International School for Advanced Studies

Protein Internal Dynamics: Coarse-grained Investigation of the Structure-Function Relationship

Thesis submitted for the degree of
Doctor Philosophiæ

Candidate
Andrea Zen

Supervisor
Cristian Micheletti

9^{th} November 2009

Contents

Introduction

Customarily, the characterization of proteins proceeds according to the following tripartite scheme (canonical paradigm of genomics):

$$sequence \overset{(A)}{\rightarrow} structure \overset{(B)}{\rightarrow} function$$

The first two steps of this logical cascade have been extensively investigated using alignments methods that allow to identify similarities respectively in the sequence (Altschul *et al.*, 1997; Chenna *et al.*, 2003; Higgins & Sharp, 1988; Thompson *et al.*, 1994) and in the three dimensional structure (Holm & Park, 2000; Holm & Sander, 1996, 1999; Konagurthu *et al.*, 2006; Micheletti & Orland, 2009; Notredame *et al.*, 2000; Shatsky *et al.*, 2004b), among different proteins. Specifically, sequence and structural alignments are also extremely useful in the investigation of the sequence and structure relationship, link (A). A classic result regarding the relationship between sequence and structure is the fact that proteins whose sequence identity is above 30% (termed homologous) adopt the same global fold (Chothia & Lesk, 1986; Chothia *et al.*, 2003; Orengo & Thornton, 2005). In two homologous proteins, the regions with the highest degree of sequence similarity are usually well super-imposable by a suitable roto-translation of one of the molecules. The issue is therefore if the structural similarity in proteins necessarily require an underlying sequence similarity. This question was tackled detecting, through efficient structural alignment algorithms, the structural similarities in huge databases of proteins. The analysis of the results highlighted that, despite a clear correlation between the similarity in sequence and structure, the same fold is sometimes adopted also by proteins with negligible sequences similarity (Holm & Sander, 1994; Murzin *et al.*, 1995; Orengo *et al.*, 1997). This behavior is typically interpreted in terms of convergent evolution of proteins structure (Andreeva & Murzin, 2006; Banavar *et al.*, 2002; Chen *et al.*, 1997; Denton & Marshall, 2001a; Krishna & Grishin, 2004; Seno & Trovato, 2007).

A broader question regards the extent to which the function of a protein is determined by structure. This question represents a matter of general interest for genomics.

Introduction

It is known that the structural organization of a protein is extremely important to understand the molecular basis of the observed biological activities performed by proteins, as it represents an important source on additional information with respect to the sequence of amino acids. However the knowledge of a single crystal structure is not sufficient to understand completely the molecular mechanisms of the biological function, as recognized in the early crystallography experiments (Perutz & Mathews, 1966). The understanding of the structure function link can be profitably advanced taking into account the structural flexibility of proteins. It is known, indeed, that proteins posses a tendency to change conformation into forms that facilitate their biological function. The overwhelming majority of biological processes relies on the capability of proteins to sustain conformational changes so to selectively recognise, bind and process other molecules, being them proteins, nucleic acids or other chemical compounds.

The elastic properties of proteins clearly reflect the characteristics of their underlying free-energy landscape in the configurational phase space that is accessible in physiological conditions (temperature, pH, etc.). The different conformations that a protein can attain correspond to local minima for the free energy. According to the suggestion of Frauenfelder *et al.* (1991), these minima are hierarchically organized and separated by free energy barriers of various height, which are expected to control the transitions among the different biologically relevant states. Biologically relevant processes, that typically occur on time scales of the order of μs to ms, involve interconversion among conformational changes that often require collective movements of large groups of atoms. This property can be observed comparing the structures of the same protein crystallised in different conditions, *e.g.* in the unliganded and liganded state (Gerstein & Krebs, 1998).

The advancements observed in the last few years in computational techniques and resources, and the increased time resolution of advanced single-molecule techniques, have allowed a multi-timescale characterization of proteins' motions (Henzler-Wildman & Kern, 2007; Henzler-Wildman *et al.*, 2007a,b). These investigations have highlighted the connection between the motion at different timescales and to the functionally oriented conformational changes. These observations, arguably related to self-similarity of the free-energy landscape (Pontiggia *et al.*, 2007, 2008), suggest that the internal dynamics of proteins is "innately" predisposed to assist the conformational changes necessary to perform their biological function (Henzler-Wildman *et al.*, 2007b; Pontiggia *et al.*, 2008).

This functionally preferred directionality of the collective large-scale movements is encoded in the fold of the protein (Henzler-Wildman *et al.*, 2007a,b; Pontiggia *et al.*, 2007, 2008), as these motions are aptly captured by topology-based elastic network

models (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2001, 2002, 2004; Tirion, 1996). These models typically rely on a coarse-grained representation of the protein's native structure, and are oblivious to the detailed chemical composition of the protein. Despite this simplification, the validation of these models versus molecular dynamics simulations and experimental data demonstrates that they are suitable to provide the salient features of protein's internal dynamics, remarkably with a minimum computational effort.

These considerations indicate that, at least for a large class of proteins and enzymes, the relationship between structure and function presumably lies in the dynamics. This suggests that it is possible to take the internal dynamics of a protein as a proxy for the function. Following this approach, we want to investigate the connection between structure and dynamics. It is known that proteins with similar structure sustain similar large-scale movements, yet it has recently emerged (Capozzi *et al.*, 2007; Carnevale *et al.*, 2006; Zen *et al.*, 2008) that similar functional movements are shared by proteins with different architecture or topology. Note that this parallels the relationship between sequence and structure. As mentioned previously, the sophisticated interplay between sequence and structure has been extensively characterized thanks to the availability of sequence and structural alignment methods. By analogy, the availability of quantitative methods for comparing the functional-oriented dynamics in proteins would allow to take to a new level the investigation of the structure/function relationship.

My research activity has been principally oriented to the development of a pairwise alignment scheme that identifies groups of amino acids that undergo similar concerted movements in proteins (Zen *et al.*, 2008). The alignment method is based on a coarse-grained elastic network model and requires as input the sole proteins' native structures. *A priori* detection of structure and sequence correspondence is not used. This dynamics-based alignment, as far as we know, represents the first attempt in the direction of aligning proteins according to their large-scale functional movements. As for the other sequential and structural alignments, the dynamics-based alignment may be used in specific applications, for the identification or prediction of functionally important residues (Zen *et al.*, 2009).

The organization of the thesis is hereafter outlined.

The first chapter of the thesis presents an overview of theoretical and computational methods and models that are commonly and successfully used to characterize proteins flexibility. In particular normal mode analysis, principal component analysis

and coarse-grained elastic network models are introduced and the advantages, disadvantages, scope of applicability in consideration of protein dynamics and thermodynamics are discussed.

The second chapter of the thesis focuses on adenylate kinases, an important enzyme whose internal dynamics is known to play a major role for its biological functionality, that is the control of the energy charge of the cell. Two extensive molecular dynamics simulations of this enzyme, starting from different initial conformational states (open and closed), are analyzed and compared. The analysis is used to investigate the salient features of the free-energy landscape and the connection between the thermally-activated structural fluctuations in the open and/or closed state and the functionally oriented motions. Finally, the essential spaces obtained from the two trajectories are compared with the low-energy modes provided by a topology-based elastic network model, in order to illustrate the viability of these simplified approaches as effective tools to characterize proteins' internal dynamics.

The third chapter of the thesis reports on a study of the structural/dynamical properties of dimeric protein complexes, aimed at gaining a further insight into some of the general mechanisms that regulate protein-protein interaction. Protein interfaces have been widely studied in literature, and they have been extensively characterized in terms of their structural and chemical properties. The aim of our study is to investigate if, and to what extent, the protein internal dynamics at dimeric interfaces can be used to classify and group dimeric complexes. Our analysis has highlighted an intriguing relationship between the structural/functional aspects of the investigated interfaces and their elasticity. An attempt to rationalize this relationship in terms of entropic effects is finally reported.

The fourth chapter of the thesis describes the dynamics-based alignment, a novel pairwise alignment tool that we have developed to identify groups of amino acids that undergo similar concerted movements in proteins. Dynamics-based alignment requires as input the sole proteins' native structures, as it relies on the collective low-energy modes provided by elastic network models. This tool is next used to perform a dynamics-based alignment and grouping of a data set of more than 70 representative enzymes covering the main functional and structural classes. One third of the statistically significant dynamics-based alignments involve enzymes that lack substantial global or local structural similarities. The analysis of specific residue-residue correspondences of these structurally dissimilar enzymes in some cases suggests a functional relationship of the detected common dynamic features.

Finally the fifth chapter of the thesis illustrates how the dynamics-based alignment can be applied to identify functionally important residues in proteins. Specifically, it

is used for predicting protein regions involved in the binding of nucleic acids on the basis of comparative large-scale dynamics. The approach is first validated considering the canonical OB-fold domains, a motif known to promote protein-nucleic acid interactions. Protein regions consensually involved in statistically-significant dynamics-based alignments are found to correlate with nucleic acids binding regions. The validated scheme is next used as a tool to predict which regions of the AXH-domain representatives, a non-canonical sub-family of the OB-fold for which no DNA/RNA complex is yet available, are putatively involved in binding nucleic acids.

The material presented in this thesis has been the object of the following publications, on which chapters two, three, four and five are largely based.

- A. Zen, V. Carnevale, A. M. Lesk, and C. Micheletti.
  Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families.
  Protein Sci. **17**: 918-929 (2008)

- F. Pontiggia, A. Zen, and C. Micheletti.
  Small and large scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics.
  Biophys J. **95**: 5901-12 (2008)

- A. Zen, C. de Chiara, A. Pastore, C. Micheletti.
  Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains.
  Bioinformatics **25**: 1876-83 (2009)

- A. Zen, O. Keskin, R. Nussinov, C. Micheletti.
  Dynamical properties of protein-protein interfaces.
  in preparation

Chapter 1

Protein Internal Dynamics: a Theoretical/Computational Perspective

1.1 Introduction

Proteins are inherently flexible biopolymers. In thermal equilibrium they typically sustain concerted internal motions and experience conformational changes, sometimes of large-scale, involving a significant displacement of many amino acids. These conformational changes are often necessary for biological function, as the large-scale collective movements of amino acids usually accompany enzymatic activity, allosteric transitions, signal transduction and various other biological processes. A well-known example of the link between functionality and structural flexibility is given by hemoglobin, whose structure was one of the first to be obtained through X-ray crystallography (Bolton & Perutz, 1970; Fermi *et al.*, 1984; Frauenfelder *et al.*, 1988; Perutz & Mathews, 1966). The analysis of the crystallographic results showed that (i) hemoglobin can assume a number of different conformations (e.g. unliganded or bound to dioxygen) and that (ii) the apo[1] conformers were too compact to possibly allow the diffusion of dioxygen towards the heme pocket, thus implying that the molecule had to open substantially to allow dioxygen to reach the binding site.

Proteins manifest their intrinsic ability to undergo functionally relevant conformational changes on a wide range of time and space scales (Henzler-Wildman & Kern,

[1] The apo structure of a macromolecule refers to enzymes without a ligand or co-factor bound. It is opposed to the holo structure, which refers to an enzyme with its ligand or a co-factor bound.

2007). For instance, recent studies on adenylate kinase (Henzler-Wildman *et al.*, 2007a) show that a connections between the dynamics at the different timescales is present, from the relatively small-amplitude atomic fluctuations on the picosecond timescale to the large domain motions on the micro- and millisecond timescale. Moreover, the large-scale motions in the substrate-free adenylate kinase have a preferential directionality, following the pathways leading to configurations capable to perform the catalysis (Henzler-Wildman *et al.*, 2007b).

In the last decades the study and the characterization of proteins' internal fluctuations and the conformational changes in thermal equilibrium have been a matter of general interest. The starting point for almost any investigation on this kind is the molecular structure of the protein, and the main experimental techniques adopted to obtain this information are X-ray crystallography and nuclear magnetic resonance (NMR). Besides providing a snapshot of the position of the atoms in one or more protein conformers, both these techniques allows to quantify the internal dynamics of the proteins and are therefore a useful reference to test the results obtained adopting theoretical/computational approaches. In particular, in a X-ray crystal structure, each atom has an associated B-factor, or temperature factor, that represents the atom's individual fluctuation in position, mainly due to its thermal motion. With NMR spectroscopy an ensemble of conformations of the protein is provided, which can be compared in order to gain a representation of the protein dynamics. Of course this representation is affected by the limitation of the experimental method: the number of conformations obtained is quite low (usually a few dozens), and the rate measurements allow to investigate usually only slow internal motions, from μs to ms. Besides X-ray and NMR in recent years other experimental techniques, as the fluorescent resonance energy transfer (FRET) (Selvin, 2000; Somogyi *et al.*, 2000), have been used to elucidate interesting aspects of protein motion at the nanoscale. In this way the gap between the timescales investigated by molecular dynamics (MD) simulation, usually up to a few tens ns, and experiments, can be bridged.

In this chapter we will present an overview of some methods and models which are used to characterize the dynamics of proteins within a theoretical/computational approach. At first we will consider the normal mode analysis (NMA) and the principal component analysis (PCA) of a MD trajectory; two approaches that rely on atomistic treatment of the potential energy of proteins. Next we will review the salient properties of the potential energy landscapes of proteins, highlighting the advantages, disadvantages and scopes of applicability of the NMA and PCA. Finally coarse-grained approaches to model proteins' flexibility will be illustrated. In particular we will focus

on the β gaussian network model (βGM), the elastic network model used principally in this thesis.

1.2 Normal Mode Analysis

Normal mode analysis (NMA) is one of the standard computational methods adopted to identify and characterize the internal motion of biological macromolecules, and proteins in particular. It was first adopted for the study of molecules of biological interests, at an atomic level of detail, in (Brooks & Karplus, 1983; Go *et al.*, 1983; Levitt *et al.*, 1985). Its important attributes, which made it an interesting and useful complement to MD simulations, were immediately recognized.

NMA provides a simple formulation of the dynamics of an underdamped system with harmonic potential energy. As it will be discussed later in this chapter, the energy landscapes in proteins are much more complex than a simple harmonic function, in that it comprises several minima. Therefore, the quadratic approximation to the potential energy required by NMA is arguably valid only within each one the local energy minima. The restricted range of validity of this quadratic approximation imposes an obvious limitation to the amplitude of the investigated motions, because displacements from the structure associated to of the energy minimum have to be small enough that the approximation holds. Later in this chapter we will also discuss about the applicability for protein of the assumption that the motion is underdamped. Despite its limitations, NMA is a widely used method to investigate proteins' properties because the insight it offers is remarkable and comes at a very affordable computational cost. Moreover, some concepts of NMA are used by some more advanced approaches that will be considered later in this chapter.

To illustrate the salient properties of NMA we shall consider, as customary, a classical interatomic potential energy function. This energy function can be expressed as a function of the Cartesian coordinates of the atoms, or a function of other internal coordinates (bond lengths, bond angles and torsion angles). For simplicity, I will consider here a potential energy $V(\mathbf{r})$ that is a function of the Cartesian coordinates $\mathbf{r} = \{\vec{r}_1, ..., \vec{r}_N\}$ of the N atoms of the protein, where $\vec{r}_i$ is the Cartesian coordinate of the atom i. Note that the μ-th Cartesian component of the i-th atom correspond to the $(3i - 3 + \mu)$-th component of the $3N$-dimensional vector $\mathbf{r}$.

The potential energy $V(\mathbf{r})$ can be expanded as a Taylor series around a reference structure $\mathbf{r}^0 = \{\vec{r}_1^0, ..., \vec{r}_N^0\}$:

$$V(\mathbf{r}) = V(\mathbf{r^0}) + \sum_{i,\mu} \left(\frac{\partial V}{\partial r_{i,\mu}} \right)_{\mathbf{r}=\mathbf{r}^0} (r_{i,\mu} - r^0_{i,\mu}) + \frac{1}{2} \sum_{i,j,\mu,\nu} \left(\frac{\partial^2 V}{\partial r_{i,\mu} \partial r_{j,\nu}} \right)_{\mathbf{r}=\mathbf{r}^0} (r_{i,\mu} - r^0_{i,\mu})(r_{j,\nu} - r^0_{j,\nu}) + ... \tag{1.1}$$

where $r_{i,\mu}$ is the μ-th Cartesian component of the i-th atom in the current conformation of the protein, and $r^0_{i,\mu}$ in the reference conformation. Summations are taken over all the atoms and the Cartesian components.
If the reference structure $\mathbf{r}^0$ is a local energy minimum, the gradient of $V(\mathbf{r})$ calculated in $\mathbf{r}^0$ is zero. Furthermore, the energy can be defined relative to the reference structure, such that $V(\mathbf{r}^0) = 0$. It is clear from 1.1 that, for small displacements from the minimum-energy reference structure, the leading contribution of the potential energy is given by the second order term, and the higher order terms can be neglected.
The quadratic approximation of the energy landscape is therefore computed by first identifying the energy-minimum reference structure $\mathbf{r}^0$ (e.g. using a conjugate gradient minimization starting from an experimental crystallographic structure). The second derivatives of the energy function, calculated in $\mathbf{r}^0$, give the force constants $F_{ij,\mu\nu}$ of the interaction between the μ-th coordinate of i-th atom and the ν-th coordinate of j-th atom:

$$F_{ij,\mu\nu} = \left(\frac{\partial^2 V}{\partial r_{i,\mu}\partial r_{j,\nu}}\right)_{\mathbf{r}=\mathbf{r}^0} \tag{1.2}$$

and the harmonic potential function is, therefore:

$$E_P = \frac{1}{2}\sum_{i,j=1}^{N}\sum_{\mu,\nu=1}^{3} F_{ij,\mu\nu}(r_{i,\mu} - r^0_{i,\mu})(r_{j,\nu} - r^0_{j,\nu}) = \frac{1}{2}(\mathbf{r}-\mathbf{r^0})^T\mathbf{F}(\mathbf{r}-\mathbf{r^0}) \tag{1.3}$$

where the rightmost expression of 1.3 is in matrix form, being $\mathbf{F}$ the $3N \times 3N$ Hessian matrix of $V(\mathbf{r})$ and ^{T} the transpose operator.

The Newtonian equations of motion for N classical particles that interact with potential 1.3 are:

$$m_i\ddot{r}_{i,\mu} = -\sum_{j,\nu} F_{ij,\mu,\nu}(r_{j,\nu} - r^0_{j,\nu}) \qquad i = 1,...,N \qquad \mu = 1,2,3 \tag{1.4}$$

where m_i is the mass i-th atom. The equations can we written more synthetically in matrix notation:

$$\mathbf{M}\ddot{\mathbf{r}} = -\mathbf{F}(\mathbf{r}-\mathbf{r}^0) \tag{1.5}$$

where $\mathbf{M}$ is the $3N \times 3N$ diagonal matrix $diag\{m_1\mathbf{I_3},...,m_N\mathbf{I_3}\}$, being $\mathbf{I_3}$ the identity matrix of order 3.
The solution of this equation[1] is:

$$\mathbf{r} = \mathbf{r}^0 + \mathbf{M}^{-1/2}\sum_{i=1}^{3N}\mathbf{a}_i\xi_i\cos(\omega_i t + \phi_i) \tag{1.6}$$

[1] Note that, using the mass-weighted coordinates $\tilde{\mathbf{r}} = \mathbf{M}^{1/2}\mathbf{r}$, the equation 1.5 rewrites as: $\ddot{\tilde{\mathbf{r}}} = -\tilde{\mathbf{F}}(\tilde{\mathbf{r}} - \tilde{\mathbf{r}}^0)$, where $\tilde{\mathbf{F}} = \mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2}$ is the mass-weighted force constant matrix.

where the $3N$-dimensional normal vectors $\{\mathbf{a}_i\}$ are the solutions of the eigenvalue equation:

$$\tilde{\mathbf{F}}\mathbf{a}_i = \lambda_i \mathbf{a}_i \qquad i = 1, ..., 3N \tag{1.7}$$

being $\tilde{\mathbf{F}} = \mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2}$. The angular frequencies ω_i in 1.6 are related with the eigenvalues λ_i through the relation $\omega_i = \sqrt{\lambda_i}$. The amplitudes ξ_i and the phases ϕ_i depend on the position and velocity at time $t = 0$. It is worth mentioning that there are six zero eigenvalues for $\tilde{\mathbf{F}}$, which correspond to the rotational ad translational degrees of freedom of the overall system.

The combination of a $3N$-dimensional eigenvector $\mathbf{a}_i$ and its eigenvalue λ_i is called vibrational normal mode. The physical interpretation arise from 1.6: the eigenvalue λ_i determines the vibrational frequency ω_i along the direction determined by the eigenvector $\mathbf{a}_i$. Moreover, the set of $3N$ orthonormal eigenvectors $\{\mathbf{a}_i\}$ form a new basis set for the coordinates of the system. The change from the Cartesian coordinates $\mathbf{r}$ to the normal mode coordinates $\{x_1, ..., x_{3N}\}$ is obtained through the relation: $\mathbf{M}^{1/2}(\mathbf{r} - \mathbf{r}^0) = \sum_{i=1}^{3N} \mathbf{a}_i x_i$. Notice that the potential and kinetic energies have a much simpler form if written in terms of the normal mode coordinates:

$$E_P = \frac{1}{2}\sum_{i=1}^{3N} {\omega_i}^2 {x_i}^2 \qquad E_K = \frac{1}{2}\sum_{i=1}^{3N} \dot{x}_i^2 \tag{1.8}$$

since they are sums of squares of x_i and $\dot{x}_i$, and in E_K there are no mass coefficients. The expression of E_P also shows that a displacement from the reference structure along a normal mode i has an energetic cost that is proportional to its frequency squared ${\omega_i}^2$. Moreover, being the motion along each normal mode i oscillatory and periodic, with period $\frac{2\pi}{\omega_i}$ according to 1.6, the potential and kinetic energies associated to each normal mode are also oscillatory and periodic. From 1.8 and 1.6 we obtain that potential (and also kinetic) energy associated to normal mode i, averaged for a time interval $\tau \gg \frac{2\pi}{\omega_i}$, is $\epsilon_i \overset{\tau\to\infty}{\to} \frac{\omega_i^2 \xi_i^2}{4}$.

Through the simple diagonalization of a matrix, all the vibrational frequencies of a protein around an energy minimum and the direction of the oscillatory motion associated to each frequency are obtained. The vibrational frequency spectrum, i.e. the number of modes per frequency interval, is therefore immediately available. The set of vibrational normal modes provide a readily accessible simple description of the conformational motion of the protein (within the limit of the approximations). From a biophysical viewpoint, the most interesting modes are the ones associated to the lowest frequencies. In fact, it can be observed that the motion along them have a collective character (i.e. most of the amino acids of the protein undergo large concerted motions)

and can be used to give a description of functionally important motions experienced by the proteins, as it was recognized in Brooks & Karplus (1985) for the well known classical example of the hinge closing motion in lysozyme.

NMA has also been used, starting from the studies of Brooks & Karplus (1983); Levitt *et al.* (1985), to give an estimate of the thermal atomic fluctuations calculated from MD simulations and which contribute to the temperature factors observed in X-ray crystallography. However it has to be observed that the oscillatory motion of the model system described by 1.6 is deterministic, depending entirely on the initial conditions of the system (which determine the values of the amplitudes ξ_i and of the phases ϕ_i). It is possible to calculate, using 1.6, that the squared displacement from the reference position of atom i, coordinate μ, averaged over a time interval $\tau \to \infty$ (i.e. $\tau \gg \frac{2\pi}{\omega_i}$ for each non-zero frequency mode i), is:

$$\left\langle \left(r_{i,\mu}(t) - r^0_{i,\mu}\right)^2 \right\rangle_t \sim \frac{1}{2} m_i \sum_{k=1}^{3N} \left(\mathbf{a}_k^{(i,\mu)} \xi_k\right)^2 \tag{1.9}$$

where $\mathbf{a}_k^{(i,\mu)}$ indicates the component $(3i + \mu - 3)$ of the k-th eigenvector $\mathbf{a}_k$, i.e. the μ-th Cartesian component of the i-th atom. We stress however that the average atomic squared displacement in 1.9 does not describe equilibrium properties of the motion, because within this model system there is no notion of equilibrium and temperature, indeed the amount of the average displacement depends not only on the "potential" but also on the initial conditions of the system (through ξ_k). In order to estimate the atomic fluctuation it is therefore necessary to make further assumptions. In Brooks & Karplus (1983); Levitt *et al.* (1985) it is assumed that each internal normal mode has a time-average potential (and kinetic) energy of $K_B T/2$, where K_B is the Boltzmann's constant and T is the temperature of the system for which the estimation of the atomic thermal fluctuation is made. This assumption implies a condition on the amplitudes ξ_i, which have to satisfy the relation $\xi_i^2 = \frac{2K_B T}{\omega_i^2}$. Substituting this last expression in 1.9 yields to an expression for the average square displacement that is no more dependent on the initial configuration. The idea that underlies the assumption just introduced is that a thermalized system, in equilibrium at a given temperature T, with an energy of the form 1.8, would have an average energy of $K_B T/2$ for each degree of freedom, according to the equipartition theorem. As already highlighted, in the model of the system treated here there is no mechanism that establishes thermodynamic equilibrium, though it can be done by generalizing the equations of motion in a Langevin scheme. However, *a posteriori* the estimated quantities approximate, in their relative values, the experimental X-ray temperature factor and/or the thermal atomic fluctuations calculated from MD simulations.

Finally, we mention that, dealing with normal modes, it arises naturally the question if classical mechanics is a suitable framework for this investigations. Quantum effects could be important for modes with high-frequency. The criterium of applicability of classical mechanics (Kubo, 1967) is $\hbar\omega \ll K_B T$, that at $300K$ correspond to consider frequencies $\nu \ll 6\text{ps}^{-1} \sim 200\text{cm}^{-1}$, and modes with higher frequencies should be treated within a quantum mechanics framework. The transformation to normal mode coordinates is valid also in a quantum context; only the interpretation must be revised. However, from a pragmatic point of view, it has to be noted that usually the most interesting modes are associated to low-frequencies, for which the classical mechanics framework is expected to hold.

1.3 Principal Component Analysis of MD simulations

Atomistic MD simulations provide fundamental informations for the study and characterization of proteins' flexibility. Similarly to the way experiments are performed, in MD simulation it is possible to follow the time-evolution of a protein at a temperature T, in vacuo or in explicit solvent, reasonably close to equilibrium. Assuming that the MD simulation time is long enough to sample a suitable fraction of the relevant phase space, the thermodynamic properties of the system are calculated from averages over the conformations sampled in the trajectory. Nowadays typical MD simulations cover timescales of the order of several tens of ns, for proteins or protein complexes with hundreds of amino acids.

A common way to examine and characterize the protein dynamics is the employment of collective coordinates (Kitao & Go, 1999), that yield an optimal choice for the most relevant degrees of freedom describing proteins' fluctuations. The number of internal degrees of freedom of a protein with N atoms in a Cartesian coordinate space is $(3N-6)$, i.e. three degrees of freedom for each atom minus the six roto-traslational degrees of freedom of the overall protein. The appropriate choice of a smaller number of collective coordinates is sufficient to describe the important features of the dynamics of a protein, defining a low dimensional subspace in which the majority of the fluctuation of the protein take place.

A transparent way to identify this collective coordinate set is the principal component analysis (PCA) (Amadei *et al.*, 1993; Garcia, 1992; Garcia & Harman, 1996; Karplus & Kushick, 1981; Kitao *et al.*, 1991; Levy *et al.*, 1984) , that consists of the diagonalization of the fluctuation covariance matrix obtained from MD simulations. The first step is the elimination from the MD trajectory of the overall rotational and translational motion, that is irrelevant for the internal motion of the protein. It is

worth mentioning that this operation is conceptually not trivial, because a protein is not a rigid body, therefore there is a degree of ambiguity in this elimination. Different methods, that are not completely equivalent, can be used, but fortunately they usually do not give significantly different results. A widely used method is the following: take a conformation of the protein (for instance the first of the trajectory) as a reference structure, and structurally align all the other conformations of the trajectory respect to this reference structure. In this way the roto-translated conformations define a new trajectory. The average structure (or better, the conformation closest to it) of the new trajectory is defined as the new reference structure. The alignment of the conformations of the trajectory respect to the reference structure and the redefinition a new reference structure is then repeated iteratively until convergence is reached (usually very few iterations are sufficient). For definiteness, we shall assume hereafter that with the term trajectory we make reference to the trajectory where the roto-translational degrees of freedom have been eliminated.

Let us indicate the conformation of the system at time t of the MD trajectory with the generalized Cartesian variable $\mathbf{r}(t) = \{\vec{r}_1(t), ..., \vec{r}_N(t)\}$, being $\vec{r}_i(t) = \{r_{i,1}(t), r_{i,2}(t), r_{i,3}(t)\}$ the Cartesian coordinate of atom i at time t. Representing $\mathbf{r}(t)$ as a $3N$-dimensional column vector, the fluctuation covariance matrix (or second-moment matrix) is defined as:

$$\mathbf{C} = \left\langle \left(\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t\right) \left(\mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t\right)^T \right\rangle_t \tag{1.10}$$

being its elements:

$$C_{ij,\mu\nu} = \left\langle (r_{i,\mu}(t) - \langle r_{i,\mu}(t) \rangle_t)(r_{j,\nu}(t) - \langle r_{j,\nu}(t) \rangle_t) \right\rangle_t \tag{1.11}$$

where $\langle \cdot \rangle_t$ represents the time average over the configurations visited during the simulation. The symmetric matrix $\mathbf{C}$ can always be diagonalized by an orthogonal matrix $\mathbf{V}$:

$$\mathbf{V}^T \mathbf{C} \mathbf{V} = \Lambda \qquad \mathbf{V}^T \mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I_{3N}} \tag{1.12}$$

where Λ is the $3N \times 3N$ diagonal matrix $diag\{\lambda_1, ..., \lambda_{3N}\}$, and $\mathbf{I_{3N}}$ is the $3N$-dimensional identity matrix. The i-th column of $\mathbf{V}$, $\mathbf{v}_i$, is the eigenvector of $\mathbf{C}$ with eigenvalue λ_i. The complete set of these eigenvectors represent the (orthonormal) basis set for the collective coordinates $\{q_1, ..., q_{3N}\}$, defined as $q_i = \mathbf{v}_i \cdot (\mathbf{r} - \langle \mathbf{r} \rangle_t)$. Observe that the mean square fluctuation (MSF) of the trajectory projected along direction $\mathbf{v}_i$ is $\langle q_i^2 \rangle_t = \lambda_i$. The overall MSF experienced by the protein in the trajectory is the trace of the covariance matrix $Tr(\mathbf{C}) = \sum_i \lambda_i$. It is worth mentioning that $\mathbf{C}$ has six zero eigenvalues associated to the roto-translational degrees of freedom which have been eliminated from the trajectory.

As we have anticipated, this definition of the collective coordinates reflects a criterion of optimality for the representation of the global internal collective atomic fluctuations of the protein in terms of a set of directions. Following Garcia (1992), an optimal direction $\mathbf{m}$ (being this a $3N$-dimensional normal vertor) that represents the motion can be defined by using the following ansatz: minimize the mean square distances of the configurations $\mathbf{r}(t)$ normal to $\mathbf{m}$, such that most of the fluctuations will be then along $\mathbf{m}$. This amounts to minimize the following functional:

$$f(\mathbf{m}, \mathbf{y_0}, \alpha) = \left\langle (\mathbf{r}(t) - \mathbf{y_0})^2 \right\rangle_t - \left\langle [(\mathbf{r}(t) - \mathbf{y_0}) \cdot \mathbf{m}]^2 \right\rangle_t + \alpha \left(\mathbf{m} \cdot \mathbf{m} - 1 \right) \tag{1.13}$$

with respect to $(\mathbf{m}, \mathbf{y_0}, \alpha)$, where $\mathbf{y_0}$ is the $3N$-dimensional vector that optimally represent the structure of the protein in the trajectory, and α is a Lagrange multiplier that impose the normalization of $\mathbf{m}$. Minimizing with respect of α and $\mathbf{y_0}$ yields respectively to $\mathbf{m} \cdot \mathbf{m} = 1$ and $\mathbf{y_0} = \langle \mathbf{r}(t) \rangle_t$. Minimizing with respect to $\mathbf{m}$ and using this expression for $\mathbf{y_0}$ yields to the eigenvalue problem: $\mathbf{C}\mathbf{m} = \alpha \mathbf{m}$, where $\mathbf{C}$ is the fluctuation covariance matrix defined in 1.10. Note that this problem is equivalent to 1.12, and the eigenvectors $\{\mathbf{v}_i\}$ of $\mathbf{C}$ represent the optimal directions to represent the motion. The mean square fluctuation normal to the direction provided by the eigenvector $\mathbf{v}_i$, which has an associated eigenvalue λ_i, is given by $f(\mathbf{v}_i, \langle \mathbf{r}(t) \rangle_t, \lambda_i) = Tr(\mathbf{C}) - \lambda_i$, therefore the most representative directions are the eigenvectors with largest eigenvalue.

Without loss of generality, we can consider the eigenvalues λ_i sorted in decreasing order, so that the first eigenvector represents the direction where the largest fluctuation is observed, and so on. In Amadei *et al.* (1993); Garcia (1992) it was observed that a very limited number of eigenvectors is sufficient to define a space, called "essential space", where most of the overall fluctuation take place. The motion along this directions, usually called "essential dynamics" (Amadei *et al.*, 1993), is slow and collective, involving the concerted motion of many atoms simultaneously. The distributions of the projections of the motion along these directions loose progressively the Gaussian character as the simulation time increase. Differently, the projections of the motion along the remaining (non essential) vectors have a systematical Gaussian distribution that is much narrower than the speed of the essential modes, therefore this space can be considered "physically constrained".

The anharmonic character of the motion along some collective coordinates was observed already in Levy *et al.* (1984), using an analysis that is termed "quasi-harmonic" analysis (QHA). In QHA the fluctuation covariance matrix $\mathbf{C}$ obtained from MD simulation is used to construct an harmonic potential model. The idea behind the approach is the following: for a quadratic potential of the form 1.3, the equilibrium probability density for the atomic positions $\mathbf{r}$ is $P(\mathbf{r}) \propto \exp\left(-\frac{(\mathbf{r}-\mathbf{r^0})^T \mathbf{F} (\mathbf{r}-\mathbf{r^0})}{2 K_B T} \right)$, being K_B the

Boltzmann's constant and T the absolute temperature[1]. This yields a correlation matrix of the position fluctuation $\left\langle (\mathbf{r} - \langle \mathbf{r} \rangle)(\mathbf{r} - \langle \mathbf{r} \rangle)^T \right\rangle = K_B T \mathbf{F}'^{-1}$, where the average here is the canonical one and $\mathbf{F}'^{-1}$ is the pseudo-inverse[2] of the matrix $\mathbf{F}$ (which has six zero eigenvalues corresponding to the six roto-translational degrees of freedom). An expression for $\mathbf{F}$ can be obtained calculating the pseudo-inverse $\mathbf{C}'^{-1}$ of the covariance matrix $\mathbf{C}$ obtained form MD simulation, assuming that the simulation time is long enough that the time average over the trajectory is equivalent to the canonical average. The projections along the eigenvectors of the lowest non-zero eigenvalues of $\mathbf{F}$, that clearly correspond to the eigenvectors of the largest eigenvalues of the covariance matrix, highlight the anharmonic character of the motion.

Despite the difference in the spirit of PCA and of QHA, the first aimed at the research of the optimal degrees of freedom for the system, the second oriented to an estimation of a quadratic potential that takes into account of the complexity of the protein landscape, we note that both the methods ultimately rely with the projection of the motion along the eigenvectors of the covariance matrix $\mathbf{C}$, and in this respect they can be considered equivalent Kitao & Go (1999). In literature PCA is also termed "molecule optimal dynamics coordinates" (Garcia, 1992; Garcia & Harman, 1996) or "covariance analysis" (Hess, 2000). The method to obtain the collective coordinates is basically always the same, a part minor differences as considering the atoms mass weighted or not. The use of mass-weighted coordinates is maybe preferable in case of comparison with NMA[3].

[1] Note that, in writing the probability density, the temperature has been introduced and consequently it is implicit that the system has to be thermalized in some way. A possible underlying physical model which provide this equilibrium behavior is the Langevin dynamics (introduced later in the chapter).

[2] If a matrix $\mathbf{A}$ has one or more zero eigenvalues, then its inverse do not exist. However it is possible to define a matrix, pseudo-invese matrix, which has some properties of the inverse matrix. We are interested in cases where the matrix $\mathbf{A}$ is symmetric, therefore we can assume this property. According to the spectral theorem, it is possible to write $\mathbf{A}$ in terms of its eigenvalues α_k and eigenvectors $|k\rangle$, that we assume to be orthonormal, in the following way: $\mathbf{A} = \sum_k \alpha_k |k\rangle \langle k|$ (we are adopting the Dirac notation). The pseudo-inverse of $\mathbf{A}$ is defined as: $\mathbf{A}'^{-1} = \sum'_k \alpha_k^{-1} |k\rangle \langle k|$, where the prime in the summation indicates the sum over the k so that $\alpha_k \neq 0$. Obviously this matrix is the real inverse when $\mathbf{A}$ is invertible (i.e. all the eigenvalues are $\neq 0$). Note that: (i) $\mathbf{A}\mathbf{A}'^{-1} = \sum'_k |k\rangle \langle k|$, which is the projection onto the vectorial space defined by the eigenvectors of $\mathbf{A}$ relative to non zero eigenvalues, and (ii) the pseudo-inverse of the pseudo-inverse of $\mathbf{A}$ is $\mathbf{A}$, because $\sum'_k \alpha_k^{-1} |k\rangle \langle k| = \sum_k \alpha_k^{-1} |k\rangle \langle k|$.

[3] The covariance matrix in mass-weighted coordinates $\tilde{\mathbf{r}} = \mathbf{M}^{1/2}\mathbf{r}$ is $\tilde{\mathbf{C}} = \left\langle \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T \right\rangle_t$, where $\tilde{\mathbf{x}} = \tilde{\mathbf{r}} - \langle \tilde{\mathbf{r}} \rangle_t$. Note that $\tilde{\mathbf{C}}$ is related with the covariance matrix $\mathbf{C}$ in 1.10 through $\tilde{\mathbf{C}} = \mathbf{M}^{1/2}\mathbf{C}\mathbf{M}^{1/2}$, therefore the estimated force constant matrix in mass-weighted cooridinates $K_B T \tilde{\mathbf{C}}'^{-1}$ is $\tilde{\mathbf{F}} = \mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2}$, whose eigenvalues and eigenvectors are directly comparable with the outcomes of NMA (see 1.7).

1.4 NMA, PCA and Free Energy Landscapes of Proteins

In the previous sections we outlined two approaches that are commonly used to investigate the flexibility of proteins. The first approach, NMA, relies on the *a priori* assumption that the system dynamics can be described in terms of the harmonic underdamped oscillations. The second approach, of PCA and QHA, relies on the *a posteriori* analysis of the positions covariance matrix obtained from an all-atom dynamical trajectory of the protein at a constant temperature. The two approaches are clearly complementary; NMA has the considerable notable property to be very cheap computationally, but PCA has two great advantages: (1) to provide results that account for the deviations from the harmonic potential (for long enough MD trajectories) and (2) to allow for the transparent account for explicit solvent effects.

An informative study where the reliability of the outputs of NMA have been compared with MD is given in Janezic *et al.* (1995). Janezic *et al.* observed, comparing the lowest five vibrational frequencies from NMA for the minimized trajectory frames of a protein for $10ps$ by $0.1ps$ and for $1ns$ by $10ps$, that the molecule oscillates in one well, correspondent to what is called a conformational substate (Elber & Karplus, 1987; Frauenfelder *et al.*, 1988), for a very limited time span of the order of a few tenth of picosecond, and then jumps into another. It is often observed that the system returns to the same well after multiple transitions on the sub-picosecond scale, but never returns to the same well on longer time scale. The frequency values and the directions associated to the lowest five vibrational modes provided by NMA for different conformational substates tend to be quite in agreement[1]. Remarkably, this agreement is observed also between conformational substates corresponding to frames of MD trajectory separated by hundreds of picoseconds.

In the same study Janezic *et al.* have also carried out QHA of atomic MD trajectories of different time-length, from $2ps$ to $1ns$, and compared the outcomes, also with those from NMA. Interestingly, the lowest energy modes obtained from NMA and from QHA of the $1ns$-long trajectory are well consistent, with a degree comparable to the consistency of modes from outcomes of QHA for $1ns$-long trajectory and $200ps$-long trajectory.

[1] The degree of similarity between the directions of two sets of five modes (in this case the lowest five vibrational modes coming from NMA of two distinct conformational substates), was quantified through the following measure (Brooks *et al.*, 1995): Overlap $= \frac{1}{6}\sum_{i=1}^{5}\sum_{j=1}^{3}(\mathbf{v}_i \cdot \mathbf{w}_j + \mathbf{v}_j \cdot \mathbf{w}_i)$, where $\mathbf{v}_i$ and $\mathbf{w}_i$ represent the i-th eigenvector for respectively the first and the second set. The overlap is 1 if the two sets are identical, and is 0 if they are orthogonal. This scheme was developed mainly to compare the lowest three modes, which indeed dominate in the overlap, however it is observed that the ordering of the modes can change in different calculations, and this measure is less susceptible to re-ranks of the modes.

Furthermore, there is not a good agreement between QHA for the $1ns$-long trajectory and those of $40ps$ or shorter duration, indicating that these suffer from an insufficient sampling of the free energy landscape.

These resultes, besides clarifying the limits of NMA, highlight many interesting properties of the free energy landscape of a protein. It emerges in particular that this landscape is constituted by a great number of local minima, where proteins fluctuate for a time-interval of the order of picoseconds, before jumping to another minima. The shape of the local minima appears similar, given the agreement of the outcomes of NMA in different minima.

What about the characterization of the jumps between different minima? In Kitao *et al.* (1998) MD trajectories of $1ns$ are studied using PCA and the jumping-across-minima (JAM) model, introduced to separate the contributions to the internal motion arising from the structural fluctuations within the conformational substates and among them. If the overall MD trajectory sample a number M of conformational substates, spending in each substate k a fraction of time $f_t = \frac{t_k}{t}$ of the overall simulation time t, then the covariance matrix $\mathbf{C}$ of the overall trajectory can be expressed as:

$$\mathbf{C} = \mathbf{C}^{\text{intra}} + \mathbf{C}^{\text{inter}} \tag{1.14}$$

$$\mathbf{C}^{\text{intra}} = \sum_{k=1}^{M} f_k \left\langle \left(\mathbf{r} - \langle \mathbf{r} \rangle_k\right) \left(\mathbf{r} - \langle \mathbf{r} \rangle_k\right)^T \right\rangle_k \tag{1.15}$$

$$\mathbf{C}^{\text{inter}} = \sum_{k=1}^{M} f_k \left(\langle \mathbf{r} \rangle_k - \langle \mathbf{r} \rangle_t\right) \left(\langle \mathbf{r} \rangle_k - \langle \mathbf{r} \rangle_t\right)^T \tag{1.16}$$

where $\langle \cdot \rangle_k$ is the average taken over the conformations of the substate k, and $\langle \cdot \rangle_t$ is the average taken over all the conformations of the entire trajectory. $\mathbf{C}^{\text{intra}}$ represents the contribution to $\mathbf{C}$ arising from fluctuations within each substate, indeed it is the weighted average of the covariance matrices of each substate. $\mathbf{C}^{\text{inter}}$ represents the contribution to $\mathbf{C}$ arising from jumping among different conformational substates. Kitao *et al.* (1998) have shown that the protein motions (in the investigated length scale of $1ns$) consists of three types of collective modes: multiply hierarchical modes, singly hierarchical modes and harmonic modes. Interestingly, the multiply hierarchical modes are the leading contributions to the MSF, besides being only 0.5% of the modes. The intra-substate motion is observed to be nearly harmonic and mutually similar. The inter-substate motions are observed only in a small-dimentional subspace spanned by the axes of multiply hierarchical and singly hierarchical modes.

These kind on investigation has been further extended to longer time-scales in the studies of Pontiggia *et al.* (2007, 2008), where all-atom MD simulations in explicit

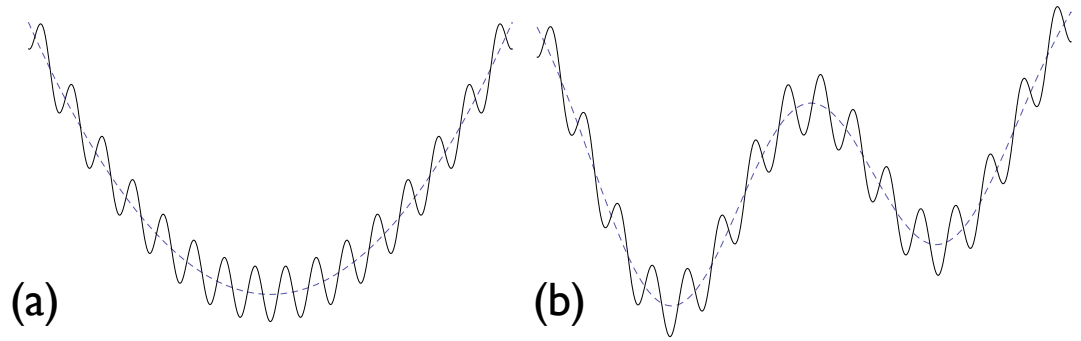

Figure 1.1: Schematic one-dimensional representation of the free energy surface, for: (a) one stable conformational state and many conformational substates; (b) two stable conformational states and many conformational substates. Dashed lines represent the smoothed out potential.

solvent for protein G (first study) and for adenylate kinase (second study) were followed for a simulation time of $400ns$ and $100ns$ respectively, and analyzed using PCA and JAM. In protein G it was observed that, while the quasiharmonic character of the free energy is found to degrade in a few ns, the essential modes display a very mild dependence on the trajectory duration. This property originates from a striking self-similarity of the free-energy landscape embodied by the consistency of the directions of the essential modes of the local minima and of the virtual jumps connecting them. Similar results are valid also for adenylate kinase, but this case is in some respects more interesting because this enzyme can interconvert spontaneously (i.e. in the absence of ligands) between the open and closed forms. The duration of the simulation was sufficiently long to reveal a partial conversion from the open to the close from, that proceeds through jumps between structurally different substates. It was observed that, despite the structural heterogeneity of the visited conformers, the generalized directions accounting for conformational fluctuations within and across the substates are mutually consistent and can be described by a limited set of collective modes. Part of this study is the subject of the second chapter.

The results reported clarify that the free energy landscape of proteins manifest a variegate multiscale structure, depending on the time and space scale used to probe it, and the conformations that are considered stable from a structural point of view comprise a great number of conformational substructures (see Fig. 1.1). At the picosecond timescale the proteins fluctuate around conformational substates and jump among them, with movements that involve mostly rearrangements of the sidechains and only

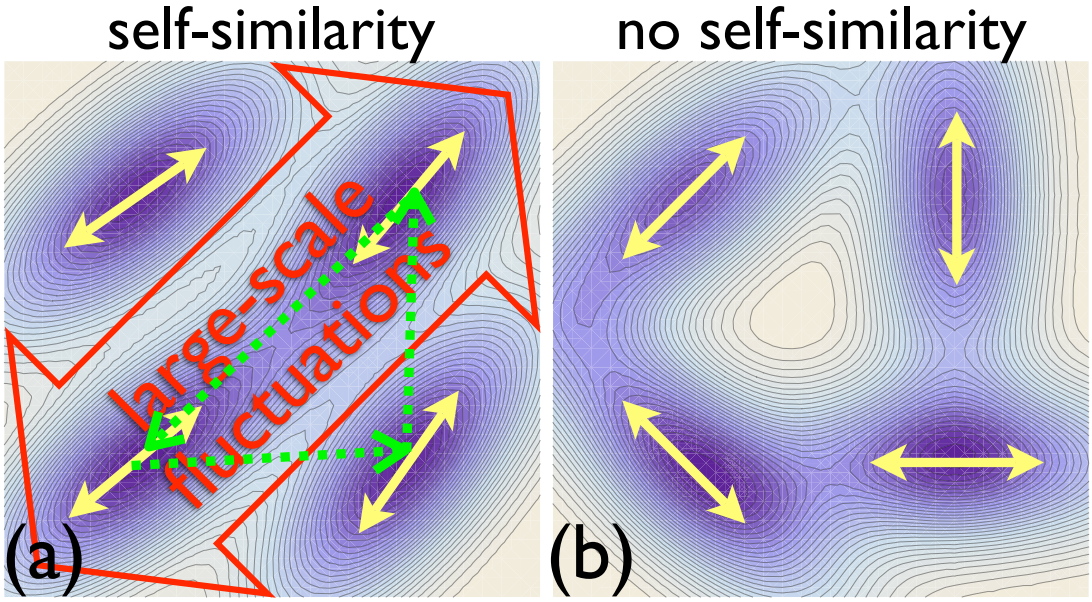

Figure 1.2: Schematic two-dimensional representation of a multiscale free energy surface, with self-similarity, in panel a, and without, in panel b. Yellow arrows represent the direction of the lowest energy models in each local minima, while the green arrows in panel a represents jumps among different minima.

minimally the sidechain. Over timescales of nanoseconds also the backbone of proteins manifests structural rearrangements, moving around stable conformations that seems to be local minima of a smoothed-out potential, however with a dynamics which is far from being vibrational. Over longer timescales proteins jump across these local minima and the quasi-harmonic character along the essential spaces disappears. Note that there is another extremely interesting property that emerge and link the description at the various levels of the multistructural free energy landscapes: their striking self-similarity (see Fig. 1.2). This self-similarity is intended in two ways: the consistency of the principal components of fluctuations calculated in different local minima, and the consistency of the essential direction of motion observed at different timescales (and consequently spacial scales).

Note that NMA is in its *a priori* expected range of validity only when used to describe local movements around a specific conformational substate. However numerous normal mode studies on proteins can reproduce experimental informations related to large-scale motions, such as domain motions. Therefore there is *a posteriori* evidence that NMA provides lowest-frequency vibrational modes with directions that are in

agreement with the directions of the movements observed over time- and length- scales beyond its *a priori* expected validity . This fact is clearly ascribable to the self-similarity of the free energy landscape. Nevertheless we want to remark that the dynamical behavior predicted by NMA completely fails in describing the long time large-scale motions observed in MD simulations.

There is another very interesting property of the large scale motions of proteins that needs to be highlighted: their collective nature. Indeed large scale displacements in proteins are the result of correlated motions of large groups of atoms, residues, sometimes entire subdomains. Concerning this point it should mentioned an interesting result obtained from Hinsen *et al.* (2000), where a MD trajectory of $1.5ns$ of a protein dimer in water solution has been decomposed into three contributions that have been shown to be almost independent: (1) the global motion of the backbone; (2) the rigid-body motions of the sidechains relative to the backbone; (3) internal deformations of the sidechain. Interestingly, the motion of the backbone and the connected rigid body motion of the sidechains gives the largest contributions to the overall large scale motion of the protein.

The collective nature of these motions lead to the following conclusion: in order to reproduce with a simple physical model some features of the long time large-amplitude motions of a protein we do not have necessarily to consider all the atoms in the protein, but it can be sufficient a coarse-gained description of the motion of the backbone. The usual approach is to take the C_α atom of each residue as representative of the overall residue, and the mass associated to each C_α atom is the mass of the overall residue represented. The interaction between the C_α atoms around a stable conformation is regulated by a coarse-grained smoothed-out potential, which can be derived from MD simulations trajectories or can be obtained more simply through coarse-grained models (that have been validated versus MD simulations). These coarse-grained models, the elastic network models, will be introduced later in this chapter, after having first provided a suitable physical framework able to deal with the main properties of this complex system, the Langevin dynamics.

1.5 Langevin Dynamics

Arguably, the simplest framework to describe the motion of a protein in thermal equilibrium is provided by the Langevin dynamics (Chandrasekhar, 1943; Doi, 1996; Wang & Uhlenbeck, 1945). Following this approach, for each C_α atom i of the protein we can write the following stochastic differential equation:

$$m_i\ddot{\vec{r}}_i = -\vec{\nabla}_i U(\mathbf{r}) - \gamma_i\dot{\vec{r}}_i + \vec{\xi}_i(t) \tag{1.17}$$

where $\vec{r}_i$ is the Cartesian coordinate of the C_α atom and m_i is its effective mass. Within this coarse grained description, the C_α atom represents the overall residue, therefore m_i is the sum of the masses of all the atoms in the residue i. The terms $-\vec{\nabla}_i U(\mathbf{r})$, $-\gamma_i \dot{\vec{r}}_i$ and $\vec{\xi}_i(t)$ in 1.17 represent respectively an external field of force, a friction force and a stochastic (or random) force that act on atom i. Note that the last two terms, the friction and the stochastic force, are intrinsically related, as they both contribute to establish of the thermal equilibrium. Let us consider in detail the three terms.

$U(\mathbf{r})$ is an effective (smoothed-out) potential that describes the effective interaction between residues in a protein. It is expressed as function of the Cartesian coordinates $\mathbf{r} = \{\vec{r}_1, ..., \vec{r}_N\}$ of all the N C_α atoms in the protein. The gradient of $U(\mathbf{r})$ respect to $\vec{r}_i$ yields an effective force $\vec{F}_i = -\vec{\nabla}_i U(\mathbf{r})$ for atom i.

The friction force $-\gamma_i \dot{\vec{r}}_i$ introduce a damping in the system, being systematically opposed to the motion and proportional to the velocity of the atom. The idea behind the introduction of this force is to implicitly include the average effective influence of the many particles that we are neglecting in our simplified treatment for the problem. Indeed, we are not explicitly including most the protein atoms nor the solvent. The physics behind the friction term is that the neglected particles interfere with the C_α atoms and the average effect is a decrease of the velocity[1]. As a further argument to justify the necessity to introduce a damping term, let us consider that, as we have previously explained, the energy landscapes of proteins comprise a huge number of local conformational substates. The jump from a conformational substate to a close one induces small displacements of the backbone atoms but implies sidechain rearrangements and crossings of the barriers that are smoothed-out in the effective potential $U(\mathbf{r})$.

The stochastic force $\vec{\xi}_i(t) = \{\xi_{i,1}(t), \xi_{i,2}(t), \xi_{i,3}(t)\}$ acting on atom i is used to introduce thermal fluctuations in the system. It is assumed that this stochastic force satisfy the following conditions (Chandrasekhar, 1943; Wang & Uhlenbeck, 1945):

$$\langle \xi_{i,\mu}(t) \rangle = 0 \tag{1.18}$$

$$\left\langle \xi_{i,\mu}(t) \xi_{j,\nu}(t') \right\rangle = 2\, K_B T\, \gamma_i\, \delta_{ij} \delta_{\mu\nu}\, \delta(t - t') \tag{1.19}$$

where $\langle \cdot \rangle$ is an average over an ensemble of realizations, δ_{ij} is the Kronecker delta and $\delta(\cdot)$ is the Dirac's delta function. The first condition establishes that the random noise

[1] It is worth mentioning that, for a system of coupled variables $\{x_1, ..., x_n\}$, the most general form of the friction force acting on x_i has the form: $-\sum_j^n \gamma_{ij} \dot{x}_j$. In our treatment we are neglecting the elements γ_{ij} with $i \neq j$ because, as we have highlighted, the friction is an effective effect, therefore the friction coefficients have to be determined in some way. We will see that they can be determined *a posteriori* from an MD trajectory (Hinsen *et al.*, 2000), provided that we do the assumption we have introduced.

does not have any preferential direction; the second condition specifies that $\vec{\xi}(t)$ is uncorrelated over time (i.e. it is a white noise signal) and its average amplitude satisfy the fluctuation-dissipation relationship, i.e. the noise add on average just as much energy to the system as is taken out by the friction term. The two conditions ensures, in the long run, the onset of canonical thermal equilibrium, so that the equilibrium probability $\mathcal{P}_{eq}(\mathbf{r})$ to observe the structure $\mathbf{r}$ of the protein is proportional to the Boltzmann's factor:

$$\mathcal{P}_{eq}(\mathbf{r}) \propto \exp\left(-\frac{U(\mathbf{r})}{K_B T}\right) \tag{1.20}$$

Note that, introducing the following notation:

$$\begin{aligned}
&\text{mass matrix} && \mathbf{M} = diag\{m_1\mathbf{I_3}, ..., m_N\mathbf{I_3}\} && (1.21)\\
&\text{friction matrix} && \mathbf{\Gamma} = diag\{\gamma_1\mathbf{I_3}, ..., \gamma_N\mathbf{I_3}\} && (1.22)\\
&3N\text{-dim. stochastic force} && \Xi(t) = \{\vec{\xi}_1(t), ..., \vec{\xi}_N(t)\} && (1.23)\\
&3N\text{-dim. gradient} && \nabla = \{\vec{\nabla}_1, ..., \vec{\nabla}_N\} && (1.24)
\end{aligned}$$

it is possible to represent the set of coupled stochastic differential equations 1.17 for the overall system, in the following simple in matrix form:

$$\mathbf{M}\ddot{\mathbf{r}} + \mathbf{\Gamma}\dot{\mathbf{r}} + \nabla U(\mathbf{r}) = \mathbf{\Xi}(t) \tag{1.25}$$

Given the stochastic nature of these Langevin equations, a solution for 1.25 is given by the probability distribution to find the system in a particular position of the phase phase, in function of the time and the initial conditions. The moments of the coordinates and of the velocities for this distribution have to satisfy 1.25.
An equivalent formulation of this problem is to consider directly the equation of motion for the probability distribution, which is called Fokker-Planck equation (Chandrasekhar, 1943; Risken, 1996; Wang & Uhlenbeck, 1945). In the particular case of 1.25, we have the Brownian motion of particles in an external field. The equation of motion for their distribution function in position and velocity space is the Kramers equation.

We assume that the potential $U(\mathbf{r})$ has a quadratic form $\frac{1}{2}(\mathbf{r}-\mathbf{r}^0)^T\mathbf{F}(\mathbf{r}-\mathbf{r}^0)$ centered around a reference conformation $\mathbf{r}^0$. This assumption is motivated by the fact that the projections of a MD simulation trajectory along the principal components have a nearly Gaussian character, if the considered time-length is smaller than a given value (e.g. $\sim 1ns$ for the protein G (Pontiggia *et al.*, 2007)). Note that the Gaussianity of these distributions is exactly what the equilibrium distribution 1.20 imply for a quadratic potential. The deviations from Gaussianity, in long MD simulations, is a

clear indication that the quadratic assumption is only a first order approximation, which, however, has the advantage of allowing for a transparent analytical treatment of 1.25.

The quadratic form of the potential yields to the following Langevin equation:

$$\mathbf{M}\ddot{\mathbf{r}} + \mathbf{\Gamma}\dot{\mathbf{r}} + \mathbf{F}(\mathbf{r} - \mathbf{r^0}) = \mathbf{\Xi}(t) \tag{1.26}$$

which has been solved in Lamm & Szabo (1986) (considering the equivalent formulation of the problem in terms of a Fokker-Planck equation).

In Langevin dynamics the motions are in general damped oscillations plus random diffusion. MD simulations show that the dynamics of proteins is severely *overdamped*: there are no periodic oscillations and the dominant aspect of the motion is constituted by the random displacements, with preferential movements towards the minimum. This means that the friction terms are the leading ones in equation 1.26. In this conditions we are allowed to neglect the kinetic term $\mathbf{M}\ddot{\mathbf{r}}$ and consider the overdamped Langevin equation:

$$\mathbf{\Gamma}\dot{\mathbf{r}} + \mathbf{F}(\mathbf{r} - \mathbf{r}^0) = \mathbf{\Xi}(t) \tag{1.27}$$

which has for solution the high friction (diffusion) limit of the solution of 1.26 (Lamm & Szabo, 1986). The Fokker-Planck equation corresponding to 1.27 is an equation for the distribution function of the positions, which is called Smoluchowski equation (Risken, 1996), and it has been solved in Hinsen *et al.* (2000); Kneller (2000).

It results, for the overdamped langevin equation 1.27, that the dynamical properties of the system can be described in terms of the eigenvalues and eigenvectors of the friction-weighted force constant matrix $\hat{\mathbf{F}} = \mathbf{\Gamma}^{-1/2}\mathbf{F}\mathbf{\Gamma}^{-1/2}$. Let us define $\hat{\lambda}_k$ and $\hat{\mathbf{u}}_\mathbf{k}$ as follows:

$$\hat{\mathbf{F}}\hat{\mathbf{u}}_k = \hat{\lambda}_k\hat{\mathbf{u}}_k \tag{1.28}$$

and assume that the eigenvectors $\hat{\mathbf{u}}_\mathbf{k}$ are orthonormal. Note that the roto-translational degrees of freedom imply that $\hat{\mathbf{F}}$ has six zero-eigenvalues, which are the transformed of the roto-translations.

Let us indicate with $\vec{x}_i = \vec{r}_i - \vec{r}_i^0$ the displacement of the of the i-th atom respect to its reference position $\vec{r}_i^0$.

The average correlation among the displacements $x_{i,\mu}(t)$ and $x_{j,\nu}(t')$ is:

$$\left\langle x_{i,\mu}(t)x_{j,\nu}(t')\right\rangle = K_BT\sum_k{}' \frac{\hat{\mathbf{u}}_k^{(i,\mu)}\hat{\mathbf{u}}_k^{(j,\nu)}}{\sqrt{\gamma_i\gamma_j}}\left(\frac{e^{-\hat{\lambda}_k|t-t'|}}{\hat{\lambda}_k}\right) \tag{1.29}$$

where the prime indicates the omission from the sum of the indexes k associated to zero eigenvalues, and $\hat{\mathbf{u}}_k^{(i,\mu)}$ represents the $(3i-3+\mu)$-th component of the 3N-dimensional eigenvector $\hat{\mathbf{u}}_k$, i.e. the μ-th Cartesian component of the i-th atom.

The auto-correlation of the displacement of atom i along the μ-th Cartesian component can be easily calculated from 1.29, and it yields:

$$\langle x_{i,\mu}(t)x_{i,\mu}(0)\rangle = K_BT\sum_k{}' \left|\mathbf{u}_k^{(i,\mu)}\right|^2 \left(\frac{e^{-\hat{\lambda}_k t}}{\hat{\lambda}_k}\right) \tag{1.30}$$

where $\mathbf{u}_k = \mathbf{\Gamma}^{-1/2}\hat{\mathbf{u}}_k$ and $\mathbf{u}_k^{(i,\mu)}$ is its $(3i-3+\mu)$-th coordinate. Note that $\hat{\mathbf{u}}_k^{(i,\mu)} = \sqrt{\gamma_i}\mathbf{u}_k^{(i,\mu)}$.
This expression elucidates the meaning the eigenvalues and eigenvectors of $\hat{\mathbf{F}}$: each eigenvalue $\hat{\lambda}_k$ is associated to a Brownian mode k of structural relaxation along the direction $\mathbf{u}_k$, with a relaxation time $\tau_k = 1/\hat{\lambda}_k$.

Using 1.29 it is easy to calculate the mean square displacement[1] (MSD), averaged over the initial conformation, in a time interval t of the i-th atom along the μ-th coordinate:

$$\left\langle (r_{i,\mu}(t) - r_{i,\mu}(0))^2 \right\rangle = 2\,K_BT\sum_k{}' \frac{\left|\hat{\mathbf{u}}_k^{(i,\mu)}\right|^2}{\gamma_i}\left(\frac{1-e^{-\hat{\lambda}_k t}}{\hat{\lambda}_k}\right) \tag{1.31}$$

and if the considered time interval is much longer than the relaxation times of each mode (i.e. $t \gg \tau_M$, where τ_M is the maximum relaxation time[2]), then the MSD is:

$$\left\langle (r_{i,\mu}(t) - r_{i,\mu}(0))^2 \right\rangle \overset{t\gg\tau_M}{\sim} 2\,K_BT\sum_k{}' \left|\mathbf{u}_k^{(i,\mu)}\right|^2 \left(\frac{1}{\hat{\lambda}_k}\right) \tag{1.32}$$

showing that the structural relaxation of the mode k, along $\mathbf{u}_k$, is inversely proportional to the eigenvalue $\hat{\lambda}_k$. This clearly highlights that the modes associated to the lowest non-zero eigenvectors are those that mostly describe the motion.

Observe that for a short time interval t (i.e. $t \ll \tau_m$, where τ_m is the minimum relaxation time) the MSD of the i-th atom is:

$$\left\langle (\vec{r}_i(t) - \vec{r}_i(0))^2 \right\rangle \overset{t\ll\tau_m}{\sim} \frac{6K_BT}{\gamma_i}t + \mathcal{O}(t^2) \tag{1.33}$$

Note that these quantities that can be measured from MD simulations, and used to provide an estimation for the friction constants γ_i. Using this method, Hinsen *et al.* (2000) proved that the friction constant can be well described by a linear function of the average density of the protein atoms inside a sphere of radius $15\mathring{A}$ centered around the atom of interest. In a typical globular protein, the average density within this length

[1] Observe that the displacement $r_{i,\mu}(t) - r_{i,\mu}(0) = x_{i,\mu}(t) - x_{i,\mu}(0)$.

[2] Clearly not considering the modes relative to zero eigenvalues, that are not included in the summation and that would have an infinite relaxation time.

scale is quite uniform, therefore the variations in the values of the friction constants are due to surface effects. Atoms close to the surface are indeed surrounded in great part by water, whose density is much smaller than that of the protein itself.

Finally we can use 1.29 to get an expression for the covariance matrix $\mathbf{C}$ introduced in the previous section. It yields[1]:

$$C_{ij,\mu\nu} = \langle x_{i,\mu}(t) x_{j,\nu}(t) \rangle = K_B T \frac{\sum'_k \hat{\lambda}_k^{-1} \hat{\mathbf{u}}_k^{(i,\mu)} \hat{\mathbf{u}}_k^{(j,\nu)}}{\sqrt{\gamma_i \gamma_j}} = K_B T F'^{-1}_{ij,\mu\nu} \tag{1.34}$$

This result is expected, because of the form of the equilibrium distribution 1.20 for a quadratic potential.

Concluding it should be mentioned that, while the above Langevin scheme is highly transparent and amenable to analytical treatment, it is not adequate to capture a number of salient features of protein's internal dynamics, for which more sophistical theoretical schemes have been devised (Kneller & Hinsen, 2001, 2004; Kou & Xie, 2004; Min *et al.*, 2005).

1.6 Elastic Network Models

In the previous sections we have seen how the large scale amplitude motions in the proteins rely on the shape "global" of the multistructural free energy landscape, which presents a striking self-sililarity at various level of detail. The "global" shape of the potential is often approximated through a quadratic function of the coordinates, around a reference structure. This approximation, beside leading to a great simplification of the problem, yields results that are in some respects in good agreement with the phenomenology observed in real and computational experiments. We have already mentioned that the effective matrix of interaction of the quadratic potential can be estimated from a MD trajectory (QHA approach). In this section we will introduce a different approach to estimate this matrix without the use of MD simulations. This approach consists in the use of elastic network models (ENM) (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2001, 2002, 2004; Tirion, 1996): "native centric" models which rely on simplified force fields and have proved useful to reproduce collective modes in proteins.

[1] Recognizing that $\sum'_k \hat{\lambda}_k^{-1} \hat{\mathbf{u}}_k^{(i,\mu)} \hat{\mathbf{u}}_k^{(j,\nu)}$ is the element $\hat{F}'^{-1}_{ij,\mu\nu}$ of the pseudo-inverse of matrix $\hat{\mathbf{F}}$, and observing that it is related with the pseudo-inverse of matrix $\mathbf{F}$ through the relation: $\hat{\mathbf{F}'}^{-1} = \mathbf{\Gamma}^{1/2} \mathbf{F}'^{-1} \mathbf{\Gamma}^{1/2}$.

The approach of ENM was stimulated by the seminal paper of Tirion (1996), where it was showed that the low-frequency spectrum of globular proteins is almost insensitive to the local details of the atomic composition of the structure and of the specific interaction between them. Specifically, the lowest dynamics predicted using a NMA on a standard atomistic MD force field was shown to be reproducible in good detail by the use of a simplified single parameter potential. Tirion's simplified potential energy of the overall molecule is given by:

$$E_P = \sum_{(i,j)}{}' E(\vec{r}_i, \vec{r}_j) \tag{1.35}$$

where $\vec{r}_i$ indicates the Cartesian coordinate of the i-th atom of the protein and the prime indicate that the sum is restricted to the atom pairs (i, j) separated by less than a cut-off distance R_c. $E(\vec{r}_i, \vec{r}_j)$ represents the interaction between atoms i and j, and it is modeled through a simple Hookean pairwise potential:

$$E(\vec{r}_i, \vec{r}_j) = \frac{C}{2}\left(\left|\vec{d}_{ij}\right| - \left|\vec{d}^{\,0}_{ij}\right|\right)^2 \tag{1.36}$$

being $\vec{d}_{ij} = \vec{r}_i - \vec{r}_j$ the difference vector between the positions of the atoms i and j. The zero superscript indicates the reference structure. The strength of the coupling constant C is a phenomenological constant assumed to be the same for all the interacting pairs. As usual, the potential 1.36 has been expanded in Taylor series relatively to the displacement $\Delta\vec{d}_{ij} = \vec{d}_{ij} - \vec{d}^{\,0}_{ij}$ of the vector $\vec{d}_{ij}$ respect to the reference structure, and the contributions beyond the second order have been neglected, yielding:

$$E(\vec{r}_i, \vec{r}_j) \sim \frac{C}{2}\left(\frac{\vec{d}^{\,0}_{ij} \cdot \Delta\vec{d}_{ij}}{\left|\vec{d}^{\,0}_{ij}\right|}\right)^2 \tag{1.37}$$

Substituting 1.37 in 1.35, the potential energy of the overall protein can be recasted as a quadratic function of the displacements $\vec{x}_i = \vec{r}_i - \vec{r}^{\,0}_i$ of the i-th atom from its reference position:

$$E_P = \frac{C}{2}\sum_{i,j,\mu,\nu} K_{ij,\mu\nu} x_{i\mu} x_{j\nu} = \frac{C}{2}\mathbf{x}^T\mathbf{K}\mathbf{x} \tag{1.38}$$

where indices i and j run over all the atoms in the protein and μ and ν over the three Cartesian components. The rightmost expression is in matrix form. Notice that the elements of the $3N \times 3N$ matrix $\mathbf{K}$ depends exclusively from the reference structure of the protein and the value of the cut-off distance R_c. The coupling constant C and the cut-off distance R_c are related, as for each matrix $\mathbf{K}$ relative to a particular value of

R_c, the value of C can be adjusted in order to obtain an optimal fit with the spectral density obtained for instance from NMA of detailed force fields. The model provides, for vales of the cut-off distance R_c ranging from 1.1 to 2Å, estimates of the lowest energy density of states and of the root mean square deviation of the mainchain C_α atoms from their reference position. A remarkable accord of these estimates is observed with predictions obtained by NMA of detailed force fields.

The remarkable reliability of low-frequency (low-energy) outcomes of a model so simplistic lies in the fact that slow vibrational modes involve collective motion of several amino acids. The effective force opposing large scale oscillations stems from the combined effect of numerous interacting atom pairs. The sum of these interactions approaches a universal form, that reflects the fundamental properties of proteins' structural architecture such as the secondary and tertiary organization, regardless of the details of individual pair potentials (which however are essential for stabilizing specific minimum energy configuration). Hence, for slow vibrations these details could be neglected.

These considerations have stimulated the further development of simplified models for capturing proteins' large scale fluctuations. In fact, the detailed atomistic force field can be replaced by simplified quadratic interactions limited to a reduced number of interaction centers, typically the C_α ones, in place of all pairs of contacting atoms. The viability of these models, generally referred to as ENM (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2004), has been largely verified *a posteriori* against both general dynamical data obtained from experiments, such as the mean-square fluctuations of each residues measured by the crystallographic Debye-Waller factors, and also against more specific dynamical properties such as the principal direction of motions or the covariance matrix obtained from MD simulation (Atilgan *et al.*, 2001; Micheletti *et al.*, 2004).

In this thesis we have extensively used a particular ENM, the beta gaussian network model (Micheletti *et al.*, 2004), which is illustrated in detail in the following subsection.

1.6.1 Beta Gaussian Network Model

The beta gaussian network model (βGM) is a simplified ENM in which the protein is represented by means of two-centroid per amino acid, one for the main-chain, coinciding with the C_α atom, and one for the side-chain. Following a geometrical rule akin to the one introduced by Park & Levitt (1996) we construct the latter interaction center as a fictitious C_β centroid:

$$\vec{r}_{CB}(i) = \vec{r}_{CA}(i) + l\frac{2\,\vec{r}_{CA}(i) - \vec{r}_{CA}(i+1) - \vec{r}_{CA}(i-1)}{|2\,\vec{r}_{CA}(i) - \vec{r}_{CA}(i+1) - \vec{r}_{CA}(i-1)|} \tag{1.39}$$

where $l = 3$Å and $\vec{r}_{CA}$ indicates the coordinates of the i-th C_α centroid. For amino acids at the beginning/end of the peptide chain(s) or for GLY the construction of eqn. 1.39 is not applicable and hence the effective C_β centroid is taken to coincide with the C_α one. A schematic view of the coarse graining procedure is given in Figs. 1.3.

The potential governing the interaction between the centroids is obtained by introducing quadratic penalties for displacing two centroids, i and j from their reference positions, $\vec{r}_i^0$ and $\vec{r}_j^0$, to generic ones, $\vec{r}_i$ and $\vec{r}_j$. The energetic cost of a displacement is precisely the same introduced by Tirion and given in 1.37.

The quadratic form of 1.37 is at the heart of the widely-used elastic or Gaussian network approaches (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2001, 2002, 2004; Tirion, 1996), which typically adopt a single-centroid amino acid description. The effective free energy function introduced in Micheletti *et al.* (2004) and used here includes, instead, pairwise contributions from all pairs of centroids, be they of the C_α or C_β type, whose reference distance falls within a given interaction cutoff, as pictorially illustrated in Fig. 1.3c. Accordingly, the resulting free energy of a trial structure, Γ, takes on the form:

$$E_{\beta GM}(\Gamma) = 2\sum_i V(\vec{d}_{i,i+1}^{CA-CA}) + \sum_{i<j}{}' V(\vec{d}_{i,j}^{CA-CA}) + \sum_{i,j}{}' V(\vec{d}_{i,j}^{CA-CB}) + \sum_{i<j}{}' V(\vec{d}_{i,j}^{CB-CB}) \tag{1.40}$$

where i and j run over the residue indices, $\vec{d}_{i,j}^{X-Y} = \vec{r}_i^X - \vec{r}_j^Y$ denotes the distance vector of the centroids of type X and Y of residues i and j, respectively, and the prime denotes that the sum is restricted to the pairs whose reference separation is below the cutoff distance of 7.5 Å. Consistently with the spirit of ENM and other approaches Tirion (1996), the last three terms in eqn. 1.40 have the same strength irrespective of the identity of the amino acids. The first term, on the other hand, accounts for the protein chain connectivity and has a double strength to reflect the geometrical constraints of the peptide chain.

As the positions of the C_β centroids depend linearly on the coordinates of the C_α ones, it is possible to analytically recast the expression 1.40 in the following quadratic form involving simply the C_α degrees of freedom, retaining the same computational complexity of the single centroids model:

$$E_{\beta GM}(\Gamma) = \frac{C}{2}\sum_{i,j,\mu,\nu} M_{ij,\mu\nu}\, x_{i,\mu}\, x_{j,\nu} = \frac{C}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} \tag{1.41}$$

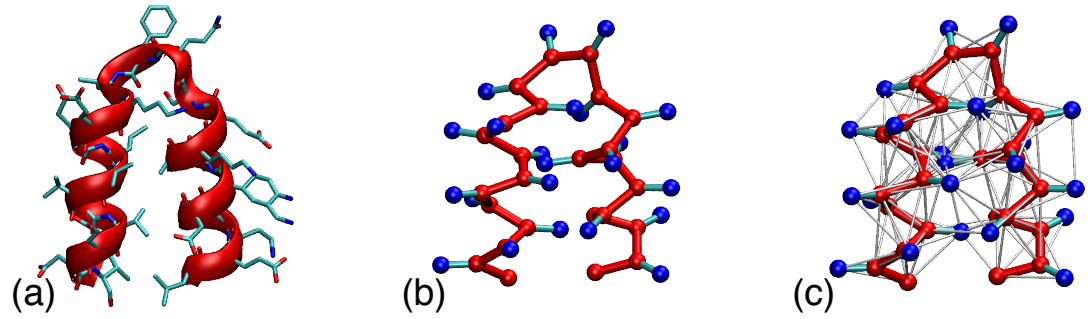

Figure 1.3: Pictorial representation of the coarse graining procedure: (a) atomic representation of a two-helix bundle (backbone highlighted as a ribbon); (b) simplified structural representation in terms of the C_α atoms for the backbone and the C_β atoms for the for the sidechains; (c) all pairs of non-consecutive centroids within 7.5 Å interact through an harmonic potential, schematically shown as a thin bond.

where $\vec{x}_i = \vec{r}_i^{CA} - \vec{r}_i^{0CA}$ is the deviation of i-th C_α centroid from the reference position, $\mathbf{M}$ is a symmetric matrix whose linear size is three times the number of residues in the protein, and the coefficient C is the phenomenological parameter controlling the strength of the quadratic coupling.

As previously discussed, the most suitable framework to interpret 1.41 is in the context of overdamped Langevin equation 1.27. For the case where the friction coefficients of the C_α atoms take the same value γ, the eigenvectors of matrix $\mathbf{M}$ provide the independent modes of structural relaxation in the protein, and the associated eigenvalues are inversely proportional to the relaxation times. The eigenvectors of $\mathbf{M}$ relative to the lowest non-zero eigenvalue are usually called low-energy modes.

The outcomes of βGM have been tested versus all-atom molecular dynamics simulation, in different contexts, and providing good results (Carnevale *et al.*, 2007b; Cascella *et al.*, 2005; De Los Rios *et al.*, 2005; Micheletti *et al.*, 2004).

Chapter 2

Functional Structural Changes and Internal Dynamics: the case of Adenylate Kinases

2.1 Introduction

In this chapter we will illustrate many of the concepts previously introduced, applying PCA and ENM on the study of an important enzyme, adenylate kinase, whose internal dynamics is know to play a major role in the accomplishment of its biological function. This reason make this enzyme an ideal case study for the investigation of the connection between the functional dynamics and the intrinsic features of the free energy landscape.

Adenylate kinase(Adk) is a monomeric enzyme which regulates the energy charge of the cell by balancing the relative abundance of AMP, ADP and ATP. The concentration of the three nucleotides is controlled by the enzyme through the catalysis of the phosphoryl transfer reaction:

$$ATP + AMP \overset{Mg^{2+}}{\Longleftrightarrow} 2ADP \ .$$

The differences in structural arrangement between the free E. Coli adenylate kinase (AKE) and the enzyme complexed with an inhibitor mimicking both ATP and AMP are illustrated in Fig.2.1 (Müller & Schulz, 1992; Müller *et al.*, 1996). By comparing the two portrayed crystal structures it is apparent that the formation of the ternary complex stabilizes the enzyme in a form where the mobile Lid and AMP-binding subdomains (highlighted in Fig. 2.1) close over the remainder core region. This rearrangement of the two mobile subdomains is necessary for the accommodation of the nucleotides in

an optimal catalytic geometry and the resulting closed enzyme conformation provides a solvent-free environment for the phosphoryl transfer.

The conformational change sustained by adenylate kinase upon complexation with ATP and AMP, and its reopening upon unbinding of the processed nucleotides, represents the rate-limiting step in the reaction turnover (Kern *et al.*, 2005). A large number of experimental studies have consequently addressed the functional implications of Adk structural elasticity (Han *et al.*, 2002; Hanson *et al.*, 2007; Henzler-Wildman *et al.*, 2007b; Kern *et al.*, 2005; Müller & Schulz, 1992; Müller *et al.*, 1996; Shapiro & Meirovitch, 2006; Shapiro *et al.*, 2000, 2002; Sinev *et al.*, 1996; Wolf-Watz *et al.*, 2004). In particular, recent investigations based on a wide range of techniques, have provided converging evidence for the fact that, even in the absence of the bound nucleotides, the free enzyme is capable of interconverting between the open and closed forms. These investigations have lead to formulating the hypothesis that evolutionary pressure has endowed Adk, and arguably other enzymes (Beach *et al.*, 2005; Eisenmesser *et al.*, 2005), with the innate ability to interconvert between the open and catalytically-potent forms.

These observations have stimulated a numerical study of the dynamical evolution of the free (apo) AKE molecule in solution (Pontiggia *et al.*, 2008), where I have collaborated with Francesco Pontiggia and Cristian Micheletti. By means of two extensive MD simulations started from the available crystal structures we have characterized, over various time scales, the conformational fluctuations sustained by the enzyme and analyzed the extent to which they indicate the suggested innate predisposition to connect the open and closed forms.

Several computational investigations of the flexibility of AdK exist and include both mesoscopic and atomistic approaches. Coarse grained models have, for instance, been applied to model the pathways connecting the open and closed forms of the enzyme (Chennubhotla & Bahar, 2007; Chu & Voth, 2007; Maragakis & Karplus, 2005; Miyashita *et al.*, 2003). Atomistic simulations have instead been used to probe the free energy landscape in the neighborhood of several known enzyme conformers, as in the recent investigations by Lou & Cukier (2006), Arora & Brooks (2007) and Henzler-Wildman *et al.* (2007b). In the first study (Lou & Cukier, 2006), an advanced sampling technique was used to show that the enzyme populated conformations compatible with the holo-form geometry, as probed by FRET experiments (Sinev *et al.*, 1996). Arora & Brooks (2007) further showed that the free energy landscape along a pre-assigned reaction coordinate connecting the open-closed forms of AKE is approximately flat for the apo-form while, upon ligand binding, it changes favoring the closed state. Finally, in the study of Henzler-Wildman *et al.* (2007b), carried out on Adk extracted from hyperthermophile *Aquifex Aeolicus*, a variety of experimental and computational probes

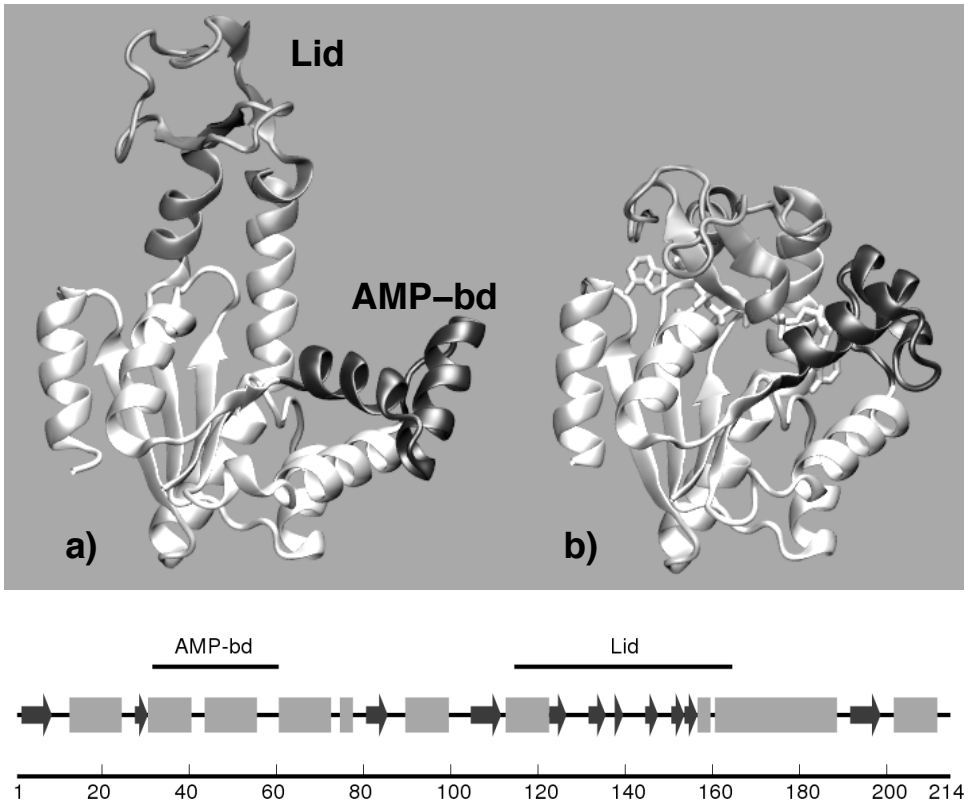

Figure 2.1: Cartoon representation (Humphrey *et al.*, 1996a) of crystallographic structures of E.Coli adenylate kinase in: (a) the open apo form and (b) the closed holo form. The PDB codes for the two structures are 4ake and 1ake, respectively (Müller & Schulz, 1992; Müller *et al.*, 1996). The flexible Lid (amino acids 114-164) and AMP-binding (amino acids 31–60) domains are colored in gray and black, respectively. The succession of secondary elements is shown in the bottom panel. Helices are indicated as grey boxes while β-strands are shown as black arrows.

have indicated the existence of several metastable configurations bridging open and closed states.

In order to illustrate the picture on protein flexibility given in the previous chapter, I will report here some of the results obtained in the study of this protein published in Pontiggia *et al.* (2008). The long MD trajectories of AKE (obtained by Francesco Pontiggia) were analyzed with the PCA and others specifically developed tools. We first highlighted that the MD trajectories show a structural heterogeneity of the visited conformational phase space. However the internal dynamics of the protein, observed at different time scales and relatively to different conformations of the protein, results surprisingly homogeneous. These findings confirm the striking self-similarity of the multiscale free-energy landscape, that we anticipated also in the previous chapter. In the final part of the chapter we will make contact between the behavior of AKE obtained from extensive MD simulations at the atomic level and the coarse-grained characterizations of the system. As already mentioned, these approaches rely on harmonic approximations of the free energy around the dominant states, in this case the open and closed conformations. We will see that some aspects of the dynamics, and in particular the directions of the collective functionally oriented movements of the protein, are reproduced with remarkable accord by simplified topology based ENM (in this case the βGM). This will serve as an *a posteriori* justification of the broad use of this model that will be done in the rest of the thesis.

2.2 Molecular dynamics simulations

The data illustrated in this chapter come from the atomistic MD evolution of E. Coli adenylate kinase, AKE, followed starting from two distinct initial structures, corresponding to the open and closed form of the enzyme. More precisely, the initial conformation of the first simulation was the free (apo form) enzyme from the 4akeA PDB crystal structure(Bernstein *et al.*, 1977a). The second simulation followed, instead, the evolution of the free closed form of the enzyme obtained by removing the Ap5A inhibitor from the 1akeA PDB structure file. In the following, for simplicity, we shall refer to the two simulations as the "open" and "closed" trajectories. The nomenclature is only meant to remind of the starting configuration as, in fact, for both trajectories a partial conversion to the complementary (open or closed) state is observed. The computational details about MD simulations are provided in Pontiggia *et al.* (2008). Both the trajectories here analyzed cover a simulation time of 40ns.

2.3 Structural fluctuations of the amino acids

The two trajectories were first analyzed to assess the level of elasticity shown by the protein during the overall time evolution. To this purpose we have considered the overall mobility of individual amino acids in each trajectory. This was characterized by means of the root mean square fluctuation (RMSF) profile of their α-carbon atoms. The RMSF of the i-th C_α, whose instantaneous coordinate at time t is indicated by $\vec{r}_i(t)$, is given by $\sqrt{\langle|\vec{x}_i|^2\rangle}$ where the brackets denote the time average and $\vec{x}_i(t) \equiv \vec{r}_i(t) - \langle\vec{r}_i\rangle$ is the instantaneous displacement from its time-averaged (reference) position. The average was taken after removing the rigid-body motions of the enzyme, exactly as it was explained in section 1.3.

The RMSF profiles for the open and closed trajectories are shown in Fig. 2.2. Observe that most of the mobility of the protein is due to the fluctuation of the Lid and AMP-bind domains, while the core is, by converse, very stable. The rigidity of this region is consistent with NMR and Xray studies, as well as with previous topology-based characterizations of the protein's elasticity (Chennubhotla & Bahar, 2007; Maragakis & Karplus, 2005; Miyashita *et al.*, 2003, 2005; Whitford *et al.*, 2007).

Moreover, we observe that the open trajectory manifests a greater degree of mobility respect to the closed trajectory, despite their lengths are the same. Indeed the total MSF of the open trajectory is 18.04 nm^2, while the MSF of the closed trajectory is 7.72 nm^2. This aspect is clearly due to the more compact structure of the closed conformations respect to the open ones, as the most compact is the structure, the most the degrees of freedom of the amino acids are bounded.

2.4 Structural heterogeneity

Next we want to investigate the variety of structural conformations sampled in the two MD trajectories, and to follow how these conformations change over the time. To this purpose we have divided the trajectory in intervals of 1ns, and calculated the average position of the C_α atoms in each interval. The density plot of the root mean square distance (RMSD) between the average positions of the C_α atoms for each pair of intervals is reported in Fig. 2.3, separately for the open and closed trajectory.

The block character of the matrix suggests that distinct conformational groups are explored during the dynamical evolution, and the system evolution proceeds by visiting distinct conformational substates through which the systems hops with rapid "transitions" (e.g. in the open trajectory we can easily localize these transitions after ∼9ns, ∼19ns, and ∼27ns). This indicates that the system meets some local free energy

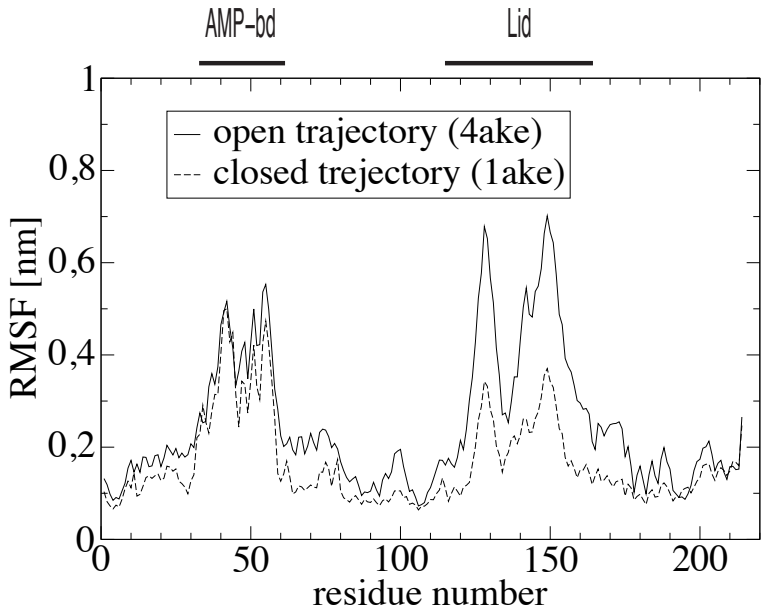

Figure 2.2: Root mean square fluctuations of the C_α atoms observed in the 40-ns long "open" and "closed" trajectory. The fluctuations have been calculated after an optimal structural superposition of the C_α trace. The flexible Lid (amino acids 114-164) and AMP-binding (amino acids 31–60) domains are highlighted.

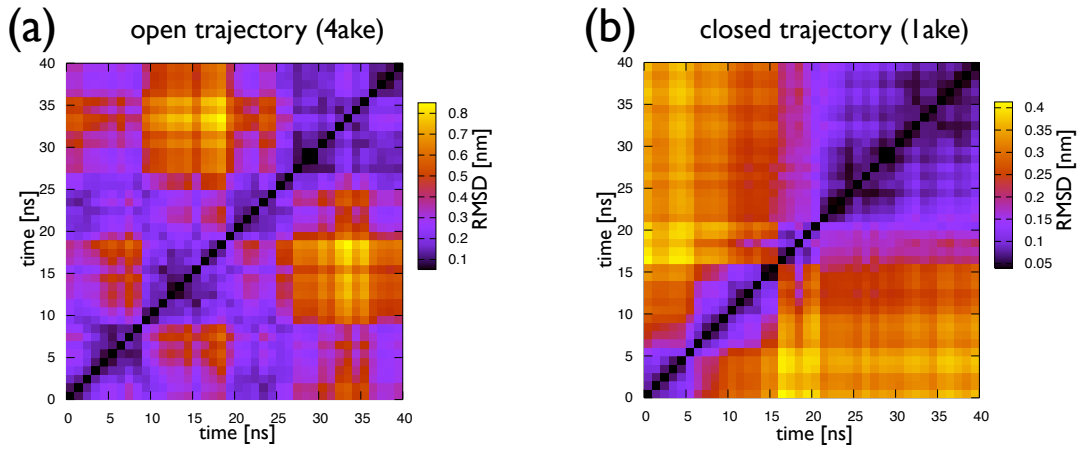

Figure 2.3: Density plot of the pairwise RMSD between the average structures of 1-ns long intervals from the (a) open and (b) closed trajectory (time labels are shown on both axes).

minima and settles there for a while, jumping then from one minimum to another. Note that this behavior is more evident in the time evolution of the system starting from the structure close to the inhibitor bound crystal structure, indeed from a visual inspection of Fig. 2.3 it is readily seen that the repertoire of structures generated by the evolution of the close conformation is much more heterogeneous than for the open one. This behavior was expected as the closed structure is simulated without the bound ligand, however it is interesting to observe that the system does not completely open up, at least during the simulated time. In both the trajectories we observe that the enzyme encounters locally stable states, where it dwells for about ten nanoseconds.

2.5 PCA of the open and closed trajectories

2.5.1 Fluctuations along the principal components

The next step of our investigation was to consider the principal components of the trajectories, calculated as illustrated in section 1.3. Note that in this study we are mainly focussed on the mobility of the backbone, therefore the covariance matrix that we considered here is that of the positions of the C_α atoms.

In Fig. 2.4 we have reported the largest eigenvalues of the covariance matrix of the open and closed trajectory. Let us remind that the value of and eigenvalue correspond

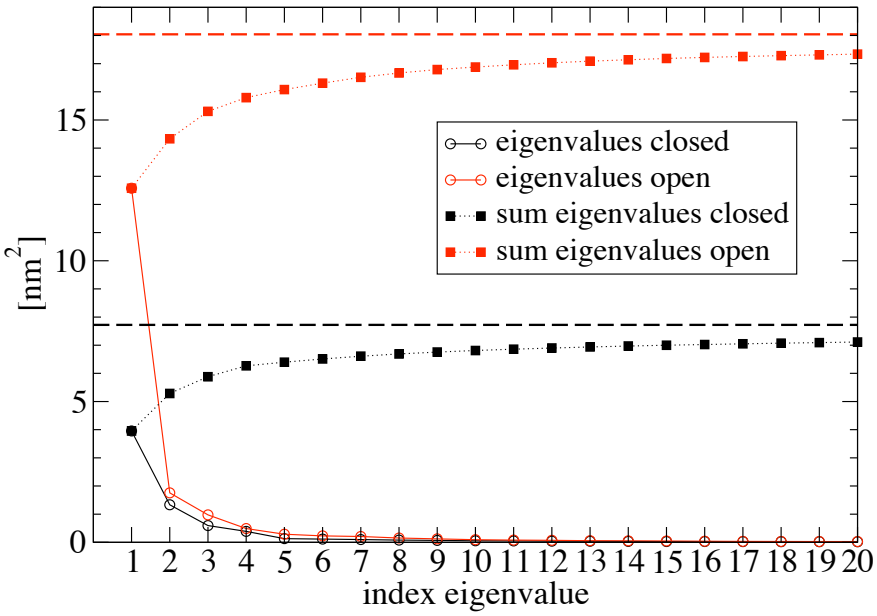

Figure 2.4: Black and red circles represent the eigenvalues, in decreasing order of magnitude, obtained from C_α coordinates covariance matrix from respectively the closed and open trajectory. The squares represent the cumulative sum of the largest eigenvalues, until the corresponding index in abscissa. The sum of all the eigenvalues, that corresponds to the total mean square fluctuation observed in the trajectory, is represented as a dashed line and it is 7.72 nm^2 for the closed trajectory (black) and 18.04 nm^2 for the open trajectory (red).

to the MSF of the trajectory projected along the relative eigenvector. Note that the MSF along the projection of the first eigenvector of the open trajectory in 12.57 nm^2. This value has to be considered in relation to the fact that it is only one of the more that 600 internal degrees of freedom of the protein, and that, as mentioned, the overall MSF observed in the overall trajectory is 18.04 nm^2. An analogous consideration is true also for the closed trajectory.

The sum of the first n eigenvalues, also reported in Fig. 2.4, gives the MSF that is accounted by the first n eigenvectors. For instance we have that, in the closed trajectory, the first three eigenvectors are enough to account for the ∼77% of the overall MSF, while in the open trajectory they account for ∼85%. Defining, as usual (Amadei *et al.*, 1999), the essential space of the protein as the one given by the first ten eigenvectors of the protein, that is enough to account for more than 80% of the fluctuation for both

the open and the closed trajectory.

2.5.2 Principal components and opening/closing motion

At this point we have investigated the relationship between the principal components of the trajectories, and the overall the opening/closing motion of the protein. To this purpose we have computed the average structure of the open trajectory $\mathbf{r}_{\text{op}}$ and that of the closed trajectory $\mathbf{r}_{\text{cl}}$. Assume that we have rigidly roto-translate one of the two structures in order to minimize the RMSD between $\mathbf{r}_{\text{op}}$ and $\mathbf{r}_{\text{cl}}$. We can calculate the difference vector $\mathbf{d}_{\text{op/cl}} = \mathbf{r}_{\text{op}} - \mathbf{r}_{\text{cl}}$, whose normalized vector $\mathbf{u}_{\text{op/cl}} = \mathbf{d}_{\text{op/cl}}/\|\mathbf{d}_{\text{op/cl}}\|$ provides the direction of the opening/closing motions. As the eigenvectors $\{\mathbf{v}_i\}$ of the covariance matrix represent an orthonormal base, the normalized difference vector can be written as: $\mathbf{u}_{\text{op/cl}} = \sum_i c_i \mathbf{v}_i$, being $c_i = \mathbf{v_i} \cdot \mathbf{u}_{\text{op/cl}}$. Clearly ${c_i}^2$ represents the fraction of the opening/closing difference vector captured by the principal component $\mathbf{v}_i$.

In this way we have calculated that the fraction ${c_1}^2$ of the direction that leads to bridging the open/closed structures of the enzyme captured by the first eigenvector of the covariance matrix for the open trajectory is $\sim$0.77. This result represents an strong indication that the fluctuations of the protein are not random, but that they are specifically oriented towards the closed, catalytically potent, state. As a consequence of this, the free-energy landscape is organized so to facilitate the spontaneous interconversion (in thermal equilibrium) of the protein from to open to the closed state, also in absence of a ligand that induces the closing mechanism.

The same analysis applied to the eigenvectors of the closed trajectory provides a trend less marked than for the open trajectory, however it results that the fraction $\sum_i^{10} {c_i}^2$ of the opening/closing difference vector captured by the first 10 eigenvectors is $\sim$0.7. It results therefore that also the closed protein follows the principal components to open up, but the correlation with the direction of the first principal components is not so strict as observed in the case of the open protein. This is probably due to the fact that there is not a unique and well characterized opened structure, but there is an ensemble of them (indeed in Fig. 2.3a we observed displacements up to 8Å RMSD between different structures sampled in the open trajectory).

2.5.3 Comparison of the essential spaces of the open and closed trajectories

We want now to compare the essential spaces of the open and closed trajectory. The simplest way to perform this comparison is to consider the scalar products $\mathbf{v}_i \cdot \mathbf{w}_j$ between all the possible pairs of eigenvectors $\{\mathbf{v}_i\}$ from the open trajectory and the

$\{\mathbf{w}_i\}$ form the closed one (assuming that the eigenvectors are indexed in decreasing order of their eigenvalues). The scalar products between the first 10 eigenvectors (which define the essential space) have been computed and reported in the density plot Fig. 2.5. It can be seen that there is no precise one-to-one matching between the modes.

The measure that is typically adopted to compare two essential dynamical spaces $\{\mathbf{v}\} = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\{\mathbf{w}\} = \{\mathbf{w}_1, ..., \mathbf{w}_n\}$ is the root mean square inner product (RMSIP):

$$\text{RMSIP} \equiv \sqrt{\frac{1}{n} \sum_{i,j=1}^{n} (\mathbf{v}_i \cdot \mathbf{w}_j)^2} \ , \tag{2.1}$$

which ranges from 0, for complete orthogonality of the $\{\mathbf{v}\}$ and $\{\mathbf{w}\}$ spaces, to 1 in case of their perfect overlap. Note that the RMSIP (2.1) is a measure that compares the vectorial spaces defined by the vectors $\{\mathbf{v}\}$ and $\{\mathbf{w}\}$, and it is indeed invariant for change of the base that defines the spaces. Hereafter we will assume that $n = 10$ in (2.1), excepts where explicitly indicated, because, as previously mentioned, the essential space is defined by the top 10 eigenvectors.

The RMSIP between the essential spaces of the open and close trajectories is 0.77, that is indicative of a good agreement (Amadei *et al.*, 1999). The RMSIP between the space defined by only the first three principal components of the trajectories is 0.69, indicating that also these small spaces are quite in agreement despite the lack of precise matching between the eigenvectors (see Fig. 2.5).

2.5.4 Consensus dynamical space

To go in further detail in the difference between the leading fluctuations for the open and closed trajectory, we proceeded to identify the consensus set of collective modes that best capture the common structural fluctuations of AKE encountered in the two trajectories. The essential dynamics analysis applied to the two merged trajectories is not adequate to this purpose as it is not designed to extract the dynamical features that are shared by the two separate trajectories.

Expression (2.1) provides an *average* measure of accord of two essential dynamical spaces, as the top 10 eigenvectors of the covariance matrix are treated on equal footing (degeneracy). This implies that the same value of RMSIP may be attained with different detailed levels of accord of two spaces.

To characterize with a finer resolution the consistency of two sets of modes we introduce a variational scheme that identifies their maximally-consistent (or inconsistent) subspaces. The scheme, explained in detail in the Appendix A, is used to redefine two new bases $\{\mathbf{v}'\} \equiv \{\mathbf{v}'_1, ..., \mathbf{v}'_{10}\}$ and $\{\mathbf{w}'\} \equiv \{\mathbf{w}'_1, ..., \mathbf{w}'_{10}\}$ for the *same* linear spaces

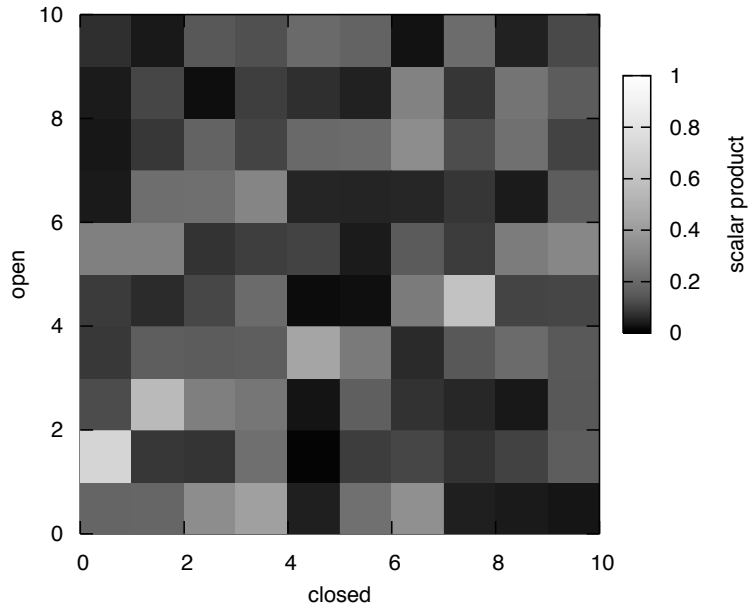

Figure 2.5: The elements of the matrix represent the scalar products (in absolute value) between the eigenvectors associated to the ten largest eigenvalues of the covariance matrix obtained from the open trajectory, and from the closed trajectory. The values of the scalar products are color coded from black (orthogonal vectors) to white (parallel vectors).

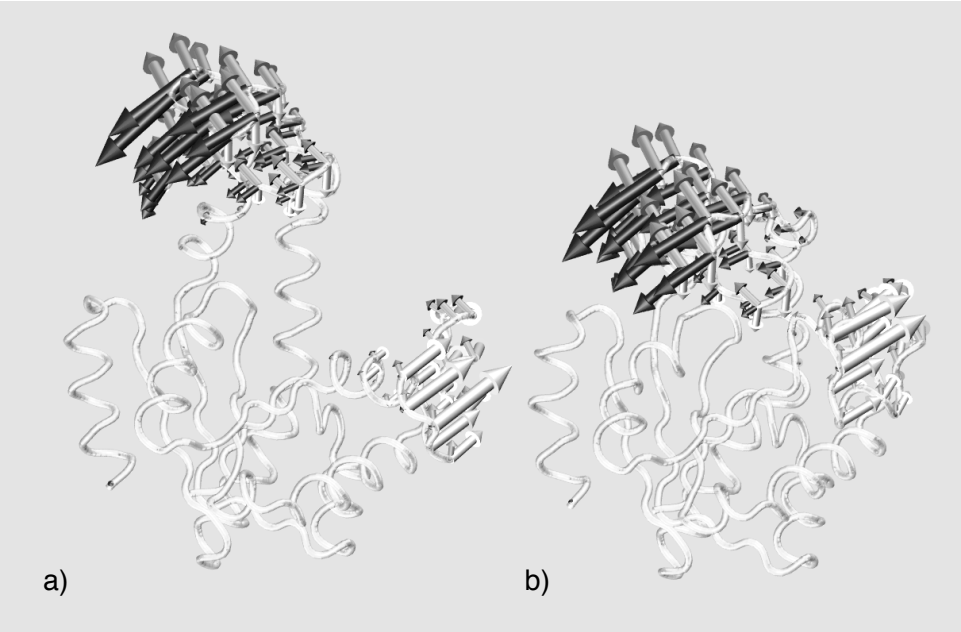

Figure 2.6: The three consensus modes of the open and closed trajectories are shown respectively with black, grey and white arrows superposed to the average structure of the (a) open and (b) closed trajectories.

described by $\{\mathbf{v}\}$ and $\{\mathbf{w}\}$. The redefined bases, $\{\mathbf{v}'\}$ and $\{\mathbf{w}'\}$, possess two notable properties: (i) a basis vector of one set is orthogonal to all basis elements of the other set except the one with the same index and (ii) the index provides a natural ordering of the basis vectors in terms of decreasing mutual consistency. Notice that the RMSIP of the new basis vectors is the same of the original one.

The method provides an optimal redefinition of the basis vectors in the two sets of modes which are returned in order of decreasing mutual consistency. We stress that the new bases span the same linear spaces of the original sets so that the original RMSIP, equal to 0.77, is unaltered by the redefinition.

It was found that the 10 lowest-energy modes of the two trajectories share, with almost perfect overlap, a three-dimensional subspace. In fact, the scalar products of the fist, second and third pair of redefined modes have scalar products greater than 0.9. These consensus modes are shown in Fig. 2.6.

2.6 Consistency of the internal dynamics

In order to investigate the consistency of the outcomes of the PCA in trajectories of length smaller than the total simulation time, we have subdivided each MD trajectory in 40 intervals, each of duration of 1 ns (and hence comprising 2000 frames). For each interval we built the covariance matrix and extracted the first 10 eigenvectors. A pairwise comparison of the dynamics is finally carried out computing the RMSIP of the two sets of 10 eigenvectors. The results for any two pairs of intervals for are shown in the density plot in Fig. 2.7. The degree of consistency appears to be extremely high throughout both the trajectories. The range of RMSIP values recorded is comprised between ~ 0.6 and ~ 0.8. This has to be compared with the standard reference value of 0.7 which typically accompanies the good consistency of essential dynamical spaces in multi-ns MD trajectories on medium-size proteins (Amadei *et al.*, 1999; Pontiggia *et al.*, 2007).

The observed degree of consistency is very striking in comparison with the level of structural heterogeneity encountered during the evolution (see Fig. 2.3). The same feature was previously observed by Pontiggia *et al.* (2007) in the context of a globular protein, protein G, where the high consistency of the space of the 10 essential eigenvectors was shown to result from a peculiar self-similar organization of the several minima that, at various scales, result in the free energy landscape. Also for this much larger protein it appears confirmed, *a posteriori*, that a low dimensional space of collective variables is sufficient to account for the system dynamics over a time-span much larger than the residence time in each of the salient free energy minima.

2.7 Comparison with the predictions of βGM

In the previous chapter we anticipated that topology based ENMs can be used to predict the low-energy modes, which correspond to the directions of maximal fluctuations, of a protein.

We can validate here the predictions of the βGM, introduced in section 1.6.1, considering the overlap between the low-energy modes of the model and the essential spaces obtained from MD simulation. We observed that the RMSIP of the top 10 eigenvectors of the covariance matrix with the top 10 modes predicted by the model, using the average structure as input, is 0.84 for the open trajectory; 0.79 for the closed one. Both these numbers are very high, indeed MD simulations of the same protein starting from different initial conditions have essential spaced with RMSIP$\sim$ 0.7, as shown in this chapter or in (Amadei *et al.*, 1999; Pontiggia *et al.*, 2007, 2008). The RMSIP of the top

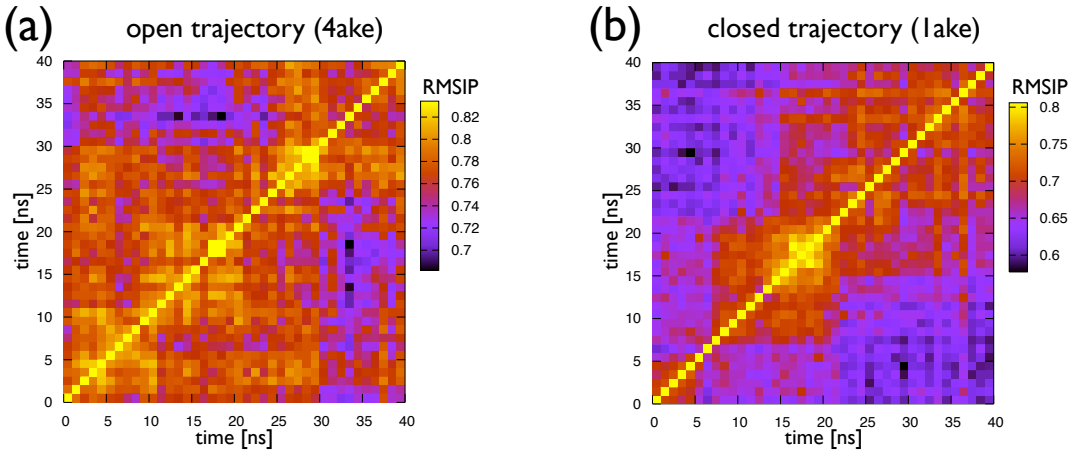

Figure 2.7: Density plot of the RMSIP between essential dynamical spaces of 1-ns long intervals of the (a) open trajectory and (b) closed trajectory (time labels are shown on both axes).

10 modes predicted applying the model to the open and closed structures is 0.84, also very high. Finally the contribution of the first mode of the model to the jump between the open to closed configuration is $\sim 75\%$ for the open structure, $\sim 30\%$ for the closed, in good agreement with the results obtained previously for MD.

Moreover, some of the results obtained here from the analysis of the MD simulations are very interesting in the perspective to understand *a posteriori* the use of these simplified models. In first place, the observed homogeneity of the dynamics, that arise from the striking self-similarity of the free-energy landscape, represents an important point in favor of the topology based models. Indeed, if the directions of the fluctuations would change as the protein visits different structural conformers, the interpretation of the low-energy modes provided by ENM would be much more complex.

In second place, the observed correspondence between the principal modes (or low-energy modes of the model) and the direction of opening/closing of the protein highlights that these collective motions have a precise directionality that appears, at least in this case, functionally oriented.

2.8 Conclusions

In this chapter we have discussed, starting from extensive molecular dynamics simulations of adenylate kinase, the predisposition of this enzyme to undergo major conforma-

tional changes. The analysis of the data has exposed interesting functionally-oriented characteristics of the internal dynamics of the enzyme and of the organization of its free energy landscape.

During the free dynamical evolution, the enzyme populates distinct conformational substates. The ensemble of different conformers is structurally heterogeneous, reflecting the pronounced mobility of the AMP-binding and Lid subdomains.

We have carried out a covariance analysis of structural fluctuations recorded over a temporal range wide enough to cover both the collective small scale fluctuations within the substates and the larger-scale ones associated to inter-substate transitions. Strikingly, irrespective of the probed time-scale, essential dynamical spaces turned out to be highly consistent. The functional relevance of this consistency, which does not originate from unspecific properties of overall amino acid mobility, is underscored by the high overlap that the essential dynamical spaces have with the deformation vector connecting the open and closed structures.

The analysis indicates that the free enzyme can be driven through various conformational substates bridging the inactive and catalytically potent states through the thermal excitation of a limited number of collective modes. These results show a functionally oriented nature of the self-similar organization of the free energy landscape (coherently with the observations on the G protein in Pontiggia *et al.* (2007)).

The results support the recent suggestion of (Adén & Wolf-Watz, 2007) that functionally-oriented conformational fluctuations are innate properties of the free (apo) Adk. In fact, the consistency of the salient features of the enzyme's internal dynamics leads to speculate about the fact that these property may have been promoted by evolutionary pressure.

Finally we have validate the low-energy modes provided by the βGM. The overlap of these modes with the principal components provided by the MD simulation is remarkable. Therefore the validity of the outcomes of the βGM, tested versus MD simulations also in (Carnevale *et al.*, 2007b; Cascella *et al.*, 2005; De Los Rios *et al.*, 2005; Micheletti *et al.*, 2004), poses this method as a valuable instrument for the investigation, also systematic, of the functionally-oriented conformational fluctuations.

2. FUNCTIONAL STRUCTURAL CHANGES AND INTERNAL DYNAMICS: THE CASE OF ADENYLATE KINASES

Chapter 3

Protein-protein Complexes: a Dynamics-based Characterization

3.1 Introduction

Characterizing the physico-chemical processes that regulate protein-protein interactions has always been a primary aim of molecular biology, as the majority of the biological processes are regulated through association and dissociation of protein molecules. Examples of these processes range from enzyme-substate binding, to antigen-antibody recognition, hormone-receptor binding, signal transduction, etc. The importance of systematically characterize and classify the complex nature of protein interactions is widely recognized, and has been addressed in a number of studies (Bogan & Thorn, 1998; Chakrabarti & Janin, 2002; Conte *et al.*, 1999; Jones & Thornton, 1996; Keskin *et al.*, 1998; Ma *et al.*, 2001; Tsai *et al.*, 1998; Valdar & Thornton, 2001; Zhang *et al.*, 2003). These studies have important implications in applications aiming to predict the conformations of multimeric assemblies and the cellular pathways, beyond being important for drug design and protein docking. Despite the large number of studies, many aspects of the mechanisms of protein-protein interactions are not yet completely understood.

Protein-protein interfaces have been characterized in terms of their structural and physical properties (size, shape, complementarity and packing) and their chemical nature (amino acid composition, chemical group distributions, hydrophobicity/ hydrophilicity, electrostatic interactions, hydrogen bonding and interactions with water) (Arkin *et al.*, 2003; Jones & Thornton, 1996; Katchalskikatzir *et al.*, 1992; Nooren & Thornton, 2003; Todd *et al.*, 2002; Wallis *et al.*, 1998).

3. PROTEIN-PROTEIN COMPLEXES: A DYNAMICS-BASED CHARACTERIZATION

Recently Keskin *et al.* (2004) have selected all the interfaces between two protein chains obtained from protein-protein complexes in the Protein Data Bank. The interfaces have been next grouped according to the degree of similarity of their architectures, and filtered to eliminate redundancy. The final set of clusters contains member proteins as diverse as enzymes, antibodies, viral capsids, etc. Each cluster was assigned to one of three main types of interfaces, which we briefly review as they will play an important role in the study described hereafter. Type I, gathers clusters which share not only the similarity of the interface, but also of the non-interface region. Type II, includes clusters whose members share only the interface region (both sides); the members of these groups, therefore, have a different overall structural organization. Finally, type III groups are characterized by the fact that their members share only one side of the interface region (i.e. a semi-interface). They have observed that the parental proteins of members of the same type I cluster belong to the same functional family, while the parental proteins of the members of the same type II or III cluster may belong to different functional families.

This study has stimulated the investigation that we report in this chapter. The question that we pose is whether different types of interfaces, according to the classification of Keskin *et al.* (2004), are characterized also by specific dynamical properties. To this purpose we provide here a detailed characterization of the equilibrium dynamical properties of a comprehensive set of dimeric protein complexes, selected starting from the database of Keskin *et al.* (2004). In order to characterize the flexibility of the complexes we have used an ENM, as it provides, with a minimum computational effort, the salient features of proteins' internal dynamics. Moreover this approach has two remarkable properties: to be systematically applicable; and to allow to account for the influence of the binding parter in the mobility of the amino acids at the interface of a monomer.

The chapter is organised as follows. The first part is devoted to describe the creation of the dataset of protein-protein constructs. In the second part the salient structural traits of the complexes and their interfaces are presented. The third part presents a detailed account, organized per interface category, of the dynamical properties of amino acids at the interface of the dimeric subunits. Finally we discuss how the findings can be interpreted considering the expected role of conformational entropy to the stability of dimeric interfaces.

3.2 Dataset selection

The dataset of dimeric complexes was compiled so to capture and represent the largest possible diversity of protein-protein constructs. The starting point for compiling the dataset was the database of protein-protein interfaces of Keskin *et al.* (2004). This comprehensive dataset was culled on the basis of two criteria. First, considerations are restricted to dimeric protein interfaces. Next, we considered only dimers whose fluctuation dynamics can be adequately captured by elastic network models.

In the following subsections we will describe in details the structurally nonredundant dataset of two-chain protein-protein interfaces obtained by Keskin *et al.* (2004) and the subsequent filtering of this database for the selection of the representative dimers which we investigated in this study. Finally we will characterize the protein-protein interaction among the selected dimers in terms of obligate and non-obligate interactions.

3.2.1 Structurally nonredundant dataset of two-chain protein-protein interfaces

We recall there that the database of Keskin *et al.* (2004) was obtaiend through to the following procedure:

1. all the multichain PDB entries in the Protein Data Bank (Berman *et al.*, 2000) were analyzed in order to get all the possible combinations of two-chains.

2. These were subsequently filtered, keeping only two-chain complexes where the number of amino acids in each side of the interface was ≥ 10.

3. The selected complexes where then grouped into clusters. This was carried out using a heuristic iterative procedure (Keskin *et al.*, 2004) to cluster the complexes according to the degree of structural similarity among their interfaces. The similarity was quantified through a sequence-order-independent structural comparison algorithm: the Geometric Hashing algorithm (Nussinov & Wolfson, 1991; Tsai *et al.*, 1996).

4. The redundancy in sequence among members of a same cluster was later removed, comparing the complete sequences of each member of a cluster, and if two or more complexes shared more than 50% similarity only one was kept.

5. Only clusters with 5 or more members (which amount to a number of 103 clusters) were kept and classified as type I, II or III according to the degree of structural similarity of the members of the cluster.

This final classification was carried out as follows. If in the structural alignment of the interfaces, members of a same cluster have only one side of the interface that is structurally aligneable, the cluster is classified of type III. If otherwise the structural alignment of interfaces involves both sides, the cluster is classified of type I or II, depending on the global structures of the two-chain complexes. The global folds of each two-chain complex was evaluated, according to SCOP database (Murzin *et al.*, 1995). If all the members of a cluster have the same global fold, the cluster is classified of type I; otherwise it is classified of type II. A total of 43, 13 and 47 clusters are classified as type I, II and III respectively[1].

Interestingly, it was observed that the parental chains members of the same cluster belong to the same functional family for clusters of type I, while they may belong to different functional families for members of type II or III clusters (Keskin & Nussinov, 2005, 2007).

3.2.2 Selection of dimeric non-homogeneous protein-protein complexes

In the database of Keskin *et al.* there are many two-chain complexes that are part of larger complexes, with more than two chains and consequently with many interfaces. The influence of all the binding partners has to be taken into account in order to study the dynamics of a particular chain. Consequently, the investigation of the dynamical properties of a chain that is part of dimeric complex is a problem with a lower level of complexity than the investigation of the dynamics of a chain that is part, for example, of a trimer. In this study we have decided to keep the analysis at the first level of complexity, so our investigations were limited to the dimeric proteins.

Consequently, the database of Keskin *et al.* (2004) was further filtered keeping only the PDBs:

1. with complete structural information;
2. consisting of only two chains;
3. relative to proteins classified as dimers, according to the informations in the PDB file and/or consulting the UNIPROT database (UniProt Consortium, 2008, 2009);
4. corresponding to structures for which the β Gaussian network model gives only six zero-energy motions (the roto-translational degrees of freedom).

[1] The data set is available at:
http://home.ku.edu.tr/∼okeskin/INTERFACE/INTERFACES.html.

Note that the last condition is motivated by the fact that we will evaluate the mobility of the amino acids using the βGM, as we will explain in section 3.4.1. The βGM, as all the topology-based ENM, gives reliable results for globular proteins. The existence of additional zero-energy modes beyond the ordinary six associated to the roto-translational degrees of freedom indicates that in the protein there are exposed loops or other parts, that in thermal equilibrium are expected to undergo diffusive-like motion that topology-based models are not suitable to treat.

The application of the described filters yielded to the selection of 12 interfaces of type I, from 8 different clusters; 8 of type II, form 6 clusters; 9 of type III, from 8 clusters. In order to avoid redundant or correlated data, we took only one representative for each cluster, the one ranked first in the clustering of Keskin *et al.* The selected proteins are reported in Table 3.1.

3.2.3 Protein-protein interaction types in the selected dimers

The above-mentioned classification scheme is aptly complemented by the notion of whether a given protein-protein complex observed in the PDB is biological or not, i.e. it corresponds to a biologically relevant interaction or to non-specific crystal packing contact. Furthermore, biological complexes can be obligatory or non-obligatory. In the first case the individual monomers that constitute the complex are not stable on their own; in the second case they are stable and can be found in the *free* form (non complexed).

To distinguish between obligate, non-obligate and crystal packing interactions we used here a recently developed automatic classification method, NOXclass (Zhu *et al.*, 2006). The classification is not trivial, therefore NOXclass do not provide a univocally characterization, but assigns a probability to each possibility. In Table 3.1 we provided the classification of the dimers in our dataset, reporting for each entry the most probable interaction type according to NOXclass.

Observe that overall there are: fourteen obligate interfaces (seven of type I, three of type II and four of type III); six non-obligate interfaces (one of type I, two of type II and three of type III); and two non-biological interfaces (one of type II and one of type III). Consequently, most dimers in Table 3.1 are obligate complexes.

The above facts prompt several observations in consideration that the subunits of a dimer have in general a different structure in the *bound* form (i.e. when complexed) and in the *free* form. For non-obligate interfaces, the *bound-free* conformational rearrangements are typically modest and are mostly localised at the interface region. For obligate complexes, however, the changes are expected to be dramatic. In fact, the

				Chain Size (a.a.)			Interface		Semi-int.		NOXclass
PDB	Chains		Classification	1^{st}	2^{nd}	both	$\mathring{A}^2$	a.a.	1^{st}	2^{nd}	interface
Interfaces of Type I											
1a03	A	B	calcium-binding protein	90	90	180	3071	82	37	45	obligate
1gdh	A	B	oxidoreductase	320	320	640	6254	148	74	74	obligate
1a05	A	B	oxidoreductase	357	357	714	5258	125	62	63	obligate
1ag1	O	T	isomerase	249	249	498	3025	75	37	38	obligate
1hii	A	B	hydrolase	99	99	198	3436	83	41	42	obligate
1mdi	A	B	electron transport/peptide	105	13	118	1340	34	24	10	non-obligate
2gsa	A	B	chlorophyll biosynthesis	427	427	854	9178	233	118	115	obligate
1ger	A	B	oxidoreductase	448	448	896	6777	172	86	86	obligate
Interfaces of Type II											
1azy	A	B	glicosyltransferase	440	440	880	1744	42	21	21	non-biological
1a0a	A	B	transcription factor	63	63	126	1860	44	22	22	obligate
1mv4	A	B	de novo protein	37	37	74	1833	43	21	22	obligate
1a93	A	B	leucine zipper	32	32	64	1565	34	17	17	non-obligate
1a15	A	B	chemokine	67	67	124	1480	38	18	20	obligate
1q6a	A	B	circadian clock protein	107	107	214	1810	46	23	23	non-obligate
Interfaces of Type III											
2fhw	B	A	signaling protein	27	24	51	1153	29	15	14	obligate
1jm7	A	B	antitumor	103	97	200	2769	72	39	33	obligate
1tmz	A	B	tropomyosin	32	32	64	1586	35	17	18	obligate
1an2	A	C	DNA-binding protein	86	86	172	2571	64	32	32	obligate
2sic	E	I	proteinase/inibitor	275	107	382	1617	45	31	14	non-obligate
1lfa	A	B	cell adhesion	183	183	365	628	21	10	11	non-biological
1shc	A	B	signal transduction/peptide	195	11	206	1456	34	25	9	non-obligate
2a93	A	B	leucine zipper	32	32	64	1568	34	16	18	non-obligate

Table 3.1: **Protein-protein interfaces investigated.** The columns represent respectively: the PDB ID; the chains that form the investigated interface; the classification, according to PRISM; the number of amino acids for the each chain and the total number; the size of the interface in $\mathring{A}^2$ and in number of amino acids involved; the number of interface amino acids in each chain; the classification of the interface into obligate/non-obligate/non-biological.

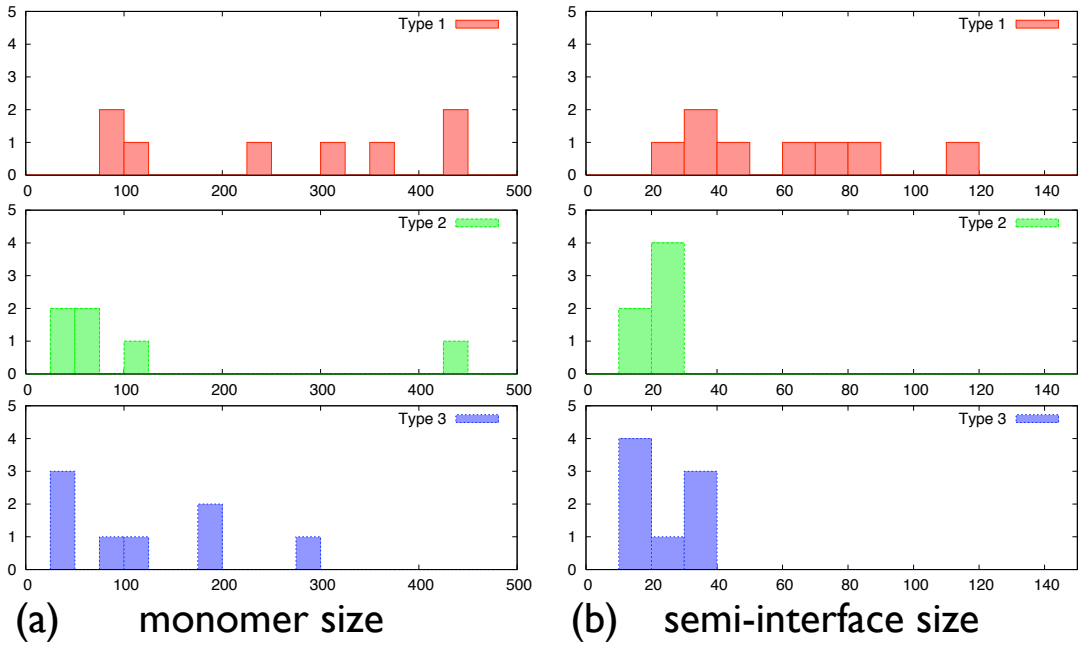

Figure 3.1: Distribution of sizes of (a) the first chain of the dimers reported in Table 3.1, and (b) their semi-interfaces formed with the second chain.

free form of the monomers may not even correspond to a well-defined structure. In view of these facts, to have a common term of comparison for the cases of obligate and non-obligate interfaces we shall base our considerations exclusively on the *bound* forms.

3.3 Structural properties of dimers and their interfaces

Before considering the dynamical properties of the dimers, we shall discuss their structural properties. At the most fundamental level the first quantity to consider is the size of the dimers, that is their length in terms of number of amino acids. The inspection of the complexes size, see Table 3.1, indicates that the dataset covers a wide range of lengths, from a minimum of 64 amino acids [PDB:1a93, 1tmz, 2a93] to 896 amino acids for oxidoreductase [PDB:1ger]. Most of the largest complexes are of type I. In fact, five of the seven complexes which have monomers comprising more than 200 residues are of type I, one is of type II and one of type III. This property is readily perceived in Fig. 3.1a, which reports the histogram of size of the first chain reported in Table 3.1 for each complex.

Further properties to consider are related to the monomeric interfaces. Well-established procedures exist to identify the interfaces. In this study, following Jones & Thornton (1997), we rely to the following definition: the *interface* in a protein-protein complex is the set of residues for which the accessible surface area (ASA), computed for the isolated components (*unbound* form) and for the complex (*bound* form), differs by more than 1Å^2. Since the considered complexes are constituted by two chains, the residues at the interface are divided into those which belong to the first and second chain, which constitute two *semi-interfaces*. The *interface size* is the number of residues that constitute the interface, and analogously the semi-interface size is the number of residues that constitute the semi-interface. The interface size is clearly the sum of the semi-interface sizes. For symmetric complexes the semi-interface sizes are equal, but for asymmetric ones they can be different. The *interface area* is defined, according to Conte *et al.* (1999), as the area of the accessible surface on both the partners that becomes inaccessible to the solvent due to the protein contacts. It is calculated as the sum of the ASA of the isolated components minus that of the complex. In this study the ASA per residue as well as the total ASA were obtained using the program NACCESS (Hubbard & Thornton, 1993), with a probe spere of radius 1.4Å. The interface size and area, and the semi-interface sizes, for each of the investigated dimers are shown in Table 3.1.

It is particularly interesting to relate the interface area and the number of residues that constitute the interface or the semi-interfaces. As for the sizes of the dimers, the sizes of the interfaces too span a wide range of values: from 21 amino acids (interface area: 628Å^2) of the cell adhesion protein [PDB:1lfa] to 233 amino acids (interface area: 9178Å^2) of chlorophyll biosynthesis [PDB:2gsa]. The largest interfaces are of type I, in fact seven of the eight interfaces of type I are larger than 3000Å^2, while none of the interfaces of type II and III are so large. All the six interfaces of type II and four of type III have area ranging from 1400Å^2 and 1900Å^2, which is indicative of interfaces of medium size (Conte *et al.*, 1999).

It is worth pointing out that for most of the type I and II interfaces, the number of residues in the two semi-interfaces is identical, while this is not true for type III dimers which (unlike cases I and II) typically consists of differently-sized monomeric units. Fig. 3.1b illustrates the size distribution of the semi-interface of the first chain reported in Table 3.1 for each complex, for type I, II and III separately. We observe that most of the semi-interfaces of type I consist of more than 30 amino acids, while those of types II and III involve less than 30 amino acids.

The inspection of Table 3.1 indicates that the largest interfaces pertain to large complexes. While this may be intuitively expected, it is interesting to notice that

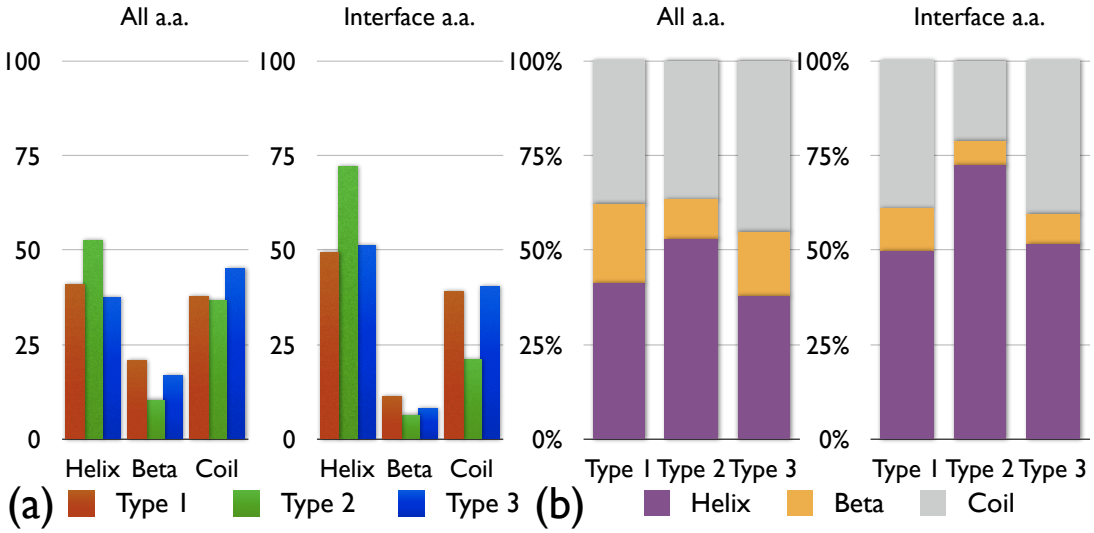

Figure 3.2: Distribution of secondary elements in the considered monomers and semi-interfaces, separately for each interface type.

for medium-sized interfaces there is not a simple correlation between complex size and interface size. For instance the leucine zipper [PDB:1a93], that is a complex of 64 amino acids, has an interface of 1565Å^2, which is larger than the interface of chemokine [PDB:1a15], a complex of 124 amino acids.

We conclude the structural characterization by discussing the secondary-structure content of the complex and their the semi-interfaces. This is assigned using the DSSP program (Kabsch & Sander, 1983), which defines seven secondary structure states: H (alpha helix), B (residue in isolated beta-bridge), E (extended strand, participates in beta ladder), G (3-helix), I (5 helix), T (hydrogen bonded turn) and S (bend). We then performed a subdivision of the amino acids in terms of helix (H, G, I), strand (B, E) and coil (T, S, or blank space).

The data, subdivided according to the three types of interfaces are reported in Fig. 3.2. More than 50% of the amino acids of type II dimers take part to helices, while the percentage decreases to 40% for type I and III cases. The fraction of amino acids in coil state is close to 40% for the complexes of the three types, and the fraction of amino acids taking part to strands is smaller than 20%. Considering the semi-interfaces, it is noticed an increase of helical content; more than 70% of the amino acids are in helical conformation for type II entries, while for type I and III the percentage is about 50%. A related decrease of coil and strand content is consequently observed.

3.4 Dynamical properties of dimers and dimeric interfaces

The investigation of the mobility of the amino acids at the dimeric interfaces was carried out from a two-fold perspective. On one hand, it is interesting to compare how the mobility of amino acids at the surface of a dimeric subunit depends on whether they take part to the dimer interface or not. Next, from the comparison of properties of several dimers we wish to establish which salient features (such as secondary and tertiary organization etc.) impact on the interface mobility. Before illustrating the results of our investigation, we will first show how it is possible to use the ENM, and βGM in particular, to evaluate the mobility of amino acids in one of the monomer, keeping into account of the influence of the binding partner

3.4.1 Evaluation of Amino Acids Mobility

For completeness we briefly recall here key facts about ENM approaches.

As we discussed at length in the first chapter, the internal large-scale concerted movements that proteins sustain around their native state in thermal equilibrium can be adequately captured by coarse-grained elastic network models (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2004; Sułkowska *et al.*, 2008; Tirion, 1996). These models typically take as input the C_α positions of the protein native state, and estimate the energy cost of deviations form the native state by adopting harmonic approximations.

In particular the βGM, as described in section 1.6.1, relies on the following quasi-harmonic approximation of the free energy $\mathcal{F}$ of the protein: a displacement $\mathbf{x} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ from the native state of the protein (where $\vec{x}_i$ is the displacement of the i-th C_α atom and N is the number of amino acids in the protein) is penalized by an increment of free-energy:

$$\mathcal{F}(\mathbf{x}) = \frac{C}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} \tag{3.1}$$

where $\mathbf{M}$ is a $3N \times 3N$ symmetric matrix which account for the pairwise interaction between amino acids, whose terms are calculated as explained in section 1.6.1. Note that this model has only one phenomenological parameter: the constant C, which can be set so that the outcomes of the model optimally fit experimental data (e.g. temperature factors in X-ray crystallography) or results of MD simulations.
In thermal equilibrium, the probability of occurrence of a deviation $\mathbf{x}$ is proportional to the Boltzmann factor:

$$P(\mathbf{x}) \propto \exp\left(-\frac{\mathcal{F}(\mathbf{x})}{K_B T}\right) \tag{3.2}$$

being K_B the Boltzmann constant and T the temperature.

As highlighted in the first chapter, the motion of a protein in thermal equilibrium can be described in terms of the overdamped langevin dynamics (see section 1.5). Within this context, and assuming here that the friction coefficients are the same for each C_α atom, it results that the collective large-scale movements of the system correspond to the low-energy modes of (3.1), obtained diagonalizing the matrix $\mathbf{M}$. This matrix has six null eigenvalues that correspond to rigid-body rotations and translations. For some proteins there are extra null eigenvalues, which are usually due to the fact that the protein is not globular and some amino acids experience diffusive motion. In this case the predictions are not trustable since the model is outside its field of applicability. For this reason dimers with more than six zero eigenvalues have been excluded from our dataset.

Let us indicate the normalized eigenvectors of $\mathbf{M}$ as $\mathbf{v}^\alpha = \{\vec{v}_1^\alpha, \vec{v}_2^\alpha, ..., \vec{v}_N^\alpha\}$ and the relative eigenvalues λ_α (in increasing order for $\alpha = 1, ..., (3N-6)$, having removed the roto-translations). Eigenvectors associated to the lowest eigenvalues give the directionality of the low-energy motions; the magnitude of the fluctuation along an eigenvector is directly proportional to the inverse of the relative eigenvalue. In particular it results (from equation 1.29 and using the assumption that we are here doing that the friction coefficients are the same for each C_α atom) that the average displacement of the amino acid i is:

$$\left\langle \|\vec{\mathbf{x}}_i\|^2 \right\rangle \propto \sum_{\alpha=1}^{3N-6} \frac{\|\vec{v}_i^\alpha\|^2}{\lambda_\alpha} \tag{3.3}$$

where the constant of proportionality depends on the temperature T of the system and on the value of the constant C. In this study we are mainly interested in changes of mobility, therefore we have fixed K_BT and C to 1, so that the estimated fluctuations are expressed in a common unit scale for all the proteins.

This yields to the following expression for the root mean square fluctuation (RMSF) of the i-th amino acid:

$$\mathrm{RMSF}(i) = \sqrt{\sum_{\alpha=1}^{3N-6} \frac{\|\vec{v}_i^\alpha\|^2}{\lambda_\alpha}} \tag{3.4}$$

which has been used as measure of its degree of mobility.

It is worth noticing that fluctuations predicted in this way will refer only to the backbone motion, not to the side-chain one, since only C_α positions are used to calculate the free energy. Finally, let us notice that the βGM, as others elastic network models, is "native-centric", i.e. it models around the input structure and gives predictions in its center of mass.

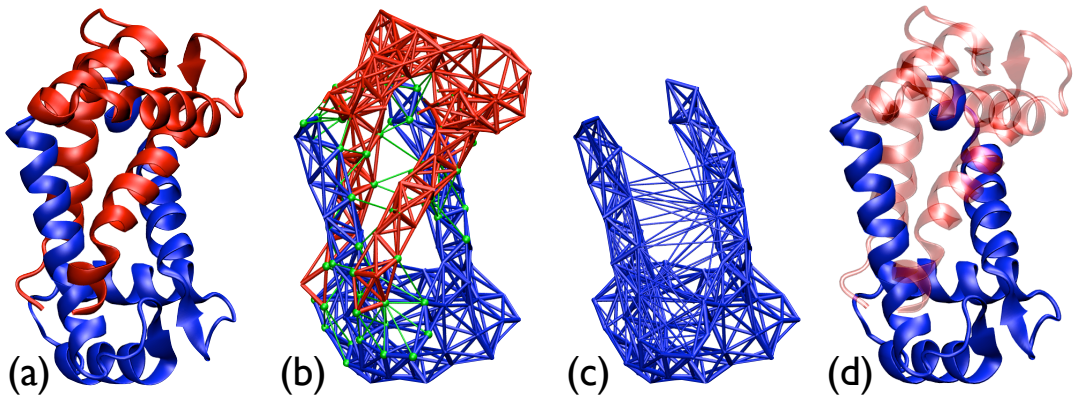

Figure 3.3: **Thermodynamic integration.** (a) A dimer is constituted by two chains A and B, respectively colored in blue and red. (b) Within an ENM approach, the free energy of the dimer can be written in terms of the interaction among the amino acids in chain A (blue), in chain B (red), and the coupling interaction between chains A and B (green). (c,d) As described in the text, equation 3.7, it is possible to compute the effective free energy governing the thermodynamics of subchain A alone (after integration of the degrees of freedom of subchain B).

3.4.1.1 Thermodynamic Integration

The complexes considered here are constituted by two monomers that we shall distinguish with the labels A and B, as shown in Fig. 3.3a. We want to study the mobility of monomer A in the *bound* form, i.e. we want to compute the degree of mobility of amino acids of monomer A, in its center of mass, taking into account the presence of B (Fig. 3.3d).

Within the βGM, we can model the free energy of the overall dimer as described by $\mathcal{F}(\mathbf{x})$ in equation 3.1. Note that a generic displacement $\mathbf{x}$ of the N amino acids of the dimer can be non-ambiguously decomposed into the displacement $\mathbf{x}_A$ of the N_A amino acids of chain A plus the displacement $\mathbf{x}_B$ of the N_B amino acids of chain B. Consequently the matrix $\mathbf{M}$ in 3.1 can be trivially reindexed so that the free energy cost $\mathcal{F}(\mathbf{x}_A, \mathbf{x}_B)$ of a displacement $\mathbf{x}_A$ of the amino acids in the first chain and $\mathbf{x}_B$ of the amino acids in the second chain is given by:

$$\mathcal{F}(\mathbf{x}_A, \mathbf{x}_B) = \frac{C}{2} \begin{pmatrix} \mathbf{x}_A^T & \mathbf{x}_B^T \end{pmatrix} \begin{pmatrix} \mathbf{M}_A & \mathbf{G} \\ \mathbf{G}^T & \mathbf{M}_B \end{pmatrix} \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix} \tag{3.5}$$

where the $3N_A \times 3N_A$ interaction matrix $\mathbf{M}_A$ represents the internal coupling between the coordinates $\mathbf{x}_A$, and analogously the $3N_B \times 3N_B$ interaction matrix $\mathbf{M}_B$ represents

the internal coupling between the coordinates $\mathbf{x}_B$. The $3N_A \times 3N_B$ matrix $\mathbf{G}$ and its transpose $\mathbf{G}^T$ represent the coupling between the coordinates $\mathbf{x}_A$ and $\mathbf{x}_B$. This is graphically illustrated in Fig. 3.3b, where $\mathbf{M}_A$, $\mathbf{M}_B$ and $\mathbf{G}$ are represented as the interactions colored respectively in blue, red and green.

As prescribed by equation 3.2, a displacement of $\{\mathbf{x}_A, \mathbf{x}_B\}$ has a probability $\mathcal{P}(\mathbf{x}_A, \mathbf{x}_B)$ to be observed which is proportional to the Boltzmann factor for the free energy $\mathcal{F}(\mathbf{x}_A, \mathbf{x}_B)$ written as in equation 3.5. Therefore the probability $\tilde{\mathcal{P}}(\mathbf{x}_A)$ to observe a displacement $\mathbf{x}_A$ in protein A, independently on the displacement in protein B, is obtained integrating $\mathcal{P}(\mathbf{x}_A, \mathbf{x}_B)$ over all the possible displacements of $\mathbf{x}_B$ (Carnevale *et al.*, 2006; Hinsen *et al.*, 2000), and it yields:

$$\tilde{\mathcal{P}}(\mathbf{x}_A) = \int \mathcal{P}(\mathbf{x}_A, \mathbf{x}_B)\, d\mathbf{x}_B \propto \exp\left(-\frac{C}{2\,K_B T}\, \mathbf{x}_A^T(\mathbf{M}_A - \mathbf{G}\mathbf{M}_B{}^{-1}\mathbf{G}^T)\mathbf{x}_A\right) \tag{3.6}$$

Thus the comparison between (3.2) and (3.6) gives that the effective free energy $\tilde{\mathcal{F}}(\mathbf{x}_A)$ governing the effective interaction among the N_A amino acids of chain A have still a quadratic form, and can be written as:

$$\tilde{\mathcal{F}}(\mathbf{x}_A) = \frac{C}{2}\, \mathbf{x}_A^T\, \tilde{\mathbf{M}}_A\, \mathbf{x}_A \tag{3.7}$$

where $\tilde{\mathbf{M}}_A = (\mathbf{M}_A - \mathbf{G}\mathbf{M}_B{}^{-1}\mathbf{G}^T)$ is the interaction matrix of monomer A that opportunely accounts for the influence of B. This is represented graphically in Fig. 3.3c.

The degree of mobility of the amino acids in monomer A, calculated considering also the influence of the binding partner B, is therefore given by equation (3.4), where the eigenvalues and eigenvectors are those of matrix matrix $\tilde{\mathbf{M}}_A$.

3.4.2 Mobility of the amino acids at the interface

It is important to point out that the characterization of the mobility of amino acids in a protein depends critically on the "reference frame" that is used for its description. Usually, the adopted reference frame is the one where the equilibrium mean square displacement of *all* amino acids are minimised. In such reference frame the center of mass of the entire protein (complex) of interest remains fixed in space. For multimeric or multidomain proteins this choice is not necessarily appropriate, as an appreciable relative motion of the protein subparts can lead to artifactual results (Henzler-Wildman *et al.*, 2007b).

In the present context, where protein dimers are considered, the consideration of the appropriate reference frame is therefore very important. In particular, when using the above-mentioned criterion there are two possible natural choices for the reference

frame for characterize the amino acid motion namely to minimize the mean-square displacements of the *entire* protein or of *one of the monomeric subunits*.

The second choice is the one that will be adopted hereafter. In fact, considering the "subunit frame of reference" appears appropriate in view of the comparisons that will be carried out between the *bound*/*unbound* forms of the dimers. For subunits in the *bound* state, the fluctuation dynamics of the amino acids will be calculated by properly taking into account the presence of the partner monomer, as illustrated in the previous sections.

For a dimer of type I or II the choice of which subunit to consider is arbitrary, since they are usually identical and both the semi-interfaces are shared by members of the same cluster. The situation is different for type III dimers, where only one side of the interface is shared by members of the same cluster (and belongs to the first chain reported in Table 3.1). We will therefore use the first chain reported in Table 3.1 to define the "subunit frame of reference", for all our dimers.

As anticipated, we shall first compare the mobility of residues in the semi-interface of interest with that of other surface residues. We have defined, following (Jones & Thornton, 1997), the surface residues as those having a relative accessible surface area (RASA) greater than 5%. As for the ASA, the RASA per residue was obtained using the program NACCESS (Hubbard & Thornton, 1993).

Considering all the proteins in our dataset, the total number of residues in the monomeric units is 3774; $\sim$ 21% of them are at the semi-interface, and $\sim$ 57% are surface but not interface residues. If the partner monomer were not present then most of the residues at the semi-interface ($\sim$ 96%) would be classified as surface residues.

By inspecting Fig. 3.4a it is possible to compare the distribution of the root mean square fluctuations (RMSF) of residues at the semi-interface (for all the studied proteins) with that of surface residues which are not at the semi-interfaces. The difference of the two distributions is more readily perceived after normalization, see Fig. 3.4b. Residues at the semi-interface appear to be, on average, less mobile than residues at the surface. The graph further indicates that the fraction of resides with a RMSF lower than 1 (in the units of the elastic network model) is $\sim$ 60% for semi-interface residues, and it is $\sim$ 44% for surface residues. Finally, the graph offers a useful indication of the typical range of amino acid mobility; in particular an RMSF value of 2 is indicative of a rather large degree of mobility, as only about 10% of the amino acids at the semi-interface have an RMSF value that overcomes this threshold.

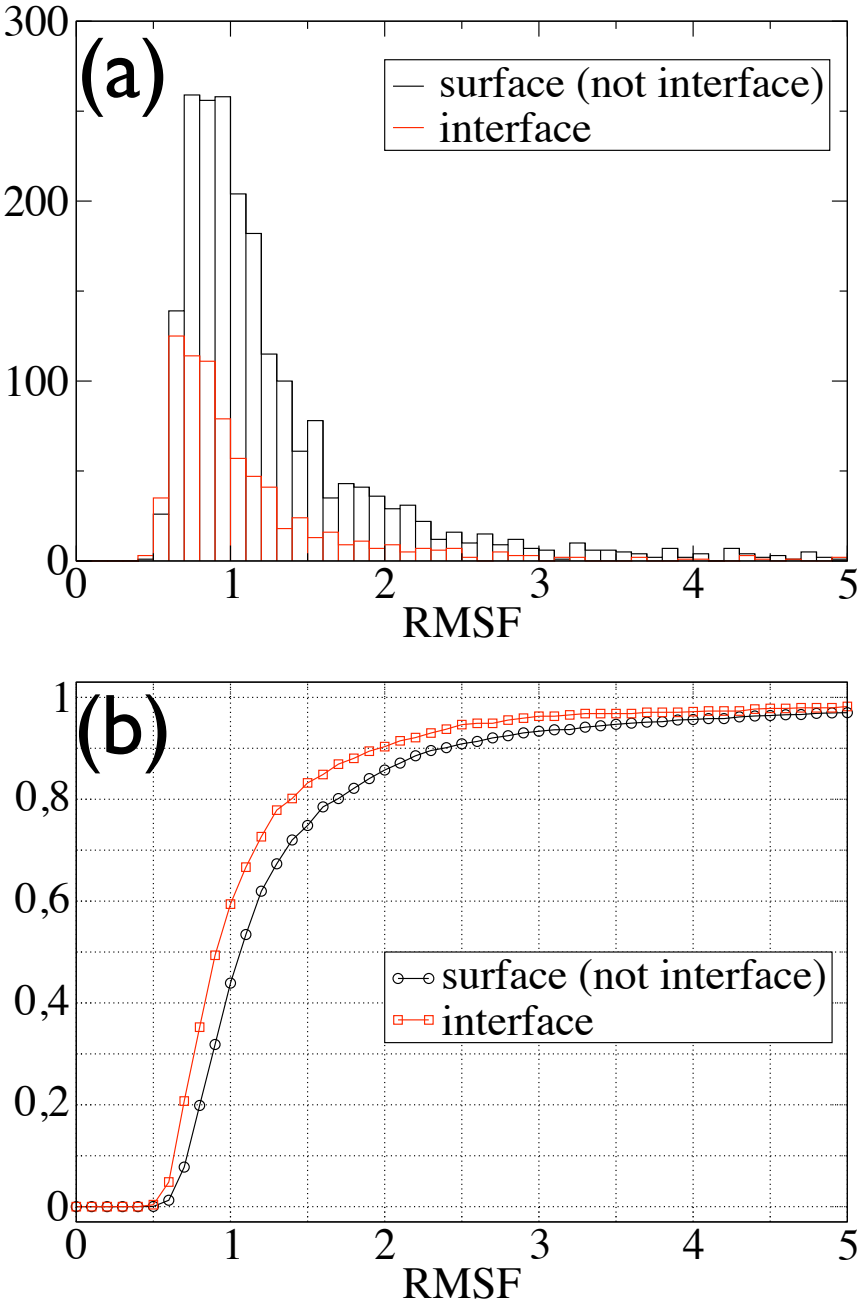

Figure 3.4: (a) Distribution and (b) normalized cumulative distribution of the RMSF for the resides at the semi-interface, and at the surface but not at the semi-interface.

3.4.3 Factors affecting the mobility of the amino acids at the interface

A priori, the observed diminished mobility of amino acids at the semi-interfaces, compared to other surface residues, could be ascribed to two main factors:

(i) the intrinsic structural architecture of the monomeric unit (such as the locality of the inter-residue contacts), and/or

(ii) the contact interactions with the partner monomer, which act as a mobility-limiting constraint.

Aspects related to the interplay of these two factors, were considered before in connection with the mobility of *free* and *bound* forms of monomeric units (Rajamani *et al.*, 2004; Smith *et al.*, 2005; Yogurtcu *et al.*, 2008). As discussed, the scope of the *free*/*bound* comparison is essentially restricted to non-obligate interfaces. Most of the interfaces considered here are obligate, implying that a well-structured *free* form of the monomer may not necessarily exist

For the purpose of the present study it is important to point out that to understand the interplay between (i) and (ii) it is not necessary to consider the *free* monomer. In fact, considerable insight can be gained by using an elastic network model to study the fluctuation dynamics of the monomer of interest and comparing the behaviour when the partner monomer is present and when it is absent. If a realistic force-field was employed to study the dynamics, the removal of the partner monomer would lead to a rapid loss of structural organization of the subunit of interest; this is because the isolated monomer would not correspond to a minimum of the free energy. Instead, by resorting to an elastic network model, it is possible to study the "intrinsic" fluctuation dynamics of the subunit of interest because the ENM approach amounts to introducing a model free energy that, by construction, has a minimum in the input reference structure, in this case the "virtual", *unbound* structure.

In conclusion, the relative role of factors (i) and (ii) can be ascertained by an ENM calculation of the fluctuation dynamics of the monomer of interest in the abence of the partner monomer and in its presence. In the latter case, a suitable thermodynamic integration of the degrees of freedom of the partner monomer needs to be carried, as explained previously.

In can be anticipated that three possible cases can emerge from the comparison:

(a) the fluctuation of the semi-interface residues are small both in the *bound* and in the virtual *unbound* forms;

(b) the fluctuation of the semi-interface residues are small in the *bound* form and large in the *unbound* one;

(c) the fluctuation of the semi-interface residues are large both for the *bound* and the virtual *unbound* forms.

As the partner monomer acts as a constraint for the mobility of the monomer semi-interface, the fourth case, where the semi-interface is more mobile in the *bound* than the *unbound* form cannot occur.

Case (a) would indicate that factor (i), i.e. the structural architecture of the monomer, is the main one influencing the mobility (the low degree of mobility, in this case) of the residues. On the contrary, case (b) indicatives that the main factor is (ii), i.e. the contact with the monomer partner. Case (c) is subtler as it would indicate that neither factors (i) and (ii) are responsible for the observed diminished mobility of the semi-interface residues, compared to other surface amino acids. It would be particularly interesting to observe that semi-interface fluctuations in the *bound* and virtual *unbound* forms were similar, as this would indicate that the binding partner has an interface organised so to not alter the "intrinsic" fluctuations of the monomer of interest.

The scatter plots of the *bound* fluctuation and the the virtual *unbound* fluctuation, for the residues at the first semi-interface of the studied complexes, are given in Fig.3.5. In these graphs, residues that are representative of behavior (a) would appear in proximity of the origin; residues representing behavior (b) would occur along the x axis; and residues representing behavior (c) would distribute in the region $y < x$. In the notable case where the partner does not influence the intrinsic fluctuations of the first monomer, the points would be distributed along the line $y = x$.

Fig. 3.5 indicates that all the three possible behaviors are present, albeit with different weight. The first conclusion is, therefore, that there is not a typical relative strength of factors (i) and (ii).

By inspecting Fig. 3.5, it emerges that the best examples of semi-interface mobility that is diminished by interaction with the partner unit (case b), are provided by the oxidoreductases 1a05 and 1gdh, the hydrolase 1hii and the calcium-binding protein 1a03. All these complexes have a type I interface. The best examples of fluctuations not appreciably affected by the partner (case c) are observed for the chemokine 1a15, the transcription factor 1a0a, and the leucine zipper 1a93. All these complexes are of type II.

It is very interesting to notice in Fig. 3.6 that, while examples of case (a) are found in complexes of type I, II and III, the behavior of type (b) is found only in type I,

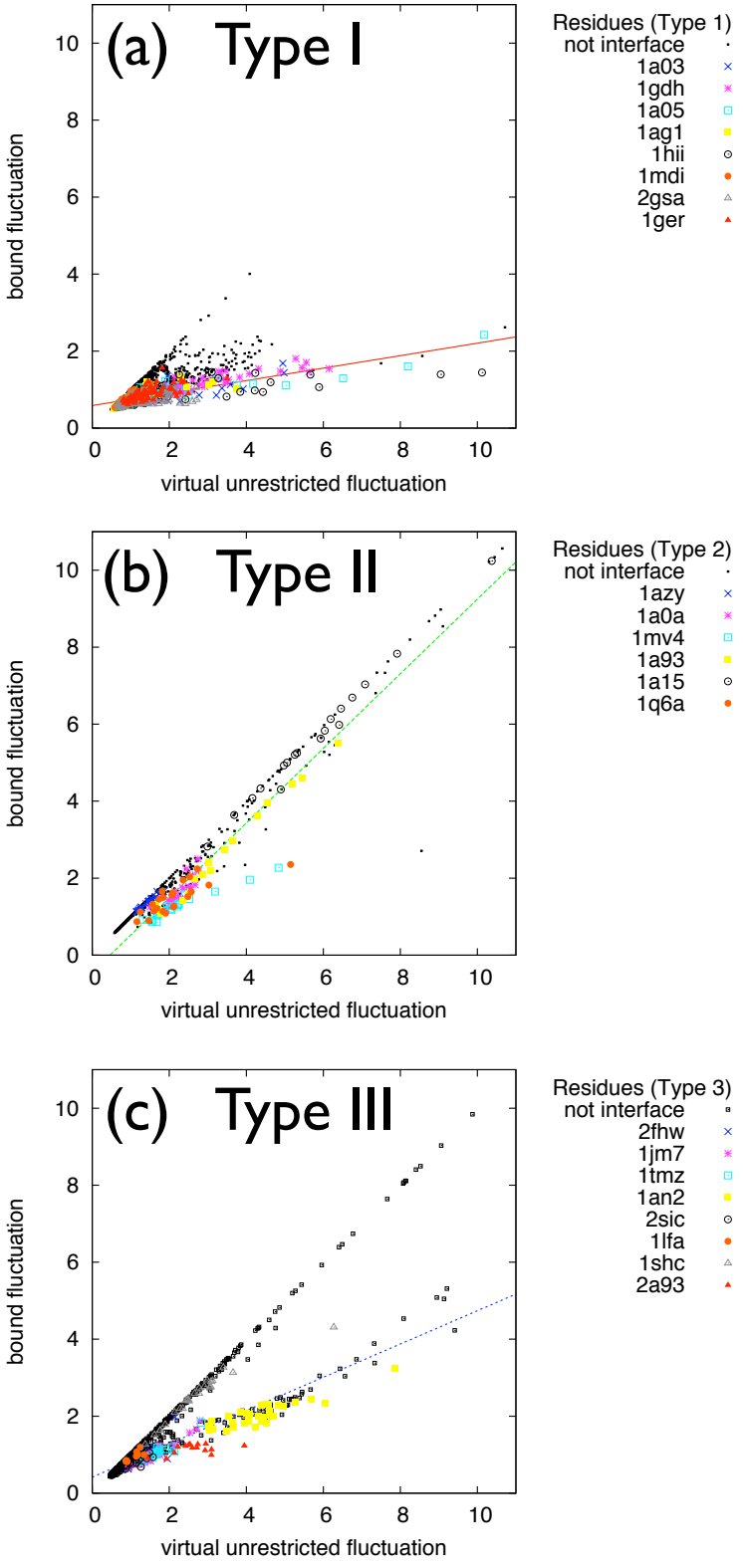

Figure 3.5: Scatter plot of the *bound* fluctuation and of the virtual *unbound* fluctuation for the amino acids in the first chain in the investigated proteins reported in Table 3.1. The amino acids and the semi-interface have been represented differently for each protein, and reported in panel (a), (b) of (c) in relation to their interface type (according to the classification of Keskin *et al.*).

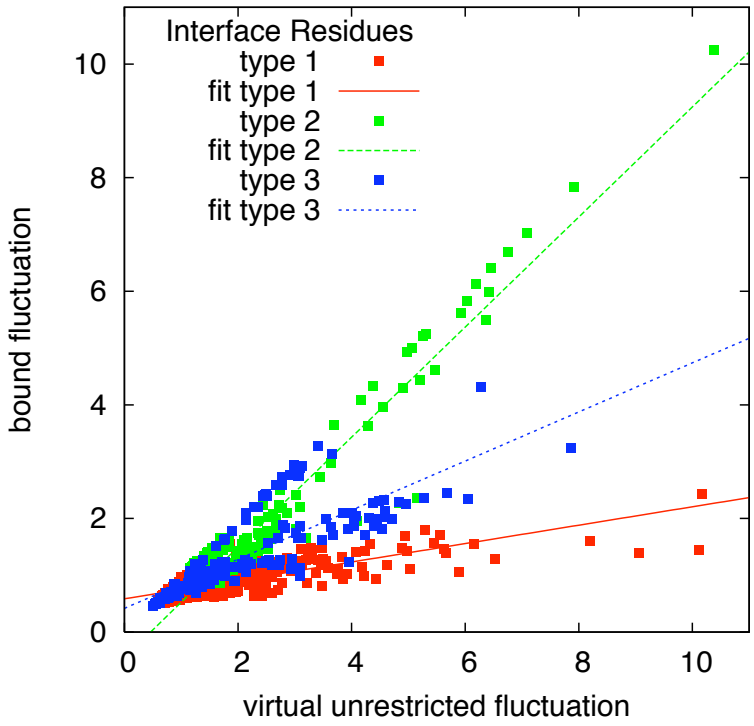

Figure 3.6: Scatter plot of the *bound* fluctuation and of the virtual *unbound* fluctuation for the amino acids at the semi-interface. Interfaces of type I, II and III have been colored respectively in red, green of blue. Interpolating lines for each type of interface are also shown.

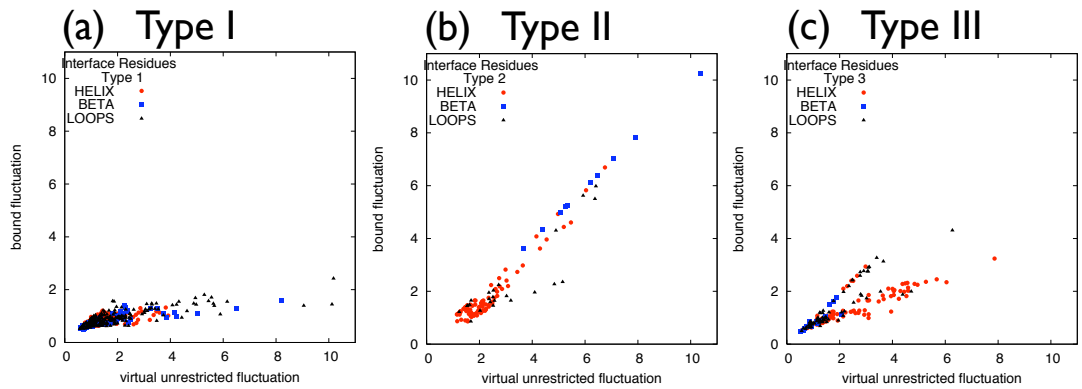

Figure 3.7: Scatter plot of the *bound* fluctuation and of the virtual *unbound* fluctuation for the amino acids at the semi-interface, colored according to their secondary structure content.

and the behavior (c) is found only in types II and III. Many complexes of type I are examples of behavior (b), as highlighted by the interpolating line (very low angular coefficient). Furthermore, in most of type II complexes, the partner seems not to affect the intrinsic fluctuations of the monomer residues (as highlighted in the figure by the interpolating line, with angular coefficient close to 1). By converse, in complexes of type III it appears that the partner partially influence the amount of fluctuation of the semi-interface residues (as highlighted by the interpolating line, with angular coefficient close to 1/2), although a definite conclusion cannot be drawn in this case due to the limited size of the sample.

It can also be observed that the examples of behavior (b) are obligate complexes, while examples of behavior (c) came either from obligate or non-obligate complexes. Furthermore, the size of the interfaces seems to be correlated to the observed behavior. In fact, all the mentioned examples of behavior (b) have an interface area larger than 3000Å^2, while behavior (c) is observed in interfaces of medium size. Behavior (a) is observed both in large and medium interfaces.

Finally, it is worth considering the secondary-structure content of the semi-interfaces and its correlation with the three types of behaviour. To address this point, the same data of Fig. 3.5 have been reproduced in Fig. 3.7 using a color scheme depending on the secondary structure to which each amino acid belongs to. Observing the plot for type I (Fig. 3.7a), it can be noticed that the highest virtual "unrestricted" fluctuation is observed for interface residues belonging to loops and beta strands. These amino acids experience the largest variation of fluctuation due to the influence of the partner monomer. In the plot for type II (Fig. 3.7b) it emerges that the interface residues are

mainly in alpha-helices (excluded some beta strands that came from [PDB:1a15]). In the plot for type III (Fig. 3.7c) most of the more fluctuating residues are alpha-helices.

3.5 Discussion

It was observed that for some complexes the residues at the interface have a low degree of mobility, and that this is an intrinsic characteristic of the structural architecture of the monomer units. In other cases it is the binding partner that causes the interface residues to have a limited mobility. This behavior is observed in large interfaces of type I, according to the classification of Keskin *et al.*. There are also cases where the interface residues do not have a low mobility, despite the presence of the partner monomer. This behavior is observed in medium-sized interfaces which are usually of type II.

An attempt to rationalise these observations can be made considering the relative role that enthalpic and entropic effects are expected to have on the formation of protein dimers.

It is known that entropy plays a fundamental role in the binding processes, and several studies have shown that the entropy associated to the fluctuations (usually called vibrational energy) makes a substantial contribution to the association free energy of a complex (Daniel *et al.*, 2003; Tidor & Karplus, 1994). The precise calculation of this entropy is a challenging task even for non-obligate complexes, and therefore it is beyond the scope of the present study to attempt a quantitative estimate of this contribution to the formation of the (often obligate) complexes considered here.

Nevertheless, the following euristic argument can be helpful to interpret the results from a simple perspective.

If the presence of the monomeric partner diminishes the mobility of residues at the semi-interface, this indicates that the partner unit provides a significant limitation to the conformational space of the semi-interface. This will come at an entropic cost which must be compensated by an enthalpy gain for the formation of the dimer.

This simple observation provides some clues as to the different behaviour observed for large versus small interfaces. In fact, a large surface of interaction can more easily lead to a large enthalpy gain and hence these complexes can afford to have an appreciable loss of conformational entropy upon binding (and hence a decrease of semi-interface mobility upon going from the *unbound* to the *bound* form). Consistently with this observation it can be noticed that this behavior was observed only in type I interfaces, which are very specific and characterised by a large gain in enthalpy upon dimerization.

Furthermore, if the residues at the semi-interface have an intrinsically low mobility then binding will not appreciably modify the fluctuation amplitude of the semi-interface. Consequently, at variance with the previous case, there is not an appreciably entropic cost to be compensated by enthalpy. This explains why this behavior was found in all the different typologies of the interfaces, also in medium and non specific ones.

Analogously, it is expected that there is not an appreciably entropic cost to be compensated by enthalpy also for the semi-interfaces whose mobility is largely unaffected by the presence of the partner monomer. Such kind of behavior is likely to be observed for semi-interfaces with a great degree of dynamical affinity. Consistently with these observations, examples of this behavior are found most frequent for type II interfaces and less so for type III. In fact interfaces of type II have both of the sides that are specific, being conserved among members of the same cluster, while in type III only one side is conserved and consequently less specific.

Chapter 4

Dynamics-based Alignment: a Pairwise Comparison of Low-energy Modes in Proteins

4.1 Introduction

Proteins are customarily characterized according to their sequence, structure and function. Available alignment tools detect similarities among different proteins in the sequence (Altschul *et al.*, 1997; Chenna *et al.*, 2003; Higgins & Sharp, 1988; Thompson *et al.*, 1994) and in the structure (Holm & Park, 2000; Holm & Sander, 1996, 1999; Konagurthu *et al.*, 2006; Micheletti & Orland, 2009; Notredame *et al.*, 2000; Shatsky *et al.*, 2004b). These tools have helped to clarify the sequence and structure relationship. It is known that high degree of sequence similarity (sequence identity above about 30%) reflects into a structural similarity (Chothia & Lesk, 1986; Chothia *et al.*, 2003; Orengo & Thornton, 2005). On the other end, it has been observed that the same fold is sometimes adopted also by proteins with negligible sequence similarity (Holm & Sander, 1994; Murzin *et al.*, 1995; Orengo *et al.*, 1997). This behavior is typically interpreted in terms of convergent evolution of proteins structure (Andreeva & Murzin, 2006; Banavar *et al.*, 2002; Chen *et al.*, 1997; Denton & Marshall, 2001a; Krishna & Grishin, 2004; Seno & Trovato, 2007).

The sequence of a protein encodes and determines both its structural and functional properties. Note however that the knowledge of the three dimensional structure represents an important source of additional information with respect to the sequence code alone for the understanding of the molecular mechanisms that regulate the biolog-

ical function of a protein. These considerations and the availability of structures with no biochemical annotations have motivated attempts to predict protein function from sequence and structural information (Redfern *et al.*, 2008; Sadowski & Jones, 2009). The determination of the function is "easy" when the investigated protein has striking similarities in sequence with other proteins for which the function is well known, as this denotes common evolutionary pathways. However, when no striking homologies are observed, determining function directly from tertiary structure has proven to be a highly challenging problem. The problem is related to the fact that, despite a significant correlation is observed between certain folds and some specific functions, the same structure can be used by different proteins to perform different functions, and it is not necessary to adopt a particular structure to carry out a particular function (Ausiello *et al.*, 2007; Bork *et al.*, 1993; Carnevale *et al.*, 2006; Russell, 1998).

In this chapter we introduce and apply a general quantitative scheme to address functional relationships of enzymes by extending the alignment procedures to the dynamical properties. We focus on a common although not universal feature of enzymatic function, namely internal large-scale concerted movements. A large body of evidence links these movements to the structural changes that often accompany protein functions, as we have seen in the second chapter for the specific case of adenylate kinase. Moreover, the displacements involved in allosteric changes in many proteins occur along the collective coordinates corresponding to the low-energy modes of the two biologically relevant states (Alexandrov *et al.*, 2005; Delarue & Sanejouand, 2002; Falke, 2002; Ming & Wall, 2005; Rod *et al.*, 2003; Smith *et al.*, 2005; Zheng *et al.*, 2007).

As we have already discussed in the first chapter, the collective and large-scale character of these fluctuations has justified their characterization by means of simplified approaches, typically elastic network models (ENM) (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2004; Sułkowska *et al.*, 2008). We recall that these models rely on a simplified free-energy function with quadratic dependence on displacements of amino acids from their reference position. Linear combinations of the ten lowest-energy modes predicted by ENMs are generally sufficient to describe most of the conformational fluctuations observed in extensive MD simulations as well as functionally-oriented changes between apo and holo forms of enzymes (Alexandrov *et al.*, 2005; De Los Rios *et al.*, 2005; Delarue & Sanejouand, 2002; Falke, 2002; Ming & Wall, 2005; Rod *et al.*, 2003; Smith *et al.*, 2005; Zheng *et al.*, 2007), as we have seen specifically for adenylate kinase in the second chapter.

Here we apply the collective low-energy modes of amino acids determined by these simplified models, to protein alignment. Unlike in structural alignments, matched amino acids need show only loose spatial proximity. The spatial tolerance is such that

the relative movements in the two enzymes are well defined, yet sufficiently generous to establish correspondences between, for example, different types of secondary structure elements.

This study extends the recent surveys of Capozzi *et al.* (2007); Carnevale *et al.* (2006), where common features were detected among the low-energy modes of proteolytic enzymes and EF-hand motifs, the structural alignments of which were known. Here we avoid the asymmetric treatment of structural and dynamical features by using a novel optimization scheme that identifies the set of amino acids which has the highest consistency of large-scale displacements, within tolerant structural correspondence. Combining structural and dynamical criteria on an equal footing appears to be necessary to detect general analogies of the internal motion of biomolecules. A pure dynamical alignment, *i.e.,* rewarding the consistency of the low-energy modes' directionality in two sets of amino acids regardless of their relative spatial relationship would, in fact, not necessarily identify *regions* that undergo analogous dynamical modulations. At the same time, the matching of the ENM-derived low-energy modes does not simply establish correspondences of simple local geometric features of two protein structures. The algorithm, in fact, goes beyond capturing correspondences between the profiles of amino acid mobility, which largely reflect static local structural (density) features (Halle, 2002), and promotes the accord of non-local correlations of amino acid displacements in thermal equilibrium. In view of the collective, non-local, nature of the ENM-derived equilibrium fluctuations exploited by the algorithm it appears justified to term the alignment as *dynamics-based*.

The chapter is organized as follows. The first part is devoted to a detailed description of the dynamics-based alignment method. In the second part the alignment procedure is applied to all pairs from a set of 76 enzymes which represent the main functional families with minimal structural redundancy. The alignment score of $\sim$30 enzyme pairs was found to be outstanding by standard criteria of statistical significance. Two thirds of such alignments reflect global or partial correspondences in the fold architecture. Notably, the remaining third involve proteins with only loose analogies of secondary and tertiary structural elements but with precisely-matching large-scale dynamics. Even for structurally-dissimilar pairs of enzymes the dynamics-based alignment can induce a remarkable spatial superposition of functionally-relevant regions. This suggests a biological rationale underlying specific common concerted movements. Further development of tools capable of detecting such dynamical correspondences is expected to provide novel elements and perspectives to address the relationship between sequence, structure and function of enzymes.

4.2 Dynamics-based alignment

The pairwise alignment method that we have developed for the comparison of the low-energy modes in proteins, is aimed at establishing correspondences between groups of amino acids experiencing similar (large-scale) motions in two given proteins. As in other contexts (Altschul *et al.*, 1997; Chenna *et al.*, 2003; Holm & Sander, 1996; Konagurthu *et al.*, 2006; Lesk, 2004; Notredame *et al.*, 2000; Shatsky *et al.*, 2004a,b) an alignment is a one-to-one pairing among a subset of "marked" amino acids in the two proteins. The number n of marked amino acids range up to the maximum length dictated by the shorter protein. Given a particular value of n, the alignment procedure, schematically represented in Fig. 4.1, is based on an iterative scheme that proceeds in the following way:

(a) generation of a tentative alignment of n amino acids, i.e. selection of a subset of n amino acids in each protein to be put in a one-to-one correspondence;

(b) identification of the low-energy modes of the selected amino acids, calculated within an elastic network approach;

(c) evaluation of the spatial/dynamical consistency of the tentative alignment through an *alignment score* that measures the accord of *both* the spatial position *and* the concerted movements of amino acids in pairwise correspondence.

These steps are repeated within a stochastic optimization method for maximizing the alignment score over thousands possible correspondences of n amino acids, finally obtaining the best scoring alignment of length n.

For each protein pair we apply this procedure for several values of n. We associate to each best scoring alignment of n amino acids a *homogeneous score* that allows the comparison of alignments of different lengths. The optimal alignment for the specific protein pair is finally found by taking as "winning" alignment the one with maximum homogeneous score.

A detailed description of the mentioned steps of the alignment follows.

4.2.1 Calculation of low-energy modes of marked amino acids

We start by considering a given tentative alignment of n amino acids between two proteins. As mentioned, we need to identify the lowest-energy modes of the amino acids marked for the alignment, and to compare them in a common Cartesian reference frame. We shall accordingly assume that one of the two proteins has been roto-translated to minimize the root mean square distance between the matching amino acids. After the

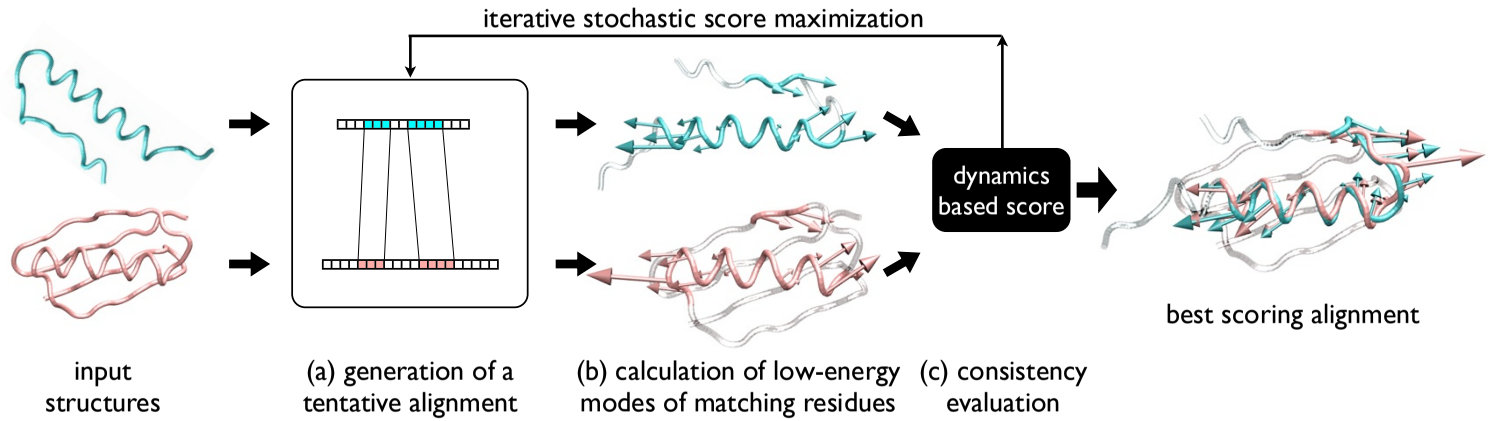

Figure 4.1: **Schematic diagram of the dynamics-based alignment.** For each pair of protein structures (left) thousands of tentative alignments, i.e. one-to-one correspondences of amino acids, are considered. For each alignment (a) the aligned amino acids are superposed, and (b) for each structure, the low-energy modes of the aligned amino acids are calculated within the elastic network model. A numerical score, to evaluate the quality of each alignment considered (c), measures the consistency between the structural alignment, and the dynamics reflected in the low-lying normal modes. That is, a particular residue-residue correspondence in the alignment contributes favourably to the score if the amino acids are *both* well-superposed spatially, *and* show similar patterns of displacement in the low-lying modes. The optimal alignment is identified by maximizing this score through a stochastic optimization loop. Images of protein structures and low-energy modes were produced with the VMD graphical package (Humphrey *et al.*, 1996b).

optimal superposition, a model free energy is introduced to characterize the thermal equilibrium fluctuations of the marked amino acids. To this purpose we adopted the well-established elastic network approach (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2004). As described in the first chapter, we recall that the free-energy $\mathcal{F}$ associated to a displacement $\delta\vec{x}_k$ of the k-th C_α from its reference position is:

$$\mathcal{F} = \frac{1}{2}\sum_{i,j} \delta\vec{x}_i \cdot M_{ij}\delta\vec{x}_j. \tag{4.1}$$

To calculate the entries of matrix M, in the dynamics-based alignment we have implemented the β-Gaussian network model of Micheletti *et al.* (2004), illustrated in section 1.6.1. Note however that the dynamics-based alignment method could be used in conjunction with other ENMs for the evaluation the matrix M, or use the interaction matrix obtained by QHA of an MD trajectory.

The model energy $\mathcal{F}$ is used to compute, for each protein, the effective matrix, $\tilde{M}$, providing the quadratic potential of mean force acting on the sole degrees of freedom of interest, that is, the positions of the n C_α's marked for alignment. In the following we shall assume that the amino acids have been re-indexed so that the first n amino acids (out of a total of N amino acids) correspond to the marked ones. To illustrate how $\tilde{M}$ is calculated it is useful to divide the M matrix into blocks reflecting the distinction of the degrees of freedom that we wish to retain (the displacement of the first n amino acids), from the rest:

$$M = \begin{pmatrix} M^a & V \\ V^T & M^b \end{pmatrix} \tag{4.2}$$

where the superscript T denotes the transpose. The physical interpretation of the blocks is straightforward: M^a corresponds to the interactions among the first n amino acids themselves; M^b contains the interactions within the remaining $N-n$ amino acids and V contains the interactions between the two groups.

The problem of finding $\tilde{M}$ is analogous to the calculation of the effective matrix of interaction of a monomer in contact with another monomer, as illustrated in the previous chapter (section 3.4.1.1). Owing to the simple quadratic nature of $\mathcal{F}$ in eqn. (4.1), the calculation of the effective energy $\tilde{\mathcal{F}}(\delta\vec{x}_1, \delta\vec{x}_2, ...\delta\vec{x}_n)$ governing the effective interaction among the first n amino acids can be done explicitly (Carnevale *et al.*, 2006, 2007a; Hinsen *et al.*, 2000) yielding

$$\tilde{\mathcal{F}} = \frac{1}{2}\sum_{i,j=1}^{n} \delta\vec{x}_i \cdot \tilde{M}_{ij}\delta\vec{x}_j \equiv \frac{1}{2}\sum_{i,j=1}^{n} \delta\vec{x}_i \cdot \left[M^a_{ij} + \Delta M_{ij}\right]\delta\vec{x}_j \quad , \tag{4.3}$$

where $\Delta M = -V[M^b]^{-1}V^T$, being $[M^b]^{-1}$ is the pseudoinverse of M^b. The lowest-energy non-zero modes of the matching amino acids are identified as the eigenvectors associated with the smallest non-zero eigenvalues of $\tilde{M}$. In the following we shall indicate with $\{\vec{v}_i^{\,\alpha}\}_{i=1,\ldots,n}$ and $\{\vec{w}_i^{\,\alpha}\}_{i=1,\ldots,n}$ the α-th low-energy mode of the marked amino acids for the first and second protein, respectively.

Note that the lowest-energy modes have to be recalculated for each different choice of the amino acids marked for the alignment. This calculation is computationally expensive, as it involves matrix inversion, multiplication and diagonalization.

4.2.2 Spatial/dynamical consistency of an alignment of n amino acids

As customary we shall assume that the ten lowest-energy modes are sufficient to account for the essential dynamics of the aligned amino acids (Amadei *et al.*, 1999). Accordingly, the quality of each tentative alignment involving n amino acids is quantified with the following combined measure of spatial and dynamical consistency:

$$q_n = \sqrt{\max\left\{0\,,\,\frac{1}{10}\sum_{\alpha,\beta=1}^{10}\left[\sum_{j=1}^{n}\vec{v}_j^{\,\alpha}\cdot\vec{w}_j^{\,\beta}\right]\left[\sum_{i=1}^{n}\vec{v}_i^{\,\alpha}\cdot\vec{w}_i^{\,\beta}\,f(d_i)\right]\right\}} \tag{4.4}$$

where d_i is the distance between the C_α positions of the ith aligned residue of the two proteins, and

$$f(d) = \frac{1}{2}\left[1 - \mathrm{tgh}\left(\frac{d-d_c}{2}\right)\right]$$

is a distance weighting factor interpolating the asymptotic values of 0 and 1 for distances respectively much larger and smaller than $d_c = 4$ Å. Observe that q_n is bounded between 0 and 1, and its value is 1 in case of a perfect correspondence between both low-energy modes and distances of aligned amino acids (which we recall have been previously optimized by the superposition of the aligned amino acids).

The measure (4.6) does not depend on the choice of the basis of the lowest-energy modes and generalises the familiar root mean square inner product,

$$\mathrm{RMSIP} = \sqrt{\frac{1}{10}\sum_{\alpha,\beta=1}^{10}\left|\sum_{i=1}^{n}\vec{v}_i^{\,\alpha}\cdot\vec{w}_i^{\,\beta}\right|^2} \tag{4.5}$$

used to measure the consistency of the dynamical spaces of the same protein in two different MD trajectories. The inclusion of the structural modulation $f(d_i)$ is sought here as we wish to promote not the mere overall dynamical correspondence of matching amino acids *per se* but only when these also have a good space proximity.

4.2.3 Stochastic exploration of the space of amino acids correspondences

The space of *all* possible alignments of two proteins is too large for an exhaustive exploration. We therefore relied on a stochastic exploration of partial, yet statistically significant correspondences between subsets of n amino acids between the two proteins. The stochastic search of putative alignments of fixed length n is performed by partitioning the n amino acids in blocks of at least 10 amino acids for each protein. The block assignment is done independently for each protein; as a result the number of blocks and their lengths are generally different for the two proteins. The amino acids taking part in the blocks are numbered sequentially from the N to the C terminus. The alignment is defined as the pairwise correspondence between the marked amino acids with the same index. This pairing scheme implies that the explored alignments must follow the sequential order of the amino acids. This condition, common to other alignment methods (Holm & Park, 2000), rules out the possibility to establish correspondences between groups of amino acids that have different block order in the two proteins.

Given a tentative alignment, i.e. an initial block assignment in the two proteins, a new tentative alignment, or *trial alignment*, is generated by modifying its block sequence by merging, splitting or shifting the blocks. Each trial alignment is accepted/rejected with the standard Metropolis criterion, within a replica-exchange scheme (Tesi *et al.*, 1996), to promote the maximum score q_n. In particular, we have six different replicas $R_i(t)$, $i = 1, ..., 6$, (which correspond to different correspondences of amino acids between the two proteins) for each time-step t, with an associated score $q(R_i(t))$ calculated using to equation 4.4. At every timestep t each replica $R_i(t)$ is used to generate a trial alignment $\tilde{R}_i$, which has an associated score $q(\tilde{R}_i)$. The i-th replica at time $t+1$, $R_i(t+1)$, is posed equal to the trial alignment $\tilde{R}_i$ with a probability:

$$\mathcal{P}_{R_i(t+1)=\tilde{R}_i} = \min\left\{1, \exp\left(\frac{q(\tilde{R}_i) - q(R_i(t))}{T_i}\right)\right\}$$

where T_i is a fictitious temperature associated to the i-th replica, otherwise $R_i(t+1) = R_i(t)$ (Metropolis criterion). Observe that, if the fictitious temperature is zero, only trial alignments which increase the score are accepted. In our implementation each replica has a different temperature, and after every interval of five time-steps, swaps between the different replicas are tried. The swap between replicas i and j is accepted/rejected depending on the scores of the replicas and their temperatures, according to the following probability to accept the swap:

$$\mathcal{P}_{R_i \leftrightarrow R_j} = \min\left\{1, \exp\left[\left(\frac{q(R_i)}{T_j} + \frac{q(R_j)}{T_i}\right) - \left(\frac{q(R_i)}{T_i} + \frac{q(R_j)}{T_j}\right)\right]\right\}$$

(replica exchange criterium). Within this scheme, and having appropriately fixed the temperatures, we observe that we obtain an alignment that maximize the alignment score q_n after a few thousand time-steps.

4.2.4 Comparison of alignments of different lengths

Observe that the quantity q_n, defined in 4.4, can be used to compare two different alignments, only provided that they have the same number of amino acids n into correspondence. Indeed the larger the length n of the alignment the lower (in average) the quantity q_n.

In order to quantify how much q_n is affected by the length of the alignment, dynamics-based alignments of different lengths were carried out among the 56 representatives with minimal structural relatedness (highlighted in Table 4.1). The stochastic search of optimal alignments was performed by maximising q_n separately for each considered length n =75, 100, 125,... For each explored value of n we obtained the statistics of the optimal q_n over all pairwise alignments and computed the first two moments of the distribution, $\langle q_n \rangle$ and $\delta q_n^2 \equiv \langle q_n^2 \rangle - \langle q_n \rangle^2$.

The values of $\langle q_n \rangle$ and δq_n are represented in Fig. 4.2 by the filled circles and the error bars, respectively. The trend of $\langle q_n \rangle$ is well captured by the single exponential fit, $f(n) = 0.5115 * \exp(-n/336)$ indicated with the continuous red line. This exponential trend was subtracted from the "raw" quantity q_n, allowing a homogeneous comparison of alignment scores of different length. Accordingly, the score of a given alignment of n amino acids was computed as:

$$s_n = q_n - f(n) \quad . \tag{4.6}$$

Notice that the dispersion of q_n (denoted by the errorbars) is visibly constant for all explored values of n (except at the largest lengths owing to poorer statistics). Consequently, it is not necessary to include in eq. (4.6) a correction for "regularizing" also the n-dependent breadth of the score distribution.

Given a protein pair, its optimal alignment is the alignment of length n associated to the maximum score s_n.

4.2.5 Statistical significance of an alignment

Given a protein pair, the dynamics-based alignment provides an optimal alignment which involves n amino acids and has an associated alignment score s_n. When is the alignment score s_n large enough to consider the alignment significant? The usual way to tackle this question is to perform a significance analysis by comparing the score

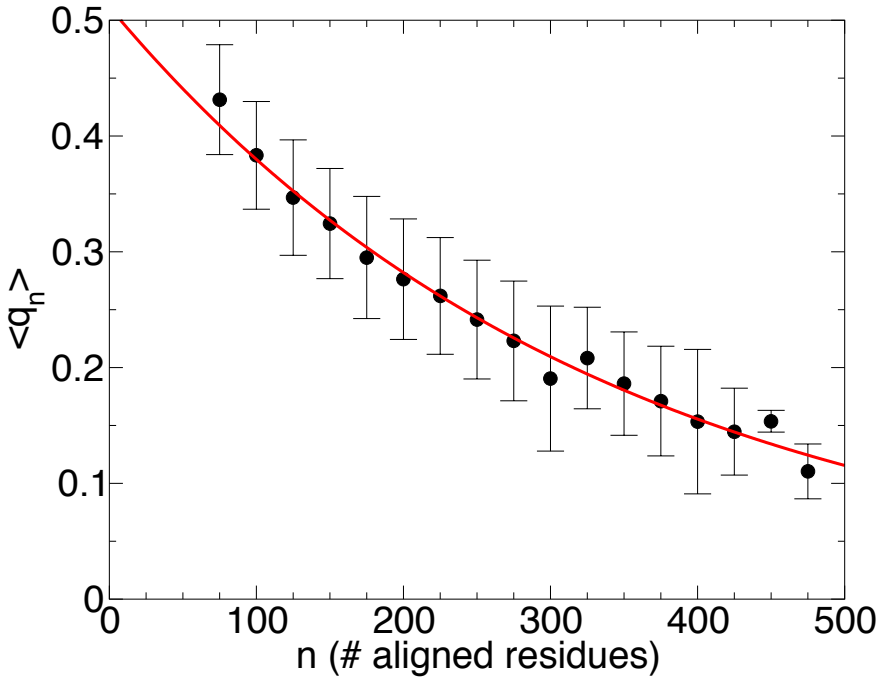

Figure 4.2: Average and dispersion of q_n from pairwise alignments within the subset of enzymes with minimal structural redundancy. The red line provides the exponential best fit to the data points and corresponds to the function $f(n) = 0.5115\ \exp(-n/336)$.

s_n with a reference distribution of scores recorded over a set of proteins which are not expected *a priori* to lead to a sizeable number of meaningful alignments. This reference set was assembled by selecting one representative protein, the longest, for each of the 56 different topologies in the data set of enzymes reported in Table 4.1. The resulting distribution of the 1540 alignment scores was compared against standard statistical distributions arising in alignment contexts (Levitt & Gerstein, 1998; Taylor, 2006) including the extreme value (Gumbel) and the Gaussian distributions, see Fig. 4.3. Assuming a Poissonian uncertainty of the height of the histogram the χ^2 associated to the Gumbel distribution is 3.7, while that of the Gaussian distribution is 1.1. As visible in Fig. 4.3, the Gaussian distribution appears to provide a good fit to the data set within three standard deviations to the left and right of the mean value.

The latter distribution was consequently taken *a posteriori* as providing the so called "null distribution" of the alignment score s_n, i.e. the reference distribution for alignment scores of unrelated protein pairs. In this way, to each alignment it is associated a z-score or, equivalently, a p-value. The former is a measure of how

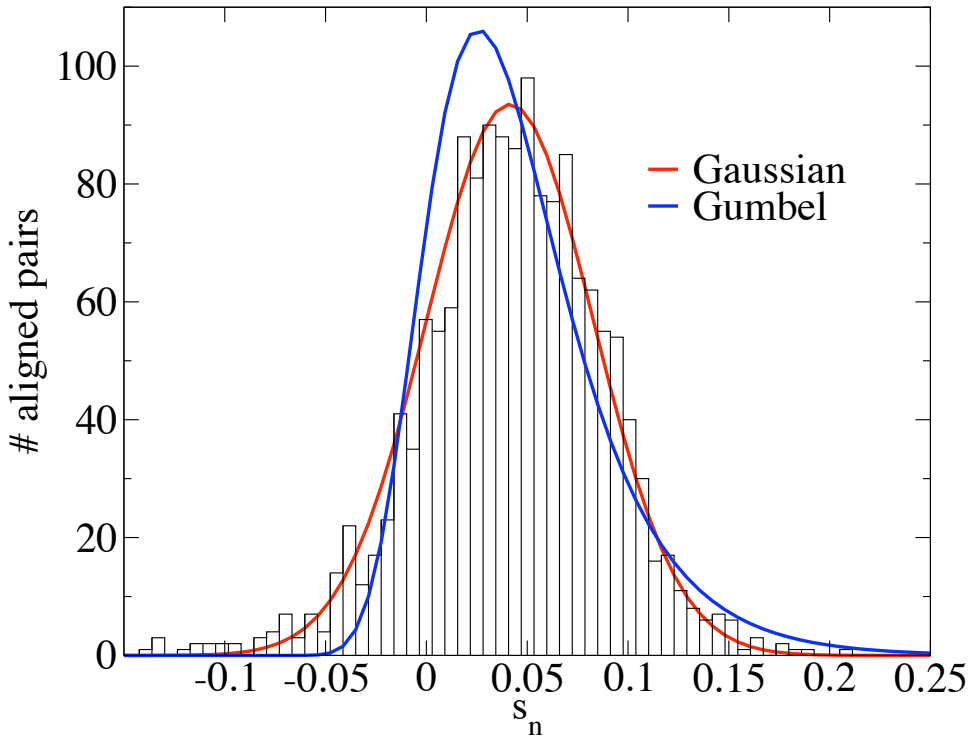

Figure 4.3: Histogram of the alignment scores, s_n, collected over the 1540 distinct pairings of 56 protein representatives *a priori* unrelated. The red and blue curves represent respectively the best fits using the Gaussian and Gumbel distributions. Parameters (mean and spread) of the best-fitting Gaussian: $\mu = 0.0414$ $\sigma = 0.0415$.

distant (in terms of standard deviations) is the obtained score from the average random reference case. The p-value, instead, corresponds to the probability that an alignment of n amino acids of two unrelated proteins returns a score higher than the one actually observed. The lower is the p-value (i.e. the higher the z-score), the more atypical, and hence significant, is the alignment.

4.2.6 Graphical representation of corresponding modes

The score q_n of eqn. (4.4), and consequently also the alignment score s_n in eqn. (4.6), are invariant upon replacing the orthonormal set of the $\vec{v}$'s (or $\vec{w}$'s) with another one obtained by their suitable linear combination. This property is used to convey in a graphically optimized way the consistency of two sets of low-energy modes. The first optimized basis vector in each set, $\vec{v}'_1 = \sum_{j=1...10} a_j \vec{v}_j$ and $\vec{w}'_1 = \sum_{j=1...10} b_j \vec{w}_j$ is

found by optimizing the linear weights, a's and b's so that the scalar product $\vec{v}_1' \cdot \vec{w}_1'$ is maximum (the unit norm of $\vec{v}_1'$ and $\vec{w}_1'$ is implied). The procedure is iterated to define the remaining vectors of the new basis which must be orthogonal to those already identified. This is exactly the procedure that we have described in detail in appendix A in the context of comparison of essential dynamical spaces.

4.2.7 A test case: dynamics-based alignment of HIV-1 protease and BACE

The dynamics-based alignment is applied in the following section of this chapter, to cases of single-chain and single-domain proteins. However we want to stress that the applicability of the dynamics-based alignment scheme goes beyond these cases. Comparisons between multimeric proteins (i.e. constituted by two or more chains) are possible, considering all the possible orderings of the chains in the protein. For a given chain ordering, the amino acids of the entire multimer are numbered consecutively and the simple pairing procedure is applied. The optimal alignment is provided by the ordering of the chains associated to the maximum alignment score.

An example of an alignment that involves a multimeric protein is provided by the comparison between HIV-1 protease (PDB code: 1nh0) and human β-secretase (BACE, PDB code: 1er8), respectively a viral and an eukaryotic Asp proteases. The former is a homo-dimer of 198 amino acids, each subunit being composed of 99 amino acids (Baca & Kent, 1993; Fitzgerald & Springer, 1991; Hong *et al.*, 2000), while the latter is a monomer of 330 amino acids. They differ for size and structure, however it is known that they are evolutionary related (Blundell & Srinivasan, 1996; Carnevale *et al.*, 2006; Cascella *et al.*, 2005; Neri *et al.*, 2005). The result of their dynamics-based alignment is shown in Fig. 4.4a and comprises 150 amino acids with a total RMSD of 5.5 Å and RMSIP of 0.73. Note that the alignment induces the superposition of the ASP dyad (the amino acids involved in the chemical catalysis) for the two proteases, as shown in Fig. 4.4b. In this figure we also notice that the dynamical alignment highlights the correspondence between the movements sustained by the flexible flaps opposite to the ASP dyas, which delimit the active site.

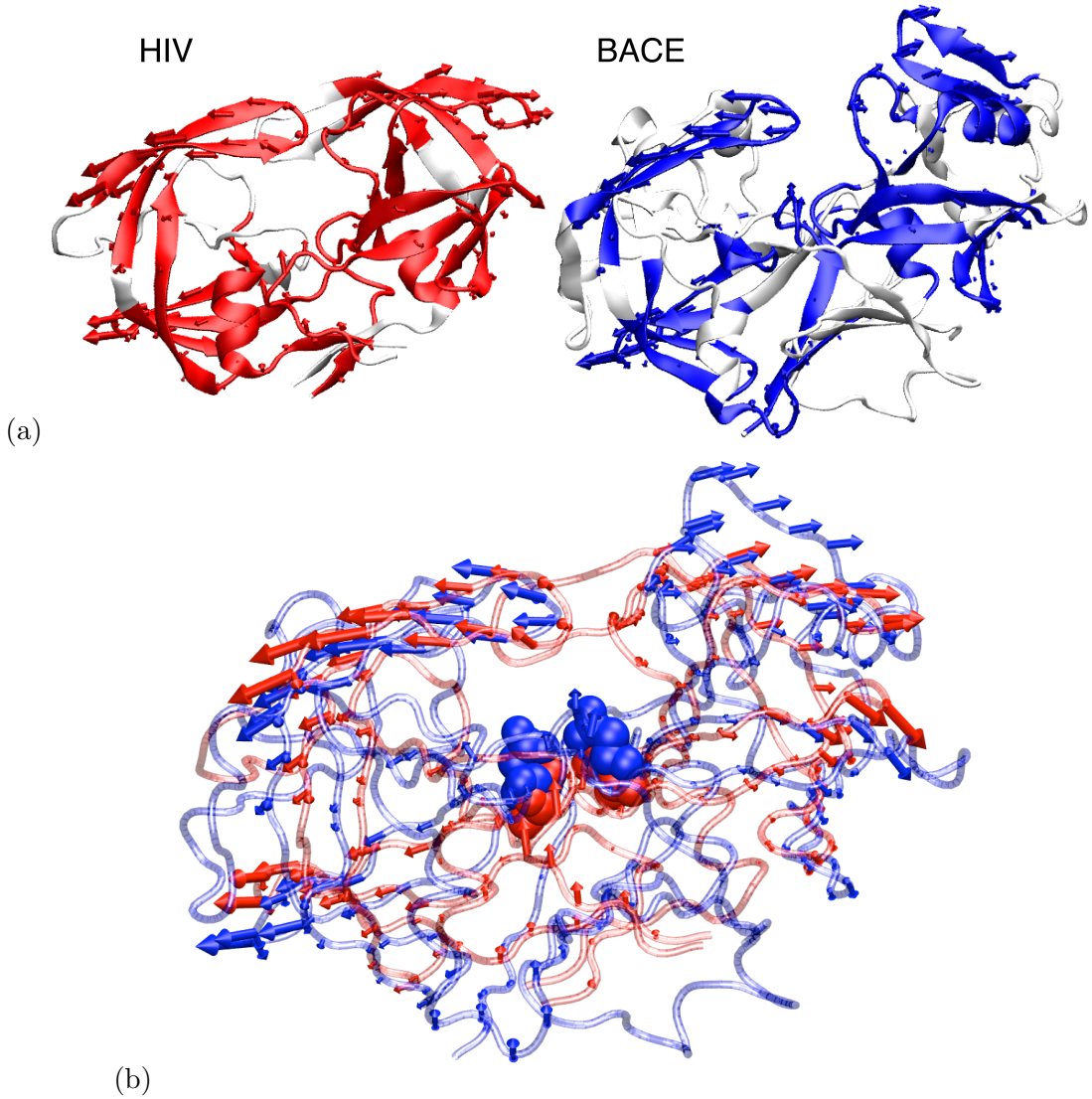

Figure 4.4: Dynamics-based alignment of the dimeric HIV-1 PR (PDB code: 1nh0) and the monomeric BACE (PDB code: 1er8) over $n = 150$ amino acids. In panel (a) the amino acids marked for the alignment have been colored in red and in blue for HIV and BACE respectively. The sets of arrows in each protein represent the two best corresponding lowest-energy modes for the aligned regions, as desdribed in section 4.2.6. In panel (b) the structures of HIV and BACE, represented respectively as red and blue transparent tubes, have been optimally superimposed according to the dynamics-based alignment. The two consensus lowest-energy modes are represented for each protein. The ASP dyan, corresponding to resides 32 and 215 of BACE and 25 (for both the chains) for HIV, has been highlighted in Van der Waals representation.

4.3 Dynamics-based comparison of enzymatic functional families

In Carnevale *et al.* (2006) and in Capozzi *et al.* (2007) it has been shown that different members of some specific enzymatic superfamilies, respectively proteases and calcium binding proteins, have similar large-scale movements in thermodynamic equilibrium, despite they do not share a striking sequence or structure similarity. In Carnevale *et al.* (2006) the similarities in the dynamics of the proteases where detected on the base of partial structural similarities, while in Capozzi *et al.* (2007) the calcium binding proteins where compared within a framework *ad hoc* developed for them.

The availability of the dynamics-based alignment allows a more general approach for the comparison of the dynamical properties of enzymatic families. As a first application of this method we have therefore selected and compared a set of enzymes, with a minimal structural similarity, which represent the main enzymatic functional families. The aim of this investigation is to study the relationship between structure, dynamics and function of proteins. Paralleling the studies that have clarified the relationship between sequence and structure, we want to use the dynamics-based alignment to highlight cases where enzymes show similarity in their dynamical properties also without an underlying striking structural similarity. Hereafter we shall describe the results of this investigation.

4.3.1 Dataset selection

The enzymes considered here were selected exploiting the hierarchical classification provided by the Enzyme Commission[1] (EC) database (Porter *et al.*, 2004). The EC functional annotation provides a transparent, though qualitative, criterion for defining an enzymatic functional distance which was used in the analysis to investigate the existence of correlations between functional and dynamics-based pairwise similarities. The reference data set was constructed by uniformly covering each of the 6 EC classes, whose enzyme-catalyzed reactions are:

EC 1 *Oxidoreductases*: catalyze oxidation/reduction reactions (i.e. transfer of H and O atoms or electrons from one substance to another);

EC 2 *Transferases*: transfer of a functional group from one substance to another (the group may be methyl-, acyl-, amino- or phosphate group);

EC 3 *Hydrolases*: formation of two products from a substrate by hydrolysis;

[1] http:/www.ebi.ac.uk/thornton-srv/databases/CSA/

EC 4 *Lyases*: non-hydrolytic addition or removal of groups from substrates (C-C, C-N, C-O or C-S bonds may be cleaved);

EC 5 *Isomerases*: intramolecule rearrangement (i.e. isomerization changes within a single molecule);

EC 6 *Ligases*: join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP.

The entire EC database was filtered to remove overall structural redundancies within each class, selecting one representative, by default the longest enzyme, of each group of molecules sharing the same structural class, architecture and topology, as defined by the CATH classification of protein structural patterns (Orengo *et al.*, 1997). Only PDBs with complete structural information and constituted by a single chain and a single domain were treated, because this is a simple criterium to select enzymes expected to work as a single monomer (note however that it is not excluded that some of them could work as part of a multimeric biological unit). The resulting set consisted of 76 enzymes, reported in Table 4.1, with the following functional distribution: oxidoreductases (8), transferases (12), hydrolases (36), lyases (12), isomerases (7), ligases (1). As the removal of structural redundancy was carried out for each EC class separately, representatives of different functional families can have the same topology, according to the CATH classification. This degeneracy, which affects only 47 of the 2850 distinct pairings of the 76 representatives, was retained as its removal would have led to an uneven representation of the distinct EC families. The total structural variability contains 56 different topologies, representing 3 CATH structural classes and 15 architectures.

As can be observed in Table 4.1, the length (in terms of number of amino acids) of the selected enzymes is heterogeneous, and the average is 245 $\pm$ 118 amino acids per protein. Pairwise sequence alignments, performed using ClustalW (Chenna *et al.*, 2003), among members of the set yield 12.2 $\pm$ 2.3 % sequence identity on average. This value indicates the absence of a strict sequential correspondence among the selected enzymes.

4.3.2 Results of the dynamics-based alignments

The resulting scores of the dynamics-based alignment for each enzyme pair are graphically represented in Fig. 4.5 in which the two matrices of panels (a) and (b) differ only in the way the entries are ordered. In (a) rows and columns appear in order of EC code; in (b), in order of CATH code. These two alternative groupings allow an intuitive perception of how functional and structural analogies are reflected by the alignment score.

PDB	EC	CATH	length	PDB	EC	CATH	length
1pb3	1.1.1.42	3.40.718.10	416	1ivb	3.2.1.18	2.120.10.10	390
2et7	1.2.3.4	2.60.120.10	201	2ayh	3.2.1.73	2.60.120.200	214
1d7o	1.3.1.9	3.40.50.720	297	1dy4	3.2.1.91	2.70.100.10	434
3cd2	1.5.1.3	3.40.430.10	206	4skn	3.2.2.3	3.40.470.10	223
1k03	1.6.99.1	3.20.20.70	399	1p7m	3.2.2.20	1.10.340.30	187
1xm0	1.8.4.6	2.170.150.20	147	8cpa	3.4.17.1	3.40.630.10	307
6pah	1.14.16.1	1.10.800.10	308	3pbh	3.4.22.1	3.90.70.10	317
1dfx	1.15.1.1	2.60.40.730	125	1avp	3.4.22.39	3.40.395.10	199
4tms	2.1.1.45	3.30.572.10	316	1qjj	3.4.24.21	3.40.390.10	200
1cia	2.3.1.28	3.30.559.10	213	1f82	3.4.24.69	3.90.1240.10	424
1h17	2.3.1.54	3.20.70.20	754	1lba	3.5.1.28	3.40.80.10	146
1cjw	2.3.1.87	3.40.630.30	166	1lqy	3.5.1.88	3.90.45.10	184
1fp9	2.4.1.25	3.20.20.80	500	1ko3	3.5.2.6	3.60.15.10	230
1qcd	2.4.2.7	3.40.50.2020	236	1rgy	3.5.2.6	3.40.710.10	360
1dtp	2.4.2.36	3.90.175.10	190	1mjz	3.6.1.1	3.90.80.10	175
1fps	2.5.1.10	1.10.600.10	348	2acy	3.6.1.7	3.30.70.100	98
1ajz	2.5.1.15	3.20.20.20	282	1l6t	3.6.3.14	1.20.20.10	79
1ax3	2.7.1.69	2.70.70.10	162	2had	3.8.1.5	3.40.50.1820	310
1ohb	2.7.2.8	3.40.1160.10	258	1ojr	4.1.2.19	3.40.225.10	274
2f65	2.7.6.3	3.30.70.560	158	2dhn	4.1.2.25	3.30.1130.10	121
4p2p	3.1.1.4	1.20.90.10	124	1yb7	4.1.2.39	3.40.50.1820	256
1u32	3.1.3.16	3.60.21.10	293	1v3w	4.2.1.1	2.160.10.10	173
2f6f	3.1.3.48	3.90.190.10	302	1v9i	4.2.1.1	3.10.200.10	261
2ffz	3.1.4.3	1.10.575.10	245	1gqn	4.2.1.10	3.20.20.70	252
1ako	3.1.11.2	3.60.10.10	268	1vbl	4.2.2.2	2.160.20.10	416
1dmu	3.1.21.4	3.40.600.20	299	1hv6	4.2.2.3	1.50.10.110	351
1vas	3.1.25.1	1.10.440.10	137	2g64	4.2.3.12	3.30.479.10	140
1goc	3.1.26.4	3.30.420.10	156	1cqh	4.2.99.18	3.40.30.10	105
2fmb	3.1.26.4	2.40.70.10	104	1fx2	4.6.1.1	3.30.70.1230	235
1bol	3.1.27.1	3.90.730.10	222	2plc	4.6.1.13	3.20.20.190	274
1k2a	3.1.27.5	3.10.130.10	136	1rtv	5.1.3.13	2.60.120.10	184
1de3	3.1.27.10	3.10.450.30	150	1h0p	5.2.1.8	2.40.100.10	182
1kab	3.1.31.1	2.40.50.90	136	1pbk	5.2.1.8	3.10.50.40	116
1b1y	3.2.1.2	3.20.20.80	500	1nsj	5.3.1.24	3.20.20.70	205
2fba	3.2.1.3	1.50.10.10	492	8cho	5.3.3.1	3.10.450.50	125
3eng	3.2.1.4	2.40.40.10	213	1mek	5.3.4.1	3.40.30.10	120
1bhe	3.2.1.15	2.160.20.10	376	1id8	5.4.99.1	3.40.50.280	137
2f47	3.2.1.17	1.10.530.40	175	1dbs	6.3.3.3	3.40.50.300	224

Table 4.1: **Enzymes dataset.** List of the 76 representatives used in this study identified by their PDB codes (Bernstein *et al.*, 1977b), EC (Porter *et al.*, 2004) and CATH (Orengo *et al.*, 1997) classification and number of amino acids. The first number of the EC field denotes the different functional classes: 1: oxidoreductases, 2: transferases, 3: hydrolases, 4: lyases, 5: isomerases, 6: ligases. For proteins 1h17 and 1avp five highly-exposed terminal amino acids, at the N- and C-terminus respectively, were omitted. Underlined PDB codes indicate the subset of enzymes with different topology (the longest enzyme was taken for each group with same topology). This subset was considered to be minimally structurally redundant.

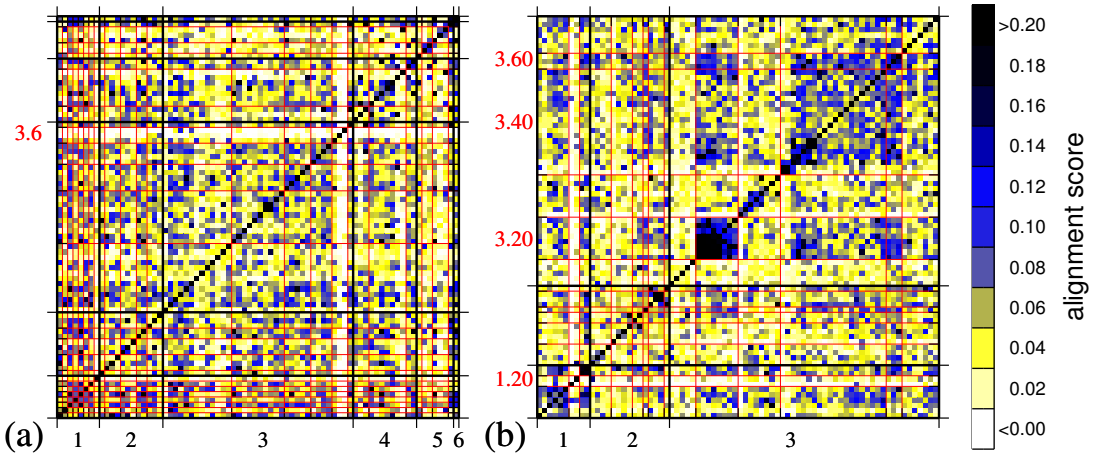

Figure 4.5: **Matrices reporting the dynamics-based score for all aligned enzyme pairs.** Good [poor] alignment scores are shown with dark blue [white] color. In (a) enzymes are ordered in each axis according to EC codes, black and red lines delimit enzymes with the same first and first two EC number, respectively. In (b) enzymes are ordered according to CATH codes. Black [red] lines separate different classes [architectures].

The qualitative appearance of the two plots is markedly different. The minimally-redundant coverage of the different EC families produces a fairly uniform scatter of good scores across various EC groups (Fig. 4.5a). It is nevertheless interesting to notice the presence of light bands corresponding to EC groups that are poorly alignable in general. The most notable of such groups comprises hydrolases acting on acid anhydrides (principal EC codes: 3.6). By contrast, the uneven representation of different structural classes, architectures and topologies in the data sets leads to a manifest inhomogeneous character of the matrix ordered by structure of Fig. 4.5b. In particular, the class with the largest proportion of good scores is the α-β one (class 3), which is also the most populated class in the set. Not all its architectural subgroups, however, display the same degree of "alignability". Both in absolute and relative terms, the most prominent architecture is the α-β barrel (principal CATH numbers: 3.20). It is also worth noting that good alignment scores are attained for several *inter*architecture alignments; a few of such cases will be discussed later. Finally a notable case of overall poor alignability is the set of the mainly-α/Up-Down Bundle (principal CATH numbers: 1.20), which shows correspondences only with enzymes belonging to the same structural group.

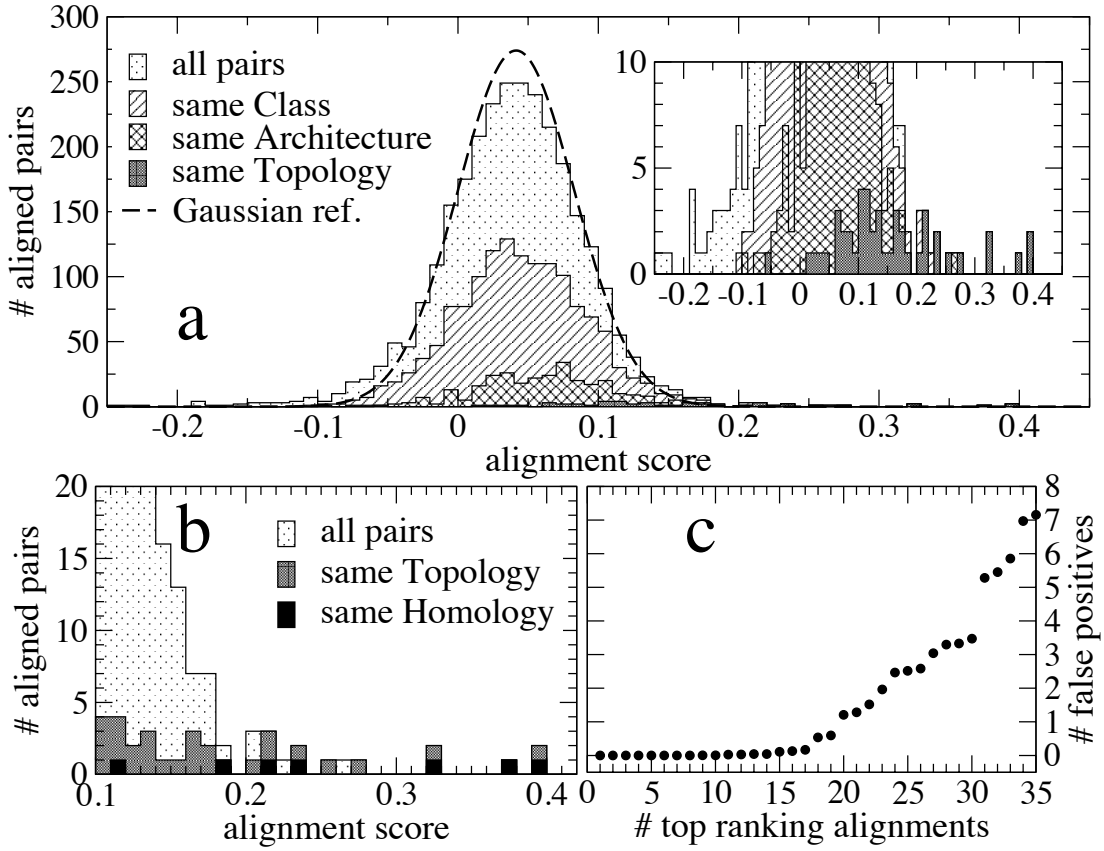

Figure 4.6: **Distribution of alignment scores.** (a) Distribution of the alignment score calculated over all 2850 enzyme pairs (the inset presents an enlargement of the histogram highlighting pairs with the same topology). The contribution of pairs with the same structural class, architecture and topology are also shown. The dashed line represents the "null" distribution (i.e. a Gaussian with mean $\mu = 0.041413$, and spread $\sigma = 0.041493$). (b) Tail of the distribution associated with the highest alignment scores. Pairs that have same topology and homologous superfamily are highlighted. (c) Number of non statistically significant (false positive) alignments expected to arise within the top-ranking alignments.

The extent to which the various degrees of structural relatedness impact on the dynamical correspondences is summarized Fig. 4.6a. The histogram portrays the distribution of optimized scores for all enzyme pairs and also pairs having the same class, architecture and topology. It is noted (see inset) that the very few pairings (47 entries) of enzymes with the same topology tend to have alignment scores distinctively better than typical enzyme pairs. On the other hand no such pronounced deviation from the average behaviour is observed for pairs with the same structural class or even architecture (that is, the two highest levels of the hierarchical structural classification in CATH).

Similarities in dynamics between structurally-related enzymes is expected. We therefore wish to focus particularly on the alignments that are highest ranking according to the dynamics-based score. The distribution of their scores is shown in Fig. 4.6b. In this figure, alignments among enzymes with the same topology (the first three CATH numbers) and same topology plus homology (the entire CATH code) have been highlighted. Among the top ∼20 alignments are 6 pairs sharing the full CATH code (the total number of such homologous pairs in the set is 8). This confirms the intuitive expectation that significant sequence and structural similarities result likely in pronounced dynamical similarities (Keskin *et al.*, 2000).

However, it is important to note that in Fig. 4.6b, besides these expected good correspondences, a fraction of the alignments approaching the tail pertain to pairs that differ at the level of class or architecture. These cases are of particular interest as they would not be singled out by criteria based solely on the CATH structural classification. A selection of these alignments, as well as other structurally-induced ones, will be discussed in the following.

4.3.3 Statistically relevant alignments

Our considerations will now concentrate on alignments that are statistically significant. As described in section 4.2.5, the statistical significance of an alignment is obtained by the comparison of its score with a null distribution. In agreement with the previous considerations, the null distribution in this case is provided by the alignments between the representative set of enzymes with different topology, highlighted in Table 4.1. Their distribution is optimally fitted by a Gaussian distribution (see Fig. 4.3). This null Gaussian distribution, renormalized for the dataset size, has been reported in Fig. 4.6a.

From the p-value analysis we estimated the number of non-significant entries (false positives) expected among the top alignments (Levitt & Gerstein, 1998; Storey & Tib-

shirani, 2003). The associated curve, shown in Fig. 4.6c, indicates that within the top 26 alignments, fewer than 10% are expected to be false positives. This threshold provides an acceptable balance between the number of entries declared significant (26) and the fraction which is deemed reliable (>90%). All further considerations will therefore be limited to the pairings in the top 26 alignments, which are reported in Table 4.2.

Within this set, the number of pairings that can be ascribed to overall similarities of the global fold topology is 16, including 6 homologous cases. A more refined and quantitative study of the level of subtler structural correspondence in the set was carried out with DALI (Holm & Park, 2000; Holm & Sander, 1996), a powerful structural alignment tool that detects *partial* similarities based on the similarity between two proteins of inter-residue distance matrices. For a consistent comparison with our results, the statistical confidence threshold on the DALI results (Sierk & Pearson, 2004) was also set to 90% (leading to 18 significant DALI pairings). It was found that 14 of our top 26 alignments had significant DALI scores. These included 12 pairs with same topology (including all the 6 homologous pairs). Of the ten pairings with different topology selected by our method only two turned out to have significant partial alignments according to DALI. These alignments were between proteins 2dhn-2g64 and 1dy4-2ayh. Importantly, within the 18 statistically-significant DALI pairings these two alignments were the only ones involving different CATH topology. Consequently, the remaining 8 of the 26 (i.e. $\sim 30\%$) dynamics-based alignments deemed significant involved pairings between enzymes whose structural relatedness is not easily detectable at the same level of statistical significance.

4.3.4 Discussion of alignment results

A selected number of significant alignments, exemplifying the sophisticated interplay of structural and dynamical features, are shown: *hydroxynitrile lyase-haloalkane dehalogenase* (200 aligned amino acids) in Fig. 4.7; *human thioredoxin-disulfide isomerase* (75 aligned amino acids) in Fig. 4.8; *dethiobiotin synthetase-phosphoribosyl anthranilate isomerase* (100 aligned amino acids) in Fig. 4.9; *exonuclease III-enoyl reductase* (175 aligned amino acids) in Fig. 4.10; *cellobiohydrolase I-endo-1,3-1,4-β-D-glucan 4-glucanohydrolase* (75 aligned amino acids) in Fig. 4.11 and *exonuclease III-human adenovirus proteinase* (75 aligned amino acids) in Fig. 4.12. The first two pairs are examples of alignments between enzymes with different functions (first EC number) but similar fold (*i.e.,* same CATH code) while the opposite is true for examples Fig. 4.11 and Fig. 4.12. Cases Fig. 4.9 and Fig. 4.10 are, instead, examples of alignments between enzymes that differ in both function and fold.

Rank	PDB1	EC	CATH	length	PDB2	EC	CATH	length	n	RMSIP	RMSD (Å)
1	1ajz	2	3.20.20.20	282	1gqn	4	3.20.20.70	252	150	0.8525	5.103
2	1cqh	4	3.40.30.10	105	1mek	5	3.40.30.10	120	75	0.8735	2.773
3	1yb7	4	3.40.50.1820	256	2had	3	3.40.50.1820	310	200	0.8090	4.552
4	1gqn	4	3.20.20.70	252	1nsj	5	3.20.20.70	205	150	0.8051	4.303
5	1ajz	2	3.20.20.20	282	1nsj	5	3.20.20.70	205	150	0.7864	4.067
6	1b1y	3	3.20.20.80	500	1gqn	4	3.20.20.70	252	175	0.7953	5.061
7	2dhn	4	3.30.1130.10	121	2g64	4	3.30.479.10	140	75	0.8261	3.014
8	1b1y	3	3.20.20.80	500	1nsj	5	3.20.20.70	205	125	0.7534	5.988
9	1k03	1	3.20.20.70	399	1nsj	5	3.20.20.70	205	100	0.7777	4.178
10	1id8	5	3.40.50.280	137	1yb7	4	3.40.50.1820	256	75	0.7964	3.606
11	1dbs	6	3.40.50.300	224	1nsj	5	3.20.20.70	205	100	0.7846	5.704
12	1bhe	3	2.160.20.10	376	1vbl	4	2.160.20.10	416	200	0.7167	10.13
13	1ajz	2	3.20.20.20	282	1k03	1	3.20.20.70	399	150	0.7319	7.226
14	1ajz	2	3.20.20.20	282	1b1y	3	3.20.20.80	500	125	0.7785	6.073
15	1gqn	4	3.20.20.70	252	2plc	4	3.20.20.190	274	125	0.7646	6.771
16	1dy4	3	2.70.100.10	434	2ayh	3	2.60.120.200	214	75	0.7493	5.251
17	1ako	3	3.60.10.10	268	1d7o	1	3.40.50.720	297	175	0.7006	8.442
18	1gqn	4	3.20.20.70	252	1k03	1	3.20.20.70	399	200	0.7196	6.302
19	1v3w	4	2.160.10.10	173	1xm0	1	2.170.150.20	147	75	0.6811	11.46
20	1v3w	4	2.160.10.10	173	2dhn	4	3.30.1130.10	121	75	0.6909	11.90
21	1ajz	2	3.20.20.20	282	1dbs	6	3.40.50.300	224	100	0.7869	7.271
22	1id8	5	3.40.50.280	137	2had	3	3.40.50.1820	310	100	0.6823	5.698
23	2f47	3	1.10.530.40	175	4tms	2	3.30.572.10	316	100	0.7159	10.25
24	1gqn	4	3.20.20.70	252	1h17	2	3.20.70.20	754	125	0.7424	9.110
25	1ajz	2	3.20.20.20	282	2plc	4	3.20.20.190	274	100	0.7256	6.446
26	1ako	3	3.60.10.10	268	1avp	3	3.40.395.10	199	75	0.7624	5.990

Table 4.2: **List of the top 26 dynamics-based alignments.** The first column is the rank of the alignment. Columns 2–10 report the PDB code, principal EC number, CATH code and length of the aligned proteins. The last three columns provide details of their optimal alignment, namely: the number of aligned amino acids, n, the RMSIP of the top 10 modes and the structural RMSD.

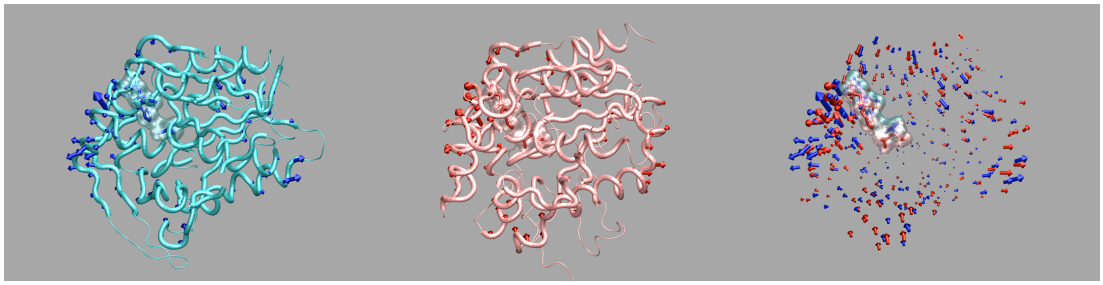

Figure 4.7: Dynamics-based alignment between *hydroxynitrile lyase* (1yb7) and *haloalkane dehalogenase* (2had). The number of aligned amino acids is 200. The rank of this alignment is 3, as reported in Table 4.2. Structural-dynamical properties of the selected alignment are graphically summarized by rendering, on the left and in the middle, in blue the first listed protein and in red the second. Aligned regions are represented as thick tubes, while non-aligned regions are represented as thin tubes. Arrows are used to indicate the directionality and magnitude of the distortions entailed by the most consistent dynamical space (section 4.2.6). The location of the catalytic residues are highlighted as Van der Waals surfaces. The rightmost panel presents the superposition of the aligned regions.

One of the enzyme pairs with the highest structural-dynamical correspondence involves *hydroxynitrile lyase* (PDB: 1yb7, length 256, EC: 4.1.2.39, CATH: 3.40.50.1820) and *haloalkane dehalogenase* (PDB: 2had, length 310, EC: 3.8.1.5, CATH: 3.40.50.1820). These differ in EC class but have the same first four CATH codes. Their best alignment, which spans 200 amino acids, covers a substantial fraction of both enzymes. Fig. 4.7 summarizes the results graphically. For clarity, the aligned regions and associated low-energy modes are shown separately for the two enzymes. Given the impossibility of conveying graphically the dynamics covered by the 10 lowest-energy modes, we have reported only the maximally consistent subspace in the two sets of modes (see section 4.2.7). The RMSD over the 200 aligned amino acids is 4.5 Å which compares well with the purely-structural DALI alignment of the same proteins: RMSD = 3.0 Å over 226 amino acids. Indeed, unlike other cases discussed in the following, this optimal alignment is also very good from a purely-structural point of view. The quality of the overall consistency of the low-energy modes is also striking, as it possesses a RMSIP of 0.81, which exceeds the reference values that typically denote good consistency of molecular dynamics trajectories of the *same* protein (Amadei *et al.*, 1999).

Another high-ranking alignment for both the dynamics-based procedure and the

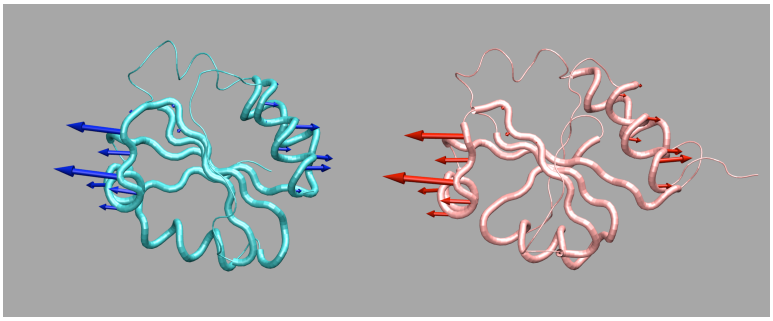

Figure 4.8: Dynamics-based alignment between *human thioredoxin* (1cqh) and *disulfide isomerase* (1mek). The number of aligned amino acids is 75. The rank of this alignment is 2, as reported in Table 4.2. Structural-dynamical properties of the selected alignment are graphically summarized by rendering, on the left and on the right, in blue the first listed protein and in red the second. Aligned regions are represented as thick tubes, while non-aligned regions are represented as thin tubes. Arrows are used to indicate the directionality and magnitude of the distortions entailed by the most consistent dynamical space (section 4.2.6).

purely-structural one is the pair of enzymes: *human thioredoxin* (PDB: 1cqh, length 105, EC: 4.2.99.18, CATH: 3.40.30.10) and *disulfide isomerase* (PDB: 1mek, length 120, EC: 5.3.4.1, CATH: 3.40.30.10) where as many as 75 amino acids correspond, with an RMSD as low as 2.8 Å and RMSIP again exceeding 0.87. Fig. 4.8 shows the high quality of the accord between structure and dynamics.

Over a third of the reliable alignments involve pairs that have dissimilar structural organization. Two notable examples appear in Fig. 4.9 and in Fig. 4.10; for the pairs: *dethiobiotin synthetase* (PDB: 1dbs, length 224, EC: 6.3.3.3, CATH: 3.40.50.300) and *phosphoribosyl anthranilate isomerase* (PDB: 1nsj, length 205, EC: 5.3.1.24, CATH: 3.20.20.70) in panel (c); and *exonuclease III* (PDB: 1ako, length 268, EC: 3.1.11.2, CATH: 3.60.10.10) and *enoyl reductase* (PDB: 1d7o, length 297, EC: 1.3.1.9, CATH: 3.40.50.720) in panel (d). Even though no strong global structural correspondences can be established between these pairs, there is a discernible consistency of the aligned regions. For the 100 aligned amino acids of the pair in Fig. 4.9 and 175 aligned amino acids in Fig. 4.10, the RMSD values are 5.7 Å and 8.4 Å respectively. The "structural tolerance" of this dynamics-based alignment is such that even elements with different secondary organization can be put in structural correspondence (*e.g.,* loops and helices). In these two cases also, low-energy modes are in very good agreement (RMSIP equal to

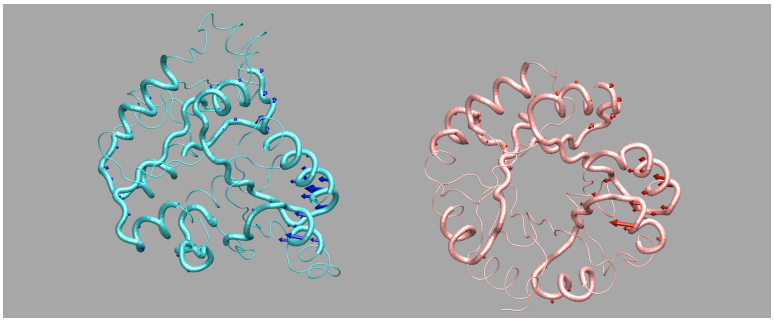

Figure 4.9: Dynamics-based alignment between *dethiobiotin synthetase* (1dbs) and *phosphoribosyl anthranilate isomerase* (1nsj). The number of aligned amino acids is 100. The rank of this alignment is 11, as reported in Table 4.2. The alignment is graphically represented as in in Fig. 4.8.

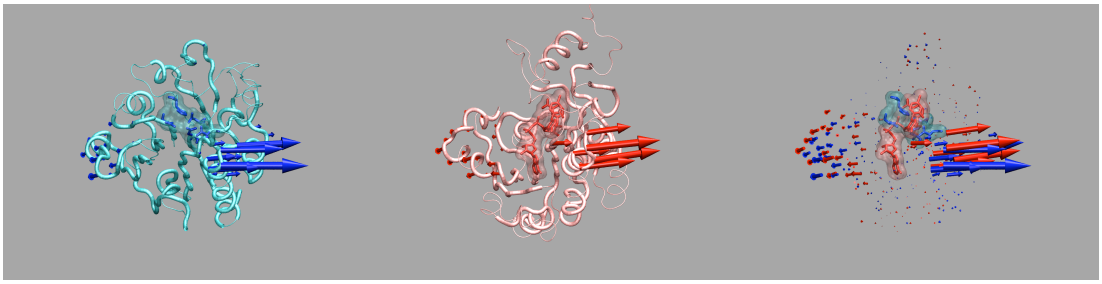

Figure 4.10: Dynamics-based alignment between *exonuclease III* (1ako) and *enoyl reductase* (1d7o). The number of aligned amino acids is 175. The rank of this alignment is 17, as reported in Table 4.2. The alignment is graphically represented as in in Fig. 4.7. The location of the catalytic residues of 1ako and of the bound ligands of 1d7o are highlighted as Van der Waals surfaces.

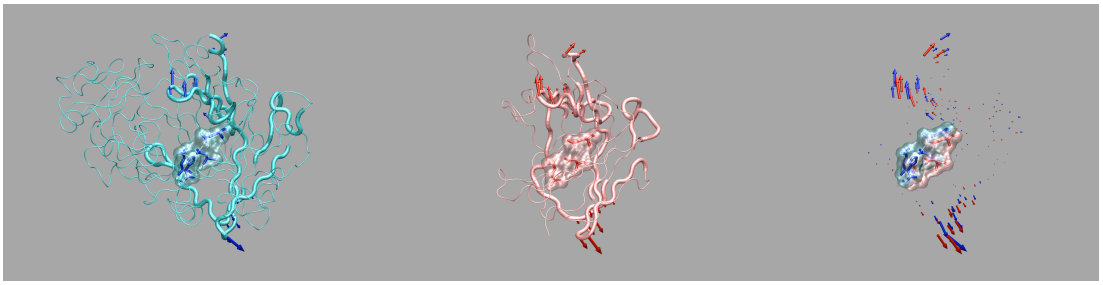

Figure 4.11: Dynamics-based alignment between *cellobiohydrolase I* (1dy4) and the *endo-1,3-1,4-β-D-glucan 4-glucanohydrolase* (2ayh). The number of aligned amino acids is 75. The rank of this alignment is 16, as reported in Table 4.2. The alignment is graphically represented as in in Fig. 4.7. The location of the catalytic residues are highlighted as Van der Waals surfaces.

0.78 and 0.70 respectively) and outline a consistent movement of fairly large compact regions in the enzyme pairs.

Another interesting observation concerns the spatial proximity of the catalytic sites induced by dynamics-based alignments. Bartlett *et al.* (2003) have shown that evolutionarily distantly related enzyme pairs that catalyze different reactions on similar structural scaffolds, retain the location of the active site and of functional structural elements, suggesting that evolution acts by changing roles and identities of amino acids at certain positions rather than recruiting new positions. Those observations raise the possibility that, besides local structural patterns, also the plasticity, *i.e.,* the conformational fluctuations, of the active sites have played a role in such conservation (Maguid *et al.*, 2006; Sacquin-Mora *et al.*, 2007). These observations prompted us to investigate whether any of the dynamics-based pairings induce correspondences of features related to catalysis or substrate binding.

Indeed, in our analysis we found that several high-ranking alignment bring active site amino acids into proximity. The rightmost panel in Fig. 4.7 shows the superposition of the 200 aligned amino acids *hydroxynitrile lyase* (PDB: 1yb7) and *haloalkane dehalogenase* (PDB: 2had), which, being evolutionarily related, are characterized by the same *four* CATH numbers: 3.40.50.1820, and belong to hydrolases (EC class 3) and lyases (EC class 4), respectively. Despite the different biological functions of the two enzymes, the positions of their catalytic residues are almost coincident. In particular HIS235, ASP207 and SER80 of *hydroxynitrile lyase* are equivalent to HIS289, ASP260 and ASP124 of *haloalkane dehalogenase* respectively.

The alignment between the *cellobiohydrolase I* (PDB: 1dy4) and the *endo-1,3-1,4-β-D-glucan 4-glucanohydrolase* (PDB: 2ayh), is also noteworthy as they differ at the CATH architecture level, though they share the same fold according to SCOP (Murzin *et al.*, 1995). The enzymes, which are both Glycosylases (EC code: 3.2.1) have analogous catalytic residues (Porter *et al.*, 2004): GLU212, HIS228, ASP214, GLU217 for the first enzyme; GLU105, ASP107, GLU109 for the second one. Despite the fact that only 20% of the larger enzyme is involved in the alignment, it is interesting to observe a remarkable space proximity of the two GLU-ASP-GLU triads (Fig. 4.11) which, in both cases, are located in an antiparallel $\beta-$sheet. Further aspects of this alignment deserve comment. For 1dy4 the active site is found in a cleft delimited by loops and which can accommodate the 1-(isopropylamino)-3-(1-naphthyloxy)-2-propanol ligand (see leftmost panel in Fig. 4.11). Also for 2ayh the active site is surrounded by loops, that form a groove which can arguably accomodate the corresponding ligand (see central panel of Fig. 4.11). The dynamics-based alignment has singled out a correspondence between the loops delimiting the binding clefts (amino acids 369 to 379 and 185 to 195 respectively for 1dy4 and 2ayh) and the directions of the matching low-energy modes are intuitively consistent with the opening/closing mechanism related to substrate binding in both enzymes (Divne *et al.*, 1998).

We now turn to specific enzyme pairing whose global/partial structural correspondences are not easily detectable, as indicated by the much higher, and more significant, dynamics-based ranking compared to the one found by purely-structural criteria (Holm & Park, 2000). Two of these pairings involve the *exonuclease III* (PDB: 1ako) which is aligned both with the *enoyl reductase* (PDB: 1d7o), and with the *human adenovirus proteinase* (PDB: 1avp). As in previous cases, the dynamics-based alignment induces a good superposition of the functionally-relevant regions of the *exonuclease III* and the *enoyl reductase*. As shown in Fig. 4.10, in fact, the active site of 1ako is well superimposed with the ligands bound by 1d7o and in both cases the corresponding low-energy modes develop an outward/inward concerted movement in the sorroundings of these regions. This relationship is plausible, given the chemical similarity of the ligands that these proteins bind (Mol *et al.*, 1995; Pidugu *et al.*, 2004; Roujeinikova *et al.*, 1999; Stockwell & Thornton, 2006).

A close relatedness of the nature of the ligands is also found for the pairing of *exonuclease III* and *human adenovirus proteinase* Fig. 4.12. Both enzymes, in fact, bind DNA (in double- and single-stranded forms, respectively). The possibility of establishing a dynamics-based connection between them is particularly interesting as they are not evolutionary related and are characterized by two different architectures, 4-Layer Sandwich (CATH: 3.60.10.10) for 1ako and 3-Layer($\alpha\beta\alpha$) Sandwich (CATH: 3.40.395.10),

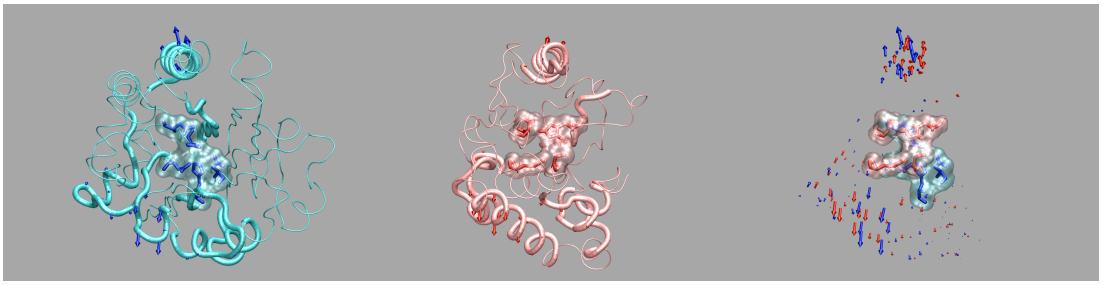

Figure 4.12: Dynamics-based alignment between *exonuclease III* (1ako) and *human adenovirus proteinase* (1avp). The number of aligned amino acids is 75. The rank of this alignment is 26, as reported in Table 4.2. The alignment is graphically represented as in in Fig. 4.7. The location of the catalytic residues are highlighted as Van der Waals surfaces.

for 1avp. Despite these features, the active site of the enzymes is well-superimposed after the dynamics-based alignment. Notably, the aligned region comprises a segments of amino acids which have been previously suggested to be involved in the binding of DNA (Gupta *et al.*, 2004; Mol *et al.*, 1995).

Finally, among the alignments involving two structurally-unrelated enzymes, we mention also the case of *dihydropteroate synthetase* (PDB: 1ajz) and *dethiobiotin synthetase* (PDB: 1dbs). Despite the differences in architecture, α-β barrel (CATH: 3.20.20.20) for 1ajz and 3-Layer ($\alpha\beta\alpha$) sandwich (CATH: 3.40.50.300) for 1dbs, and of the catalyzed reactions, the catalytic residues are found in good correspondence and one-to-one pairings can be established between the three catalytic residues of 1dbs and three of the four catalytic residues of 1aj7 (Porter *et al.*, 2004; Yang *et al.*, 1997). The C_α distances of such pairings range from 5.4 to 7.5 Å.

The specific cases discussed so far provide concrete illustrations of the biological implications of the dynamics-based alignment. They suggest particularly that functional correspondences in protein may be revealed on the basis of similarity in dynamics, thereby complementing available powerful strategies based on similarity at the level of sequence or at the level of structure. It is, in fact, well known that nonhomologous enzymes with similar mechanisms can share the spatial configuration of active site catalytic residues: On the basis of this observation it is possible to detect proteins with related functions on by identifying similar configurations of catalytic residues. The alignment scheme considered here is motivated by the fact that some enzymes undergo conformational changes as an integral part of their function. This observation has been applied in a spirit analogous to the structure-based inference mentioned above. In

particular it is considered that proteins with similar mechanisms might share not only similar configurations of catalytic residues, but also similarities in dynamics, and that these similarities might be detectable computationally. The specific cases discussed here suggest that proteins can show convergent evolution to shared dynamics related to function.

4.4 Conclusions

From the comparative analysis of large-scale movements in representatives of different functional categories of enzymes, ∼30 outstanding alignments are identified using established criteria for statistical significance. Detailed analysis of the results indicates that good dynamical similarities in enzyme pairs can arise even in the absence of strict correspondence of structure or sequence. Indeed, one third of the outstanding pairings involve enzymes with different structural organization at the global or partial fold level.

Strikingly, it is found that, even in the absence of easily-detectable structural correspondences, dynamics-based alignment can establish spatial relationships among regions involved in catalysis or substrate binding. In addition, the common dynamical features are oriented towards the structural rearrangements that arguably accompany the enzymatic functionality. This implies that a biological, function-related rationale underlies several of the outstanding alignments (though this is not necessarily true for *all* alignments, as large-scale movements are not expected to be involved in function for every enzyme).

These facts suggest that dynamics-based criteria can be profitably introduced in protein alignment contexts to expose functionally-related correspondences that would not be capturable, at the same level of significance, using purely sequence- or structure-based criteria. As a complement to these established techniques, further developments of dynamics-based approaches can contribute novel elements for exploring relationships between sequence, structure and function of enzymes.

In this respect the results reported in this chapter along with previous studies of dynamical-relatedness within specific enzymatic families (Capozzi *et al.*, 2007; Carnevale *et al.*, 2006) suggest that tools capable of exposing dynamics-based correspondences may provide a general quantitative and natural framework to group proteins according to their large-scale movements.

Furthermore, by cross-referencing results of purely structural and dynamics-based alignment it might be possible to address if, and to what extent, structural and

dynamically-related functional features have been subjected to different selective pressure. Two extreme scenarios may, in fact, be envisaged behind function-related dynamical correspondences between structurally-diverse enzymes. On one hand common large-scale dynamics may reflect features present in ancestral proteins/enzymes preserved during evolution, or they might reflect features selected by the necessity of well-defined movements for biological function (requiring only very general relationships between sequences and structures). Analogous questions have arisen about protein folds: it appears that both convergence and conservation have resulted in the limited number of available folds (Andreeva & Murzin, 2006; Chothia, 1992; Denton & Marshall, 2001b; Lupas *et al.*, 2001; Rose *et al.*, 2006).

It would therefore be most interesting to address these issues connected to the evolutionary convergence/conservation of functionally-oriented motions, for specific enzymatic families that have been the subject of thorough investigation from an evolutionary perspective (Lesk & Fordham, 1996; Scheeff & Bourne, 2005; Xu *et al.*, 1999).

4. DYNAMICS-BASED ALIGNMENT: A PAIRWISE COMPARISON OF LOW-ENERGY MODES IN PROTEINS

Chapter 5

Prediction of Nucleic Acid Binding Sites in Proteins using the Dynamics-base Alignment

5.1 Introduction

As already mentioned, the functionality of proteins and enzymes often relies on the capability of these biomolecules to sustain large-scale conformational changes (Frauenfelder *et al.*, 1991). It has been established that these concerted functional movements are typically shared by members of enzymatic superfamilies which may otherwise differ significantly by fold, oligomeric state, and even by the details of the catalytic chemistry (Capozzi *et al.*, 2007; Carnevale *et al.*, 2006). In the previous chapter we have introduced a quantitative algorithm to detect similar motions in protein pairs. The procedure is termed dynamics-based alignment because it allows the establishment of one-to-one correspondences between amino acids that experience similar large-scale movements in the two molecules (Zen *et al.*, 2008). Applying this method to a set of representatives of enzymatic functional families, we have shown that a dynamics-based alignment can result in a remarkable spatial superposition of functionally relevant regions even for structurally dissimilar families of proteins. These results suggest that specific common concerted movements may have a functional rationale (Carnevale *et al.*, 2006; Zen *et al.*, 2008).

The goal of this study is to illustrate this concept using as a model system the OB fold, a well characterized nucleic-acid binding motif for which several structures are available in the PDB database in both their free and bound forms. Most commonly,

the OB fold consists of a closed barrel formed by two three-stranded antiparallel β-sheets. $\beta 1$ is shared by both sheets whilst $\beta 3$ and $\beta 5$ close the barrel partially or completely by forming a parallel network of hydrogen bonds (Murzin, 1993; Theobald *et al.*, 2003). The structure and topology of a canonical OB-fold representative are shown in fig. 5.1a.

A relatively distant member of the OB family known to bind nucleic acids is formed by the AXH motif. So far, AXH motifs have been identified in two apparently unrelated human proteins of medical importance (Mushegian *et al.*, 1997): the HMG box transcription factor HBP1 and the polyglutamine-containing ATX1 protein (Banfi *et al.*, 1994; Lesage *et al.*, 1994). Both proteins are thought to be transcription factors (Berasi *et al.*, 2004; Tsai *et al.*, 2004). HBP1, first identified as a target for family members of the retinoblastoma tumor suppressor (Lavender *et al.*, 1997; Tevosian *et al.*, 1997), is involved in cancer signalling pathways (Paulson *et al.*, 2007). Mutations in ATX1 cause the spinocerebellar ataxia type-1 (SCA1), an autosomal-dominant neurodegenerative disorder characterized by ataxia and progressive motor deterioration (reviewed in Orr & Zoghbi (2001)).

The two AXH domains of ATX1 and HBP1 (ATX1_AXH and HBP1_AXH) share a sequence identity of ca. 30% and a homology of ca. 50% depending on the species. Though evolutionarily related, the two proteins have different domain boundaries and distinct properties (de Chiara *et al.*, 2003). ATX1_AXH, as solved by crystallography (Chen *et al.*, 2004), forms a dimer of asymmetric dimers. The structure of the dimer formed by chains A and B and topology of chain A are shown in fig. 5.1c. The corresponding region of HBP1 is a monomer in solution as assessed by nuclear magnetic resonance (NMR) (de Chiara *et al.*, 2005), and its structure and topology are shown in fig. 5.1b. Possibly because of their self-association properties and because of a long insertion in HBP1_AXH, the two domains have the same secondary structure but are not topologically equivalent (see fig. 5.1). The AXH motifs seem to play an important role in the function of the respective proteins as most of the interactions of both ATX1 and HBP1 with other molecular partners map into these regions (de Chiara *et al.*, 2003; Yue *et al.*, 2001). Both domains have been shown to bind nucleic acids in vitro, although with different specificities. ATX1_AXH binds RNA homopolymers with preference for poly(rG) and poly(rU) (de Chiara *et al.*, 2003). This preference corresponds to the same specificity observed for the full-length protein (Yue *et al.*, 2001). HBP1_AXH (de Chiara *et al.*, 2003; Yue *et al.*, 2001) binds poly(rU) and poly(rA). Weaker or no binding was observed for poly(rG) and poly(rC). No structure of an AXH complex with RNA or DNA is available, and the surface of interaction to RNA was hypothesized only

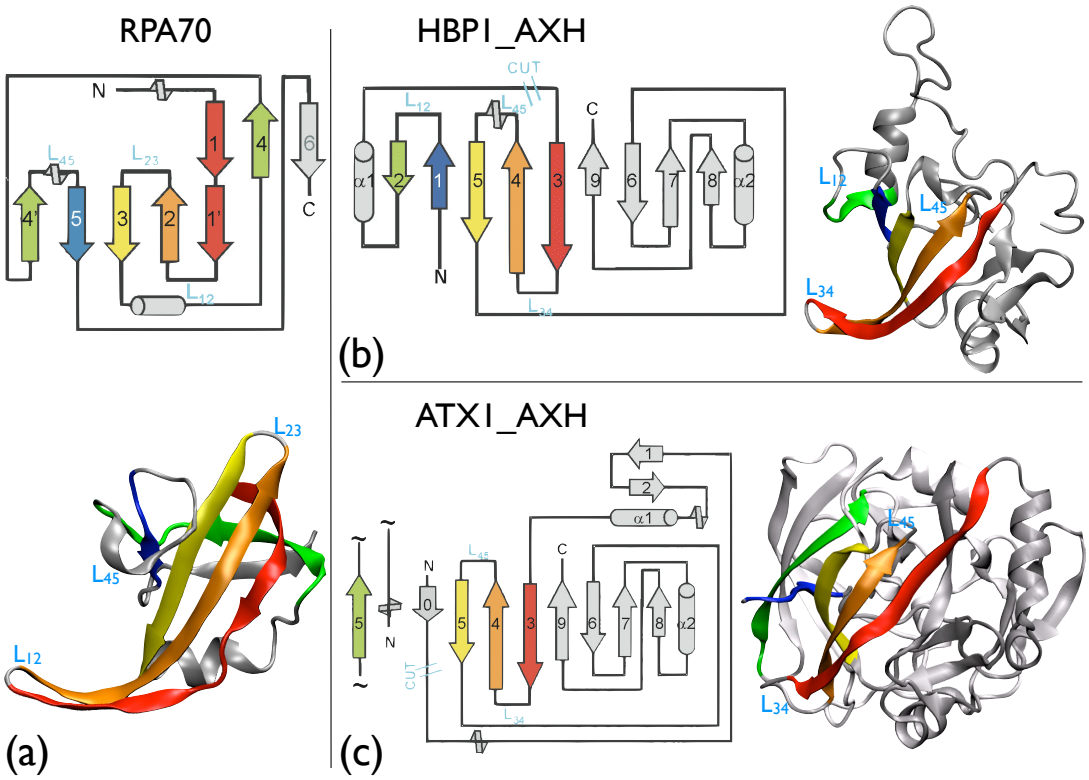

Figure 5.1: Comparison of topologies and structures of (a) the canonical OB-fold (RPA70, repeat DBD-A) and the non-canonical one of (b) HBP1_AXH and (c) the dimeric ATX1_AXH. Corresponding β-strands are indicated with the same colors thereby highlighting the different sequential order and sequence directionality of matching strands. In ataxin-1, the green strand is from the contiguous monomer (only this element is indicated). It is also worth noting that the symmetry of the dimer breaks around this region and strand β5 in monomer A corresponds to a short 3-10 helix in monomer B.

on the basis of the combined use of sequence conservation and structure-based analysis. AXH domains therefore constitute a paradigmatic example on which to test the possibilities of a dynamics-based alignment approach.

Our analysis is organized in two steps. First, the viability of a dynamics-based alignment as a scheme to predict putative binding sites, was investigated by aligning OB-fold members whose interaction surface with RNA or DNA is known. By adopting an ENM (Atilgan *et al.*, 2001; Bahar *et al.*, 1997; Delarue & Sanejouand, 2002; Hinsen, 1998; Micheletti *et al.*, 2004), the βGM (Micheletti *et al.*, 2004), we calculated the low-energy modes for members of the OB fold family. Using the dynamics-based alignment (Zen *et al.*, 2008), introduced in the previous chapter, we identified the regions sharing similar dynamics. These regions were correlated to the surfaces involved in nucleic acid binding and/or recognition. We found that the amino acids involved in several pairwise dynamics-based alignments have a good overlap with the known surface of interaction with nucleic acids. Based on this validation, the dynamics-based alignment was next used to predict the putative DNA/RNA interaction surfaces of HBP1_AXH and ATX1_AXH. The predicted sites are a subset of those previously singled-out on the basis of supervised structural alignments (de Chiara *et al.*, 2005) and do not involve positively-charged amino acids.

We propose the dynamics-based method as a new approach for predicting functional regions in protein families.

5.2 Consensus profile of dynamics-based alignments

We recall that dynamics-based alignment establishes one-to-one correspondences between groups of amino acids experiencing similar large-scale motions in two given proteins. As described in detail in previous chapters, the method is based on a stochastic exploration of the space of the correspondences of amino acids, aimed at obtaining the matches that maximize the alignment score, a quantity that measures the similarity of the large-scale motions of the amino acids into correspondence. These large scale motions of the residues marked for the alignment are evaluated using βGM (see first chapter). We have seen in the previous chapter how the ENMs' approach allows a transparent treatment of the influence of the amino acids non marked for the alignment on the mobility of the aligned ones.

The alignment score accounts for *both* the agreement between the low-energy modes of the marked amino acids *and* their good space proximity after an optimal alignment. We wish to recall here that we can assign to each alignment a statistical significance,

which is obtained from the comparison of its score against an empirical (Gaussian) reference distribution for alignment scores involving residues in *a priori* unrelated protein pairs (see section 4.2.5). In this way, each alignment between two given proteins has a corresponding z-score and p-value.

Here, we present a systematic analysis to identify the key aligned residues that recurrently appear in significant alignments (i.e. alignments with p-value larger than a suitable threshold). For each protein we calculated the *consensus profile* of dynamical accord, that is the residue-wise average contribution to the statistically-significant alignment with other OB-fold members.

The degree of dynamical involvement of the k-th amino acid of the reference protein A, in an alignment of n amino acids with a protein B, is measured as:

$$\xi_k^{[AB]} \equiv \frac{n}{10} \sum_{\alpha,\beta=1}^{10} \vec{v}_i^{\,\alpha} \cdot \vec{w}_i^{\,\beta} \left(\sum_{j=1}^{n} \vec{v}_j^{\,\alpha} \cdot \vec{w}_j^{\,\beta} \right) \tag{5.1}$$

where $\{\mathbf{v}^\alpha\}_{\alpha=1,\ldots,10}$ and $\{\mathbf{w}^\beta\}_{\beta=1,\ldots,10}$ are the 10 non-zero lowest-energy modes of the n aligned amino acids, for proteins A and B respectively; j runs over the indices of the aligned amino acids and i is the index of the matching pair to which amino acid k takes part to[1]. If the k-th amino acid of A does not take part to the alignment, $\xi_k^{[AB]}$ is set to zero.

The physical meaning of $\xi_k^{[AB]}$ is transparent as, apart from a multiplicative factor, it represents the local contribution to the mean square inner product (MSIP) of the modes of the aligned residues. Indeed, the comparison with equation 4.5 yields: $\text{MSIP} = \frac{1}{n}\sum_k \xi_k^{[AB]}$. Therefore, for a perfect matching of the modes $\{\mathbf{v}^\alpha\}_{\alpha=1,\ldots,10}$ and $\{\mathbf{w}^\beta\}_{\beta=1,\ldots,10}$, the average value of $\xi_k^{[AB]}$ per aligned amino acids is 1.

Based on this observation, the alignment consensus score ξ_k^A of the k-th amino acid of protein A is defined as the average of $\xi_k^{[AX]}$ over all the N_n^A significant alignments of length n involving A:

$$\xi_k^A \equiv \left\langle \xi_k^{[AX]} \right\rangle_X = \frac{1}{N_n^A} \sum_X{}' \xi_k^{[AX]} \tag{5.2}$$

[1] It is worth noticing that equation 5.1 rewrites, in terms of the optimally consistent lowest-energy modes $\{\mathbf{v'}^\alpha\}$ and $\{\mathbf{w'}^\beta\}$ (see sections 4.2.6 and A), in the following way:

$$\xi_k^{[AB]} \equiv \frac{n}{10} \sum_{\alpha=1}^{10} \vec{v'}_i^{\,\alpha} \cdot \vec{w'}_i^{\,\alpha} \left(\mathbf{v'}^\alpha \cdot \mathbf{w'}^\alpha \right)$$

where we have used that $\mathbf{v'}^\alpha \cdot \mathbf{w'}^\beta = \sum_{j=1}^n \vec{v'}_j^{\,\alpha} \cdot \vec{w'}_j^{\,\beta} = 0$ if $\alpha \neq \beta$. It is clear from this expression that the most consistent modes gives the biggest contribution to $\xi_k^{[AB]}$.

where the prime in the rightmost expression indicate the summation over the proteins X which have a significant alignment of length n with A.

The consensus score, which provides an indication of the dynamically most important amino acids of a protein, is used in this study to predict a set of residues putatively involved in the binding of the nucleic acids.

5.3 Validation of the dynamics-based prediction scheme

A set of canonical OB-fold representatives was compiled based on the OB-fold survey of Theobald *et al.* (2003). The detailed list of representatives is shown in Table 5.1. Note that some of the selected OB-fold domains are part of bigger proteins. The part of the protein selected for this study is reported in the last column of the table. The selected OB-folds are holo forms, i.e. they are complexed with nucleic acids. In this way we can compare our predictions, that clearly have to be obtained removing the nucleic acids, with the true binding sites, i.e. amino acids that actually bind the nucleic acids.

The holo forms are hence used to validate the dynamics-based prediction scheme. However, the ultimate purpose of the aligment is to predict the regions putatively involved in the binding of nucleic acids, in proteins for which the holo form is clearly not available. It is known that the structure of the apo form of a protein can be different from the holo form, see Fig. 5.2. Considering that the dynamics-based alignment partly relies on structural information, a fundamental question has to be addressed: is it appropriate to validate our prediction method on holo forms, and then apply that on apo forms? To justify to use the holo forms, we can *a priori* argue that the dynamics-based alignment has a spatial tolerance that allows one to establish correspondences also among different secondary elements, provided that their dynamics is similar. A further support in favor of the strategy is that we have checked *a posteriori* that when we considered also the corresponding apo forms of some of the OB-folds reported in table 5.1, we obtained results perfectly coherent with those of the holo forms.

5.3.1 Dynamics-based alignment of the OB-fold representatives

Dynamics-based alignments were carried out among all 120 distinct pairings of the 16 canonical OB-fold representatives constituted by all the domains listed in Table 5.1. The quality of each alignment is conveyed by an alignment score which rewards correspondences between amino acids that have (i) similar geometric relationships in the two proteins and (ii) sustain similar large-scale movements. The combined consideration of structural and dynamical features ensures that high-scoring alignments reflect genuine

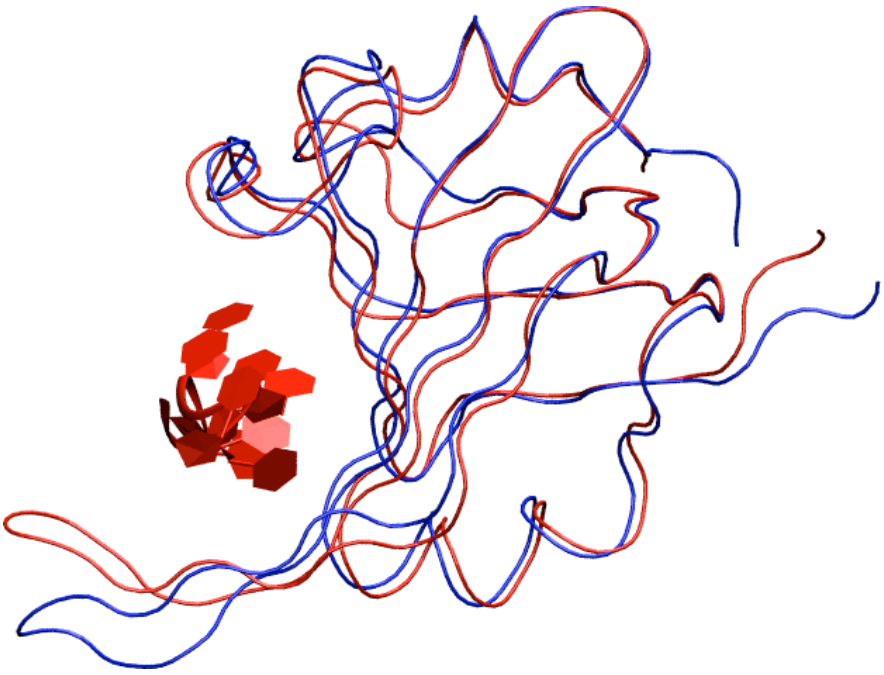

Figure 5.2: Apo and holo forms, represented respectively in blue and red, of the OB-fold domain RPA70. DNA for the holo form is represented in red. The conformational change upon ligand binding can be appreciated as the structures as been optimally superimposed minimizing the RMSD between their C_α atoms.

Table 5.1: OB-fold representatives (holo forms) considered in this study.

#	Structure	bound ligand	PDB id	domain (chain, residue range)
1	RPA70	ssDNA	1jmc	DBD-A (A, 198-289)
2	RPA70	ssDNA	1jmc	DBD-B (A, 305-402)
3	EcSSB	ssDNA	1eyg	(A, 1007-1112)
4	EcRho	ssRNA	2a8v	(A, 48-118)
5	OnTEBP α_1	ssDNA	1jb7	domain 1 (A, 36-204)
6	OnTEBP α_1	ssDNA	1jb7	domain 2 (A, 205-314)
7	OnTEBP α_2	ssDNA	1kix	domain 1 (A, 36-204)
8	OnTEBP α_2	ssDNA	1kix	domain 2 (A, 205-314)
9	OnTEBP β	ssDNA	1k8g	domain 1 (A, 36-203)
10	OnTEBP β	ssDNA	1k8g	domain 2 (A, 205-315)
11	EcAspRS	tRNA anticodon	1c0a	(A 1-104)
12	ScAspRS	tRNA anticodon	1asy	domain 1 (A, 68-201)
13	ScAspRS	tRNA anticodon	1asy	domain 2 (B, 68-201)
14	RecG	Junction DNA	1gm5	(A, 157-245)
15	S12	16S rRNA	1j5e	(L ,26-110)
16	S17	16S rRNA	1j5e	(Q, 3-102)

correspondences of large-scale rearrangements in two given proteins. The statistical significance of each alignment is quantified by comparing the score against a reference distribution of scores from a heterogeneous set of enzymes. From this comparison, we could calculate a p-value (or equivalently a z-score). Given the limited size of the database considered, we assumed as indicative of a significant alignment a z-score > 2.3, corresponding to a p-value < 0.01.

The dynamics-based scores for all pairwise alignments among the proteins in Table 5.1 are provided in the density maps of fig. 5.3a. The accompanying graph, see fig. 5.3b, summarises the dynamics-based correspondences having a statistical significance higher than the above mentioned threshold. Inspection of the graph reveals the existence of several triangular relations (i.e. protein A is in relation with proteins B and C, and also B is in relation with C). Proteins OnTEBP, RPA70 and RecG, for instance, form a completely connected subgraph. These circular relationships suggest the existence of a common alignable core among these proteins. This can be verified by inspecting fig. 5.4 which shows pileup representations of the alignments involving OnTEBP α2

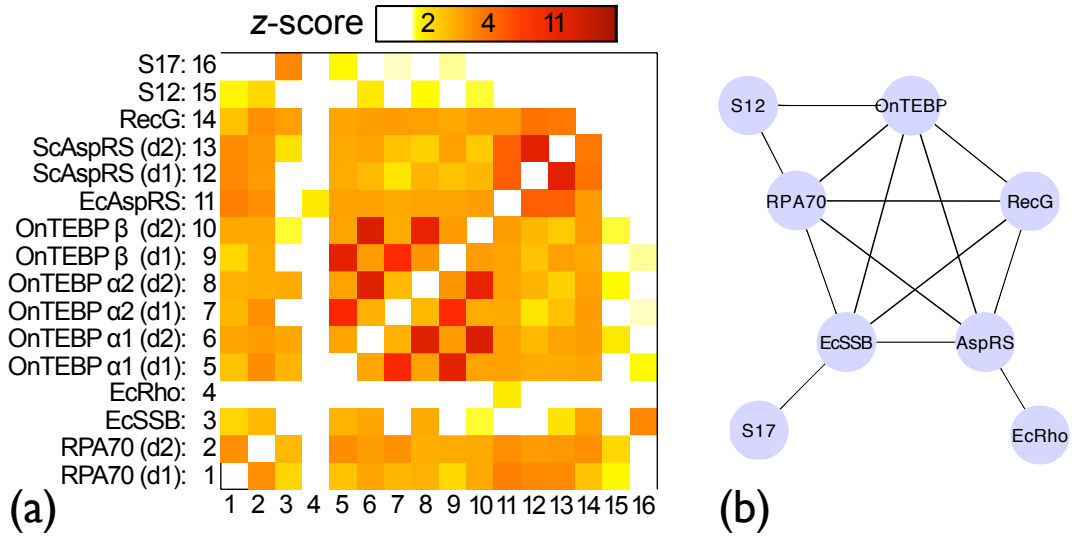

Figure 5.3: (a) Density map of the z-score for all pairwise dynamics-based alignment of canonical OB-fold representatives (indexing according to the Table 5.1). (b) Graph representation of significant pairwise alignments (z-score > 2.3).

(domain 1), RecG and RPA70 (repeat DBD-B) and the alignable partners.

The structural superposition of OnTEBP $\alpha 2$ (domain 1) with RecG and with RPA70 (repeat DBD-B) is shown in Figs. 5.5a and 5.5b, respectively. The alignable regions involve amino acids that are flexible and in proximity of the bound nucleic acid, as can be appreciated by comparison with the complexes in fig. 5.6, where the nucleic acid is represented.

This observation suggests that the set of amino acids of a given OB-fold that can be significantly aligned with several other OB-fold partners are typically located in regions involved in nucleic acid binding.

5.3.2 Performance of the dynamics-based prediction scheme

The dynamics based prediction of nucleic acid binding amino acids is compared, for validation purposes, against the sites that actually bind DNA or RNA. As in (Jones *et al.*, 2003), they are identified as the amino acids whose accessible surface area (ASA) changes by more than 1Å^2 upon omitting the nucleic acid from the available structure of the protein/DNA (or RNA) complex. The calculation of the ASA was performed with NACCESS (Hubbard & Thornton, 1993). For most of the proteins in Table 5.1, the typical fraction of residues contacting nucleic acids is 20%.

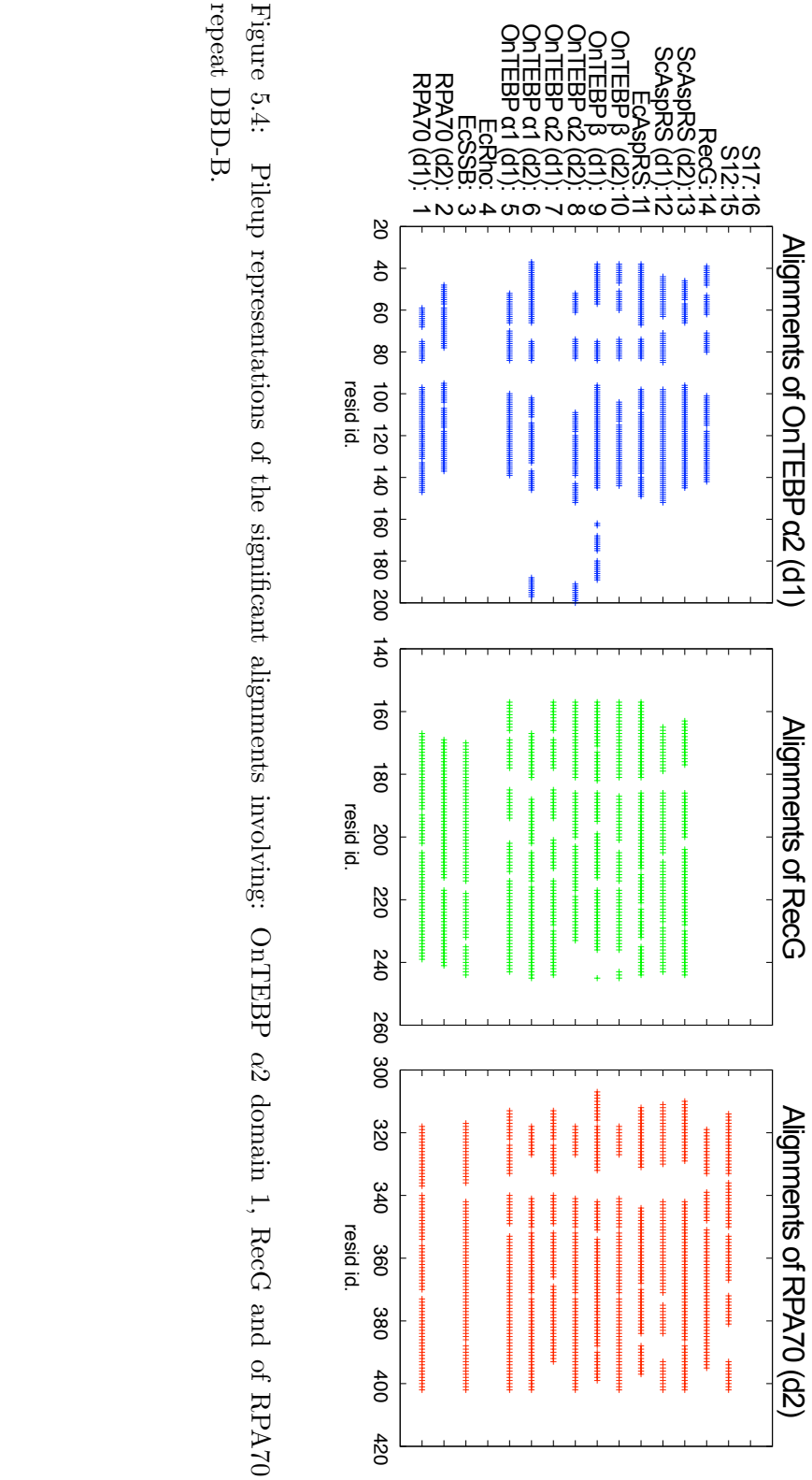

Figure 5.4: Pileup representations of the significant alignments involving: OnTEBP $\alpha 2$ domain 1, RecG and of RPA70 repeat DBD-B.

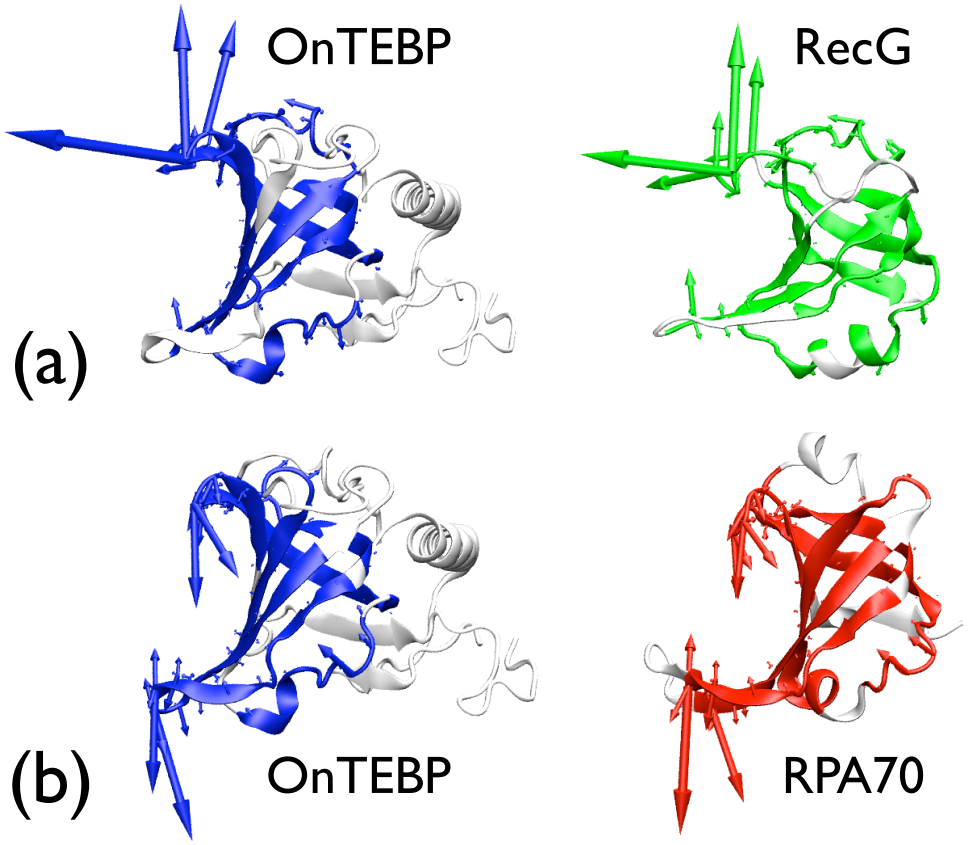

Figure 5.5: The dynamics-based alignment of OnTEBP α2 domain 1 and RecG is shown in (a), while the one between OnTEBP α2 domain 1 and RPA70 repeat DBD-B is shown in (b). Amino acids involved in alignments are colored. The arrows represent the three best corresponding lowest-energy modes for the aligned regions, see section 4.2.6.

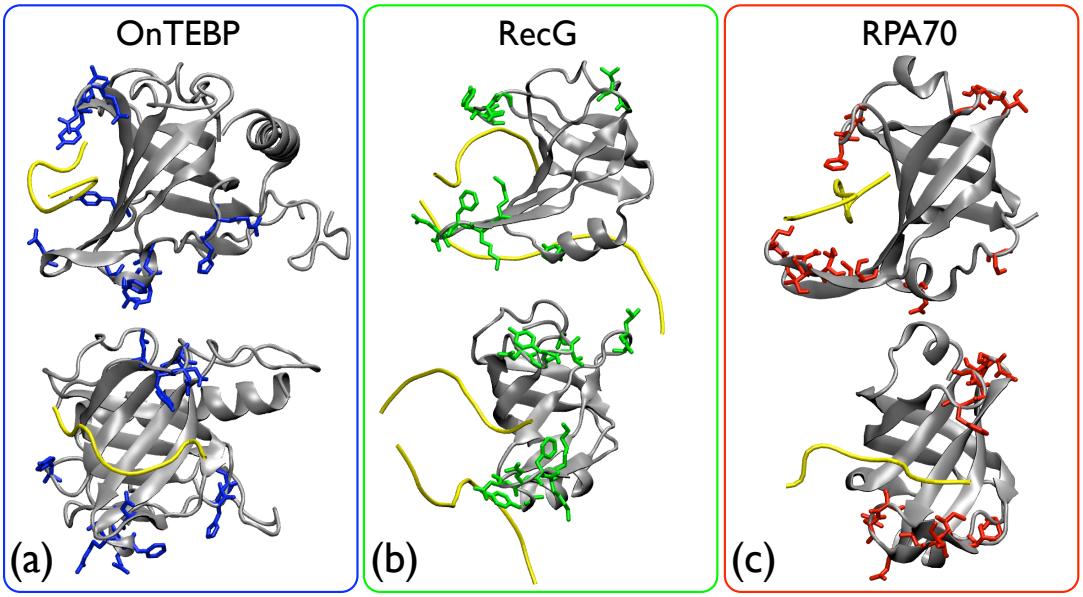

Figure 5.6: Panels (a), (b) and (c) illustrate, respectively, the consensus residues of OnTEBP $\alpha 2$ domain 1, RecG and RPA70 repeat DBD-B. Two different views are displayed, the upper one is the same adopted in fig. 5.5, the lower is rotated of 90^o around the z-axis. Nucleic acid strands are shown as yellow tubes and the sidechains of consensus residues are highlighted in color.

Amino acids are divided in those predicted to interact or not to interact with nucleic acids according to whether their consensus score is, respectively above, or below a given threshold. All possible values for the threshold were considered and the performance of the prediction was assessed by comparison against the sets of amino acids that are known to interact (or not interact) with DNA/RNA. For a given threshold value, the prediction is characterized, as customary, in terms of the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The TP are the amino acids that are correctly predicted to interact with DNA or RNA, while TN are those correctly predicted not to interact. The FP are the amino acids that are incorrectly predicted to interact with DNA or RNA, while FN are those incorrectly predicted not to interact. These basic quantities are used to define the accuracy, specificity and selectivity of the prediction (Baldi *et al.*, 2000).

The accuracy is the fraction of correct prediction for amino acids that are, or are not, contacting nucleic acids and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \ . \tag{5.3}$$

The specificity, defined as:

$$\text{Specificity} = \frac{TP}{TP + FP} \ , \tag{5.4}$$

represents the fraction of correct hits among residues predicted.
The sensitivity:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \ , \tag{5.5}$$

is the fraction of residues known to interact with DNA/RNA which are predicted to do so.

The predictive performance of the method as a function of the consensus score threshold is aptly summarized by the Receiver Operating Characteristic curve (ROC curve) obtained by plotting "hit rate" (sensitivity, TP/(TP+FN)) versus the "false alarm rate" (false positive rate, FP/(FP+TN)).

We computed the consensus alignment score for all amino acids of proteins RPA70 (repeat DBD-A), EcSSB, EcRho, OnTEBP α1 (domain 1), OnTEBP α2 (domain 1), OnTEBP β (domain 1), EcAspRS, ScAspRS (domain 1), RecG. Notice that proteins S12 and S17, that are largely surrounded by nucleic acids, were not considered for the test and that, to limit redundancy, only the N terminal domain was retained for multidomain proteins. Since in this study most of the significant alignments that we have obtained have length n=70, we have used only the alignments of 70 residues to calculate the consensus score.

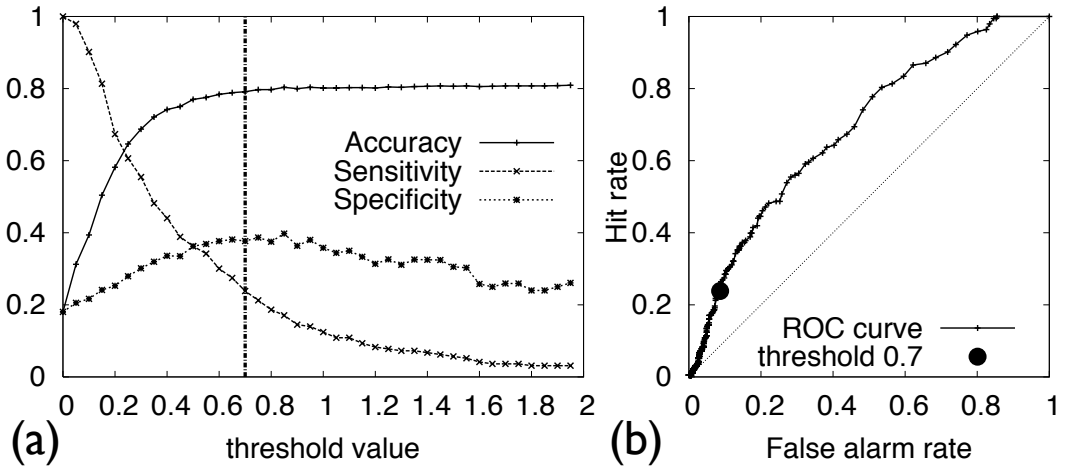

Figure 5.7: (a) Trend for the accuracy, sensitivity and specificity of dynamics-based predictions of amino acids at the protein/nucleic acid interface are shown as a function of the consensus score threshold. (b) Corresponding receiver operating characteristic (ROC) curve.

Amino acids with a sufficiently high consensus score are expected to be relevant for the functional dynamics and hence to correlate with sites involved in nucleic acid binding. To assess the extent to which the consensus score can be used to predict interaction sites with DNA/RNA we carried out the performance analysis. The results are summarised in the plots in fig. 5.7.

The plots can be used to set the threshold for the consensus score so to have a balanced predictive performance in terms of accuracy, specificity and selectivity. In fact, excessively large threshold values correspond to very few predictions for amino acids interacting with DNA/RNA and this reflects in a poor coverage of the sites that are known to interact with nucleic acids. Conversely, very small threshold values result in predicting that almost all amino acids interact with DNA/RNA thus leading to a large fraction of false positives. A balance between these two limiting situations is achieved by setting the consensus score threshold to 0.7. Examples of the consensus regions are given in fig. 5.6a-c.

The corresponding overall accuracy of the algorithm is 79%, specificity is 38% and sensitivity is 24%.

5.3.3 Comparison between dynamics-based and other prediction schemes

A useful term of reference for these values is provided by advanced sequence-based techniques for the prediction of nucleic-acids binding sites. For instance, an accuracy of 71%, a specificity of 35% and a sensitivity of 53% was calculated for the method implemented by Yan *et al.* (2006), in a different dataset of DNA-binding proteins.

In addition, on the specific dataset considered in this study, the on-line sequence-based method[1] of Hwang *et al.* (2007) for DNA binding-sites prediction had an accuracy of 63%, a specificity of 23% and a sensitivity of 45%. In fig. 5.8 it is shown the dynamics-based and sequence-based predictions for OnTEBP α2, RecG and RPA70, in comparison with the actual DNA-binding residues.

It therefore emerges that the dynamics-based approach compares well with other prediction schemes in terms of accuracy and specificity, while returns appreciably smaller values for sensitivity. This aspect is rationalised by the observation that the dynamics-based alignment will be especially promoted in correspondence of flexible amino acids, and consequently the residues close to the nucleic acid chain and with a low mobility are likely to have a low consensus score. The dynamics-based predictions are therefore particularly targeted at a specific subset of nucleic acid binding sites (the mobile ones) and this reflects in a diminished sensitivity of the algorithm compared to the complementary sequence-based methods. Additionally, regions which cannot be aligned and that are therefore not common to all OB folds may be also involved in binding and be the ones responsible for recognition specificity.

5.4 Prediction of the nucleic acid binding surface of the AXH domains

The above results indicate that, within the limits of binding specificity, the consensus residues point at regions involved in nucleic acid binding. The approach was used as a predictive tool for representatives of the AXH-domain family. The first model of the PDB file 1v06 was taken as the reference structure of HBP1_AXH, while for ATX1_AXH we considered the dimer (chains A and B) of PDB file 1oa8.

Prediction of the nucleic acid binding surface based on sequence and structural comparison with other members of the OB-fold was previously attempted (de Chiara *et al.*, 2005). However, the two families are too divergent to extract useful hints from sequence conservation, whereas a structure-based analysis was inconclusive. It was

[1] From the available on-line web server (http://lcg.rit.albany.edu/dp-bind) we selected the mode "sequence-based binary encoding" to input the protein sequence.

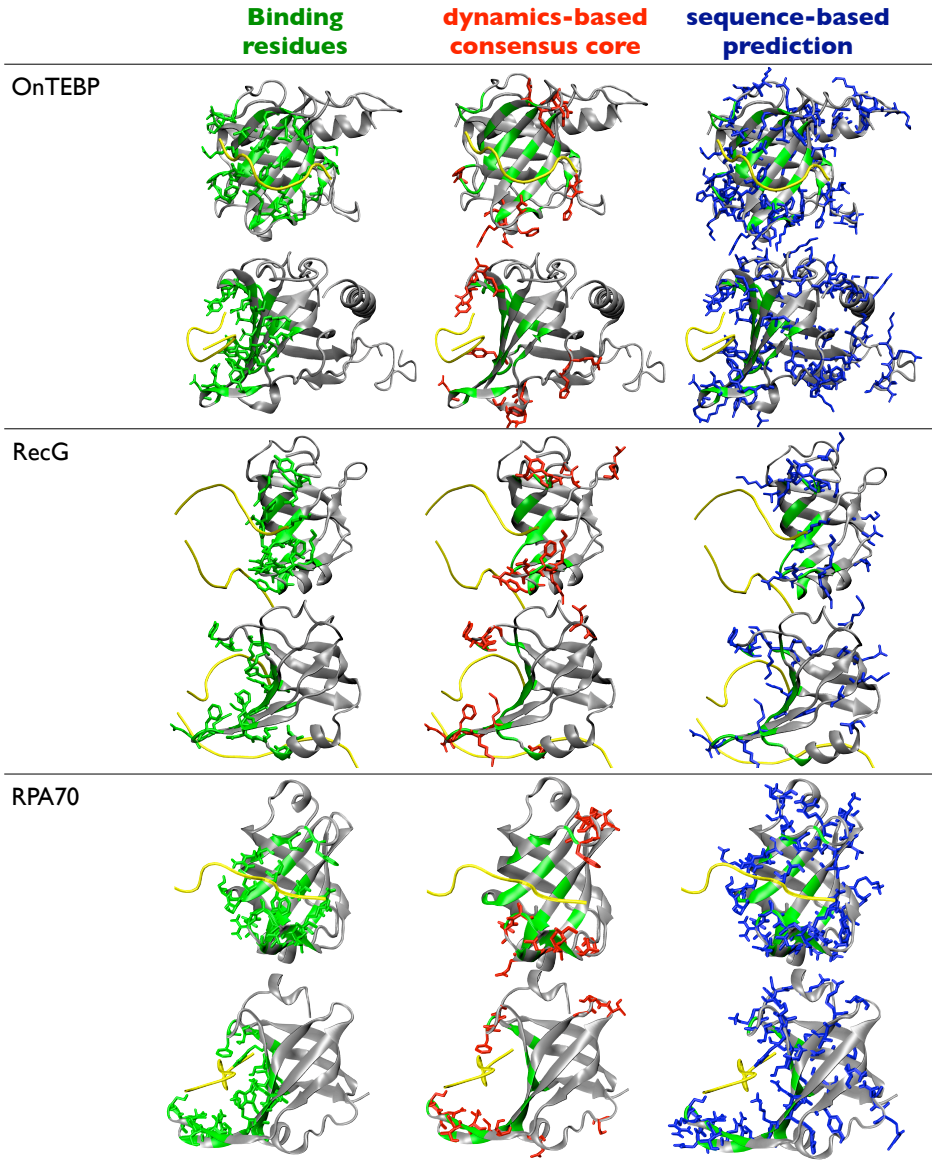

Figure 5.8: Comparison between the nucleic acid binding residues, the dynamics-based consensus residues and the sequence-based DP-bind prediction (Hwang *et al.*, 2007) for DNA binding residues. Proteins here shown are OnTEBP α2 domain 1 (top panel), RecG (middle panel) and RPA70 repeat DBD-B (bottom panel), as in fig. 5.6. DNA strands are shown as yellow tubes and residues that actually bind DNA are highlighted in green. Their sidechains are explicitly reported in green (first column). Sidechains highlighted in red (second column) corresponds to our dynamics-based consensus cose, and sidechains highlighted in blue (third column) correspond to a sequence-based prediction. Two different views are here displayed, as in figure 5.6.

only through a combined use of sequence and structural conservation that two distinct patches of conserved or semi-conserved residues could be identified. Only one of them corresponds to the surface involved in nucleic acid binding in other OB-folds. We therefore reasoned that this example would be an appropriate case for attempting a dynamics-based prediction.

5.4.1 Comparison between canonical and non-canonical OB-folds

The salient differences of the canonical and non-canonical OB-folds are illustrated in fig. 5.1. Structurally-corresponding β-strands in RPA70 and the AXH domains are shown with the same colour (and same letter in the secondary structure topologies). Strands $\beta1$, $\beta2$ and $\beta3$ of RPA70 match strands $\beta3$, $\beta4$ and $\beta5$ of HBP1. However, strands $\beta5$ and $\beta4$ of RPA70 do not correspond to $\beta7$ and $\beta6$ of HBP1, as expected for preserved β-strands succession, but with $\beta1$ and $\beta2$. The latter, in addition, have opposite sequence directionality with respect to RPA70.

Alignments where amino acids are paired sequentially from the N- to C-termini cannot set correspondences of all five β-strands in canonical and non-canonical OB-folds. As described in the previous section, the space of possible alignments of two proteins is too large for an exhaustive exploration, therefore some constraints were introduced within the dynamics-based alignment, to restrict the search space of matching residues. One of these constraints was that the one-to-one correspondences between the amino acids follow the sequential ordering. The pairing scheme was accordingly generalised by "remapping" the amino acid indices so to achieve a consistent β-strands matching on canonical and non-canonical folds. The procedure is illustrated in fig. 5.9. Amino acid reindexing was performed by (i) introducing a single "virtual cut" in HBP1, and (ii) by changing the order of the two subchains and the sequence directionality in one of the two (see diagrams at bottom of panel d). The location of the virtual cut is found by identifying which blocks of residues, and in which sequence order, can be put in loose structural correspondence by local structural alignments. This was done by structurally superposing short segments of 20 amino acids in RPA70 and HBP1. Such superpositions may induce the spatial proximity (C_α separation below 3Å) of other amino acids besides those in the two segments. Several local superpositions imply global correspondences in that they entail more than half of the residues in RPA70 are in proximity with a residue in protein HBP1. The matrix in fig. 5.9 reports the mapping of such global pairings, which being induced by local structural superpositions can capture robust global structural correspondences that are elusive to structural alignments methods employing various combinatorial explorations of matching segments. Inspection of the

mapping, allows a transparent identification of the virtual cut for both HBP1 (fig. 5.9) and ATX1.

5.4.2 Predictions and discussion

Since HBP1_AXH is monomeric and therefore easier to deal with than ATX1_AXH, we aligned it (1v06) first against OB-fold representatives using their dynamics properties. HBP1_AXH can be significantly aligned with two distinct regions of RPA70 (z-score 3.5) (fig. 5.10a). It also aligns with RecG with a z-score of 2.5.

The single stranded DNA-binding domain of human RPA70 (residues 183-420) contains two tandem OB-fold repeats. Dynamics-based alignments of HBP1_AXH against both repeats are highly consistent and involve residues 212-237 and 214-235 (including β1 and β2) with a reversed backbone orientation to regions β4 and β5 (fig. 5.1a) of DBD-A and DBD-B. The consensus regions emerging from such alignments strongly suggest that nucleic acid binding involves HBP1 residues N228, K229, E230, S270, V271, S272, F273, G274, E275, T286, V287 and E288 which correspond to the cavity formed by loops β1/β2, β3/β4 and β4/β5 of HBP1 (fig. 5.10a, left). These residues correspond to residues in direct contact with DNA in the holo-form of RPA70 (fig. 5.10a, right). The predicted residues are not positively charged, suggesting that the interaction would not be electrostatically driven but rather sequence or structural specific. They are well consistent with those previously predicted on the base of a structural alignment (fig. 5.10b, citedeChiara:2005p868).

ATX1_AXH aligns with RecG with a z-score of 3.3 (fig. 5.11a). The aligned sidechains are all exposed and do not interfere with dimer formation (fig. 5.11b).

Finally, the dynamics-based alignment between HBP1_AXH and ATX1_AXH comprises residues 257-271, 274-288, 290-339, 222-213 and 609-623, 624-638, 639-688, 565-574 respectively (fig. 5.12a). It is worth noting that the region 222-213 of HBP1_AXH, which is not topologically equivalent in the two proteins, aligns with a reverse orientation in sequence with the corresponding region of ATX1_AXH (fig. 5.1b,c). This could suggest that despite their difference, the two regions share a functional role within the context of the domain.

ATX1_AXH (monomer A) and HBP1_AXH can be superposed by structural criteria (fig. 5.12b) with an RMSD of 3.8 Å over 84 amino acids. The two folds differ for the topology of an N-terminal β1, β2 and α1 motif which packs differently in the two structures.

At the same time, the spacing between these three elements of secondary structure is different and only the regions 260-335 of HBP1_AXH and 612-684 of ATX1 AXH can be

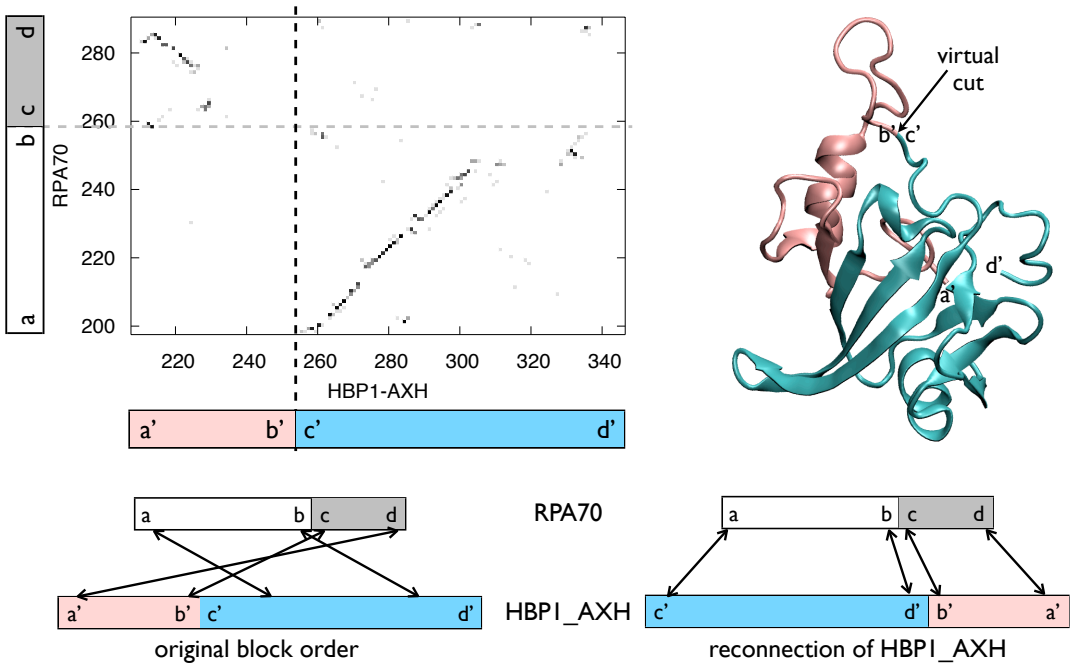

Figure 5.9: In HBP1_AXH the canonical order and directionality of β-strands is achieved (for alignment convenience) by juxtaposing the two parts separated by the virtual cut, as shown. The procedure to identify the virtual cut is described in the text.

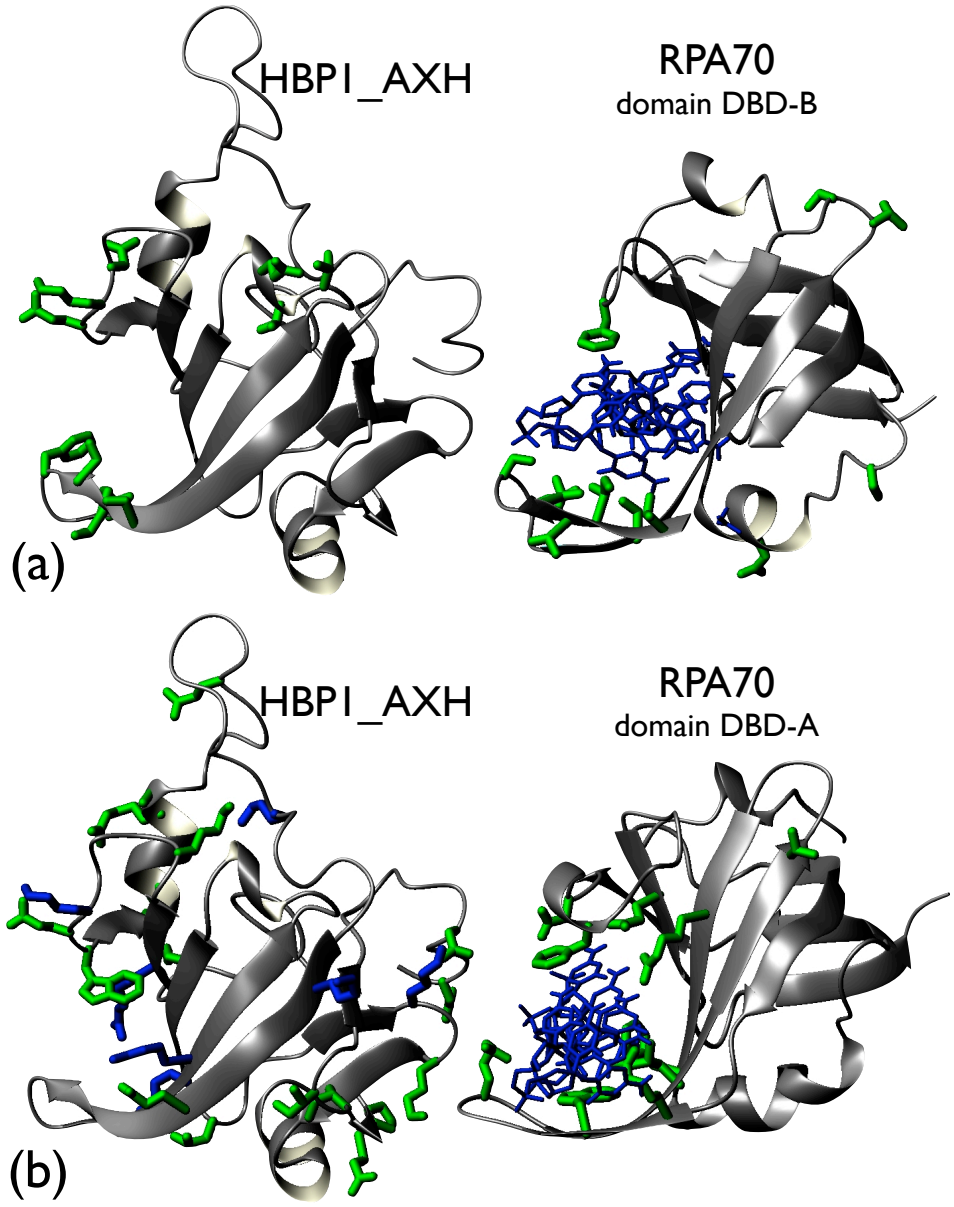

Figure 5.10: (a) Ribbon representations of HBP1_AXH (left) and of the DBD-B repeat of RPA70 (right) as dynamically aligned. The side chains of consensus residues are explicitly reported. (b) Comparison of HBP1_AXH (left) and of the DBD-A repeat of RPA70 (right) in complex with DNA (in blue) as aligned structurally (adapted from fig. 5 of de Chiara *et al.* (2005)). The side chains of completely and semiconserved residues of HBP1_AXH are indicated in green, additional lysines and arginines that could contribute to binding are shown in blue. DNA and the side chains of residues of RPA70 DBD-A in contact with DNA are indicated explicitly in blue and green respectively.

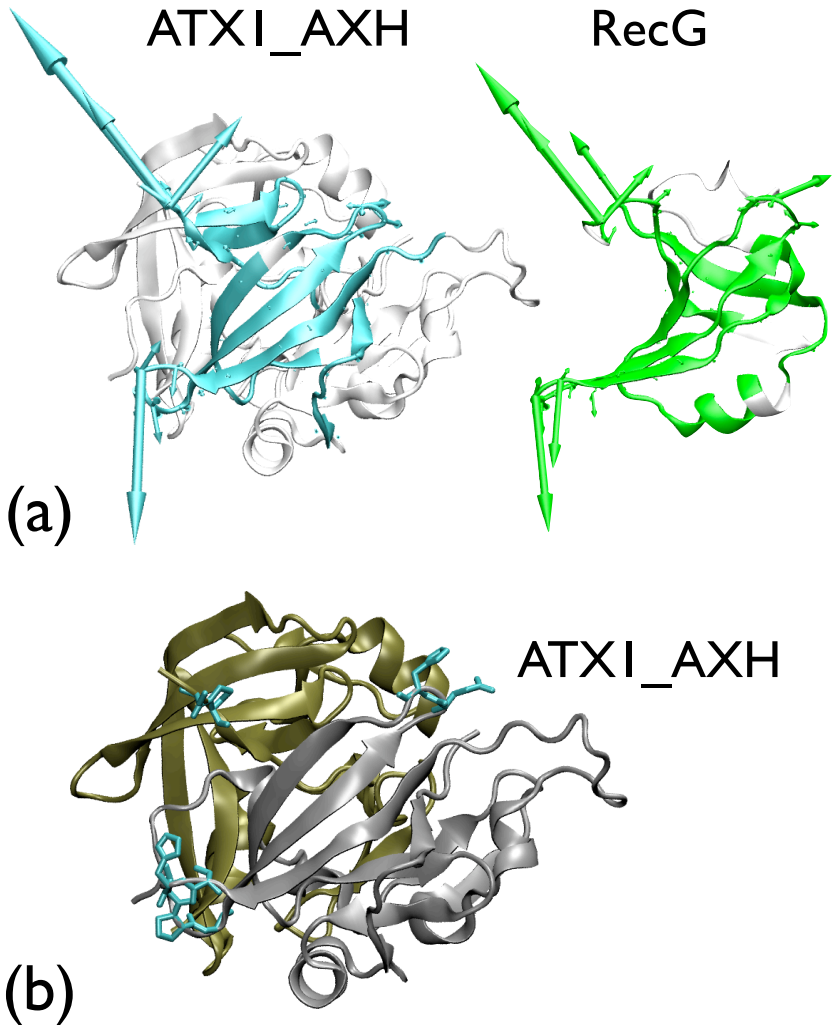

Figure 5.11: (a) Dynamics-based alignment of the ATX1_AXH dimer (left) with RecG. (right). Aligned regions are shown in cyan and green respectively. Arrows represent the three best corresponding lowest-energy modes for the aligned residues, as described in section 4.2.6. (b) The sidechians of the consensus residues are shown on the ATX1_AXH dimer in cyan. The two subunits forming the dimer are in gold and silver.

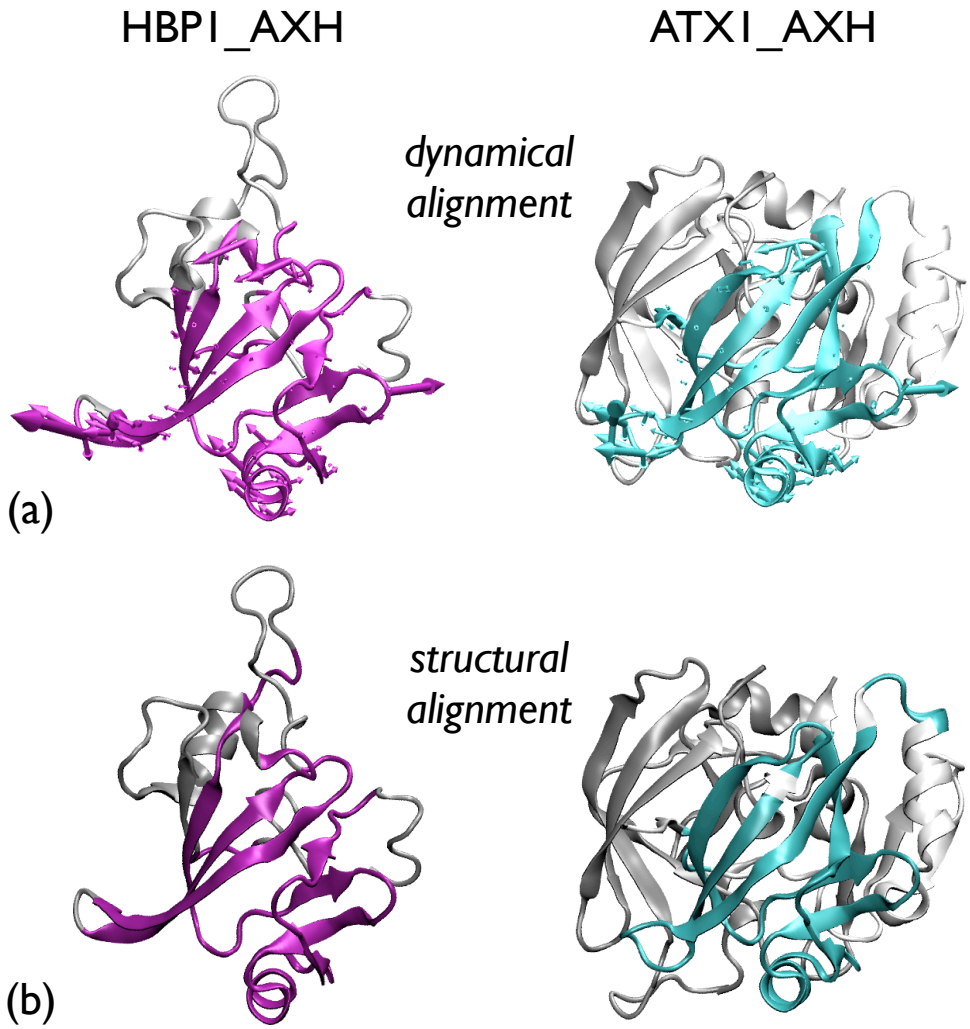

Figure 5.12: (a) Dynamics-based alignment of HBP1_AXH (left) and ATX1_AXH (right). Aligned regions are colored in purple and cyan respectively. The dynamics-based alignment involves 90 residues with an RMSD of 3.5 Å. The RMSIP of the ten lowest-energy modes (the best corresponding three are shown as arrows) as calculated using the βGaussian network model, is 0.77. (b) Structurally-based alignment of the same proteins as achieved by DALIlite. 84 residues were aligned with an RMSD of 3.8 Å.

meaningfully aligned. These regions are a subset of the residues alignable on structural considerations (de Chiara *et al.*, 2005). Exposed conserved and semiconserved residues of the AXH subfamily (corresponding to K217, E235, D236, E268, G285, P324, N344, K225, E230, W231, R239, A240, E246, E269, L298, K307, E327, L328, I330 and N341 in HBP1, fig. 5.10b) cluster near the two exposed patches that comprise or are directly contiguous to those predicted by dynamics-based alignment. Interestingly, as for the alignment of ATX1_AXH with other OB-folds, the two AXH folds would not lead to interference of nucleic acid binding with the dimerization interface of the ATX1_AXH domain, thus being well compatible with the knowledge that this domain is an obligate dimer in solution (de Chiara *et al.*, 2005).

5.5 Conclusions

Several methods, both sequence- and structure-based, exist that provide predictions for nucleic acid binding sites in proteins. While sequence-based techniques have the advantage of being applicable when structural models are not available, it is commonly recognized that exploiting structure-based information (such as surface shape, solvent accessibility, interatomic interaction potentials etc.) can significantly improve prediction. Here we introduce and discuss a new method that, while not making use of primary sequence information, identifies putative binding sites on the basis of similarities in the dynamics of a family of proteins. The new approach may be used (possibly in conjunction with other criteria) to predict the interaction surface within a protein family.

We have shown here a specific application to the OB-fold, selected because it represents an ancient fold, able to evolve to accommodate a wide range of sequences and ligand binding functions, and with a structure tolerant to mutation. A the large plethora of data is available for this domain. By comparing the dynamics of a comprehensive subset of members of the family known both in their free and bound forms, we observed that nucleic acid binding sites share common dynamical properties. This observation prompts the consideration that the large-scale movements that putatively accompany/assist biological functionality may be conserved among protein families and that can be detected using dynamics-based alignments. We then used this information to a non-canonical OB-fold, for which the putative nucleic acid binding surface could not be easily predicted from sequence or structural (static) considerations.

While still in need of further validation using different and even more divergent examples, for which sequence and structure-based alignments may be not obvious, our present results encourage us to believe that our method may develop into a useful and

powerful predictive tool. Natural applicative avenues for the method, which we plan to validate in other contexts, are structure/function genomics studies.

Concluding Remarks

In this thesis we reported on a number of investigations where statistical mechanical tools and concepts were introduced and used to characterize aspects of the relationship between structure and function.

Specifically, in chapter 2, we first considered the functionally-relevant movements in a enzyme of primary biological interest, namely adenylate kinase. Consistently with other theoretical and experimental studies (Henzler-Wildman *et al.*, 2007b), our findings indicate that the thermal fluctuations of this enzyme have a preferred directionality, arguably encoded in the protein fold, that assist the free enzyme in attaining the catalytically competent form (Pontiggia *et al.*, 2008). The third chapter was instead devoted to investigating the role of the dynamics in the protein-protein interactions. To this purpose we have studied, using an elastic network model, the mobility at the monomer-monomer interface in a dataset of dimeric proteins and highlighted how the obligatory or not nature of the complexes correlates with detectably different dynamical traits.

In the fourth chapter we introduced and applied a quantitative tool for comparing the internal dynamics of proteins (Zen *et al.*, 2008). The method, termed dynamics-based alignment, is used to gain insight into the last step of the logical ladder $sequence \rightarrow structure \rightarrow function$, which is a used guideline to characterize proteins and enzymes. The tool was applied to a dataset of enzymes covering the main functional and structural classes. Notably, it was found that a number of significative alignments involved enzymes without substantial global or local structural similarity, a fact that highlights the complex relationship between structure and function.

Finally, as described in the final chapter, the alignment method was used to predict the nucleic acid binding residues on the basis of comparative dynamics. It was found that the performance of the dynamics-based prediction scheme compared well with other existing sequence- or structure-based prediction methods (Zen *et al.*, 2009). This suggests that dynamics-based criteria may profitably be introduced to identify functionally-important protein regions in other contexts too. In particular, it would be

most interesting to introduce dynamics-based criteria, along side sequence and structure ones, to investigate the evolutionary relationship of member of various protein families.

Appendix A

Comparing Essential-Dynamics Spaces

In the first and second chapter we have shown that the fluctuations of a protein are often proficiently described by a small set of collective modes, which can arise from from NMA, from PCA of a MD trajectory or from ENM. There is often the necessity to compare the modes provided by different methods, or the essential spaces obtained from two different MD trajectories of the same protein (e.g. starting from different initial configurations). Assume here, for example, that we want to compare the vectorial spaces V and W, spanned by the top N principal components, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ and $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ respectively, of two MD trajectories of the same protein[1]. A measure that is commonly used to quantify the similarity between the spaces V and W is the root mean square inner product (RMSIP):

$$\text{RMSIP} = \sqrt{\frac{1}{N}\sum_{i,j=1}^{N} |\mathbf{v}_i \cdot \mathbf{w}_j|^2} = \sqrt{\frac{1}{N} Tr(P_V P_W)} \tag{A.1}$$

where $Tr(\cdot)$ is the trace, $P_V = \sum_i |v_i\rangle \langle v_i|$ is the projector into the vectorial space V and P_W is the projector into W. Observe that the RMSIP is 1 in case of perfect correspondence between spaces V and W, and it is 0 if they are orthogonal. Typically, the RMSIP of the top 10 principal components calculated from different MD trajectories of the same protein is ~ 0.7.

[1] Notice that each of these vectors is defined in a space that is in general much bigger than the number N of compared vectors.

We wish to establish if, or to what approximation, V and W share a common subspace. The problem amounts to find new orthonormal basis vectors for V and W, $\{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_N\}$ and $\{\mathbf{w}'_1, \mathbf{w}'_2, \ldots, \mathbf{w}'_N\}$ respectively, which are ranked with decreasing mutual consistency. In principle, this could be accomplished through an iterative procedure where the first pair of vectors, $\mathbf{v}'_1$ (belonging to V) and $\mathbf{w}'_1$ (belonging to W), is picked so to have the largest possible scalar product. This optimal selection procedure is next repeated in the remaining complementary spaces of V and W and so on. The sought pairs of vectors $\mathbf{v}'_i$ and $\mathbf{w}'_i$ are such to make stationary the following functional:

$$f(\mathbf{v}'_i, \mathbf{w}'_i) = \langle w'_i | v'_i \rangle - \alpha_i \langle v'_i | v'_i \rangle - \beta_i \langle w'_i | w'_i \rangle \tag{A.2}$$

Coefficients α_i and β_i have been introduced to enforce normalization. Let $A_{i,j}$ and $B_{i,j}$ be the two N dimensional orthogonal matrices representing the change of basis: $|v'_i\rangle = \sum_{j=1}^{N} A_{i,j}|v_j\rangle$ and $|w'_i\rangle = \sum_{j=1}^{N} B_{i,j}|w_j\rangle$; and let $\vec{a_i}$ and $\vec{b_i}$ be the rows of matrices A and B respectively. Defining the non-symmetric N-dimensional real matrix C as $C_{ij} = \langle w_i | v_j \rangle$, the functional in equation (A.2) can be rewritten as:

$$f(\vec{a_i}, \vec{b_i}) \quad = \quad \vec{b_i} \cdot C\vec{a_i} - \alpha\, \vec{a_i} \cdot \vec{a_i} - \beta\, \vec{b_i} \cdot \vec{b_i}\,. \tag{A.3}$$

The stationary condition gives the following set of eigenvalue equations:

$$C^T C\ \vec{a_i} = \lambda_i \vec{a_i} \tag{A.4}$$

$$CC^T\ \vec{b_i} = \lambda_i \vec{b_i} \tag{A.5}$$

with $i = 1, \ldots N$, $\vec{a_i}$ and $\vec{b_i}$ are vectors with unit norm, and the coefficient λ_i equals $4\alpha_i\beta_i$. It's important to note that the two solutions are not independent. Assuming we have a solution $\vec{a_i}$ for (A.4); then it's easy to see that $\vec{b_i} = \frac{1}{\sqrt{\lambda_i}} C\vec{a_i}$ is a solution for (A.5), and the scalar product of the vectors v'_i and w'_i associated to this solution is $\langle w'_i | v'_i \rangle = \sqrt{\lambda_i}$. As $C^T C$ is an NxN symmetric matrix, we have a complete solution to the eigenproblem of equation (A.4). Let's consider the non-degenerate case with $\lambda_i \neq \lambda_j\, \forall\, i \neq j$ and order the eigenvalues in descending order $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Vectors $\mathbf{v}'_i$ and $\mathbf{w}'_i$ are defined by the i-th solution of (A.4), as follows:

$$|v'_i\rangle = \sum_{j=1}^{N} A_{i,j}|v_j\rangle \qquad |w'_i\rangle = \sum_{j=1}^{N} B_{i,j}|w_j\rangle \tag{A.6}$$

and their scalar product is $\sqrt{\lambda_i}$. Notice also that $\langle w'_i | v'_j \rangle = \sqrt{\lambda_i}\delta_{ij}$ in case of no degeneration in solutions of (A.4).

References

Adén, J. & Wolf-Watz, M. (2007). NMR identification of transient complexes critical to adenylate kinase catalysis. *J Am Chem Soc*, **129**, 14003–12. 45

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D. & Gerstein, M. (2005). Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. *Protein Sci*, **14**, 633–43. 70

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402. 1, 69, 72

Amadei, A., Linssen, A.B. & Berendsen, H.J. (1993). Essential dynamics of proteins. *Proteins*, **17**, 412–25. 13, 15

Amadei, A., Ceruso, M.A. & Di Nola, A. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins*, **36**, 419–24. 38, 40, 43, 75, 90

Andreeva, A. & Murzin, A.G. (2006). Evolution of protein fold in the presence of functional constraints. *Curr Opin Struct Biol*, **16**, 399–408. 1, 69, 97

Arkin, M., Randal, M., DeLano, W., Hyde, J., Luong, T., Oslob, J., Raphael, D., Taylor, L., Wang, J., McDowell, R., Wells, J. & Braisted, A. (2003). Binding of small molecules to an adaptive protein-protein interface. *P Natl Acad Sci Usa*, **100**, 1603–1608. 47

Arora, K. & Brooks, C.L. (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc Natl Acad Sci USA*, **104**, 18496–501. 32

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, **80**, 505–15. 3, 26, 28, 29, 56, 70, 74, 102

Ausiello, G., Peluso, D., Via, A. & Helmer-Citterich, M. (2007). Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics*, **8 Suppl 1**, S24. 70

Baca, M. & Kent, S.B. (1993). Catalytic contribution of flap-substrate hydrogen bonds in "hiv-1 protease" explored by chemical synthesis. *P Natl Acad Sci Usa*, **90**, 11638–42. 80

Bahar, I., Atilgan, A.R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & design*, **2**, 173–81. 3, 26, 28, 29, 56, 70, 74, 102

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–24. 111

Banavar, J.R., Maritan, A., Micheletti, C. & Trovato, A. (2002). Geometry and physics of proteins. *Proteins*, **47**, 315–22. 1, 69

Banfi, S., Servadio, A., Chung, M.Y., Kwiatkowski, T.J., McCall, A.E., Duvick, L.A., Shen, Y., Roth, E.J., Orr, H.T. & Zoghbi, H.Y. (1994). Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat Genet*, **7**, 513–20. 100

REFERENCES

Bartlett, G.J., Borkakoti, N. & Thornton, J.M. (2003). Catalysing new reactions during evolution: economy of residues and mechanism. *Journal of Molecular Biology*, **331**, 829–60. 93

Beach, H., Cole, R., Gill, M.L. & Loria, J.P. (2005). Conservation of mus-ms enzyme motions in the apo- and substrate-mimicked state. *J Am Chem Soc*, **127**, 9167–76. 32

Berasi, S.P., Xiu, M., Yee, A.S. & Paulson, K.E. (2004). HBP1 repression of the p47phox gene: cell cycle regulation via the NADPH oxidase. *Mol Cell Biol*, **24**, 3011–24. 100

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The protein data bank. *Nucleic Acids Res*, **28**, 235–42. 49

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977a). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542. 34

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977b). The protein data bank. a computer-based archival file for macromolecular structures. *Eur J Biochem*, **80**, 319–24. 84

Blundell, T.L. & Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc Natl Acad Sci USA*, **93**, 14243–8. 80

Bogan, A. & Thorn, K. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol*, **280**, 1–9. 47

Bolton, W. & Perutz, M.F. (1970). Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 angstrom units resolution. *Nature*, **228**, 551–2. 7

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci*, **2**, 31–40. 70

Brooks, B. & Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA*, **80**, 6571–5. 9, 12

Brooks, B. & Karplus, M. (1985). Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc Natl Acad Sci USA*, **82**, 4995–9. 12

Brooks, B., Janezic, D. & Karplus, M. (1995). Harmonic analysis of large systems. I. methodology. *J Comput Chem*, **16**, 1522–1542. 17

Capozzi, F., Luchinat, C., Micheletti, C. & Pontiggia, F. (2007). Essential dynamics of helices provide a functional classification of EF-hand proteins. *J Proteome Res*, **6**, 4245–55. 3, 71, 82, 96, 99

Carnevale, V., Raugei, S., Micheletti, C. & Carloni, P. (2006). Convergent dynamics in the protease enzymatic superfamily. *J Am Chem Soc*, **128**, 9766–72. 3, 59, 70, 71, 74, 80, 82, 96, 99

Carnevale, V., Pontiggia, F. & Micheletti, C. (2007a). Structural and dynamical alignment of enzymes with partial structural similarity. *Journal of Physics*. 74

Carnevale, V., Raugei, S., Micheletti, C. & Carloni, P. (2007b). Large-scale motions and electrostatic properties of furin and HIV-1 protease. *The journal of physical chemistry A*, **111**, 12327–32. 30, 45

Cascella, M., Micheletti, C., Rothlisberger, U. & Carloni, P. (2005). Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J Am Chem Soc*, **127**, 3734–42. 30, 45, 80

Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343. 47

Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy. *Reviews of Modern Physics*. 21, 22, 23

Chen, L., DeVries, A. & Cheng, C. (1997). Convergent evolution of antifreeze glycoproteins in antarctic notothenioid fish and arctic cod. *Proc Natl Acad Sci USA*, **94**, 3817–3822. 1, 69

Chen, Y.W., Allen, M.D., Veprintsev, D.B., Löwe, J. & Bycroft, M. (2004). The structure of the AXH domain of spinocerebellar ataxin-1. *J Biol Chem*, **279**, 3758–65. 100

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, **31**, 3497–500. 1, 69, 72, 83

Chennubhotla, C. & Bahar, I. (2007). Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol*, **3**, 1716–26. 32, 35

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, **357**, 543–4. 97

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823–6. 1, 69

Chothia, C., Gough, J., Vogel, C. & Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science*, **300**, 1701–3. 1, 69

Chu, J.W. & Voth, G.A. (2007). Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophysical Journal*, **93**, 3860–71. 32

Conte, L.L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, **285**, 2177–98. 47, 54

Daniel, R.M., Dunn, R.V., Finney, J.L. & Smith, J.C. (2003). The role of dynamics in enzyme activity. *Annual review of biophysics and biomolecular structure*, **32**, 69–92. 67

de Chiara, C., Giannini, C., Adinolfi, S., de Boer, J., Guida, S., Ramos, A., Jodice, C., Kioussis, D. & Pastore, A. (2003). The AXH module: an independently folded domain common to ataxin-1 and HBP1. *FEBS Lett*, **551**, 107–12. 100

de Chiara, C., Menon, R.P., Adinolfi, S., de Boer, J., Ktistaki, E., Kelly, G., Calder, L., Kioussis, D. & Pastore, A. (2005). The AXH domain adopts alternative folds the solution structure of HBP1 AXH. *Structure*, **13**, 743–53. 100, 102, 113, 118, 121

De Los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F. & Juanico, B. (2005). Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophysical Journal*, **89**, 14–21. 30, 45, 70

Delarue, M. & Sanejouand, Y.H. (2002). Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *Journal of Molecular Biology*, **320**, 1011–24. 3, 26, 28, 29, 56, 70, 74, 102

Denton, M. & Marshall, C. (2001a). Laws of form revisited. *Nature*, **410**, 417–417. 1, 69

Denton, M. & Marshall, C. (2001b). Protein folds: laws of form revisited. *Nature*, **410**, 417. 97

Divne, C., Ståhlberg, J., Teeri, T.T. & Jones, T.A. (1998). High-resolution crystal structures reveal how a cellulose chain is bound in the 50 a long tunnel of cellobiohydrolase i from trichoderma reesei. *Journal of Molecular Biology*, **275**, 309–25. 94

Doi, M. (1996). *Introduction to Polymer Physics*. Oxford University Press, Usa. 21

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E. & Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438**, 117–21. 32

Elber, R. & Karplus, M. (1987). Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, **235**, 318–21. 17

Falke, J.J. (2002). Enzymology. a moving story. *Science*, **295**, 1480–1. 70

REFERENCES

Fermi, G., Perutz, M.F., Shaanan, B. & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 a resolution. *Journal of Molecular Biology*, **175**, 159–74. 7

Fitzgerald, P.M. & Springer, J.P. (1991). Structure and function of retroviral proteases. *Annual review of biophysics and biophysical chemistry*, **20**, 299–320. 80

Frauenfelder, H., Parak, F. & Young, R.D. (1988). Conformational substates in proteins. *Annual review of biophysics and biophysical chemistry*, **17**, 451–79. 7, 17

Frauenfelder, H., Sligar, S.G. & Wolynes, P.G. (1991). The energy landscapes and motions of proteins. *Science*, **254**, 1598–603. 2, 99

Garcia, A. (1992). Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, **68**, 2696–2699. 13, 15, 16

Garcia, A. & Harman, J. (1996). Simulations of CRP:(cAMP)(2) in noncrystalline environments show a subunit transition from the open to the closed conformation. *Protein Sci*, **5**, 62–71. 13, 16

Gerstein, M. & Krebs, W. (1998). A database of macromolecular motions. *Nucleic Acids Res*, **26**, 4280–90. 2

Go, N., Noguti, T. & Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci USA*, **80**, 3696–700. 9

Gupta, S., Mangel, W.F., McGrath, W.J., Perek, J.L., Lee, D.W., Takamoto, K. & Chance, M.R. (2004). DNA binding provides a molecular strap activating the adenovirus proteinase. *Mol Cell Proteomics*, **3**, 950–9. 95

Halle, B. (2002). Flexibility and packing in proteins. *Proc Natl Acad Sci USA*, **99**, 1274–9. 71

Han, Y., Li, X. & Pan, X. (2002). Native states of adenylate kinase are two active sub-ensembles. *FEBS Lett*, **528**, 161–5. 32

Hanson, J.A., Duderstadt, K., Watkins, L.P., Bhattacharyya, S., Brokaw, J., Chu, J.W. & Yang, H. (2007). Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc Natl Acad Sci USA*, **104**, 18055–60. 32

Henzler-Wildman, K. & Kern, D. (2007). Dynamic personalities of proteins. *Nature*, **450**, 964–72. 2, 7

Henzler-Wildman, K.A., Lei, M., Thai, V., Kerns, S.J., Karplus, M. & Kern, D. (2007a). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, **450**, 913–6. 2, 8

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., Hübner, C.G. & Kern, D. (2007b). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**, 838–44. 2, 8, 32, 59, 123

Hess, B. (2000). Similarities between principal components of protein dynamics and random diffusion. *Physical review E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, **62**, 8438–48. 16

Higgins, D.G. & Sharp, P.M. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–44. 1, 69

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–29. 3, 26, 28, 29, 56, 70, 74, 102

Hinsen, K., Petrescu, A., Dellerue, S., Bellissent-Funel, M. & Kneller, G.R. (2000). Harmonicity in slow protein dynamics. *Chemical Physics*, **261**, 25–37. 21, 22, 24, 25, 59, 74

Holm, L. & Park, J. (2000). Dalilite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–7. 1, 69, 76, 88, 94

Holm, L. & Sander, C. (1994). The fssp database of structurally aligned protein fold families. *Nucleic Acids Res*, **22**, 3600–9. 1, 69

Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603. 1, 69, 72, 88

Holm, L. & Sander, C. (1999). Protein folds and families: sequence and structure alignments. *Nucleic Acids Res*, **27**, 244–7. 1, 69

Hong, L., Zhang, X.C., Hartsuck, J.A. & Tang, J. (2000). Crystal structure of an in vivo hiv-1 protease mutant in complex with saquinavir: insights into the mechanisms of drug resistance. *Protein Sci*, **9**, 1898–904. 80

Hubbard, S.J. & Thornton, J.M. (1993). Naccess (department of biochemistry and molecular biology, university college, london). 54, 60, 107

Humphrey, W., Dalke, A. & Schulten, K. (1996a). Vmd - visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38. 33

Humphrey, W., Dalke, A. & Schulten, K. (1996b). Vmd: visual molecular dynamics. *Journal of molecular graphics*, **14**, 33–8, 27–8. 73

Hwang, S., Gou, Z. & Kuznetsov, I.B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–6. 113, 114

Janezic, D., Venable, R. & Brooks, B. (1995). Harmonic analysis of large systems. III. comparison with molecular dynamics. *J Comput Chem*, **16**, 1554–1566. 17

Jones, S. & Thornton, J.M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, **93**, 13–20. 47

Jones, S. & Thornton, J.M. (1997). Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, **272**, 121–32. 54, 60

Jones, S., Shanahan, H.P., Berman, H.M. & Thornton, J.M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res*, **31**, 7189–98. 107

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–637. 55

Karplus, M. & Kushick, J. (1981). Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **14**, 325–332. 13

Katchalskikatzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C. & Vakser, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *P Natl Acad Sci Usa*, **89**, 2195–2199. 47

Kern, D., Eisenmesser, E.Z. & Wolf-Watz, M. (2005). Enzyme dynamics during catalysis measured by nmr spectroscopy. *Meth Enzymol*, **394**, 507–24. 32

Keskin, O. & Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng Des Sel*, **18**, 11–24. 50

Keskin, O. & Nussinov, R. (2007). Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**, 341–54. 50

Keskin, O., Bahar, I., Badretdinov, A., Ptitsyn, O. & Jernigan, R. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci*, **7**, 2578–2586. 47

Keskin, O., Jernigan, R.L. & Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophysical Journal*, **78**, 2093–106. 87

Keskin, O., Tsai, C., Wolfson, H. & Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, **13**, 1043–1055. 48, 49, 50

REFERENCES

Kitao, A. & Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol*, **9**, 164–9. 13, 16

Kitao, A., Hirata, F. & Go, N. (1991). The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical Physics*, **158**, 447–472. 13

Kitao, A., Hayward, S. & Go, N. (1998). Energy landscape of a native protein: jumping-among-minima model. *Proteins*, **33**, 496–517. 18

Kneller, G. & Hinsen, K. (2001). Computing memory functions from molecular dynamics simulations. *J Chem Phys*, **115**, 11097–11105. 26

Kneller, G. & Hinsen, K. (2004). Fractional brownian dynamics in proteins. *J Chem Phys*, **121**, 10278–10283. 26

Kneller, G.R. (2000). Inelastic neutron scattering from damped collective vibrations of macromolecules. *Chemical Physics*, **261**, 1–24. 24

Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. & Lesk, A.M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–74. 1, 69, 72

Kou, S. & Xie, X. (2004). Generalized langevin equation with fractional gaussian noise: Subdiffusion within a single protein molecule. *Physical Review Letters*, **93**, 180603. 26

Krishna, S.S. & Grishin, N.V. (2004). Structurally analogous proteins do exist! *Structure*, **12**, 1125–7. 1, 69

Kubo, R. (1967). *Statistical Mechanics*. North-Holland Amsterdam. 13

Lamm, G. & Szabo, A. (1986). Langevin modes of macromolecules. *J Chem Phys*, **85**, 7334–7348. 24

Lavender, P., Vandel, L., Bannister, A.J. & Kouzarides, T. (1997). The HMG-box transcription factor HBP1 is targeted by the pocket proteins and E1A. *Oncogene*, **14**, 2721–8. 100

Lesage, F., Hugnot, J.P., Amri, E.Z., Grimaldi, P., Barhanin, J. & Lazdunski, M. (1994). Expression cloning in K+ transport defective yeast and distribution of HBP1, a new putative HMG transcriptional regulator. *Nucleic Acids Res*, **22**, 3685–8. 100

Lesk, A.M. (2004). *Introduction to Protein Science: Architecture, Function and Genomics*. Oxford University Press, UK. 72

Lesk, A.M. & Fordham, W.D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *Journal of Molecular Biology*, **258**, 501–37. 97

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci USA*, **95**, 5913–20. 78, 87

Levitt, M., Sander, C. & Stern, P.S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, **181**, 423–47. 9, 12

Levy, R., Srinivasan, A.R., Olson, W.K. & McCammon, J.A. (1984). Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers*, **23**, 1099–112. 13, 15

Lou, H. & Cukier, R.I. (2006). Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations. *The journal of physical chemistry B*, **110**, 24121–37. 32

Lupas, A.N., Ponting, C.P. & Russell, R.B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*, **134**, 191–203. 97

Ma, B., Wolfson, H. & Nussinov, R. (2001). Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr Opin Struc Biol*, **11**, 364–369. 47

Maguid, S., Fernandez-Alberti, S., Parisi, G. & Echave, J. (2006). Evolutionary conservation of protein backbone flexibility. *J Mol Evol*, **63**, 448–57. 93

Maragakis, P. & Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *Journal of Molecular Biology*, **352**, 807–22. 32, 35

Micheletti, C. & Orland, H. (2009). Mistral: a tool for energy-based multiple structural alignment of proteins. *Bioinformatics*, **25**, 2663–9. 1, 69

Micheletti, C., Banavar, J.R. & Maritan, A. (2001). Conformations of proteins in equilibrium. *Physical Review Letters*, **87**, 088102. 3, 26, 29

Micheletti, C., Lattanzi, G. & Maritan, A. (2002). Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. *Journal of Molecular Biology*, **321**, 909–21. 3, 26, 29

Micheletti, C., Carloni, P. & Maritan, A. (2004). Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins*, **55**, 635–45. 3, 26, 28, 29, 30, 45, 56, 70, 74, 102

Min, W., Luo, G., Cherayil, B., Kou, S. & Xie, X. (2005). Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Physical Review Letters*, **94**, 198302. 26

Ming, D. & Wall, M.E. (2005). Allostery in a coarse-grained model of protein dynamics. *Physical Review Letters*, **95**, 198103. 70

Miyashita, O., Onuchic, J.N. & Wolynes, P.G. (2003). Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA*, **100**, 12570–5. 32, 35

Miyashita, O., Wolynes, P.G. & Onuchic, J.N. (2005). Simple energy landscape model for the kinetics of functional transitions in proteins. *The journal of physical chemistry B*, **109**, 1959–69. 35

Mol, C.D., Kuo, C.F., Thayer, M.M., Cunningham, R.P. & Tainer, J.A. (1995). Structure and function of the multifunctional dna-repair enzyme exonuclease iii. *Nature*, **374**, 381–6. 94, 95

Müller, C.W. & Schulz, G.E. (1992). Structure of the complex between adenylate kinase from escherichia coli and the inhibitor ap5a refined at 1.9 a resolution. a model for a catalytic transition state. *Journal of Molecular Biology*, **224**, 159–77. 31, 32, 33

Müller, C.W., Schlauderer, G.J., Reinstein, J. & Schulz, G.E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, **4**, 147–56. 31, 32, 33

Murzin, A.G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J*, **12**, 861–7. 100

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**, 536–40. 1, 50, 69, 94

Mushegian, A.R., Bassett, D.E., Boguski, M.S., Bork, P. & Koonin, E.V. (1997). Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc Natl Acad Sci USA*, **94**, 5831–6. 100

Neri, M., Cascella, M. & Micheletti, C. (2005). The influence of conformational fluctuations on enzymatic activity: modelling the functional motion of beta-secretase. *J Phys-Condens Mat*, **17**, S1581–S1593. 80

Nooren, I.M.A. & Thornton, J.M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology*, **325**, 991–1018. 47

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–17. 1, 69, 72

REFERENCES

Nussinov, R. & Wolfson, H.J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA*, **88**, 10495–9. 49

Orengo, C.A. & Thornton, J.M. (2005). Protein families and their evolution-a structural perspective. *Annu Rev Biochem*, **74**, 867–900. 1, 69

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–108. 1, 69, 83, 84

Orr, H.T. & Zoghbi, H.Y. (2001). SCA1 molecular genetics: a history of a 13 year collaboration against glutamines. *Hum Mol Genet*, **10**, 2307–11. 100

Park, B. & Levitt, M. (1996). Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *Journal of Molecular Biology*, **258**, 367–92. 28

Paulson, K.E., Rieger-Christ, K., McDevitt, M.A., Kuperwasser, C., Kim, J., Unanue, V.E., Zhang, X., Hu, M., Ruthazer, R., Berasi, S.P., Huang, C.Y., Giri, D., Kaufman, S., Dugan, J.M., Blum, J., Netto, G., Wazer, D.E., Summerhayes, I.C. & Yee, A.S. (2007). Alterations of the HBP1 transcriptional repressor are associated with invasive breast cancer. *Cancer Res*, **67**, 6136–45. 100

Perutz, M.F. & Mathews, F.S. (1966). An x-ray study of azide methaemoglobin. *Journal of Molecular Biology*, **21**, 199–202. 2, 7

Pidugu, L.S., Kapoor, M., Surolia, N., Surolia, A. & Suguna, K. (2004). Structural basis for the variation in triclosan affinity to enoyl reductases. *Journal of Molecular Biology*, **343**, 147–55. 94

Pontiggia, F., Colombo, G., Micheletti, C. & Orland, H. (2007). Anharmonicity and self-similarity of the free energy landscape of protein g. *Physical Review Letters*, **98**, 048102. 2, 18, 23, 43, 45

Pontiggia, F., Zen, A. & Micheletti, C. (2008). Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophysical Journal*, **95**, 5901–12. 2, 18, 32, 34, 43, 123

Porter, C.T., Bartlett, G.J. & Thornton, J.M. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, **32**, D129–33. 82, 84, 94, 95

Rajamani, D., Thiel, S., Vajda, S. & Camacho, C.J. (2004). Anchor residues in protein-protein interactions. *Proc Natl Acad Sci USA*, **101**, 11287–92. 62

Redfern, O.C., Dessailly, B. & Orengo, C.A. (2008). Exploring the structure and function paradigm. *Curr Opin Struct Biol*, **18**, 394–402. 70

Risken, H. (1996). *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer. 23, 24

Rod, T.H., Radkiewicz, J.L. & Brooks, C.L. (2003). Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc Natl Acad Sci USA*, **100**, 6980–5. 70

Rose, G.D., Fleming, P.J., Banavar, J.R. & Maritan, A. (2006). A backbone-based theory of protein folding. *Proc Natl Acad Sci USA*, **103**, 16623–33. 97

Roujeinikova, A., Sedelnikova, S., de Boer, J., Stuitje, A.R., Slabas, A.R., Rafferty, J.B. & Rice, D.W. (1999). Inhibitor binding studies on enoyl reductase reveal conformational changes related to substrate recognition. *J Biol Chem*, **274**, 30811–7. 94

Russell, R.B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, **279**, 1211–27. 70

Sacquin-Mora, S., Laforet, E. & Lavery, R. (2007). Locating the active sites of enzymes using mechanical properties. *Proteins*, **67**, 350–9. 93

Sadowski, M.I. & Jones, D.T. (2009). The sequence-structure relationship and protein function prediction. *Curr Opin Struc Biol*, **19**, 357–62. 70

Scheeff, E.D. & Bourne, P.E. (2005). Structural evolution of the protein kinase-like superfamily. *PLoS Comput Biol*, **1**, e49. 97

Selvin, P.R. (2000). The renaissance of fluorescence resonance energy transfer. *Nat Struct Biol*, **7**, 730–4. 8

Seno, F. & Trovato, A. (2007). Minireview: The compact phase in polymers and proteins. *Physica A*, **384**, 122–127. 1, 69

Shapiro, Y.E. & Meirovitch, E. (2006). Activation energy of catalysis-related domain motion in e. coli adenylate kinase. *The journal of physical chemistry B*, **110**, 11519–24. 32

Shapiro, Y.E., Sinev, M.A., Sineva, E.V., Tugarinov, V. & Meirovitch, E. (2000). Backbone dynamics of escherichia coli adenylate kinase at the extreme stages of the catalytic cycle studied by (15)n nmr relaxation. *Biochemistry*, **39**, 6634–44. 32

Shapiro, Y.E., Kahana, E., Tugarinov, V., Liang, Z., Freed, J.H. & Meirovitch, E. (2002). Domain flexibility in ligand-free and inhibitor-bound escherichia coli adenylate kinase based on a mode-coupling analysis of 15n spin relaxation. *Biochemistry*, **41**, 6271–81. 32

Shatsky, M., Dror, O., Schneidman-Duhovny, D., Nussinov, R. & Wolfson, H.J. (2004a). BioInfo3D: a suite of tools for structural bioinformatics. *Nucleic Acids Res*, **32**, W503–7. 72

Shatsky, M., Nussinov, R. & Wolfson, H.J. (2004b). A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–56. 1, 69, 72

Sierk, M.L. & Pearson, W.R. (2004). Sensitivity and selectivity in protein structure comparison. *Protein Sci*, **13**, 773–85. 88

Sinev, M.A., Sineva, E.V., Ittah, V. & Haas, E. (1996). Domain closure in adenylate kinase. *Biochemistry*, **35**, 6425–37. 32

Smith, G.R., Sternberg, M.J.E. & Bates, P.A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *Journal of Molecular Biology*, **347**, 1077–101. 62, 70

Somogyi, B., Lakos, Z., Szarka, A. & Nyitrai, M. (2000). Protein flexibility as revealed by fluorescence resonance energy transfer: an extension of the method for systems with multiple labels. *J Photochem Photobiol B, Biol*, **59**, 26–32. 8

Stockwell, G.R. & Thornton, J.M. (2006). Conformational diversity of ligands bound to proteins. *Journal of Molecular Biology*, **356**, 928–44. 94

Storey, J.D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, **100**, 9440–5. 87

Sułkowska, J.I., Kloczkowski, A., Sen, T.Z., Cieplak, M. & Jernigan, R.L. (2008). Predicting the order in which contacts are broken during single molecule protein stretching experiments. *Proteins*, **71**, 45–60. 56, 70

Taylor, W.R. (2006). Decoy models for protein structure comparison score normalisation. *Journal of Molecular Biology*, **357**, 676–99. 78

Tesi, M., Vanrensburg, E., Orlandini, E. & Whittington, S. (1996). Monte carlo study of the interacting self-avoiding walk model in three dimensions. *J Stat Phys*, **82**, 155–181. 76

Tevosian, S.G., Shih, H.H., Mendelson, K.G., Sheppard, K.A., Paulson, K.E. & Yee, A.S. (1997). HBP1: a HMG box transcriptional repressor that is targeted by the retinoblastoma family. *Genes Dev*, **11**, 383–96. 100

REFERENCES

Theobald, D.L., Mitton-Fry, R.M. & Wuttke, D.S. (2003). Nucleic acid recognition by OB-fold proteins. *Annual review of biophysics and biomolecular structure*, **32**, 115–33. 100, 104

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80. 1, 69

Tidor, B. & Karplus, M. (1994). The contribution of vibrational entropy to molecular association. the dimerization of insulin. *Journal of Molecular Biology*, **238**, 405–14. 67

Tirion, M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, **77**, 1905–1908. 3, 26, 27, 29, 56

Todd, A., Orengo, C. & Thornton, J. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. *Structure*, **10**, 1435–1451. 47

Tsai, C.C., Kao, H.Y., Mitzutani, A., Banayo, E., Rajan, H., McKeown, M. & Evans, R.M. (2004). Ataxin 1, a SCA1 neurodegenerative disorder protein, is functionally linked to the silencing mediator of retinoid and thyroid hormone receptors. *Proc Natl Acad Sci USA*, **101**, 4047–52. 100

Tsai, C.J., Lin, S.L., Wolfson, H.J. & Nussinov, R. (1996). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *Journal of Molecular Biology*, **260**, 604–20. 49

Tsai, C.J., Xu, D. & Nussinov, R. (1998). Protein folding via binding and vice versa. *Folding & design*, **3**, R71–80. 47

UniProt Consortium (2008). The universal protein resource (uniprot). *Nucleic Acids Res*, **36**, D190–5. 50

UniProt Consortium (2009). The universal protein resource (uniprot) 2009. *Nucleic Acids Res*, **37**, D169–74. 50

Valdar, W. & Thornton, J. (2001). Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124. 47

Wallis, R., Leung, K., Osborne, M., James, R., Moore, G. & Kleanthous, C. (1998). Specificity in protein-protein recognition: Conserved im9 residues are the major determinants of stability in the colicin e9 dnase-tm9 complex. *Biochemistry-Us*, **37**, 476–485. 47

Wang, M.C. & Uhlenbeck, G.E. (1945). On the theory of the brownian motion ii. *Reviews of Modern Physics*, **17**, 323–342. 21, 22, 23

Whitford, P.C., Miyashita, O., Levy, Y. & Onuchic, J.N. (2007). Conformational transitions of adenylate kinase: switching by cracking. *Journal of Molecular Biology*, **366**, 1661–71. 35

Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E.Z. & Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol*, **11**, 945–9. 32

Xu, H., Aurora, R., Rose, G.D. & White, R.H. (1999). Identifying two ancient enzymes in archaea using predicted secondary structure alignment. *Nat Struct Biol*, **6**, 750–4. 97

Yan, C., Terribilini, M., Wu, F., Jernigan, R.L., Dobbs, D. & Honavar, V. (2006). Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262. 113

Yang, G., Sandalova, T., Lohman, K., Lindqvist, Y. & Rendina, A.R. (1997). Active site mutants of escherichia coli dethiobiotin synthetase: effects of mutations on enzyme catalytic and structural properties. *Biochemistry*, **36**, 4751–60. 95

Yogurtcu, O.N., Erdemli, S.B., Nussinov, R., Turkay, M. & Keskin, O. (2008). Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophysical Journal*, **94**, 3475–85. 62

Yue, S., Serra, H.G., Zoghbi, H.Y. & Orr, H.T. (2001). The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. *Hum Mol Genet*, **10**, 25–30. 100

Zen, A., Carnevale, V., Lesk, A.M. & Micheletti, C. (2008). Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Sci*, **17**, 918–29. 3, 99, 102, 123

Zen, A., de Chiara, C., Pastore, A. & Micheletti, C. (2009). Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to ob-fold domains. *Bioinformatics*, **25**, 1876–83. 3, 123

Zhang, Y., Kolinski, A. & Skolnick, J. (2003). Touchstone ii: A new approach to ab initio protein structure prediction. *Biophys J*, **85**, 1145–1164. 47

Zheng, W., Brooks, B.R. & Thirumalai, D. (2007). Allosteric transitions in the chaperonin groel are captured by a dominant normal mode that is most robust to sequence variations. *Biophysical Journal*, **93**, 2289–99. 70

Zhu, H., Domingues, F.S., Sommer, I. & Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27. 51