant motions in the protein*. These eigenvectors span a plane in the multidimensional space in which most of the motion takes place. The motion along any desired eigenvector can be inspected by projecting all the frames from the MD trajectory onto the specified eigenvector. A new trajectory is generated which, upon visual inspection, reveals large concerted motions of atoms.

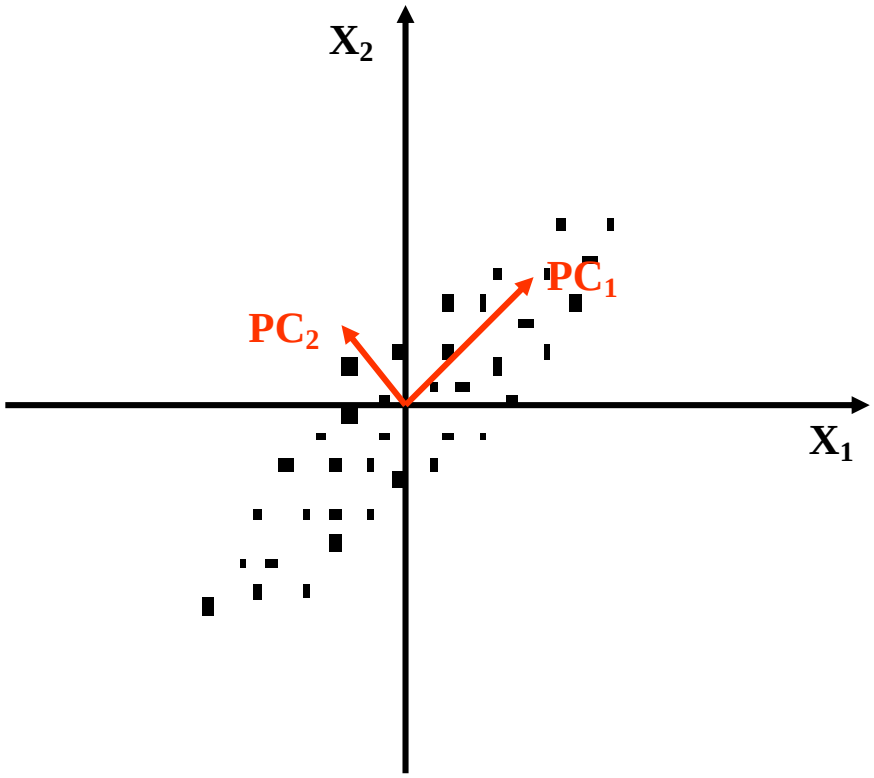


Fig. 2.5 Schematic representation of a 2D space showing the way PCA determines the important axes of motion (colored red) within a data set.

If the data is concentrated in a linear subspace, this provides a way to compress data without loosing much information and simplifying the representation. By picking the eigenvectors having the largest eigenvalues we lose as little information as possible. One can choose a fixed number of eigenvectors and their respective eigenvalues and get a consistent representation of the data. We are faced with

contradictory goals: on one hand, we should simplify the problem by reducing the dimension of the representation, on the other hand, we want to preserve as much as possible of the original information content. PCA offers a convenient way to control the trade-off between loosing the information and simplifying the problem at hand.

However, in the multivariate case when the number of variables is by far grater then 3, a plot of the cases is not possible. Using PCA, the variables can be reduced (to a number of 2 or 3, for example), so that a visual inspection is possible.

Let the variable matrix be given by the matrix X_{ij}, where i=1,n (n being the number of frames) and j=1,N (N being the number of atoms). Now we have to build the variance-covariance matrix (C_{ij}) among all N variables starting from the data set matrix (X_{ij}).

$$C_{ij} = cov\ (x_i, x_j) \tag{2.14}$$

The variance of a component indicates the spread of the component values around its mean value. The covariance between two variables, let's say x and y, is given by the equation:

$$cov\ (x,y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})\left(y_k - \bar{y}\right) \tag{2.15}$$

$$var\ (x) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 \tag{2.16}$$

where $\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$ *and* $\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$ represents the mean of one atom (x, or y) over all the frames.

The matrix is called variance-covariance matrix because on the diagonal we'll have the variance of each variable ($i = j$ so $x = y \Rightarrow cov\ (x, x) = var\ (x)$). It is evident that the variance-covariance matrix is symmetrical, due to the fact that both i and j goes from 1 to N.

Sometime principal components (PC's) are defined using the ***correlation*** matrix instead of the ***covariance*** matrix. The correlation between a pair of variables is equivalent to the covariance divided by the product of the standard deviation of the two variables:

$$\text{cor} = \frac{\text{cov}}{\sqrt{\text{var}}} = \frac{\text{cov}}{st.dev.} = cor \quad coef. \tag{2.17}$$

PCA based on covariance matrix has the potential drawback that PC's are highly sensitive to their values. If there are large differences between the variances of the variables, then the first few PC's computed with the covariance matrix are dominated by the variables with large variances. The correlation coefficient can take values between –1 and 1 so we can have anti correlation when cor=-1, no correlation when cor=0 or perfect correlation when cor=1 (see Fig. 2.6).

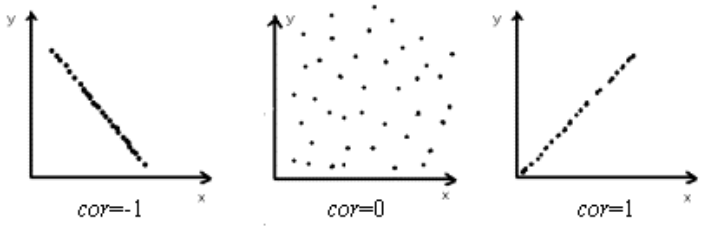


Fig. 2.6 Data arrangement corresponding to different correlation coefficients.

The eigenvalues and the eigenvectors of the variance-covariance matrix are determined by solving the equation:

$$C * a_i = \lambda_i * a_i \tag{2.18}$$

where i goes from 1 to N, the λ_i are the eigenvalues and a_i are the corresponding eigenvectors (a_i is a vector of length N). We can rewrite the equation 2.18 like:

$$(C - \lambda * I) * a_i = 0 \tag{2.19}$$

and because the values of a_i witch are zero do not present any interest, we have to calculate the determinant of a_i's coefficient:

$$\det (C - \lambda * I) = 0 \tag{2.20}$$

where I is the identity matrix (same order with C).

We can find at most N different values for λ (eigenvalues). These eigenvalues are sorted in such way that $\lambda_1 > \lambda_2 > ..... > \lambda_N$. Once the eigenvalues are sorted, we have

to determine the eigenvector (a_i) corresponding to each eigenvalue (λ). For that we have to solve the undetermined equation:

$$(C - \lambda * I) * a_i = 0 \quad (2.21)$$

and we'll have all N vectors correspondent to all N values of λ. The eigenvectors are determined by their corresponding eigenvalues only. The resulting eigenvalues give the projection of the original distribution on the new coordinate set, and the eigenvectors give the new coordinates of the original points in the new axes frame. So we'll have a matrix A, containing all determined eigenvectors, called principal component's matrix:

$$A = [\, a_1 \;\; a_2 \;\; \ldots\ldots \; a_N \,] \quad (2.22)$$

The main motivation for using PCA is to construct a low-dimensional representation of the original high-dimensional data. The idea behind this approach is that effective dimensionality of the studied systems is significantly smaller than their full dimensionality. In principal component projection the resulting eigenvalues represent the variation of the original distribution along the principal directions. When the eigenvalues and corresponding eigenvectors are sorted in decreasing order the first eigenvector represents the axis of maximal variance, the second is the axis with the second largest variance, and so forth. Projection of the distribution onto the first two or three dimensions represents the best possible planar respectively 3D projections of the distribution[17].

How does A retain the information contained in the initial matrix X?

$$Y = A\left(x_i - \overline{x_i}\right) \quad (2.23)$$

we reconstruct the original data vector X from Y by:

$$X = A^T * Y + \overline{X}_i \quad (2.24)$$

using the property of an orthogonal matrix $A^{-1} = A^T$. The original vector X vas projected on the coordinate axes defined by the orthogonal basis. The original vector was reconstructed by a linear combination of the orthogonal basis vectors. Instead of using all the eigenvectors of the covariance matrix, we may represent the data in terms of only a few basis vectors of the orthogonal basis.

How to choose the number of PC's to be visualized? The considerations are two fold: how well the PCA has condensed the variance into the small number of principal components, and how many dimensions we can handle in order to visualize our data.

The central idea in PCA is to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set. Principal components are linear transformations of the original set of variables, are uncorrelated and ordered so that the first few contain most of the variations in the original data set.

Why PCA would reduce the dimensionality? This is because most of the raw data sets are not orthogonal. Thus, we could derive a reduced number, but orthogonal (un-related), base components by linear transformation, and without introducing significant error into the visualization[18].

PCA help us to retain the strength of current multidimensional and multivariate visualization tools by significantly reducing the dimensionality. It produces better visualization because it rotates the data set to a reduced number of orthogonal principal component axes, where the maximum of the variance is visible.

2.4.2 Involvement Coefficients

Another method of particular utility is the calculation of the Involvement Coefficients, which are showing the degree of importance given by each mode of motion (obtained for example by performing PCA) with respect to a specific conformational change of the protein[19]. If X_1 and X_2 denote two distinct conformers of a protein, a unit vector **d** can be constructed:

$$d \equiv \frac{(X_1 - X_2)}{|X_1 - X_2|} \tag{2.25}$$

defining the direction of the conformational transition. The projection of the eigenvectors obtained from the PCA analysis in the direction of the transition is the inner product of the eigenvector (the selected PCA mode) with the vector **d**. We have:

$$I_k = |A_k d| \tag{2.26}$$

where I is a 3N dimensional vector with it's k^{th} component defined as the *involvement coefficient* of the k^{th} PCA mode. A schematic representation of such a projection onto the displacement vector is seen in Fig. 2.7. The components of I are the absolute values of the inner product, and their values are in the range of [0, 1]. The involvement coefficient is +1 if a particular eigenvector has a vibrational pattern that is parallel with the conformational transformation and it is 0 if it is orthogonal to the conformational difference[20]. Larger the value of the involvement coefficient indicates the fact that this specific mode is highly relevant to the conformational transition being examined[21]. A related quantity that indicates the weight of a set of modes is the cumulative involvement coefficient expressed by:

$$C_k = \sum_{k=1}^{n} I_k^2 \tag{2.27}$$

where I_k is the k^{th} involvement coefficient. If all the eigenvectors are taken in consideration C_k is expected to be 1.

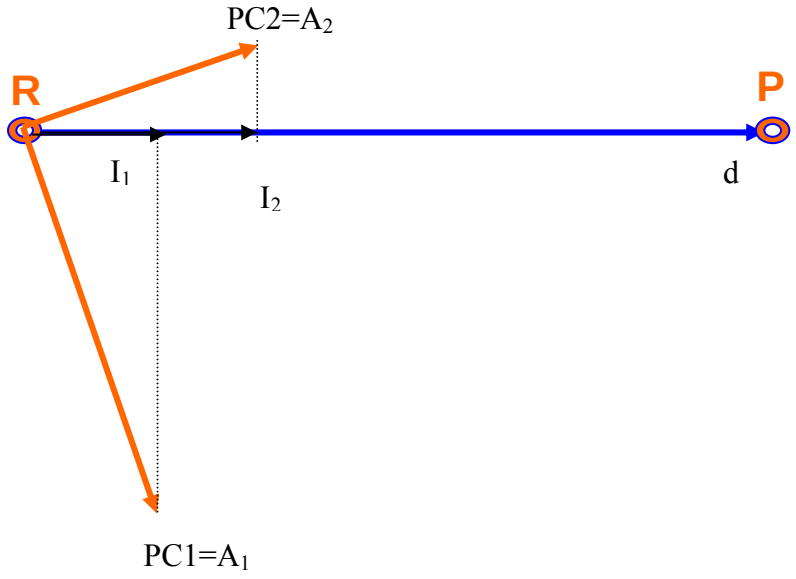


Fig. 2.7 Schematic representation of the projection of the first two eigenvectors (PC1 and PC2) obtained from PCA analysis onto the displacement vector (**d**) calculated between the reactant and the product states of a certain conformational change.

2.4.3 Sammon Mapping

Sammon's mapping method is a nonlinear alternative to the widely used PCA (being a linear projection method). The problem of nonlinear projection can bee stated as follows. Suppose we have a numerical data set of n vectors embedded in an N-dimensional space. Such database may originate from a trajectory of N frames. If there are dependencies between the features of the vectors then data are not randomly distributed and the database is said to have a structure. For example, a plane in the 3-dimensional space is a 2-dimensional space structure, if there is one dependency linking the 3 coordinates *x, y,* and *z*.

Very often, the dimension of a structure is lower than the dimension of its embedding N-dimensional space, so the structure can be projected to a lower-dimensional space. A well known projection method is the previously mentioned Principal Component Analysis which works pretty well, but only when the dependencies are strictly linear. Once you have nonlinear dependencies, another method is needed, in order to find a nonlinear mapping between the N-dimensional embedding space and the projection space.

Nonlinear mapping algorithms employ nonlinear transformations, which attempt to preserve the inherent structure of the data when the patterns are projected from a higher-dimensional space onto a lower-dimensional space. The preservation of this inherent structure is achieved by preserving the distances between patterns under projection (schematically presented in Fig. 2.8).

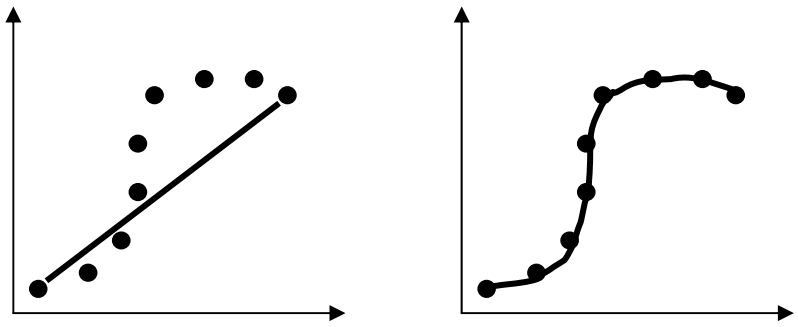


Fig. 2.8 A schematic representation of a linear (left panel) and a nonlinear (right panel) projection.

Molecular similarity is one of the most ubiquitous concepts in chemistry. It is used to analyze and categorize chemical phenomena, rationalize the behavior and function of molecules, and design new chemical entities with improved physical, chemical, and biological properties. Molecular similarity is typically quantified in the form of a numerical index derived either through direct observation, or through the measurement of a set of characteristic features, which are subsequently combined in some form of dissimilarity or distance measure. For large collections of compounds, similarities are usually described in the form of a symmetric matrix that contains all the pairwise relationships between the molecules in the collection. Unfortunately, pairwise similarity matrices do not lend themselves for numerical processing and visual inspection. A common solution to this problem is to embed the objects into a low-dimensional Euclidean space in a way that preserves the original pairwise proximities as faithfully as possible. This approach, known as multidimensional scaling or nonlinear mapping, converts the data points into a set of real-valued vectors that can be used for a variety of pattern recognition and classification tasks.

Description of the method:

Sammon's Mapping provides a mapping from a high-dimensional vector space onto a 2-dimensional output space.[22] Unless there is a high degree of redundancy in the coordinates of the data, we can never hope to achieve a perfect projection without suffering a degree of distortion in the projected representation. The basic idea is to arrange all the data points on a 2-dimensional plane in such way, that the distances between the data points in this output plane resemble the distances in vector space as defined by some metric as faithfully as possible.

A researcher is often interested in mappings that reveal the inherent structure in order to explore the data, to find possible clusters, correlations or underlying distributions[23].

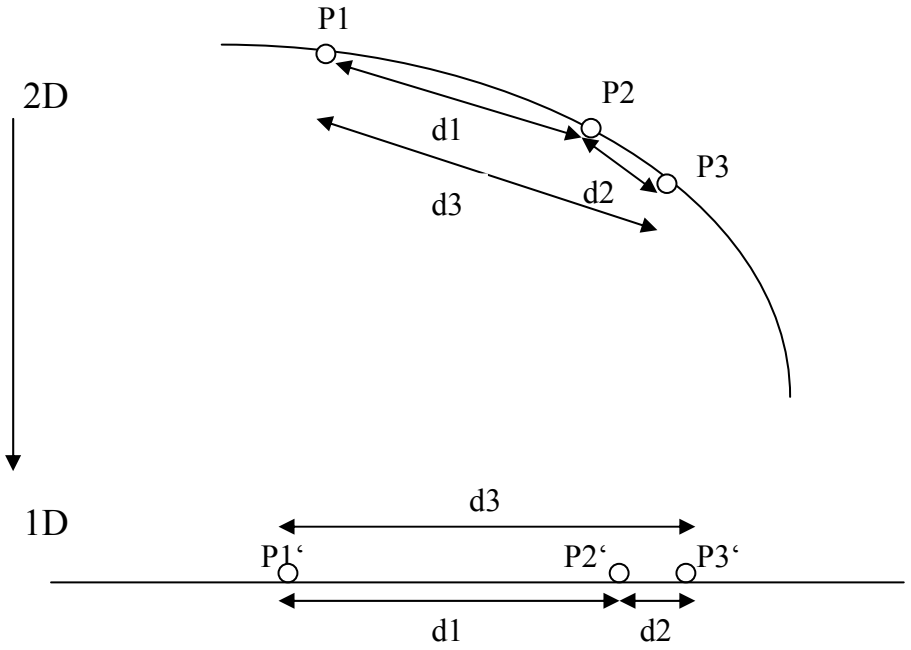
Sammon mapping comes from the area of multidimensional scaling and is an important pattern recognition tool, which reveals the structure present in a data set.

Sammon mapping has two disadvantages. Firstly, it lacks generalization, which means that new points cannot be added to the obtained map without recalculating it. Secondly, because it operates on all interpoint distances, the complexity of finding the map is very high. The solution to the first problem is used in improving the speed of the algorithm.

Understanding data is often a difficult task, especially when it refers to a complex phenomenon that is described by many variables. If the data consists of one or two variables, many simple methods are available, showing or emphasizing some of the properties or relations between objects. But, when multivariate data is examined, it is nearly impossible to comprehend its structure. This presents a need for more sophisticated techniques. The early stage of data analysis is to visualize data on a plane or in 3D space. By this, one hopes to gain some intuition about the data and to understand the relation between objects, see the intrinsic structure or possible cluster tendencies, etc. In the area of *projection methods*, nonlinear projection techniques play an important role. For a nonlinear manifold imposing e.g. the preservation of all (or some) interpoint distances in the mapping is taken. Such techniques are powerful tools for data visualization and exploration, but as iterative processes, they are time consuming. Sammon mapping is a technique of high complexity. The idea behind Sammon's map is *topology preservation*. One way to achieve this goal consists in preserving distances between samples of the numerical database:

- Two samples that are close to each other have to stay close when projected
- Two projected samples that are close to each other have to originate from two samples that were close two each other

Sammon mapping attempts to preserve as much of the original data structure as possible by positioning points in the lower-dimensional space such that the distance between any two points is a close approximation to the distance between the two corresponding points in the higher-dimensional space. This is more easily seen in the diagram below.



In this example, the points P1, P2, P3 along the two dimensional curve are to be projected onto the horizontal line. As can be seen, the aim is to locate the points P1', P2' and P3' such that the error between the interpoint distances ($d1^*$, $d2^*$, $d3^*$) on the 2D curve and their counterparts along the straight line (*d1*, *d2*, *d3* respectively) is a minimum. The process for achieving this projection from a high-dimensional space (H-D) to a lower-dimensional representation (L-D) is as follows:

1. Calculate all interpoint distances in H-D space. Will be n(n-1)/2 of them, where n is the number of points to project.
2. Generate n random points in L-D space. Let's call each projected point y_i
3. Calculate the mapping error over all the interpoint distances in L-D space. We shall call this error E.
4. If E is less than a pre-defined threshold, or the number of iterations through this loop exceeds some arbitrary count, then stop.
5. Adjust the coordinates of the points in L-D space using a function of the form $y'_i = y_i - f(E, y_i)$
6. Go to step 3

Without loss of generality, only projections onto a 2-dimensional space are studied, since our interest is in data visualization. There is a need for a criterion to decide

whether one configuration is better than another. For that purpose, the error (stress) function E is considered, which measures the difference between the present configuration of n points in the new low-dimensional space (d_{ij}) and the configuration of n points in the original high-dimensional space (d_{ij}^{*}). The problem of finding the right configuration in a low-dimensional space is an optimization problem: we are interested in obtaining such a configuration where the stress function yields the minimum. In general, this optimization problem is difficult because of the very high dimensionality of the parameter space. The stress function is optimal when all the original distances are equal to the distances of the projected points. However, this is not likely to happen exactly. Therefore, the found distances will be distorted representations of the relations within the data. Larger the stress value, grater the distortion. The stress is given by the following formula:

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^{*}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\left(d_{ij}^{*} - d_{ij}\right)^2}{d_{ij}^{*}} \tag{2.28}$$

and yields in fact a badness-of-fit measure for the entire representation.

The Euclidean distance is frequently used. Sammon's stress is a measure of how well the interpattern distances are preserved when the patterns are projected from a high-dimensional space to a lower-dimensional space. The minimum of Sammon's stress is achieved by carrying out a steepest-descent procedure. As in steepest-descent based approaches, local minima are often unavoidable. This implies that a repetitive number of experiments with different random initializations have to be performed before the initialization with the lowest stress is obtained. Sammon mapping technique approximates local geometric relationships of vectorial samples in a two–dimensional plot. In particular, given a finite set of m-dimensional samples $\{x_i, i=1,2,...,n; x_i$ is an element of $\Re^m\}$, where m is equal with 3 times the number of atoms, and n is the number of frames in a trajectory, a distance function $d_{ij}^{*} = d^{*}(x_i,x_j)$ between x_i and x_j, calculated as Euclidean distance would be expressed by:

$$d_{ij}^{*} = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2} \tag{2.29}$$

The objective is to replace d_{ij} onto the plane in such a way that their Euclidean distances approximate as closely as possible the corresponding original values d_{ij}^*. This projection, which can only be made approximately, is carried out in an iterative fashion by minimizing the error function, *E(t)*, which measures the difference between the distance matrices of the original and projected vector sets:

$$E(t) = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^*} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\left(d_{ij}^* - d_{ij}(t)\right)^2}{d_{ij}^*} \tag{2.30}$$

where t is the iteration number. The initial coordinates, d_i, are determined at random, and are updated using equation 2.31:

$$r(t+1) = r(t) - \lambda\Delta(t) \tag{2.31}$$

where λ is the learning rate parameter set at 0.3, and

$$\Delta(t) = \frac{\dfrac{\partial E(t)}{\partial r(t)}}{\left|\dfrac{\partial^2 E(t)}{\partial r^2(t)}\right|} \tag{2.32}$$

The partial derivatives in equation 2.32 are given by:

$$\frac{\partial E(t)}{\partial r(t)} = -2\frac{\sum_{j=1, j\neq i}^{k} \dfrac{d_{ij}^* - d_{ij}}{d_{ij}^* d_{ij}}(r_{ik} - r_{jk})}{\sum_{i<j}^{k} d_{ij}^*} \tag{2.33}$$

$$\frac{\partial^2 E(t)}{\partial r^2(t)} = -2\frac{\sum_{i<j}^{k} \dfrac{1}{d_{ij}^* d_{ij}}\left[\left(d_{ij}^* - d_{ij}\right) - \dfrac{\left(r_{ik} - r_{jk}\right)^2}{d_{ij}}\left(1 + \dfrac{\left(d_{ij}^* - d_{ij}\right)}{d_{ij}}\right)\right]}{\sum_{i<j}^{k} d_{ij}^*} \tag{2.34}$$

So, the nonlinear mapping is obtained by repeatedly selecting two points at random and updating their projected coordinates using equation 2.31 and 2.32. And to find a projected map, we start from the initial configuration of points (e.g. randomly

chosen), and then the stress is calculated by using equation 2.30. Next, the configuration is improved by shifting around all the points in small steps to approximate better and better the original distances (thus decreasing the stress). This process is reiterated, until the map corresponding to a (local) minimum of stress is found.

The successful outcome of using neural computing techniques for any problem depends on many factors. However, one important criterion is having a good level of understanding of the data and its origin. Often, it is found that the data represents a high-dimensional view of a process[24]. Ideally would be to use a visualization technique that can display the whole data in just two-dimensions, while still preserving its underlying structure in a meaningful form, so the role of Sammon mapping is considered.

2.4.4 DynDom

The protein architecture is based on different structure levels (i.e., primary, secondary, tertiary and quaternary) that give rise to a very variable dynamics with a wide range of time scales from protein to protein. Large proteins are generally composed of many domains whose movements are likely to be the slowest of all motions in the proteins. These motions are often found to be functionally significant[25]. The ability of different regions in a protein to move relative to each other with only a small expenditure of energy is defined as the protein *intrinsic flexibility*[14]. The two types of motion associated with intrinsic flexibility are governed by the internal packing of the interfaces between two regions in a protein. One is a *hinge mechanism* that occurs when there is no continuously maintained interface constraining the motion and the other one is a *shear mechanism* that occurs when two interfaces slide across each other in order to maintain a well-packed interface.

A key concept in the study of protein structure is the domain. If one has two conformations of a protein, or can furnish more than one conformation by doing simulation, then one has the possibility to define domains on the basis of groups of

residues moving in a concerted fashion. Such a dynamic definition of a domain is useful due to the fact that the function of many proteins can be related to the concerted motion of its parts[26].

Due to the fact that domain motions play an important role in the function of proteins, prof. Berendsen and coworkers have become interested in developing methods to analyze the "important" motions inherent to molecular trajectories. One of these methods is DynDom, a program useful when someone wants to analyze the domain motion in proteins. DynDom can be used when two conformations of the same protein are available. Given two conformations of a protein, the program will analyze the conformational change in terms of dynamics domains, hinge axis, and hinge-bending regions[26].

The underlying idea behind the method is that domain can be identified by their differing rotational properties. The conformational change may be quite complicated in detail, butit can be visualized as involving the movement of domains as quasi-rigid bodies. DynDom allows you to visualize the domain motion in terms of the rotation of one domain relative to another.

Dynamical domains are defined based on two key criteria: first that the residues within a domain must rotate around an axis with a similar orientation and second that the "amount of motion" (measured by the mean square fluctuation normalized to the size of the domain) within a domain must be smaller than the "amount of motion" between domains, i.e. domains are more rigid than the whole.

The only ingredients necessary for application of the method are a structure plus a set of displacement vectors showing how each atom or subset of atoms moves along the conformational change. Such a set of displacement vectors is given directly from a normal mode analysis, or can be determined from a molecular dynamics simulation or from a pair of X-ray conformers after performing a least-square best fit. The program DynDom determines dynamic domains, interdomain screw axes, and regions involved in the interdomain bending.

Determination of dynamic domains: After a whole protein best fit of the two conformations is made, the rotation vectors of residues are determined by calculating the curl of the displacement vector field for each residue. Each rotation vector can be

represented by a “rotation point” in the three-dimensional rotation space, whose coordinates form the rotation vector. A K-means clustering algorithm is then used to identify clusters of rotation vectors. Groups of residues forming these clusters form possible dynamic domains.

Determination of hinge axes: The next step is to determine the hinge axes and for that groups of residues are accepted for the analysis of hinge axes if they satisfy a criterion based on the ratio of the interdomain displacement to intradomain displacement with another group of residues with which there is a physical connection. If this is the case the two group of residues form dynamic domains and their interdomain motion is meaningful. The axes determined are in fact interdomain screw axes. This is based on the theorem of Chasles, which states that the general displacement of a rigid body is a screw motion. The location of the interdomain screw axis tells us something about the kind of motion allowed by the interdomain connections. It is possible for the interdomain screw axis to be located far away from the interdomain connections if they are very flexible. Only if the interdomain screw axis is located near to those residues involved in the interdomain bending can we think of the axis as a hinge axis. In such a case the axis is called "effective hinge axis" and the residues are said to be acting as "mechanical hinges”.

Determination of residues involved in interdomain bending*:* The last step of the algorithm is the determination of residues involved in interdomain bending. If one domain is fixed in space with the other rotating, then one will see a rotational transition in the connecting region between the two domains. One can define the residues involved in the interdomain bending to be those at the interdomain boundaries, as found by the clustering algorithm, plus those neighboring residues whose rotations deviate at least one standard deviation from the average of the domain to which they belong. So these regions of the backbone where a rotational transition is seen between the rotational properties of the two dynamic domains, are called “bending regions”.

Description of the needed parameters*:* For analyzing a protein some parameters need to be defined and for that their significance is needed:

"clusters" - sets the maximum number of clusters of rotation vectors the program looks for. This can be set high because only then we can be sure that we find all the possible domains.
"iterations" - sets the maximum number of iterations the clustering algorithm is allowed to determine the clusters. It normally requires far less than 100 iterations.
"window" - a sliding window is used to generate backbone segments whose rotation vectors are calculated for the clustering algorithm. The longer the window the better local intrasegment rotations, which may have nothing to do with the domain motion, are eliminated. A shorter window, however, is preferable if one wants to accurately identify those residues involved in the interdomain bending. The length of the window should be short compared to the minimum domain size. The number of residues specifies the length of the window. The length of the window must be odd, as the central residue must be defined.
"domain" - sets the minimum domain size in number of residues. It depends on how big one considers that a domain should be, to be called a domain. Should not be smaller than the window length.
"ratio" - gives the minimum value for the ratio of interdomain displacement to intradomain displacement for a domain pair. There is no clear cutoff for this value, but the lower it is for a domain pair, the less sensible it is to analyze their motion in terms of an interdomain motion. If one set the minimum to a value much less than 1.0 we may end up analyzing meaningless results.

2.4.5 Rigidity analysis

Protein motions, ranging from molecular flexibility to large-scale conformational changes, play an essential role in many biochemical processes. For example, local conformational change often occurs in binding interactions between proteins and between proteins and ligands, sugars, and other small molecules. While no consensus has been reached regarding models for protein binding, the importance of protein flexibility in the process is well established by the ample evidence that the

same protein can exist in multiple conformational states and can bind to structurally different molecules. Our understanding of molecular movement is still very limited and has not kept pace with the explosion of knowledge regarding protein structure and function.

Several computational approaches have studied rigidity and flexibility in proteins. One approach infers the protein's flexibility/rigidity by comparing different known conformations of the protein[27]. Molecular dynamics has been used to extract flexibility information from simulated motion[28-30]. A third method studies rigidity and flexibility of a single conformation[31-33].

Rigidity matrix formation:

From a MD trajectory of a protein a matrix $\mathbf{C} \in [0,1]^{N \times N}$ is computed, where N is the number of atoms used for this analysis. For all analysis made in this thesis only C_α-atoms were used meaning that in this case N is identical to the number of protein residues. **C** is a symmetric rigidity matrix, defined such that a value $C_{ij}=0$ means that atoms i and j are moving uncorrelated while when $C_{ij}=1$ means that the atoms i and j are moving together and their Euclidean distance stays perfectly constant. Given the pair-wise atomic distances in every time step, $d_{ij}(t)$, the matrix elements, C_{ij}, are obtained as follows:

$$C_{ij} = \frac{1 - \min\{\sigma(d_{ij}), \sigma_{cut}\}}{\sigma_{cut}} \tag{2.35}$$

$$\sigma(d_{ij}) = \sqrt{\left\langle d_{ij}(t) \right\rangle^2 - \left\langle (d_{ij}(t))^2 \right\rangle} \tag{2.36}$$

where $\sigma(d_{ij})$ is the standard deviation of $d_{ij}(t)$ and σ_{cut} is a user-defined cutoff – all deviations larger than σ_{cut} are mapped to 0 in matrix **C**. Obviously, the choice of σ_{cut} will affect the results of the clustering. There is no "right" or "wrong" setting for σ_{cut}, but it may rather be viewed as a parameter determining the "resolution" of the rigidity analysis. Increasing the value of σ_{cut} also leads to increasingly large domains due to the fact that the intra-domain motion is less penalized.

Identification of rigid domains:

C can be used to identify rigid clusters of atoms by finding a group of atoms $G_i=(g_1 \ldots g_{Ni})$ such that each atom within the group has a large pair-wise rigidity with all other atoms in the group, but a small pair-wise rigidity with all atoms outside the group. Formally, we attempt to identify a set of groups, or clusters, $G=(G_1,\ldots,G_i,\ldots,G_M)$, each containing a set of atom indexes $G_i=(g_{i,1},\ldots,g_{i,j},\ldots,g_{i,Ni})$. G maximizes the following target function:

$$Z = \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} z_{ij} \tag{2.37}$$

where $$z_{ij} = \begin{cases} C_{ij}, \text{if i and j are in the same cluster} \\ 1 - C_{ij}, \text{otherwise.} \end{cases}$$

This maximization problem is NP-hard[34]. This means that the computational time required finding the global maximum of Z increases exponentially with the size of **C**. For proteins, **C** is quite large and it is therefore number unrealistic to attempt finding the global maximum. However, this is not a strong restriction in the current work. Firstly this is because there are many good solutions, which are acceptable. For example some atoms do not clearly belong to any single cluster but are rather part of interface regions whose atoms have a high pair-wise rigidity with several clusters. To obtain a qualitative, overall picture of the rigid bodies in the protein it does not matter much, which clusters these atoms are assigned to, even though different assignments may lead to slightly different values of the target function. Secondly, the problem is benign because the target function, *Z*, clearly distinguishes "acceptable" and "unacceptable" partitions. Consider two rigid clusters of atoms, 1 and 2, moving as rigid bodies each. Clearly any partition which approximately identifies the two rigid bodies is much better than any partition which has atoms from both rigid clusters in a single group. This becomes clear from the way the target function *Z* is defined. Moving an atom which belongs to cluster 1 from group 1 into group 2 will tremendously reduce the target function value as all the z_{ij} between that atom and

atoms in group 2 are low and the z_{ij} between that atom and atoms in group 1 are high. Although it is very hard to identify the global maximum of Z it is easy to identify a good local maximum of Z. Here, Z is maximized using following stochastic optimization method:

1. Start with $|G|=N$ clusters, each containing a single atom. Compute $Z_{old}=Z(G)$.
2. Consider at random one the following operations:
    a. *Split*: Determine a random cluster, G_i, with at least two atoms. Split this cluster into two new clusters, G_j and G_k, assigning each member of G_i to a random one of these two. Remove old cluster G_i.
    $$Z_{new} = Z_{old} + \sum_{u\in G_j, v\in G_k} 1 - 2C_{uv}$$
    b. *Merge*: If there is more than one cluster in G, select two different clusters at random, G_i and G_j, and merge them into a new cluster G_k. Remove clusters G_i and G_j.
    $$Z_{new} = Z_{old} + \sum_{u\in G_j, v\in G_k} 2C_{uv} - 1$$
    c. *Jump*: If there is more than one cluster in G, select two different clusters at random, G_i and G_j, with $|G_i|>1$. Remove a random member u from G_i and add it to G_j.
    $$Z_{new} = Z_{old} + \sum_{v\in G_j \text{ except u}} 1 - 2C_{uv} + \sum_{v\in G_j} 2C_{uv} - 1$$
    d. *Swap*: If there is more than one cluster in G, select two different clusters at random, G_i and G_j, with $|G_i|>1$ and $|G_j|>1$. Determine a random member, u and v, in each cluster and swap them.
    $$Z_{new} = Z_{old} + \sum_{w\in G_j \text{ except u}} 2C_{vw} - 2C_{uw} + \sum_{w\in G_j \text{ except v}} 2C_{uw} - 2C_{vw}$$
3. If $Z_{new}<Z_{old}$, perform the selected operation, set $Z_{old}=Z_{new}$ and go to 2.

This Monte-Carlo procedure turns out to be very efficient. For the myosin case studied in this thesis ($N\approx800$), Z converges within 100000 steps, which takes a few seconds on a standard CPU to compute. Repeating the optimization various times gives very similar results.

References

1. Wales D.J. - Energy Lanscapes, *University Press*, Cambridge, 2003
2. Leach A. R. – Molecular Modeling Principles and Applications, *Pearson Education Limited*, Prentice Hall, 2001.
3. Pearlman d. A., Case D. A., Caldwell J. W., Ross W. R., Cheatham-III T. E., DeBolt S., Ferguson D., Seibel G. & Kollman P – AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules, *Comp. Phys. Commun.* **91**, 1-41, 1995.
4. van Gunsteren W. F. & Berendsen H. J. C. - Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry, *Angew. Chem. Int. Ed. Engl.* **29**, 992-1023, 1990.
5. Brooks B. R., Bruccolieri R. E., Olafson B. D., States D. J., Swaminathan s. & Martin Karplus – CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, *J. Comp, Chem.* **4**, 187-217, 1983.
6. Haile J. M. – Molecular Dynamics simulations: Elementary methods. New York: Wiley, 1992.
7. Allen M. P. & Tildesley D. J. – Computer simulations of Liquids. Oxford Science Publications, 1987.
8. Fischer S. & Karplus M. – Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom, *Chem Phys Let,* **194 (3)**, 252-261, 1992.
9. Fischer S, Michnick S & Karplus M. – A mechanism for rotamase catalysis by the FK506 binding protein (FKBP), *Biochemistry* **32**, 13830-13837, 1993.
10. Fischer S., Dunbrack R. L. Jr. & Karplus M. - Cis-Trans imide isomerizaion of the proline dipeptide, *J. Am. Chem. Soc.* **116**, 11931-11937, 1994.
11. Sopkova-de Oliveira S., Fischer S., Guilbert C., Lewit-Bentley A. & Smith J.C. - Pathway for large-scale conformational change in annexin V, *Biochemistry* **39**, 14065-14074, 2000.

12. Blondel A., Renaud J. P., Fischer S., Moras D. & Karplus M. - Retinoic acid receptor: a simulation analysis of retinoic acid binding and the resulting conformational changes, *J. Mol. Biol.* **29**, 101-115, 1999.
13. Dutzler E., Schirmer T., Karplus M. & Fischer S. – Translocation mechanism of long sugar chains across the maltoporin membrane channel, *Structure* **10,** 1273-1284, 2002.
14. Gerstein M., Lesk A. & Chothia C. – Structural mechanism for domain movements, *Biochemistry* **33**, 6739-6749, 1994.
15. Karplus M. & Elber R. - Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin, *Science* **235(4786)**, 318-321, 1987.
16. Amadei A., Linssen A. B. M. & Berendsen H. J. C. - Essential dynamics of proteins, *Proteins* **17**, 412-425, 1993.
17. Becker O. M. - Principal coordinate maps of molecular potential energy surfaces, *J. Comp. Chem*. **19 (11)**, 1255-1267, 1998.
18. Pan Z. - PCA Based Visualization and Human Melanoma Classification
19. Tama F. & Sanejouand Y. - Conformational change of proteins arising from normal mode calculations. *Prot. Eng.* **14,** 1-6, 2001.
20. Marques O. & Sanejouand Y. H. – Hinge-bending motion in citrate synthase arising from normal mode calculations, *Protiens: Struct. Funct. Genet.* **23**, 557-560, 1995.
21. Li G. & Cui Q. - Analysis of functional motions in Brownian Molecular Machines with an efficient Block Normal Mode Approach: Myosin-II and Ca^{2+} – ATPase. *Biophys. J.* **86**, 743-763, 2004.
22. Sammon J. W. – A non-linear mapping for data structure analysis, *IEEE Trans. Comp. C* **18 (5)**, 401-409, 1969.
23. Pekalska E., De Ridder D., Duin R. P. W. & Kraaijveld M. A. – Annual Conference of the Advanced School for Computing and Imaging, 221-228, 1999.
24. Nabney I. & Starr D. – *Networks* **27**, 2000.

25. Hayward S., Kitao A. & Berendsen H. J. C. – Model-Free Methods of Analyzing Domain Motions in Proteins From Simulations: A Comparison of Normal Mode Analysis and Molecular Dynamics Simulation of Lysozyme, *Proteins: Struct. Func. And Genetics* **24,** 425-437, 1997.
26. Hayward S. & Berendsen H. J. C. – Systematic Analysis of Domain Motions in Proteins From Conformational Change: New Results on Citrate Synthase and T4 Lysozyme, *Proteins: Struct. Func. And Genetics* **30**, 144-154, 1998.
27. Case. D. - Molecular dynamics and normal mode analysis of biomolecular rigidity. In M. Thorpe and P. Duxbury, editors, *Rigidity theory and applications*, 329–344. Kluwer Academic/Plenum Publishers, 1999.
28. Covell D. - Folding protein *α*-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.* **14(4),** 409–420, 1992.
29. Jacobs D. & Thorpe M. - Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.* **75(22),** 4051–4054, 1995.
30. Krivov S. V. & Karplus M. - Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys* **114(23)**, 10894–10903, 2002.
31. Mu˜noz V., Henry E. R., Hoferichter J. & Eaton W. A. - A statistical mechanical model for *β*-hairpin kinetics. *Proc. Natl. Acad. Sci. USA* **95,** 5872–5879, 1998.
32. Nauli S., Kuhlman B. & Baker D. - Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.* **8(7),** 602–605, 2001.
33. Prusiner S. - Prions. *Proc. Natl. Acad. Sci. USA* **95(23),** 13363–13383, 1998.
34. Oswald M. (IWR Heidelberg), personal communication.

Chapter 3

Analyzing Large-Scale Structural Changes in Proteins by Reducing the Dimensionality

Proteins are involved either directly or indirectly in all biological processes in living organisms. It is now generally accepted that conformational changes of proteins can affect their function and that any progress in modeling protein motion and flexibility is a step forward towards understanding of their biological role. While experimental methods produce rather limited information regarding protein flexibility and also computational methods such as MD may be quite slow for use of large systems (a medium sized protein may have up to few thousands of degrees of freedom), the necessity of reducing the high dimensional space to a much lower dimensional representation capturing the dominant motions of the protein has become imperative. This chapter presents results about how the complexity of modeling flexibility in proteins can be decreased by reducing the number of necessary dimensions in order to analyze important macromolecular motions. Two projection methods were tested (Sammon Mapping and Principal Component Analysis) on four different proteins and the obtained results are presented as they were published in *Proteins: Structure, Function and Bioinformatics*, 64, 210-218, 2006.

3.1 Abstract

Effective analysis of large-scale conformational transitions in macromolecules requires transforming them into a lower dimensional representation that captures the dominant motions. Two different dimensionality-reduction techniques are applied and compared, namely Principal Component Analysis (PCA), a linear method, and Sammon Mapping which is nonlinear. The two methods are used to analyze four different protein transition pathways of varying complexity, obtained by using either the Conjugate Peak Refinement (CPR) method or constrained Molecular Dynamics (MD). For the return-stroke in myosin, both Sammon Mapping and PCA show that the conformational change is dominated by a simple rotation of a rigid body. Also, the case of the T→R transition in hemoglobin, both methods are able to identify the two main quaternary transition events. In contrast, in the cases of the unfolding transition of SNase or the signaling switch of Ras p21, which are both more complex conformational transitions, only Sammon Mapping is able to identify the distinct phases of motion.

3.2 Introduction

An important aspect of the relation between protein structure and function is flexibility. For example, is well known that changes in protein conformation play a key role in many biological processes[1]. However, the representation of protein dynamics on a full 3N-dimensional potential energy surface (where N is the number of atoms) is computationally expensive and renders interpretation difficult. Consequently, the development is desirable of simplified representations of the configurational space explored that make the best use of the data available. A decrease in complexity can, in principle, be achieved by reducing the number of dimensions.

Among the techniques that allow visualization of high-dimensional data in a low-dimensional space are projection methods. Two complementary methods are chosen for analysis here: Principal Component Analysis (PCA) and Sammon Mapping. Both methods are commonly used to reduce an original space with thousands of variables into two dimensions. The PCA transformation represents directions of the greatest variation in the original space. The aim of Sammon Mapping is to preserve the distance structure of the original space while reducing the dimensionality. A major difference between the two methods is that PCA is a linear technique, in which, for example, points arranged along a line in the original space are still arranged along a line in the projected space, whereas Sammon Mapping is nonlinear. PCA has been widely used in the analysis of molecular dynamics trajectories of proteins[2-5]. In contrast, Sammon Mapping applications have been relatively rare. In biology, the Sammon Mapping method has hitherto been used mainly for indexing and mapping of proteins[6,7], and in gene expression analysis[8-10].

Conceptual differences between Sammon Mapping and PCA are illustrated in Fig. 3.1 in which a simplified 3D example is used for projection. Fig. 3.1A shows that if, along a certain path, several changes in direction are present, then Sammon Mapping is able to identify them in the reduced space, while PCA may fail in identifying some of them. Fig. 3.1B illustrates how, in some cases, a PCA projection can contain points that appear very close (overlapping) while in real space they are distant. Both cases (A and B) show that Sammon Mapping can, in certain situations, furnish more information than a PCA projection. This paper will thus introduce Sammon Mapping as a useful tool complementing PCA in protein dynamical analysis.

We investigate here situations in which PCA and Sammon Mapping are useful in extracting information concerning simulation-derived conformational transitions. The trajectories used were obtained with Molecular Dynamics (MD) simulations or, for the larger conformational changes, with the Conjugate Peak Refinement (CPR) method[11]. CPR finds a minimum-energy path (MEP) between two known end states[12-14]. Transitions in four proteins are investigated: myosin, hemoglobin, Staphylococcal nuclease and Ras p21. The transitions range from relatively simple displacements of rigid bodies relative to each other through to complex, involved rearrangements

containing many small, separable events. The paper examines under which circumstances PCA and/or Sammon Mapping can be usefully applied to identify important events along these types of structural change.

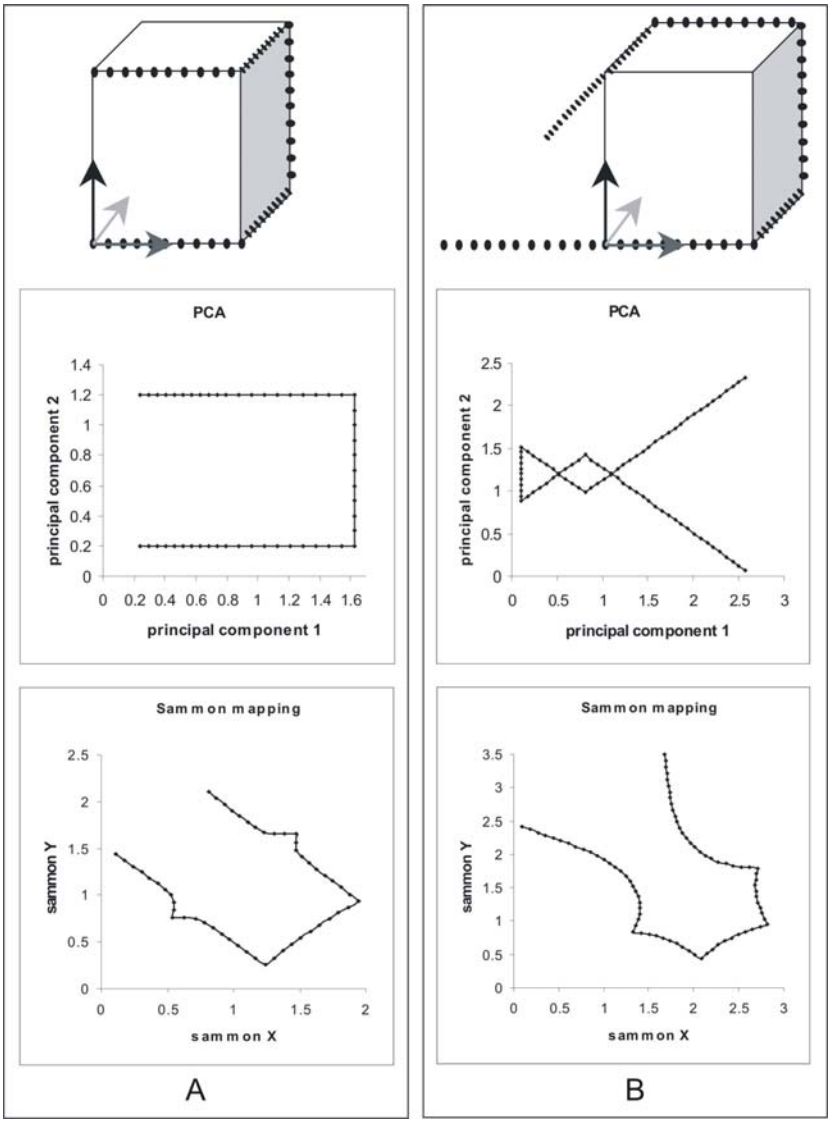


Fig. 3.1 Comparison of Sammon Mapping and PCA. Two trajectories (top) involving several changes in direction in a hypothetical 3-D conformational space (moving along the edges of a cube) are projected onto a 2D plane with the PCA (middle) and Sammon Mapping (bottom) methods. Panel **A** The PCA projection shows only two changes in direction, while Sammon Mapping identifies all four of them. Panel **B** PCA overlaps points which are not close in the original space. In contrast, Sammon mapping keeps distant points apart, at the expense of a distortion of the straight path segments.

3.3 Methods

Principal Component Analysis[15,16] is probably the most commonly used technique for dimensionality reduction in biomolecular simulation. PCA has also been referred as "quasiharmonic analysis" or "essential dynamics" and has been widely used in a variety of applications involving sampling and visualization of conformational spaces[2-5,17-25]. PCA has been also used in X-ray and NMR refinement[26-28].

PCA involves a *linear* combination of the original data points that transforms the original high-dimensional set of (possibly) correlated variables into a reduced set of uncorrelated variables – the principal components (PCs). In order to extract the internal motion from protein simulation, rotational and translational whole-molecule motions are first eliminated. The internal motion is described by a trajectory, *x(f)* where *x* is a 3N-dimensional vector (N is the number of atoms) of all atomic coordinates and *f* is the number of frames. Each step of the trajectory (each frame) is represented by a point in 3N-dimensional Cartesian space. From this the interatomic distance fluctuation covariance matrix, c_{ij} is constructed. c_{ij} is defined by:

$$c_{ij} = \left\langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \right\rangle \tag{3.1}$$

where $x_{i/j}$ are Cartesian atomic coordinates of the atom i, j and $\langle x_i \rangle$, $\langle x_j \rangle$ denote the average coordinates (taken over the trajectory). Because the size of the matrix varies with the square of the number of atoms for which the covariance is determined, in the present examples only the C_α atoms were used to construct the covariance matrix. However, it has been shown that the C_α atoms adequately sample large-scale concerted motions in proteins[3,29,30].

Upon diagonalization of the covariance matrix, a set of eigenvalues and the corresponding eigenvectors is obtained. The eigenvector represents a direction in the original 3N-dimensional space along which a concerted motion of atoms takes place. The amplitude of each motion is determined by the corresponding eigenvalue, λ_i.

$$C \cdot a_i = \lambda_i \cdot a_i \tag{3.2}$$

where a_i is the i^{th} eigenvector and λ_i is the i^{th} eigenvalue of C.

The resulting PCs are linear transformations of the original set of coordinates. They are ordered so that the first vector, $PC_{(1)}$, gives the direction of largest variation in the data set, the second vector gives the next direction of largest variation, and so on. The fraction, V_i of the total variance content in any given eigenvector, a_i is given by:

$$V_i = \frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k} \tag{3.3}$$

where n is the total number of eigenvectors.

The central interest of the PCA method is that the eigenvectors with large eigenvalues characterize correlated motions and may be important for describing functionally-significant dynamics in a protein. These eigenvectors span a plane in the multidimensional space with high variance and in which most of the motion takes place. These eigenvectors span a multidimensional sub-space with high variance and in which most of the motion takes place. A limitation consists in the fact that several types of motion (rotations, twisting or screw motion) cannot be captured in one mode, and need several eigenvectors to be represented. Usually it is hard to identify a motion spread over several modes but if the data is concentrated in a linear subspace (*e.g*, a line or a plane), this provides a way of projecting the data onto one or two directions without losing much information.

PCA can be used to separate the configurational space into two subspaces: an "essential" subspace containing only a few degrees of freedom describing the anharmonic motion that determines most of the positional fluctuations; and the remaining subspace. The essential subspace contains large-scale motions that can be described in terms of a small number of collective variables[4,20,21].

Sammon Mapping is a *nonlinear* method that provides a map of a high-dimensional vector space projected onto two dimensions[31]. In this method the data points are arranged in the two-dimensional output plane in such way that the distances between the data points in the output plane reproduce the distances in the original vector space (defined by some metric) as faithfully as possible[32].

The process for obtaining a Sammon map from a high-dimensional (HD) space in a low-dimensional (LD) representation is as follows. Firstly, all interpoint distances in the HD space are calculated (RMSD of Cartesian coordinates). Next, *n* points are randomly generated in the LD space (see Fig. 3.2). Here, different initial conditions were also tested (e.g., the *n* points were linearly generated), but this did not significantly affect the resulting map.

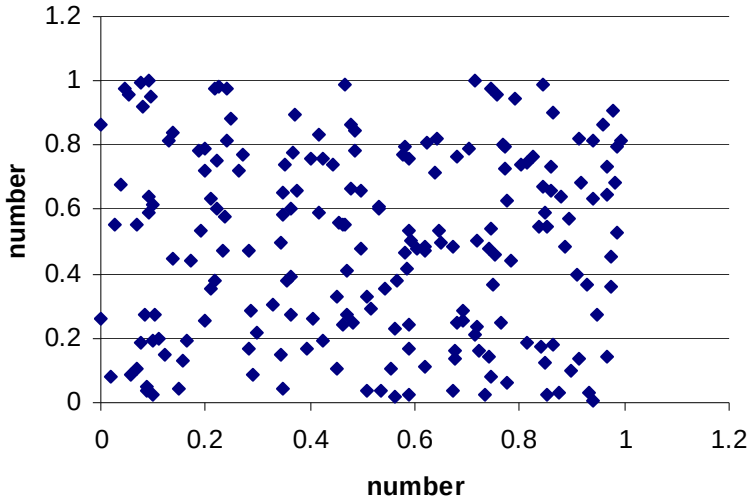


Fig. 3.2 Random generation of points in the low-dimensional space.

An error (stress) function, E is then calculated, which gives a measure of the difference between the distance matrices of the original and projected vector space:

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^{*}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\left(d_{ij}^{*} - d_{ij}\right)^2}{d_{ij}^{*}} \tag{3.4}$$

where d_{ij} is the distance between points i and j in the LD space and d_{ij}^{*} is the corresponding distance in the original HD space. Next, the configuration is changed by shifting around all the points in small steps, until the error decreases. The process is carried out in an iterative fashion using a steepest descent algorithm[6,31,33] until the map corresponding to a (local) minimum of error is found. The stress function is

optimal when all the original distances are equal to the distances of the mapped points (in which case the residual value of *E* is zero). The magnitude of the residual stress function gives an estimate of how faithfully the data can be projected onto a subspace of given dimensionality. To obtain an estimate of how many dimensions would be required to give a representation of the data to within a given error, the Sammon Mapping would have to be performed repeatedly with a varying number of dimensions.

3.4 Results

The two analysis methods (PCA and Sammon Mapping) were applied here to four systems: myosin, hemoglobin, Staphylococcal nuclease (Snase) and Ras p21, exhibiting different dynamical behavior and complexity. Molecular movies of the paths/trajectories used here can be seen or downloaded from: http://www.iwr.uni-heidelberg.de/groups/biocomp/fischer.

3.4.1 Myosin

The first protein analyzed in this study is myosin II. The structure of myosin (the S1-head) has been solved crystallographically in the two end conformations (OPEN and CLOSED) of the return stroke of muscle contraction[34,35]. Before the return stroke ATP binds to the OPEN conformer of myosin, and induces the return into the CLOSED (pre-power-stroke) conformer[36]. The largest difference between the OPEN and CLOSED structures is a rigid-body rotation of the converter domain, which carries the lever arm along and which is rotated by ~60° [37] relative to the main part of the protein. A structural pathway of the return stroke has been determined computationally using the CPR method. A detailed description of this structural mechanism of the recovery stroke is given elsewhere[14]. The motion analyzed here represents the pathway between these two states.

When a trajectory describes a large conformational change involving relative domain motion, such as in the return stroke of myosin, a PCA projection is expected to be dominated by the first eigenvector. This eigenvector gives the direction of the major movement which, for myosin, is the rotation of the lever arm. Fig. 3.3A shows that, for myosin, the first PCA eigenvector indeed contains more than 92% of the total variance of all eigenvectors and the first two eigenvectors contain ~97%. Thus, for this path, the PCA method captures the largest fluctuations using only two directions in configurational space.

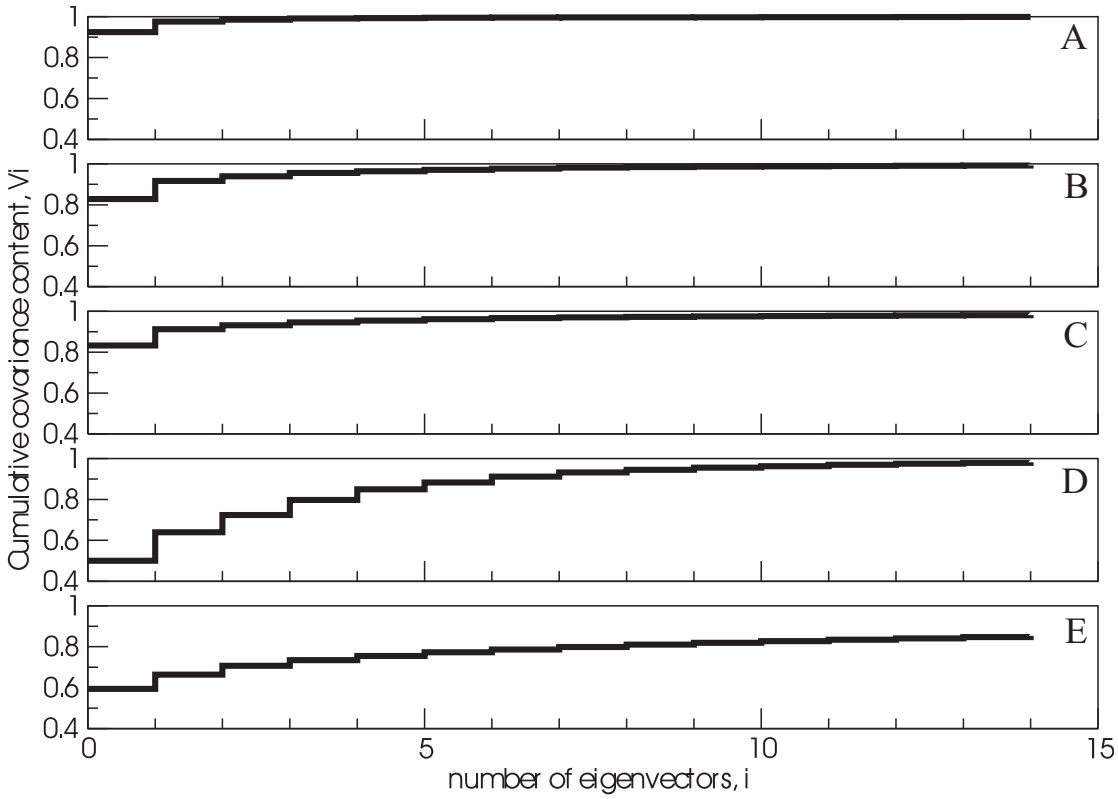


Fig. 3.3 Cumulative sum of the covariance content in the first 14 eigenvectors (V_i, equation 3.3) for the PCA of the transition in: Panel **A** Myosin recovery stroke(the first two PCs account for 97% of the total variance of all eigenvectors); Panel **B** Hemoglobin T→R (the first two PCs account for 91%); Panel **C** SNase unfolding (the first two PCs account for 91%); Panel **D** Ras P21 switch (the first two PCs account for 63%) and Panel **E** MD of SNase without constraints (the first two PCs account for 66%).

The PCA projection appears rather unidirectional and this is confirmed when the same trajectory is mapped onto the plane found with the Sammon Mapping method (see Fig. 3.4A). Another indication that the trajectory can be described by a simple motion is the fact that the value of the error function is very small (0.00771 after just 20 iterations). This means that the resulting 2D-Sammon map gives a good representation of the actual motion. Both methods indicate a simple, semi-rigid-body motion, as has been shown to be the case for the converter rotation[37]. In both plots small, localized changes can be seen superposed on the large-scale motion between λ_1 and λ_2 (see Fig. 3.4A). This myosin example shows that PCA and Sammon Mapping provide similar information when the transition involves a relatively simple type of large-scale motion.

3.4.2 Hemoglobin

The second protein analyzed in our study is hemoglobin, one of the experimentally and theoretically best-studied proteins. Hemoglobin is a tetramer consisting of two α-chains (α1 and α2) and two β-chains (β1 and β2), each chain carrying one heme group. The oxygen transport function is controlled by a conformational transition involving tertiary and quaternary structural changes. Superposition of the two end-states (deoxy = T for "tense" and oxy = R for "relaxed") shows that the T → R transition involves a large subunit rearrangement (a 15° rotation of one αβ dimer relative to the other, around a virtual axis) (see Fig. 3.4B).[38]

The trajectory used here is a CPR pathway of the hemoglobin T→R transition[12]. The movie of the path presents the large structural change as being composed of two sequential quaternary large-scale rotations (Q1 and Q2), separated by local tertiary changes (T1, T2 and T3). This occurs in a particular sequence, well identified by both PCA projection and Sammon Mapping (Fig. 3.4B). The tertiary and quaternary structural changes alternate.

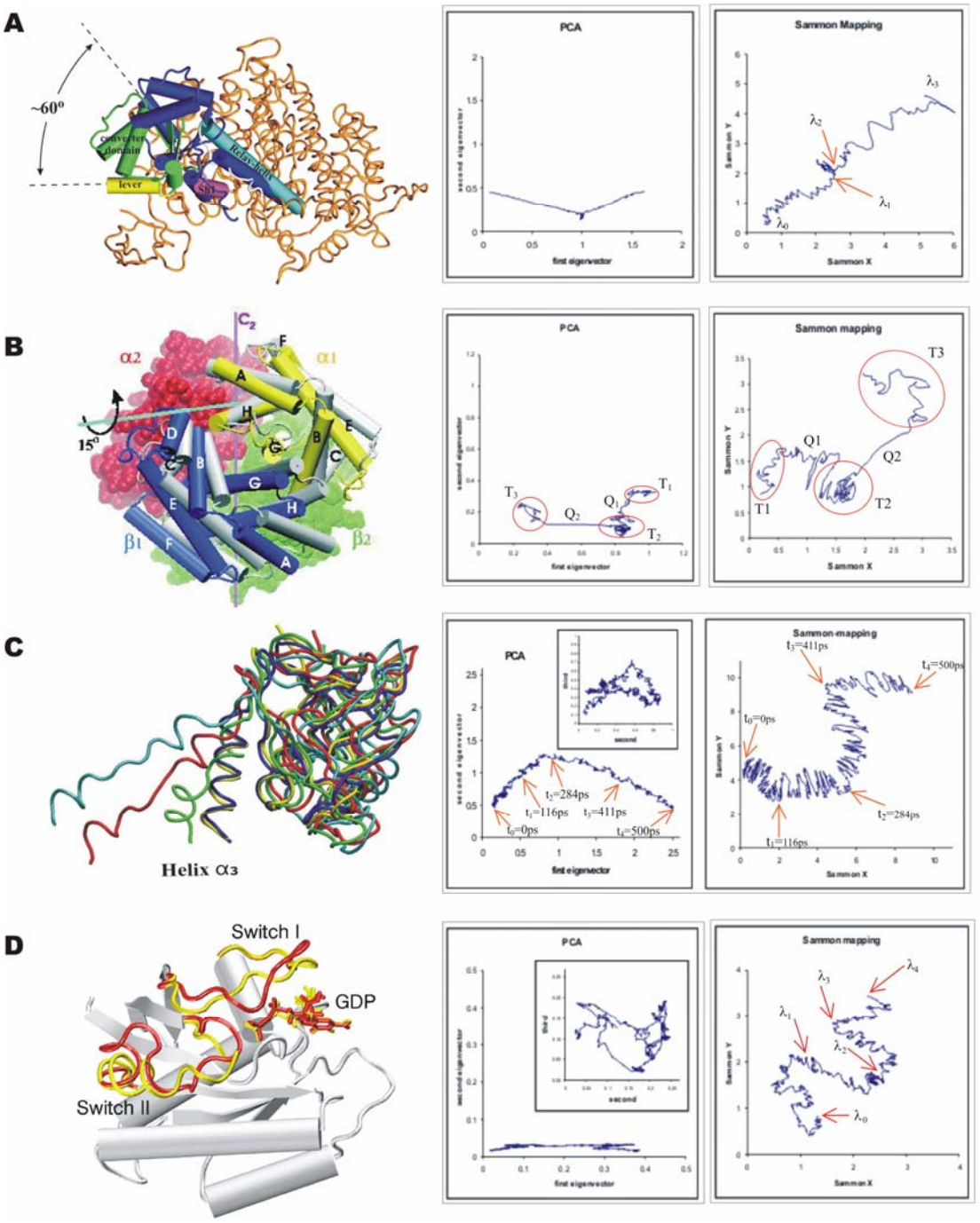


Fig. 3.4 PCA and Sammon Mapping of protein transitions. The PCA projections of the C_α atom trajectories are onto the plane defined by the two principal components with largest eigenvalues. The Sammon mapping uses all atoms for the distance metric. Panel **A**: Minimum-energy pathway of the recovery-stroke in the myosin head. Superposition of converter-domain in the end-state conformation (dark blue) onto the starting-state conformation (other colors). The main body (in orange) does not change significantly (only the starting

conformation is shown). The converter domain rotates by 60° relative to the main body. Both PCA and Sammon Mapping show the transition as following a single main direction, reflecting the fact that the motion consists mainly of the simple continuous rotation of the converter domain around a single axis. Panel **B**: Minimum energy pathway of the T→R quaternary transition in hemoglobin. The α_1 β_1 units are shown in yellow and blue for the T state and in gray for the R state, after superposing the α_2 β_2 units (red and green, shown only for the T-state). The α_1 β_1 dimer appears rotated by ~15° (the screw axis identified by DynDom is shown in cyan). The C2-symmetry axis relating the α_1 β_1 and α_2 β_2 dimers is in magenta. The overall pathway has two major quaternary events (Q1 and Q2), separating three tertiary events (T1, T2 and T3). These can be identified on both the PCA projection and on the Sammon map (the three circles show the localized T1, T2 and T3 phases). Panel **C**: Constrained MD trajectory of SNase unfolding (500 ps). The superposed structures show the progressive unfolding of the C-terminal helix α_3. The unfolding goes through distinct phases (described in text), separated by color: blue=native fold (t_0=0ps), yellow (t_1=116ps), green (t_2=284ps), red (t_3=411ps), cyan (t_4 =500ps). The corresponding times are indicated on the PCA projection and the Sammon map. The PCA shows the end of phase 2 as an inflection point on the projected trajectory (at t_2), but does not distinguish between the other phases. On the Sammon map, the ends of phase 1, 2 and 3 are visible as inflection points at t_1, t_2 and t_3. Panel **D**: Minimum energy pathway of the signaling switch in Ras p21. The gray regions are very similar in both end states. The main changes are in the Switch 1 and Switch 2 loops, shown in yellow for the starting state (GTP bound state) and in red for the end state (GDP bound state). The complex rearrangement of these loops cannot be resolved by PCA. On the Sammon map, four distinct transition phases are visible, separated by inflections of the path at λ_1, λ_2 and λ_3.

The first eigenvector obtained with PCA contains about 82% of the total variance of the data in the T→ R transition (see Fig. 3.3B), and the projection onto the first two eigenvectors retains ~91%. The residual stress value obtained for the Sammon map is 0.01693. This value is higher than in the previous example, and the number of iteration steps needed for the residual stress to converge is also higher (after 1000 iterations), underlining the fact that the motion in this case is more complex. In the case of hemoglobin the configurational space can be approximately reduced to a

plane. This illustrates how the PCA method is also effective in reducing the dimensionality when the process involves two types of sequentially-occurring motion.

To gain further insight into the hemoglobin change, the *DynDom* domain-motion analysis program was used[24,39]. Dynamic domains, interdomain screw axes, and regions involved in the interdomain bending can be determined with this program. DynDom works with two structures of the protein as input. However, due to the complexity of the hemoglobin transition, an analysis using only the two end states cannot accurately identify domains, motions or hinges along the path. Therefore, the trajectory was divided into the five transitions identified from the PCA projection and Sammon Mapping. The first and last frames of the two quaternary transitions, Q1 and Q2, were used as input. Several domains were identified (see Fig. 3.5) from which one is considered fixed (the yellow domain for the Q1 and the red domain for Q2) with the other identified domains (blue and red for Q1 and blue for Q2) moving relative to the fixed one.

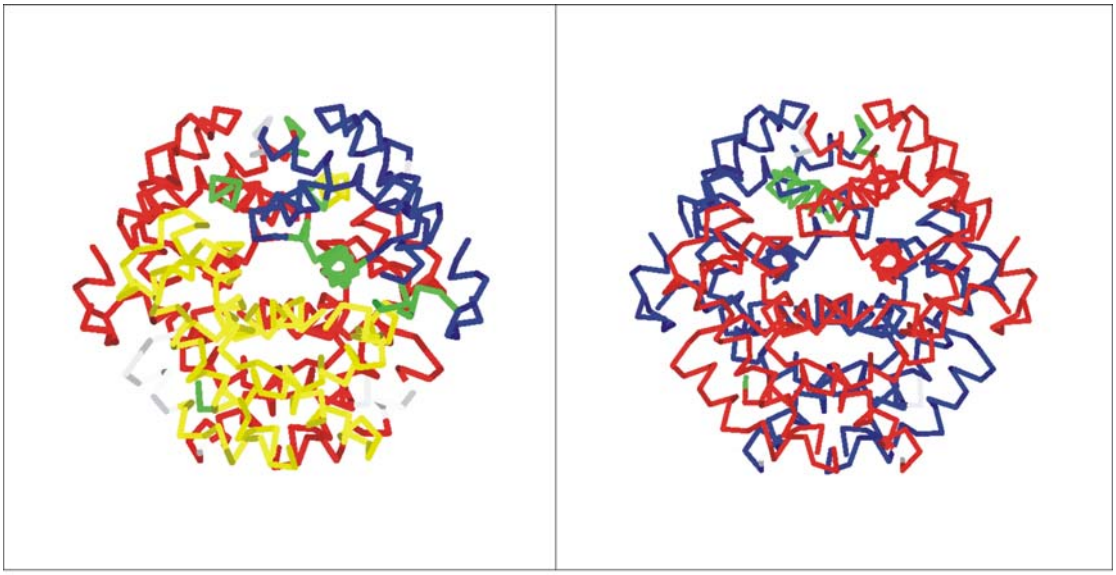


Fig. 3.5: DynDom representation of the two quarternary transitions of hemoglobin. The blue domain rotates around an axis close to the green α-helix. Panel **A**: The first quaternary transition (Q1 in Fig. 3.4B) and Panel **B**: The second quaternary transition (Q2 in Fig. 3.4B).

Of particular interest are bending regions (colored in green for both transitions), which are situated between the domains and are considered to play an important role in the motion that takes place along a path. In Q1 the most important bending region found is the αG-helix, while for Q2, the hinge is the αH-helix. These elements were previously found to act as "rotation axes" (hinges)[12] and the present analysis confirms this suggestion. The hemoglobin case illustrates how PCA and Sammon Mapping are useful in finding individual "events" along a path that can be subsequently selected and used for further analysis (*e.g,* using the DynDom program).

3.4.3 SNase

The third protein analyzed is *Staphylococcal* nuclease (SNase). SNase is a relatively small bacterial enzyme. Due to its small size and the absence of disulfide bridges, SNase constitutes a model system for experimental and theoretical studies of protein folding and function, as demonstrated by the large amount of data available on its structural and biochemical properties.

We examine here a truncated form of SNase, which is known experimentally to exist in a partially-denatured state[40-47]. A molecular dynamics (MD) simulation is analyzed here that examines unfolding of this truncated form in a 500-ps trajectory at 300 K subjected to a harmonic restraint term in the energy function that forces the protein to slowly increase its radius of gyration with time. A detailed description of the simulation has been published[48].

Here, we apply Sammon Mapping and PCA to the SNase MD trajectory. The resulting plots are shown in Fig. 3.4C. The first PCA eigenvector captures about 83% of the total variance (see Fig. 3.3C), the variance content in the second and third eigenvectors being 8% and 2%, respectively. The dominance of the first eigenvector is visible in the projections onto the first/second and second/third eigenvectors, but also from the plot of the cumulative sum of V_i (Fig. 3.3C). The overwhelming contribution of the first eigenvector arises from the fact that, as in the case of myosin, a large-scale conformational change occurs, the main motion in this simulation being

the displacement of the α_3 helix away from the body of the protein. Several key positions are defined along the trajectory (t_1, t_2, t_3). In Fig. 3.4C the differences in the structures at each of these positions can be seen. The frames corresponding to t_0 (blue) and t_1 (yellow) are very similar, and from the Sammon map it can be concluded that, until t_2, significant changes in structure do not occur. A closer look at the Sammon map shows that the trajectory can be split into several segments. In the first segment, between t_0 and t_2, the terminal helix undocks. In the second segment, between t_2 and t_3, the α_3 helix unwinds and other parts of the protein start to unfold. After the unfolding of the α_3 helix, between t_3 and t_4, the helix moves away from the center-of-mass of the protein. These segments, detected with Sammon Mapping, are observable also in the downloadable movie, and are in agreement with previous observations[48]. At the junction between these segments, the trajectory on the Sammon map shows inflections. These are not always noticeable on the PCA projection, as for example between t_2 and t_4, where the method identifies just one phase rather than the two seen with Sammon Mapping. The stress value in this case is 0.02371 after 1000 iterations, again larger than the previous values but still small given the complexity of the motion to be represented. The inability of PCA to identify these transitions arises from the fact that it is dominated by the first eigenvector. This case demonstrates that even in the presence of high informational content in the first two eigenvalues there is no guarantee that PCA will resolve simple motions well.

3.4.4 Ras p21

Finally the two methods are tested on Ras p21. Ras proteins play a pivotal role in the signal transduction pathways that control proliferation, differentiation, and metabolism[49,50]. They act as switches by cycling between their GDP-bound (inactive) and GTP-bound (active) forms. The associated conformational differences are large but relatively localized[51,52]. Without external activation, Ras p21 is found to be mainly in the GDP-bound form[53]. The trajectory that is analyzed here is a CPR path

describing the conformational change occurring after GTP hydrolysis (for more details see Ref. 13).

In Fig. 3.3D it can be observed that the PCA projection for Ras p21 does not encompass as high a percentage of the total variance as in the previous cases: only about 50% is captured in the first eigenvector, rising to 63% with the first two. These values are not high enough to conclude that the motion along these two eigenvectors is meaningful for decomposing the transition. Thus with only the PCA projection it is difficult to interpret the results. The stress value obtained for this protein is 0.03909 after 1000 iterations, again higher than in the previous cases.

It is known that the major conformational changes occur in the Switch I and Switch II regions[54]. The transition involves a complex rearrangement of the backbone fold around the nucleotide binding site (Fig. 3.4D) which, along the CPR path, can be separated into four main events. These events are delineated on the Sammon map by the λ_0, λ_1, λ_2. λ_3 and λ_4 arrows.

The phases λ_0-λ_4 are seen only when using Sammon Mapping, as this method gives a better representation for the cases involving frequent changes of spatial direction. Some phase boundaries, such as λ_2 and λ_3, fall on inflection points along the mapped trajectory, whereas λ_1 is not apparent as an inflection point. The transition that Ras p21 undergoes is very complex and requires a large number of dimensions to reasonably describe it (Fig. 3.3D). Therefore, even a 2D Sammon map cannot fully resolve the different motions. However, the Ras example illustrates how, a conformational transition that is complex and cannot be easily captured in a linear PCA subspace can, using Sammon Mapping, be usefully partitioned.

3.5 Conclusions

PCA is a well established method in biomolecular simulation and efficient algorithms for its computation with guaranteed convergence are readily available. Apart from its use in projecting a high-dimensional data set onto a low dimensional space for viewing and analysis, PCA has also been used to constrain certain motions

(conformational flooding[55,56]) and/or to define a reduced coordinate system for running molecular dynamics (essential dynamics[3]). The fact that the dynamic behavior of a simulated protein can sometimes be captured in only a few directions of configurational space can be used to improve sampling efficiency in MD simulations by driving a second MD run along eigenvectors extracted from an initial MD run[57-59]. However, PCA can detect only linear dependencies. For this reason it is useful to have another method available that can be used as a complement.

Various alternatives to PCA can be used to simplify the analysis of dynamical trajectories, including tree-based analysis[60]. The method examined here, Sammon Mapping, does not perform a coordinate transformation, due to the absence of a defined operation allowing conversion from the initial space to the Sammon map. Therefore, although the method is useful as a nonlinear alternative to PCA for dimension reduction, it cannot itself be used as a reduced coordinate system. The fact that Sammon Mapping method lacks generalization, which means that new points cannot be added to the obtained map without recalculating it, can also sometimes be disadvantageous. Also, the mode of operation (on all interpoint distances) may be problematic due to the CPU-intensive calculations involved that may require a relatively large amount of computer memory. On the other hand, whereas, in the case of PCA, if the protein is large computational limitations may reduce the analysis to the coordinates of C_α atoms so as to keep the size of the matrix manageable, in Sammon Mapping all atoms can be used for obtaining the map, even for very large proteins. A comparison between the PCA projection and the Sammon map obtained using C_α atoms (see Fig. 3.6) shows that for simple motions such as that for myosin examined here, the map is similar to that derived using all the atoms (Fig. 3.6A). However, when the conformational transition is more complex as in the case of SNase (Fig. 3.6B) then it is necessary to include all atoms in the calculation. Another difference between the two methods is that for Sammon Mapping the distances can be weighted (as is apparent from the fractional in eq. 3.4). For this reason, the local geometry is better preserved in these maps (as has been observed also for some structural modeling algorithms[61,62]).

It has been shown that Sammon Mapping is able to map the structure of proteins with similar functionality to the same region of a 2-D Sammon map[7]. Another possible use of Sammon maps could be to compare different trajectories of the same protein simulated under slightly different conditions, in order to see the influence of a certain parameter. Also, if the trajectory involves some very complex motions, the method can be used to simplify the trajectory by allowing to divide it into sub-trajectories between conformations with meaningful differences.

The present paper does not address the problem of convergence of the dynamical trajectories and associated projected representations. For the dynamics examined in the four present cases the motions involved are likely to be somewhat undersampled. Criteria for determining the convergence of PCA modes have been discussed[63].

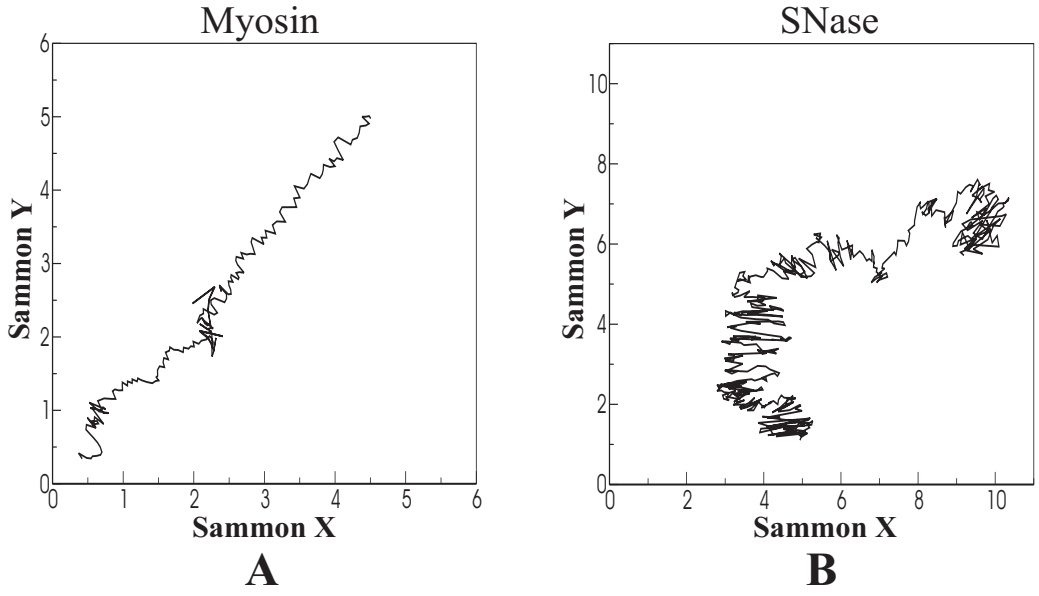


Fig. 3.6. Sammon map obtained when only C_α atoms are used for the distance metric. Panel **A**: Minimum-energy pathway of the recovery-stroke in the myosin head. The map resembles that obtained using all atoms are used for the distance metric (Fig. 3.4A). Panel **B**: Constrained MD trajectory of Snase unfolding. The phases identified when using all atoms (Fig. 3.4C, right panel) are not visible as clearly when only C_α atoms are used.

For a medium-sized protein, applying projection methods often corresponds to reducing a problem with a few thousand degrees of freedom to one with less than three. Although this inevitably leads to loss in information (quantified by the V_i value for PCA and the stress value for Sammon Mapping) it is observed here for all four cases studied that it is possible to capture the main motion in a drastically reduced space. This premise is valid if an important conformational change is captured along a path. A MD trajectory of SNase without constraints (Fig. 3.3E), is used as an example in which the space is not as well represented by a small number of eigenvectors. This path is used as a case where the contribution of each eigenvector is small relative to the constrained MD (Fig. 3.3C), and in consequence many more eigenvectors need to be taken into account when projecting with PCA.

The PCA representation of the hemoglobin T→R transition, when combined with the DynDom analysis, identifies the two kinds of large-scale motion observed earlier[12]: a rotation of the α subunits followed by a rotation of the whole αβ dimer. Sammon Mapping gives a similar picture of the different events, but reveals higher resolution in the localized phases. In some cases, such as the Ras p21 transitions, projections are not easy to interpret. However, as shown in Fig. 3.7 in these cases Sammon Mapping may be used to identify clusters along a path.

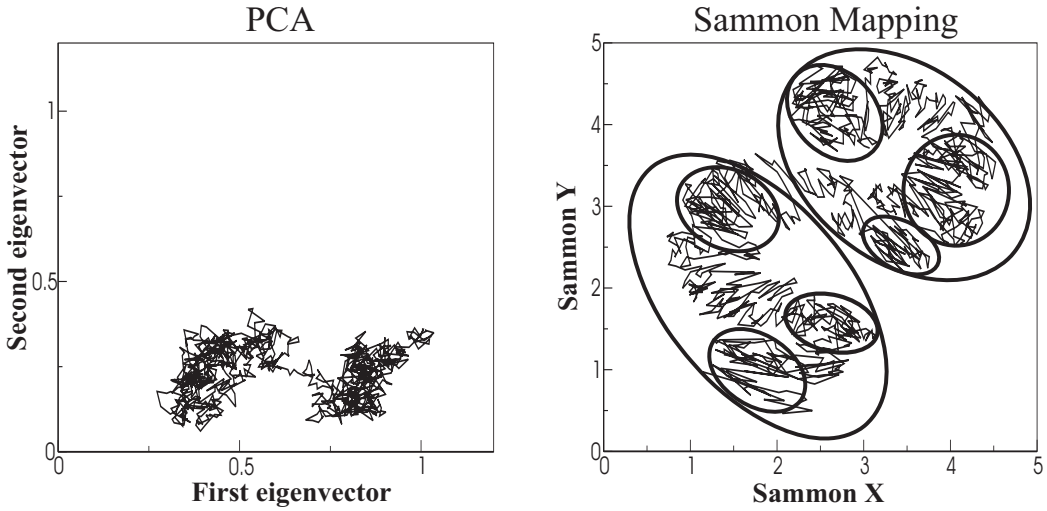


Fig. 3.7. MD of SNase without constraints. Panel **A**: PCA projection of the C_α atoms trajectory onto the plane defined by the two principal components with the largest eigenvalues. Panel **B**: Sammon Mapping

of the all atom trajectory onto a 2D map. Sammon Mapping identifies sub-clusters (shown as circles on the map) that are not resolved by PCA.

PCA and Sammon Mapping provide simplified descriptions of dynamic information obtained using different simulation techniques that cannot be provided by experimental techniques in a straightforward manner. A reduced representation of protein flexibility can thus be obtained. The spectrum of examples examined here demonstrates how judicious choice of PCA and/or Sammon Mapping can enable the sequence of small transitions (events) that make up a large conformational change to be identified and characterized.

References

1. Gerstein M., Lesk A. & Chothia C. - Structural Mechanisms for Domain Movements in Proteins, *Biochemistry* **33(22)**, 6739-6749, 1994.
2. Hayward S., Kitao A. & Go N. - Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and Principal Component Analysis, *Protein Sci*. **3**, 936-943, 1993.
3. Amadei A., Linssen A. B. M. & Berendsen H. J. C. - Essential dynamics of proteins, *Proteins: Structure, Function, and Genetics* **17(3)**, 412-425, 1993.
4. Hayward S., Kitao A., Hirata F. & Go N. - Effect of solvent of collective motions in globular protein, *J. Mol. Biol*. **234**, 1204-1217, 1993.
5. Tournier A. & Smith J. C. - Principal Components of the Protein Dynamical Transition, *Physical Review Letters* **91(20)**, 208106-4, 2003.
6. Agrafiotis D. K. - A new method for analyzing protein sequence relationships based on Sammon maps, *Protein Science* **6(2),** 287-293, 1997.
7. Apostol I. and Szpankowski W. - Indexing and Mapping of Proteins Using a Modified Nonlinear Sammon Projection, *J. Comp. Chem.* **20(10)**, 1049-1059, 1999.
8. Ewing R. M. & Cherry J. M. - Visualization of expression clusters using Sammon's non-linear mapping, *Bioinformatics* **17**, 658-659, 2001.
9. Karpen M. E., Tobias D. J. & Brooks C. L. - Statistical Clustering Techniques for the Analysis of Long Molecular Dynamics Trajectories: Analysis of 2.2-ns Trajectories of YPGDV, *Biochemistry* **32**, 412-420, 1993.
10. Duan J. & Nilsson L. - Thermal Unfolding Simulations of a Multimeric Protein-Transition State and Unfolding Pathways, *Proteins* **59**, 170-182, 2005.
11. Fischer S. & Karplus M. - Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom, *Chem. Phys. Lett*. **194**, 252-261, 1992.

12. Olsen K., Fischer S. & Karplus M. - A continuous path for the T→R allosteric transition of hemoglobin, *Biophysical J*. **78**, 394A, 2000.
13. Noe F., Ille F., Smith J. C. & Fischer S. - Automated computation of low-energy pathways for complex rearrangements in proteins: Application to the conformational switch of Ras p21, *Proteins* **59**, 534-544, 2005.
14. Fischer S., Windshuegel B., Horak D., Holmes K. C. & Smith J. C. - Structural mechanism of the recovery stroke in the Myosin molecular motor, *Proceedings Natl. Acad. Sciences USA* **102(19)**, 6873-6878, 2005.
15. Pearson K. - On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin *Philosophical Magazine and Journal of Science* **2**, 572, 1901.
16. Hotelling H. - Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441,498-520, 1933.
17. Fersht A R - Nucleation mechanisms in protein folding, *Curr. Opin. Struct. Biol.* **7**, 3-9, 1997.
18. Garcia A. N. E. - Large-amplitude nonlinear motions in proteins, *Phys. Rev. Lett.* **68**, 2696-2699, 1992.
19. Becker O. M. - Quantitative visualization of a macromolecular potential energy funnel, *J. Mol. Struct*. (THEOCHEM), 398-399, 507-516, 1997.
20. Karplus M. & Kushick J. N. - Method for estimating the configurational entropy of macromolecules, *Macromolecules* **14**, 325-332, 1981.
21. Perahia D., Levy R. M. & Karplus M. - Motions of an alpha-Helical Polypeptide: Comparison of Molecular and Harmonic Dynamics, *Biopolymers* **29**, 645-677, 1990.
22. Ichiye T. & Karplus M. - Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations, *Proteins: Structure, Function, and Genetics* **11**, 205-217, 1991.
23. De Groot B. L., Van Aalten D. M. F., Amadei A. & Berendsen H. J. C. - The consistency of large concerted motions in proteins in molecular dynamics simulations, *Biophys. J.* **71(4)**, 1707-13, 1996.

24. Hayward S., Kitao A. & Berendsen H. J. C. - Model-Free Methods of Analyzing Domain Motions in Proteins From Simulations: A Comparison of Normal Mode Analysis and Molecular Dynamics Simulation of Lysozyme, *Proteins: Structure, Function, and Genetics* **24**, 425-437, 1997.
25. Caves L. S., Evanseck J. D. & Karplus M. - Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin, *Protein Sci*. **7(3)**, 649-66, 1998.
26. Kidera A. & Go N. - Refinement of Protein Dynamic Structure: Normal Mode Refinement, *PNAS* **87**, 3718-3722, 1990.
27. Mizuguchi K., Kidera A. & Go N. - Collective motions in proteins investigated by X-ray diffuse scattering, *Proteins: Structure, Function, and Genetics* **18**, 34-48, 1994.
28. Abseher R. & Nilges M. - Are there non-trivial dynamic cross-correlations in proteins?, *Journal of Molecular Biology* 279, 911-920, 1998.
29. Van Aalten D. M. F., Findlay J. B. C., Amadei A. & Berendsen H. J. C. - Essential dynamics of the cellular retinol binding protein: evidence for ligand induced conformational changes, *Prot. Eng.* **8**, 1129-1136, 1995.
30. Van Aalten D. M. F., De Groot B. L., Berendsen H. J. C., Findlay J. B. C. & Amadei A. - A comparison of techniques for calculating protein essential dynamics, *J. Comp. Chem*. **18**, 169-181, 1997.
31. Sammon J. W. Jr. - A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers* **C-18**, 401-409, 1969.
32. Pekalska E., de Ridder D., Duin R. P. W. & Kraaijveld M. A. - A new method of generalizing Sammon mapping with application to algorithm speed-up, *Annual Conference of the Advanced School for Computing and Imaging*, 221-228, 1999.
33. Barlow T. W. & Richards W. G. - A novel representation of protein structure, *J. Mol. Graphics* **13**, 373-376, 1995.
34. Fisher A. J., Smith C. A., Thoden J., Smith R., Sutoh K., Holden H. M. & Rayment I. - X-ray Structures of the Myosin Motor Domain of Dictyostelium

discoideum Complexed with MgADP.cntdot.BeFx and MgADP.cntdot.AlF4-, *Biochemistry* **34(28)**, 8960-8972, 1995.

35. Smith C. A. & Rayment I. - X-ray Structure of the Magnesium(II)·ADP·Vanadate Complex of the Dictyostelium discoideum Myosin Motor Domain to 1.9 Å Resolution, *Biochemistry* **35(17)**, 5404-5417, 1996.
36. Lymn R. W. & Taylor E. W. - Mechanism of adenosine triphosphate hydrolysis by actomyosin, *Biochemistry* **10(25)**, 4617-4624, 1971.
37. Geeves M. A. & Holmes K. C. - Structural mechanism of muscle contraction, Ann. Rev. *Biochem.* **68**, 687-728, 1999.
38. Baldwin J. & Chothia C. - Hemoglobin: the structural changes related to ligand binding and its allosteric mechanism, *J. Mol. Biol*. **129**, 175-220, 1979.
39. Hayward S. & Berendsen H. J. C. - Systematic Analysis of Domain Motions in Proteins From Conformational Change: New Results on Citrate Synthase and T4 Lysozyme, *Proteins: Structure, Function, and Genetics* **30**, 144-154, 1998.
40. Zhou B. & Jing G. Z. - Conformational features of a truncated staphylococcal nuclease R (SNR135) and their implications for catalysis, *Arc. Biochem. Biophys*. **360**, 33-40, 1998.
41. Whitten S. T. & Garcia-Moreno B. - pH dependence of stability of staphylococcal nuclease: evidence of substantial electrostatic interactions in the denatured state, *Biochemistry* **39**, 14292-14304, 2000.
42. Alexandrescu A. T. & Shortle D. - Backbone dynamics of a highly disordered 131 residue fragment of Staphylococcal nuclease, *J. Mol. Biol.* **242**, 527-546, 1994.
43. Wrabl J. O., Shortle D. & Woolf T. B. - Correlation between changes in nuclear magnetic resonance order parameters and conformational entropy: molecular dynamics simulations of native and denatured staphylococcal nuclease, *Proteins* **38**, 123-133, 2000.

44. Ermacora M. R., Ledman D. W. & Fox R. O. - Mapping the structure of a non-native state of staphylococcal nuclease, *Nat. Struct. Biol.* **3**, 59-65, 1996.
45. Wang Yi & Shortle A. - A dynamic bundle of four adjacent hydrophobic segments in the denatured state of staphylococcal nuclease, *Protein Sci.* **5**, 1898-1906, 1996.
46. Shortle D. & Meeker A. K. - Residual structure in large fragments of staphylococcal nuclease: effects of amino acid substitution, *Biochemistry* **28**, 936-944, 1989.
47. Flanagan J. M., Kataoka M., Shortle D. & Angelman D. M. - Truncated staphylococcal nuclease is compact but disordered, *Proc. Natl. Acad. Sci. USA* **89**, 748-752, 1992.
48. Gruia A. D., Fischer S. & Smith J. - Molecular dynamics simulations reveals a surface salt bridge forming a kinetic trap in unfolding of truncated staphylococcal nuclease, *Proteins* **50**, 507-515, 2003.
49. Barbacid M. - Ras Genes, *Ann. Rev. Biochem*. **56**, 779-827, 1987.
50. Lowy D. R. & Willumsen B. M. - Function and regulation of Ras, Ann. Rev. *Biochem.* **62**, 851-891, 1993.
51. Milburn M. V., Tong L., deVos A. M., Brűnger A., Yamaizumi Z., Nishimura S. & Kim S. H. - Molecular switch for signal transduction: Structural differences between active and inactive forms of protooncogenic Ras proteins, *Science* **247**, 939-945, 1990.
52. Wittinghofer A. & Pai E. - The structure of ras protein: a model for a universal molecular switch, *Trends Biochem. Sci.* **16**, 382-387, 1991.
53. Wittinghofer A. & Herrmann C. - Ras-effector interactions, the problem of specificity, *FEBS Lett*. **369**, 52-56, 1995.
54. Ma J. and Karplus M. - Molecular switch in signal transduction: Reaction paths of the conformational changes in ras p21, *Proc. Natl. Sci. USA* **94**, 11905-11910, 1997.
55. Grubmüller H. - Predicting slow structural transitions in macromolecular systems: Conformational flooding, *Phys. Rev. E* **52**, 2893-2906, 1995.

56. Schulze B. G., Grubműller H. & Evanseck J. D. - Functional Significance of Hierarchical Tiers in Carbonmonoxy Myoglobin: Conformational Substrates and Transitions Studied by Conformational Flooding Simulations, *J. Am. Chem. Soc.* **12**, 8700-8711, 2000.
57. Amadei A., Linssen A. B. M., De Groot B. L., Van Aalten D. M. F. & Berendsen H. J. C. - An efficient method for sampling the essential subspace of Proteins, *J. Biom. Str. Dyn.* **13(4)**, 615-626, 1996.
58. De Groot B. L., Amadei A., Scheek R. M., Van Nuland N. A. J. & Berendsen H. J. C. - An extended sampling of the configurational space of HPr from E.coli, *Proteins* **26**, 314-322, 1996.
59. De Groot B. L., Amadei A., Van Aalten D. M. F. & Berendsen H. J. C. - Towards an exhausive sampling of the configurational spaces of the two forms of the peptide hormone guanylin, *J. Biomol. Str. Dyn*. **13**, 741-751, 1996.
60. Ota M., Ikeguchi M. & Kidera A. - Phylogeny of protein-folding trajectories reveals a unique pathway to native structure, *PNAS* **101(51)**, 17658-17663, 2004.
61. Aszodi A. & Taylor W. T. - Homology modeling by distance geometry, *Folding Design* **1**, 325-334, 1996.
62. Aszodi A., Munro R. E. J. & Taylor W. T. - Distance geometry based comparative modeling, *Folding Design* **2**, S3-S6, 1997.
63. Hess B. - Convergence of sampling in protein simulations, *Phys. Rev. E* **65**, 031910, 2002.

Chapter 4

The molecular motor Myosin II

The present thesis focuses on one particular molecular motor that enables skeletal muscle contraction: myosin. Since its discovery, many experimental and theoretical studies have attempted to elucidate its mechanism at atomic detail. For a better understanding of the relevance of the theoretical studies presented in the following chapters, a brief description of different myosin classes is given below, with more detailed information about myosin II isolated from the slime mold *Dictyostellium discoideum* as it is the best experimentally-characterized among myosins and for which a large amount of structural and functional data exists. For this reason, this myosin constitutes the protein model analyzed in the studies described in Chapters 5, 6 and Appendix.

4.1 Molecular motors

All forms of movement in the living world are powered by protein machines known as molecular motors. Molecular motors are proteins that utilize energy from the hydrolysis of a nucleotide triphosphate and convert it into mechanical work. They are involved in a large variety of cellular tasks and operate by small increments, converting changes in protein conformation into direct motion. Orderly motion across distances requires a track that steers the motion of the motor assembly. There are three families of molecular motors involved in motion along filaments: myosins, kinesins, and dyneins. A compilation of motor protein organization, function, and regulation is given in Ref.1 and 2 and strategies for the intracellular transport of the molecular motors are reviewed in Ref. 3.

Apart from the motor proteins that move along predefined tracks there are molecular motors that couple proton motive force to hydrolysis of nucleotide triphosphate. Examples for this type of motors are the bacterial flagellar motor[1] or the F_1 motor of ATP synthase.[4-9] Molecular motors are also involved in RNA and DNA polymerization (polymerases) and unwinding of nucleic acids (helicases).

A focus on the myosin family of molecular motors is made in Section 4.1.1, and some of their movement strategies are presented in Section 4.1.2.

4.1.1 Myosin classes

The first motor protein identified was skeletal muscle myosin, which is responsible for generating the force for muscle contraction (in more detail presented in Section 1.2). It was initially thought that myosin was present only in muscle, but in 1970's, researchers found that a similar myosin protein was also present in nonmuscle cells, including protozoan cells. Subsequently, many other myosin types were discovered.

The myosin super-family is classified into 18 different myosin classes.[10,11] Relationships between different myosins as judged by sequence similarity are shown in Fig. 4.1.

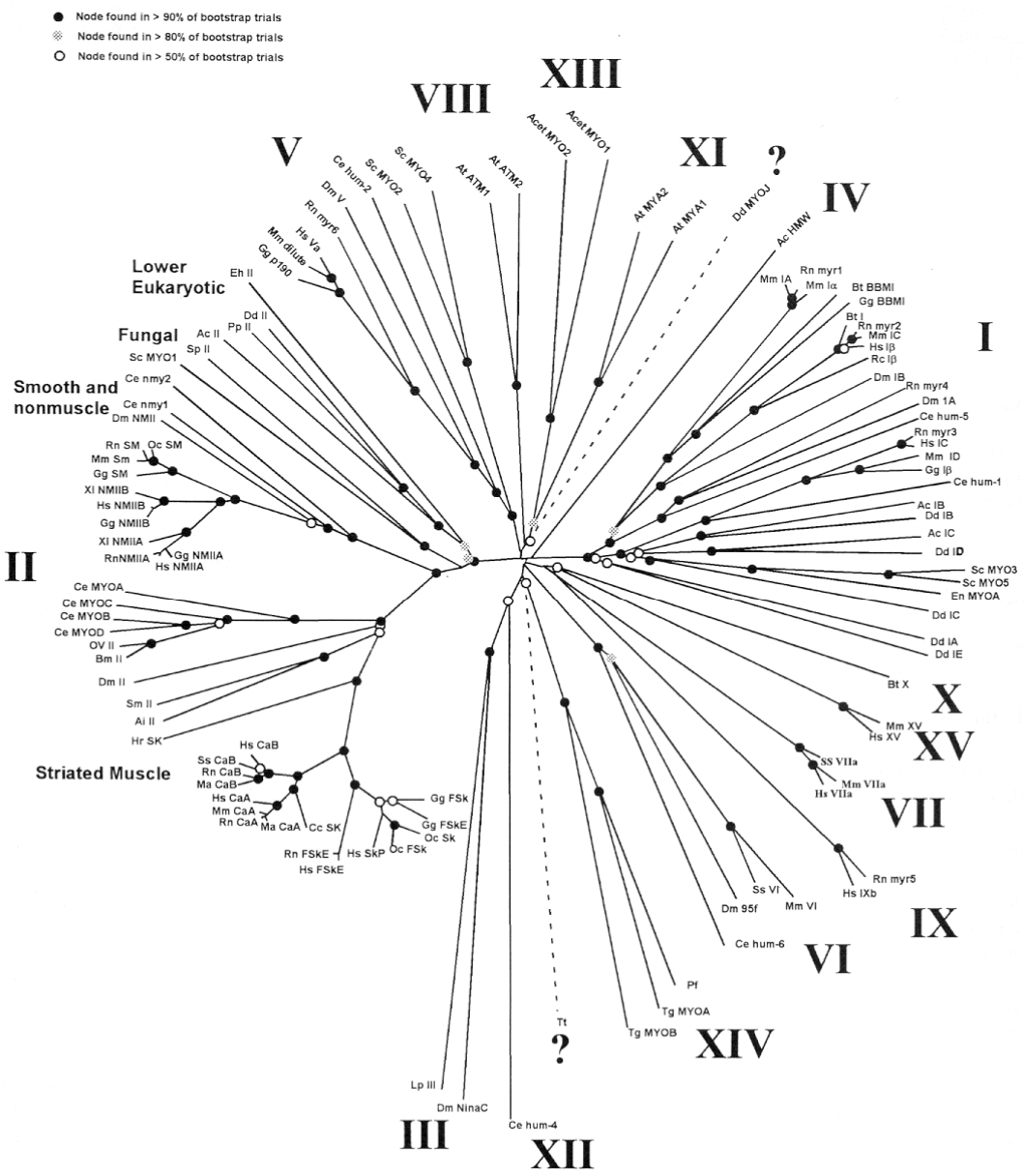


Fig. 4.1 Unrooted phylogenetic tree obtained from the myosin motor domain sequences. The figure is taken from Ref. 10.

Myosin is an elongated protein formed from two heavy chains and two copies of each of the two light chains. Each of the heavy chains has a globular head domain at its N-terminus that contains the force generating machinery, followed by a long amino acid sequence that forms an extended coiled-coil (see Fig. 4.2) Small structural changes at the catalytic site in the motor domain are converted into a large swing of the light-chain binding domain that thus serves as a lever arm.

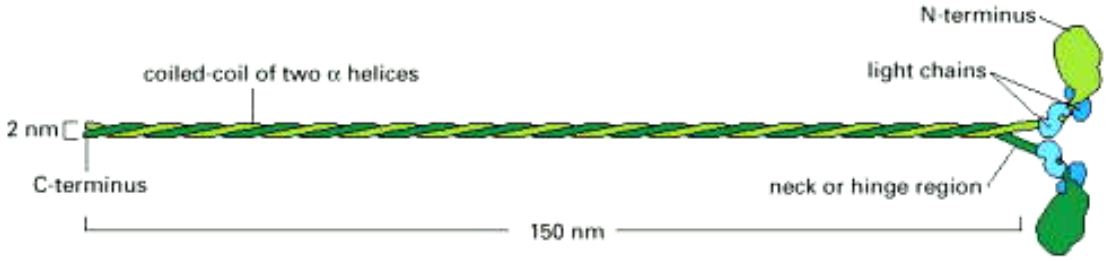


Fig. 4.2 A myosin II molecule composed of two heavy chains (each about 2000 amino acids long (green color) and four light chains (blue color). The figure is adapted from Ref. 12.

The best characterized myosins are the class II myosins (also called conventional myosins) presented in more detail in Section 4.2. The unconventional myosins and their receptors are involved in diverse tasks such as in organelle translocation and cytoskeletal reorganization. Myosin I for example has been shown to produce its working stroke in two steps[13-18]. Cytokinesis, or nonmuscle myosin II, is responsible for maintenance of cell integrity and structure. Myosin III[19] is involved in cell signaling, whereas myosin V is involved in vesicle transport and membrane trafficking and its mechanochemical coupling has very recently been studied.[20-22] Particle transport and anchoring is present in myosin VI[23], myosin VII is involved in cell adhesion, hearing, and maintenance of balance, whereas signaling has been shown to involve myosin IX, phyllopod extension myosin X, and gliding motility myosin XIV. An overview of myosin functions and diseases due to myosin dysfunction is given in Ref. 24.

In spite of the accumulating amount of structural and functional information about all types of myosin, their mechanism is not yet known or understood in atomic detail.

4.1.2 Movement strategies

Movement strategies used by different motors depends mainly on the function of the motor, which can adopt different strategies regarding the way of moving along their associated filaments. Motor proteins use the energy of ATP hydrolysis to move along microtubules or actin filaments. All known motor proteins that move on actin filaments are members of the myosin super-family. The motor proteins that move on microtubules are members of either the kinesin super-family or the dynein family.

Motor proteins can move either processively or nonprocessivly. Processivity is defined as the average number of steps taken per diffusional encounter between a motor and its filament.[25] Another key feature characteristic to a specific motor is the directionality.[26,27] All myosins except one, move toward the plus end of an actin filament, although they do so at different speeds. The exception is myosin VI, which moves toward the minus end.[12] Myosin V motors for instance are involved in intracellular trafficking having vesicles as cargo to be transported along the actin cytoskeleton. It has been shown that the myosin V dimer utilizes a mechanism in which one head of the myosin dimer remains attached to the track at all times (high-resolution fluorescence imaging techniques showed the presence of a hand-over-hand mechanism[28]).

A classical example of a non-processive motor is Myosin II that detaches from its filament at each ATP turnover. This is adequate for myosin II because many myosin II monomers self-assemble into filaments that slide relative to its actin filaments. The “cargo” in this case is the myosin filament itself.

Although myosin and kinesin walk along different tracks and use different mechanisms to produce force and movement by ATP hydrolysis, they share a common structural core. The fact that myosins and kinesins seem to use the same

hydrolysis mechanism is also suggested by crystal structure analysis of different motors revealing that the motor domain of kinesin is structurally similar to the myosin motor domain[29].

4.2 The molecular motor Myosin II

Details on myosin motors with particular emphasis on conventional myosins are presented in this section. A description of the organization of skeletal muscle is made in Section 4.2.1. The actomyosin interaction is presented through the sliding filament model (see Section 4.2.2), followed by a description of the structural features of myosin II (Section 4.2.3). Finally, a short description of different conformational states of myosin is done in Section 4.2.4.

4.2.1 Skeletal muscle

Skeletal muscle is one type of striated muscle, attached to the skeleton. The microanatomy of vertebrate skeletal muscle comprises bundles of fibers sheathed in connective tissue (see Fig. 4.3). Each skeletal muscle fiber is a single cylindrical muscle cell. An individual skeletal muscle may be made up of hundreds, or even thousands, of muscle fibers bundled together and wrapped in a connective tissue covering. A connective tissue sheath called the epimysium surrounds each muscle. Portions of the epimysium project inward to divide the muscle into compartments. Each compartment contains a bundle of muscle fibers. Each bundle of muscle fiber is called a fasciculus and is surrounded by a layer of connective tissue (the perimysium). Within the fasciculus, each individual muscle cell, called a muscle fiber, is surrounded by a connective tissue (the endomysium). Each fiber contains around thousand myofibrils, the rod-like organelles responsible for contraction occupying up to 80% of

its volume. A striking feature of each fiber, when examined under a microscope, is the presence of cross bands arising from aligned striations on each myofibril (Fig. 4.4).

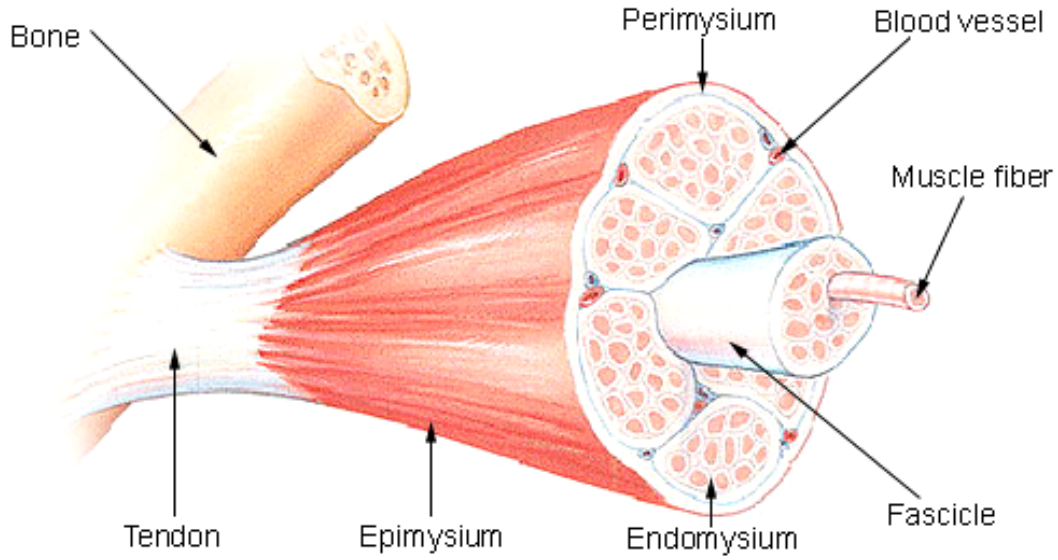


Fig. 4.3 Structure of a Skeletal Muscle. Figure taken from Ref. 30.

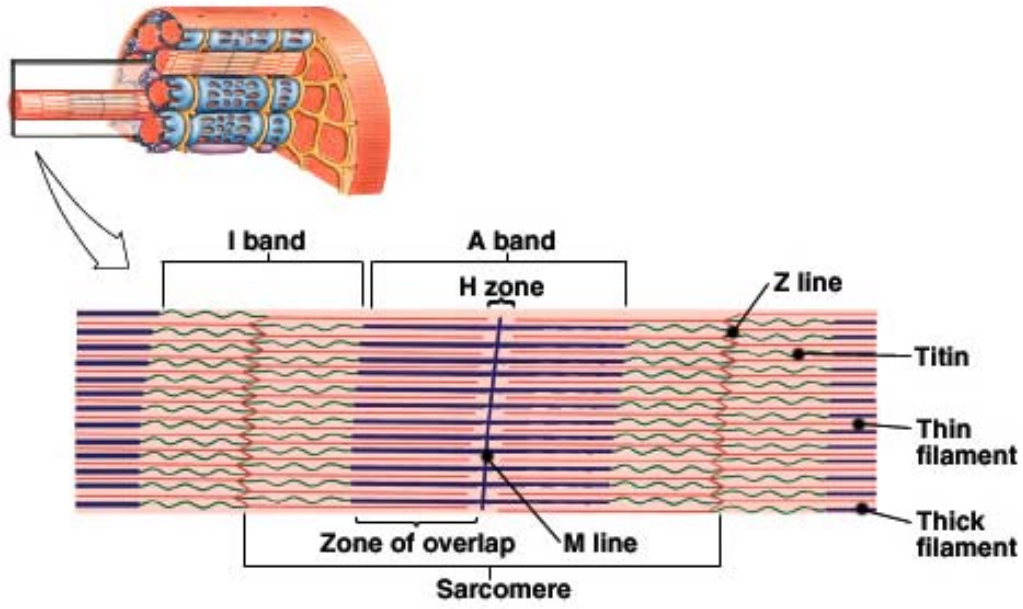


Fig. 4.4 Bands and lines in the contractile apparatus of skeletal muscle.

This distinct repeating pattern of bands (also known as the sarcomere), formed by the Z, I, A, H, and the M bands, is the product of the interaction between the proteins that form the thick filaments and the thin filaments. Under the polarizing microscope the protein dense bands are seen to be anisotropic with respect to the refractive index (A-band), while the less dense regions between them are relatively isotropic (I-band). In the center of the I-bands a highly refractive Z-line is observed, while in the middle of the A-band is less dense, forming the H-zone. Thin filaments emanate from Z-line and make up the I-band, while the thick filaments make up the A-band. The M-line is revealed in the middle of the H-zone holding the thick filaments. The unit between two Z-lines is defined as a sarcomere.

4.2.2 The sliding filament model

Using interference microscopy to view living muscle fibers[31] showed that on stretching or shortening, the A-bands remain at constant length, while the I-bands change (Fig. 4.5). Phase contrast microscopy[32] revealed that when isolated myofibrils were induced to contract by addition of ATP, the I-band and the H-zone shortened in unison.

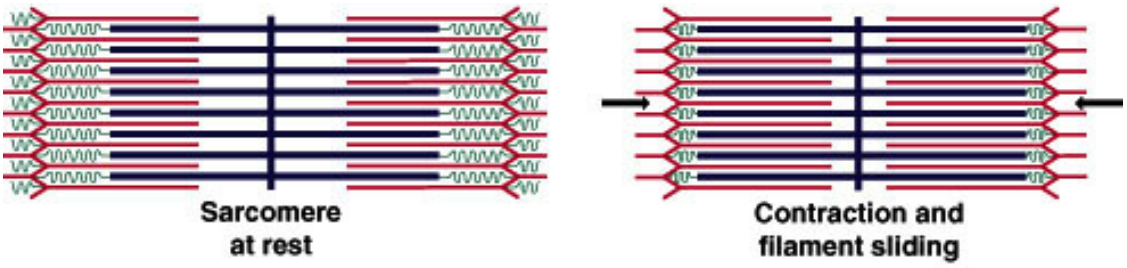


Fig. 4.5 Change in the size of some bands with the contraction of the sarcomere.

These findings, taken in conjunction with electron micrographs, which demonstrated the underlying basis for the striations, led to the sliding filament theory in which contraction was proposed to occur solely by interdigitation of thick and thin filaments.

The thick filament comprises mainly myosin and the thin filament mainly actin, as shown by selective extraction of the A- and I-bands with salt solutions. The motion is initiated by major structural rearrangements in myosin resulting in a hinge-like bending of the myosin head. The necessary energy is provided by ATP hydrolysis. The sliding filament theory first suggested by Hugh Huxley[32] is now widely accepted. The sliding filament model assumes four different structural states of the actomyosin interaction (reviewed for example in Refs. 33 and 34). A schematic representation of this model is shown in Fig. 4.6.

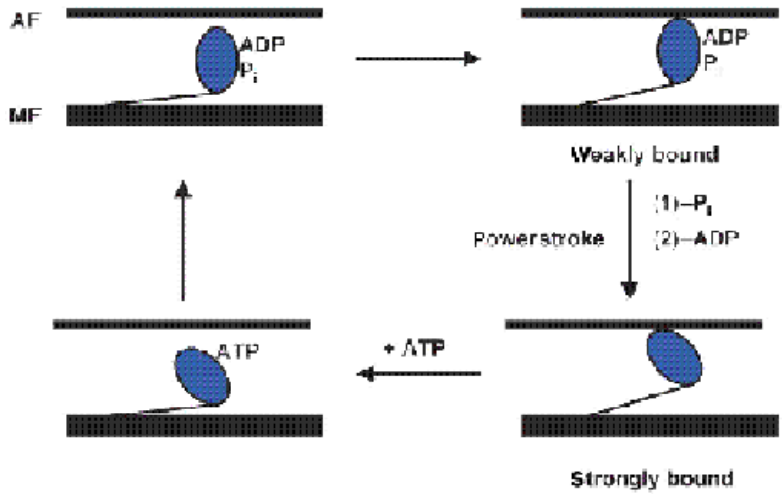


Fig. 4.6 The sliding filament model of muscle contraction. The figure was taken from Ref. 35.

Myosin heads (also called cross-bridges) are detached from actin in the post-hydrolysis state (upper left in Fig. 4.6) in which the reaction products ADP and Pi are bound. Weak actin binding results in formation of the pre-power-stroke state (upper right) and triggers product release. Along with product release, the power-stroke occurs in which myosin moves past actin. After the power-stroke the post-power-stroke state is formed in which myosin strongly binds actin (lower right). This state is also called the rigor state. ATP binding dissociates the actomyosin complex, leading to the pre-recovery-stroke state (lower left). A large conformational change called the recovery stroke reverses the conformational change of the power stroke while myosin

remains detached from actin. The recovery stroke is coupled to the hydrolysis of ATP. Thus, reaching the post-hydrolysis state again and completing the cycle.

The sliding filament model implies that tiny changes in the myosin head are amplified and result in a large motion of the neck domain serving as a rigid lever arm. Consequently, longer the lever arm larger is the step size taken. The step-size that muscle myosin II takes is dependent on the load present and has been measured (in intact muscle fibers) to be between 8 and 13 nm in each interaction with actin.[36]

4.2.3 Structural features of Myosin II

Global organization of myosin. Myosin II or conventional myosin is a hexameric protein. A monomer of myosin II consists of one heavy chain and two light chains. Each myosin heavy chain can be split into one light meromyosin (LMM) and one heavy meromyosin (HMM) schematically represented in Fig. 4.7. HMM can further be split into an N-terminal globular subfragment (S1, also called myosin head fragment), and 1 rod-shaped subfragment (S2). The rod like tail sequence is highly repetitive, showing cycles of a 28-residue that are characteristic for alpha-helical coiled coils.[37] S1 consists of a globular domain at the N-terminus from which a α-helical "tail" extends. The essential (ELC) and the regulatory (RLC) light chain are stabilizing the beginning of the tail.

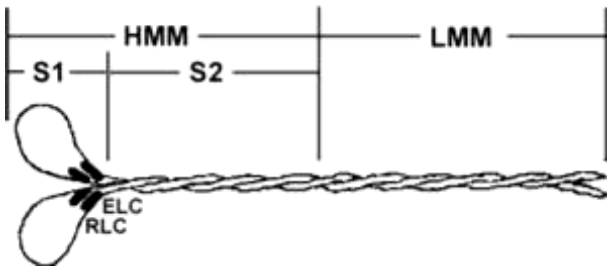


Fig. 4.7 Diagram representation of the myosin II protein and its subfragments (Source:http://www.cytoskeleton.com/products/actinbind/my02.html).

The C-terminal α-helices belonging to the two heavy chains form a coiled-coil structure (a homodimer), which self-assemble into filaments. Further limited proteolysis breaks the N-terminal globular domain into several fragments that are named after their apparent molecular masses: 25 kDa (N-terminal), 50 kDa (middle) and 25 kDa (C-terminal)[38] hosting the catalytic site and the actin-binding region. It is therefore referred to as the catalytic or motor domain. The myosin motor domain alone (consisting of the first 759 amino acids), has been shown to be functional[39].

Subdomain organization of the globular domain. There are four specific subdomains (see Fig. 1.8) that make up the heavy chain molecule of myosin: the N-terminal domain (25 kDa, residues 1 to 200) which is flanked by the 50 kDa domain that is split into the upper (residues 201 to 475) and lower (residues 476 to 613) 50 kDa domains and the converter domain (residues 711 to 781). Between the upper and lower 50 kDa domains there is a cleft that is supposed to be closed upon actin binding. Numerous α–helices surrounding a 7-stranded β-sheet, form a deep cleft extending from the nucleotide-binding site to the actin-binding face. At the tip of the cleft two loops that are known to play major roles in actin binding[40-42] are present: the cardiomyopathy loop (HCM loop, belonging to the upper 50 kDa domain, Ile398-Val405) and the loop-2 (G519-G525). At the bottom of the cleft a third loop termed the strut loop (Asp590-Gln593), connects the upper and the lower 50 kDa domains. The α-helical neck region (also called light-chain binding domain or LCBD) provides the binding sites to which the essential (ELC) and regulatory (RLC) light chains are bound. The α-helical tail responsible for filament formation is not shown in Fig. 4.8.

The rigid light-chain binding domain is believed to serve as a lever arm that amplifies the slight structural changes at the nucleotide-binding pocket into large movements. It has been shown that the native LCBD can be replaced by an artificial domain of similar rigidity and dimensions without loss of functionality.[44]

Conformational changes in the nucleotide-binding pocket are induced to the reactive thiol region that got its name from two cysteine residues that can be chemically cross-linked. Chemical modification of them would produce significant alterations of the ATPase activity and actin binding affinity. The reactivity of the thiol

groups can also be used to introduce spin labels as has been done to measure electron paramagnetic resonance (EPR) spectra.[45,46]

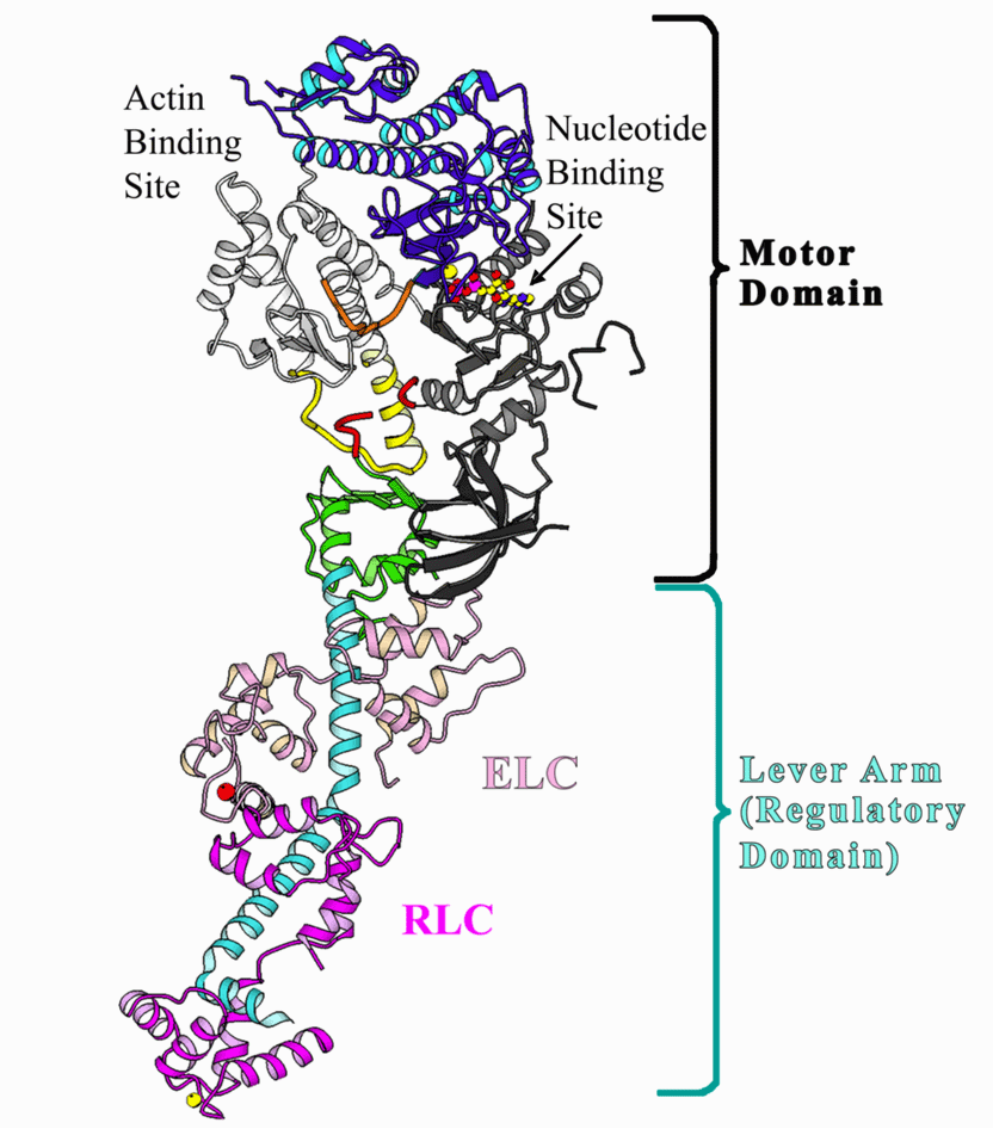


Fig. 4.8 Three-dimensional structure of chicken fast skeletal muscle myosin S1. The figure is taken from Ref. 43.

Binding sites. The nucleotide binding site is located near the 25kDa-50 kDa fragment boundary (see Fig. 4.9) and it is composed of three loops that are conserved not only among motor proteins but also among the G-proteins. These are the P-loop that is a

common feature of a large number of enzymes that bind nucleotide[47], the switch-1 loop and the switch-2 loop. Together they form the so-called phosphate tube. The switch-1 and switch-2 loops are located in the upper respectively lower 50 kDa domains, whereas the P-loop belongs to the N-terminal 25 kDa domain. The two switch loops got their names from the observation that they can adopt different conformation, thus serving as a switch for information transduction by changing their conformation.

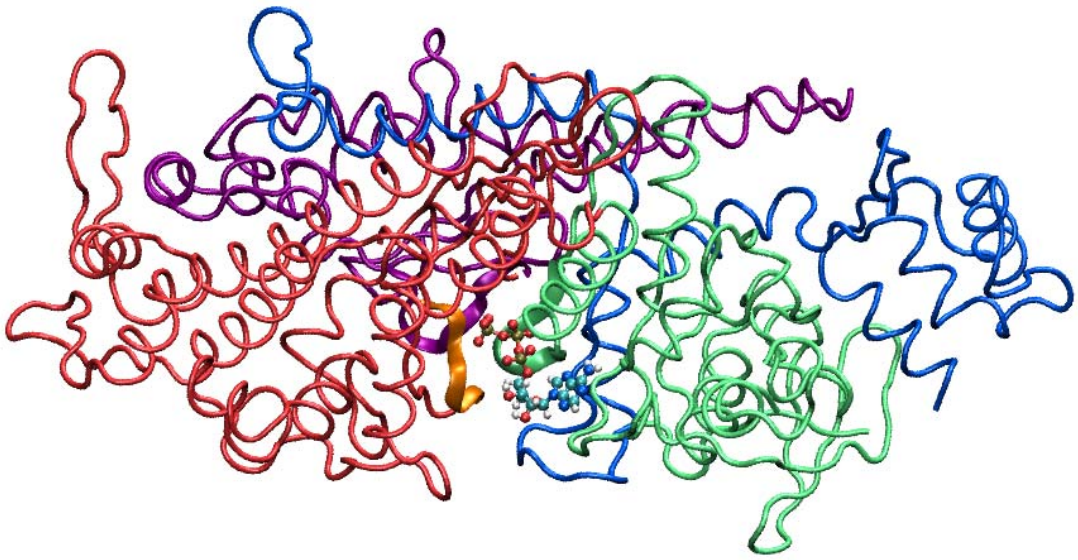


Fig. 4.9 Localization of the nucleotide binding pocket in the catalytic domain of *Dictyostelium discoideum* myosin II. The figure was prepared from the structure in Ref. 48. The N-terminal 25 kDa domain is shown in green, the upper 50 kDa domain in red, the lower 50 kDa domain in purple, the C-terminal 20 kDa domain (including the converter domain) in blue, the P-loop in green, the switch-1 loop in orange, the switch-2 loop in purple, and the nucleotide in CPK representation.

In the closed conformation of the nucleotide pocket when both switch loops are located close to each other and in close spatial proximity to the P-loop, there is a salt bridge between Arg238 (switch-1) and Glu459 (switch-2). This salt bridge is only formed in the closed conformation and when the switch-2 loop swings away, the nucleotide pocket adopts its open conformation. Fig. 4.10 shows an overlap of the

open and closed nucleotide pockets as have been observed in crystallographic studies[48,49]. It is obvious from the picture that the positions of the P-loop, switch-1 loop and nucleotide are almost identical in both structures, whereas the switch-2 loop (colored purple) has moved by about 4 Å thus breaking the salt bridge in the open structure (showed transparent in Fig. 4.10).

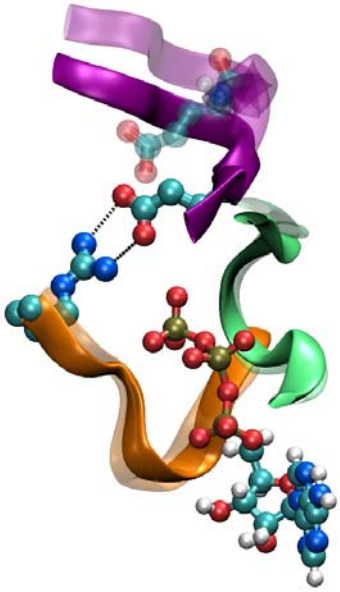


Fig. 4.10 Overlap between the closed (opaque colors, ref1[49]) and open (transparent colors, prepared from PDB code 1MMD[48]) form of the nucleotide-binding pocket elements (P-loop is shown in green, switch-1 loop in orange, switch-2 loop in purple). ADP.BeF_3 and the salt-bridge between Arg238 (switch-1) and Glu459 (switch-2) are in CPK representation.

The positions of the switch-1 and switch-2 loops can be used to classify different conformational states of myosin. Conformations corresponding to switch-1 closed / switch-2 closed (C/C), switch-1 closed / switch-2 open (C/O), and switch-1 open / switch-2 open (O/O) have been identified by crystallography (an overview of this conformations is given in Table 4.1). Binding of ATP induces a conformational change towards the C/C state of the nucleotide-binding pocket, characterized by a well-defined γ-phosphate binding site that can be occupied either by phosphate itself

or by γ-phosphate analogs such as vanadate, beryllium fluoride, or magnesium fluoride.[50]

PDB ID	resolution [Å]	ligand	state	Ref.
1D0X	2.00	m-nitrophenyl aminoethyl.PP_i.BeF_3	C/O	(51)
1D0Y	2.00	o-nitrophenyl aminoethyl.PP_i.BeF_3	C/O	(51)
1D1A	2.00	o,p-dinitrophenyl aminoethyl.PP_i.BeF_3	C/O	(51)
1D1B	2.00	o,p-dinitrophenyl aminopropyl.PP_i.BeF_3	C/O	(51)
1D1C	2.30	N-methyl-o-nitrophenyl aminoethyl.PP_i.BeF_3	C/O	(51)
1FMV	2.10	-	C/O	(52)
1FMW	2.15	Mg.ATP	C/O	(52)
1G8X	2.80	Mg.ADP	C/O	(53)
1LVK	1.90	Mg.BeF_3.*mant*ADP	C/O	(54)
1MMA	2.10	Mg.ADP	C/O	(55)
1MMD	2.00	Mg.ADP.BeF_3	C/O	(48)
1MMG	1.90	Mg.ATPγS	C/O	(55)
1MMN	2.10	Mg.AMPPNP	C/O	(55)
1MND	2.60	Mg.ADP.AlF_4	C/C	(48)
1MNE	2.70	Mg.PP_i	C/O	(56)
1Q5G	1.90	-	O/O	(57)
1VOM	1.90	Mg.ADP.VO_4	C/C	(58)
REF1		Mg.ADP.BeF_3	C/C	(49)

Table 4.1: Available crystal structures for *Dictyostelium discoideum* myosin II. The conformational states of switch-1/switch-2 are indicated. ``O'' refers to the open state, whereas ``C'' refers to the closed state. Table taken from Ref. 59.

Nucleotide binding occurs from the solvent-exposed side of the pocket also termed as the "front-door" as revealed by a crystal structure in which the nucleotide is partially bound[60]. Usually, product release occurs *via* the same route as substrate

binding. In myosin, however, P_i release has been shown to occur *prior* to ADP release[61,62]. As long as the nucleotide remains bound this front door is locked. Thus, phosphate release has been postulated to occur *via* a back-door mechanism. One possible back door consists in the opening of the phosphate tube and it is discussed in detail based on molecular dynamics simulations in Ref. 63.

A detailed mechanism how the conformational changes at the catalytic site are coupled to conformational changes of the converter domain has recently been proposed based on computational investigations. The structural elements: relay helix (Phe466 to Lys498), relay-loop (Ile499 to Asp509), and SH1-helix (Val681 to Lys690) have been found to play significant roles as mediators[64].

The ATP-binding site is about 4.0 nm away from the actin-binding site (see Fig. 4.8). Like the nucleotide-binding pocket, the actin-binding face of myosin must account for several properties crucial to the function of this motor. Primarily, it must form a strong interaction with actin during the power stroke that allows putative conformational changes in myosin to result in production of force and translocation of the actin filament. In addition, the interaction of this face of the myosin with actin must signal the release of hydrolysis products from the active site. The actin-binding site also must lower its affinity for actin in response to ATP binding.

Binding of nucleotide and of actin to myosin is antagonistic: ATP binding to actomyosin causes actin dissociation and actin binding accelerates P_i and ADP release. It has been shown that the 50 kDa cleft undergoes structural changes upon actin binding[65]. Thus, if phosphate is released into the 50 kDa cleft *via* the back door it may directly induce conformational changes of the 50 kDa cleft thereby influencing the affinity for actin binding. An alternative view is suggested by a recent nucleotide-free crystal structure of myosin II in which both the switch-1 and switch-2 loops have moved away from their positions in the CLOSED structure[57].

4.2.4 Myosin conformers

Myosin can adopt different conformational states some of which have been trapped in crystallographic studies. Since the first crystal structure of the motor domain published in 1993[66], a large number of X-ray crystallographic structures have been reported. An overview of the currently available crystal structures of *Dictyostelium discoideum* myosin II was presented in Table 4.1. Also a third actin-detached structural state has been identified for scallop myosin subfragment 1.[43,67,68]

The closure of the switch-2 domain causes a ~60˚ rotation of the converter domain (highlighted in Fig. 4.11) with a fulcrum at the distal end of the SH1-helix (purple in Fig. 4.11). As the converter domain is directly connected to the α-helical neck region, its rotation results in an amplified movement of the light chain binding domain (~11 nm). Motions of the latter (lever arm) have been resolved using EPR spectroscopy.[69]

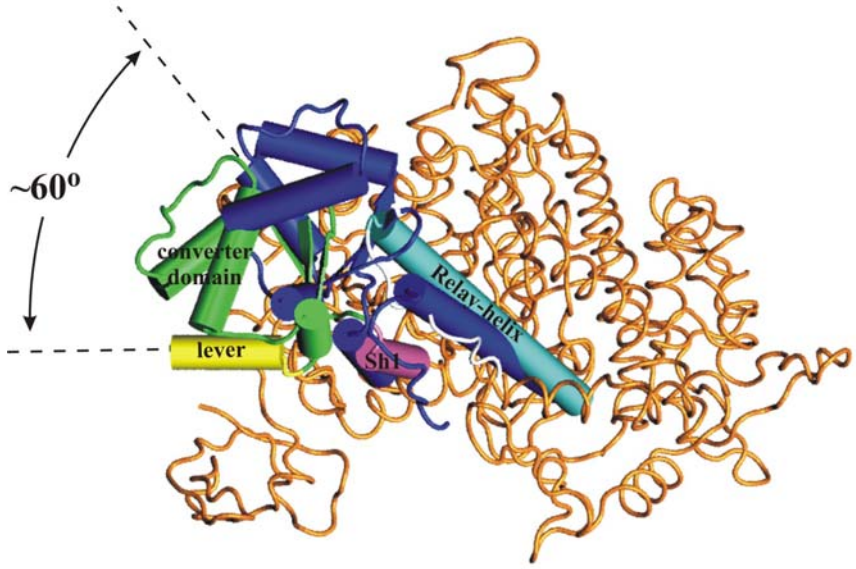


Fig. 4.11 The OPEN and the CLOSED conformations of myosin head representing the end states if the recovery-stroke. Superposition of the converter-domain in CLOSED structure (dark blue) onto the OPEN structure (other colors). The main body (orange) does not change significantly (only the OPEN conformation is shown).

The large change in conformation corresponding to the beginning (OPEN) and the end (CLOSED) of lever arm's movement is driven through molecular cogs and gears by a small (0.5 nm) change in the active site. Thus, if switch-2 is in its OPEN state, the lever arm is in its *down* position and *vice versa*. Therefore, it seems likely that the myosin power and respectively return stroke are produced by switching between these two conformations.

A schematic summary of the structural states observed in different myosins together with their corresponding positions of switch-1, switch-2, and the lever arm as well as their corresponding actin affinity is shown in Fig. 4.12. How these structural states can be mapped along the ATPase cycle of actomyosin is shown in Fig. 4.13.

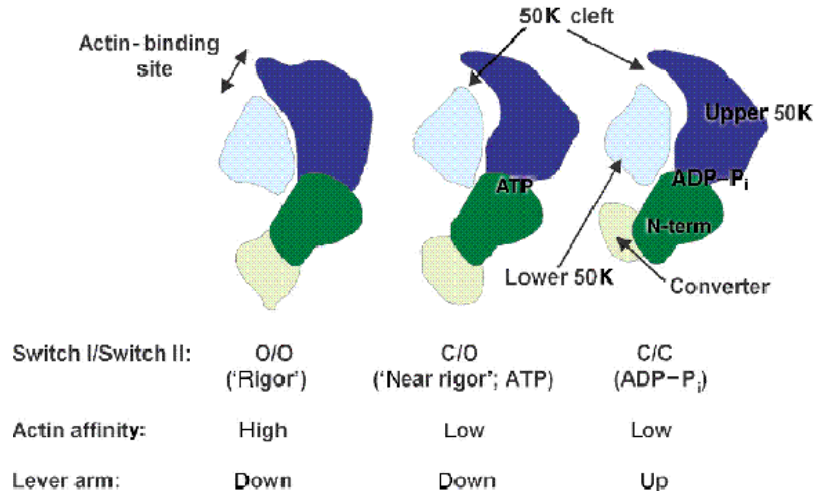


Fig. 4.12 Proposed subdomain rearrangements in the myosin motor domain and corresponding switch-1, switch-2, and the lever arm positions as well as actin affinity. The figure was taken from Ref. 35.

As it is shown in the above description, a clear understanding of the action mechanism of myosin II does not exist to date, in spite of the large amount of information coming from both experimental data and theoretical studies. In the following chapters, theoretical studies of myosin II dynamics are performed with the aim of identifying the essential motions in the function of this protein. The results obtained analyzing different MD trajectories of the two existent crystal structures of

the myosin head (i.e., the OPEN and CLOSED conformations) are present in Chapter 5. The essential motions of myosin are extracted out of these trajectories by performing PCA and then calculating the corresponding Involvement Coefficients, and they are confirming previously postulated mechanochemical couplings regarding a possible communication mechanism between the active sites.

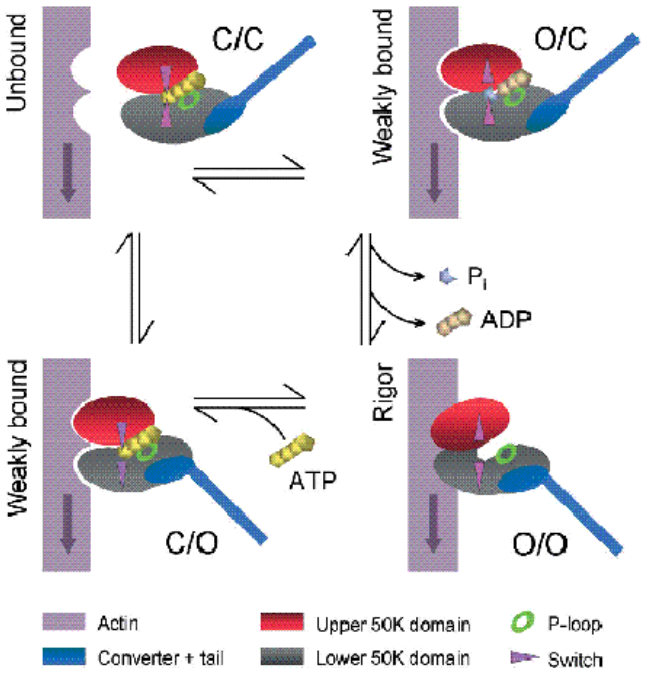


Fig. 4.13 Structural model for the actin-activated myosin II ATPase cycle. The figure was taken from Ref. 55.

References

1. Schliwa M. - Molecular motors, Wiley-VCH, Weinheim, 2003.
2. Schliwa M. & Woehlke G. - Molecular motors, *Nature* **422**, 759-765, 2003.
3. Vale R. D. - The molecular motor toolbox for intracellular transport, *Cell* **112**, 467-480, 2003.
4. Oster G. & Wang H. - Reverse engineering a protein: the mechanochemistry of ATP synthase, *Biochim. Biophys. Acta* **1458**, 482-510, 2000.
5. Diez M., Zimmermann B., Bőrsch M., Kőnig M., Schweinberge E., Steigmiller S., Reuter R., Felekyan S., Kudryavtsev V., Seidel C. A. M. & Gräber P. - Proton-powered subunit rotation in single membrane-bound F_oF_1-ATP synthase, *Nature Struct. Mol. Biol.* **11**, 135-141, 2004.
6. Itoh H., Takahashi A., Adachi K., Noju H., Yasuda R., Yoshida M. & Kinosita Jr. K. - Mechanically driven ATP synthesis by F_1-ATPase, *Nature* **427**, 465-458, 2004.
7. Ma J., Flynn T. C., Cui Q., Leslie A. G. W., Walker J. E. & Karplus M. - A dynamic analysis of the rotation mechanism for conformational change in F_1ATPase, *Structure* **10**, 921-931, 2002.
8. Antes I., Chandler D., Wang H. & Oster G. - The unbinding of ATP from F_1-ATPase, *Biophys. J.* **85**, 695-706, 2003.
9. Gao Y. Q., Yang W., Marcus R. A. & Karplus M. - A model for the cooperative free energy transduction and kinetics of ATP hydrolysis by F_1-ATPase, *Proc. Natl. Acad. Sci.* **100**, 11339-11344, 2003.
10. Sellers J. R. - Myosins: a diverse superfamily, *Biochim. Biophys. Acta* **1496**, 3-22, 2000.
11. Berg J. S., Powell B. C. & Cheney R. E. – A millennial myosin census, *Mol. Biol. Cell* **12**, 780-794, 2001.
12. Alberts B., Johnson A., Lewis J., Raff M., Roberts K. & Walter P. - Molecular Biology of the Cell, *Garland Science*, 29 West 35th Street, New York, NY 10001-2299, 2002.

13. Barylko B., Binns D. D. & Albanesi J. P. - Regulation of the enzymatic and motor activities of myosin I, *Biochim. Biophys. Acta* **1496**, 23-35, 2000.
14. Fujita-Becker S., Dürrwang U., Erent M., Clark R. J., Geeves M. A. & Manstein D. J. – Changes in Mg^{2+} -ion concentration and heavy chain phosphorylation regulate the motor activity of a class-I myosin.
15. Geeves M. A., Perreault-Micale C. & Coluccio L. M. - Kinetic analyses of a truncated mammalian myosin I suggest a novel isomerization event preceding nucleotide binding, *J. Biol. Chem.* **275**, 21624-21630, 2000.
16. Jontes J. D. & Milligan R. A. – Brush border myosin-I structure and ADP-dependent conformational changes revealed by cryelectron microscopy and image analysis, *J. Cell. Biol.* **139**, 683-693, 1997.
17. Ostap M. E. & Pollard T. D. - Biochemical kinetic characterization of the *Acanthamoeba* myosin-I ATPase, *J. Cell. Biol.* **132**, 1053-1060, 1996.
18. Veigel C., Coluccio L. M., Jontes J. D., Sparrow J. C., Milligan R. A. & Molloy J. E. - The motor protein myosin-I produces its working stroke in two steps, *Nature* **398**, 530-533, 1999.
19. Bähler M. - Are class III and class IX myosins motorized signaling molecules?, *Biochim. Biophys. Acta* **1496**, 52-59, 2000.
20. Uemura S., Higuchi H., Olivares A. O., De la Cruz E. M. & Ishiwata S. – Mechanochemical coupling of two substeps in a single myosin V motor, *Nature Struct. Mol. Biol.* **11**, 877-883, 2004.
21. Reck-Peterson S., Provance Jr. D. W., Mooseker M. S. & Mercer J. A. - Class V myosins, *Biochim. Biophys. Acta* **1496**, 36-51, 2000.
22. Tang F., Kauffman E. J., Novak J. L., Nau J. J., Catlett N. L. & Weisman L. S. - Regulated degradation of a class V myosin receptor directs movement of the yeast vacuole, *Nature* **422**, 87-92, 2003.
23. Altmann D., Sweeney H. L. & Spudich J. A. - The mechanism of myosin VI translocation and its load-induced anchoring, *Cell* **116**, 737-749, 2004.
24. Kieke M. C. & Titus M. A. - The myosin superfamily: an overview, In Manfred Schliwa, editor, *Molecular motors,* 3-44, Wiley-VCH, Weinheim, 2003.

25. Molloy J. E. & Veigel C. - Myosin motors walk the walk, *Science* **300**, 2045-2046, 2003.
26. Wells A. L., Lin A. W., Li-Qiong Chen, Safer D., Cain S. M., Hasson T., Carragher B. O., Milligan R. A. & Sweeney H. L. - Myosin VI is an actin-based motor that moves backwards, *Nature* **401**, 505-508, 1999.
27. Schliwa M. - Myosin steps backwards, *Nature* **401**, 431-432, 1999.
28. Yildiz A., Forkey J. N., McKinney S. A., Ha T., Goldman Y. E. & Selvin P. R. - Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization, *Science* **300**, 2061-2065, 2003.
29. Kull F. J., Sablin E. P., Lau R., Fletterick R. J. & Vale R. D. - Crystal structure of the kinesin motor domain reveals a structural similarity to myosin, *Nature* **380**, 550-559, 1996.
30. http://training.seer.cancer.gov/module_anatomy/unit4_2_muscle_structure.html
31. Huxley A. F. & Niedergerke R. M. - Structural Changes in Muscle During Contraction: Interference Microscopy of Living Muscle Fibers, *Nature* **173**, 971-973, 1954.
32. Huxley A. F. & Hanson J. - Changes in the Cross-Striations of Muscle during Contraction and Stretch and their Structural Interpretation, *Nature* **173**, 973-976, 1954.
33. Geeves M. A. & Holmes K. C. - Structural mechanism of muscle contraction, *Ann. Rev. Biochem.* **68**, 687-728, 1999.
34. Holmes K. C. & Geeves M. A. - The structural basis of muscle contraction, *Philos. Trans. R. Soc. B* **355**, 419-431, 2000.
35. Goody R. S. - The missing link in the muscle cross-bridge cycle, *Nature Struct. Biol.* **10**, 773-775, 2003.
36. Reconditi M., Linari M., Lucii L., Stewart A., Yin-Biao Sun, Boesecke P., Narayanan T., Sischetti R. F., Irving T., Piazzesi G., Irving M. & Lombardi V. - The myosin motor in muscle generates a smaller and slower working stroke at higher load, *Nature* **428**, 578-581, 2004.

37. http://ca.expasy.org/cgi-bin/niceprot.pl?P08799, SwissProt entry for *Dictyostelium discoideum* myosin II heavy chain (accession number P08799).
38. Mornet D., Pantel P., Audemard E. & Kassab R. - The limited tryptic cleavage if chymotryptic S-1: an approach to the characterization of the actin site in myosin heads, *Biochem. Biophys. Res. Comm.* **89**, 925-932, 1979.
39. Itakura S., Yamakawa H., Toyoshima Y. Y., Ishijima A., Kojima T., Harada, T. Yanagida Y., Wakabayashi T. & Sutoh K. - Force-generating domain of myosin motor, *Biochem. Biophys. Res. Comm.* **196**, 1504-1510, 1993.
40. Joel P. B., Sweeney H. L. & Trybus K. M. - Addition of lysines to the 50/20 kDa junction of myosin strengthens weak binding to actin without affecting the maximum ATPase activity, *Biochemistry* **42**, 9160-9166, 2003.
41. Sasaki N., Ohkura R. & Sutoh K. - *Dictyostelium* myosin II as a model to study the actin-myosin interactions during force generation, *J. Muscle Res. Cell Mot.*, **23,** 697-702, 2002.
42. Uyeda T. Q. P., Patterson B., Mendoza L. & Hiratsuka Y. - Amino acids 519-524 of *Dictyostelium* myosin II form a surface loop that aids actin binding by facilitating a conformational change, *J. Muscle Res. Cell Mot.* **23**, 685-695, 2002.
43. Houdusse A., Kalabokis V. N., Himmel D., Szent-Györgyi A. & Cohen C. - Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head, *Cell* **97,** 459-470, 1999.
44. Anson M., Michael A. Geeves S. E. Kurzawa & Manstein D. J. - Myosin motors with artificial lever arms, *EMBO J.* **15**, 6069-6074, 1996.
45. Seidel J. C. - The effects of nucleotides and Mg^{2+} on the electron spin resonance spectra of myosin spin lebeled at the S_2 thiol groups, *Arch. Biochem. Biophys.* **152**, 839-848, 1972.
46. Seidel J. C., Chopek M. & Gergely J. - Effects of nucleotides and pyrophosphate on apin labels bound to S_1 thiol groups of myosin, *Biochemistry* **9**, 3265-3272, 1970.

47. Smith C. A. & Rayment I. - Active site comparisons highlight structural similarities between myosin and other P-loop proteins, *Biophys. J.* **70**, 1590-1602, 1996.
48. Fisher A. J., Smith C. A., Thoden J., Smith R., Sutoh K., Holden H. M. & Rayment I. - X-ray Structures of the Myosin Motor Domain of *Dictyostelium discoideum* Complexed with MgADP.BeF_x and MgADP.AlF^-_4, *Biochemistry* **34**, 8960-8972, 1995.
49. Kull F. J., Schlichting I., Becker A., Kollmar M., Manstein D. J. & Holmes K. – An alternate conformation for MgADP-berilium fluoride bound myosin II.
50. Park S., Ajtai K. & Burghardt T. P. - Inhibition of myosin ATPase by metal fluoride complexes, *Biochim. Biophys. Acta.* **1430**, 127-140, 1999.
51. Kliche W., Fujita-Becker S., Kollmar M., Manstein D. J. & Kull F. J. - Structure of a genetically engineered molecular motor, *EMBO J.* **20**, 40-46, 2001.
52. Bauer C. B., Kuhlman P. A., Bagshaw C. R. & Rayment I. - X-ray crystal structure and solution fluorescence characterization of Mg.2'(3')-O-(N-methylanthraniloyl) nucleotides bound to the *Dictyostelium discoideum* myosin motor domain, *J. Mol. Biol.* **274**, 394-407, 1997.
53. Gulick A. M., Bauer C. B., Thoden J. B. & Rayment I. - X-ray Structures of the MgADP, MgATPγS, and MgAMPPNP Complexes of the *Dictyostelium discoideum* Myosin Motor Domain, *Biochemistry* **36**, 11619-11618, 1997.
54. Smith C. A. & Rayment I. - X-ray Structure of the Magnesium(II)-Pyrophosphate Complex of the Truncated Head of *Dictyostelium discoideum* Myosin to 2.7 Å Resolution, *Biochemistry* **34**, 8973-8981, 1995.
55. Reubold T., Eschenburg S., Becker A. , Kull F. J. & Manstein D. J. - A structural model for actin-induced nucleotide release in myosin, *Nature Struct. Biol.* **10**, 826-830, 2003.
56. Smith C. A. & Rayment I. - X-ray Structure of the Magnesium(II).ADP.Vanadate Complex of the *Dictyostelium discoideum* Myosin Motor Domain to 1.9 Å Resolution, *Biochemistry* **35**, 5404-5417, 1996.

57. Schwarzl S. M. – Understanding the ATP Hydrolysis mechanism in Myosin using Computer Simulation Techniques, Mensch & Buch Verlag, ISBN 3-86664-044-7, 2006.
58. Risal D., Gourinath S., Himmel D. M., Szent-Györgyi A. G. & Cohen C. - Myosin subfragment 1 structures reveal a partially bound nucleotide and a complex salt bridge that helps couple nucleotide and actin binding, *Proc. Natl. Acad. Sci.* **101**, 8930-8935, 2004.
59. Lawson J. D., Pate E., Rayment I. & Yount R. G. - Molecular dynamics analysis of structural factors influencing back door Pi release in myosin, *Biophys. J.* **86**, 3794-3803, 2004.
60. Goldman Y. E. - Kinetics of the actomyosin {ATPase} in muscle fibers, *Ann. Rev. Physiol.* **49,** 637-654, 1987.
61. Trentham D. R., Eccleston J. F. & Bagshaw C. R. - Kinetic analysis of ATPase mechanisms, *Q. Rev. Biophys.* **9,** 217-281, 1976.
62. Rayment I., Rypniewski W. R., Schmidt-Bäse K., Smith R., Tomchick D. R., Benning M. M., Winkelmann D. A., Wesenberg G. & Holden H. M. - Three-dimensional structure of myosin subfragment-1: a molecular motor, *Science* **261**, 50-58, 1993.
63. Houdusse A., Szent-Györgyi A. G. & Cohen C. - Three conformational states of scallop myosin S1, *Proc. Natl. Acad. Sci.* **97**, 11238-11234, 2000.
64. Fischer S., Windshügel B., Horak D., Holmes K. C. & Smith J. C. – Structural mechanism of the recovery stroke in the Myosin molecular motor, *Proc. Natl. Acad. Sci*. **102**, 6873-6878, 2005.
65. Conibear P. B., Bagshaw C., Fajer P. G., Kovács M. & Málnási-Csizmadia A. - Myosin cleft movement and its coupling to actomyosin dissociation, *Nature Struct. Biol.* **10**, 831-835, 2003.
66. Himmel D. M., Gourinath S., Reshetnikova L., Shen Y., Szent-Györgyi A. G. & Cohen C. - Crystallographic findings on the internally uncoupled and near-rigor states of myosin: further insights into the mechanics of the motor, *Proc. Natl. Acad. Sci.* **99**, 12645-12650, 2002.

67. LaConte L. E. W., Baker J. E. & Thomas D. D. - Transient kinetics and mechanics of myosin's force-generating rotation in muscle: Resolution of millisecond rotational transitions in the spin-labeled myosin light-chain domain, *Biochemistry* **42**, 9797-9803, 2003.
68. Gulick A. M., Bauer C. B., Thoden J. B., Pate E., Yount R. G. & Rayment I. - X-ray structures of the *Dictyostelium discoideum* myosin motor domain with six non-nucleotide analogs, *J. Biol. Chem.* **275**, 398-408, 2000.
69. Bauer C. B., Holden H. M., Thoden J. B., Smith R. and Rayment I. - X-ray structures of the Apo and MgATP-bound states of *Dictyostelium discoideum* myosin motor domain, *J. Biol. Chem.* **275**, 38494-38499, 2000.

Chapter 5

Analyzing Large-Scale Structural Changes in Proteins by Reducing the Dimensionality

In the present chapter are presented results from a Principal Component Analysis applied on MD trajectories of the OPEN and CLOSED states of myosin. The aim of the study was to explore the dynamics of these two conformations that occur along the myosin cycle and to identify essential motions that contribute to the mechanics of the coupling between the ATP hydrolysis and active movement of myosin on the actin filaments. The results confirmed the previously proposed model of myosin activation during the recovery stroke and provided atomic detailed description of the most important domain motions contributing to the mechanism of the recovery stroke. Together with a thorough description of the structural details of the recovery stroke mechanism, the results represent a step forward towards understanding of how chemical energy released by ATP hydrolysis is transformed into mechanical energy at structural level, thereby highlighting the importance of simulation methods in the investigation of complex biological processes. The results are presented below as they were accepted for publication in *Journal of Molecular Biology*, 2007.

5.1 Abstract

Muscle contraction is driven by a cycle of conformational changes in the myosin II head. After myosin binds ATP and releases from the actin fibril, myosin prepares for the next power stroke by rotating back the converter domain that carries the lever arm by 60°. This recovery stroke is coupled to the activation of myosin's ATPase by a mechanism that is essential for an efficient motor cycle. The mechanics of this coupling have been proposed to occur via two distinct and successive motions of the two helices that hold the converter domain: in a first phase a see-saw motion of the relay helix, followed by a piston motion of the SH1 helix in a second phase. To test this model, we have determined the principal motions of these structural elements during equilibrium molecular dynamics simulations of the crystallographic end states of the recovery-stroke by using Principal Component Analysis. This reveals that the only principal motions of these two helices that make a large amplitude contribution towards the conformational change of the recovery stroke are indeed the predicted seesaw and piston motions. Moreover, the results demonstrate that the seesaw motion of the relay helix dominates in the dynamics of the pre-recovery stroke structure, but not in the dynamics of the post-recovery stroke structure, and *vice versa* for the piston motion of the SH1 helix. This is consistent with the order of the proposed two-phase model for the coupling mechanism of the recovery stroke.

5.2 Introduction

At the molecular level, muscle contraction is generated by cyclic interactions of myosin heads with actin filaments.[1-3] This is fueled by ATP hydrolysis and involves large conformational changes in the myosin head, which contains all three functional units: the ATP and actin binding sites, and the "converter/lever arm" domain. The sequence of interactions between actin, myosin and ATP that leads to the production of mechanical force is described by the Lymn-Taylor cycle[4] (see Fig. 5.1).

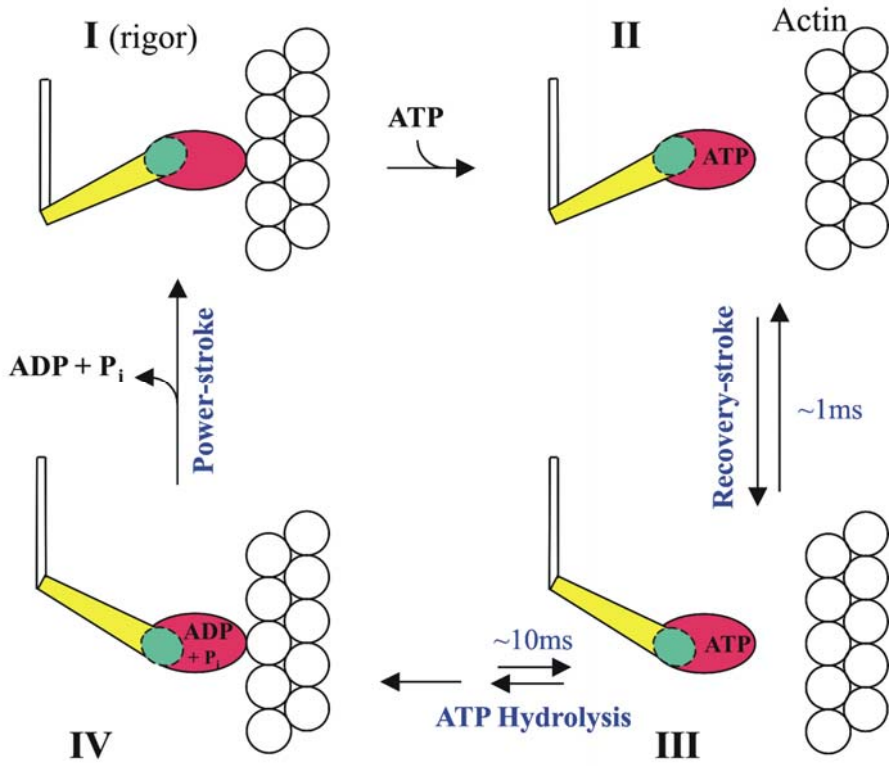


Fig. 5.1 Lymn-Taylor cycle. The structural domains of myosin are: myosin head (red); lever arm (yellow); converter domain (green). The actin filament is shown as white spheres.

In this cycle, myosin first binds strongly to actin in the absence of ATP (the rigor conformation, State I in Fig. 5.1). ATP binding leads to the dissociation of myosin from actin (State II). Myosin then undergoes a large reversible transition (the "recovery stroke") that brings the lever arm in the pre-power stroke orientation and activates the ATPase function (State III). After ATP hydrolysis, myosin rebinds to actin (State IV) and performs the "power stroke" to return back to the rigor state. An essential requirement for the Lymn-Taylor cycle to function is that myosin must couple small changes in the catalytic ATPase site with large conformational changes in the actin-binding and the converter domains. This relies on well-defined communication mechanisms that ensure that these changes are correlated in the protein so as to efficiently produce mechanical work. The current study is realized on the communication pathway responsible for passing structural information between

the ATPase site and the 40 Å distant converter domain during the recovery-stroke (State II → State III).

Myosin II has been crystallized with different ATP analogues under various conditions[5-7] in the absence of actin and was found in two conformations that were assigned to State II and State III of the Lymn-Taylor cycle. The largest structural difference between these states is in the orientation of the converter/lever-arm domain, which is rotated by ~60° relative to the rest of myosin head (see Fig. 5.2A and D). Another difference is that the ATP binding site is partially open in the pre-recovery conformation (Fig. 5.2C), while it is closed by the Switch-2 loop in the post-recovery state (Fig. 5.2F), thereby switching on the catalytic ATPase function.[1] To understand how the structural information is passed between the Switch-2 loop and the converter domain in a reversible manner (the recovery stroke is reversible in presence of ATP), it is necessary to have atomic structures of intermediates along this large conformational transition. Unfortunately, this information is not easily obtained experimentally. An alternative is to use the crystallographic structures of the transition end points to compute pathways over the protein energy landscape. This approach was taken recently and led to a structural model that describes the coupling between the closing of Switch-2 and the rotation of the converter domain.[8] This mechanism is initiated (see Fig. 5.2B) by a movement of Gly457 on Switch2 towards ATP (to make a hydrogen bond with the γ–phosphate) which is transmitted as a pull on the so called relay helix (residues 466-498) through a hydrogen bond between the Gly457 peptide group and the side chain of Asn475 located on the N-terminal half of the relay helix. This pull causes a "seesaw"-like motion of the relay helix (which pivots around Phe652) transmitted to its C-terminal end where it is connected to the converter domain, which reacts with a rotation of ~25° (see reference 8 for more details). After this "seesaw" phase, the second part of the coupling between Switch-2 motion and further rotation of the converter domain is due to a piston-like translation of the "SH1 helix" (residues 681-691). The SH1 helix is the second helix to which the converter domain is attached, its motion causing a further 40° rotation of the converter domain, see Fig. 5.2E.[9] The SH1 helix translates in response to a wedging against its N-terminal end by loop 572-574. This "wedge loop" moves to accompany the final

closing of Switch-2 because the side chain of Phe458 (which forms a hydrogen bond with Ser181 on the P-loop) is making tight hydrophobic interactions with the wedge-loop (for ex. with Tyr573, see Fig. 5.2E). Thus, in the above model of the coupling mechanism, the rotation of the converter domain is controlled in two phases by successive motions of the two helices that hold the converter domain (i.e., the relay helix and the SH1-helix), involving first a seesaw motion of the relay helix, followed by a combined wedge/piston motion of the wedge-loop/SH1-helix.

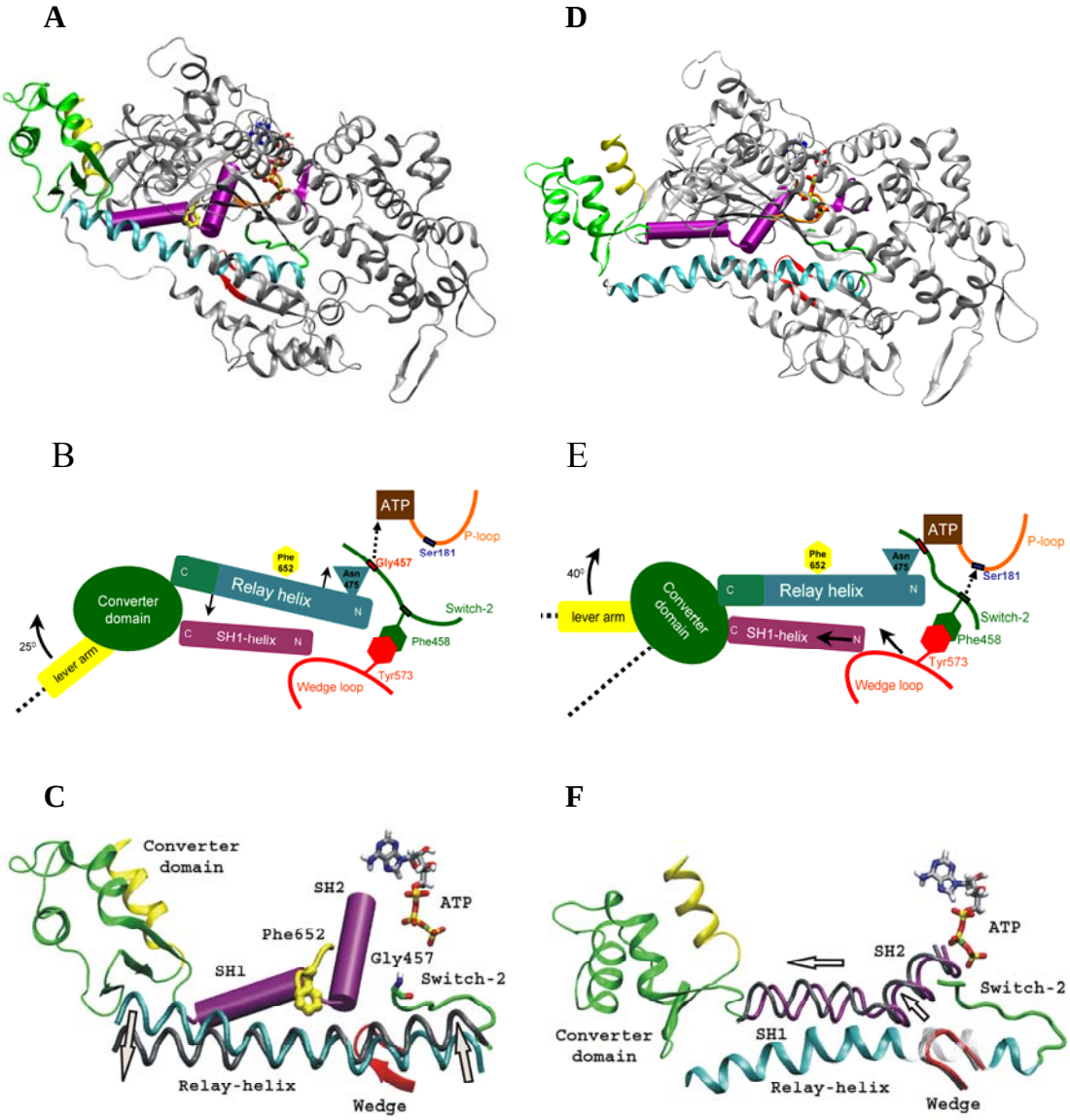


Fig. 5.2 The two phases of the recovery stroke mechanism. A), B) and C): The first phase. D), E) and F): The second phase. In all panels, the relay helix is colored in cyan; SH1 and SH2 helices in

purple; converter domain in green; lever arm in yellow; P-loop in orange; Switch-1 loop in purple; Switch-2 loop in green; Wedge loop in red. **A)** The pre-recovery conformation (State II in Fig. 5.1).[6] **D)** The post-recovery conformer (State III).[7] **B)** The coupling mechanism during phase I: Gly457 moves towards ATP (dotted arrow), pulling the relay helix, which reacts by a seesaw motion (shown as solid straight arrows at the C and respectively N terminal ends of the relay helix) that is accompanied by an initial 25° rotation of the converter/lever-arm domain. **E)** The coupling mechanism during phase II: Phe458 moves towards the P-loop (dotted arrow) and pulls the Tyr573-loop, which wedges against the N-terminal end of the SH1-helix that moves piston-like (solid straight arrows) to provoke a further 40° rotation of the converter/lever-arm domain. **C)** The seesaw motion of the relay helix during MD of State II (PC_2 in Figure 5.4B): Starting position of the helix in cyan; final position in gray (the arrows indicate the seesaw movement, also shown as molecular movie). **F)** The wedge/piston motion during MD of State III (PC_3 in Fig. 5.6D): Starting position of the wedge-loop in red and of the SH1/SH2 helices in purple; final positions in grey (the arrows indicate the wedge/piston movement also shown as molecular movie under http://www.iwr.uni-heidelberg.de/groups/biocomp/fischer).

The goal of the present analysis is to check this coupling model by inspecting the motions of the main structural elements involved in it (i.e., the converter domain, the relay and the SH1 helices, and the wedge loop) during equilibrium dynamics of the protein in the crystallographic end-states of the recovery stroke, and to examine whether/how these motions contribute to the recovery-stroke transition. If the coupling model is correct, these structural elements can be expected to exhibit coherent oscillatory motions of significant amplitude that correspond to those predicted in the model. Moreover, if the coupling mechanism occurs in two successive phases with distinct motions, the relevant motions in the dynamics of the pre- and post-recovery states will differ and correspond respectively to the predicted motions of the first and second phases of the model.

Molecular Dynamics (MD) simulations were performed, of the fully solvated myosin II head of *Dictyostelium discoideum* (henceforth 'myosin') in the two crystallographic conformations of State II and III. The protein conformers generated by each MD simulation were analyzed using Principal Component Analysis (PCA). PCA is a data analysis tool that allows to characterize the deformations of largest

amplitude that occur in a sample of conformers[10-12] It has been used to analyze the conformations generated in a variety of protein MD's (see, e.g., references 13-25). PCA generates a complete and orthogonal set of atomic displacement vectors, each vector corresponding to a particular type of correlated motion of all atoms with their individual amplitudes (see Methods section). Each vector describes what is called a principal motion (or principal component) in the molecule. In proteins, the principal motions with largest amplitude correspond to collective motions, which can be functionally important.[25,26] The principal motions of individual domains or sub-fragments of a protein can be separated from the overall protein motions by performing the PCA on only the atoms of the fragment. This was done here to obtain the principal motions of the structural elements implicated in the coupling model. The principal motions most involved in the recovery stroke transition were identified by computing their Involvement Coefficients[27,28], which measures the contribution from a particular principal motion towards a conformational transition of interest (see Methods). The results reveal that only two principal motions of large amplitude have a dominant Involvement Coefficient. In one, the relay helix exhibits precisely a "seesaw" motion of large amplitude during the MD simulation of the pre-recovery conformation (State II, see Fig. 5.2C). This "seesaw" motion is not seen to be dominant in the dynamics of the post-recovery stroke conformation (State III), which is consistent with the idea that this motion initiates the mechanism of the recovery-stroke. The other dominant principal motion is the wedge/piston motion of wedge loop/SH1-helix, which is found in the MD simulation of State III, but not in State II, again confirming the model according to which this motion is taking place during the second phase of the recovery stroke. The present results thus provide strong evidence in favor of the proposed mechanics of the recovery stroke and the related coupling model.

5.3 Results

5.3.1 Converter domain rotation

The converter domain rotates by 60-65° between the crystallographic end-states of the recovery stroke. To test whether the present PCA approach can detect some partial amount of this rotation in the dynamic fluctuations of the converter domain, PCA was performed on a sub-fragment of the protein consisting mostly of the converter domain, as well as the relay and SH1 helices. The resulting Principal Component (PC) vectors were each projected onto the difference vector between the two end-state structures. This projection is called the Involvement Coefficient (I_k, see Methods) and identifies the principal motions that are precursors of a larger conformational change. When doing this for the conformational sample obtained from the MD of State II, the first PC vector has a significantly larger Involvement Coefficient (I_1^2=34%) than the other PC vectors (Fig. 5.3A). Note that the sum over all I_k^2 converges to 1 (see inset of Fig. 5.3A), so that I_k^2 can be considered as the relative contribution from a given PC_k (i.e., the atomic motions along PC_1 contribute ~34% of the total conformational variance). Because the first PC is also (by definition) the PC of largest amplitude, the amplitude weighted Involvement Coefficient is even larger for PC_1 than for the other PC's (not shown). Thus the atomic motions of PC_1 clearly participate in the motion of the converter region in the early stages of the recovery stroke. These atomic motions are shown in Fig. 5.3B. They consist in a partial rotation of the converter domain (by ~10 degrees) around the axis of the SH1 helix. This rotation is accompanied by a translation of the C-terminus of the relay helix.

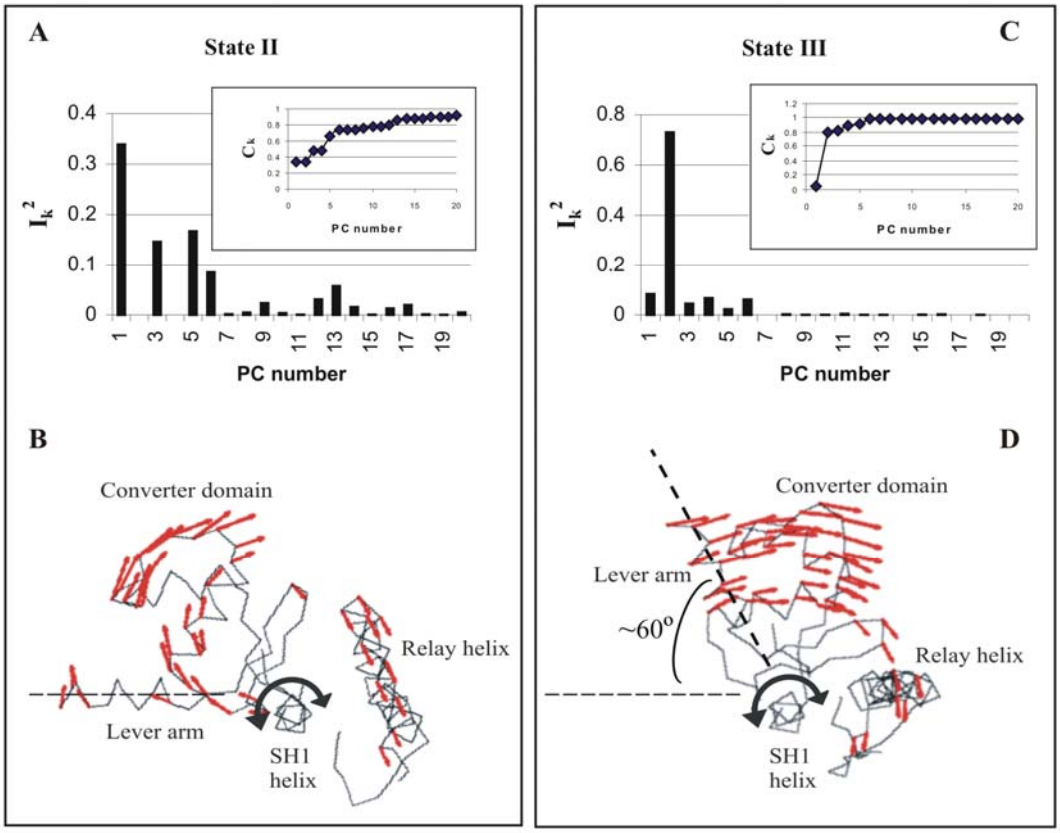


Fig. 5.3 Principal Component Analysis of the converter domain and the SH1 and relay helices. A) and B): during MD of pre-recovery stroke State II. C) and D): During MD of post-recovery stroke State III. A) and C): Squared Involvement Coefficients (I_k^2) and the Cumulative Involvement Coefficients (C_k, inset) of the first 20 Principal Component (PC) eigenvectors, see methods section. B) and D): Atomic displacements (red arrows) in the PC with the dominant Involvement Coefficient (PC_1 in panel B and PC_2 in panel D). Drawing generated by MOLSCRIPT.[29]

The natural amplitude in the MD at room temperature of this concerted motion of relay helix and converter domain in PC_1 can best be seen in a molecular movie, available under http://www.iwr.uni-heidelberg.de/groups/biocomp/fischer. It shows that the motions of PC_1 correspond to the expected rotation of the converter domain and swinging of the lever-arm. Thus, the present approach of selecting the principal motion with a dominant Involvement Coefficient was able to identify the functional motion of the converter domain during the recovery stroke. Applying this approach to

State III is equally successful: The Involvement Coefficients of the PC's obtained from the PCA of the MD of the post-recovery stroke conformation are plotted in Fig. 5.3C. Again, choosing the PC with the largest I_k^2 (PC_2, I_2^2=73%) and analyzing its atomic displacements (shown in Fig. 5.3D, and also available as molecular movie) reveals a motion of the converter region that corresponds to its expected rotation in State III.

PCA of the converter region was also performed on the MD of the pre-recovery conformation without ATP, or in the MD of the post-recovery conformation with ADP·Pi bound. It is interesting to note that, in these two states which are not the natural end-states of the recovery stroke, no single PC involving converter rotation could be captured by a dominant Involvement Coefficient (see Fig. 5.4). This would be consistent with the fact that ATP is not yet hydrolyzed during the Lymn-Taylor recovery-stroke, but could also be do to a slower convergence in these particular trajectories and will require further investigation.

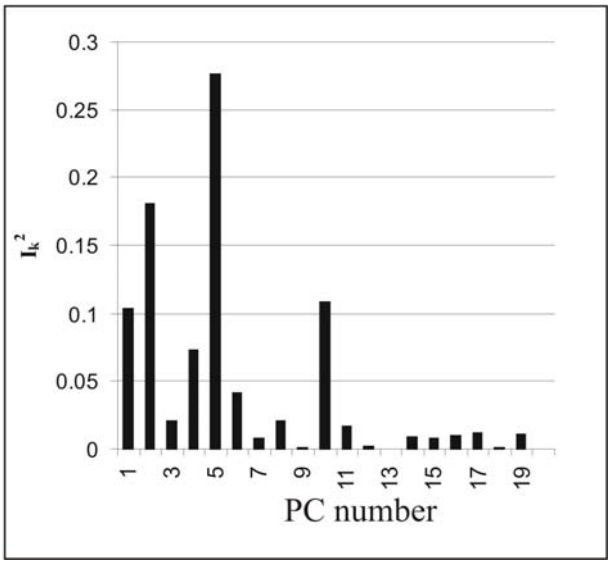


Fig. 5.4 Squared Involvement Coefficients for the MD of the post recovery state with bound ADP·Pi, PCA done on the converter domain plus SH1 and relay helices.

5.3.2 Seesaw motion of the relay helix

To obtain more detailed information about the motion of local elements that may be involved in transmitting and/or amplifying conformational changes, the PCA was focused onto the relay helix. The role of the relay helix has been proposed to be especially important in the first half of the recovery stroke[8], where it is thought to undergo a seesaw-like motion in response to the movement of the Gly457/Ser456 peptide group toward the γP (pulling on Asn475 of the relay helix, see the schematic view of this event in Fig. 2B). Gly457 forms an H-bond with the γP that is essential for the myosin ATPase function.[30] Because the seesaw motion leads to a translation of the C-terminal end of the relay helix, where it is connected to the converter domain, it is a key element of the coupling mechanism between converter rotation and ATPase activation. One might expect to find small oscillatory displacements along this seesaw motion in the MD of State II.

The results of the PCA of the dynamics of the relay helix are shown in Fig. 5.5. In State II, the first principal component, which captures ~60% of the total fluctuations (Fig. 5.5A), is a rather incoherent motion of the relay helix, which has a relatively small involvement in the recovery stroke transition (I_1^2=18%, Fig. 5.5B). In contrast, the second principal component (PC_2) has a dominant Involvement Coefficient (I_2^2=53%). The atomic motions in PC_2 correspond to a coherent seesaw motion of the relay helix (see the inset in Fig. 5.5B), in which the atoms at one end of the helix swing in the direction opposite to the direction of the atoms at the other end, while the stationary point of the relay helix is located in the middle of the N-terminal half of the helix, where it is in contact with Phe652. Phe652 (colored yellow in Fig. 5.2A, B and C) is the pivoting point for the relay helix seesaw movement of PC_2 (see Fig. 5.2C and the molecular movie). It is located on the central β–sheet of the main body and is tightly interlocked with relay helix residues Phe481 and Phe482, thus forming the molecular fulcrum of the see-saw.[8]

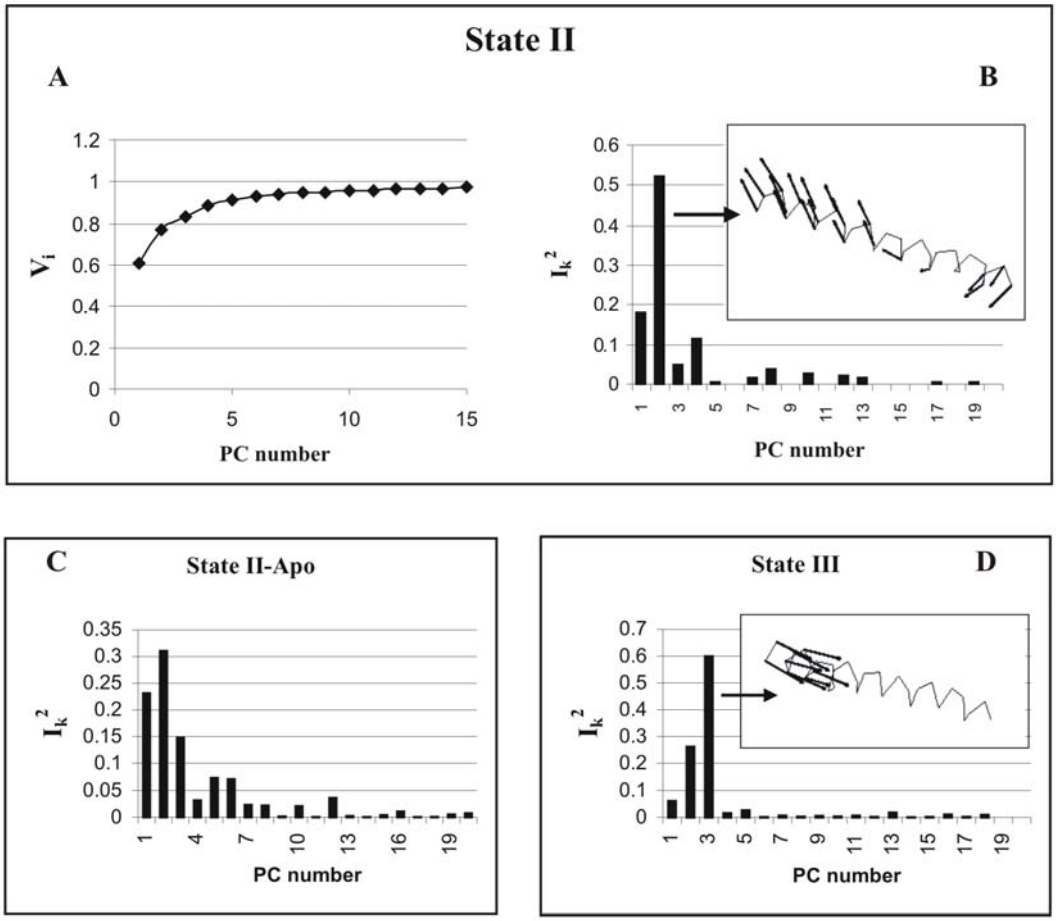


Fig. 5.5 PCA of the relay helix A) and B): During the MD of the pre-recovery State II. **A)** Cumulative sum of the covariance content (V_i, equation 5.5) in the first 15 PC eigenvectors. **B)** The Squared Involvement Coefficients (I_k^2). The atomic motions of PC_2 are shown as arrows in the inset. **C)** I_k^2 obtained for the MD of the pre-recovery conformer in absence of ATP (Apo-State II). **D)** I_k^2 obtained for the MD of the post-recovery State III. The atomic motions of PC_3 are shown in the inset.

The correlation during the MD of State II between the seesaw motion of the relay helix and the rotation of the converter domain is clearly visible when comparing the evolution in time of the variation in the amplitude of these two motion. Fig. 5.6A shows the amplitude *p(t)* of the seesaw displacements of the relay helix along the Principal Component PC_2. *p(t)* is obtained by projecting a given vector L_k (here PC_2) onto the displacement vector taken at different times of the MD trajectory:

$$p(t) = \frac{L_k}{|L_k|}(X(t) - X(0)) \tag{5.1}$$

where *X(t)* is the coordinate-set vector at time *t* along the MD trajectory and *X(0)* is the starting coordinate frame at *t* = 0. The dot-product in Eq. 5.1 is taken over the desired selection of atoms, for example in Fig. 5.6A only the atoms of the relay helix. Figure 5.5B shows the corresponding amplitude changes *p(t)* for the converter domain with L_k taken as PC_1 of the MD of State II (shown in Fig. 5.3A). Comparing Fig. 5.6A and 6B shows that the amplitudes of the relay helix seesaw and the converter domain rotation tend to co-vary along the trajectory. This correlation is more apparent when plotting these two amplitude changes against each other (Fig. 5.6C), yielding a correlation coefficient of 0.75. This shows that the seesaw motion of the relay helix and the rotation of the converter domain are clearly coupled in State II, consistent with the predictions of the model for the first phase of the recovery stroke.

In a PCA of the post-recovery conformer with bound ATP (State III), the relevant motion is that of PC_3 (I_3^2=60%, Fig. 5.5D). This motion corresponds to a breaking of the C-terminal end of the relay helix (see inset of Fig. 5.5D). This is consistent with the formation of the "kink" in the relay helix seen in the crystal structure of post-recovery structure, as the relay helix locally unwinds in the second phase of the recovery stroke.[8] This unwinding is due to the fact that the whole C-terminal third of the relay helix is embedded in the converter domain and accompanies converter rotation. To accommodate this rotation, the helix breaks where it is embedded neither in the converter domain nor in the main body of myosin (shown in gray in Fig. 5.2A and D). This suggests that relay helix unwinding is rather a consequence than a cause of the converter domain rotation that occurs during the second phase of the recovery stroke.

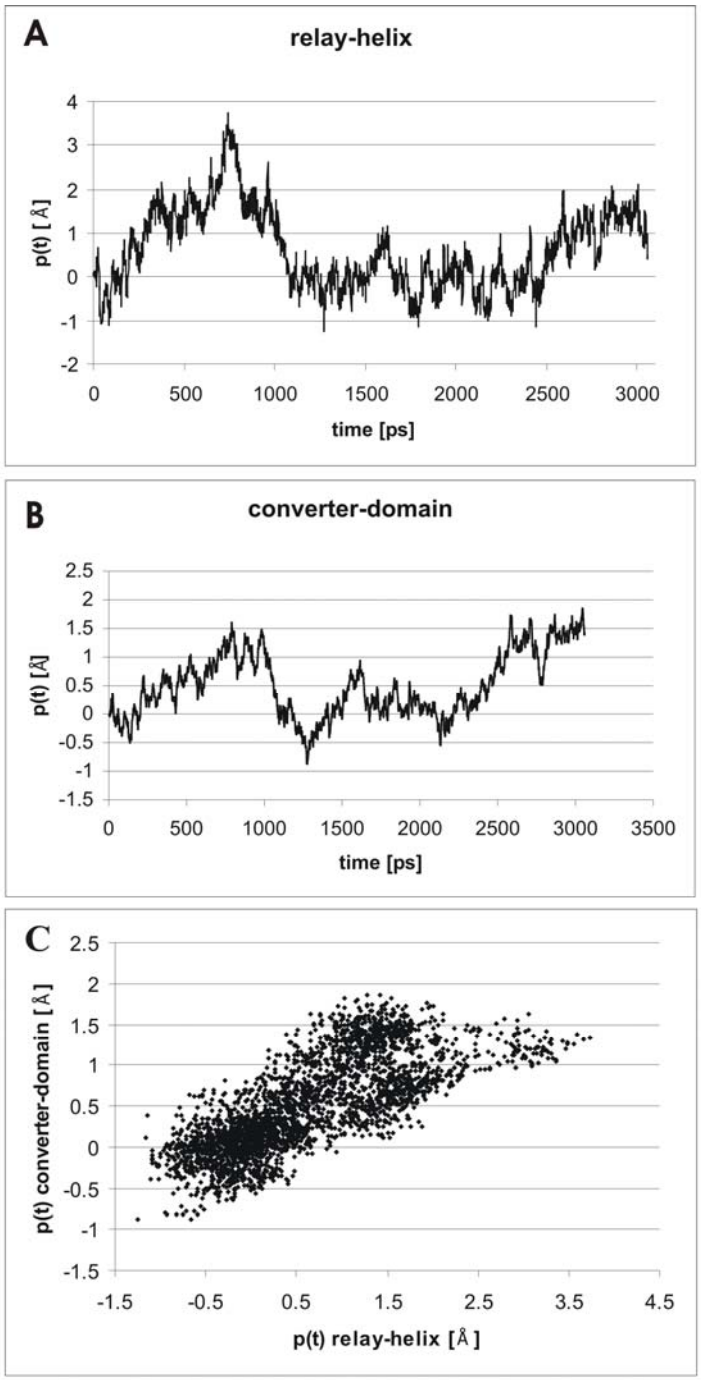


Fig. 5.6 Correlation between the relay helix seesaw and converter rotation in the MD of State II. The amplitude of atomic displacements along the dominant PC vector (*p(t)*, equation 5.1) is plotted as a function of the time of the MD simulation. A): Displacements of the relay helix projected onto the PC vector shown in Fig. 5.4B. B): Displacements of the converter domain projected onto the PC vector shown in Fig. 5.3B. C): Correlation plot of *p(t)* from panel B against *p(t)* from panel A.

The MD of the pre-recovery conformer without a bound nucleotide (i.e., State II-apo) was analyzed. The first three PC's do not correspond to a coherent motion and, as described above for the PCA of the converter domain in the absence of the ATP, no single principal motion of the relay helix has a clearly dominant Involvement Coefficient (the first two PC's both having I_k^2 in the 27±4% range, Fig. 5.5C). In PCA of the MD of the post-recovery conformation with bound ADP·Pi the principal motions of the relay helix are incoherent and have no dominant Involvement coefficient (Fig. 5.7).

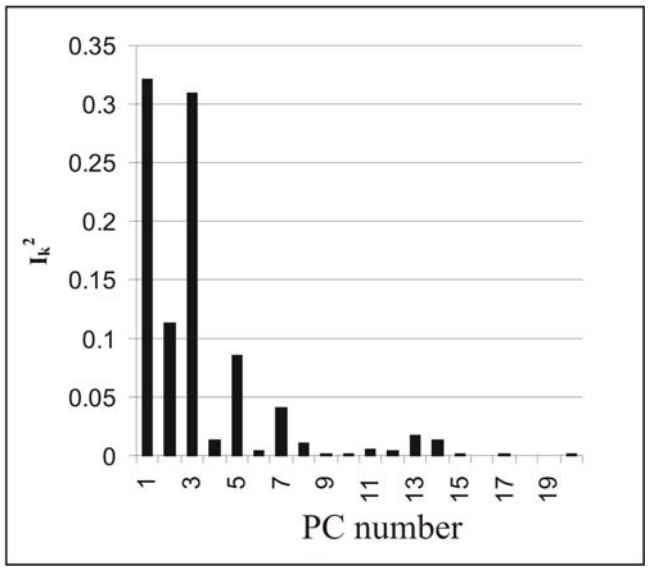


Fig. 5.7 Squared Involvement Coefficients for the MD of the post recovery state with bound ADP·Pi, PCA done on the relay helix.

All these findings about the relay helix indicate that, at the beginning of the recovery transition and in the presence of ATP, the relay helix undergoes a "seesaw" motion that couples its position at the N-terminal end (near the ATPase site) to its position at the C-terminal end (near the converter domain). At a later stage of the recovery transition, the relay helix plays less the role of a coupling element, but rather provides a mechanical hinge point by unwinding locally to allow the final converter rotation.

5.3.3 Wedging motion against the SH1 helix

PCA was performed on another sub-fragment of myosin that has been proposed to be instrumental in the coupling mechanism during the second phase of the recovery stroke: the SH1 and SH2 helices and the wedge loop (Fig. 5.2E).[9] In this model the "wedge" loop (residues 572-574) moves by ~4 Å towards the ATP and therefore wedges against the loop connecting the SH1 and SH2 helices. The SH1 helix responds by a longitudinal piston-like translation, thus pushing into the converter domain and thereby provoking its final 40º rotation. The motion of the wedge loop is due to the formation of a hydrogen bond between the Phe458 carbonyl group on Switch-2 and the Ser181 amide group of the P-loop binding ATP (dotted line in Fig. 5.2E), Switch-2 pulling the wedge loop along because the side chain of Phe458 is embedded in a hydrophobic cradle formed by the wedge loop (His572-Tyr573-Ala574).

If the correlated wedge/piston motions of the wedge-loop/SH1-helix are involved in the second phase of the recovery stroke, one might expect to find them in the MD of the post-recovery conformation (State III). This is indeed the case, as shown by the PCA of the MD in State III: The atomic motions along third principal component, which has the dominant Involvement Coefficient (I_3^2=33%, Fig. 5.8D), correspond precisely to the proposed "wedge/piston" motions (inset in Fig. 8D), the atoms of the wedge-loop moving towards the corner between the SH1 and SH2 helices, whose atoms move together towards the converter domain (see Fig. 5.2F and the molecular movie).

In contrast, in the PCA of the MD in State II, there is no single Principal Component with a dominant Involvement coefficient (PC_4, PC_6 and PC_7 having their I_k^2 in the 20-25% range, Fig. 5.8B). The same lack of involvement is seen for the principal motions in the post-recovery conformation with ADP·Pi bound (results shown in Fig. 5.9). Thus, the proposed piston/wedging motion is found only in the MD of State III, consistent with the model that this motion occurs towards the end of the recovery stroke and in the presence of ATP.

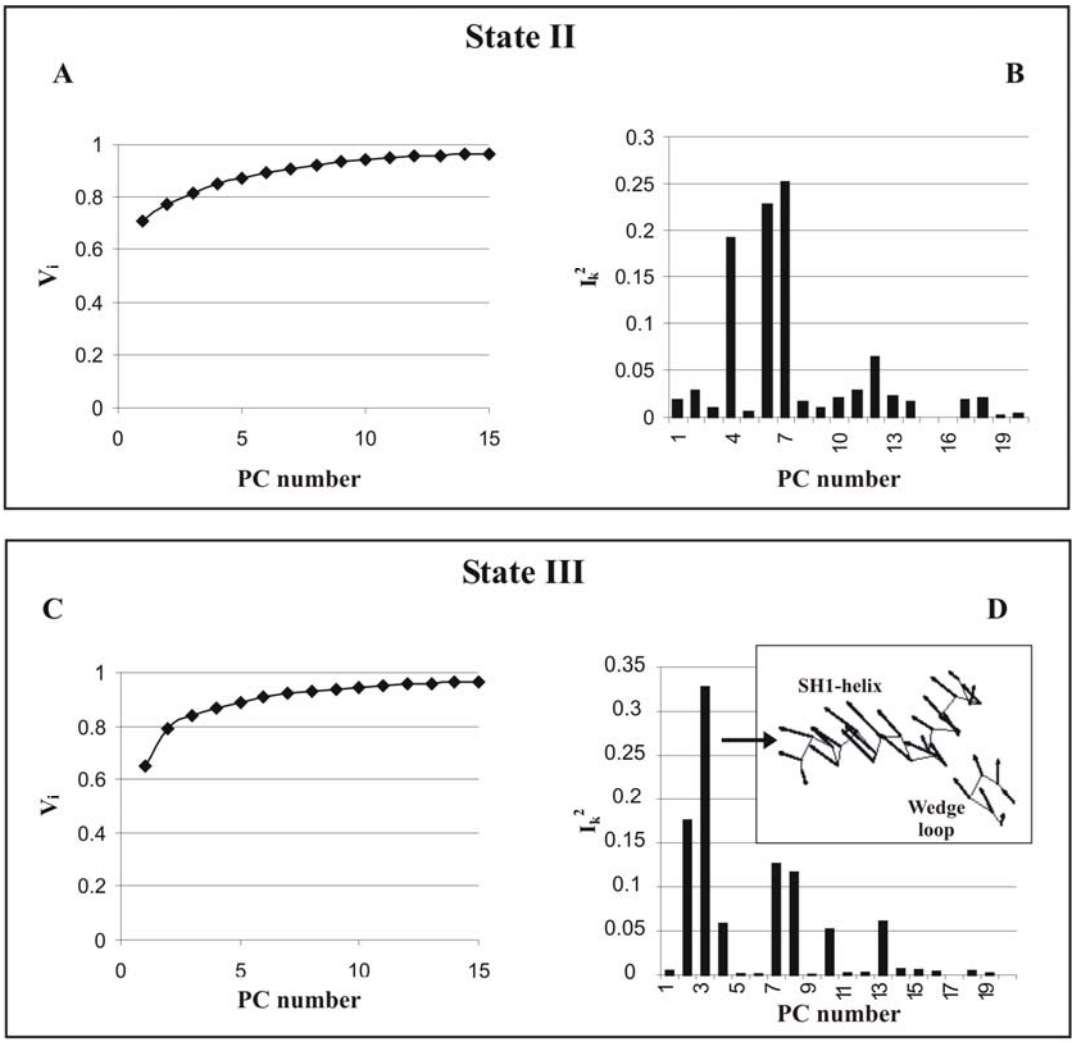


Fig. 5.8 PCA of the wedge loop and SH1 and SH2 helices. A) and B) MD of the pre-recovery State II. C) and D) MD of the post-recovery State III. **A)** and C) Cumulative sum of the covariance content (V_i) in the first 15 PC eigenvectors. **B)** and D) Squared Involvement Coefficients (I_k^2). The atomic motions of PC_3 in State III are shown in the inset of panel D.

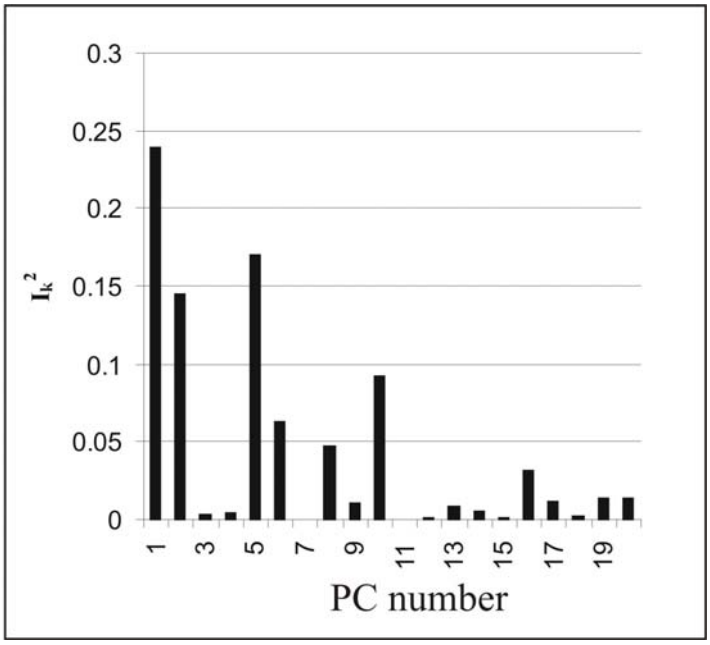


Fig. 5.9 Squared Involvement Coefficients for the MD of the post recovery state with bound ADP·Pi, PCA done on the wedge loop plus SH1 and SH2 helices.

5.4 Discussions

The structural changes of the switch-2 loop near the ATP binding site are amplified and transmitted so as to induce a large rotation of the converter domain. The results from the present analysis of MD simulations in the end conformations of the recovery stroke are consistent with a two-phase coupling model between ATPase activation and the recovery stroke. The first, phase (Fig. 5.2B) is initiated by the pulling of Gly457 towards ATP, leading to the seesaw motion of the relay helix and a partial 25º rotation of the converter domain. A molecular fulcrum anchors the relay helix on the main body of the myosin head (at Phe652) and provides a pivoting point for the seesaw motion (Fig. 5.2C). In this coupling model, the seesaw motion id dependent on the formation of the H-bond between the Ser456/Gly457 peptide group and γ–P of ATP. This is consistent with the phenotype of point mutational study which shows that mutation of the strictly conserved Ser456 to a Leucine is reducing

the step size of myosin walking along the actin filament.[31] Indeed, the model suggests that a larger side chain, *i.e.* Leucine in the mutant S456L, should hinder the movement of the Gly457/Ser456 peptide group towards γP during the seesaw phase, leading to a reduced pull on Asn475 and thus to a reduced upswing of the relay helix. This would cause a smaller rotation of the converter domain and explain the observed reduction in step size.

The second, "piston" phase is characterized by the wedging motion of the Wedge loop against the N-terminal end of SH1-helix (Fig. 5.2E), at the junction between the SH1 and SH2 helices around Gly680. This pushes the Sh1 helix towards the converter domain, which gives way by rotating a further 40º. The Wedge loop[9] and the SH1/SH2 junction are well-conserved loop structures in the myosin family.[32-35] Hydrophobic interactions between Switch-2 and the Wedge loop pull the latter element towards the ATP upon formation of the second H-bond of Switch-2, between Phe458 and Ser181 on the P-loop. Two studies involving the mutation of the highly conserved Gly680 to either Valine[36] or Alanine[37] have shown lower ATPase activity and significantly lower in-vitro mobilities. This is consistent with the coupling model, since in both mutants the presence of a large side chain would impede the insertion of the wedge loop between the SH1-SH2 helix junction and Switch-2. This prevents the recovery stroke from proceeding beyond the seesaw phase and prevents the complete closure of Switch-2 over the ATP binding site, thus leading to a lower mobility and less efficient ATP hydrolysis.

The structural flow of information has been presented here in the direction ATP binding site to converter domain, but the recovery stroke is reversible in presence of ATP[38], so that the coupling mechanism, could have been presented just as well in the reverse direction: Thermal fluctuation in the rotation angle of the converter domain transmit via the SH1 and relay helices to facilitate the closing of the Switch-2 loop over the ATP. The coupling mechanism simply implies that whenever the lever arm is in post-recovery orientation, the ATPase function is switched on, and when it is in the pre-recovery position, the ATPase function is switched off. Moreover, it is possible that some structural elements responsible for the coupling mechanism in the recovery stroke, such as the relay and the SH1 helices might also be implicated during

the power stroke (although this does not imply micro-reversibility, due to the fact that the power stroke occurs in a different, actin bound, conformation).

The present analysis of myosin MD trajectories has revealed that the principal motions implicated in the recovery stroke are dependent on the presence of ATP. For the MD trajectories in which nucleotide is absent the functional motions becomes less marked. This suggests that the ATP induces small but essential changes in the dynamical behavior of the protein. This is consistent with results from fluorescence experiments[38], electron-density maps[39] and cryo-electron micrographs[40], all of which suggest that the recovery stroke occurs predominantly when ATP is bound. It remains to be investigated how the differences in the nucleotide state induces these small but essential changes in the dynamical behavior of the protein, and to verify that these effects are not simply coincidental by running multiple trajectories.

As more structures of macromolecular complexes are solved in different conformations, it has become apparent that flexibility is an inherent part of their biological function. Filtering out and understanding those protein motions that have functional role is a major challenge in structural molecular biology. Here, the fact that large amplitude motions in a protein can be captured by PCA in only a few principal motions has lead to the identification of the functional motion involved in the recovery stroke. Because present-day the MD simulations of large proteins are short (a few ns) compared to the time needed for a large-scale conformational transition like the recovery stroke step (~1ms), the functional motions of interest are embedded in stochastic fluctuations. Here we have shown that when the primary elements involved in a given structural change can be somehow guessed, one can by filtering the relevant principal motions with the help of the Involvement Coefficient analysis extract their functional motions. Which structural elements may be primarily involved in a transition can be determined in various ways: by determining rigid protein domains from MD[41], by determining the hinge regions from direct comparison of the crystallographic end-states[23], or by analyzing minimum-energy pathways of the transition (which led to the present model for myosin[8]). When the implicated sub-fragments are relatively small their motions occur on much faster time-scale than the

overall transition. This was clearly seen for the relay helix (residues 466–498) or the converter domain (residues 691-760), whose functional motions undergo stochastic oscillations on the nanosecond time-scale (Fig. 5.5A and 5.5B). While these oscillations do not have the full amplitude of the motion that these elements undergo in the complete recovery stroke, the PCA/Involvement-Coefficient approach is able to distinguish whether these elements participate in a functional motion or not.

5.5 Methods

5.5.1 Molecular Dynamics Simulations

The X-ray crystal structure of the truncated Myosin II head from *Dictyostelium discoideum* complexed with Mg-ADP-BeF_3, a non-hydrolyzing ATP analog (PDB code: 1MMD[6]) was used as a starting point for the MD simulations (of the pre-recovery conformer). A missing segment in 1MMD (residues 501-507) was modeled based on the 2MYS structure.[42] For the MD of State II, the ATP-analogue was replaced by ATP. For the MD of State II-apo, the ATP-analogue was deleted and replaced with 14 water molecules. The Mg-ADP-BeF_3 structure used for the post-recovery conformation was obtained from Jon Kull and is very similar to the PDB structure 1VOM[7], but provides coordinates for the relay loop, which are missing in 1VOM. For MD of State III, the ATP-analogue was replaced with ATP. For MD of the post-hydrolysis state, the ATP-analogue was replaced with ADP and Pi and deleting water molecule 29 (considered to be the attacking water of hydrolysis[30]). The γ–P was positioned at the location of the BeF_3 and energy optimized. 31 buried crystal waters that are resolved in most Myosin II structures were included. The protein was placed in a box (125x90x75 $Å^3$) and solvated with 27000 water molecules, using the TIP3P model of CHARMM.[43] The production runs were performed under the same conditions for each of the four trajectories, at constant pressure (1 bar) and temperature (300 K). Details of the simulation protocol have been

reported elsewhere.[9] All MD simulations and energy minimizations were performed using CHARMM.[44]

5.5.2 Principal Component Analysis

PCA was carried out on the four different MD trajectories of the myosin motor domain described above. PCA starts from the variance-covariance matrix of interatomic fluctuations, **C**, whose elements, $c_{ij,}$ are calculated from the Cartesian coordinates of the atoms r_i (i = 1 to 3N; N=number of atoms included in the PCA) as

$$c_{ij} = \left\langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \right\rangle = \left\langle r_i r_j \right\rangle - \left\langle r_i \right\rangle \left\langle r_j \right\rangle \tag{5.2}$$

and where $\langle ... \rangle$ denotes an average taken over the MD trajectory. Only the Cα atoms were included in the present PCA. The translations and rotations of the protein along the trajectories were removed by always orienting all structures so as to obtain the least-squares best fit of the main body (i.e., the myosin head excluding the converter domain). The eigenvalues λ_i and eigenvectors L_i of

$$C \cdot L_i = \lambda_i \cdot L_i \tag{5.3}$$

are obtained by diagonalizing the matrix **C**. Each of the 3N normalized vectors L_i corresponds to a collective motion of atoms whose average amplitude is proportional to $2\sqrt{\lambda_i}$ (i.e., twice the standard deviation of the fluctuations along the vector L_i). L_i is called a Principal Component (PC) of the protein motions. The PC's are sorted in order of decreasing amplitude: The vector L_1 with largest eigenvalue λ_1 is called the first PC, which corresponds to the motion of largest average amplitude during a given MD run. For example in the MD of State III, the amplitudes of the 25 first PC's are plotted in Fig. 5.10A.

The fraction v_i of the total variance content in a given eigenvector L_i is given by:

$$v_i = \frac{\lambda_i}{\sum_{k=1}^{3N} \lambda_k} \tag{5.4}$$

v_i measures the contribution of a given PC towards the overall protein fluctuations. The cumulative contribution from a set of PC's is

$$V_n = \sum_{i=1}^{n} v_i \quad , \quad V_{3N} = 1 \tag{5.5}$$

The number n of PC's needed to capture most of the total variance, for example $V_n \approx 80\%$, is a measure of the variability of the motions in the protein. The first few PC's generally capture most of the intramolecular fluctuation of a protein.[10,14,20,25,45-47] This can be seen for example when doing PCA on the whole protein (N=744 residues) for the MD of State III: The motions in the 10 first PC's are already sufficient to account for 80% of the total variance (grey bars in Fig. 5.10B). While this is a very significant reduction in the dimensionality (i.e., 10 out of a total of 3N=2232 dimensions), the motions of individual protein sub-fragments (such as the ones implicated in the present coupling model) are still a combination of these 10 PC's, making it difficult to characterize them. Therefore, PCA was performed for individual sub-fragments of myosin. For example, when doing the PCA on only the main structural elements implicated in the coupling model (the converter/lever arm domain, the relay and SH1 helices, the Switch-2, the relay and the P- loops, for a total of 138 residues), the three first PC's give more than 80% of the total fluctuations (black bars in Fig. 5.10B). Note that this reflects a true reduction in the variability of the motions that inherent to these structural elements, and is not due to a more reduction in the size of the covariance matrix (which, having fewer eigenvectors, could be expected to result in a faster increase of V_n). Indeed, simply doing the PCA on a reduced number of the residues without regard to the structural elements, for example by choosing every fifth residue along the sequence (yielding 148 residues, and leaving 10 residues out in order to have the same total of 138 residues), 9 PC's are needed to capture 80% of the variance (white bars in Fig. 5.10B), nearly as many

as when all 744 residues are used. The other MD trajectories display similar results (not shown).

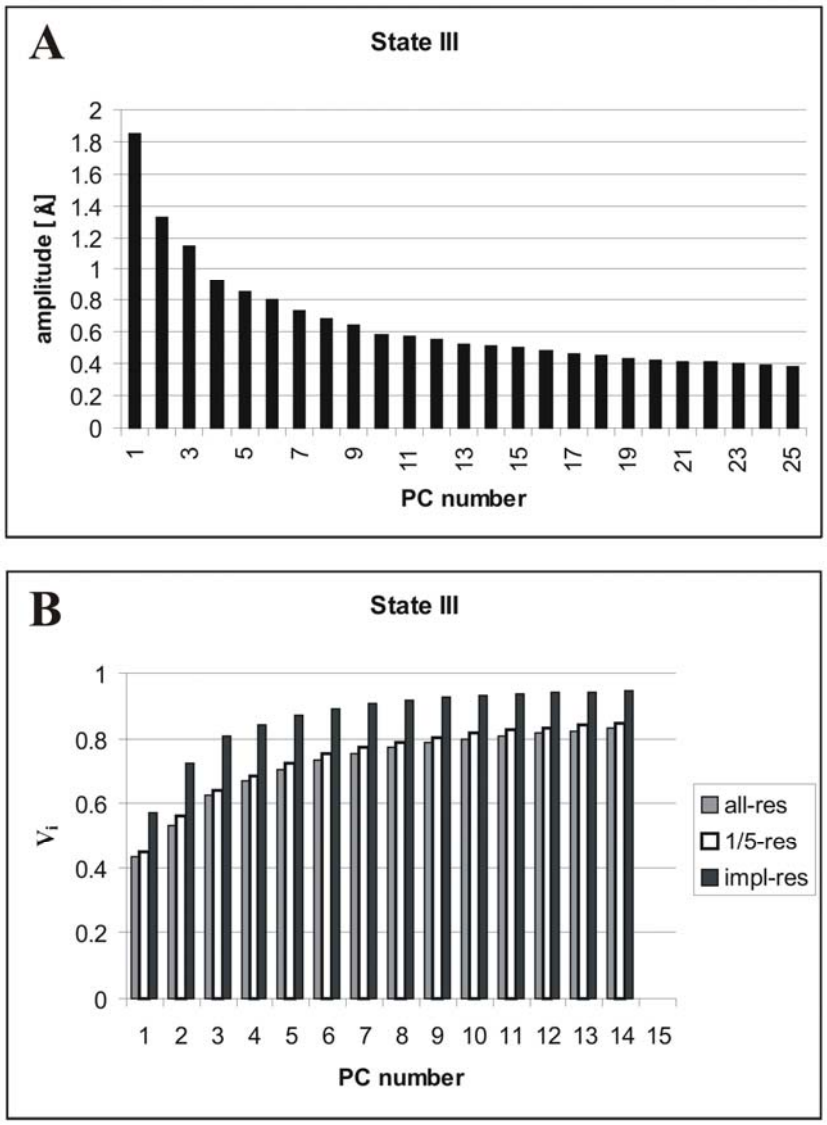


Fig. 5.10 PCA of the MD simulation of myosin in State III. **A)** Amplitudes of the first 25 Principal Components (out of a total of 2232 eigenvectors). **B)** Cumulative sum of the covariance content (V_i, equation 5.5) for the first 14 Principal Component. Grey bars: using all 744 residues of the protein to build the covariance matrix. White: using only 138 residues equally spaced along the polypeptide chain. Black: using the sub-fragments implicated in coupling mechanism (138 residues, see text).

5.5.3 Involvement Coefficients

The Involvement Coefficient of a particular PCA eigenvector L_k is calculated by projecting L_k onto the normalized vector connecting the end states of the conformational transition:

$$I_k = \frac{(X_1 - X_2)}{|X_1 - X_2|} \cdot L_k \tag{5.6}$$

where $\vec{X}_1 - \vec{X}_2$ is the displacement vector between two end-state conformations (in our case corresponding to states II and III of myosin). A large value of the Involvement Coefficient indicates that the motion of this PC is highly relevant to the conformational change being examined. A related quantity is the Cumulative Involvement Coefficient that measures the contribution from a set of eigenvectors:

$$C_n = \sum_{k=1}^{n} I_k^2 \tag{5.7}$$

Since the PCA eigenvectors form a complete orthonormal set for the 3N-dimensional space, the total Cumulative Involvement Coefficient (C_{3N}) is 1, meaning that I_k^2 is the fractional contribution of motion along the k^{th} PC to the main conformational difference between the two end states.

5.5.4 Convergence of L_k and I_k for sub-fragments

It is known that the largest amplitude PC's of a whole protein can suffer from convergence problems.[48-50] This is particularly true for short MD simulations of large proteins. This problem is alleviated here by focusing the PCA on sub-fragments with a small number of residues. For a short MD simulation the first few principal components of a large protein can reflect structural drift, rather than the actual oscillations of a long MD that has converged.[49] However, the motions of smaller sub-fragments of the protein converge faster, as can be seen for example for the relay

helix or the converter domain in Figures 5.6A and 5.6B. Rather than displaying a drift, these plots show repeated exchanges within a certain amplitude range (in a diffusive manner), indicating that the sampling of local substrates explored by the corresponding PC eigenvectors has converged. Further evidence that these localized motions have converged in the present study is obtained by splitting a given MD trajectory into two and applying the PCA to each half. The result for the behavior of the relay helix in the MD of State II is shown in Fig. 5.11A and 5.11B: In each trajectory half, the second PC has a markedly higher Involvement Coefficient and corresponds to the same seesaw motion as described in Results.

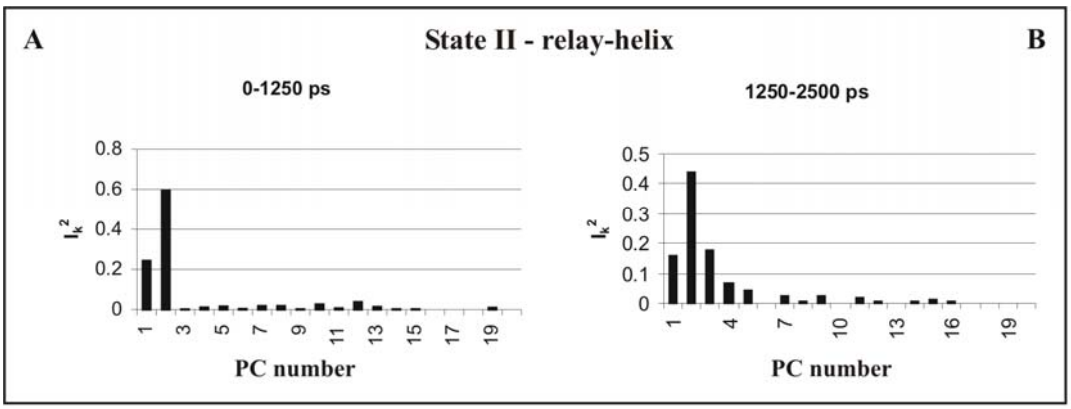


Fig. 5.11 Convergence of the Involvement Coefficients. I_k^2 of the PCA on the relay helix for the A) first and B) second half of the MD trajectory of State II. Both halves have PC_2 as dominantly involved in the recovery stroke, same as found for the whole trajectory (Fig. 5.5B)

References

1. Geeves M. A. & Holmes K. C. - Structural mechanism of muscle contraction. *Annu. Rev. Biochemistry* **68**, 687−728, 1999.
2. Holmes K. C. & Geeves M. A. - The structural basis of muscle contraction. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* **355**, 419-431, 2000.
3. Geeves M. A. & Holmes K. C. - The molecular mechanism of muscle contraction. *Adv. Protein Chem.* **71**, 161-193, 2005.
4. Lymn R. W. & Taylor E. W. - Mechanism of adenosine triphosphate hydrolysis by actomyosin. *Biochemistry* **10**, 4617-4624, 1971.
5. Gulick A. M., Bauer C. B., Thoden J. B. & Rayment I. - X-ray structures of the MgADP, MgATPgS, and MgAMPPNP complexes of the Dictyostelium discoideum myosin motor domain. *Biochemistry* **36**, 11618-11619, 1997.
6. Fisher A. J., Smith C. A., Thoden J. B., Smith R., Sutoh K., Holden H. M. & Rayment I. - X-ray Structures of the Myosin Motor Domain of Dictyostelium discoideum Complexed with MgADPBeFx, and MgADPAlF_4^-. *Biochemistry* **34(28)**, 8960-8972, 1995.
7. Smith C. A. & Rayment I. - X-ray structure of the Magnesium(II)ADP Vanadate complex of the dictyostelium myosin motor domain to 1.9 A resolution. *Biochemistry* **35**, 5404-5417, 1996.
8. Fischer S., Windshuegel B., Horak D., Holmes K. C. & Smith J. C. - Structural mechanism of the recovery stroke in the Myosin molecular motor. *Proc. Natl. Acad. Sci. USA* **102(19)**, 6873-6878, 2005.
9. Koppole S., Smith J. C. & Fischer S. - Simulations of the myosin II motor reveal a nucleotide-state sensing element that controls the recovery stroke. *J. Mol. Biol.* **361**, 604-616, 2006.
10. Kitao A., Hirata F. & Go N. - The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and vacuum. *J. Chem. Phys.* **158**, 447-472, 1991.

11. Brooks B. R., Janežič D. &. Karplus M. - Harmonic analysis of large systems. I: Methodology. *J. Comp. Chem.* **16**, 1522-1542, 1995.
12. Higo J., Sugimoto Y., Wakabayashi K. & Nakmura H. - Collective motions of myosin head derived from backbone molecular dynamics and combination with X-ray resolution scattering data. *J. Comp. Chem.* **22**, 1983-1994, 2001.
13. Fersht A. R. - Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9, 1997.
14. Garcia, A. N. E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68**, 2696-2699.
15. Hayward S., Kitao A. & Go N. - Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and Principal Component Analysis. *Protein Sci.* **3**, 936-943, 1994.
16. Becker O. M. - Quantitative visualization of a macromolecular potential energy funnel. *J. Mol. Struct. (THEOCHEM)* **398-399**, 507-516, 1997.
17. Karplus M. & Kushick J. N. - Method for estimating the configurational entropy of macromolecules. *Macromolecules* **14**, 325-332, 1981.
18. Perahia D., Levy R. M. & Karplus M. - Motions of an α-helical Polypeptide: Comparison of Molecular and Harmonic Dynamics. *Biopolymers* **29**, 645-677, 1990.
19. Ichiye T. & Karplus M. - Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins: Str. Func. and Gene.* **11**, 205-217, 1991.
20. Amadei A., Linssen A. B. M. & Berendsen H. J. C. - Essential dynamics of proteins. *Proteins: Str. Func. and Gene.* **17(3)**, 412-425, 1993.
21. Hayward S., Kitao A., Hirata F. & Go N. - Effect of solvent on collective motions in globular protein. *J. Mol. Biol.* **234**, 1204-1217, 1993.
22. De Groot B. L., Van Aalten D. M. F., Amadei A. & Berendsen H. J. C. - The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophys. J.* **71(4)**, 1707-1713, 1996.
23. Hayward S., Kitao A. & Berendsen H. J. C. - Model-Free Methods of Analyzing Domain Motions in Proteins From Simulations: A Comparison of

Normal Mode Analysis and Molecular Dynamics Simulation of Lysozyme. *Proteins: Str. Func. and Gene.* **24**, 425-437, 1997.

24. Caves L. S., Evanseck J. D. & Karplus M. - Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7(3)**, 649-666, 1998.
25. Tournier A. & Smith J. C. - Principal Components of the Protein Dynamical Transition. *Phys. Rev. Lett.* **91(20)**, 208106-4, 2003.
26. Ota N. & Agard D. A. - Enzyme specificity under dynamic control II: Principal component analysis of -lytic protease using global and local solvent boundary conditions. *Protein Sci.* **10**, 1403-1414, 2001.
27. Tama F. & Sanejouand Y. - Conformational change of proteins arising from normal mode calculations. *Prot. Eng.* **14**, 1-6, 2001.
28. Li G. & Cui Q. - Analysis of functional motions in Brownian Molecular Machines with an efficient Block Normal Mode Approach: Myosin-II and Ca^{2+} – ATPase. *Biophys. J.* **86**, 743-763, 2004.
29. Kraulis P. J. - Molscript - a program to produce both detailed and schematic plots of protein structures. *J Appl. Cryst.* **24**, 946-950, 1991.
30. Schwarzl S. M., Smith J. C. & Fischer S. - Insights into the Chemomechanical Coupling of the Myosin Motor from Simulations of its ATP Hydrolysis Mechanism. *Biochemistry* **45**, 5830-5847, 2006.
31. Murphy C. T., Rock R. S. & Spudich J. A. - A myosin II mutation uncouples ATPase activity from motility and shortens step size. *Nature Cell Biology* **33**, 311-315, 2001.
32. Jamie M., Cope T. V., Whisstock J. Rayment I. & Kendrick-Jones J. - Conservation within the myosin motor domain: implications for structure and function. *Structure* **4**, 969-987, 1996.
33. Sellers J. R. - Myosins: A diverse superfamily. *Biochim. Biophys. Acta* **1496**, 3-22, 2000.
34. Hodge T. & Cope M. J. T. V. - Myosin Motor Domain Sequence Alignment. http://www.mrc-lmb.cam.ac.uk/myosin/trees/colour.html., 2000.

35. Hodge T. & Cope M. J. T. V. - A myosin family tree. *J. Cell Sci.* **113**, 3353-3354, 2000.
36. Patterson B., Ruppel K. M., Wu Y. & Spudich J. A. - Cold-sensitive mutants g680v and g691c of Dictyostelium myosin II confer dramatically different biochemical defects. *J. Biol. Chem.* **272**, 27612-27617, 1997.
37. Batra R., Geeves M. A. & Manstein D. J. - Kinectic analysis of Dictyostelium discoideum myosin motor domains with glycine-to-alanine mutations in the reactive thiol region. *Biochemistry* **38**, 6126-6134, 1999.
38. Málnási-Csizmadia A., Pearson D. S., Kovács M., Woolley R. J., Geeves M. A. & Bagshaw C. R. - Kinetic Resolution of a Conformational Transition and the ATP Hydrolysis Step Using Relaxasion Methods with a Dictyostelium Myosin II Mutant Containung a Single Tryptophan Residue. *Biochemistry* **40**, 12727-12737, 2001.
39. Bauer C. B., Holden H. M., Thoden J. B., Smith R. & Rayment I. - X-ray structures of the APO and Mg-ATP-bound states of *Dictyostelium discoideum* myosin motor domain. *J. Biol. Chem.* **275**, 38494-38499, 2000.
40. Holmes K. C., Schröder R. R., Eschenburg S., Sweeney H. L. & Houdusse A. - The structure of the rigor complex and its implications for the power-stroke. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* **359**, 1819-1828, 2004.
41. Gaudin F., Lancelot G. & Genest D. - Search for rigid subdomains in DNA from molecular dynamics simulations. *J. Biomol. Struct Dyn.* **15(2)**, 357-367, 1997.
42. Rayment L., Rypniewski W. R., Schmidt-Base K., Smith R., Tomchick D. R., Benning M. M., Winkelmann D.A., Wesenberg G. & Holden H.M. - Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261(5117)**, 50-58, 1993.
43. Jorgensen W. L., Chandrasekhar J., Medura J. D., Impey R. W. & Klein M. L. - Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79(2)**, 926-935, 1983.

44. Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S. & Karplus M. - CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187-217, 1983.
45. Horiuchi T. & Go N. - Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme. *Proteins: Str. Func. and Gene.* **10**, 106-116, 1991.
46. van Aalten D. M., Amadei A., Linssen A. B., Eijsink V. G., Vriend G. & Berendsen H. J. - The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins* **22(1)**, 45-54, 1995.
47. Mesentean S., Fischer S. & Smith J. C. - Analyzing Large-Scale Structural Change in Proteins: Comparison of Principal Component Projection and Sammon Mapping. *Proteins: Str. Func. and Bioinf.* **64**, 210-218, 2006.
48. Balsera M. A., Wriggers W., Oono Y. & Schulten K. - Principal component analysis and long time protein dynamics. *J. Phys. Chem.* **100**, 2567-2572, 1996.
49. Hess B. - Convergence of sampling in protein simulations. *Phys. Rev. E* **65**, 031910-10, 2002.
50. Meinhold L. & Smith J. C. - Fluctuations and correlations in Crystalline Protein Dynamics: A Simulation Analysis of Staphylococcal Nuclease. *Biophys. J.*, **88**, 2554-2563, 2005.

Chapter 6

Formation and Dissociation of Rigid Domains in Myosin's Functional Cycle

Proteins are dynamic structures that can undergo considerable conformational changes. Large proteins are normally built of smaller domains, which are generally autonomous folding units and they may be quite rigid. Domains move with respect to each other, but the overall structure of single domains remains most of the time unaffected. Proteins utilize domain motions for function. They can be triggered by events like substrate- or DNA-binding, changes in ion concentration, or by protein-protein interaction to enumerate a few examples. Proteins that bind different substrates have been crystallized in open (substrate-free) and closed (substrate-bound) conformations leading to a good view of their static structure. However, high-resolution crystal structures allow the formulation of hypothesis on the mechanism of domain motion in the absence of dynamical information. Due to the fact that a complex communication pathway is also governing the previous presented interaction between myosin and actin, we show in this chapter how with a relatively simple method rigid identification and relative motion of myosin II domains can be investigated. The results presented here are part of a research article that is to be submitted for peer-review. However, for simplicity reasons, references to previous chapters of the thesis are often made.

6.1 Abstract

Myosin II head is a molecular motor that transforms chemical energy derived from ATP hydrolysis into mechanical work, such as muscle contraction, by cyclically interacting with actin filaments. In the myosin functional cycle, the binding of ATP to myosin dissociates the myosin-actin rigor complex and a subsequent large-scale conformational change known as recovery stroke brings the myosin head into a position in which it can rebind to actin and perform the power-stroke, thus exerting productive mechanical work. Recent computer simulations have shed light on the detailed structural mechanism of the recovery stroke and a model was proposed in which nearly rigid domains, such as the relay helix, the wedge loop or the SH1/SH2 helices, move in a well-defined way relative to each other so as to establish the overall conformational change. Whether these rigid domains are always present in myosin or actually change during the conformational and chemical changes in the functional cycle was previously not investigated. Here we present results obtained by applying a method which automatically identifies relative rigid domains on four different states of myosin: 1) pre-recovery stroke conformation without nucleotide, 2) pre-recovery stroke conformation (or State II), 3) post-recovery stroke conformation (or State III) and 4) the post-recovery stroke conformation with ADP·P_i. The rigidity analysis tested here is successful in automatically detecting rigid domains which are known to be functionally relevant. Furthermore, it is found that the rigid domains significantly differ in the four different states of myosin, in particular, the relay and the SH1-SH2 helices which have been identified to be relevant for the mechanical coupling between the binding site and the converter domain become rigid only after the binding of ATP and dissociate subsequent to the recovery stroke. Thus, the ATP binding event is not only a signal to initiate the recovery stroke, but it actually presets three-dimensionally all the structural elements needed for the recovery stroke to occur in the first place.

6.2 Introduction

The muscle contraction is possible due to cyclic interactions of the myosin heads with the actin filaments and relies on a well-defined sequence of conformational changes in the motor domain. The myosin head (or myosin cross-bridge) contains several units known to be relevant for mechanical function, including the nucleotide-binding site, the actin-binding site and the "converter/lever arm" domain (also called the force generating domain). The ability of myosin to couple small changes in the catalytic ATPase site with large conformational changes in both the actin-binding and the force-generating domains requires well-defined communication mechanisms ensuring that these changes are correlated such as to efficiently produce mechanical work. A well-known theoretical model of the interaction of actin and Myosin II is the Lymn-Taylor cycle[1] (see Fig. 5.1 in Chapter 5). According to this cycle myosin binds strongly to actin in the absence of ATP in the so-called rigor conformation (State I in Fig. 5.1). Nucleotide binding leads to the subsequent dissociation of myosin and actin and to conformational changes, with the formation of the post-rigor state (denoted here as the "pre-recovery" or State II). Myosin then undergoes a reversible transition (named "recovery stroke") to the pre-power stroke conformation (denoted here "post-recovery" or State III). In State III, ATP hydrolysis occurs and myosin rebinds to actin (State IV). By performing the "power-stroke" and releasing the products of ATP hydrolysis (ADP·Pi), myosin returns to the rigor state. In order to understand the process in atomic detail, Myosin II has been crystallized with different ATP analogues under various experimental conditions[2-4]. While no crystal structures of the myosin-actin complex are available, two myosin structures in the absence of actin have been found and were assigned to States II and III in the Lymn-Taylor cycle, respectively. The largest structural difference between these states is the orientation of the converter/lever-arm domain which is rotated by ~60° relative to the rest of myosin head (see Fig. 6.1). Another difference consists in the fact that the ATP binding site is partially opened in the pre-recovery conformation while in the post-recovery one it is closed, thus switching on the catalytic ATPase function[5].

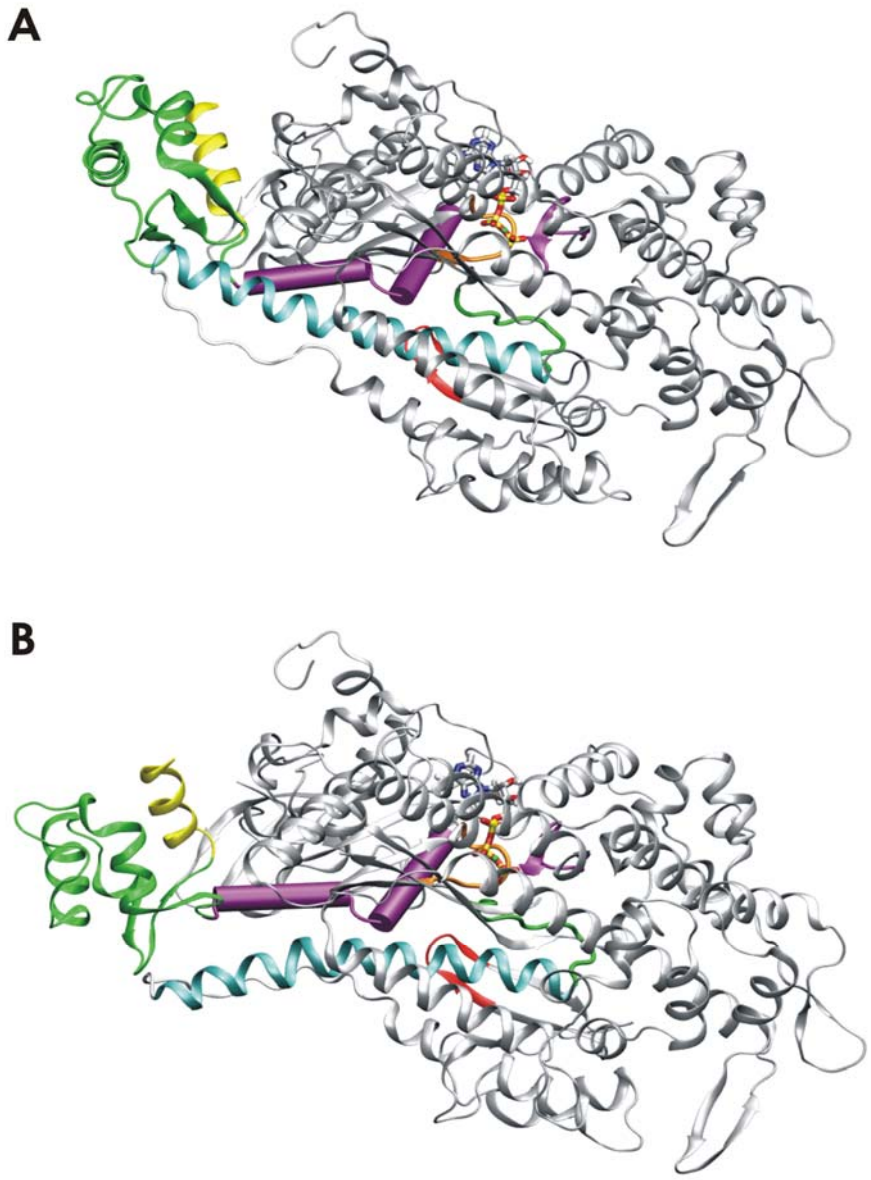


Fig. 6.1 End states of the recovery stroke. Structural elements highlighted: Relay helix in cyan; SH1 and SH2 helix in purple; Converter domain in green; Lever arm in yellow; P-loop in orange; Switch-1 in purple; Switch-2 in green; Wedge loop in red. Rest of the protein is colored in gray. Panel A) representation of the pre-recovery conformation (State II) and B) of the post-recovery conformation (State III).

Recently, a structural model of the recovery stroke was proposed based on a computer simulation in which a minimum-energy pathway was calculated between the pre- and post-recovery stroke conformations[6]. In this model, the motion is initiated

by the movement of the Gly457 on "Switch-2 loop" (residues 454-459) towards the γ–phosphate which is transmitted as a pull on the "Relay helix" (residues 499-509) through a hydrogen bond between the Gly457 peptide group and the side chain of Asn475 located on the N-terminal half of the relay helix. This pull initiates a "see-saw" motion of the relay helix transmitted to its C-terminal end connected to the converter domain. Therefore, the converter domain reacts with a partial rotation (see Ref. 6 for more details). Another element was lately postulated to control the orientation of the converter domain after the "see-saw" motion of the relay helix, namely the "SH1-helix" (residues 681-691), which translates lengthwise parallel to the relay helix and towards the converter[7] thus determining the final rotation of the converter/lever arm domain. This motion is produced due to the fact that in the post-recovery conformation Tyr573 wedges against the N-terminal end of the SH1-helix translating it relative to the relay helix. The hydrogen bond formed between Asn475 and Tyr573 locks the SH1 and relay helices together with the converter (and inherent the lever arm) in the post-recovery conformation.

The myosin recovery stroke is an example of a large-scale functional transition, which seems to be controlled by relative movements of nearly rigid domains, such as the converter domain, relay helix, or SH1 helix. Often, such domains are inferred from static information, such as the differences between crystallographic end-states of a transition, either by visual inspection or using algorithms that identify groups of atoms comprising little internal rearrangements[8,9]. In Ref. 6, the functional domains were identified by visual inspection of a minimum energy pathway connecting the pre- and post-recovery stroke states. In Ref. 7 the same kind of analysis was applied to molecular dynamics (MD) simulations of the myosin head. Hayward[8] also proposed to identify rigid domains by clustering similar rotation vectors in the first few components from a harmonic or quasi-harmonic analysis. In Ref. 10, a binary rigidity matrix was defined so as to yield a nearly block-diagonal structure. Based on the latter idea, a method in which a matrix with continuous rigidity values formulated as an optimisation problem to identify the groups of atoms with maximum intra-group rigidity and inter-group flexibility is used

in this thesis. Good solutions to its optimisation are generated with a Monte-Carlo maximization procedure.

Rigid domains are not only specific to a given protein, but also to the conformation and the chemical state (e.g. type of nucleotide in the binding pocket). The question how do rigid- and possibly functional-domains change when the conformation or chemical state changes has been largely disregarded in the past, but is explicitly addressed in the present work. Using our method, we identify rigid domains based on MD simulations in four different states: In the pre-recovery stroke conformation without nucleotide bound, modelling the situation just before the ATP binds and the recovery stroke can occur, in the pre-recovery stroke conformation, in the post-recovery stroke conformation and in the post-recovery stroke conformation with ADP·Pi, modelling the state just subsequent to hydrolysis. The rigidity analysis is successful in automatically detecting rigid domains known to be functionally relevant. More importantly, it points out the differences between different conformational and chemical states. For example, the relay and the SH1-SH2 helices, which have been identified to be relevant for the mechanical coupling between the binding site and the converter domain, become rigid only after the binding of ATP and dissociate subsequent to the recovery stroke. Thus, the rigid domains are actually formed just for performing the functional motion they are relevant for. Moreover, it is found that small changes in the chemical state, such as replacing the ATP by ADP·Pi can dramatically affect the rigid domains in the protein. The present analysis focuses on the analysis of the recovery stroke step in the myosin functional cycle, but the methodology presented here can be generally useful to understand the association and dissociation of rigid domains in conformational and chemical changes of macromolecules. Thus, our approach may be widely used to identify functionally relevant structural elements and understand how they are coupled in order to perform their function.

6.3 Methods

6.3.1 Simulation setup and force field

The truncated Myosin II head from *Dictyostelium discoideum*, Myosin, sub-fragment I has been crystallized in the absence of actin with different ATP analogs in the active site. The X-ray crystal structure of Myosin II complexed with Mg-ADP-BeF_3, a non-hydrolyzing ATP analog (PDB code: 1MMD[3]) was used as a starting point for two of the MD simulations (corresponding to the pre-recovery conformer of myosin with ATP included and without ATP). A missing segment in 1MMD (residues 501-507) was modeled based on the 2MYS structure[11]. To mimic the apo-state of myosin, the pre-recovery structure with no ATP was created by deleting the ATP and adding 14 water molecules in the ATP binding site. One of the structures used for the post-recovery conformation is similar to the PDB structure 1VOM[4], but also provides the coordinates for the relay loop, which are missing in 1VOM. The post-hydrolysis state of the post-recovery conformation was also modeled by replacing ATP with added ADP and Pi in the active site and deleting water molecule 29 (considered to be the attacking water of hydrolysis[12]). In all states ATP was modeled by replacing BeF_3 of the bound MG·ADP·BeF_3 with a phosphate group. The 31 crystal waters that were resolved in most Myosin II structures to date were included using the TIP3P model[13]. This various structures were then used as starting points for the MD simulations analyzed in this thesis. All MD simulations and energy minimizations were performed using CHARMM[14]. Details of the simulation methodology and protocol are reported elsewhere[7]. The production runs were performed under the same conditions for each of the four trajectories.

6.3.2 Rigidity matrix

From a MD trajectory of the protein we compute a symmetric matrix $\mathbf{C} \in [0,1]^{N \times N}$, where N is the number of atoms used for this analysis. Here, only

C_α-atoms are used for the analysis. **C** is defined such that a value C_{ij}=0 means that atoms *i* and *j* are moving uncorrelated while C_{ij}=1 means that the atoms *i* and *j* are moving together. Given the pair-wise atomic distances in every time step, $d_{ij}(t)$, the matrix elements, C_{ij}, are obtained as follows:

$$C_{ij} = \frac{1 - \min\{\sigma(d_{ij}), \sigma_{cut}\}}{\sigma_{cut}} \tag{6.1}$$

where $\sigma(d_{ij})$ is the standard deviation of $d_{ij}(t)$ and σ_{cut} is a user-defined cutoff . All deviations larger than σ_{cut} are mapped to 0 in matrix **C**. Obviously, the choice of σ_{cut} will affect the results of the clustering. In the examples studied here we used a value of σ_{cut} = 1Å, which turned out to be appropriate for identifying the functionally relevant domains of myosin. Increasing the value of σ_{cut} also leads to increasingly large domains due to the fact that the intra-domain motion is less penalized. In the present case, values of σ_{cut} >2.5 Å produce a single large cluster encompassing nearly the complete protein, apart from some solvent-exposed loops which are very flexible (data not shown). Modifications of σ_{cut} close to the value of 1 Å, however, do not have a strong effect on the results as shown in Fig. 6.2, which visualizes two partitions into rigid domains performed with σ_{cut} = 1 Å and σ_{cut} = 1.25 Å. The size of the domains is slightly different but there are no crucial changes in the identified clusters.

6.3.3 Identifying rigid domains

C can be used to identify rigid clusters of atoms by finding a group of atoms such that each atom within the group has a large pair-wise rigidity with all other atoms in the group, and a small pair-wise rigidity with all atoms outside the group. Formally, we attempt to identify a set of clusters by maximizing the following target function:

$$Z = \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} z_{ij} \tag{6.2}$$

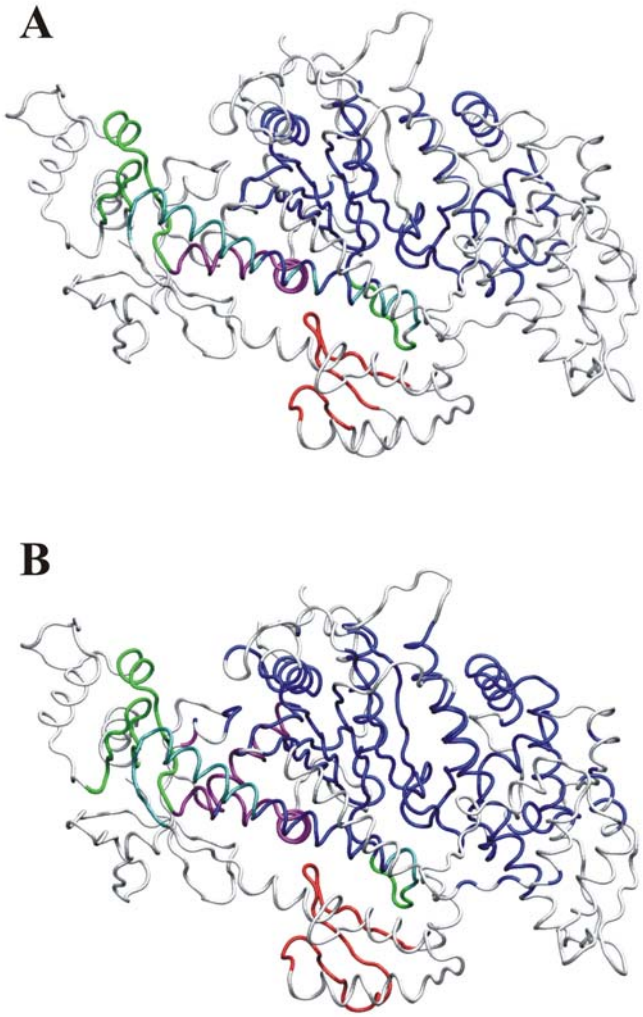


Fig. 6.2 Rigid domains obtained from the analysis of MD trajectory of Sate II with a cutoff: of 1 in panel A and of 1.25 in panel B.

Although it is very hard to identify the global maximum of Z it is easy to identify a good local maximum of Z, which is maximized using a stochastic optimization method (in detail described in Chapter 2). The main steps during this methodology are:

1. Start with N clusters, each containing a single atom.
2. Consider a random one the following operations:

a. *Split*: Randomly determine a cluster with at least two atoms. Split this cluster into two new clusters assigning each member to a random one of these two. Remove the old cluster.
b. *Merge*: If there is more than one cluster in the original one, select two different clusters at random and merge them into a new cluster. Remove the two selected clusters.
c. *Jump*: If there is more than one cluster in the original one, select two different clusters at random. Remove a random member from one of them and add it to the other one.
d. *Swap*: If there is more than one cluster in the original one, select two different clusters at random. Determine a random member in each cluster and swap them.

3. In each iteration, a random feasible operation is selected. A move is accepted, if it improves the target function Z if not return to step 2.

This Monte-Carlo procedure turns out to be very efficient. For the myosin case tested here (with $N \approx 800$), Z converges within 100000 steps, which takes a few seconds on a standard CPU to compute. Repeating the optimization various times gives very similar results.

6.4 Results

Different MD simulations of fully solvated myosin head structures (Myosin II) in *pre-recovery conformation* without nucleotide bound (corresponding to the Apo-State) or with ATP bound (corresponding to State II) and in *post-recovery conformation* with ATP bound (corresponding to State III) or with ADP·Pi bound (corresponding to the Post-hydrolysis State) were analyzed with the aim to detect rigid clusters relevant for the motion underlying the recovery stroke. A rigidity analysis (see Methods) was then applied onto the MD simulations in order to identify nearly rigid atom clusters. The cluster partition obtained represents a simplified mechanical model that captures the functional motions relevant to the recovery stroke.

The main results of the clustering analysis are depicted in Fig. 6.3. Several domains known to be functionally relevant could be identified as rigid regions and they can be seen in Fig. 6.1. In all four states there is a main (large) rigid cluster identified in the core of the protein known to be quasi rigid[16-18] (blue cluster in Fig. 6.3). The other rigid clusters found in the different states and their functional relevance are discussed in the following.

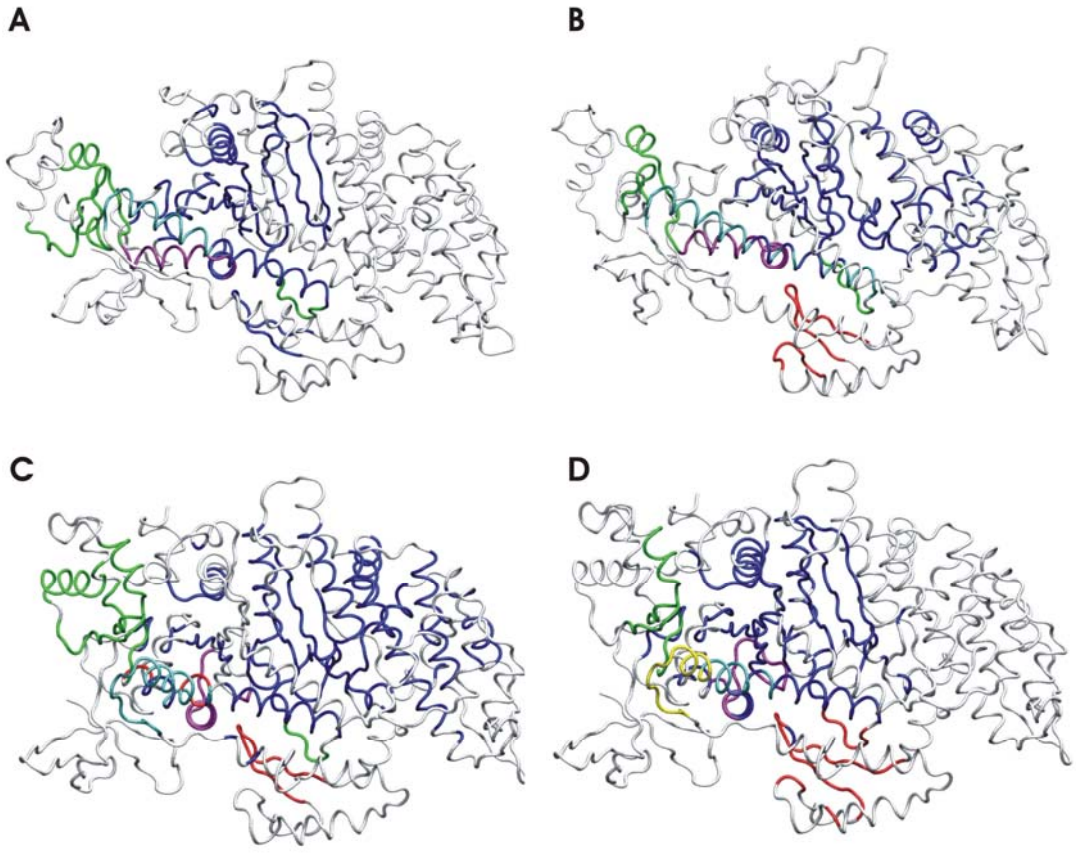


Fig. 6.3 Rigidity analysis on MD trajectories. A): State II without nucleotide bound and with a cutoff of 1. B): State II with a cutoff of 1. C): State III with a cutoff of 1. D): Post hydrolysis State (State III with ADP·Pi bound) with a cutoff of 1.

6.4.1 State II, without nucleotide.

There is no crystal structure available for the nucleotide-free rigor state of myosin. Therefore, in order to simulate such a state (i.e. previous to ATP binding) the starting conformation was modeled by using the State II crystal structure and replacing the ATP by water molecules (see Methods). The cluster partition (see Fig. 6.3) shows that there is no direct mechanical coupling between the binding site and the converter domain as both the relay helix (cyan in Fig. 6.3) as well as the SH1-SH2 helices (purple in Fig. 6.3A) are broken in two separate rigid domains each. This finding is consistent with the well-known fact that no converter motion occurs in the nucleotide-free state.

6.4.2 State II, ATP bound

Using the State II structure (where ATP is bound), the situation changes drastically and the rigidity analysis automatically identifies several rigid elements known to play key roles in the large conformational change performed between State II and III along the Lymn-Taylor cycle (see Fig 6.3B). Switch 2 belongs to the nucleotide-binding site and together with the P-loop and the Switch1 forms the so-called phosphate tube. It has been hypothesized that the movement of the Switch2 is a direct response to the binding of ATP and carries this signal on by forming a hydrogen bond, which pulls on the relay helix[6]. In contrast to the previous state, most residues in the relay-helix (~80%) are assigned to a single, separate domain (colored cyan in Fig. 6.3). Thus, the binding of ATP rigidifies the relay helix and enables it to transmit the mechanical force from the binding site to the converter. This reinforces the "seesaw-phase" theory[6] where the movement of the relay helix generated by Switch 2 is carried to the converter domain, thus initiating converter's large-scale rotation. The main part of the converter (colored green in Fig. 6.3) is forming a completely separate domain, indicating a flexible coupling between relay helix and converter (similar to a hinge-coupling). Moreover, in contrast to the previous state, the SH1 and SH2 helices have merged into a single cluster. Interestingly, these helices

have been observed as playing a role in the stability and the degree of rotation that is allowed for the converter domain[6]. The so-called wedge-loop, now identified as a separate cluster (shown in red), has been speculated to push the SH1 helix into the direction of the converter[7]. Due to the merging of SH1 and SH2 into a single cluster that is separated from the main cluster formed by the protein core, this motion can now be carried on to the converter and thereby initiate the second phase of the converter rotation.

6.4.3 State III, ATP bound

Subsequent to the recovery stoke, both the relay helix and the SH1-SH2 helices are again broken into two rigid clusters each, thus mechanically decoupling the converter domain from the binding site (Fig 6.3C). It is possible that this decoupling stabilizes the converter domain in its new conformation by switching off correlated motions that would promote unproductive back-rotations. As shown in Fig. 6.4 first rigid domains are found and then rigid clusters are identified out of them. These clusters are made such that each atom within the group has a large pair-wise rigidity with all other atoms in the group, and a small pair-wise rigidity with all atoms outside the group.

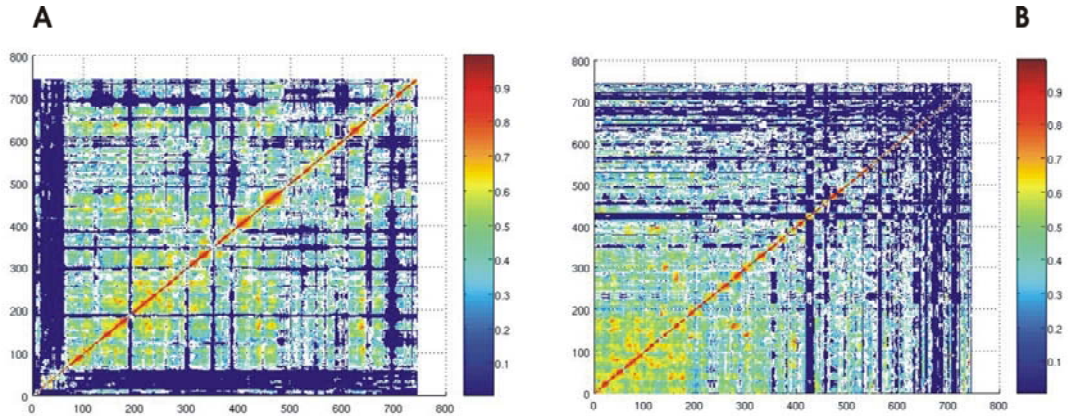


Fig. 6.4 Rigidity matrix obtained for State III with a cutoff of 1. Panel A shows the matrix before the clustering and panel B after clustering.

Another mechanism that is likely to stabilize the post-recovery stroke state is the wedge loop being locked in an "up" position, thereby preventing the SH1-SH2 helices to move back, thus blocking a back-rotation of the converter. The rigidity analysis seems to confirm this by coupling the wedge loop with nearby residues in the SH1 helix (colored in red in Fig. 6.3C). Another potentially interesting finding is that the rigid domain in the converter region has enlarged compared to the previous states in the Lymn-Taylor cycle (colored in green).

6.4.4 State III, after hydrolysis

The hydrolysis of ATP into ADP·Pi seems to be a relatively small event, but its effects on the rigid clusters of the protein are quite large. For example, the relay helix breaks into three rigid clusters, and the mechanical coupling between the wedge loop and the SH1-SH2 helices is lost, the wedge loop now being in a rigid cluster with other surrounding residues. As it was the case in the nucleotide-free State II, the rigid clusters in the post-hydrolysis state does not reflect the mechanisms involved in the recovery stroke, but are more likely to conform the (yet unknown) conformational rearrangements that may be required for the rebinding of myosin to actin in the transition III→IV along the Lymn-Taylor cycle.

6.4.5 Clustering in a CPR path

We also performed a rigidity analysis on a CPR path (see Chapter 2 and 3 for trajectory details) of myosin (i.e., going from State II towards State III along the Lymn-Taylor cycle) and the two main rigid domains found for this case are, as expected, the converter (red in Fig. 6.5) and most of the remaining myosin head (blue colour in Fig. 6.5).

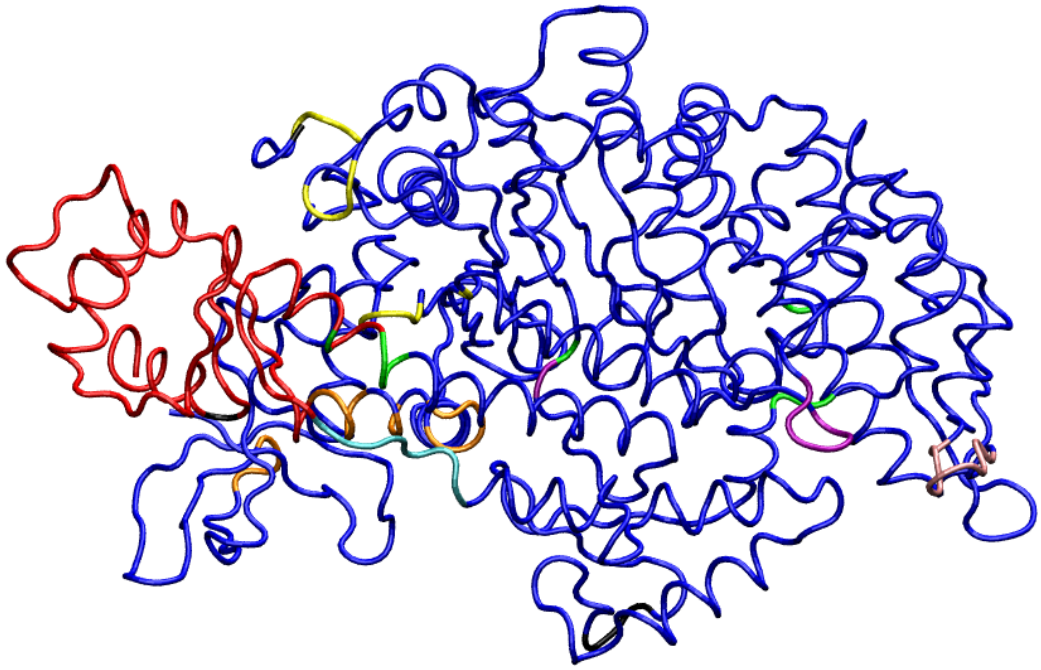


Fig. 6.5 Rigidity domains obtained for the CPR trajectory of myosin. Cutoff is 1.

6.5 Conclusions

The myosin recovery stroke is an example of a large-scale functional transition, which as shown seems to be controlled by relative movements of nearly rigid domains like the converter domain, the relay helix, or the SH1 helix. Here we present a method in which a matrix with continuous rigidity values formulated as an optimisation problem is calculated in order to identify groups of atoms with maximum intra-group rigidity and inter-group flexibility.

We showed here that rigid domains of myosin are specific to chemical states differing in what regards the type of nucleotide in the binding pocket. We show with the current work how do rigid and possibly functional domains change when the conformation or chemical state changes. For the particular case of myosin it seems that the presence of the nucleotide is crucial when identifying the functional rigid

clusters of the protein. The rigid parts found are closely related to the big rotation of the converter domain, which is the most obvious change when crystal structures are compared.

To conclude the present analysis on different myosin conformers allowed the identification of several rigid elements known to play key roles in the large conformational change between State II and State III. These findings appear to be consistent with the coupling mechanism of the recovery stroke step in the myosin cycle, as presented in the previous chapters of this thesis.

References:

1. Lymn, R. W. & Taylor, E. W. - Mechanism of adenosine triphosphate hydrolysis by actomyosin. *Biochemistry* **10**, 4617-4624, 1971.
2. Gulick A. M., Bauer C. B., Thoden J. B. & Rayment I. - X-ray structures of the mg-adp, mg-atp-γs, and mg-amp-pnp complexes of the *dictyostelium discoideum* myosin motor domain. *Biochemistry* **36**, 11618-11619, 1997.
3. Fisher A. J., Smith C. A., Thoden J. B., Smith R., Sutoh K., Holden H. M. & Rayment I. - X-ray Structures of the Myosin Motor Domain of *Dictyostelium discoideum* Complexed with MgADPBeFx, and MgADPAlF_4^-. *Biochemistry* **34(28)**, 8960-8972, 1995.
4. Smith C. A. & Rayment I. - X-ray structure of the Magnesium(II)ADP Vanadate complex of the dictyostelium myosin motor domain to 1.9 A resolution. *Biochemistry* **35**, 5404-5417, 1996.
5. Geeves M. A. & Holmes K. C. - Structural mechanism of muscle contraction. *Annu. Rev. Biochemistry* **68**, 687−728, 1999.
6. Fischer S., Windshuegel B., Horak D., Holmes K. C. & Smith J. C. - Structural mechanism of the recovery stroke in the Myosin molecular motor. *Proc. Natl. Acad. Sci. USA* **102(19)**, 6873-6878, 2005.
7. Koppole S., Smith J. C. & Fischer S. - Simulations of the myosin II motor reveal a nucleotide-state sensing element that controls the recovery stroke. *J. Mol. Biol.* in press, 2006.
8. Hayward S., Kitao A., Berendsen H. J. C. - Model-Free Methods of Analyzing Domain Motions in Proteins From Simulation: a comparison of normal mode analysis and molecular dynamics simulation of Lysozyme. *Proteins* **27**, 425-437, 1997.
9. Hayward S. & Berendsen H. J. C. - Systematic Analysis of Domain Motions in Proteins From Conformational Change: New Results on Citrate Synthase and T4 Lysozyme, *Proteins* **30**, 144-154, 1998.

10. Gaudin F., Lancelot G. & Genest D. - Search for rigid subdomains in DNA from molecular dynamics simulations. *J. Biomol. Struct Dyn.* **15(2)**, 357-367, 1997.
11. Rayment L., Rypniewski W. R., Schmidt-Base K., Smith R., Tomchick D. R., Benning M. M., Winkelmann D. A., Wesenberg G. & Holden H. M. - Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261**(5117), 50-58, 1993.
12. Schwarzl S. M., Smith J. C. & Fischer S. - Insights into the Chemomechanical Coupling of the Myosin Motor from Simulations of its ATP Hydrolysis Mechanism. *Biochemistry* **45**, 5830-5847, 2006.
13. Jorgensen W. L., Chandrasekhar J., Medura J. D., Impey R. W. & Klein M. L. - Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79(2), 926-935, 1983.
14. Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S. & Karplus M. - CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4, 187-217, 1983.
15. Oswald M. (IWR Heidelberg), personal communication.
16. Houdusse A., Szent-Györgyi A. G. & Cohen C. - Three conformational states of scallop myosin S1. *Proc. Natl. Acad. Sci. USA* **97,** 11238–11243, 2000.
17. Houdusse A., Kalabokis V. N., Himmel D., Szent-Györgyi A. G. & Cohen C. - Atomic Structure of Scallop Myosin Subfragment S1 Complexed with MgADP: A Novel Conformation of the Myosin Head. *Cell* **97,** 459–470, 1999.
18. Himmel D. M., Gourinath S., Reshetnikova L., Shen Y., Szent-Györgyi A. G. & Cohen C. - Crystallographic findings on the internally uncoupled and near-rigor states of myosin: Further insights into the mechanics of the motor. *Proc. Natl. Acad. Sci. USA* **99,** 12645–12650, 2002.

Chapter 7

Conclusions and future perspectives

In the present thesis the dynamical behavior of functionally and structurally different proteins is investigated, as well as the identification and contribution of essential motions to protein function.

Proteins are complex molecular machines synthesized as long strings of aminoacids that fold within minutes into a well-defined three dimensional structure that is functionally competent. Knowledge of the complicated three-dimensional architecture of many proteins has been gained from X-ray crystallography techniques that have provided in the last five decades an increasing number of conformations for cytoplasmic and membrane proteins. This technique made it possible to study in atomic detail the fine spatial interactions between different protein domains, as well as their position in space relative to each other. However, the picture emerging from these crystal structures was a static one, in which proteins were believed to be rigid structures. Contrary to this belief, results coming from more sophisticated experimental and theoretical studies have clearly shown in the past years that proteins, in their three dimensional configurations are very dynamic systems constantly undergoing atomic fluctuations and transitions between slightly different configurations separated by small energy barriers on the energy surface. These transitions are very fast on a nanosecond time-scale, are small amplitude motions and can be captured with advanced experimental techniques in terms of their timescale, but not in atomic detail. Some proteins exist in many stable conformations that are drastically different from one another and their function is related to the change from

one conformation to another. The function of such proteins involves a large, activated conformational change that is characterized by the overcoming of a high energy barrier and large amplitude motions. These transitions are inaccessible to experimental techniques in atomic detail due to the high thermodynamic instability of the transition state. And while crystallography provides invaluable information on equilibrium end-states and protein intermediate structures, the experimental characterization of transition pathways is much more difficult to obtain due to their non-equilibrium nature. Therefore, simulations techniques of both short timescale transitions and large conformational changes have been proven as being powerful tools in the analysis of protein motions in the past.

In spite of the accumulating knowledge about protein dynamics, there is still a lack of understanding of the role of protein motions in protein function. One question is whether all protein motions contribute together, as a sum to the function of a protein, or whether slow protein motions, involving large portions or domains moving together in a correlated manner are the dominant motions during function. Another interesting question is whether the dominant motions can be clearly separated from the background, "noise-like" ones and if so, are they due to the correlated movement of flexible large domains or are they a consequence of concerted motions of rigid parts of the proteins. The present thesis addresses these questions by means of theoretical approaches and it makes use of trajectories generated using Molecular Dynamics (MD) simulations and Conjugate Peak Refined (CPR) methods to investigate protein motions.

In order to separate the significant protein motions from the non-significant ones, especially when it comes to complex proteins that undergo complicated large-scale conformational changes one challenge consists in reducing the high dimensionality of the configurational space as protein dynamics represented in such a space is computationally demanding. Therefore, a first step towards an effective analysis of such conformational transitions in macromolecules requires the extraction of the dominant motions by lowering the dimensionsionality. Two different

dimensionality reducing techniques were applied and compared in this thesis, namely Principal Component Analysis (PCA) and Sammon Mapping. The two methods were firstly used to analyze four different protein transition pathways of varying complexity obtained by using either the CPR method or constrained MD to test whether they are really effective in extracting the essential protein motions from the simulated pathways. The first analyzed case involved the study of the return-stroke in the myosin contraction cycle. In this case, both applied methods revealed that this conformational change is clearly dominated by a simple rotation of a rigid body. The next analyzed case was the T→R conformational change in hemoglobin where two main quaternary transitions could be identified. In contrast, in the cases of the unfolding transition of Staphylococcal nuclease (SNase) or the signaling switch in Ras p21, which are both more complicated conformational transitions, only Sammon Mapping was able to identify the distinct phases of the transitions. For a medium-sized protein, applying projection methods often corresponds to reducing a problem with a few thousand degrees of freedom to one with less than three. The main goal when reducing the number of degrees of freedom in the simulated systems is to preserve as much information as needed in order to accurately describe the dynamics of native proteins. It was shown here with this analysis that even beyond the harmonic approximation, protein dynamics is dominated by a limited number of collective coordinates. PCA and Sammon Mapping can provide simplified descriptions of the dynamic information obtained using different simulation techniques that cannot be provided by experimental techniques in a straightforward manner. A reduced representation of protein flexibility could thus be obtained. The spectrum of examples examined in this thesis demonstrates how judicious choice of PCA and/or Sammon Mapping can enable the identification, extraction and analysis of the sequence of small transitions (events) that make up a large conformational. Therefore, the results of this study show that dimensionality reducing methods, if chosen well, can provide important information about the essential motions governing conformational transitions.

Conclusions and future perspectives

One major challenge in structural molecular biology is to understand the large conformational change of myosin by filtering out the non-essential motions. Thus, an analysis of the return stroke in the myosin cycle was thoroughly performed in the frame of this thesis. The fact that large amplitude motions in a protein can be captured by PCA in only a few principal motions has lead to the identification of the functional motion involved in the recovery stroke of myosin. In this work, by analyzing different MD trajectories of the two existent crystal structures of myosin head (corresponding to beginning and end of the recovery stroke step along the Lymn-Taylor cycle), the previously postulated mechanochemical coupling regarding a possible communication mechanism between the active sites could be confirmed. The principal motions governing the proposed two-phase mechanism behind the recovery stroke are extracted by performing PCA and they are also consistent with experimental data available from mutational studies.

Because MD simulations of large proteins are short (up to few ns due to computer power limitations) compared to the time needed for a large-scale conformational transition like the recovery stroke step (~1ms), the functional motions of interest are embedded in stochastic fluctuations. It was shown in this thesis that, when the primary elements involved in a given structural change can be defined up to some degree, one can extract their functional motions by filtering the relevant principal motions with the help of the Involvement Coefficient analysis. When the implicated sub-fragments are relatively small their motions occur on much faster time-scale than the overall transition. This was clearly seen here in myosin for several sub-fragments whose functional motions undergo stochastic oscillations on the nanosecond time-scale. While these oscillations do not have the full amplitude of the motion that these elements undergo in the complete recovery stroke, the PCA/Involvement-Coefficient approach is able to distinguish whether these elements participate in a functional motion or not. The present analysis of myosin MD trajectories has also revealed that the principal motions implicated in the recovery stroke are dependent on the presence of ATP. For the MD trajectories in which the

nucleotide is absent the functional motions become less marked. This suggests that the ATP induces small but essential changes in the dynamical behavior of the protein (results also experimentally confirmed), thus permitting the achievement of the recovery stroke.

Another aim of this thesis was to identify rigid domains of myosin, whose locally-restricted motions may determine the overall conformational change that myosin undergoes during muscle contraction. With the present analysis, several rigid elements previously observed as playing key roles in the large conformational change between pre-recovery stroke and the post-recovery stroke conformation could be identified and seem to confirm once again the coupling mechanism present behind the recovery stroke step of myosin.

To conclude, the methods described in this thesis used either to reduce the dimensionality of a system or to extract the essential motions of a protein or even to find rigid and correlated domains constitute a realistic and easy to apply strategy when one is interested in protein dynamics from simulation-generated trajectories. Taken together, the results presented in this thesis show the successful applicability of two dimensionality reducing methods to large conformational changes in proteins. These methods not only provided valuable information regarding the recovery stroke step in the myosin cycle, which represents a major contribution towards understanding in atomic detail the muscle contraction mechanism, but also confirmed their suitability in analyzing and dissecting the dynamical transitions of some proteins, thus emphasizing the importance of theoretical studies in complementing the experimental observations.

Future perspectives

Computational techniques are nowadays widely used for the study of conformational properties of biological macromolecules, and their range of application will only grow in the future. After the refinement of experimental structures to the *ab initio* folding of proteins, computer simulation techniques have proven to be valuable tools that can complement insights obtained from experiment. Also the reverse order could constitute an advance in science due to the fact that computer modeling and simulation when verified by direct connection to detailed experimental work could be improved.

In the present thesis no effort has been made towards the application of PCA and/or Sammon Mapping to other known large conformational changes in order to get a clear statistical picture of the suitability of these methods to all transitions in general. This is subject to future studies. Also, even though a quite extensive analysis of myosin motions during the recovery stroke step has been performed in this thesis, the complexity of the mechanochemical coupling between the myosin motion along actin filaments and ATP hydrolysis still remains to be investigated as not all questions have been answered. For instance, many aspects of the domain motions relative to each other, as well as the correlation between the ATP binding/hydrolysis and myosin conformational changes during muscle contraction are subject to further investigation and analysis. Also, the canonical correlation analysis performed here with the aim of detecting if motions performed during the two phase mechanism of the recovery stroke are determined by a specific rigid domain motion that may or may not be correlated did not yield significant results. With this kind of analysis, clear correlations between well defined domains could not be detected. A future, more extended analysis of the already identified clusters may reveal important motions which may be in a not known yet way related to the function of myosin.

When talking about myosin and its flexibility and dynamics, despite the fact that it is one of the most studied proteins, there remains much to be investigated and understood.

Appendix

Analyzing Essential Motions of Myosin Through a Correlation Analysis

A thorough understanding of the structure-function relationship of a protein requires detailed information of its biological function, as well as knowledge of its primary structure (amino acid sequence) or a description of secondary, tertiary or quaternary structural features available through physical methods such as X-ray crystallography. Identification of functional domains and/or conserved regions of the protein may reveal important details regarding its mechanism of function. Therefore, this aspect has been the focus of many theoretical studies in the structural computational field. More interestingly however, and more difficult to access experimentally is the concerted and/or correlated movement of groups of atoms, thereby defining a functional moving domain. In this Appendix, a method (Canonical Correlation Analysis) for interpreting structural data and analysis of functional motions was used in the attempt to identify possible correlated atomic motions, eventually leading to the identification of correlated-moving domains or just clusters of atoms. The method was applied on MD trajectories of myosin II.

A.1 Canonical Correlation

Since early 90's, atomic positional fluctuations and correlated motions between atom pairs, as observed in MD simulations of proteins, have been analyzed to further the understanding of a variety of dynamical properties of proteins and their biological functions.[1-3] These analysis include identification of groups of atoms moving in concert, interrelationship between different crystallographic conformers, domain-domain communication[4] and conformational changes occurring in dynamics of different conformers of the same protein. Usually a standard correlation analysis is used to extract more information about atomic movements in MD simulations by extrapolating the motions along directions in which the major changes should occur. Nowadays different methods have been used to determine correlations in proteins[5-7]. Given that with the classical cross-correlation (or Pearson correlation[8]) method the perpendicular correlations are basically lost[3,5,9], we tested a canonical correlation approach (which avoids the above mentioned limitation) in order to observe if this type of correlations are representative or not for myosin. The method, through which the canonical correlation coefficient between two variables (atoms belonging to the same or different domains) is determined, has been previously described in Ref. 10 and 11, and it is already implemented in the program called TECOR which was also used here[6]. A schematic description of this program is presented here in Fig. A.1, and a short overview of the method behind the above mentioned algorithm is following it. This analysis consists in calculating the canonical-correlation coefficients in order to identify correlated and/or concerted atomic motions.

Correlations between pairs of atoms or between groups of atoms can be determined based on a canonical statistical analysis used for comparing different groups of variables, in our case atoms[12]. Here we consider only correlations between pairs of atoms belonging to the same group.

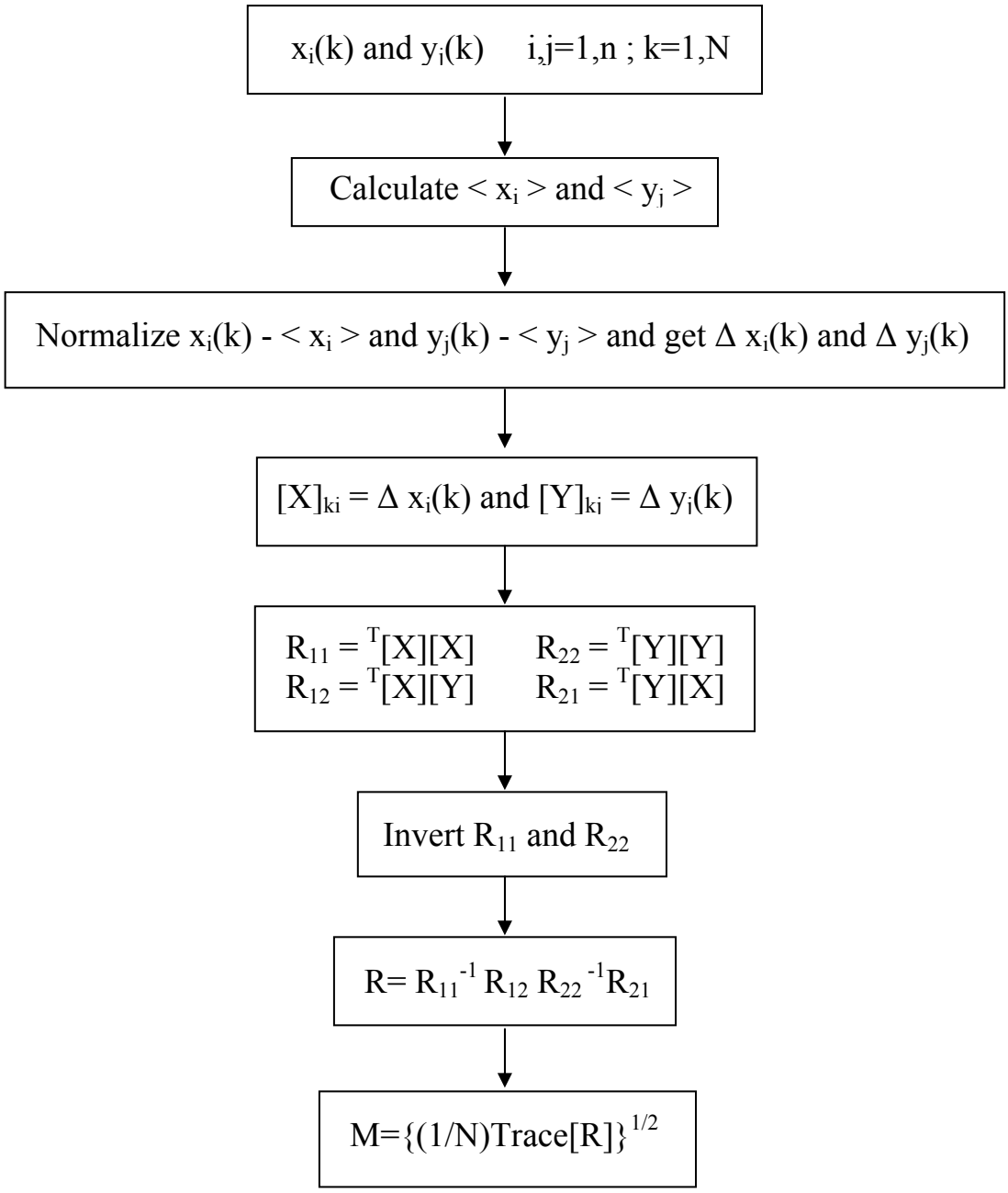


Fig. A.1 Overview of the method for calculating the canonical correlation coefficient between two variables x and y sampled by N values. Figure taken from Ref. 6.

If our group contains *n* atoms, then the variables are the normalized fluctuation components, δu_i^k, defined by:

$$\delta u_i^k = \frac{\left(u_i^k - \langle u_i^k \rangle\right)}{\left[\left\langle \left(u_i^k - \langle u_i^k \rangle\right)^2 \right\rangle\right]^{\frac{1}{2}}} \tag{A.1}$$

where $u_i^k = x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, z_1, z_2, \ldots, z_n$, for $k = 1,2,\ldots,n, n+1, \ldots 3n$ if i labels the group (of n atoms). Each variable represents an N-dimensional vector, N being the number of the simulation frames. The angular brackets represent averages over N. This procedure is similar with the calculation of the normalized covariance matrix which measures how/if the movement of a given residue i correlates with that of residue j.[5] The elements of this matrix are also known as being the cross-correlation coefficients (see Chapter 2 for more details) and they vary between –1 (completely anticorrelated motion) and +1 (completely correlated motion). When a coefficient is close to 1 then the two residues are highly correlated, i.e. they concertedly move in the same direction in space and when it is closer to –1 then the two residues are anti-correlated, i.e. they are coupled but move in opposite directions. Unlike the covariance matrix elements, these coefficients do not bear any information about the magnitude of the motion; therefore, it may be that small local oscillations are expressed by the same correlation coefficient as large-scale collective motions[3,5]. This matrix will only reflect correlation of displacements along a straight line. When the coefficients tend to be zero then the two residues are uncorrelated. This means for the cross-correlation analysis either that they do not move at all, or that the two atoms have fluctuations of the same period and in the same phase but the displacements are oriented along perpendicular lines. The expressions defining the variables δu_j^l corresponding to a group j are identical to the ones for group i. For our case $l=k=3n$. The correlation coefficient between any pair of variables related to atoms i or j is an element of one of the matrices:

$$R_{ii}(k.l) = \left\langle \delta u_i^k \delta u_i^k \right\rangle \tag{A.2}$$

$$R_{jj}(k.l) = \left\langle \delta u_j^k \delta u_j^k \right\rangle \tag{A.3}$$

$$R_{ji}(k.l) = \left\langle \delta u_j^k \delta u_i^k \right\rangle \tag{A.4}$$

$$R_{ij}(k.l) = \left\langle \delta u_i^k \delta u_j^k \right\rangle \tag{A.5}$$

To suppress the problem present in the cross-correlation method related to the orientation dependence, new sets of variables (X_i^k, X_j^l) are defined. They are called canonical variables and are linear combinations of the initial variables δu_i^k and δu_j^l (are spanning the same sub-spaces):

$$\left\langle X_i^k X_i^l \right\rangle = \delta_{kl} \tag{A.6}$$

$$\left\langle X_j^k X_j^l \right\rangle = \delta_{kl} \tag{A.7}$$

$$\left\langle X_i^k X_j^l \right\rangle = q_k \delta_{kl} \tag{A.8}$$

$$\left\langle X_j^k X_i^l \right\rangle = q_k \delta_{kl} \tag{A.9}$$

where the q_k is called canonical coefficient and it can be shown that its square values are the eigenvalues of the square matrix $R = R_{ii}^{-1}R_{ij}R_{jj}^{-1}R_{ji}$.[12] Therefore, a correlation coefficient, M, can be defined as being an average correlation between group i and j:

$$M = \left[\frac{1}{N} Trace(R) \right]^{1/2} \tag{A.10}$$

For perfect correlated or anticorrelated motions we will have $M = 1$, whereas for uncorrelated ones $M = 0$ (see Fig. A.2).

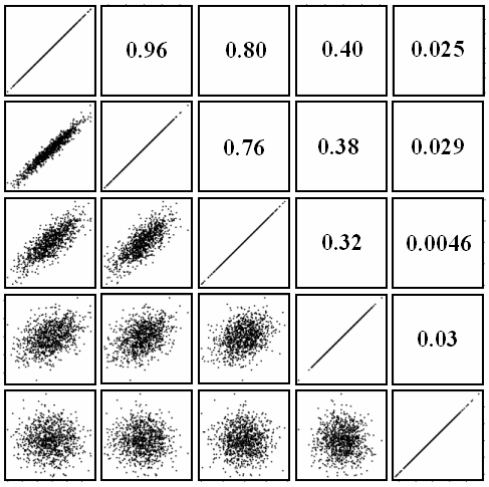


Fig. A.2 Schematic representation of a correlation matrix. On the diagonal M is 1 and the rest of the correlations (lower left side) are representing according to their coefficients (upper right side).

Different Molecular Dynamics (MD) trajectories were analyzed in order to detect if the previously identified motions along the recovery stroke step of myosin (see Chapter 5) are correlated in any way. Details about the generation of these trajectories are given in Chapter 2 and 5.

Because of technical reasons the algorithm cannot be used to perform the correlation analysis for all Cα atoms (myosin having ~750 residues) in all the frames (which for some MD trajectories used are more than 5500). Therefore we delete solvent-exposed loops that are very floppy; to reduce the number of atoms to 525 and in the same time we read every 10^{th} frame (10 ps) instead of reading every frame (i.e., we will have just 550 frames out of 5500).

A.2 Results

Correlated motions in biomolecules are often essential for their function[13] and they remain difficult to access experimentally. It seems that for the moment, Molecular Dynamics simulations is the most common method used for identify them. The obtained maps after performing the correlation analysis are shown in Fig. A.3 and they represent the correlations of the 525 selected C_α atoms along the MD trajectories corresponding to the pre-recovery (Fig. A.3A) and post-recovery conformation (Fig. A.3B). On the diagonal the self-correlations are represented, but of interest here are the off diagonal ones which can provide useful information about possible regions of myosin moving in a correlated manner. Three such regions could be depicted based on our analysis (shown by colored circles in Fig. A.4), which seems to be self-correlated but also inter-correlated with each other. The protein domains corresponding to these regions are almost identical for both analyzed trajectories, suggesting that these domains of myosin may play an important role during myosin's recovery stroke. This information results also when a simpler cross-correlation analysis is performed (see Fig. A.3C). In Fig. A.4 the regions of the protein corresponding to the correlated domains are highlighted in the same color with the

correlation matrix in order to better identify the possible connection to the previously identified events (see Chapter 5).

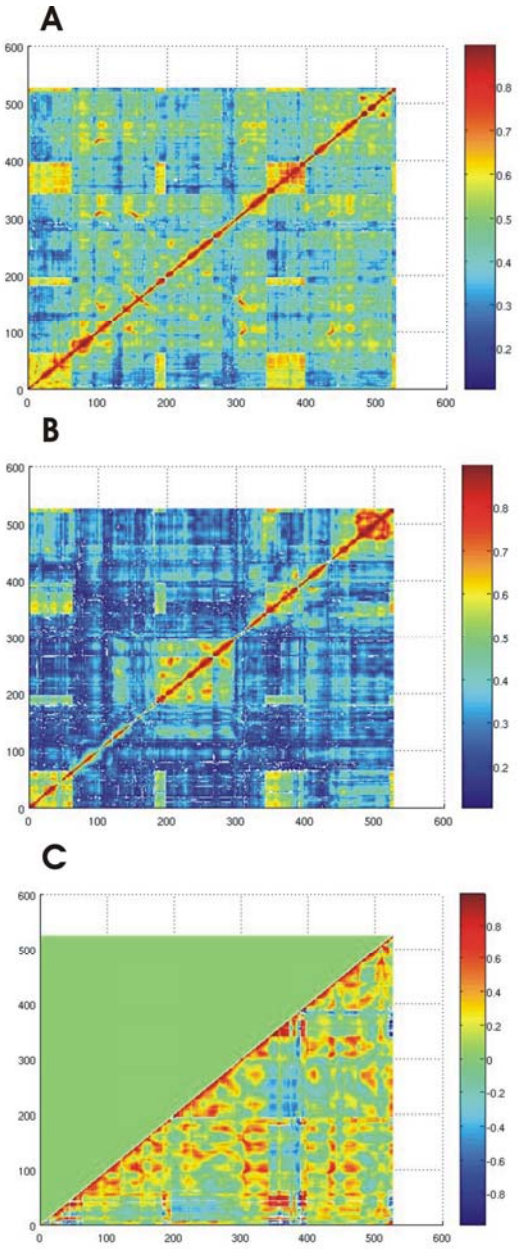


Fig. A.3 Correlation analysis. **A):** Canonical correlation matrix of the pre-recovery stroke conformation MD (State II, 525 residues taken in consideration). **B):** Canonical correlation matrix of the post-recovery stroke conformation MD (State III, 525 residues). **C):** Cross correlation analysis of State II (525 residues).

The question that now arises for this paradox is then: why correlated protein domains that are not known (up to date) to contribute significantly to the large conformational changes in myosin are resulting out of this analysis? A possible explanation may be the fact that the C-terminal of the relay helix is coupled to a certain extent to the wedge loop during the second phase of the coupling mechanism. Therefore they may cluster and thus move in a correlated manner (blue region in Fig. A.4). Another domain related to the movement of the converter seems to be the green block (see Fig. A.4), which was observed to strengthen the movement of the SH1 helix and that way indirectly determining the final rotation of the converter domain. Are these correlations just an artifact of the MD trajectory? Are they consistent along the path? In order to find an answer to these questions we split the MD trajectory of pre-recovery state (State II) into two halves and as shown in Fig. A.5 the same kind of correlations could be identified in both of them.

Even if possible plausible connections between these correlated domains may be made, a clear explanation why they are correlated could not be found; therefore this remains an open question to be addressed in future research.

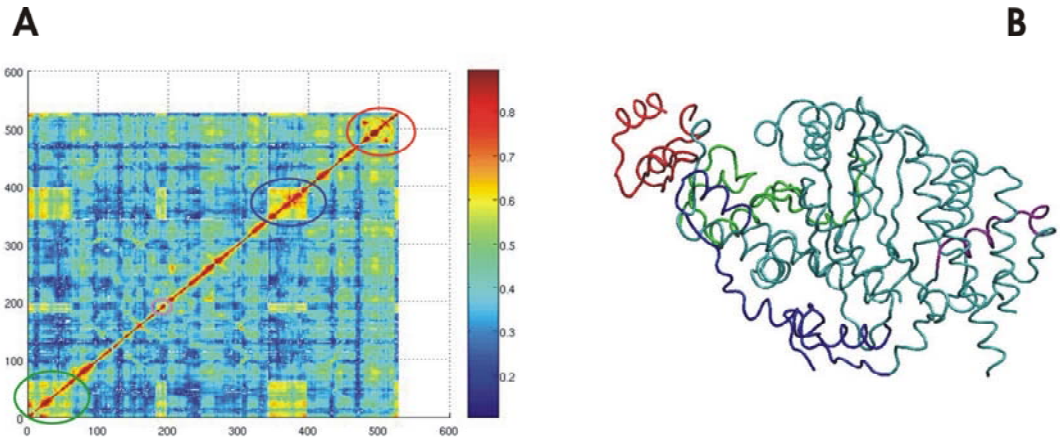


Fig. A.4 Canonical correlation analysis of State III. A): The correlated domains are highlighted with colored circles. B): The domains corresponding to the correlated areas are represented in the same color as the corresponding circles in panel A.

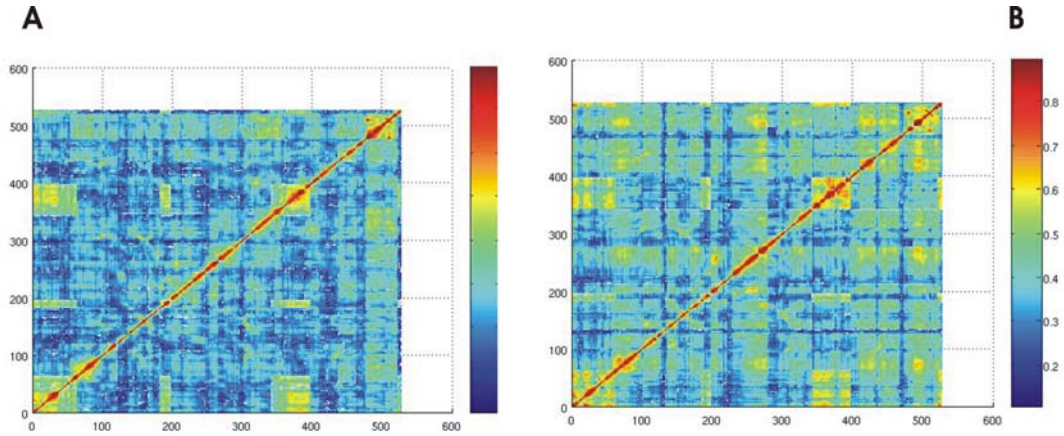


Fig. A.5 Canonical correlation of State II. **A):** First half of the trajectory is taken in consideration (first 2.5 ns). **B):** Second half is taken in consideration (last 2.2 ns).

In order to check if the reduction of data (made at the beginning for computational reasons) is having an influence onto the regions highlighted as correlated, we tested both reduction criteria we made. An even bigger reduction of the number of frames (from 10 to 100 for example) does not affect the results in any obvious way (not shown). Taking in consideration all myosin Cα residues (due to less trajectory frames) leads to different results (as observed in Fig. A.6). The previously correlated domains vanished. New regions are dominating the plot but like in the previous case (except the converter domain which is expected to appear in both of them and to be rigid) a real connection between this new domains and myosin function could not be found.

A.3 Conclusions

The present study was aimed at the identification of possible domains moving in a correlated fashion so that they may possibly produce and/or propagate essential motions contributing to the large conformational changes in the myosin cycle. We found correlated domains in the analyzed MD simulations of different myosin conformational states by using the Canonical Correlated Analysis method. The domains found to move in a correlated manner are different than the ones

expected based on previous analysis (see Chapter 5 and 6). However, these domains may still reveal important motions which may be in a not yet known way related to the myosin function. Even though the Canonical Correlation Analysis has been shown to be an useful tool in detecting correlations in proteins, its application to large conformational changes is not very clear based on our preliminary results shown here. Therefore, more detailed analysis on more proteins may be needed, but this is subject to future research.

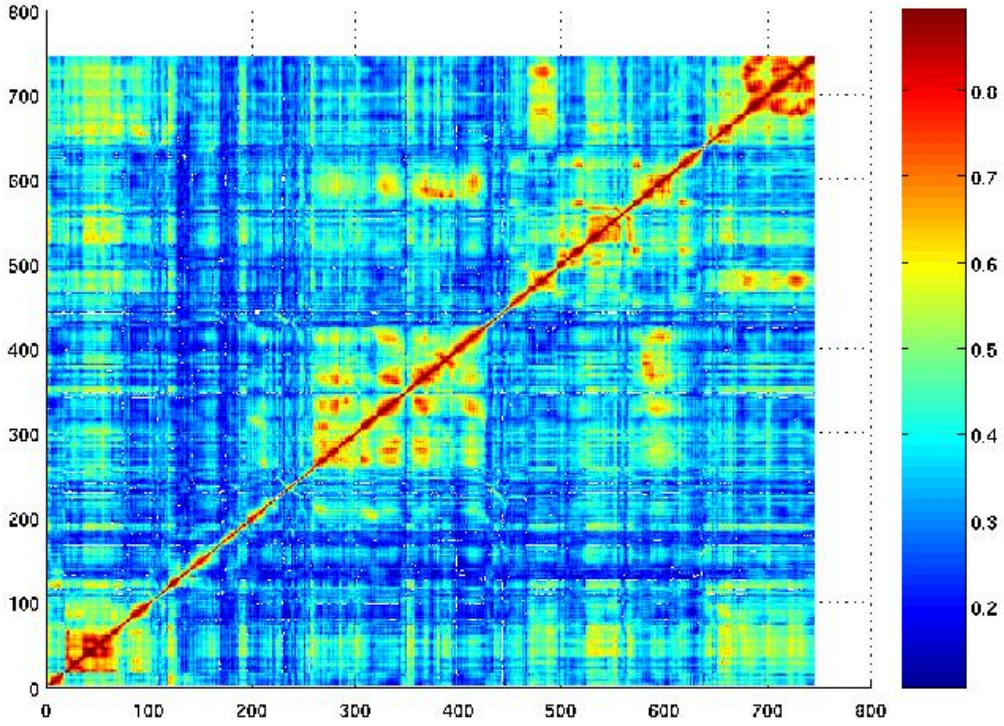


Fig. A.6 Canonical correlation matrix of State III when all residues are taken in consideration.

References

1. McCammon J.A. & Harvey S.C. – Dynamics of proteins and Nucleic Acids, *Cambridge University Press*, Cambridge, 1987.
2. Brooks C.L., Karplus M. & Pettitt B. M. – Proteins: a theoretical perspective of dynamics, structure, and thermodynamics, *Advan. Chem. Phys.* **71**, 1-259, 1988.
3. Ichiye T. & Karplus M. – Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations, *Proteins: structure, Function, and Genetics* **11**:205-217, 1991.
4. Harte W.E.jr, Swaminathan S., Mansuri M.M., Martin J.C., Rosenberg I.E. & Beveridge D. L. – Domain communication in the dynamical structure of human immunodeficiency virus 1 protease, *Proc. Natl Acad. Sci. USA* **87**, 8864-8868, 1990.
5. Hűnenberger P.H., Mark A. E. & van Gunsteren W.F. – Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations, *J. Mol. Biol.* **252,** 492-503, 1995.
6. Genest D. – Correlated Motions Analysis from Molecular Dynamics Trajectories: Statistical Accuracy on the Determination of Canonical Correlation Coefficients, *J. Comput. Chem.*, **20(14)**, 1571-1576, 1999.
7. Lange O. F. & Grubműller H. – Generalized Correlation for Biomolecular Dynamics, *Proteins* **62,** 1053-1061, 2006.
8. Barlow R. J. - Statistics: A Guide to the Use of Statistical Methods in the *Physical sciences*. John Wiley and Sons, Chichester, UK, 1989.
9. Arcangeli C., Bizzarri A. R. & Cannistraro S. – Molecular dynamics simulation and essential dynamics study of mutated plastocyanin: structural, dynamical and functional effects of a disulfide bridge insertion at the protein surface, *Biophysical Chemistry* **92**, 183-199, 2001.

10. Briki F. & Genest D. – Canonical Analysis of Correlated Atomic Motions in DANN from Molecular Dynamics Simulation, *Biophys. Chem*. **52**, 35-43, 1994.
11. Briki F. & Genest D. – Rigid-Body Motions of Sub-Units in DANN: A Correlation Analysis of a 200 ps Molecular Dynamics Simulation, *J. Biomol. Struct. Dyn.* **12(5)**, 1063-1082, 1995.
12. Saporta G. – Probabilités, Analyses de Données et Statistiques, Technip., Paris, Chapt. 9, 187-190, 1990.
13. Agarwal P.K., Billeter S.R., Rajagopalan P.T.R., Benkovic S.J. & Hammes-Schiffer S. - Network of Coupled Promoting Motions in Enzyme Catalysis, *Proc. Natl. Acad. Sci. USA* **99**, 2794-2799, 2002.

EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich, dass ich die Arbeit selbst verfasst habe und mich keiner anderen als der ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

Sidonia E. Mesentean,
Heidelberg, 07.05.2007