# Coarse-grained modelling of protein structure and internal dynamics: comparative methods and applications


A thesis submitted for the degree of

*PhilosophiæDoctor*

October 2010

Candidate
Raffaello Potestio

Supervisor
Cristian Micheletti

Statistical and Biological Physics sector

Ph.D. course in Physics and Chemistry of Biological Systems
International School for Advanced Studies - SISSA
Trieste

To Gloria

*I was born not knowing and have only had a little time to change that here and there.*

Richard Feynman

# Contents

# List of Figures

# List of Tables

# Introduction

Traditionally, the characterisation of the properties of proteins and enzymes is articulated according to the tripartite scheme *sequence* → *structure* → *function*.

The investigation of the first two elements of this scheme and the relation between them largely benefited from the use of alignment methods. In particular, sequence alignment algorithms have been widely used to identify evolutionary relationships in different proteins by detecting common stretches of the primary sequence. Similarly, structure alignment methods provided further evolutionary insight through the comparison of the architectural organisation of protein structures. The combined use of these techniques allowed to broaden our understanding of the relation between sequence and structure: a striking result concerning this relation is the fact that proteins with primary sequence similarity above 30% typically adopt the same structural organisation (1; 2; 3). In the light of this result, the question naturally rose, of whether a relevant structural similarity can be shared by proteins having low sequence identity: the use of structural superposition algorithms to compare large datasets of proteins, and the subsequent comparison of the primary sequences of the aligned structures, revealed that the same fold can occasionally be shared by proteins having markedly different sequences (4; 5; 6). This result is commonly interpreted in the light of evolutionary convergence (7; 8; 9; 10; 11; 12).

Experimental results (13), as well as computational studies (14; 15), completed the picture highlighting the impact of structural features on a protein's biological function. Specifically, while on the one hand the catalytic activity of an enzyme relies on the chemical details of the active site, the influence of the overall molecule structural architecture to carry on its biological functionality has become more and more evident in a growing number of cases. The architecture of the protein, in fact, not only determines the shape of the interface the molecule exposes to the substrates, but in many enzymes it also influences the internal dynamics properties which play a central role for the performance of the biological activity (16; 17; 18).

This role can be played in a twofold way: in the vicinity of the active site, the catalytic geometry poses tight constraints on the possible local structure of the molecule and its flexibility, which in turn determine which ligands can bind to the enzyme. On

a global scale, on the other hand, the possibility to perform concerted conformational changes is of paramount biological importance to many enzymes.

A prototypical example of the aforementioned relation between structure, internal dynamics and function is given by adenylate kinase (Adk). This 214-residue-long monomeric protein regulates the energy balance of the cell by converting AMP, ADP and ATP molecules according to the relation:

$$MgATP + AMP \Longleftrightarrow 2ADP + Mg^{2+}$$

Adk is composed by a central core and two domains, the ATP binding domain (Lid) and the AMP binding one. These domains are highly mobile: in the available 'closed' crystallographic state (PDB code: 1ake) they are displaced towards the core by more than 7 Å RMSD with respect to the 'open' crystal structure (PDB code: 4ake), as shown in Fig. 1.

The catalytic activity of Adk relies on a conformational change, bridging the two structures, which involves the collective displacement of a large number of amino acids. In fact, the ligands require to be processed in a water-free environment: the open structures largely populate the ensemble in absence of the ligands, but when the latter are present the population of conformers is shifted in favour of the closed conformations, thus excluding water molecules from the catalytic region.

The structural architecture of Adk strongly affects its internal dynamics (14; 15): the functionally-oriented fluctuations proved to be consistent among adenylate kinase molecules from a variety of organisms with low sequence similarity (13; 14), thus suggesting that the relative positioning of the secondary and tertiary structure elements shapes the collective dynamics of a protein.

This relation, which bridges the three-dimensional structure of a protein and its functionally-relevant collective dynamics, represents the main topic of the present thesis.

The functional relevance of proteins' internal dynamics was not evident at the early stage of protein science. The first hypothesis on the functioning of enzymes, formulated by Emil Fischer in 1894 and commonly known as the *lock-and-key* model (see Fig. 2), depicts these biomolecules as rigid units with a defined shape. The interaction between two enzymes was thought to depend on the complementarity in the shape of the surfaces.

The lock-and-key model represented a brilliant molecular intuition about enzyme catalysis, since no protein structure was available at the time. Nonetheless, the limitations of this model became evident when the very first crystallographic structures of globins were resolved. In fact, the unbound structure of hemoglobin (see Fig. 3) showed that the channel connecting the surface to the heme groups is too narrow to allow the oxygen transit. This observation made clear that a more accurate protein

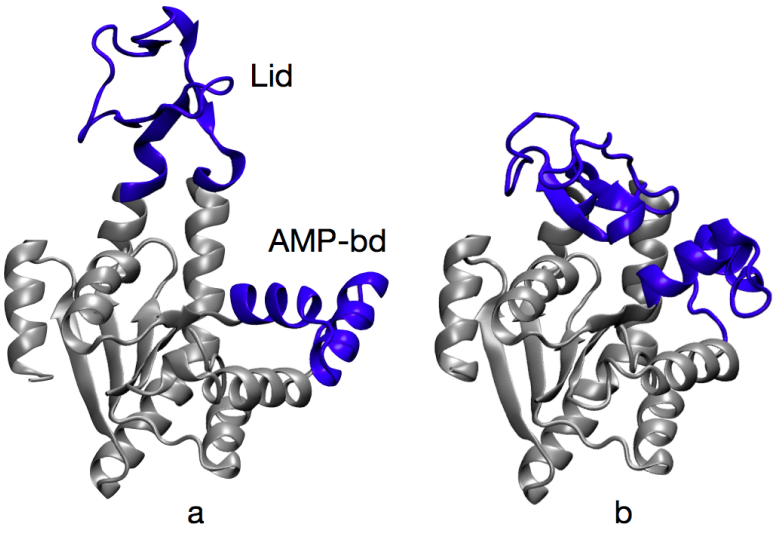


**Figure 1: Open and closed crystallographic structures of adenylate kinase** - The open (panel a) and closed (panel b) crystallographic structures of 214-residue-long *E. Coli* adenylate kinase are here shown in cartoon representation. This molecule is customarily subdivided in three domains: the core (in gray), the Lid and the AMP-binding domain (both in blue). The two crystallographic structures have a RMSD of about 7Å.

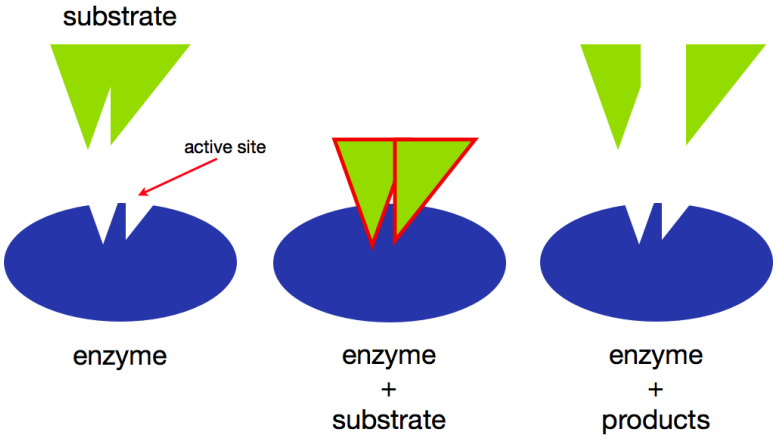


**Figure 2: Lock-and-key model of protein-ligand interaction** - The cartoon illustrates schematically the lock-and-key model of protein-ligand interaction. In the first stage, the ligand fits in a pocket of the protein having a complementary shape; the catalysis then takes place, and the products are released.

model had to include conformational plasticity and allow for conformational changes of the structure.

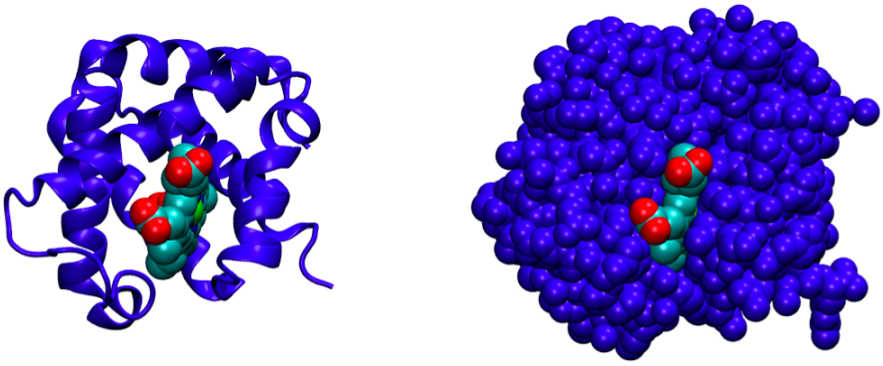

**Figure 3: Crystallographic structure of hemoglobin** - The crystallographic structure of bound-state hemoglobin (in cartoon and Van der Waals representation) is shown. The heme group is buried in a pocket the oxygen cannot reach, due to steric interactions. A conformational change is therefore needed to widen the aperture.

Consistently with the previous considerations, the apo and holo forms of several proteins and, particularly, enzymes, are found in different crystallographic conformers. This fact supports the relevance of conformational plasticity/elasticity for function. In particular, several enzymes possess, in physiological conditions, two main conformations -usually called *open* and *closed*- and interconvert between these two states to perform the catalytic activity: adenylate kinase (Fig. 1) is a typical example of a *molecular switch*. The open and closed conformers are commonly associated to *minima* or *basins* of a free energy profile in a multi-dimensional coordinate space describing the conformations of the protein: the interconversion between the two types of conformers is therefore interpreted as arising from the overcoming of the free energy barrier separating the minima.

The presence of a partner substrate (another protein or a different type of molecule) can modify the free energy profile, and in particular the relative depth of the basins and the height of the barriers. In this case, a set of functionally-active conformations can become more populated with respect to the equilibrium state, favouring the protein-ligand interaction. This picture is usually termed *induced fit model*.

The internal dynamics of adenylate kinase has been considered for a long time an example of induced fit. It was thought that the interconversion between these two conformations was triggered only by the presence of the ligand. However, as single-molecule experiments (19) have later shown, the two stable conformations of the molecule can be well populated even in absence of the substrate. This points to the predisposition of Adk's internal dynamics to bridge the open/closed conformations and to its capability

to overcome the free-energy barriers separating the two reference states, consistently with indications from atomistic molecular dynamics simulations (15; 20; 21). In this case, the protein spontaneously undergoes thermally-activated large-scale conformational changes by jumping across different free energy minima; the ligand does not trigger the structural changes, but rather tends to bind to those molecules already having an appropriate conformation. This model is called *conformational selection*.

Induced fit and conformational selection are two of the most common and studied mechanisms of protein-substrate interaction; nonetheless, all intermediate cases can occur, depending on the effective number of free energy minima associated to equilibrium structures and their relative depth, and also on the influence of the ligand on the properties of the free energy profile.

The paramount importance of internal dynamics for the biological function of a large number of proteins stimulated the development of many different strategies to analyse the flexibility properties of these molecules. Specifically, a variety of tools have been designed to investigate protein internal dynamics in terms of collective degrees of freedom, aiming at a simplified picture of the structure and dynamics; coarse-grained descriptions, in fact, can be valuable resources to understand the salient aspects of the relation between structure, internal dynamics and biological function of a protein.

A large fraction of the work I carried out during my Ph.D. has been devoted to the study of the flexibility properties of globular proteins, and their relation with their structure and biological function. In particular, I focused on the development of coarse-graining strategies to subdivide a protein in a few groups of amino acids, based on their concerted movements. The resulting simplification of the protein structure is in turn used to characterise the large-scale fluctuations of the molecule focusing on the collective properties of the motion.

The similarity of the concerted movements among structurally different proteins motivated the investigation of the possibility to ascertain nontrivial structural similarities on the basis of good dynamical consistency. Dynamics-based alignment methods have been, in this respect, the pivot of the study of the dynamics-mediated relation between structure and biological function.

As a side topic, I worked on the possible sequence and structural relations occurring among proteins in different topological states. The vast majority of the available protein structures are unknotted. Nonetheless, a small but non-negligible fraction of the available chains unambiguously show to be knotted. The use of sequence and structure alignment methods supply the appropriate framework to perform a dataset-wide comparative analysis, providing insight into the unusual properties of knotted proteins.

The material presented in this thesis is organised as follows:

The first chapter is devoted to a brief summary of the basic techniques commonly used to characterise protein's internal dynamics, and to perform those primary analyses which are the basis for our further developments. To this purpose we recall the basics of Principal Component Analysis of the covariance matrix of molecular dynamics (MD) trajectories. The overview is aimed at motivating and justifying *a posteriori* the introduction of coarse-grained models of proteins.

In the second chapter we shall discuss dynamical features shared by different conformers of a protein. We'll review previously obtained results, concerning the universality of the vibrational spectrum of globular proteins and the self-similar free energy landscape of specific molecules, namely the G-protein and Adk. Finally, a novel technique will be discussed, based on the theory of Random Matrices, to extract the robust collective coordinates in a set of protein conformers by comparison with a stochastic reference model.

The third chapter reports on an extensive investigation of protein internal dynamics modelled in terms of the relative displacement of quasi-rigid groups of amino acids. Making use of the results obtained in the previous chapters, we shall discuss the development of a strategy to optimally partition a protein in units, or *domains*, whose internal strain is negligible compared to their relative fluctuation. These partitions will be used in turn to characterise the dynamical properties of proteins in the framework of a simplified, coarse-grained, description of their motion.

In the fourth chapter we shall report on the possibility to use the collective fluctuations of proteins as a guide to recognise relationships between them that may not be captured as significant when sequence or structural alignment methods are used. We shall review a method to perform the superposition of two proteins optimising the similarity of the structures as well as the dynamical consistency of the aligned regions; then, we shall next discuss a generalisation of this scheme to accelerate the dynamics-based alignment, in the perspective of dataset-wide applications.

Finally, the fifth chapter focuses on a different topic, namely the occurrence of topologically-entangled states (knots) in proteins. Specifically, we shall investigate the sequence and structural properties of knotted proteins, reporting on an exhaustive dataset-wide comparison with unknotted ones. The correspondence, or the lack thereof, between knotted and unknotted proteins allowed us to identify, in knotted chains, small segments of the backbone whose 'virtual' excision results in an unknotted structure. These 'knot-promoting' loops are thus hypothesised to be involved in the formation of the protein knot, which in turn is likely to cover some role in the biological function of the knotted proteins.

The material presented in this thesis is largely based on the following publications:

- R. Potestio, F. Pontiggia and C. Micheletti, *Coarse-grained description of proteins' internal dynamics: an optimal strategy for decomposing proteins in rigid subunits*, Biophysical Journal **96**, June 2009, pp. 4993–5002

- T. Aleksiev, R. Potestio, F. Pontiggia, S. Cozzini and C. Micheletti, *PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains*, Bioinformatics **25(20)**, October 2009, pp. 2743-4

- R. Potestio, F. Caccioli and P. Vivo, *Random matrix approach to collective behavior and bulk universality in protein dynamics*, Physical Review Letters **103**, December 2009, p. 268101

- R. Potestio, T. Aleksiev, F. Pontiggia, S. Cozzini and C. Micheletti, *ALADYN: a web server for aligning proteins by matching their large-scale motion*, *Nucleic Acids Research - web server issue*, May 2010

- R. Potestio, C. Micheletti and H. Orland, *Knotted vs. unknotted proteins: evidence of knot-promoting loops*, *PlOS Comput. Biol.*, 2010; 6(7)

- R. Potestio, F. Pontiggia, V. Carnevale and C. Micheletti, *Bridging the atomic and coarse-grained descriptions of collective motions in proteins*, invited contribution for the book *Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies*, edited by A. Kolinski, Springer, in press

# Part I

# Coarse-grained modelling of protein structure and dynamics

# 1

# Protein internal dynamics: from all-atom simulations to coarse-grained models

## 1.1 Computational methods to investigate protein internal dynamics

*Molecular Dynamics* (MD) approaches are among the most common computational strategies used to characterise the thermodynamics and kinetics of proteins. These methods have experienced an outstanding growth, allowed by the availability of fast computers and reliable and free software packages. The transferable formulation of the force-fields permits to simulate many different classes of biomolecules for constantly increasing time spans, gaining insight in their physical and chemical properties with unprecedented detail.

A Molecular Dynamics simulation consists in the numerical integration of the equations of motion of a molecular system, which can be performed with different degrees of structural detail and interaction force fields. The simulations can be *quantum* or *classical*; the latter scheme, in particular the atomistic approach, had a deep development in the last decades.

In atomistic MD all atoms of the molecules are taken into account; the interactions among them are reproduced with empiric force fields which incorporate Van der Waals forces, dihedral penalties, screened electrostatics etc. The time evolution can be performed integrating the Newton equations of motion (constant energy simulation of the NVE ensemble); nonetheless, it is usually preferred to introduce some degree of stochasticity making use of *thermostats* to keep the temperature constant (NVT ensem-

ble). In a present-day MD simulation the constant temperature dynamical evolution of a protein can be followed for time scales ranging from the hundredth of ns to the millisecond.

Solvent molecules can be accounted for explicitly (with one or more atoms per water molecule) or implicitly, introducing an effective interaction to take into account its effects; a very simple and used scheme to perform implicit-solvent simulations is given by *Brownian dynamics*, in which each atom of the system is subject to a stochastic force, mimicking the resultant of the solvent molecules' random impacts, and to a velocity-dependent friction force absorbing the excess momentum, consistently with the fluctuation-dissipation theorem.

A single MD trajectory produces a considerable amount of information, since the full atomistic detail of the molecule is taken into account. A powerful yet standard way to extract from this wealth of information the few collective degrees of freedom that most account for the molecule's structural fluctuations is the calculation and analysis of the *covariance matrix*.

The first step of this procedure consists in the alignment of all the frames of the simulation: in fact, during the time evolution, unless properly constrained the simulated molecule performs a diffusive motion in space. This motion, consisting of a roto-translation of the protein, introduces a spurious mobility which must be removed; it is worth to highlight, nonetheless, that this roto-translational motion can be unambiguously removed only for a true rigid body.

Once all the frames are aligned, it is possible to identify a *reference structure* $\vec{r}^0$, for example the average structure or, more properly, the instantaneous structure closest to it. The covariance matrix is then given by the following expression:

$$\mathcal{C}_{ij,\mu\nu} = \langle (r_{i,\mu} - r^0_{i,\mu})(r_{j,\nu} - r^0_{j,\nu}) \rangle \equiv \langle \delta r^\mu_i \delta r^\nu_j \rangle \tag{1.1}$$

where $r_{i,\mu}$ is the $\mu$-th cartesian component of the $i$-th amino acid position, and the angular brackets indicate the time average.

A widely used technique to analyse the covariance matrix is the *principal component analysis* (PCA) (22; 23), consisting in the diagonalisation of the matrix and the study of its modes of fluctuation. The latter satisfy the equation:

$$\mathcal{C}\vec{v}_\ell = \lambda_\ell \vec{v}_\ell \tag{1.2}$$

where $\vec{v}_\ell$ are the eigenvectors of the covariance matrix and $\lambda_\ell$ are the corresponding eigenvalues. The mean square fluctuation of the protein, given by MSF $= \sum_i \langle |\vec{r}_i - \vec{r}^0_i|^2 \rangle$ is by definition the *trace* of the covariance matrix; therefore, an equivalent way to obtain the MSF is to sum the covariance eigenvalues:

$$\mathrm{MSF} = \sum_i \langle |\vec{r}_i - \vec{r}_i^{\,0}|^2 \rangle \equiv \sum_\ell \lambda_\ell \tag{1.3}$$

The modes having the largest covariance eigenvalues describe those degrees of freedom which best capture the large-scale conformational changes. It has been shown (23) that a few modes of the covariance are sufficient to take into account a large fraction of the protein mobility. The small subspace of these vectors, usually called *essential space*, is therefore the starting point of many analyses aimed at characterising the molecule's flexibility and internal dynamics.

A possible way to characterise the motion of the molecule in terms of the eigenmodes of the covariance is to project onto these vectors the instantaneous displacement:

$$p^\ell(t) = \delta\vec{r}(t) \cdot \vec{v}^\ell \tag{1.4}$$

where $\delta\vec{r}(t) \equiv \vec{r}(t) - \vec{r}^{\,0}$, and $\vec{v}^\ell$ is the $\ell$-th vector of the covariance matrix. Go (24) showed that the histograms of the collective variables $p^\ell(t)$ are characterised by a unimodal distribution for time spans of the order of the ns; this property is progressively lost as longer time intervals are considered. As an example, Fig. 1.1 shows the normalised distribution of the projection, on the first mode of the covariance, of increasingly long time intervals of a 50-ns long simulation of adenylate kinase performed by Pontiggia *et al.* (15).

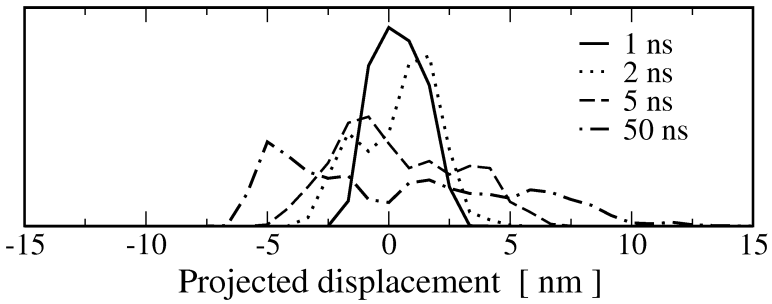


**Figure 1.1: Normalised distribution of a MD trajectory of Adk projected on the first mode of the covariance** - The unimodal character of the distributions, which is preserved for time spans up to 10 ns, is progressively lost in favour of broad, multimodal distributions as the time interval is increased. Picture taken from ref. (25).

In the Introduction we mentioned the tight relation between collective conformational changes and the functional-oriented movements performed by many proteins. The collective modes of the covariance matrix play a central role in understanding

a protein's internal dynamics and its biological function. Unfortunately, in order to obtain reliable information from the covariance matrix it is required that the MD simulation has sampled a sufficient fraction of the relevant phase space; this condition can be difficult to realise due to the important computational effort required by simulations covering a long enough time span and/or involving large proteins. This limitation can be overcome making use of coarse-grained models of protein structure and interaction.

## 1.2 Coarse-grained models of proteins

The collective character of the conformational changes occurring in many proteins opens *a priori* the possibility that the functional-oriented internal dynamics can be captured by simplified models of the structure and interactions. In the following, we shall give a brief account of the methods commonly employed to characterise the low-frequency and low-energy fluctuations of proteins under the assumptions of local deviations from a reference structure.

### 1.2.1 Normal Modes Analysis

The short-time internal dynamics of proteins is characterised by high-frequency vibrations in local minima of the energy (26). These minima describe small-scale deformations of the molecule, such as rotational motions of the side-chains about dihedral angles; the energetic barriers separating two minima are very small (below $K_BT$) and are easily overcome by the molecule, which moves from a minimum to a nearby one on a timescale of the order of the ps (26). The effective dimensionality of the configurational space, in which these minima are embedded, is so large that the probability to visit twice the same minimum is negligible; the protein is therefore assumed to perform a diffusive motion in this space.

For a given reference structure, the usual way to obtain the local vibrational spectrum is the following. First, the energy is minimised with a steepest-descent procedure in order to bring the original configuration to a new one in the minimum of the energy. This practice is required due to the fact that crystallographic structures do not correspond to minima of the empiric force field used to parametrize the molecule internal energy; moreover, they might also be affected by crystal packing effects resulting, for example, in unnaturally stretched bonds.

The energy $E$ is next expanded in terms of the displacements from the minimum structure $\vec{r}^0$, according to the Taylor formula:

$$
\begin{aligned}
E(\vec{r}) &= E(\vec{r}^0) + \sum_{i,\mu} \left.\frac{\partial E}{\partial r_{i,\mu}}\right|_{\vec{r}^0} (r_{i,\mu} - r^0_{i,\mu}) \\
&+ \frac{1}{2}\sum_{i,\mu}\sum_{j,\nu} \left.\frac{\partial^2 E}{\partial r_{i,\mu}\partial r_{j,\nu}}\right|_{\vec{r}^0} (r_{i,\mu} - r^0_{i,\mu})(r_{j,\nu} - r^0_{j,\nu}) + \mathcal{O}(\vec{r} - \vec{r}^0)^3
\end{aligned}
\tag{1.5}
$$

The constant term $E(\vec{r}^0)$ can be neglected since it simply represents a shift of the energy level; on the other hand, the extremality condition requires the first derivatives of the energy to vanish at $\vec{r}^0$. Therefore, the first non-zero contribution to the energy in the neighbourhood of the minimum is given by the quadratic term:

$$
E(\vec{r}) \sim \frac{1}{2}\sum_{i,\mu}\sum_{j,\nu} \left.\frac{\partial^2 E}{\partial r_{i,\mu}\partial r_{j,\nu}}\right|_{\vec{r}^0} \delta r_i^\mu \; \delta r_j^\nu \tag{1.6}
$$

where, as customary, $\delta r_i^\mu \equiv r_{i,\mu} - r^0_{i,\mu}$. In the most widely used force fields the energy is a sum of pairwise interactions between atoms, and the single terms depend only on the *distance* separating two atoms. In particular, in the neighbourhood of the minimum, the reference structure allows to consider the variations of the distance with respect to the rest configuration: without any loss of generality, we shall assume that the pairwise energy depends on the absolute value of the distance variation. Therefore one has:

$$
\begin{aligned}
E(\vec{r}) &= \sum_{ij} E_{ij}(|d_{ij} - d^0_{ij}|) \\
&\sim \frac{1}{2}\sum_{i,\mu}\sum_{j,\nu} \left.\frac{\partial^2 E_{ij}(x)}{\partial x^2}\right|_{x=0} \frac{\partial |d_{ij} - d^0_{ij}|}{\partial \delta r_i^\mu} \; \frac{\partial |d_{ij} - d^0_{ij}|}{\partial \delta r_j^\nu} \delta r_i^\mu \; \delta r_j^\nu \\
&= \frac{1}{2}\sum_{i,\mu}\sum_{j,\nu} C_{ij} \; \mathcal{M}_{ij}^{\mu\nu} \delta r_i^\mu \; \delta r_j^\nu = \frac{1}{2}\sum_{i,\mu}\sum_{j,\nu} \delta r_i^\mu \; H_{i,j}^{\mu,\nu} \; \delta r_j^\nu
\end{aligned}
\tag{1.7}
$$

The following shorthand notations have been used:

$$
\begin{aligned}
&\left.\frac{\partial^2 E_{ij}(x)}{\partial x^2}\right|_{x=0} \equiv C_{ij} \\
&\frac{\partial |d_{ij} - d^0_{ij}|}{\partial \delta r_i^\mu} \; \frac{\partial |d_{ij} - d^0_{ij}|}{\partial \delta r_j^\nu} = (2\delta_{ij} - 1)\hat{d}^{0\mu}_{ij}\hat{d}^{0\nu}_{ij} \equiv \mathcal{M}_{ij}^{\mu\nu} \\
&C_{ij} \; \mathcal{M}_{ij}^{\mu\nu} \equiv H_{i,j}^{\mu,\nu}
\end{aligned}
\tag{1.8}
$$

where the symbol $\delta_{ij}$ is the Kronecker delta, and $\hat{d}^0_{ij}$ is the unit distance vector between atoms $i$ and $j$. The Hessian matrix $H$ contains all the relevant information about the energy profile in the proximity of the reference structure. If the higher order terms of the expansion are neglected, the dynamics is governed by Newtonian equations of motion:

$$M\ddot{\vec{r}} = -H\vec{r} \tag{1.9}$$

where $M$ indicates the mass matrix. The general solution of these equations is given by a linear superposition of eigenvectors of the mass-weighted Hessian matrix, $M^{-1/2}HM^{-1/2}$. These modes represent the collective degrees of freedom of the protein's vibrations; the eigenvectors associated to the lowest eigenvalues oscillate with the slowest frequency, and represent the concerted vibrations involving the simultaneous displacements of a large number of atoms. The study of these modes is commonly known as *Normal Mode Analysis (NMA)*.

Unfortunately, the procedure described above to obtain the low-energy vibrational spectrum of a protein (energy minimisation, calculation of the all-atom Hessian, diagonalisation) requires a considerable computational effort to characterise minima of the energy in which the system dwells for a few femtoseconds. In her seminal paper, M. Tirion (27) suggested a dramatic simplification of the energy; in particular, she proposed to replace the atomistic pairwise interaction terms with simple Hookean springs between all atoms within a given cutoff distance $R_c$:

$$\begin{aligned} &E_{ij}(\vec{r}_i,\vec{r}_j) \equiv \frac{1}{2}K\ \Delta_{ij}\ (d_{ij}-d^0_{ij})^2 = \frac{1}{2}\delta\vec{r}\ H\ \delta\vec{r} + \mathcal{O}(\delta\vec{r})^3 \\ &\Delta_{ij} = \theta(R_c - d^0_{ij}) \end{aligned} \tag{1.10}$$

This model is related to the second order Taylor expansion in Eq. 1.7 by the substitution:

$$C_{ij} \quad = \quad K\Delta_{ij} \tag{1.11}$$

The low-energy modes of vibration obtained with this very simple model remarkably reproduced, after appropriate tuning of the elastic constant, those resulting from the diagonalisation of the Hessian. More interestingly, it was found that the low-energy part of the spectrum was essentially unchanged if different values of the cutoff radius $R_c$ were used, ranging from 1.1 to 2.5 Å. This result suggested the hypothesis that the salient features of a protein's collective internal dynamics are not sensitive to the fine-grained

detail of the interaction. The possibility to obtain important informations about the concerted, functionally-relevant motions of a protein making use of simple harmonic force fields, has been thoroughly investigated (28; 29; 30; 31; 32; 33) and represents the basis for a number of applications, including those that shall be discussed in this thesis.

### 1.2.2 Elastic Network Models

The time-scales probed by the protein dynamics described in terms of the Hessian matrix are of the order of the picosecond (34; 35). In a larger time scale, the protein visits many energy minima and explores the corresponding equilibrium conformations. The quadratic expansion of the energy is therefore not applicable in this case.

Nonetheless, in many cases MD simulations have shown that the protein undergoes pronounced fluctuations around a well-defined reference structure, indicating that the several 'tiny' local energy minima are embedded in a broader minimum of the free energy, as pictorially sketched in Fig. 1.2 (36).

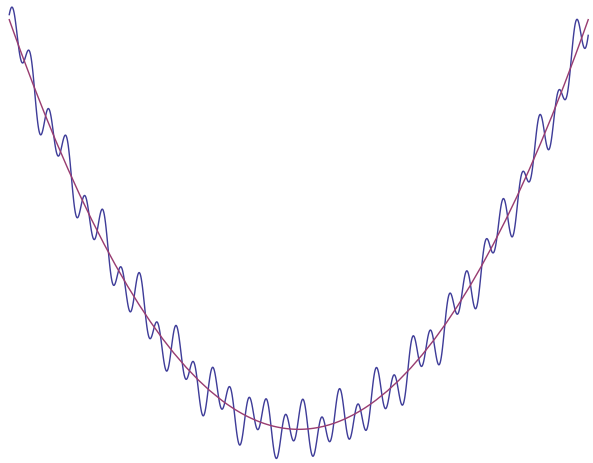

**Figure 1.2: Pictorial representation of the free energy landscape of a protein** - The conformational space of a protein, here assumed to be described by a single coordinate, is characterised by a large number of small energy minima, in which the molecule undergoes small-scale vibrations. The convolution of these minima forms a large well of the free energy, which can be approximated with a quadratic potential.

This minimum can be parametrized in the same spirit of the simplification of the energy suggested by Tirion, i.e. assuming pairwise interactions between atom pairs quadratically penalising the deviations of the relative distance from the reference value.

Moreover, a further approximation can be done, taking into account the result shown in the previous paragraph concerning the mild dependence of the low-frequency spectrum from the atomistic detail. In fact, the conformational rearrangements that the protein undergoes during the time-scales of interest are much larger than those explored in an energy minimum: in this case, the small variations of a side chain dihedral angle will not change appreciably the free energy; at the same time, the latter cannot take into account such minutiæ. Therefore, the fine atomistic details of the protein can be neglected, and a single atom can be retained to represent an amino acid - typically the $C_\alpha$ atom. These phenomenological, coarse-grained models of a protein are commonly known as *Elastic Network Models (ENMs)*.

The elastic network free energy (the *Hamiltonian*) of the protein is given by:

$$\mathcal{H} = \frac{C}{2} \sum_{ij,\mu\nu} \delta r_i^\mu \ \mathcal{M}_{ij}^{\mu\nu} \ \delta r_j^\nu \tag{1.12}$$

where the variables $\delta r_i^\mu$ indicate the displacement of the $i$-th $C_\alpha$ from its reference position, $r_i^0$, along the $\mu$-th direction. The Hamiltonian $\mathcal{H}$ presents a strict formal analogy with both the Hessian matrix and the simplified quadratic model of Tirion; The difference with respect to these models is that the $\mathcal{M}$ matrix is built *a priori* on the basis of the sole structural information, and its detailed form and properties depend on the specific elastic network model under exam.

The equilibrium dynamical properties of a protein described as an elastic network can be obtained with the use of the partition function formalism. In fact, the covariance matrix is calculated as an equilibrium average:

$$\begin{aligned} \mathcal{C}_{ij}^{\mu\nu} &\equiv \langle \delta x_i^\mu \delta x_j^\nu \rangle = \frac{1}{\mathcal{Z}} \int \mathcal{D}[\delta\vec{x}] e^{-\beta \frac{1}{2} \delta\vec{x}\mathcal{M}\delta\vec{x}} \delta x_i^\mu \delta x_j^\nu \\ \mathcal{Z} &= \int \mathcal{D}[\delta\vec{x}] e^{-\beta \frac{1}{2} \delta\vec{x}\mathcal{M}\delta\vec{x}} \end{aligned} \tag{1.13}$$

where $\mathcal{Z}$ indicates the partition function of the system, and $\beta = 1/K_B T$ is the inverse temperature. The calculation of the covariance matrix elements can be performed exactly introducing an auxiliary field in the Hamiltonian and differentiating the partition function, as follows:

$$\begin{aligned} \mathcal{Z}[\vec{J}] &= \int \mathcal{D}[\delta\vec{x}] e^{-\beta\left[\frac{1}{2}\delta\vec{x}\mathcal{M}\delta\vec{x} + \vec{J}\delta\vec{x}\right]} \\ \mathcal{C}_{ij}^{\mu\nu} &= \frac{1}{(-\beta)^2} \frac{1}{\mathcal{Z}} \frac{\partial^2 \mathcal{Z}[\vec{J}]}{\partial J_i^\mu \partial J_j^\nu}\bigg|_{\vec{J}=0} \equiv -\frac{1}{\beta} \frac{\partial^2 F[\vec{J}]}{\partial J_i^\mu \partial J_j^\nu}\bigg|_{\vec{J}=0} \\ F[\vec{J}] &= -\frac{1}{\beta} \log(\mathcal{Z}[\vec{J}]) \end{aligned} \tag{1.14}$$

Performing the substitution:

$$\delta\vec{x} = \delta\vec{y} + \vec{J}\mathcal{M}^{-1} \tag{1.15}$$

one obtains the following form for the partition function (observe that the integral is not affected by the change of variables since the extremes of the integrations are $\pm\infty$):

$$\begin{aligned} \mathcal{Z}[\vec{J}] &= \int \mathcal{D}[\delta\vec{y}] e^{-\beta\left[\frac{1}{2}\delta\vec{y}\mathcal{M}\delta\vec{y} - \frac{1}{2}\vec{J}\mathcal{M}^{-1}\vec{J}\right]} \equiv \mathcal{Z}\ e^{\frac{\beta}{2}\vec{J}\mathcal{M}^{-1}\vec{J}} \\ F[\vec{J}] &= -\frac{1}{2}\vec{J}\mathcal{M}^{-1}\vec{J} \end{aligned} \tag{1.16}$$

The derivation of the last equation, according to the Eq. 1.14, thus gives:

$$\mathcal{C}_{ij}^{\mu\nu} = -\frac{1}{\beta}\frac{\partial^2 F[\vec{J}]}{\partial J_i^\mu \partial J_j^\nu}\bigg|_{\vec{J}=0} = \frac{K_B T}{2}\mathcal{M}^{-1} \tag{1.17}$$

The above calculation shows that the covariance matrix of amino acid displacements can be exactly calculated, in the framework of a quadratic ENM, simply performing an inversion of the interaction matrix $\mathcal{M}$ (provided that the null modes are excluded from the spectral decomposition). The eigenvalues of the interaction matrix are related to the relaxation time of the collective deformations of the molecule described by the corresponding modes: smaller eigenvalues of $\mathcal{M}$, or equivalently larger eigenvalues of $\mathcal{C}$, are related to large-scale, collective fluctuations involving the coherent motion of a large number of amino acids.

### 1.2.3 The $\beta$-Gaussian Model

In this paragraph we describe the specific elastic network model, namely the $\beta$-Gaussian model ($\beta$-GM hereafter) introduced by Micheletti *et al.* (37), that will be largely used throughout this thesis.

In this model the protein is described in a two-centroids representation per amino acid: a $C_\alpha$ carbon and an effective $C_\beta$ representing the side-chain. The use of two atoms per amino acid determines a negligible increase in the computational complexity; on the other hand, the mobility of the residues calculated with this approach reproduces with higher accord the experimental B-factors with respect to one-centroid models. In fact, in the latter formulation a large and unphysical cutoff radius of $\sim 10$ Å must be used, in order to prevent the system from having more than 6 null modes. Contrarily,

if two centroids are used the network is more constrained, and the cutoff radius can be lowered to $\sim 7$ Å without introducing non-rototranslational null modes.

In the $\beta$-GM, the motion of the $C_\beta$ is constrained by geometrical relations to follow the displacement of the neighbouring atoms. The position $\vec{r}_{CB}(i)$ of the $i$-th sidechain centroid is in fact expressed by (38):

$$\vec{r}_{CB}(i) = \vec{r}_{CA}(i) + l \, \frac{2\vec{r}_{CA}(i) - \vec{r}_{CA}(i+1) - \vec{r}_{CA}(i-1)}{|2\vec{r}_{CA}(i) - \vec{r}_{CA}(i+1) - \vec{r}_{CA}(i-1)|} \tag{1.18}$$

The parameter $l$, fixing the distance from the $C_\alpha$ and its effective sidechain atom, is usually assigned a value of 3 Å. For amino acids at the beginning/end of the peptide chain(s), or for glycine atoms, the construction of Eq. 1.18 is not applicable: in these cases the effective centroid is taken to coincide with the $C_\alpha$ atom.

The effective energy among the atoms of the model is built as a sum of pairwise Hookean interactions, as in Eq. 1.10. The total energy is given by:

$$E_{\beta GM} = \tag{1.19}$$
$$2\sum_i E(d_{i,i+1}^{CA-CA}) + \sum_{i<j}{}' E(d_{i,j}^{CA-CA}) + \sum_{i,j}{}' E(d_{i,j}^{CA-CB}) + \sum_{i<j}{}' E(d_{i,j}^{CB-CB})$$

The first term in Eq. 1.19 takes into account the chain connectivity; in order to reflect the strength of the covalent peptide bond a factor 2 is put in front of the summation. The following terms describe the interaction between non-consecutive atoms, and the summations are restricted to those pairs whose distance lies within the cutoff radius, as indicated by the primed sums. In the spirit of Tirion (27), all terms have the same strength.

The pairwise interaction terms, which depend only on the coordinates of the $C_\alpha$ atoms, admit an expansion as in Eq. 1.8, leading to a quadratic Hamiltonian:

$$E_{\beta GM} = \frac{C}{2} \sum_{ij,\mu\nu} \delta r_i^\mu \; H_{ij}^{\mu\nu} \; \delta r_j^\nu \tag{1.20}$$

This model allows for an efficient and reliable calculation of the low-energy space of a protein, and has been applied to a number of cases in different contexts (37; 39; 40). In the present thesis, the $\beta$-GM will be a pivotal gear of many algorithms, whose application will provide a deeper understanding of the internal dynamics of proteins.

# 2

# Common features in protein internal dynamics and identification of relevant collective variables

In the previous chapter it was shown that salient features of protein internal dynamics can be captured by coarse-grained models, where the small-scale details of structure and interactions are neglected. This fact, which was proven *a posteriori*, prompts the following question: is it possible to identify conserved dynamical features among different conformational states of the same protein?

The covariance matrix of a MD simulation and the analysis of its principal components are among the most commonly used methods to extract the relevant information from the ensemble of protein conformers. Garcia (22) and, later, Amadei (23) showed that a few top eigenvectors of the covariance matrix, having the largest eigenvalues, account for a large fraction of the protein's mobility. Moreover, the studies of refs. (15; 23; 24) highlighted the non-harmonic character of these modes, in contrast with the essentially Gaussian behaviour of higher energy fluctuations. These observations suggested that a small subset of collective variables, namely the top-ranking modes of the covariance, capture the relevant, functionally-oriented properties of a protein's internal dynamics.

Nonetheless, this reduction of the effective internal dynamics space can be reliably carried out only for long enough simulations. To show this, Hess (41) considered the PCA of a high-dimensional random diffusion, consisting in 120 independent Brownian processes in Euclidean space. Statistical features of this simple reference model were

found in MD simulations of proteins of about 200 residues, covering time spans of a few ns. In both systems, for example, the amplitude of the top modes of the covariance were cosines with frequency proportional to the mode rank.

This similarity showed that for short simulated time intervals robust properties of the protein internal dynamics might not be captured by a PCA. In particular, if only a small fraction of the conformational space has been sampled, the motion of the molecule is mainly due to noise, in spite of the fact that the first few modes of the covariance are sufficient to capture a large fraction of the mean square fluctuation.

A method to validate the robustness of the essential space is to compare the essential dynamics spaces of two subsets of the same MD trajectory (23; 41), i.e. to measure the degree of internal consistency between the top-ranking modes of the two sub-trajectories' covariance matrices.

Various measures have been proposed to estimate the similarity of sets of vectors rather than single pairs; one of the most simple and widely used is the *Root Mean Square Inner Product*:

$$\mathrm{RMSIP} = \sqrt{\frac{1}{n}\sum_{\ell,m=1}^{n}(\vec{v}^{\ell}\cdot\vec{w}^{m})^{2}} \tag{2.1}$$

Notice that in Eq. 2.1 the consistency of the linear space spanned by the first $n$ modes of the covariance is measured without reference to the difference in their eigenvalues, which are thus treated as degenerated.

This quantity represents an extension of the scalar product to subspaces of vectors having the same dimensionality. If the two sets describe the same manifold the RMSIP is equal to one, while it is zero for orthogonal subspaces. The consistency of two subspaces can be easily ascertained comparing the value of their RMSIP with a reference distribution, obtained for example comparing unrelated vectors (15).

Unfortunately, no criterion exist for establishing *a priori* the number $n$ of top modes that should be retained to describe well the protein internal dynamics: the value of $n$ is, in fact, customarily taken equal to 10 by a pure convention. Moreover, the results discussed so far showed that the eigenvalues alone are not sufficient to decide *a priori* if a given set of these modes comprise the robust dynamics, nor to be sure that this information results from a converged simulation.

The study discussed in this chapter is aimed at establishing which subset of a protein's low energy modes is robustly shared by the various configurations that altogether describe a conformational sub-state. The problem will be tackled from a 'conformational ensemble' perspective, in that we shall suitably compare the properties of a large collection of protein structures. This approach therefore differs in spirit and formulation from other existing schemes where the detailed kinetic history of the protein is

analysed to ascertain if a given MD trajectory is sufficiently long to allow for a confident determination of the low-energy modes. In the following section we shall review a few results discussed in the literature to lay the ground of our method.

## 2.1 Common features in protein internal dynamics

### 2.1.1 Universal density of vibrational modes in globular proteins

The bulk density of globular proteins is typically comparable with that of molecular crystals. General features of the latter, such as the vibrational spectral density at low frequencies, depend essentially on their effective dimensionality. The curiosity might thus arise, if proteins show 'universal' vibrational features similar to molecular crystals.

One of the first attempts to identify shared features among different globular proteins was performed by Ben Avraham (42). In this work, the number densities of vibrational modes, $g(\omega)$, from five different globular proteins were compared. The spectra, obtained through a normal mode analysis of the quadratic approximation of the potential energy, showed a striking superposition of the $g(\omega)$ (see Fig. 2.1) for proteins whose difference in length spanned one order of magnitude.

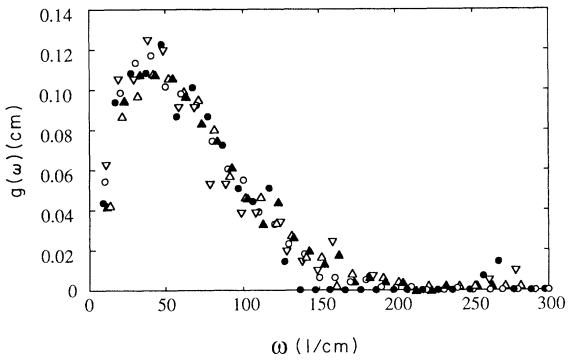


FIG. 1. Density of vibrational normal modes, $g(\omega)$, of $g$ actin (○) lysozyme (△), ribonuclease I (▲), BPTI (▽), and crambin (●) as a function of frequency.

**Figure 2.1: Density of vibrational normal modes of different proteins** - Figure reproduced from the paper by Ben Avraham (42).

Another peculiar result is given by the spectral dimension of the 'universal' vibrational spectra. From the analysis of the number densities of the modes it turned out that $g(\omega) \sim \omega$. According to the definition of spectral dimension, $d_s$:

$$g(\omega) = \omega^{d_s - 1} \tag{2.2}$$

one obtains that $d_s = 2$. Consequently, the low-frequency vibrational spectrum of proteins is comparable to the one of two-dimensional crystals. A possible explanation of this fact can be found in the tight backbone bonds, which reduce the effective number of degrees of freedom to 2 per amino acid, at least for the low-energy modes of fluctuation.

These results indicate that general properties of low-energy vibrations are shared by different proteins. It is therefore expected that local fluctuations of a protein with respect to different conformers, as they can be obtained in a MD simulation, show a large consistency. The lack of the latter, in turn, is likely to reveal features of the molecule's internal dynamics that cannot be ascribed to the noise, that is the unavoidable stochasticity that accompanies MD simulations of large molecular systems.

### 2.1.2 Self-similarity of free-energy landscape of G-protein and adenylate kinase

One of the most striking properties of many proteins is the innate character of their internal dynamics, depending only on general features of the structure (13).

In (43) Pontiggia *et al.* performed four different 100-ns long MD simulations of immunoglobulin binding domain of protein G (GB1). The analysis of the trajectories showed that the distributions of the motion projected on the top eigenvectors of the covariance matrix have a unimodal, quasi-harmonic character at the beginning of the simulation, which is lost after a few ns. Accordingly, the time evolution of the eigenvalues resulted in a progressive decrease of the effective frequency of the slow modes, suggesting a broadening of the quadratic well approximating the free energy landscape explored progressively by the trajectory.

These results are suggestive of the limited range of applicability of the harmonic approximation of the free energy. Nonetheless, if on the one hand the harmonic approximation of the free energy profile showed to be valid only for the first ns of the simulation, on the other hand a remarkable consensus was found among the principal directions of fluctuation explored corresponding to the low-energy modes of the protein.

In order to quantify the degree of similarity shared by two subspaces, each constituted by $N$ orthonormal vectors, Pontiggia *et al.* calculated the RMSIP between the top ten modes of the covariance obtained from the first ns of the first simulation and those obtained considering increasing time intervals of all four trajectories. The resulting time-dependent RMSIP spanned values ranging between 0.6 and 0.7: the statistical significance of this result was established by comparison with the distribution of RMSIP of two randomly picked sets of orthonormal bases, which returned a value of $0.24 \pm 0.02$.

The robustness of the low-energy, collective directions of fluctuations among the four trajectories of protein domain GB1 also emerged from the analysis of a 50-ns long MD simulation of Adk (15). During this 50-ns simulation the molecule explored many

different free energy minima, or *sub-states*. The latter are defined as structurally homogeneous, continuous intervals of the MD trajectory: within each sub-state, the protein fluctuates around a well-defined average structure. The identification of the optimal subdivision of the trajectory in a given number of sub-states is performed minimising, over all possible partitions in continuous intervals, the mean square fluctuation internal to each sub-state.

A remarkable consensus of the dynamics of this protein was found, not only among the fluctuations internal to the structurally-homogeneous trajectory sub-states identified, but also among the modes *connecting* the different sub-states. The functional relevance of this self-similarity of the free energy profile was assessed by the high degree of overlap which was found between the essential dynamical spaces and the difference vector connecting the open and closed crystallographic structures of the protein.

The successful prediction of a protein's modes of fluctuation making use of coarse-grained ENM's thus results from the fact that the essential dynamics is generally shared among the sub-states; as a consequence, all sub-states can be used to obtain the essential spaces. On the other hand, the amplitudes of the modes are out of direct control, so that the most robust feature is the *directionality* of the essential modes.

## 2.2 Identification of relevant modes of fluctuation in adenylate kinase internal dynamics: a Random Matrix Theory approach

The cases discussed so far showed that robust features of the internal dynamics can coexist with general properties common to many different proteins. In fact, on the one hand the collapse of the vibrational spectra observed by Ben Avraham is suggestive that the coarse features are generally shared by all globular proteins; on the other hand, the robustness of the collective fluctuations, corresponding to the top-ranking covariance matrix modes of GB1 Adk, indicates that some internal dynamics properties of proteins are conserved throughout the dynamical evolution.

Here we carry out a novel analysis, aimed at identifying which subset of a protein's low energy modes of fluctuation is robust with respect to differences in the reference structure. In particular, we consider an ensemble of conformations of adenylate kinase, obtained in the atomistic MD simulation of ref. (15). In order to characterise the internal dynamics of these conformers, we make use of the $\beta$-Gaussian ENM as a proxy for the Hessian, in the spirit of ref. (27). Each instantaneous configuration of Adk is thus taken as a reference structure for the $\beta$-GM: the resulting ensemble of covariance matrices describe the local fluctuation space of different, though structurally homogeneous conformers of Adk.

The statistical properties of the corresponding low-energy spaces will be compared to those of a reference ensemble of random matrices: those modes, whose eigenvalues markedly deviate from the universal distribution, are identified as the most robust with respect to small structural differences.

In the following section a brief account of the basics of Random Matrix Theory is given, to familiarise with the concepts that will be later used.

### 2.2.1 A brief account of Random Matrix Theory

The foundations of Random Matrix Theory date back to the 1950's. Originally, ensembles of matrices having randomly-distributed entries were introduced as effective models to describe the excited states of certain atomic nuclei.

At that time, in fact, no well-established theory accounted for a valid *dynamical* model of the nucleus which could explain the cross-section spectra obtained experimentally. In order to circumvent this limitation, Wigner (44) suggested a *statistical* approach, relying on the complexity of the nuclear spectra. This proposition might appear in contradiction with the fact that the time evolution of these quantum systems is governed by a well-defined Hamiltonian; nonetheless, statistical concepts can be useful - and often the only available tools to describe the *average* properties of complex systems.

Wigner proposed a statistical theory which is somewhat diverse from the standard approaches of Statistical Physics. In the latter discipline, in fact, one usually assumes systems to be governed by a given Hamiltonian, and considers the time evolution of the system starting from many different initial condition. On the other hand, Wigner considered ensembles of systems governed by *different* Hamiltonians, sharing the same symmetry properties. The basic assumption is therefore that we know nothing about the system interactions but a few general constraints (e.g. conservation laws) which are to be enforced. The basic models of this Theory consider finite-size $N \times N$ matrices satisfying suitable constraints. The size $N$ is kept finite for sake of computational feasibility; nonetheless these matrices model quantum Hamiltonians in an infinite-dimensional Hilbert space, therefore the limit $N \to \infty$ has to be taken at some stage, akin to the thermodynamic limit.

Using early group–theoretical results by Wigner, Dyson showed that in the framework of standard Schroedinger theory, there are three generic ensembles of random matrices, defined in terms of the symmetry properties of the Hamiltonian. These ensembles are characterised by:

- Orthogonal symmetry, with time-reversal and rotational invariance
- Hermitian symmetry, where time-reversal invariance is violated

- Time-reversal invariance, halfinteger spin and broken rotational symmetry

These three models, which are fundamentally group-irreducible, lay the basis of the Theory.

The matrices of these ensembles are drawn from a probability distribution, whose functional form must be compatible with the aforementioned symmetry requirements. A flat, uniform distribution satisfies the constraints, but leads to divergent integrals. Dyson therefore assumed the trace of the squared Hamiltonians to be Gaussian distributed:

$$P_{N\beta}(H) \propto exp\left(-\frac{N\beta}{c}Tr[H^2]\right) \tag{2.3}$$

In the probability distribution of Eq. 2.3 a normalisation factor has been neglected; the factor $N$ has been introduced to ascertain that the spectrum remains bounded in the limit $N \to \infty$, and the constant $c$, independent of $N$, defines the width of the distribution.

The choice of the $\beta$ parameter and the appropriate symmetries define the ensemble: for $\beta = 1$ we have the Gaussian Orthogonal Ensemble (GOE); $\beta = 2$ defines the Gaussian Unitary Ensemble (GUE); finally, for $\beta = 4$ we have the Gaussian Symplectic Ensemble (GSE). All three ensembles are discussed in the context of Gaussian Random Matrix Theory (GRMT).

Due to the Gaussian damping of the eigenvalues, in the thermodynamic limit the support of the spectrum of these matrices is bounded in the interval $-2c \leq \lambda \leq 2c$. Wigner derived the distribution of the eigenvalues, which has the shape of a semicircle (the *semicircle law*):

$$\rho(\lambda) = \frac{N}{\pi c}\sqrt{\left(1 - \frac{\lambda}{2c}\right)^2} \tag{2.4}$$

Another important quantity which is investigated in the theory of random matrices is the *local spacing statistics*; for eigenvalues $\lambda_k$ ranked in decreasing order, one is interested in the the distribution of the Individual Eigenvalue Spacing (IES) $s_k$, defined as (45):

$$s_k = \frac{\lambda_k - \lambda_{k+1}}{\langle \lambda_k - \lambda_{k+1} \rangle} \tag{2.5}$$

The average $\langle \cdot \rangle$ is taken over the matrix ensemble where each element is generated with the weight of Eq. 2.3 and, clearly, $\langle s_k \rangle = 1$ for any $k$.

## 2. COMMON FEATURES IN PROTEIN INTERNAL DYNAMICS AND IDENTIFICATION OF RELEVANT COLLECTIVE VARIABLES

Wigner (44) proposed a form for the distribution $p(s)$ of individual eigenvalue spacings; this *Wigner surmise*, originally stated for $\beta = 1$, has the form:

$$\begin{aligned} p_\beta^{WS}(s) &= a_\beta s^\beta \exp(-b_\beta s^2) \\ a_\beta &= 2\frac{\Gamma^{\beta+1}((\beta+2)/2)}{\Gamma^{\beta+2}((\beta+1)/2)} \\ b_\beta &= 2\frac{\Gamma^2((\beta+2)/2)}{\Gamma^2((\beta+1)/2)} \end{aligned} \tag{2.6}$$

The Wigner surmise shows a strong, $\beta$–dependent level repulsion at small spacings and a Gaussian decay for large values of $s$.

An important extension to the Wigner surmise has been introduced by Brody (46):

$$\begin{aligned} p_q(s) &= c_q(1+q)s^q \exp(-c_q s^{1+q}) \\ c_q &= \Gamma^{1+q}((2+q)/(1+q)) \end{aligned} \tag{2.7}$$

For the orthogonal case ($\beta = 1$) this expression interpolates between the Poisson and the Wigner distributions; the mixing between the two functional forms is parametrized by the phenomenological parameter $q$.

Eq. 2.7 proved particularly effective to reproduce the spacing distribution of a class of GOE matrices of particular interest, the Wishart-Laguerre (WL) (47; 48) ensemble. This ensemble includes $N \times N$ covariance matrices $\mathcal{W}$ of the form:

$$\mathcal{W} = \frac{1}{T}\sum_k \mathcal{X}_{ik}\mathcal{X}_{jk} \equiv \frac{1}{T}\mathcal{X}\mathcal{X}^t \tag{2.8}$$

where $\mathcal{X}$ is an $N \times T$ matrix containing $N$ time series of $T$ independent elements, and the superscript $t$ indicates transposition. The elements of the $\mathcal{X}$ matrix are drawn from a Gaussian distribution with zero mean and *fixed* variance, i.e. $\mathcal{X}_{ij} \sim \mathcal{N}(0, \sigma^2)\ \forall\ i, j$.

Since $\mathcal{W}$ is the covariance matrix of a maximally random data-set, it is ideally suited to serve as a term of reference to establish the non-random character of a covariance matrix calculated from a MD trajectory. In particular, it can be used to ascertain how many of the components are robust, i.e. non significantly affected by the 'noise' of the MD simulation. This approach has been previously applied to analyse financial data (49), internet routers networks (50), EEG data (51) and atmospheric correlations (52) among others. It appears therefore appropriate to investigate the statistical properties of protein covariance matrices, $\mathcal{C}$. Nonetheless, a generalisation of this model is needed, in order to account for the symmetry properties of the Gaussian Orthogonal Ensemble which are not shared by protein covariance matrices, as discussed in the following section.

### 2.2.2 Reference stochastic model of covariance matrices

The covariance matrix, $\mathcal{C}$, of a protein, as appears in Eq. 1.1, can be equivalently written in terms of a matrix $\mathcal{X}$ defined as:

$$\mathcal{X}_{ik} = \delta r_i(k) - \langle \delta r_i \rangle \tag{2.9}$$

where $k$ is a time index, and $i$ labels both the residue and the Cartesian coordinate. The covariance matrix $\mathcal{C}$ is thus given by:

$$\mathcal{C}_{ij} \equiv \langle (\delta r_i - \langle \delta r_i \rangle)(\delta r_j - \langle \delta r_j \rangle) \rangle = \frac{1}{T} \sum_k \mathcal{X}_{ik} \mathcal{X}_{jk} \tag{2.10}$$

In the WL ensemble, the diagonal elements of the covariance matrix

$$\mathcal{C}_{ii} \equiv \left\langle (\delta r_i - \langle \delta r_i \rangle)^2 \right\rangle \tag{2.11}$$

have the same value - namely, they are normalised to unity. The enforcement of this constraint in the WL ensemble allows for the exact solvability of the model (53). Nevertheless, it makes inappropriate the comparison with protein covariance matrices, where the variances corresponding to the $\mathcal{C}_{ii}$ entries are not forced to be equal.

We thus turned to a *generalised* $\sigma$-WL non-invariant model. Specifically, we considered an ensemble of $N \times T$ random matrices $\mathcal{X}$ describing $N$ time-series with zero average and standard deviation drawn from a uniform distribution. This choice, which represents one of the simplest deviations from the WL ensemble, breaks the invariance of the model under orthogonal transformations; this in turn prevents the applicability of standard techniques to solve the problem analytically. However, the latter can be straightforwardly tackled computationally, so to obtain numerically the distributions of the spacings for this model.

In spite of the introduction of unequal diagonal elements of the covariance matrices, the Brody distribution provides a valid model for the spacings of this ensemble's eigenvalues. In Figs. 2.2 – 2.3 we plotted the histograms of the first 3 and the 9th spacings, obtained from a set of $10^4$ rectangular matrices with $N = 10$ and $T = 20$. From the inspection of Figs. 2.2 – 2.3 it can be seen that the histograms are well approximated by a one-parameter fit with the Brody distribution, with the $q$ value ranging from 1.2 for the first spacing to $q = 0.84$ for all the remaining 8 spacings. The Brody distribution may therefore be used as the stochastic reference against which we can compare the spacing distributions of covariance matrices lacking orthogonal invariance.

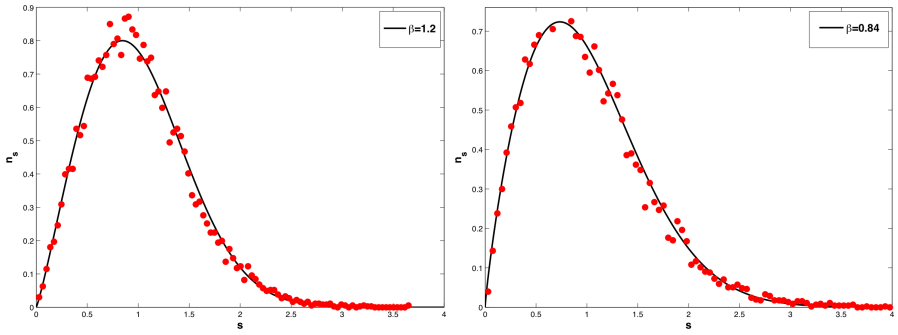


**Figure 2.2: Distributions of the 1st and 2nd spacing of the $\sigma$-WL ensemble** - The distributions of the 1st and 2nd spacings here shown have been obtained from a sample of $10^4$ random matrices of rank 10 built according to the $\sigma$-WL model.

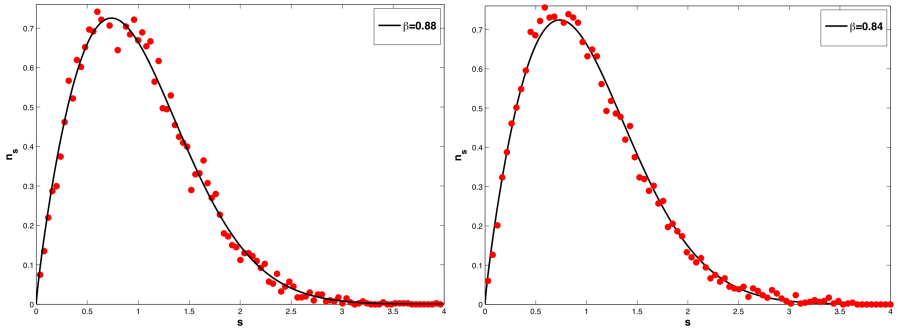


**Figure 2.3: Distribution of the 3rd and 9th spacing of the $\sigma$-WL ensemble** - Distributions of the 3rd and 9th spacings.

### 2.2.3 RMT analysis of an ensemble of covariance matrices

We applied our analysis to the 4000 MD 'frames' of the shortest (2 ns) structurally-homogeneous interval of the simulation of Adk discussed in (15). Each frame was taken as the reference structure for the $\beta$-GM, so to obtain the corresponding ensemble of covariance matrices.

For each matrix we computed the eigenvalues and level spacings, whose statistical properties were compared with the predictions of random correlation matrices.

Together with the spectra of the 'bare' eigenvalues $\{\lambda_\ell^{(j)}\}$ (the $\ell$-th eigenvalue of $j$-th CM sample, ranked in decreasing order), we also considered the eigenvalues normalised to the trace of the covariance matrix:

$$\mu_\ell^{(j)} := \frac{1}{(3N-6)} \frac{\lambda_\ell^{(j)}}{\mathrm{Tr}[\mathcal{C}^{(j)}]} \tag{2.12}$$

where $N = 214$ is the length of the protein. The $\mu$'s are thus normalised so that their sum reproduces the number of degrees of freedom.

The first quantity we analysed is the fraction of motion captured by the first $n$ modes:

$$f_n = \Big\langle \frac{1}{\mathrm{Tr}[\mathcal{C}]} \sum_{\ell=1}^{n} \lambda_\ell \Big\rangle \equiv \frac{1}{3N-6} \Big\langle \sum_{\ell=1}^{n} \mu_\ell \Big\rangle \tag{2.13}$$

and plotted in Fig. 2.4. As expected, the very first eigenvalues capture more than 70% of the protein's overall mobility; this feature is consistently preserved throughout the simulated time span, as indicated by the relatively small error bars.

In order to statistically characterise the $k$-th ranking eigenvalue we plot, in Fig. 2.5, its relative dispersion (stdev/mean) *vs.* its index $\ell$: a low ratio signals a strong localisation. The $\mu$'s display a constant value, suggesting stability in the distribution of the fraction of total mobility captured by each mode. In comparison, the relative dispersion of the un-normalised $\lambda$'s rapidly decay to very low values after crossing the range spanned by the $\mu$'s approximately between the 3rd and the 4th eigenvalue. The broad dispersion of the $\lambda_k$ for low $k$ suggests that the amount of internal dynamics, in absolute value, captured by the first few modes of the covariance can vary depending on the conformation that is taken as the ENM reference structure. On the other hand, the low and fairly constant relative dispersion of the $\mu$'s indicates that the *fraction* of the MSF, which is captured by the low-energy modes, is much less sensitive to small structural differences. This result points at a possible discrepancy between the statistical properties of the two eigenvalue sets.

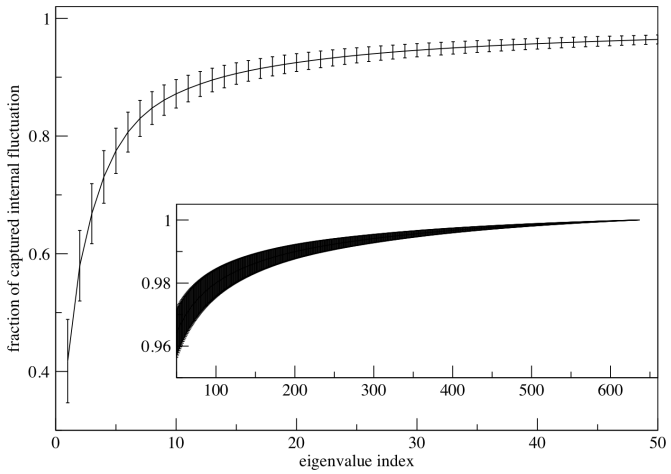


**Figure 2.4: Cumulative fraction of the internal fluctuation of Adk as a function of the number of modes** - The figure shows the average fraction of the protein's fluctuation $f_n$ -defined in Eq. 2.13- as a function of the first $n$ eigenvalues; the error bars are calculated as standard deviations. The inset shows the detail of the curve for values of $f_n$ larger than 0.96: the error bars are of the order of 1%.

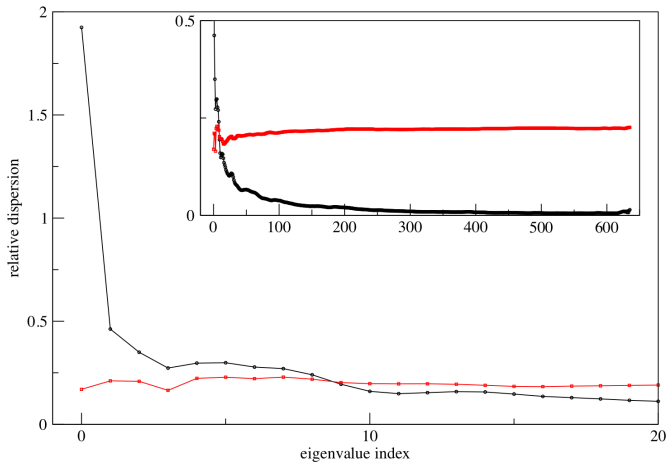


**Figure 2.5: Relative dispersion of the eigenvalues of Adk** - The relative dispersion of the eigenvalues (defined as standard deviation in units of the mean) is shown for the bare (black curve) and normalised (red curve) eigenvalues. The former show a rapid drop to small values, while the normalised ones remain almost constant.

In contrast with the latter expectation, the spacing distributions in the 'bulk' of the eigenspace show a remarkably universal pattern. In Figs. 2.6 and 2.7, the distributions of the $\lambda$'s and $\mu$'s are fitted with a Brody distribution. With a $\chi^2$ test, the consistency with the null-hypothesis reference distribution can be rejected with high confidence (1% level) for the first 3 spacings in both cases. The subsequent ones instead, give overall quite a good agreement with the Brody distribution fit, with a fit parameter $q = 0.8 \pm 0.1$[1]. Note that the same $q$, within the statistical bounds, fits the spacing distribution for the null $\sigma$-WL model, which is remarkable given the simplicity of the reference model.

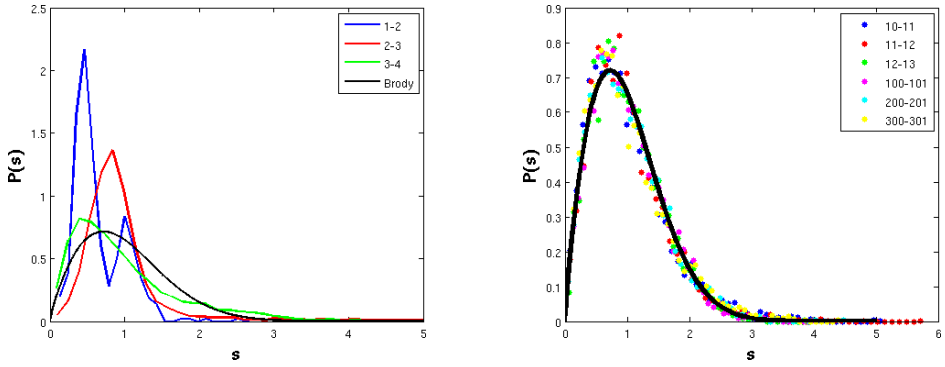


**Figure 2.6: Level spacing distributions of the bare ($\lambda$) eigenvalues** - Left panel: level spacing distributions of the first 3 $\lambda_k$. Right panel: samples of $\lambda_k$ spacings from the bulk.

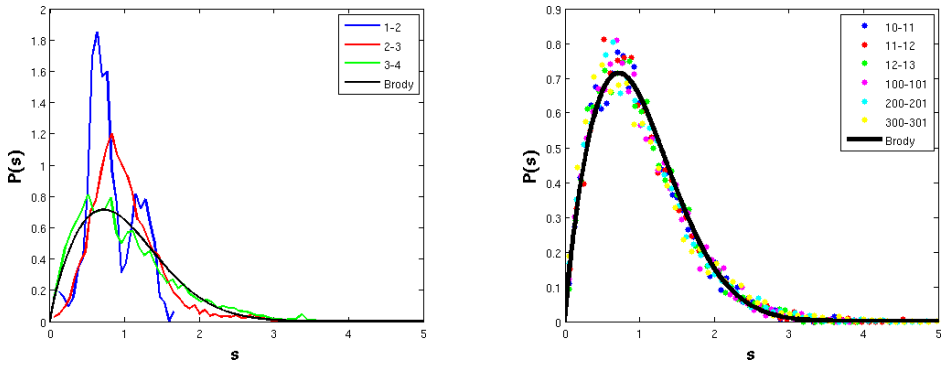


**Figure 2.7: Level spacing distributions of the normalised ($\mu$) eigenvalues** - Left panel: level spacing distributions of the first 3 $\mu_k$. Right panel: samples of $\mu_k$ spacings from the bulk.

[1]Standard error among the first 100 spacings.

The relevance of the first four modes, as indicated by the non-standard statistics of the corresponding three spacings, is further supported by the high degree of consistency between the essential spaces calculated from two halves of the trajectory. We applied the method introduced in (15) to determine an optimal redefinition of the orthonormal basis vectors of the two essential spaces, in order to quantify the degree of overlap between the two sets. Specifically, the redefined basis vectors in one set are ranked in order of decreasing overlap with the linear space spanned by the vectors in the other set. The consistency of the top modes of the two covariance matrices is confirmed by the high overlap of the first few optimal eigenvector pairs. This criterion, based on the properties of the essential space vectors rather than the eigenvalues, identifies about 4 conserved modes having scalar product of about 0.9 (see Fig. 2.8).

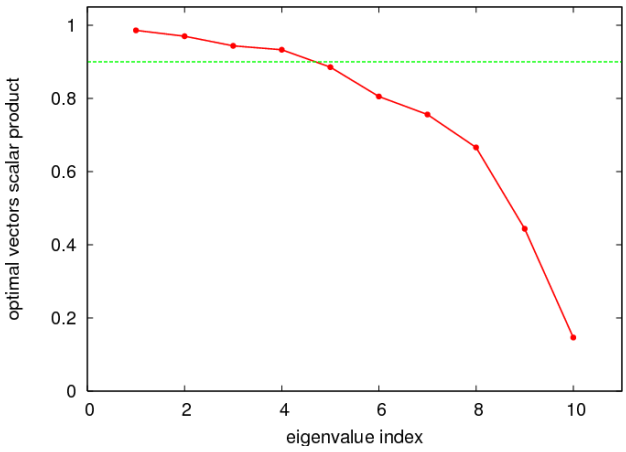


**Figure 2.8: Scalar products among top-ranking modes of the covariances from two halves of the trajectory** - The pairwise scalar products among two sets of modes are here shown, which are eigenvectors of the covariance matrices obtained from two halves of the MD trajectory under exam. These modes result from an optimal linear combination which maximises the pair scalar products between the two sets. The green line, indicating the 0.9 threshold, is shown as a guide to the eye.

The analysis of level spacings was completed with a Kolmogorov-Smirnov (KS) test among all pairs of spacing histograms. Fig. 2.9 shows the colour-coded values of the KS distances between the cumulative distributions. The distributions of the spacings from the $4^{th}$ onwards are well-superposable and, in particular, are well-fitted by the Brody distribution. On the other hand, the first 3 level spacing distributions show pronounced discrepancies with respect to the reference curve.

The investigation thus allowed to establish that, for the specific ensemble of conformations under consideration, one can identify a subspace spanned by very few collective

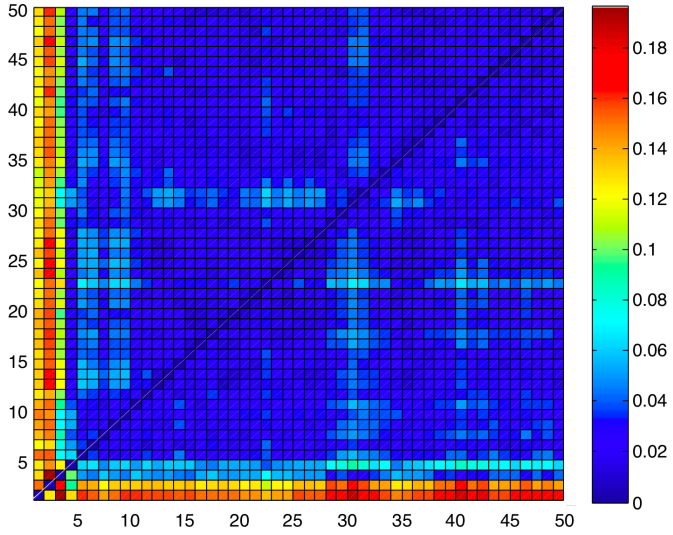


**Figure 2.9: KS distances among the $\lambda$ spacing distributions** - Kolmogorov-Smirnov distances among spacing distributions of the $\lambda$ (top triangle) and $\mu$ (bottom triangle).

modes that have a non-standard statistics. The other modes, instead, share the same quasi-universal distribution.

## 2.3 Summary

The functional-oriented character of collective dynamics for many proteins and enzymes prompted an intense study of these motions in many different ways. To this end, PCA, ENM normal modes analysis, sub-states identification and analysis, and RMSIP calculations are among the most widely and proficiently used tools.

In addition to well-established 'traditional' analyses, we discussed a novel investigation method based on the theory of Random Matrices. Specifically, we considered the ensemble of conformers from a 2-ns long interval of a MD simulation of adenylate kinase. Each conformer was used as a reference structure for the $\beta$-Gaussian elastic network model: the latter allowed us to calculate the low-energy eigenspaces of fluctuation of en ensemble of structurally-homogeneous conformers of Adk. The statistical properties of these eigenspaces, namely the distribution of the eigenvalue spacings, were then compared with universal RMT predictions, such as the Brody distribution.

This study highlighted signatures of 'universality' and random-like behaviour shared by all but the first few eigenvectors of the analysed covariance matrix ensemble. The consequence is a *quantifiable* separation between the 'bulk' modes of the covariance, whose spacing distributions could be properly fitted by the reference Brody distributions, and the few top low-energy modes, characterised by their own peculiar statistics.

This property might provide a novel framework to characterise the internal dynamics properties of many globular proteins; possible implications include a more precise identification of the collective variables describing the large-scale, functionally relevant fluctuations of biological molecules, with applications to accelerated MD schemes.

# 3

# Quasi-rigid domains in proteins

In the previous chapters we discussed aspects related to the large-scale dynamics of proteins. In particular, we have argued that the collective low-energy fluctuations can be captured by simple structure-based models, such as elastic network models. This fact suggests that the relevant information about these concerted movements, characterised by the simultaneous and coherent displacement of large groups of amino acids, is encoded in the overall architecture of the molecule. Experiments (14) have also shown that these 'innate' modes of structural fluctuation are limitedly affected by the differences in the amino acid sequence across wild-type conformers or mutants, indicating that the changes in the chemical detail can have a mild effect on the large-scale internal dynamics.

These conformational changes play a major role in the function of many proteins and enzymes (21; 22; 27; 54; 55; 56; 57; 58; 59; 60; 61; 62): it is therefore of great interest to understand the dynamical features bridging the structural organisation of the molecule to the biological function. A possible strategy to gain insight into the relation between collective internal dynamics and biological activity consists in the identification of domains, in the protein structure, which move approximately as rigid bodies. A simplified, modular description of the molecule in terms of quasi-rigid domains could thus help the understanding and the description of the functionally-relevant motions of biomolecules.

In the present chapter we shall discuss a scheme to optimally subdivide a protein in a preassigned number of domains on the basis of its collective fluctuations. Specifically, the internal dynamics of a protein, in the form of a MD trajectory, covariance matrix PCA or ENM modes, is used to group amino acids in clusters having the least possible internal distance fluctuation. The method, which is based on the simple and effective idea of implementing the very definition of rigid body, is developed, tested and applied to specific test-cases of single proteins and in the context of dataset-wide investigations.

## 3.1 Features and limitations of available dynamical domain decomposition methods

The problem of identifying groups of amino acids in a protein, which undergo a collective displacement in the molecule's equilibrium fluctuations, has been already addressed in the literature. Only a limited number of methods have been developed so far (31; 63; 64; 65; 66; 67) to decompose a protein in domains on the basis of their internal dynamics, making use of different domain definitions and domain-decomposition algorithms. In the following paragraphs we shall review some of the most common approaches.

### 3.1.1 Geometrical deformation

One of the first methods introduced to identify quasi-rigid regions of a protein is the analysis of the degree of deformation internal to the molecule (31). In this scheme, two different protein conformations are compared to find those parts which experience a relatively high geometrical deformation, or *strain*. The starting point is the calculation of a *deformation energy*, given by:

$$E_i = \frac{1}{2} \sum_{j=1}^{N} K(|\vec{d}_{ij}^{0}|) \left[ |\vec{d}_{ij}^{0} + \vec{v}_i - \vec{v}_j| - |\vec{d}_{ij}^{0}| \right]^2 \tag{3.1}$$

where $\vec{d}_{ij}^{0}$ indicates the distance between residues $i$ and $j$ in the reference conformation, and $\vec{v}_i$ is the displacement vector of residue $i$. The 'elastic constant' $K(|\vec{d}_{ij}^{0}|)$ depends on the distance between the residues, and is usually taken to decrease exponentially with $|\vec{d}_{ij}^{0}|$.

The different structures that are compared can, in principle, come from crystallographic experiments or from the deformation of a reference structure along a normal mode of fluctuation. In particular, the difference vector connecting two conformations (for example, the open and closed structures of adenylate kinase) results in a *finite* displacement, while modes describe *infinitesimal* motions. In the latter case, the expression in Eq. 3.1 admits a simple quadratic approximation (see Eq. 1.8) leading to:

$$E_i = \frac{1}{2} \sum_{j=1}^{N} K(|\vec{d}_{ij}^{0}|) \frac{|(\vec{v}_i - \vec{v}_j) \cdot \vec{d}_{ij}^{0}|^2}{|\vec{d}_{ij}^{0}|^2} \tag{3.2}$$

The analysis of the elastic strain profile can give indications of those regions of the protein which undergo an important local deformation. This information can prove useful to identify quasi-rigid domains in the molecule structure, since it is assumed that

a part of the protein which moves approximately as a rigid body does not experience a high strain. On the contrary, if a residue shows a high value of the deformation energy it is likely that it belongs to a *hinge* between two domains.

This scheme is simple and intuitive; nonetheless, it suffers several limitations. First, the experimental error in crystallographic structures can result in high strain energies scattered throughout the molecule, making it difficult to identify a sharp transition from a rigid to a flexible region; moreover, it does not make use of the information about the proximity of the residues, thus requiring the direct inspection of the structure and the local deformation values. On top of that, the lack of a high strain region does not necessarily imply the absence of a hinge region, since a sharp transition from one domain to the next can be not captured by the deformation measure.

### 3.1.2 Rigid-body motion of amino acids groups

In order to overcome these limitations, Hinsen (31; 63) proposed an approach which identifies protein regions undergoing a coherent motion by using the criterion of searching for subregions whose motion is described by similar rotational-translational parameters.

The algorithm of Hinsen first performs a subdivision of the protein in groups of spatially-close residues; for example the space is partitioned with a three-dimensional cubic grid, and the residues falling in the same cube form a group; those groups consisting of less than three residues are neglected.

The elimination of groups with one or two residues is required by the fact that the six rigid-body parameters of the motion are well-defined only for objects of at least three non-collinear points. These six parameters, which are obtained from a least-square fit of the reference structure onto the 'deformed' structure, characterise a given group in terms of the six-dimensional space of roto-translations. The metric introduced by Hinsen to measure the distance between two groups in the rigid-body motion parameter space is given by:

$$S_{ij} = 3\frac{|\vec{\phi}_i + \vec{\phi}_j|}{|\vec{\phi}_i - \vec{\phi}_j|} + \frac{|\vec{t}_i + \vec{t}_j|}{|\vec{t}_i - \vec{t}_j|} \tag{3.3}$$

where $\vec{\phi}$ and $\vec{t}$ indicate, respectively, the rotation angle and the translation vector of a residue group; the factor 3 in front of the rotation component of the metric is due to the empirical observation that the former is a better domain identifier with respect to the translation component. This distance is used in the context of a clustering algorithm, which is next applied to gather those groups having similar parameters and therefore undergoing similar motions. The number and dimension of the resulting clusters depend on a coarseness parameter $c$, which fixes the minimum value of the $S$

distance for two groups to be assigned to the same cluster. The dynamical domains are finally identified with the amino acid clusters.

This approach takes as input two structures, or a reference structure and a deformation vector. The identification of the domains requires the specification of the coarseness parameter. The number of resulting clusters is not fixed a priori, since it depends on this clustering tolerance and on the very properties of the protein's internal dynamics.

### 3.1.3 DynDom

A clustering strategy related to the ones described so far is employed by the DynDom (64) web-server. In this case a comparison is done between two protein structures. The building blocks consist of short segments of the protein backbone, that are lumped together based on the similarity of their rigid-body motion parameters; the latter are identified making use of a clustering algorithm. In order to ascertain if the residue grouping is physically meaningful, the hinge axes of two groups are compared: the groups therefore form a dynamical domain if close-by regions perform a continuous deformation. The number of domains identified by the algorithm is determined as the largest for which the backbone-connectedness is preserved and a user-defined criterion is satisfied, related to the ratio between inter- and intra-domain motion.

As for the methods described so far, DynDom requires the input of two different conformers. However, two crystallographic conformers of the same protein may not always be available, or in the case of MD trajectories, there may be too many conformers among which to choose the two structures.

### 3.1.4 Methods based on the amino acid correlation

Other dynamics-based grouping schemes of amino acids have been devised, based on positive correlations of amino acid displacements entailed by a single low-energy mode (65) or from pairwise correlations in the covariance matrix itself (66). The basic idea behind these algorithms is that amino acids, or groups of amino acids, performing a rigid-body motion, are displaced approximately in the same direction: therefore, they must show a positive correlation of the dynamics. On the other hand, negative or quasi-zero correlations are assumed to indicate a small degree of coherence in the motion.

In the *Hierarchical Clustering Correlation Pattern* (HCCP) scheme, for example, the dynamical domains of a protein are identified with a clustering procedure gathering residues on the basis of their covariance matrix elements. At the first step of the algorithm each residue represents a cluster itself; next, two residues are grouped in a

single cluster if their pairwise correlation $c_{ij}$ is above a given threshold. Once a 'temporary' set of clusters has been defined, the covariance among the clusters is recalculated according to:

$$c_{ij}^{'} = \frac{1}{m_i m_j} \sum_{k \in M_i} \sum_{l \in M_j} c_{kl} \tag{3.4}$$

where $M_i$ is the $m_i$-dimensional vector indexing the residues in the $i$-th cluster. This clustering procedure is iterated until the whole protein is a unique cluster itself: the identification of the optimal decomposition is then done performed inspecting the domain partitions at the various stages of the iteration procedure.

Further insight is obtained by the calculation of the Hierarchical Clustering of the Correlation Patterns: the latter consists in the linear correlation among two full rows of the covariance matrix, obtained as:

$$p_{ij} = \frac{\frac{1}{N} \sum_{k=1}^{N} c_{ik} c_{kj} - \bar{c}_i \bar{c}_j}{\sigma_i \sigma_j} \tag{3.5}$$

where $\bar{c}_i$ and $\sigma_i$ are, respectively, the mean and the standard deviation of the covariance matrix row $c_i$. The matrix $p_{ij}$,which undergoes the same clustering procedure described above, provides information about the correlations among groups of residues, rather than pairwise correlations. Therefore, it is assumed to be more robust than the covariance matrix $c_{ij}$ for the identification of the domains.

Schemes like the HCCP, or other methods based on the positive correlation of their motion, are inherently affected by a conceptual limitation. In fact, if on the one hand a positive correlation reasonably suggests a similarity of the motion, it cannot be concluded *a priori* that a negative correlation between two residues implies that they do not move as a rigid body. For example, consider a group of amino acids performing a rigid-body rotation about an axis internal to the group: pairs of points, lying at the two opposite edges of the body will be negatively correlated (see Fig. 3.1), even if their motion is definitely rigid. The aforementioned methods are insensitive to the anti-correlation of pivotal motions, and would identify as disconnected regions of the molecule groups of amino acids which fluctuate in a genuinely coherent manner.

### 3.1.5 Translation/Libration/Screw-like motion (TLS)

A further interesting approach is offered by the TLS (translation, libration, screw-like motion) analysis introduced by Schomaker and Trueblood (68). The method permits the determination of the local mean square fluctuations compatible with the rigid-body

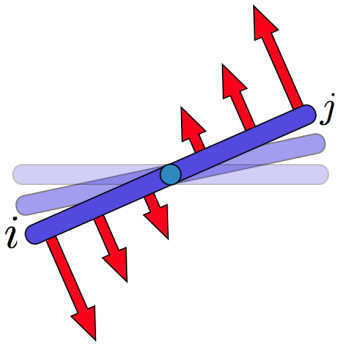


**Figure 3.1: Example of anti-correlated rigid body motion** - The rotation of an object around a pivot is undeniably a rigid-body motion; nonetheless, the scalar product of the displacement vectors on the two sides of the body is negative. Similarly, an anti-correlated motion with negative covariance matrix entry does not necessarily imply a non-rigid coherent displacement.

motion of a group of atoms. The pattern of these fluctuations can, in turn, be compared with crystallographic data.

Originally, it was developed to evaluate the reliability of small molecule crystallographic structures. The effectiveness of the method motivated its use to identify those regions of a protein whose rigid-body motion is optimally compatible with the fluctuation pattern described by the B-factors (69).

This method represents an interesting framework to characterise the rigid-like motion of a protein's subparts. Unfortunately, for reasons of computational efficiency, the combinatorial space of the possible assignments of amino acids to various rigid-like domains must be explored in a stochastic (non-exhaustive) manner. In particular, the TLSMD (70) web server explores all the possible domain subdivision of a protein chain in a given number of uninterrupted segments (from two on), and returns the one which best fits with the B-factors. On the other hand, it is not possible to identify groups of amino acids constituted by disconnected segments of the sequence. Moreover, the B-factors are isotropic quantities which do not provide any information about the direction of the motion, but are limited to its global extension.

## 3.2 Optimal subdivision of a protein in quasi-rigid domains

In the following of this chapter we shall discuss and apply a variational scheme for the identification of nearly-rigid protein subparts. The rigid-like character of the groups, or dynamical domains, is identified directly from a variational principle where no prior

assumption is made on the proximity in sequence or space of the grouped amino acids. The number of domains, $Q$, in which the molecule is partitioned, is assumed to be pre-assigned; the optimal choice of $Q$ can be made on the basis of criteria that will be later discussed.

### 3.2.1 Identification of quasi-rigid domains from a MD trajectory

We start by discussing an algorithm to optimally partition a protein into a pre-assigned number $Q$ of domains, based on the data collected in an atomistic MD simulation. The central idea is that the optimal subdivision is the one for which the motion of the domains is as close as possible to a rigid-body motion.

Consider therefore an atomistic MD simulation of a protein, of which we shall neglect the non-$C_\alpha$ atoms for simplicity; and its reference structure, i.e. the 'frame' of the simulation closest to the average.

Our criterion to subdivide the protein into domains can be stated as follows: *we are interested in assigning the amino acids of the protein to a given number $Q$ of domains; the optimal partition is the one maximising, over all possible assignments, the contribution to the displacement of the putative domains which can be ascribed to a rigid-body motion.*

To do so, for each tentative partitioning of the amino acids we consider the instantaneous displacement vector, $\vec{v}_q(t)$, of a putative domain; this vector connects the coordinates of the domain in the reference structure, $\vec{r}_q^{\,0}$, to the corresponding coordinates in a given MD trajectory frame:

$$\vec{r}_q(t) = \vec{r}_q^{\,0} + \vec{v}_q(t) \tag{3.6}$$

where the subscript $q$ labels the putative domain. The vector $\vec{v}_q(t)$ can, in turn, be separated in two contributions: $\vec{v}_q^{rb}(t)$, corresponding to a rigid roto-translation of the domain, and $\Delta\vec{v}_q(t)$, which takes into account the fluctuations internal to the domain:

$$\vec{v}_q(t) = \vec{v}_q^{rb}(t) + \Delta\vec{v}_q(t) \tag{3.7}$$

The rigid-body component $\vec{v}^{rb}(t)$ can be decomposed in a translation vector $\vec{\tau}_q(t)$ and a rotation parametrized by the matrix $\mathcal{R}$ and the vector $\vec{\omega}_q(t)$:

$$\vec{v}_q^{rb}(t) = \vec{\tau}_q(t) + \mathcal{R}[\vec{\omega}_q(t)](\vec{r}_q^{\,0} - \vec{R}_q) \tag{3.8}$$

where $\vec{R}_q$ are the coordinates of the $q$-th domain's centre of mass.

The matrix $\mathcal{R}$ can be calculated by means of the Kabsch algorithm (71), which finds the optimal rotation of to sets of points minimising the RMSD between them.

The extremality conditions that are imposed to find $\vec{\tau}_q(t)$ and $\vec{\omega}_q(t)$ guarantee the orthogonality between the rigid-body displacement $\vec{v}_q^{rb}(t)$ and the internal fluctuation term $\Delta\vec{v}_q(t)$:

$$\vec{v}_q^{rb}(t) \cdot \Delta\vec{v}_q(t) = 0 \ \forall t \tag{3.9}$$

Eq. 3.9 allows to decompose the total mean square fluctuation of the molecule in two contributions:

$$\begin{aligned} \mathrm{MSF} &\equiv \sum_{q=1}^{Q} \langle |\vec{r}_q - \vec{r}_q^{\,0}|^2 \rangle = \sum_{q=1}^{Q} \langle |\vec{v}_q|^2 \rangle = \\ &= \sum_{q=1}^{Q} \langle |\vec{v}_q^{rb}|^2 + |\Delta\vec{v}_q|^2 \rangle = \mathrm{MSF}^{\parallel} + \mathrm{MSF}^{\perp} \end{aligned} \tag{3.10}$$

where:

$$\begin{aligned} \mathrm{MSF}^{\parallel} &\equiv \sum_{q=1}^{Q} \langle |\vec{v}_q^{rb}|^2 \rangle \\ \mathrm{MSF}^{\perp} &\equiv \sum_{q=1}^{Q} \langle |\Delta\vec{v}_q|^2 \rangle \end{aligned} \tag{3.11}$$

A natural and intuitive definition of the optimal protein partition into a pre-assigned number $Q$ of domains, is the one which maximises, over all possible partitions, the rigid-body component of the mean square fluctuation, $\mathrm{MSF}^{\parallel}$, or equivalently, the partition which minimises the internal fluctuation $\mathrm{MSF}^{\perp}$.

This method is completely general, since no assumption is made on the contiguity in space or sequence of the residues of the domains: in principle, any possible assignment of the residues to the domains is tried for a given $Q$, and the optimal choice is performed on the quantitative basis of the internal distance fluctuations of the domain.

A further simplification can be done, in order to characterise, in the simplest possible terms, the salient features of the protein internal dynamics. Specifically, we can look for the simplest constrained motion of the domains which best captures the mean square fluctuation of the protein.

As an example, we can force the motion of the domains to be composed exclusively by rigid rotations about fixed axes passing through hinge points $\vec{\rho}_q$: this amounts at modifying Eq. 3.8 as follows:

$$\vec{v}_q^{rb}(t) = \mathcal{R}[\omega_q(t); \hat{n}](\vec{r}_q^0 - \vec{\rho}_q) \tag{3.12}$$

Note that the rotation matrix now depends parametrically on the rotation axis $\hat{n}$, and the instantaneous angle $\omega_q(t)$ is the one which maximises the rigid-like component of the domain MSF, $|\vec{v}_q^{rb}(t)|^2$.

For a given domain $q$, the corresponding optimal rotation axis $\hat{n}_q$ is found as the one maximising the captured MSF of the domain, defined as:

$$\mathrm{MSF}_q^{\parallel} = \langle |\vec{v}_q^{rb}|^2 \rangle \tag{3.13}$$

The total amount of the overall mobility, which is captured by the rigid rotation of the the domains, is calculated as in Eq. 3.11.

The quantity $\mathrm{MSF}^{\parallel}$ of Eq. 3.13 represents the amount of the protein internal motion which can be captured if the moving parts are allowed to perform only rigid-body rotations.

The specific characteristic of this scheme is that the motion of a quasi-rigid domain is described by the rotation angle about the axis with respect to the reference frame. Once the optimal axes are known, it is possible to characterise the internal dynamics of a protein in terms of nonlinear, continuous coordinates, namely the rotation angles of the domains. Examples of insight provided by this scheme will be given, in the following section, for two proteins, namely adenylate kinase and HIV1 protease.

### 3.2.2 Simplification of the algorithm making use of the essential spaces

The optimal domain partition scheme described so far requires a substantial computational effort: in fact, the number of possible partitions grows exponentially with the number of domains $Q$ and the protein length $N$. Even for small proteins and a small number of blocks the calculation of $\mathrm{MSF}^{\parallel}$ is slow, due to the fact that the instantaneous roto-translational parameters, namely $\vec{\tau}_q(t)$ and $\vec{\omega}_q(t)$ in Eq. 3.8, must be obtained for each domain and each frame of the MD simulation.

In order to reduce the calculations, one can simplify the algorithm replacing the time average of the rigid-body component of the instantaneous fluctuation with a weighted average of the rigid-body component of the covariance matrix eigenmodes. In fact, the relevant information to subdivide a protein in large groups of quasi-rigid domains is mainly encoded in the concerted movements entailed by the collective modes of the covariance; it is therefore possible to adapt the scheme devised so far to find the

partition of the protein which maximises the rigid-body component of the low-energy modes of fluctuation. The quantity to be maximised is therefore given by:

$$f \equiv \frac{\mathrm{MSF}^{\parallel}}{\mathrm{MSF}} = \frac{\sum_{\ell=1}^{n} \lambda_\ell \, |\vec{v}_\ell^{\,rb}|^2}{\sum_{\ell=1}^{n} \lambda_\ell} \tag{3.14}$$

where $\vec{v}_\ell^{\,rb}$ and $\lambda_\ell$ are, respectively, the $\ell$-th mode rigid-body component and the eigenvalue of the covariance matrix. The number of modes $n$, which is left as an input parameter, can be chosen so to capture a preassigned fraction of the overall internal fluctuation. If the full eigenspace of the covariance were considered for the sum in Eq. 3.14, this scheme would be completely equivalent to the one described in the previous paragraph (provided that the $\mathrm{MSF}^{\parallel}$ is normalised to the total MSF).

Also the optimal axes of rotation can be identified making use of the essential spaces: the only requirement is that the rigid rotation component which is maximised in Eq. 3.12 is obtained from the modes of the covariance.

It is worth noting that the use of the modes leads to an important generalisation of the method. In fact, one can apply the algorithms discussed so far to low-energy modes obtained from elastic network models: it is therefore possible to perform a domain subdivision of a protein, and the identification of its optimal rotation axes, also when MD simulations are not available. This possibility is particularly relevant if one wants to investigate the internal dynamics of large proteins, which obviously require time-consuming MD simulations.

### 3.2.3 Optimised strategy: mapping to a Potts-model colouring problem

In the previous paragraph we discussed a simplified strategy to optimally identify quasi-rigid domains making use of the essential dynamical space. This method can appreciably reduce the computational effort required by the frame-per-frame calculations of the MD-trajectory based scheme. Nonetheless, the large number of possible partitions of the protein makes it impossible to perform a fast computation: the number of repeated matrix operations involved is so large that it is not computationally convenient to optimise directly the quantity $f$.

A more effective strategy is to perform a preliminary exploration of the configuration space by optimising, with respect to the tentative partitions of the molecule, a simple objective function, in order to efficiently identify a candidate subdivision over which $f$ is finally evaluated and maximised.

The objective function that we consider is the following one:

$$F(\{\sigma\}) = \frac{1}{2}\sum_{i\neq j}\delta_{\sigma_i,\sigma_j}\sum_{\ell=1}^{n}\lambda_\ell[(\vec{v}_\ell(i)-\vec{v}_\ell(j))\cdot\vec{d}^0_{ij}]^2 + \frac{\alpha}{2}\sum_{i\neq j}(1-\delta_{\sigma_i,\sigma_j})\frac{1+\tanh(R_c-|\vec{d}^0_{ij}|)}{2} \tag{3.15}$$

where $\sigma_i = 1 \cdots Q$ labels the group to which amino acid $i$ belongs to, $\delta$ is the Kronecker delta, $\vec{d}^0_{ij}$ is the distance vector of amino acids $i$ and $j$ in the reference conformation, $R_c$ is an interaction cutoff distance set equal to 7 Å and $n = 10$. The sought optimal grouping of amino acids is the one which *minimizes* $F$: for systems consisting of truly-rigid subparts this will be analogous to *maximizing* $f$ of Eq. 3.14.

The first term in the sum represents the cost of the average elastic energy associated to the internal deformation of the molecule. This term penalises fluctuations in the distance of any two points belonging to the same putatively-rigid group, consistently with the definition of rigid bodies. The second term introduces a penalty, controlled by the parameter $\alpha \geq 0$, for dynamical domains consisting of regions that are disconnected in space. Upon increasing $\alpha$, in fact, the term disfavours the number of pairs of neighbouring amino acids (those closer than the cutoff distance $R_c = 7$ Å) that belong to different groups. The optimisation of $F$, therefore, leads to group assignments that minimise the interface area between the groups, while not strictly enforcing the spatial compactness of the domains. The minimisation of $F$ can be straightforwardly performed within a simulated annealing protocol, with elementary moves corresponding to changes of the group assignment of individual amino acids. The corresponding changes of $F$ only require the summation of $N-1$ pre-calculated quantities ($N$ being the number of amino acids in the protein) corresponding to the interaction terms among the re-assigned amino acid and all the other ones.

The search for the optimal solution is carried out separately for increasing values of $\alpha$. Eventually, for a large enough value of $\alpha$ the presence of boundaries is forbidden and a single dynamical domain is returned by the minimisation of $F$: therefore, in the intermediate range of values the algorithm could find solutions having fewer groups than $Q$, which are discarded. The solution with $Q$ domains corresponding to the largest value of $f$ will be taken as the one corresponding to the best subdivision.

## 3.3 Applications

As a first example we shall apply the rigid block decomposition method to adenylate kinase and HIV-1 protease. The decomposition is performed on the basis of data from atomistic molecular dynamics simulations (for Adk, the MD simulation performed by

Pontiggia *et al.* (15)) and from elastic network models (for HIV-1 PR). In both cases, the analyses will be complemented with the application of the optimal axis identification scheme. Subsequently, the method will be applied for a rigid-subunit decomposition of two sets of proteins. The first set consists of monomeric enzymes, representing of the main CATH structural classes (6) of hydrolases (class 3 according to EC (72)). For these enzymes we investigate the existence of systematic biases in the location of the known catalytic site with respect to the boundaries separating primary dynamical subdomains. We conclude the analysis by investigating the extent to which the optimal subdivision returns groups of residues that span uninterrupted stretches of the primary sequence or occupy compact regions in space.

### 3.3.1 Test-case i: Adenylate Kinase

The rigid-block decomposition scheme, that we shall apply here to Adk, provides a natural and objective scheme for assessing if, and to what extent, the molecule's internal dynamics can be described in terms of few parts that move as nearly-rigid units. We begin by considering the fluctuations within the sub-state where the 50ns-long trajectory started from the open structure, 4ake, dwelled for about 10 ns (15). The reference structure for the sub-state, which is the most populated of the MD trajectory, is provided in Fig. 3.2, along with the representation of the lowest energy mode. The mobility the Lid and of the AMP-binding subdomains, corresponding to regions 117-164 and 30-64 respectively, is evident.

The $n = 10$ lowest energy modes within the sub-state were used to subdivide the enzyme into $Q = 2, 3...10$ dynamical domains. A representation of the subdivisions into 3 and 4 groups are provided in Figs. 3.3a-b. The fraction of essential dynamics motion, see Eq. 3.14, captured by the various subdivisions is shown in Fig. 3.4.

The graph indicates that a very limited number of dynamical domains is already sufficient to account for most of the essential dynamics. In fact, subdivisions into $Q = 2$, 3 and 4 blocks capture as much as 52%, 77% and 83% of the fluctuations entailed by the $n = 10$ essential modes (which account for the 80% of the overall mobility).

The subdivision for $Q = 2$ identifies region 122-156 as an approximately-rigid, but highly mobile, unit. The region overlaps well with the Lid indicated before. The less mobile AMP-binding domain is identified as a distinct unit when using $Q = 3$. In fact, for $Q = 3$, the regions corresponding to the two mobile nearly-rigid subdomains are 122-158 and 32-59, and are compatible with the customary tripartite subdomain division of Adk. If the entire 50-ns long trajectory is used rather than the most populated sub-state it is found that the boundary of the AMP-binding domains is virtually unaltered (sequence interval 32-60). The larger configurational space spanned by the more mobile Lid domain instead reflects into an extension of the both the left and right subdomain boundaries by about ten residues, thus covering the interval 112-167.

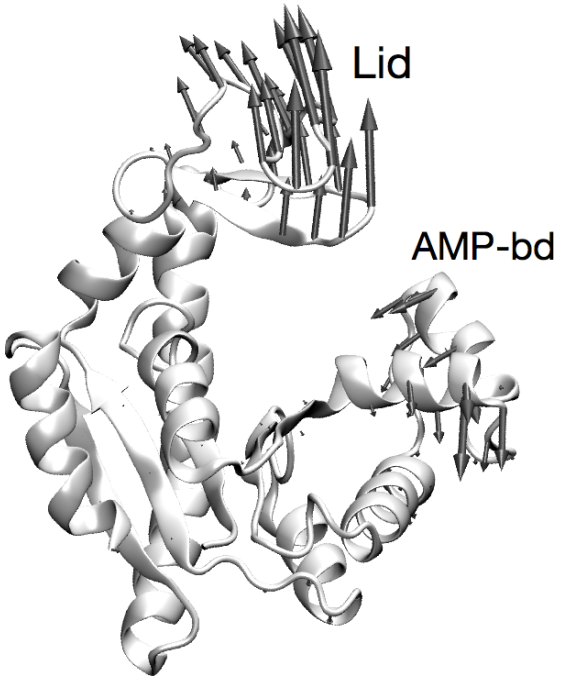


**Figure 3.2: First essential mode of E.Coli adenylate kinase** - The length of the arrows has been enhanced for sake of clarity.

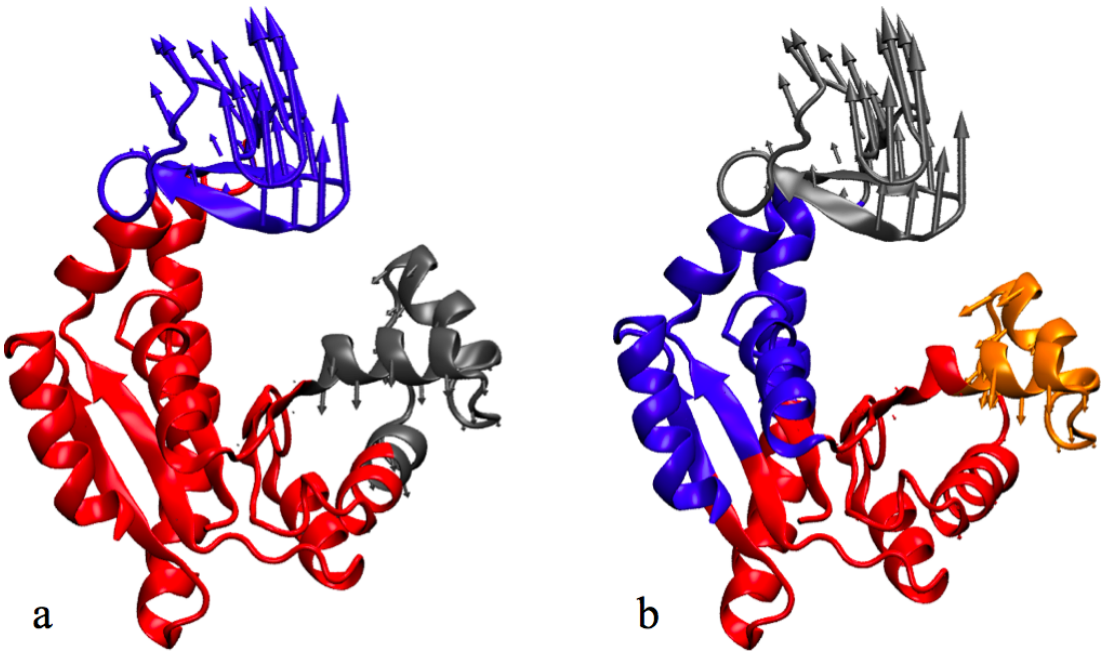


**Figure 3.3: Quasi-rigid domain decomposition of adenylate kinase** - Figure *a* shows the decomposition in three quasi-rigid domains, while figure *b* shows the partition in four blocks.

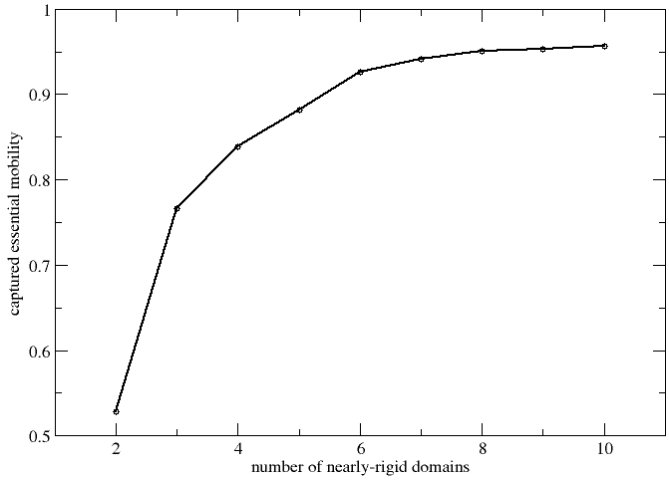


**Figure 3.4: Fraction of essential dynamics captured by the quasi-rigid domain decomposition of Adk** - The fraction of essential dynamical motion (see Eq. 3.14) captured by the subdivision of Adk into $Q = 2 \dots 10$ rigid domains is here shown.

Decomposing Adk in $Q = 3$ units 5 sequence intervals are found: one for each of the two mobile domains and three for the nearly-fixed core. It is interesting to compare this dynamics-based subdivision with the one provided by the TLS (70) analysis of crystallographic data which enforces the sequence continuity of each rigid-block. The TLS decomposition of 4ake into five intervals (as many as those found with $Q = 3$) returns the following segments: 1-27, 78-116 and 171-214, identifiable with the core, and 28-77 and 117-170, compatible with the AMP-bd and Lid subdomain, respectively. With the exception of one of the AMP-bd/core boundaries, the TLS subdivisions and those of our analysis of the full MD trajectory are mismatched by only about 5 residues. They are hence generally consistent, despite the differences not only in method but also for the nature of the input data (crystallographic B-factors for TLS and MD data for our method). An important distinction between the two results is, however, that the three segments constituting the core regions are encompassed in a single rigid unit by the present variational method, while are treated as independent ones within the TLS scheme.

Our optimal 3-domain subdivision was compared also with the one returned by the DynDom server (64) which requires the input of two structures representing the conformational variability of the molecule of interest. Accordingly, from the set of MD-sampled conformers we selected the pair with the largest RMSD. DynDom returned a subdivision in two domains, the smallest corresponding to the Lid (sequence interval 110-169) plus a small loop (residues 6–12) and the other to the core plus the AMP-binding domain. Interestingly, this latter subdomain is recognised as a separate dynamical domain if the open and closed crystallographic conformers of Adk (1ake, 4ake) are used as input structures.

We conclude the analysis of the partitions of Adk considering $Q = 4$ domains. With respect to the $Q = 3$ case, the boundaries of the two mobile domains are only slightly adjusted to 35-60 and 118-159, respectively. However, a new domain, comprising several sequence segments 7–25; 108–117; 160–174; 195–214, is identified at the interface between the core and the Lid. This group of 'hinge' residues have consistently been shown, by independent methods (15; 73), to be subject to a significant strain during the free enzyme dynamical evolution. In Fig. 3.5 we show the pile-up of the sequence subdivisions in $Q = 2, 3$ and 4 domains: it is worth to note that the sequence interruptions between two contiguous domain segments are preserved within a few amino acids.

The application of the quasi-rigid domain identification scheme to the adenylate kinase returned reasonable results. The partitions of the molecule in different numbers of domains proved to be consistent with those identified by methods previously discussed in the literature. Moreover, the large fraction of essential dynamics which is captured

by the optimal decompositions indicates that the large-scale motion of this molecule can be reliably described as the rigid-body motion of a small number of subparts.

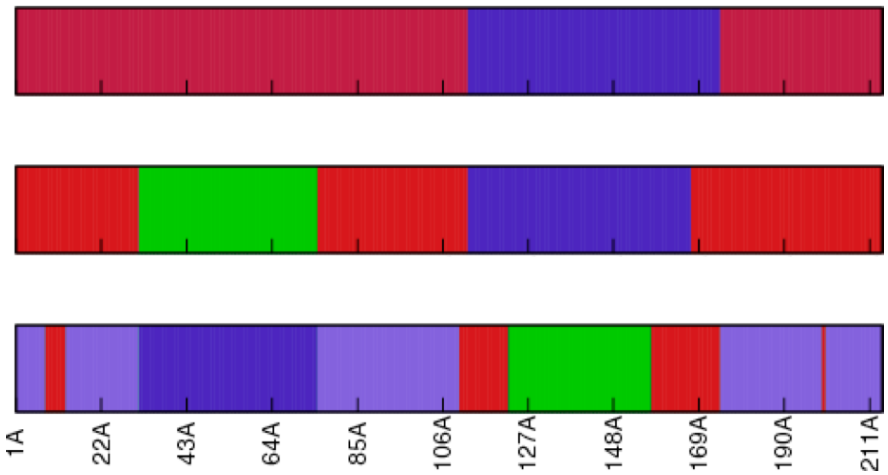


**Figure 3.5: Sequence partition of Adk in 2, 3 and 4 domains** - The domains of the Adk subdivisions in 2, 3 and 4 quasi-rigid blocks are indicated on the amino acid sequence in different colours. It is worth noting that the interruptions about residues 33, 74, 110 and 170 are preserved among the different partitions.

The aforementioned simplification of the dynamics can be proficiently performed with the application of the scheme to optimally identify the domain axes of rotation. In order to highlight the effectiveness of the method in capturing large-scale motions, we perform this analysis on the whole 50 ns-long MD trajectory discussed in (15). The chosen partition of the protein is the one of the customary $Q = 3$ domains case previously discussed: this results in a fixed core and two mobile satellite domains coinciding with the AMP binding domain and the Lid.

The fraction of the total mean square fluctuation, which is captured forcing the AMP-bd and the Lid to move as rigid rotating bodies, amounts to $\sim$ 80%. This remarkable result is compatible with the 3 blocks quasi-rigid domain decomposition based on the PCA, which captures 77% of the top 10 modes of fluctuation, and confirms the modular character of this molecule structure. Even more interesting are the rotation axes, shown in Fig. 3.6: the open-close movement of the molecule can be readily perceived by the axes orientation. It is worth noting that at both AMP-bd and Lid interfaces with the core, the rotation axes passes through two $C_\alpha$ atoms: this can be attributed to the fact that the moving domains are connected to the core by two similarly-bending hinges.

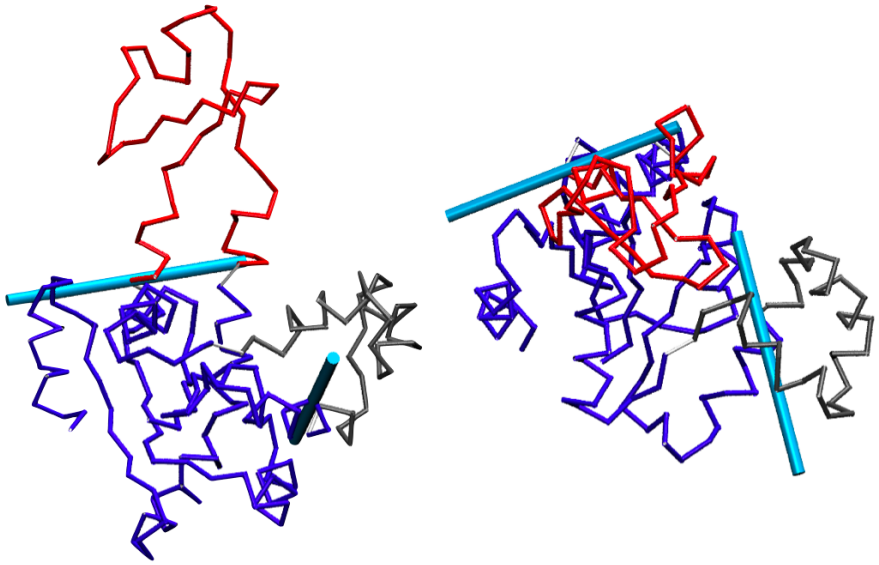

**Figure 3.6: Optimal axes of rotation of Adk** - The reference structure of Adk is shown along with the optimal rotation axes.

### 3.3.2 Test-case ii: HIV-1 PR

As a second example of the dynamics-based decomposition we consider the HIV-1 protease dimer complexed with a peptide substrate.

HIV-1 PR is a 198-residue-long homodimeric aspartic protease (see Fig. 3.7). It is essential for the life cycle of the human immunodeficiency virus (HIV): in fact, its function is to cleave the polyproteic chains to allow the newly synthesised viral proteins to fold. Being a central actor in the life cycle of HIV virus, this protein has been the subject of numerous studies; in particular, it has been pinpointed as a target of inhibiting drugs hindering the cleavage of viral proteins and preventing the infection to spread. Unfortunately, the high mutation rate of this protein strongly limits the efficacy of these drugs, which must necessarily be used together with other pharmaceuticals to attack simultaneously various facets of the virus' replication cycle.

One of the earliest all-atom MD simulations of HIV-1 PR was performed by Piana *et al.* (56). This 10-ns long simulation shed light on the motion performed by the *flaps* of the dimer, and the functional role of these collective displacements of about 25 amino acids in an open-close fashion. In fact, the protein acts as a molecular scissor both chemically and mechanically, by first opening the flaps to accommodate the substrate in the catalytic pocket, then closing to stretch the polyprotein in a $\beta$-

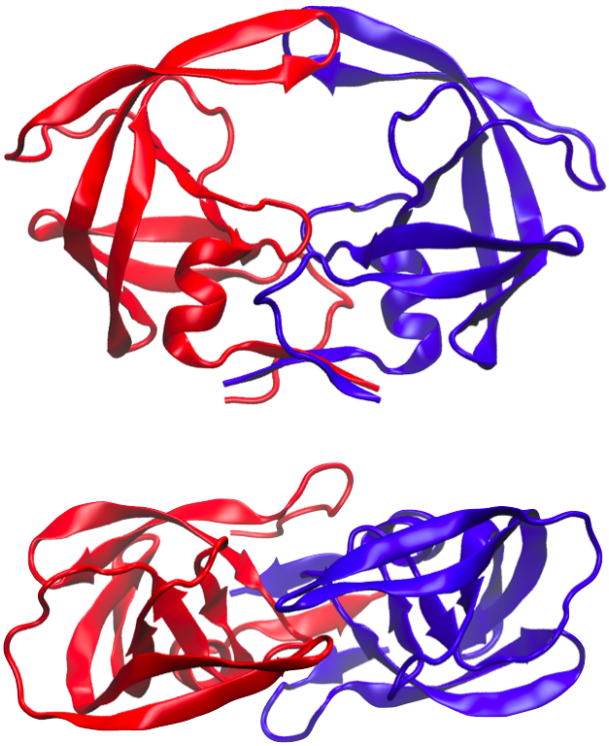

**Figure 3.7: Crystal structure of the HIV-1 protease** - Front and top views of the HIV-1 protease. The two 99-residue-long dimers are shown with different colours.

extended conformation to perform the cleavage reaction. Moreover, the analysis of the MD trajectory put in evidence a mechanical coupling between the catalytically active aspartic dyad and a distal site, separated by about 25 Å(37; 56). This long-range coupling plays a central role in the viral resistance to inhibiting drugs: in fact, mutations, which reduced dramatically the drug efficiency were found in the proximity of residues mechanically coupled with the aspartic dyads (37; 56; 74).

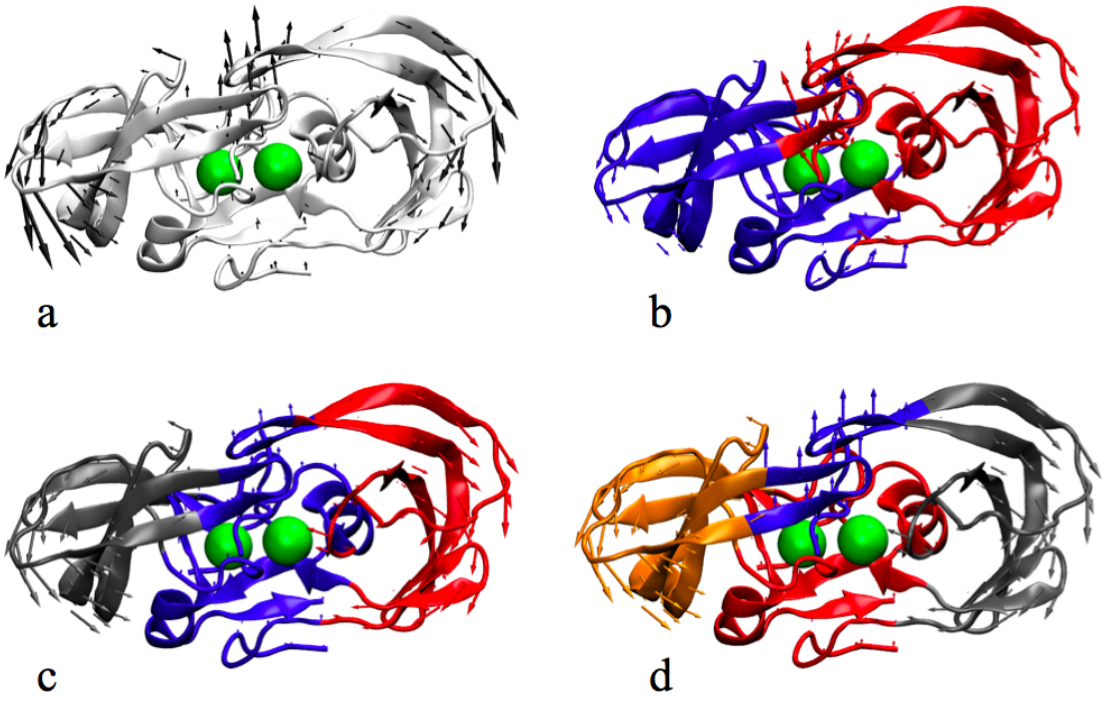


**Figure 3.8: Quasi-rigid domain decomposition of adenylate kinase** - Figure *a*: first essential mode of HIV-1 protease; figure *b*, *c*, *d*: decomposition of HIV-1 PR in two, three and four quasi-rigid domains, respectively.

To illustrate the applicability of the domain decomposition method in the absence of data from atomistic simulations, we obtained the essential dynamical spaces from the $\beta$-GM approach (37). The complex shown in Fig. 3.8a, which corresponds to the equilibrium structure of the MD study of ref. (56), was subdivided from 2 to 10 domains, see Fig. 3.8b-d. The fraction of internal dynamics captured by the various decompositions is shown in Fig. 3.9. The curve has a slightly slower increasing trend compared to Adk (3.4). In fact, when $Q = 3, 4$ domains are used, about 72% and 79%, respectively, of the essential dynamical fluctuations is captured for the HIV-1 protease/substrate complex.

The subdivisions into $Q = 2, 3, 4$ approximately-rigid units are represented in Fig. 3.8b-d. As for Adk, the units are compactly-organised in space but do not cover a single

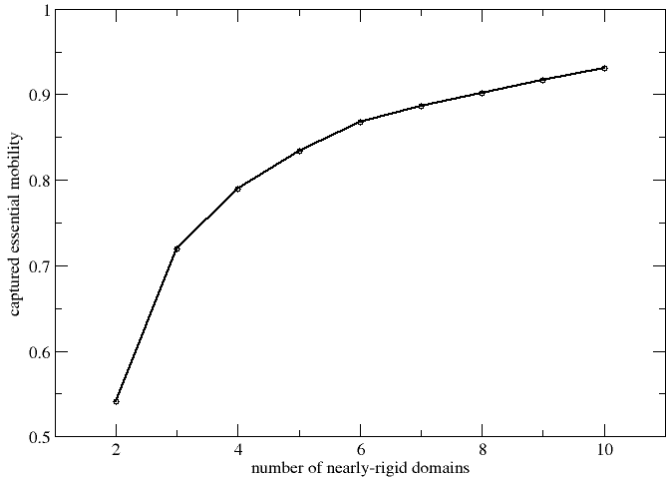


**Figure 3.9: Fraction of essential dynamics captured by the quasi-rigid domain decomposition of HIV-1 PR** - The fraction of essential dynamical motion (see Eq. 3.14) captured by the subdivision of HIV-1 protease into $Q = 2 \dots 10$ rigid domains is here shown.

stretch of the primary sequence. The sequence-disconnected nature of the domains does not lend itself simply to a detailed comparison with the TLS decomposition for the same number of blocks. We shall therefore restrict ourselves to consider the primary 'hinge points', represented by amino acids 20, 35, 57, 70, that emerge from the pre-calculated subdivision offered by the TLS web-server of the HIV-1 PR monomer (PDB structure 1t3r) in 5-7 intervals. The first three hinges fall within 3 amino acids (along the primary sequence) from boundaries identified for the optimal subdivision in two primary domains, $Q = 2$ (see Appendix A), suggestive of good consistency.

We also compared our domain decomposition with the pre-calculated subdivision of the HIV-1 PR monomer offered by the DynDom server (based on structures 1aid and 1hsg). The returned subdivision consisted of two domains, the smaller one comprising segments 32-60, 75-77, and broadly corresponding to the monomer flap. Though it should be bourne in mind that the subdivision might depend on the fact that only one monomer is considered because multimers are not accepted by the DynDom server, the identified modular nature of the flaps is compatible with salient aspects of the TLS and variational decomposition.

The inspection of the optimal subdivisions in Fig. 3.8b-d prompts two considerations. The motion of the flaps is largely consistent with a coordinated rotatory movement around the central fulcrum regions. It is evident from the $Q = 3, 4$ cases that within the nearly-rigid parts comprised by the flaps the points at the two extremes are displaced in opposite directions. On one hand this feature illustrates that the mo-

tion of nearly-rigid units in proteins can be sufficiently general to permit the presence of anti-correlated motion within its constituent parts. On the other hand, the analysis supports the qualitative description of the flap motion first given by Piana *et al.* (56) based on the visual inspection of the first essential mode of a multi-ns MD simulation (see Fig. 6b in ref. (56)). Indeed, if the first mode only is used for decomposing the proteins into $Q = 2$ blocks, it is found that each entire flap is identified as a nearly-rigid unit (see Fig. A.1 in the Appendix).

The second observation regards the location of the catalytic site of HIV-1 protease with respect to the 'primary dynamical boundaries'. By the latter we mean the boundaries separating the most prominent rigid-like regions in a protein (i.e. when using $Q = 2$ or $Q = 3$). By inspecting Fig. 3.8b-d it is seen that the highlighted catalytic aspartic dyad, which has low mobility, straddles the rigid-domains interface for $Q = 2$ and is close to one or more domain boundaries for $Q = 3$ and 4. This suggests that the proximity of the catalytic amino acids to the primary dynamical boundary is instrumental for sustaining functionally-oriented large-scale fluctuations, as required by the cleavage reaction (75).

We conclude this analysis with the identification of the optimal axes of rotation of the HIV-1 protease flaps during the 10-ns long MD trajectory of the HIV1-PR dimer performed by Piana *et al.* (56). The 3 domain subdivision was considered. In order to allow the flaps to move freely, the tips, which were ascribed to the core by the partition method (see Fig. 3.8), were separated. The domains are thus defined as follows:

- **core** 1–33, 81–132, 180–198
- **flap 1** 34–80
- **flap 2** 133–179

Fig. 3.10 shows the result of the optimal rotation axis identification scheme. The fraction of internal dynamics that is captured when forcing the rotation of the flaps about the optimal axes amounts to 39 % of the overall mobility. This value reflects a lower degree of collectivity of the internal dynamics of HIV1-PR, compared to adenylate kinase; yet, it still amounts to a substantial portion of the overall MSF when it is considered that only *two* degrees of freedom (the instantaneous rotation angles) are used to describe the protein's internal dynamics.

Nonetheless, it is interesting to discuss the *direction* of the optimal rotation axes. The latter, in fact, present an orientation which is compatible with the one suggested by Piana *et. al* (56). The dynamics of the flaps is described as a rotation about axes perpendicular to the transverse plane (see Fig. 3.11): this picture, which was based on the direct visual inspection of the protein motion, is here supported by the position and orientation of the optimal axes of rotation found by our algorithm.

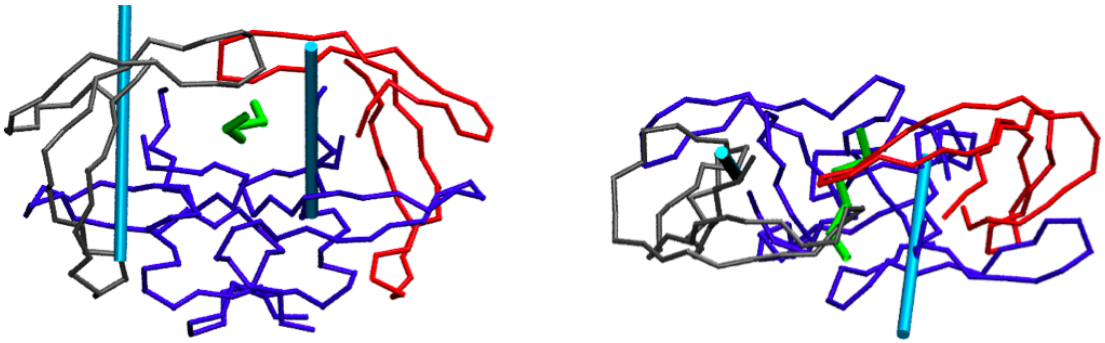

**Figure 3.10: Optimal rotation axes of HIV1-PR** - The reference structure of HIV1 protease is shown along with the optimal rotation axes. The orientation of the axes suggest a coherent rotatory motion of the flaps.

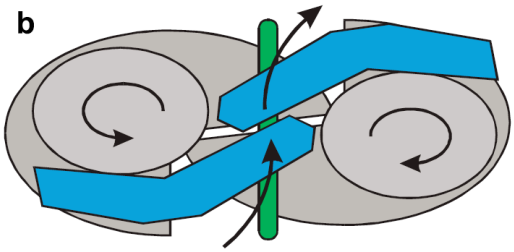


**Figure 3.11: Schematic description of the motion occurring in HIV1-PR** - The picture, reproduced from ref. (56), describes the concerted rotation of the HIV1-PR flaps. This motion is suggested by the largest-eigenvalue mode of the covariance matrix, obtained from the 10-ns atomistic MD simulation performed by Piana *et al.* (56).

### 3.3.3 Catalytic site location of EC3 class enzymes

Modulations of the region neighbouring the active site, analogous to those highlighted by the optimal axes of rotation of HIV-1 protease, have been found for several other proteolytic enzymes differing by catalytic chemistry and structural architecture (40; 76; 77). One is therefore led to ask wether this particular location of the active site near a primary dynamical boundary can be a common trait of other enzymes as for HIV1-PR.

The question appears particularly appealing also in view of considerations made by Del Sol *et al.* (61) that functional sites in proteins with allosteric behaviour are preferentially located at the boundary between regions that are very modular in terms of contacting amino acids.

The existence of common large scale movements in different proteases was taken as the starting point for examining what relationship, if any, exists between the location of the cleavage sites and the proximity of primary dynamical boundaries for other enzymes belonging to the class of hydrolases. The latter are particularly interesting in this context, since the chemical reaction which takes place at the active site is accompanied by an open-close motion, which is suggestive of a possible role of this modulation for the correct performance of the enzymatic activity.

We addressed the problem within a rather comprehensive framework, where proteolytic enzymes are considered along with other members of class 3 of the Enzyme Nomenclature Commission (72). The enzymes were taken from the list of 76 representatives of the main EC and CATH groups singled out in the study of ref. (77). The list, restricted for simplicity to monomeric enzymes (following the indication in annotated UNIPROT (78) entries) is reported in Table 3.1 along with the EC and CATH code and with the indication of the amino acids constituting the catalytic site.

For each entry, the boundary between the two primary dynamical domains was identified from the $Q = 2$ subdivision. To measure the separation of a catalytic residue from the primary dynamical boundary, we considered the distance of its $C_\alpha$ from the nearest $C_\alpha$ belonging to the other dynamical domain. The normalised distribution of these distances is shown with a thick line in Fig. 3.12 along with the reference distribution (dashed line) of the boundary separation of every amino acid in the 15 proteins.

The two distributions present appreciable and informative differences: the reference distribution of residue distances from the primary interface appears spread in a range from 2 to $20-25$ Å. On the other hand, the catalytic site residues show, with respect to the previous curve, a marked peak at less than 10 Å. This discrepancy indicates that, despite differences in structural organisation and nature of the bound substrate, the catalytic site of these enzymes is found to be preferably located at a particular subregion of the primary interface. At the *hinge* between the two domains, in fact, the active site experiences a low mobility, reflected in a negligible structural deformation which is

| | PDB | length | catalytic site |
|---|---|---|---|
| a) | 4p2p | 124 | H48, D99 |
| b) | 1ako | 268 | N7, D151, N153, D229, H259 |
| c) | 1vas | 137 | T2, R22, Q23, R26 |
| d) | 2fmb | 104 | D25 |
| e) | 1bol | 222 | H46, E105, H109 |
| f) | 1k2a | 136 | H15, H129 |
| g) | 1de3 | 150 | H137, E96, H50 |
| h) | 1kab | 136 | R35, R87 |
| i) | 3eng | 213 | D10, D121 |
| j) | 2f47 | 175 | E11, D20 |
| k) | 2ayh | 214 | E105, E109 |
| l) | 4skn | 223 | N145 |
| m) | 1avp | 204 | H54, E71, C122 |
| n) | 1qjj | 200 | E93 |
| o) | 1lqy | 184 | E154 |

**Table 3.1: Monomeric members of the EC class 3 enzymes (hydrolases)** - The enzymes were taken from the representative list of ref. (77) which covers the main CATH groups. To avoid excessive dispersion in length, only enzymes with 100 to 270 amino acids were considered. The amino acids constituting the catalytic site were taken from the catalytic site atlas (79) when literature evidence was available, otherwise they were obtained by intersecting the catalytic site atlas and Uniprot data.

necessary to preserve the catalytic geometry. On the other hand, and consistent with HIV-1 protease, the motion of the rigid units delimiting the active region are found to be generally compatible with functionally-oriented movements leading to the binding or processing of the substrate.

The overall indication of catalytic-site/boundary proximity in hydrolases conveyed by Fig. 3.12 was complemented by a case-by-case analysis of the 15 enzymes in Table 3.1. This detailed investigation was necessary in view of the fact that the cumulated data in Fig. 3.12 reflect properties of a group of enzymes with a certain heterogeneity in length, structural architecture, and number of catalytic sites.

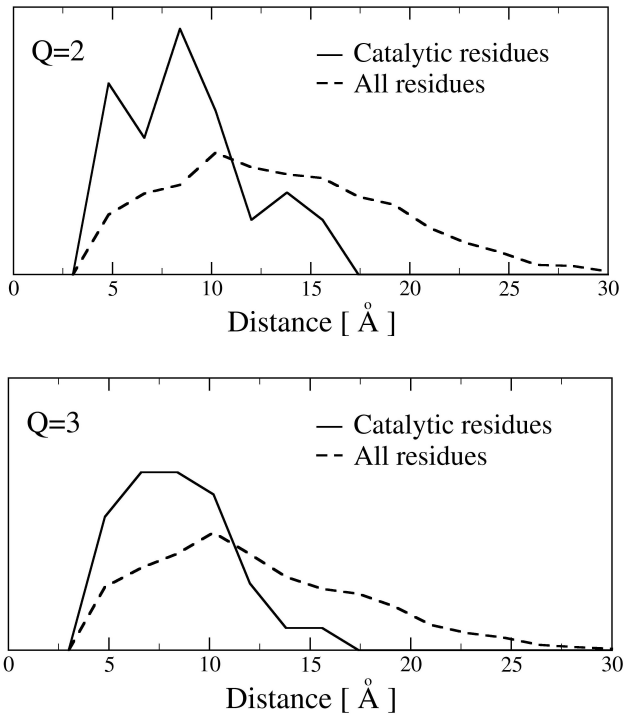


**Figure 3.12: Distribution of amino acid distances from the boundary separating the two (top) and three (bottom) primary dynamical subdomains.** - The dashed line indicates the distribution of boundary distances for all 2690 amino acids in the data set of Table 3.1, while the thick line gives the distribution only for the 34 catalytic amino acids. Both distributions are normalised.

The $Q = 2$ subdivisions of the 15 enzymes consistently reveal the good proximity of the cleavage sites with the boundaries between the dynamical domains. Here we limit the discussion to three enzymes whose dynamical role in the functional cleavage of peptides or nucleic acids has been previously considered (80; 81; 82; 83; 84),

namely: *exonuclease III* (PDB 1ako), *human adenovirus proteinase* (PDB 1avp) and *endo-1,3-1,4-β-D-glucan 4-glucanohydrolase* (PDB 2ayh). Their $Q = 2$ subdivision is representated in Fig. 3.13a,b,c.

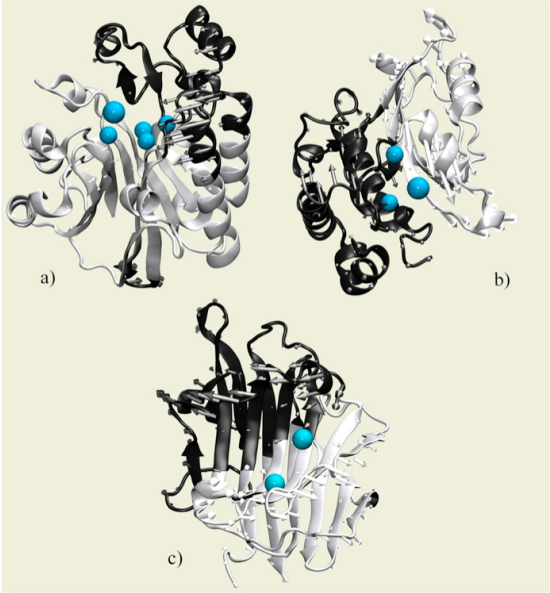


**Figure 3.13: EC3-class enzymes two-domains decompositions.** - Subdivision into $Q = 2$ dynamical domains (represented in different colors) of exonuclease III (panel a), human adenovirus proteinase (b), and endo-1,3-1,4-$\beta$-D-glucan 4-glucanohydrolase (c). The decomposition was performed taking into account the ten lowest-energy modes. For clarity, only the rigid-body approximation to the first mode is shown. Catalytic residues are shown as spheres.

Exonuclease III and adenovirus proteinase bind DNA in double- and single-stranded forms, respectively. In ref. (77) a dynamics-based connection between them was established, which is particularly interesting as they are not evolutionary related and are characterised by two different architectures, 4-Layer Sandwich (CATH: 3.60.10.10) and 3-Layer ($\alpha\beta\alpha$) Sandwich (CATH: 3.40.395.10) respectively. In both cases the catalytic residues are found to be located at the primary boundary. As visible in Fig. 3.13, the low-energy modes have a common character as they entail an outward/inward concerted movement between the two blocks in the surroundings of the catalytic sites, with the latter at the centre. The analysis carried out on endo-1,3-1,4-$\beta$-D-glucan 4-glucanohydrolase also shows that the the two catalytic residues of the enzyme are located in proximity of the interface between the two primary dynamical domains, both surrounded by loops that form a groove which can arguably accommodate the corresponding ligand.

The consistent indication of Fig. 3.13 is that the catalytic site is located close to the primary interface. This fact appears particularly interesting when considering how the primary boundary is modulated by the lowest-energy modes of the domains (which are compatible with the opening/closing of the catalytic cleft (80; 83; 84)). By comparison to non-interface amino acids, it is found that interface residues cover a fairly large range of values both for overall mobility and for the degree of distortion of the local structural environment (see Appendix A). As anticipated, the catalytic site is accommodated at, or close to, an interface sub-region having both low mobility and low-structural deformation. While these properties are consistent with the expected rigidity of the catalytic region, the relevant observation is that they can take place in proximity of the primary dynamical boundary, where appreciable elastic strain can be built up due to the relative motion of the dynamical domains.

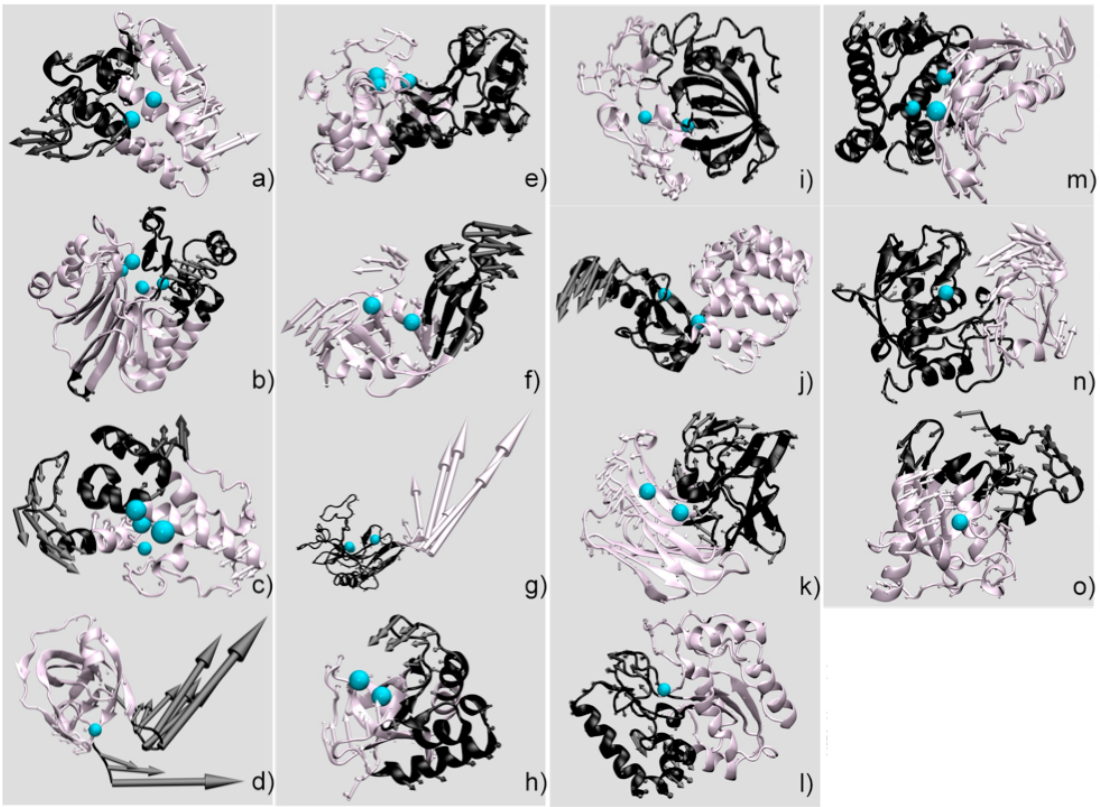


**Figure 3.14: Two-domain subdivision of the 15 EC3 representatives** - The 15 representatives of the hydrolases class discussed in the text are here shown. The two quasi-rigid domains, in which the proteins are subdivided, are shown with different colours, together with their first low-energy mode of fluctuation. The catalytic residues are highlighted in Van der Waals representation. The labels refer to the proteins listed in Table 3.1.

### 3.3.4 Comparison between dynamical and CATH domains

As a final application of the quasi-rigid domain identification scheme we performed a dataset-wide survey of the sequence compactness of the dynamical domains obtained with our method. In particular, we carried out a systematic evaluation of the extent to which the subdivision of a protein into a limited number of approximately-rigid units results in dynamical domains that are compact in space and/or cover uninterrupted regions of the primary sequence. The interest in this question is two-fold. On one hand it can provide indications on the viability, for computational efficiency, of enforcing *a priori* the proximity in sequence or space of the group amino acids (as it is done, for example, in the TLS (70) scheme). On the other hand, it can shed some light on the existence of consistent modular organisations of proteins at the level of sequence, structure and dynamics.

An interesting general context where these questions can be posed is provided by multi-domain proteins. For an appropriate, supervised definition of domain we resorted to the CATH (6) database. This classification scheme groups proteins in a hierarchy of categories: *Class*, *Architecture*, *Topology* and *Homology*. The Class level is the most general, and takes into account the secondary structure element content of a protein; Architecture and Topology classify the overall shape and connectivity of the elements; at the Homology level also the primary sequence is considered. A protein is assigned a CATH number specifying its location in this hierarchy by an automated procedure, with a minimal human intervention in ambiguous cases.

For our analysis, we considered a data set of 90 protein monomers, listed in Table A.1, with overall sequence identity below 90%, and each comprising from 3 to 6 CATH domains consisting of a single sequence interval. Each protein was subdivided into a number of dynamical domains equal to the number of CATH domains.

We were interested in characterising the robust properties of dynamical domains, and in particular considering if they preserve the sequentiality of the backbone. In order to reduce the 'noise' due to excessively short sequence fragments, the rigid-unit subdivisions were post-processed to eliminate segments smaller than 1/20th of the protein length (and in any case no longer than 10 amino acids); the amino acids in these fragments were therefore re-assigned to the nearest flanking unit.

The resulting dynamical domains subdivisions (along the primary sequence) were compared with the ones provided by CATH. It was found that only for 30 proteins out of 90 the number of sequence intervals matched. Therefore, in two-thirds of the cases the dynamical domain subdivisions gathered regions that were disconnected along the primary sequence, at variance with the CATH subdivisions. In fact, for 30 cases the dynamical subdivisions were very well consistent with the CATH ones. Out of the 91 domain boundaries in the primary sequence occurring in the 30 proteins, as many as

71 occurred at a separation of less than 10 residues of the CATH ones. By commonly-employed criteria (85) this reflects a very strong agreement of the subdivisions. It is worth noting that also for the 60 non-matching proteins, most of the CATH subdivisions fall within 10 residues from the dynamical ones, which are however more numerous.

An example of different CATH and dynamical-domain subdivision is reported in Fig. 3.15. The picture shows the CATH (panels c-d) and dynamics-based domains (panels e-f) of protein chain 1ordB. A good consensus can be found in the overall partition, especially for two of the four domains (in red and blue in the picture). The major differences are concentrated in two CATH domains (gray and orange in Fig. 3.15c-d) which are grouped a single dynamical domain by our method. At the same time, a small cluster of amino acids associated to the blue CATH domain of Fig. 3.15c-d is identified as a dynamical domain by itself: this partition can be attributed to the relatively loose connection of this residue group to the rest of its CATH domain, a structural property that is likely to result in a high collective mobility. The dynamics-based partition thus appears more reasonable than the CATH subdivision when the conformational plasticity of the molecule is taken into account.

For all the 90 proteins we checked the extent to which the non-postprocessed dynamical domains, despite possibly comprising segments that are not contiguous in sequence, occupy compact regions in space. The compactness of a domain was ascertained by measuring the diameter of the graph given by the contact map of the residues (with a contact cutoff distance of 7.5 Å). A finite value of the diameter, which measures the minimum number of graph edges that need to be traversed for connecting any two nodes in the graph, indicates the spatial compactness of the domain. It was found that less than 5 dynamical domains out of 308 comprised disconnected, though nearby, regions. These cases, however, can be neglected, due to the fact that the dynamical domains partition is excessively influenced by small fluctuating loops. These short, exposed regions of the protein are characterised by an unnaturally high mobility, which forces the partition algorithm to identify them as a single domain and assign faraway residues to the same cluster: this reduces the effective number of large clusters, and determines a mismatch with the CATH domains. An example of this effect is given in Fig. 3.16, where the quasi-rigid domain decomposition of protein chain 2ex3C is shown. With the exception of a few 'pathological' cases, the compactness measure of the dynamical domains provides an *a posteriori* indication of the fact that rigid-like units comprise amino acids that occupy spatially-connected regions.

Finally, in spite of the differences in the sequence partition of the dynamical and CATH domains we performed a test to quantify the *overlap* between these two decompositions. This quantity was calculated by exploring the combinatorial space of the possible one-to-one pairings of the $Q$ CATH and $Q$ dynamical domains. For each combination of paired domains we computed the number $n_q$ of amino acids that are

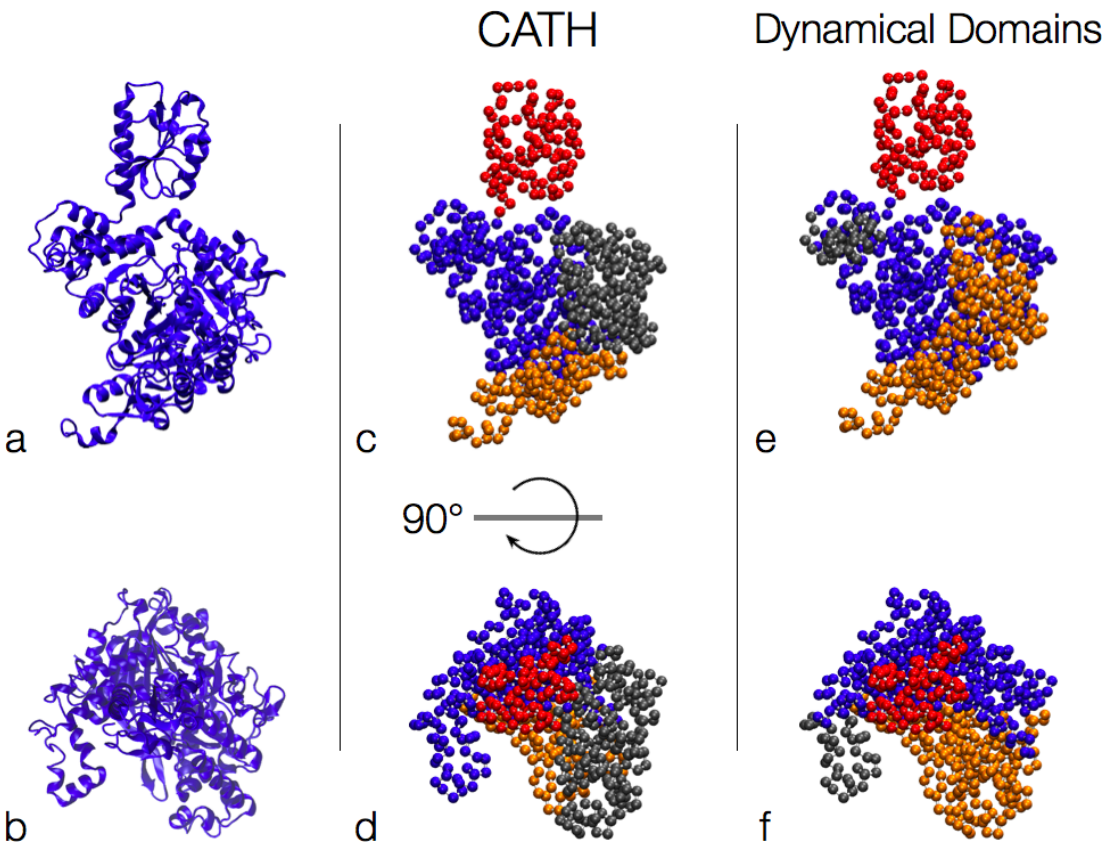


**Figure 3.15: CATH and dynamics-based partition of protein chain 1ordB** - Panels a-b show orthogonal view of protein chain 1ordB in cartoon representation. The CATH domain subdivision is similarly shown in panels c-d, while the dynamical decomposition is shown in panels e-f. The colours of the domains are chosen so to highlight the similarities in the partitions, as well as the major differences. The latter are mainly located in the dynamical subdivision of the orange CATH domain of panels c-d, and the small cluster of residues, in gray in panels e-f, which is identified as a separate dynamical domain.

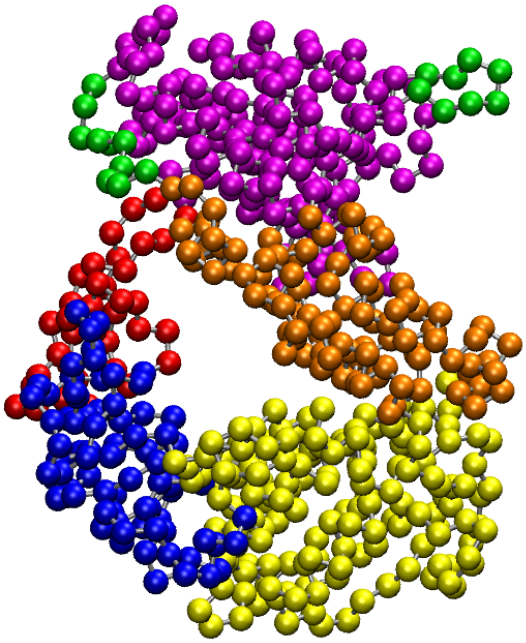

**Figure 3.16: Example of non-connected quasi-rigid domain** - The quasi-rigid domain subdivision of protein chain 2ex3C showed a block composed by two separate clusters of amino acids (here coloured in green). Cases like these can occur in presence of exposed, highly mobile loops or termini whose diffusive motion biases the rigid block assignment.

shared by the $q$th corresponding pair of CATH and dynamical domains. The only set of domain pairing that is further considered is the one yielding the largest number of shared amino acids:

$$n_q^* = \max_{pairings} \sum_{q=1}^{Q} n_q \tag{3.16}$$

The overlap of the the $q$th CATH and dynamical domains, comprising $N_q^C$ and $N_q^D$ amino acids respectively, is then defined as:

$$\mathrm{W} = \frac{2\, n_q^*}{N_q^C + N_q^D} \tag{3.17}$$

This quantity represents the number of residues shared by the two domains, normalised to the mean size of the latter. It provides an estimate of the degree of similarity of the two domains' volumes: if it is close to unity, the residues contained in a domain are present also in the other.

In the analysis of our dataset, we found that the mutual one-to-one overlap of the CATH domains and rigid units was, on average, 80%: this degree of similarity underscores a non-trivial, albeit not perfect, consistency between the two subdivision criteria. A very good consensus of the domain subdivisions is found, in particular, for architectures 2.40, 3.30 and especially for 3.40; architectures 1.10 and 2.60 show an overlap shifted towards smaller values. The distributions of the $W$ parameter, specialised for the two principal CATH codes (class and architectures) is shown in Fig. 3.17.

## 3.4 Web-server implementation

The optimal scheme to identify quasi-rigid domains in proteins has proven to be efficient: in fact, a protein as long as 200 residues can be partitioned in 2 to 20 domains in less than 2 minutes on standard desktop computers and laptops. Moreover, the results of the test-case subdivisions performed insofar proved to be consistent with previous findings discussed in the literature.

These considerations prompted us to implement the method as a web server: this freely-accessible resource, called PiSQRD[1] (after Protein Structure Quasi-Rigid Domain decomposition) allows users to perform an efficient partitioning of the desired structure and easily access and download the results.

[1] http://pisqrd.escience-lab.org/

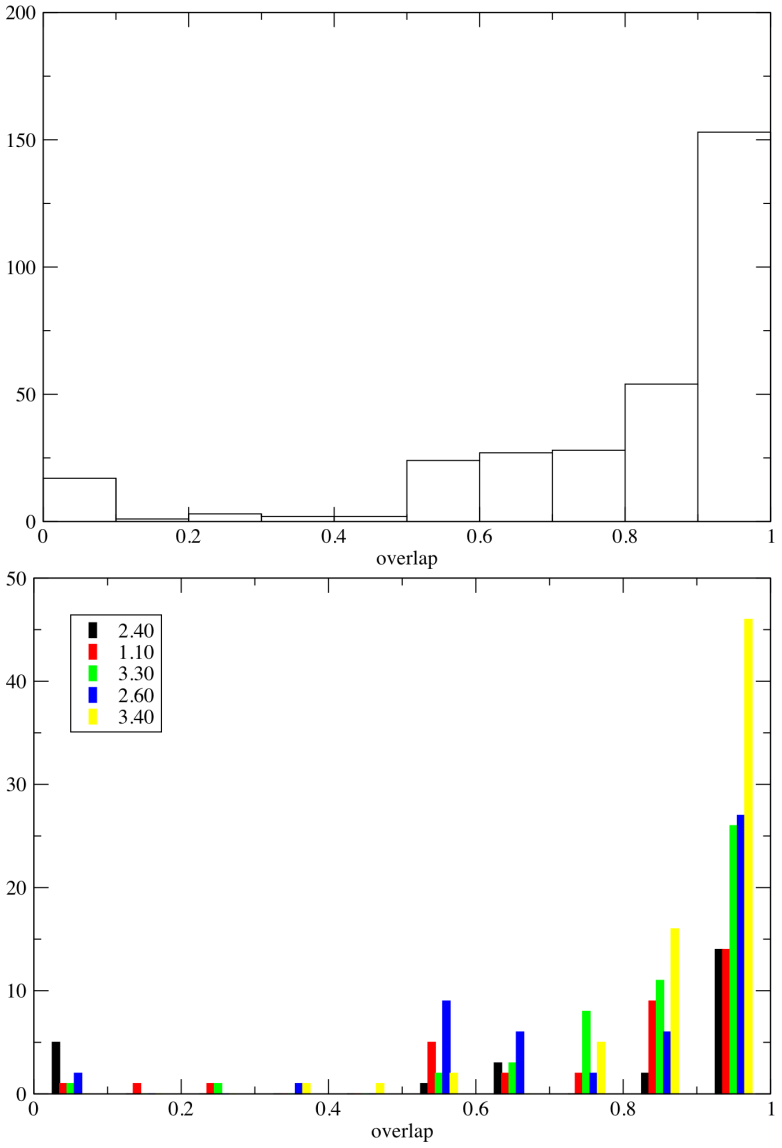


**Figure 3.17: Histograms of the overlap between CATH and dynamical domains** - Top: histogram of the overlap of all the CATH domains in the dataset with the dynamical domains (quasi-rigid subunits). Bottom: histogram of the overlap of the CATH and dynamical domains subdivided according to the CATH architecture.

The user, in fact, is simply required to provide the PDBid of the protein of interest, which is automatically retrieved from the protein data bank (PDB (86)). Alternatively, the user can upload a PDB structure file, which can contain the sole $C_\alpha$ coordinates. In both cases it is possible to specify wether a single chain of the molecule is to be taken into account: if no chain indicator is provided the subdivision is performed on the whole structure.

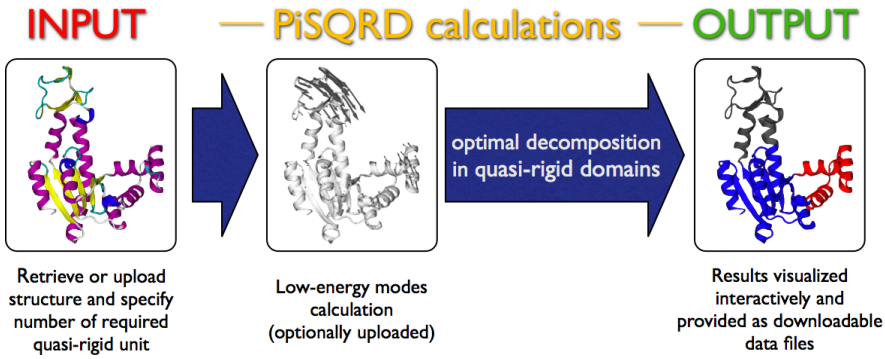


**Figure 3.18: Graphical summary of the PiSQRD flowchart** - Users provide the input structure (4akeA in the example), and the desired threshold fraction of internal essential dynamics, $\bar{f}$. The $n = 10$ lowest-energy essential modes are next calculated and used to decompose the protein from 2 to 20 quasi-rigid domains. The returned optimal decomposition is the one having the smallest $Q$ that is sufficient to capture the preassigned threshold of internal essential mobility (three domains in the example shown, using the default values $\bar{f} = 80\%, n = 10$).

The server relies on the $\beta$-GM (37) elastic network model to calculate the low-energy essential space; nonetheless, users can seamlessly overcome this step and provide a 'zip' file containing the low-energy modes of fluctuation. By default the top 10 modes will be used for the calculation, but in the *advanced options* section a form is provided where it is possible to change this number.

The server automatically subdivides the protein into 2 to 20 quasi-rigid domains. The first result page returned to the user is the one corresponding to the partition with the smallest number of domains, having a captured mobility $f$ (Eq. 3.14) above a given threshold: the latter is set to 80%, but a different value can be specified by the user before running the calculation. The result page contains several informative interfaces: a textual panel provides the data about the calculation; a Jmol applet allows the user to visualise the subdivided protein in which the domains are coloured differently; a graph shows the fraction of captured motion $f$ as a function of the number of domains. Finally, textual data files containing the subdivision details can be downloaded clicking on the relative link, allowing users to store the results of the PiSQRD calculations.

## 3.5 Summary

In the present chapter we discussed the characterisation of a protein's internal dynamics in terms of the motion of a few subparts; these large groups of amino acids, or *dynamical domains*, are assumed to move approximately as rigid bodies. A modular description of the protein structure presents many advantages: first, the large-scale internal dynamics, which is often related to the biological function of many enzymes, is more readily and intuitively perceived in terms of the relative displacement of a few blocks. Second, the notion of quasi-rigid domain allows to separate the protein's motion in two contributions: one due to the rigid-body displacement of the domains, and one ascribed to the strain internal to the domain. It is therefore possible to estimate the degree of 'dynamical modularity' of the protein from the fraction of total motion which is captured by a given subdivision.

We introduced an optimal scheme to identify the best partition of a protein in a given number of domains, which entails the largest fraction of internal dynamics. This method, in principle, can be applied to a MD trajectory; nonetheless, the computational effort required to evaluate the effectiveness of a tentative subdivision while exploring a large combinatorial space pushed us to devise more efficient schemes. Specifically, we introduced a method to identify the optimal domains minimising a cost function which penalises the distance fluctuations among residues in the same group. This function is calculated on a small set of collective modes obtained from a PCA or from elastic network models. With respect to other schemes, the method does not enforce any sequence proximity of the grouped amino acids, which are assigned to quasi-rigid domains with a criterion which is based on the very definition of rigid body.

The partition of a protein in terms of a few blocks can be used to characterise their motion in a MD simulation, performing a constrained rigid-body fit of the domains. In addition, we introduced a further scheme where all fluctuations of the domains, except a rigid rotation about fixed axes, are suppressed; the location and direction of these axes are found maximising the mean square fluctuation that is captured by the constrained rotation of the domains. This method represents a further simplification of the protein internal dynamics, thus highlighting the salient features of the collective motions of few subparts of the molecule.

The domain subdivision method, and the optimal axis identification scheme, can be valuable tools for many purposes. Besides the aforementioned advantages in the analysis, description and interpretation of protein dynamics data, it comes natural, in fact, to use the quasi-rigid domains to improve the computational efficiency of available algorithms. Examples are the use of the domains as collective variables in accelerated MD schemes or, in conjunction with the identification of optimal rotation axes, to speed up protein-protein docking methods.

The viability of this tool has been illustrated in a number of cases. Specifically, the domain subdivision and optimal axes analysis were performed with the benchmark cases of adenylate kinase and HIV-1 protease, whose internal dynamics have been extensively studied. The results proved in good accord with previous works, and triggered further investigations on the similarity of functional-related dynamics in a class of hydrolytic enzymes. In fact, 2-domains subdivisions of 15 representatives of the EC3 class were performed. The active sites of these molecules showed a preferential location in proximity of the boundary separating the two domains, suggesting a common dynamical trait among proteins differing for both fold and architecture. Similar features, which have been ascertained here on the basis of specific indicators (the active site distance from the primary boundary), might be general properties of structurally-different proteins and enzymes. The investigation of this similarity is the topic of the next chapter, where we shall discuss a method to quantitatively compare the internal dynamics of two proteins.

# 4

# Similar collective dynamics in structurally different proteins

The characterisation of proteins is commonly organised according to the tripartite scheme *sequence* → *structure* → *function*. In the Introduction we mentioned that our understanding of the relation bridging the first two steps of this logical cascade has been extensively investigated using sequence and structure alignment schemes. These methods helped to unveil similarities between proteins differing for amino acid sequence or structural organisation. A remarkable result, for example, is that a degree of sequence identity larger that 30% generally implies a high degree of structural similarity, but also that proteins sharing low sequence identity can nonetheless have similar folds (4; 5; 6).

In the previous chapter we have further seen that in a dataset of different EC3 representatives a functional property -that is, the location of the active site- could be related to a common structural/dynamical feature: namely, the fact that the active site is found preferentially at the interface between the two primary dynamical domains. This result is suggestive of the possibility that the second link in the logic ladder of protein characterisation, i.e. the relation between structure and function, may be mediated by dynamical properties: could the similarity of motion among two different objects help to find their hidden shared features (see Fig. 4.1)? It would be of great interest to investigate this relation between structure and function on the basis of protein internal dynamics.

This possibility can be explored with *dynamics-based alignment methods*, i.e. algorithms which superpose two protein structures taking into account not only the similarity of their structures, but also the consistency of the aligned regions' motion. In the past, various approaches were investigated to compare protein structures on the basis of their internal dynamics (76; 77; 87). In particular, Zen *et al.* (77) recently developed a method to align proteins by optimally matching their structures and dynamics.

This scheme relies on a comparison of the low-energy collective motions of the aligned regions. It has been shown, in fact, that for many proteins the concerted fluctuations described by the lowest-energy modes entail a large fraction of the functionally-oriented conformational changes (13; 15). The alignment is thus performed so to maximise the degree of similarity of the tentatively superposed regions.

In the present chapter we discuss a dynamics-based alignment scheme, related to the one of Zen *et al.*, of which our algorithm represents a generalisation. Our method relies on an approximated scoring function to identify major traits of structural similarity and dynamical consistency of the aligned regions, thus reducing the exploration of the possible alignments before looking for more exact matches. The resulting alignments show a very good, albeit non perfect agreement with the ones obtained with the 'exact' method; on the other hand, the fast performance of the calculations makes possible the application to large dataset-wide surveys, which were not feasible in a short time with the original algorithm.

In what follows, we shall briefly describe the original method introduced by Zen *et al.*; then, the optimised scheme will be discussed and applied to test-cases.

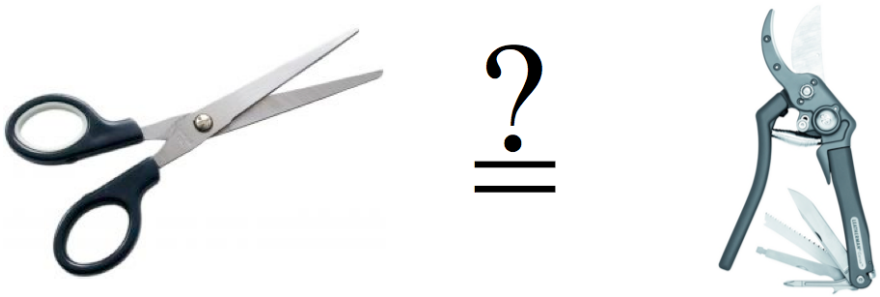


**Figure 4.1: Different fold, similar motion** - Can objects of different shape have similar internal movements? Does this dynamical consistency say something about their function?

## 4.1 Dynamics-based alignment of proteins

The dynamics-based alignment method introduced by Zen *et al.* (77) aims at establishing correspondences among regions of the two aligned proteins which undergo similar collective motions. As it was anticipated at the beginning of this chapter, the concerted movements entailed by low-energy modes are often related to the functionally-oriented internal dynamics of a protein. It is therefore natural to investigate whether a correspondence exists, and to what extent, among the collective fluctuations of different proteins: in fact, as structure is more evolutionarily conserved than sequence, it was speculated (77) that the internal dynamics could be more conserved that the structure

which sustains it. This conserved dynamical features might unveil relations, in the biological function of the two proteins, which are not evident by the inspection of the structures alone.

The dynamical similarities among two proteins are searched as pairwise correspondences of marked amino acids, i.e. structural alignments. The dynamics of the marked residues are compared, and the optimal alignment is sought, which maximises a scoring function rewarding spatial proximity and dynamical similarity of the residue pairs.

Specifically, the algorithm of ref. (77) works as follows. The first step is the exploration of one-to-one correspondences between amino acids of two proteins. The number of aligned residues, $n$, ranges from 1 to the length of the shortest protein. This tentative alignment divides the residues of each protein into two classes: the $n$ residues marked for the alignment, and the non-pairing ones.

The second step is the calculation of the internal dynamics of the marked amino acids. This is done making use of the $\beta$-GM (37) elastic network model where, as customary, only the $C_\alpha$ atoms are taken into account.

In order to compare on equal footing the fluctuations of the sole marked residues, the motion of the non-aligned ones is integrated out. Due to the quadratic nature of the interaction matrix, in fact, this integration can be performed exactly. Consider the following elastic network Hamiltonian for a protein:

$$E_{\beta GM} = \frac{1}{2} \sum_{i,j=1}^{N} \vec{\delta x}_i H_{ij} \ \vec{\delta x}_j \tag{4.1}$$

The displacement vector can be written as:

$$\vec{\delta x} \equiv \{\vec{\delta x}^{\,a}_1, \vec{\delta x}^{\,a}_2, \cdots \vec{\delta x}^{\,a}_n, \vec{\delta x}^{\,b}_1, \cdots \vec{\delta x}^{\,b}_{N-n}\} \tag{4.2}$$

where the first $n$ vectors refer to the marked amino acids ($a$ superscript), and the second $N - n$ vectors refer to the non-marked ones ($b$ superscript). The Hamiltonian can be decomposed in blocks, as:

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

The diagonal terms $M^a$ and $M^b$ describe the interactions among the aligned and non-aligned residues, respectively, while the off-diagonal term $V$ takes into account the interaction between the two groups (the superscript $T$ indicates transposition).

Following the example of ref. (76), the non-aligned degrees of freedom can be traced out, leading to an effective Hamiltonian which describes the interactions among the sole aligned residues:

$$E^{eff}_{\beta GM} = \frac{1}{2} \sum_{i,j=1}^{n} \vec{\delta x}_i^{\,a} \left( M^a_{ij} + \Delta M_{ij} \right) \vec{\delta x}_j^{\,a} \tag{4.3}$$
$$\Delta M = -V[M^b]^{-1}V^T$$

where $[M^b]^{-1}$ is the pseudoinverse of $M^b$. This procedure, applied to the two tentatively aligned proteins, returns two sets of orthonormal vectors, $\{\vec{v}_\ell\}$ and $\{\vec{w}_\ell\}$, describing the motion of the two proteins' aligned residues alone.

The third and last step of the algorithm described in ref. (77) consists in the calculation of the scoring function for each tentative alignment. The score, which compares the first 10 low-energy modes of the two aligned subsets, is defined as:

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

where $d_i$ is the distance separating the $i$-th aligned residues, and

$$f(d) = \frac{1}{2} \left[ 1.0 - \tanh \left( \frac{d - R_0}{1\text{Å}} \right) \right] \tag{4.5}$$

is a weight function penalising those paired residues whose separation exceeds the cutoff $R_0$, usually put at $\sim 4$ Å.

The scoring function in Eq. 4.4 is not sensitive to a redefinition of the basis vectors $\{\vec{v}_\ell\}$ and $\{\vec{w}_\ell\}$, because it depends only on the full linear space spanned by the two sets of vectors. Notice that Eq. 4.4 represents a distance-weighted generalisation of the Root Mean Square Inner Product (RMSIP), which is given by:

$$\text{RMSIP} = \sqrt{\frac{1}{10} \sum_{\ell,m=1}^{10} \left| \sum_{i=1}^{n} \vec{v}_i^{\,\ell} \cdot \vec{w}_i^{\,m} \right|^2} \tag{4.6}$$

The score in Eq. 4.4 ranges from 0 to 1, indicating respectively null or perfect correspondence between the marked residues low-energy spaces, as in the case of RMSIP; on the other hand, at the contrary with respect to the latter, the score $q_n$ incorporates also the structural information, which is encoded in the distance weighting of the scalar products.

For a given alignment length, $n$, the optimal pairing is the one maximising, over all tentative superpositions, the score of Eq. 4.4. The choice of the optimal alignment length is finally made introducing a suitable normalisation in Eq. 4.4.

## 4.2 Optimised strategy of dynamics-based alignment for large-scale applications

The dynamics-based alignment method described so far was proven to identify and highlight relevant structural and dynamical correspondences among proteins having markedly different structures. Most notably, the use of this method allowed to establish spatial relationships between structurally-dissimilar proteins involved in similar catalytic reactions. Unfortunately, this scheme requires, for a single alignment, a considerable amount of computational resources. In fact, for each of the tentative alignments the degrees of freedom of the non-aligned residues have to be traced; the combinatorial space of possible alignments is so large to make the exhaustive exploration of all possible matchings unfeasible.

Optimised schemes have been thus applied, such as a stochastic search making use of a replica exchange method (88) to efficiently explore the combinatorial space. Moreover, the pairing scheme was forced to match segments of at least 10 amino acids rather than single residues, and a sequentiality constraint was enforced.

Even with this restrictions, the computation of a single alignment of two proteins takes too much time (about 15 minutes on standard computers) to allow for the efficient application to large protein datasets. The need to improve the efficiency of the calculation motivated us to develop an optimised, approximate scheme to perform dynamics-based alignments. In the following, we shall describe this optimised algorithm, its limitations and strengths, and apply it on two test cases.

### 4.2.1 The algorithm

The main bottleneck of the original formulation of the previously discussed scheme is the repeated identification of the aligned amino acids and the tracing out of the non-marked residues. This procedure is nonetheless required to correctly compare the internal dynamics of two sets of residues having the same length.

This limitation can be overcome, though in an approximate fashion, by relaxing the requirement of a pairwise comparison of the residues' motion: in the following we shall describe a tolerant scoring function whose minimisation, over all the possible alignments of two proteins, rewards the dynamical similarity of corresponding protein *regions* that can be well superposed structurally.

Consider the following scoring function:

$$s = -\frac{N_1\, N_2}{10} \sum_{\ell,m=1}^{10} \left[ {\sum_{\substack{i=1\ldots N_1 \\ j=1\ldots N_2}}}' \vec{v}_i^{\,\ell} \cdot \vec{w}_j^{\,m}\, f(\Delta_{ij}) \right]^2 \tag{4.7}$$

where indices $i$ and $j$ run, respectively, over the $N_1$ amino acids of the first protein and the $N_2$ of the second one.

The quantity in Eq. 4.7 closely resembles a Mean Square Inner Product. The main difference lies in the scalar product: in place of a sum of $\vec{v}_i^{\ell} \cdot \vec{w}_i^{m}$ terms running on the pair index $i$, the vector $\vec{v}_i^{\ell}$ of a protein's residue $i$ is compared with *all* the displacements $\vec{w}_j^{m}$ of the other protein. These scalar products are weighted with a decreasing function $f(d)$ of the distance (see Eq. 4.5) in order to exclude from the comparison residues exceeding the cutoff distance $R_c$, which in this case is set at a value of 7Å.

The scoring function 4.7 is more tolerant, in the comparison of the internal dynamics of two proteins, with respect to Eq. 4.4: in fact, the latter matches only residue pairs, while in the approximate method the overall motion of two *regions* (enclosed by spheres of 7Å radius) are compared. Therefore, an appropriate notion of distance between the two regions has been introduced. Specifically, the sigmoidal function $f$ is not calculated on the simple distance between an amino acid pair: an effective distance $\Delta_{ij}$ is rather used, measuring the spatial separation of the fragments $[i-1, i, i+1]$ and $[j-1, j, j+1]$. *A priori* the latter could be matched with either the same or opposite sequence orientation. For the two cases the segments distance is defined respectively as:

$$\begin{aligned} d_{ij}^{+} &= \max\{d_{i-1,j-1}, d_{i,j}, d_{i+1,j+1}\} \\ d_{ij}^{-} &= \max\{d_{i-1,j+1}, d_{i,j}, d_{i+1,j-1}\} \end{aligned} \tag{4.8}$$

with $d_{ij}$ being the Euclidean distance of amino acids $i$ and $j$. The most appropriate sequence orientation is chosen *a posteriori* by setting $\Delta_{ij} = \min(d_{ij}^{+}, d_{ij}^{-})$.

As discussed above, the 'generalised' scalar product in Eq. 4.7 compares the 'field of motion' of entire protein regions rather than single amino acid pairs. Because the effective number of compared residues does not reflect in the normalisation of the eigenvectors, which satisfy the relation

$$\sum_{i=1}^{N} \vec{v}_i^{\ell} \cdot \vec{v}_i^{m} = \delta_{\ell m} \tag{4.9}$$

we introduced the factors $N_1 N_2$ in order to account for this effect. Assuming uniform displacement vectors, in fact, the normalisation condition requires that each residue component of a mode has length $|\vec{v}_i| = 1/\sqrt{N}$: the multiplication by the proteins' lengths returns single residues displacement vectors of unit length on average. This correction is further enforced excluding from the computations those regions having an atypically large mobility (e.g. exposed loops or termini). The latter, in fact, could introduce artefacts in the comparison of the dynamics, due to the large module of their

displacement vectors. The score contributions are thus restricted to pairs, indicated by the primed sum, whose residues' square mobility does not exceed by a factor 4 the average one per amino acid.

The minimisation of the score $s$ of Eq. 4.7 over the relative rotations and translations of the two molecules of interest is carried out using the engine of the MISTRAL structural alignment program (89). Specifically, the two proteins are first superposed by optimally aligning segments of up to 50 amino acids. This initial superposition is next optimised by minimising $s$ over the possible relative orientations of the molecules. The list of equivalent amino acids is finally computed using a 'seed and grow' search for matching segments: in the applications, we used a seed threshold equal to 4.5 Å and a tolerance equal to 5 Å (89; 90).

Once the optimal superposition minimising the score $s$ is found, and the aligned pairs of residues have been marked, we take apart the approximation and use the strict, though computationally more onerous, measure of dynamical consistency. This can be done following the tracing procedure of the non-aligned residues described in the previous section, and calculating the RMSIP between the two equal-length sets of low-energy modes. This calculation is performed only once after the effective alignment, resulting in a considerable saving of computational resources. In Fig. 4.2 the flow chart of the original algorithm introduced by Zen *et al.* is compared with the optimised, approximate scheme here discussed.

### 4.2.2 Assessing the statistical significance of the alignments

The approximate dynamics-based alignment algorithm described above returns the optimal superposition corresponding to the highest consistency of two proteins' internal dynamics, as measured by the scoring function $s$. The latter, though, does not provide any information about the *statistical significance* of an alignment, i.e. the probability to obtain a given value of the optimal score in the alignment of two unrelated protein entries.

The measure of the statistical significance of an optimal superposition can be assessed comparing the corresponding $s$ score with a reference distribution, obtained from a pool of alignments among unrelated proteins whose dynamical consistency is expected to be low.

We calculated the distribution of the $s$ score (in absolute value) performing pairwise alignments of the representative protein dataset of Sierk and Pearson (92): it is expected *a priori* that only a negligible fraction of the alignments in this set will correspond to true positive correspondences. From this set we randomly picked $10^5$ pairs of non-homologous and structurally dissimilar proteins, differing at the level of CATH architecture.

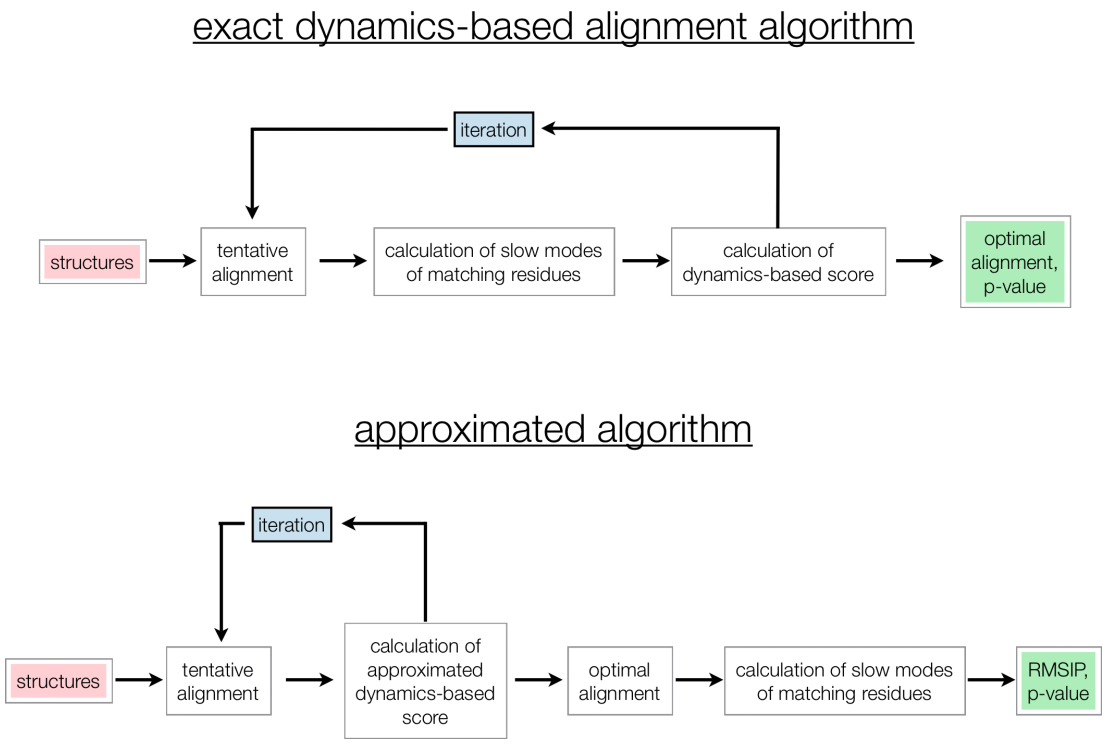


**Figure 4.2: Flow chart of the exact and approximate dynamics-based alignment algorithms** - The schemes here reproduced illustrate the basic steps of the algorithms introduced by Zen *et al.* (77) and Potestio *et al.* (91), in the exact and approximated version respectively, to perform a dynamics-based alignment of proteins. Note, in particular, that the integration of the non-aligned degrees of freedom is performed at each iteration step in the exact scheme, while it appears only once in the approximated algorithm, after the optimal alignment has been found.

In order to take into account the dependence of the score on the proteins' size, we grouped the $s$ values according to the largest protein's length:

$$\begin{aligned} &\{s\}_n \equiv \{s(N_1, N_2) : \max_{N_1,N_2} \in [n-\Delta, n+\Delta]\} \\ &n = \Delta \cdot (2k+2),\ k = 0, 1, 2... \end{aligned} \tag{4.10}$$

where $N_1, N_2$ are the two proteins' lengths, and $\Delta$ was chosen to be equal to 25.

For each group $\{s\}_n$ we fitted the resulting distribution with a Gumbel extremal statistics. This particular distribution arises when the maximum of a random set of values is considered. The analytic expression of the Gumbel distribution is:

$$P(z) = \frac{ze^{-z}}{\beta},\ \text{with}\ z = e^{-\frac{x-\mu}{\beta}} \tag{4.11}$$

The parameters $\mu$ and $\beta$ are related to the mean and variance of the distribution as follows:

$$\begin{aligned} &\langle x \rangle = \mu - \beta\gamma \\ &\langle x^2 \rangle - \langle x^2 \rangle = \mu - \beta \ln(\ln(2)) \end{aligned} \tag{4.12}$$

where $\gamma \sim 0.577$ is the Euler-Mascheroni constant.

The fits of the various sets (see, for example, the fit of the group with $n = 200$ in Fig. 4.3) allowed us to obtain the Gumbel parameters as functions of the largest protein size, $n$:

$$\begin{aligned} \mu(n) &= a_1 + \gamma a_2 \\ \beta(n) &= \sqrt{\frac{\pi^2}{6 a_1 a_2}} \end{aligned} \tag{4.13}$$

where:

$$\begin{aligned} a_1(n) &= 1.25 + n e^{-\left(\frac{n}{218.45}\right)^2} \\ a_2(n) &= 13.45 + 0.38n \end{aligned} \tag{4.14}$$

For a given value of the largest protein length, the Gumbel distribution with the appropriate parameters is used as a 'null' reference distribution against which we measure

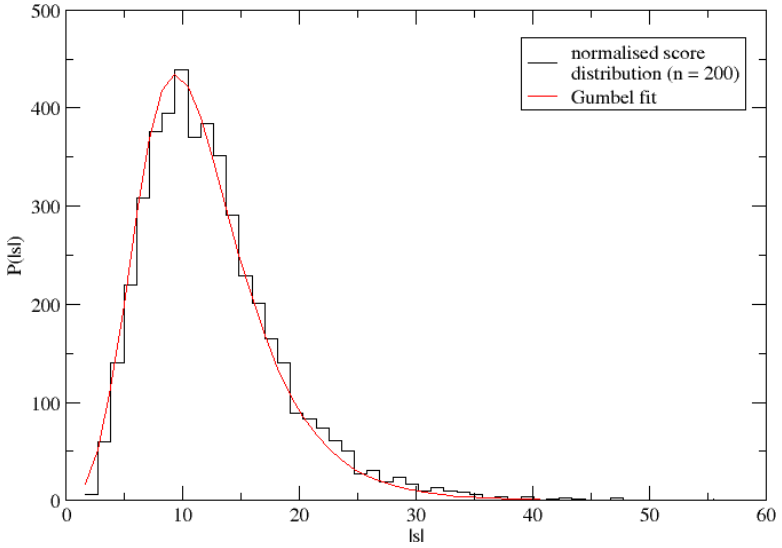


**Figure 4.3: Probability distribution of the optimal alignment score** - In this graph we plotted the absolute values of the score $s$, associated to the alignments involving proteins having maximum length between 175 and 225. The normalised histogram was fitted with a Gumbel distribution (red curve).

the statistical significance of the alignment. This is quantified in terms of the p-value or, equivalently, by the z-score, defined as follows:

$$\begin{aligned} p(s_n) &= \int_{|s_n^*|}^{\infty} P(x;\mu(n),\beta(n))\, dx \\ z(s_n) &= \frac{|s_n^*| - \langle |s_n| \rangle}{\sigma_n} \end{aligned} \tag{4.15}$$

where $\sigma_n$ is the standard deviation of the $n$-distribution; as anticipated, the score is taken in absolute value for sake of simplicity. The p-value corresponds to the probability to obtain a given optimal score $|s_n|$ or higher aligning two unrelated proteins. A low p-value thus indicates a high statistical significance; similarly, the z-score measures the distance of the alignment score from the average, in units of the standard deviations: the largest the z-score, the more unlikely, and hence significant is the alignment.

## 4.3 Test-case application of the dynamics-based alignment method

We now discuss two applications of the approximate dynamics-based alignment. The reliability of the scheme has been tested with two benchmark cases, previously treated in

the literature: human beta-secretase (BACE) vs. HIV-1 protease (93), and exonuclease III vs. human adenovirus proteinase (94; 95).

### 4.3.1 HIV-1 PR and beta secretase

The additional insight offered by the dynamics-based alignment with respect to 'static' alignment approaches is aptly illustrated by the comparison of HIV-1 PR (PDBid: 1aid) and human beta-secretase (PDBid: 3hvgA). The two enzymes, which are both aspartic proteases, present major structural differences. In fact, HIV-1 PR is a 198-amino-acid-long homo-dimer, and is almost entirely composed of $\beta$ sheets. On the contrary, beta-secretase is a monomeric enzyme consisting of 379 amino acids and rich in $\alpha$ helices. Despite the differences in symmetry, oligomeric state, length and secondary structure content the two enzymes share several segments of the primary sequence and are hence believed to be evolutionarily related (93). In fact, they admit a partial, but significant, structural superposition: their DALIlite alignment (96) returns 94 corresponding residues with an associated RMSD of 3.4 Å, while the MISTRAL alignment returns 128 equivalent amino acids at 2.4 Å RMSD. In addition to the partial structural correspondence, previous studies, based on atomistic MD simulations had highlighted the similarity of the low-energy modes of the two molecules (39; 40).

The dynamics-based alignment returned by our method is statistically significant, as the associated $p$-value is appreciably smaller than the conventional threshold of 0.05. and is fully consistent with the above-mentioned findings. The alignment consists of more than 140 amino acid pairs at an RMSD smaller than 4 Å. The good correspondence of the modes is highlighted by the large RMSIP value of the matching modes, which is about 0.8.

The functional relevance of the dynamics-based alignment is underscored by the following facts. First, the returned alignment superposes the catalytic dyads of the two enzymes. This is a non-trivial aspect in consideration that no information about the chemical composition (such as the primary sequence) was used. The second observation regards the consensus movements in the two proteins, which entail the modulation of the region accommodating the peptide chain to be cleaved. It is known that in order for the proteolytic reaction to occur, both BACE and HIV-1 PR must "stretch" the substrate in a $\beta$-extended conformation (39; 40), and the consensus motion captured by the method, see Fig. 1, is consistent with the required deformation (97).

The dynamics-based alignment therefore vividly illustrates the existence of a fundamental similarity underlying the internal dynamics of these enzymes, which is instrumental to produce analogous, functionally-oriented, deformation patterns in spite of the overall structural differences.

### 4.3.2 Exonuclease III and human adenovirus proteinase

Exonuclease III (PDB: 1ako) and the human adenovirus proteinase (PDB: 1avp) are not evolutionary related and are structurally dissimilar at the CATH architecture level. Their structural alignment has a $p$-value larger than 0.1 according to MISTRAL and, similarly, it is ruled out as 'not significant' by DALIlite.

Despite these differences, the enzymes process chemically similar substrates. In fact, both exonuclease III and human adenovirus proteinase bind DNA (in double- and single-stranded forms, respectively). In the study of Zen *et al.* (77) the dynamics-based alignment of the enzymes was found to have a good statistical significance. As for the case of BACE and HIV-1 PR, the functional relevance of the dynamical correspondence was underscored by the fact that the known active sites of the proteins (79) were spatially-superposed by the alignment and by the fact that the consensus motion was compatible with the expected functionally-oriented structural changes (94; 95).

All the above established results are reproduced by the new alignment scheme which employs a more general search scheme than the method of (77). As visible in Fig. 4.4, the two proteins align over more than 90 amino acids, at an RMSD smaller than 4 Å. The consistency of the dynamics of the aligned regions is high (RMSIP value larger than 0.7). It is readily noticed that the alignment yields a good spatial overlap of the active sites of the two enzymes. In accord with previous findings (77), the latter are located in a region at the interface between two oppositely-moving 'domains'. As suggested for other enzymes, e.g. the EC3 hydrolase representatives discussed in the previous chapter (98), this characteristic ought to preserve the catalytic geometry at the active site, while facilitating the accommodation/processing of the substrate.

## 4.4 Web-server implementation

The computational efficiency of this dynamics-based algorithm allowed us to implement it in a freely-accessible web-server, named ALADYN[1] (after Dynamical Alignment method).

The server interface is kept at a minimum level of complexity, as it only requires the input of the two proteins of interest (as PDB id's, or PDB files to be uploaded).

The algorithm's running time scales approximately proportionally to the product of the lengths of the input proteins. In fact, the time required for the alignment of two proteins of about 250 amino acids is typically less than one minute on the modern multicore server that hosts ALADYN, while two proteins of about 500 amino acids are completed in about 4 minutes.

[1] http://aladyn.escience-lab.org

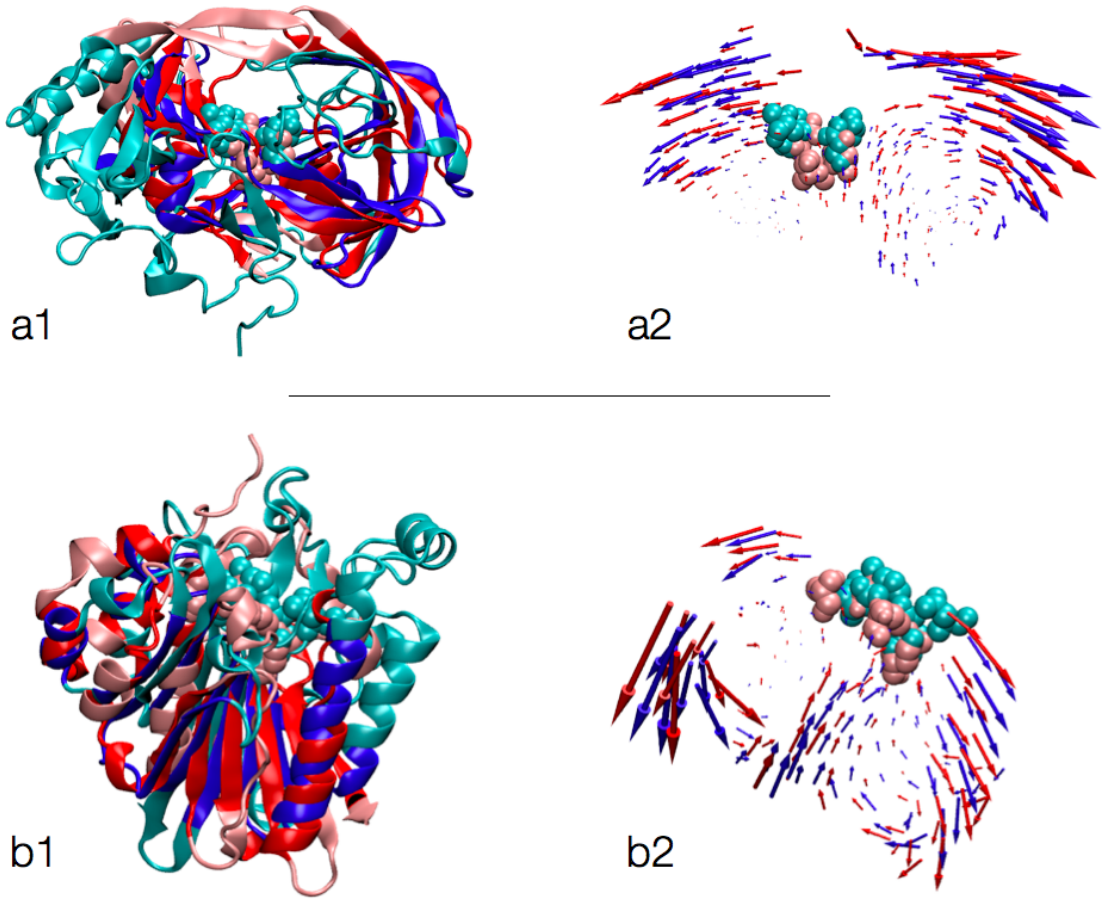


**Figure 4.4: Examples of dynamics-based alignments** - The structural correspondences and the consistency of the fluctuation dynamics of the aligned regions are shown side-by-side for each of the test cases discussed in the *Test cases* section. **(a)** The alignment of HIV-1 protease (pink/red) and beta-secretase (cyan/blue) are shown in panels a1 and a2. **(b)** The alignment of human adenovirus proteinase (pink/red) and exonuclease III (cyan/blue) are shown in panels b1 and b2. Aligned regions are shown with saturated colours (i.e. red and blue), while the active sites are highlighted using a Van der Waals representation.

Upon successful completion, users are finally directed to an interactive graphical representation of the superposed proteins, based on the Jmol (99) applet, which is complemented by a summary of the salient properties of the alignment, number of aligned amino acids, RMSIP, RMSD and the statistical significance conveyed by the $z$-score and $p$-value. The links provided at the bottom of the results page allow users to download data-files containing all details of the alignment output.

## 4.5 Summary

The dynamics-based alignment of two proteins represents an important complement of the available sequence- and structure-alignment schemes. The possibility of recognising a similar modulation of the collective dynamics between structurally-different proteins, in fact, can widen our understanding of the biological function of these biomolecules and the strategies they adopt to perform their activity.

In this chapter we revised a dynamics-based alignment scheme previously introduced by Zen *et al.* (77), and discussed a possible extension. Our method relies on the continuity and smoothness of the modes to estimate the similarity of the motion without establishing a one-to-one correspondence of the proteins' amino acids, which represents the bottleneck of the original algorithm. This approximation is justified *a posteriori* by the high degree of collectivity of the low-energy modes, whose modulation is compared between the two proteins under exam. The simplification adopted in this 'tolerant' scheme can naturally lead to non-perfect alignments, which are seamlessly performed by the original method; on the other hand, the drastically reduced computational effort of the calculations allows for a thorough and efficient exploration of the alignment configurational space and, consequently, in dataset-wide applications which were out of reach otherwise.

In the previous chapters a large body of evidence has been collected that large-scale, functionally-relevant conformational changes of proteins and enzymes are encoded in their structure. Specifically, the robustness of the low-energy fluctuations and the modular properties of many protein structures -highlighted by the dynamical domain decomposition- have shown that the structural properties of a protein often reflect into the modulation of concerted displacements; the latter, in turn, can assist the biological function of the molecule.

The application of the dynamics-based alignment can allow us to gain further insight into this dynamics-mediated relation between structure and function. In particular, we have seen that a given set of collective fluctuations is not tied to a unique structure: on the contrary, many different folds can perform similar motions. The possibility to ascertain this dynamical similarity among structurally-different proteins gives us new instruments for investigating the structure-function relation in proteins.

# Part II

# Comparing knotted and unknotted proteins

# 5

# Knotted-unknotted protein pairs: evidence of knot-promoting loops

The protein fold space is an important concept which plays a pivotal role in the organisation of protein structural data and their relation with sequence and function. The constantly increasing number of available structures, in fact, imposes the necessity to classify these informations, in order to recognise correspondences as well as differences among proteins.

In spite of its importance, the notion of protein fold and its definition are still matter of controversial debate (100). Two of the most used protein structure classification resources, namely SCOP (5) and CATH (6), make use of a *hierarchical* scheme; for example, in CATH the top level (*Class*) organises proteins according to the secondary structure element content: mainly $\alpha$, mainly $\beta$, mixed $\alpha-\beta$ and unstructured. According to these schemes, the protein structure space consists of discrete non-overlapping 'islands' (folds): two proteins belonging to different folds can only share those features which characterise the common parent classification level.

Clearly, relevant structural similarities can be found even among members of different folds. An example was given in chapter 4, where significant dynamics-based alignments were found for proteins with different CATH topology. In addition, it was recently suggested that the fold space is *continuous* (100; 101), and different folds can be connected with a path of progressively similar structures. In a continuous space the 'distance' between two folds is not given by discrete classifiers (like the $\alpha/\beta$ content) but is instead quantified by continuous measures of structural similarity (RMSD could be one).

Obviously, both descriptions of the fold space -discrete and continuous- are legitimate, and both can contribute different and complementary points of view to broaden our understanding of the structural features of protein folds. Nonetheless, a feature

exists, which usually is not incorporated in either of the two schemes: the topology of the protein chain.

In fact, the topological properties of proteins -namely, the knotted or unknotted state of the backbone- are difficult to include in a structure classification scheme since, by its very definition, topology is a geometric invariant. This means that proteins having completely different structures (according to both the discrete and continuous view of the fold space) can share the same topology, and symmetrically two almost identical protein structures can differ by a few angstroms, sufficient to give place to a knotted and an unknotted structure. An example of the latter possibility is given by the SOTCase discussed in ref. (102), whose oligomeric chains contain a trefoil knot. Sequence-related proteins with the same CATH code of the knotted SOTCase can be found, which differ by the latter for less than 2 Å RMSD though not having a knotted structure.

The topological state can thus be depicted as a 'discrete quantum number' in protein fold space which can group together structurally far instances and segregate in different classes proteins with a RMSD close to zero.

In the present chapter, we shall tackle the problem of knotted topologies in proteins with a comparative approach: a very small and non-redundant set of knotted proteins is found, and relations with unknotted proteins are sought making use of both sequence and structure alignment schemes: the evolutionary and structural relationships between geometrically similar but topologically different protein pairs can help to shed light on this issue.

## 5.1 Knots in proteins

### 5.1.1 The knotted protein puzzle

The presence of knots in proteins, their formation and their role, represent one of the major puzzles of nowadays protein science. The existence of protein chains naturally occurring with a nontrivial topology has been suggested since when the very first crystal structures were resolved (103); nonetheless, this possibility was readily deemed as impossible: it was in fact assumed that a knotted state of the polypeptide chain could only represent a hindrance of the folding process or a kinetic trap, preventing the protein from reaching the functionally active native state.

This idea was shared by the vast majority of the scientific community, and it reflects the fact that numerous structure prediction algorithms exclude knotted structures by default (104; 105). It was therefore quite a surprise when the first deeply knotted protein structure was discovered (103; 106), proving that a knotted native state was not incompatible with the folding process *in vivo*.

The existence of knotted proteins obviously poses important conceptual challenges, both in the comprehension of the role played by knots to carry on the biological activity and, in particular, for what concerns the *formation* of the knot itself, being this an extreme case of folding complexity.

### 5.1.2 Knots in biopolymers: chance or necessity?

But what makes knots in proteins so unexpected, and knotted proteins so special? A vast variety of diverse physical and biophysical examples show that knots are not rare at all: in biopolymers such as DNA filaments, on the contrary, they form extremely often and with a broad range of complexity (107; 108; 109; 110). Several experiments have in fact shown that DNA strands can form highly complex knots, and computer simulations have broadly investigated the occurrence of knotted DNA topologies in various physical situations, e.g. viral DNA confined in a capsid (107; 108; 109; 111; 112).

Indeed, from the physical point of view, the formation of a knot in a sufficiently long flexible chain is a certain event which follows precise statistical laws (113; 114; 115; 116; 117), and the occurrence of a specific type of knot can be predicted with probabilistic methods.

Proteins on the other hand, do not behave like random flexible polymers, as they have evolved to fold reproducibly in a well-defined native state in physiologic conditions (118). As anticipated, the formation of a knot in a protein chain is usually assumed to represent a kinetic trap. Moreover, a substantial difference exists between the knots observed in proteins and those occurring in other biopolymers like DNA, which is *reproducibility*: the appearance of a given type of knot in a given point of a DNA strand can be assessed only in a statistical sense, while a knotted protein folds *always* with the same knot type in the same location. It is in this sense that a knotted protein differs from a knotted polymer as much as a tied shoelace differs from an entangled cord in a bag.

### 5.1.3 Knotted protein folding - a series of fortunate events

The characterisation of the folding process of a knotted protein is a difficult task, due to the complexity introduced by the topology. How does a knot form? By which succession of events the molecule entangles itself to form a topologically nontrivial structure?

A possible scheme has been suggested (106; 119), according to which the protein partially folds forming a loop, through which a terminus is later threaded to form the knot. It is worth noting nevertheless, that experiments indicate (120) that proteins fold in a two-state process, in which the molecule exists in a *all-or-none* fashion. The rapid transition from denatured to native state is therefore not supportive of a knot formation scheme assuming a partial folding of a protein region.

An alternative possibility has been recently explored (121; 122), that the formation of the knot occurs at the early stages of the folding process, when the molecule is still in a swollen configuration. This hypothesis was supported by experiments on fusion proteins (121; 122) and circularised, denatured knotted proteins (123). Nonetheless, it should be bourne in mind that the presence of the knot was not observed directly in these complexes, but inferred indirectly.

Presently, the sequence of events which guide a protein to fold in a knotted native state is far from being understood, and the reasons why a given protein needs a knot to perform its function are still unclear for many cases.

### 5.1.4 Identification and classification of protein knots

The study of protein knots requires a proper and unambiguous definition of the latter. In fact, from the mathematical point of view a knot is well defined only for *closed curves*; proteins, on the other hand, are linear open chains [1]. The proper identification of a knot in a protein therefore requires that the two termini are joined by a suitable *closure procedure*.

Different algorithms have been introduced to accomplish this task in an automated and un-subjective way. One possibility is to extend the termini outwards with respect to the protein's centre of mass: this procedure reduces the possibility of crossing the bulk of the molecule, but does not guarantee the uniqueness of the closure. Recently Millet (124) introduced a *statistical* scheme for the protein closure: a large number of directions for prolonging the termini are drawn randomly. The protein is then closed and the knot type is identified: the true knot type is the one occurring with largest probability.

In our work we adopted a simpler and more stringent scheme, which is also computationally efficient. The method is based on the idea that an unambiguous closure can be introduced if the termini are sufficiently exposed on the protein surface: in geometrical terms, this means that a terminus can be safely prolonged away from the protein if there exist a plane passing through the terminus and leaving all the protein on one side. When this condition is met for both termini, the latter are extended in a direction orthogonal to the plane in the outward direction, then joined through an arc. Because the identification of the knot type can be complicated by the presence of several coplanar bonds in the closed proteins, a small perturbation is added to the arc joining the termini. In Fig. 5.1 a closed globular protein (PDB code 1yveI) is shown.

The identification of the planes is performed with the method of the *perceptron* (125). In its original formulation, the perceptron is a simple mathematical model of a

[1] In what follows, we shall take into account only the backbone chain and the knots it forms; other knots or pseudo-knots formed by covalent bonds (e.g. disulfide bonds) will not be discussed.

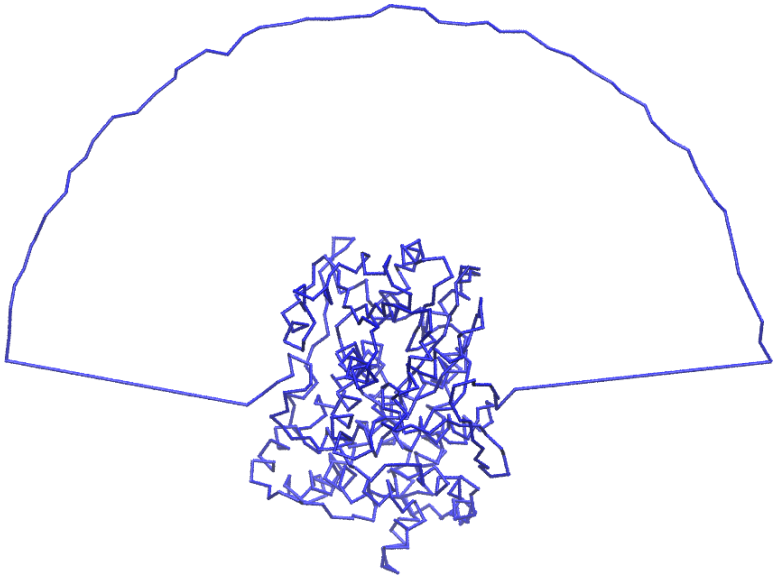

**Figure 5.1: Example of protein chain closure** - The knotted protein chain 1yveI is here shown in trace representation after the closure process. The arc, connecting the protrusions of the protein termini, is not smooth but a small random drift is added in order to prevent problems in the knot type identification (see text).

neural cell: it takes an $N$-dimensional vector $\vec{n}$ of input values and returns a response $f$. In formulæ:

$$f(\vec{n}) = \vec{r} \cdot \vec{n} + \vec{b} \tag{5.1}$$

where $\vec{r}$ is a *weight pattern*, and $\vec{b}$ is a bias. In our formulation, we search for the unit vector $\vec{n}$ applied to the terminus of the protein, having negative scalar product with *all* the vectors $\vec{r}_i$ joining the terminus and the other residues. If this vector exists, the terminus can be extended unambiguously in its direction; otherwise, the protein cannot be safely closed.

Once the protein chain has been circularised, the topological state must be identified. This classification is performed making use of *topological invariants*, i.e. properties which do not depend on the exact geometry of the ring: typical invariants used to characterise a knot are the Alexander polynomial, the Jones polynomial and the HOMFLY polynomial (126). The determination of the invariants require the projection of the three-dimensional chain on a plane in order to calculate the number and type of the crossings; this procedure can be extremely complex and ambiguous in the case of long and entangled chains like proteins: therefore, it is typically preceded by a simplification of the structure, consisting in a smoothening of the chain and a reduction of its

length. The procedure we applied in our work was based on a set of routines written by Micheletti *et al.* (107; 108), and on the Knotfind algorithm (127).

## 5.2 Sequence and structure comparison of proteins having different topology

In this section, we shall discuss a quantitative, comparative analysis of proteins having different topological state. This investigation is performed on a minimally redundant dataset of knotted and unknotted proteins.

### 5.2.1 Identification of the knotted and unknotted representatives

The Protein Data Bank as of December 2009 contained 6.2 $10^4$ entries: each of these was parsed into single chains, which were processed separately. In order to avoid incomplete or badly resolved structures (whose indetermination could result in an incorrect knot identification) we retained only those chains with length matching the nominal one, provided in the SEQRES PDB field, to within 25 amino acids. Also those chains, whose length was shorter than 50 a.a. or larger than 1000 a.a., were eliminated, as well as those with missing $C_\alpha$ coordinates. This sieving procedure returned 1.2 $10^5$ chains.

The closing procedure previously discussed was applied to these chains, out of which 6.4 $10^4$ could be circularised. For proteins constituted by identical monomeric chains, only one representative chain was considered, reducing the number of considered entries to 4.5 $10^4$.

Finally, this dataset was further processed to establish the knot topology of each entry; only 247 protein chains, listed in Table 5.1, were found to have nontrivial topology. The sets of knotted and unknotted proteins were affected by a large sequence redundancy: for example, as many as 194 of the 229 knotted proteins, are carbonic anhydrases. The primary sequence comparison of the entries revealed that less than 50 chains are non-identical in sequence.

The datasets were hence processed to achieve a uniform and minimally-redundant coverage in sequence space. The redundancy of the knotted protein set was removed at the stringent 10% sequence identity level using the web tool developed by Cedric Notredame [1]. The culling procedure returned the 11 representatives shown in Table 5.2. No significant structural relatedness was found among any pair of these representatives.

The large set of unknotted proteins was too large to be culled with the Notredame web tool; we therefore resorted to the standalone UniqueProt (128) program to efficiently remove the overall sequence similarity. Its iterative application with default parameters returned 2.4 $10^3$ unknotted representatives.

[1] Unpublished. The web address of the tool is **http://www.expasy.ch/tools/redundancy**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1a42A | 1am6A | 1azmA | 1bcdA | 1bicA | 1bnqA | 1bnuA | 1bnwA | 1bv3A | 1bzmA |
| 1cahA | 1caiA | 1cajA | 1cakA | 1calA | 1camA | 1cayA | 1cazA | 1cilA | 1cngA |
| 1craA | 1czmA | 1dmxA | 1dmyA | 1eouA | 1fljA | 1fqnA | 1fqrA | 1fr4A | 1fsqA |
| 1fsrA | 1g0eA | 1g0fA | 1g1dA | 1g3zA | 1g45A | 1g46A | 1g48A | 1g4jA | 1g4oA |
| 1g52A | 1g54A | 1gz0A | 1gz0B | 1gz0D | 1gz0F | 1gz0H | 1hcbA | 1heaA | 1hecA |
| 1huhA | 1i8zA | 1i90A | 1i91A | 1i9lA | 1i9mA | 1i9nA | 1i9oA | 1i9pA | 1i9qA |
| 1if4A | 1if5A | 1if6A | 1if7A | 1if8A | 1if9A | 1ipaA | 1j9wA | 1jv0A | 1keqA |
| 1kwqA | 1kwrA | 1lg5A | 1nxzA | 1nxzB | 1o6dA | 1oq5A | 1p7lA | 1qmgA | 1qmgD |
| 1rayA | 1razA | 1rg9A | 1rj5A | 1rj6A | 1rzaA | 1rzcA | 1rzdA | 1rzeA | 1s1hI |
| 1t9nA | 1tb0X | 1tbtX | 1te3X | 1teuX | 1tg3A | 1tg9A | 1th9A | 1thkA | 1ttmA |
| 1ugaA | 1ugcA | 1ugdA | 1ugeA | 1ugfA | 1uggA | 1urtA | 1v9eA | 1v9iC | 1vh0A |
| 1x7pA | 1xd3A | 1xd3C | 1xegA | 1xevA | 1xevB | 1xpzA | 1xq0A | 1yddA | 1yh1A |
| 1yo0A | 1yo1A | 1yo2A | 1yveI | 1z97A | 1zgeA | 1zgfA | 1zh9A | 1zjrA | 1zsaA |
| 1zsbA | 2aw1A | 2ax2A | 2cbaA | 2cbbA | 2cbcA | 2cbdA | 2cbeA | 2efvA | 2egvA |
| 2etlA | 2eu2A | 2eu3A | 2ez7A | 2fg6C | 2fg6D | 2fg6Z | 2fg7C | 2fg7X | 2fnkA |
| 2fnmA | 2fnnA | 2foqA | 2fosA | 2fovA | 2foyB | 2g7mC | 2g7mX | 2gehA | 2h15A |
| 2ha8A | 2hd6A | 2hfxA | 2hfyA | 2hkkA | 2hl4A | 2hocA | 2nmxA | 2nmxB | 2nn1A |
| 2nn1B | 2nn7A | 2nngA | 2nnoA | 2nnsA | 2nnvA | 2nwoA | 2nwpA | 2nwyA | 2nwzA |
| 2nxrA | 2nxsA | 2nxtA | 2o9cA | 2obvA | 2osfA | 2osmA | 2p02A | 2pouA | 2povA |
| 2qmmA | 2qo8A | 2qp6A | 2rh3A | 2vvbX | 2wegA | 2wehA | 2wejA | 3b4fA | 3bbdA |
| 3bbeA | 3bbhA | 3betA | 3bjxB | 3bl0A | 3c2wC | 3c2wH | 3c7pA | 3cajA | 3czvB |
| 3d0nA | 3d93A | 3d9zA | 3da2A | 3dbuA | 3dc9A | 3dccA | 3dcsA | 3dd0A | 3dd8A |
| 3dv7A | 3dvbA | 3dvcA | 3dvdA | 3eftA | 3f4xA | 3f8eA | 3ffpX | 3gz0A | 3hkqA |
| 3hkuA | 3hs4A | 3iaiA | 3iaiB | 3iaiD | 3ibiA | 3iblA | 3ibnA | 3ibuA | 3ic6A |
| 3iefA | 3ilkA | 3k2fA | 3ktyB | 3ktyC | 4cacA | 5cacA | | | |

**Table 5.1: Knotted protein list** - List of the 247 knotted protein chains found in the December 2009 PDB release.

| name | PDB | knot type | CATH | EC | knotted region |
|---|---|---|---|---|---|
| hypothetical protein | 2efvA | $3_1$l | | | 6-86 |
| plasmid pTiC58 VirC2 | 2rh3A | $3_1$l | | | 82-194 |
| N-succinyl-L-ornithine transcarbamylase (SOTCase) | 2fg6C | $3_1$r | 01:3.40.50.1370 **02:3.40.50.1370** | | 149-257 |
| methyltransferase (MT) domain of human TAR (HIV-1) RNA binding protein (TARBP1) | 2ha8A | $3_1$r | | | 83-167 |
| alpha subunit of human S-adenosyl-methionine synthetase (SAM-S) | 2p02A | $3_1$r | 01:3.30.300.10 02:3.30.300.10 03:3.30.300.10 | 2.5.1.6 | 38-328 |
| human carbonic anhydrase II (CA2) | 5cacA | $3_1$r | 3.10.200.10 | 4.2.1.1 | 11-260 |
| acetohydroxyacid isomeroreductase | 1qmgA | $4_1$ | 01:3.40.50.720 **02:1.10.1040.10** | 1.1.1.86 | 302-553 |
| photosensory core domain of aeruginosa bacteriophytochrome (PaBphP) | 3c2wH | $4_1$ | | | 5-302 |
| ubiquitin carboxy-terminal hydrolase (UCH) | 2etlA | $5_2$l | 3.40.532.10 | 3.4.19.12 | 1-233 |
| group I haloacid dehalogenase | 3bjxB | $6_1$r | | 3.8.1.10 | 46-288 |
| ribosomal 80S-eEF2-sordarin complex | 1s1hI | $3_1$r | | | 78-125 |

**Table 5.2: List of the knotted protein representatives** - CATH (6) and EC (72) codes are indicated where available; the knotted region refers to the PDB residue numbering. The chirality is indicated with a *l* or *r* tag appended to the knot type. CATH domains containing the knot are highlighted in boldface for multidomain proteins. The knot region is defined by taking the strictly knotted protein segment returned by the Protein Knot server (129) and extending it by 20 amino acids on both sides. For protein chain 2p02A, which is not recognised as knotted by the server, the strictly knotted protein segment was identified using the method of ref. (130). The knot in the last entry (1s1hI) has a probably artifactual origin, see Results and Discussion.

### 5.2.2 Knots spectrum and knot chirality

Before discussing the comparative analysis, we report in this section the results of our investigation about the properties of knotted proteins in the dataset.

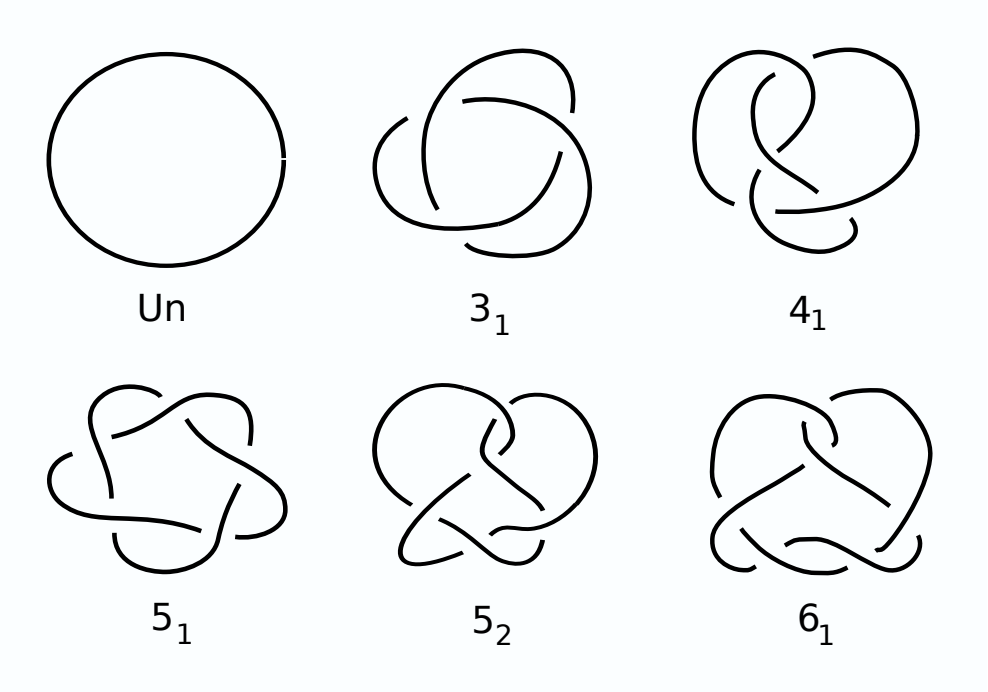


**Figure 5.2: Knot diagrams of the simplest knots** - The knot diagrams of the knots discussed in the text are here shown. The number labelling each knot is related to the number of crossings; clearly, the Unknot has zero crossings.

The simplest knot type, $3_1$ (see Fig. 5.2), also known as trefoil knot, is by far the most abundant knot type in the initial redundant set, and is also the most abundant in the representative list of Table 5.2. Indeed, 7 of the 11 entries are trefoils.

Among the trefoil representatives in Table 5.2 we have identified the shortest known knot, consisting of only 10 amino acids. The knot is found in the cryo-em resolved PDB entry 1s1hI (ribosomal 80S-eEF2-sordarin complex) (131). Several clues point to its possible artifactual nature: the knotted region (from a.a. 98 to 105) is listed in the structure file as having highly non-standard stereochemical parameters. Furthermore, the associated temperature-factor values are in excess of 100, and are hence indicative of poor compliance with the electron-density map. For these reasons the knot in entry 1s1hI is probably artifactual and has been excluded from the comparative analysis.

More complex knot types, $4_1$ and $5_2$ (see Fig. 5.2), are represented by two and one entries respectively in Table 5.2 and, in any case, by very few chains in the redundant

set. The survey of the December 2009 PDB release did not return knots more complex than the $6_1$ type (see Fig. 5.2), which was reported in ref. (106).

It is interesting to observe a parallel between the chronological succession of the first PDB release of the various types of protein knots and the complexity of the knots. In fact, the first structures containing $3_1$, $4_1$, $5_2$ and $6_1$ knots were resolved or released, respectively, in 1988 (PDB entries 4cac and 5cac (132)), 1996 (PDB entry 1yve (133)), 2004 (PDB entry 1xd3 (134)) and 2007 (PDB entry 3bjx (135)). Although the steady increase of the PDB cannot be viewed as resulting from the repeated addition of structures sampled uniformly in "protein structure space", it is natural to assume that the chronological succession of the knots "discovery" is inversely correlated to the abundance of the various knot types.

This qualitative consideration is supported by the fact that, in compact flexible polymers, the abundance of the simplest knot types decreases with knot complexity (107; 136). One notable point of these polymeric reference systems is that, for entropic reasons, the knot type $5_1$ is appreciably less abundant than $5_2$, which has the same nominal complexity (107; 108). The absence of the $5_1$ knot in presently-available proteins (a fact previously also related to the unknotting number (103)), may thus reflect the still limited pool of known knotted proteins and might hence populate in the future.

Finally, we discuss the extent to which knots of different handedness occur among knotted proteins. Apart from the $4_1$ knot which is achiral, knots $3_1$, $5_2$ and $6_1$ can exist in left- and right-handed versions. Previous observations made on a redundant set of proteins folded in trefoil knots concluded that, except for a single protein entry, all other ones were right-handed trefoils. For the most numerous family of knotted proteins, namely carbonic anhydrases, the bias towards right-handed knots was related to the intrinsic chirality of the $\beta\alpha\beta$ motif adopted by such enzymes (103).

The investigation of the handedness in this latest dataset, where sequence redundancy has been removed, provides a novel context for examining the problem. As reported in Table 5.2, the balance between right- and left-handed knots is 5 to 3, respectively. The near equality of the populations is thus compatible with the null hypothesis that left- and right-handed protein knots occur in equal proportion (after removal of the biases of representation due to sequence redundancy of otherwise detectable evolutionary relationships).

### 5.2.3 Sequence → structure relationship

As previously anticipated, the ability of a protein to fold in a knotted state represents a still unsolved problem. A systematic comparison of knotted proteins with unknotted ones, sharing with the former sequence or structure relations, represents a promising strategy to understand the peculiarities of topologically entangled protein chains.

In this subsection we tackle one facet of the problem. Specifically, we discuss how primary-sequence similarities reverberate in relatedness of the knotted/unknotted topological state. To this purpose, for each of the 11 representatives in Table 5.2 we performed a PDB-wide BLAST (137) search for related sequences. The search was restricted to sequences of proteins of known structure (i.e. contained in the PDB) because without the structural data it would not be possible to compare the knottedness of pairs with related primary sequences. The sequence comparison analysis, mainly performed by C. Micheletti, started by first running the PDB-wide BLAST queries using a stringent E-value threshold (0.1). False positives are hence not expected to occur appreciably among the returned entries. Only for three protein chains, namely 5cacA, 2fg6C and 2ha8A, the number of significant matches was larger or equal to 10. Incidentally we mention that, consistently with the probable artifactual origin of the knot in entry 1s1hI, all the 10 significant BLAST matches of 1s1hI were unknotted protein chains.

All the returned matches for the 5cacA human carbonic anhydrase and the 2ha8A methyltransferase domain of the human TAR RNA binding protein (TARBP1-MTd), consisted exclusively of a dozen knotted proteins, all with the same knot type. These matches were therefore not informative for the purpose of understanding if and how differences in sequence reverberate into differences of knotted state. On the contrary, the BLAST matches of the trefoil-knotted N-succinyl-ornithine transcarbamylase (SOTCase), associated to the PDB entry 2fg6C (138), proved particularly interesting as only 7 of the tens of matching entries are knotted (all in a trefoil knot).

To advance the understanding of the precise type of sequence relatedness of the SOTCase and its knotted and unknotted homologs, the matching BLAST sequences were used as input for a CLUSTALW multiple sequence alignment (139). The results were used, in turn, to establish a phylogenetic relationship between the related proteins using a neighbour-joining bootstrapping algorithm (140). The method associates to each branch of the phylogenetic tree a percent confidence estimated from the occurrence of the branch in 1000 repeated phylogenetic reconstructions using only a subset of the aligned amino acids.

The phylogenetic tree for the SOTCase is represented in Fig. 5.3a. The tree shows that the knotted entries appear in two terminal branches sharing a common root. Each branch gathers entries that are highly similar in sequence; in fact their sequence identity (computed by dividing the number of aligned identical amino acids by the average length of the two compared proteins) is not smaller than 90%. The sequence identity across the two branches has the much smaller, but still significant, average value of 40%. The homology relation among all members of the phylogenetic tree is further confirmed by the fact that those, for which CATH (6) code is known, belong to the same CATH family. On the other hand, the robustness of the separation of

the knotted sequence subgroup from the unknotted one is strongly suggested by the bootstrap algorithm, with a confidence level larger than 99%.

Among the knotted and unknotted entries, the average level of sequence identity is about 20%, with a standard deviation of 7%. Indeed, it is interesting to observe that few knotted/unknotted pairs can have a level of mutual sequence identity even larger than knotted pairs. For example the knotted chain 2g68A has a sequence identity of 33% and 38% respectively, against 1js1X (knotted) and 1pvvA (unknotted).

The present results offered a novel insight into the possible mechanisms that have led to the appearance of knotted proteins. In particular, the phylogenetic tree structure suggests the existence of a simple evolutionary lineage between the sets of knotted and unknotted proteins shown in Fig. 5.3a. In fact, both groups of trefoil knotted proteins, which have a limited mutual sequence identity, appear to have commonly diverged from the main tree of unknotted entries.

The implications are twofold. On the one hand, the robust conservation of the knotted fold in the two sequence-diverged knotted groups suggests the functionally-oriented characteristics of the knotted topology. Indeed, it had already been pointed out for one member of this family (102) that the active site is located close to the knotted region, a fact that led to speculate that knottedness would confer a necessary mechanical rigidity to the protein as a whole or to the active site (141; 142). On the other hand, the existence of a single knotted branch indicates that the knot appearance, and its subsequent conservation, are rare evolutionary events.

Further clues about the biological rationale behind the evolutionary pathways that have led to the emergence/conservation of the knotted structures in Fig. 5.3a ought to be addressed using more powerful tools than the present sequence-based analysis: in particular, a more general reconstruction of the phylogenetic relatedness should be accomplished within a genome-wide perspective for the organisms involved.

### 5.2.4 'Knot-promoting' loops in SOTCase

Valuable insight into the fundamental similarities and differences in the entries appearing in the tree of Fig. 5.3a can be obtained by inspecting their structural alignment. In this case it appears particularly appropriate the use of a multiple non-sequential structure alignment method: in fact, correspondences are sought between proteins with different knotted state, and hence with expected differences in fold organisation.

To this purpose we used the MISTRAL (89) multiple structure alignment method. The alignment tool was used for two reasons. First, it has been shown to yield a reliable estimate of the statistical significance of a given alignment and, secondly, it can detect structurally-corresponding regions that do not have the same succession or directionality along the primary sequence of the input proteins. The necessity to

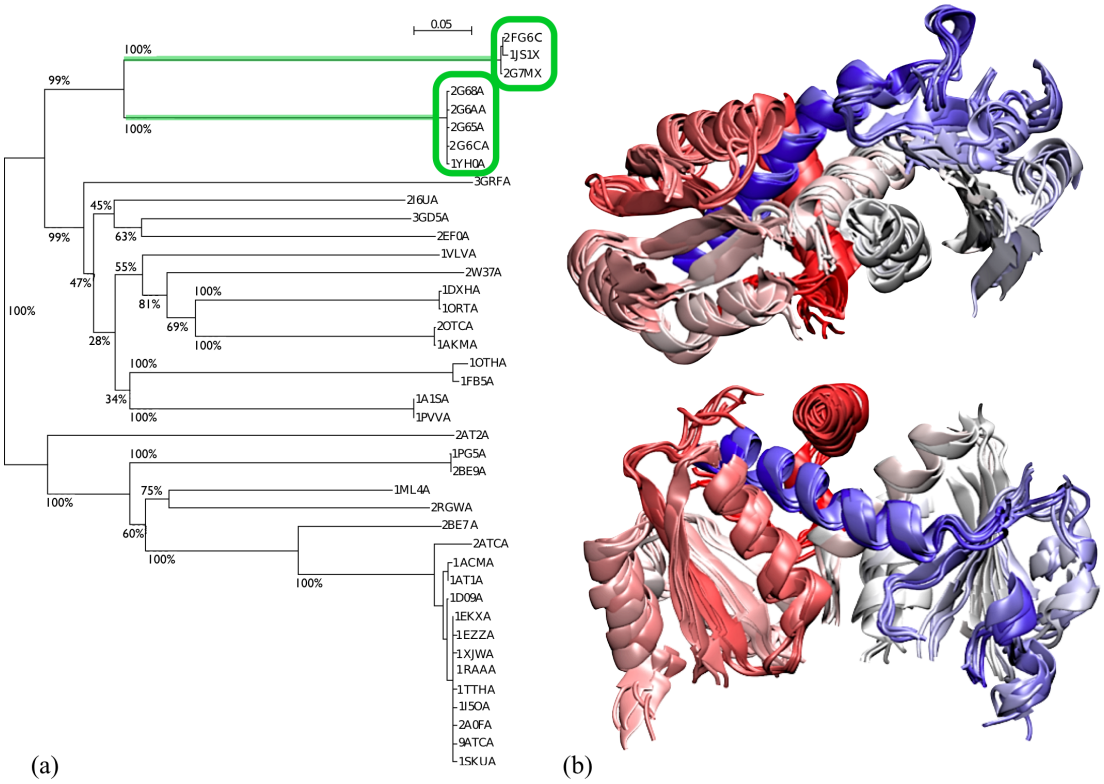


**Figure 5.3: SOTCase and homologous proteins: phylogenetic tree and structural alignment core** - (a) The phylogenetic tree was obtained by applying a neighbour joining algorithm (140) to the CLUSTALW multiple sequence alignment of SOTCase and its sequence homologs. The branches' length reflects the percentage sequence dissimilarity (5% gauge shown at the top). The numbers at the nodes, calculated by the bootstrap algorithm, indicate the percent robustness of the separation of two bifurcating branches. The two branches involving knotted proteins (all trefoils) are highlighted in green. (b) Two orthogonal views of the MISTRAL alignment core of six representatives of the SOTCase homologous proteins, namely 2fg6C (knotted), 2i6uA, 2g68A, 2at2A, 1pg5A and 1ortA. These proteins are 313 amino acids long on average. Their alignment core consists of 212 amino acids at an average RMSD of 1.9Å. The colour scheme red → white → blue follows the N to C sequence directionality.

account for such generalised relationships in proteins has emerged from recent analysis of protein evolutionary relationships (143).

All pairwise structural alignments between the representatives of the unknotted and knotted proteins were computed. Among those with a $p$-value smaller than $5.0\,10^{-3}$ we singled out those which involved at least 40% of the protein region that encompasses the knot. The latter is defined by taking the chain portion that is strictly occupied by the knot according to the criterion of ref. (129) and extending it by 20 amino acids on both sides of the primary sequence (unless a terminus is closer): all the selected alignments are provided in table 5.3.

The proteins appearing in the phylogenetic tree can be all simultaneously structurally-aligned. Their aligned core consists of as many as 192 amino acids, which is a substantial fraction of the full proteins (which have an average length of about 310 a.a.). Over the core region, the average RMSD of any pair of matching amino acids is less than 2 Å. The good structural superposability of the protein set (which we recall includes protein pairs with average mutual sequence identity of about 20%) is exemplified in Fig. 5.3b where the alignment of 6 proteins taken from the various primary branches of the phylogenetic tree is shown.

The detailed pairwise structural comparison indicates that members of the two knotted branches admit a good structural superposition over the full protein length (and, in particular, over the knotted region). To highlight the salient differences between the knotted and unknotted entries in the tree we analysed all the pairwise structural superpositions of the knotted SOTCase with the unknotted homologs. This investigation generalises the structural comparative inspection of two specific instances of knotted and unknotted carbamylases carried out in ref. (102).

The results are best illustrated considering the closest matching pair, namely the SOTCase and PDB entry 1ortA. In spite of their limited mutual sequence identity, which is about 25%, these proteins admit a very good structural superposition, see Fig. 5.5a,b. Indeed, as many as 246 of their amino acids (which are 321 and 335 in total for chains SOTCase and chain 1ortA, respectively) can be superposed with an RMSD as small as 2.5Å. The alignment respects the overall sequence directionality of the chains. The few non-matching regions are typically insertions in exposed stretches of the sequence, corresponding to small loops protruding out of the surface of the molecule, which have no particular bearing on the protein topology.

The case is different for two regions of the SOTCase: the proline-rich segment comprising amino acids 174–182, and the segment 235–255; both regions are located in proximity of the active site (residues 176-178, 252). As shown in Fig. 5.5a, these loops, which do not contain highly hydrophobic segments (see Fig. 5.4), have a particular mutual concatenation which directly impacts on the protein knotted state. In fact, the

virtual excision (bridging) of these two segments, which both have a small end-to-end separation, results in the elimination of the knot from SOTCase.

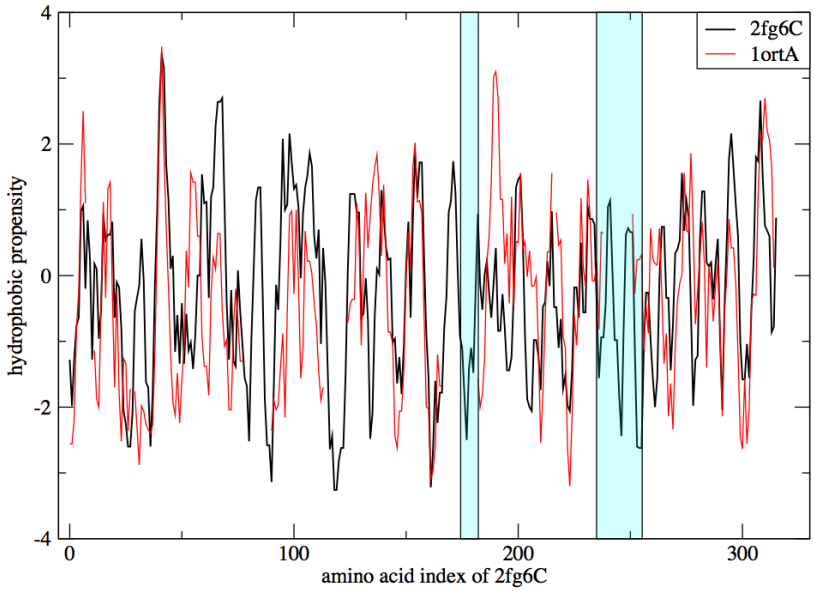


**Figure 5.4: Hydrophobicity profile - 2fg6C** - Hydrophobicity profiles of the knotted protein 2fg6C and the structurally-matching amino acids of the unknotted partner 1ortA. The knot-promoting segments (174–182, 235–255) are highlighted by the light blue boxes. The hydrophobicity was calculated using the Kyte and Doolittle scale and an averaging window of 5 amino acids.

Virnau *et al.* (102) had observed that the knottedness of the transcarbamylase of *X. Campestris* was probably due to the excess length of the region comprising residue 176 with respect to the human analog. This observation is reinforced by the present general sequence- and structure-based systematic comparison which additionally points out the systematic absence of a second loop segment 235–255 in the unknotted homologs of the SOTCase. The results provide a quantitative basis for suggesting that some light on the process of protein knot formation can be shed by targeting these regions in suitable mutagenesis experiments. It would be particularly interesting to analyse whether both of the identified 'knot-promoting' loops need to be excised to produce an unknotted native state, or if only one would suffice.

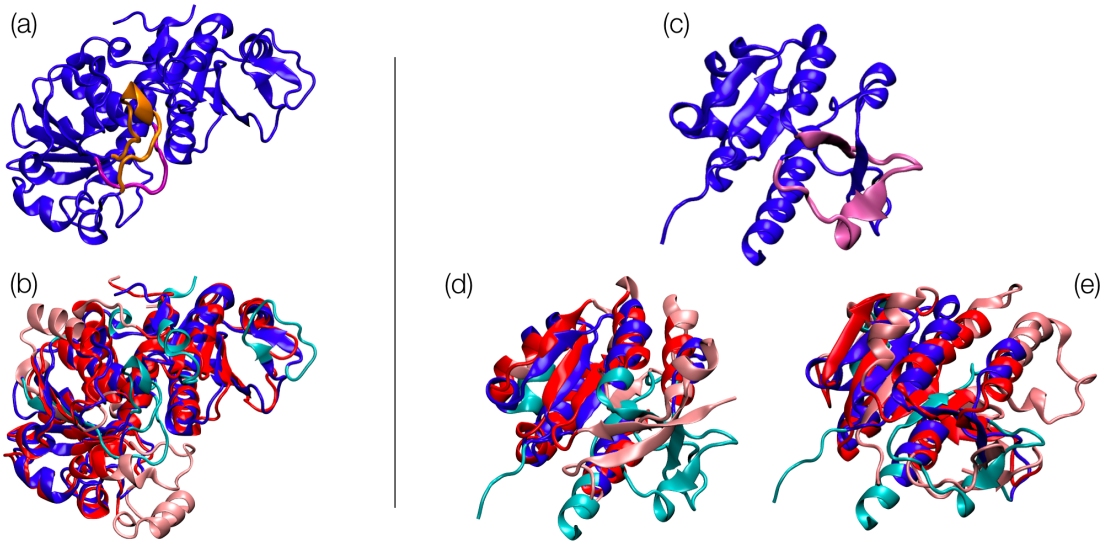


**Figure 5.5: Structural alignment of knotted and unknotted proteins** - SOTCase (a) is shown in cartoon representation; the knot-promoting loop segments are highlighted in orange and purple. The MISTRAL alignment with unknotted entry 1ortA is shown in panel (b): aligned residues are colored in blue and red, respectively, while non aligned residues are correspondingly colored in cyan and pink. Knotted protein TARBP1-MTd is shown in panel (c) with the knot-promoting loop segment highlighted in purple. The MISTRAL alignments of TARBP1-MTd with the unknotted proteins 1b93A and 1hdoA are shown in panels (d) and (e), respectively.

### 5.2.5 Knot-promoting loops in other proteins

The results discussed in the previous subsection indicate that knotted proteins appear to be sparsely distributed in sequence space. In fact, only for one of the representatives in Table 5.2, it was possible to establish significant sequence-based relationships with unknotted proteins.

We now discuss structural similarities between knotted and unknotted proteins, irrespective of the level of primary sequence relatedness. The search was performed by carrying out MISTRAL structural alignments of each of the knotted representatives in Table 5.2, against an extensive set of about $2.4\,10^3$ unknotted protein chains. The top-ranking alignments are reported in Table 5.3.

Hereafter we focus on a limited number of cases which, regardless of their ranking in alignment quality, can be aptly used to highlight interesting relationships between knotted and unknotted pairs. In particular, they might possibly be used to shed light on important kinetic or thermodynamic mechanisms that guide or otherwise favour the formation of knots in naturally occurring proteins.

We first discuss the limited number of cases where the alignment suggests the presence of knot-promoting loop segments, analogously to the case of the SOTCase and chain 1ortA. These segments are identified using two main criteria: (i) the segments' ends must be sufficiently close that they could be virtually bridged by very few amino acids; (ii) the bridging/excision operation should lead to an unknotted conformation.

The automated search for such segments returned positive matches for three representatives. One of them was the same SOTCase chain, which we discussed in previous sections. The other chains were the aforementioned TARBP1-MTd and the photosensory core module of *Pseudomonas aeruginosa* bacteriophytochrome (PaBphP, PDBid 3c2wH).

#### 5.2.5.1 TARBP1 methyltransferase domain

TARBP1-MTd aligns well with two unknotted protein representatives that have very different overall structural organisation. Despite the differences, discussed hereafter, the alignments consistently indicate that loop 101-123 is a knot-promoting loop for chain A of TARBP1-MTd.

The alignment against the unknotted protein chain 1b93A (144) comprises 87 amino acids (at 3.5 Å RMSD) and covers the entire knotted region with the exception of the above mentioned segment. The fact that the ends of the segments are less than 5Å apart, readily suggests that the excision of the fragment ought to result in an unknotted protein with structure analogous to the 1b93A chain. The inspection of the hydrophobicity profile based on the Kyte and Doolittle scale (145) (see Fig. 5.6) indicates that one of the regions with high hydrophobicity falls within the knot-promoting loop. In

| knotted protein PDBid | unknotted protein PDBid | fraction of aligned knot residues |
|---|---|---|
| 2fg6C | 1ortA | 0.706 |
| 2ha8A | 2qipA | 0.647 |
| 2ha8A | 1b93A | 0.565 |
| 2ha8A | 1f51E | 0.494 |
| 1vh0A | 1b93A | 0.596 |
| 2efvA | 2rjiA | 0.531 |
| 2ha8A | 3fhkA | 0.412 |
| 3bbeA | 1aoxA | 0.443 |
| 2ha8A | 1u7oA | 0.459 |
| 2ha8A | 2b98A | 0.494 |
| 2fg6C | 1a4iA | 0.560 |
| 1vh0A | 1d0iA | 0.426 |
| 1vh0A | 2z5vA | 0.468 |
| 2qmmA | 1hdoA | 0.651 |
| 2ha8A | 1hdoA | 0.553 |
| 2ha8A | 1d0iA | 0.518 |
| 2ha8A | 3gpgA | 0.518 |
| 2efvA | 1d8jA | 0.494 |
| 2ha8A | 1c25A | 0.647 |
| 3ktyC | 121pA | 0.413 |

**Table 5.3: Top ranking knot-unknot alignments** - Top ranking MISTRAL alignments involving representatives of knotted and unknotted chains. In order to account for the different topology of the compared proteins, the alignments were obtained with the following non-default MISTRAL parameters: the alignment tolerance was set to 6.0 Å; the minimum segment length was set to 10 amino acids. For each alignment we report, in the third column, the percentage of the knotted region (defined in Table I of the main article) that takes part in the structural alignment. The listed pairs include only significant MISTRAL alignments (p-value $\leq 5\,10^{-3}$) where the percentage of the aligned knotted region is larger than 40.

analogy with what suggested in ref. (146) for YibK, it is therefore possible that the kinetic accessibility of the knotted state is enhanced by contacts that this region forms with other parts of the protein.

The topologically-important role of the segment is further highlighted by the alignment with the 1hdoA chain. At variance with the case of 1b93A, the good alignment does not involve regions that have the same succession, along the primary sequence, in the two proteins. This is readily ascertained by the inspection of the structural diagram of Fig. 5.7a,b where it is possible to appreciate the different "rewiring" of several corresponding secondary structure elements. In this case too, the alignment comprises the knotted region with the exception of the previously mentioned segment. This reinforces the previous suggestion that the removal of the segment ought to result in an unknotted folded configuration.

#### 5.2.5.2 PaBphP photosensory core module

The "figure-of-eight" knot in protein PaBphP (147) spans a very large portion of the photosensory core module of PaBphP (a.a. 24 to 282). This protein is composed of three domains: named PAS (Per-ARNT-Sim), GAF (cGMP phosphodiesterase/adenyl cyclase/FhlA) and PHY (phytochrome) domains. The GAF domain is known to be present in several sequence-unrelated proteins and, in fact, it represent the core region of the good alignment of PaBphP photosensory core module with the non-homologous chain 2b18A (148).

The alignment singles out the segment of amino acids 203 to 256 as a knot-promoting loop. Indeed, while the knot length is very large, the knot appears to result from the "threading" of the N-terminal domain through the above mentioned loop. As for SOTCase, the hydrophobicity profile (see fig. 5.8) does not provide a definite indication that the loop region takes part to contacts aiding the kinetic accessibility of the knotted native state.

The removal of the loop, as readily seen from Fig. 5.9, leads to an unknotted structure, and therefore suggests that, like the other cases, it could be profitably targeted in mutagenesis experiments to ascertain its role in the process of knot formation.

### 5.2.6 Other correspondences of knotted and unknotted proteins

The analysis performed and discussed so far was based on the identification of knot-promoting regions suggested by significant alignments of the knotted representatives in Table 5.2 against unknotted representatives. Only for the three representatives discussed above it was possible to identify such correspondences on the basis of available structural data.

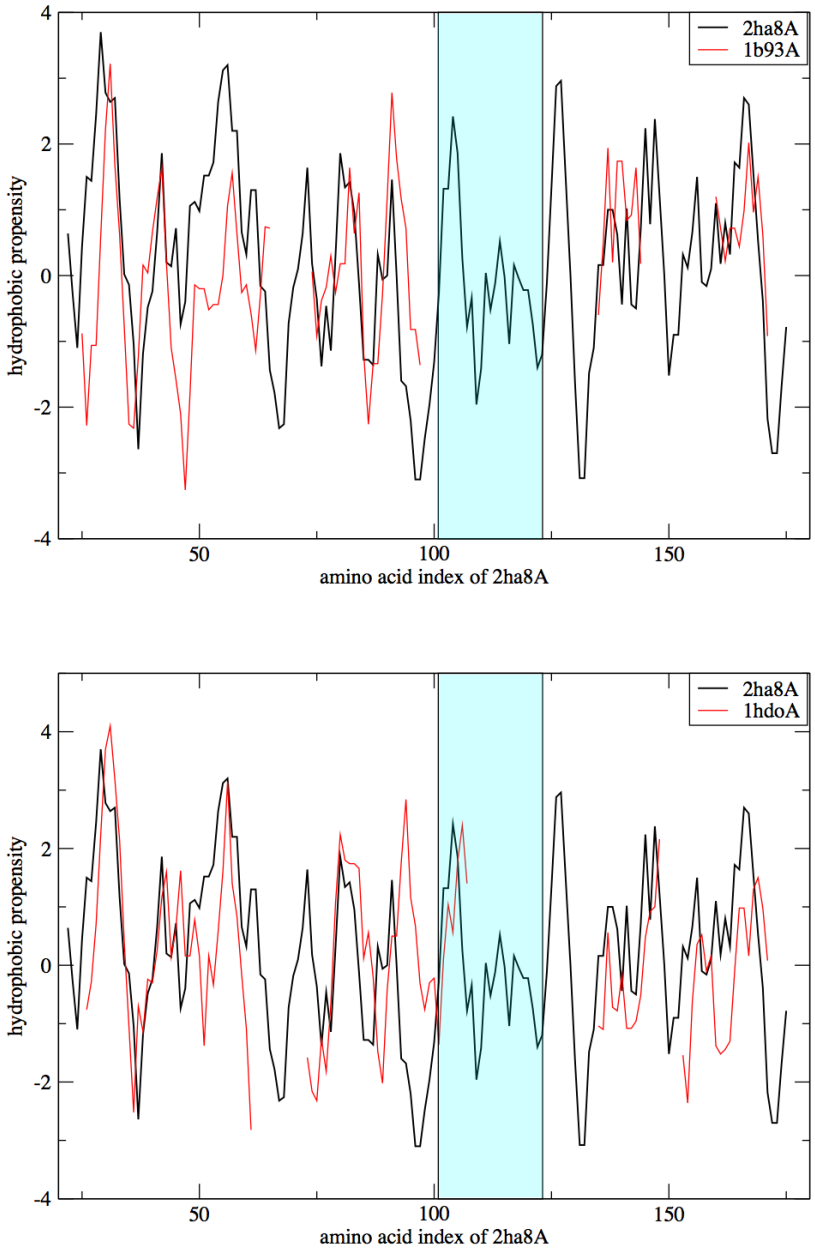


**Figure 5.6: Hydrophobicity profile - 2ha8A** - Top: hydrophobicity profiles for the knotted chain 2ha8A and the MISTRAL structurally-matching amino acids of the unknotted chain 1b93A. The knot-promoting segment (101–123) is highlighted by the light blue box. Bottom: hydrophobicity profiles for the knotted chain 2ha8A and the MISTRAL structurally-matching amino acids of the unknotted chain 1hdoA. The knot-promoting segment (101–123) is highlighted by the light blue box. Notice that, at variance with the case in the previous figure, the MISTRAL alignment of 2ha8A and 1hdoA is non-sequential and shows two gaps, one of which is the knot-promoting segment. The hydrophobicity profiles show very similar patterns for the aligned regions of the two proteins.

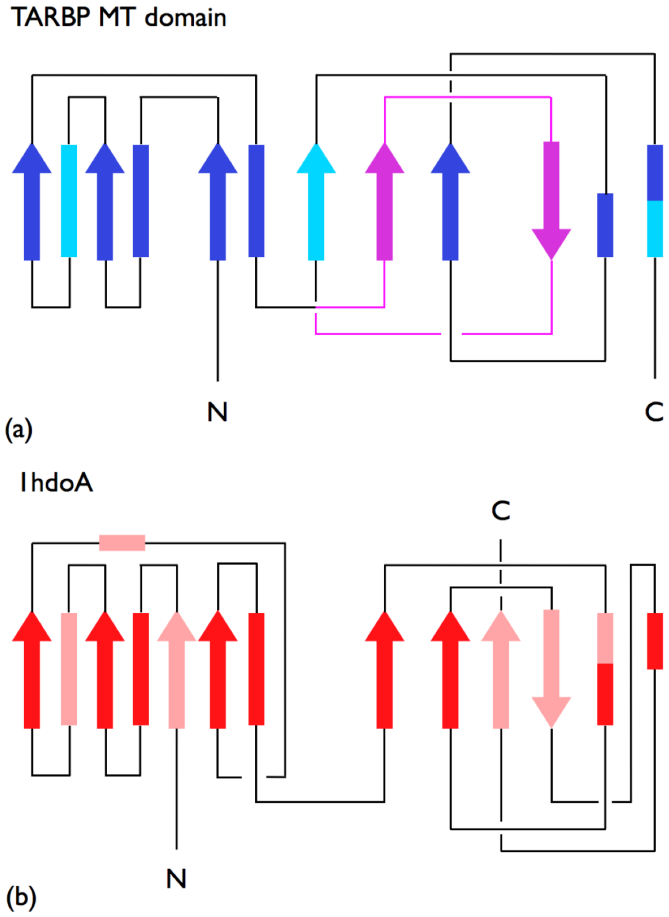


**Figure 5.7: Two-dimensional diagrams of the secondary and tertiary organisation of the knotted TARBP1-MTd and unknotted counterpart** - Two-dimensional schematic diagrams of the secondary and tertiary organisation of the knotted TARBP1-MTd (PDBid 2ha8A) (a) and unknotted protein chain 1hdoA (b), which admit a significant structural superposability (see Fig. 5.5). The colour-coding of the aligned and non-aligned secondary elements and of the knot-promoting loop follows the one in Fig. 5.5. The overall correspondence of the secondary elements is manifest, despite noticeable differences in their "wiring" which reflect in (i) a different fold organisation and (ii) a different knotted state.

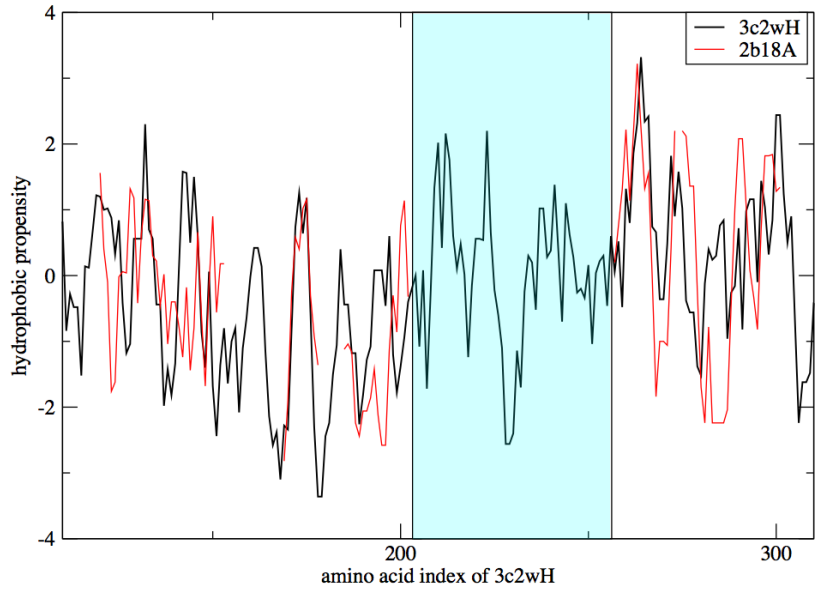


**Figure 5.8: Hydrophobicity profile - 3c2wH** - Hydrophobicity profiles for the knotted chain 3c2wH and the MISTRAL structurally-matching amino acids of the unknotted chain 2b18A. The knot-promoting segment (203–256) is highlighted by the light blue box.

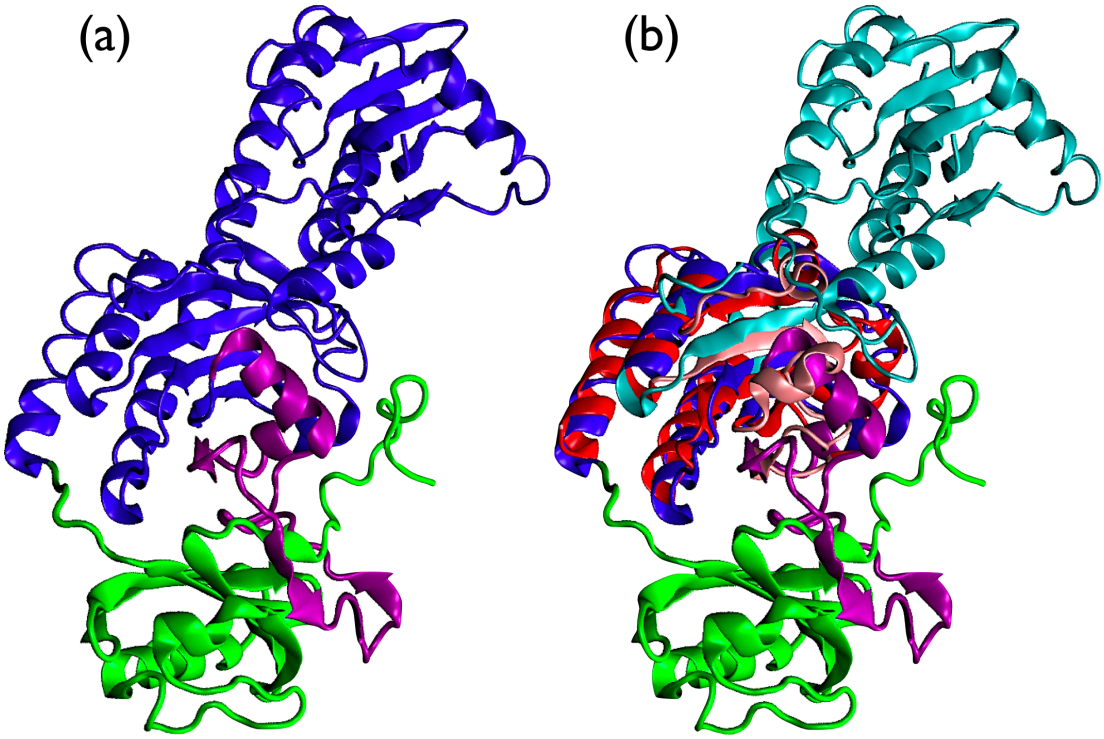


**Figure 5.9: Knotted photosensory core module of PaBphP** - Knotted photosensory core module of PaBphP (a) and its alignment with the unknotted chain 2b18A (b). In the knotted structure the knot-promoting loop is highlighted in purple, while the N-terminal domain, which threads through the loop, is shown in green. In panel (b), the aligned residues of knotted and unknotted proteins are coloured in blue and red, respectively, while non aligned residues are correspondingly coloured in cyan and pink. The N-terminal PAS domain (green) and C-terminal PHY domain (cyan) are well-separated by the aligned region, which instead covers almost completely the central GAF domain of PaBphP photosensory core module.

Yet, it is interesting to point out that for two other representatives, namely chains 2etlA (ubiquitin carboxy-terminal hydrolase, UCH) and 2p02A (alpha subunit of human S-adenosylmethionine synthetase, hereafter $\alpha$-SAM-S), good structural matches involving the knotted region were found against unknotted structures. At variance with previous cases, however, these matches do not suggest the possibility to unknot the protein by a simple excision operation. Yet, they are interesting for the purpose of understanding how continuous is the structure space between knotted and unknotted PDB entries.

The two examples are shown in Fig. 5.10. Panel (b) presents a superposition of the knotted UCH (149), which is the only $5_2$ knot representative, against the unknotted entry 1aecA (150). The alignment, though not spanning the entirety of the protein structures, highlights a good correspondence of secondary and tertiary structure elements.

Analogous considerations, hold for the alignment of $\alpha$-SAM-S (151) and 2bx4A (152) (Fig. 5.11), whose mutual sequence identity is less than 10%. The alignment highlights the threefold symmetry of the knotted protein, which however, builds on a non-trivial domain organization which results in a trefoil knot.

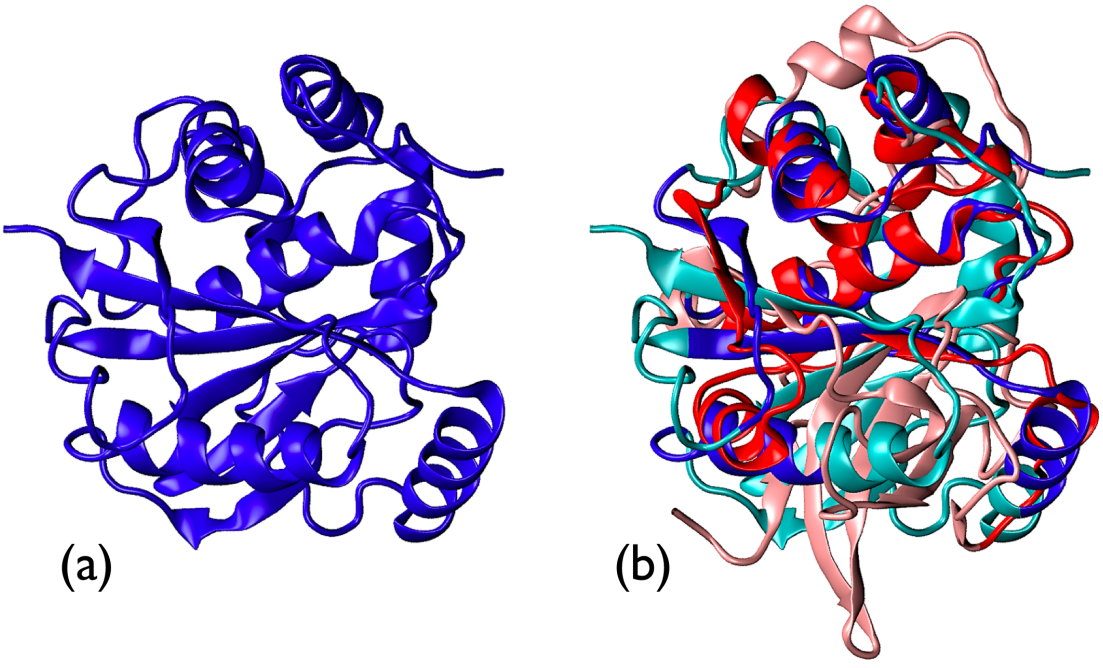


**Figure 5.10: Knotted protein UCH** - Knotted protein UCH (a) and its alignment with the unknotted chain 1aecA (b). The aligned residues of the knotted and unknotted protein are colored in blue and red, respectively while unsaturated colors (cyan and pink) are used for non-aligned residues.

## 5.3 Summary

The topology of protein chains represents an interesting open problem: the existence of proteins that fold in a knotted native state provide a most interesting avenue to characterise the interplay of kinetic and thermodynamic effects in protein folding.

In this chapter we discussed proteins in different topological states sharing a relevant sequence or structural similarity. Specifically, we performed a dataset-wide search, among knotted and unknotted representatives, for sequence-related and/or structurally superposable protein pairs having different knotted states.

A sequence alignment among the proteins of the dataset allowed us to identify a specific SOTCase, namely 2fg6C, whose phylogenetic tree comprises both knotted and unknotted entries. Interestingly, the few knotted homologs -less than 1/6 of the whole set- were gathered in two commonly-rooted sub-branches of the phylogenetic tree, while the remaining branches were occupied by unknotted proteins. From this fact one may argue that the appearance of a knot in a protein lineage is a rare evolutionary event. On the other hand, it must be noticed that the knotted sub-branches did not contain unknotted entries: this points to a possible role of the knot in the biological activity of the proteins, because of which it has been preserved by evolutionary pressure.

The members of the SOTCase family showed a considerable degree of structural similarity. In order to pinpoint their differences in fold organisation we performed a structure-based alignment. We found that the knotted domains differed from the unknotted counterparts, for the presence of two additional short segments with a small end-to-end separation. The bridging of these knot-promoting loop segments, that is their removal from the primary sequence, ought to result in an unknotted native state equivalent to the one of the unknotted homologs.

Prompted by the identification of these knot-promoting loops in SOTCases, we performed structural alignments among knotted and unknotted pairs of proteins in the dataset: at variance with the sequence, the structural investigation revealed several significant knotted/unknotted correspondences. In an appreciable number of instances, these correspondences involved a substantial fraction of the region where the knot is accommodated. Also in these cases, knotted proteins appeared to differ from the unknotted partner by the presence of knot-promoting segments analogous to those identified in the alignments involving the SOTCase. The results therefore point to the key role that these specific protein segments play for the global knotted topology of the folded protein: they might thus represent ideal candidates for mutagenesis or excision experiments, to monitor the impact of these regions on the process of knot formation. Moreover, the quantitative comparison of the internal fluctuations of aligned knotted/unknotted regions indicated a possible role of the knot in modulating the dynamics of these proteins, suggesting that further investigations in this directions could provide a major insight in the biological function of the knot.

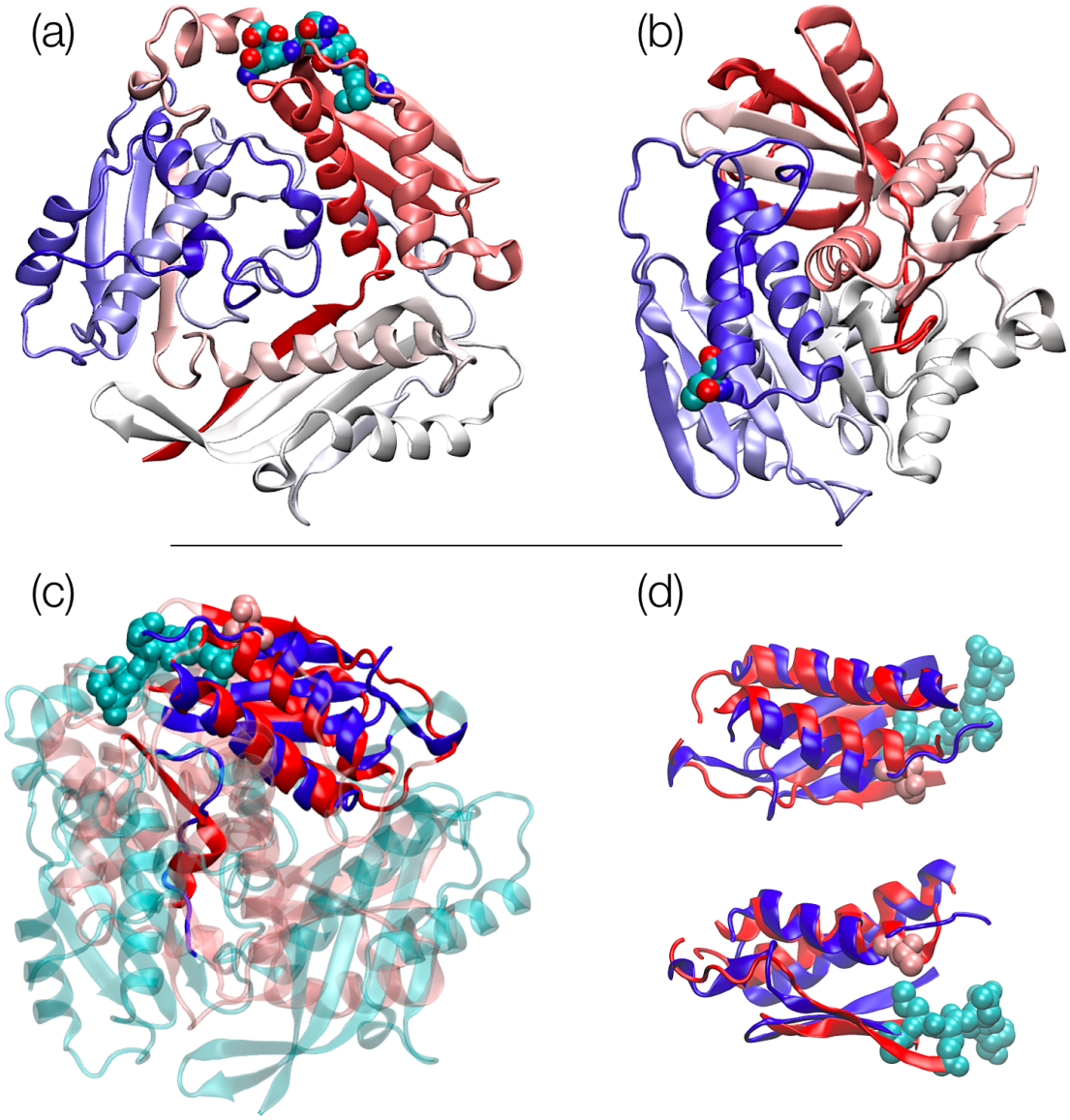


**Figure 5.11: Knotted protein $\alpha$-SAM-S** - Knotted protein $\alpha$-SAM-S (a) and unknotted protein 1bx4A (b), colored according to the residue index (red-white-blue); bottom, the structural superposition of these two entries where the aligned residues of knotted and unknotted proteins in the bottom row are coloured in blue and red, respectively, while non-aligned residues are correspondingly coloured in cyan and pink. Panel (c) shows the whole structures, while in panel (d) two orthogonal views of the sole aligned regions are presented. In all panels catalytic residues are included in Van der Waals representation. In panel (a), the knotted topology of $\alpha$-SAM-S can be readily perceived following the colouring of the chain.

# 6

# Concluding remarks

This thesis was largely focussed on the internal dynamics of globular proteins, and in particular on its relationship with protein structure.

The characterisation of the internal dynamics of proteins can be made with a variety of different methods, ranging from atomistic MD simulations to coarse-grained models, each contributing with its specific peculiarity. These tools provide a general and comprehensive picture of the motions occurring in proteins, from the small-scale vibrations of a residue side-chain to the collective fluctuations involving a large number of amino acids.

These concerted movements, which often accompany and support the biological activity, are at the heart of our investigations. In particular, these large-scale conformational changes have been used to identify quasi-rigid domains in proteins and highlight dynamical consistencies among structures lacking major similarities.

The possibility to identify, in a protein structure, those regions undergoing minor internal fluctuations while performing large-scale displacements relative to the protein centre of mass represents the basic objective of many methods of investigation: accelerated schemes to efficiently explore the conformational space, and reliable coarse-grained descriptions of protein structures to be used in docking algorithms are among the noteworthy possible applications.

The *reductio ad essentiam* of the structural and dynamical features of proteins can also be used to highlight those properties that are shared by molecules which appreciably differ by both sequence and structural organisation. This dynamical similarity encoded in a variety of architectures motivated the development of specific algorithms to perform dynamics-based alignments, i.e. to superpose two proteins maximising not only their structural similarity but also the consistency of their internal dynamics. This method complements the available sequence- and structure-alignment tools to investigate the relation between sequence, structure and function in proteins.

Finally, we turned our attention to the topological properties of proteins, namely, the presence of knots in their folded state. In order to advance our understanding of the properties of knotted proteins, we performed a comparative analysis involving both sequence and structure alignments of protein pairs differing by topological state. Our investigation led to two main conclusions. First, the formation of a knot in a set of SOTCases appeared to be an evolutionarily rare, but functionally-oriented event; secondly, for many cases of well-superposable knotted-unknotted protein pairs the structural difference between the two partners merely consisted in the presence of small segments, whose excision from the knotted protein resulted in the unravelling of the knot. These results could be used in further computational/experimental studies to design new means of probing the salient steps that lead to knot formation during the folding process.

# Appendix

# Appendix A

# Quasi-rigid domain decomposition

## Domain decomposition of HIV-1 PR based on the lowest energy mode

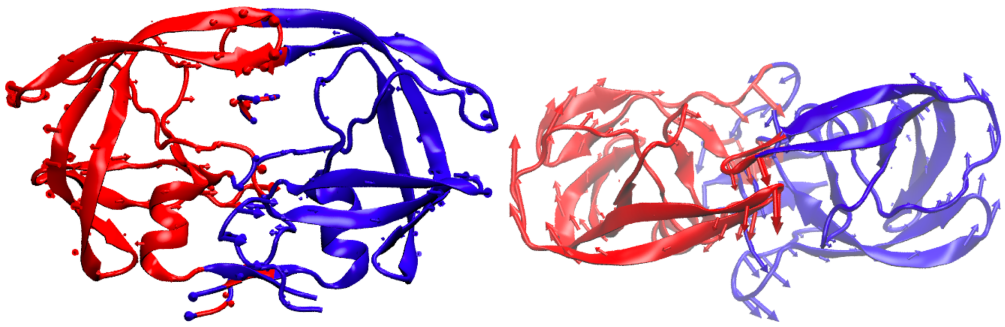

**Figure A.1: Decomposition of HIV-1 protease in 2 dynamical domains** - Only the first low-energy mode was used for the rigid-blocks decompositions ($n = 1$ in Eq. 3.14 of the paper). Left: front view; right: top view. The captured fraction of essential motion is 81.7%.

## Decomposition of HIV-1 PR in rigid subunits.

Chain A: 1-99
Chain B: 100-198
Peptide: 199-204

- Subdivision in $Q = 2$ blocks

Domain 1. 1-48, 53-107, 194-200

Domain 2. 49-52, 108-193, 201-204

- Subdivision in $Q = 3$ blocks

Domain 1. 1-10, 23-31, 48-53, 84-109, 122-130, 147-152, 183-204

Domain 2. 110-121, 131-146, 153-182

Domain 3. 11-22, 32-47, 54-83

- Subdivision in $Q = 4$ blocks

Domain 1. 47-54, 146-153, 199-204

Domain 2. 1-10, 23-31, 84-109, 122-130, 183-198

Domain 3. 110-121, 131-145, 154-182

Domain 4. 11-22, 32-46, 55-83

## Amino acids mobility and local structural deformations

In the following figures the mean square fluctuation of all amino acids of proteins 1ako, 1avp and 2ayh is reported against the degree of deformation of their local structural environment (again resulting from thermal fluctuations).

The degree of local structural deformation, hereafter also termed geometric strain, for the $i$-th residue is defined, in analogy to ref. (15) as:

$$s_i = \sum_j f(|\vec{d}^0_{ij}|)\langle(\vec{d}_{ij} - \vec{d}^0_{ij})^2\rangle \tag{A.1}$$

where $\vec{d}_{ij}$ is the distance vector of the $C_\alpha$ atoms of amino acids $i$ and $j$; a superscript 0 is used to denote quantities calculated for the average reference structure. $f(d) = \frac{1}{2}(1 - \tanh(d - d_{cut}))$ is a sigmoidal function weighting the average spatial proximity of the two amino acids. Its point of inflection is set at the cutoff distance of $d_{cut} = 7.5$ Å; the brackets indicate the canonical average. The mean square fluctuations and the geometrical strain plotted in the subsequent figures are expressed in the units of the Beta-Gaussian elastic network model (37).

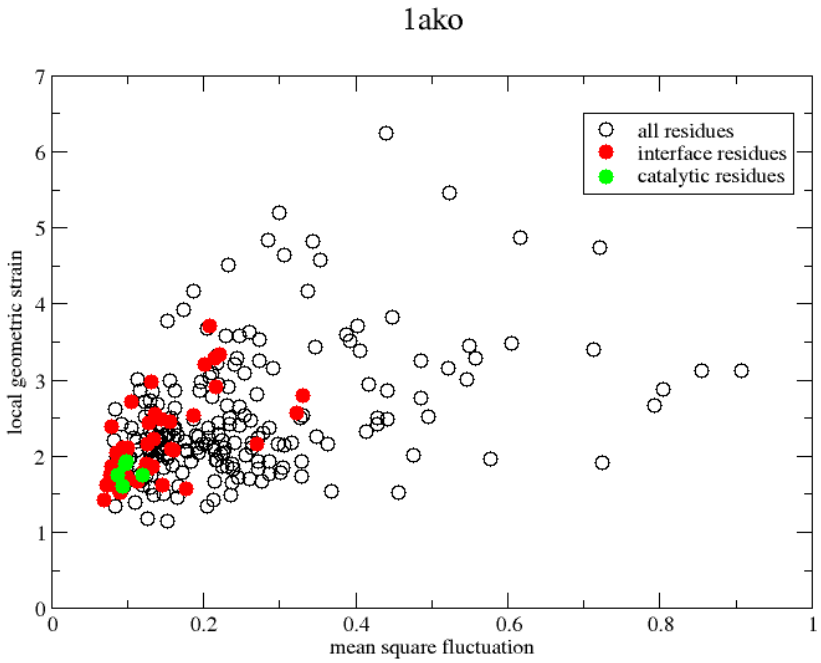


**Figure A.2: Strain vs. MSF - 1ako** - Scatter plot of the local geometric strain versus the mean square fluctuations for all amino acids of protein 1ako. Residues at the primary dynamical boundary are shown in red while catalytic ones are shown in green.

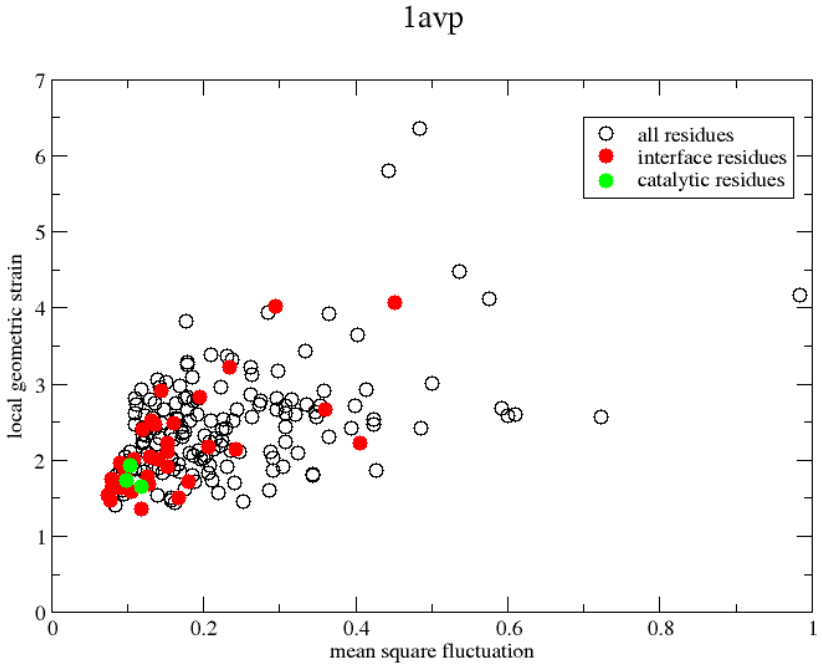


**Figure A.3: Strain vs. MSF - 1avp** - Scatter plot of the local geometric strain versus the mean square fluctuations for all amino acids of protein 1avp. Residues at the primary dynamical boundary are shown in red while catalytic ones are shown in green.

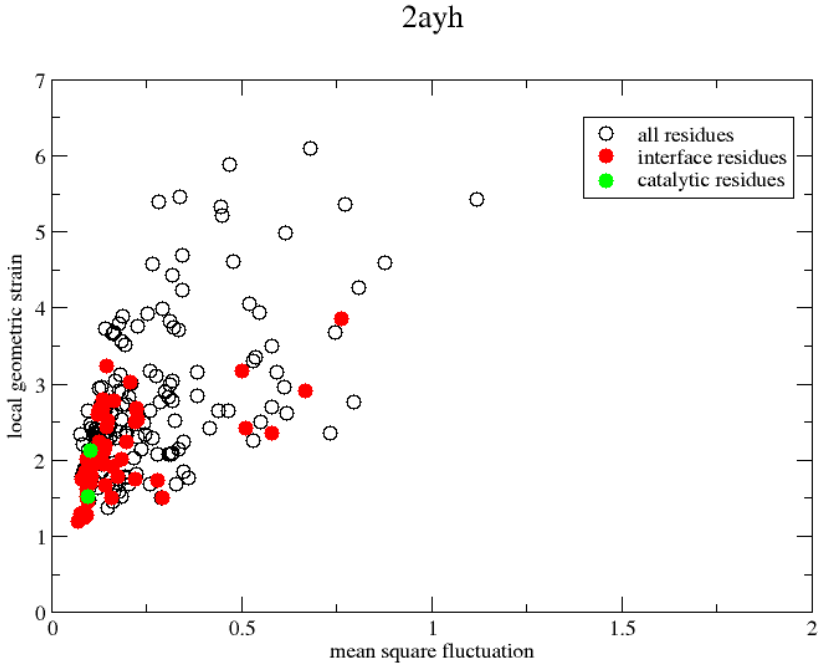


**Figure A.4: Strain vs. MSF - 2ayh** - Scatter plot of the local geometric strain versus the mean square fluctuations for all amino acids of protein 2ayh. Residues at the primary dynamical boundary are shown in red while catalytic ones are shown in green.

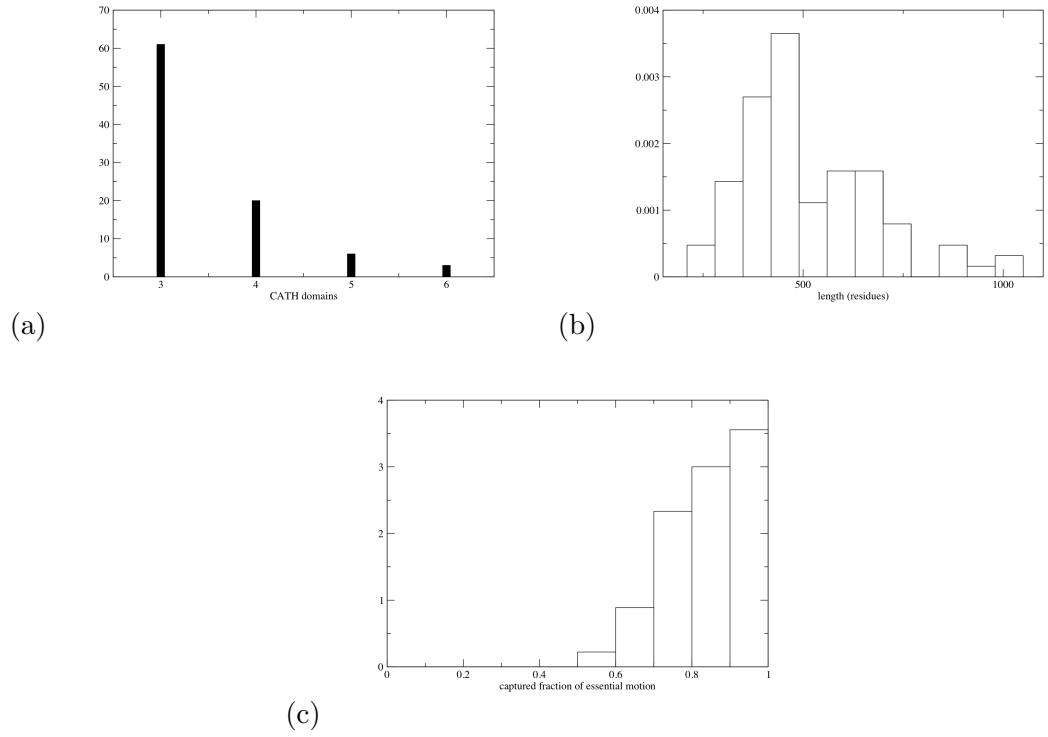


**Figure A.5: Histograms from the CATH study** - Histogram of (a) the number of CATH domains, (b) protein lengths and (c) captured fraction of essential motion, for entries in table A.1. A large part of the dataset (almost 60%) is populated by proteins composed by three CATH domains. Since the average number of residues per protein is about 500, this results in a small size of CATH domains. This fact reverberates on the fraction of essential motion captured by the rigid decomposition, which is higher than 70% for the largest fraction of the dataset.

## A. QUASI-RIGID DOMAIN DECOMPOSITION

| PDB | L | D | F | PDB | L | D | F | PDB | L | D | F |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1mw7A | 220 | 3 | 0.9154 | 2f8xC | 424 | 3 | 0.8538 | 1l7rA | 573 | 3 | 0.6963 |
| 1ep3B | 261 | 3 | 0.7157 | 1h3eA | 427 | 3 | 0.9751 | 1j2bB | 576 | 4 | 0.8949 |
| 1vlfX | 274 | 3 | 0.9428 | 2banB | 427 | 4 | 0.9103 | 1k7yA | 577 | 4 | 0.8125 |
| 1f02I | 282 | 3 | 0.9681 | 1cvrA | 432 | 3 | 0.7949 | 1fuiA | 591 | 3 | 0.7737 |
| 2dln0 | 306 | 3 | 0.7554 | 1f20A | 435 | 3 | 0.8790 | 2bzdA | 601 | 3 | 0.9079 |
| 1ixsB | 315 | 3 | 0.8391 | 8ohm0 | 435 | 3 | 0.9669 | 1f7vA | 606 | 3 | 0.7214 |
| 2pia0 | 321 | 3 | 0.8037 | 1s3sD | 436 | 4 | 0.9315 | 2j6hA | 608 | 3 | 0.9089 |
| 1krhA | 337 | 3 | 0.8455 | 2b7cA | 437 | 3 | 0.8527 | 1bhgB | 611 | 3 | 0.6609 |
| 1jr3D | 338 | 3 | 0.9849 | 1tubA | 440 | 3 | 0.5205 | 1i7dA | 620 | 4 | 0.9340 |
| 1obaA | 338 | 3 | 0.8304 | 1lwjB | 441 | 3 | 0.7566 | 1uh4A | 637 | 3 | 0.7454 |
| 1e4eB | 340 | 3 | 0.6766 | 1heiA | 443 | 3 | 0.9393 | 1t2xA | 639 | 3 | 0.8319 |
| 2btvT | 349 | 3 | 0.9741 | 1opkA | 449 | 4 | 0.8929 | 1kwkA | 644 | 3 | 0.7309 |
| 1oxxK | 352 | 3 | 0.9442 | 1xzqA | 449 | 3 | 0.9138 | 1ps9A | 671 | 3 | 0.8432 |
| 1b6sB | 355 | 3 | 0.7814 | 1w25B | 454 | 3 | 0.7207 | 4sli0 | 679 | 3 | 0.8936 |
| 1g292 | 372 | 3 | 0.9487 | 1nj6A | 463 | 3 | 0.7278 | 1ciu0 | 683 | 4 | 0.8065 |
| 1hwiC | 374 | 3 | 0.9318 | 1kmhB | 467 | 3 | 0.7376 | 9cgtA | 684 | 4 | 0.7847 |
| 2scuE | 385 | 3 | 0.9935 | 1w0kD | 467 | 3 | 0.7374 | 1qhpA | 686 | 4 | 0.8043 |
| 2fx3A | 387 | 3 | 0.8335 | 2cv2A | 468 | 5 | 0.9712 | 1v3mA | 686 | 4 | 0.7950 |
| 2fpgB | 393 | 3 | 0.9944 | 1gqyB | 469 | 3 | 0.8809 | 2dij0 | 686 | 4 | 0.8014 |
| 1okeB | 394 | 4 | 0.9622 | 1skyE | 470 | 3 | 0.7432 | 1gqkB | 708 | 3 | 0.7684 |
| 1svb0 | 395 | 4 | 0.9712 | 2hgsA | 472 | 5 | 0.8355 | 1ru3A | 728 | 5 | 0.9902 |
| 1gvhA | 396 | 3 | 0.8128 | 1h4sB | 473 | 3 | 0.6571 | 1ordB | 730 | 4 | 0.8907 |
| 1hfeM | 396 | 3 | 0.6684 | 1mdfA | 536 | 4 | 0.8038 | 1bf20 | 750 | 3 | 0.7429 |
| 1d2eC | 397 | 3 | 0.7960 | 1x6nA | 539 | 3 | 0.9159 | 1w0pA | 753 | 3 | 0.9103 |
| 2c78A | 397 | 3 | 0.8872 | 1gn9B | 543 | 4 | 0.6812 | 1qbb0 | 858 | 4 | 0.7816 |
| 1dljA | 402 | 3 | 0.8125 | 1kzhA | 550 | 3 | 0.8074 | 1kc7A | 872 | 6 | 0.9582 |
| 1cqxB | 403 | 3 | 0.7244 | 1e1dA | 553 | 4 | 0.6394 | 1vbgA | 874 | 6 | 0.9895 |
| 1psdA | 404 | 3 | 0.9417 | 1uok0 | 558 | 3 | 0.5862 | 1xc6A | 971 | 5 | 0.8801 |
| 2dcuA | 407 | 3 | 0.6887 | 2hmiA | 558 | 5 | 0.9837 | 2f7pA | 1014 | 5 | 0.8050 |
| 1sqgA | 424 | 4 | 0.9361 | 2ex3C | 570 | 6 | 0.9083 | 1ulvA | 1019 | 4 | 0.9833 |

**Table A.1: Dataset of proteins used for CATH domain study** - Data set of protein monomers with overall sequence identity below 90% used used in the comparison of dynamical subdomains and CATH domains. Each CATH domain occupies an uninterrupted sequence interval. Entries are sorted by increasing length. The entries in the column correspond to: PDB = protein data bank accession code; L = length (residues); D = number of CATH domains; F = captured fraction of essential motion.

| Arch. | D |
|---|---|
| 3.40 | 71 |
| 2.60 | 53 |
| 3.30 | 52 |
| 1.10 | 35 |
| 2.40 | 25 |

**Table A.2: List of the most populated CATH domain architectures** - List of the most populated CATH domain architectures (Arch.) in the dataset and the corresponding number of domains (D).

| PDB id | CATH id |
|---|---|
| 1bhgB | 2.60.40.320 |
| 1e1dA | 1.20.1270.20 |
| 1f20A | 2.40.30.10 |
| 1f7vA | 1.10.730.10 |
| 1g292 | 2.40.50.140 |
| 1gn9B | 1.20.1270.20 |
| 1j2bB | 3.90.1020.10 |
| 1kc7A | 1.20.80.30 |
| 1kwkA | 2.60.40.1180 |
| 1mdfA | 2.30.38.10 |
| 1ordB | 3.90.1150.10 |
| 1oxxK | 2.40.50.140 |
| 1vbgA | 1.20.80.30 |
| 2b7cA | 2.40.30.10 |
| 2dcuA | 2.40.30.10 |
| 2ex3C | 4.10.80.20 |
| 2hgsA | 3.30.470.20 |

**Table A.3: Details of the CATH domains** - Details of the CATH domains (and parent proteins) for which the lowest overlaps with dynamical subdivisions are observed.

# References


[1] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–826, April 1986. xiii

[2] C. Chothia, J. Gough, C. Vogel, and S.A. Teichmann. Evolution of the protein repertoire. *Science*, 300:1701–3, 2003. xiii

[3] C A Orengo and J M Thornton. Protein families and their evolution-a structural perspective. *Annu Rev Biochem*, 74:867–900, 2005. xiii

[4] L Holm and C Sander. The fssp database of structurally aligned protein fold families. *Nucleic Acids Res*, 22(17):3600–3609, Sep 1994. xiii, 65

[5] A G Murzin, S E Brenner, T Hubbard, and C Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995. xiii, 65, 81

[6] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8):1093–1108, August 1997. xiii, 40, 56, 65, 81, 88, 91

[7] A Andreeva and A G Murzin. Evolution of protein fold in the presence of functional constraints. *Curr Opin Struct Biol*, 16(3):399–408, 2006. xiii

[8] J.R. Banavar, A. Maritan, Micheletti, C., and A. Trovato. Geometry and physics of proteins. *Proteins*, 47:315–22, 2002. xiii

[9] L Chen, A L DeVries, and C H Cheng. Convergent evolution of antifreeze glycoproteins in antarctic notothenioid fish and arctic cod. *Proc Natl Acad Sci U S A*, 94(8):3817–3822, 1997. xiii

[10] M Denton and C Marshall. Protein folds: laws of form revisited. *Nature*, 410(6827):417–417, 2001. xiii

[11] S S Krishna and N V Grishin. Structurally analogous proteins do exist! *Structure*, 12(7):1125–1127, 2004. xiii

[12] F Seno and A Trovato. Minireview: The compact phase in polymers and proteins. *Physica A: Statistical Mechanics and its Applications*, 384(1):122 – 127, 2007. xiii

[13] K A Henzler-Wildman, V Thai, M Lei, M Ott, M Wolf-Watz, T Fenn, E Pozharski, M A Wilson, G A Petsko, M Karplus, C G Hübner, and D Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007. xiii, xiv, 16, 66

[14] K A Henzler-Wildman, M Lei, V Thai, S J Kerns, M Karplus, and D Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 2007. xiii, xiv, 29

[15] 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(12):5901–5912, Dec 2008. xiii, xiv, xvii, 5, 13, 14, 16, 17, 23, 26, 40, 43, 44, 66, 113

[16] X Li, O Keskin, B Ma, R Nussinov, and J Liang. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol*, 344(3):781–795, 2004. xiii

[17] O N Yogurtcu, S B Erdemli, R Nussinov, M Turkay, and O Keskin. Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. *Biophys J*, 94(9):3475–3485, 2008. xiii

[18] A Zen, C Micheletti, O Keskin, and R Nussinov. Comparing interfacial dynamics in protein-protein complexes: an elastic network approach. *BMC Struct Biol*, 10:26–26, 2010. xiii

[19] D Kern, E Z Eisenmesser, and M Wolf-Watz. Enzyme dynamics during catalysis measured by nmr spectroscopy. *Methods Enzymol*, 394:507–524, 2005. xvi

[20] H F Lou and R I Cukier. Molecular dynamics of apoadenylate kinase: a principal component analysis. *J. Phys. Chem. B*, 110(25):12796–12808, Jun 2006. xvii

[21] K Arora and C L Brooks. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc Natl Acad Sci U S A*, 104(47):18496–18501, Nov 2007. xvii, 29

[22] A.E. Garcia. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68:2696–2699, 1992. 4, 13, 29

[23] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993. 4, 5, 13, 14

[24] N Go, T Noguti, and T Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A*, 80(12):3696–3700, 1983. 5, 13

[25] R Potestio, F Pontiggia, V Carnevale, and C Micheletti. Bridging the atomic and coarse-grained descriptions of collective motions in proteins, *invited contribution for the book* Multiscale approaches to protein modelling: structure prediction, dynamics, thermodynamics and macromolecular assemblies*, edited by A. Kolinski.* Springer, in press. 5

[26] B. Brooks, D. Janezic, and M. Karplus. Harmonic analysis of large systems. iii. comparison with molecular dynamics. *Journal of Computational Chemistry*, 1995. 6

# REFERENCES


[27] M. M. Tirion. Large amplitude elastic motions in proteins from a single–parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996. 8, 12, 17, 29

[28] A R Atilgan, S R Durell, R L Jernigan, M C Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*, 80(1):505–515, 2001. 9

[29] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2(3):173–181, 1997. 9

[30] M Delarue and Y H Sanejouand. Simplified normal mode analysis of conformational transitions in dna-dependent polymerases: the elastic network model. *J Mol Biol*, 320(5):1011–1024, 2002. 9

[31] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998. 9, 30, 31

[32] Cristian Micheletti, Jayanth R. Banavar, and Amos Maritan. Conformations of proteins in equilibrium. *Phys. Rev. Lett.*, 87(8):088102, Aug 2001. 9

[33] C Micheletti, G Lattanzi, and A Maritan. Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. *J Mol Biol*, 321(5):909–921, 2002. 9

[34] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235(4786):318–321, 1987. 9

[35] H Frauenfelder, F Parak, and R D Young. Conformational substates in proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):451–479, 1988. 9

[36] P Doruker, A R Atilgan, and I Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, 40(3):512–524, 2000. 9

[37] C Micheletti, P Carloni, and A Maritan. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins*, 55(3):635–645, May 2004. 11, 12, 47, 62, 67, 113

[38] B Park and M Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–392, 1996. 12

[39] V Carnevale, S Raugei, C Micheletti, and P Carloni. Large-scale motions and electrostatic properties of furin and HIV-1 protease. *J Phys Chem A*, 111:12327–12332, 2007. 12, 75

[40] M Cascella, C Micheletti, U Rothlisberger, and P Carloni. Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J Am Chem Soc*, 127(11):3734–3742, Mar 2005. 12, 51, 75

[41] Berk Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62(6):8438–8448, Dec 2000. 13, 14

[42] Daniel Ben-Avraham. Vibrational normal-mode spectrum of globular proteins. *Physical Review B*, 47(21):14559+, June 1993. 15

[43] F Pontiggia, G Colombo, C Micheletti, and H Orland. Anharmonicity and self-similarity of the free energy landscape of protein g. *Phys. Rev. Lett.*, 98(4):048102–048102, Jan 2007. 16

[44] E P Wigner. *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press, 1959. 18, 20

[45] Markus Müller, Yurytzy López Jiménez, Christian Rummel, Gerold Baier, Andreas Galka, Ulrich Stephani, and Hiltrud Muhle. Localized short-range correlations in the spectrum of the equal-time correlation matrix. *Phys. Rev. E*, 74(4):041119, Oct 2006. 19

[46] T A Brody. A statistical measure for the repulsion of energy levels. *Lett. Nuov. Cim.*, 7:482, 1973. 20

[47] M L Mehta. *Random Matrices*. Academic Press, third edition, 2004. 20

[48] J Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20:32, 1928. 20

[49] L. Laloux, P Cizeau, J P Bouchaud, and M Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467, 1999. 20

[50] M Barthélemy, B Gondran, and E Guichard. Large scale cross-correlations in internet traffic. *Phys. Rev. E*, 66:056110, 2002. 20

[51] P Šeba. Random matrix analysis of human eeg data. *Phys. Rev. Lett.*, 91:198104, 2003. 20

[52] M S Santhanam and P K Patra. Statistics of atmospheric correlations. *Phys. Rev. E*, 64:016102, 2001. 20

[53] T Guhr, A Mueller-Groeling, and HA Weidenmueller. Random matrix theories in quantum physics: Common concepts. *arXiv:cond-mat/9707301v1*, 1997. 21

[54] H. Frauenfelder, H.A. Sligar, and P.G. Wolynes. The energy landscape and motions of proteins. *Science*, 254:1598, 1991. 29

[55] J J Falke. Enzymology. a moving story. *Science*, 295(5559):1480–1481, Feb 2002. 29

[56] S Piana, P Carloni, and M Parrinello. Role of conformational fluctuations in the enzymatic reaction of hiv-1 protease. *J Mol Biol*, 319(2):567–583, May 2002. 29, 45, 47, 49, 50

[57] M Garcia-Viloca, J Gao, M Karplus, and D G Truhlar. How enzymes work: analysis by modern rate theory and computer simulations. *Science*, 303(5655):186–195, Jan 2004. 29

[58] A. L. Perryman, J-H. Lin, and J. A. McCammon. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Prot. Sci.*, 13:1108–1123, 2004. 29

[59] M Wolf-Watz, V Thai, K Henzler-Wildman, G Hadjipavlou, E Z Eisenmesser, and D Kern. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol*, 11(10):945–949, Oct 2004. 29

[60] J A Hanson, K Duderstadt, L P Watkins, S Bhattacharyya, J Brokaw, J W Chu, and H Yang. Illuminating the mechanistic roles of enzyme conformational dynamics. *Proc Natl Acad Sci U S A*, 104(46):18055–18060, Nov 2007. 29

[61] A. del Sol, M. J. Arauzo-Bravo, D. Amoros, and R. Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory likages. *Genome Biology*, 8, 2007. 29, 51

[62] M J Bradley, P T Chivers, and N A Baker. Molecular dynamics simulation of the escherichia coli nikr protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J Mol Biol*, 378(5):1155–1173, May 2008. 29

[63] K. Hinsen, A. Thomas, and M. J. Field. Analysis of domain motion in large proteins. *Proteins*, 34:369–382, 1999. 30, 31

[64] S Hayward, A Kitao, and H J C Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins: Structure, Function, and Genetics*, 27:425–437, 1997. 30, 32, 43

[65] S. Kundu, D. C. Sorensen, , and Jr. G. N. Phillips. Automatic domain decomposition of proteins by a gaussian network model. *Proteins*, 57:725–733, 2004. 30, 32

[66] S. O. Yesylevskyy, V. N. Kharkyanen, and A. P. Demchenko. Hierarchical clustering of correlation patterns: new method of domain identification in proteins. *Biophysical Chemistry*, 119:84–93, 2006. 30, 32

[67] H. Golhlke and M. F. Thorpe. A natural coarse graining for simulating large biomolecular motion. *Biophysical Journal*, 91:2115–2120, 2006. 30

[68] V. Schomaker and K. N. Trueblood. On rigid-body motion of molecules in crystals. *Acta Cryst.*, B 24:63–76, 1968. 33

[69] C. Chaudhry, A. L. Horwich, A. T. Brunger, and P. D. Adams. Exploring the structural dynamics of the e.coli chaperonin groel using translation-libration-screw crystallographic refinement of intermediate states. *Journal of Molecular Biology*, 342(1):229–245, 2004. 34

[70] J. Painter and E. A. Merritt. Optimal description of a protein structure in terms of multiple groups undergoing tls motion. *Acta Cryst.*, D 62:439–450, 2006. 34, 43, 56

[71] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. 35

[72] E. C. Webb. *Enzyme nomenclature. Reccomendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.* Academic Press, New York, 1992. 40, 51, 88

[73] P Maragakis and M Karplus. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J Mol Biol*, 352(4):807–822, Sep 2005. 43

[74] S Piana, P Carloni, and U Rothlisberger. Drug resistance in hiv-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci*, 11(10):2393–2402, 2002. 47

[75] J. D. Tyndall, T. Nall, and D. P. Fairlie. Proteases universally recognize beta strands in their active sites. *Chem. Rev.*, 105:973–999, 2005. 49

[76] V Carnevale, S Raugei, C Micheletti, and P Carloni. Convergent dynamics in the protease enzymatic superfamily. *J Am Chem Soc*, 128(30):9766–9772, Aug 2006. 51, 65, 67

[77] 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(5):918–929, May 2008. 51, 52, 54, 65, 66, 67, 68, 72, 76, 78

[78] "The UniProt Consortium". The universal protein resource (uniprot). *Nucl. Acids Res.*, 36(suppl1):D190–195, 2008. 51

[79] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Acids Res.*, 32:D129–D133, 2004. 52, 76

[80] C.D. Mol, C-F. Kuo, M.M. Thayer, R.P. Cunningham, and J.A. Tainer. Structure and function of the multifunctional dna-repair enzyme exonuclease iii. *Nature*, 374:381–386, 1995. 53, 55

[81] G.R. Stockwell and J.M. Thornton. Conformational diversity of ligands bound to proteins. *J Mol Biol*, 356:928–944, 2006. 53

[82] A. Roujeinikova, S. Sedelnikova, G. de Boer, A.R. Stuitje, A.R. Slabasi, J.B. Rafferty, and D.W. Rice. Inhibitor binding studies on enoyl reductase reveal conformational changes related to substrate recognition. *J. Biol. Chem.*, 274:30811–30817, 1999. 53

[83] L.S. Pidugu, M. Kapoor, N. Surolia, A. Surolia, and K. Saguna. Structural basis for the variation in triclosan affinity to enoyl reductases. *J Mol Biol*, 343:147–155, 2004. 53, 55

[84] S. Gupta, W.F. Mangel, W.J. McGrath, J.L. Perek, D.W. Lee, K. Takamoto, and M.R. Chance. Dna binding provides a molecular strap activating the adenovirus proteinase. *Mol. & Cell. Proteom.*, 3:950–959, 2004. 53, 55

[85] O C Redfern, A Harrison, T Dallman, F M Pearl, and C A Orengo. Cathedral: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol*, 3(11), Nov 2007. 57

# REFERENCES


[86] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000. 62

[87] F Capozzi, C Luchinat, C Micheletti, and F Pontiggia. Essential dynamics of helices provide a functional classification of EF-hand proteins. *J Proteome Res*, 6:4245–4255, 2007. 65

[88] M. Tesi, E. Rensburg, E. Orlandini, and S. Whittington. Monte carlo study of the interacting self-avoiding walk model in three dimensions. *Journal of Statistical Physics*, 82:155–181, 1996. 69

[89] C Micheletti and H Orland. MISTRAL: a tool for energy-based multiple structural alignment of proteins. *Bioinformatics*, 25:2663–2669, 2009. 71, 92

[90] M. Shatsky, R. Nussinov, and H.J. Wolfson HJ. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56:143–156, 2004. 71

[91] R Potestio, T Aleksiev, F Pontiggia, S Cozzini, and C Micheletti. Aladyn: a web server for aligning proteins by matching their large-scale motion. *Nucleic Acids Res*, 38 Suppl:41–45, 2010. 72

[92] M L Sierk and W R Pearson. Sensitivity and selectivity in protein structure comparison. *Protein Sci*, 13:773–785, 2004. 71

[93] T.L. Blundell and N. Srinivasan. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. USA*, 93:14243–14248, 1996. 75

[94] C D Mol, C F Kuo, M M Thayer, R P Cunningham, and J A Tainer. Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature*, 374:381–386, 1995. 75, 76

[95] S Gupta, W F Mangel, W J McGrath, J L Perek, D W Lee, K Takamoto, and M R Chance. DNA binding provides a molecular strap activating the adenovirus proteinase. *Mol Cell Proteomics*, 3:950–959, 2004. 75, 76

[96] L Holm, S Kääriäinen, P Rosenström, and A Schenkel. Searching protein structure databases with DaliLite v.3. *Bioinformatics*, 24:2780–2781, 2008. 75

[97] J D Tyndall, T Nall, and D P Fairlie. Proteases universally recognize beta strands in their active sites. *Chem Rev*, 105:973–999, 2005. 75

[98] R Potestio, F Pontiggia, and C Micheletti. Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys J*, 96(12):4993–5002, Jun 2009. 76

[99] Jmol: an open-source java viewer for chemical structures in 3d. http://www.jmol.org/, 2010. 78

[100] R I Sadreyev, B H Kim, and N V Grishin. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol*, 19(3):321–328, Jun 2009. 81

[101] R Kolodny, D Petrey, and B Honig. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr Opin Struct Biol*, 16(3):393–398, Jun 2006. 81

[102] P Virnau, L A Mirny, and M Kardar. Intricate knots in proteins: Function and evolution. *PLoS Comput Biol*, 2(9), Sep 2006. 82, 92, 94, 95

[103] W R Taylor. Protein knots and fold complexity: some new twists. *Comput Biol Chem*, 31(3):151–162, Jun 2007. 82, 90

[104] F Khatib, M T Weirauch, and C A Rohl. Rapid knot detection and application to protein structure prediction. *Bioinformatics*, 22(14):252–259, Jul 2006. 82

[105] A Tramontano, R Leplae, and V Morea. Analysis and assessment of comparative modeling predictions in casp4. *Proteins*, Suppl 5:22–38, 2001. 82

[106] D Bölinger, J I Sulkowska, H P Hsu, L A Mirny, M Kardar, J N Onuchic, and P Virnau. A Stevedore's protein knot. *PLoS Comput Biol*, 6(4), 2010. 82, 83, 90

[107] C Micheletti, D Marenduzzo, E Orlandini, and D W Sumners. Knotting of random ring polymers in confined spaces. *Journal of Chemical Physics*, 124(6):64903–64903, Feb 2006. 83, 86, 90

[108] C Micheletti, D Marenduzzo, E Orlandini, and D W Sumners. Simulations of knotting in confined circular DNA. *Biophys J*, 95(8):3591–3599, Oct 2008. 83, 86, 90

[109] D Marenduzzo, E Orlandini, A Stasiak, D W Sumners, L Tubiana, and C Micheletti. DNA-DNA interactions in bacteriophage capsids are responsible for the observed DNA knotting. *Proc Natl Acad Sci U S A*, 106(52):22269–22274, Dec 2009. 83

[110] D Marenduzzo, C Micheletti, and E Orlandini. Biopolymer organization upon confinement. *J. Phys.: Condens. Matter*, 22(28):283102, 2010. 83

[111] V V Rybenkov, N R Cozzarelli, and A V Vologodskii. Probability of DNA knotting and the effective diameter of the DNA double helix. *Proc Natl Acad Sci U S A*, 90(11):5307–5311, Jun 1993. 83

[112] Z Liu, L Zechiedrich, and H S Chan. Local site preference rationalizes disentangling by DNA topoisomerases. *Phys Rev E*, 81(3 Pt 1):031902–031902, Mar 2010. 83

[113] M. Delbruck. Knotting problems in biology. *Mathematical Problems in the Biological Sciences, edited by R. E. Bellman, Proceedings of Symposia in Applied Mathematics,American Mathematical Society, Providence, Rhode Island*, 14:55–63, 1962. 83

[114] H. L. Frisch and E. Wasserman. Chemical topology. *Journal of American Chemical Society*, 83:3789–3795, 1961. 83

[115] A. V. Vologodskii, A. V. Lukashin, M. D. Frank-Kamenetskii, and V. V. Anshelevich. The knot problem in statistical mechanics of polymer chains. *Sov. Phys.-JETP*, 39:1059–1063, 1974. 83

[116] D. W. Sumners and S. G. Whittington. Knots in self-avoiding walks. *Journal of Physics A-Mathematical and General*, 21:1689–1694, 1988. 83

[117] E. J. Janse van Rensburg, D. W. Sumners, E. Wasserman, and S. G. Whittington. Entanglement complexity of self-avoiding walks. *Journal of Physics A-Mathematical and General*, 25:6557–6566, 1992. 83

[118] C Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–5, 1968. 83

[119] J I Sulkowska, P Sulkowski, and J Onuchic. Dodging the crisis of folding proteins with knots. *Proc Natl Acad Sci U S A*, 106(9):3119–3124, Mar 2009. 83

[120] S E Jackson and A R Fersht. Folding of chymotrypsin inhibitor 2. 1. evidence for a two-state transition. *Biochemistry*, 30(43):10428–10435, Oct 1991. 83

[121] A L Mallam, S C Onuoha, J G Grossmann, and S E Jackson. Knotted fusion proteins reveal unexpected possibilities in protein folding. *Mol Cell*, 30(5):642–648, Jun 2008. 84

[122] A L Mallam. How does a knotted protein fold? *FEBS J*, 276(2):365–375, Jan 2009. 84

[123] AL Mallam, JM Rogers, and SE Jackson. Experimental detection of knotted conformations in denatured proteins. *Proc Natl Acad Sci U S A*, 107(18):8189–8194, 2010. 84

[124] K Millett, A Dobay, and A Stasiak. Linear Random Knots and Their Scaling Behavior. *Macromolecules*, 38(2):601–606, 2005. 84

[125] F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. 84

[126] E Orlandini and S Whittington. Statistical topology of closed curves: Some applications in polymer physics. *Rev. Mod. Phys.*, 79(2):611–642, Apr-Jun 2007. 85

[127] F Khatib, MT Weirauch, and CA Rohl. Rapid knot detection and application to protein structure prediction. *Bioinformatics*, 22(14):252–259, 2006. 86

[128] S Mika and B Rost. Uniqueprot: Creating representative protein sequence sets. *Nucleic Acids Res*, 31(13):3789–3791, Jul 2003. 86

[129] G Kolesov, P Virnau, M Kardar, and L A Mirny. Protein knot server: detection of knots in protein structures. *Nucleic Acids Res*, 35(Web Server issue):425–428, Jul 2007. 88, 94

[130] B Marcone, E Orlandini, A L Stella, and F Zonta. Size of knots in ring polymers. *Phys Rev E*, 75(4 Pt 1):041105–041105, Apr 2007. 88

[131] C M Spahn, M G Gomez-Lorenzo, R A Grassucci, R Jorgensen, G R Andersen, R Beckmann, P A Penczek, J P Ballesta, and J Frank. Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J*, 23(5):1008–1019, Mar 2004. 89

[132] A E Eriksson, T A Jones, and A Liljas. Refined structure of human carbonic anhydrase II at 2.0 A resolution. *Proteins*, 4(4):274–282, 1988. 90

[133] V Biou, R Dumas, C Cohen-Addad, R Douce, D Job, and E Pebay-Peyroula. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 A resolution. *EMBO J*, 16(12):3405–3415, Jun 1997. 90

[134] S Misaghi, P J Galardy, W J Meester, H Ovaa, H L Ploegh, and R Gaudet. Structure of the ubiquitin hydrolase UCH-L3 complexed with a suicide substrate. *J Biol Chem*, 280(2):1512–1520, Jan 2005. 90

[135] J W Schmidberger, J A Wilce, A J Weightman, J C Whisstock, and M C Wilce. The crystal structure of DehI reveals a new alpha-haloacid dehalogenase fold and active-site mechanism. *J Mol Biol*, 378(1):284–294, Apr 2008. 90

[136] JPJ Michels and FW Wiegel. On the topology of a polymer ring. *Proc. R. Soc. Lond.*, A403:269–284, 1986. 90

[137] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990. 91

[138] D Shi, H Morizono, J Cabrera-Luque, X Yu, L Roth, M H Malamy, N M Allewell, and M Tuchman. Structure and catalytic mechanism of a novel N-succinyl-L-ornithine transcarbamylase in arginine biosynthesis of Bacteroides fragilis. *J Biol Chem*, 281(29):20623–20631, Jul 2006. 91

[139] J D Thompson, D G Higgins, and T J Gibson. 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(22):4673–4680, Nov 1994. 91

[140] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987. 91, 93

[141] J I Sulkowska, P Sulkowski, P Szymczak, and M Cieplak. Tightening of knots in proteins. *Phys Rev Lett*, 100(5):058106–058106, Feb 2008. 92

[142] Joanna I. Sulkowska, Piotr Sulkowski, P. Szymczak, and Marek Cieplak. Stabilizing effect of knots on proteins. *Proc Natl Acad Sci U S A*, 105(50):19714–19719, 2008. 92

[143] L Xie and P E Bourne. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A*, 105:5441–5446, 2008. 94

[144] D Saadat and D H Harrison. The crystal structure of methylglyoxal synthase from Escherichia coli. *Structure*, 7(3):309–317, Mar 1999. 97

[145] J Kyte and R F Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982. 97

[146] S Wallin, K B Zeldovich, and E I Shakhnovich. The folding mechanics of a knotted protein. *J Mol Biol*, 368(3):884–893, May 2007. 99

[147] X Yang, J Kuk, and K Moffat. Crystal structure of Pseudomonas aeruginosa bacteriophytochrome: photoconversion and signal transduction. *Proc Natl Acad Sci U S A*, 105(38):14715–14720, Sep 2008. 99

[148] V M Levdikov, E Blagova, P Joseph, A L Sonenshein, and A J Wilkinson. The structure of CodY, a GTP- and isoleucine-responsive regulator of stationary phase and virulence in gram-positive bacteria. *J Biol Chem*, 281(16):11366–11373, Apr 2006. 99

[149] C Das, Q Q Hoang, C A Kreinbring, S J Luchansky, R K Meray, S S Ray, P T Lansbury, D Ringe, and G A Petsko. Structural basis for conformational plasticity of the Parkinson's disease-associated ubiquitin UCH-L1. *Proc Natl Acad Sci U S A*, 103(12):4675–4680, Mar 2006. 104

[150] K I Varughese, Y Su, D Cromwell, S Hasnain, and N H Xuong. Crystal structure of an actinidin-e-64 complex. *Biochemistry*, 31(22):5172–5176, Jun 1992. 104

[151] J M Mato, F J Corrales, S C Lu, and M A Avila. S-adenosylmethionine: a control switch that regulates liver function. *FASEB J*, 16(1):15–26, Jan 2002. 104

[152] I I Mathews, M D Erion, and S E Ealick. Structure of human adenosine kinase at 1.5 A resolution. *Biochemistry*, 37(45):15607–15620, Nov 1998. 104

# Acknowledgements

The object of this thesis is the work that I carried out during the last four years in SISSA. The object of this page, instead, are those people who made it possible.

The first person that I have to thank is Cristian. It is not rhetoric to say that I am really grateful to him for his teaching, support, mentorship and friendship.

My deepest gratitude goes to all those people who had the patience to work with me; in particular, Francesco Pontiggia deserves a special mention for being so supportive during my first months of real work, and later on.

Since the very beginning of this story I've been blessed by the friendship of a number of marvellous persons. Carmelo, my beloved flatmate; Daniele, Elena, Fabio and Serena, the *fab four* of Statistical Physics; Barbie, Donata, Eleonora, Giulia, Goff (zce'), Il Nardecchione, Lucalepori, Lucadiluzio, Marcolino, Miriam, Pierpaolo, Pratika, Schirom, Valentina: all of you I hug and take with me wherever I shall go.

I sincerely thank all the members of my Sector, in particular Andrea Zen and my officemate Luca; Riccardo, Federica and Barbara, top-ranking officiers of the Administrative army; and the bar staff, who provided the substantial amount of coffee that I need to stay alive.

Many persons who do not live here in Trieste have been nonetheless a constant and fundamental presence in the last four years: my friends Alessandro, Alessandra, Carmine, Claudia, Kristian, Mimmo, Monica; and my family, warm and caring soil of my still growing roots.

Finally, I thank Gloria, to whom this thesis is dedicated. I have so many reasons to thank you that I could spend the whole night writing them down. Therefore, in brief: I love you.