

Structural bioinformatics

ResiCon: a method for the identification of dynamic domains, hinges and interfacial regions in proteins

Maciej Dziubiński¹, Paweł Daniluk^{1,2,*} and Bogdan Lesyng^{1,2}

¹Department of Biophysics and CoE BioExploratorium, Faculty of Physics, University of Warsaw, 02-089 Warsaw, Poland and ²Bioinformatics Laboratory, Mossakowski Medical Research Centre, Polish Academy of Sciences, 02-106 Warsaw, Poland

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on April 9, 2015; revised on August 1, 2015; accepted on August 21, 2015

Abstract

Motivation: Structure of most proteins is flexible. Identification and analysis of intramolecular motions is a complex problem. Breaking a structure into relatively rigid parts, the so-called dynamic domains, may help comprehend the complexity of protein's mobility. We propose a new approach called ResiCon (Residue Contacts analysis), which performs this task by applying a data-mining analysis of an ensemble of protein configurations and recognizes dynamic domains, hinges and interfacial regions, by considering contacts between residues.

Results: Dynamic domains found by ResiCon are more compact than those identified by two other popular methods: PiSQRD and GeoStaS. The current analysis was carried out using a known reference set of 30 NMR protein structures, as well as molecular dynamics simulation data of flap opening events in HIV-1 protease. The more detailed analysis of HIV-1 protease dataset shows that ResiCon identified dynamic domains involved in structural changes of functional importance.

Availability and implementation: The ResiCon server is available at URL: <http://dworkowa.imdik.pan.pl/EP/ResiCon>.

Contact: pawel@bioexploratorium.pl

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Proteins are not static. Nuclear magnetic resonance (NMR) spectroscopy (Martin and Zektzer, 1988) and the spin-echo spectroscopy (Bu and Callaway, 2011) experiments show that in several cases it was proven that flexibility may be crucial to protein functionality (Farago *et al.*, 2010; Hamelberg and McCammon, 2005). Although experimental methods provide only general clues about intramolecular motions, molecular dynamics (MD) simulations extend their reach by giving a higher-resolution picture—both in space and time—of protein mobility. By studying an NMR ensemble or MD trajectory one may notice that it is composed of relatively rigid structural parts, often referred to as *dynamic domains* (Hayward *et al.*, 1997).

Domains in traditional sense are regarded as parts of the protein which are: conserved (in terms of evolution), autonomous (in terms of folding) and/or compact (in terms of tertiary structure). Such 'static' domains are identified through sequence homology, structural analysis of a single configuration or both. (For conventional methods of identifying protein domains based on multiple sequences or a single structure see e.g. Bork, 1991; Richardson 1981.) Conversely, dynamic domains depend on structural transitions performed by the protein.

A number of methods for identification of dynamic domains have been developed. The simplest procedures are based on normal mode analysis, which assumes a harmonic approximation of the potential energy function (Hinsen, 1998). More advanced approaches

use the Gaussian Network Model and analyze correlations in motions between the residues of the protein (Yesylevskyi et al., 2006). Many other approaches have also been developed (Bahar et al., 1997; Bernhard et al., 2010; Genoni et al., 2012; Potestio et al., 2009; Wrighers and Schulten, 1997), but all of them anticipate dynamic domains by analyzing a single structure of a protein. Another class of methods for dynamic domains assignment requires exactly two configurations (see Hayward and Berendsen, 1998; Lee et al., 2003; Ye and Godzik, 2003). However, the assumption that two ‘representative’ structures encompass all relevant motions is rather speculative.

Experimental and *in silico* methods reach beyond single-structure representation, and are capable of producing numerous configurations of a given protein. Rather than inferring dynamic domains from one or two structures, a more natural approach would be to determine them based on an ensemble of configurations. GeoStaS is the only method known to us that analyzes a whole ensemble of configurations and assigns each residue to a dynamic domain (Romanowska et al., 2012). Although GeoStaS can analyze not only proteins, but also nucleic acids, it fails to discover dynamic domains whenever they rotate with respect to each other. Alternative methods of analyzing ensembles of configurations assign residues to a static ‘core’ or unstructured bundle (see Kirchner and Guntert, 2011; Snyder and Montelione, 2005).

The purpose of this study was to develop a novel methodology named ResiCon, capable of extracting dynamic domains from an ensemble of protein’s configurations. ResiCon analyzes strengths of contacts between residues based exclusively on geometrical changes occurring in the provided set of structures. The set may be an NMR ensemble of configurations, or snapshots produced in the course of an MD simulation. ResiCon’s main functionality is to identify dynamic domains, but it can also find hinges and interfacial (interdomain) regions.

2 Approach

ResiCon starts with identifying pairs of residues which are *in contact*. There are several definitions of contacts between amino-acid residues in the literature. We used the definition presented in Daniluk and Lesyng (2011) and adapted it to the case when more than one structure is given (see also Daniluk and Lesyng 2014).

Next, ResiCon constructs a virtual scaffold, connecting residues which are in contact with bars. Stiffness of a given bar reflects the estimated strength of the corresponding contact.

Finally, to identify dynamic domains, ResiCon carries out a partitioning (by computing minimal cuts) of the scaffold, cutting weaker and preserving stiffer bars. This partitioning is carried out by applying a spectral clustering algorithm presented in the following section.

The fundamental underlying assumption is that stability of rigid parts results from stable interactions between its residues. However, in our approach we do not analyze physical interactions between residues—they may be hydrophobic, electrostatic or significant in some other way. We simply assume that the measure of strength of a contact between residues is reflected by their geometrical variability across a given sample of structures.

3 Methods

Throughout this article we use terms: *model*, *configuration*, *structure* and *conformation* interchangeably. We will refer to a set of

structures acquired from an MD trajectory or NMR experiment as the *ensemble of configurations* or simply: *an ensemble*. Let us denote the number of structures in an ensemble by S .

3.1 Residue contact

For each pair of residues in every model we compute distances between C_α atoms (d_x) and between geometrical centers of side-chains R_C (d_C) (for glycine $R_C = C_\alpha$, and for alanine $R_C = C_\beta$). We say that two residues are *in contact*, if at least one configuration in the ensemble satisfies the condition:

$$(d_x \leq 6.5 \text{ \AA}) \text{ or } (d_C \leq 8 \text{ \AA} \text{ and } d_x - d_C \geq 0.75 \text{ \AA}) \quad (1)$$

Threshold values are the same as in definition of contact presented in Daniluk and Lesyng (2011) and relate to the range of distances in which physical interactions between residues occur. The second sub-condition favors residues whose side-chains point toward each other (see Fig. 1). Residues that are sequential neighbors are not taken into account.

We assign a quantitative value to the strength of a contact in terms of geometrical variability of the structural part associated with that contact. Such measure is required to capture not only changes in d_x , but also rotational shifts and alterations in the backbone in the vicinity of both residues that are in contact. To do so, we constructed structural parts, comprising sequential neighbors of the two residues in contact. ResiCon assigns a numerical value to the geometric variability by using the least root mean square deviation (RMSD) (Kabsch, 1976). Before elaborating on the details, we proceed with the following definitions.

Elements

An *element* is a structural part of a protein centered around a given residue. It comprises five points, corresponding to the positions of the C_α atoms of the central residue, and its four sequential neighbors (two preceding and two following). For each model s in an ensemble, and each residue i , an element—denoted by E_i^s —is constructed. Residues for which an element cannot be built (e.g. N- and C-termini) are omitted.

Geometrical variability

We consider pairs of elements, $E_{ij}^s := (E_i^s, E_j^s)$, and express structural deviation of a contact between two configurations r and s by RMSD of E_{ij}^r and E_{ij}^s . We use the following function to express the strength of a contact in terms of the whole ensemble:

$$G(i, j) := \max_{\text{pairs of states } (r, s)} \text{RMSD}(E_{ij}^r, E_{ij}^s).$$

The smaller the geometric variability, the stronger the contact.

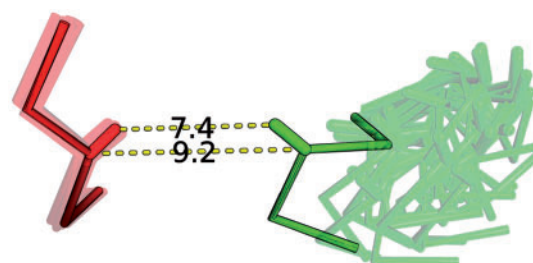


Fig. 1. For each configuration in an ensemble, pairs of elements are constructed. In this figure, residues are in contact because there exists a configuration for which the condition (1) is satisfied (the distance $d_C = 7.4 \text{ \AA}$, and $d_x - d_C = 1.8 \text{ \AA}$)

We tested several statistics based on RMSDs of pairs of elements. This particular definition of G assumes that a strong contact is not ‘broken’ in any pair of models. Conversely, a contact whose structural stability is breached at least once is assumed to be weak and not contributing to the stability of a given dynamic domain.

Note that conformational transitions may be rapid or insufficiently sampled. Therefore, defining geometrical variability in terms of some averaging statistic (e.g. mean, median) might lead to omitting significant structural changes occurring in a protein.

3.2 Contact matrix

We now describe a matrix representation of an edge-weighted graph, in which nodes correspond to residues. Because we used a spectral clustering algorithm which required that weights in the graph were in the interval $[0,1]$, we needed to renormalize the geometrical variability. We calculated the weight between nodes i and j using the following *contact function*:

$$D(i, j) := \begin{cases} 1 & |i - j| \leq 1 \\ L_{\alpha, \beta}(G(i, j)) & \text{residues } i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases}$$

where $L_{\alpha, \beta}$:

$$L_{\alpha, \beta}(x) := (1 + e^{\frac{x-\alpha}{\beta}})^{-1}$$

is a logistic function. We refer to matrix $[D(i, j)]_{ij}$ as the *contact matrix*.

Parameters $\beta > 0$ and $\alpha > 0$ allow for the customization of ResiCon. The logistic function, $L_{\alpha, \beta}$, can be thought of as a rescaling transformation for the measure of geometrical variability, G . It has a simple interpretation—values of G exceeding α are smoothly cut-off, where the degree of ‘smoothness’ is determined by β . All results presented in this article were acquired with default values of α and β :

$$\beta_{\text{default}} := \frac{1}{\sigma(G)} \quad \alpha_{\text{default}} := \mu(G) \quad (2)$$

where the σ and μ stand for standard deviation and mean taken over values of G for all pairs of elements.

Another feature of the logistic function becomes apparent if we consider an ensemble composed of (nearly) identical structures. This has two possible interpretations: the protein is very stable and no conformational changes exist, or that the provided ensemble does not represent such changes. Thus, ResiCon will assume that the contacts are strong—nearly as strong as the peptide bonds ($G \approx 0 \Rightarrow L_{\alpha, \beta} \approx 1$). Consequently, the contact matrix becomes a so-called *contact map*, assigning binary values to pairs of residues (i.e. 1 if a contact occurred at least once, and 0 otherwise).

3.3 Clustering

The contact matrix may be treated as a *similarity matrix*, denoted W , and be used as input in a clustering procedure. Thus, we consider residues to be similar if they are likely to belong to the same dynamic domain. The identified clusters would then correspond to quasi-rigid structural parts.

The choice of a clustering algorithm is not a trivial task and for the identification of dynamic domains two crucial requirements need to be met. First, contact matrices for various proteins vary in dimension and density and the clustering algorithm needs to perform well regardless of these variabilities. Second, the algorithm should facilitate an automated method of choosing the optimal number of clusters.

Agglomerative hierarchical clustering algorithms are one of the most popular approaches to clustering (Han *et al.*, 2006). They follow a greedy scheme to construct a dendrogram encoding distances between clusters. This dendrogram can be cut at a certain height, which determines the number of clusters. If the height parameter could be set so that for all similarity matrices we would obtain high-quality clusters, the hierarchical clustering would have been a good candidate for a clustering procedure. However, as we explain in the [Supplementary Materials](#), estimation of this parameter is difficult, and to determine the number of clusters we would need to extend the conventional hierarchical clustering with a measure of cluster quality.

This was one of the reasons we have chosen a spectral clustering algorithm, which has an inherent indicator of a partitioning’s quality.

Spectral clustering

Clustering algorithms based on finding the eigensystem of the similarity matrix (or more often a matrix derived from it) are termed *spectral algorithms*. They perform a clustering by minimizing the cost of cutting a graph into subgraphs, which agrees with our intuition about finding quasi-rigid parts based on a contact matrix. Optimal clustering is achieved by discarding contacts with the lowest total weight (as few and as weak as possible) to achieve a partitioning into unconnected regions.

In the case of the clustering algorithm used in ResiCon optimal partitioning is decoded from eigenvectors corresponding to the largest eigenvalues of a stochastic matrix $D^{-1}W$, where $D := \text{diag}(d(1), \dots, d(n))$ and $d(i) := \sum_{j=1, j \neq i}^n w(i, j)$ (see Weber *et al.*, 2004). This transformation of the similarity matrix ensures that the identified clusters tend to have similar sizes, which prevents from identifying singular nodes as clusters. Spectral algorithms make no assumptions on the shape of clusters, and, in contrast to the greedy procedures, are insensitive to the ordering of vertices.

In [Figure 2](#), we present a 150×150 similarity matrix with rows and columns ordered in two different ways. The ordering on the right is dictated by the spectral clustering—elements 1–50 were assigned to the first cluster, 51–90 to the second and 91–150 to the third. To find these three clusters we use the first three eigenvectors (corresponding to three largest eigenvalues) of the stochastic matrix $D^{-1}W$. The first eigenvector always has a constant value in all positions and corresponds to a trivial clustering into a single group. The second eigenvector encodes a partitioning into two groups: nodes 51–90 in the first, and the remaining nodes in the second group. The third eigenvector allows to discern the third cluster, composed of nodes 1–50.

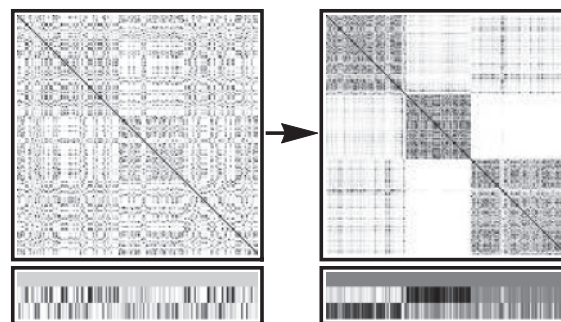


Fig. 2. An example of a similarity matrix W and three eigenvectors of the stochastic matrix $D^{-1}W$ with two different orderings

Clustering algorithm

We used a spectral clustering algorithm in which partitioning of a graph is expressed in terms of a membership matrix χ . A short description of the procedure is presented below, for details refer to Weber et al. (2004).

Let Y denote the matrix containing k eigenvectors of the $D^{-1}W$ stochastic matrix, corresponding to k largest eigenvalues. The procedure computes a linear mapping \mathcal{A} from the eigenvectors Y to the membership matrix:

$$\mathcal{A}Y = \chi$$

The element χ_{ij} of this matrix represents the membership of the i th node in the j th cluster. Therefore, if n is the number of nodes in the graph and k is the number of clusters, then $\chi \in \mathbb{R}^{n \times k}$.

This algorithm has two important features:

- it computes the membership matrix χ which allows for overlapping clusters and
- it offers an indicator, called χ_{\min} , used to determine the optimal number of clusters.

Number of clusters

ResiCon first checks if $k = 1$. To do so, a partitioning into two clusters is carried out. The spectral algorithm presented above finds an optimal cut (Weber et al., 2004) leading to clusters A and B (two sets of indices, that correspond to nodes). We express the cost of such cut by:

$$f := \frac{\sum_{i \in A} \sum_{j \in B} w_{ij}}{(\sum_{k,l \in A} w_{kl})(\sum_{k,l \in B} w_{kl})},$$

where w_{ij} is the weight of the edge between nodes i and j . The validity of a clustering into clusters A and B was checked by asserting that its cost is less than a given threshold. If the criterion was met, we assumed that a clustering into two or more clusters existed. Default threshold for f used to produce all results presented in this study was 0.1. Above this value we observed that compact static proteins were partitioned into short (less than four residues long) segments, which we regarded as improper dynamic domains. On the other hand, lower values resulted in a single dynamic domain assignment in several cases where two domains were apparent.

If $k > 1$, the optimal number of clusters is determined with the use of the indicator presented in Weber et al. (2004). Here, we give a short overview of the properties of χ_{\min} and propose a simple procedure for computing the optimal number of clusters. The indicator $\chi_{\min}(k)$ is defined as the minimal element of the membership matrix χ found by partitioning it into k clusters.

In the case of $k = 2$ the indicator is always zero, $\chi_{\min}(2) = 0$. For $k > 2$ the indicator is less than zero, however if $\chi_{\min}(k) \approx 0$, the clustering into k clusters is the optimal one (Fig. 3). Let us recall the notion of visualizing a clustering by a block-like similarity matrix. Roughly speaking, the value of $\chi_{\min}(k)$ resembles the deviation of the similarity matrix from the ‘pure’ block-like form. However, it is difficult to decide which values of χ_{\min} are sufficiently close to zero to indicate the optimal number of clusters. Therefore, the problem at hand is: does the optimal number of clusters equal two, or more?

In our first approach, the optimal number of clusters was chosen based on a threshold—the optimal k was the one for which χ_{\min} was

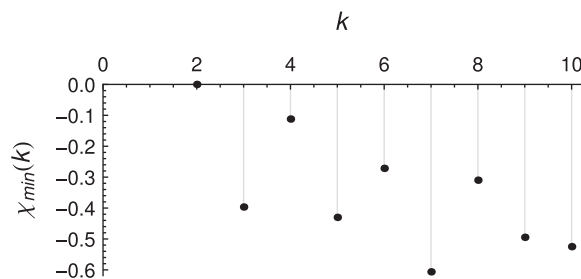


Fig. 3. Values of the χ_{\min} indicator for $k = 2, \dots, 10$, computed for HIV-1 protease. Because $\chi_{\min}(4)$ is closer to 0 than $\chi_{\min}(6)$, the optimal number of clusters is 4

above a certain value. However, it was difficult to find the right threshold because values of χ_{\min} strongly depend on the number of nodes in the graph. Therefore, the following procedure was adapted in ResiCon:

1. Determine the cost of the optimal cut. If it exceeds the 0.1 threshold, the optimal number of clusters is $k = 1$. Otherwise, assume that $k > 1$ and continue the procedure.
2. Compute the values of χ_{\min} for partitionings into 3, 4, \dots , M clusters.
3. Find $k_1 > 2$ for which χ_{\min} is closest to 0, and $k_2 > 2$ for which χ_{\min} is the second closest to 0.
4. If $\chi_{\min}(k_1) > 0.5\chi_{\min}(k_2)$, then the number of clusters is $k = k_1$. Otherwise, $k = 2$.

The 0.5 constant in the fourth point means that we choose k_1 as the number of clusters, if $\chi_{\min}(k_1)$ is closer to 0 than to the next best χ_{\min} (i.e. $\chi_{\min}(k_2)$). The maximal number of clusters M is set to 10 by default, but the user can specify a different number. All results presented in this article were computed with $M = 10$.

In other words, ResiCon chooses $k > 2$ for which the indicator χ_{\min} is ‘relatively close’ to 0. When no such value exists, a partitioning into two clusters is assumed to be optimal.

3.4 Hinges and interfacial regions

Hinges

We define hinges as parts of the structure satisfying both of the following conditions:

1. they do not belong conclusively (in terms of membership as explained below) to any dynamic domain and
2. they are sequentially located between dynamic domains.

The first condition is tested in terms of membership: if a residue membership in any cluster does not exceed certain threshold χ_{hinge} it may belong to a hinge. The default value of the parameter is 0.65, but the user can specify a different value.

Interfacial regions

A residue is assumed to compose an interfacial region if two conditions are met:

1. it does not belong to any hinge and
2. it was in contact with a residue that does not belong to the same dynamic domain at least once.

3.5 Results comparison

According to our knowledge, no expert curated database of dynamic domains exists. Also, we are not aware of any quality measure for

the dynamic domains assignment. Therefore, we compared different methods by analyzing agreement between their results.

We used the measure presented in Meila (2007) called *Variation of Information* (\mathcal{VI}) to analyze the results compatibility. It has the advantage of being a metric in the space of all partitionings of a given dataset. The downside of \mathcal{VI} is that its values do not lie in a fixed interval (e.g. $[0, 1]$), but instead have an upper bound that depends on the size of data. In our case data size equals the number of residues in a given protein. This means that values of \mathcal{VI} for partitionings of one protein are not directly comparable to the values acquired for partitionings of another protein, with a different number of residues. Nonetheless, when considering a particular protein, the \mathcal{VI} metric quantifies the agreement between different assignments of dynamic domains. Here we give an outline of the method, for details see Meila (2007).

Let us denote a clustering by \mathcal{C} . It is composed of clusters—mutually disjoint subsets C_1, \dots, C_k . That is, $\mathcal{C} = \{C_1, \dots, C_k\}$ such that $C_i \cap C_j = \emptyset$ for all pairs i, j . Assume that the numbers of points in consecutive clusters are n_1, \dots, n_k . Then, the probability that a random point from the dataset belongs to the i th cluster is

$$P(i) := \frac{n_i}{n},$$

where n is the number of all points in the set. Note that $\sum_{i=1}^k n_i = n$. Analogously, let another clustering of the same set of points $\mathcal{C}' = \{C'_1, \dots, C'_{k'}\}$ be composed of clusters with $n'_1, \dots, n'_{k'}$ points. By $n_{ij'}$ we will denote the number of points assigned to cluster i in clustering \mathcal{C} and cluster j' in \mathcal{C}' . Then,

$$P(i, j') := \frac{n_{ij'}}{n}$$

is the probability of randomly choosing a point that belongs to both clusters.

The \mathcal{VI} measure is defined in terms of entropy and joint entropy of the probability distributions defined above. That is, if the entropy of clustering \mathcal{C} is expressed by:

$$H(\mathcal{C}) := -\sum_{i=1}^k P(i) \log_2 P(i),$$

and the joint entropy of two clusterings is given by

$$H(\mathcal{C}, \mathcal{C}') := -\sum_{i=1}^k \sum_{j'=1}^{k'} P(i, j') \log_2 P(i, j'),$$

then the variation of information of the two clusterings is defined as

$$\mathcal{VI}(\mathcal{C}, \mathcal{C}') := 2H(\mathcal{C}, \mathcal{C}') - H(\mathcal{C}) - H(\mathcal{C}')$$

3.6 Quality of dynamic domains

We consider dynamic domains to be structural parts of the protein, which move with respect to each other, but remain internally rigid. In order to assess which method for dynamic domains identification is better, a measure was required that would quantify the quality of a given assignment. We did not find such a scoring function in the literature, and propose the following geometrical measure called *total geometrical variability*:

$$Q := \sum_{i=1}^k \max_{\text{pairs of states } (r,s)} \text{RMSD}(D_i^r, D_i^s),$$

where D_i^s is the set of C_α atoms comprising the domain D_i in the state s , and k is the number of domains. Smaller values of Q indicate higher quality.

Note that typically, if a domain is structurally rigid, the maximal RMSD for that domain is smaller than the sum of maximal RMSDs of its two subsets. Therefore, the proposed measure favors large, compact domains. It is also worth noting that for a trivial dynamic domain (single residue) the RMSD is undefined. We set its value to 0, although this artificially reduces Q (see section ‘Quality analysis’).

4 Results and discussion

We compared ResiCon with two recent methods: GeoStaS (Romanowska *et al.*, 2012) and PiSQRD (Potestio *et al.*, 2009). The latter method represents the class of methods which identify dynamic domains by analyzing a single structure. We used a test set comprising 30 NMR-resolved protein structures exhibiting significant mobility, which was previously used in Snyder and Montelione (2005), Kirchner and Guntert (2011) and Romanowska *et al.* (2012). The set was initially proposed to examine the efficacy of a method of identifying structurally stable cores in flexible proteins. These structures often contain a single rigid core with significant geometrical distortions present in peripheral regions (as indicated in Snyder and Montelione, 2005 and also observable in ResiCon’s results).

We have also used ResiCon to analyze a canonical test case—the HIV-1 protease molecule—using an MD trajectory computed with a coarse-grained force field RedMD (Gorecki *et al.* 2007, 2009) as input data. This example shows that ResiCon is capable of finding dynamic domains of a protein whose functionality depends on flexibility and mobility of its rigid parts.

We also explain the so-called *zebra effect*—a peculiar result produced by ResiCon, observed in several cases. This effect is especially strong when a suboptimal number of clusters is chosen.



























































































4.1 Comparative analysis

ResiCon and GeoStaS are both designed to work on an ensemble of structures, and produce a single partitioning into dynamic domains. Both methods impose no assumptions about sampling and order of provided conformations. ResiCon uses maximal local distortions computed over all pairs of frames, indifferent to over- or undersampling of configurations, as long as they are present in the ensemble.

However, the PiSQRD server by default analyzes a single structure and estimates the so-called *low-energy modes*, which are the eigenvectors of the structural covariance matrix (under the canonical ensemble, i.e. assuming Boltzmann distribution of configurations). These low-energy modes are assumed to carry the information relevant to dynamic domains identification. We observed that the choice of an input structure influences the results significantly, but there is no definite criterion for choosing the *right* structure for the analysis. The PiSQRD server provided with a PDB file containing NMR models by default finds dynamic domains for the first model. This might introduce a bias unfavorable for PiSQRD’s performance. We, therefore, decided to examine results produced by PiSQRD for every structure in the ensemble in order to compare ResiCon against its full capacity. The best dynamic domains were chosen and presented together with the results produced by ResiCon and GeoStaS in Table 1.

PiSQRD also gives a possibility of providing a set of low-energy modes extracted from a structural covariance matrix estimated from a set of structures, but this procedure is not straightforward and requires additional assumptions (see [Supplementary Materials](#)). We scrutinize the quality of dynamic domains found by PiSQRD from user-provided low-energy modes in the section ‘Quality analysis’.

Table 1. Summary of results produced by GeoStaS, ResiCon and PiSQRD. Dynamic domains shown for PiSQRD are those for which the lowest value of Q was achieved

PDB code	Method	Clustering	Q	PDB code	Method	Clustering	Q
2rgf	GeoStaS		1.430	2k3c	GeoStaS		5.678
	ResiCon		1.275		ResiCon		3.676
	PiSQRD		1.421		PiSQRD		3.422
2pas	GeoStaS		4.926	1cfc	GeoStaS		12.658
	ResiCon		1.398		ResiCon		3.780
	PiSQRD		6.141		PiSQRD		2.294
1aey	GeoStaS		1.520	1a67	GeoStaS		4.478
	ResiCon		1.520		ResiCon		4.548
	PiSQRD		4.455		PiSQRD		5.887
1pkt	GeoStaS		1.670	3egf	GeoStaS		5.700
	ResiCon		1.670		ResiCon		5.713
	PiSQRD		5.147		PiSQRD		3.499
4a5v	GeoStaS		0.678	2pni	GeoStaS		16.195
	ResiCon		1.698		ResiCon		5.994
	PiSQRD		2.561		PiSQRD		6.993
1pit	GeoStaS		2.177	1zda	GeoStaS		9.610
	ResiCon		2.209		ResiCon		6.549
	PiSQRD		2.171		PiSQRD		6.798
2vil	GeoStaS		6.329	1adr	GeoStaS		6.241
	ResiCon		2.414		ResiCon		7.417
	PiSQRD		4.146		PiSQRD		7.139
1aiw	GeoStaS		8.066	1yug	GeoStaS		8.479
	ResiCon		2.636		ResiCon		8.230
	PiSQRD		4.527		PiSQRD		9.033
2ktf	GeoStaS		2.822	1d1d	GeoStaS		19.862
	ResiCon		2.693		ResiCon		8.368
	PiSQRD		0.636		PiSQRD		7.313
1vve	GeoStaS		3.299	2114	GeoStaS		10.284
	ResiCon		2.767		ResiCon		8.525
	PiSQRD		2.767		PiSQRD		10.411
3mef	GeoStaS		3.770	1bf8	GeoStaS		7.947
	ResiCon		3.086		ResiCon		9.153
	PiSQRD		4.765		PiSQRD		4.519
1vvd	GeoStaS		5.439	2htg	GeoStaS		10.514
	ResiCon		3.110		ResiCon		10.514
	PiSQRD		2.821		PiSQRD		8.665
2spz	GeoStaS		3.244	1qo6	GeoStaS		11.403
	ResiCon		3.212		ResiCon		11.057
	PiSQRD		8.071		PiSQRD		7.985
11eb	GeoStaS		6.885	2k0e	GeoStaS		12.771
	ResiCon		3.234		ResiCon		11.604
	PiSQRD		5.874		PiSQRD		7.927
2ait	GeoStaS		5.370	2kr6	GeoStaS		56.578
	ResiCon		3.310		ResiCon		54.160
	PiSQRD		3.871		PiSQRD		26.834

Because of a large number of models, it was not possible to provide a graphical representation for each partitioning. We therefore focused on values of the agreement measure \mathcal{VI} between results produced by the three methods.

To familiarize the reader with values of the measure \mathcal{VI} , we take a look at two most distant partitionings produced by PiSQRD (denoted by P1 and P2) for an exemplary 1d1d protein, and how they relate to the results given by ResiCon (R) and GeoStaS (G) (Fig. 4). The clustering P2 seems to be very similar to the one produced by ResiCon, while P1 gives a sliced-and-diced picture of the protein's mobility. It seems that a partitioning produced by PiSQRD depends on physical properties embedded in a particular configuration. Although NMR models carrying information about





	R	G	P1	P2
R		0	1.47	3.29
G		1.47	0	3.76
P1		3.29	3.76	0
P2		0.21	1.51	3.29

Fig. 4. Values of \mathcal{VI} for partitionings of the 1d1d protein molecule. A pair of results produced by PiSQRD which had the highest \mathcal{VI} are denoted by P1 and P2

dynamic domains may exist, others lead to chaotic partitionings. In order to produce a single clustering, PiSQRD would need a procedure for interpreting physical properties based on an ensemble of configurations.

Table 2. Discrepancies in assignments expressed by radii of balls encompassing 25, 50 and 75% of results

	ResiCon vs. GeoStaS $\mathcal{V}\mathcal{I}$	ResiCon vs. PiSQRD			GeoStaS vs. PiSQRD			PiSQRD vs. PiSQRD		
		$r_{25\%}$	$r_{50\%}$	$r_{75\%}$	$r_{25\%}$	$r_{50\%}$	$r_{75\%}$	$\langle r_{25\%} \rangle$	$\langle r_{50\%} \rangle$	$\langle r_{75\%} \rangle$
2rgf	0.62	0.45	0.48	0.57	0.26	0.26	0.32	0.01	0.14	0.20
2pas	1.94	2.43	2.45	2.47	2.00	2.08	2.11	0.61	0.96	1.23
1aey	0.00	1.99	2.33	2.37	1.99	2.33	2.37	0.77	1.10	1.52
1pkt	0.00	2.07	2.24	2.36	2.07	2.24	2.36	1.38	1.63	1.84
4a5v	1.00	1.08	1.12	1.19	1.53	1.54	1.60	0.24	0.29	0.38
1pit	1.59	1.76	2.05	2.14	2.42	2.55	2.74	1.35	1.70	1.94
2vil	1.89	1.19	1.39	1.89	2.31	2.49	2.67	1.11	1.36	1.57
1aiw	1.64	1.66	1.80	2.13	2.58	2.69	2.77	1.30	1.53	1.90
2ktf	0.80	1.69	2.01	2.06	2.15	2.36	2.48	0.93	1.17	1.74
1vve	1.05	0.00	0.00	0.00	1.05	1.05	1.05	0.62	0.62	0.62
3mef	0.32	1.30	1.43	1.50	1.21	1.32	1.40	0.64	0.86	1.09
1vvd	1.24	0.21	0.21	0.21	1.26	1.26	1.26	0.31	0.31	0.31
2spz	0.13	1.42	1.62	1.94	1.37	1.59	1.96	1.28	1.60	1.73
1leb	0.61	2.05	2.27	2.41	2.33	2.37	2.45	0.92	1.21	1.52
2ait	0.71	1.26	1.65	1.82	1.49	1.67	1.99	0.99	1.31	1.52
2k3c	1.86	0.82	0.97	0.98	1.61	1.69	1.69	0.16	0.42	0.63
1cfc	1.60	0.41	0.44	0.97	1.41	1.48	1.86	0.77	0.92	1.30
1a67	1.07	1.90	1.96	2.06	2.22	2.38	2.69	1.58	1.93	2.12
3egf	0.14	1.30	1.40	1.43	1.29	1.41	1.45	0.63	0.91	1.14
2pni	1.60	1.16	1.58	1.78	2.00	2.25	2.35	1.18	1.44	1.65
1zda	0.79	1.37	1.55	1.96	0.83	1.28	1.78	1.26	1.54	1.75
1adr	0.84	0.80	1.30	1.81	1.11	1.64	2.22	1.19	1.57	1.94
1yug	1.79	0.90	1.00	1.31	1.98	2.12	2.15	0.70	0.99	1.29
1d1d	1.47	0.00	0.21	3.15	1.47	1.51	3.80	1.08	3.08	3.17
2l14	0.55	0.49	0.57	0.75	0.91	0.98	1.18	0.61	0.71	0.95
1bf8	1.61	0.74	1.27	3.22	1.69	1.89	4.20	1.13	1.74	2.99
2htg	0.00	1.36	1.58	1.91	1.36	1.58	1.91	0.84	1.13	1.30
1qo6	0.88	0.63	0.72	0.88	0.87	0.98	1.18	0.57	0.71	0.94
2k0e	0.97	0.59	0.66	0.82	1.05	1.16	1.31	0.72	0.87	1.03
2kr6	1.11	0.79	0.81	0.84	0.85	0.87	0.93	0.61	0.68	0.82

The ordering of the results is the same as in Figure 5, i.e. best-scoring results are first.

Nearly 50% of the assignments produced by PiSQRD for 1d1d are coherent with the clustering given by ResiCon (see [Supplementary Materials](#)). On the other hand, the result given by GeoStaS differs from all results produced by PiSQRD. In fact, ResiCon and PiSQRD are more coherent than GeoStaS and PiSQRD, which can be expressed by the mean of $\mathcal{V}\mathcal{I}$. However, it would be naive to use the mean value as an indicator of self-consistency of PiSQRD. Among partitionings produced by PiSQRD, $N(N - 1)/2$ agreements were calculated (where N is the number of models of 1d1d), whereas the comparison of ResiCon or GeoStaS with PiSQRD gave N values of $\mathcal{V}\mathcal{I}$. Consequently, the average value of $\mathcal{V}\mathcal{I}$ between PiSQRD clusterings is not directly comparable with the average value of $\mathcal{V}\mathcal{I}$, e.g. for ResiCon versus PiSQRD.

Therefore, to give a better picture of the divergence of results we exploit the fact that $\mathcal{V}\mathcal{I}$ is a metric in the space of all clusterings. For a given partitioning we computed the radius of a ball centered at that partitioning, encompassing a certain fraction of the results. In [Table 2](#) we provide values of radii of balls which encompass 25, 50 and 75% of the results. In the case of PiSQRD, we constructed balls (for a given percentage of results) centered at each partitioning, and computed the mean values of their radii. The mean of radii does not carry the bias mentioned earlier. In addition, histograms of $\mathcal{V}\mathcal{I}$ for each protein can be found in [Supplementary Materials](#).

4.2 Quality analysis

The box-and-whisker plot in [Figure 5](#) provides a concise picture of quality scores Q (see Section ‘Methods’) of the dynamic domains assignments. It clearly shows that beyond a few exceptions ResiCon gives the best results. There are seven notable exceptions: 4a5v, 2ktf, 3egf, 1adr, 1bf8, and 2kr6.

The single dynamic domains found by PiSQRD (blue in [Fig. 5](#)) were produced using low-energy modes, computed as eigenvectors of a structural covariance matrix estimated by superimposing all models in an ensemble on a representative structure. In the [Supplementary Materials](#), we explain how this representative structure is chosen, and show qualities of dynamic domains found by PiSQRD using different methods of estimating the structural covariance matrix. It should be emphasized, however, that these dynamic domains strongly depend on the method of estimating the structural covariance matrix, which is not part of PiSQRD’s functionality. Therefore, these results should only be treated as an additional insight into what the user can expect from a more complex analysis of NMR structures.

In the case of 1adr GeoStaS gave partitionings with lower Q , by introducing a trivial domain (cutting off C- and N-termini—see [Supplementary Materials](#)). On the other hand, for 4a5v GeoStaS identified a single domain, which gave a lower value of Q than any other partitioning.

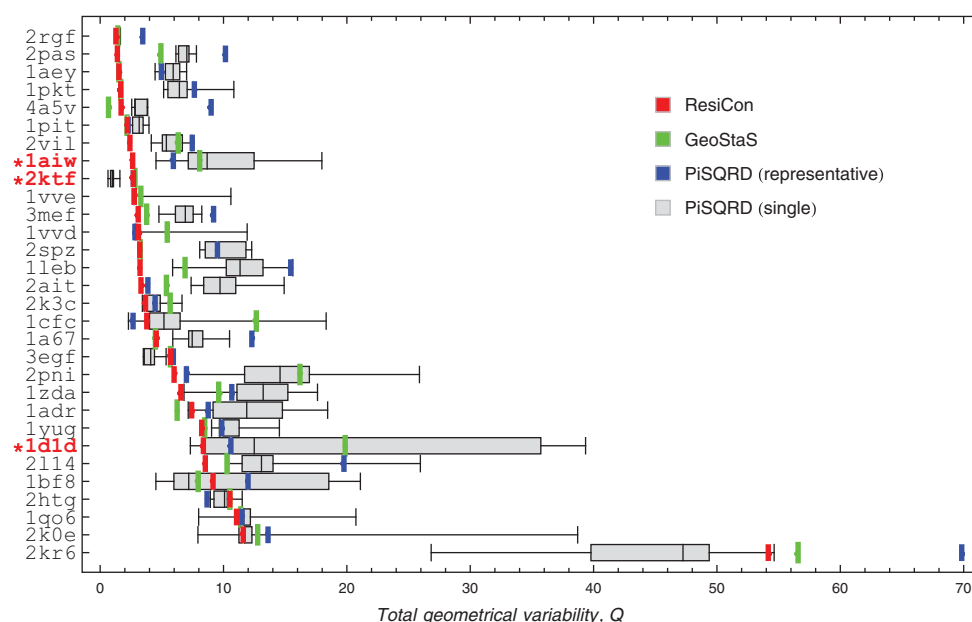


Fig. 5. Box and whiskers plot of the dynamic domains quality score Q for ResiCon, GeoStaS, PiSQRD. In blue are Q values for PiSQRD's dynamic domains determined from structural covariance matrices (see [Supplementary Materials](#))

For proteins 2ktf, 3egf and 2htg, PiSQRD produced results with lower Q than ResiCon, by finding numerous small quasi-rigid fragments. However, for 2ktf >25% of assignments found by PiSQRD contained a trivial, single-residue domain (indicated in red, with an asterisk). Although our measure does not penalize for this, we consider such behavior undesirable. Also, note that proteins 3egf and 2htg comprise 53 and 27 residues accordingly. It seems that in case of small proteins ResiCon often identifies a single domain, which does not necessarily result in the lowest Q .

The 2kr6 protein is an interesting example. It contains a flexible linker composed of >30 residues. As a consequence, partitionings with high values of Q are observed. Only PiSQRD was able to produce lower values of Q , by cutting the linker into many shorter parts. The example of the 2kr6 protein shows that ResiCon does not consider unstructured regions to be separate dynamic domains. Instead, residues constituting linkers and lacking long-distance contacts are assigned to dynamic domains which are closest in sequence.

4.3 Comments

Although PiSQRD's capability of analyzing a single structure may be considered an advantage, results presented in [Table 2](#) and [Figure 5](#) show that there is a large discrepancy for different configurations of the same protein. Nevertheless, based on the histograms of \mathcal{V} presented in [Supplementary Materials](#) and the values of radii in [Table 2](#), we conclude that in most cases results given by ResiCon and PiSQRD are mutually more coherent, than GeoStaS and PiSQRD (e.g. 1cfc, 1qo6, 1vvd, 1vve, 1yug, 2k3c). Notable exceptions are: 2rgf, 2pas, 3mef and 1zda. For these proteins ResiCon did not find any significant structural transitions and achieved the lowest value of Q (see [Fig. 5](#)) by assigning a single domain.

Dynamic domains found by ResiCon are generally larger (particularly: 2k3c, 1d1d, 2k0e and 2kr6). Conversely, GeoStaS often allocates flexible N- and C-terminal parts (e.g. 1adr, 1qo6, 1vve, 3egf) as quasi-rigid parts. The size of dynamic domains is also the main difference between ResiCon and PiSQRD. In case of small, static proteins (such as 1aey, 1pkt, 1pit, 2spz, 2ait, 3egf and 1zda), PiSQRD identifies numerous small and

often trivial dynamic domains. ResiCon on the other hand detects no significant conformational changes (by analyzing an ensemble), and assigns a single dynamic domain. We observe that in many cases ResiCon identified a single domain which had the lowest value of Q among all partitionings (see 2rgf, 2pas, 1aey, 1pkt, 1pit, 2vil, 1aiw, 3mef, 2spz, 1leb, 2ait, 2pni, 2114). In these cases ResiCon correctly detected that no significant transitions were present in the protein. Therefore, unlike PiSQRD and GeoStaS, ResiCon can reliably indicate whether conformational changes occur in an ensemble of structures.

It is also noteworthy that ResiCon does not employ any post-processing procedures. This keeps the algorithm simple and clean, but results in a discontinuity of certain partitionings (see 4a5v, 1bf8 and 1vvd), which we refer to as the *zebra effect*. Although this seems to be an artifact of the clustering algorithm, we will take a closer look at this effect and show that it may also carry valuable information referring to the protein's dynamics.

4.4 HIV-1 protease

An analysis of an MD trajectory of the HIV-1 protease showcases ResiCon's capabilities. This protein undergoes substantial conformational changes associated with opening/closing of its structural parts, so-called *flaps* ([Hamelberg and McCammon, 2005](#)). A database of X-ray-resolved structures, representing configurations which the protease can attain, is available ([Vondrasek and Wlodawer, 2002](#)).

We examined a set of configurations of the HIV-1 protease acquired from a simulation carried out using the RedMD package ([Gorecki et al., 2009](#)). This coarse-grained force field was designed to simulate intramolecular motions in proteins and nucleic acids. It has correctly predicted the flap-opening motion in the HIV-1 protease, which is known biological fact ([Hamelberg and McCammon, 2005](#)), and was independently confirmed by an all-atom MD simulation ([Sadiq and De Fabritiis, 2010](#)). Roughly, ~1% of the trajectory seems to exhibit significant conformational transitions (flap opening events). This is a typical scenario in MD simulations, where a transition between meta-stable states is swift and fairly short. We needed a set of representative structures in order to find

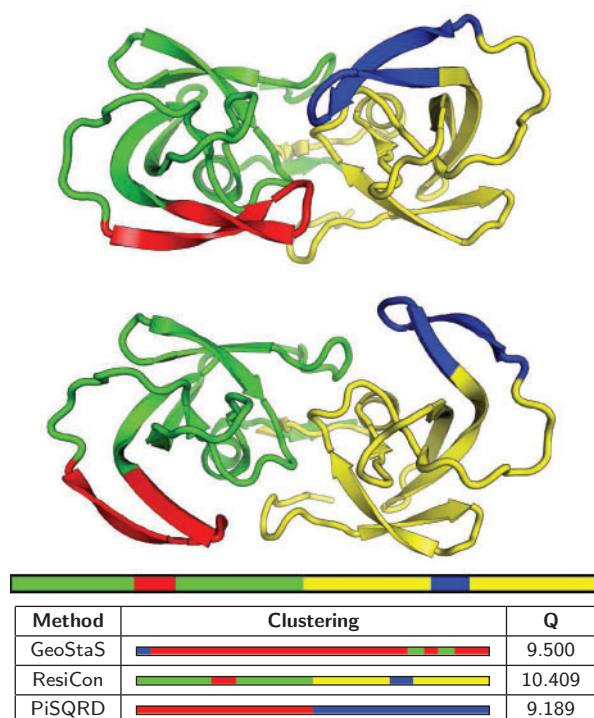


Fig. 6. Two representative conformations of HIV-1 protease with flaps being closed and open, and summary of results produced by GeoStaS, ResiCon and PiSQRD. Structures are colored according to the dynamic domains detected by ResiCon

dynamic domains using ResiCon. From the whole trajectory, we extracted a set of 200 configurations using a generic procedure based on the Principal Component Analysis, implemented in the R programming language (in the `bio3d` package—Grant *et al.*, 2006)—see [Supplementary Materials](#).

Values of χ_{\min} for different numbers of clusters are given in [Figure 3](#). The optimal number of clusters according to our procedure is 4. [Figure 6](#) depicts the dynamic domains found by ResiCon and two representative configurations of the protease, as well as results from GeoStaS and PiSQRD. Because the sample of configurations was drawn according to the Boltzmann distribution, we were able to straightforwardly estimate the structural covariance matrix and provide PiSQRD with well-founded low-energy modes, and acquire a high-quality partitioning into two domains. Results produced by GeoStaS were acquired from the whole trajectory of the protease.

Dynamic domains identified by ResiCon have the highest value of the Q measure. Note that ResiCon does not try to minimize Q , but to produce a clustering with the optimal value of the χ_{\min} indicator. Consequently, at the cost of a slightly higher Q (ca. 1 Å) we arrive at a partitioning which corresponds to the biologically relevant sub-division of the protein. It is also worth mentioning that ResiCon's partitioning into $k=2$ clusters incidentally leads to an identical assignment as the one produced by PiSQRD.

Motions of the flaps between the closed and open states are crucial in the functionality of the HIV-1 protease. It is also known that throughout their motion they remain quasi-rigid (Freedberg *et al.*, 2002). Therefore, using ResiCon, we successfully extracted a simplified picture of the mobility of HIV-1 protease, which agrees with experimental knowledge.

Alternative partitioning into six clusters (see [Fig. 7](#)) also deserves interest. Apart from the flaps, additional dynamic domains

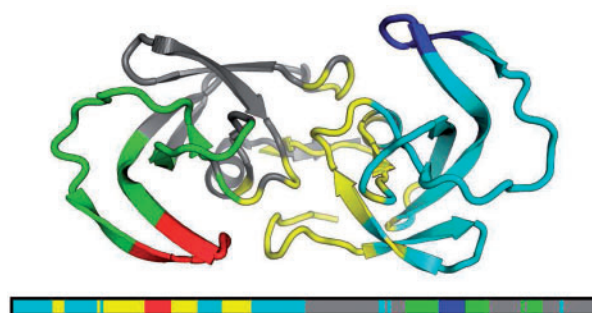


Fig. 7. More subtle division emerges when the number of clusters is set to 6.

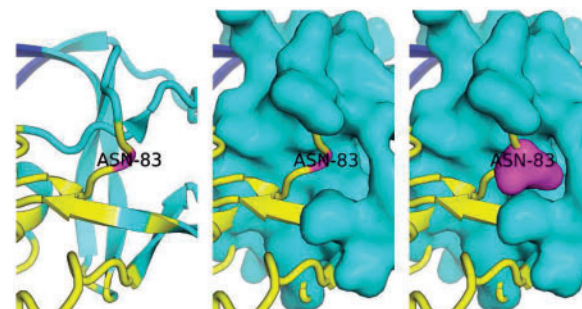


Fig. 8. From the analysis carried out by ResiCon one can see that throughout the trajectory ASN-83 not only remains close to residues in the N-terminal arm domain but also fills a cavity in that dynamic domain

resembling ‘arms’ carrying the flaps can be observed. Note that conversely to the case of the four dynamic domains, the quasi-static regions are similar, but not ideally reflective. This indicates that motions of the two centrally symmetric sub-units of the HIV-1 protease in the provided trajectory were slightly different.

It can be also seen that domains are discontinuous. Especially, residues belonging to the N-terminal lobe (yellow) and to the arm (cyan) are interleaved. In the following section we will analyze this effect, and show that this partitioning also carries valuable information.

4.5 The zebra effect

Dynamic domains found by ResiCon may include residues that seem to be pulled out from another quasi-rigid fragment. This is indicated as discontinuities (stripes) in [Figure 7](#). Let us take a look at an example of such an extracted residue and try to understand the source of this effect. ResiCon assigned the ASN-83 residue to the arm (gray) dynamic domain (see [Fig. 8](#)). However, its sequential neighbors belong to the lobe domain (yellow). The reason of this discontinuity is that, although ASN-83 has peptide bonds with its yellow neighbors, they are outweighed by contacts with the gray residues (see [Fig. 8](#)). The discontinuity suggests that throughout the trajectory ASN-83 remained docked in the cavity of the arm domain, while its peptide bonds formed an axis of a hinge.

It seems that the zebra effect is not accidental, and that it can be used to find residues which act as ‘pivot points’. However, this effect is volatile and depends on small fluctuations in the input. Residues are assigned to dynamic domains based on the fuzzy membership matrix. At certain positions values of membership to different domains may be almost equal, and when discretized cause emergence of stripes. Therefore, to strengthen the identification of these ‘pivot point’ residues, a more thorough analysis of the membership matrix is required.

5 Conclusions

Typically, protein structures are flexible and mobile, and their conformational changes may be crucial in facilitating signaling and metabolic processes in which they participate. Breaking a structure into dynamic domains may be compared with discovering gears and levers connected by cogs and pegs analogous to those found in classical machines. Such analyses allow us to better understand molecular mechanisms responsible for biological functions (e.g. Taylor *et al.*, 2013), facilitate MD simulations in large time-scales with simplified forcefields (e.g. Sinititskiy *et al.*, 2012), or discover potential binding sites when designing inhibitors (e.g. Zhang *et al.*, 2009).

We have presented a universal tool for discovering dynamic domains in proteins. ResiCon is capable of analyzing a single structure, or an NMR ensemble of structures provided in a PDB format. It can also be applied to the set of independently obtained structures (e.g. X-ray crystallographic structures obtained under different conditions). In any case, it provides a complete set of results comprising partitioning of the molecule into dynamic domains with additional highlighting of residues composing hinges and interfacial regions. ResiCon also provides an indicator of partitioning quality, and suggests the optimal number of domains.

We tested our method using the reference set of NMR structures. It is comparable or better than the recently developed GeoStaS and PiSQRD. Apart from giving more compact dynamic domains, it is also capable of distinguishing structures composed of a single quasi-rigid region.

We have made ResiCon available online (<http://www.dworakowa.imdik.pan.pl/EP/ResiCon>). To make the analysis feasible and limit the number of uploads, queries may be rerun with changed parameters. Also, all queries and results are stored on our server. (These two features require registration for security reasons.)

Acknowledgements

We thank Joanna Trylska and Adam Górecki for providing MD trajectories of the HIV-1 protease.

Funding

These studies were supported by the Biocentrum-Ochota project (POIG.02.03.00-00-003/09), the research grant (DEC-2011/03/D/NZ2/02004) of the National Science Centre of Poland and research funds of the Faculty of Physics, University of Warsaw (BST-170000/BF/34).

Conflict of Interests: none declared.

References

Bahar, I. *et al.* (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173–181.

Bernhard, S. and Noe, F. (2010) Optimal identification of semi-rigid domains in macromolecules from molecular dynamics simulation. *PLoS one*, **5**, e10491.

Bork, P. (1991) Shuffled domains in extracellular proteins. *FEBS Lett.*, **286**, 47–54.

Bu, Z. and Callaway, D.J.E. (2011) Proteins move! Protein dynamics and long-range allostery in cell signaling. *Adv. Protein Chem. Struct. Biol.*, **83**, 163–221.

Daniluk, P. and Lesyng, B. (2011) A novel method to compare protein structures using local descriptors. *BMC Bioinformatics*, **12**, 344.

Daniluk, P. and Lesyng, B. (2014) *Theoretical and Computational Aspects of Protein Structural Alignment*. In: *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*. Springer, Berlin, pp. 557–598.

Farago, B. *et al.* (2010) Activation of nanoscale allosteric protein domain motion revealed by neutron spin echo spectroscopy. *Biophys. J.* **99**, 3473–3482.

Freedberg, D.I. *et al.* (2002) Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci.*, **11**, 221–232.

Genoni, A. *et al.* (2012) Identification of domains in protein structures from the analysis of intramolecular interactions. *J. Phys. Chem. B*, **116**, 3331–3343.

Gorecki, A. *et al.* (2007) Causality and correlation analyses of molecular dynamics simulation data. *Comput. Biophys. Syst. Biol.*, **36**, 25–30.

Gorecki, A. *et al.* (2009) RedMD-reduced molecular dynamics package. *J. Comput. Chem.*, **30**, 2364–2373.

Grant, B.J. *et al.* (2006) Bio3D: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.

Hamelberg, D. and McCammon, J.A. (2005) Fast peptidyl cis-trans Isomerization within the Flexible Gly-Rich Flaps of HIV-1 Protease. *J. Am. Chem. Soc.*, **127**, 13778–13779.

Han, J. *et al.* (2006) *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan Kaufmann, Burlington, MA.

Hayward, S. and Berendsen, H.J.C. (1998) Systematic analysis of domain motions in proteins from conformational change; new results on citrate synthase and T4 lysozyme. *Proteins*, **30**, 144–154.

Hayward, S. *et al.* (1997) Modelfree methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins*, **27**, 425–437.

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

Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **32**, 922–923.

Kirchner, D.K. and Guntert, P. (2011) Objective identification of residue ranges for the superposition of protein structures. *BMC Bioinformatics*, **12**, 170.

Lee, R. *et al.* (2003) The DynDom database of protein domain motions. *Bioinformatics*, **19**, 1290–1291.

Martin, G.E. and Zektzer, A.S. (1988) *Two-dimensional NMR Methods for Establishing Molecular Connectivity*. Wiley-VCH.

Meila, M. (2007) Comparing clusterings—an information based distance. *J. Multivariate Anal.*, **98**, 873–895.

Potestio, R. *et al.* (2009) Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.*, **96**, 4993–5002.

Richardson, J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.

Romanowska, J. *et al.* (2012) Determining geometrically stable domains in molecular conformation sets. *J. Chem. Theory Comput.*, **8**, 2588–2599.

Sadiq, S.K. and De Fabritiis, G. (2010) Explicit solvent dynamics and energetics of HIV1 protease flap opening and closing. *Proteins*, **78**, 2873–2885.

Sinititskiy, A.V. *et al.* (2012) Optimal number of coarse-grained sites in different components of large biomolecular complexes. *J. Phys. Chem. B*, **116**, 8363–8374.

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

Taylor, D. *et al.* (2013) Classification of domain movements in proteins using dynamic contact graphs. *PLoS One*, **8**, e81224.

Vondrasek, J. and Wlodawer, A. (2002) HIVdb: a database of the structures of human immunodeficiency virus protease. *Proteins*, **49**, 429–431.

Weber, M. *et al.* (2004) Perron cluster analysis and its connection to graph partitioning for noisy data. *Konrad-Zuse-Zentrum für Informationstechnik Berlin*, 1–20.

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

Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(suppl 2), ii246–ii255.

Yesylevskyy, S.O. *et al.* (2006) Dynamic protein domains: identification, interdependence, and stability. *Biophys. J.*, **91**, 670–685.

Zhang, Z. *et al.* (2009) Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys. J.*, **97**, 2327–2337.