Knowledge-based identification of functional domains in proteins

ma per seguir virtute e canoscenza

SISSA

A thesis submitted for the degree of
Philosophiæ Doctor

(October 2016)

Candidate
Luca Ponzoni

Supervisor
Cristian Micheletti

Molecular and Statistical Biophysics sector
PhD course in Physics and Chemistry of Biological Systems

Contents

Overview

The characterization of proteins and enzymes is traditionally organised according to the sequence-structure-function paradigm.

The investigation of the inter-relationships between these three properties has motivated the development of several experimental and computational techniques, that have made available an unprecedented amount of sequence and structural data. The interest in developing comparative methods for rationalizing such copious information has, of course, grown in parallel.

Regarding the structure-function relationship, for instance, the availability of experimentally resolved protein structures and of computer simulations have improved our understanding of the role of proteins' internal dynamics in assisting their functional rearrangements and activity. Several approaches are currently available for elucidating and comparing proteins' internal dynamics. These can capture the relevant collective degrees of freedom that recapitulate the main conformational changes. These collective coordinates have the potential to unveil remote evolutionary relationships between proteins, that are otherwise not easily accessible from purely sequence- or structure-based investigations.

Starting from this premise, in the first chapter of this thesis I will present a novel and general computational method that can detect large-scale dynamical correlations in proteins by comparing different representative conformers. This is accomplished by applying dimensionality-reduction techniques to inter-amino acid distance fluctuation matrices. As a result, an optimal quasi-rigid domain decomposition of the protein or macromolecular assembly of interest is identified, and this facilitates the functionally-oriented interpretation of their internal dynamics.

Building on this approach, in the second chapter I will discuss its systematic application to a class of membrane proteins of paramount biochemical interest, namely the class A G protein-coupled receptors. The comparative analysis of their internal dynamics, as encoded by the quasi-rigid domains, allowed us to identify recurrent patterns in the large-scale dynamics of these receptors. This, in turn, allowed us to single out a number of key functional sites. These were, for the most part, previously known – a fact that at the same time validates the method, and gives confidence for the viability of the other, novel sites.

Finally, for the last part of the thesis, I focussed on the sequence-structure relationship. In particular, I considered the problem of inferring structural properties of proteins from the analysis of large multiple sequence alignments of homologous sequences. For this purpose, I recasted the strategies developed for the dynamical features extraction in order to identify compact groups of coevolving residues, based only on the knowledge of amino acid variability in aligned primary sequences.

Throughout the thesis, many methodological techniques have been taken into considerations, mainly based on concepts from graph theory and statistical data analysis (clustering). All these topics are explained in the methodological sections of each chapter.

The material presented in chapter 1 is largely based on the published paper Ponzoni L., Polles G., Carnevale V., Micheletti C., *SPECTRUS: a dimensionality reduction approach for identifying dynamical domains in protein complexes from limited structural datasets*, Structure, 2015. A manuscript describing the results presented in Chapter 2 has been submitted in August 2016 to PLOS Computational Biology (Ponzoni L., Rossetti G., Carloni P., Micheletti C., *Unifying view of mechanical and functional hotspots across class A GPCRs*). Chapter 3 includes preliminary results of my latest research done in collaboration with Daniele Granata, Vincenzo Carnevale and Cristian Micheletti, and a manuscript is in preparation, with expected submission in November 2016.

Chapter 1

Quasi-rigid domain decomposition of proteins and viral capsids

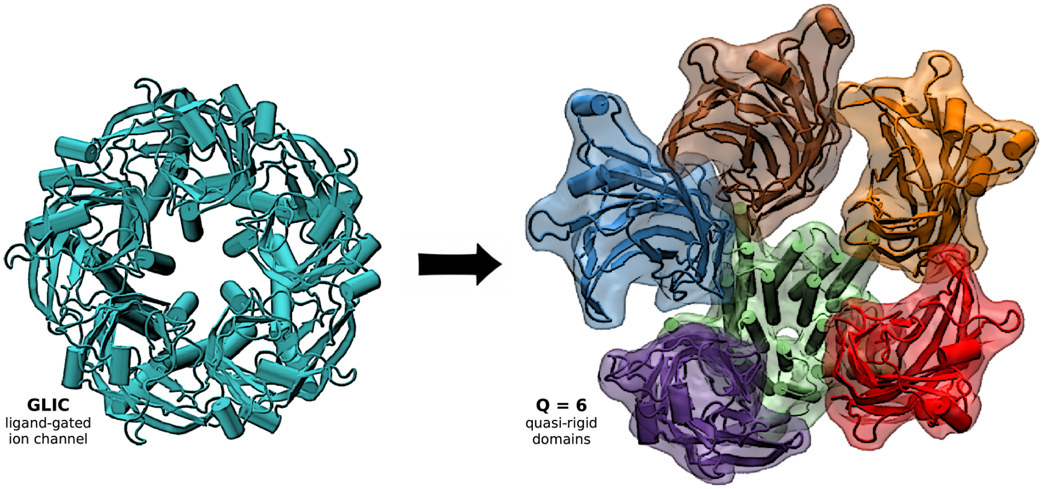
Proteins' functionality is promoted and assisted by the conformational changes their structures undergo both spontaneously, from the thermal excitation of innate fluctuations encoded by its three-dimensional organization, and/or as a response to external effects, like the binding of ligands.

Elucidating the conformational dynamics of macromolecules can therefore be a valuable tool to better understand the structure-function relationship in proteins and enzymes, to detect common large-scale features across superfamilies, as well as expose evolutionary relationships that would otherwise be elusive from the sequence and structural point of view.

For this purpose, many analysis tools have been devised, like normal mode analysis of elastic network models or principal component analysis of MD trajectories. These tools allow to detect the few collective degrees of freedom that capture most of the structural fluctuations characterizing the internal dynamics of a protein.

However, in order to obtain a simple insight and a transparent interpretation, the collective structural changes represented by the modes can be described in a more natural way, as resulting from the relative motion of few subdomains. Although these elementary functional units can be intuitively perceived by the visual inspection of modes, the problem emerges of defining a quantitative and rigorous scheme to derive such representation.

In this chapter, I will present a general computational method, named



SPECTRUS, that I developed as my main PhD project, and that is aimed at giving an objective identification of quasi-rigid domains in single proteins or macromolecular assemblies.

The method uses advanced dimensionality-reduction techniques to single out the basic quasi-rigid, functional units of proteins complex from the sole analysis of alternative conformers, even very few of them. No prior assumptions are made on the molecules' properties and hence the method applies equally well to individual proteins or very large complexes, and to structural sets comprising thousands of conformers sampled from molecular dynamics simulations or just very few crystal structures. A key element is the introduction of a quality score parameter which guides the selection of the most significant, innate subdivisions, thus solving one of the main difficulties in providing an objective criterion for quasi-rigid domain decompositions.

The present discussion is largely based on the paper where this algorithm was first introduced [1]. My contributions regarded all phases of this work, from the design and implementation of the algorithm, to its application to the practical examples illustrated below, and to the creation of an online webserver.

1.1 Introduction

The functional proficiency of proteins as molecular machines often relies on their internal structural dynamics. In fact, the innate conformational fluctuations of these biomolecules are often primed to favor and assist the interconversion between different substates, such as activated and inactivated forms of an enzyme or the open and closed states of pores and channels [2–15].

The fact that these structural changes are typically of large amplitude and have a collective character [16–22] has naturally posed the challenge of developing suitable methods for describing these rearrangements in terms of rigid-like displacements (rotations and translations) of a limited number of quasi-rigid domains [23–39]. By these means, in fact, one can achieve a parsimonious identification of the few degrees of freedom that suffice to describe and explore the conformational space accessible to a given protein.

The applicative avenues of quasi-rigid domain decomposition strategies are several and diverse. For instance, they can be used to extend the analysis of molecular dynamics trajectories beyond the linear superposition of essential dynamical spaces [40, 41], for preconditioning enhanced sampling techniques [42], and for comparing the functional dynamics of proteins with different degrees of sequence and structural similarity [3, 6, 38, 43–47]. Other applications include the selection of a manageable parameter space for inferential or maximum-likelihood structure determination [48, 49] as well as the inference of the basic mechanical and assembly units of large macromolecular complexes, such as viral capsids [50–57].

Available quasi-rigid domain decomposition methods build either directly or indirectly on the notion that, for genuinely rigid bodies, the distances between any two constitutive points are strictly preserved during the motion in space [28, 34, 36, 37, 58]. Accordingly, a common starting point is the calculation of the distance fluctuations for each pair of amino acids, a and b:

$$f_{a,b} = \sqrt{\langle d_{a,b}^2 \rangle - \langle d_{a,b} \rangle^2} \ , \tag{1.1}$$

where $d_{a,b}$ is the C$_\alpha$ atoms distance and the $\langle\rangle$ brackets denote the average over representative conformers from available crystal structures or sampled from molecular dynamics trajectories. A model f matrix can also be computed from a single reference structure by using elastic networks.

The entries of the distance fluctuation matrix give a quantitative measure of the likelihood that two amino acids belong to the same rigid domain, and hence provide a natural metric for their grouping by using generic clustering algorithms.

The implementation of this general and transparent strategy is, however, limited by two main factors.

First, because the quasi-rigid character of biomolecular domains holds only approximately, the subdivisions can depend significantly on the clustering algorithm and on the number of conformers used to derive the distance fluctuation matrix, especially when only few conformers are used.

The second challenge is the definition of an objective, quantitative criterion for choosing the most significant subdivision among those obtained by clustering amino acids in an increasing number of domains. In specific contexts, this challenge can be overcome by using *ad hoc* auxiliary parameters. For instance, considerations of domain shape homogeneity and structural integrity have been recently used to identify viable decompositions of viral capsids [56].

Because such strategies are intrinsically tailored to specific systems it remains open the issue of identifying suitable order parameters for ranking the significance of various subdivisions using general criteria that are internal to the clustering procedure itself. In this regard we note that although virtually all current decomposition strategies entail the minimization of the total intra-domain distance fluctuations, the latter quantity is generally not useful for singling out the optimal quasi-rigid domains. In fact, it attains its global minimum for the trivial subdivision where each amino acid corresponds to a single domain.

Here we introduce, validate and apply a self-contained quasi-rigid domain subdivision strategy, termed SPECTRUS after SPECTral-based Rigid Units Subdivision, that allows for overcoming both difficulties.

Specifically, the consistency of quasi-rigid subdivisions with respect to the number of available conformers as well as the adopted clustering method is achieved through the Laplacian spectral projection. This is a data preconditioning technique which provides an optimal dimensional reduction of the phase space describing the distance fluctuations of all amino acids pairs, thus providing the required partitioning robustness.

Furthermore, we show that the very same properties of the reduced-dimensionality space can be seamlessly used to define a quality score which can pinpoint significant subdivisions based on the balance of intra- and inter-domain distance fluctuations compared to a random reference case. To our knowledge, this balance, which is increasingly recognized as crucial for optimal clustering strategies [59], has not been exploited yet in the context of protein rigid domain decomposition, where intra-domain compactness is usually the sole quantity being optimized. Therefore, the proposed quality score arguably represents a first general, quantitative criterion that is internal to the decomposition method and allows for assessing the statistical significance of subdivisions and hence single out the innate one(s).

The effectiveness of SPECTRUS as a general and transferable strategy for quasi-rigid domain decomposition has been evaluated and ascertained by comparing it to various alternative subdivision schemes applied to a wide range of proteins and molecular assemblies. In particular, starting from the familiar validation case of adenylate kinase we next consider two membrane protein complexes, namely GLIC and NavAb, whose basic functional, quasi-rigid domains have been suggested only recently, based on the supervised inspection of novel experimental and numerical data [60–62]. Finally, the capability of the method to operate with a "high dynamic range" of domains and molecular sizes is illustrated for two viral capsids, namely those of the satellite tobacco mosaic virus and Triatoma virus, for which it correctly pinpoints the several tens of functional units as established from molecular dynamics simulations or AFM nano-indentation experiments [51, 52, 55]. For these or even larger macromolecular assemblies, which may be too onerous to simulate with atomistic molecular dynamics, we further show that the distance fluctuation matrix can be viably obtained from computationally-effective elastic network models. This further illustrates the applicability of the decomposition method even in limiting cases where a single crystal structure is available.

SPECTRUS is made available as an online tool, accessible at the address http://spectrus.sissa.it/, from where it is also possible to download the source code.

1.2 Domain subdivision strategies

In practical contexts, only a few structural data are usually available for the conformational analysis of a given protein. Moreover, even when extensive molecular dynamics simulations have been carried out, in most cases there is no guarantee that all relevant conformers have been sampled. When dealing with this general issue of the scarcity of structural information, it is therefore even more crucial to make sure that the tools used for the analysis are as robust with respect to noisy data as possible.

For this purpose, we conducted a comparative analysis of the most used clustering schemes in order to establish if they meet the requirements of robustness and reliability needed for producing meaningful results. We then considered both hierarchical schemes, here represented by the complete-linkage and group-average agglomerative clusterings, and flat schemes, represented by k-medoids and cut-based clusterings.

These techniques, whilst widely used and algorithmically very simple, are found to provide inconsistent results on particularly challenging datasets. We will then discuss a more sophisticated clustering method, the spectral clustering, which can be used in combination with any of the previously mentioned algorithms and can be described as a sort of preprocessing step on the original dataset. By performing a dimensional reduction, it is able to enhance the dominant features of the underlying data, making them more recognizible for the subsequent step of clustering.

In the following sections, we first provide a detailed description of the algorithms and then discuss their application.

1.2.1 Reference clustering schemes

In their general formulation, the various clustering schemes take as inputs a matrix of pairwise similarities, σ, or dissimilarities, δ, between the elements. In our context, the elements are the amino acids and the average fluctuations of pairwise C_α distances, f, defined in eq. (1.1), provide a natural measure of dissimilarity, i.e. $\delta \equiv f$.

- *Agglomerative clustering.* Hierarchical agglomerative schemes start by assigning each element (amino acid) to a separate cluster and proceed by iteratively merging the two least dissimilar clusters [63]. For the complete-linkage scheme, the dissimilarity Δ of two clusters, C_i and C_j, is given by the most dissimilar pair of their elements:

  $$\Delta_{C_i,Cj} = \max_{a\in C_i;b\in C_j} \delta_{a,b} \ . \tag{1.2}$$

  For the group-average scheme one instead considers the dissimilarity averaged over all possible pairs of elements from the two clusters:

  $$\Delta_{C_i,Cj} = \frac{1}{n_{C_i}\, n_{C_j}} \sum_{a\in C_i;b\in C_j} \delta_{a,b} \ , \tag{1.3}$$

  where n_{C_i} is the number of elements in cluster C_i. In both cases, the process ends with the subdivision into two clusters, when a full tree of partitions is obtained. The complete-linkage usually returns more compact clusters than the group-average one. However, the latter is clearly less sensitive to the presence of outliers.

- *k-medoids clustering.* In the k-medoids scheme [64], the subdivisions into a given number Q of clusters are obtained by assigning Q representative elements $\{r_1, r_2, \ldots, r_Q\}$, called medoids, and the members of each cluster, so to minimize the total intra-cluster dissimilarity:

$$\sum_{i=1}^{Q} \sum_{a \in C_i} \delta_{a,r_i} \ . \qquad (1.4)$$

The minimization of the score in eq. (1.4) is carried out iteratively starting from an initial tentative choice of the Q representatives. The other elements are next assigned to the cluster of their nearest representative. Each representative is then replaced by the element for which the sum of the dissimilarities from the other cluster members is smallest, and so on until convergence. The subdivision associated to the lowest total dissimilarity over several different initial conditions is retained as the optimal partitioning.

- *Cut-based clustering.* The cut-based clustering aims at simultaneously maximizing the intra-cluster similarity and minimizing the inter-cluster one. As such it is naturally formulated in terms of the matrix of pairwise similarities, σ, rather than the one of dissimilarities, δ. A convenient mapping between corresponding entries of the two matrices is obtained with a Gaussian weighting function, $\sigma_{a,b} = \exp(-\delta^2_{a,b}/2\bar{\delta}^2)$, where $\bar{\delta}$ is a conservative measure for intra-cluster dissimilarities [65]. For our purposes, since the local network formed by amino acids in close contact is expected to be typically rigid, $\bar{\delta}$ is computed as the average dissimilarity (i.e. the average distance fluctuation) of C_α pairs that are closer than 10 Å. Moreover, in order to improve performance by taking advantage of sparse matrices properties, entries corresponding to amino acid pairs that are farther than 10 Å apart are set equal to 0.

The clustering is performed by using a stochastic optimization method (e.g. simulated annealing) to identify the partitioning which minimizes

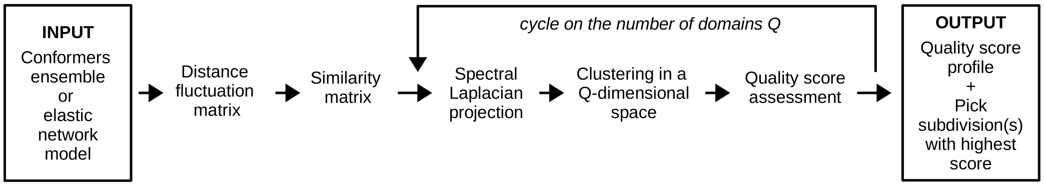

INPUT
Conformers ensemble or elastic network model
Distance fluctuation matrix
Similarity matrix
cycle on the number of domains Q
Spectral Laplacian projection
Clustering in a Q-dimensional space
Quality score assessment
OUTPUT
Quality score profile + Pick subdivision(s) with highest score

Figure 1.1: SPECTRUS flowchart.

the cost function:

$$\sum_{i=1}^{Q} \frac{\sum_{a\in C_i}\sum_{b\notin C_i}\sigma_{a,b}}{\sum_{a,b\in C_i}\sigma_{a,b}} \,. \tag{1.5}$$

1.2.2 Spectral projection and clustering

As it is shown in the flowchart of Fig. 1.1, the SPECTRUS subdivision of amino acids into quasi-rigid domains relies on a preconditioning step involving the spectral dimensional reduction of the distance fluctuation matrix. More specifically, starting from the similarity matrix σ, the partitioning of the N amino acids into Q clusters is achieved through the following steps [66], which are heuristically explained further below:

1. Calculation of the $N \times N$ symmetric Laplacian matrix, $L = I - D^{-1/2}\,\sigma\, D^{-1/2}$, where I is the identity matrix and D is a diagonal matrix with elements equal to $D_{a,a} = \sum_b \sigma_{a,b}$.

2. Calculation of the Q lowest eigenvectors of L (i.e. those associated to the Q smallest eigenvalues), $\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^Q$.

3. Construction of the auxiliary $N \times Q$ matrix X, whose columns are the Q lowest eigenvectors of L. Accordingly, the matrix entries are defined as $X_{i,j} = v^j_i$.

4. Normalization of each row of X. The normalized arrays represent the coordinates of N points on the Q-dimensional unit sphere. These points provide the projection of the original N elements in the lower-dimensional spectral space.

5. The N projected points are finally grouped in Q clusters with a method of choice, such as the k-medoids scheme, which is computational efficiency and has a simple formulation. Because the projected points are

in one-to-one correspondence with the original elements, this clustering straightforwardly translates into the partitioning of the amino acids into Q quasi-rigid domains.

The outlined strategy can be intuitively understood by recalling that the Laplacian matrix describes how a probability density, defined on the nodes of a graph, evolves by diffusion on the graph itself. Accordingly, the eigenvectors at step 2 embody the slowest modes of relaxation to the steady state probability distribution defined on the network (graph) connecting similar elements (that is, in the present case, amino acid pairs experiencing the least distance fluctuations). The usefulness of these eigenvectors in clustering contexts readily emerges when considering a graph consisting of Q disconnected subparts. Because the probability distribution evolves independently on each of the uncoupled subgraphs, one has that the support of each one of the top Q Laplacian eigenvectors is associated to a different subgraph. Accordingly, in the more general context of a fully-connected graph, the norm of the top eigenvectors of the Laplacian is expected to be concentrated on the subgraphs with the least inter-connection between each other.

When dealing with protein subdivisions into Q quasi-rigid domains, it is therefore sufficient to restrict considerations to the subspace spanned by the orthonormal set of the top Q eigenvectors of the Laplacian of the similarity (rigidity) matrix, see steps 1-3 above. Moreover, since the graph is typically fully connected, it can be proved that the first eigenvector has a definite sign [66]. The actual clustering (steps 4-5), therefore, is finally performed on elements which are represented as unit vectors whose first component is always positive (or negative), i.e. the projected points in the Q-dimensional space lie on the surface of half the Q-dimensional unit sphere.

1.2.3 Quality and significance of spectral partitions

In ideal clustering cases, where the elements have neatly-separated groupings, the correct number of clusters can be identified by the presence of a gap in the Laplacian spectrum [65, 66]. This sharp criterion is generally not applicable in practical contexts, including subdividing protein in domains whose rigid character holds only approximately.

As a robust criterion to guide the identification of the innate number and type of partitions, we introduce an order parameter, which we term quality score, which quantifies how compact and well separated are the clusters respect to a random reference case. The calculation of the quality score is described hereafter for partitions obtained at step 5 with the k-medoids clustering method, that we elected to use after the comparative tests described in the Results section.

To measure the compactness and separation of the Q returned clusters we take

$$\rho(Q) = \text{median}_{a=1,\ldots,N}(\delta_{a,\nu_a}/\delta_{a,\mu_a}). \tag{1.6}$$

In the above expression, a is the index of one element, μ_a is its representative medoid, i.e. the nearest of the medoids, and ν_a is the second nearest medoid. For distance (or dissimilarity) of two elements, δ, we take their arclength separation on the surface of the Q-dimensional unit (hemi)sphere where the points lie. The quantity ρ therefore captures how typically distant are the elements from the closest alternative cluster compared to the distance of their own cluster representative. It therefore provides an apt measure of clustering quality since it is simultaneously informative about intra-cluster compactness and inter-cluster separation. The use of the median in place of the average over all elements confers further robustness against the presence of outliers.

For an equal footing comparison of the clustering quality across different values of Q we finally normalize $\rho(Q)$ dividing it by its value computed over a collection of N points that are randomly distributed on the unit Q-dimensional hemi-sphere and clustered in Q groups. This normalization factor is straightforwardly computed numerically.

Significant subdivisions are clearly associated to values of the quality score that are appreciably larger than 1. This, in fact, implies that the clusters are substantially more compact and better separated than for the random case.

The present introduction of the quality order parameter represents a valuable contribution to spectral clustering approaches in general and, as we shall discuss, proves very valuable for the problem of protein domain decomposition, in particular.

1.3 Applications

1.3.1 Adenylate kinase

For a first assessment and validation of the SPECTRUS strategy we consider the case of *E. coli* adenylate kinase. This is a monomeric phosphotransferase enzyme of about 200 amino acids which balances the energy charge of the

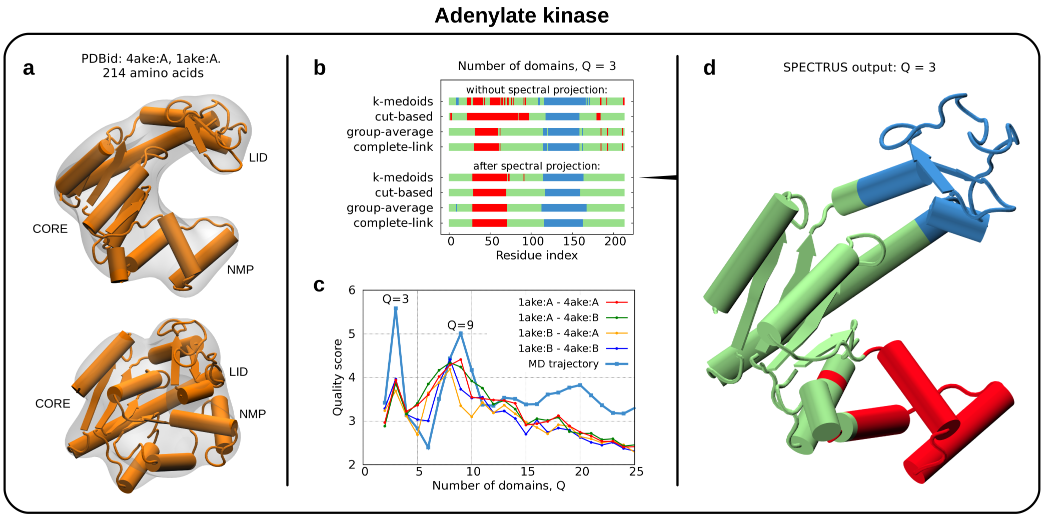


Figure 1.2: (a) Conformations of *Escherichia coli* adenylate kinase in the open and closed state. (b) Color-coded representation on the sequence of the subdivisions into $Q = 3$ domains, obtained by using various clustering methods, before and after performing the spectral projection step. (c) Quality score profiles, as a function of the number of domains Q, returned by SPECTRUS for different pairs of open/closed conformations of adenylate kinase. The optimal subdivisions into few domains is achieved for $Q = 3$ while $Q = 8, 9$ should be used for finer subdivisions, see Fig. 1.4b. The robustness of this result is underscored by the parallel behavior of the quality score profile computed from thousands of conformations sampled with extensive MD simulations [22] and shown in blue in the same panel. The associated $Q = 3$ and $Q = 9$ subdivisions are provided in Fig. 1.4a and are in very good agreement with the two-structures case. (d) Structural representation of the $Q = 3$ subdivision.

cell by catalyzing the conversion of ATP and AMP into two molecules of ADP. The enzyme is capable of spontaneously interconverting between the open conformation and the closed, catalytically competent one [2, 22].

Consistently with the noticeable structural differences of the two forms, which are shown in Fig. 1.2a, the enzyme functional mechanics is usually rationalized in terms of the relative movements of two main subparts, the ATP binding domain (LID) and the AMP binding domain (NMP) around the central core.

We shall accordingly start the study of adenylate kinase by focusing on the subdivisions into $Q = 3$ domains obtained by using the various types of clustering strategies mentioned before. For all of them, we used as input the distance fluctuation matrix computed from the sole open and closed conformers of Fig. 1.2a. The difference between these two conformers captures the enzyme's functional motion to the fullest extent. We accordingly use them to check whether rigid domains can be reliably identified when the input dataset is limited and yet representative of the biologically-relevant conformational ensemble.

We first apply the direct partitioning approach, that is without the preconditioning step of the spectral Laplacian projection. The resulting subdivisions are shown with one-dimensional color-coded representations in the upper part of Fig. 1.2b. Although these partitions display an overall accord for identifying the core, NMP (residues ∼30 to ∼60) and LID (residues ∼115 to ∼160) as the main quasi-rigid units, there are also significant qualitative differences across them. This is readily conveyed by the varying degree of fragmentation of the domain assignment along the sequence. In this regard, we note that, while a high degree of sequence-wise fragmentation is not plausible *a priori*, it is not directly penalized by the clustering schemes. Therefore, checking for overall sequence-wise domain integrity can be a useful *a posteriori* criterion for evaluating a subdivision's viability.

Exactly the same analysis was next repeated after preconditioning the data with the Laplacian spectral projection. The results obtained after such dimensional reduction are shown in the lower section of Fig. 1.2b. The dramatic improvement of the degree of consistency and sequence-wise domain integrity is readily noticed. In fact, it is seen that all methods now give a practically unanimous consensus for the location of the domain boundaries and the sequence-wise domain discontinuity is now minimal across all methods. The consensus domain boundaries agree with those returned by other quasi-rigid decomposition methods such as DynDom, PiSQRD, TLSMD and CYRANGE, as detailed in Fig. 1.3a. For CYRANGE, which aims at identifying rigid-like domains that are recurrent across NMR models, we however point out that only the core and the LID regions are recognised as being quasi-rigid.

The advantages of the spectral projection technique, however, better emerge after lifting the restriction to the $Q = 3$ case. In fact, the data preconditioning step can help identifying the innate number and type of domains based on a quality score, previously introduced, which captures the statistical significance of the subdivisions.

We accordingly computed the quality score for all adenylate kinase subdivisions from $Q = 2$ to $Q = 30$ quasi-rigid domains, using separately as input each of the four possible pairings of chains from PDB entries 1AKE

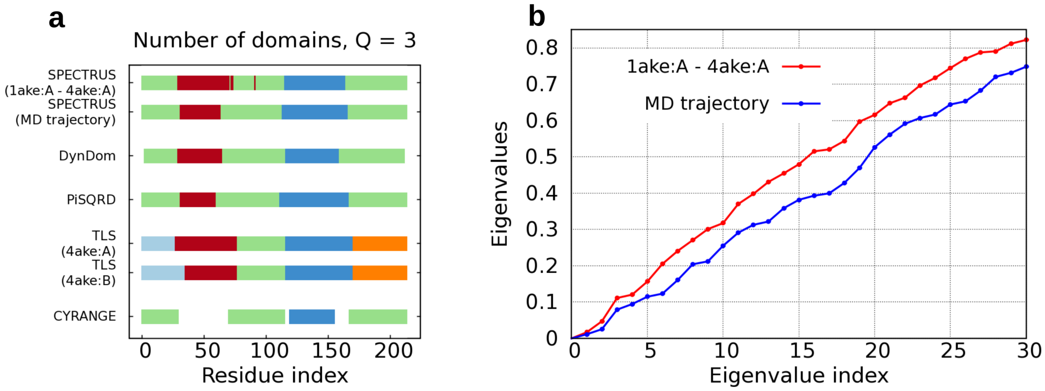


Figure 1.3: (a) Subdivisions of AKE into $Q = 3$ domains, obtained with SPECTRUS and with other rigid-domain partitioning tools. The latter include: (i) DynDom, which exclusively operates on two alternative conformers, here the open (4AKE:A) and the closed (1AKE:A) conformers; (ii) PiSQRD, applied to the same MD trajectory used before; (iii) TLS Motion Determination, based on the analysis of the B-factors of the 4AKE open structure (this tool enforces the sequence-continuity of domains, thus we accordingly considered 5 segments); (iv) CYRANGE, which, as a preliminary step, identifies a set of core atoms that are next grouped into domains (the atoms not considered are shown as blank gaps in the sequence). The overall consensus of the domains is apparent. (b) Eigenvalue spectrum of the Laplacian matrix for AKE. In ideal contexts, the viable numbers of eigenvectors (clusters) to be used for the projection step is indicated by the presence of gaps in the spectrum. The approximate quasi-rigid character of the dynamical domains does not allow for a clear identification of such gaps.

and 4AKE. For reasons of efficiency, for each value of Q we considered the subdivision returned by the k-medoids scheme after the Laplacian spectral projection.

The resulting profile of the quality score is shown in Fig. 1.2c. Across the wide range of considered number of domains, two peaks emerge clearly in the profiles. The first peak is associated to the above discussed subdivision into $Q = 3$ domains, shown on the open conformation in Fig. 1.2d. It is noteworthy and pleasing that this intuitive and customary subdivision consistently emerges as the optimal one among those involving only few domains. The second, and most prominent peak occurs for the finer subdivision into $Q = 8$ or 9 domains. The associated subdivisions, which are hierarchically related to the $Q = 3$ one, are shown in Fig. 1.4b.

The robustness of these findings was ascertained by repeating the analysis using as input the distance fluctuation matrix obtained from extensive MD trajectories rather than the minimalist set of the sole open and

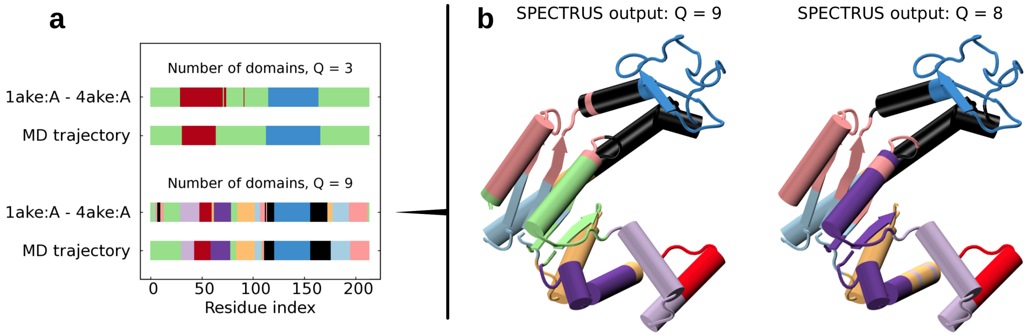


Figure 1.4: (a) AKE subdivisions into $Q = 3, 9$ domains, computed from both the sole two 1AKE:A and 4AKE:A structures, and the extensive MD simulations reported in [22]. The consistency is manifest: moreover, most of the domains for $Q = 9$ appear to be nested inside the coarser domains for $Q = 3$. (b) Subdivision of AKE into 9 and 8 domains, shown on the open conformation.

closed conformers. To this purpose we used thousands of conformations from previously-published 50ns-long atomistic simulations started from both the open and closed state of adenylate kinase [22]. This duration suffices to yield reliable essential dynamical spaces (the cosine content of the top ten principal components is < 0.5 [67]).

Fig. 1.2c shows that the resulting quality score profile has a trend that parallels the one obtained for the two-structures case. In particular, the subdivisions for $Q = 3$ and $Q = 9$ stand out as the most significant ones, as in the previous case, and are in very good agreement with the subdivisions obtained from the crystal structures (Fig. 1.4a).

The main difference is in the relative quality score of the subdivisions based on only two conformers, since the most significant one now corresponds to $Q = 3$.

We finally note that the possibility to single out only few outstanding subdivisions among tens of possible ones is made possible by the quality score definition which, with a single parameter, can convey how compact and well-separated are the clusters compared to a null reference case. These features, for instance, cannot be straightforwardly gleaned by a standard analysis of the Laplacian matrix spectrum, see Fig. 1.3b.

1.3.2 Robustness of subdivisions and quality score

To better illustrate the robustness of SPECTRUS, we considered the impact of progressively impoverishing the dataset on the subdivision consistency. Specifically we considered: (i) removing the residues with the largest B-factors from 1AKE and 4AKE so to simulate structural gaps due to non-resolved residues, (ii) using separately the two halves of the above-mentioned MD trajectories of AKE, (iii) using alternative open and closed forms of the homologous *Streptococcus pneumoniae* adenylate kinase, (iv) using various combinations of the structural representatives for three enzymes, whose internal dynamics and functional domains were previously characterized in [68]. The results of all these cases are detailed in Appendix A and show that impoverishing the input datasets usually has minor effects on the domain assignments, while the sharpness of the quality score peaks can degrade. The latter observation prompted us to complemented the profile of the quality score median (as defined above in the section on methodology) with the profiles of the 40th and 60th percentiles. In this way, the consistency (or lack thereof) of the three percentile trends provides valuable elements to assess *a posteriori* whether a sharp indication of the intrinsic number of domains emerges from the available structural data.

1.3.3 GLIC, a ligand-gated ion channel

We now turn to the case of GLIC, a pentameric ligand-gated ion channel [69]. The ongoing efforts to characterize its molecular mechanism of ion permeation have been significantly aided very recently by the successful X-ray determination of the open and closed forms of the channel [60, 61]. The root-mean-square distance (RMSD) of the 1555 amino acids present in both crystal structures is less than 2.0 Å. This overall small structural deviation makes the identification of quasi-rigid domains in GLIC very challenging in general.

As a matter of fact, the mechanistic bases of the GLIC gating action have ultimately been elucidated by Sauguet *et al.* with a meticulous and laborious supervised procedure [61].

To this same purpose, we carried out the quasi-rigid domain decomposition of GLIC. The corresponding distance fluctuation matrix was computed using the five available pentameric conformers of GLIC: four from the closed structure (PDB ID: 4NPQ) and one from the open structure (PDB ID: 4HFI),

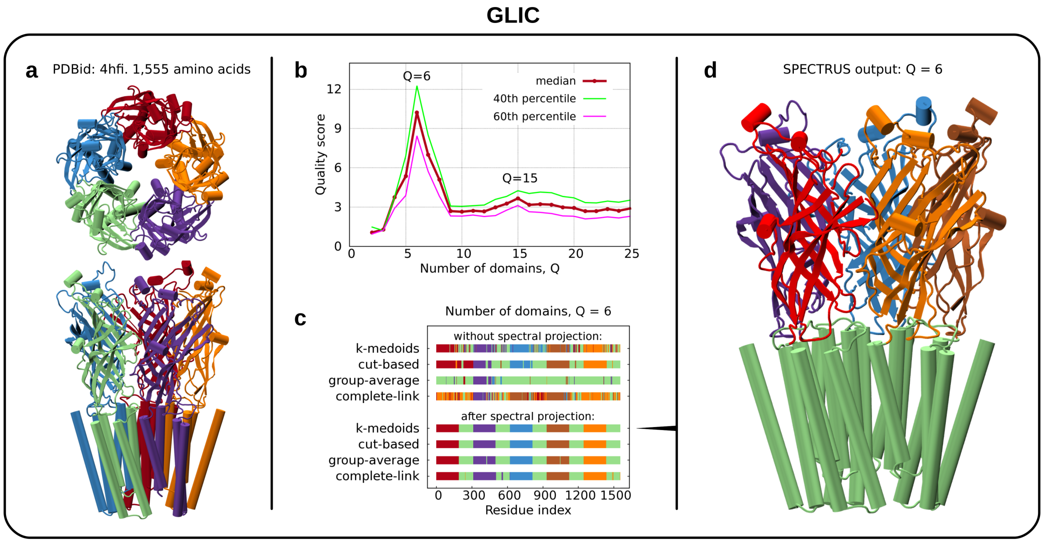


Figure 1.5: Quasi-rigid domain decomposition of the GLIC ion channel based on the distance fluctuation matrix computed from the only two crystal structures that are available for it: 4HFI and 4NPQ [60, 61]. The crystal structure of conformer 4HFI is shown in panel (a) and its five constitutive monomers are highlighted with different colors. (b) The quality score profile of the SPECTRUS subdivisions indicates that the primary partitioning involves $Q = 6$ quasi-rigid domains. The robustness of the $Q = 6$ partitioning respect to the clustering method is shown in panel (c) and is contrasted by the results obtained without the spectral projection preconditioning step. The structural representation of the $Q = 6$ subdivision is shown in panel (d).

see Fig. 1.5a. The matrix was constructed by omitting the few amino acids which were solved for only one of the two forms. The resulting linear size of the matrix was 1555, corresponding to 311 amino acids per each monomer of the pentameric channel. The matrix was next used as input for the same combination of clustering methods previously discussed for adenylate kinase.

We start by directly discussing the SPECTRUS output, that is the results obtained with the spectral projection preconditioning step. The quality score profile, computed for the computationally-efficient k-medoids scheme, is shown in Fig. 1.5b and presents a very prominent peak for a subdivision in $Q = 6$ domains. The corresponding sequence-wise amino acid partitioning is shown in the lower half of Fig. 1.5c.

The same panel illustrates that, as for adenylate kinase, after the projection preconditioning step the subdivisions of GLIC are practically independent of the clustering method. Accordingly, for reasons of efficiency, in the following we shall exclusively consider the k-medoids partitioning.

Such robustness is particularly notable when considering two aspects. The first one regards the contrast of the rather large size of the complex, more than 1500 amino acids, and the minimal number of conformers used for the analysis. Such combination might be expected *a priori* to negatively affect the sensitive discrimination of domain boundaries, thus making subdivisions too dependent on the adopted decomposition strategy. This is, in fact, what is observed when the clustering is applied directly to the unprojected distance fluctuation matrix, as shown in Fig. 1.5c. The difference with the projection case, in terms both of sequence-wise domain continuity and consistency across the methods, is striking. The second aspect regards the fact that several clustering methods tend to balance the size of the domains. This effect is discernible in the subdivisions, particularly the k-medoids one of Fig. 1.5c, but is strikingly absent after the introduction of the spectral projection.

The last point is particularly important in connection with the mechanistic interpretation of GLIC gating action. In fact, the $Q = 6$ subdivisions consensually indicate that about half the channel is encompassed in a single quasi-rigid domain. This domain corresponds to the lower part of the pentamer, see Fig. 1.5d, which, in turn, is the channel portion that is surrounded by the lipid membrane. The remaining half of each monomer is, instead, assigned to a different quasi-rigid domain. Pleasingly, this subdivision is well-consistent with the partitioning obtained by the PiSQRD web-server when constrained to return 6 quasi-rigid domains for each of the two GLIC conformers. We recall that PiSQRD differs from SPECTRUS for the use of essential dynamical spaces in place of the matrix of distance fluctuations and for the lack of a non-monotonic quality score. The other decomposition methods were not applicable to GLIC for the aforementioned limits of the input protein size or oligomeric state.

The SPECTRUS subdivision for $Q = 6$ has a straightforward interpretation as it indicates that the top halves of the five monomers (which is the extracellular part of the channel) can move relatively to each other while being hinged or anchored to the same pentameric quasi-rigid core. This agrees with the fact that the binding sites for GLIC ligands are located in the extracellular part of the monomers which must hence be mobile with respect to the intra-membrane core to trigger the pore response.

This mechanistic view, which blurs the boundaries between structural and dynamical domains, is in very good agreement with the conclusions drawn by Sauguet *et al.* based on the supervised inspection and comparison of the

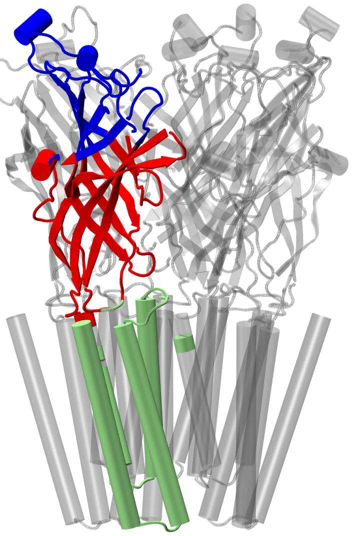


Figure 1.6: Domain decomposition of GLIC into $Q = 15$ domains, corresponding to the secondary peak of the quality score, shown in Fig. 1.5b. For visual clarity the subdivision is shown for a single monomer as it is identically replicated for the other monomers. The comparison with the $Q = 6$ case of Fig. 1.5d exposes the hierarchical quasi-rigid organization of the individual GLIC monomers.

available crystal structures [61].

This successful comparison represents a further validation of the spectral decomposition method and hence gives confidence for applying the method to capture the finer aspects of the channel mechanical articulation, which would be particularly challenging to establish with supervised techniques. In particular, we note that the profile of the quality parameter in Fig. 1.5b features a secondary peak for $Q = 15$. This partitioning involves finer subdivisions of the $Q = 6$ domains, see Fig. 1.6, and hence provides valuable insight into the functional mechanics of GLIC and its connection with the hierarchical quasi-rigid organization of its five constitutive monomers.

1.3.4 NavAb, a voltage-gated ion channel

As a further challenging case we considered the NavAb channel, whose gating action is controlled by the polarization state of the embedding membrane and whose structure has been recently solved [70]. The structural organization of the tetrameric NavAb complex is given in Fig. 1.7a. It features a pore domain, assembled from the last two transmembrane helices of each monomer (conventionally referred to as S5 and S6), and four separate structural domains, each comprising the first four helices of a monomer (S1 to S4), which act as voltage sensors.

When the polarization state of the membrane is altered, the voltage sensing domains undergo a conformational change which displaces the S4 helix

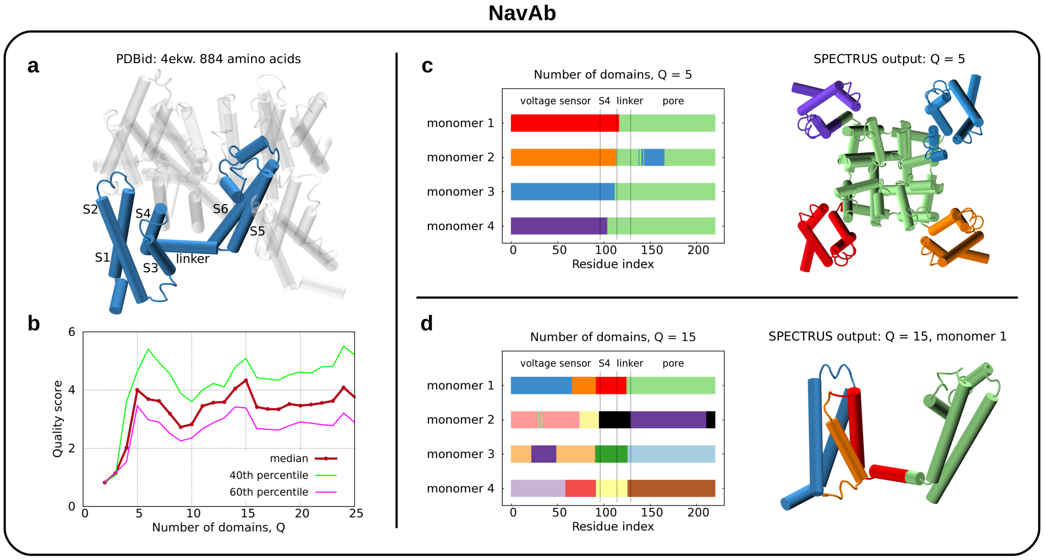


Figure 1.7: Quasi-rigid domain decomposition of the NavAb voltage-gated ion channel based on the distance fluctuation matrix computed from atomistic simulations. The representative structure of NavAb is shown in panel (a) along with the standard labeling of the helices for one of its constitutive monomers. The quality score profile of the SPECTRUS subdivisions shown in panel (b) indicates that NavAb is ideally subdivided into $Q = 5$ main quasi-rigid domains or into $Q = 15$ finer ones. The subdivision into $Q = 5$ domains is illustrated in both sequence-wise and structural representation in panel (c), while the one for the $Q = 15$ case is given in panel (d).

perpendicularly to the membrane [71–74]. The displacement is next propagated allosterically via the S4-S5 linker to the tetrameric pore causing its opening or closing. One standing issue regards how these displacements are coupled.

To address this point, we applied the spectral domain decomposition analysis to the NavAb complex using as input the distance fluctuation matrix computed from a set of six extensive atomistic molecular dynamics simulations sampled by Amaral *et al.* [62]. Four of these trajectories sampled the main steps along the voltage sensor activation pathway, but with the pore still closed, while the remaining ones featured respectively a partially-open and an open pore.

The results of the SPECTRUS analysis are summarized in Fig. 1.7. The quality score profile in Fig. 1.7b features a peak for $Q = 5$ quasi-rigid domains. Such subdivision corresponds to the intuitive partition of the tetrameric complex into the four separate voltage sensing domains plus the core.

The interesting and informative point regards the location of the boundary between the pore domain and the voltage sensing ones. It is seen that for all four monomers this boundary occurs systematically in the loop connecting the S4 helix to the S4-S5 linker between residues 105 and 118, see Fig. 1.7c.

This is a very valuable clue for the mechanics underpinning the pore gating mechanism. In fact, it indicates that a hinge is present allowing displacements of the voltage sensor domain with respect to the pore domain and that the S4 helix is more rigidly connected to the voltage-sensor domain than to the pore domain. The intervening S4-S5 linker is instead co-opted in the quasi-rigid pore macrodomain in all four monomers. Because the latter are treated independently, i.e. no symmetry of the subdivision across the monomers is enforced *a priori*, one concludes that the mechanical coupling of the linker and the pore domain is robust. As a matter of fact, the linker and the pore domain section of each monomer are recognized as two distinct quasi-rigid units when one considers the finer subdivision in $Q = 15$ domains, corresponding to the second peak of the quality score, see Fig. 1.7d. Other aspects of this finer subdivision, however, are less conserved across the monomers, arguably due to the still imperfect sampling of the MD trajectories. Specifically, while the voltage sensing domain is subdivided in two domains in all monomers, there appears to be two alternative locations for the boundary. For the aforementioned reasons, the profiles of the median, 40th and 60th percentiles of quality score do not follow exactly the same trend.

Importantly, the results from the $Q = 15$ domains subdivision suggest an activation mechanism in which the displacement of S4 results in a motion of the linker that releases the steric hindrance exerted by the latter on the pore domain in the resting/closed state; this observation disfavors the alternate scenario in which the linker exerts an active pulling on the pore domain. These quantitative indications ought to be valuable for designing future studies aimed at elucidating by more direct means the mechanical workings of the pore complex.

1.3.5 Viral capsids: STMV and TrV

As a last applicative avenue, we discuss the quasi-rigid domain decomposition of viral capsids. Identifying the mechanical, quasi-rigid units of viral

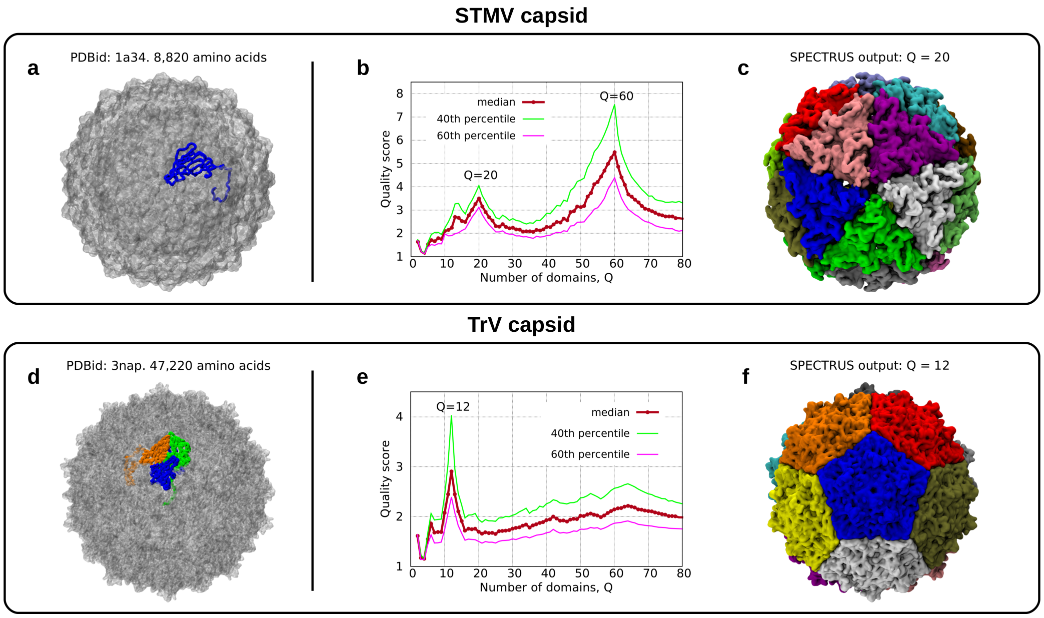


Figure 1.8: Quasi-rigid domain decompositions of the capsids of the satellite tobacco mosaic virus (STMV) and of the Triatoma virus (TrV). The crystal structure of STMV is shown in panel (a) where the protein defining the asymmetric unit is highlighted. The quality-score profile of the SPECTRUS subdivisions shown in panel (b) indicates that STMV is ideally subdivided into $Q = 60$ domains, which correspond to the intuitive partitioning into the individual constitutive proteins. The next, coarser subdivision, corresponds to $Q = 20$ quasi-rigid domains. Their structural representation reveals that they correspond to trimers. The reference structure and SPECTRUS quality score of the Triatoma virus are shown in panels (d) and (e), respectively. The optimal partition into $Q = 12$ units is represented in panel (f), which clarifies that the units are all pentameric.

shells has several practical ramifications: it is important for singling out the fundamental assembly or disassembly units and for identifying the functional blocks that preside structural changes such as those involved in the maturation steps [56, 75–77].

Even the smallest viral capsid is much larger than any of the complexes considered so far and is constituted by dozens of proteins. Accordingly, the subdivision task is significantly more challenging than the previously discussed cases, even for obtaining the basic input data for general domain decompositon strategies, that is the matrix of pairwise amino acid distance fluctuations, f. In fact, alternative crystal structures are usually not available and the large capsids size makes it largely impractical to compute the f

matrix using atomistic molecular dynamics simulations.

To tackle the problem, we accordingly resorted to the use of elastic network models (ENMs) which, thanks to the specific properties of proteins [16] and of their free energy landscape, can reliably reproduce the equilibrium structural fluctuations of proteins and protein assemblies starting from the input of a single reference crystal structure [25, 77–79]. The applicability of these models to viral capsids has been previously demonstrated in the context of viral capsid maturation [75–77] and for the supervised identification of geometrically-stable blocks through auxiliary parameters related to their shape homogeneity and integrity [56].

Here, we use a β-Gaussian ENM, which accounts for both main- and side-chains in proteins, to obtain a model distance fluctuation matrix, f, for the two icosahedral capsids of two viruses: namely the satellite tobacco mosaic virus (STMV) and the Triatoma virus (TrV). The STMV was chosen for validation purposes. In fact, it is the smallest known viral capsid and the first one for which the mechanical stability of the protein shell has been probed by all-atom molecular dynamics simulation [52, 80]. The functional domains of TrV, instead, have so far been probed only by nano-indentation experiments and characterized from the inspection of rupture debris [55, 81]. For this still relatively unexplored system, the identification of the quasi-rigid domains by theoretical/computational means can add valuable insight about the nature of the mechanical domains.

We accordingly start by discussing the SPECTRUS application to the STMV capsid consisting of 60 identical proteins for a total of 8820 residues, see Fig. 1.8a. The distance fluctuation matrix derived from ENM data (see Appendix B for details about its calculation) was next used as input for the spectral decomposition from $Q = 2$ to $Q = 80$ quasi-rigid domains. This range was chosen because it covers from the coarsest possible subdivisions to finer ones that are smaller than the constitutive proteins.

The profile of the quality order parameter is shown in Fig. 1.8b. It is seen that there exist a prominent peak for $Q = 60$ and a secondary one for $Q = 20$ domains. The same peak profiles are seen if a plain anisotropic elastic network and a different interaction cutoff are used, see Fig. 1.9a.

The subdivision into $Q = 60$ domains practically coincides with the physically-viable partitioning into single constitutive proteins. Although this result is intuitive, we stress that it is obtained by using only the ENM-based internal structural fluctuation of the capsid, with no explicit reference to the sequence or structural boundaries of the individual proteins.

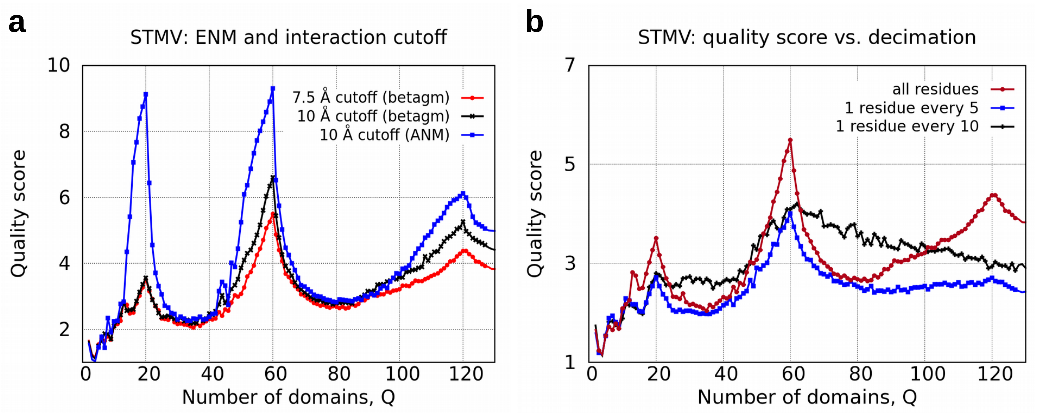


Figure 1.9: (a) The use of different interaction cutoffs for the ENM (here, we considered 7.5 Å and 10 Å) doesn't significantly affect the overall partitioning and quality score. The relative heights of the peaks in the score profile, however, could differ when using, for instance, ANM instead of the β-Gaussian ENM. For ANM, a cutoff of 10 Å has been used, which is about the minimum interaction range yielding only six null modes (corresponding to roto-translations). (b) Quality score from SPECTRUS for the STMV capsid, plotted against the ones obtained by retaining in the analysis only one residue every 5 or 10. The cutoff value $\bar{\delta}$ has been set to 20 Å to make up for the increased separation of proximal amino acids in the reduced set. The fact that for the 1-every-5 case the main peaks in the score profile are preserved suggests the feasibility of the reduction scheme, which can be important for keeping computational costs at a manageable level for particularly large systems.

The coarser subdivision into $Q = 20$ domains is therefore more informative and relevant for the multimeric mechanical units of the capsid. Fig. 1.8c shows the associated subdivision which is very symmetric despite the fact that no *a priori* symmetry was enforced. It is readily seen that each domain corresponds to trimeric assemblies of the individual capsid proteins. These trimers are, indeed, the correct mechanical units for STMV, as it has been previously established by computational studies on STMV structural stability [52] and in our previous ENM-based analysis of functional units based on the analysis of several order parameter specifically tailored for viral shells [56]. We also note that, as for the case of AKE, in this context too the innate character of the subdivisions into $Q = 60$ and $Q = 20$ domains eludes the analysis based on the inspection of the Laplacian matrix spectrum, see Fig. 1.10. This latter fact is not surprising, because the approximate quasi-rigid character of the domains makes it particularly challenging to detect genuine gaps in the spectrum, especially at relatively high indices of the ranked eigenvalues.

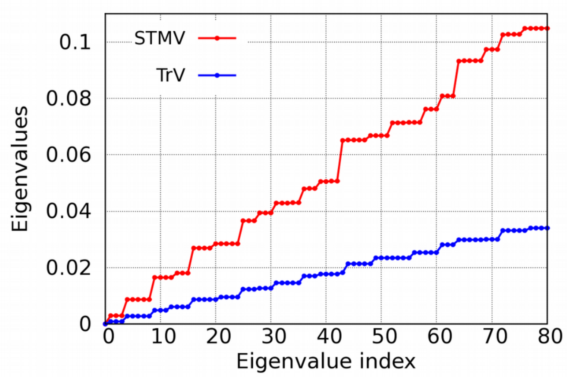


Figure 1.10: Eigenvalue spectrum of the Laplacian matrix for STMV and TrV. In ideal contexts, the viable number of egenvectors (clusters) to be used for the projection step is indicated by the presence of one or more gaps in the spectrum. The approximate quasi-rigid character of the dynamical domains does not allow for a clear, objective identification of significant gaps in the spectrum. In particular, no discernible features are observed in correspondence of the subdivisions for $Q = 20$, 60 for STMV and $Q = 12$ for TrV, while they are readily singled-out from the quality score profiles.

These considerations underscore the discriminatory capability of the quality score even for a wide dynamic range of Q values.

Building on the successful validation of the STMV domain decomposition, we next considered the still largely unexplored case of TrV. The recently-solved structure of the TrV capsids is shown in Fig. 1.8d: it is formed by 180 proteins which come in three structurally-nonequivalent types, colored differently in the panel, for a total of 47,220 amino acids.

In this case too the capsid was subdivided in a number of domains ranging from 2 to 80 and the profile of the associated quality order parameter is shown in Fig. 1.8e. The most prominent peak corresponds to the $Q = 12$ solution illustrated in Fig. 1.8e. This subdivision involves twelve identical pentagonal units, each formed by 15 proteins. The markedly high value of the quality score for this subdivision, compared to that for more or fewer domains, is a strong indication of the robust, innate character of these multimers as the fundamental mechanical blocks of the capsid.

As a matter of fact, this result is fully consistent with the conclusions of Snijder *et al.* [55] who, by inspecting the debris of TrV capsids ruptured by an AFM tip, concluded that the most plausible mechanical blocks were the same pentagonal units observed here. The present result therefore reinforces, from an independent quantitative perspective, the earlier conclusions based on nano-indentation experiments.

This consensus, in turn, adds confidence to the viability of the present approach for identifying the quasi-rigid units from individual proteins to large macromolecular assemblies.

For the latter, computationally onerous contexts, the scope of elastic networks can be extended by using a coarse-grained representation of the capsid where, for instance, only one in five amino acids are retained. As shown in Fig. 1.9b, this structural simplification procedure, if kept within reasonable limits, can significantly speed up the numerical calculation without impairing the overall correct identification of the rigid domains.

1.4 Conclusions

In conclusion, we introduced and used a transparent and transferable method for identifying both the number and type of quasi-rigid domains in proteins and protein complexes involving from few hundreds to tens of thousands amino acids.

The method, named SPECTRUS, takes as input the matrix of pairwise distance fluctuations of amino acids, which can be obtained from various sources: it can be either computed from a limited number of available crystal structures, or from conformations sampled with extensive molecular dynamics trajectories, or derived from elastic network models, when a single conformation of the molecule of interest is available. The partitioning into quasi-rigid domains is recasted as a clustering problem. A key element of the strategy is the preconditioning step where the distance fluctuation matrix is projected (via the essential spaces of its Laplacian) in a space of low dimensionality which is ideally suited to expose the innate groupings of amino acids taking part to different quasi-rigid domains. This step has two major advantages. First, it is instrumental for making the subdivision robust and practically independent of the specific method of clustering. Secondly, it is crucial for providing a quantitative basis to assess the quality of a given subdivision, that is the extent to which the observed intra-cluster compactness and inter-cluster separation differ from equivalent random subdivisions. By these means, it is possible to single out the most significant number and type of subdivisions of a given protein and protein complex.

The viability of the SPECTRUS approach was ascertained by applying it to a number of well-characterized cases, which are used for validation purposes, as well as open and debated ones. The former include adenylate kinase, the GLIC channel and the STMV capsid, for which the customary, supervised subdivisions are all correctly reproduced. For the more challenging and open cases we instead considered the NavAb voltage-gated ion channel and the viral capsid of the Triatoma virus, for which the decomposition provides valuable insight regarding their still relatively unexplored functional mechanics.

The source code of SPECTRUS is made freely-available, upon request, for academic use.

1.5 Appendix A: robustness of domain decompositions

To simulate the effect of missing/unresolved residues, we applied SPECTRUS after removing the 10 and 20 most mobile amino acids (those with highest B-factors) from the open and closed conformers of adenylate kinase. The 10 and 20 excluded residues have indexes: 41, 75-79, 127, 129-130, 157, and 41, 74-79, 127, 129-131, 142-143, 151-152, 157, 160, 211, 213-214, respectively.

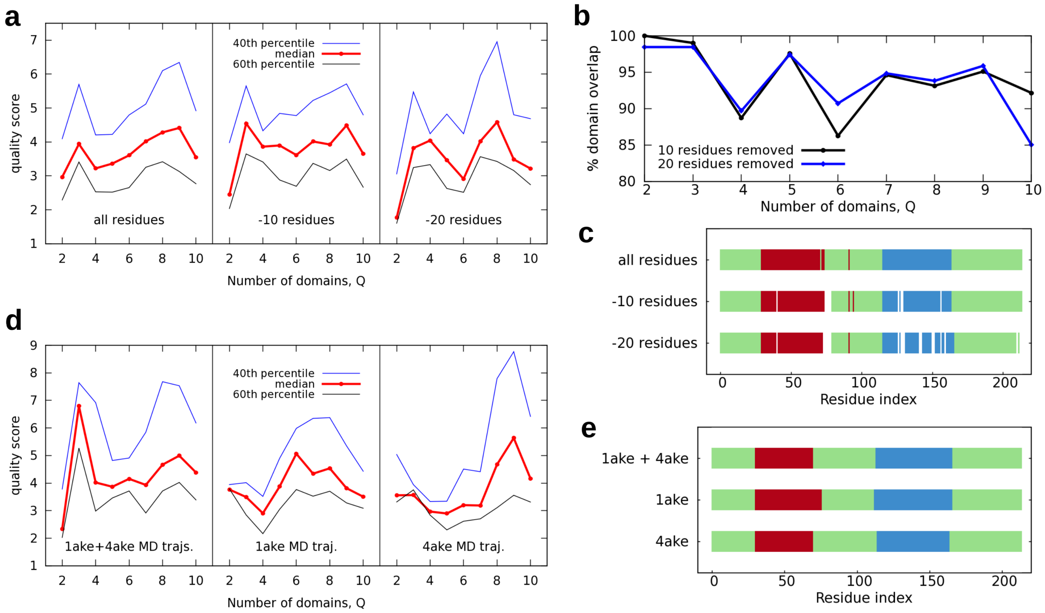


Figure 1.11

We observed that the progressive removal of residues can diminish the sharpness of the quality score peaks (Fig. 1.11a), which can be assessed *a posteriori* by comparison with those of the 40th and 60th percentiles. The consistency (or lack thereof) of the three percentile trends allows to judge whether a clear indication of the number of innate domains emerges from the available data. It is seen that after removing 20 amino acids the peak for $Q = 3$ is still discernible, though it has lost its original prominence. Nevertheless, the actual subdivision is much more robust, see Fig. 1.11c (blank gaps refer to missing residues). The degree of consistency (percentage of residues assigned to the same domains, see Fig. 1.11b), with respect to the all-residues partitioning up to 10 domains, is of 94.1% on average (with a minimum value of 86.3%) for the first case, and of 93.8% on average (minimum: 85.1%) for the second case.

As a further test of the domain decomposition robustness against the input dataset size, we repeated the analysis of adenylate kinase by considering separately each of the two MD trajectories (one started from 1AKE, the other from 4AKE) that were used jointly in the main text (Fig. 1.11d). The decompositions performed on each individual trajectory, while still producing the same domains (at least 97.7% of domain overlap, see Fig. 1.11e), present noticeable differences in the quality score profiles. The ability of the score in detecting the innate domain number can be qualitatively assessed *a posteriori* by comparing the profiles of the 40th and 60th percentiles with respect to the median. Based on this criterion, the higher discriminatory power of the two combined trajectories relative to each of them emerges very clearly.

Another interesting analysis consisted in computing the quasi-rigid domain decomposition for an homologous enzyme, namely *Streptococcus pneumoniae D39* adenylate kinase (see Fig. 1.12a), using its 4 structures available in the PDB: 4W5H, 4NTZ (open), and 4W5J, 4NU0 (closed). The quality score shows the same two prominent peaks in $Q = 3$ and $Q = 9$ seen for the pair of open and closed *E. coli* adenylate kinase structures, 1AKE and 4AKE, discussed in the main text. As it is shown by the rightmost panel, the $Q = 3$ partitioning for the two homologous kinases are practically identical.

Finally, we considered three additional cases, whose functional domains had been already detailed in a recent work [68]. In Figs. 1.12b–d we show the domain decompositions of HIV-1 RT, p38 MAP kinase and cyclin-dependent kinase 2 based on datasets consisting of various numbers and combinations of their structural representatives (identified as those highlighted in Figs. 1a, 2a, 3a of ref. [68]).

HIV-1 RT (Fig. 1.12b) is an enzyme composed of two subunits, p66 and p51. We focused on the p66 unit, which is customarily partitioned into 5 functional domains. Based on its 6 representatives structures (1HQE, 1N6Q, 1RTJ, 1TVR, 1VRT, 3DOK) the optimal decomposition is for 5 quasi-rigid domains, which are in good correspondence with the functional ones. These subdivisions remain practically unchanged if various combinations of only two of the 6 representatives are used: their average overlap with the reference (6-representatives) subdivision is always larger than 95% (98.0% on average). For some pairs, however, a degradation of the quality score peak is seen.

Analogous results, i.e. robustness of the subdivisions for fixed number of

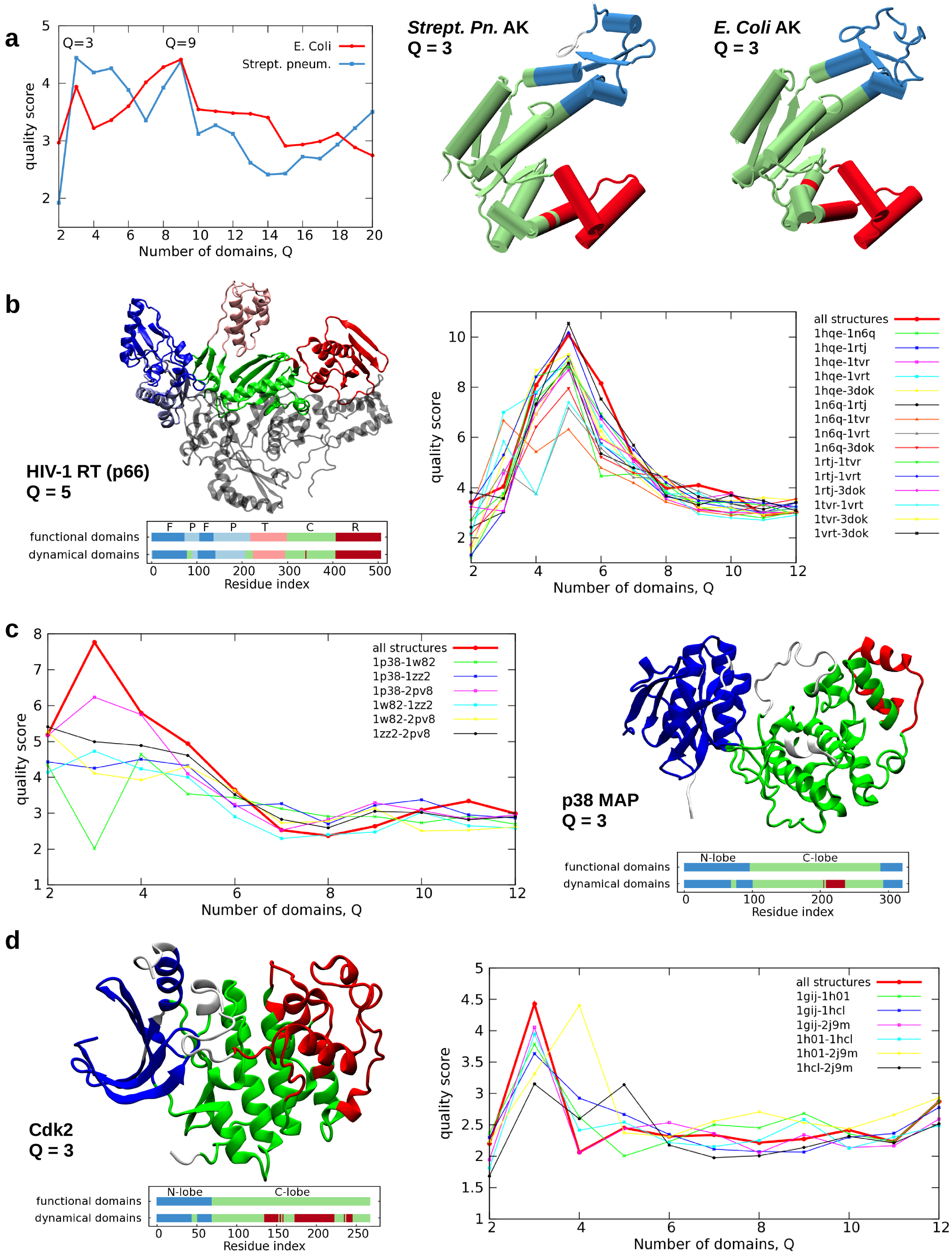

a
Q=3
Q=9
E. Coli
Strept. pneum.
quality score
Number of domains, Q
Strept. Pn. AK
Q = 3
E. Coli AK
Q = 3
b
HIV-1 RT (p66)
Q = 5
functional domains
dynamical domains
F P F P T C R
Residue index
all structures
1hqe-1n6q
1hqe-1rtj
1hqe-1tvr
1hqe-1vrt
1hqe-3dok
1n6q-1rtj
1n6q-1tvr
1n6q-1vrt
1n6q-3dok
1rtj-1tvr
1rtj-1vrt
1rtj-3dok
1tvr-1vrt
1tvr-3dok
1vrt-3dok
quality score
Number of domains, Q
c
all structures
1p38-1w82
1p38-1zz2
1p38-2pv8
1w82-1zz2
1w82-2pv8
1zz2-2pv8
quality score
Number of domains, Q
p38 MAP
Q = 3
functional domains
dynamical domains
N-lobe
C-lobe
Residue index
d
Cdk2
Q = 3
functional domains
dynamical domains
N-lobe
C-lobe
Residue index
all structures
1gij-1h01
1gij-1hcl
1gij-2j9m
1h01-1hcl
1h01-2j9m
1hcl-2j9m
quality score
Number of domains, Q

Figure 1.12

domains, and occasional degradation of the sharpness of the quality score peaks are seen upon impoverishing the datasets of representative structures for p38 MAP kinase (Fig. 1.12c) and Cdk2 (Fig. 1.12d). For the former, the used representatives are: 1P38, 1W82, 1ZZ2, 2PV8 (1OZA was omitted because significantly shorter) and the $Q = 3$ subdivisions with the reduced datasets have average domain overlap of 91.9% (with a minimum value of 70.4%) with the global one. For Cdks the used representatives are 1GIJ, 1HCL, 2J9M, 1H01, and the $Q = 3$ subdivisions with the reduced datasets have average domain overlap of 97.7% (minimum value: 92.2%). For both p38 MAP kinase and Cdk2, the largest quasi-rigid domains are in good correspondence with the two functional ones, although the SPECTRUS subdivisions are finer as they feature a third, smaller domain.

1.6 Appendix B: distance fluctuation matrix from elastic network models

A model distance fluctuation matrix, f, can be computed from a single reference structure by using elastic network models. To this purpose we used the β-Gaussian network model which uses a two-centroid amino acid description (mainchain and sidechain) and has been successfully validated against atomistic MD simulations [78, 79].

Given a set of ENM (non-zero) eigenmodes, $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n\}$, and their corresponding eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, one constructs an auxiliary, "meta-ensemble" of conformations, each of which is obtained by adding to the reference protein structure a deformation vector obtained by a linear stochastic superposition of the modes: $\epsilon \sum_{i=1,\ldots,n} \eta_i \, \vec{v}_i$, where η_i is drawn from a Gaussian distribution with zero mean and variance equal to $1/\lambda_i$. The f matrix is next straightforwardly computed from the conformational ensemble.

Notice that, because the entries of f are proportional to ϵ, the relative magnitude of different entries of this matrix (which is what controls the clustering scheme) are independent of ϵ, which then can be set to any convenient value, e.g. equal to 1.

The construction of the auxiliary ensemble can be sped up by restricting the stochastic superposition to the lowest energy modes which are those that account for most of the structural fluctuations (e.g., 10 modes suffice *a posteriori* to yield a converged f for capsids of ~ 9000 amino acids, see Fig.

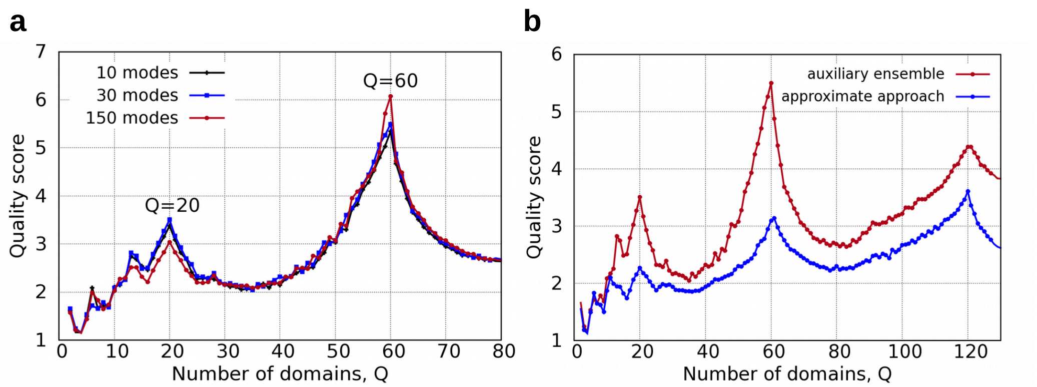


Figure 1.13: (a) Quality score profiles for STMV, illustrating the dependence on the number of low-energy elastic modes used for computing the covariance matrix. It is striking that 10 modes suffice to identify unambiguously the main peaks in the profile. (b) Quality scores for STMV relative to both the approaches that can be used to derive the distance fluctuation matrix from ENM. In the first approach, used for the cases discussed in the Applications section, an auxiliary ensemble of conformations is produced by a stochastic superposition of the lowest-energy eigenmodes. In the second one, the distance fluctuations are calculated directly from the ENM covariance matrix using an approximate expression.

1.13a). The results presented in the text were obtained by using 30 modes.

For a faster, albeit approximate approach, one can obtain f from: $f_{ij} = \sqrt{\sum_\alpha C_{ii}^{\alpha\alpha} + C_{jj}^{\alpha\alpha} - 2C_{ij}^{\alpha\alpha}}$, where $C_{ij}^{\alpha\beta}$ is the ENM covariance matrix element pertaining to centroids i and j and Cartesian components α and β. This approximate method yields consistent results for the location of the quality score peaks, although not necessarily for their relative height, see Fig. 1.13b.

Chapter 2

Unifying view of mechanical and functional hotspots across class A GPCRs

G protein-coupled receptors (GPCRs) are the largest superfamily of signaling proteins. Their activation process is accompanied by conformational changes that have not yet been fully uncovered. Here, we carry out a novel comparative analysis of internal structural fluctuations across a variety of receptors from class A GPCRs, which currently has the richest structural coverage. We infer the local mechanical couplings underpinning the receptors' functional dynamics and finally identify those amino acids whose virtual deletion causes a significant softening of the mechanical network. The relevance of these amino acids is demonstrated by their overlap with those known to be crucial for GPCR function, based on static structural criteria. The differences with the latter set accordingly point to sites whose functional role is more clearly detected by considering dynamical and mechanical properties. Of these sites with a genuine mechanical/dynamical character the top ranking is amino acid 7x52, which we accordingly point out as a previously unexplored, and experimentally verifiable key site for GPCR conformational response to ligand binding.

The work presented here was done in collaboration with Giulia Rossetti[1], Paolo Carloni[1] and Cristian Micheletti, and a manuscript is currently under review in the PLOS Computational Biology journal.

[1] IAS/INM, Forschungszentrum Jülich, Jülich, Germany

2.1 Introduction

Mammalian G protein-coupled receptors (GPCRs) are the largest family of signaling proteins, with approximately ∼850 unique members up to now identified in the human genome [82,83]. Given the size of this family, their ubiquitous expression, and their involvement in virtually every (patho)physiological process in mammals, it is not surprising that human GPCRs are targeted by more than half of current drugs [84].

GPCRs share a distinctive structural signature, namely seven α-helical transmembrane (TM) domains [85]. Such common structural organization strongly contrasts with the structural diversity of the agonists: these range from subatomic particles (a photon), to ions (H^+ and Ca^{++}), to small organic molecules, to peptides and proteins [85]. The presence of an agonist (or a photon in the case of rhodopsin) triggers specific downstream G protein-dependent signaling pathways.

The mechanisms that precisely control GPCR agonist binding and the following receptor activation have until very recently been hindered by a lack of crystallized active receptor states and receptor-ligand complexes. However, significant advances in crystallization has recently permitted the structural determination of several class A receptors in active state. Moreover, several mutagenesis and assay procedures were performed in an attempt to identify functionally important residues [86], along with specific micro-switches, i.e. small groups of residues that undergo conformational change during receptor activation [87, 88].

Despite a consolidated list of residues important for binding and/or function emerged, the findings are limited by their individualized nature [89].

Indeed, GPCRs are not rigidly switching between the alternative agonist-bound and inactive forms. They rather adopt a series of intermediate conformations influenced not only by association with ligands, but also by other receptors, signaling and regulatory proteins, by post-translational modifications, and by environmental cues [83]. The capability of GPCRs to engage with such diverse signaling machinery strongly depends on their conformational flexibility. All these diverse signaling events are indeed accompanied by dynamic conformational changes. Each state is likely represented by an ensemble of conformations [90].

A characterization of the conformational and structural dynamics of these proteins is therefore critical for understanding the molecular mechanisms underlying their function. A suitable comparative analysis of the available structures for these receptors ought to give insight into their structure–function relationship by clarifying the functional-oriented character of their internal dynamics.

While the inspection of GPCRs' and G proteins' structures has been essential to map out the accessible distinct signaling states, our knowledge is still limited regarding the internal dynamics of such states and the pathways that link them [91].

To our knowledge this problem has not yet been addressed systematically. The reason for its challenging character lies, at least in part, in the high structural heterogeneity of the conformers that bridge between the active and inactive forms. Such structural diversity, for instance, limits *a priori* the scope of general methods, such as elastic networks and normal mode analysis, which can otherwise be profitably used to identify low-energy collective modes from near-native fluctuations [92, 93].

Here, we introduce and apply a novel comparative tool that can single out those sites that act as hubs in the network of mechanical connections between the receptor residues, i.e. that are crucial for maintaining the integrity of the protein's large-scale dynamics and mechanics.

We present and discuss this strategy, which is otherwise general and transferable, to the members of a specific GPCR class, namely the class A. This functional group was chosen precisely because of its well-populated and structurally diverse repertoire of conformers.

We analyze the structural fluctuations across representative conformers to identify those residues that are central for the network of mechanical couplings, and hence the functional dynamics, of the receptors. Such sites have good overlap with known key residues, including those established by purely static structural considerations, but involve additional sites whose functional relevance, that is experimentally verifiable, emerges more clearly from a dynamical perspective.

2.2 Results and discussion

We focus on GPCRs belonging to the rhodopsin-like class A. This class has currently the broadest structural coverage spanning between active, or partially active, and inactive forms. The set includes six different types of receptors, namely: A_{2A} adenosine, β_2 adrenergic, M_2 muscarinic, μ-opioid, neurotensin NTS1 and rhodopsin (see Table 2.2).

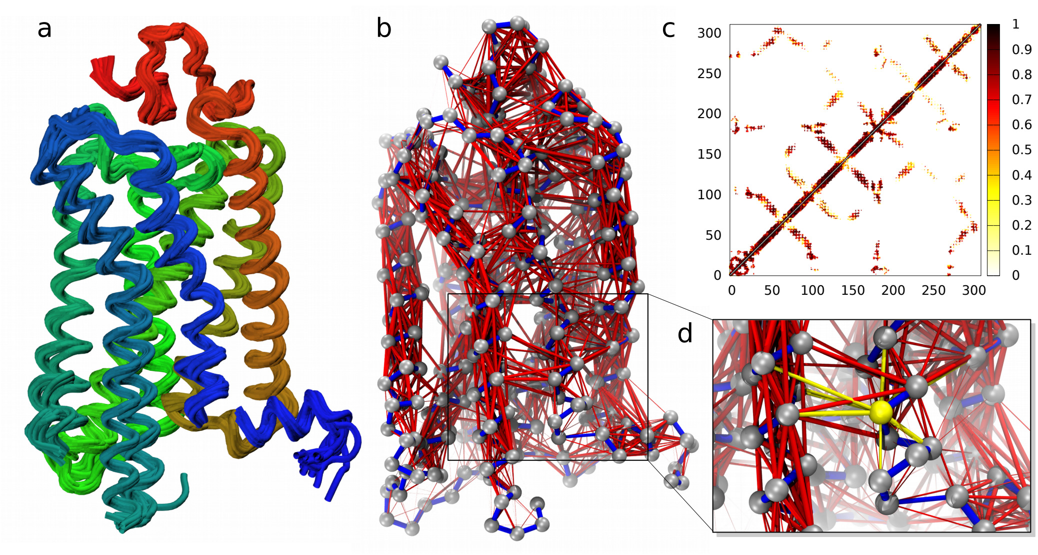


Figure 2.1: Scanning GPCRs' mechanical network for key sites. The structural ensemble of a G protein-coupled receptor, see panel (a) for rhodopsin, is used to compute the distance fluctuations for all pairs of amino acids. (b) The pairings in the local mechanical network (C_α distance < 12Å) are highlighted with red bonds with thickness proportional to the observed rigidity; only the strongest links are shown here, while the full network is shown in Fig. 2.5. The network is represented as a color-coded contact map in panel (c). Key residues for the overall mechanical integrity of the network are identified by measuring how the link connectedness varies when one removes all the links of a node corresponding to non-covalent bonds (highlighted in yellow in panel d).

2.2.1 Identifying the mechanical hubs

The mechanical hubs of these receptors were identified with a three-step strategy described below and sketched in Fig. 2.1, see Methods for further details.

As a first step, for each receptor we first retrieved all available PDB structures covering its conformational repertoire (Fig. 2.1a). Next, for each pair of residues in spatial proximity (within 12Å on average), we computed their distance fluctuations over the structural set. The fluctuation amplitude is a measure of rigidity, and the residues' pairwise distance variance can be used as an inverse measure of residues mechanical couplings [1, 28, 34, 36, 37, 58]. Hence, this step allows to define the local mechanical network that underpins the receptors functional dynamics (Figs. 2.1b–c).

In the final step, each amino acid is profiled based on how much its virtual "mutation", performed by deleting from the network its local mechanical interactions, changes the network's connectivity, an approach similar and alternative to measuring the centrality of a particular node in a network (Figs. 2.1d). The higher is the perturbation induced on the network, the higher is the dynamical impact of the considered amino acid. The returned quantity is a measure of the relevance of each residue in establishing indirect couplings between structural fluctuations across distant parts of the receptors. For this reason we shall refer to it as the "mechanical bridging score".

As we shall discuss later, amino acids with high mechanical bridging score are typically located at the hinge or interface regions between quasi-rigid protein domains and are accordingly well-suited to affect the long-range propagation of structural fluctuations, including functionally-oriented ones. Note that, because we consider intrinsically dynamical properties (structural fluctuations), our notion of bridging score can aptly complement previous GPCRs' profiling based on networking properties defined from single, static, structures [94, 95].

For a robust identification of the aforementioned mechanical hubs, we combined the six mechanical bridging profiles of the different receptors (Fig. 2.6 and 2.7) into a single, average one. The average was taken over the set of corresponding residues (with same GPCRdb numbers [96]) that are shared by all considered receptors.

The resulting average profile is shown in Fig. 2.2. One observes that the highest average bridging scores are found at the interface between transmembrane helices that are known to be relevant for the receptor activation, namely: TM3, TM6 and TM7 [88, 97].

2.2.2 Validating the mechanics-based profiling

The functional relevance of sites with high average bridging score can be shown more stringently by cross-referencing them with the list of currently known key residues for class A receptors based on the survey of Tehan *et al.* [97]. This list of residues was recently compiled by combining sequence- and structure-based selection criteria, that is by singling out residues that are both highly conserved as well as located along the pathway that structurally connects the orthosteric site and the G protein docking site. This connecting region coincides with a hydrophobic core that is central to the helix bundle. The top ranking sites for the average bridging score and those reported in ref. [97] are given in Table 2.1.

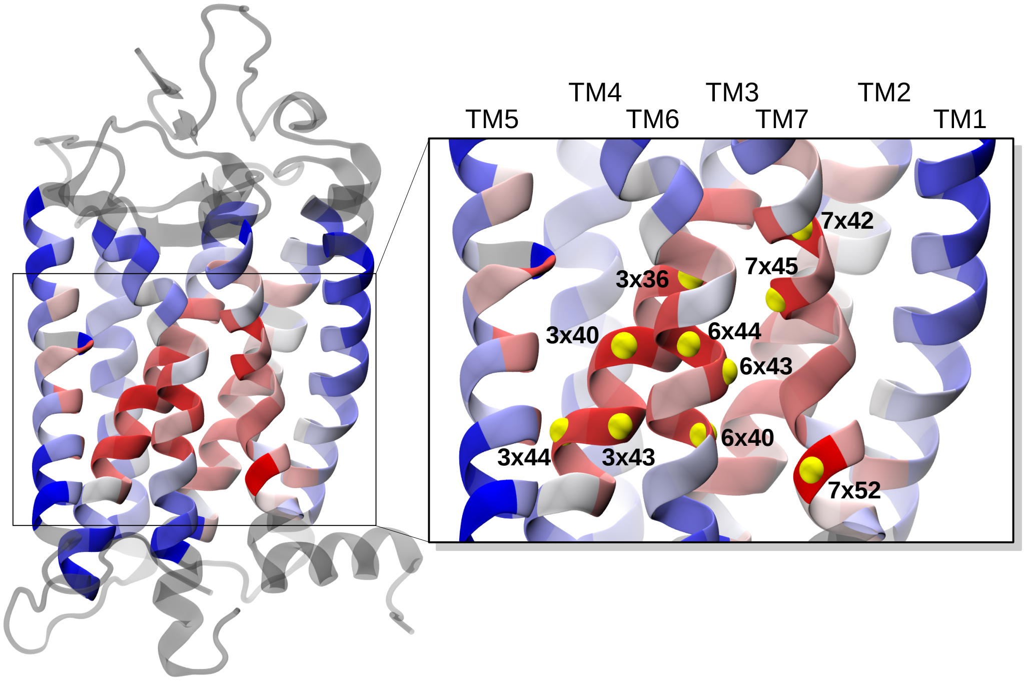


Figure 2.2: Color-coded profile of the average bridging score. Amino acids in a reference GPCR structure (rhodopsin, PDB ID: 1F88) are color-coded according to the mechanical bridging score averaged over all receptors (blue to red from low to high scores). Residues shown in grey are those with no equivalent positions across the receptors' ensemble. The top ten ranking sites, listed in the first column of Table 2.1, are labelled and highlighted with yellow beads in the inset.

The overlap between our top ranking sites and the known key functional residues reported by Tehan *et al.* [97] was assessed by using the receiver operating characteristic (ROC) curve in Fig. 2.3a. The curve shows that by running through our ranked list of residues, the "discovery" of the known functional sites occurs at a significantly higher rate than expected for a random reference case (the plot diagonal).

This is an indication that the average bridging score is able to capture with a significant degree of sensitivity those residues likely to be involved in the functionality of class A GPCRs.

This conclusion is further supported by comparing the ranking based on the average bridging score with one based on a purely static structural criterion. To this end, we ranked the amino acids based on their number

top sites for average bridging score	key functional sites (Tehan *et al.* [97])	
7x52	1x50	**6x40**
3x40	2x46	6x41
7x42	2x50	**6x44**
6x44	**3x40**	6x48
7x45	**3x43**	6x50
3x43	3x50	7x49
3x36	4x50	7x50
3x44	5x50	7x53
6x43	5x58	
6x40	6x30	

Table 2.1: Key mechanical and functional sites. The first column provides the ranked list of sites with the highest mechanical bridging score averaged over all receptors of class A. The list of known key functional sites for the same class is shown in the second column. Residues present in both lists are highlighted in boldface.

of contacts. This allows for a transparent and equal-footing comparison, since the criterion exclusively considers the average amino acid connectedness, regardless of whether a contact is associated to a strong (i.e. rigid-like) mechanical coupling or not. This structure-based ranking criterion is inspired by previous works on GPCRs [94, 95] that demonstrated a correlation between sites with functional relevance and graph properties of the static contact map build on single receptor structures, a fact that is confirmed by the marked deviation of the corresponding ROC curve from the diagonal in the plot of Fig. 2.3a. The key observation that is relevant here is that the average bridging score ROC curve is well in line with the structure-based one, thus underscoring the functional significance of the mechanics-based ranking criterion. In addition, it prompts to understand the different insight that it can offer over pure structural approaches.

To clarify the latter point, we show in Figs. 2.3b-c and 2.8 the profiling of residues according to the dynamical or structural criteria. The comparative inspection indicates that the differences are mostly localized at specific portions of TM6 and TM7, which are high ranking for the mechanical bridging score, but not for the structural one. These regions, therefore, appear to have a key role across class A members that is genuinely tied to the receptors' functional mechanics and hence cannot be detected from static structural observables.

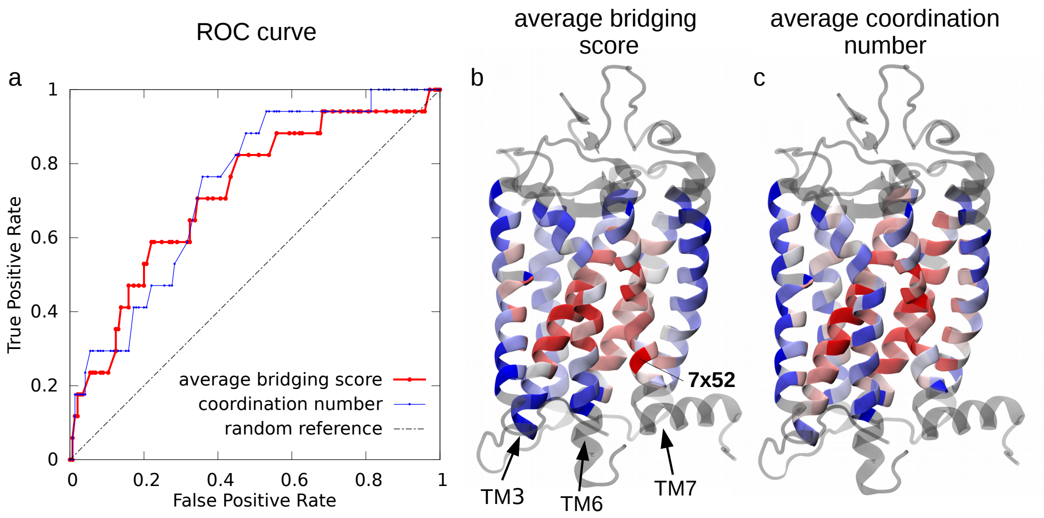


Figure 2.3: Functional profiling of key sites for GPCR's mechanical and structural networks. (a) The list of known GPCRs functional sites in Table 2.1 are used for the ROC curve profiling of the top mechanical sites in Table 2.1 (red) and of those that have the highest structural coordination (number of contacts) across the receptor ensemble (blue). For reference, the performance of a random classifier is shown by the dashed black line. Color-coded representations of the average bridging score and of the average coordination number are shown for rhodopsin (PDB ID: 1F88) in panels (b) and (c), respectively. The representation in panel (b) is the same as in Fig. 2.2. The coordination number averaged over the six receptors shown in panel (c) ranges from 18.7 (blue) to 47.4 (red).

2.2.3 Functional role of key mechanical hubs

The 10 sites with the highest average bridging score (Table 2.1) include residues forming the so-called hydrophobic hindering mechanism (HHM: 6x44, 3x43 and 6x40). Mutagenesis experiments have shown that this conserved hydrophobic triplet, that is contacted by other listed residues, namely 3x40, 6x43 and 3x44, is essential for the activation process of class A GPCRs [97]. The HHM triplet plus the proximal site 3x40, which has the second highest score, all take part to the structural rearrangements bridging the inactive and active state. The latter, in fact, depends on the HHM opening for establishing the water channel in the active conformation [97]. Residue 3x40 additionally participates to the transmission switch [88] and is highly conserved as a large hydrophobic residue as well [97].

Residue 7x42 is, instead, involved in a different molecular switch, i.e. the TM3-TM7 lock [88]. This is the main mechanism responsible for activation in rhodopsin and possibly one of the first switches triggered by ligand binding in other GPCRs. Position 7x45 is one of the most conserved residue in TM7 [88]. Finally, the 3x36 position, though not conserved, was shown by site-directed mutagenesis experiments to have a stabilizing role for the inactive state [88].

Most of the top scoring residues listed in Table 2.1 are therefore sites with a demonstrated involvement in class A GPCRs activity. This validates the viability of dynamical profiling approaches in general, and the mechanical bridging score in particular, for singling out functionally important residues and providing a rationale for their relevance. Given these premises, of particular interest are those sites that have a high bridging score, but are not yet known as functionally relevant.

This is the case for site 7x52, that has the highest score in our analysis. This amino acid is part of the well-conserved motif NPxxY$(x)_{5,6}$F, but is otherwise not particularly central in the static network of contacts, see Fig. 2.3c and Fig. 2.8. Its functional relevance therefore has not been fully investigated before, though its possible participation in stabilising the TM6–TM7 interhelical interaction has been suggested by [98]. Mutations at position 7x52 were shown to constitutively activate the TSH (thyroid stimulating hormone) receptor [86, 99] by possibly disrupting the packing between TM6 and TM7. We therefore suggest site 7x52 as a putative novel site crucial for functionality. Again, the fact that its relevance does not emerge from structural considerations indicates that its role is likely to be a genuinely dynamical, or mechanical one.

We finally note that the highest scoring sites in Fig. 2.2 are immediately adjacent to the region that the latest studies of refs. [100, 101] have identified as the most structurally affected by the activation/inactivation transitions. In particular, by comparing class A GPCRs with different activation states, Venkatakrishnan *et al.* [101] identified three G protein-coupling residues, 3x46, 6x37 and 7x53, whose contacts are disrupted during activation, and that are exposed to the G protein-binding pocket by the dislocation of the cytoplasmic side of TM6 away from the helix bundle.

2.2.4 Analysis of μ-opioid MD simulation and receptors' rigid domains

The conclusions of the previous section are supported by two complementary extensions of the analysis above. Specifically, we first repeated the bottom-

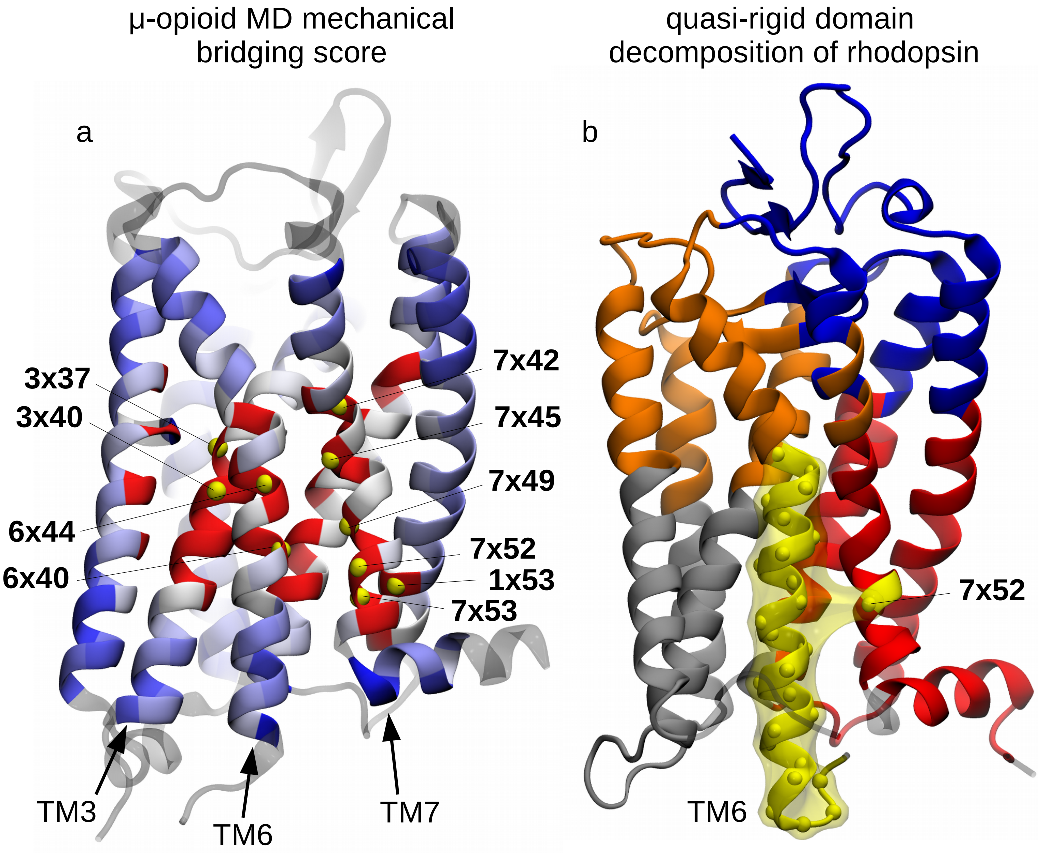

µ-opioid MD mechanical bridging score
quasi-rigid domain decomposition of rhodopsin
a
b
3x37
3x40
6x44
6x40
7x42
7x45
7x49
7x52
1x53
7x53
TM3
TM6
TM7
7x52
TM6

Figure 2.4: Functional role of site 7x52: MD simulations and quasi-rigid domain decomposition. (a) Amino acids of the μ-opioid receptor (PDB ID: 4DKL) are color-coded according to the mechanical bridging score computed from atomistic molecular dynamics simulations. The color convention is the same as in Fig. 2.2, with the top 10 ranking residues being labelled and highlighted with yellow beads, corresponding to the following sites, in decreasing order of score: **6x40**, 7x52, 7x45, **3x40**, 1x53, **7x49**, 7x42, **7x53**, 3x37, **6x44** (in boldface, the key functional sites also present in the list of Tehan *et al.* [97]). Panel (b) shows the optimal SPECTRUS [1] decomposition of rhodopsin into 5 quasi-rigid domains. The TM6-based domain is highlighted in yellow and it notably includes residue 7x52 from TM7. Analogous decompositions for the other receptors are shown in Fig. 2.9.

up mechanical profiling of residues for a single receptor using an ensemble of structures obtained from a molecular dynamics simulation. Finally, we examined the mechanical role of residue 7x52 by using a top-down approach based on the quasi-rigid domain decomposition of all receptors.

For the first extension, we applied our protocol to conformers sampled by extensive atomistic molecular dynamics (MD) simulations of the μ-opioid receptor [102] started from both the inactive state and the ligand-bound active one. The MD ensemble provides a richer sampling of the active and inactive conformers and hence allows to capture the internal dynamics and mechanics with greater fidelity than from the sole pair of available crystal structures.

The results of the single-residue analysis for the μ-opioid (Fig. 2.4a) are well consistent with those of Fig. 2.2, based on the cumulated profiles of all six receptors. Specifically, the highest scoring residues, highlighted in Fig. 2.4a and listed in the caption, include conserved residues of helices TM3, TM6 and TM7, two residues of the HHM (6x40 and 6x44) and, again, site 7x52.

We finally turn to the top-down analysis based on the quasi-rigid domain decomposition of the six class A receptors. To this purpose we used the SPECTRUS webserver [1]. This performs an optimal domain decomposition based on the internal distance fluctuations across a set of representative structures. The analysis, an example of which is illustrated in Fig. 2.4b for rhodopsin, presented two salient features that recurred across the different receptors.

First, the intracellular half of TM helix 6 was systematically identified as a quasi-rigid domain, consistent with its role in the internal rearrangements accompanying the receptors' activation [97].

The second feature is that residue 7x52 is often assigned to the same rigid domain as TM6. Such domain association is interesting because intuitively one would otherwise always assign 7x52 to the TM7-based domain, to which it structurally belongs, see Fig. 2.4b. As a matter of fact, site 7x52 is recognised part of the TM6 dynamical domain in a sizeable fraction ($\sim$ 25%) of the subdivisions from 2 to 10 domains of the receptors, including the μ-opioid MD simulations, see Fig. 2.9. This means that the displacements of 7x52, unlike other sites in TM7, are appreciably coupled with those of the cognate helix, TM6. Accordingly, 7x52 appears to act as an interface, bridging site between the two distinct mobile TM6- and TM7-based domains, as it is illustrated in Fig. 2.4b for rhodopsin.

The recurrent difference of the dynamics- and structure-based assignment is consistent with the other evidence presented above that residue 7x52, whose functional role is still largely unexplored, is likely relevant for the mechanical response of class A GPCRs.

2.2.5 Concluding remarks

The current understanding of GPCRs functionality, and primarily the response to ligand binding, has been significantly shaped by the analysis of the growing number of their structures solved with X-ray or NMR [103]. Though such structures give valuable clues for the active state of GPCRs, they still include a limited set of snapshots of the likely conformational states induced by agonist and G protein binding. In addition, both experiments and atomistic MD simulations indicate that the receptors are capable of adopting multiple conformations, depending on the nature of the bound ligand. Our insight into the agonist- and G protein-initiated conformational changes is therefore still limited.

As a step towards clarifying this open problem, we devised and applied a strategy for identifying key sites presiding the functional dynamics and mechanics of class A GPCRs. This is the largest subclass and it has arguably the widest structural coverage, with conformers from 6 different receptor types (including rhodopsin) in different activation states. We analysed the internal structural fluctuations across the dataset. In particular, we focussed on the pairwise distance fluctuations of corresponding amino acids which were used to infer the network of local mechanical couplings that underpin the large scale, and arguably functionally-oriented conformational changes. The mechanical network was finally analyzed to locate the few sites that most contribute to GPCR's collective mechanics. To do so we identified the residues whose virtual deletion leads to the strongest softening of the overall mechanical response.

The viability of the approach to single out the most relevant functional sites was validated by the significant overlap between key sites for mechanical response and those known to be crucial for function based on independent and different criteria.

On the one hand, this result provides a concrete and vivid illustration of the relevance of dynamics- and mechanics-based criteria for locating key sites for enzyme functionality and hence prompts their use in combination with other more established structure-based static criteria.

On the other hand, the validation revealed that mechanically-relevant sites at interface between transmembrane helices 6 and 7 were not included in the list of previously known functionally-relevant positions. This was particularly the case for site 7x52, which is among the highest ranking ones for the mechanical response, and whose relevance is supported by the analysis of both atomistic MD simulations of the μ-opioid receptor as well as the analysis of GPCR's rigid-domain decompositions.

Based on these convergent indications, we conclude that site 7x52 likely plays a key role in the conformational dynamics of class A GPCRs. Its functional relevance, as well as that of other sites in the central region of the transmembrane helical bundle, ought to be experimentally verifiable, e.g. with site-directed mutagenesis experiments.

2.3 Methods

2.3.1 Network of dynamical similarities

The receptors' mechanical network was inferred from the analysis of distance fluctuations between pairs of amino acids. These, in fact, are key elements to define the subparts of the proteins that interact in such a concerted manner that they behave as quasi-rigid domains [1]. The distance fluctuation $f_{a,b}$ between two residues a and b is computed as the standard deviation of the distances $d_{a,b}$ between their C$_\alpha$ atoms over two or more structures (PDB entries or snapshots from MD simulations):

$$f_{a,b} = \sqrt{\langle d^2_{a,b} \rangle - \langle d_{a,b} \rangle^2}. \tag{2.1}$$

The strength (rigidity) of the pairwise mechanical couplings is then quantified with a Gaussian weighting of the corresponding distance fluctuations

$$\sigma_{a,b} = \exp(-f^2_{a,b}/2\bar{f}^2), \tag{2.2}$$

Because we are interested to define the receptors' mechanical network in terms of physical, local coupling between amino acids, we set $\sigma_{a,b} = 0$ for amino acids whose C$_\alpha$'s are at an average distance larger than 12Å, see Fig. 2.10. The value of the sensitivity parameter, $\bar{f}$, in eq. 2.2 is then set as the average of $f_{a,b}$ over the amino acids pairs closer than 12Å.

2.3.2 Mechanical bridging score

To define the key mechanical bridging sites, or hubs, of the receptors, we resort to the spectral clustering analysis of the mechanical network [65, 66].

Specifically, given the matrix, σ, of couplings between N amino acids, we characterize the spectrum of the symmetric Laplacian matrix,

$$L = I - D^{-1/2}\,\sigma\,D^{-1/2}, \tag{2.3}$$

where I is the identity matrix and D is the degree matrix $D_{a,b} = \delta_{a,b}\sum_c \sigma_{a,c}$. Its non-negative eigenvalues $0 = \lambda_0 \leq \ldots \leq \lambda_i \leq \ldots \leq \lambda_{N-1}$ provide information about how well the network is neatly partitioned in distinct clusters (mechanical domains) and, accordingly, are typically used to define optimal subdivisions of the network.

Here, the eigenvalues will be used for a different goal, namely to ascertain how important is each node to maintain the overall mechanical connectedness of the network. This amounts to measuring how much the network Laplacian spectrum changes when the connections, or couplings, of a node with its neighbors (excluding the connections corresponding to bonded interactions) are deleted. This response for residue k is given by the mechanical bridging score:

$$\Delta_k = \Omega_k - \Omega^0. \tag{2.4}$$

where $\Omega^0 = \tilde{\sum}_{i=1}^{N-1} \frac{1}{\lambda_i}$ is the sum of the inverse eigenvalues (the tilde superscript denotes the omission of zero eigenvalues) for the full network, and Ω_k is the same quantity but calculated for the network where the couplings relative to the kth node have been deleted.

The bridging score profile is computed separately for each receptor using its available structural representatives. The average bridging score is then obtained by averaging the bridging score over all equivalent positions of the various receptors.

2.3.3 Class A GPCRs database

The structures used for the analysis are listed in Table 2.2. Among the receptors whose structure is reported in the Protein Data Bank, we selected those for which both active and inactive conformations were known. These include the following: A_{2A} adenosine, β_2 adrenergic, M_2 muscarinic, μ-opioid, neurotensin NTS1, rhodopsin. Moreover, we applied the same analysis on an

receptor	PDB ID	state	organism	receptor	PDB ID	state	organism
A_{2A} adenosine	3RFM	inactive	H. sapiens	μ-opioid	4DKL	inactive	M. musculus
	3VG9	inactive	H. sapiens		5C1M	active	M. musculus
	3EML	inactive	H. sapiens	neurotens. NTS1	4GRV	active (?)	R. norvegicus
	3QAK	active	H. sapiens		4BUO	inactive	R. norvegicus
	2YDO	p. active	H. sapiens		4BV0	inactive	R. norvegicus
	3PWH	inactive	H. sapiens		4BWB	inactive	R. norvegicus
	3REY	inactive	H. sapiens		4XEE	active	R. norvegicus
	2YDV	p. active	H. sapiens		4XES	active	R. norvegicus
	3VGA	inactive	H. sapiens	rhodopsin	1F88	inactive	B. taurus
	4EIY	inactive	H. sapiens		1GZM	inactive	B. taurus
	4UHR	active	H. sapiens		1L9H	inactive	B. taurus
	3UZA	inactive	H. sapiens		1U19	inactive	B. taurus
	3UZC	inactive	H. sapiens		3DQB	active	B. taurus
	4UG2	active	H. sapiens		4A4M	active	B. taurus
β_2 adrenergic	2RH1	inactive	H. sapiens		3PXO	meta-II	B. taurus
	3D4S	inactive	H. sapiens		2X72	active	B. taurus
	3PDS	inactive	H. sapiens		3PQR	active	B. taurus
	3NY8	inactive	H. sapiens		2J4Y	inactive	B. taurus
	3P0G	active	H. sapiens		2I37	active	B. taurus
	3SN6	active	H. sapiens		2I35	inactive	B. taurus
	3NY9	inactive	H. sapiens		2I36	inactive	B. taurus
	3NYA	inactive	H. sapiens		3CAP	inactive	B. taurus
	4LDE	active	H. sapiens		1HZX	inactive	B. taurus
	4LDL	active	H. sapiens		2HPY	p. active	B. taurus
	4LDO	active	H. sapiens		2G87	inactive	B. taurus
	4QKX	active	H. sapiens		2PED	inactive	B. taurus
M_2 muscarinic	3UON	inactive	H. sapiens		3C9L	inactive	B. taurus
	4MQS	active	H. sapiens		3C9M	inactive	B. taurus
	4MQT	active	H. sapiens		3OAX	inactive	B. taurus

Table 2.2: Structures' dataset. List of PDB structures of the six receptors considered for the bridging score profiling.

MD trajectory as well, obtained by merging two simulations of the μ-opioid receptor [102], starting from the inactive state (PDB ID: 4DKL [104]) and the active state bound to the agonist BU72 (PDB ID: 5C1M [105]).

Each of the six receptors included in our dataset had a minimum of two crystal structures (μ-opioid) and a maximum of 21 (rhodopsin), including both active and inactive conformations.

The GPCRdb numbering scheme [96] has been used to match the residue positions common to all receptors. This scheme consists of the combination of two numbers in the form AxBB, where the first one is the helix number, while the second one is a progressive number chosen so that the most conserved residue in each helix has the value of 50.

When defining the set of common positions, those residues, close to the intra- and inter-cellular regions, for which the process of cutting the surrounding connections could lead to unwanted disconnections of the network, were not included. The remaining set of positions correspond to the transmembrane region of the receptors, with numbering: 1x36 - 1x56, 2x40 - 2x63, 3x24 - 3x54, 4x42 - 4x61, 5x38 - 5x60, 6x34 - 6x57, 7x36 - 7x43, 7x45 - 7x55.

2.4 Appendix: additional figures

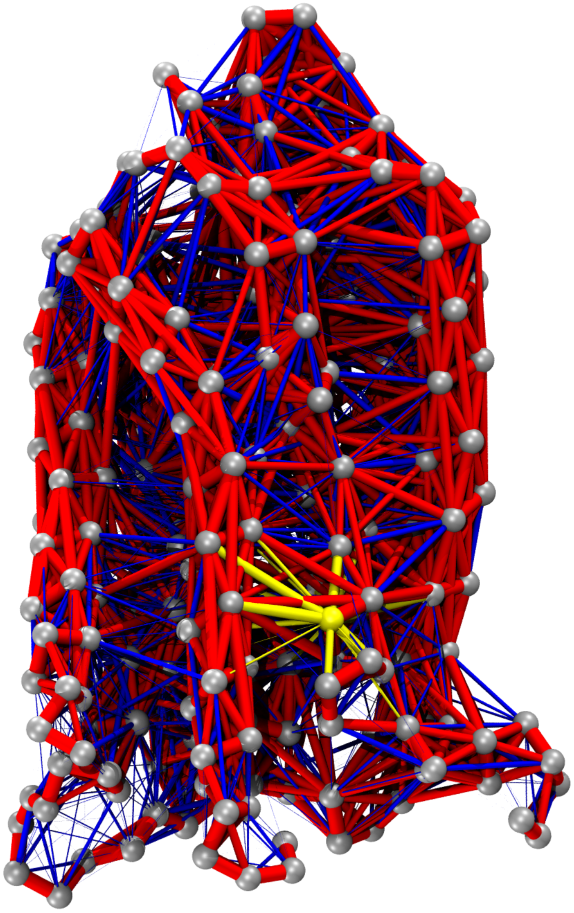


Figure 2.5: Mechanical network of rhodopsin. The full mechanical network for rhodopsin is shown here, connecting all C_α's whose distance is less than < 12Å on average. The strongest links are colored in red, while the weakest ones are in blue. For residue 7x52, the links corresponding to non-covalent bonds are highlighted in yellow.

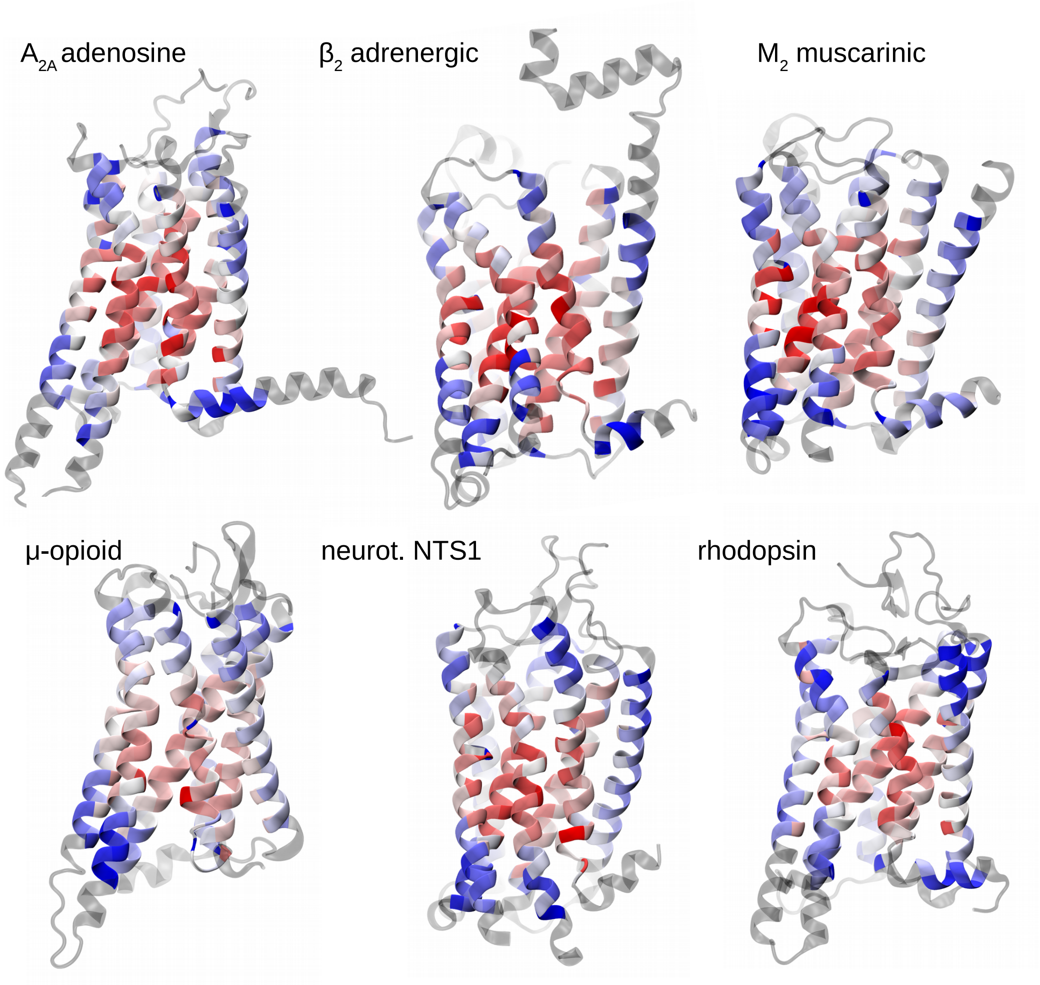


Figure 2.6: Mechanical bridging score for the six GPCRs considered in this work. The score is presented with a decreasing color scale from red to blue, on one of their conformations (PDB IDs: 4UHR, 4QKX, 4MQT, 5C1M, 4XES, 4A4M). The structures are oriented with TM6 and TM7 up front.

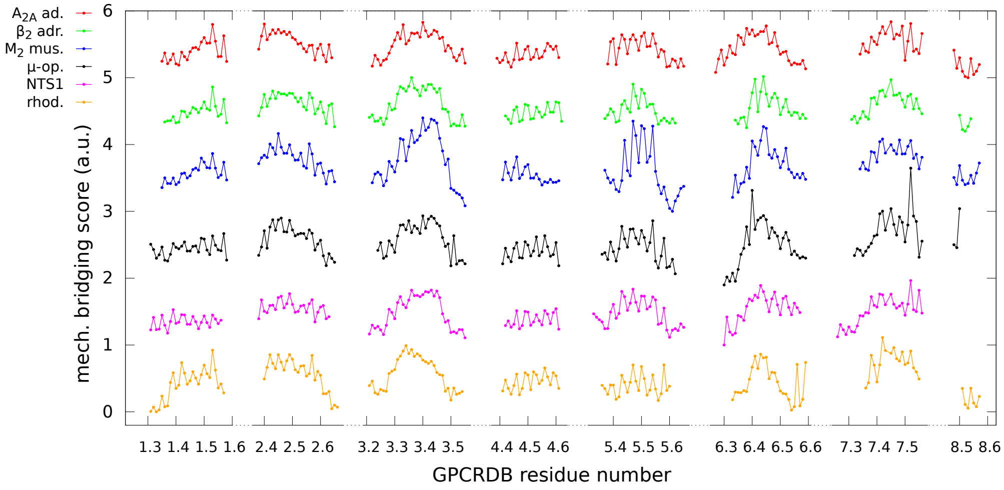


Figure 2.7: Sequence profiles of the mechanical bridging score. The residues are numbered according to the GPCRdb scheme.

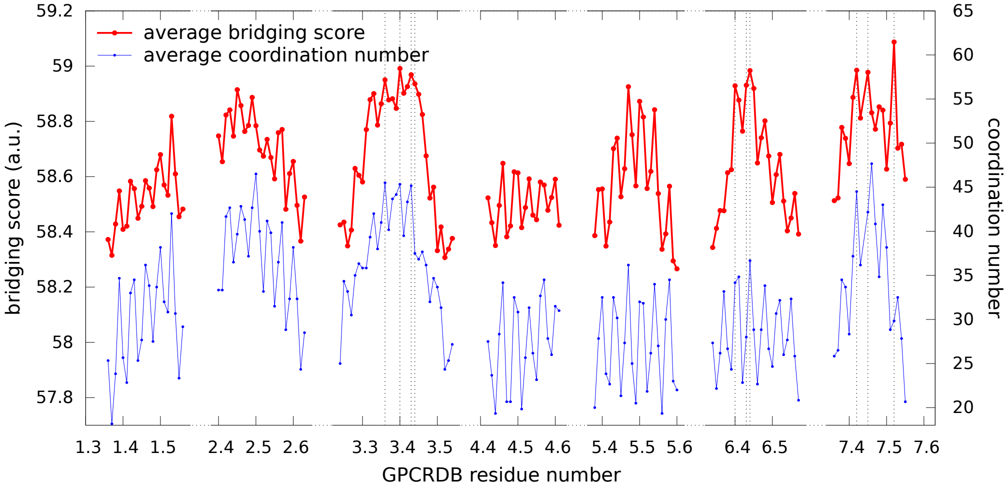


Figure 2.8: Average bridging score (red) and coordination number (blue). The positions of the 10 top ranking residues of the bridging score, as listed in the first column of Table 2.1, are indicated by the dashed vertical lines. The residues are numbered according to the GPCRdb scheme.

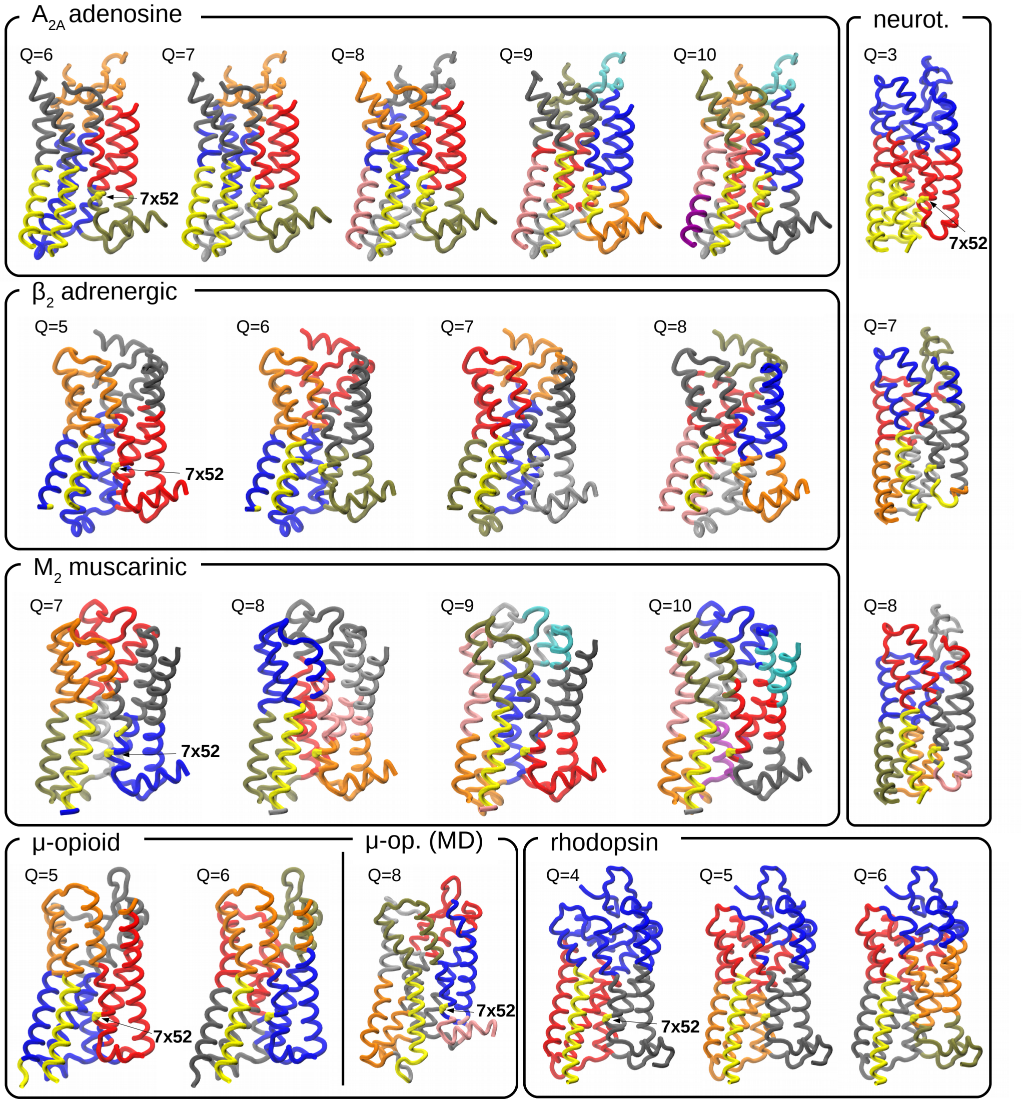


Figure 2.9: Quasi-rigid domain decompositions of receptors. The decompositions into quasi-rigid domains of the six GPCRs were produced by the SPECTRUS webserver (http://spectrus.sissa.it) [1]. Those decompositions including a TM6+7x52 domain have been selected, and this particular domain is highlighted in yellow. For μ-opioid, a subdivision based on conformations from the MD simulation is shown as well.

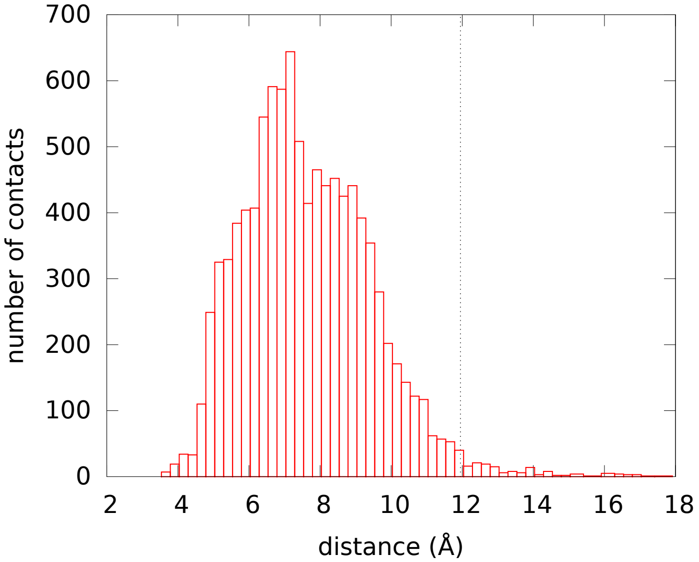


Figure 2.10: Distribution of interhelix C_α-C_α distances. Histogram of the distances, for all six receptors considered in our analysis, between each residue's C_α and its nearest neighbor one, belonging to the closest facing helix. A threshold of 12Å, corresponding to the 98.5th percentile (dashed vertical line), guarantees that the great majority of connections is included, without disrupting contacts between helices.

Chapter 3

From sequence to structure: identifying protein domains by coevolution-based clustering

Recent methodological developments in bioinformatics have led to an unexpected and striking use of the rapidly-growing databases of sequenced genomes, namely the inference of inter-residue contacts from the analysis of large multiple sequence alignments.

Here we elaborate on the possibility of deriving structural insight from sequence-based coevolutionary information by further using the SPECTRUS dimensionality-reduction approach of Chapter 1.

The material presented in this chapter illustrates some preliminary results from a work done in collaboration with Daniele Granata[1]and under the supervision of Cristian Micheletti and Vincenzo Carnevale[1]. A manuscript is currently in preparation.

3.1 Introduction

In recent years, the steady increase in the amount of available protein sequence data and the development of sequence homology detection methods allowed for the construction of very accurate multiple sequence alignments (MSA). A single MSA can contain tens of thousands of non-redundant amino acid sequences, which collectively capture the evolutionary pathway of a

[1] ICMS, Temple University, Philadelphia, USA

protein. As a result, the patterns of amino acid substitutions at analogous sequence positions in a MSA have the potential to unveil the constraints imposed by the protein three-dimensional structure and biological function [106, 107].

Many techniques have been devised to address the problem of inferring structural features from coevolutionary correlation between homologous proteins. The rationale is that mutations of one amino acid are likely accompanied by compensatory mutations of its neighbours. Accordingly, systematic co-mutations of amino-acid pairs ought to correlate with their spatial proximity. Approaches implementing such contact-prediction strategies are referred to as direct coupling analysis (DCA) methods.

Hinging on this concept, we try to go beyond the "pointwise" contact inference strategy and, instead, focus on patterns of multiple, or cooperative coevolutionary relationships. Specifically, we aim at using DCA approaches to detect *coevolutionary domains*, that is groups of residues that presumably underwent similar structural and functional selection mechanisms.

3.2 Methods

An overview of the step-by-step procedure discussed here is sketched in Fig. 3.1.

The core of the procedure consists in interpreting the norm of statistical couplings J_{ij}, obtained by direct coupling analysis (DCA) of the protein MSA, as a measure of evolutionary proximity between residues i and j. Such similarity measure is then processed by a clustering algorithm, which is an adapted version of the SPECTRUS clustering scheme of Chapter 1 [1].

3.2.1 The dataset

For our analysis we used the extensive dataset of $\sim$ 800 MSAs compiled in [108]. These alignments were computed by using the homology detection method HHblits [109], and are characterized by a heterogeneous number of sequences (16–65535) and positions (30–494). The dataset was specifically aimed at testing the contact prediction capability of DCA methods. Accordingly, each MSA in the dataset is associated to a target structure, that is the known conformation of the most representative member in the MSA. These structures will be used here *a posteriori* to examine the structural character-

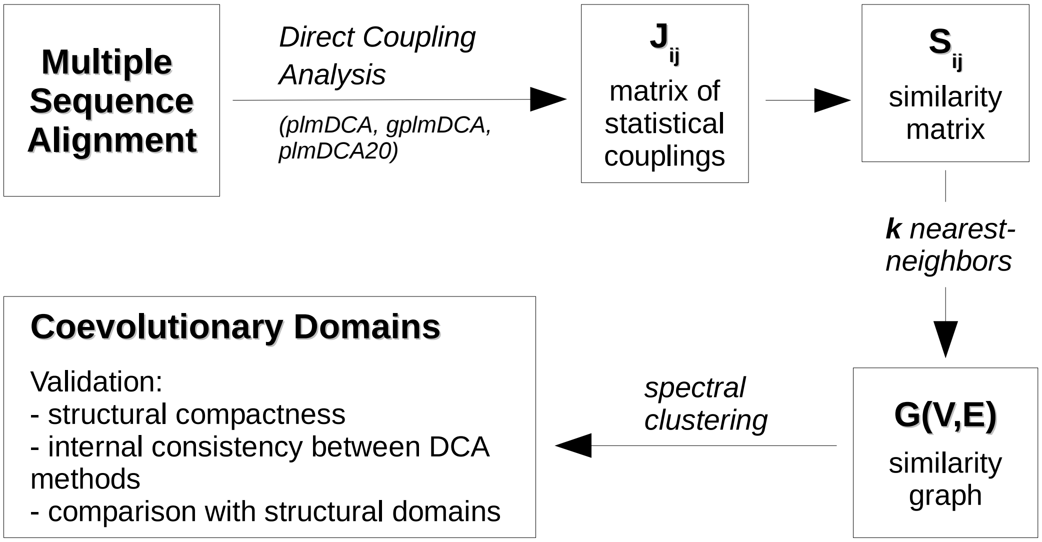


Figure 3.1: Schematic workflow of the steps for identifying the coevolutionary domains from an initial MSA. The procedure is applied for each of the ~ 800 entries in the dataset.

istic of the "coevolutionary domains" that we will infer on the basis of the MSA information only.

3.2.2 Direct coupling analysis

We present here a brief overview, based on [108], of the most recent computational methods implementing the direct coupling analysis (DCA). This term indicates a family of methods to predict contact between amino acid pairs from the analysis of correlated mutations between sequence positions in a MSA.

The common characteristic of these methods is the idea of disentangling direct from indirect couplings between residue positions, by eliminating spurious correlations. By using a global statistical approach, in fact, it is possible to recognize the cases in which a high correlation between two residues is due to the fact that both of them are directly correlated to a third variable.

From a statistical point of view, a MSA can be considered as a collection of M samples $\left\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(M)}\right\}$ from an unknown probability distribution. Each row of the MSA, corresponding to one protein, is then one of the q^N possible realizations of a set of N random variables, where each random variable represents one position in the alignment, that can take $q = 21$ different values (the amino acid type or the gap symbol):

$$\mathbf{a} = (a_i, \dots, a_N), \quad a_i \in 1, \dots, q. \tag{3.1}$$

In its original formulation [110,111], the DCA assumes that the underlying probability distribution generating the samples can be described by the Potts model of statistical physics:

$$\begin{aligned} P_{\text{Potts}}(\mathbf{a}) &= \frac{e^{-H_{\text{Potts}}(\mathbf{a})}}{\mathcal{Z}}, \\ H_{\text{Potts}}(\mathbf{a}) &= -\sum_{i<j} J_{ij}(a_i, a_j) - \sum_i h_i(a_i), \end{aligned} \tag{3.2}$$

where $h_i \in \mathbb{R}^{21}$ and $J_{ij} \in \mathbb{R}^{21\times 21}$ are the couplings to be inferred.

The inference process is based on the maximum likelihood principle, where one selects the optimal probability distribution in a class which minimizes a negative-log-likelihood function

$$L = -\frac{1}{M} \sum_{m=1}^{M} \log P_{\text{Potts}}(\mathbf{a}^{(m)}). \tag{3.3}$$

However, since the partition function $\mathcal{Z}$ in eq. 3.2 cannot be evaluated exactly, the minimization can only be done approximately.

Several approaches have been proposed to tackle this challenge, including the mean-field approximation [112], the sparse inverse covariance methods (PSICOV) [113] and the pseudo-likelihood optimization. In the following, we will focus on the latter approach, which has been shown to outperform the others [114,115].

In [116], Ekeberg *et al.* first adopted the weaker criterion of pseudo-likelihood maximization for solving the inverse Potts model. In this approximation, the probability in eq. 3.3 is replaced by the conditional probability of observing one variable a_r, given the observations of all the other variables $\mathbf{a}_{\backslash r}$, that is

$$\begin{aligned} &P\left(a_r = a_r^{(m)} \,\middle|\, \mathbf{a}_{\backslash r} = \mathbf{a}_{\backslash r}^{(m)}\right) = \\ &\qquad = \frac{\exp\left[-\sum_{i\neq r} J_{ri}\left(a_r^{(m)}, a_i^{(m)}\right) - h_r\left(a_r^{(m)}\right)\right]}{\sum_{l=1}^{q} \exp\left[-\sum_{i\neq r} J_{ri}\left(l, a_i^{(m)}\right) - h_r\left(l\right)\right]}, \end{aligned} \tag{3.4}$$

where, for notational convenience, $J_{ri}(l,k)$ means $J_{ir}(k,l)$ when $i < r$.

Given an MSA, the conditional likelihood is then maximized by minimizing

$$L_r(J_r, h_r) = -\frac{1}{M}\sum_{m=1}^{M} \log P\left(a_r = a_r^{(m)} \,|\, \mathbf{a}_{\backslash r} = \mathbf{a}_{\backslash r}^{(m)}\right). \tag{3.5}$$

This expression now depends only on the parameters $J_r = \{J_{ir}\}_{i \neq r}$ and h_r relative to position r, and by repeating the process for each position r it is possible to determine all parameters.

In this way, however, there could be different predictions for J_{rs} and J_{sr}, that must be reconciled. A possible way consists in taking their average $\frac{1}{2}(J_{rs} + J_{sr}^T)$. Alternatively, one can avoid such inconsistencies by directly minimizing a new objective function, obtained by summing up all contributions:

$$L_{\text{pseudo}} = \sum_{r=1}^{N} L_r(J_r, h_r). \tag{3.6}$$

The minimization of L_{pseudo} does not predict the same parameters as minimizing the true likelihood function 3.3. However, replacing L with L_{pseudo} makes the problem computationally tractable, because it does not require the evaluation of the full partition function $\mathcal{Z}$.

This procedure, described in [116], is named *pseudo-likelihood maximization DCA* (plmDCA), and it has been shown to achieve very accurate predictions, when validated against experimentally determined protein structures.

In a more recent paper, Feinauer *et al.* [108] noted that the outcome from this method can be affected negatively by the presence of long gap-rich portions in the MSA. This issue can be traced back to the choice of the rather simple Potts model of eq. 3.2 for representing the (unknown) probability distribution. In this model, in fact, the gap symbol is considered on the same footing as any other amino acid type. This assumption is therefore quite distant from reality, since a real MSA, which usually includes sequences with long stretches of a same variable (the gap symbol), cannot be regarded as a set of independent realizations drawn randomly from such probability distribution.

They therefore suggested to use a refined version for the inference model, in which an additional set of parameters is introduced in the original Potts model 3.2 in order to deal with the possible presence of gaps:

$$P_{\text{Gap-Potts}}(\mathbf{a}) = \frac{e^{-H_{\text{Potts}}(\mathbf{a})-H_{\text{Gap}}}}{\mathcal{Z}}. \tag{3.7}$$

The new term H_{Gap}, which describes the propensity of each position to be the beginning of a gap of variable length, introduces a maximum of NL additional parameters to be learned, where L is the largest gap length found in a given alignment. This is a relatively small increment in the computational burden, compared to the previous number of parameters in 3.2, which is about $\frac{1}{2}q^2N^2$. On the other hand, the modified inference model, which is referred to as *gap-enhanced pseudo-likelihood maximization DCA* (gplmDCA) has been shown to improve the overall contact prediction [108].

The outcome of learning a model like 3.2 or 3.7 is a set of pairwise interaction coefficients J_{ij} for each pair of positions (i, j). Each coefficient, in turn, consists in a 21×21 matrix. In order to quantify the strength of a coupling between two positions, one can then compute the norm of the relative coefficient matrix:

$$c_{ij} = ||J_{ij}||^2. \tag{3.8}$$

Here, we will use the Frobenius norm augmented by the average-product correction (APC), as introduced in [116], which mitigates the bias due to insufficient sampling.

By introducing a small modification in the latter step, it is also possible to derive an alternative method of handling the gaps. The procedure consists in simply ignoring the gap terms when computing the couplings c_{ij}, by calculating the Frobenius norm on the 20 sub-matrix not involving the gap variables. Such method is named *plmDCA20* [108], and its contact prediction performance has been found to be slightly better than the one of gplmDCA.

In the following, we will use the couplings c_{ij}, computed from all three inference methods, plmDCA, gplmDCA and plmDCA20, as the inter-residue similarity measure to be processed by the clustering algorithm.

3.2.3 Spectral clustering

The clustering strategy adopted in this context is the same as the one described in the previous chapters of this thesis, and in particular for SPECTRUS, namely the spectral clustering [66, 117].

This algorithm takes in input the pairwise similarities S_{ij} between the elements to be clustered and returns a set of subdivisions into a variable number of clusters $Q = Q_{\min}, \ldots, Q_{\max}$. For our purpose, the similarities between the residues are given by the DCA couplings c_{ij} defined in eq. 3.8.

The quality score, introduced with SPECTRUS [1], will be also used for the analysis of single examples. We recall that the purpose of this score is to help identifying the optimal number of clusters Q_{best}, among the ones spanned by the clustering algorithm.

3.2.4 *k*-nearest neighbor graph and clustering coefficient

In the context of spectral clustering, the pairwise similarities S_{ij} between the elements to be clustered are used for building a graph representation $G(V, E)$, where $V = \{v_i\}$ is a set of vertices, corresponding to the original elements, and $E = \{e_{ij}\}$ the edges connecting them. The goal of such representation is to map the initial similarities into local neighborhood relationships between the graph vertices [117].

This is usually done by either defining a cutoff value for the similarities or choosing a maximum number of edges connecting each vertex to its nearest-neighbors. The SPECTRUS algorithm, for instance, was based on a similarity graph built by connecting residues with an edge only if their reciprocal distance on the structure was less than 10 Å. In the present work, however, the graph vertices represent sequence positions, and, as such, it is not possible to introduce a spatial cutoff. We therefore consider a k-nearest neighbor graph, obtained by connecting each vertex to the top k most similar neighbors, with an edge weighted according to the relative similarity ($e_{ij} = S_{ij}$).

The final graph must be symmetric, therefore an edge between v_i and v_j is drawn if either e_{ij} or e_{ji} is one of the k strongest edges. Moreover, in order to be sure to always work on a completely connected graph, we also force the inclusion of edges between consecutive residues on the sequence. As a result, the effective average number of nearest-neighbors in the final graphs is always larger than the initial (nominal) k (see Fig. 3.2a for examples from the dataset).

The parameter k controls the sparsity of the connections in the graph: a high value of k corresponds to a greater overall graph connectivity and is more suitable for identifying larger clusters, while a smaller one might be used for more granular decompositions.

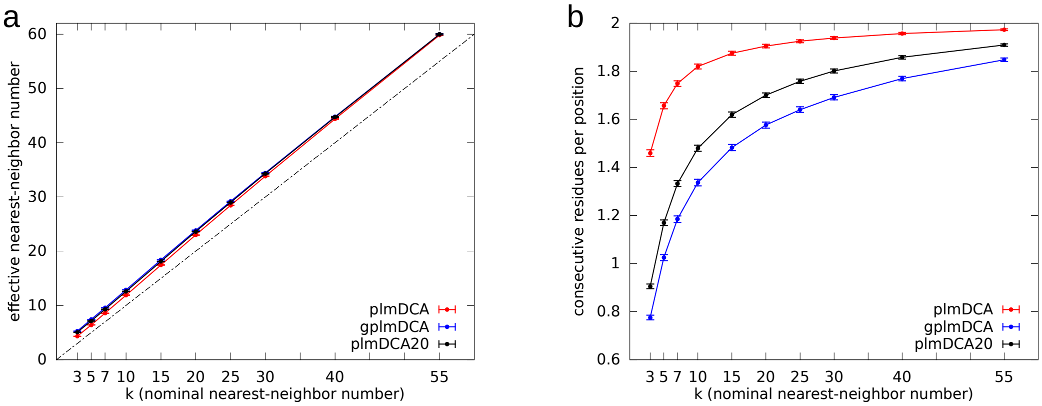


Figure 3.2: (a) Average effective number of nearest-neighbors in the graphs obtained from the three DCA approaches, as a function of the nominal number k. (b) Average number of sequence consecutive residues per position included in the set of top k residues with strongest couplings. The red line shows that plmDCA tends to assign stronger couplings to the pairs of consecutive residues on the sequence. In both figures, error bars = s.e.m.

To choose k, we use an internal criterion based on the clustering coefficient C [118], also known as "cliqueness". This quantity measures the average probability that two neighbors of a vertex be also connected between themselves. More precisely, given a vertex v_i with n_i neighbours, the local clustering coefficient is computed as

$$C_i = \frac{t_i}{n_i(n_i-1)/2}, \tag{3.9}$$

where t_i is the number of links between the neighbors of v_i. The global clustering coefficient is then defined as the average of the local coefficients of vertices with more than 1 neighbor:

$$C = \langle C_i \rangle_{n_i>1}. \tag{3.10}$$

A high clustering coefficient means that the probability for two vertices to be connected by an edge is higher if they have a mutual neighbor, and for real-world networks it typically ranges from a few percent to about 50% or more [119].

For comparison, it is informative to derive the clustering coefficient for random graphs, also known as Erdös-Rényi graphs. In these models, the probability p of edge formation is defined *a priori*, and it is independent for each vertex, so that the clustering coefficient is simply equal to p. Given the number N of vertices, the average number of neighbours (or average degree) $\bar{k}$ is also easily computed as the total number of edges, $N(N-1)/p$, divided by N, *i.e.* $\bar{k} = (N-1)p$. It then follows that:

$$C_{\text{rand}} = p = \frac{\bar{k}}{(N-1)}. \tag{3.11}$$

The difference of the clustering coefficient with respect to the random case,

$$\Delta C = C - C_{\text{rand}}, \tag{3.12}$$

provides then a good measure for the intrinsic clustering “propensity” of a graph.

In the Results section, we will use this criterion to guide our choice of the parameter k.

3.2.5 Cluster compactness

Once the clusters of the graph $G(V, E)$ are obtained, we will assess their spatial compactness on the protein structure with a suitable order introduced hereafter.

In case of a compact cluster q of n_q vertices, each vertex can be connected with any other one in the same cluster by at least one path joining only vertices within the cluster itself. When a single vertex v_i is completely disconnected from the relative cluster, the number of intra-cluster “broken paths” b_q is equal to $n_q - 1$ (counting symmetric paths once). We then say that there are $d_q = b_q/(n_q - 1)$ disconnected vertices, which in the latter case amount to 1. We note that this definition distinguishes between a “fuzzy” cluster, where the disconnected vertices are scattered and do not group together, and a cluster split in multiple sub-clusters. In particular, the case in which half of the cluster vertices are scattered over the graph in isolated positions has a worse count ($b_q = (n_q/2)(n_q - 1)$, $d_q = n_q/2$) than the case where a cluster is split in two compact sub-clusters ($b_q = (n_q/2)^2$, $d_q \simeq n_q/4$).

We, then, finally compute an intensive quantity, called compactness, defined as:

$$\Omega = 1 - \frac{1}{N} \sum_q d_q, \tag{3.13}$$

where N is the total number of vertices.

3.2.6 Adjusted mutual information

When discussing the results obtained by our method, we will need a quantitative way to compare two cluster partitionings in order to assess their similarity. In the context of information theory, one of the way is to estimate the amount of shared information between the two partitionings, *i.e.* by computing their mutual information (MI).

The information content of a partitioning P of N elements into Q clusters is quantified by the entropy:

$$H(P) = -\sum_{q=1}^{Q} \frac{n_q}{N} \log \frac{n_q}{N}, \tag{3.14}$$

where n_q is the number of elements in cluster q.

Given a second partitioning P', the average amount of information needed to describe it, once the first partitioning P is already known, is given by the conditional entropy

$$H(P'|P) = -\sum_{q'}^{Q'} \sum_{q}^{Q} \frac{n_{q'q}}{N} \log \frac{n_{q'q}}{n_q}, \tag{3.15}$$

where $n_{q'q}$ counts the elements belonging to both cluster q' and q.

The mutual information is then defined as the difference between the entropy associated with P' and the conditional entropy $H(P'|P)$:

$$\mathrm{MI}(P', P) = H(P') - H(P'|P). \tag{3.16}$$

For our purpose, we will adopt the related concept of adjusted mutual information (AMI). This quantity is defined as:

$$\mathrm{AMI} = \frac{\mathrm{MI} - \langle \mathrm{MI} \rangle}{\max(\mathrm{MI}) - \langle \mathrm{MI} \rangle}, \tag{3.17}$$

where $\langle \mathrm{MI} \rangle$ is the expected value of MI over pairs of random partitions, for which an analytical expression has been derived [120].

The AMI has two desirable properties: (i) it is normalized, *i.e.* two identical partitions return an AMI equal to 1, and (ii) it is adjusted-for-chance, namely two random clusterings produce on average an AMI equal to 0.

3.3 Results and discussion

We present and discuss here the main findings from the application of the scheme detailed in Fig. 3.1 to the dataset of ~ 800 MSAs. The final goal is to test the feasibility of the decomposition strategy for obtaining domains of evolutionarily-related residues and hence assess its usefulness for studying the sequence-structure relationship. Throughout this analysis, we will also present a comparison between the three DCA methods (plmDCA, gplmDCA, plmDCA20) briefly recalled in the Methods section.

We start by describing the analysis performed on the matrix of statistical couplings between residues, as obtained from the DCA, in order to prepare the ground for the clustering procedure. In fact the latter, for optimal discriminatory performance, is best applied not to the matrix of couplings, but to a sparser graph obtained by retaining only the strongest couplings for each amino acid. This ensures that the partitioning will capture the strongest signals of coevolutionary relationships. The number of strongest couplings to retain, k, is a parameter that will itself be optimised.

In the subsequent sections, we will then present a dataset-wide survey of some properties of the domain decompositions obtained from coevolutionary information. A few notable cases will be discussed in detail as well.

3.3.1 Clustering propensity of coevolutionary signals

As detailed in the Methods section, the input of the clustering algorithm is a similarity graph, $G(V, E)$, whose vertices V are the sequence sites (*i.e.* the residues in the consensus sequence) and the pairwise edges E are the k strongest coevolutionary couplings for each site.

The optimal value of the connectivity parameter k is set by generating an ensemble of graphs G_k for a few values of k, ranging from 3 to 55, and for each MSA in the database. We then evaluate their corresponding clustering coefficients, adjusted with respect to a random reference, according to the definition of eq. 3.12. We recall that the clustering coefficient has been introduced in the context of graph theory to measure the "clustering propensity" of a graph, that is the tendency of vertices in a graph to aggregate locally.

Fig. 3.3a shows the histogram of the parameters k giving a maximum value of the adjusted clustering coefficient $\Delta C = C - C_{\text{rand}}$. In most cases, the maximum occurs for relatively small values of k, *i.e.* about 7. The same properties hold for the average value of ΔC as well (see Fig. 3.3b). Based

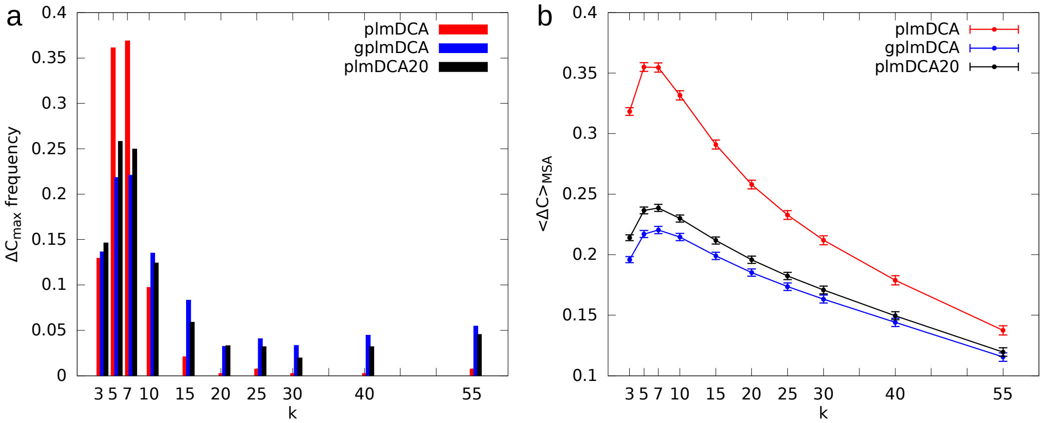


Figure 3.3: (a) Histogram of the parameter k for which the highest adjusted clustering coefficient ΔC_{max} is observed, over the MSA dataset. (b) Average adjusted clustering coefficient for different nearest-neighbor numbers k (error bars = s.e.m.).

on these typical, general features of the graphs, we set k to be equal to 7 and, for definiteness, we will keep this parameter fixed for each entry in the dataset.

By comparing the curves in Fig. 3.3 one also observes that, among the three considered DCA methods, the higher clustering propensity is given by plmDCA. It is possible that this compliance to clustering results from the fact that plmDCA takes into higher account the influence of peptide chain connectivity on coevolutionary mutations (see Fig. 3.2b), which may favour the formation of stronger couplings between mutual neighbors of the same residue.

In the following, we will demonstrate a correlation between the adjusted clustering coefficient ΔC computed on a similarity graph and some desirable properties of the associated clusters of residues, like their compactness on the structure.

3.3.2 Structural compactness of coevolutionary domains

Following the scheme of Fig. 3.1 we next use the spectral clustering scheme to partition the graph in groups of residues that are strongly related from the evolutionary point of view. As for the SPECTRUS approach, the number of partitions, Q, is first varied in the 2–10 range and then the optimal value of Q is set for each MSA based on the quality score profile.

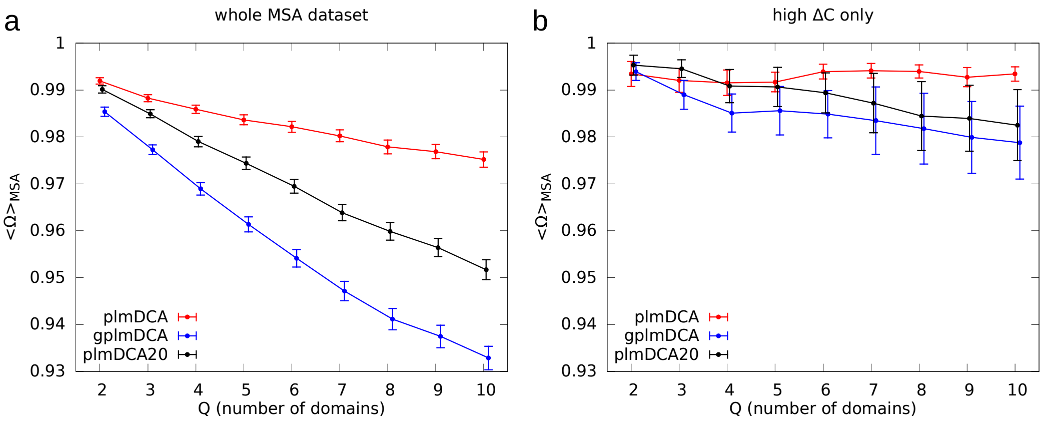


Figure 3.4: (a) Domain compactness, averaged over the MSA dataset, at fixed number of domains, for the three inference methods. (b) Average compactness computed over the top 40 MSAs which feature the highest adjusted clustering coefficient ΔC. In both figures, the error bars represent the standard error of the mean.

One of the key elements of our approach is assessing the degree to which the coevolutionary domains are compact in space. This is a particularly relevant question because, we recall, the degree of coevolutionary relationship is inferred from the sole sequence alignment, with no reference to the proteins' actual three-dimensional structure.

As a first step in this endeavour, we compute the degree of domain compactness Ω (see Methods section) averaged over all MSAs ($\langle\Omega\rangle_{\mathrm{MSA}}$), for each value of Q.

The results are shown in Fig. 3.4a and clarify that, indeed, the clusters are typically compact structural domains. In fact, according to the definition of compactness in eq. 3.13, a value of Ω larger than 0.9 means that less than 10% of residues do not belong to spatially compact domains. Of course, this significant overall compactness tends to decrease as the number of clusters increases.

Again, we observe that plmDCA performs better than gplmDCA and plmDCA20, yielding more compact domains.

For each single MSA, we then consider the degree of domain compactness, averaged over all number of considered clusters Q as well ($\langle\Omega\rangle_Q$). This is done to obtain a comprehensive view that is robust and not tailored to specific domain numbers.

By inspecting Figs. 3.5, one observes that low values of the Q-averaged

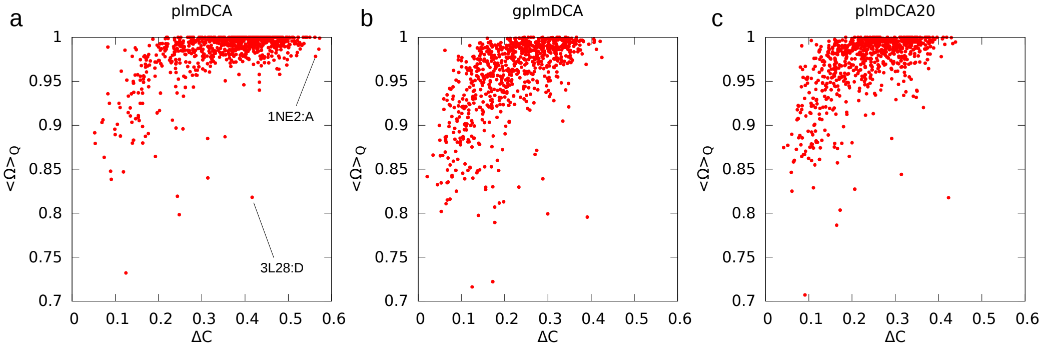


Figure 3.5: Scatter plots of the domain compactness, averaged over the subdivisions into $Q = 2, \ldots, 10$ domains of a single MSA, *vs* the adjusted clustering coefficient, for the three inference methods.

compactness are typically observed for MSAs with low adjusted clustering coefficient. The converse is also true: if we restrict considerations to the top 40 cases with highest ΔC, corresponding to $\sim 5\%$ of dataset, their compactness degree is typically never below 0.97 for any of the three DCA methods, see Fig. 3.4b.

To better illustrate the implications of the results in Figs. 3.5 we shall discuss a set of specific plmDCA decompositions, namely those marked in Fig. 3.5a.

As a first example, we describe the case of Thermoplasma Acidophilum 1320 (PDB ID: 1NE2), whose MSA includes more than 65,000 sequences (corresponding to 14,000 homology reduced sequences, with a maximum of 90% sequence identity) and which features a relatively high ΔC.

Its coevolutionary domain decomposition is typical of most plmDCA decompositions, as almost 78% of MSAs have a comparable or higher compactness score. The quality score computed on the spectral clustering results gives a clear indication of the optimal number of domains for this protein, located at $Q = 7$, see Fig. 3.6a. The relative coevolutionary domains, shown on the protein structure in Fig. 3.6b, are extremely well-defined from a structural point of view, comprising uninterrupted stretches of secondary elements as α-helices or β-sheets. The compactness is therefore very high ($\Omega \simeq 0.99$).

For completeness, we additionally provide in Fig. 3.6c the subdivision of the same protein into $Q = 4$ coevolutionary domains. This subdivision has a compactness of 0.94, which is the lowest in the $Q = 2, \ldots, 10$ range, though clearly still being very good. The individual domains are again compact in

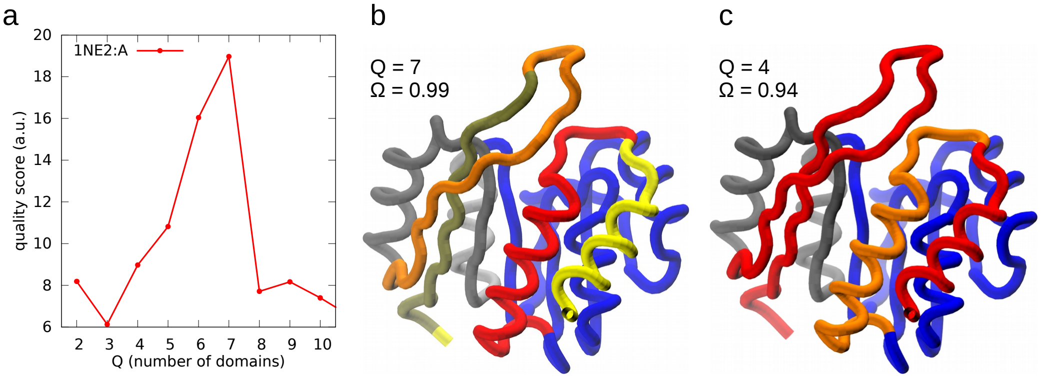


Figure 3.6: Coevolutionary domain decomposition of Thermoplasma Acidophilum 1320 (PDB ID: 1NE2), based on plmDCA. (a) Quality score from the spectral clustering. (b) Decomposition into $Q = 7$ domains, corresponding to the maximum in the quality score. (c) Decomposition into $Q = 4$ domains. The red domain is the "gap domain", which includes nearly all amino acid positions with a gap content higher than 87 %, localized at the two termini (see also Fig. 3.7a)

both structure and sequence, with only one domain, the one comprising the termini, that appears split in two. This particular domain assignment is found in a significant fraction of decompositions in the MSA dataset, and is related to the presence of a high percentage of gaps in those sequence sites (see Fig. 3.7a).

It is known, in fact, in the context of DCA, that the couplings between gap-rich positions tend to be overestimated. This issue has been addressed, for instance, in the case of plmDCA, by the introduction of the two improved versions gplmDCA and plmDCA20, as discussed in the Methods section. While the latter approaches help reducing the bias induced by gaps on the contact prediction [108], we do not have a definitive answer, at the moment, to whether they mitigate this phenomenon of a "gap domain" assignment. Similar subdivisions as the one in Fig. 3.6c from plmDCA are, for example, obtained from gplmDCA and plmDCA20 as well. On the other hand, it is noteworthy that such domain assignment is clearly penalized by the quality score in Fig. 3.6a.

We conclude by illustrating an atypical case, specifically the Zaire Ebola viral protein 35 (PDB ID: 3L28), which is an outlier in the scatter plot of Fig. 3.5a. This entry has an atypically low Q-averaged compactness, despite having a good clustering propensity. It is relevant to note that this entry has the lowest number of sequences in the dataset (9.15 homology reduced

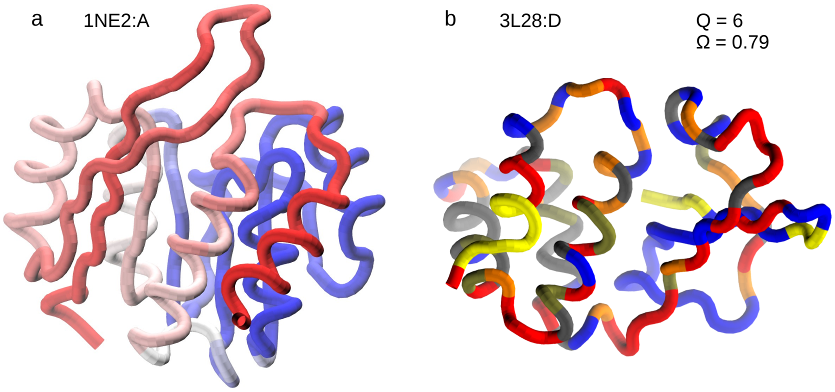


Figure 3.7: (a) Percentage of gaps per position in the aligned sequences of 1NE2:A, from 0% (blue) to almost 100 % (red), reported on the protein structure. (b) Coevolutionary domain decomposition of Zaire Ebola viral protein 35 into $Q = 6$ domains (plmDCA), with compactness $\Omega = 0.79$. We note that the choice of always retaining the couplings between sequence nearest-neighbors, while useful to ensure a global connectivity of the coupling matrix, does not prevent the occurrence of this kind of fragmented clusters.

sequences), which clearly makes it more noisy and less robust. Fig. 3.7b shows its decomposition into 6 coevolutionary domains, corresponding to a compactness close to its average value of 0.8.

The analysis of structural compactness, then, gives us confidence about the method's viability. The coevolutionary domains identified by the decomposition strategy are, indeed, generally compact on the structure. The emergence of cases with an unusually low degree of domain compactness is clearly linked to problems with the input data statistics, that can be kept under control by looking, for instance, at the clustering coefficient of the similarity graph or at the number of sequences of the initial MSA.

3.3.3 Comparison of DCA methods

To gain further insight on the domain analysis, we look at the internal consistency between the subdivisions obtained from the three inference methods (plmDCA, gplmDCA and plmDCA20). Such consistency measure is carried out by calculating the Adjusted Mutual Information (AMI) between the respective partitions P and P' at the same number of domains Q, and by taking

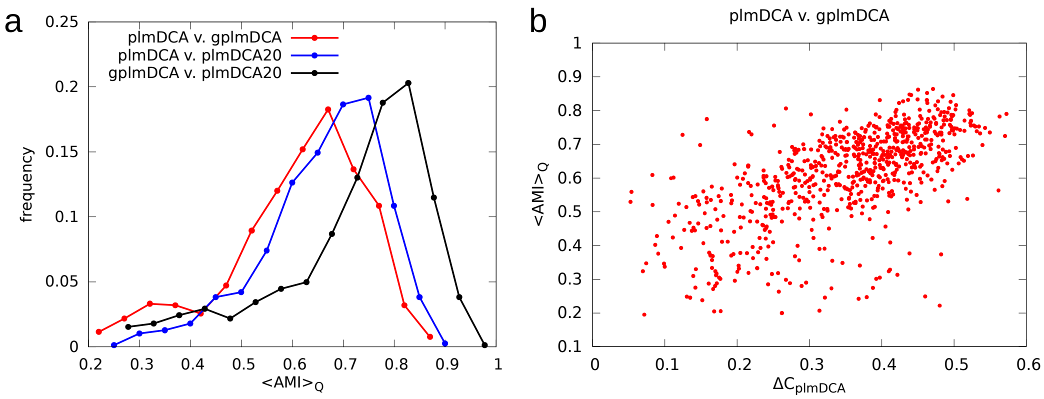


Figure 3.8: (a) Histograms of the Q-averaged Adjusted Mutual Information (AMI) between the partitions into coevolutionary domains derived from the three inference methods. (b) Plot of the Q-averaged AMI between plmDCA and gplmDCA, as a function of the adjusted clustering coefficient ΔC, computed for plmDCA. Analogous plots for the other method comparisons show a similar behavior.

the average:

$$\langle \mathrm{AMI} \rangle_Q = \frac{1}{9} \sum_{Q=2}^{10} \mathrm{AMI}(P(Q), P'(Q)). \tag{3.18}$$

The AMI, described in detail in the Methods section, quantifies in a rigorous way the overlap between two subdivisions, producing a score normalized within the range $[0, 1]$, where 0 is the value expected by chance agreement between two random partitions, and 1 the value associated with identical partitions. Since the shorter MSA is only 30 positions long, the average is computed on decompositions into up to a maximum of 10 domains.

In principle, the quality score derived from the spectral clustering, previously employed for instance in the analysis of 1NE2, should allow to pick out the optimal number of domains Q_{best} for a decomposition. Here, we decide, however, to consider the AMI averaged over the number of domains because this makes more intuitive the comparison between decompositions with different Q_{best}.

As shown by the histograms in Fig. 3.8a, a general agreement between the three methods emerges quite clearly, with typical values of $\langle \mathrm{AMI} \rangle_Q$ within the range $[0.65, 0.85]$ (see the example in Fig. 3.9 for an intuitive term of reference for the parameter range).

In particular, gplmDCA and plmDCA20, the two improved versions of

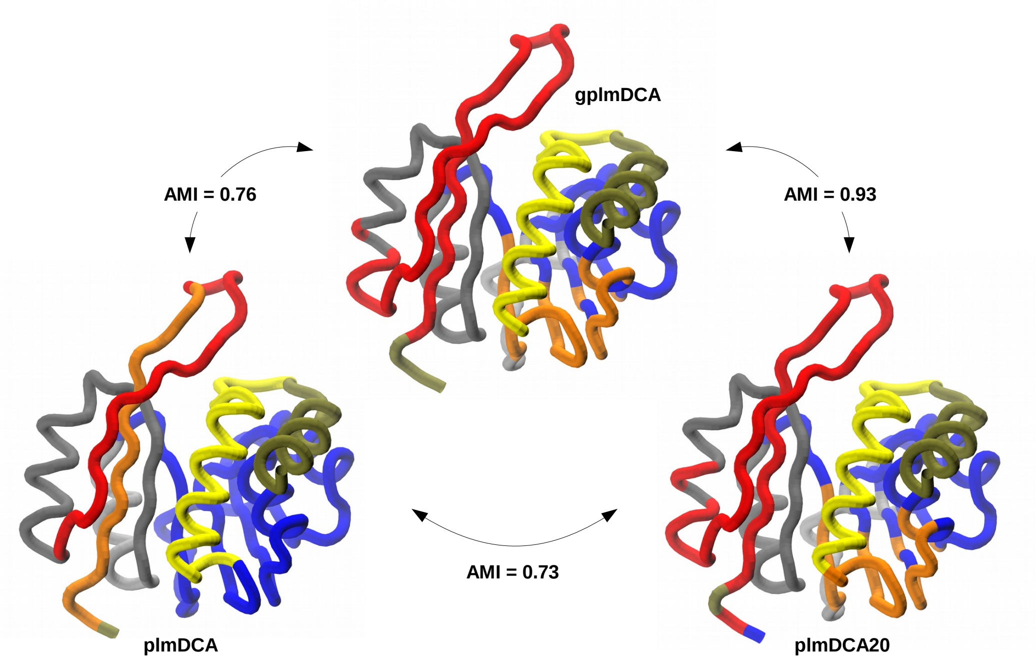


Figure 3.9: Comparison between the decompositions of Thermoplasma Acidophilum 1320 (PDB ID: 1NE2) into 7 coevolutionary domains, obtained from the three DCA methods. The AMI between the three partitionings is shown as well.

the original plmDCA show a quite large AMI between them, larger than the respective comparison with plmDCA. This suggests that the modifications introduced by these two methods for handling gap-rich sequence segments in MSAs makes them distinctly different from plmDCA and yet similar among themselves.

Fig. 3.8b shows, as an example, the adjusted clustering coefficient computed on plmDCA, plotted against the AMI between plmDCA and gplmDCA partitionings. The correlation observed here reaffirms what has been previously noted from the compactness analysis, *i.e.* that the clustering propensity measured initially on the matrix of statistical couplings is a good indicator of the robustness of the final results, here measured from the comparison between DCA methods.

3.3.4 Connection with structural domains

The accuracy of contact predictions, in the context of DCA methods validation, is customarily evaluated by computing the fraction of contacts correctly identified on the true contact map, for cases where the relative three-dimensional structure is known [108]. Here, for the assessment of a protein coevolutionary domain decompositions, we adopt the approach of directly comparing our findings with the structural domains, obtained by clustering the C_α-C_α distance matrix (within a cutoff of 10 Å, as done in SPECTRUS with the distance fluctuation matrix). This structural clustering, similarly to the sequence-based clustering, with which it shares the same algorithmic engine, produces a hierarchy of subdivisions into a variable number of spatially compact domains.

The similarity between the sequence- and structure-based partitionings is measured by $\langle \mathrm{AMI} \rangle_Q$, as introduced before in eq. 3.18. In Fig. 3.10a, we then compare the $\langle \mathrm{AMI} \rangle_Q$ for the three inference methods. From the histograms in the figure, one can see that the performances of plmDCA, gplmDCA and plmDCA20 are comparable, an observation in accord with the consistency test of the previous section. Hereafter, we will then mainly refer to the results relative to plmDCA.

Fig. 3.10b demonstrates once again the usefulness of measuring the adjusted clustering coefficient associated with an MSA in order to rule out those cases that are presumably affected by low statistics, and which are more prone to give inconsistent partitionings.

As an example, we detail in Fig. 3.11 the case of the NSE1 domain pro-

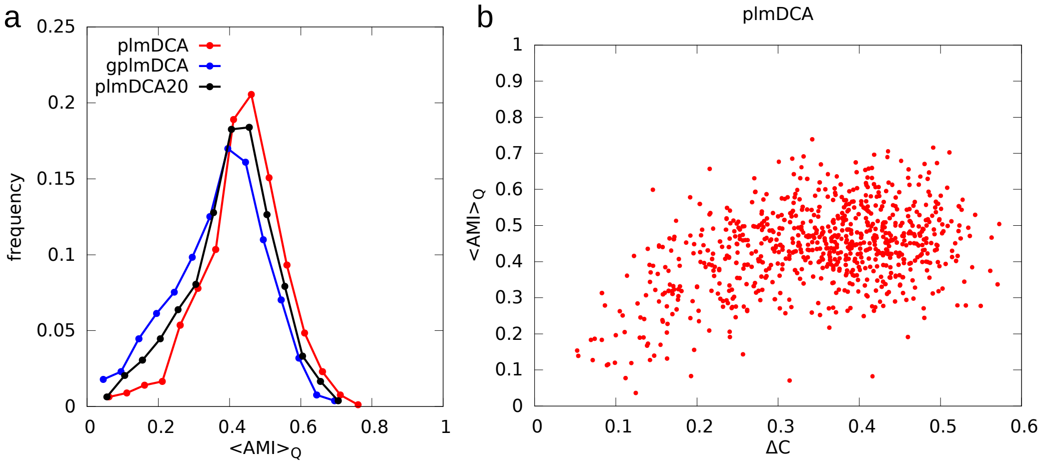


Figure 3.10: (a) Histograms of the Q-averaged AMI between the partitions into coevolutionary domains, derived from the three inference methods, and the corresponding structural ones. (b) Plot of $\langle \text{AMI} \rangle_Q$ for each MSA *vs.* the relative adjusted clustering coefficient ΔC, based on plmDCA.

tein (PDB ID: 3NW0:A), which features the highest clustering propensity for plmDCA ($\Delta C = 0.57$). This is a typical example of a sequence-based decomposition with a good overall overlap with the corresponding structural partitioning. The highest value of AMI (0.64), restricted to a few domains, is obtained for $Q = 3$, see Fig. 3.11a. From the structural comparison, we can see how the coevolutionary domains correctly predict the disposition of residues into the three subdomains.

Without having any prior knowledge of the protein structure, one can be guided in the choice of an optimal number of coevolutionary domains by the quality score, introduced with SPECTRUS. In the case of NSE1 domain protein, the quality score computed for the sequence-based decomposition, shown in Fig. 3.12, has a first, significant maximum at $Q = 4$. This is different from the optimal Q found for the structural decomposition, which is for 3 domains. We note that such discrepancy, although legitimate, given the different protein aspects the two decompositions are based on, might affect the estimate of the overlap between them. A more accurate calculation of the AMI may then be obtained from a comparison at different number of domains. Fig. 3.12d shows, for example, that the AMI between the optimal sequence- and structure-based decompositions, into 4 and 3 domains respectively, is larger than the one computed at the same number of domains,

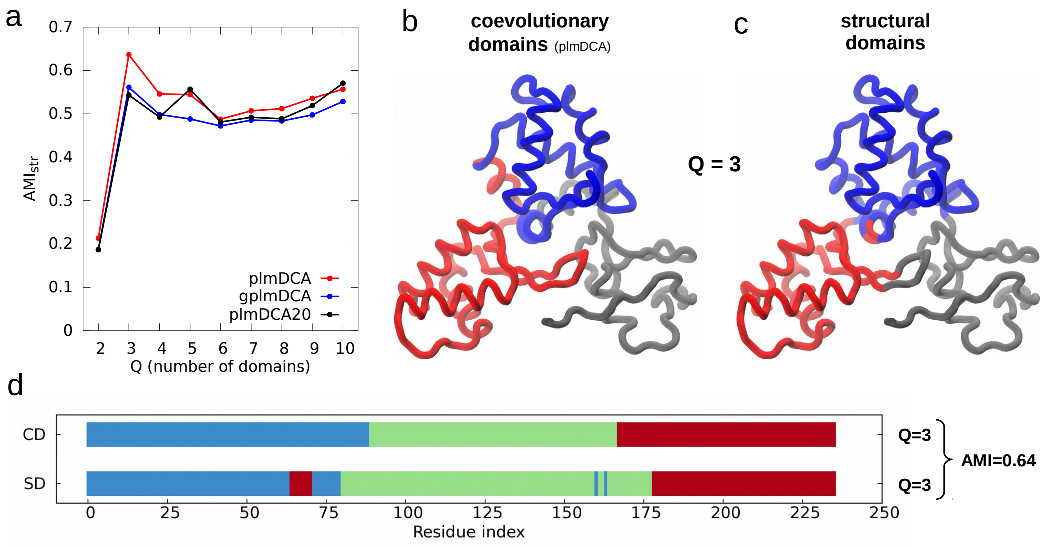


Figure 3.11: Analysis of NSE1 domain protein (PDB ID: 3NW0:A). (a) The AMI with the relative structural decomposition has a first maximum for $Q = 3$ domains, for all three inference methods. Panels (b) and (c) show on the structure the relative sequence- and structure-based decompositions (for plmDCA). In (d), the corresponding partitionings are shown on the sequence (with different colors with respect to panels (b) and (c)).

$Q = 4$. The overlap between optimal decompositions, however, can only be computed by inspecting individually each single case. For the dataset-wide analysis presented here, the unsupervised Q-averaged AMI is better suited.

In the Appendix, we illustrate a few selected examples of decompositions, chosen in the dataset among the cases with highest clustering propensity ΔC and a sufficiently large number of sequences in the MSA. A characteristic observed in most of these examples is that the coevolutionary domains are generally compact not only in structure, but also in sequence, that is each domain is formed by a single uninterrupted amino acid stretch. However, we note that this is not always the case: a notable example is *E. coli* SbmC protein (PDB ID: 1JYH:A), whose decomposition for $Q = 4$ shown in Fig. 3.13 features structurally compact domains formed by several sequence segments connected by tertiary contacts. The agreement with the corresponding structural decomposition into 4 domains, that is one of the optimal partitioning according to the structural quality score, is remarkable (AMI $= 0.82$).

We can then conclude that the coevolutionary domains decomposition is a very promising tool for identifying the residues that are more likely to form

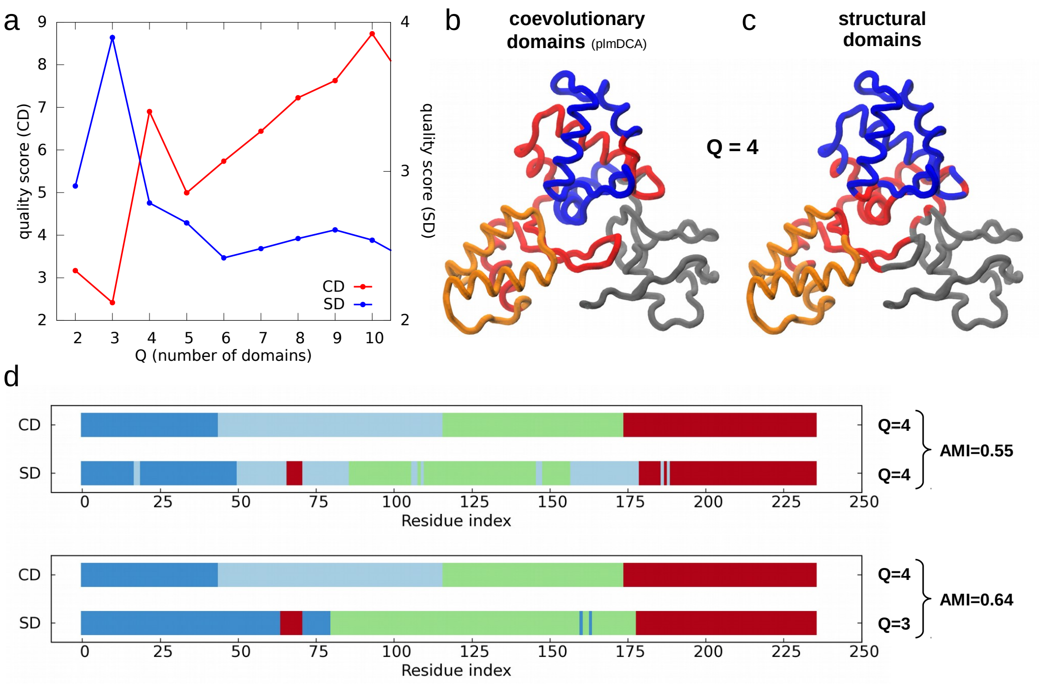


Figure 3.12: (a) Quality scores of the coevolutionary (CD) and structural (SD) domain decompositions of NSE1 domain protein (PDB ID: 3NW0:A). Panels (b), (c) and (d) show a few notable domain decompositions, both on the structure and on the sequence. The color legend is not preserved between the two representations.

spatially compact clusters in the folded structure.

3.4 Conclusions

In this chapter, we have summarized some preliminary results from the application of the coevolutionary domain decomposition strategy.

The method has been tested on an extensive dataset of almost 800 multiple sequence alignments. A particular attention has been put in recognizing and controlling the effect of insufficient sampling in MSAs, that can be detected *a priori* by measuring intrinsic properties of the coevolutionary inter-residue coupling matrix.

Measures of internal consistency as well as a comparison with experimental structural information suggest that the method is indeed capable of providing robust and meaningful results.

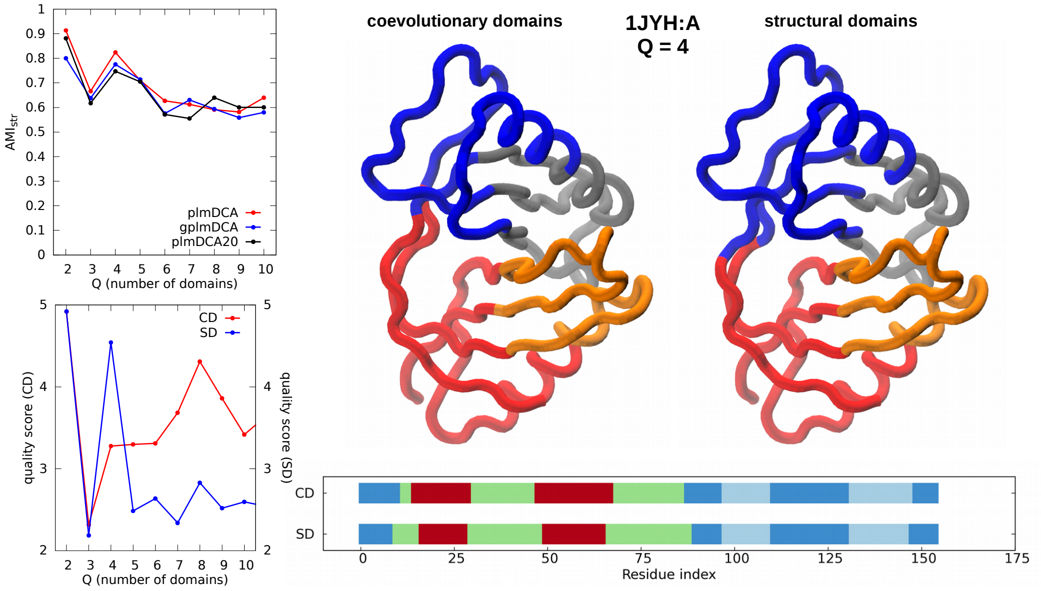


Figure 3.13: Coevolutionary (CD) and structural (SD) domain decompositions of *E. coli* SbmC protein (PDB ID: 1JYH:A).

This tool may be therefore useful to interpret and integrate the results from well established techniques of contact prediction from coevolutionary information.

3.5 Appendix: additional figures

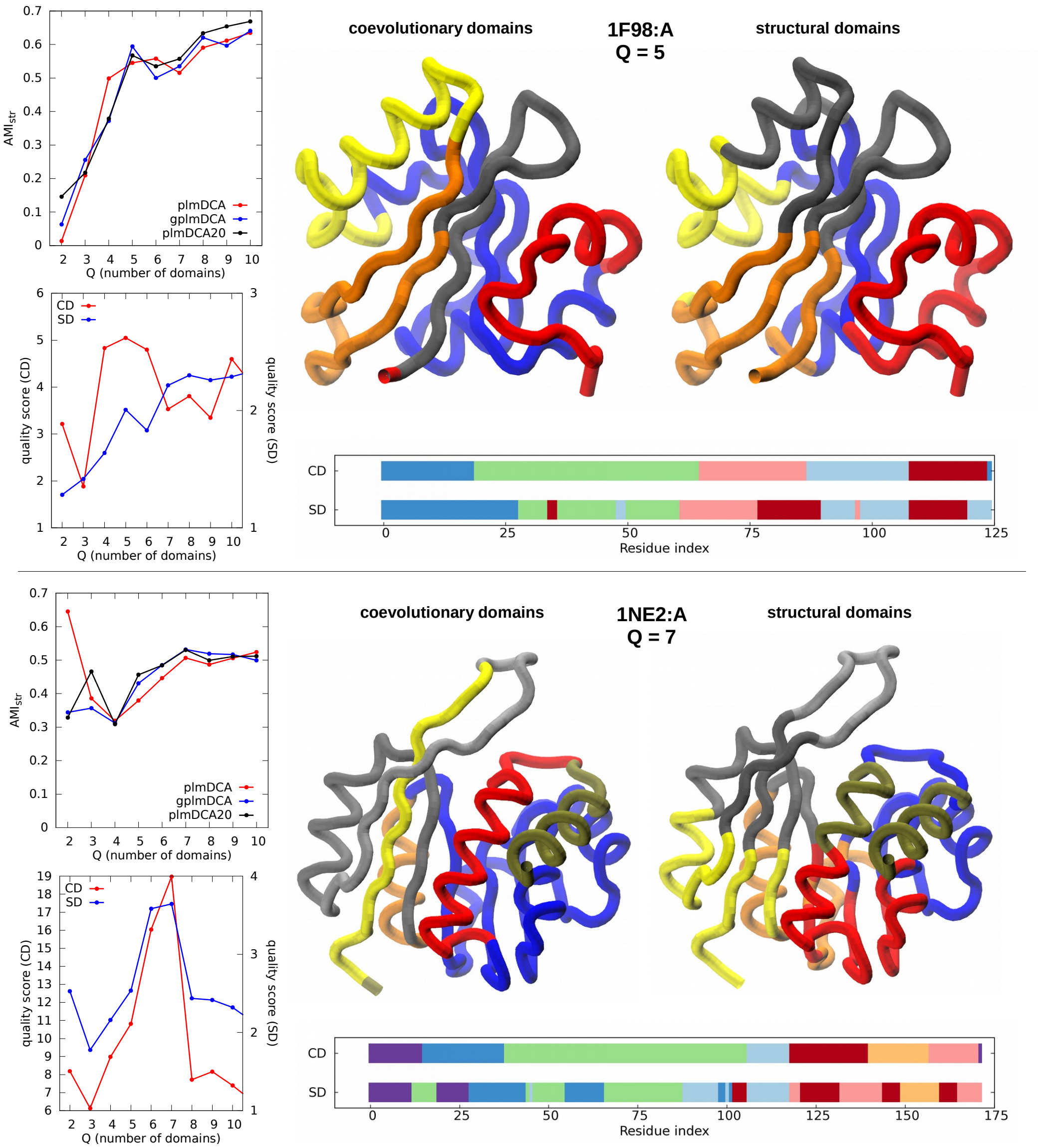


Figure 3.14: Typical examples of coevolutionary domain decompositions, compared with the corresponding structural domains. The number of domains selected for the representation on the structure is one of the maxima in the sequence-based quality score. Different colors are used for the corresponding representation on the sequence.

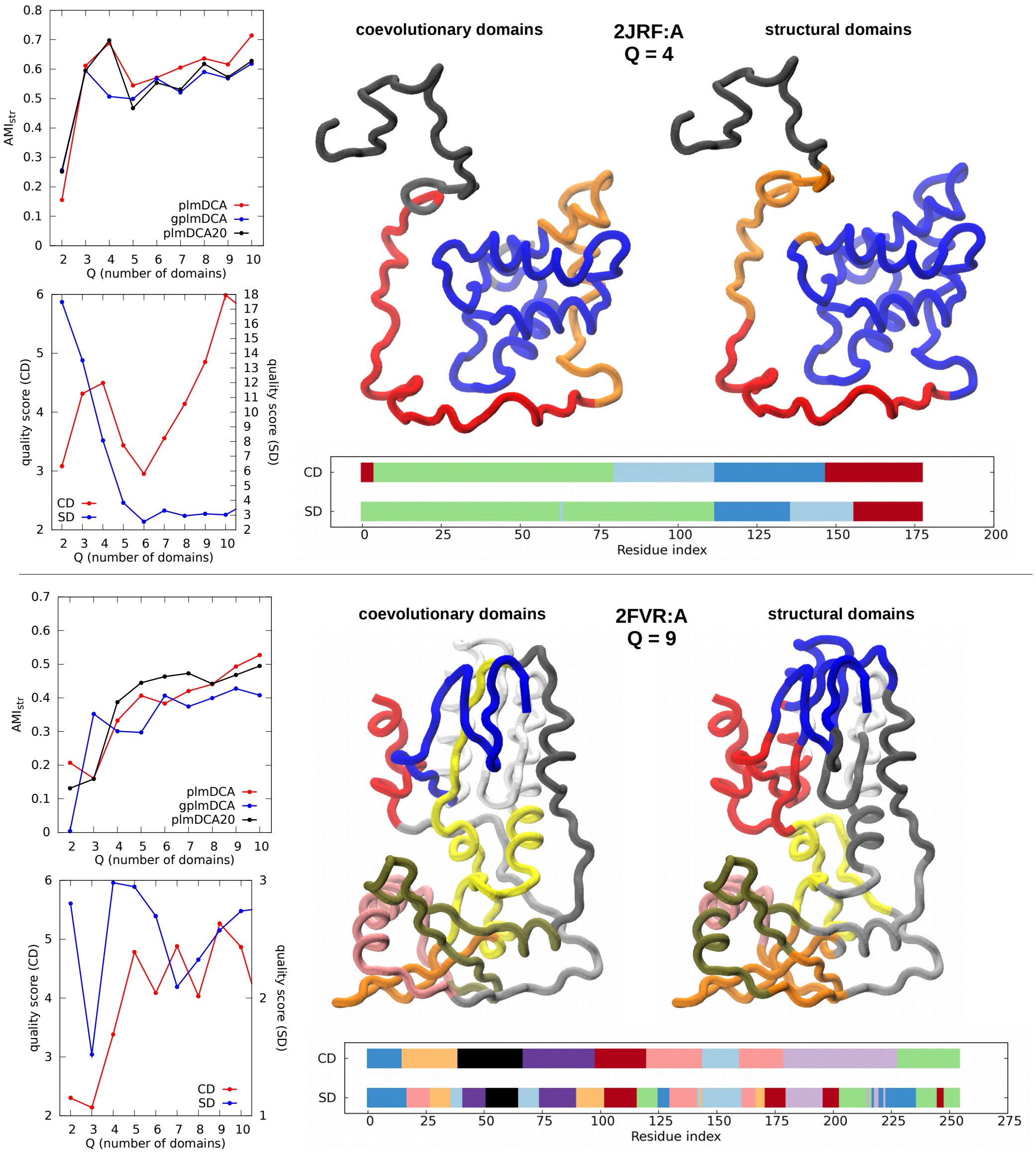

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
AMI$_{str}$
plmDCA
gplmDCA
plmDCA20
2 3 4 5 6 7 8 9 10
Q (number of domains)
quality score (CD)
quality score (SD)
6 5 4 3 2
18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2
CD
SD
coevolutionary domains
2JRF:A
Q = 4
structural domains
CD
SD
0 25 50 75 100 125 150 175 200
Residue index
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
AMI$_{str}$
plmDCA
gplmDCA
plmDCA20
2 3 4 5 6 7 8 9 10
Q (number of domains)
6 5 4 3 2
3 2 1
quality score (CD)
quality score (SD)
CD
SD
coevolutionary domains
2FVR:A
Q = 9
structural domains
CD
SD
0 25 50 75 100 125 150 175 200 225 250 275
Residue index

Figure 3.15

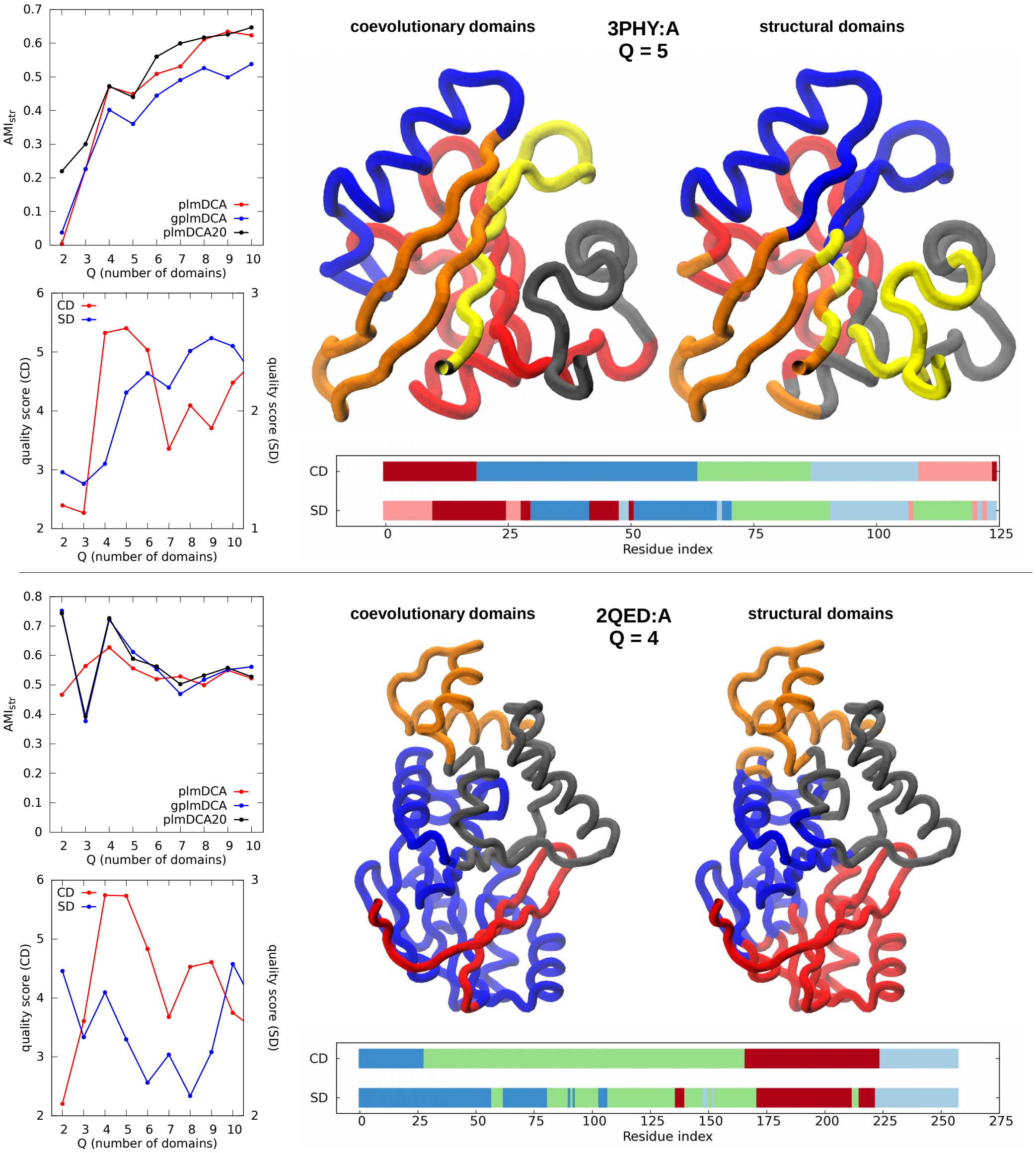

coevolutionary domains
3PHY:A
Q = 5
structural domains
AMI$_{str}$
Q (number of domains)
plmDCA
gplmDCA
plmDCA20
CD
SD
quality score (CD)
quality score (SD)
Residue index
0 25 50 75 100 125
coevolutionary domains
2QED:A
Q = 4
structural domains
0 25 50 75 100 125 150 175 200 225 250 275

Figure 3.16

PDB ID	N	N_{seq}	$N_{\text{eff.seq}}$	ΔC_{plm}	ΔC_{gplm}	ΔC_{plm20}
1F98:A	125	65147	23111	0.52	0.42	0.44
1NE2:A	172	65535	14080	0.56	0.39	0.41
2FVR:A	255	52373	9025	0.56	0.30	0.38
2JRF:A	178	51412	13794	0.53	0.35	0.37
2QED:A	258	65496	18268	0.49	0.34	0.36
3PHY:A	125	65182	23212	0.52	0.28	0.43

Table 3.1: Number of residues N, number of (effective) sequences $N_{\text{(eff.)seq}}$ in the MSA and adjusted clustering coefficients ΔC for the examples shown in previous figures.

Bibliography

[1] Luca Ponzoni, Guido Polles, Vincenzo Carnevale, and Cristian Micheletti. SPECTRUS: A dimensionality reduction approach for identifying dynamical domains in protein complexes from limited structural datasets. *Structure*, 23(8):1516–1525, 2015.

[2] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

[3] A. Ramanathan and P. K. Agarwal. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. *PLoS Biol*, 9, 2011.

[4] Arvind Ramanathan, Andrej Savol, Virginia Burger, Chakra S Chennubhotla, and Pratul K Agarwal. Protein conformational populations and functionally relevant substates. *Accounts of chemical research*, 47:149–156, 2014.

[5] S. Piana, P. Carloni, and U. Rothlisberger. Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. *Prot. Sci.*, 11:2393–2402, 2002.

[6] C. Micheletti. Comparing proteins by their internal dynamics: Exploring structure–function relationships beyond static structural alignments. *Physics of life reviews*, 10:1–26, 2013.

[7] V. C. Nashine, S. Hammes-Schiffer, and S. J. Benkovic. Coupled motions in enzyme catalysis. *Curr Opin Chem Biol*, 14:644–651, 2010.

[8] A. del Sol, C. J. Tsai, B. Ma, and R. Nussinov. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure*, 17:1042–1050, 2009.

[9] Y. Liu and I. Bahar. Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, 29:2253–2263, 2012.

[10] P. K. Agarwal. Role of protein dynamics in reaction rate enhancement by enzymes. *J. Am. Chem. Soc.*, 127:15248–15256, 2005.

[11] H. Li, S. Sakuraba, A. Chandrasekaran, and L. W. Yang. Molecular Binding Sites Are Located Near the Interface of Intrinsic Dynamics Domains (IDDs). *J. Chem. Inf. Model.*, 54:2275–2285, 2014.

[12] C. Chennubhotla and I. Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput. Biol.*, 3:1716–1726, 2007.

[13] S. Sacquin-Mora, O. Delalande, and M. Baaden. Functional modes and residue flexibility control the anisotropic response of guanylate kinase to mechanical stress. *Biophys. J.*, 99:3412–3419, 2010.

[14] G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, 332:234–238, 2011.

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

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

[17] I. Bahar and R. L. Jernigan. Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, 38:3478–3490, 1999.

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

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

[20] J. J. Falke. Enzymology: A moving story. *Science*, 295:1480 – 1481, 2002.

[21] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA*, 100:6980–6985, 2003.

[22] F. Pontiggia, A. Zen, and C. Micheletti. Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. *Biophys. J.*, 95:5901–5912, 2008.

[23] S. Hayward, A. Kitao, and H. J. C. Berendsen. Model-free methods of analyzing domain motions in proteins from simulation. *Proteins*, 27:425–437, 1997.

[24] W. Wriggers and K. Schulten. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins*, 29:1–14, 1997.

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

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

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

[28] D. A. Snyder and G. T. Montelione. Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins*, 59:673–686, 2005.

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

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

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

[32] J. Painter and E. A. Merritt. Tlsmd web server for the generation of multi-group tls models. *Journal of Applied Crystallography*, 39:109–111, 2006.

[33] J. Camps, O. Carrillo, A. Emperador, L. Orellana, A. Hospital, M. Rueda, D. Cicin-Sain, M. D'Abramo, J. L. Gelpí, and M. Orozco. Flexserv: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, 25:1709–1710, 2009.

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

[35] T. Aleksiev, R. Potestio, F. Pontiggia, S. Cozzini, and C. Micheletti. PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains. *Bioinformatics*, 25:2743–2744, 2009.

[36] S. Bernhard and F. Noé. Optimal identification of semi-rigid domains in macromolecules from molecular dynamics simulation. *PloS one*, 5:e10491, 2010.

[37] D. K. Kirchner and P. Güntert. Objective identification of residue ranges for the superposition of protein structures. *BMC bioinformatics*, 12:170, 2011.

[38] G. Morra, R. Potestio, C. Micheletti, and G. Colombo. Corresponding functional dynamics across the hsp90 chaperone family: insights from a multiscale analysis of md simulations. *PLoS Comput Biol*, 8:e1002433, 2012.

[39] J. Romanowska, K. S. Nowinski, and J. Trylska. Determining geometrically stable domains in molecular conformation sets. *Journal of Chemical Theory and Computation*, 8:2588–2599, 2012.

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

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

[42] H. Vashisth, G. Skiniotis, and C. L. Brooks. Collective variable approaches for single molecule flexible fitting and enhanced sampling. *Chemical Reviews*, 114:3353–3365, 2014.

[43] M. Cascella, C. Micheletti, U. Rothlisberger, and P. Carloni. Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J. Am. Chem. Soc.*, 127:3734–3742, 2005.

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

[45] O. Keskin, R. L. Jernigan, and I. Bahar. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.*, 78:2093–2106, 2000.

[46] M. Münz, R. Lyngso, J. Hein, and P. C. Biggin. Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. *BMC Bioinformatics*, 11:188–188, 2010.

[47] E. Fuglebakk, S. P. Tiwari, and N. Reuter. Comparing the intrinsic dynamics of multiple protein structures using elastic network models. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2014.

[48] J. Zhang, P. Minary, and M. Levitt. Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images. *Proc. Natl. Acad. Sci. USA*, 109:9845–9850, 2012.

[49] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J. Carazo. Disentangling conformational states of macromolecules in 3d-EM through likelihood optimization. *Nat Meth*, 4:27–29, 2007.

[50] Florence Tama and Charles L Brooks. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *Journal of molecular biology*, 318(3):733–747, 2002.

[51] A. Arkhipov, P. L. Freddolino, and K. Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 14:1767–77, 2006.

[52] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437–449, 2006.

[53] M. Cieplak and M. O. Robbins. Nanoindentation of virus capsids in a molecular model. *The Journal of Chemical Physics*, 132:015101, 2010.

[54] W. H. Roos, M. M. Gibbons, A. Arkhipov, C. Uetrecht, N. R. Watts, P. T. Wingfield, A. C. Steven, A. J. R. Heck, K. Schulten, W. S. Klug, and G. J. L. Wuite. Squeezing Protein Shells. *Biophys. J.*, 99:1175–1181, 2010.

[55] J. Snijder, C. Uetrecht, R. Rose, R. Sanchez-Eugenia, G. Marti, J. Agirre, D. Guérin, G. Wuite, A. Heck, and W. Roos. Probing the biophysical interplay between a viral genome and its capsid. *Nat. Chem.*, 5:502–509, 2013.

[56] G. Polles, G. Indelicato, R. Potestio, P. Cermelli, R. Twarock, and C. Micheletti. Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition. *PLoS Computational Biology*, 9:e1003331, 2013.

[57] M. Cieplak and M. O. Robbins. Nanoindentation of 35 Virus Capsids in a Molecular Model: Relating Mechanical Properties to Structure. *PLoS ONE*, 8:e63640, 2013.

[58] S. Menor, A. M. De Graff, and M. Thorpe. Hierarchical plasticity from pair distance fluctuations. *Physical biology*, 6:036017, 2009.

[59] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 911–916. IEEE, 2010.

[60] L. Sauguet, F. Poitevin, S. Murail, C. Van Renterghem, G. Moraga-Cid, L. Malherbe, A. W. Thompson, P. Koehl, P. Corringer, and

M. Baaden. Structural basis for ion permeation mechanism in pentameric ligand-gated ion channels. *The EMBO journal*, 32:728–741, 2013.

[61] L. Sauguet, A. Shahsavar, F. Poitevin, C. Huon, A. Menny, À. Nemecz, A. Haouz, J. Changeux, P. Corringer, and M. Delarue. Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation. *Proc. Natl. Acad. Sci. USA*, 111:966–971, 2014.

[62] C. Amaral, V. Carnevale, M. L. Klein, and W. Treptow. Exploring conformational states of the bacterial voltage-gated sodium channel navab via molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 109:21336–21341, 2012.

[63] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[64] 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, 2009.

[65] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[66] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[67] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65:031910, 2002.

[68] A. Bakan and I. Bahar. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *PNAS*, 106:14349–14354, 2009.

[69] Nicolas Bocquet, Lia Prado de Carvalho, Jean Cartaud, Jacques Neyton, Chantal Le Poupon, Antoine Taly, Thomas Grutter, Jean-Pierre Changeux, and Pierre-Jean Corringer. A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. *Nature*, 445:116–119, 2006.

[70] J. Payandeh, T. Scheuer, N. Zheng, and W. A. Catterall. The crystal structure of a voltage-gated sodium channel. *Nature*, 475:353–358, 2011.

[71] M. Ø. Jensen, V. Jogini, D. W. Borhani, A. E. Leffler, R. O. Dror, and D. E. Shaw. Mechanism of voltage gating in potassium channels. *Science*, 336:229–233, 2012.

[72] Lucie Delemotte, Mounir Tarek, Michael L. Klein, Cristiano Amaral, and Werner Treptow. Intermediate states of the kv1.2 voltage sensor from atomistic molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 108:6109–6114, 2011.

[73] E. Vargas, F. Bezanilla, and B. Roux. In search of a consensus model of the resting state of a voltage-sensing domain. *Neuron*, 72:713–720, 2011.

[74] U. Henrion, J. Renhorn, S. I. Börjesson, E. M. Nelson, C. S. Schwaiger, P. Bjelkmar, B. Wallner, E. Lindahl, and F. Elinder. Tracking a complete voltage-sensor cycle with metal-ion bridges. *Proc. Natl. Acad. Sci. USA*, 109:8552–8557, 2012.

[75] A. J. Rader, D. H. Vlad, and I. Bahar. Maturation dynamics of bacteriophage HK97 capsid. *Structure*, 13:413 – 421, 2005.

[76] C. Chennubhotla, A. J. Rader, L. Yang, and I. Bahar. Elastic network models for understanding biomolecular machinery. *Physical Biology*, 2:S173, 2005.

[77] Wouter H Roos, Ilya Gertsman, Eric R May, Charles L Brooks, John E Johnson, and Gijs JL Wuite. Mechanics of bacteriophage maturation. *Proc. Natl. Acad. Sci. USA*, 109:2342–2347, 2012.

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

[79] E. Fuglebakk, N. Reuter, and K. Hinsen. Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *J. Chem. Theory Comput.*, 9:5618–5628, 2013.

[80] D. S. D. Larsson, L. Liljas, and D. van der Spoel. Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Comput. Biol.*, 8:e1002502, 2012.

[81] WH Roos, R Bruinsma, and GJL Wuite. Physical virology. *Nature Phys.*, 6:733–743, 2010.

[82] Robert Fredriksson, Malin C Lagerström, Lars-Gustav Lundin, and Helgi B Schiöth. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular pharmacology*, 63(6):1256–1272, 2003.

[83] Pierangelo Geppetti, Nicholas A Veldhuis, TinaMarie Lieu, and Nigel W Bunnett. G protein-coupled receptors: Dynamic machines for signaling pain and itch. *Neuron*, 88(4):635–649, 2015.

[84] Sabine Schlyer and Richard Horuk. I want a new drug: G-protein-coupled receptors in drug development. *Drug discovery today*, 11(11):481–493, 2006.

[85] Brian K Kobilka. G protein coupled receptor structure and activation. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1768(4):794–807, 2007.

[86] James AR Dalton, Isaias Lans, and Jesús Giraldo. Quantifying conformational changes in GPCRs: glimpse of a common functional mechanism. *BMC bioinformatics*, 16(1):1, 2015.

[87] Rie Nygaard, Thomas M Frimurer, Birgitte Holst, Mette M Rosenkilde, and Thue W Schwartz. Ligand binding and micro-switches in 7tm receptor structures. *Trends in pharmacological sciences*, 30(5):249–259, 2009.

[88] B Trzaskowski, D Latek, S Yuan, U Ghoshdastider, A Debinski, and S Filipek. Action of molecular switches in GPCRs–theoretical and experimental studies. *Current medicinal chemistry*, 19(8):1090–1109, 2012.

[89] John A Salon, David T Lodowski, and Krzysztof Palczewski. The significance of G protein-coupled receptor crystallography for drug discovery. *Pharmacological reviews*, 63(4):901–937, 2011.

[90] James Gumbart, Fatemeh Khalili-Araghi, Marcos Sotomayor, and Benoît Roux. Constant electric field simulations of the membrane potential illustrated with simple systems. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1818(2):294–302, 2012.

[91] Anita M Preininger, Jens Meiler, and Heidi E Hamm. Conformational flexibility and structural dynamics in GPCR-mediated G protein activation: a perspective. *Journal of molecular biology*, 425(13):2288–2298, 2013.

[92] F Tama and Y-H Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein engineering*, 14(1):1–6, 2001.

[93] Cristian Micheletti, Paolo Carloni, and Amos Maritan. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins: Structure, Function, and Bioinformatics*, 55(3):635–645, 2004.

[94] Samuel Sheftel, Kathryn E Muratore, Michael Black, and Stefano Costanzi. Graph analysis of β_2 adrenergic receptor structures: a "social network" of GPCR residues. *In silico pharmacology*, 1(1):1–15, 2013.

[95] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, 2(1), 2006.

[96] Vignir Isberg, Chris de Graaf, Andrea Bortolato, Vadim Cherezov, Vsevolod Katritch, Fiona H Marshall, Stefan Mordalski, Jean-Philippe Pin, Raymond C Stevens, Gerrit Vriend, et al. Generic GPCR residue numbers–aligning topology maps while minding the gaps. *Trends in pharmacological sciences*, 36(1):22–31, 2015.

[97] Benjamin G Tehan, Andrea Bortolato, Frank E Blaney, Malcolm P Weir, and Jonathan S Mason. Unifying family A GPCR theories of activation. *Pharmacology & therapeutics*, 143(1):51–60, 2014.

[98] Cedric Govaerts, Supriya Srinivasan, Astrid Shapiro, Sumei Zhang, Franck Picard, Karine Clement, Cecile Lubrano-Berthelier, and Chris-

tian Vaisse. Obesity-associated mutations in the melanocortin 4 receptor provide novel insights into its function. *Peptides*, 26(10):1909–1919, 2005.

[99] Ya-Xiong Tao. Constitutive activation of G protein-coupled receptors and diseases: insights into mechanisms of activation and therapeutics. *Pharmacology & therapeutics*, 120(2):129–148, 2008.

[100] Vaclav Cvicek, William A Goddard III, and Ravinder Abrol. Structure-based sequence alignment of the transmembrane domains of all human GPCRs: Phylogenetic, structural and functional implications. *PLoS Comput Biol*, 12(3):e1004805, 2016.

[101] A. J. Venkatakrishnan, Xavier Deupi, Guillaume Lebon, Franziska M. Heydenreich, Tilman Flock, Tamara Miljus, Santhanam Balaji, Michel Bouvier, Dmitry B. Veprintsev, Christopher G. Tate, Gebhard F. X. Schertler, and M. Madan Babu. Diverse activation pathways in class A GPCRs converge near the G protein-coupling region. *Nature*, advance online publication, Aug 2016. Letter.

[102] Xiaojing Cong, Pablo Campomanes, Achim Kless, Inga Schapitz, Markus Wagener, Thomas Koch, and Paolo Carloni. Structural determinants for the binding of morphinan agonists to the μ-opioid receptor. *PloS one*, 10(8):e0135998, 2015.

[103] Søren GF Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski, et al. Crystal structure of the β_2 adrenergic receptor-gs protein complex. *Nature*, 477(7366):549–555, 2011.

[104] Aashish Manglik, Andrew C Kruse, Tong Sun Kobilka, Foon Sun Thian, Jesper M Mathiesen, Roger K Sunahara, Leonardo Pardo, William I Weis, Brian K Kobilka, and Sébastien Granier. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature*, 485(7398):321–326, 2012.

[105] Weijiao Huang, Aashish Manglik, AJ Venkatakrishnan, Toon Laeremans, Evan N Feinberg, Adrian L Sanborn, Hideaki E Kato, Kathryn E

Livingston, Thor S Thorsen, Ralf C Kling, et al. Structural insights into μ-opioid receptor activation. *Nature*, 524(7565):315–321, 2015.

[106] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[107] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[108] Christoph Feinauer, Marcin J. Skwark, Andrea Pagnani, and Erik Aurell. Improving Contact Prediction along Three Dimensions. *PLoS Comput Biol*, 10(10):e1003847, oct 2014.

[109] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

[110] Alan S Lapedes, Bertrand G Giraud, LonChang Liu, and Gary D Stormo. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*, pages 236–256, 1999.

[111] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[112] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[113] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[114] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341–356, 2014.

[115] Marcin J Skwark, Abbi Abdel-Rehim, and Arne Elofsson. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, 29(14):1815–1816, 2013.

[116] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.

[117] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, aug 2007.

[118] DJ Watts and SH Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–2, Jun 1998.

[119] Mark E. J. Newman. Random graphs as models of networks. In *From the Genome to the Internet*, pages 35–68. Wiley-Blackwell, dec 2004.

[120] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

Acknowledgements

I would like to express my deepest and sincere gratitude to my supervisor, Cristian Micheletti, for supporting me during my PhD and for giving me invaluable advice. His example in the working environment and outside inspired me in all these years and will guide me in the future.

I would also like to thank the people I collaborated with: Vincenzo Carnevale, Paolo Carloni, Giulia Rossetti and Alejandro Giorgetti. A special mention goes to Guido Polles and Daniele Granata: both helped me in many ways with their experience and their friendship, and I owe them a lot.

My gratitude goes to the whole SBP staff as well, in particular Angelo Rosa, Alessandro Laio and Giovanni Bussi, who together created a wonderful group in SISSA, and a friendly environment for us, students and postdocs.

I must also thank all the SBP people. In particular, my officemates, Alejandro and Lorenzo; Francesca and Giovanni, who started with me this adventure four years ago; and the ping-pong-after-lunch club.

A big thanks to Mauro, Alessandra and Giulia: in these years, they really have been like a second family to me, a small piece of Franciacorta in Trieste.

And finally, my family: my parents, my brother and my uncles, that were always present for me. Thank you.