Simpler protein domain identification using spectral clustering

Frédéric Cazals *and Jules Herrmann* and Edoardo Sarti*

February 10, 2024

Abstract

The decomposition of a biomolecular complex into domains is an important step to investigate biological functions and ease structure determination. A successful approach to do so is the SPECTRUS algorithm, which provides a segmentation based on spectral clustering applied to a graph coding interatomic fluctuations derived from an elastic network model.

We present SPECTRUS++ , which makes three straightforward and useful additions to SPECTRUS. For single structures, we show that high quality partitionings can be obtained from a graph Laplacian derived from pairwise interactions–without normal modes. For sets of homologous structures, we introduce a Multiple Sequence Alignment mode, exploiting both the sequence based information (MSA) and the geometric information embodied in experimental structures. Finally, we propose to analyse the clusters/-domains delivered using the so-called D-family-matching algorithm, which establishes a correspondence between domains yielded by two decompositions, and can be used to handle fragmentation issues.

Our domains compare favorably to those of the original SPECTRUS, and those of the deep learning based method Chainsaw. In particular, a comparison of SPECTRUS++ and Chainsaw on the manually curated domain classification ECOD as a reference shows that high quality domains are obtained without using any evolutionary related piece of information.

SPECTRUS++ is provided in the Structural Bioinformatics Library at http://sbl.inria.fr.

Keywords: protein domains, conformational changes, protein dynamics; spectral clustering, graph cut, diffusion models.

1 Introduction

1.1 Protein segmentation and dynamic domains

Molecular functions, motions, and domains. Molecular motions within proteins span ~ 15 orders of magnitude [1], and their prediction is key to shed light on complex biological functions. A classical strategy to study these complex phenomena consists of abstracting a molecular system in a hierarchical fashion, with rigid domains connected by linkers. While this decomposition may be carried out in a recursive fashion, the simplest model consists of a flat hierarchy–rigid domains connected by linkers. Such a model corresponds to a segmentation such that interactions within domains are tight, while those across domains are looser. Identifying domains based on such interactions requires taking into account the atomic packing, the presence of disulfide bonds, hydrogen bonding, and naturally atomic movements. In modeling a biomolecule as a graph whose units are amino acids or bases, with an edges between interacting units, the previous physical intuition calls for methods reminiscent of graph clustering and/or community detection. We now briefly review two classes of methods for this task.

Domains via (non) supervised approaches. Spectral graph partitioning techniques form a rich class of methods which may be interpreted in terms of graph min cuts as well as random walks and Markov chains [2].

*Inria - Université Côte d'Azur. Email: frederic.cazals@inria.fr, edoardo.sarti@inria.fr

The domain partitioning algorithm SPECTRUS [3] falls in this vein, as it relies on spectral clustering based on a similarity matrix derived from fluctuations and normal modes–details thereafter. A similar three step partitioning procedure is SWORD2 [4]. The first step performs an assignment of secondary structure elements (SSE) with eight conformational states. The second one is based on a contact map defined from a logistic function applied to inter-residue distances this step identifies protein units (PUs) using an iterative hierarchical clustering of the contact probability map. The final step reconstructs domains by gradually merging PUs, maximizing the compactness of domains and their separation. The algorithm UniDoc defines domains by maximizing intra-domain interactions and minimizing inter-domain interactions. Inter/intra domain scores are defined from a logistic function on the pairwise C_β (C_α for glycine) distances [5]. In a first step, the protein is split into fragments (using so-cal. led continuous and discontinuous splits) based on a graph min cut. The second one iteratively merges fragments into domain to maximize the interaction inside each domain.

Supervised approaches were also recently developed. The method FUPred [6] couples contact maps yielded by deep residual neural networks and coevolutionary precision matrices. Chainsaw [7] uses a residual convolutional neural network to predict the aforementioned adjacency matrix over residues. The structure is converted into five feature channels (pairwise C_α distances, and four channels for predicted SSE (helix vs strand, within vs boundary of SSE)). The features are converted into a pairwise probability matrix using a deep convolutional network initially designed for structure prediction–a variant of trRosetta [**?**]. The learning process minimizes the cross entropy between the predicted soft adjacency matrix and the one representing the probability of residue co-occurrence in the same domain.

The Domain Parser for Alphafold models (DPAM) targets AlphaFold2 models [8]. This raises two new challenges: the large number of structures, and the requirement to handle non domain regions which are unsuitable for domain classification (IDR, linkers, coiled coil, etc). Of particular interest in this context is the Predicted Aligned Error (PAE) provided by AlphaFold2. A *combined probability* for pairs of residues is defined, based on the PAE, the inter-residue distance, the HHsuite support, and the Dali support. The weights are learned to match a library of known domains. Using a partitioning of the protein into residues segments of length five, an iterative clustering of residues is performed to obtain domains.

1.2 Contributions

We present three straightforward and useful simplifications and extensions to SPECTRUS.

First, for single structures, we show that high quality partitioning can be obtained from a graph Laplacian derived from pairwise interactions of several types (covalent, non-covalent, hydrogen bonds, salt-bridges, etc). This mode, termed the *diffusion map* (DM) mode is a significant simplification of SPECTRUS algorithm [3], which derives atomic fluctuations from the normal modes of an elastic network model (ENM) [9]. The DM mode indeed relies on a single eigenproblem (used in spectral clustering) instead of two (normal modes, spectral clustering). Denoting n the number of residues, the fact that the latter operates on a matrix of size $n \times n$ instead of $3n \times 3n$ (for normal modes) yields a speedup of at least one order of magnitude–considering a quadratic lower bound for sparse eigenproblems.

Second, for sets of structures, we introduce a multiple sequence alignment (MSA) mode, in which fluctuations are derived from valid positions within the MSA. Using various structural measures (lRMSD from flexible structural alignments and combined RMSDs), we show in particular that the domains obtained are useful to study conformational changes and molecular motions.

Third, we propose to analyse the clusters/domains delivered in a hierarchical fashion, using an edit distance between clusters, and we use this technique to reduce domain fragmentation issues faced by most if not all previous methods.

We show that our domains compare favorably to those output by two reference methods, the original SPECTRUS, and those of the deep learning based method Chainsaw. We show in addition that our clusters are more coherent with those from the reference database ECOD, despite the fact that we do not use any evolutionary related piece of information.

The software SPECTRUS++ is made available in the Structural Bioinformatics Library, see https://sbl.inria.fr/applications/ and https://sbl.inria.fr/doc/Spectrus-user-manual.html.

2 Methods: SPECTRUS and SPECTRUS++

2.1 The original SPECTRUS algorithm

For the sake of completeness and exposure, let us review briefly the e SPECTRUS algorithm [3]. The method relies on the following intuition to define domains: atomic fluctuations within (resp. across) domains are tame (resp.large). For a predefined number of clusters/domains k, the algorithm is as follows:

1. Compute a fluctuation matrix assessing the distance variation between any two residues in the chain. Denoting d_{ij} the distance between the C_α carbons of amino acids i and j, the quantity of interest is the standard deviation
$$f_{ij} = \sqrt{< d_{ij}^2 > - < d_{ij} >^2}. \tag{1}$$
For a single structure [3], fluctuations are derived from from normal modes associated with a harmonic model [9]. For several homologous chains, one estimates Eq. (1) from the distances observed in the conformations processed.

2. Convert the fluctuations f_{ij} into similarities w_{ij} via thresholding and exponentiation. That is, with σ a conservative measure for intra-cluster/domain dissimilarity and $d_{\max}$ a distance threshold:
$$w_{ij} = \begin{cases} \exp(-f_{ij}/2\sigma^2) & \text{if } d_{ij} \leq d_{\max}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$
This results in a weighted graph whose vertices are the individual residues. Intuitively, the weight w_{ij} on edge i to j can be seen as the probability for residues i and j to be in the same cluster/domain.

3. Apply spectral clustering to the previous graph [10]. This step involves clustering vectors on the unit sphere S^{k-1}, which is done with k-medoids in SPECTRUS.

To identify suitable values of k, the previous steps are repeated. A classical way to assess the significance of a clustering is to resort to a null model [11, 12]. For SPECTRUS, the quality of the domains obtained for each k is measured by a score ($\text{Score}_{k,n}$, Eq. (4)) based on a ratio involving a null model consisting of random points on S^{k-1}. See details in Sec. S1. The boxplot of this score as a function of k is termed the *quality plot*. The values of k interest are the local maxima of the quality plot with low variance of the score $\text{Score}_{k,n}$. In increasing k, one expects this variance to grow: a larger k entails more partitions, some with less pertinent boundaries, whence a larger score and variance. As we shall see, this behavior is indeed observed in practice and is useful in picking appropriate values of k.

2.2 SPECTRUS++

We now present our three additions to SPECTRUS.

Direct segmentation of a single model: the diffusion map (DM) mode. The original SPECTRUS for single structures uses spectral clustering based on similarities defined from the normal modes of an elastic network model (ENM). A more direct route consists of defining the weights w_{ij} directly from the spring constants of the ENM, thus removing the normal modes calculation. Consider three stiffness constants $\gamma_{ij}^{\text{Cov}}(= 10), \gamma_{ij}^{\text{NCov}}(= 1)$, and $\gamma_{ij}^{\text{HB}}(= 1/2)$, for covalent, non covalent, and hydrogen bonds respectively. (We note in passing that additional terms *e.g.* for salt bridges can easily be accommodated to weight contacts.) Because non covalent contacts are detected using a distance threshold, denoting d_{ij} the distance between the C_α carbons of two interacting residues, we soften the stiffnesses γ_{ij} as follows:
$$w_{ij} = \gamma_{ij} * exp(-\frac{\mid d_{ij} - d_0 \mid}{\sigma}). \tag{3}$$

We take $d_0 = 4.5$Å as typical distance between consecutive C_αs along the backbone, and $\sigma = 1/2$.

Note that two amino acids may interact in different ways. For example, two a.a. on the same generatrix of a α-helix interact non-covalently, but also via hydrogen bonding. We therefore define two modes to define the weight w_{ij} of two residues with multiple interaction types, the `Weight-Max` (resp. `Weight-Sum`) mode in which we take the max (resp. the sum) of individual stiffnesses. In our helix example, one gets $\gamma_{ij} = \max(\gamma^{\text{NCov}}, \gamma^{\text{HB}})$, and $\gamma_{ij} = \gamma^{\text{NCov}} + \gamma^{\text{HB}}$ respectively. The resulting weights are directly passed to spectral clustering based on the symmetric Laplacian [10].

This diffusion map (DM) mode preserves the anisotropy of fluctuations embodied in normal modes (if any) via the anisotropy of contacts (if any). But it has two direct advantages over the ENM mode (original `SPECTRUS`): (i) the segmentation requires contacts/distances only rather than coordinates used to derive the Hessian of the ENM, (ii) the matrix diagonalization operates on a sparse matrix of size $n \times n$ rather than $3n \times 3n$.

Remark 1 *In processing PDB files, hydrogen bonding information is retrieved from secondary structures. A mode loading a* `DSSP` *file is available, but was not used for experiments.*

`SPECTRUS` **for multiple homologous structures: the MSA mode.** Consider multiple structures for which a MSA is available. We cover the case of conformations of the same protein using this model also, since experimental structures of the same protein often harbors addition/deletion at the N-tear and/or C-ter of the chain. A position in the MSA is termed *valid* provided that (i) the corresponding chains have no gap at this position in the MSA, and (ii) the corresponding residues have been solved in the individual experimental structures. The fluctuation of Eq. (1) can now be computed for each pair of valid indices in the MSA, and used for spectral clustering again. One obtains a partitioning of the structures into domains which are *compact/rigid*, but move with respect to one another.

While conceptually straightforward, the MSA mode requires a careful treatment of all indices involved (resids in structures which are not contiguous in general, sequence indices, alignment indices). See Sec. S1.2.

Remark 2 *The ability to handle homologous structures via a MSA deserves a comment in terms the sparsity of alignments and structures handled. If too many residues are invalid (Section S1, Defs. 2 and 3), the similarity matrix is too sparse, which results in a large number of clusters. Intuitively, one should recall that the eigenspace associated to the value 0 of the graph Laplacian encodes connected components of the graph.*

Fragmentation issues. The spectral decomposition may yield fragmentation issues, when a region (*e.g.* a region from an helix) is assigned to a nearby domain. For `SPECTRUS` or `SPECTRUS++` , this happens when the distance d_{ij} between two residues is small, which results in large weight w_{ij} – Eq. (3).

We handle such issues using an algorithm identifying the correspondences between the domains of two candidate decompositions, called the D-family-matching algorithm [13]. Let $\mathcal{C}_1 = \{C_{1,i}\}$ and $\mathcal{C}_2 = \{C_{2,j}\}$ be two decompositions into domains associated to two values k_1 and k_2. In short, the D-family-matching creates groups of clusters (domains in our case) within $\mathcal{C}_1$ and likewise within $\mathcal{C}_2$, called *meta-clusters*, with a one-to-one correspondence between these meta-clusters. The method relies on the complete edge-weighted bipartite graph $\mathcal{C}_1 \times \mathcal{C}_2$, the weight of edge $(C_{1,i}, C_{2,j})$ being equal to the number of points (amino-acids in our case) shared by $C_{1,i}$ and $C_{2,j}$. It is governed by a scale parameter D counting the maximum number of steps (zigzags) one can make to join a cluster from $\mathcal{C}_1$ to a cluster of $\mathcal{C}_2$–formally the diameter of the subgraph induced by meta-clusters. The number of meta-clusters obtained is dictated by the combinatorial optimization problem solved. See [13] for details, as well as the software in the Structural Bioinformatics Library `https://sbl.inria.fr/doc/D_family_matching-user-manual.html`.

Consider a chain which faces a fragmentation issue (Figs. S2 and S3). Let us consider two decompositions $\mathcal{C}_1$ and $\mathcal{C}_2$ of this chain into domains. Further decompose each domain $C_{1,i} \in \mathcal{C}_1$ (and likewise $C_{2,j} \in \mathcal{C}_2$) into contiguous stretches along the sequence of the chain, yielding $C_{1,i} = \cup_j s_{1,i,j}$ – with i (resp. j) the index of the domain (resp. stretch). To possibly re-assign a mis-assigned stretch $s_{1,i,j}$ to the correct domain of $\mathcal{C}_1$, we proceed in three steps. (NB: the method is mutatis mutandis for a stretch from a domain from $\mathcal{C}_2$.)

First, using D-family-matching with $D = 2$, we compute meta-clusters $M_k, k = 1, \ldots, K$ to match groups of $\{C_{1,i}\}$ with groups of $\{C_{2,j}\}$. The clusters of $\mathcal{C}_1$ (resp. $\mathcal{C}_2$) aggregated by M_k are denoted $M^1_{k,l}$ (resp. $M^2_{k,l}$). Second, we compute the intersection size $|s_{A,i;j} \cap \cup_l M^2_{k,l}|, k = 1, \ldots, K$. Let $o \in 1, \ldots, K$ be the index of the meta-cluster yielding the best score for $s_{1,i,j}$. Consider the set $M^1_{o,l}$ of clusters/domains from $\mathcal{C}_1$ associated with this meta-cluster M_o. Because $s_{1,i,j}$ is contiguous along the sequence, it is flanked by exactly one or two domains from $\mathcal{C}_1$. Let the *flanking number* the number of such domains from $M^1_{o,l}$. If the flanking number is one, we assign $s_{1,i,j}$ to that domain. If not, we do not perform any update: a value of zero would mean that the stretch is not contiguous to a domain from $M^1_{o,l}$; and a value > 1 would trigger the merge of two domains identified by `SPECTRUS++` .

Remark 3 (Clustering on S^{k-1} and quality plots.) *Spectral clustering requires clustering unit vectors on the unit sphere S^{k-1}. We use* `k-means++` *instead of* `k-medoids` *in* `SPECTRUS` *[14]. Due to the hardness of* `k-means` *clustering and the random initialization of seeds in* `k-means++`*, several clusterings are in general obtained for a given value of k. Whence the boxplot associated to each value of k on our quality plots.*

In practice, the default range of values explored by `SPECTRUS++` *is $[k_{min} = 2, k_{max} = 10]$. On the quality plots, the values of k such that primary sequence size divided by the number of domains is less than 35 amino acids (default size for a helix) are shaded, as decomposing a structure into SSE is not the goal pursued.*

2.3 Implementation.

The implementation is provided in the Structural Bioinformatics Library [15]. See Sec. S1 (Algo. 1) for the pseudo-code, and `https://sbl.inria.fr/doc/Spectrus-user-manual.html`.

3 Experiments

We present three complementary experiments: Sec. 3.1 illustrates the ability of `SPECTRUS++` to handle challenging case for competing methods; Sec. 3.2 compares the MSA and DM modes of `SPECTRUS++` to study conformational changes and molecular motions; Sec. 3.3 studies the coherence between the domains found by `SPECTRUS++` and those of the `ECOD` database [16].

Experiments in the DM mode for single structures use the `Weight-Sum` mode for weights–Sec. 2.2, and exploit hydrogen bonding information from SSE elements in PDB file headers. Default values were used for the stiffness constants, that is $\gamma^{\text{Cov}}_{ij} = 10, \gamma^{\text{NCov}}_{ij} = 1, \gamma^{\text{HB}}_{ij} = 1/2$.

3.1 Comparisons between `SPECTRUS`, `Chainsaw`, and `SPECTRUS++`

Rationale: assessing partitioning and fragmentation. In order to illustrate the efficacy of `SPECTRUS++` , we present in this section three examples of increasing difficulty: the first is extracted from the `SPECTRUS-ori` test set, the second from the `Chainsaw` test set, and the third is a system we recently studied and for which we have clear dynamical information. The complete test sets from the `SPECTRUS-ori` and `Chainsaw` works are included in the SBL package.

Results: Escherichia coli adenylate kinase. The Escherichia coli adenylate kinase (PDBid 1ake and 4ake), studied in [3], is a monomeric enzyme of about 200 amino acids, which freely interconverts between open and close conformations (Fig. 1). When `SPECTRUS-ori` is given the two conformations, it finds the correct number of clusters $k = 3$, where two clusters correspond to the ATP binding domain (LID) and the AMP binding domain (NMP) whereas the third includes the rest of the structure. This decomposition incurs two spurious mis-assignment due to fragmentation issues (Fig. 1, `SPECTRUS-ori`, circled areas). `SPECTRUS++` in its MSA mode reproduces this result, but does not suffer from these mis-assignments. `SPECTRUS++` also finds the same partition from the sole open conformation (PDBid 4ake), which confirms the robustness of the method. On the same system, `Chainsaw` was not able to detect multiple domains.

Results: human serum transferrin. The human serum transferrin (PDBid 1a8e) presents two domains one of which is discontinuous (i.e. formed by the union of two continuous polypeptide segments). This poses a challenge for algorithms that are intrinsically based on sequence such as `Chainsaw`, but does not pose additional complexity for graph-based approaches such as `SPECTRUS++` . `Chainsaw` and `SPECTRUS++` both correctly predict the partition (Fig. 2), but `Chainsaw` incorrectly assigns the last residue of the C-terminal helix. `SPECTRUS-ori` also correctly finds $k = 2$, but misassigns the same C-terminal helix as well as one of the two beta sheets linking the two domains.

Results: acriflavine resistance B. Acriflavine resistance B (AcrB, PDBid 4zit) is a transmembrane drug/proton antiporter serving as drug efflux pump in Gram-negative bacteria. It is a homotrimer where each monomer cycles between three consecutive states that compose the alternating access mechanism. These conformational changes rely on the interplay of 7 quasi-rigid dynamical domains and 1 disordered domain. `SPECTRUS++` finds a partition in 4 clusters that are unions of the real quasi-rigid domains (Fig. 3). So does `Chainsaw` (although its partition is not symmetric) and `SPECTRUS-ori` (although its domains include several incorrect fragments). `SPECTRUS++` is the only one to suggest the correct partition as second-best candidate.

Remark: quality score variance as a function of k It is interesting to note that in all three examples, the quality score variance increases with increasing k. Although qualitative, this trend reveals that finer partitions are on average less stable than coarser ones – larger values of k increase the number of possible partitions, some of which incur a large variance.

3.2 Domains: prediction in case of conformational changes

Rationale: modeling conformational changes. We wish to compare the DM and MSA modes in case of conformational changes. In the sequel, we consider three cases from the *molecular motions DB* [17], see `http://www.molmovdb.org/cgi-bin/browse.cgi`.

The MSA mode is of special interest since fluctuations directly exploit conformations. It requires a MSA, which we take for granted. For a pair of structures, the nature of conformational changes can be studied using three complementary measures:

- A lRMSD can be computed using the mapping between valid positions provided by the MSA.
- Irrespective of the MSA, a structural alignment can be computed for the two chains. We do so using the structural aligner `Kpax` [18], re-implemented within the `Kpax` package of the `SBL`.
- Assume domains identified by `SPECTRUS++` have been obtained. A lRMSD can be computed for each such domain, and these can be combined to provide the *combined RMSD* ($\text{RMSD}_{\text{Comb.}}$) based on the per-domain changes [19]. We use the package Molecular_distances_flexible package from the SBL, see `https://sbl.inria.fr/doc/Molecular_distances-user-manual.html`.

Results: glutamine-binding protein (GlnBP). GlnBP (226 a.a.) is a protein of ellipsoidal shape consisting of two domains–[20] and Fig. 4. It has been crystallized alone and with its ligand (Gln). Upon binding, GlnBP-Gln exhibits a large-scale movement of the two hinges connecting the two globular domains, with two significant angular changes in the two hinges (41.1 deg . in the Φ angle of Gly89; 34.3 deg . in the ψ angle of Glu181) [20]. The two hinges and the two domains are unambiguously detected by `SPECTRUS++` (Fig. 4), both in the MSA and DM modes (even if this latter mode uses a single structure). The values for the triple (lRMSD , lRMSD for the `Kpax` alignment, $\text{RMSD}_{\text{Comb.}}$) clearly establishes the rigidity of the two domains along this hinge motion (Table 1).

Results: yeast elongation factor (eEF2). eEF2 (842 a.a.) is a GTP binding protein mediating the translocation of peptidyl-tRNA from the A site to the P site on the ribosome. In yeast, sordarin is a selective inhibitor of eEF2. The conformational changes between unbound and bound (with sordarin) of eEF2 have

been studied in detail [21]. To analyse the domains identified, let us recall that six domains are usually considered within eEF2-see [21] and Fig. 5: three N-ter domains: 2-218 or 329-345 (domain I or G-domain), 219-328 (G' domain), 346-481 (domain II); three C-ter domains: 482-558 (domain III), 559-726 or 801-842 (domain IV), 727-800 (domain V).

The comparison unbound to bound provided by the MSA mode of SPECTRUS++ suggests inspecting three decompositions – $k = 2, 4, 6$ (Fig. 5). The first ($k = 2$) shows two meta-domains moving relatively to one another. The three N-ter domains (I, G', II) are grouped together by SPECTRUS++ , which makes sense since they move as a rigid body [21]. The second local maximum ($k = 4$) splits the second meta-domain into the known C-ter domains (III, IV, V). This is coherent with the previous studies, where it is shown along the conformational change unbound to bound, complex motions occur between domains (III, IV, V), and of course with respect to the three N-ter domains. Finally, for $k = 6$, the MSA mode fully recovers the 6 domains as described in the literature. On this example too, the various structural measures indicate that the domains identified are pertinent to study the relative motions at different scales (Table 1).

On such a complicated case, inferring domains from a single structure with the DM mode is more challenging (Fig. 7). The analysis of the unbound case is not very informative. However, that of the bound structure is so. With $k = 2$, the split into three N-ter and C-ter domains is obtained. Also, the last value of k with low variance ($k = 6$) identifies the essential features of the six domains.

Altogether, these results show the ability of SPECTRUS++ to describe molecular motions at different scales, in particular when exploiting several conformations.

	PDBid/chain	PDBid/chain	#aa	lRMSD	#aa	lRMSD Kpax	#aa	$\text{RMSD}_{\text{Comb.}}$	k	Fig.
GBP	1ggg/A	1wdn/A	220	5.34	164	3.27	215	1.43	2	4
EF2	1n0u/A	1n0v/C	818	13.60	537	4.17	818	4.57	2	5
							818	3.90	4	5
							818	3.87	6	5

Table 1: **Fluctuations/MSA mode: structural comparisons for pairs of structures.** Columns 3-4: lRMSD computed from the sequence alignment; Columns 5-6: lRMSD computed using the structural aligner Kpax; Columns 7-8: combined RMSD $\text{RMSD}_{\text{Comb.}}$ computed from the domains identified by SPECTRUS++ . NB: NA values for the (1eps/A, 1g6s/A) owe to the fact that 1eps contains CA only, so that the canonical representation used by Kpax (based on N, C_α, C atoms) cannot be computed.

3.3 Domains: comparison against a manually curated domain classification

Rationale: large-scale domain detection on an evolutionary domain classification. Topology-based structural domain classifications such as CATH and SCOP are often incomplete with respect to homology-based approaches. For this reason, we chose to test our method against ECOD [16], whose collection of structural domain has been found to be the most extensive, while remaining consistent with SCOP on the shared domains [22].

Results: variation of information on ECOD domains. For each ECOD family featuring multi-domain systems, we chose a representative structure and thus collected a set of ~ 3000 proteins partitioned in their domains. Since ECOD domains do not in general cover the whole sequence, for each partitioned protein, we define an extra domain containing all the residues that ECOD did not assign to any of its domains.

We then used the D-family-matching algorithm again in order to relate ECOD and SPECTRUS++ domains: each of the two partitions was re-annotated basing on the metacluster indices, then the variation of information VI was calculated on the metacluster partitions. (For X_1 and X_2 two discrete random variables–describing here the assignment of amino acids to domains, denoting H the entropy and I the mutual information, recall that $\text{VI}(X, Y) = H(X_1) + H(X_2) - 2I(X_1, X_2)$. VI is null if and only if the two clusterings are equal [23].) The variation of information depends on the number of meta-clusters, but also on the relative

values of the number of meta-clusters and the number of ECOD domains. For a domain partitioning method (SPECTRUS++ , Chainsaw), we therefore consider the scatter plot #meta-clusters/#ECOD clusters versus VI (Fig. 6). We first note that VI is less that 0.25 for both methods. A one-sided Mann-Whitney U test (a non-parametric two-sample test) to compare the individual values $\mathrm{VI}_{\mathrm{SPECTRUS++}}$ versus $\mathrm{VI}_{\mathrm{Chainsaw}}$ yields a $\mathrm{p-value} \sim 10^{-11}$, showing that the former values are lower than the latter. Remarkably, the partitioning returned by SPECTRUS++ is more coherent than that of Chainsaw with respect to ECOD, without using any evolutionary piece of information.

4 Outlook

Molecular motions are critical for the activity of biomolecules. They are at play at different scales, ranging from subtle vibrational properties contributing to the free energy of the system, to large amplitude conformational changes repositioning whole structural domains. The automatic identification of such domains is of particular interest as a first step towards the mechanistic explanation of complex mechanisms. Rooted in the vast realm of *graph partitionning / community detection methods*, the original SPECTRUS algorithm provides a robust and physically grounded method to identify structural domains. Our work simplifies this method by replacing the computation of normal modes by a simpler *diffusion map* mode relying on a graph Laplacian directly coding pairwise interactions between amino acids. The DM mode significantly reduces the complexity of the method while retaining the physically appealing rationale. We also propose a fully fledged *multiple sequence alignment* mode (MSA), leveraging sequence alignments and fluctuations in experimental structures. The comparison of the DM (single molecule) and MSA (multiple structures) modes stressed the robustness of the method and the coherence of the two modes. Finally, we show how to combat so-called fragmentation issues using a matching between clusters of clusters. Our battery of tests, which handles challenging cases for competing methods, shows that SPECTRUS++ compares favorably to contenders. We anticipate that its integration to the Structural Bioinformatics Library, along with a variety of complementary structure analysis tools, will contribute to its dissemination and broad use.

Two ongoing developments are of particular interest. The first relates to the prediction of mechanisms for macromolecular machines involving a large number of domains (say 10 or more). When these domains move in a rigid fashion with respect to one another, their identification coupled to the ability to sample coherently multiple loop conformations paves the way to deciphering these complex dynamics. This perspective is backed up by our recent experience with resistance-nodulation-cell division (RND) transporters, and membrane proteins in general. The second concerns the ability to identify partially disordered domains, which often occur with deep structure predictors. So far, a direct consequence of the spectral clustering strategy employed in SPECTRUS++ is the assignment of all amino acids to domains. Incorporating into SPECTRUS++ a disorder predictor and/or a special geometric/topological processing to identify partially disordered regions is a route of choice to handle such cases.

5 Artwork

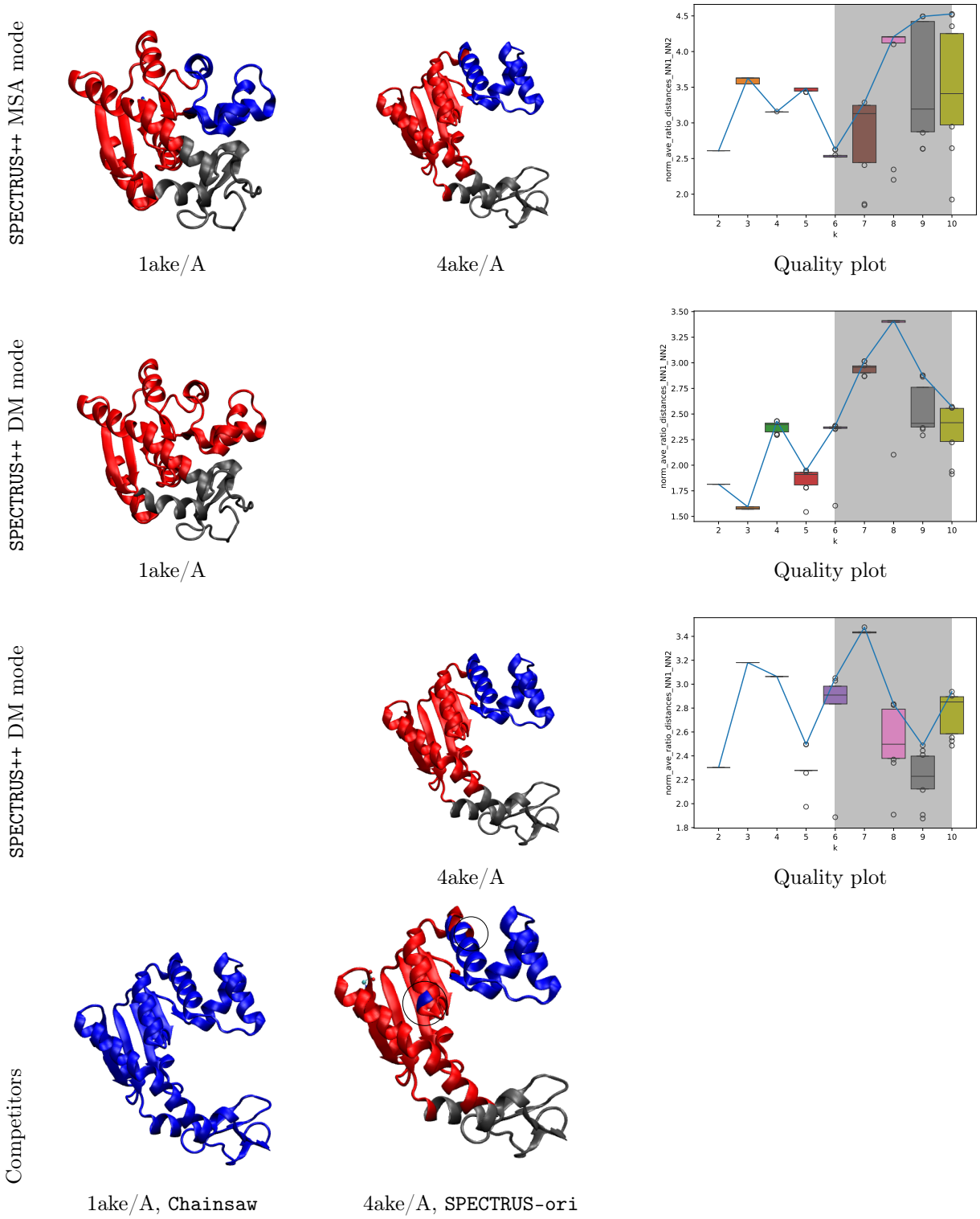

Figure 1: **Escherichia coli adenylate kinase: comparison of the SPECTRUS++ modes MSA and DM.** close and open conformations (PDBid 1ake, 4ake). Crystal structures: closed: 1ake/A; open: 4ake/A. (Row 1) MSA mode applied to chains A of 1ake and 4ake. (Row 2) DM mode for 1ake/A. (Row 3) DM mode for 4ake/A. (Row 4) Comparison with SPECTRUS-ori and Chainsaw

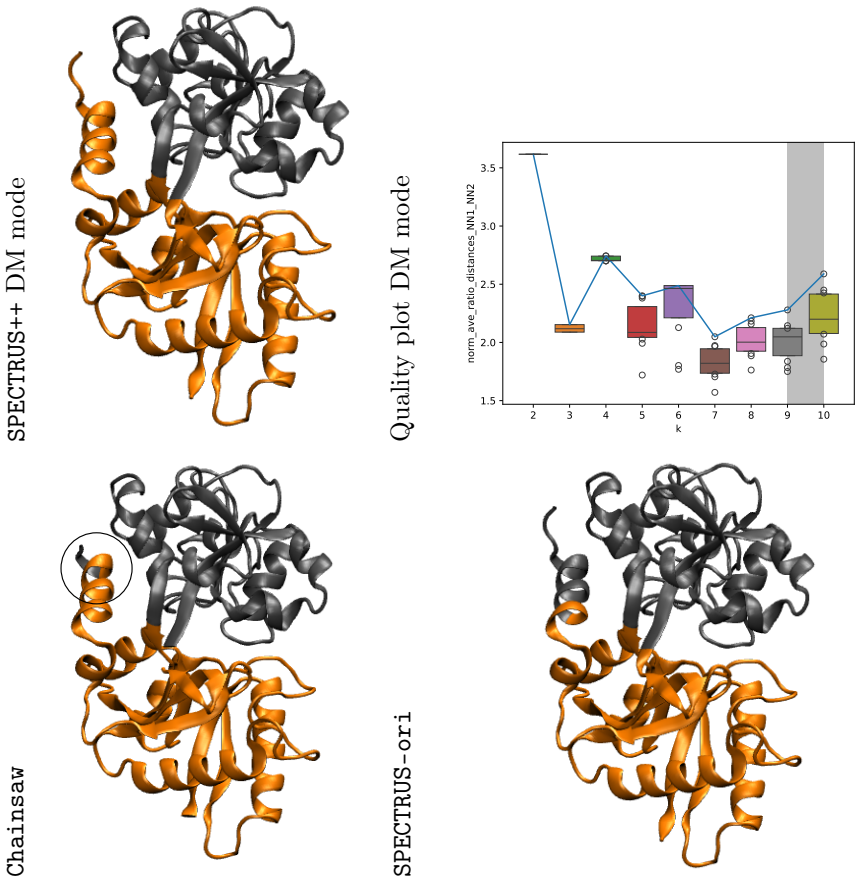

Figure 2: **Human serum transferrin (PDBid 1a8e/A), DM mode.** The protein has two dynamical domains one of which composed by non-contiguous segments. `SPECTRUS++` , `Chainsaw`, and `SPECTRUS-ori` recover the correct number of clusters, with slight differences in the partition on the terminal helix.

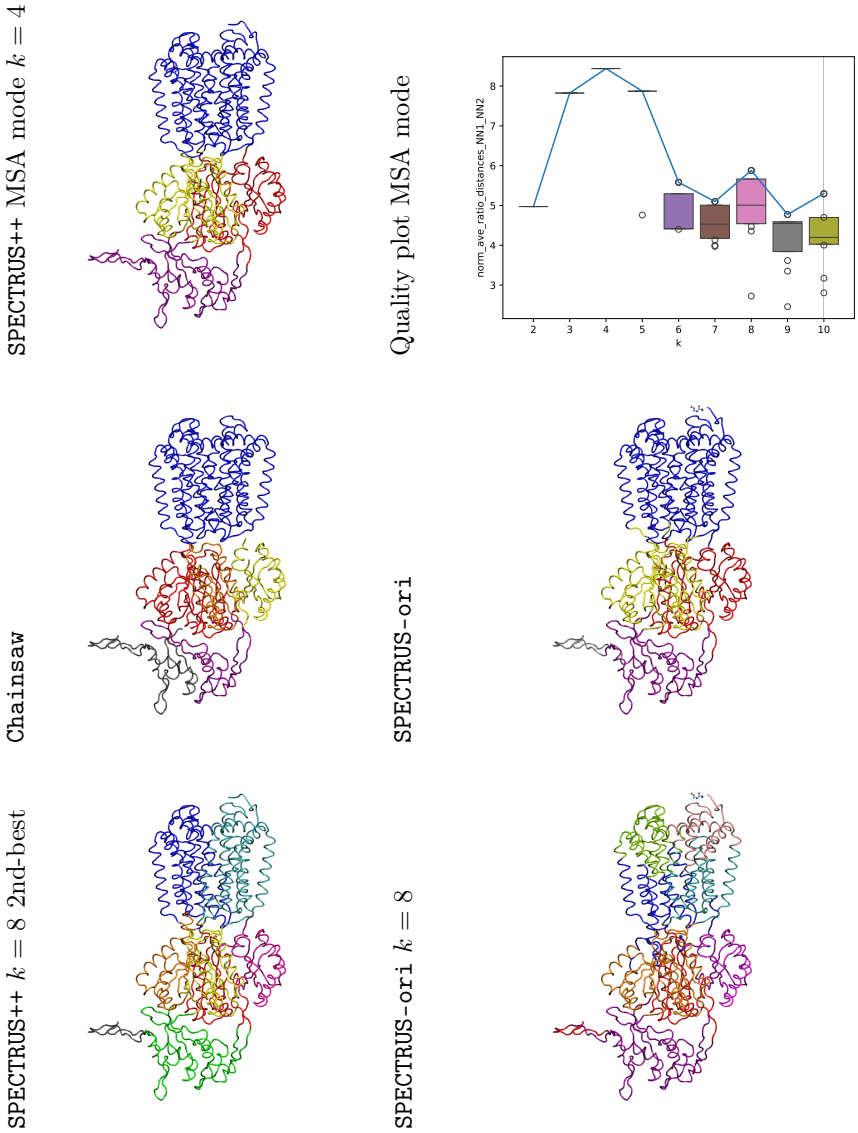

Figure 3: **Monomer of the trimeric Acriflavine resistance B (AcrB, PDBid 4zit/A), MSA mode.** The monomer can be decomposed in 7 quasi-rigid domains plus 1 disordered domain. `SPECTRUS++` is the only algorithm to suggest the correct repartition as second-best choice, the best one still being correct yet more coarse-grained.

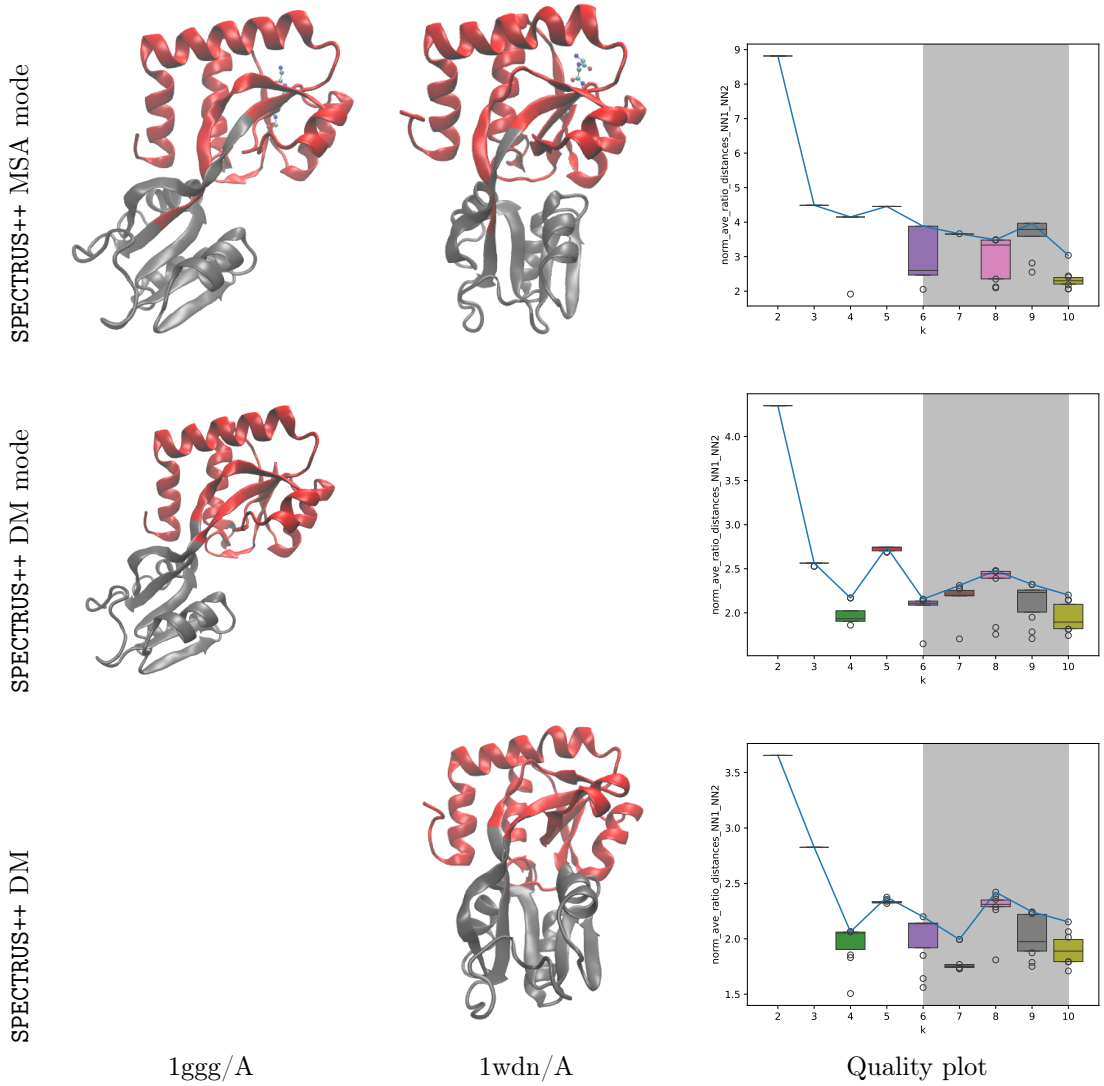

Figure 4: **Glutamine-binding protein (GlnBP): comparison of the SPECTRUS++ modes MSA and DM.** Crystal structures: unbound: 1ggg/A; bound: 1wdn/A. **(Top)** MSA mode applied to chains A of 1ggg and 1wdn. **(Middle)** DM mode for 1ggg/A. **(Bottom)** DM mode for 1wdn/A.

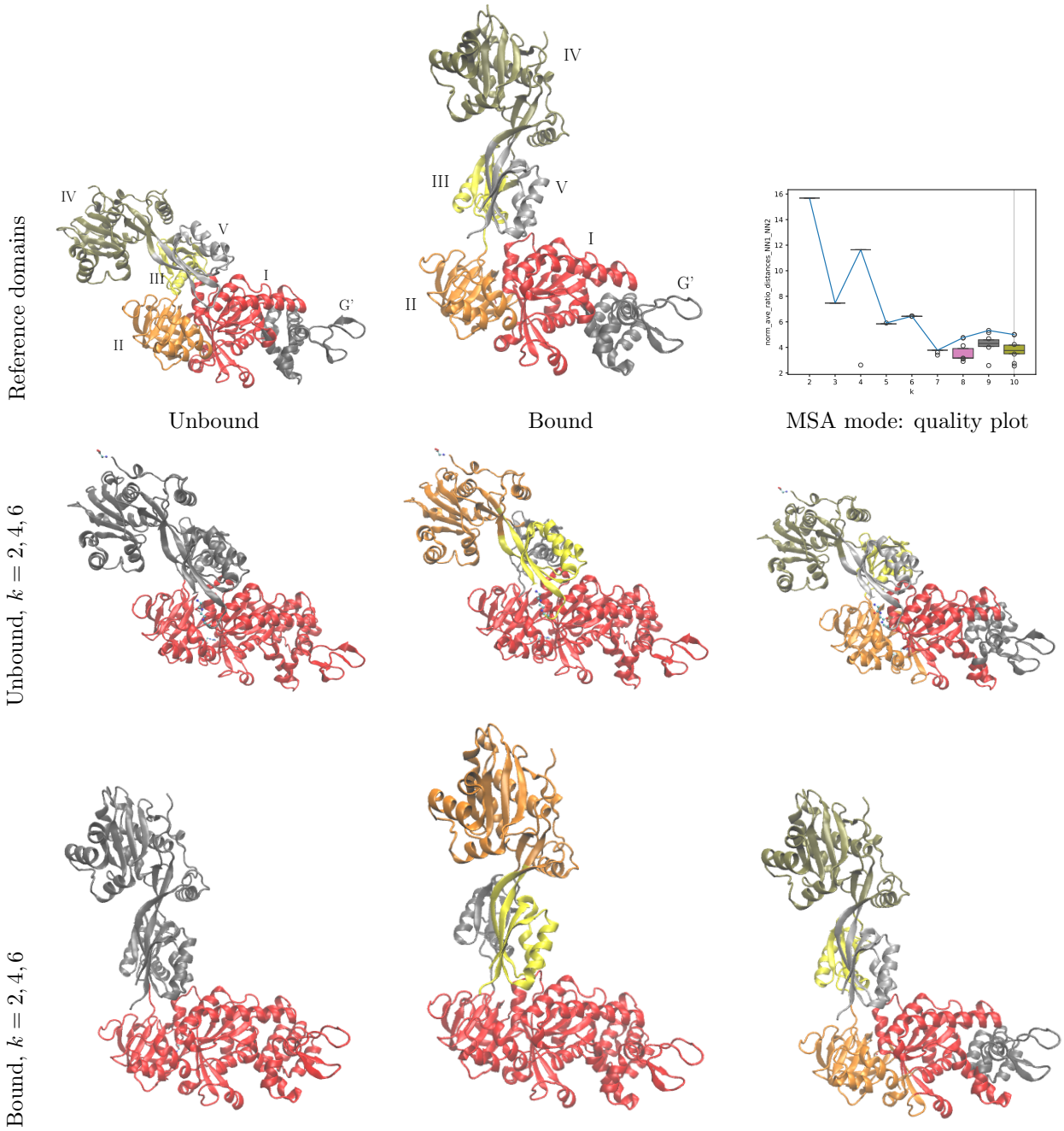

Figure 5: **Yeast elongation factor eEF2: analysis with SPECTRUS++ in MSA mode.** Crystal structures: unbound: 1n0v/C; bound: 1n0u/A. Values of k of interest: $k = 2, 4, 6$ – local maxima with low score variance. **(Top)** Reference domains [21] in for the unbound (1n0v) and bound structures (1n0u), and quality plot for the MSA mode. **(Middle)** Domains of the unbound structure. **(Bottom)** Domains of the bound structure.

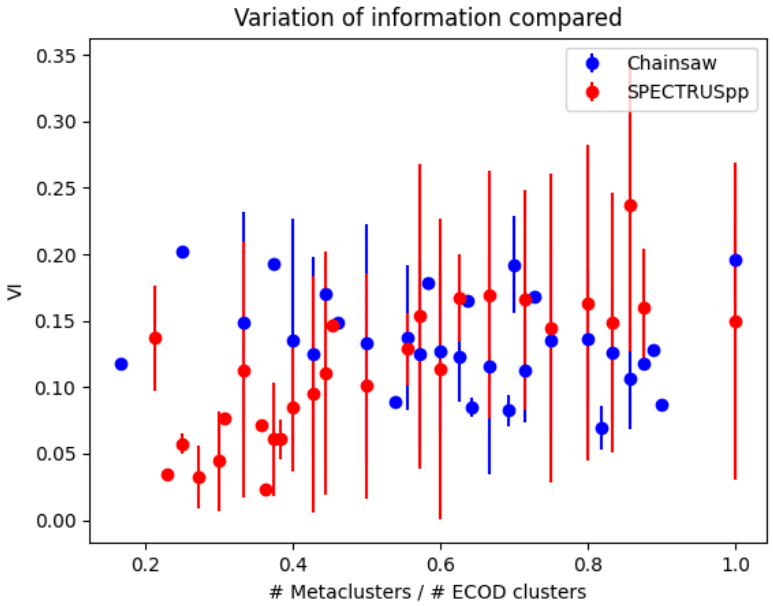

Figure 6: **Variation of information (VI) between the clusters of `SPECTRUS++` and `Chainsaw` against those from `ECOD`.** The x-axis reflects the dependency of the VI on the number of labels (metaclusters). When the ratio #metaclusters / #`ECOD` increases, the average VI and its deviation become larger–the coherence with `ECOD` decreases. The one-sided Mann-Whitney U test (a non-parametric two-sample test) shows that the VI values from `SPECTRUS++` are smaller than those from `Chainsaw`– $p - \text{value} \sim 10^{-11}$.

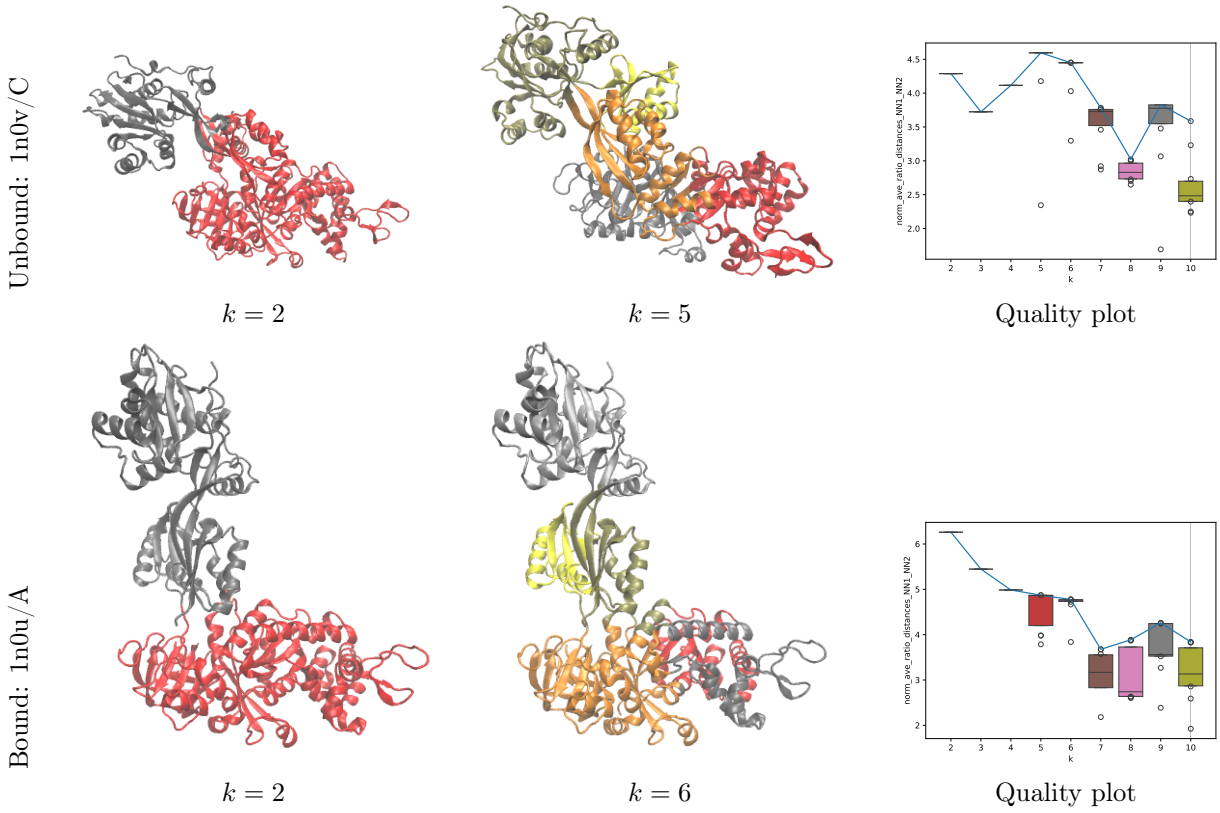

Figure 7: **Yeast elongation factor eEF2: analysis with SPECTRUS++ in DM mode.**

References

[1] S.A. Adcock and A.J. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615, 2006.

[2] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[3] 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.

[4] Gabriel Cretin, Tatiana Galochkina, Yann Vander Meersche, Alexandre G de Brevern, Guillaume Postic, and Jean-Christophe Gelly. Sword2: hierarchical analysis of protein 3d structures. *Nucleic Acids Research*, 50(W1):W732–W738, 2022.

[5] Kun Zhu, Hong Su, Zhenling Peng, and Jianyi Yang. A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinformatics*, 39(2):btad070, 2023.

[6] Wei Zheng, Xiaogen Zhou, Qiqige Wuyun, Robin Pearce, Yang Li, and Yang Zhang. Fupred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics*, 36(12):3749–3757, 2020.

[7] Jude Wells, Alex Hawkins-Hooker, Nicola Bordin, Christine Orengo, and Brooks Paige. Chainsaw: protein domain segmentation with fully convolutional neural networks. *bioRxiv*, pages 2023–07, 2023.

[8] Jing Zhang, R Dustin Schaeffer, Jesse Durham, Qian Cong, and Nick V Grishin. Dpam: A domain parser for alphafold models. *Protein Science*, 32(2):e4548, 2023.

[9] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[10] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14: Proceeding of the 2001 Conference*, pages 849–856, 2001.

[11] Allan D Gordon. Null models in cluster validation. In *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer, 1996.

[12] Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(1–28), 2017.

[13] F. Cazals, D. Mazauric, R. Tetley, and R. Watrigant. Comparing two clusterings using matchings between clusters of clusters. *ACM J. of Experimental Algorithms*, 24(1):1–42, 2019.

[14] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SODA*, page 1035. Society for Industrial and Applied Mathematics, 2007.

[15] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[16] Hua Cheng, R Dustin Schaeffer, Yuxing Liao, Lisa N Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V Grishin. ECOD: an evolutionary classification of protein domains. *PLoS computational biology*, 10(12):e1003926, 2014.

[17] Samuel Flores, Nathaniel Echols, Duncan Milburn, Brandon Hespenheide, Kevin Keating, Jason Lu, Stephen Wells, Eric Z Yu, Michael Thorpe, and Mark Gerstein. The database of macromolecular motions: new features added at the decade mark. *Nucleic acids research*, 34(suppl_1):D296–D301, 2006.

[18] D. Ritchie, A. Ghoorah, L. Mavridis, and V. Venkatraman. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

[19] F. Cazals and R. Tetley. Characterizing molecular flexibility by combining lRMSD measures. *Proteins: structure, function, and bioinformatics*, 87(5):380–389, 2019.

[20] Yuh-Ju Sun, John Rose, Bi-Cheng Wang, and Chwan-Deng Hsiao. The structure of glutamine-binding protein complexed with glutamine at 1.94 å resolution: comparisons with other amino acid binding proteins. *Journal of molecular biology*, 278(1):219–229, 1998.

[21] Rene Jørgensen, Pedro A Ortiz, Anne Carr-Schmid, Poul Nissen, Terri Goss Kinzy, and Gregers Rom Andersen. Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. *Nature Structural & Molecular Biology*, 10(5):379–385, 2003.

[22] RD Schaeffer, LN Kinch, J Pei, KE Medvedev, and NV Grishin. Completeness and consistency in structural domain classifications. *ACS Omega*, 6(24):15698–15707, 2021.

[23] M. Meilă. Comparing clusterings–an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.