SISSA/ISAS - International School for Advanced Studies

Protein Structure and Functionally-oriented Dynamics: From Atomistic to Coarse-grained Models

Thesis submitted for the degree of Doctor Philosophiæ

Candidate
Francesco Pontiggia

Supervisor
Cristian Micheletti

27^{th} October 2008

Contents

Outline

The overwhelming majority of biological processes relies on the capability of proteins to sustain conformational changes so to selectively recognise, bind and process other molecules, being them proteins, nucleic acids or other chemical compounds.
The paradigmatic tripartite characterization of proteins in terms of sequence→structure→function has served to interpret the above-mentioned capability as being encoded in proteins' native structures, which is in turn determined by the amino acidic sequence [6].

Arguably, the earliest quantitative attempts to relate proteins' functionality to their structural flexibility have been prompted by the pioneering crystallographic studies on heme proteins. In fact, the analysis of the very first X-rays resolved structures showed [111, 45] that (i) myoglobing and hemoglobin can assume a number of different conformations (e.g. unliganded or bound to dioxygen) and that (ii) the apo conformers were too compact to possibly allow the diffusion of dioxygen towards the heme pocket, thus implying that the molecule had to open substantially to allow dioxygen to reach the binding site. Both observations indicated that proteins are endowed with an unsuspected degree of elasticity which permits the visiting of alternative conformers which are local minima of the free energy. It was further suggested by Frauenfelder that these minima are hierarchically organised and separated by free energy barriers of various depth (see e.g. ref [46] and references therein). This hierarchical free energy organization is expected to reflect in a multiplicity of time scales regulating transitions among the different biologically relevant states[56]. Most of biologically relevant processes, occurring on time scales of the order of μs to ms, are expectedly controlled by the activation time required to cross free energy barriers separating the various biological conformers. The conformational changes involved by such interconversions often have a collective scale, as suggested by the comparison of conformers of the same protein crystallised in different conditions, typically unbound or bound to a ligand [50]. Traditionally, such interconversions are believed to be triggered or induced, by binding[82]. Each of the "biological conformer" comprises in turn several substates which differ by sidechain orientations, local dihedral angles, and are hence separated by smaller free energy barriers.

Nowadays, processes occurring on timescales of the order of μs to ms can be captured by a number of experimental methods. The interconversion among the local substate, however, occurs on much smaller timescales, and is hence elusive for traditional experimental techniques. A valuable complement to experimental investigations in probing such small timescales ($\leq \mu$s) is represented by atomistic molecular dynamics (MD) simulations, which allow one to follow the detailed dynamical evolution of the systems over small time scales that are usually in the range 1-10 ns.

In the last few years, the advancements in computational techniques and resources, and the increased time resolution of advanced single-molecule techniques, have allowed to almost bridge gap between the different timescales probed by computation and experiment. By relating these two means of investigation it has been possible to achieve a multi-timescale characterization of proteins' conformational changes and their relevance for biological function (see e.g. ref [56] and references therein). For a growing number of proteins and enzymes, these approaches have clarified that the molecule's internal dynamics is "innately" predisposed to assist the interconversion between the various functional substates even in the absence of substrates or triggering events[56]. It has been further hypothesised that the innate functional-oriented character of proteins' dynamics has been evolutionarily promoted as binding by selection of already fitting conformers is more efficient that the induced-fit mechanism[89, 86, 17, 56, 1].

In this thesis we shall discuss several aspects relating proteins' internal flexibility to functionally-oriented conformational changes.

In the first part I shall present our investigations of the near native free energy landscape of two globular protein, the immunoglobulin binding domain of protein G and E.Coli adenylate kinase. The multiminima organization of the near-native free energy landscape is probed by means of extensive atomistic molecular dynamics simulations (covering hundreds of ns) and the results are interpreted in terms of simplified coarse-grained models. The notable aspect emerging for both system is that the generalized directions corresponding to the low-energy conformational deformations of the individual substates and of the virtual "jumps" that connect them are remarkably similar and reflect in an unsuspected robustness of the essential dynamical spaces as a function of the trajectory duration. The functionally-oriented character of these principal directions is finally discussed.

In the second part we investigate aspects regarding the connection between protein structure and internal dynamics. In particular, we first compare the essential dynamics of a large number of proteins that are members of the calcium-binding superfamily. It is found that the modes of tens of EF-hand domains can serve as basis for classifying the calcium-binding proteins into a limited number of groups having distinct internal dynamics. Notably, this dynamics-based grouping is found to gather proteins belonging to the same functional family. A generalised scheme for the comparison of essen-

tial modes is finally applied to the case of two proteases which, despite a partial structural similarity, differ by length, number of secondary elements and catalytic chemistry. The sophisticated interplay between structural and dynamical analogies are discusses in details.

Both parts contains an introduction clarifying the concepts discusses in the following chapters.

The material presented have been the object of the following publications:

1. F. Pontiggia, G. Colombo, C. Micheletti and H. Orland,
   "Anharmonicity and Self-Similarity of the Free Energy Landscape of Protein G",
   Phys. Rev. Lett. 2007; 98:048102.

2. 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" ,
   Biophysical Journal: In Press.

3. F. Capozzi, C. Luchinat, C. Micheletti and F. Pontiggia,
   "Essential Dynamics of Helices Provide a Functional Classification of EF-Hand Proteins",
   J. Proteome Res.; 2007; 6(11):4245.

4. V. Carnevale, F. Pontiggia and C. Micheletti,
   "Structural and dynamical alignments of enzymes with partial structural similarity",
   J. Phys.: Condens. Matter 2007; 19(28):285206

Part I

Free Energy Landscape of Proteins

Chapter 1

Introduction

The equilibrium population of a typical protein expectedly comprises several different conformers, through which the molecules spontaneously interconverts owing to its internal dynamics [38, 57]. The presence of different biologically-relevant substates can be inferred, for example, by the different crystallographic conformations of the same protein free or bound to its substrate, or by different arrangements of the free protein in the asymmetric unit of the crystal [104, 103]. Variability in the various substates can range from localized differences, such as dihedral angles isomers, to much larger deformations, reflecting the concerted displacement of several elements of secondary structures, possibly also involving changes in the overall tertiary organization.

Computational approaches have proved useful for probing these changes at various levels of structural and temporal detail [72]. It is now well established that these motions can be efficiently captured in terms of a small number of collective coordinates [79, 78], describing and predicting the motion in a low dimensional effective phase space. Several computational approaches have been devised to extract these few relevant coordinates either through "a posteriori" analysis of MD simulation data [48, 4, 20, 78] or predicting them a priori with simplified models [123, 10, 59, 92, 8, 95, 36, 93].

A transparent way of identifying these generalized directions of motions is through a principal component analysis of the fluctuation covariance matrix obtained from molecular dynamics simulations.

The covariance matrix is defined as

$$C_{ij,\mu\nu} = \langle (r_{i,\mu} - \langle r_{i,\mu} \rangle) \cdot (r_{j,\nu} - \langle r_{j,\nu} \rangle) \rangle \tag{1.1}$$

where $r_{i,\mu}$ indicates the μth Cartesian coordinate of the ith atom and $< \cdot >$ represents the time average over the configurations visited during the simulation. For a sufficiently long constant-temperature simulation the time average is equivalent to the canonical average. However, for simulation of limited duration, the visited phase space will be a portion of the available one. Currently available computational resources allow to investigate

timescales at most of the order of $\sim 100ns$, which is typically at least two order of magnitudes less than relevant biomolecular timescales. Nevertheless we shall discuss how specific features of the near native free energy, are robust even over these "limited" timescales.

In calculating covariance matrices, the overall rototranslational motion is usually eliminated. Eliminating the overall rigid-body motion of the molecule is essential to analyse the proteins' internal dynamics [70].

The set of eigenvectors and eigenvalues of the matrix C, corresponding to the collective principal directions of motion and the associated fluctuation amplitudes, can be determined solving the standard eigenvalue problem

$$CV = V\Lambda \tag{1.2}$$

Where C is the covariance matrix of equation 1.1, V is the matrix whose columns contain the eigenvectors of C and Λ is the diagonal matrix of the eigenvalues. The coordinates $r_{i,\mu}$ appearing in the definition of matrix C can be weighted according to their masses. In this case the results of principal component analysis (PCA) of the covariance matrices, can be straightforwardly compared to results obtained from normal mode analysis (NMA), in which the matrix C can be computed starting from the matrix of the second derivatives of the potential energies, evaluated in the energy minimum configuration. It has been shown that, when a protein's dynamical evolution is followed for a very limited time span ($\sim$ ps), the motion occurs within a single energy minimum and the two approaches gives compatible results [20, 68, 79, 78].

When the time window of observation is increased, several local energy minima are visited and the overall motion can be efficiently described as a composition of a diffusion within the individual minima occasionally followed by jumps among them [68, 79, 78]. Following Kitao et al. [79] in a model called JAM (Jumping Among Minima), the covariance matrix may be decomposed in intra-minima and inter-minima contributions:

$$C_{ij,\mu\nu} = \sum_l \omega_l C^l_{ij,\mu\nu} + \sum_l \omega_l [\langle r_{i,\mu}\rangle_l - \langle r_{i,\mu}\rangle][\langle r_{j,\nu}\rangle_l - \langle r_{j,\nu}\rangle] \tag{1.3}$$

$\langle\cdot\rangle_l$ representing an average restricted to the sole configurations pertaining to the substate l, whose weight ω_l is the fraction of the simulation time spent in it.

On a timescale of the order of the hundreds of picoseconds, the protein explores several minima, whose typical residence time is of the order of $\sim$ 10 ps. The comparison of the spectral state densities computed with a PCA within each of the energy minimum substates bore striking similarities. Thus, within the timescale of $\sim$ 1 ns, the system fluctuations can be captured and described efficiently by a limited number of degrees of freedom. We will

discuss in the next Chapters to what extent this result holds when the simulation time is extended to hundreds of ns.

The principal component analysis can be performed both on a detailed atomistic level, keeping into account all degrees of freedom of the system, or within a simplified description, where for example only a single centroid per amino acid (typically coincident with the position of the Cα atom) is retained [4, 78].

In both cases, with a different level of details, the low energy, large scale modes typically correspond to concerted motion of several elements of nearly rigid secondary structures.

As a complement to the PCA analysis of MD simulations, "native centric" models have been introduced. These models, which rely on simplified force fields, have proved useful to reproduce collective modes in proteins [123, 10, 59, 92, 8, 95, 36, 93].

In her seminal paper, Tirion [122]showed that the low-frequency spectrum of globular proteins is almost insensitive to the local details of the atomic composition of the structure and of the specific interaction between them. Specifically, she compared the frequencies obtained from a NMA using a standard atomistic MD force field with those obtained from a simplified force field where all heavy atoms within a certain cutoff distance R where connected by springs of equal strength.

Her analysis started from considering a generic pairwise potential between atoms i and j, $V(d_{ij})$ which is held to be a function of the interatomic distance $\vec{d_{ij}} \equiv \vec{r_i} - \vec{r_j}$. It is further assumed that the potential minimum is attained in correspondence of the atoms separation: d^0_{ij}.

The potential can be quadratically expanded close to the energy minimum and recasted as a function of the displacements of the individual atoms from their reference positions $\delta\vec{r_i} \equiv \vec{r_i} - \vec{r}^{\,0}_i$

$$V = V_0 + \sum_{i,j,\mu,\nu} K_{i,j} \frac{d^{0,\mu}_{ij}\, d^{0,\nu}_{ij}}{|\vec{d}^{\,0}_{ij}|^2} \delta r^{\mu}_{ij}\, \delta r^{\nu}_{ij} \equiv \frac{1}{2} \sum_{i,j,\mu,\nu} M^{\mu,\nu}_{ij}\, \delta r^{\mu}_{i}\ \delta r^{\nu}_{j} \tag{1.4}$$

where $\delta r^{\mu}_{ij} = \delta r^{\mu}_{i} - \delta r^{\mu}_{j}$.
The coupling constant $K_{i,j}$ represents the strength of the quadratic interatomic interaction. Tirion [122]compared the results obtained using coupling constants coming from a parametrized force field with those obtained assuming all $K_{i,j}$ equal to the constant K for all pairs of atoms closer than the cutoff distance R and zero otherwise. The two free parameters of the model, the constant K and the cutoff distance R, can be tuned to optimally reproduce the spectral density obtained with the detailed force field. The accord on the lowest energy density of states is remarkable.

This observation can be justified in terms of the fact that the low-frequency (low energy) excitations involve correlated displacement of entire

groups of amino acids. The observation implies that the atomistic details of the structure, essential for stabilizing specific minimum energy configuration, do not significantly influence the large scale features of the collective excitations.

These considerations have stimulated the further development of simplified models for capturing proteins' large scale fluctuations. In fact, the detailed atomistic force field can be replaced by simplified quadratic interactions (as in eqn. 1.4) limited to a reduced number of interaction centers, typically the Cα ones, in place of all pairs of contacting atoms.

The viability of these models (generally referred to as Elastic Network Models [10, 59, 8, 36, 93]) has been largely verified *a posteriori* against both general dynamical data obtained from experiments, such as the mean-square fluctuations of each residues measured by the crystallographic Debye-Waller factors, and also against more specific dynamical properties such as the principal direction of motions or the covariance matrix obtained from MD simulation [8, 93].

Arguably, the simplest framework for interpreting the simplified motion described by this simplified coordinates, is provided by the overdamped Langevin dynamics system [37, 66]. The resulting stochastic equations of motion for each amino acid subject to the thermodynamic potential of eqn. (1.4) and the surrounding medium [73, 119, 19, 59, 60] are:

$$\gamma_i \dot{\delta r}_{i,\mu}(t) = -\sum_{j,\nu} M_{ij}^{\mu\nu} \delta r_{j,\nu}(t) + \eta_{i,\mu}(t) \ , \tag{1.5}$$

where γ_i is the effective viscous friction coefficient acting on the ith particle, and $\eta_{i,\mu}(t)$ is a stochastic noise whose first and second moments satisfy the usual fluctuation-dissipation relationships [37]

$$\langle\ \eta_{i,\mu}(t)\ \rangle = 0 \ , \quad \langle\ \eta_{i,\mu}(t)\eta_{j,\nu}(t')\ \rangle = \delta_{ij}\,\delta_{\mu\nu}\ \delta(t-t')\ 2K_B\,T\ \gamma_i \ , \tag{1.6}$$

and $K_B\,T$ is the thermal energy. From eqns. 1.5 and 1.6 the average correlations among the displacements of various pairs of residues can be calculated [28]. For the case where the various viscous coefficients in eqn (1.5) take on the same value, γ, one has:

$$\langle \delta r_{i,\mu}(t)\, \delta r_{j,\nu}(t+\Delta t)\rangle \ = \ K_B T \sum_{l}{}' \vec{v}_i^{\,l} \vec{v}_j^{\,l}\, \frac{1}{\lambda_l}\, e^{-\frac{\lambda_l}{\gamma}\Delta t} \ . \tag{1.7}$$

where $\vec{v}^{\,l}$ and λ_l are, respectively, the l-th eigenvector and the l-th eigenvalue of the matrix M and the prime indicates the omission from the sum of the six eigenspaces associated to the zero eigenvalues of M (roto-translational degrees of freedom) [93]. The eigenvectors of M represent the independent modes of structural relaxation in the protein while the associated eigenvalues are inversely proportional to the relaxation times. In fact one has:

$$\langle \delta r_{i,\mu}(t)\, \delta r_{j,\nu}(t) \rangle \;\; = C_{ij,\mu\nu} = \;\; K_B T \sum_{l}{}' \vec{v}_i^{\,l} \vec{v}_j^{\,l} \, \frac{1}{\lambda_l} \; = K_B T M'^{-1}_{i,j,\mu,\nu} \, . \qquad (1.8)$$

While the above Langevin scheme is highly transparent and amenable to analytical treatment, it should be mentioned that it is not adequate to capture a number of salient features of protein's internal dynamics (for which more sophistical theoretical schemes have been devised) [80, 81, 83, 97]

The proved utility of the elastic network models for describing proteins' internal dynamics poses a number of questions that have been addressed in the following Chapters. In particular: what light can be shed by extensive MD simulations on the effective harmonic character of the near-native free energy landscape? Can extensive MD simulations provide clues to develop improved elastic network models(for example capable of accounting for amino acid specificity?). In a number of points, these questions overlap with those posed by Kitao et al. in their seminal study of the JAM model [79]. However, in that context, the near-native free-energy landscape was probed by means of sub-ns short MD trajectories. The analysis that is carried out here is based on the collections of MD data over hundreds of ns, and this confers to the investigation a much broader scope and hence qualitatively-different perspective.

We will start our investigation analyzing the evolution of a well studied globular protein, B1 binding domain of protein G. On the timescale of $\sim$ 100 ns, the system visit a number of local free energy minima. A systematic analysis of the essential dynamical spaces calculated within and across the different substates encountered during the dynamical evolution, has revealed an unexpectedly simple self-similar structure of the free energy, which reflects in the unexpected robustness of the system essential dynamical spaces at all probed time-scales. By converse, the amplitudes projected along the essential spaces depend strongly on the number of visited minima as well as on their depth. As a result, the dynamical projections have a strong dependence on the duration of the simulation, a fact that accounts for the observed inconsistency of the covariance matrix entries calculated over time intervals of different duration. The implications of these results for the development and use of elastic network models are discussed.

Finally, these concepts are further developed in the context of another widely studied enzyme, Adenylate Kinase. First, we briefly discuss the most relevant recent experimental and computational studies that have elucidated the major role of conformational fluctuations in the realization of the biological function for this specific enzyme. Then we will describe how the system exploits its innate flexibility to explore the landscape and efficiently bridge the important conformations.

As in the previous study, irrespective of the probed time-scale, all intra-

and inter-substate essential dynamical spaces turned out to be highly consistent. Moreover, the functional relevance of this consistency is underscored by the high overlap that the essential dynamical spaces have with the deformation vector connecting the known ligand-bound and free structures of AKE.

The analysis indicates that the free enzyme can be driven through various conformational substates bridging the inactive and catalytically potent states through the thermal excitation of a limited number of collective modes.

Chapter 2

Immunoglobulin Binding Domain of Protein G (GB1)

Our first investigation is performed on the β1 Immunoglobulin Binding Domain of Streptococcal Protein G [47], a protein living on the surface of the bacterial cell. Its role is to bind the constant region of immunoglobulin thus allowing the bacterium to elude the immune system of the host organism.

This protein domain has been the object of a large number of experimental and computational studies [87, 43, 116, 44] as, despite consisting of only 56 amino acids, it is stably folded and possesses a non-trivial α/β tertiary organization. We have chosen this protein domain as a test case for our free-energy investigation as its moderate length makes it amenable to extensive MD simulations. A cartoon representation of the three dimensional structure of the protein is portrayed in Fig 2.1.

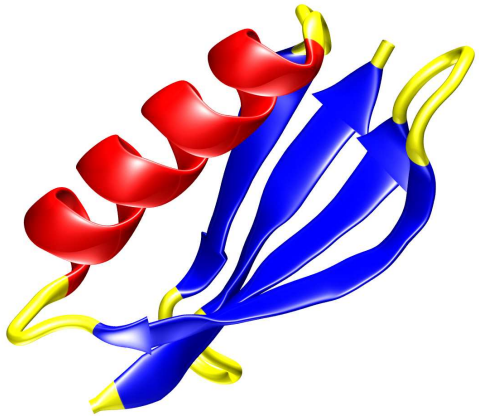


Figure 2.1: Cartoon representation of the Immunoglobulin Binding Domain of Protein G, pdb code 1PGB [47].

We have followed the dynamical evolution of the system in four independent molecular dynamics runs, covering a total time span of 0.4 μs. We analyzed the visited conformers, assessing several features of the near native free energy landscape, described in terms of the collective motion of the Cα atoms. We first address the viability of adopting an effective harmonic approximation to the free energy. A quadratic character of the free energy (in terms of the principal components of the protein's internal fluctuations) is suggested by the effectiveness of elastic network models (ENM) in capturing proteins large scale movements. Understanding the key aspects underlying the supposedly quadratic character of the free energy would have several conceptual and practical ramifications, including the possibility of leading to improved ENM.

As a first mean for probing the viability of the harmonic approximation we shall analyze the robustness of the covariance matrix entries calculated over simulations of increasing duration. It is found that, as soon as the evolution of the protein is followed for time spans larger than few nanoseconds, a marked multiminima character of the principal components projections arises, thus indicating that well separated conformational substates are visited. A detailed investigation of the free energy landscape is performed to characterize the relatedness of "small-scale" structural fluctuations within the substates and the "large-scale" ones associated to the hopping between substates. It is found that the principal directions of the local free energy minima and the directions connecting them are extremely robust upon extension of the exploration time. This remarkable robustness of the principal components provide a new perspective through which validity of simplified models can be rationalized.

2.1 Molecular Dynamics Simulation

The crystallographic structure deposited in the Protein Data Bank [15] with pdb code 1PGB [47] was taken as the starting point of four MD runs in explicit solvent each of 100 ns. The system has been solvated in an octahedral box containing a solvent layer of 1.2 nm. The protein has been parametrized with GROMOS 96 (G43a1) [126] force field and the solvent with the SPC water model [13]. After energy minimization and progressive heating to 300K, the density was adjusted by a 100ps-long MD in NPT conditions by weak coupling to a bath of constant pressure (P = 1 bar, coupling time τ=0.5 ps) [14]. Subsequently the four different 100-ns long trajectories were started with different sets of Maxwellian (T=300 K) initial atomic velocities. The bond vibrations of all hydrogens have been constrained using the Lincs algorithm [58] and all other system degrees of freedom have been integrated with the standard Verlet algorithm with an integration time step of 2 fs, using the GROMACS simulation package [125]. The particle mesh

Ewald method was used for treating electrostatics [34, 39] and temperature was controlled with the algorithm proposed by Berendsen (coupling time τ=0.1ps) [14].

The first 10 ns of each trajectory have been removed from the analysis to reduce the effect of correlations due to having used the same configuration as starting structure for the four runs.

2.2 Consistency Analysis of Covariance and Effective Coupling Matrices

To ascertain the degree of robustness of the covariance matrices of the Cα displacements we compare their corresponding elements computed over uncorrelated fragments of trajectory (also of different duration). Prior to calculating the covariance matrix entries, the conformers ("frames") collected during the dynamical evolution have been aligned to optimally remove the rigid body motion of the molecule [70]. The energy-minimized starting configuration was taken as the initial reference structure for the alignment. The average structure of the aligned frames was next taken as reference for a second alignment round. This iterative procedure converges very rapidly and leads to the maximal removal of rigid-body motions in the system (in the sense that it minimizes the overall fluctuation ascribed to the protein internal dynamics). The covariance matrix is finally computed (see eqn. 1.1).

In Fig. 2.2 we show the comparison between covariance matrices built from frames taken from an interval of duration $\Delta t = 1$ ns, extracted at the end of two distinct trajectories, so to exclude as much as possible intrinsic correlation. In panel (a) of the figure, the reduced covariance matrix elements are represented:

$$\tilde{C}_{i,j} = \sum_{\alpha} C_{ij,\alpha\alpha} \tag{2.1}$$

From the analysis of the accord between reduced covariance matrices (invariant under rototranslations of the proteins) it is evident that there is a fair degree of correlation, but the dispersion of the data pairs is substantial: the linear correlation coefficient over the $\sim$ 1500 distinct entries ($\frac{N(N-1)}{2}$ where N=56 is the number of protein residues) is 0.65. Also, it is interesting to notice that the "cloud" of Fig. 2.2 does not appear to be oriented along the $y = x$ line. Indeed, there is roughly a factor 4 in the breadth of the covariance matrix spanned by the two runs. This is indicative that the width of the visited minima is only roughly similar, despite the equal simulation time.

The dispersion of the points increases substantially if we consider the comparison of full covariance matrices entries (2.2b), taking into account

also the directionality of the fluctuations. The correlation over the $\sim$ 14000 entries ($\frac{3N(3N-1)}{2}$) decreases to 0.29.

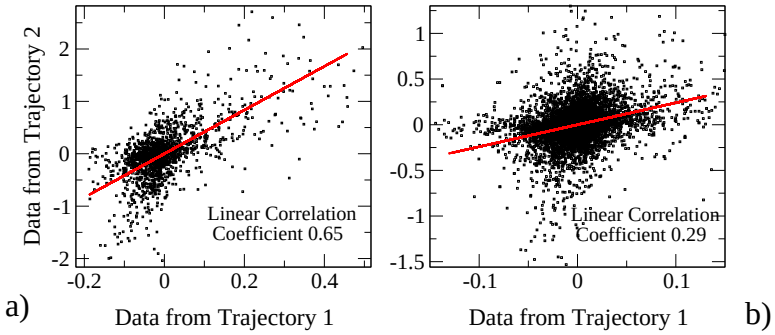


Figure 2.2: Scatter plot of the covariance matrices relative to the last ns of 2 different trajectories. In panel a are shown the $\sim$ 1500 distinct entries of the reduced covariance matrices while the scatter plot of panel b is referred to the $\sim$ 14000 entries of the full covariance matrices.

If we analyse the covariance matrices calculated over progressively longer time spans, the degree of correlation degrades very rapidly.

The poor consistency of the covariance matrices is expected to be reflected in a similar lack of robustness of the effective coupling matrices M, (see eqn. 1.8) obtained by pseudo inversion of the covariance matrices.

To verify this we have systematically compared pairs of M matrices from the different runs and/or for increasing Δt. Their similarity is measured by means of the Kendall's correlation coefficient (τ) computed over the $\sim$ 14000 corresponding entries [112]. Using the Kendall's τ has the advantage of reducing the influence of data outliers. More importantly, it provides a robust measure of data association with no prior assumption of the parametric dependence of the examined data sets. As visible in the figure 2.3, extending the duration of the trajectory, be it 1 or 16 ns, by four or more times leads to a substantial deviation of corresponding entries of M (with $\tau \sim 0.30$). Consistently, halving each trajectory and considering the correlation among the two halves gives values of τ between 0.17 and 0.25. The consistency of M matrices of different trajectories is, furthermore, much poorer and almost independent on the compared time spans. In fact, pairwise comparisons of different runs over the first ns or over the entire 90-ns duration trajectories yield values of τ in the [0.05- 0.15] range.

These values indicate a substantial degree of heterogeneity in corresponding matrix entries and hence point to the impossibility of having a robust definition of M (even over hundreds of ns) to be used e.g. as phenomenological parametrization of elastic network models.

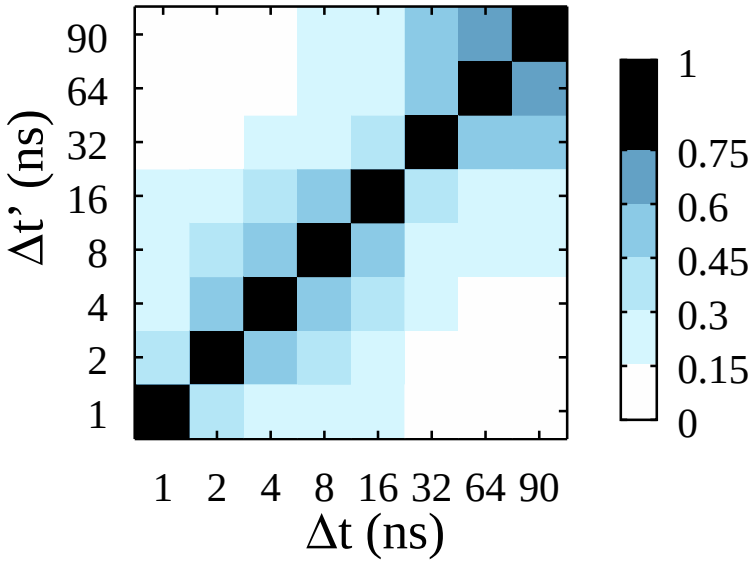


Figure 2.3: Color-coded plot of the Kendall's correlation coefficient, τ, between corresponding elements of the coupling matrices, M, calculated over the first Δt and $\Delta t'$ ns of the first trajectory. Matrix elements along the diagonal or pertaining to consecutive C_α's were omitted in the calculation of τ.

2.3 Simple Robust Features of the Coupling Matrices

Despite the fact that no asymptotic value appears to be reached by M (or C) entries, the latter display some general property that appears to be rather robust against the increase of simulation time span, Δt.

We begin by considering an effective free energy quadratic in the displacement of C_α atoms

$$F \approx \frac{1}{2} \sum_{i,j,\alpha,\beta} M_{ij}^{\alpha,\beta} \, \delta r_i^\alpha \; \delta r_j^\beta \; . \tag{2.2}$$

Notice that the strength of the coupling between the pair i, j of amino acids is aptly given by (see eqn 1.4):

$$K_{ij} = -\sum_\alpha M_{ij}^{\alpha\alpha} = -\sum_\alpha C_{ij}^{-1\;\alpha\alpha}, \tag{2.3}$$

where the sign convention is chosen so that positive [negative] entries correspond to attractive [repulsive] interactions. We have used this relationship to compute the strength of the harmonic pairwise couplings between

all amino acids.

Below we plot the histogram of values of the couplings, K_{ij}, for pairs of consecutive C_α's (bonded interactions), see Fig 2.4a, and for pairs of non-consecutive C_α's with a native distance smaller than 7.5 Å, see Fig 2.4b.

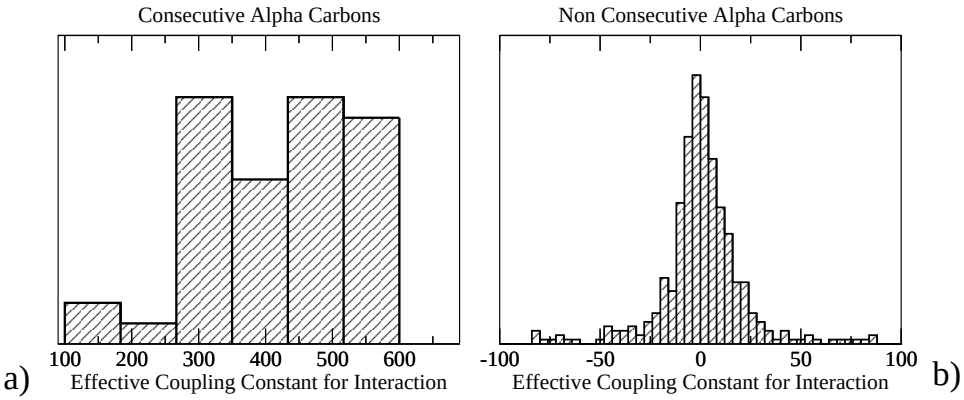


Figure 2.4: Histogram of effective couplings for (a) consecutive Cα and (b) contacting but not consecutive Cα

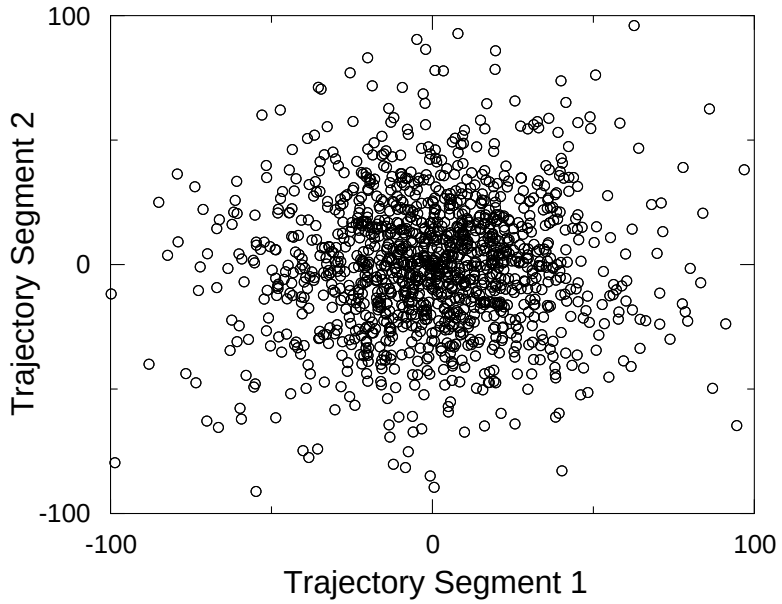


Figure 2.5: Scatter plot of the K_{ij} of pairs of not consecutive Cα with native distance below 7.5 Å obtained from 1ns of simulation in independent trajectories.

The results shown in Fig 2.4 indicate that the effective coupling for amino

acids that are consecutive in sequence is about two orders of magnitude larger than the strength of contacting, but non-consecutive residues. The enhanced interaction couplings in Fig 2.4a reflect the fact that the virtual bond connecting consecutive C_α's undergoes fluctuations that are very limited. This is consistent with what already observed in previous studies in the contest of Hessian analysis[60]. A second interesting feature is that the distribution in panel (b) covers also negative values ("repulsive" interactions). They are however outweighted by the positive ones, as the system would otherwise be unstable. The distribution is nevertheless very different from the idealised one commonly employed in elastic network models, which consists of a sharp peak reflecting the uniform strength of all harmonic couplings between contacting amino acids.

A natural question that emerges is whether the disperse character of the distribution in Fig 2.4 reflects any sequence-specific properties of the system. In fact, one may expect that different types of amino acids may be associated with different interaction strengths. While we cannot rule out that specific pairs of amino acids may interact with definite strength (or sign) [102], a further analysis indicates that no such information can be easily extracted from the data in Fig. 2.4.

To ascertain this, we have computed the effective amino acid couplings using MD data from 1ns-long intervals from two distinct simulation. Corresponding entries are plotted one against the other in Fig. 2.5. The figure strikingly exhibits a lack of correlation between the two data sets. On one hand, this fact points at the impracticality of using an MD-based analysis to develop "phenomenological" (sequence-specific) elastic network models. On the other hand it stimulates the further investigation of why, despite the apparent lack of robustness of the covariance matrix entries (and hence of the individual a.a. pairwise couplings) the essential dynamical spaces observed in MD trajectories are usually in good accord with those predicted by simplified ENM approaches.

2.4 Anharmonic Character of the Low Energy Modes

The lack of robustness of M matrices upon increase of the simulated time Δt is presumably rooted in a complex character of the underlying free energy. We will dissect the structure of the free energy by analyzing explicitly principal directions of motion and their corresponding amplitudes. For a system governed by a purely quadratic free energy, the equilibrium fluctuations around the average structure have a Gaussian probability distribution along any of the eigenvectors of M. The width of the Gaussian is largest for the lowest energy mode which, corresponding to the direction of least curvature of F, mostly accounts for the system fluctuations in thermal equilibrium. A valuable insight into the effective free energy landscape described by $C(\Delta t)$

is hence obtained by considering the distribution of the projections along the lowest energy mode of the conformational fluctuations in trajectories of increasing duration. Typical results are shown in Fig. 2.6.

It can be noticed that, in accord with what established in previous studies (see e.g. [4]), for $\Delta t \sim 1$ ns, the distributions have a unimodal character with a fair degree of Gaussianity, while as time spans increases the multi-modal character of the distributions becomes evident.

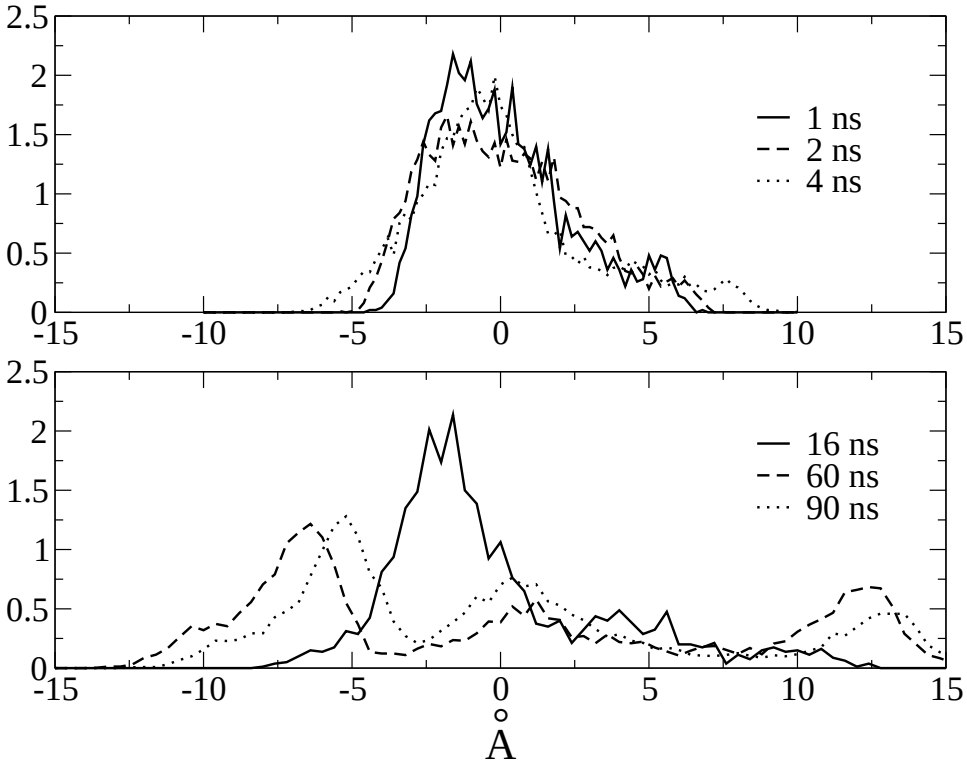


Figure 2.6: Normalised distribution of the projection along the first slow mode of the conformational fluctuations encountered in intervals of increasing length from one of the trajectories.

A natural and useful quantifier for assessing the progressive deviation from Gaussianity of the distribution for increasing Δt is the normalized kurtosis $\kappa = (\langle x^4 \rangle - 3\langle x^2 \rangle^2)/\langle x^4 \rangle$ (x being the projection). For Gaussian distributions the normalized kurtosis is 0. An analysis on the distributions of our projections shows that the initially low kurtosis value ($\kappa \approx 0.15$ for $\Delta t < 1$ ns) progressively increases with Δt and attains values of $\kappa \approx 1$ for $\Delta t \approx 10$ ns.

The considerations of the non-Gaussian character of the principal components projections can be extended to the fluctuations of the individual amino acids. We have considered the distributions of the trajectories of

each C_α atom projected onto its own principal directions of motion. The deviations from gaussianity for these distributions are particularly marked in specific protein regions, for example for residues 30–40. Figure 2.7 shows a typical behavior for one of these distribution. It can be seen that the corresponding C_α atom takes on two different positions.

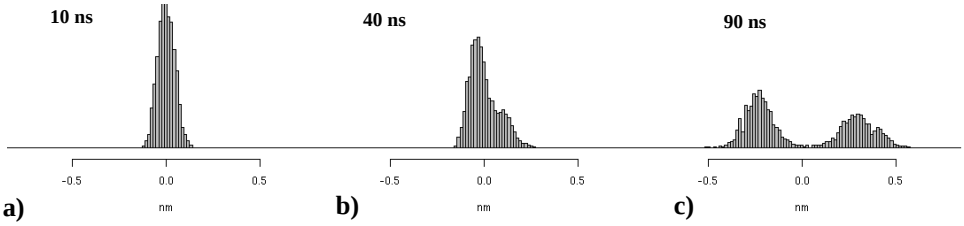


Figure 2.7: Normalised distribution of the projection along the first mode for one of the most mobile residues.

2.5 Progressive Softening of the Underlying Effective Potential

The analysis has shown that the width of the visited configurational space increases progressively with the simulation time and leads to an increasing anharmonic character of the explored free energy surface.

An interesting counterpart of this property is that the eigenvalue of M associated to the lowest energy mode is a decreasing function of Δt. This can be interpreted as a progressive weakening of the strength of the quadratic free-energy well upon widening of the explored phase space. This trend can be illustrated by means of a simple, but transparent, dynamical variational approach. More precisely, for a fixed simulated time span, we shall consider the projection of a trajectory along its slowest mode $\vec{v}$ having components $\{\vec{v}_1, \vec{v}_2, ...\}$ and attempt to describe its time-varying amplitude with a deterministic harmonic modulations of the amino acids displacements.

If we indicate with $\langle \vec{r}_i \rangle$ the average position of the ith centroid, the model dynamics (indicated with the superscript m) is therefore governed by the Hamiltonian:

$$\mathcal{H}^m = \frac{\lambda}{2}\Big(\sum_i \vec{v}_i \cdot (\vec{r}_i{}^m - \langle \vec{r}_i \rangle)\Big)^2 \qquad \sum_i \|\vec{v}_i\|^2 = 1 \tag{2.4}$$

The corresponding Newton's equations of motion are:

$$m_i \ddot{\vec{r}}_i{}^m = -\frac{\partial \mathcal{H}^m}{\partial \vec{r}_i{}^m} = -\vec{v}_i \sum_j \lambda \vec{v}_j \cdot (\vec{r}_j{}^m - \langle \vec{r}_j \rangle) \tag{2.5}$$

We consider for simplicity the effective mass of all centroids equivalent, $m_i = m$, and define $\frac{\lambda}{m} = \omega^2$, so that Newton's equation read:

$$\ddot{\vec{r}}_i^m = -\vec{v}_i \sum_j \omega^2 \vec{v}_j \cdot (\vec{r}_j^m - \langle \vec{r}_j \rangle) \tag{2.6}$$

We now introduce the auxiliary variable: $y^m = \sum_i \vec{v}_i \cdot \vec{r}_i^m$. Using the condition of normalization of the eigenvector $\vec{v}$ we obtain from the equations of motion:

$$\ddot{y}^m = -\omega^2 (y^m - \overline{y}) \tag{2.7}$$

where $\overline{y} = \sum_i \vec{v}_i \cdot \langle \vec{r}_i \rangle$. The equation 2.7 readily gives:

$$y^m(t) = \overline{y} - A\cos(\omega t + \phi) \tag{2.8}$$

Returning to the original variables $\{\vec{r}_i\}$ we obtain $\vec{r}_i^m(t) = (y^m(t) + D_i t)\vec{v}_i$ subject to the condition $\sum_i D_i \vec{v}_i = 0$. We propose a solution like

$$\vec{r}_i^m(t) = C_i - A\vec{v}_i \cos(\omega t + \phi) \tag{2.9}$$

subject to the initial conditions $C_i = \vec{r}_i^0 + A\vec{v}_i \cos(\phi)$, so that the oscillators coordinates coincide with those of the real system at $t = 0$ (indicated with $\vec{r}_i^0$).

We now introduce a criterion of optimality to fix the free parameters ω, A and ϕ. They are chosen so to minimize the time-averaged total square deviation of the model trajectory from the real one:

$$\langle \sum_i \| \vec{r}_i^m(t) - \vec{r}_i(t) \|^2 \rangle \tag{2.10}$$

We have adopted this scheme to model the dynamics of several intervals of various duration taken from the recorded trajectories. As visible in Fig. 2.8, for time spans of up to fractions of a ns the oscillator is able to account satisfactorily for the evolution of the true trajectory (manifestly overdamped over $\Delta t >$ 0.5 ns).
The plot of Fig. 2.9, presents the trend of the average optimal frequency ω, as a function of the length of the time interval. It can be perceived how rapidly the curvature of the effective quadratic free energy decreases as a function of Δt. The softening reflects the complexity and anharmonicity of the free energy landscape which, as confirmed by the multimodal character of the distributions of Fig. 2.6, is constituted by broad minima of varying depth which are progressively explored as the dynamics advances.

The analysis conducted in this and the preceding sections has highlighted the progressive erosion of the harmonic character of the free-energy as the phase space visited by MD trajectories is enlarged. In the next section we

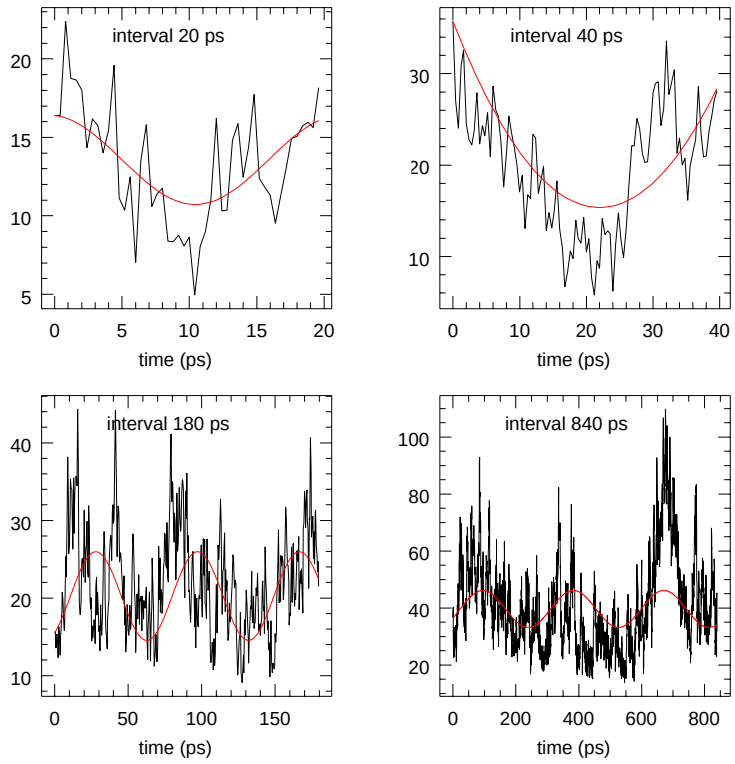


Figure 2.8: Time evolution of the projection along the 1st eigenvector and of the variational deterministic oscillator for intervals of different duration.

shall discuss how, despite the substantial degree of anharmonicity the near-native free energy possesses a remarkable self-similar organization (which can profitably exploited in the context of simplified, coarse-grained models).

2.6 Effective Dimension of the Visited Landscape

As already mentioned, in ref. [79] Kitao et al. discuss the idea of separating in the calculation of covariance, the contribution due to the diffusion within the local energy minima visited, from the jumping among them. This is done by considering the covariance matrix as made of contributions from fluctuations within each of the various minima and the structural displacements of the various minima from the reference (time-averaged) structure. The expression of equation 1.3 can be recasted as

$$C = \sum_l w_l \{ C^l + \sum_{k,m} w_m \, w_k \, |\vec{d}_{l,k}\rangle\langle\vec{d}_{l,m}| \} \tag{2.11}$$

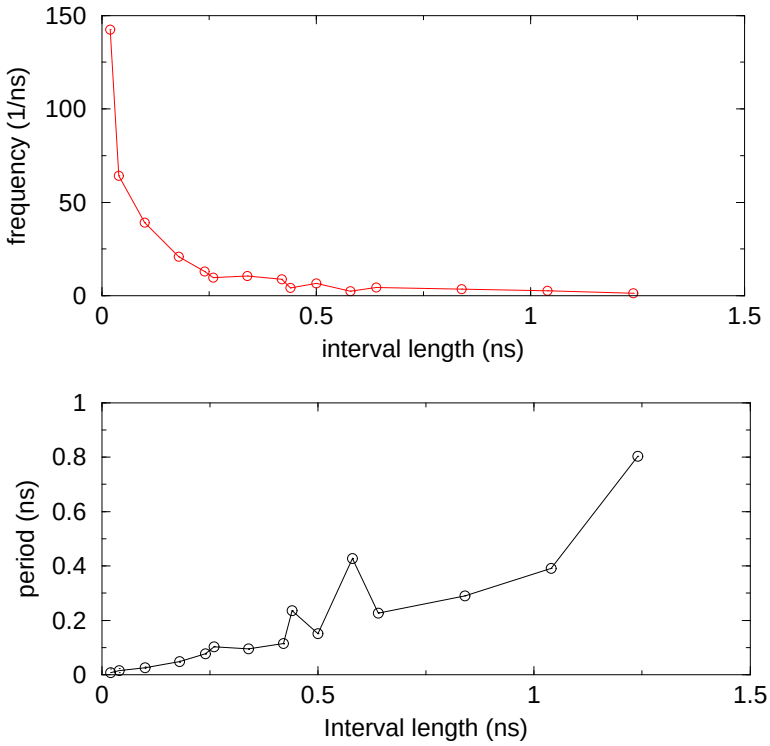


Figure 2.9: Average value of the variational oscillator frequency as a function of the interval duration, Δt.

where w_l is the weight of the lth cluster, that is the fraction of time spent by the system in it, C^l is the covariance matrix of the lth cluster, and $\vec{d}_{l,m}$ is the distance vector of the representative (average) structures of clusters l and m.

We shall specifically characterize the salient minima performing a cluster analysis of the structures generated in all four trajectories and considered the 10 most populated sets. As a measure of structural similarity of any two conformers we consider the root mean square deviation (RMSD) of corresponding Cα atoms after an optimal structural superposition. The distribution of all pairwise RMSD distances of 40000 configurations, recorded once every 10 ps from the cumulated trajectories is portrayed in fig 2.10. The double peak character of the distribution is an indication of natural groups present in the sample. The location of the left-most peak provides an estimate of the typical intra-group pairwise distance while the second can give a measure of the expected inter cluster separation[96]. We have thus performed a cluster analysis by grouping structures closer than 1.5 Å RMSD. To identify the groups and the corresponding representatives, we

first count the number of neighbors of each structure, within the selected cutoff distance. The configuration having the largest number of neighbors defines the most populated cluster. The procedure is iterated over the remaining conformations[35, 96]. In the analysis we will consider the 10 most populated groups. Each cluster, containing conformations at less than 1.5 Å RMSD from the representative, gathers configurations from time-intervals of very different length, from 1 to tens of ns, originating from one to three of the four trajectories. By carrying out a structural covariance analysis (with the time average of eqn. 1.1 being replaced with an average over cluster members) we identified the 10 principal components describing the largest conformational changes in each cluster.

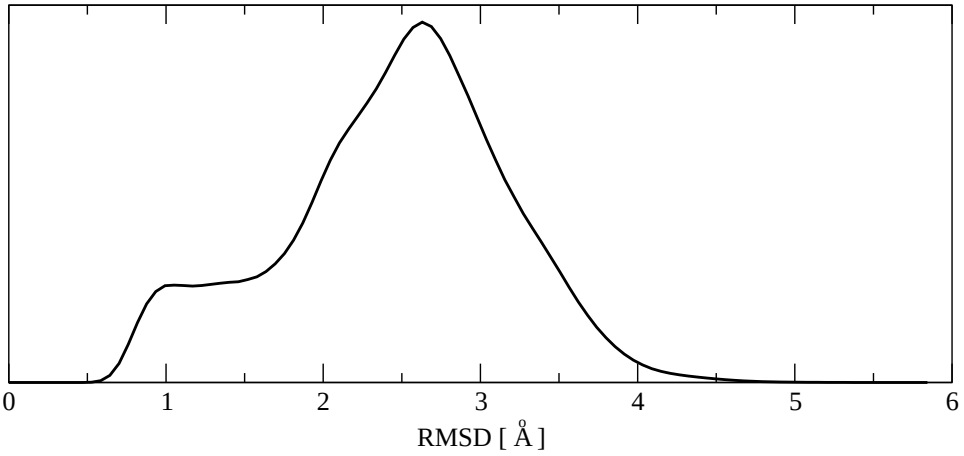


Figure 2.10: Distribution of RMSD of 40000 conformations extracted from the 4 trajectories.

The sets of the most relevant (typically the top ten eigenvectors are considered) principal directions corresponding to the largest eigenvectors of different covariance matrices are compared to assess their common overall orientation.

We shall indicate two such sets of essential dynamical spaces as $\{v\} \equiv \{\vec{v}_1, \vec{v}_2, ..., \vec{v}_{10}\}$ and $\{w\} \equiv \{\vec{w}_1, \vec{w}_2, ..., \vec{w}_{10}\}$. Their common orientation, induced by the superposition of the corresponding reference structures, will also be assumed. The consistency of $\{v\}$ and $\{w\}$ was quantified, as customary, via the root mean square inner product, (RMSIP) [3]:

$$RMSIP = \sqrt{\frac{1}{10} \sum_{i,j=1}^{10} (\vec{v}_i \cdot \vec{w}_j)^2} \ , \tag{2.12}$$

which ranges from 0, for complete orthogonality of the $\{v\}$ and $\{w\}$ spaces, to 1 in case of their perfect coincidence.

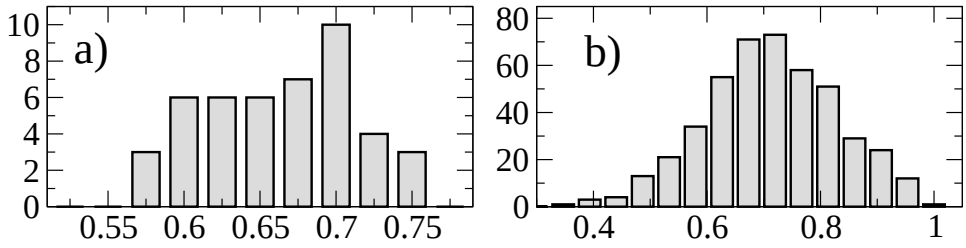


Figure 2.11: (a) Histogram of the RMSIP values calculated for the principal components of all pairs of top structural clusters. (b) Histogram of the norm of the projection of all pairwise distance vectors between cluster representatives along the top 10 principal components of the largest cluster.

By means of the RMSIP we finally compared the 10 principal components of all distinct pairs of clusters. The resulting distribution of RMSIP values indicated a very strong consistency of the sets of principal directions, see Fig. 2.11a. Indeed, for the considered protein length, numerical results indicate that if the set of principal components were completely unrelated, the expected RMSIP value would be 0.24 ± 0.02. The distribution of values in Fig. 2.11 is sufficiently distant from this random reference value to convey the significance of the observed consistency of the principal components of different clusters. Strikingly, it was also found that the difference vectors, $\vec{d}_{ij}$ connecting the representatives of any pair of different clusters i and j are also well described by the principal components of any of the clusters. For example, as shown in Fig. 2.11b, the top 10 principal components of the largest cluster are usually sufficient to account for most of the norm of the "virtual jumps" connecting the representatives of any two top clusters.

This observation extend and complements the considerations of Kitao et al. [79] on the similarity of the energy minima. Not only the deep free-energy minima corresponding to the main clusters have similar principal components, but also the virtual "jumps" connecting their representatives are describable in the same low-dimensional space. That being so, we can assume a single set of top principal component can be used to expand both intra- and inter- substate terms appearing in the decomposition of the covariance matrix of eqn 2.11.

According to this approximation, the principal eigenvectors of C, and hence the slow modes of M, would coincide with $\vec{v}_1$,... $\vec{v}_n$ and thus remain unchanged over all time scales. On the contrary, the associated eigenvalues would explicitly depend on the MD duration through the time-dependent number and weight of the visited clusters.

This remarkable consistency of the principal directions is further address comparing the essential spaces of different time intervals from the simulated trajectories. The results are illustrated in Fig. 2.12 which portrays the RMSIP calculated between the essential spaces of the 1st ns for trajectory

1 with larger and larger time spans for the same and other trajectories. It is seen that the top slowest modes are very robust against increasing Δt and remain consistent even increasing the simulation time by two orders of magnitude (from 1 to 90 ns). The statistical significance of this result is highlighted by the difference of RMSIP ranges in Fig. 2.12 from the aforementioned random reference value of 0.24. As a further comparison we also considered the RMSIP value calculated over all pairs of 10 mid-ranking eigenvectors of the last 1 ns of all four trajectories. Also this more stringent test indicates that the RMSIP's in Fig. 2.12 exceed the control value by at least 4 standard deviations and hence have a high statistical significance.

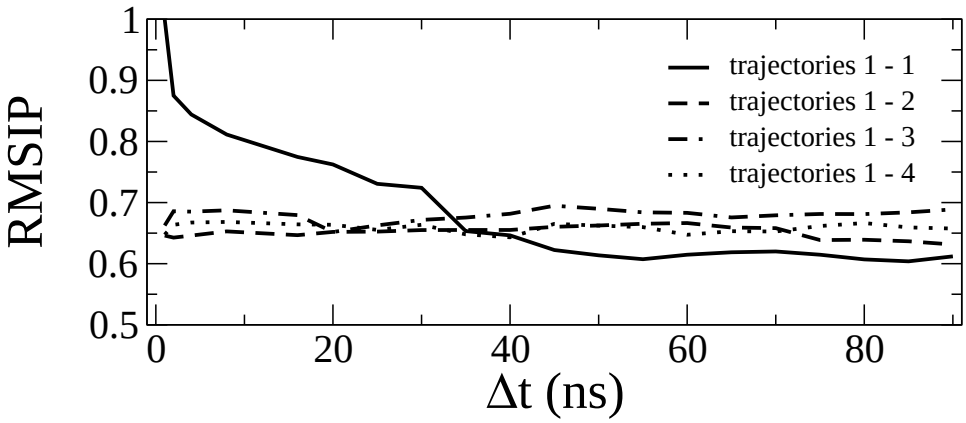


Figure 2.12: RMSIP between the top essential dynamical spaces of the 1st ns of trajectory 1 and intervals of longer duration, Δt, from the same and other trajectories.

2.7 Summary

We have shown that the near-native free-energy of GB1 possess a simple self-similar structure reflected by the high consistency of the principal directions of the various local minima and of the virtual jumps that connect them. The analysis complements and extends previous investigations of free-energy organization that were based on MD trajectories having durations two order of magnitude smaller than the present analysis. This remarkable feature reflects into the exceptional robustness of the essential dynamical spaces (slow modes) calculated over trajectories with very different duration. However, the typical amplitudes projected along the slowest modes depends on the number of visited minima as well as on their depth. As a result, the dynamical projections have a strong dependence on the duration of the simulation, a fact that accounts for the observed inconsistency of the coupling (or covariance) matrix entries. The observed properties, besides

elucidating general features of the free energy landscape of one particular protein, have important practical ramifications. In particular they provide a first perspective, for understanding the scope and viability of coarse-grained elastic network models as well as short MD simulations. Accordingly, it is expected that the directionality of the slow modes/essential dynamical spaces can be determined with considerable more confidence than the amplitude of the associated dynamical projections. These considerations provide a strong motivation for investigating the validity and transferability of the present analysis to other protein contexts.

Chapter 3

Subdomains Motion and Mechanics of Adenylate Kinase

Stimulated by the previous findings, we extended the investigation of the structure of the free energy landscape to another important enzyme, adenylate kinase, whose internal dynamics is known to play a major role in the accomplishment of the biological function. For this reason this enzyme is an optimal case study to further investigate the connection between protein functional dynamics and intrinsic features of the free energy landscape.

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

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

The differences in structural arrangement between the free E. Coli adenylate kinase (AKE) and the enzyme complexed with an inhibitor mimicking both ATP and AMP are illustrated in Fig.3.1 [104, 103]. By comparing the two portrayed crystal structures it is apparent that the formation of the ternary complex stabilizes the enzyme in a form where the mobile Lid and AMP-binding subdomains (highlighted in Fig. 3.1) close over the remainder core region. This rearrangement of the two mobile subdomains is necessary for the accommodation of the nucleotides in an optimal catalytic geometry and the resulting closed enzyme conformation provides a solvent-free environment for the phosphoryl transfer.

The conformational change sustained by adenylate kinase upon complexation with ATP and AMP, and its reopening upon unbinding of the processed nucleotides represents the rate-limiting step in the reaction turnover [75]. A large number of experimental studies have consequently been devoted to elucidating the functional implications of Adk structural elasticity

[104, 103, 117, 115, 113, 54, 128, 75, 114, 57, 55]. In particular, recent investigations based on a wide range of techniques, have provided converging evidence for the fact that, even in the absence of the bound nucleotides, the free enzyme is capable of interconverting between the open and closed forms [57, 55]. These investigations have lead to formulating the hypothesis that evolutionary pressure has endowed Adk, and arguably other enzymes [12, 38], with the innate ability to interconvert between the open and catalytically-potent forms [1, 57].

These observations have stimulated the present numerical study of the dynamical evolution of the free (apo) AKE molecule in solution. By means of two MD simulations started from the available crystal structures we have characterized, over various time scales, the conformational fluctuations sustained by the enzyme and analyzed the extent to which they indicate the suggested innate predisposition to connect the open and closed forms.

Several previous computational investigations of the flexibility of AdK exist and include both mesoscopic and atomistic approaches. Coarse grained models have, for instance, been applied to model the pathways connecting the open and closed forms of the enzyme[100, 91, 30]. Atomistic simulations have instead been used to probe the free energy landscape in the neighborhood of several known enzyme conformers, as in the recent investigations by Lou *et al.*[90], Arora *et al.* [7] and Henzler-Wildman *et al.* [57]. In the first study [90], an advanced sampling technique was used to show that the enzyme populated conformations compatible with the holo-form geometry, as probed by FRET experiments. [117]. Arora *et al.* [7] further showed that the free energy landscape along a pre-assigned reaction coordinate connecting the open-closed forms of AKE is approximately flat for the apo-form while, upon ligand binding, it changes favoring the closed state. Finally, in the study of ref. [57], carried out on Adk extracted from hyperthermophile *Aquifex Aeolicus*, a variety of experimental and computational probes provided converging evidence for the existence of several metastable configurations bridging open and closed states.

In the present study, we have analyzed the recorded trajectories with a series of specifically-developed tools. Prompted by the previously discussed results of our study of GB1, we first assess the level of structural heterogeneity of the visited conformational phase space, identifying local metastable substates. The internal dynamics within the substates and the discontinuous jumps across them are analyzed in detail in order to elucidate the what relationship, if any, exists between (i) the directions of largest structural variability within the substates, (ii) the difference vectors that connect the substate representatives and (iii) the functional conformational change associated to the deformation vector bridging the available apo/holo crystal structures. These questions are at the heart of the "multiscale" spirit of the present analysis aimed at characterizing the connection between the system conformational fluctuations at the smallest scale (within the substates) and

at the largest, functional one, embodied by the open/close conformational change.

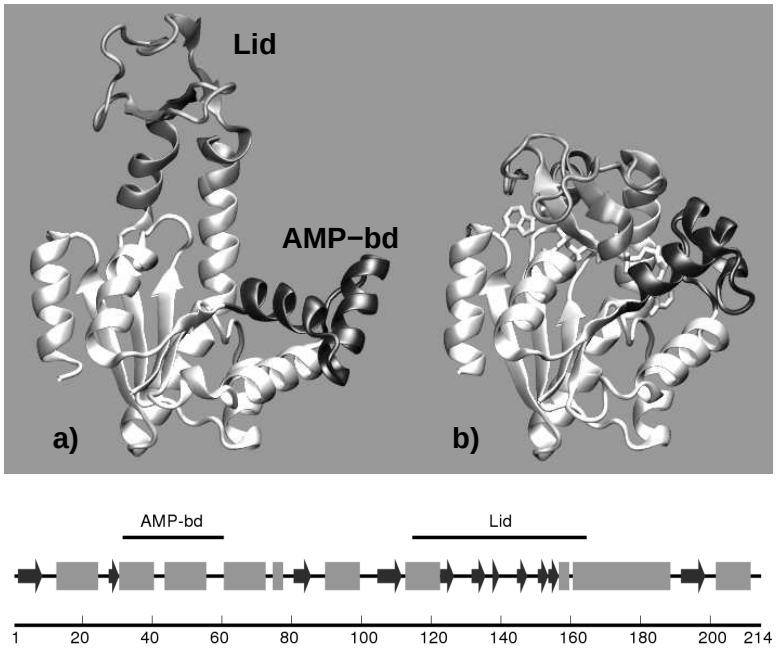


Figure 3.1: Cartoon representation [67] of crystallographic structures of E.Coli adenylate kinase in: (a) the open apo form and (b) the closed holo form. The PDB codes for the two structures are 4ake and 1ake, respectively [104, 103]. The flexible Lid (amino acids 114-164) and AMP-binding (amino acids 31–60) domains are colored in gray and black, respectively. The succession of secondary elements is shown in the bottom panel. Helices are indicated as gray boxes while β-strands are shown as black arrows.

3.1 Molecular Dynamics Simulation

The atomistic molecular dynamics evolution of E. Coli adenylate kinase, AKE, was followed starting from two distinct initial structures, corresponding to the open and closed form of the enzyme. More precisely, the initial conformation of the first simulation was the free (apo form) enzyme from the 4akeA PDB crystal structure. The second simulation followed, instead, the evolution of the free closed form of the enzyme obtained by removing the Ap5A inhibitor from the 1akeA PDB structure file. In the following, for simplicity, we shall refer to the two simulations as the "open" and "closed" trajectories. The nomenclature is only meant to remind of the starting con-

figuration as, in fact, for both trajectories a partial interconversion to the complementary (open or closed) state is observed.

Each system was parametrized with OPLSS-(AA)/L force field [69, 71] and was energy minimized after solvation by 17694 simple point charge (SPC) water molecules in a cubic box. Periodic boundary conditions were applied and the overall charge neutrality was ensured by the presence of four Na+ cations. The system was gradually heated up to 300 K. The temperature was next adjusted, along with the system density, in a 500-ps long MD simulation at constant temperature (300K) and pressure (1 bar). The coupling times to the Nose-Hoover thermostat [106, 65] and Berendsen barostat [14] were 0.2 ps and 0.5 ps, respectively. After equilibration, the barostat was removed and the system dynamics was followed in the NVT ensemble with a cubic simulation box of side l =8.35 nm for 52-ns. The dynamics was integrated with the GROMACS software (version 3.3.1)[125] with an integration time-step of 1fs. Constraints on bond lengths were enforced with the Lincs algorithm[58] and water internal degrees of freedom were controlled with the Settle algorithms[99]. Long-range electrostatic interaction was treated with the particle mesh Ewald method (PME)[34, 39]. The initial 2ns of each trajectory were not considered for analysis, which was instead performed on the subsequent 50-ns long production runs. The sampling time for the structural data (atomic coordinates of the enzyme and water) was equal to 0.5 ps for a total of 10^5 frames.

In the following we will present a detailed analysis of the MD results for AKE. We shall primarily focus on the 50-ns long simulation started from the crystallographic structure (PDB:4ake) of the open free enzyme. The resulting salient properties will be compared with the second simulation started from a closed, again ligand-free, configuration (prepared starting from the structure in PDB:1ake).

3.2 Structural fluctuations and a phenomenological model for mechanical strain

The recorded trajectories were first analyzed to assess the level of overall conformational heterogeneity encountered during the time evolution. The structural differences between the two starting crystal structures reflect the different orientation of the Lid and AMP-binding subdomains (corresponding to residues 114 to 164 and 31 to 60, respectively). The remainder Core region, consisting of 133 amino acids, presents minor differences in the two crystal structures (1.61 Å RMSD). The RMSD of the full C_α trace of 1ake and 4ake is 8.14 Å (see Fig 3.1).

The overall mobility of individual amino acids in each trajectory was characterized by means of the root mean square fluctuation (RMSF) profile of their α-carbon atoms. The RMSF of the ith C_α, whose instantaneous co-

ordinate at time t is indicated by $\vec{r}_i(t)$, is given by $\langle|\vec{x}_i|^2\rangle$ where the brackets denote the time average and $\vec{x}_i(t) \equiv \vec{r}_i(t) - \langle\vec{r}_i\rangle$ is the instantaneous displacement from its time-averaged (reference) position. The average was taken after removing the rigid-body motions of the enzyme. Following ref. [57] each recorded frame was oriented so to align the rigid core (for definition of the domains see caption of Fig. 3.1) against the core of the open crystal structure.

Fig.3.2 shows the RMSF of each amino acid calculated for the entire 50ns-long open trajectory after removing the rigid-body motions of the Core region. The structural deviations are accumulated in correspondence of the Lid and AMP-binding regions. The core is, by converse, very stable as its amino acids have root-mean-square fluctuations (RMSFs) of less than 2 Å. The rigidity of this region is consistent with NMR and Xray studies, as well as with previous topology-based characterizations of the protein's elasticity [100, 101, 91, 127, 30]. Analogous results emerge from the analysis of the fluctuations in the closed simulation.

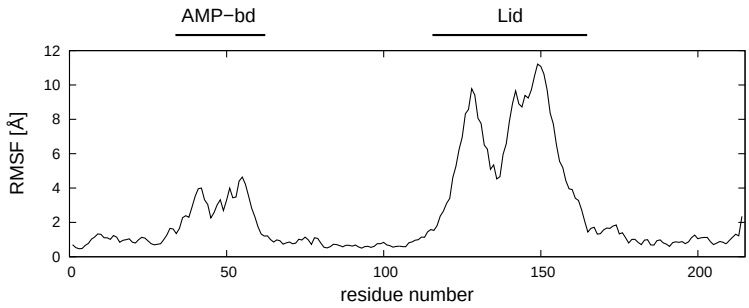


Figure 3.2: Root mean square fluctuations of the C_α atoms observed in the 50-ns long "open" trajectory. The fluctuations have been calculated after an optimal structural superposition of the C_α trace of the Core region of the enzyme.

The rearrangements experienced by the two mobile subdomains are aptly summarized by the time evolution of two independent geometric parameters which discriminate between the open and closed configurations of the Lid and AMP-binding regions, respectively. The degree of bending of the Lid towards the core was measured by the angle formed by two virtual bonds connecting the C_α's of amino acids (152,162) and (162,173), see Fig. 3.3a, and its time evolution is shown in Fig. 3.3c. The arrangement of the AMP-binding subdomain relative to the core was captured by the distance between the C_α's of amino acids 55 and 169, as illustrated in Fig. 3.3b. The time evolution of this second parameter, which was previously considered also in FRET experiments and computational studies for AKE [117, 90], is shown in panel d of Fig. 3.3. By comparison with the initial structure, during

the second half of the simulation, the Lid is bent towards the core and the Lid-Core geometric parameter frequently takes on values that are compatible with the closed (holo) form (dashed reference line in panel c). The AMP-bd–Core distance, instead, fluctuates within a fairly constant range throughout the trajectory, see panel d.

Consistently with previous reports on the approximately-independent motion of the two subdomains on the ns-scale [57], also from the present analysis no significant correlation emerges among the two time evolutions of panels c and d of Fig. 3.3. In fact, the Kendall correlation coefficient of the data set constituted by 100 pairs at equal times of the two geometric parameters (i.e. sampled at 0.5 ns) was equal to τ =0.065. The probability to observe a Kendall's τ having modulus smaller or equal to 0.065 in random sets of 100 elements is equal to 67% [112].

To characterize with finer detail the structural fluctuations of AKE we investigated how extensive, as a function of time, are the changes to the local structural environment of each amino acid (represented by the C_α atoms). In particular we quantify the changes to the set of distances between one amino acid and the one in close proximity (within 7.5 Å). The distortions of the contact network of the ith amino acid is quantified with the "geometric strain" parameter, q_i, providing a measure of how much its instantaneous distances with neighboring amino acids differ with respect to the time-average:

$$q_i(t) = \sum_j f(\langle d_{ij} \rangle) \; (d_{ij}(t) - \langle d_{ij} \rangle)^2 \tag{3.1}$$

where d_{ij} is the distance of the C_α atoms of amino acids i and j and $f = (d) = \frac{1}{2}(1 - \tanh(d - d_{cut}))$ is a sigmoidal function weighting the average spatial proximity of the two amino acids. Its point of inflection is set at the cutoff distance $d_c = 7.5$Å.

By analyzing the time evolution of the geometric strain profile it is possible to identify those regions of the enzyme that undergo a rigid-like motion. The geometrical distances for any pair of amino acids within such regions would be highly conserved in time, regardless of the amplitude of the motion of the region with respect to a fixed reference frame. Consequently, by cross-referencing the RMSF and the geometric strain analysis it is possible to identify *a posteriori* the amino acids (if any) that act as hinges for the articulated motion for AKE. The time evolution of the geometric strain profile q_i is shown in the bottom panel of Fig 3.4, along with the profile of the cumulative strain of the full protein chain (top panel of the same figure).

Fig. 3.4 illustrates two notable features of adenylate kinase dynamics which are hereafter discussed in detail. First, the geometric strain is mostly concentrated on specific regions of the protein chain and, secondly, the patterns of geometric strain evolve discontinuously in time.

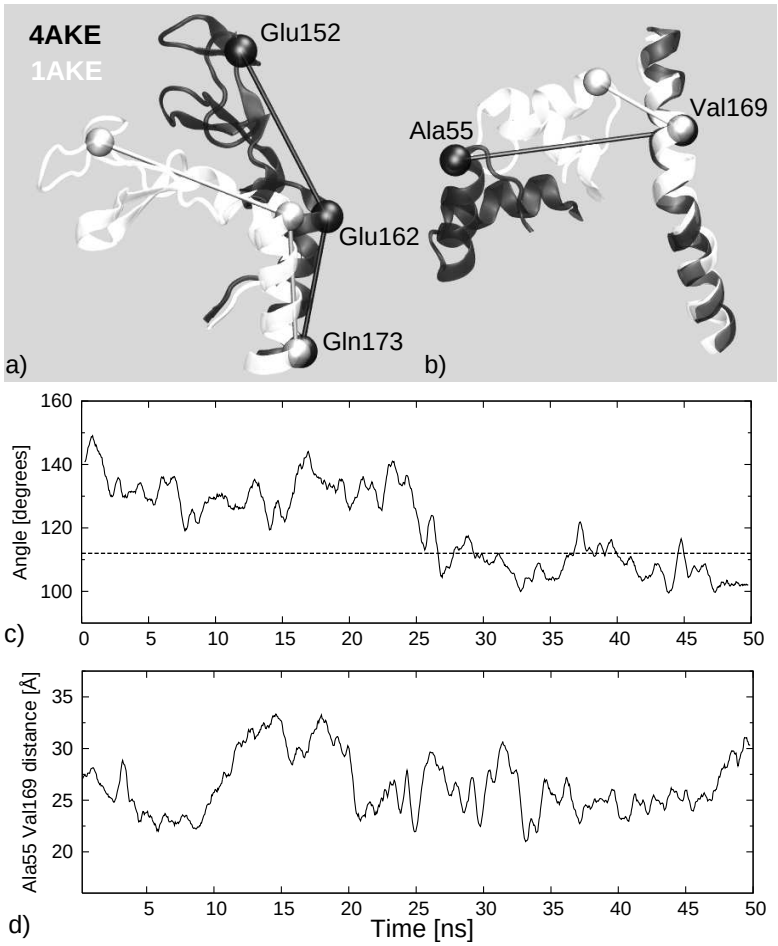


Figure 3.3: To discriminate between open (black) and closed (white) conformations of the Lid and AMP-binding subdomains two independent geometric parameters are introduced. (a) The Lid orientation is captured by the angle between the virtual bonds formed by the C_α's of amino acids 152,162,173. (b) The AMP-bd arrangement is captured by the distance between C_α's of Ala55 and Val169, which attains the value of 12 Å for the closed conformation. The time series of the Lid and AMP-bd geometric parameters during the open trajectory are shown in panels (c) and (d), respectively. Open [closed] conformations of the AMP-bd and Lid subdomains are associated to values of the parameters greater [smaller] than 12 Å and 116 degrees (dashed line in panel c), respectively. The lack of correlation of the time series in panels (c) and (d) suggest the independent motion of the two subdomains.

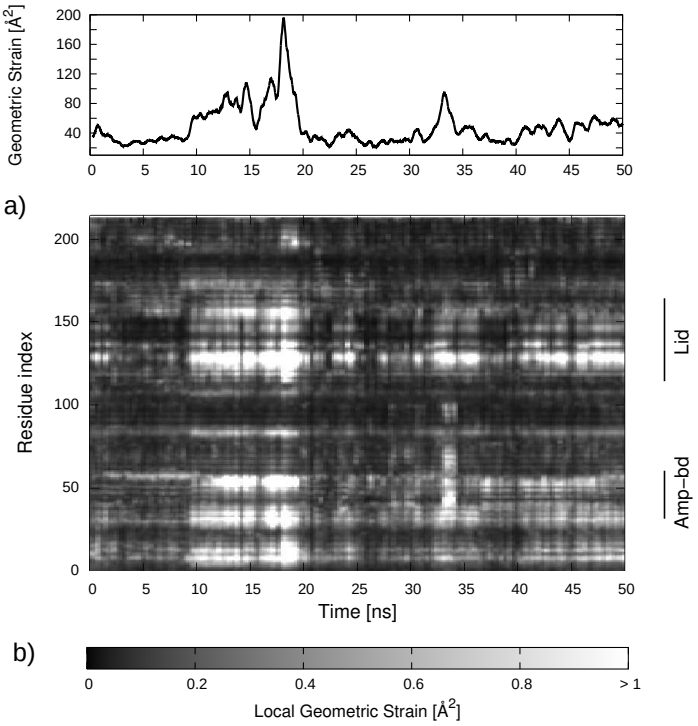


Figure 3.4: (a) Time evolution of the geometric strain (see eqn.3.1) summed over all amino acids. The strain of each individual amino acid is shown in panel b.

Six sets of amino acids,labeled a–f, are associates with significant strain: namely group (a): amino acids 8 to 15 ; (b): 28 to 35; (c): 53 to 60 ; (d): 123 to 130 ; (e): 135 to 137 ; (f): 153 to 158. In particular, group (d) correspond to the Lid-Core interface, while sets (b) and (c) to the AMP-Core one; these groups of residues hence act as primary hinges for the motion of the two mobile subdomains. It is worth noticing that a further region of high geometric strain (group e) is found in the middle of the Lid subdomain, indicating an articulated motion of the latter around this joint.

We now turn to the observation that the build-up/release of geometric deformations of these regions is discontinuous in time. For example, at t=9 ns there is a rapid increase of the geometric strain in correspondence of all above-mentioned groups which persists up to t=19 ns. At this time another coordinated change of these regions is observed.

These facts indicate that the system evolution proceeds by visiting distinct conformational substates through which the systems hops with rapid "transitions" signaled by discontinuities in the geometric strain profiles.

3.3 Structural clustering

This conclusion is supported by the analysis of the density plot in Fig 3.5a which provides the RMSD between each pair of conformations sampled from the open trajectory. The block character of the matrix suggests that distinct conformational groups are explored during the dynamical evolution. Consistently with this qualitative observation, the analysis of the distribution of the pairwise distances [96] suggests that the system populates conformational basins where the internal structural heterogeneity is about 2.5 Å RMSD while the RMSD of conformations in different basins is mostly in the range [4-7 Å] (see Fig 3.6). As a quantitative method to identify the conformational basins visited by the trajectory we have applied and compared results of two structural clustering schemes.

First the standard K-medoids clustering scheme [74] was used to partition each trajectory in structurally-homogeneous groups. The input of this algorithm consists of the pairwise RMSD distances between all pairs of the 10^3 recorded structures (one every 50 ps). The returned output consists of the grouping of the structures in a pre-assigned number of non-empty clusters, K. A representative conformation for each cluster is also provided. The clusters and their representatives are identified by minimizing the dissimilarity score obtained by summing the RMSD of each structure from its cluster representative. The method is commonly implemented in an iterative fashion through the following steps: (i) the members of the K clusters are first assigned randomly; (ii) the cluster representatives are next identified by picking in each cluster the element with smallest total distance from the other cluster members; (iii) the clusters are finally redefined by assigning each data-set member to the closest representative. Steps (ii) and (iii) are repeated until the dissimilarity score does not decrease anymore. To avoid trapping in local minima of the dissimilarity score, the method is repeatedly applied for several initial random groupings.

We emphasize that the clustering returned by standard K-medoids scheme described above is based solely on the input of the RMSD distance of any pair of structures (aligned over the core region) and hence does not consider their succession in time along the trajectory.

The clustering was performed by varying K from 2 to 15. Values of K larger than 6 resulted in a noticeable "intermittent" assignment to different clusters of structures contained in time intervals smaller than 0.5 ns. This effect was taken as indicating an excessively fine subdivision of the conformational substates. For values of $K \leq 6$, instead, each cluster comprised structures covering, with only sporadic outliers, continuous time intervals of duration not smaller than 2 ns.

Accounting for the time order of the structures is essential for partitioning the recorded trajectories in a succession of progressively-visited substates. The K-medoids scheme was accordingly modified to ensure that each

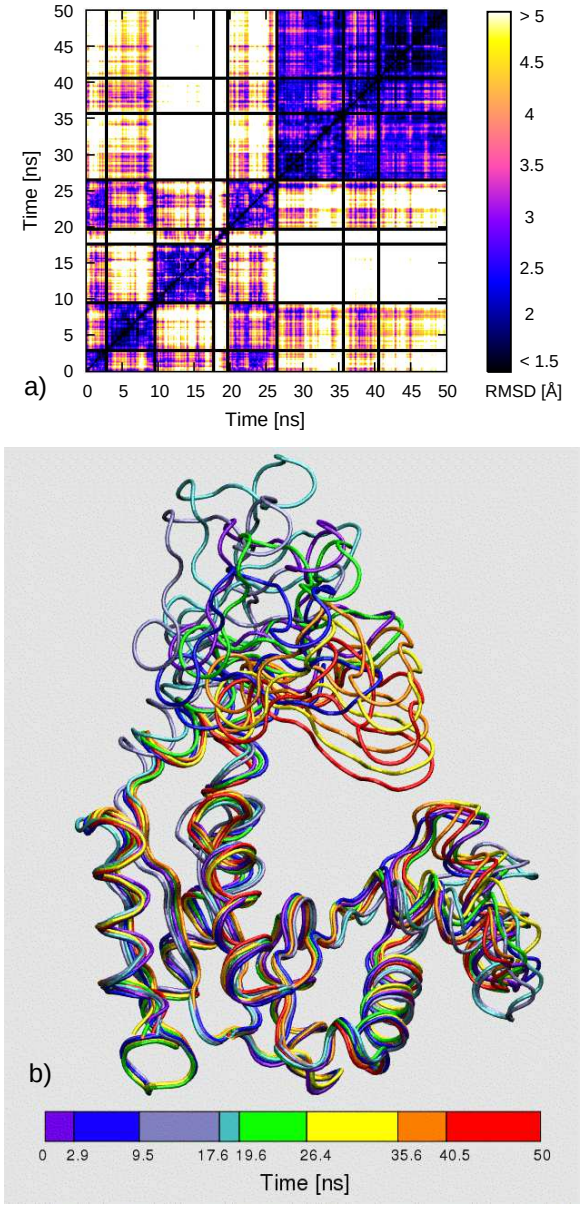


Figure 3.5: (a) Density plot of the pairwise RMSD of 1000 time-equispaced conformations from the open trajectory (time labels are shown on both axes). Solid black lines are used to separate the substates identified with the structural clustering procedure. The structures representing the substates are shown in panel (b) and are colored according to the time subdivision (clustering) shown at the bottom.

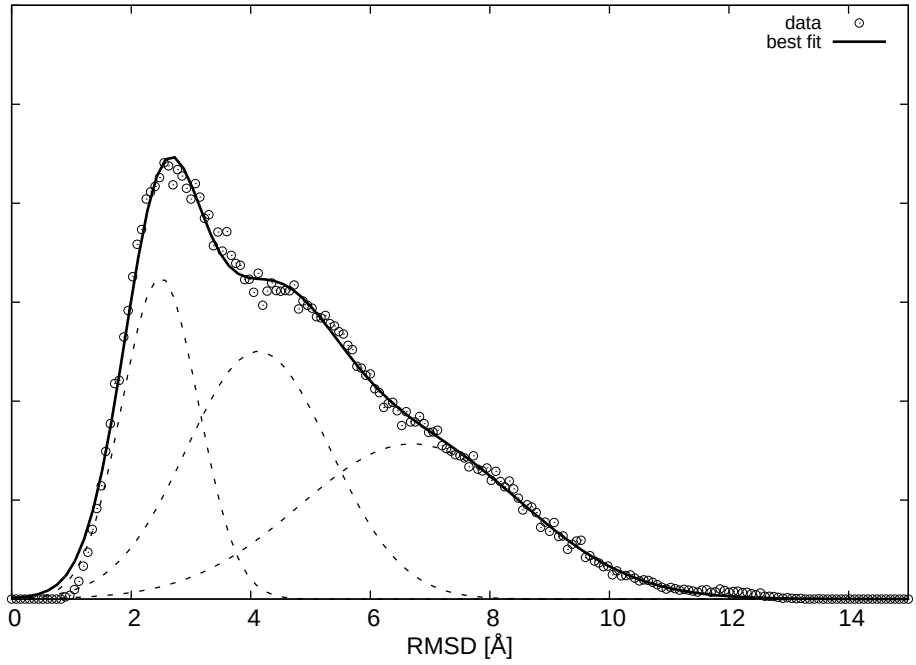


Figure 3.6: Distribution of RMSD distances of each pair of conformers of the simulation started from the open crystallographic structure (circles). The data are well fitted by a sum of three gaussian functions (solid line). The three gaussian are shown as dashed lines. The lowest peak ($\sim$ 2.5 Å) provides a measure of the tipical intra-basin distance of the structures, while the broader background distribution is a measure of the distances between structures belonging to distinct substates.

cluster gathered structures spanning an uninterrupted time interval of the simulation. The introduction of the "time continuity" constraint simplifies the definition of the cluster members, which are unambiguously specified by introducing $K-1$ time subdivisions of the trajectory.

The minimization of the dissimilarity score subject to the "time continuity" constraint is performed within a greedy stochastic minimization scheme. Given the $K-1$ time-subdivisions (initially equispaced), the representative of each cluster is defined as in step (ii) of the standard K-medoids scheme and the resulting dissimilarity score is computed. At variance with step (iii) of the original method, a new clustering is *proposed* by randomly re-assigning one or more of the $K-1$ subdivisions and ensuring that no two subdivisions coincide as empty-clusters would result. The new cluster representatives are found and the new dissimilarity score is calculated. The proposed clustering is kept if it leads to a decrease of the dissimilarity score, otherwise the previous one is retained and a new partitioning is pro-

posed. The procedure is iterated until convergence of the dissimilarity score (10^5 iterations were typically sufficient to reach convergence for partitioning 1000 structures in K=10 clusters, requiring a few minutes of computation on present-day personal computers.)

The partitioning obtained with the two methods have been compared (see Fig 3.7), and a good consistency was obtained using $K = 8$ non-overlapping intervals. The emerging consensus time subdivision, shown in Fig.3.5a, was consequently used to identify the most prominent conformational substates explored by the trajectory.

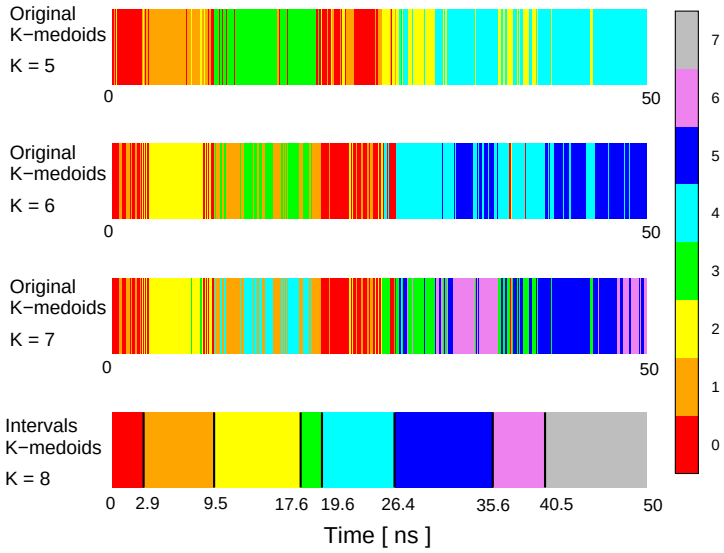


Figure 3.7: Partitioning of the trajectory started from the open conformation in different conformational basins (frames belonging to different clusters are represented as lines of different colors). The grouping has been performed with the standard K-medoids algorithm with several values of K (here we show only results for K=5,6,7), and with a modified version of the algorithm that requires the elements of each group to span an uninterrupted time interval (lower panel).

The typical RMSD of structures belonging to the same cluster was equal to 1.9 Å while structures belonging to different substates differed from 3 Å up to 12 Å RMSD (after alignment of the core region). The representatives of the clusters are shown in Fig.3.5b. The figure conveys the large variability of conformations encountered; nonetheless the average structure of the whole trajectory is at only 2.2 Å RMSD from the starting crystallographic conformation.

The trajectory started from the closed structure presented qualitative parallels with the above described results. In particular, as can be perceived from Fig 3.8, also the "closed trajectory" visits about 8 substates, characterized by residence times of the order of 5-10 ns. However, due to its more compact arrangement, the structural fluctuations of the closed enzyme are smaller than for the open form. This reverberates in a smaller global mean square fluctuation (i.e. summed over all C_α's), which is 2838 and 908 $Å^2$ for the open and closed trajectory, respectively. Also the RMSD distances between the substate representatives are smaller, typically about 3 Å.

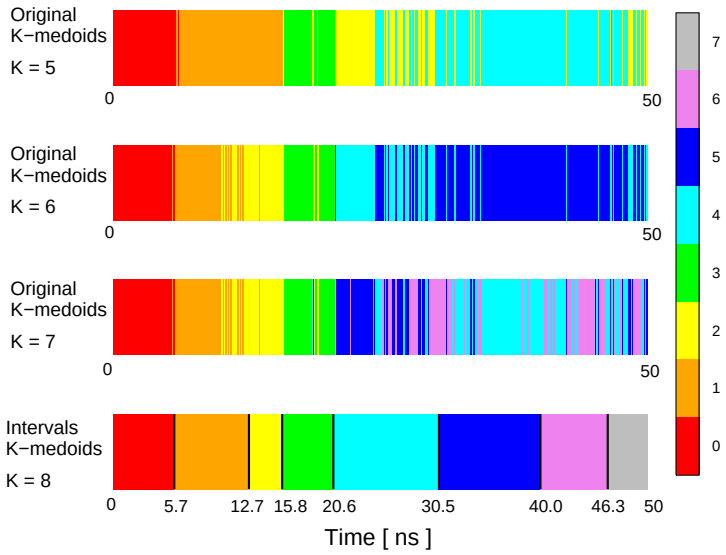


Figure 3.8: Partitioning of the trajectory started from the closed conformation in different conformational basins (frames belonging to different clusters are represented as lines of different colors). The grouping has been performed with the standard K-medoids algorithm with several values of K (her we show only results for K=5,6,7), and with a modified version of the algorithm that requires the element of each group to span an uninterrupted time interval (lower panel).

Though the ensembles of conformations explored by the two simulations do not strictly overlap, it is noteworthy to notice that the RMSD between pairs of structures in the two trajectories can be as low as 2.5 Å.

3.4 Intra- and Inter-Substate Fluctuations

Recalling the results obtained for GB1 domain of protein G, we wish to analyze in details the structure of the various conformational substates.

We will in particular first assess the relative extent to which structural fluctuations within the substates and across them impact on the breadth of visited conformational space.

To this purpose we first analyzed the intra- and inter-substates contributions to the global mean square fluctuation (MSF) of the molecule (i.e. the sum of the mean squared displacements of each C_α).

The decomposition of eqn. 2.11 of the covariance matrix in contribution arising from the structural fluctuations within the substates and across them can be rewritten as

$$C_{i,j} = C_{ij}^{intra} + C_{ij}^{inter} \tag{3.2}$$

$$C_{ij}^{intra} = \sum_l \omega_l C_{i,j}^l \tag{3.3}$$

$$C_{ij}^{inter} = \sum_l \omega_l [\langle r_i \rangle_l - \langle r_i \rangle][\langle r_j \rangle_l - \langle r_j \rangle] \tag{3.4}$$

where, analogously to 2.11, l is an index referring to the substates, w_l is the weight of the lth substate, that is the fraction of simulation time spent by the system in it, $\langle\rangle_l$ denotes the average taken over the conformations of the lth substate, and $C_{ij}^l = \langle [r_i - \langle r_i \rangle_l][r_j - \langle r_j \rangle_l] \rangle_l$ is the covariance matrix of the lth substate itself.

It may be anticipated that the relative interplay of the intra- and inter-substates fluctuations depends on the duration of the simulations (which e.g. affects the number of visited substates). We have accordingly computed the intra- and inter-substates contributions to the MSF, for increasing duration of the trajectory, that is at each of the time subdivisions indicated at the bottom of panel b in Fig.3.5 . The global MSF calculated at all stages of the trajectory is also reported in Fig. 3.9a.

Over the 50-ns long trajectory, the fraction of global MSF accounted for by the 7 inter-substates hops is 70 %. The result is striking as the inter-substates contribution is computed merely on the basis of the eight structures which represent the visited substates and their representation weight (i.e. the time-intervals duration).

Finally, besides indicating that the jumps across substates represent a key aspect of the equilibrium dynamics of the system, the increasing trend of the global MSF in Fig. 3.9 indicates that the progressive broadening of the visited configuration space is still ongoing after 50 ns. This aspect is consistent with the experimental indication that an exhaustive exploration of the available structural space of the apo form of AKE occurs over time scales that largely exceed the one covered by the simulation [128, 55, 57].

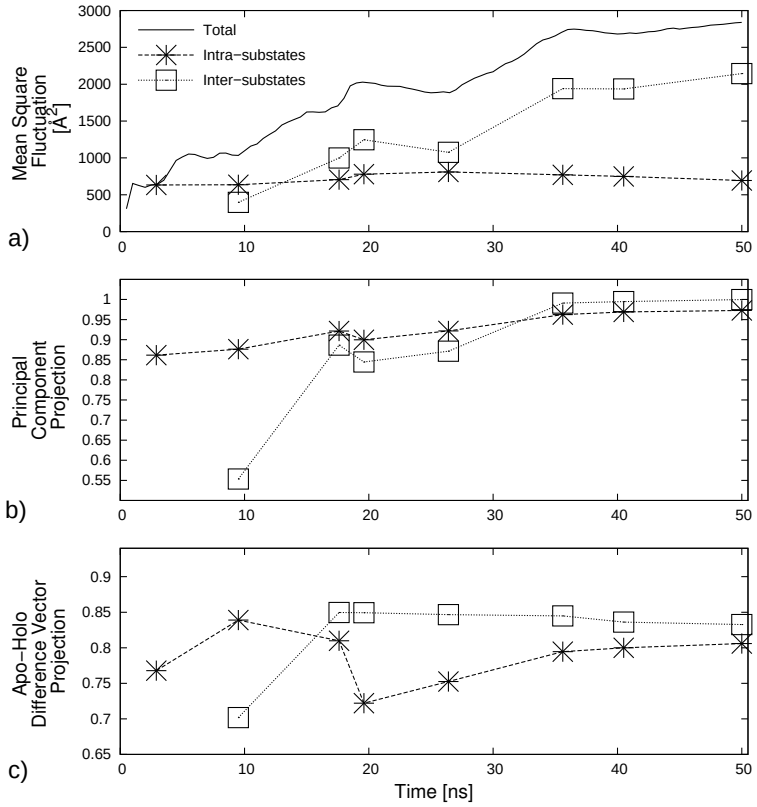


Figure 3.9: (a) Time evolution of the total mean square fluctuation of the enzyme and of its intra- and inter-substates contributions. The other panels portray the evolution of the scalar product (in modulus) between the first intra- and inter-substates essential modes and: the principal mode of the entire trajectory (panel b) and the normalised apo/holo difference vector (panel c).

3.5 Robustness of the Lowest Energy Modes

The intra/inter substate decomposition discussed above stimulates the investigation of which relationship, if any, exist among the generalized coordinates which correspond to the essential dynamical spaces calculated separately for each of the 8 substates as well as the different vectors between the representative structures of the clusters. In the analysis of GB1 it was found that a common limited set of directions can be used to describe both the principal components within the various substates separately, and the direction of the "jumps" among them. Being able to extend similar considerations also in the case of Adenylate Kinase, would be extremely appealing in that it would shed light on the functional-oriented character of the internal dynamics. To

address this issue we have first calculated the RMSIP between the essential dynamical spaces of the 28 distinct possible pairs of substates. The average RMSIP was equal to 0.83 with a dispersion of 0.03. The value indicates a very high degree of consistency of the fluctuations within the various clusters. Such consistency remarkably extends also to the inter-cluster structural fluctuations. A stringent verification of this point is given in Fig. 3.9b which shows, as a function of time, the modulus of the scalar product between the first principal component of the intra- and inter-substates matrix, and the first principal component of the covariance of the entire trajectory. Despite considering only *one* space, the principal direction of the intra-cluster covariance matrix computed with as few as 2 substates is already well aligned with the one of the total trajectory (scalar product equal to about 0.9). The quality of the accord does not deteriorate as more and more substates are visited.

The consistency of the directionality of the structural fluctuations within and across the substates appears remarkable in consideration of the increasing breadth and diversity of the visited conformational space (see Fig. 3.9a).

The above considerations further prompt the conclusion that a limited set of collective coordinates, indicated by the consensus of the inter/intra-cluster essential dynamics, would be adequate to describe the salient conformational fluctuations over a range of time scale wide enough to capture the (sub-ns) dynamics within substates and the transitions across them (occurring at the multi-ns level).

This expectation is verified and illustrated in Fig. 3.10 which shows the highly consistent RMSIP between the essential dynamical spaces of pairs of intervals of 0.5 ns or 5ns from the trajectory (that is with time-subdivisions covering a wide temporal range and unrelated to the substates partitioning).

We have devised a test to establish the extent to which the RMSIP of two spaces is likely to have arisen by a mere consistency of "unspecific" dynamical features, such as the overall mobility of amino acids. In fact, to a certain extent, the mean square fluctuation of amino acids in a protein correlates with the local density (the higher the density the lower the mobility) [53]. On the basis of this observation it could be anticipated that certain regions of AKE are more or less mobile than others. This poses the question of whether a given RMSIP value simply reflects the consistency of the salient aspects of the local density profile of two intervals of the trajectory.

A specific test was devised for addressing this issue. It consists of computing the distribution of RMSIP values that arise when the essential dynamical spaces of $\{v\}$ and $\{w\}$ are modified so to (i) preserve the normalized mean-square fluctuation profiles of each mode while (ii) retaining the orthonormal relationships within the new sets $\{v'\}$ and $\{w'\}$ and (iii) ensuring the orthogonality of $\{v'\}$ and $\{w'\}$ with the zero-energy modes associated to translations and rotations of the system.

The algorithm used for this purpose is described hereafter. For each

amino acid we performed a random reorientation of its three-dimensional displacement appearing in all modes of each set (i.e. the randomization is carried separately for $\{v\}$ and $\{w\}$). It is important to stress that the rotation differs from site to site but the same rotation is applied to the displacements of one particular site (amino acid) in all 10 modes.

This random reorientation procedure realizes requirements (i) and (ii). In fact, it preserves the normalized mean-square fluctuation profiles of each mode and the randomized modes are still orthonormal. However, the new modes, in general, will have non-zero overlap with the six zero-energy modes associated to the rigid-body motions of the system. From a practical point of view, the orthogonality condition can be enforced in an approximate way by retaining, out of a large number of randomly-generated sets of modes the ones where each mode had a projection smaller than 0.05 on the six-dimensional linear space of the zero-energy modes.

In this way, the RMSIP value of two original sets of modes, $\{v\}$ and $\{w\}$, can be compared against the distribution of RMSIP values of randomized sets that, mode by mode, still possess the same RMSF profile. This reference distribution therefore provides an indication of how much the mere specification of the normalized RMSF profiles of all the modes, constrains the possible RMSIP values.

In fact, the distribution of RMSIP values that would follow from specifying only the MSF profiles of each mode is provided in Fig.3.10 and covers a region of values much lower that those observed here, confirming the statistical significance of the observed RMSIP values.

3.6 Functional Oriented Character of Low Energy Modes

A naturally emerging question is whether the observed degree of consistency is functionally oriented, i.e. related to the prominent structural rearrangement between the open and closed forms of the enzyme. To establish this property, and in analogy with the analysis of Fig.3.9b, we have computed, as a function of time, the fraction of the norm of the difference vector between the apo-holo crystal structures projected onto the first essential eigenvector of the intra and inter-substate covariance matrices. The results are plotted in Fig.3.9c and indicate that the fraction of captured norm is about 0.8 throughout the trajectory, supporting the functional relevance of the molecules' internal dynamics.

Fig.3.9c reveals the interesting aspect that the inter-substate hops are typically better aligned along the apo/holo conformational changes than the intra cluster principal directions. Over the entire 50ns-long trajectory, 82% of the open/closed conformational change is already captured by the the first low-energy mode while considering the top ten modes captures 96%

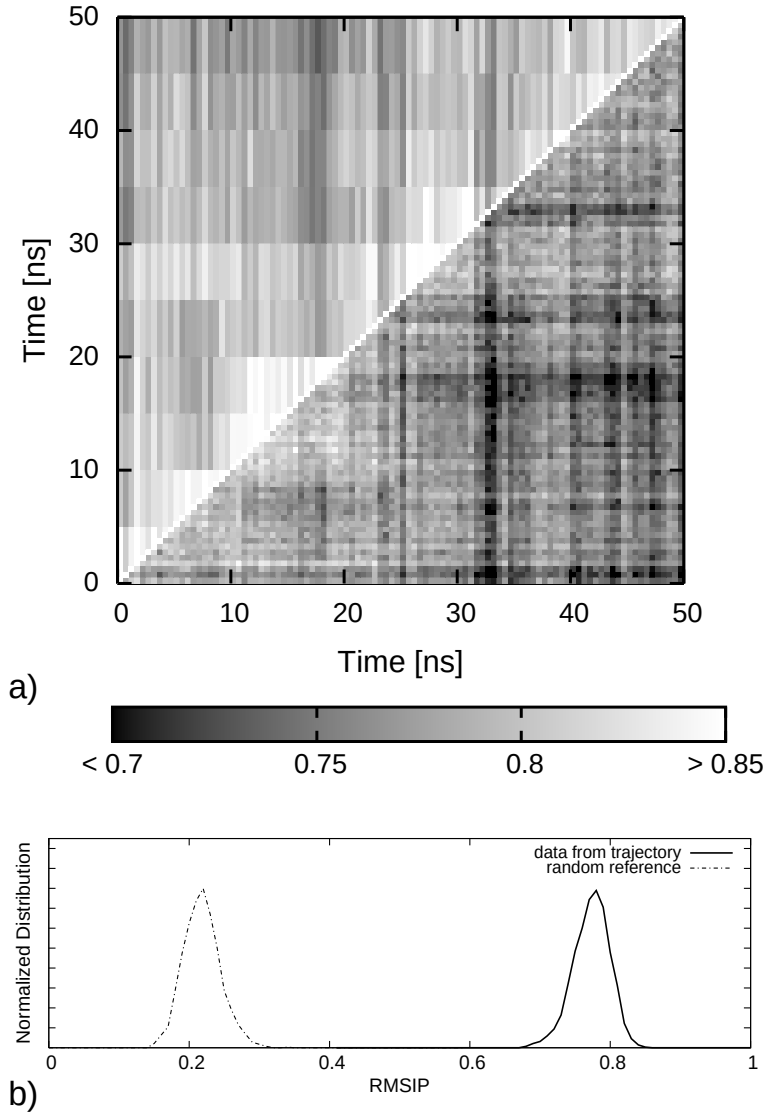


Figure 3.10: (a) Density plot of the RMSIP between essential dynamical spaces (EDS) of 0.5-ns long intervals of the open trajectory (time labels on the abscissas) against themselves, lower triangle, and against the EDS of 5ns-long intervals, upper triangle. (b) The distribution of RMSIP values from the lower triangle of the matrix (pairs of 0.5ns-long intervals) is shown with a solid curve. The dashed curve shows the distribution of RMSIP values obtained by randomizing (in a way which preserves the mobility profiles of the amino acids) the original EDS.

of the norm. The fact that the conformational changes within and across the substates occur mostly along the direction bridging the open and closed conformers is illustrated in the scatter plots of Fig. 3.11. In panel a the substates visited by the open trajectory are represented, with different colors, in the space of the three lowest-energy modes. The trajectory started from the closed structure is also shown for comparison. For clarity, two bidimensional projections of the scatter plot are provided in panel b. The discrete character of the clouds associated to each substates is readily perceivable, as wells as their preferential elongation along the apo/holo difference vectors. This anisotropy is particularly apparent for the trajectory as a whole.

It is important to notice that, despite the good orientation of the principal dynamical directions along the apo/holo change, the succession of visited substates does not proceed in a directed manner in that no constant progression for e.g. the open to the closed conformation is seen. As a consequence of the non-directed character of the dynamics it is expected that the full interconversion occurs over much longer timescales than those accessed here, consistently with experimental indications [55].

3.7 Consensus Dynamical Space

As a final stage of our analysis we proceeded to identifying the consensus set of collective modes that best capture the common structural fluctuations of AKE encountered in the two 50ns-long trajectories. The essential dynamics analysis applied to the two merged trajectories is not adequate to this purpose as it is not designed to extract the dynamical features that are shared by the two separate trajectories.

Expression (5.4) provides an *average* measure of accord of two essential dynamical spaces, as the top 10 eigenvectors of C are treated on equal footing (degeneracy). This implies that the same value of RMSIP may be attained with different detailed levels of accord of two spaces.

To characterize with a finer resolution the consistency of two sets of modes we introduce a variational scheme that identifies their maximally-consistent (or inconsistent) subspaces. The scheme, explained in detail in the Appendix A, is used to redefine two new bases $\{v'\} \equiv \{\vec{v}'_1, \vec{v}'_2, ..., \vec{v}'_{10}\}$ and $\{w'\} \equiv \{\vec{w}'_1, \vec{w}'_2, ..., \vec{w}'_{10}\}$ for the *same* linear spaces described by v and w. The redefined bases, $\{v'\}$ and $\{w'\}$, possess two notable properties: (i) a basis vector of one set is orthogonal to all basis elements of the other set except the one with the same index and (ii) the index provides a natural ordering of the basis vectors in terms of decreasing mutual consistency. Notice that the RMSIP of the new basis vectors is the same of the original one.

The method provides an optimal redefinition of the basis vectors in the two sets of modes which are returned in order of decreasing mutual consistency. We stress that the new bases span the same linear spaces of the

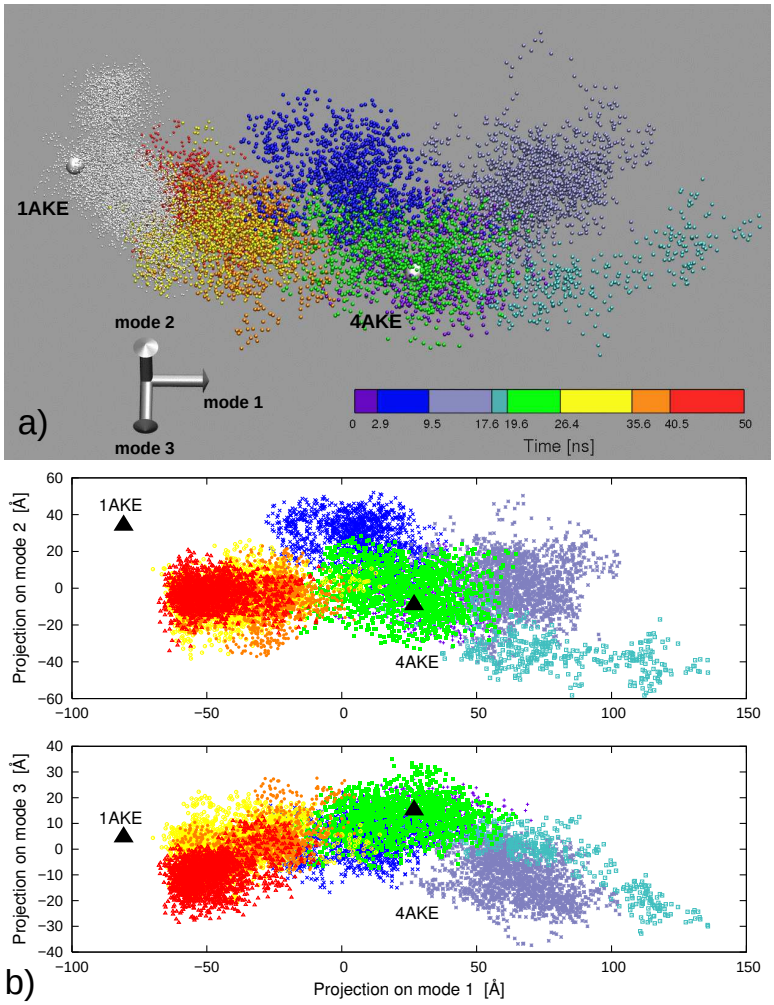


Figure 3.11: (a) Representation of the "open" and "closed" dynamical trajectories in the space of the three lowest-energy modes of the open trajectory. Each point represents an MD configuration and its co-ordinates are obtained by projecting the instantaneous deformation vector from the average structure of the open trajectory onto the three modes. White points are used to represent the closed trajectory, while different colors are used to distinguish the substates of the "open" trajectory. The open and closed crystal structures are represented as large spheres. Two bi-dimensional projections (mode 1 vs. mode 2 and mode 1 vs. mode 3) of the three-dimensional scatter plot are provided in panel (b)

original sets so that the original RMSIP, equal to 0.786, is unaltered by the redefinition.

It was found that the 10 lowest-energy modes of the two trajectories

share, with almost perfect overlap, a three-dimensional subspace. In fact, the scalar products of the fist, second and third pair of redefined modes have scalar products equal to 0.972, 0.951 and 0.925, respectively.

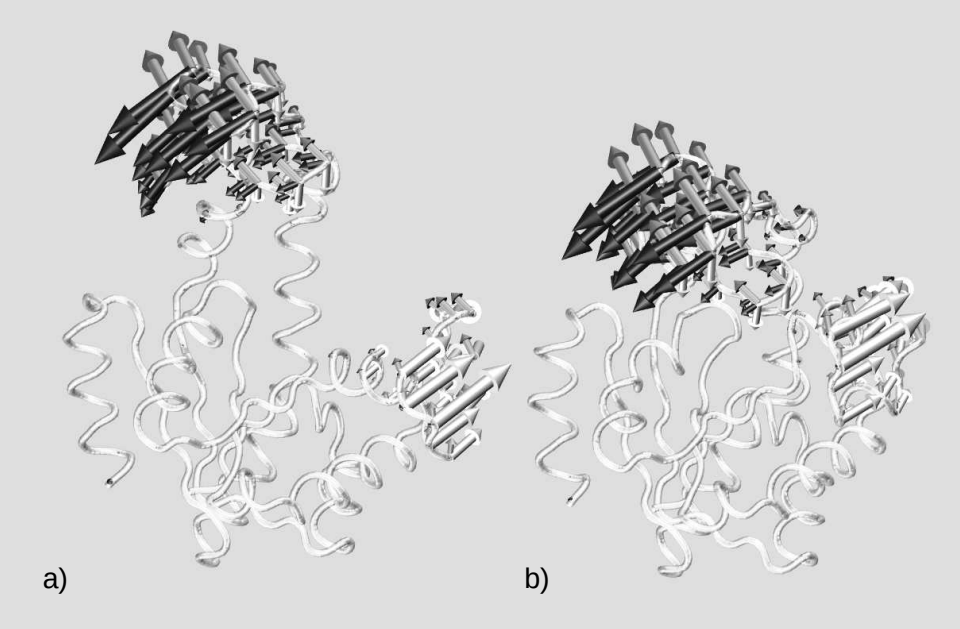


Figure 3.12: The three consensus low-energy modes of the open and closed trajectories are shown respectively with black, gray and white arrows superposed to the average structure of the (a) open and (b) closed trajectories.

For each trajectory, this three-dimensional consensus space is sufficient to account for more than (i) 57% of the total mean square fluctuations, (ii) 50 % of the intra-substate MSF and (iii) 60 % of the inter-substate MSF. It also captures (iv) 77 % of the norm of the apo/holo difference vector.

The consensus modes thus constitute an extremely limited set of generalized coordinates that account for the system internal dynamics over a wide range of time scales (encompassing both intra and inter-substates fluctuations) and indicates their relatedness to the major functional conformational change between the open and closed structures.

From a practical point of view the findings also suggest the use of the consensus collective modes as natural candidates for profiling the free energy of the system in terms of a reduced number of generalized variables.

3.8 Summary

The predisposition of adenylate kinase to undergo major, functionally-oriented, conformational changes was investigated through extensive molecular dynamics simulations of the free enzyme. Available crystal structures of AKE were taken as starting points for two MD simulations covering a total time-span of 100ns. The analysis of the data collected over this previously-unaccessed MD time-scale has exposed interesting functionally-oriented characteristics of the internal dynamics of the enzyme and of the organization of its free energy landscape.

During the free dynamical evolution, the enzyme populates distinct conformational substates with residence times of 5-10 ns. The ensemble of different conformers is structurally heterogeneous (with inter-substates differences up to 12 Å RMSD), reflecting the pronounced mobility of the AMP-binding and Lid subdomains.

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

The analysis indicates that the free enzyme can be driven through various conformational substates bridging the inactive and catalytically potent states through the thermal excitation of a limited number of collective modes. These results extend our consideration on GB1, elucidating a functionally oriented nature of the self-similar organization of the free energy landscape.

This observation supports recent suggestion of Aden *et al.*[1] that functionally-oriented conformational fluctuations are innate properties of the free (apo) Adk. In fact, the consistency of the salient features of the enzyme's internal dynamics within and across substates leads to speculate about the fact that these property may have been promoted by evolutionary pressure.

Part II

Comparing Proteins' Internal Dynamics

As anticipated, for an increasing number of proteins end enzymes it has been shown, by means of both experiments and computation, that the internal dynamics influences and assists the biological functionality. For the two specific proteins considered in the previous chapters, namely Immunoglobulin binding domain of protein G and Adenylate Kinase, it has been shown that they are endowed with an "innate" ability to sustain specific (and arguably functionally-oriented) large-scale movements. The most vivid indication of this dynamical predisposition is conveyed by the fact that a low-dimensional space of collective variables is sufficient to account for the essential spaces of the discrete visited substates and of the difference vectors between them. The innate functionally-oriented movements, which can nowadays be probed by single molecules techniques, are aptly captured also by simplified coarse-grained models (elastic networks) which are oblivious to the detailed chemical composition of the protein as they are based on a simplified representation of a protein's native state. This provides a direct indication that salient features of proteins' internal dynamics reflect fundamental properties of proteins' structural architecture such as the secondary and tertiary organization.

The elucidation of the link between proteins structure and their functional large scale movements can be viewed as a facet of the traditional tripartite organization of proteins' characterization in terms of: sequence $\rightarrow$ structure $\rightarrow$ function [18, 33, 41, 88]. A number of key questions concerning the second step of this ladder emerge naturally after considering the well-characterized connection between sequence and structure [49, 77, 85, 121, 110, 51]. Arguably, the best-known general result regards the fact that proteins whose similarity in chemical composition is above 30% adopt the same fold [6, 32, 107]. For such homologous proteins, the common amino acids (identified by sequence alignment), are expected to be highly superimposable by a suitable roto-translation of the molecules. The availability of structural alignment algorithms [62, 64] have added notable elements to the picture by demonstrating that the same fold can be adopted by proteins with unrelated chemical composition. Typically this is interpreted in terms of convergent evolution of proteins structure [29, 85, 5]. The observations prompt analogous questions concerning the connection between structure and functionally oriented dynamics. Namely: to what extent does structural relatedness reverberate in related concerted movements? Can analogous patterns of concerted movements be sustained by proteins with substantially different overall architecture?

These questions are at the heart of the ongoing effort to establish a general scheme for aligning proteins according to dynamics-based criteria ([26, 22, 25, 130]). The mentioned issues have provided the motivation for two investigations that fall within the scope of the present thesis and are the object of the following two chapters.

We have first considered the connection between structure and dynam-

ics in a superfamily of calcium binding proteins, called EF-hand proteins [105, 16]. Proteins belonging to this superfamily share a common structural motif, the so called EF-hand domain, composed by four helices arranged in a particular tertiary way. This is the most common motif for binding calcium and its internal structural rearrangement upon calcium binding is responsible for triggering the calcium signaling cascade [105]. Members of the superfamily, though having in common the EF-Hand domain, are spread across several different functional families and differ for the overall tertiary and quaternary organization . Notably, even the details of the internal arrangement of the local motif are subject to a certain degree of variability. These facts have stimulated a number of previous investigations, aimed at clarifying the connection between the structural variability of the domain and the functional role of the protein the domain belongs to [105, 129, 16, 9, 52]. However the presence of different functional families does not to reflect in a clear way in any naturally emerging groups where detailed characterization of the structural variability is attempted [9]. Nevertheless, the specific internal arrangement and flexibility of the EF-hand domains, which is crucial for biological functionality, is expected to be different across the various biological families [105, 21].

We have addressed these aspects by performing an analysis of the essential dynamics in the space of the generalized coordinates of the angles among the four helices constituting the EF-hand domains for a data-set of more than 150 domains.

The use of this set of generalized coordinates, common to all proteins of the EF-hand superfamily, and already used in structural classification studies [9], has allowed us to compare in a well-defined way their internal dynamics. We have investigated the relation between structural and dynamical similarity, and established a connection between the dynamical and functional grouping. Interesting correlations emerges between the essential dynamics of the domains and their functional family classification while only a loose relationship exists between the degree of structural and dynamical similarity. The non-simple connection between local structural and dynamical correspondence is presumably ascribable both to the detailed local structural differences of the domains and to the influence of the different global arrangements of the proteins.

This stimulated a more general question regarding how a structural correspondence among the regions of two proteins reverberates on the accord of their essential motions. Partial structural alignment methods allow to highlight significant local correspondences also in proteins having different global organization[61, 63, 64]. The identification of a common super-imposable core can be due either to sequence homology or to convergent evolution of the structures. In both cases the problem of how the evolutive selection of a specific part of sequence reverberates in specific structural arrangement of that part of the structure is not trivial as the folding process usually in-

volves the protein as a whole. Analogously, as the spatial correspondence highlighted by structural alignment pertains only to a common structural core (the remaining part of the two proteins can differ both for length and spatial arrangement) it is not obvious what degree of dynamical accord should be expected over the common core. In fact, the motion of these super-imposable core regions will be influenced also by the remaining part of the molecule. As described in the case of EF-proteins, this reflects in a non-simple connection between local structural similarity and dynamical consistency.

Besides the case of the EF-Hand domains, the above questions are investigated also for two members of the protease superfamily[11], carboxypeptidase A[76] and pyroglutamyl peptidase[120]. Though the overall architecture of the two biomolecules is analogous, the differences in length and number of secondary elements is sufficiently large to require the use of a non-subjective and quantitative scheme to detect and quantify the dynamical correspondences. Moreover the comparison of their large scale dynamics is of particular interest given that the two enzymes rely on different catalytic chemistries and belong to different protease clans[11]. Hence, the possibility to establish a consistency in their large scale fluctuation dynamics provides valuable insight into general aspects of the enzymes's structural modulations that accompany the proteolysis reaction.

Chapter 4

EF-Hand Superfamily - A Playground for Dynamics-Based Comparison

In this chapter we shall report on the internal dynamics of members of a superfamily of proteins primarily involved in the calcium signaling pathways. Calcium is known to be a universal regulator of many cellular processes, from duplication to apoptosis [40, 23, 24]. It interacts with a large number of proteins having different biological function. Yet the repertoire of configurations of calcium binding sites is limited [105, 16]. In fact calcium's regulatory role relies on the transduction of the signals carried out by a number of calcium binding proteins. Most of the latter share a common building block, the EF-Hand motif, composed by two helices connected by a short linker. The name of the motif originates from the classification of the structural elements in parvalbumin, where it was first discovered [84]. Parvalbumin is composed by three such motifs, and in particular the geometry of the last two helices in the chain, helices E and F, whose reciprocal arrangement recall the position of forefinger and thumb in our hand (Fig. 4.1), was taken as a reference for representing this kind of motifs. The peculiar arrangement of the helices was later discovered to be highly recurrent, and is arguably the most common pattern coordinating and binding calcium. For example, more than 270 entries of the human genome database correspond to EF-hand sequences [21].

The minimal functional unit, called EF-hand domain, is constituted by a pair of these EF-hand motifs connected by a short linker and packed face to face (Fig. 4.2) [52]. Each EF-hand domain is capable of binding two calcium ions through specific interactions with conserved amino acids and backbone carbonyl groups of the two loops. In spite of sharing virtually the same structural topology and architecture [109], EF-hand domains display a multiplicity of arrangements of the four helices, in three main architectures:

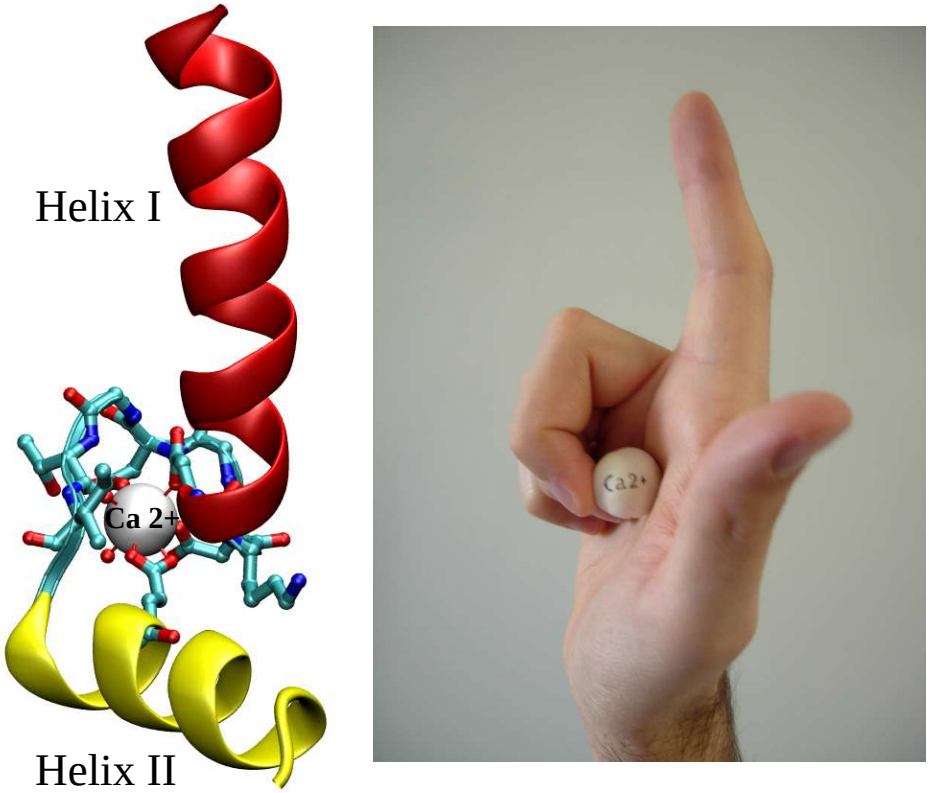


Figure 4.1: The EF-hand motif is constituted by two helices connected by a 12 or 14 amino acid long interhelical loop whose highly conserved residues are arranged in an optimal geometry for coordinating one calcium ion. The geometry of the helix-loop-helix arrangement resembles our hand's thumb and forefinger.

antiparallel bundle, orthogonal bundle and chair bundle. Their cartoon representations are shown in Figure 4.3, adapted from ref [9]. The variability of the internal arrangement is believed to be crucial for the diversification of the signal transduction. In fact, upon calcium binding, EF-hand domains undergo different degrees of conformational changes, related to the variety of different functions of the proteins they belong to. The differences in the dynamical response are in turn expected to reflect specific elastic properties of the domain [105, 129, 16, 9, 52].

The large conformational diversity and variety of responses to calcium binding of EF-hand domains has posed major difficulties for their grouping and classification in terms of structural features. Recent advances in simplifying the structural representation of EF-hand domains have been made by Babini et al. [9], who recently proposed a description of the domains based on the six angles formed by the 4 helices. It was found that only two linear

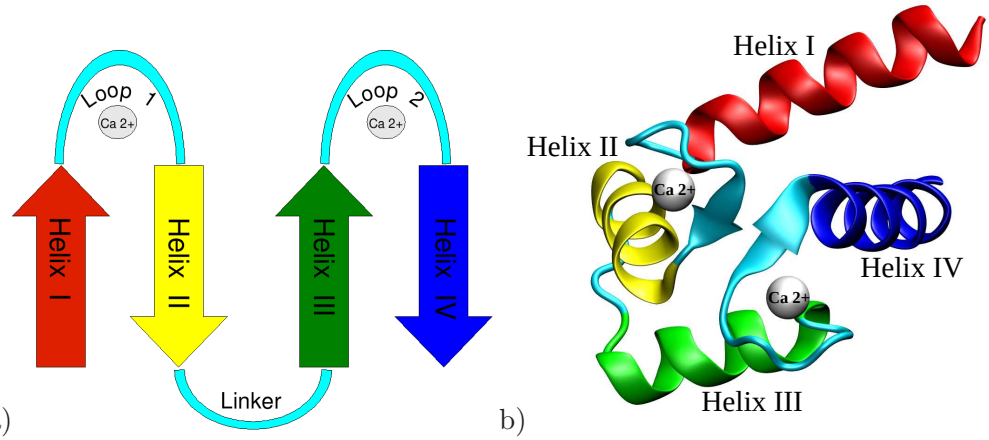


Figure 4.2: (a) A schematic representation of the EF-hand domain. It is composed by two helix-loop-helix motifs, connected by a linker. The loops of the two EF-hand motifs individuates the two binding sites for calcium. Panel (b) portrays a typical three dimensional arrangement of the domain.

combinations of the angles suffice to capture most of the observed structural variations, and thus provide a concise and quantitative framework for representing the structural diversity. Yet, owing to the almost continuous repertoire of structures, no simple unsupervised criterion emerged for partitioning the proteins in neatly-separated groups corresponding to the functional families. Following the definition ref [9], we will consider EF-Hand domains grouped in "functional families" according to the specific protein family they belong to, the specific terminus of location, and the metal ion or ligand binding state. In figure 4.3, adapted from ref [9], protein domains are represented as points whose coordinates are the position on the first 2 principal components of the six inter-helix angles, hereafter indicated as PC1 and PC2. The domains are colored according to the functional family of appartenence by a knowledge-based assignment.

In collaboration with F. Capozzi and C. Luchinat, two of the authors of the study of ref [9], we have undertaken an investigation of the unifying traits within the EF-hand superfamily by examining and comparing the directions of the concerted interhelical movements that the different EF-hand domains can sustain upon thermal excitation. The fact that the domain is shared among such a large number of proteins, sets a natural spatial reference frame for comparing the internal fluctuation and represents an obvious starting point for our investigations of relations between structural and dynamical features.

To characterize the large-scale helical movements in distinct EF-hand domains we followed a strategy articulated over three main steps: (i) creation of a database of viable EF-hand domains selected among all non-redundant

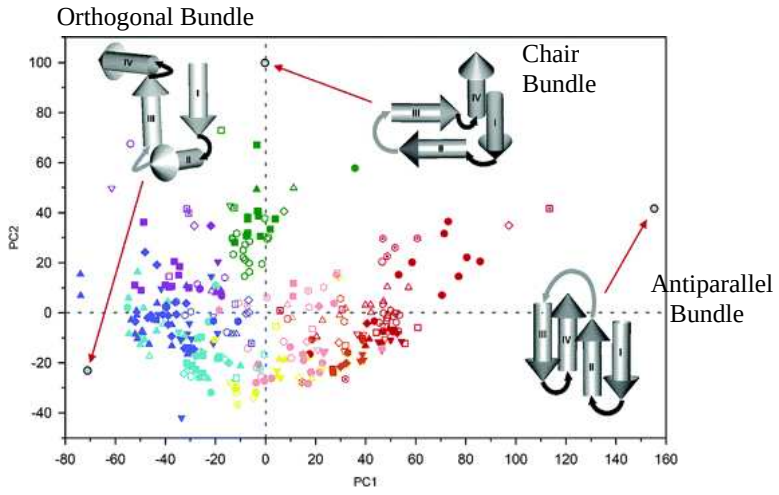


Figure 4.3: EF-hand domains are represented as points in the first 2 principal components of the six inter helix angles. Points are colored according to their functional family. Three idealized structures representing extreme arrangements of the helices are depicted as schematic cartoons. Figure adapted from ref [9]

available structures; (ii) use of suitable mesoscopic models for systematic identification of the essential dynamics of each selected EF- hand domains; (iii) detection of statistically significant similarities of the concerted interhelical movements in distinct EF-hand domains and discussion of the connection with the structural similarities.

4.1 Selection of the Dataset for the Analysis

We have considered a selection of 308 EF-hand domains analyzed in the structural comparative study by Babini et al. [9]. This dataset was originally compiled from a comprehensive set of X-ray or NMR resolved structures ensuring the widest representation of the different biological families and the various possible structural/chemical contexts of EF-hand domains. The dataset included both apo and holo forms, both N-terminal and C-terminal domains, and each of them both in the presence or absence of bound ligands. This database was sieved for this study to remove entries (i) with incomplete or ambiguous (alternate locations for CA atoms) structural information or containing non-standard amino acids, (ii) where any of the EF-hand domain helices was too short to reliably determine its axial orientation (see below) for the purposes of the present calculations or (iii) where the number of heavy atoms was too large (greater than 2000) for an efficient numerical

calculation of the model interhelical dynamics. Accordingly, we did not consider PDB entries with missing residues or with helices spanning less than six amino acids (helices defined according to ref. [9]). This selection procedure singled out 185 EF- hand domains.

4.2 Essential Dynamics in the Interhelical Angle Space

To describe essential dynamics of proteins belonging to the selected dataset, we will make use of an elastic network model akin the one of Tirion [122], discussed in the introduction to Part 1. Details of both the internal structural arrangement of the domain and of the interaction with the rest of the protein and the bound metal ion are expected to influence the internal dynamics of the EF-hand domains. For this reason, to compute the essential dynamics of the domains, we have specifically devised a scheme keeping into account both aspects. We have decided to maintain an atomistic description of the systems and to include into the model also other regions of the protein (if present), not limiting to the truncated domain. For technical reasons related to computational efficiency, we restrict our considerations to the full protein chain comprising the EF-hand domain of interest. Therefore we discarded the interaction with the other chains in multimeric proteins. Accordingly, our model free energy will gather contributions from the pairwise interactions among all heavy atoms, within a cutoff distance of 5Å belonging to the chain containing the EF-Hand domain. All pairwise interaction are controlled by the same coupling constant, except those involving calcium atoms, that are enhanced by a factor of 5. This aims at capturing in a simplified manner the strong electrostatic interactions between calcium and the protein atoms coordinating it.

It is worth to specify that, as expected for this kind of models, the precise value of the harmonic coupling constant does not particularly affect the low energy modes. For example, upon changing the enhancement factor of the interaction between protein heavy atoms and metal ions from 2 to 5, the eigenvalues of the interaction matrix associated to the slowest mode of protein 1qlk change by only 3%. The change approaches 5% if the enhancement factor is set to 20 or more.

The adopted elastic network model combines the reliability in capturing direction of the concerted low-energy protein internal motions [122, 10, 59, 8, 94, 42], with the advantage of being computationally less intensive than molecular dynamics simulations. This makes it a suitable tool for investigating the internal dynamics of a large number of proteins. The traditional description of the slow modes in the space of atomic displacements is not immediately suited for comparing global movements in different EF-hand domains. Owing to the differences in length, domain organization, cofactors

etc. it is not possible, nor desirable, to establish a pervasive one-to-one correspondence of the displacements of all heavy atoms in different proteins. A natural way to circumvent this difficulty is to consider the large-scale movements of the four helices composing the domains as described by the fluctuation of the six angles formed by the helices axes. In fact, the space of the angle among the helices provides a well-defined common framework for dynamical comparison. Moreover, the fact that this set of generalized degrees of freedom has already been used for structural characterizations of these proteins, allows us to investigate in a precise manner the connection between structural arrangement and dynamical behavior, comparing structural and dynamical similarities.

Starting from the description of the free energy as quadratic in the displacement of the atomic coordinates $x_{i,\alpha} = \delta r_{i,\alpha}$, we need to extract information on the lowest energy direction of motion in the space of the angles among the helices composing the EF-Hand domain.

4.2.1 Thermodynamical Integration

Before addressing the specific problem, we recall a general result concerning dimensional reduction and Gaussian integrals, already discussed and used in the contest of proteins in several studies [20, 60, 98]. We start assuming the free energy to be quadratic in the set of system coordinates $\vec{x}$. Let us consider a reduced set of coordinates $\vec{y}$ obtained by a linear orthogonal transformation of a subset $\vec{x}_1$ of the original coordinates while the orthogonal subset $\vec{x}_2$ has been eliminated by thermodynamical integration. It can be easily shown that a quadratic character of the free energy is maintained for the coordinate $\vec{y}$.

More precisely, suppose that the original degrees of freedom $\vec{x}$ can be subdivided in two independent sets $\vec{x}_1$ and $\vec{x}_2$, and suppose that the free energy of the system in the original degrees of freedom $\vec{x}$ can be expressed as

$$\mathcal{F}(\vec{x}) = \vec{x}^T \mathbf{F} \vec{x} = (\vec{x}_1, \vec{x}_2)^T \begin{pmatrix} \mathbf{F}_1 & \mathbf{W} \\ \mathbf{W}^T & \mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix} \tag{4.1}$$

where $\mathbf{F}_1$ is the matrix that governs the interaction between coordinates $\vec{x}_1$, $\mathbf{F}_2$ between $\vec{x}_2$ and $\mathbf{W}$ is the matrix containing the coupling between $\vec{x}_1$ and $\vec{x}_2$. The free energy expressed as a function of a set of coordinates $\vec{y}$ obtained from an orthogonal transformation $\mathbf{A}$ of the coordinates $\vec{x}_1$, can be expressed as

$$\mathcal{F}(\vec{y} = \mathbf{A}\vec{x}_1) = \vec{y}^T \tilde{\mathbf{F}}_y \vec{y} = \vec{y}^T (\mathbf{F}_y - \mathbf{W}_y^T \mathbf{F}_2^{-1} \mathbf{W}_y) \vec{y} \tag{4.2}$$

where we have indicated $\mathbf{F}_y = \mathbf{A}\mathbf{F}_1\mathbf{A}^{-1}$ and $\mathbf{W}_y = \mathbf{W}\mathbf{A}^{-1}$ (See Appendix B for further details).

The low energy modes of the system within this new set of coordinates $\vec{y}$ are obtained identifying the eigenvectors corresponding to the lowest part of the spectrum of the matrix effective interaction matrix $\tilde{\mathbf{F}}_y$, that contains the contribution due to the presence of the integrated coordinates $\vec{x}_2$.

4.2.2 Interhelical-angles Dynamics

In our case, the original degrees of freedom controlling the quadratic free-energy of the protein containing the EF-hand domain of interest are the displacements of the heavy atoms from their equilibrium positions. The reduced degrees of freedom are, instead, the deviations from the average of the six interhelical angles formed by any distinct pair of the four EF-hand helical axes.

To project the high-dimensional free-energy in the space of angles among the helices, we need first to build an orthogonal transformation to a set of coordinates including the displacements of the interhelical angles, and then performing a thermodynamic integration over all other degrees of freedom.

As first step we have calculated the orientation of the axes of the four helices. A helix axis is taken as the normalized distance vector between the residues at the end and those at the beginning of the helix. A robust and simple scheme is to express the axis as a weighted linear combination of the position of the Cα in the helix :

$$\vec{a}_l = \frac{\sum_i \nu_i^l \vec{r}_i}{\| \sum_i \nu_i^l \vec{r}_i \|^2} \tag{4.3}$$

where l labels the index of the helix and i runs over the index of Cα's belonging to the l^{th} helix. A natural choice for the weights is given by setting $\nu_i^l = +1[-1]$ for the 4 residues of the C [N] helical terminus, and zero otherwise.

Despite the simplicity of definition 4.3, the interhelical-angles are typically in accord within 5% with those observed with more sophisticated method such as the one used in the structural study of Babini et al [9]. A further motivation for adopting definition 4.3 is that it straightforwardly lends to calculating the deviation of the angles upon perturbative displacements of the atoms coordinates. This allows a simple derivation for the transformation matrix $\mathbf{A}$ of eqn 4.2.

We shall denote with $\theta_{lm} = \arccos(\vec{a}_l \cdot \vec{a}_m)$ the angle formed by the lth and mth helices. The relative orientation of the four helices will be summarised with the vector $\vec{\theta}$ having components: $\vec{\theta} = \{\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}\}$.

To calculate how the angle between two helical axes depends perturbatively on the displacements of the atoms coordinates we use an expansion to the first order of the arccosine and obtain.

$$\delta\theta_{lm} = \sum_{j,\beta} \frac{\partial\theta_{lm}}{\partial r_{j,\beta}} \delta r_{j,\beta} \tag{4.4}$$

where

$$\frac{\partial\theta_{lm}}{\partial r_{j,\beta}} = -\frac{1}{\sqrt{1-\vec{a}_l \cdot \vec{a}_m}} \left(\sum_{\alpha} \frac{\partial a_l^{\alpha}}{\partial r_{j,\beta}} a_m^{\alpha} + \sum_{\alpha} \frac{\partial a_m^{\alpha}}{\partial r_{j,\beta}} a_l^{\alpha}\right) \tag{4.5}$$

and

$$\frac{\partial a_n^{\gamma}}{\partial r_{m,\mu}} a = \frac{\nu_m^n (\delta_{\mu,\gamma} - a_n^{\gamma} a_n^{\mu})}{\| \sum_i \nu_i^l \vec{r}_i \|^2} \tag{4.6}$$

It should be noted that a nonlinear relationship exists between the six interhelical angles:

$$A^2 + D^2E^2 + 2ADE - BC = 0 \tag{4.7}$$

where

$$\begin{aligned}
A &= \cos\theta_{13}\cos\theta_{14} - \cos\theta_{12}\cos\theta_{23}\cos\theta_{14} - \cos\theta_{12}\cos\theta_{13}\cos\theta_{24} + \cos\theta_{23}\cos\theta_{24} \\
B &= \left(\cos^2\theta_{12} - 2\cos\theta_{13}\cos\theta_{23}\cos\theta_{12} + \cos^2\theta_{13} + \cos^2\theta_{23} - 1\right) \\
C &= \left(\cos^2\theta_{12} - 2\cos\theta_{14}\cos\theta_{24}\cos\theta_{12} + \cos^2\theta_{14} + \cos^2\theta_{24} - 1\right) \\
D &= \cos^2\theta_{12} - 1 \\
E &= cos\theta_{34}
\end{aligned}$$

This reflects at a perturbative level in a linear relationship among the six angular deviations (a consequence of the three-dimensional embedding of the four helical axis). This prevents from adopting a simple orthogonal transformation involving all six equations 4.4.

However we need to maintain the description in terms of six angles, both for consistency with previously published results on the structural characterization of the domains, and because there is no natural way of selecting 5 independent angles out of the six. We need to include explicitly the constraint accounting for the relationship among them. This difficulty is overcome by (i) calculating the reduced free-energy in terms of 5 of the 6 displacement of the angles, and (ii) then extending it to the six-dimensional parameter space by adding a very large free-energy penalty for deviations not satisfying the linearised constraint of equation 4.7.

We will thus retain 5 of the 6 relationship expressed by the equations 4.4. We complement the set of relations including additional linear relationships among the coordinate displacements. These relationships can be conveniently chosen among the identity relationship among heavy atoms displacements. 3N-5 such relationships must be used, making sure that not all

equalities involving Cα atoms defining helical axes are used. The complete 3N set of relations provides a linear transformation from the starting set of coordinates $\vec{x}$ to a new set of coordinates $Y = \mathbf{A}\vec{x}$ whose first 5 coordinates are the displacement of 5 of the 6 interhelical angles. We then operate an integration on all but the first 5 (angular) degree of freedom. By doing so it is obtained a free-energy that is quadratic in terms of the displacements of 5 of the 6 interhelical angles.

As mentioned above the most convenient way to recover the information in the sixth dimension is to embed the 5 dimensional space described by the matrix $\mathbf{F}_\theta$ in the more natural six-dimensional interhelical angle space, enforcing numerically the constraint among displacement of the angles, due to the relationship among the helices axis. The five lowest energy modes in this six dimensional interhelical angle space represents the modes of relaxations in decreasing order of amplitude of fluctuations. They span the vectorial space orthogonal to the constraint of eqn 4.7.

To summarize, starting from computing the interaction matrix in the space of the Cartesian coordinates of all atoms, after applying a linear transformation plus an integration of degrees of freedom we obtain a description of the independent relaxation modes in the space of the six angles between the helices, including the relationship between the angle variation as a constraint enforced in the calculation of the essential modes. The sixth mode, representing the geometrical constraint among the angles displacement, does not contribute to the angle fluctuations.

The previous equation allows one to express the fluctuations in thermal equilibrium of any pair of angles in terms of the corresponding element of the inverse of $\mathbf{F}_\theta$:

$$< \delta\theta_i \delta\theta_j > \propto \mathbf{F}_{\theta\, i,j}^{-1} \tag{4.8}$$

Also in this case the independent modes of decay of thermal excitations are provided by the six eigenvectors of the $\mathbf{F}_\theta$ matrix, $\vec{v}_1,\ldots,\vec{v}_6$ (with associated eigenvalues,$\lambda_1,\ldots\lambda_6$). As a consequence of 4.8 the fluctuations along each of the six principal directions will be proportional to the inverse of the corresponding eigenvalues. This fact is conveniently exploited to establish the relative contribution, ω_i, of the ith slow mode to the overall square fluctuation dynamics:

$$\omega_i = \frac{\lambda_i^{-1}}{\sum_{j=1}^{6} \lambda_j^{-1}} \tag{4.9}$$

Due to the progressive decrease of the weight of the slow modes with their ranking, our results and discussion will be restricted to the top two slow modes which are indeed sufficient to account for most of the angular fluctuations.

4.2.3 Application to the Dataset

For each member of the comprehensive structural selection of 185 EF-hand domains, we identified the concerted interhelical angular low energy modes that occur in thermal equilibrium and established their relative contribution to the overall angular fluctuation dynamics. In particular, for each considered EF-hand domain we calculated the fraction of angular fluctuation dynamics captured by the first slow mode alone and by the combination of the first and second slow modes. The distribution of such fractional dynamical contributions over the whole dataset is shown in Figure 4.4. It is seen that the first mode alone (Figure 4.4a) is typically sufficient to cover 38% of the total angular fluctuations. As illustrated by Figure 4.4b, the inclusion of the second mode raises the fraction of captured angular fluctuations whose average is $\bar{\omega} = 0.671$ with dispersion $\sigma_\omega = 0.055$. Lower fractional values of the total fluctuation reflect a rather flat free energy landscape where it is hence difficult to have a clear-cut ranking of the different essential dynamical spaces. To avoid these ambiguous situations we omitted from further analysis the proteins in the leftmost tail of the distribution of Figure 4.4b. More precisely, we discarded the 31 entries whose weight of the first two modes was smaller than $\bar{\omega} - \sigma_\omega$. In the following we shall thus exclusively consider the 154 proteins that result from this filtering of the original data set.

The complete dataset of the analyzed domains, along with the corresponding information on definition of the helices and the fraction of fluctuation captured by the first two modes has been deposited as Supplementary Information to Journal of Proteome Research and is available through the website http://pubs.acs.org/journals/jprobs/index.html

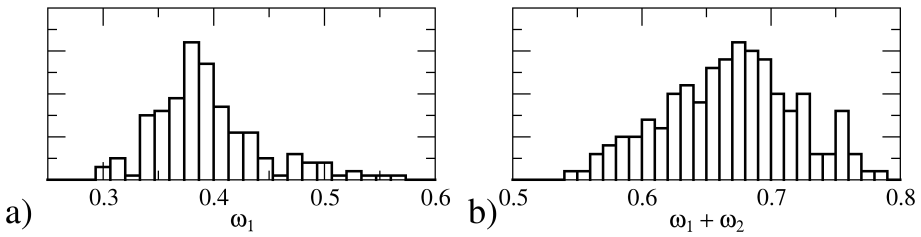


Figure 4.4: Normalized distribution of the fraction of angular fluctuations captured by the first mode (a) and by the first plus second modes (b).

4.3 Comparison of interhelical angles and their fluctuation dynamics in different proteins.

Two different notions of distances will be used to compare the reference (crystallographic or NMR) interhelical angles and their slow modes in all distinct pairs of proteins (distinguished by the superscript a or b). The structural (static) distance used to capture the similar relative orientation of the four EF-hand domain helical axes is naturally defined as the Euclidian distance in the space of angular vectors:

$$d^{struct}(a,b) = \sqrt{\sum_{i=i}^{6}(\theta_i^a - \theta_i^b)^2} \tag{4.10}$$

The RMSIP introduced introduced in Chapter 2 (see eqn 5.4) has been used to define the (non-euclidean) "dynamical distance" between two sets of essential angular modes :

$$d_n^{dyn}(a,b) = 1 - \sqrt{\frac{1}{n}\sum_{i=i}^{n}|\vec{v}_i^{\,a} \cdot \vec{v}_j^{\,b}|^2} \tag{4.11}$$

where we indicate with $\vec{v}_i^{\,a}$ [$\vec{v}_j^{\,b}$] the ith [jth] slowest mode of protein a [b].

As in the cases discussed in the previous chapters we have complemented the pairwise distances in dynamical space with a control on the likelihood that they could have arisen by chance (i.e. in the absence of any significant correlation between the slow modes of two proteins). This was accomplished by comparing the observed distance values against a control distance distribution expected for sets of orthonormalized vectors randomly picked in the six-dimensional space. More precisely, we generate random basis sets for the six-dimensional space by repeatedly applying the Graham-Schmidt orthogonalization procedure to a succession of unit vector uniformly picked on the six dimensional unit sphere following the procedure outlined in Allen and Tildesley [2]. The resulting distribution of $d_n^{dyn}(a,b)$ values thus provided the required statistical reference. It should be noted that other criteria for choosing the reference distribution could be adopted, possibly also accounting for the non-linear relationship that ties the six interhelical angles. The one adopted here was chosen for its transparency and simplicity of application.

Our first aim is to establish the existence of statistically-significant analogies of angular fluctuations in distinct EF-hand domains and, if so, cluster the dataset into groups with similar dynamics. We hence started by measuring the dynamical distances of the first slow mode for all distinct pairs of proteins. The resulting distribution of 11781 pairwise distances is shown in Figure 4.5a. The distribution shows an increase for distance values smaller

than 0.1. This distribution should be compared with the reference distribution of distance values expected from random choices of the slow modes (dotted curve in Figure 4.5a). It is immediately apparent that the experimental distribution has only a small overlap with the random distribution, thus indicating the existence of statistically significant correlations among the slowest angular modes of distinct EF-hand domains. The experimental distribution, however, lacks the bimodal signature that should accompany a clear clustering of the domains in distinct groups of comparable population and diameter". In the latter case, in fact, the cumulative distribution should present a peak at small distances arising from intra-cluster pairwise comparisons, and one at larger distances corresponding to the most probable inter-cluster distance.

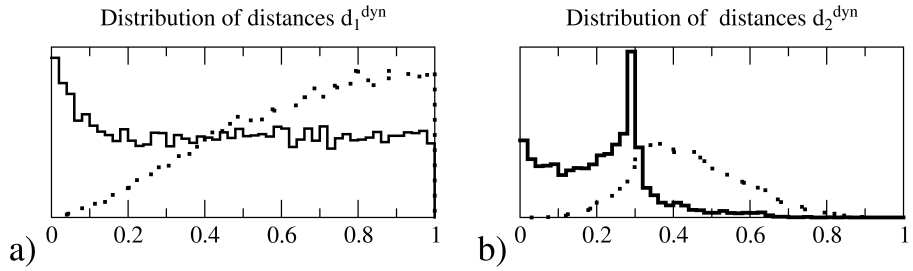


Figure 4.5: Normalized distributions of the dynamical distance $d_n^{dyn}(a,b)$ for n=1 (a) and n=2 (b) (solid lines), compared with the random reference distributions (dotted line).

Conversely, a bimodal character is clearly visible when the angular dynamical distance is calculated over both the first and second slow mode (n = 2 in Equation 4.11), see Figure 4.5b. Again, the increase at $d_n^{dyn}(a,b)$ values smaller than 0.1, which measures the dynamical consistency (radius) within the dominant clusters, has a good statistical significance, as it does not appreciably overlap with the reference distribution. The good consistency of the fluctuation dynamics within the clusters is further accompanied by the separation of the dominant clusters, denoted by the peak at $d_n^{dyn} \sim 0.3$ (which captures the typical dynamical distance of members of distinct dominant clusters). In conclusion, the distribution obtained including two slow modes indeed indicates the existence of dynamical clusters.

4.4 Dynamics-based Grouping of Functional Families

The inspection of the distribution of dynamical distances for n = 2 exhibits the distinctive hallmark of the presence of clusters with very similar

interhelical angle fluctuations. Of the several available clustering schemes we adopted the standard K-medoids algorithm[74] (suitable also for non-Euclidean metrics), already discussed briefly in Chapter 3. For a given number of desired clusters, K, the algorithm optimizes the choice of cluster representatives and cluster members so that the summed distance of each protein from its cluster representative is the smallest possible. The dataset clustering was initially carried out for K from 2 to 15. Inspection of the resulting clusters and the matrix of pairwise distances of inter-cluster and intra- cluster members revealed that a balanced subdivision could be achieved for K = 4. Larger values of K lead to a small inter-cluster pairwise distance compared to the natural cluster separation indicated by the overall distribution of pairwise distances. On the other hand, smaller values of K yielded excessively large intracluster pairwise distances compared to the natural cluster "radius", again indicated by the overall distance distribution. For the subdivision in K=4 groups sizeable populations and typical inter- and intra-cluster distances consistent with the features of the distribution of Figure 4.5 was obtained. The number of members in the four clusters is respectively, 66 (CL0), 48 (CL1), 21 (CL2) and 19 (CL3). The four representatives are: CL0: the holo-form of the rat S100B protein (1qlk_A0); CL1: the holo-form of the C-terminal human CaM bound to a target peptide (1nwd_A2); CL2: the holo-form of the N-terminal human nucleobindin 1 (1snl_A2); and CL3: the apo-form of the C-terminal small subunit of rat calpain (1aj5_A2). The quality of the clustering can be visually appreciated through the density plot of Figure 4.6.

The figure portrays a color-coded distance matrix of all pairs of domains, which were re-indexed so that the first 66 entries are the members of the first cluster, the subsequent 48 belong to the second cluster, etc. A detailed analysis of the distances of each entry from the four cluster representatives indicates that the dynamical similarity is high across members of the two dominant clusters and less marked for the remaining two which, indeed, contain some members that are close to CL0. Nevertheless, the marked block nature of the matrix suggests that the distinct character of the clusters should persist upon addition of EF-hand domains unrelated to the ones used here. We have verified this expectation by repeatedly reducing the data set by 20% (that is by leaving out 31 members) and identifying the four cluster representatives. All members of the non-reduced set (i.e. including entries not used for the identification of the representatives) were then assigned to the cluster of the nearest representative. The resulting clusters were then compared with the reference ones (CL0, CL1 etc.). On average, the random data set reduction results in only 16 members out of 154 to be assigned to a cluster different from the reference one. This indicates that (i) the partitioning in four groups is not labile even when the number of removed items compares with the combined population of the two smallest clusters and (ii) newly-available entries unrelated to those used for clustering can be

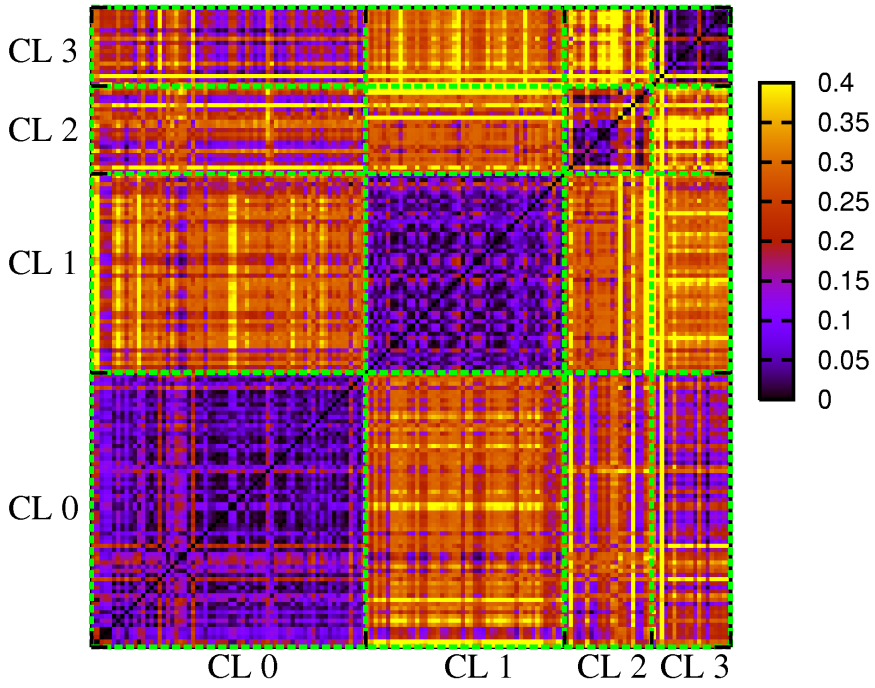


Figure 4.6: Density plot of pairwise dynamical distances. Proteins have been grouped according the subdivision in four clusters (CL0,, CL3).

reliably attached to existing clusters using the criterion of minimal dynamical distance to the nearest representative. The partitioning of all examined EF-hand domains into the four CL0-CL3 clusters, along with the dynamical distance of each member of the clusters from its representatives, is available as Supplementary Information on the website of Journal of Proteome Research through the link http://pubs.acs.org/journals/jprobs/index.html .

4.4.1 Functional and Dynamical Groups

A significant correspondence exists between each functional family and the dynamics of its domain. It appears that each functional family has its own characteristic movements, describable by a pair of slow modes which are shared by other members of the same functional family.

In Figure 4.7, the dynamical distances between domains belonging to the 12 most populated families are reported in a color-coded matrix. Visual inspection of this figure confirms that the dynamical distances within each functional family are on average very small. It also appears that EF-hand domains containing a bound peptide show a larger variability of behavior,

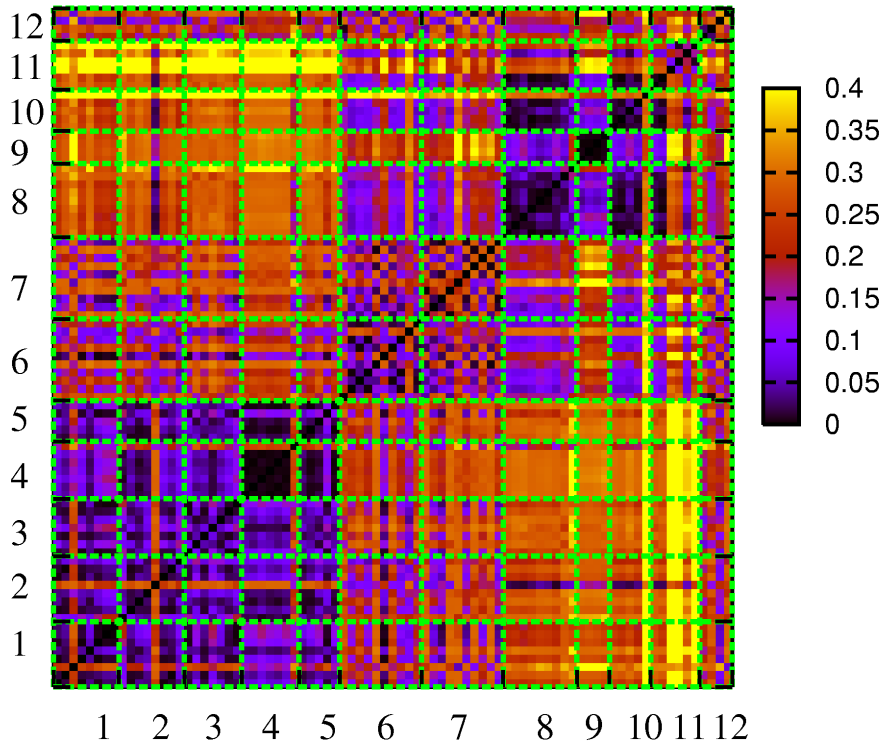


Figure 4.7: Density plot of pairwise dynamical distances. Only members of the most populated functional families have been represented. Keys to the families: 1) calmodulin N 2Ca P0; 2) calmodulin-N 2Ca P2; 3) Myosin regulatory light chain C1 Mg 1 P2; 4) calmodulin C2 Ca P0; 5) calmodulin C 2Ca P2; 6) myosin regulatory light chain C0 Ca P2; 7) myosin essential light chain C 0Ca P2; 8) S100 N2 Ca P0; 9) calpain small subunit C 2Ca p0; 10) skeletal troponin C N 0Ca P0; 11) S100 N 0 Ca P0; 12) calmodulin C 0CA P0. For further details see Table in Appendix C

consistent with the structural diversity of the EF-hand domains complexed with various ligands. We should however recall that our model takes into account only interactions amog atoms within the protein chain that comprises the EF-Hand domain. Specific considerations on the motion in the presence of the ligands are out of the scope of the model used. The following analysis will thus concentrate on EF-hand domains free of bound ligands. The relative information is summarized in Table 1.

In summary: the apo-forms of the N-terminal domains, or of single-domain proteins, are characterized by essential modes clusterized in CL0, CL2 and CL3. CL0 comprises some S100 proteins, most of the skeletal troponins C (skTpCs), the penta-EF-hand (PEF) proteins, including calpains,

Table 4.1: Structure-dynamics clusterization of functional domains within the EF-hand superfamily; Legend:
[a] : N-term, N-terminal domain/motif pair; M-pair, middle motif pair; C-term, C-terminal domain/motif pair
[b] : A: antiparallel bundle; O: orthogonal bundle; C: chair bundle
[c] : The number on the right indicates which motif is occupied by Ca within the EF-hand pair. EPS15 has a calcium ion in either motif 1 or motif 2.

Protein	Domain[a]	APO	Struct[b]	CL	1Ca[c]	Struct	CL	2Ca	Struct	CL
CaM	N-term	0Ca	A	3				2Ca	O	1
	C-term	0Ca	A/C	2				2Ca	O	1
skTpC	N-term	0Ca	A	0/3				2Ca	O	1
	C-term							2Ca	O	1
MELC	N-term	0Ca	A	3						
	C-term	0Ca	C	2						
Recoverin	N-term	0Ca	A	3	1Ca2	O	0			
	C-term				1Ca1	O	0			
Neurocalcin D	N-term				1Ca2	O	0			
	C-term							2Ca	O	0
Neuronal CS	N-term				1Ca2	A	1			
	C-term									
cTpC	N-term				1Ca2	A	0			
	C-term									
Calcineurin B-like	N-term							2Ca	O	0
	C-term				1Ca2	O	0	2Ca	O	0
Ca Vector Prot	N-term	0Ca	A	2						
	C-term									
KchiP	N-term	0Ca	O	0						
	C-term							2Ca	O	1
EhCaBP	N-term							2Ca	A	2
	C-term									
Calpain Small	N-term	0Ca	A	0				2Ca	A	0
	C-term	0Ca	A	3				2Ca	A	3
Sorcin	N-term	0Ca	A	0						
	C-term	0Ca	A	0						
Grancalcin	N-term	0Ca	A	0						
	C-term	0Ca			1Ca1	A	0			
ProgCellDeath	N-term							2Ca	O	0
	C-term	0Ca	A	0						
Calbindin D28K	N-term				1Ca1	O	0			
	M-pair							2Ca	O	1
	C-term				1Ca1	O	0			
S100		0Ca	A	0/2				2Ca	A	0
Calbindin D9K		0Ca	A	0				2Ca	A	3
Nucleobindin1								2Ca	C	2
EPS15					1CA1(2)	A	0			

and the KChIP. The remaining S100 proteins are grouped in CL2, while CL3 collects CaM, MELC, recoverin and one skeletal troponin C. It is apparent that there is no clear relationship of the partitioning between CL0, CL2 and CL3 with the conformational state of the domains. All domains are in the antiparallel bundle conformation, except KChIP which has an orthogonal bundle conformation despite being an apo- domain. This domain falls in CL0, together with several other domains with clear antiparallel bundle structure. The apo-forms of the C-terminal domains fall in CL0, CL2 and CL3. CL0 comprises all PEF protein domains but calpains, CL2 comprises CaM, MELC and calcium vector proteins, and CL3 comprises calpains (small subunit). Domains binding only one calcium ion are all clustered in CL0, independently of whether they are N- or C- terminal domains or whether the calcium ion is bound in the first or the second loop of the

domain. Finally, the di-calcium (or holo) forms may fall into any of the four clusters. CL0 comprises neurocalcin, calcineurin B, the dimeric S100 proteins, and the N-terminal domains of the PEF calpain (small subunit) and programmed cell death protein; CL1 hosts N- and C- terminal domains of CaM and skeletal troponin C, and the C-terminal domains of KchIP; CL2 contains only the single EF-hand domain of nucleobindin1, and CL3 contains calbindin D9k and the C-terminal domain of the PEF calpain (small subunit).

4.4.2 Representatives Dynamics

To help appreciating the differences in essential dynamics among the four different clusters, we have selected the representative domain for each of them and visualized the first essential mode (EM1) direction in the form of arrows applied to a cilinder representing the axes of each helices (Fig. 4.8). The same can be done for the second essential mode (EM2). Animated motion of the first two essential modes of the representatives have been deposited as Supplementary Information to Journal of Proteome Research and are available through the website http://pubs.acs.org/journals/jprobs/index.html

Each couple of essential modes, EM1 and EM2, characterizing the essential dynamics of each representative, is describable by looking at the effect of the movement on the interhelical angles. For example, a parallel, consensus movement of two helices has a negligible effect on their interhelical angle (e.g., H2-H4 in EM1 of the CL0 representative), while a scissor-type, anticorrelated movement has a large effect on their interhelical angle (e.g., H1-H3 in EM2 of the same CL0 representative). Analogously, a stationary helix determines an intermediate effect on the interhelical angle when the other helix of the pair is fluctuating (e.g., H1-H2 in EM2 of the CL0 representative). The slowest modes of the four cluster representatives indicate different mobility for the four helices. For example, in CL0, EM2 shows much wider fluctuations for H1 and H3 than for H2 and H4. A dissimilar behavior of helical displacements is also observed in the other three clusters. In CL1, EM1 shows H2 and H3 more displaceable than H1 and H4, and EM2 shows H3 and H4 more displaceable than H1 and H2. In CL2, EM1 has a practically immobile H2, and H3 is much less moveable than H1 and H4. The last case of asymmetry in movements of the four helices is encountered in EM2 of CL3, where H3 is the only helix being almost immobile. The description of the slow modes of all representatives is concluded by looking at EM1 of CL0, EM2 of CL2, and EM1 of CL3, where all helices are capable of large movements. However, a different combination of correlated and anticorrelated fluctuations gives origin to three distinct pairs of modes. EM1 of CL0 has scissor-type displacements, mainly for H1-H2 and for H3-H4; EM2 of CL2 has the same type of displacement mainly for H1-H3 and for H2-H4; finally, EM1 of CL3 has a scissor-type displacement mainly for H1-H3 and

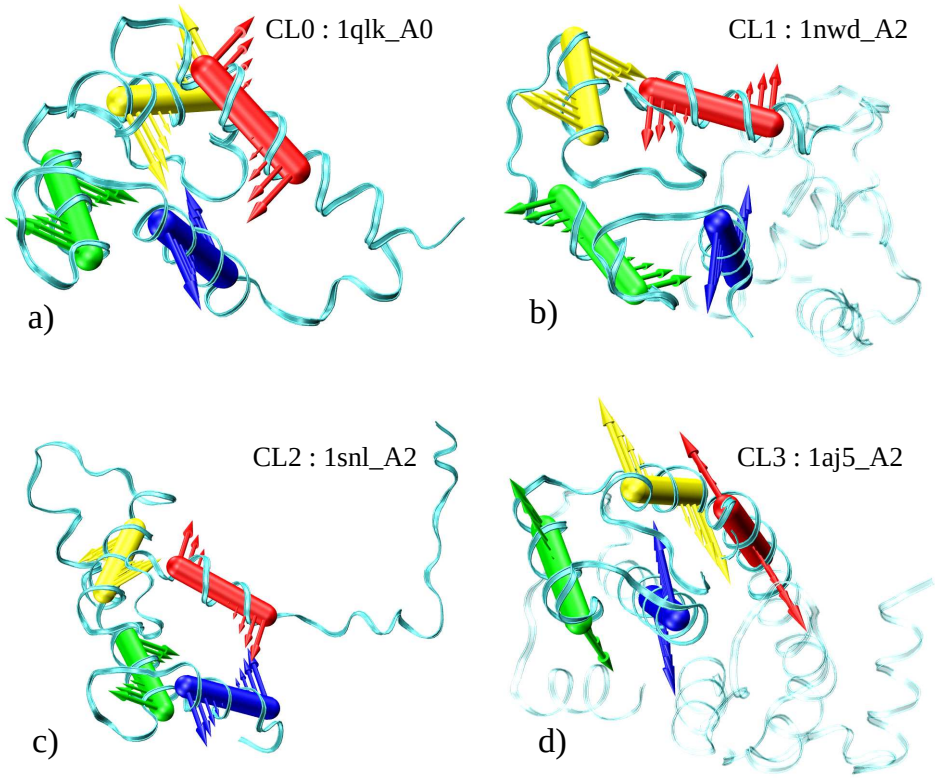


Figure 4.8: The four helices H1-H4 are represented as ribbons harboring colored cylinders. From H1 to H4, they are colored in red, yellow, green and blue, respectively. The lowest energy mode is represented as arrows describing the direction of motion of the helix.

H3-H4.

Overall, the pairs of slow modes characteristic of each cluster are sizeably different from one another. It is therefore surprising that domains with analogous structure (e.g. N- and C-terminal domains of apo CaM, Table 1) are assigned to different clusters (CL3 and CL2, respectively), and even more surprising that domains that differ sizably in structure, such as, for example, apo and calcium bound forms of S100 proteins (Table 1), belong to the same cluster (CL0). In general, domains that undergo a sharp transition between an antiparallel bundle and an orthogonal bundle structure upon calcium binding (such as CaM-like proteins), move from either CL2 or CL3 in the apo form to CL1 in the calcium-bound form, while domains that undergo still sizable but more localized conformational rearrangements, such as S100 proteins, usually belong to CL0 in the apo form and remain in the same cluster in the calcium bound form.

4.4.3 Connection among Structural and Dynamical Similarities

Apparently, the dynamical clustering reflects only to a modest extent the "structural context" of the 4 helices under consideration. We recall that the slow modes calculations, though analyzed for the 4 helices only, are performed on the whole protein chain. In summary, there is no strict correlation of the dynamical distance between two functional domains expressed with the $d_n^{dyn}(a,b)$ values, which measure the consistency of the space spanned by the two lowest-energy modes of both proteins, with the structural distance measured as the Euclidian distance of the six interhelical angles. The relationship between the dynamical and angular distances can be also appreciated in the scatter plot of Figure 4.9, where we have reported simultaneously both quantities for each of the 11781 distinct proteins pairs. The overall trend of the distribution indicates a fair degree of correlation between the interhelical dynamics and the helical spatial arrangements, but the dispersion is substantial. As already observed, pairs of proteins with very similar interhelical angles can differ significantly in dynamics, and vice-versa.

The histograms flanking the scatter plot in Figure 4.9, showing the distribution of the dynamical distances and of the angular distances, prompt further considerations. The presence of several pairs of members of different clusters contributing to the $d_n^{dyn}(a,b)$ peak at 0.3 is clearly visible (the values attainable by $d_n^{dyn}(a,b)$ range from 0 to 1). The double peaked character of the $d_n^{dyn}(a,b)$ histogram in Figure 4.9 should be compared with the broader distribution of angular distances. Although, as elucidated by previous studies, some grouping of the proteins can also be discerned in terms of angular distances, the distinction among such groups is less pronounced than for the dynamical features. The qualitative difference of the two histograms emphasizes the non complementarity of the insights offered by comparison of static and dynamic features of the interhelical angles. In particular, it is noteworthy that, despite the almost continuous repertoire of interhelical arrangements, it is possible to identify groups of proteins that have common and distinctive dynamical traits. These features are conveniently illustrated by exploiting a reduced representation for the interhelical angles based on their first two principal components that was introduced to account concisely for the observed structural diversity of EF-hand domains. Angular movements can be represented as arrows in the 2D space of the principal components described above. Due to the near degeneracy of the first two slow modes we decided to define for each protein a new orthonormal base (V1 and V2) in the space generated by its first two modes so that the first vector of the base is the one that maximizes the scalar product with the top mode of its cluster representative. In the plot of Figure 4.10 the interhelical arrangements are represented as points in the principal component (PC) space of Fig 4.3, while the segments indicate the new, optimized, basis vec-

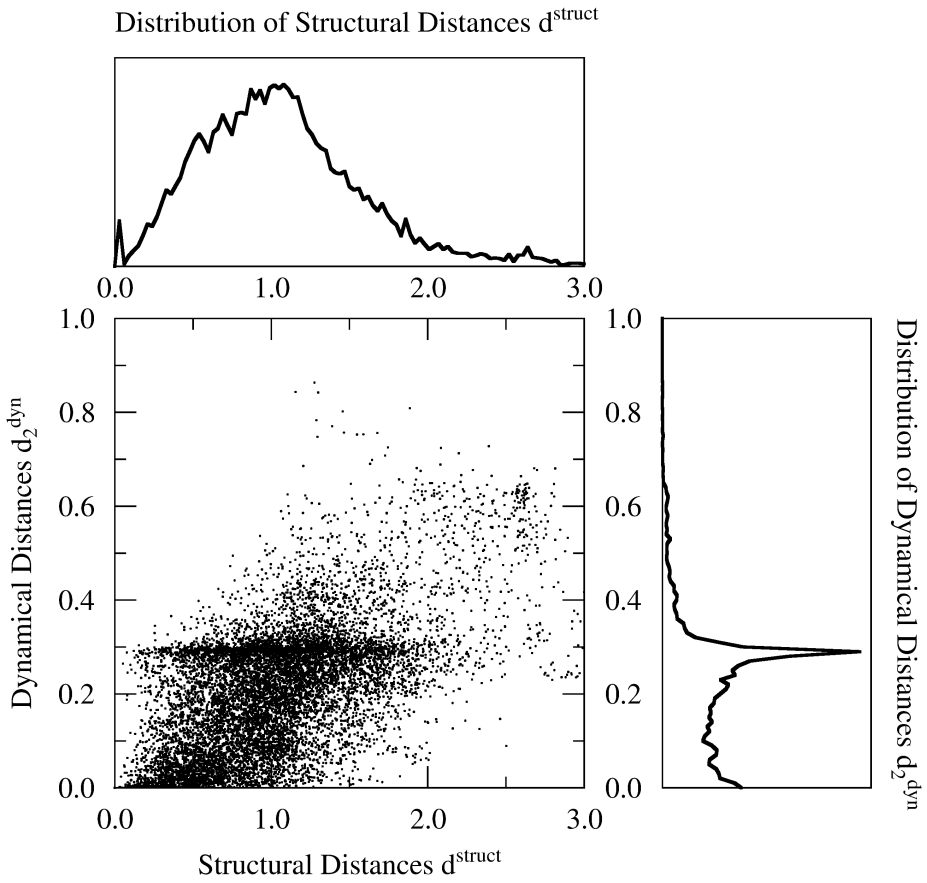


Figure 4.9: Scatter plot and projected histograms of structural and dynamical distances for all distinct pairings of the 154 proteins.

tors directions. For clarity, V1 and V2 are represented in separate graphs. The color code of the points and segments reflect the dynamical cluster of appartenence, the fainter the color the higher its dynamical distance from the cluster representative. These plots provide a vivid illustration of the features emerging from the previous structural/dynamical analysis. In particular, members of the same dynamical cluster may occupy a fairly large region of the PC space, and yet their intercluster dynamical consistency, perceivable by the projected directions of V1 and V2, is very high (indeed all members of any cluster have mutual dynamical distance typically below 0.2).

It is also apparent that, in the absence of the dynamical clues (i.e. the segments in Figure 4.10) it would be very difficult to introduce any clear-cut objective criterion for grouping the points, given the diffuse repertoire of interhelical angles. The dynamical criterion, on the contrary, leads to a

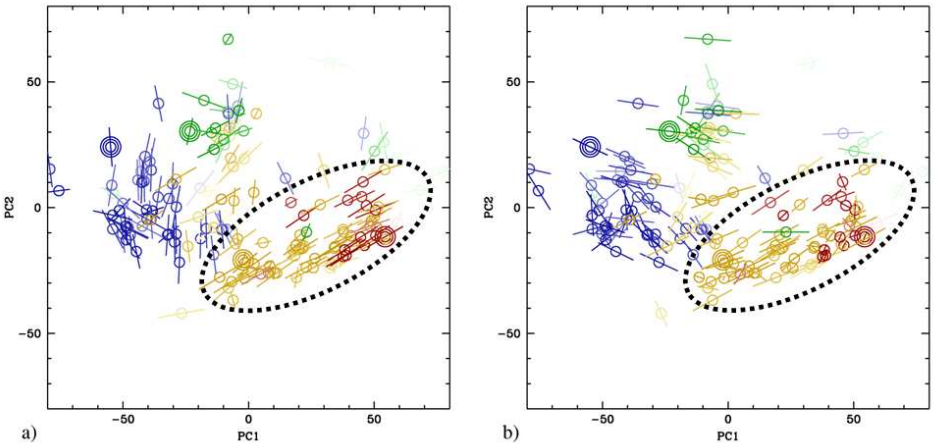


Figure 4.10: Projections of the (a) first and (b) second optimized basis vector of the slow modes on the two-dimensional PC space. Members of the four clusters are colored in: orange (CL0), blue (CL1), green (CL2), red (CL3). Cluster representatives are shown with thick circles.

sharper distinction among the most populated clusters. Though the representation of Figure 4.10 is in a reduced PC space, it can be appreciated how pairs of proteins with very different dynamics can be close in angle space and vice versa. From the detailed analysis of Figure 4.10 several interesting structural/dynamical considerations can be made. First, for all members of CL3 (colored in red) and for the large fraction of those of CL0 (colored in orange) that lie in the lower right area of the PC space (indicated by a dashed ellipse in Figure 4.10), the two projected directions of motion are almost parallel to one another and oriented along the direction of spreading in the PC space of the corresponding EF-hand family members. As pointed out before, these domains comprise mostly apo domains and di-calcium ones that do not open up completely upon calcium binding, i.e. those mainly in the antiparallel bundle form. This fact strongly suggests that the considerable structural variation observed among the individual domains along the major axis of the ellipse in Figure 4.10 reflect a progressive distortion along the easy directions of motion indicated by the segments. It is appealing to observe that individual domains within this large subgroup are mostly scattered precisely along the common easy axis thus covering the space of easily excitable conformational changes available to the individual members. Conversely, the directions of motion of the members of CL1 and CL2 as projected in the PC space are both different for the two modes and also different within each cluster. CL1 and CL2 host many di-calcium domains that are in the orthogonal or chair bundle form, i.e. those that are ready to recognize and bind target proteins. It is conceivable that their motion is

less predictable, being possibly also dictated by the rather broad range of interactions that some of them (e.g. calmodulin) are able to perform. The above considerations hint at a relationship between interhelical dynamics and biological function.

Further considerations and the biological implications of the findings have been made by C. Luchinat and F. Capozzi that have been our partners in the clustering study of calcium-binding proteins. Exposition of these details is beyon the scope of the present thesis and is covered in ref [22].

4.5 Summary

We have analysed the essential modes of motion of the four helices of 187 domains members of the EF-hand protein superfamily, using a novel approach that describes the motions directly in the six-dimensional interhelical angle space. It is found that, for as many as 154 of them, the dynamics the two lowest energy modes account for more than 65% of the global motions. The formulation of the problem in interhelical angle space provides the common framework for comparing the concerted helical movements sustained in all possible pairs of EF-hand domains, regardless of the degree of structural similarity. This represents a systematic quantitative attempt to elucidate the connection between the large structural variability with EF-hand domains and large- scale concerted movements, which typically shape the conformational changes that assist or accompany the functional activity. It is found that the distance in dynamical space of two EF-hand domains is only loosely related to the spatial (static) difference of the helices orientation. In particular, the analysis of the distribution of both static and dynamic distances shows that the former is essentially a continuum, while the latter is clearly bimodal, indicating that a "natural" clustering of the type of motions of EF-hand domains occurs. This dynamical grouping, which aptly complements previous structurally-related subdivisions, appears adequately captured at the level of four dynamical clusters. The robust nature of the dynamical grouping, which highlights the presence of highly-corresponding interhelical movements in otherwise different domains, hints at the functional relevance of the observed modes.

Chapter 5

Dynamical Comparison of Proteins with Partial Structural Similarity

The comparative analysis of essential dynamical spaces has proved useful in characterizing the dynamical behavior of EF-hand domains of proteins belonging to several functional families. The interplay between structural and dynamical similarities presents interesting aspects. In fact, while there is a certain degree of correlation between structural and dynamical similarity, it is found that domains with different structural arrangement can share common large-scale fluctuations whereas similar structures can have different dynamical behavior. This complexity in the relation among structure and dynamics is contained both in the detailed differences of the domains composition and in the effect on dynamics of the remaining part of the protein chains, whose length and structure can vary appreciably from protein to protein.

We shall address here the connection between structural and dynamical similarities for two members of the protease superfamily [11], carboxypeptidase A [76] and pyroglutamyl peptidase [120]. The overall architecture of the two biomolecules is analogous according to the CATH classification [109]; their sequence homology is, in fact, as high as 45%. Yet, the enzymes differ significatively by length (208 and 307 residues, respectively) and number of secondary elements. The two proteases rely on different catalytic chemistries and belong to different clans (clans MC and CF for carboxypeptidase A and pyroglutamyl-peptidase, respectively). The comparison of the large scale fluctuation dynamics of the two enzymes is therefore interesting as it can shed light on possible general and transferable aspect of the structural modulations that accompany the proteolysis reaction. This investigation extends and complements a recent comparative study aimed at elucidating common dynamical features shared by members of protease superfamily [26]. Unlike

the case of EF-hands, the two proteases under consideration do not possess detailed structural correspondences encompassing the enzymes in their entirety. This prevents us from using an intuitive structural inspection for identifying their common degrees of freedom to be used in the dynamical comparison. However though differing for length and secondary content, the two molecules are still partially superimposable. This allows us to rely to a structural comparison algorithm to identify a structurally relevant matching region. The question we wish to address is how the statistically-relevant spatial correspondence of a common structural core reverberates in a correspondence of the essential dynamics of the aligned regions.

To compute the essential dynamical space in the aligned regions we will follow the methodology recently developed in our group [26].

In particular the adopted method is articulated over a few steps. First significant partial structural correspondences between the two enzymes are identified by means of structural alignment techniques. The aligned core provides the sought objective reference frame for comparing the enzyme's dynamics. Next, the coarse-grained Beta Gaussian Model of ref [93], described in Appendix D, is used to compute (and hence compare) the large-scale fluctuations of the aligned amino acids. Notice that this step entails a thermodynamic integration over the non-aligned residues. Finally, we discuss the statistical relevance of the dynamical accord and elaborate on its biological implications.

5.1 The case at study : two Proteases

Proteases play crucial roles in the life cycle of all organisms as they affect a wide spectrum of physiological processes such as cell growth, cell death, blood clotting, immune defense and secretion. At a molecular level they act as "scissors" capable of cleaving polypeptide chains, that is other proteins and enzymes. The repertoire of known proteases covers a wide range of:

- (a) catalytic/reactive mechanisms and substrate specificity (the hydrolysis reaction leading to the cleavage of the peptide bond can involve different catalytic residues, such as Ser, Asp, Cys, Glu and Thr or even Zn metal ions) [11],
- (b) structural folds, as the approximately 2,000 proteases of known structure can be assigned to as many as thirteen distinct folds [124].

This chemical and structural diversity is so significant that, prior to the identification of common fluctuation dynamics [26], the various classes were thought to be related only by the fact that the peptide substrate in the binding cleft adopts an extended beta conformation [124].

The two enzymes considered here, carboxyl peptidase A [76] and pyroglutamyl peptidase [120], provide an example of the differences in protease

catalytic chemistry and, to a lesser extent, of the structural traits. A ribbon representation of the two enzymes is given in Fig. 5.1. According to the MEROPS classification of proteases [11], carboxypeptidase A belongs to the M14 family of metallo proteases, whose function in mammals is related to alimentary digestion. It acts by cleaving a single C-terminal amino acid (particularly aromatic ones or residues with branched side chains). Its fold is constituted by a 3 layer alpha/beta/alpha sandwich with an antiparallel beta-sheet of eight strands. The active site is on the beta layer and is composed by: a single catalytic zinc ion, tetrahedrally coordinated by two histidines (His69 and His196), a glutamate (Glu72) and the catalytic water molecule (responsible for the nucleophilic attack), Arg127 and Glu270. His69 Arg71 and Glu72, arranged in a sequence pattern well conserved among the family, are in a loop connecting one beta strand and one of the helices; Arg127 is also in a loop connecting two helices, His 196 and Glu270 are placed on adjacent strands. Another key residue, Tyr248, which is strictly conserved within the family is located in a loop surmounting the active site; experimental evidence has indicated its important role in substrate binding and/or catalysis [31].

The second proteases is a bacterial pyrrolidone carboxyl peptidase, from hyperthermophile, and belongs to the cysteine protease enzymatic class. Its molecular function consists of removing one pyroglutamate residue from the N-terminus of a peptide. As in almost all cysteine proteases, the active site is constituted by a catalytic triad, namely a glutamate, a histidine and a cysteine (the latter being the nucleophilic agent) [11]. Akin to carboxypeptidase A also the pyroglutamyl peptidase is organized as a alpha-beta-alpha sandwich with four long parallel and two short antiparallel beta-strands surrounded by three helices on one side and two on the other [118]. However, its crystallographic quaternary organization can be very different for different organisms: monomeric for mammals and bacteria as opposed to tetrameric for archaea. The oligomeric state of the enzyme in solution is still a matter of debate. However, as the active site is contained completely within a monomer we shall consider here the enzyme in its monomeric form. Interestingly, the fold is very different from that of any other cysteine peptidase, but structural similarities were detected between members of its clan and those of clans of metallopeptidases. In fact, despite the fact that MEROPS [11] classification scheme assigns it to a different clan from the metallo carboxypeptidases the core of both enzymes presents visible analogies (see Fig 5.1). Moreover, as for carboxypeptidase A, the active site is located on the beta layer with the Glu 79 and His166 positioned on two adjacent strands of the layer and the nucleophilic Cys142 located on a alpha-helix flanking the beta layer.

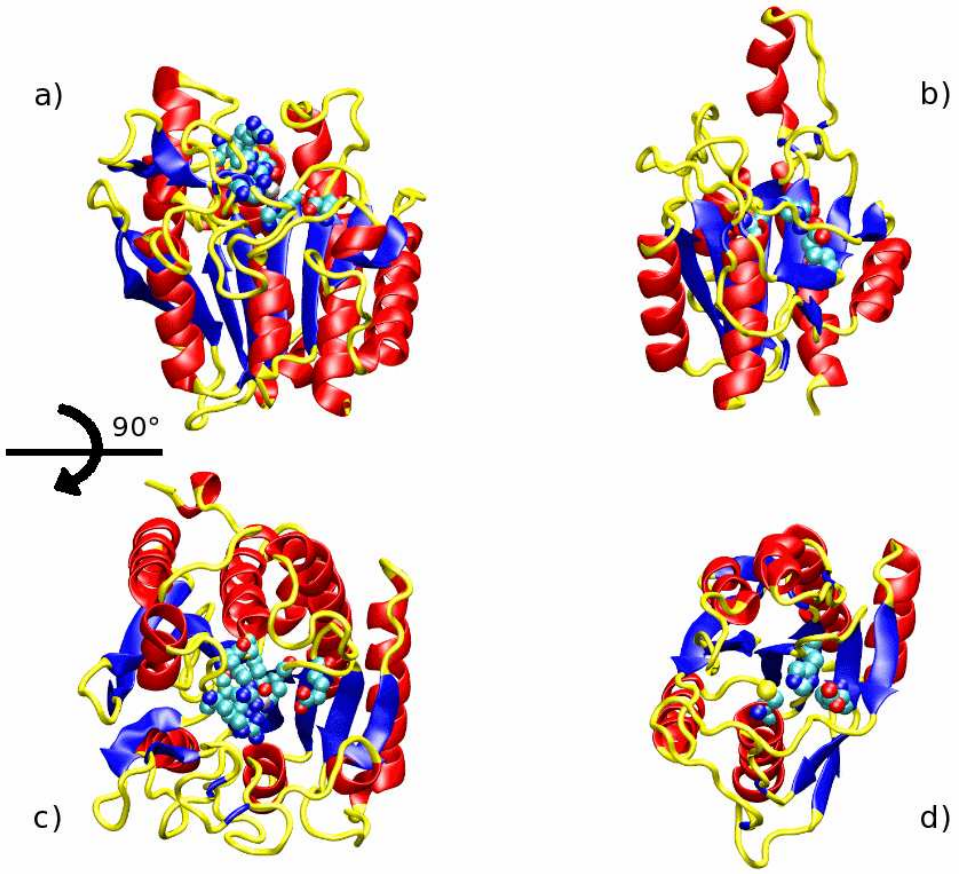


Figure 5.1: Cartoon representations of the two considered protease representatives: (a) carboxypeptidase A and (b) pyroglutamyl peptidase, PDB codes 8cpa [76] and 1iof [120], respectively. The bottom panel presents a different view of the same structures. The catalytic residues are shown with atomistic detail.

5.2 Dynamics-based comparison of the two enzymes

We shall now discuss in detail how the native structure of the two enzymes can be exploited to gain insight into their putative functional movements and to highlight correspondences in the dynamical behavior of the two enzymes.

We shall first describe the collective movements of the two enzymes analyzed separately. Then, by means of a structural alignment procedure (DALI [62]) we will select two subset of the residues to be put in one-to-one correspondence and we will compute the lowest energy modes regarding that regions of the two proteins, performing a thermodynamic integration on the not-matching residues as explained in the previous chapter. We will then

analyse the comparison devising a measure that gives the contribution of the single amino acids to the dynamical matching and highlighting the most relevant consensus fluctuations.

5.2.1 Large scale movements of the single enzymes

To compute the essential modes we will make use of the Beta Gaussian Model introduced in ref [93] and briefly discussed in Appendix D. Through this model we have calculated separately the lowest energy modes of carboxypeptidase A and pyroglutamyl peptidase. The two lowest energy modes thus identified for each enzyme are illustrated in Fig. 5.2.

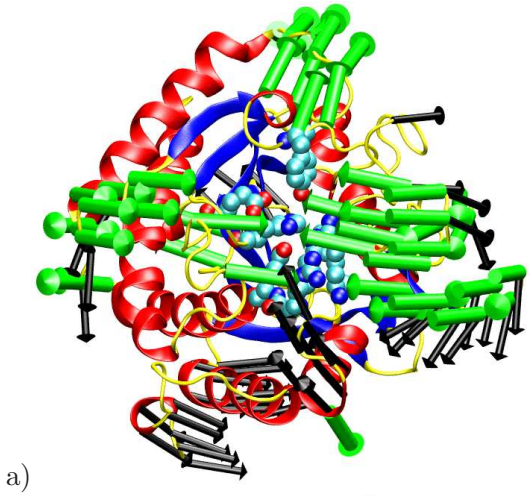
For carboxypeptidase A, the modes appear mostly localized in the regions delimiting the binding site. In particular they result in a contraction/elongation of the distances between the two unstructured regions (from Ser157 to Tyr169 and from Thr274 to Phe279) and between the top of one helix (from Phe118 to Leu125) and the the loop from Ile244 to Gln249. It is interesting to notice that the latter is constituted by residues involved in enzyme/substrate interactions. The overall behavior suggests the pictorial representation of a "breathing" motion of the binding site.

In pyroglutamyl peptidase six long loops embrace the binding pocket (three on each side) resulting in an active site cleft flanked by two lobes. Most of the movements entailed by the lowest energy modes are concentrated in these regions. The concerted motion of two lobes can be aptly described in terms of a bend (first mode) and shear (second mode) motion. It is interesting to note that, as for carboxypeptidase A, there is a modulation of the overall shape of the cleft.

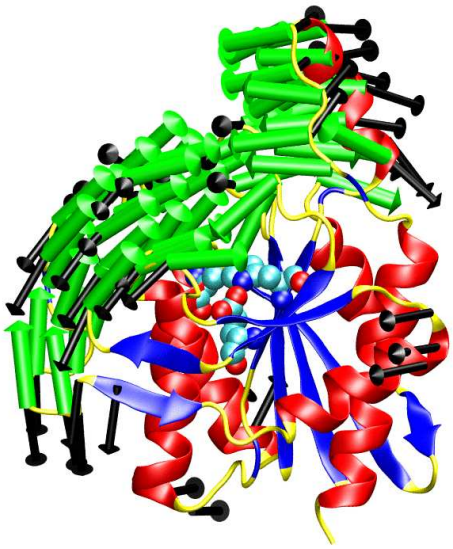
We will address in a systematic way these apparent qualitative similarities within the scheme outlined above.

5.2.2 Partial structural alignment

The first step of the dynamical comparison procedure relies on the identification of sets of amino acids in the two enzymes that can be put in structural correspondence. For the two enzymes considered here, the conformational difference, though not impacting on the overall architecture, is significant both for the secondary content and enzyme length. This is sufficient to call for an automated scheme to single out the structural analogies which, by necessity, cannot encompass the biomolecules in their entirety. This will identify a common structural core used for comparing the fluctuation dynamics. We shall then address the extent to which structural correspondences impact on dynamical analogies and discuss the biological implications. As already discussed both aspects are far for being a trivial consequence of the partial structural matching. In fact the structural correspondence of a subset of the two structures does not guarantee their



a)



b)

Figure 5.2: The most mobile residues in the first [second] lowest energy mode for (a) carboxypeptidase A and (b) pyroglutamyl peptidase are shown with green thick [black thin] arrows.

dynamical accord, as the low energy modes are significantly influenced by the remaining part of the protein. The search for the best *partial* structural match of the enzymes is aptly performed with the DALI algorithm [62]. The algorithm is based on a scoring function (formulated in terms of the matrix of pairwise residue distances of each enzyme) that quantifies the geometrical consistency of any given set of corresponding residues in the two enzymes. The best partial structural alignment is identified by a stochastic optimization of the scoring function. Typical enzymatic DALI structural alignments

lead to identify several corresponding blocks of consecutive residues, each involving 10-15 amino acids. It is clear that, owing to the pervasive presence of secondary motifs in proteins, the search of partial structural matches in two proteins will almost always be successful. Discriminating between meaningful structural alignments and "accidental" ones is therefore a key point of the analysis which entails a statistical significance test. Indeed, for each DALI alignment, the optimised value of the scoring function is compared against a reference distribution of scores expected for two generic enzymes of length equal to the assigned ones. The statistical significance of the best DALI alignment is finally summarised in a Z-score which measures the number of standard deviations by which the optimal score exceeds the reference one. The better the alignment, the larger the Z-score.

For pyroglutamyl peptidase and carboxypeptidase A, their DALI Z-score was equal to 8.9, which indicates that they can be aligned with high statistical confidence. The optimal partial alignment involves 151 residues out of 208 [307] for pyroglutamyl peptidase [carboxypeptidase A]. The alignment has a simple character in that aligned blocks have the same succession in the two enzymes (and no inversion of sequence directionality occurs). This simple organization allows to represent the alignment with the simple graphical plot of Fig. 5.3.

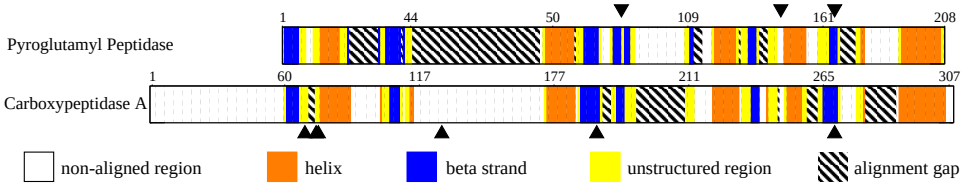


Figure 5.3: Primary structures alignment induced by the optimal DALI structural superposition of the two proteases. Each primary sequence is colored with a scheme related to the secondary structure (helices in orange, beta strands in blue and unstructured regions in yellow). Blank regions represent non-aligned regions while alignment gaps are represented with diagonal lines. The triangles mark the location of the catalytic triad for pyroglutamyl peptidase and of the five residues coordinating the Zn ion for carboxypeptidase A

.

As visible in the figure the first 60 residues of carboxypeptidase A are excluded from the alignment. The first, third, fourth, fifth, eight, ninth beta strand of pyroglutamyl peptidase matches respectively with the third, fourth, fifth, sixth, seventh and eight of carboxypeptidase A; helices 1, 2,3,4,6 of pyroglutamyl peptidase are in correspondence with helices 2,5,6,8,9 of carboxypeptidase A; helix 5 of pyroglutamyl peptidase has a partial match

with helix 9 of carboxypeptidase A. Also, there are non-trivial correspondence between different secondary structure elements: beta strand 2 of pyroglutamyl peptidase matches with a fragment of helices 2 and 3 of carboxypeptidase A, parts of beta strands 3, 5, 6 and 7 of pyroglutamyl peptidase match into loops in the partner structure, as well as segments of helices 2, 4,7 and 9 of carboxypeptidase A. A view of the aligned structures is given in Fig. 5.4.

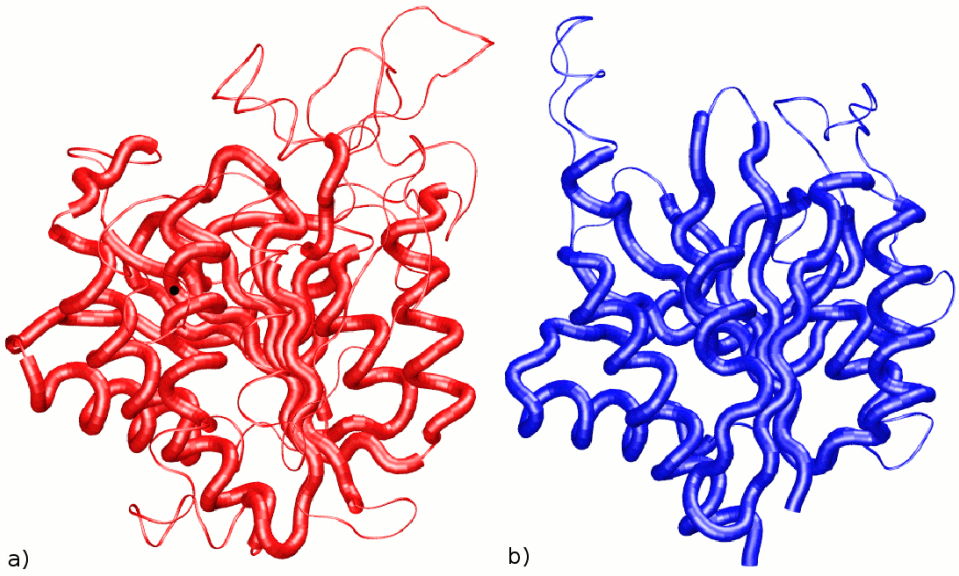


Figure 5.4: Backbone traces of (a) carboxypeptidase A and (b) pyroglutamyl peptidase. The matching regions are highlighted with a thick CA trace.

5.2.3 Dynamical Comparison

We are interested in calculating the concerted displacements of the sole structurally-aligned residues, yet taking into account the dynamical influence exerted over them by the non-aligned ones. We assume that the proteins are represented in the Cartesian reference frame providing the optimal structural superposition of the DALI matching regions. For each of the two proteins, following the procedure and the notation outlined in the previous chapter, we will identify with $\vec{x}_1$ the displacement of the coordinates of the structurally matching residues, and with $\vec{x}_2$ the remaining degrees of freedom f the system. The effective free energy $\mathcal{F}(\vec{x}_1)$ as a function of the coordinates $\vec{x}_1$ reads:

$$\mathcal{F}(\vec{x}_1) = \vec{x}_1^T \tilde{\mathbf{F}}_{x_1} \vec{x}_1 = \vec{x}_1^T \left(\mathbf{F}_{x_1} - \mathbf{G}\, \mathbf{F}_{x_2}^{-1} \mathbf{G}^\dagger \right) \vec{x}_1 \ . \tag{5.1}$$

The eigenvectors associated with the smallest eigenvalues of $\tilde{\mathbf{F}}_{x_1}$ represent the integrated lowest energy modes of the matching regions.

The eigenvectors of $\tilde{\mathbf{F}}_{x_1}$, calculated separately for the two enzymes after an optimal structural superposition of the DALI matching regions, can be directly compared component by component. To quantify the agreement of the integrated dynamics we applied the following heuristic procedure which straightforwardly leads to define a novel measure of dynamical accord which generalises the RMSIP value [3].

Indicating the effective free energy for the two proteins A and B as $\tilde{\mathcal{F}}_A$ and $\tilde{\mathcal{F}}_B$ we introduce infinitesimal harmonic couplings between corresponding residues in the two proteins. More precisely, we consider the following effective free energy function the combined system $\mathcal{F}_{A+B}$:

$$\mathcal{F}_{A+B} = (\vec{x}_{1,A}^{\,T}\ \vec{x}_{1,B}^{\,T}) \begin{pmatrix} \tilde{\mathbf{F}}_{x_1,A} & \epsilon\,\mathbf{1} \\ \epsilon\,\mathbf{1} & \tilde{\mathbf{F}}_{x_1,B} \end{pmatrix} \begin{pmatrix} \vec{x}_{1,A} \\ \vec{x}_{1,B} \end{pmatrix} \tag{5.2}$$

where ϵ indicates the strength of the harmonic coupling between corresponding residues and $\mathbf{1}$ indicates the identity matrix. We shall be concerned with the limit $\epsilon \to 0$, where the harmonic coupling becomes a weak perturbation. The coupling of eqn. (5.2) has an abstract character in that it introduces isotropic interactions between corresponding residues without reference to any particular notion of space proximity of residues in the two proteins. Due to its minimalistic nature, the free energy function of eqn. (5.2) does not have full spatial invariance properties, though it is still invariant for rotations of both proteins around the common center of mass of the matching regions (we recall that in our analysis we consider as only degrees of freedom of the system centroids coincident with the Cα atoms, assigning to all residues the same mass).

The introduction of the coupling between the matching residues facilitates the detection of similar large-scale fluctuations of corresponding residues in the two proteins. The information on the extent to which these correlations exist is aptly conveyed by the covariance matrix obtained by inverting the effective matrix. To leading order in ϵ the covariance matrix is given by

$$\begin{pmatrix} \tilde{\mathbf{F}}^{-1}_{x_1,A} & -\epsilon\tilde{\mathbf{F}}^{-1}_{x_1,A}\tilde{\mathbf{F}}^{-1}_{x_1,B} \\ -\epsilon\tilde{\mathbf{F}}^{-1}_{x_1,A}\tilde{\mathbf{F}}^{-1}_{x_1,B} & \tilde{\mathbf{F}}^{-1}_{x_1,B} \end{pmatrix}$$

According to this expression, the degree of dynamical correlation of the displacements of pairs of corresponding residues is provided by the diagonal terms of the submatrix, $(-\epsilon\tilde{\mathbf{F}}^{-1}_{x_1,A}\tilde{\mathbf{F}}^{-1}_{x_1,B})$. To turn this observation into a practical procedure it is convenient to express $\tilde{\mathbf{F}}_{x_1,A}$ and $\tilde{\mathbf{F}}_{x_1,B}$ in terms of their eigenvalues and eigenvectors. Indicating with $\vec{v}_i$ and $\vec{w}_i$ the ith eigenvector of the first protein and second protein, respectively (with associated

eigenvalues λ_i and μ_i) we have

$$\tilde{\mathbf{F}}^{-1}_{x_1,A} = \sum_l \lambda_l^{-1} \vec{v}_l^{\dagger} \vec{v}_l \qquad \tilde{\mathbf{F}}^{-1}_{x_1,B} = \sum_l \mu_l^{-1} \vec{w}_l^{\dagger} \vec{w}_l$$

Hence, the sum of the diagonal terms of $-\epsilon \tilde{\mathbf{F}}^{-1}_{x_1,A} \tilde{\mathbf{F}}^{-1}_{x_1,B}$ is equal to $\epsilon \sum_{l,m} \lambda_l^{-1} \mu_m^{-1} |\vec{v}^{\,l} \cdot \vec{w}^{\,m}|^2$. In case of perfect correspondence of both sets of ranked eigenvectors and eigenvalues, the previous quantity attains its maximum value that is $\sum_l \lambda_l^{-1} \mu_l^{-1}$

This observation allows to introduce a novel normalized measure for the agreement between two dynamical spaces, that we shall term the root weighted square inner product, RWSIP,

$$RWSIP = \sqrt{\frac{\sum_{l,m} \frac{1}{\lambda_l} \frac{1}{\mu_m} |\vec{v}_l \cdot \vec{w}_m|^2}{\sum_l \frac{1}{\lambda_l \mu_l}}} \tag{5.3}$$

By comparison with the familiar root mean square inner product (RMSIP) expression,

$$RMSIP = \sqrt{\sum_{l,m=1,..,10} |\vec{v}_l \cdot \vec{w}_m|^2 / 10} \ , \tag{5.4}$$

it can be observed that the RWSIP includes information about the eigenvalues of the effective energy matrix (i.e. the inverse eigenvalues of the covariance matrix). Hence it provides a more stringent and comprehensive account of the degree of accord of two dynamical spaces, and avoids the introduction of a subjective limit to the number of essential eigenvectors to keep.

For the two enzymes under consideration the RWSIP of the DALI aligned region was equal to 0.875. To assess the statistical significance of this value we have compared it with a reference distribution. The latter is aptly taken as the distribution of RWSIP values resulting by randomly choosing the residues in set 1, that is for arbitrary choices of the blocks of corresponding residues in the two structures. Accordingly, we stochastically generated 1000 "decoy" sets of matching residues involving the same number of amino acids (151) as the optimal DALI alignment of carboxypeptidase A and pyroglutamyl peptidase. Also the typical size of DALI matching blocks (10-15 residues) is respected in the control alignments. For each stochastic alignment we carried out numerically the dynamical integration described above and hence obtained the corresponding RWSIP value from equation (5.3). By processing the results of the 1000 decoy alignments we calculated the average value and dispersion of the control RWSIP distribution, $\langle \text{RWSIP} \rangle$ and ΔRWSIP. These quantities were used to define the *dynamical Z-score*: $(\text{RWSIP}_{DALI} - \langle \text{RWSIP} \rangle)/\Delta\text{RWSIP}$. In analogy to the structural Z-score, it provides a measure of how unlikely it is that the RWSIP of the DALI matching regions could have arisen by chance. The value obtained for RWSIP in

the DALI alignment was accordingly found to yield a dynamical Z-score of 8.98. The control distribution of RWSIP values is shown in Fig. 5.5.

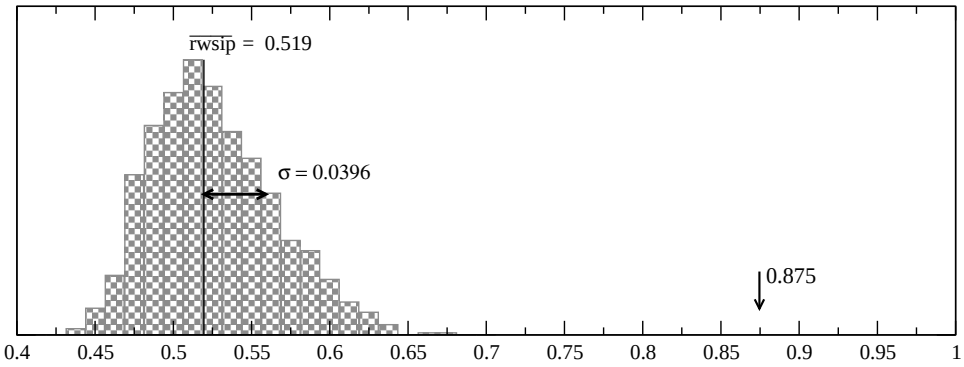


Figure 5.5: Normalized probability distribution of the RWSIP values computed on the 1000 randomly generated alignments. The arrow indicates the RWSIP of the DALI alignment.

This suggest an "a posteriori" validation of the choice of the matching set. The statistically significant structural matching provided by DALI algorithm reflects in a significant overall accord of the essential dynamical spaces of the matching region. This observation contributes to support the functional significance of the DALI alignment and gives first indication of the functional importance of the common dynamics. We will now address in more detail the particular importance of the motion of some specific residues important for catalysis, and discuss the connection of the dynamical accord of the single residues with their spatial pairwise correspondence.

The heuristic approach followed here to derive the RWSIP measure lends also to a transparent criterion for isolating the individual contributions of corresponding amino acids in the protein. In fact, the weighted inner product is $WSIP \propto \sum_i q_i$ where

$$q_i = \sum_{l,m} \frac{1}{\lambda_l} \frac{1}{\mu_m} \vec{v}_l{}^i \cdot \vec{w}_m{}^i (\vec{v}_l \cdot \vec{w}_m) / \sum_l \frac{1}{\lambda_l \mu_l} \ . \tag{5.5}$$

Where we have indicated with $\vec{v}_l{}^i$ the three dimensional vector, part of the lth eigenvector of $\tilde{\mathbf{F}}_{x_1,A}$, referring to the residue i. For the two enzymes under consideration, the non-normalised profile of q_i is illustrated in Fig. 5.6. The profile provides a qualitative account of the dynamical importance of the residues contributing to the dynamical accord. Values of q_i are plotted in Fig. 5.6 with colors reflecting the subdivision in three sets of relevance. The high inhomogeneity of the profile complements appropriately the information about the structural alignment where all aligned residues are treated on equal footing.

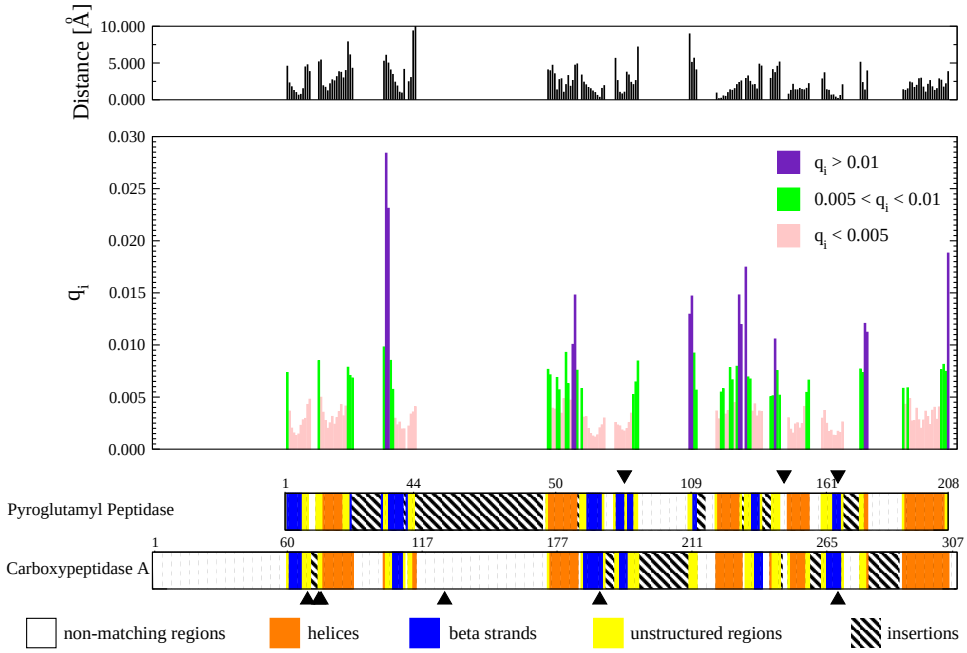


Figure 5.6: The three dimensional distance of the aligned residues in the two proteins is shown in the upper panel. The q_i profile over DALI aligned residues (middle panel) is shown along with the primary alignment induced by the optimal DALI superposition. The following color code was used to partition residues according to the value of q_i: $q_i \leq 0.005$ (pink), $0.005 < q_i \leq 0.01$ (green) and $q_i > 0.01$ (violet).

It is interesting to compare the dynamical matching q_i of the single residues, with the corresponding three dimensional distance of the residues in the two proteins. The profile of the pairwise spatial distances is portrayed in top panel of Fig 5.6. No obvious correspondence is perceivable between the peaks of dynamical similarities and the regions in close space proximity. This provides a heuristic indication of the impact that non-aligned residues have over the dynamics of the aligned ones.

We will address the biological implications of the observed dynamical accord. In Fig. 5.7 the first two relevant eigenvectors of the reduced covariance matrix are represented as arrows (green thick and black thin respectively), for the top 50 fluctuating residues of each structure. In each of the two enzymes it is possible to identify two halves of the enzymes which undergo rotatory fluctuations. The resulting shear motion may consequently produce a mechanical stress of the bound substrate.

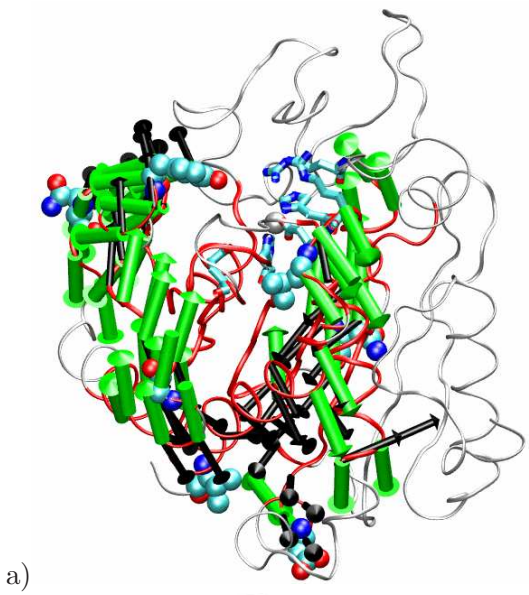
In the picture we highlighted Van der Waals volumes of seven residues for each structure, representing the seven most relevant peaks in the dynamical alignment (in violet in Fig. 5.6). It's interesting to note that among them

there are the most mobile residues, and some residues close to the catalytic site that modulates the shape of the active site, in particular Tyr238 in carboxypeptidase A, whose role has been proved to be crucial for the substrate binding and/or processing [31], that matches Gly138 of pyroglutamyl peptidase, which is in a crucial position (it surmounts the catalytic cysteine) and contributes to enzyme/substrate interactions [11] In particular, this Glycine is in the exact position of the Glycine of the carboxyanion hole in serine proteases, thus suggesting the key functional role played by this residue also for pyroglutamyl peptidase.

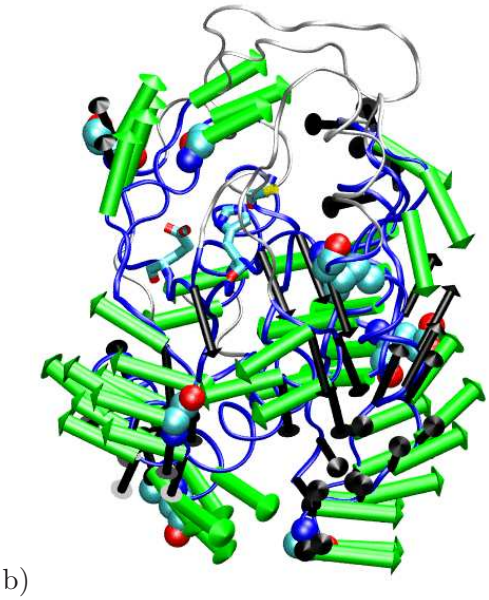
5.3 Summary

We have presented a quantitative scheme through which the essential dynamical spaces of proteins with partial different structural arrangement can be compared. The method relies on the detection of a partial structural correspondences between two biomolecules. This is motivated by the necessity of finding a common reference frame, whose internal dynamics could be functionally relevant. By means of a number of techniques, partly developed specifically for this problem, the large scale concerted movements of the structurally aligned regions have been calculated and finally compared in terms of a novel quantitative measure. This dynamical measure lends straightforwardly to identifying the extent to which different residues (all in structural correspondence) contribute to the overall dynamical consistency.

This scheme was applied to two representatives of the protease enzymatic superfamily, carboxypeptidase A and pyroglutamyl peptidase. Considering these enzymes is particularly instructive as, besides having different sequence, length and secondary content, they also rely on a different catalytic chemistry. The first feature emerging from the partial structural alignment is that the 151 residues taking part to the optimal superposition occupy a region within $\sim$30 Å of active site cleft. The lowest energy modes calculated over this region are in remarkable accord across the two proteins (despite including the dynamical influence of the regions that are not in structural correspondence). The largest contribution to the dynamical accord arises from two patches of residues that appear to be capable of modulating the cleft of the active site. As previously elucidated for other members of the protease enzymatic superfamily [27, 26], the findings provide a strong indication of how the biological selective pressure for efficient cleavage of peptide substrate has promoted not only similar structural architectures in the neighborhood of the active site, but also consistent concerted movements that putatively accompany and facilitate the substrate recognition or cleavage.



a)



b)

Figure 5.7: The most mobile residues in the first [second] lowest energy mode for the *aligned residues* of (a) carboxypeptidase A and (b) pyroglutamyl peptidase are shown with green thick [black thin] arrows. The structurally-aligned regions are shown with a thick CA trace.

Summary

In the first part of the thesis we have discussed the MD-based investigation of the near-native free energy landscape for two well-known proteins: GB1 and adenylate kinase. By analysis of the salient internal dynamics features collected over hundreds of ns it has been found that the near-native free-energy possesses a self-similar structure in that principal directions of the explored free-energy local minima and the virtual jumps that connect them are highly consistent. This remarkable feature reflects into the exceptional robustness of the essential dynamical spaces of the protein. These consensus generalised directions of internal motion, oblivious of the structural details differentiating the various substates, reflect an intrinsic protein of the protein, arguably encoded in its structural architecture. The observed consistency provides a very efficient means for the system to exploiting thermal fluctuations to diffuse within and across the substates. The functional relevance of the consesus essential dynamical spaces is suggested by the fact that these "innate" essential dynamics has a very good overlap with the difference vector connecting the available free/bound crystal structures of adenylate kinase. The functionally-oriented character of proteins internal dynamics, robustly encoded in their structure, leads to speculate that this property may have been promoted by evolutionary pressure, consistently with recent suggestions made on the basis of experimental data.

These results have lead us to investigate further how various degree of overall or partial structural similarities in proteins reflect in the consistency of their internal dynamics. The findings of our investigations have been outlined in the second part of the thesis.

Specifically, we have quantitatively compared structural and dynamical similarities of a calcium binding domain, the so called EF-Hand domain, which is shared by a large number of proteins. We have performed an analysis of the internal essential dynamics of the EF-hand domains for a data-set of more than 150 domains coming from structurally- or functionally-different proteins. Interesting correlations emerges between the essential dynamics of the domains and their functional classification while a non-stringent relationship exists between the degree of structural and dynamical similarity. This second aspect is presumably ascribable both to the detailed local structural differences of the domains and to the influence of the different

global arrangements of the proteins.

We have finally discussed a general scheme through which the essential dynamics of proteins with only partial structural correspondence can be compared. In particular we have calculated and compared the large-scale concerted movements of structurally-alignable portions of two proteins belonging to the protease superfamily. The two proteins, despite sharing a common structural core, differ in length and in the overall structural arrangement, and moreover belong to different clans of the superfamily, and hence rely on different catalytic chemistry. The lowest-energy modes calculated over the structurally-corresponding region are in remarkable accord across the two proteins despite including the dynamical influence of the regions that are not in structural correspondence. The consensus directions of motion describe a modulation of the cleft of the active site of the enzymes. The findings provide indications of the fact that these consistent concerted movements, that putatively accompany and facilitate the substrate recognition or cleavage, may have been evolutionary selected.

Appendix A

Comparing sets of Essential-Dynamics Spaces: Optimal Identification of their Consensus Subspace

Consider the vectorial spaces V and W, spanned by the top N essential dynamical spaces, $\{v_1, v_2, \ldots, v_N\}$ and $\{w_1, w_2, \ldots, w_N\}$ respectively, of two MD trajectories of the same protein, e.g. starting from different initial configurations. We wish to establish if, or to what approximation, V and W share a common subspace. The problem amounts to find new orthonormal basis vectors for V and W, $\{v'_1, v'_2, \ldots, v'_N\}$ and $\{w'_1, w'_2, \ldots, w'_N\}$ respectively, which are ranked with decreasing mutual consistency. In principle, this could be accomplished through an iterative procedure where the first pair of vectors, v'_1 (belonging to V) and w'_1 (belonging to W), is picked so to have the largest possible scalar product. This optimal selection procedure is next repeated in the remaining complementary spaces of V and W and so on. The sought pairs of vectors, v'_i and w'_i are such to make stationary the following functional

$$f(v'_i, w'_i) = \langle w'_i | v'_i \rangle - \alpha_i \langle v'_i | v'_i \rangle - \beta_i \langle w'_i | w'_i \rangle \tag{A.1}$$

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

$$f(\vec{a_i}, \vec{b_i}) \;=\; \vec{b_i} \cdot C \vec{a_i} - \alpha \, \vec{a_i} \cdot \vec{a_i} - \beta \, \vec{b_i} \cdot \vec{b_i} \, . \tag{A.2}$$

The stationary condition gives the following set of eigenvalue equations:

$$C^T C\ \vec{a_i} = \lambda_i \vec{a_i} \tag{A.3}$$
$$CC^T\ \vec{b_i} = \lambda_i \vec{b_i} \tag{A.4}$$

with $i = 1, \ldots N$, $\vec{a_i}$ and $\vec{b_i}$ are vectors with unit norm, and the coefficient λ_i equals $4\alpha_i\beta_i$.

It's important to note that the two solutions are not independent. Assuming we have a solution $\vec{a_i}$ for (A.3); then it's easy to see that $\vec{b_i} = \frac{1}{\sqrt{\lambda_i}} C\vec{a_i}$ is a solution for (A.4), and the scalar product of the vectors v_i' and w_i' associated to this solution is $\langle w_i'|v_i'\rangle = \sqrt{\lambda_i}$. As $C^T C$ is an NxN symmetric matrix, we have a complete solution to the eigenproblem of equation (A.3).

Let's consider the non-degenerate case with $\lambda_i \neq \lambda_j \,\forall\, i \neq j$ and order the eigenvalues in descending order $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Vectors v_i' and w_i' are defined by the i^{th} solution of (A.3), as follows:

$$|v_i'\rangle = \sum_{j=1}^{N} A_{i,j}|v_j\rangle \qquad |w_i'\rangle = \sum_{j=1}^{N} B_{i,j}|w_j\rangle \tag{A.5}$$

and their scalar product is $\sqrt{\lambda_i}$. Notice also that $\langle w_i'|v_j'\rangle = \sqrt{\lambda_i}\delta_{ij}$ in case of no degeneration in solutions of (A.3).

Appendix B

Thermodynamical Integration of Degrees of Freedom in Elastic Network Models

Let $P(\vec{x})$ be a probability distribution for the values of the system coordinates $\vec{x}$ and suppose we are interested in calculating the marginal distribution of a set of coordinates $\vec{y}$ obtained as an orthogonal transformation $\mathbf{A}$ of a subset $\vec{x}_1$ of the system coordinates $\vec{x}$.

$$P(\vec{y}(\vec{x}_1)) = \int d\vec{x}_2 \frac{P(\mathbf{A}^{-1}\vec{y}, \vec{x_2})}{|Det(\mathbf{A})|}. \tag{B.1}$$

Let us further assume that the system thermodynamics is governed by a free energy $\mathcal{F}$ that can be represented as a quadratic function of the original coordinates $\vec{x}$,

$$\mathcal{F}(\vec{x_1}, \vec{x_2}) = \vec{x}^T \mathbf{F} \vec{x} = (\vec{x_1}, \vec{x_2})^T \begin{pmatrix} \mathbf{F}_1 & \mathbf{W} \\ \mathbf{W}^T & \mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \vec{x_1} \\ \vec{x_2} \end{pmatrix} \tag{B.2}$$

The blocks in the interaction matrix $\mathbf{F}$ have a straightforward meaning: $\mathbf{F}_1$ [$\mathbf{F}_2$] represent the internal coupling between the subset of coordinates $\vec{x_1}$ [$\vec{x_2}$] while $\mathbf{W}$ couple coordinates between the subset $\vec{x_1}$ and $\vec{x_2}$. The probability distribution $P(\vec{x})$ for the coordinates $\vec{x}$ is connected to the Free Energy through the Boltzmann distribution:

$$P(\vec{x}) \propto \exp(-\frac{\mathcal{F}}{K_B T}) \tag{B.3}$$

Let us consider the expression for $P(\mathbf{A}^{-1}\vec{y}, \vec{x_2})$ and expand explicitly $\mathcal{F}(\mathbf{A}^{-1}\vec{y}, \vec{x_2})$

$$\mathcal{F}(\mathbf{A}^{-1}\vec{y},\vec{x_2}) = \vec{y}^T\mathbf{A}\mathbf{F}_1\mathbf{A}^{-1}\vec{y} + \vec{y}^T\mathbf{A}\mathbf{W}\vec{x_2} + \vec{x_2}^T\mathbf{W}^T\mathbf{A}^{-1}\vec{y} + \vec{x_2}^T\mathbf{F}_2\vec{x_2} \quad \text{(B.4)}$$

The expression can be recasted as

$$\mathcal{F}(\mathbf{A}^{-1}\vec{y},\vec{x_2}) = \vec{y}^T\mathbf{F}_y\vec{y} + 2\vec{y}^T\mathbf{W}_y\vec{x_2} + (\sqrt{\mathbf{F}_2}\vec{x_2})^T(\sqrt{\mathbf{F}_2}\vec{x_2}) \quad \text{(B.5)}$$

where $\mathbf{F}_y = \mathbf{A}\mathbf{F}_1\mathbf{A}^{-1}$ and $\mathbf{W}_y = \mathbf{A}\mathbf{W}$. Now, simply noting that $\vec{y}^T\mathbf{W}_y\vec{x_2}$ is equal to $(\sqrt{\mathbf{F}_2}^{-1}\mathbf{W}_y^T\vec{y})^T(\sqrt{\mathbf{F}_2}\vec{x_2})$, we obtain by completion of the square

$$\mathbf{F}_y = \vec{y}^T(\mathbf{F}_y - \mathbf{W}_y^T\mathbf{F}_2^{-1}\mathbf{W}_y)\vec{y} - (\sqrt{\mathbf{F}_2}^{-1}\mathbf{W}_y^T\vec{y} + \sqrt{\mathbf{F}_2}\vec{x_2})^T(\sqrt{\mathbf{F}_2}^{-1}\mathbf{W}_y^T\vec{y} + \sqrt{\mathbf{F}_2}\vec{x_2}) \quad \text{(B.6)}$$

The Boltzmann factor can be factorized as

$$e^{\frac{-\vec{y}^T(\mathbf{F}_y - \mathbf{W}_y^T\mathbf{F}_2{}^{-1}\mathbf{W}_y)\vec{y}}{K_BT}} \cdot e^{-\frac{(\sqrt{\mathbf{F}_2}^{-1}\mathbf{W}_y^T\vec{y} + \sqrt{\mathbf{F}_2}\vec{x_2})^T(\sqrt{\mathbf{F}_2}^{-1}\mathbf{W}_y^T\vec{y} + \sqrt{\mathbf{F}_2}\vec{x_2})}{K_BT}} \quad \text{(B.7)}$$

When we integrate with respect to $\vec{x_2}$ to obtain $P(\vec{y})$ the second exponential term in the above expression is a Gaussian integral contributing only by a multiplicative constant. Thus, apart a from multiplicative normalization factor, $P(\vec{y})$ reads

$$P(\vec{y}) \propto \exp\left[-\left(\frac{\vec{y}^T(\mathbf{F}_y - \mathbf{W}_y^T\mathbf{F}_2{}^{-1}\mathbf{W}_y)\vec{y}}{K_BT}\right)\right] \quad \text{(B.8)}$$

The effective free energy described as a function of the new reduced variables $\vec{y}$ is still a quadratic function of the coordinates $\vec{y}$.

Appendix C

Functional Groups of EF-HAND Domains

Detailed information on members of functional groups reported in Fig. 4.7
Key:
Column 1: Cluster ID
Column 2: pdb code, chain index and domain
Column 3: Protein terminus, calcium state and peptide binding state
Column 4: Group ID as per indexing of the density plot of Fig. 4.7
Column 5: Protein name according to Swiss Protein database
Column 6: Complete protein name

Table C.1: Members of the 12 mostly populated functional families reported in Fig. 4.7.

Cluster ID	PDB	Structural Information	Family ID	Swiss prot	Protein name
1	1cll01	N 2Ca P0	1	CALM_HUMAN	Calmodulin
1	1clm01	N 2Ca P0	1	CALM_PARTE	Calmodulin
1	1ggzA1	N 2Ca P0	1	CALL_HUMAN	Calmodulin-like
1	1j7oA0	N 2Ca P0	1	CALM_HUMAN	Calmodulin
1	1oojA1	N 2Ca P0	1	CALM_CAEEL	Calmodulin
1	1osa01	N 2Ca P0	1	CALM_PARTE	Calmodulin
1	1rfjA1	N 2Ca P0	1	CALM_SOLTU	Calmodulin
2	1cffA1	N 2Ca P0	1	CALM_HUMAN	Calmodulin
0	1ckkA1	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	1cdlA1	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	1g4yR1	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	1iq5A1	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	1mxeA1	N 2Ca P2	2	CALM_DROME	Calmodulin
1	1nwdA1	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	2bbmA1	N 2Ca P2	2	CALM_DROME	Calmodulin
1	3cln01	N 2Ca P2	2	CALM_HUMAN	Calmodulin
1	1b7tY1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1kk7Y1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1kk8B1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1kwoB1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1l2oB1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1qviY1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
1	1wdcB1	N 1Mg1 P2	3	MLR_AEQIR	Myosin regulatory light chain
0	1cmg00	C 2Ca P0	4	CALM_HUMAN	Calmodulin
1	1cll02	C 2Ca P0	4	CALM_HUMAN	Calmodulin
1	1clm02	C 2Ca P0	4	CALM_PARTE	Calmodulin
1	1ggzA2	C 2Ca P0	4	CALL_HUMAN	Calmodulin-like
1	1j7pA0	C 2Ca P0	4	CALM_HUMAN	Calmodulin
1	1osa02	C 2Ca P0	4	CALM_PARTE	Calmodulin
1	1rfjA2	C 2Ca P0	4	CALM_SOLTU	Calmodulin
1	1k90D2	C 2Ca P2	5	CALM_HUMAN	Calmodulin
1	1k93D2	C 2Ca P2	5	CALM_HUMAN	Calmodulin

Continued on next page

Table C.1 – continued from previous page

Cluster ID	PDB	Structural Information	Family ID	Swiss prot	Protein name
1	1mxeA2	C 2Ca P2	5	CALM_DROME	Calmodulin
1	1nwdA2	C 2Ca P2	5	CALM_HUMAN	Calmodulin
1	1sk6D2	C 2Ca P2	5	CALM_HUMAN	Calmodulin
0	1kk7Y2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
0	1l2oB2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
0	1s5gY2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
0	1sr6B2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
1	1kwoB2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
1	1wdcB2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
2	1b7tY2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
2	1dfkY2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
2	1kqmB2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
2	1qviY2	C 0Ca P2	6	MLR_AEQIR	Myosin regulatory light chain
0	1kqmC2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
0	1kwoC2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
0	1l2oC2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
0	1s5gZ2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
0	1sr6C2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
2	1b7tZ2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
2	1dfkZ2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
2	1kk7Z2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
2	1kk8C2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
2	1wdcC2	C 0Ca P2	7	MLE_AEQIR	Myosin essential light chain
0	1e8aA0	N 2Ca P0	8	S112	S100
0	1gqmA0	N 2Ca P0	8	S112_HUMAN	S100
0	1k96A0	N 2Ca P0	8	S106_HUMAN	S100
0	1k9kA0	N 2Ca P0	8	S106_HUMAN	S100
0	1mho00	N 2Ca P0	8	S10B_BOVIN	S100
0	1mr8A0	N 2Ca P0	8	S108_HUMAN	S100
0	1qlkA0	N 2Ca P0	8	S10B_RAT	S100
0	1uwoA0	N 2Ca P0	8	S10B_HUMAN	S100
3	1b1gA0	N 2Ca P0	8	S10D_BOVIN	S100
3	1alvA2	C 2Ca P0	9	CANS_PIG	Calpain small subunit
3	1dviA2	C 2Ca P0	9	CANS_RAT	Calpain small subunit
3	1nx1A2	C 2Ca P0	9	CANS_PIG	Calpain small subunit
3	1nx2A2	C 2Ca P0	9	CANS_PIG	Calpain small subunit
0	1skt00	N 0Ca P0	10	TPCS_CHICK	Troponin C skeletal
0	1tnp00	N 0Ca P0	10	TPCS_CHICK	Troponin C skeletal
0	1zac00	N 0Ca P0	10	TPCS_CHICK	Troponin C skeletal
0	1trf00	N 0Ca P0	10	TPCS_MELGA	Troponin C skeletal
3	4tnc01	N 0Ca P0	10	TPCS_CHICK	Troponin C skeletal
0	1a4pB0	N 0Ca P0	11	S110_HUMAN	S100
0	1clb00	N 0Ca P0	11	S100 monomeric	S100
0	1nshA0	N 0Ca P0	11	S111_RABIT	S100
2	1b4cA0	N 0Ca P0	11	S10B_RAT	S100
2	1cfpA0	N 0Ca P0	11	S10B_BOVIN	S100
2	1k2hA0	N 0Ca P0	11	S10A1_RAT	S100
1	1dmo02	C 0Ca P0	12	CALM_HUMAN	Calmodulin
2	1cfd02	C 0Ca P0	12	CALM_HUMAN	Calmodulin
2	1f71A0	C 0Ca P0	12	CALM_HUMAN	Calmodulin
2	1g4yR2	C 0Ca P0	12	CALM_HUMAN	Calmodulin

Appendix D

Beta Gaussian Model

The Beta Gaussian model is a simplified Elastic Network Model in which the protein is represented by means of two-centroid per amino acid, one for the main-chain, coinciding with the CA atom, and one for the side-chain. Following a geometrical rule akin to the one introduced by Park and Levitt [108] we construct the latter interaction center as a fictitious CB centroid:

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

where $l = 3$Å and $\vec{r}_{CA}$ indicates the coordinates of the ith CA centroid. For amino acids at the beginning/end of the peptide chain(s) or for GLY the construction of eqn. D.1 is not applicable and hence the effective CB centroid is taken to coincide with the CA one.

A schematic view of the coarse graining procedure is given in Figs. D.1a, b and c.

The potential governing the interaction between the centroids is obtained by introducing the following quadratic penalties for displacing two centroids, i and j from their reference positions,$\vec{r}_i^{\,0}$ and $\vec{r}_j^{\,0}$ to generic ones, $\vec{r}_i$ and $\vec{r}_j$:

$$V(\vec{r}_{ij}) = K\sum_{\mu,\nu}\frac{r^0_{ij,\mu}\,r^0_{ij,\nu}}{|\vec{r}^{\,0}_{ij}|^2}\;\delta r_{ij,\mu}\,\delta r_{ij,\nu} \tag{D.2}$$

where $\vec{r}^{\,0}_{ij} \equiv (\vec{r}^{\,0}_i - \vec{r}^{\,0}_j)$ is the native distance vector of the centroids, $\delta\vec{r}_{ij}$ is the distance vector change, $\delta\vec{r}_{ij} \equiv (\vec{r}_i - \vec{r}^{\,0}_i) - (\vec{r}_j - \vec{r}^{\,0}_j)$, μ and ν run over the three Cartesian components and k is a parameter controlling the strength of the quadratic coupling.

The quadratic form of eqn. (D.2), already discussed in Introduction to Part I (see 1.4), is at the heart of the widely-used elastic or Gaussian network approaches [123, 10, 59, 92, 8, 95, 36, 93], which typically adopt a single-centroid amino acid description. The effective free energy function introduced in ref. [93] and used here includes, instead, pairwise contributions

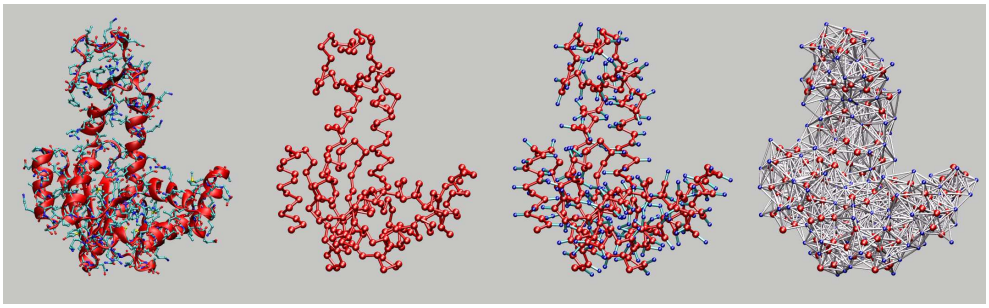


Figure D.1: Pictorial representation of the coarse graining procedure: (a) atomic representation of Adenylate Kinase (backbone highlighted as a ribbon); (b) the CA trace is considered as starting point for the coarse graining procedure; (c) simplified structural representation in terms of the CA centroids for the backbone and the CB ones for the sidechains; (c) all pairs of non-consecutive centroids within 7.5 Å interact through an harmonic potential, schematically shown as a thin bond.

from all pairs of centroids, be they of the CA or CB type, whose reference distance falls within a given interaction cutoff, as pictorially illustrated in Fig. D.1c. Accordingly, the resulting free energy of a trial structure, Γ, takes on the form:

$$\mathcal{F}(\Gamma) = 2\sum_{i} V(\vec{r}_{i,i+1}^{CA-CA}) + {\sum_{i<j}}' V(\vec{r}_{i,j}^{CA-CA}) + {\sum_{i,j}}' V(\vec{r}_{i,j}^{CA-CB}) + {\sum_{i<j}}' V(\vec{r}_{i,j}^{CB-CB}) \tag{D.3}$$

where i and j run over the residue indices, $\vec{r}_{i,j}^{X-Y}$ denotes the distance vector of the centroids of type X and Y of residues i and j, respectively, and the prime denotes that the sum is restricted to the pairs whose reference separation is below the cutoff distance of 7.5 Å. Consistently with the spirit elastic network models and other approaches [122], the last three terms in eqn. D.3 have the same strength irrespective of the identity of the amino acids. The first term, on the other hand, accounts for the protein chain connectivity and has a double strength to reflect the geometrical constraints of the peptide chain.

As the positions of the CB centroids depend linearly on the coordinates of the CA ones, it is possible to analytically recast the expression (D.3) in the following quadratic form (akin the one discussed in Introduction to Part I) involving simply the CA degrees of freedom, retaining the same computational complexity of the single centroids model:

$$\mathcal{F} = \frac{1}{2} \sum_{ij,\mu\nu} \delta r_{i,\mu} \, M_{ij}^{\mu\nu} \, \delta r_{j,\nu} \, , \tag{D.4}$$

where $\delta\vec{r}_i = \vec{r}_i^{CA} - \vec{r}_i^{0CA}$ is the deviation of i-th CA centroid from the reference position and M is a symmetric matrix whose linear size is three times the number of residues in the protein. As already discussed (see eqn 1.8), within a Langevin scheme, the independent modes of structural relaxation in the protein corresponds to the eigenvectors of the matrix M.

Bibliography

[1] J. Adén and M. Wolf-Watz, *Nmr identification of transient complexes critical to adenylate kinase catalysis*, J Am Chem Soc, 129 (2007), pp. 14003–14012.

[2] M. P. Allen and D. J. Tildsley, *Computer Simulation of Liquids*, Clarendon Press, New York, NY, USA, first ed., 1988.

[3] A. Amadei, M. A. Ceruso, and A. Di Nola, *On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations*, Proteins, 36 (1999), pp. 419–424.

[4] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Essential dynamics of proteins*, Proteins, 17 (1993), pp. 412–425.

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

[6] C. Anfinsen, *Principles that govern the folding of protein chains*, Science, 181 (1973), pp. 223–230.

[7] 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 (2007), pp. 18496–18501.

[8] 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 (2001), pp. 505–515.

[9] E. Babini, I. Bertini, F. Capozzi, C. Luchinat, A. Quattrone, and M. Turano, *Principal component analysis of the conformational freedom within the ef-hand superfamily*, J Proteome Res, 4 (2005), pp. 1961–1971.

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

[11] A. J. Barrett, N. D. Rawlings, J. F. Woessner, and eds, *Handbook of Proteolytic Enzymes*, Elsevier, Amsterdam, second ed., 2004.

[12] H. Beach, R. Cole, M. L. Gill, and J. P. Loria, *Conservation of mus-ms enzyme motions in the apo- and substrate-mimicked state*, J Am Chem Soc, 127 (2005), pp. 9167–9176.

[13] H. Berendsen, J. Postma, W. Van Gunsteren, and J. Hermans, *Intermolecular forces*, B. Pullman (Ed.), 1981.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *Molecular dynamics with coupling to an external bath*, The Journal of Chemical Physics, 81 (1984), pp. 3684–3690.

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

[16] S. Bhattacharya, C. G. Bunick, and W. J. Chazin, *Target selectivity in ef-hand calcium binding proteins*, Biochim Biophys Acta, 1742 (2004), pp. 69–79.

[17] H. R. Bosshard, *Molecular recognition by induced fit: how fit is the concept?*, News Physiol Sci, 16 (2001), pp. 171–173.

[18] C. Branden and J. Tooze, *Introduction to protein structure*, Garland Publishing, New York, 1991.

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

[20] B. R. Brooks, D. Janezic, and M. Karplus, *Harmonic analysis of large systems I. metodology*, 16 (1995), pp. 1522–1542.

[21] F. Capozzi, F. Casadei, and C. Luchinat, *Ef-hand protein dynamics and evolution of calcium signal transduction: an nmr view*, J Biol Inorg Chem, 11 (2006), pp. 949–962.

[22] 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 (2007), pp. 4245–4255.

[23] E. Carafoli and C. B. Klee, *Calcium as Cellular Regulator*, Oxford University Press, New York, 1999.

[24] E. Carafoli, L. Santella, D. Branca, and M. Brini, *Generation, control, and processing of cellular calcium signals*, Crit Rev Biochem Mol Biol, 36 (2001), pp. 107–260.

[25] V. Carnevale, F. Pontiggia, and C. Micheletti, *Structural and dynamical alignment of enzymes with partial structural similarity*, Journal of Physics: Condensed Matter, 19 (2007), p. 285206 (14pp).

[26] V. Carnevale, S. Raugei, C. Micheletti, and P. Carloni, *Convergent dynamics in the protease enzymatic superfamily*, J Am Chem Soc, 128 (2006), pp. 9766–9772.

[27] 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 (2005), pp. 3734–3742.

[28] S. Chandrasekhar, *Stochastic problems in physics and astronomy*, Rev. Mod. Phys., 15 (1943), pp. 1–89.

[29] 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 (1997), pp. 3817–3822.

[30] C. Chennubhotla and I. Bahar, *Markov methods for hierarchical coarse-graining of large protein dynamics*, J Comput Biol, 14 (2007), pp. 765–776.

[31] J. H. Cho, D. H. Kim, K. J. Lee, and K. Y. Choi, *The role of tyr248 probed by mutant bovine carboxypeptidase a: insight into the catalytic mechanism of carboxypeptidase a*, Biochemistry, 40 (2001), pp. 10197–10203.

[32] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, *Evolution of the protein repertoire*, Science, 300 (2003), pp. 1701–1703.

[33] T. Creighton, *Proteins, structure and molecular properties*, W.H.Freeman and Company, New York, second ed., 1993.

[34] T. Darden, D. York, and L. Pedersen, *Particle mesh ewald: An n [center-dot] log(n) method for ewald sums in large systems*, The Journal of Chemical Physics, 98 (1993), pp. 10089–10092.

[35] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. van Gunsteren, and A. Mark, *Peptide folding: When simulation meets experiment*, Angewandte Chemie-International Edition, 38 (1999), pp. 236–240.

[36] 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 (2002), pp. 1011–1024.

[37] M. Doi, *Introduction to Polymer Physics*, Clarendon Press, Oxford, UK, first ed. ed., 1996.

[38] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, *Intrinsic dynamics of an enzyme underlies catalysis*, Nature, 438 (2005), pp. 117–121.

[39] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *A smooth particle mesh ewald method*, The Journal of Chemical Physics, 103 (1995), pp. 8577–8593.

[40] J. Evenäs, A. Malmendal, and S. Forsén, *Calcium*, Curr Opin Chem Biol, 2 (1998), pp. 293–302.

[41] A. R. Fersht, *Structure and mechanism in proteinscience: a guide to enzyme catalysis and protein folding*, W.H. Freeman, New York, 1999.

[42] S. Flores, N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E. Z. Yu, M. Thorpe, and M. Gerstein, *The database of macromolecular motions: new features added at the decade mark*, Nucleic Acids Res, 34 (2006), pp. 296–301.

[43] M. Frank, G. Clore, and A. Gronenborn, *Structural and dynamic characterization of the urea denatured state of the immunoglobulin binding domain of streptococcal protein-g by multidimensional heteronuclear nmr-spectroscopy*, Protein Science, 4 (1995), pp. 2605–2615.

[44] W. Franks, D. Zhou, B. Wylie, B. Money, D. Graesser, H. Frericks, G. Sahota, and C. Rienstra, *Magic-angle spinning solid-state NMR spectroscopy of the beta 1 immunoglobulin binding domain of protein G (GB1): N-15 and C-13 chemical shift assignments and conformational analysis*, Journal of the American Chemical Society, 127 (2005), pp. 12291–12305.

[45] H. Frauenfelder, F. Parak, and R. D. Young, *Conformational substates in proteins*, Annu Rev Biophys Biophys Chem, 17 (1988), pp. 451–479.

[46] H. Frauenfelder, H. Sligar, and P. Wolynes, *The energy landscape and motions of proteins*, Science, 254 (1991), p. 1598.

[47] T. Gallagher, P. Alexander, P. Bryan, and G. L. Gilliland, *Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr*, Biochemistry, 33 (1994), pp. 4721–4729.

[48] A. Garcia, *Large-amplitude nonlinear motions in proteins*, 68 (1992), pp. 2696–2699.

[49] J. A. Gerlt and P. C. Babbitt, *Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies*, Annu Rev Biochem, 70 (2001), pp. 209–246.

[50] M. Gerstein and W. Krebs, *A database of macromolecular motions*, Nucleic Acids Res, 26 (1998), pp. 4280–4290.

[51] M. E. Glasner, J. A. Gerlt, and P. C. Babbitt, *Evolution of enzyme superfamilies*, Curr Opin Chem Biol, 10 (2006), pp. 492–497.

[52] Z. Grabarek, *Structural basis for diversity of the ef-hand calcium-binding proteins*, J Mol Biol, 359 (2006), pp. 509–525.

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

[54] Y. Han, X. Li, and X. Pan, *Native states of adenylate kinase are two active sub-ensembles*, FEBS Lett, 528 (2002), pp. 161–165.

[55] 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 (2007), pp. 18055–18060.

[56] K. Henzler-Wildman and D. Kern, *Dynamic personalities of proteins*, Nature, 450 (2007), pp. 964–972.

[57] 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 (2007), pp. 838–844.

[58] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, *A linear constraint solver for molecular simulations*, Journal of Computational Chemistry, 18 (1997), pp. 1463–1472.

[59] K. Hinsen, *Analysis of domain motions by approximate normal mode calculations*, Proteins, 33 (1998), pp. 417–429.

[60] K. Hinsen, A. J. Petrescu, S. Dellerue, M. C. Bellissent-Funel, and G. Kneller, *Harmonicity in slow protein dynamics*, Chem. Phys., 261 (2000), pp. 25–37.

[61] L. Holm and C. Sander, *Dali: a network tool for protein structure comparison*, Trends Biochem Sci, 20 (1995), pp. 478–480.

[62] ——, *Alignment of three-dimensional protein structures: network server for database searching*, Methods Enzymol, 266 (1996), pp. 653–662.

[63] ——, *Mapping the protein universe*, Science, 273 (1996), pp. 595–603.

[64] ——, *Protein folds and families: sequence and structure alignments*, Nucleic Acids Res, 27 (1999), pp. 244–247.

[65] W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Phys Rev A, 31 (1985), pp. 1695–1697.

[66] J. Howard, *Mechanics of motor proteins and the cytoskeleton*, Sinauer Associates, Sunderland, MA, 2001.

[67] W. Humphrey, A. Dalke, and K. Schulten, *Vmd - visual molecular dynamics*, J. Mol. Graph., 14 (1996), pp. 33–38.

[68] D. Janezic, R. Venable, and B. R. Brooks, *Harmonic analysis of large systems iii. comparison with molecular dynamics*, 16 (1995), pp. 1544–1556.

[69] W. Jorgensen, D. Maxwell, and J. Tirado-Rives, *Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids*, Journal of the American Chemical Society, 118 (1996), pp. 11225–11236.

[70] W. Kabsch, *A discussion of the solution for the best rotation to relate two sets of vectors*, Acta Crystallographica Section A, 34 (1978), pp. 827–828.

[71] G. Kaminski, R. Friesner, J. Tirado-Rives, and W. Jorgensen, *Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides*, Journal of Physical Chemistry B, 105 (2001), pp. 6474–6487.

[72] M. Karplus, *Molecular dynamics simulations of biomolecules*, Acc. Chem. Res., 35 (2002), pp. 321–323.

[73] M. Karplus and D. L. Weaver, *Protein-folding dynamics*, Nature, 260 (1976), pp. 404–406.

[74] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons. Wiley's Series in Probability and Statistics., New York, 2005.

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

[76] H. Kim and W. N. Lipscomb, *Comparison of the structures of three carboxypeptidase a-phosphonate complexes determined by x-ray crystallography*, Biochemistry, 30 (1991), pp. 8171–8180.

[77] L. N. Kinch and N. V. Grishin, *Evolution of protein structures and functions*, Curr Opin Struct Biol, 12 (2002), pp. 400–408.

[78] A. Kitao and N. Go, *Investigating protein dynamics in collective coordinate space*, Curr Opin Struct Biol, 9 (1999), pp. 164–169.

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

[80] G. R. Kneller and K. Hinsen, *Computing memory functions from molecular dynamics simulations*, The Journal of Chemical Physics, 115 (2001), pp. 11097–11105.

[81] ——, *Fractional brownian dynamics in proteins*, The Journal of Chemical Physics, 121 (2004), pp. 10278–10283.

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

[83] S. C. Kou and X. S. Xie, *Generalized langevin equation with fractional gaussian noise: Subdiffusion within a single protein molecule*, Physical Review Letters, 93 (2004), p. 180603.

[84] R. H. Kretsinger and C. E. Nockolds, *Carp muscle calcium-binding protein. ii. structure determination and general description*, J Biol Chem, 248 (1973), pp. 3313–3326.

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

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

[87] J. Kuszewski, G. Clore, and A. Gronenborn, *Fast folding of a prototypic polypeptide - the immunoglobulin binding domain of streptococcal protein-g*, Protein Science, 3 (1994), pp. 1945–1952.

[88] A. M. Lesk, *Introduction to Protein Science: Architecture, Function and Genomics*, Oxford University Press, UK, 2004.

[89] W. N. Lipscomb, *Structure and catalysis of enzymes*, Annu Rev Biochem, 52 (1983), pp. 17–34.

[90] H. Lou and R. I. Cukier, *Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations*, J Phys Chem B, 110 (2006), pp. 24121–24137.

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

[92] C. Micheletti, J. R. Banavar, and A. Maritan, *Conformations of proteins in equilibrium*, Phys Rev Lett, 87 (2001), pp. 088102–088102.

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

[94] C. Micheletti, A. Laio, and M. Parrinello, *Reconstructing the density of states by history-dependent metadynamics*, Phys Rev Lett, 92 (2004), pp. 170601–170601.

[95] 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 (2002), pp. 909–921.

[96] C. Micheletti, F. Seno, and A. Maritan, *Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies*, Proteins, 40 (2000), pp. 662–674.

[97] W. Min, G. Luo, B. J. Cherayil, S. C. Kou, and X. S. Xie, *Observation of a power-law memory kernel for fluctuations within a single protein molecule*, Physical Review Letters, 94 (2005), p. 198302.

[98] D. Ming and M. E. Wall, *Allostery in a coarse-grained model of protein dynamics*, Phys. Rev. Lett., 95 (2005), p. 198103.

[99] S. Mityamoto and P. A. Kollman, *Settle: An analytical version of the shake and rattle algorithm for rigid water models*, Journal of Computational Chemistry, 13 (1992), pp. 952–962.

[100] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, *Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins*, Proc Natl Acad Sci U S A, 100 (2003), pp. 12570–12575.

[101] O. Miyashita, P. G. Wolynes, and J. N. Onuchic, *Simple energy landscape model for the kinetics of functional transitions in proteins*, J Phys Chem B, 109 (2005), pp. 1959–1969.

[102] S. Miyazawa and R. L. Jernigan, *Residue-residue potentials with a favorable contact pair term an unfavorable high packing density term, for simulation and threading*, 256 (1999), pp. 623–644.

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

[104] C. W. Müller and G. E. Schulz, *Structure of the complex between adenylate kinase from escherichia coli and the inhibitor ap5a refined at 1.9 a resolution. a model for a catalytic transition state*, J Mol Biol, 224 (1992), pp. 159–177.

[105] M. R. Nelson and W. J. Chazin, *Structures of ef-hand ca(2+)-binding proteins: diversity in the organization, packing and response to ca2+ binding*, Biometals, 11 (1998), pp. 297–318.

[106] S. Nosè, *A molecular dynamics method for simulations in the canonical ensemble*, Molecular Physics, 52 (1984), pp. 255–268.

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

[108] B. Park and M. Levitt, *Energy functions that discriminate x-ray and near-native folds from well-constructed decoys*, 258 (1996), pp. 367–392.

[109] F. M. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, and C. A. Orengo, *The cath database: an extended protein family resource for structural and functional genomics*, Nucleic Acids Res, 31 (2003), pp. 452–455.

[110] S. C. Pegg, S. Brown, S. Ojha, C. C. Huang, T. E. Ferrin, and P. C. Babbitt, *Representing structure-function relationships in mechanistically diverse enzyme superfamilies*, Pac Symp Biocomput, (2005), pp. 358–369.

[111] M. Perutz and F. Mathews, *An x-ray study of azide methaemoglobin*, Journal of Molecular Biology, 21 (1966), pp. 199–&.

[112] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, CUP, Cambridge, 1999.

[113] Y. E. Shapiro, E. Kahana, V. Tugarinov, Z. Liang, J. H. Freed, and E. Meirovitch, *Domain flexibility in ligand-free and inhibitor-bound escherichia coli adenylate kinase based on a mode-coupling analysis of 15n spin relaxation*, Biochemistry, 41 (2002), pp. 6271–6281.

[114] Y. E. Shapiro and E. Meirovitch, *Activation energy of catalysis-related domain motion in e. coli adenylate kinase*, J Phys Chem B, 110 (2006), pp. 11519–11524.

[115] Y. E. Shapiro, M. A. Sinev, E. V. Sineva, V. Tugarinov, and E. Meirovitch, *Backbone dynamics of escherichia coli adenylate kinase at the extreme stages of the catalytic cycle studied by (15)n nmr relaxation*, Biochemistry, 39 (2000), pp. 6634–6644.

[116] F. Sheinerman and C. Brooks, *Molecular picture of folding of a small alpha/beta protein*, Proceedings of the National Academy of Sciences of the United States of America, 95 (1998), pp. 1562–1567.

[117] M. A. Sinev, E. V. Sineva, V. Ittah, and E. Haas, *Towards a mechanism of amp-substrate inhibition in adenylate kinase from escherichia coli*, FEBS Lett, 397 (1996), pp. 273–276.

[118] M. Singleton, M. Isupov, and J. Littlechild, *Crystallization and preliminary x-ray diffraction studies of pyrrolidone carboxyl peptidase from the hyperthermophilic archaeon thermococcus litoralis*, Acta Crystallogr. Sect. D, 55 (1999), pp. 702–703.

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

[120] H. Tanaka, M. Chinami, T. Mizushima, K. Ogasahara, M. Ota, T. Tsukihara, and K. Yutani, *X-ray crystalline structures of pyrrolidone carboxyl peptidase from a hyperthermophile, pyrococcus furiosus, and its cys-free mutant*, J. Biochem. (Tokyo), 130 (2001), pp. 107–118.

[121] D. L. Theobald and D. S. Wuttke, *Divergent evolution within protein superfolds inferred from profile-based phylogenetics*, J Mol Biol, 354 (2005), pp. 722–737.

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

[123] M. M. Tirion and D. ben Avraham, *Normal mode analysis of g-actin*, 230 (1993), pp. 186–195.

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

[125] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, *Gromacs: fast, flexible, and free*, J Comput Chem, 26 (2005), pp. 1701–1718.

[126] W. Van Gunsteren, S. Billeter, A. Eising, P. Hunenberger, P. Kruger, A. Mark, W. Scott, and I. Tironi, *Biomolecular simulation: the GROMOS96 manual and user guide*, Vdf Hochschulverlag AG an der ETH Zrich, Zurich, 1996.

[127] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, *Conformational transitions of adenylate kinase: switching by cracking*, J Mol Biol, 366 (2007), pp. 1661–1671.

[128] 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 (2004), pp. 945–949.

[129] K. L. Yap, J. B. Ames, M. B. Swindells, and M. Ikura, *Diversity of conformational states and changes within the ef-hand protein superfamily*, Proteins, 37 (1999), pp. 499–507.

[130] 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 (2008), pp. 918–929.