



**HAL**  
open science

# Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions

Alessandro Genoni, Giulia Morra, Giorgio Colombo

► **To cite this version:**

Alessandro Genoni, Giulia Morra, Giorgio Colombo. Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *Journal of Physical Chemistry B*, 2012, 116 (10), pp.3331-3343. 10.1021/jp210568a . hal-02196462

**HAL Id: hal-02196462**

**<https://hal.univ-lorraine.fr/hal-02196462>**

Submitted on 10 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This document is the Accepted Manuscript version of a Published Work that appeared in final form in *The Journal of Physical Chemistry B*, copyright © American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see <https://pubs.acs.org/doi/10.1021/jp210568a>.

1  
2  
3 **Identification of Domains in Protein Structures from the**  
4  
5 **Analysis of Intramolecular Interactions**  
6  
7  
8  
9

10  
11  
12 Alessandro Genoni <sup>(1,2)\*</sup>, Giulia Morra <sup>(1)</sup>, Giorgio Colombo <sup>(1)\*</sup>  
13  
14

15 (1) Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco  
16  
17 9, 20131 Milano, Italy.  
18  
19

20 (2) Equipe de Chimie et Biochimie Théoriques, UMR 7565 SRSMC – CNRS,  
21  
22 Université de Lorraine, Boulevard des Aiguillettes BP 70239, 54506  
23  
24 Vandoeuvre-lès-Nancy Cedex, France.  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42

43  
44 \_\_\_\_\_  
45 \* Correspondence to:

46 Giorgio Colombo, Istituto di Chimica del Riconoscimento Molecolare – CNR, Via  
47 Mario Bianco 9, 20131 Milano, Italy. E-mail: [giorgio.colombo@icrm.cnr.it](mailto:giorgio.colombo@icrm.cnr.it). Phone:  
48  
49 +39-02-28500031; Fax: +39-02-28901239;  
50  
51

52  
53 Alessandro Genoni, Equipe de Chimie et Biochimie Théoriques, UMR 7565 SRSMC  
54  
55 – CNRS, Université de Lorraine, Boulevard des Aiguillettes BP 70239, 54506  
56  
57 Vandoeuvre-lès-Nancy Cedex, France. E-mail: [Alessandro.Genoni@cbt.uhp-nancy.fr](mailto:Alessandro.Genoni@cbt.uhp-nancy.fr).  
58  
59 Phone: +33-0(3) 83 68 43 77; Fax: +33-(0)3 83 68 43 71.  
60

**Abstract**

The subdivision of protein structures into smaller and independent structural domains has a fundamental importance in understanding protein evolution and function, in the development of protein classification methods as well as in the interpretation of experimental data. Due to the rapid growth in the number of solved protein structures, the need of devising new accurate algorithmic methods has become more and more urgent. In this paper, we propose a new computational approach that is based on the concept of domain as a compact and independent folding unit and on the analysis of the residue-residue energy interactions obtainable through classical all-atom force field calculations. In particular, starting from the analysis of the non-bonded interaction energy matrix associated with a protein, our method filters out and selects only those specific subsets of interactions that define possible independent folding nuclei within a complex protein structure. This allows grouping different protein fragments into energy clusters that are found to correspond to structural domains. The strategy has been tested using proper benchmark datasets and the results have shown that the new approach is fast and reliable in determining the number of domains in a totally *ab initio* manner and without making use of any training set or knowledge of the systems in exam. Moreover, our method, identifying the most relevant residues for the stabilization of each domain, may complement the results given by other classification techniques and may provide useful information to design and guide new experiments.

**Key Words:** molecular simulations, force-field calculations, protein domain, protein stability, independent folding unit, non-bonded interactions.

## Introduction

The partitioning of complex proteins into simpler structural components, known as domains, is of paramount importance since this has become a crucial step in protein classification and in functional and evolutionary studies. The concept of structural protein domains dates back to the paper published by Wetlaufer more than thirty years ago, where the domains have been considered for the first time not only as separate regions of the molecule but also as fundamental units for the protein architecture and as independent nucleation sites in protein folding<sup>1</sup>.

Nevertheless, it must be noted that, so far, it has been impossible to give a unique definition of structural domain. Indeed, to accomplish this task, different criteria may be used: compactness, structural stability, presence of a hydrophobic core, independent folding, occurrence in combinations with a variety of other domains and presence of function. Therefore, in this context, several computational strategies to identify domains in proteins have appeared and each method gives prominence to one or more of the concepts listed above. This may sometimes lead to dissenting domain assignments that affect the consequent classification of protein structures<sup>2</sup>. Furthermore, in light of the observations reported above, it is important to observe that it is not possible to decide *a priori* which is the best computational technique and relying only on the prediction of a single algorithm can be misleading. A possibility to overcome this drawback consists in building a consensus assignment based on the outputs provided by different strategies, as recently proposed by Bourne and coworkers<sup>3</sup>.

The methods for the decomposition of proteins into structural domains are usually subdivided into two main groups: the ones that require human expertise for the evaluation of each structure (from now on, *manual strategies*) and the ones that are

1  
2  
3 completely automated (from now on, *algorithmic techniques*). It is worthwhile to note  
4 that, although each method for domains identification is necessarily “constrained” by  
5 the domain definition that it tries to fulfill, new strategies continue to appear at a  
6 constant pace and this is particularly true for the latter category. In fact, as the number  
7 of solved protein structures exponentially increases, the manual techniques become a  
8 bottleneck in keeping the domain databases up to date and, therefore, reliable and  
9 fully automated methods are highly required.

10  
11 Among the strategies based on the human expertise, SCOP<sup>4</sup> and CATH<sup>5</sup> are the most  
12 relevant and popular. The former is completely manual and mainly uses evolutionary  
13 and structural relationships to define domains, while the latter is a combination of  
14 algorithmic and manual procedures. Indeed, it consists in using three different  
15 automated decomposition programs, but, at the final step, an expert decides to accept  
16 or reject the possible consensus reached among the algorithmic strategies and, if  
17 necessary (namely, in absence of consensus or if the consensus is rejected), assigns  
18 the domains manually. Finally, in this group of methods, it is also possible to include  
19 the technique recently proposed by Majumdar *et al.*<sup>6</sup> who have constructed a database  
20 of manually defined domains for a set of topologically complex proteins taking into  
21 account structural, functional, sequence and evolutionary aspects after a careful study  
22 of relevant literature and analysis of domains obtained by means of the existing  
23 automated methods.

24  
25 Considering the algorithmic strategies, they are mainly compactness-based  
26 approaches that rely on protein structural features (e.g., the contacts density or the C<sub>α</sub>-  
27 C<sub>α</sub> distance maps) and whose underlying principle consists in the fact that, under a  
28 correct domain assignment, the number of inter-domain interactions is lower than the  
29 intra-domain interactions<sup>7</sup>. In this context we can consider the Domain Parser method  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 where a graphic-theoretic approach is used to perform the domain decomposition<sup>8, 9</sup>.  
4  
5 In particular, after associating each residue with a node of a connected network and  
6  
7 all the significant residue-residue contacts with an edge characterized by a capacity  
8  
9 value that depends on the type of the interaction between the two involved  
10  
11 aminoacids, this technique aims at dividing the network into two or more connected  
12  
13 parts in such a way that the total edge capacities across the divisions are minimized<sup>8</sup>.  
14  
15 Furthermore, in order to improve the accuracy of the decompositions when there is  
16  
17 not an exact *a priori* knowledge of the number of domains, the previous method has  
18  
19 been afterwards improved including the hydrophobic moment profile and neural  
20  
21 networks to check the quality of the identified domains<sup>9</sup>. Along these lines, the PDP  
22  
23 (Protein Domain Parser) and DDomain approaches are based on the assumptions that  
24  
25 each structural domain is associated with a set of continuous protein fragments and  
26  
27 identify the domains using a normalized domain-domain contact-interaction profile<sup>10</sup>.  
28  
29 <sup>11</sup>. Other two interesting techniques that exploit the residue-residue contacts as basic  
30  
31 tool to identify structural domains in proteins are PUU<sup>12</sup> and Domak<sup>13</sup>. The former is  
32  
33 a fast algorithm that is able to deal with multiple-segment domains and that tries to  
34  
35 determine potential folding units through an eigenvalue analysis of the contacts map,  
36  
37 while the latter, starting from the already mentioned fundamental principle that a  
38  
39 domain has more residue-residue contacts within than without, is a strategy that  
40  
41 introduces secondary structure information and other factors in order to improve the  
42  
43 agreement with a training set of protein structures with well-established domains  
44  
45 definitions. In the spirit of the techniques briefly described above, Dodis<sup>14</sup> takes into  
46  
47 account only the C<sub>α</sub> atoms and the matrix of the C<sub>α</sub>-C<sub>α</sub> distances is clustered using an  
48  
49 agglomerative hierarchical procedure that separates the C<sub>α</sub> atoms into distinct  
50  
51 domains. It is worthwhile to observe that this method does not aim at providing the  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 best partition, but it produces many possible decompositions of a protein structure  
4  
5 into one or more domains. Finally, in the framework of the compactness-based  
6  
7 strategies, it is interesting to mention the approach proposed by Taylor<sup>15</sup>, a technique  
8  
9 that basically consists in an Ising model and in which the protein residues (namely,  
10  
11 the structural elements of the model) change their state (in this case, a multi-value  
12  
13 label) in function of the state of the close residues (i.e., the state of the neighbors).  
14  
15

16  
17 All the methods described so far are based on the concept of compactness, but they  
18  
19 ignore recurrence criteria, namely the recognition that many identified domains are  
20  
21 common to different proteins at a high level of statistical significance. Therefore, in  
22  
23 order to overcome this drawback, Holm and Sander have proposed DALI<sup>16</sup>, an  
24  
25 automated method that is able to identify recurrent domains in proteins computing the  
26  
27 recurrence in terms of statistical significance of structural similarity for many pairs of  
28  
29 substructures.  
30  
31

32  
33 A completely different way to automatically identify structural domains is based on  
34  
35 the observation that the internal dynamics of proteins can be accurately described in  
36  
37 terms of movements of approximately rigid subunits. In this context, examples are the  
38  
39 translation-libration-screwlike (TLS) motion analysis developed by Schomaker and  
40  
41 Trueblood<sup>17</sup>, the DynDom technique introduced by Hayward *et al.*<sup>18, 19</sup>, the strategy  
42  
43 devised by Hinsen and coworkers<sup>20, 21</sup>, where clusters of aminoacids are identified as  
44  
45 quasi-rigid blocks taking into account only low-frequency modes of fluctuations, and  
46  
47 the approach proposed by Yesylevskyy *et al.*<sup>22</sup>, which perform a hierarchical  
48  
49 clustering of the correlation patterns. Furthermore, it is worthwhile to mention the  
50  
51 method recently introduced by Micheletti and coworkers that, considering only the  
52  
53 essential dynamic spaces of a protein, cluster the aminoacids minimizing a  
54  
55 phenomenological strain-energy function by means of a variational scheme, so  
56  
57  
58  
59  
60



1  
2  
3 ensuring that the fluctuations within each single group are as small as possible  
4  
5 compared to those across different groups<sup>23</sup>.  
6

7  
8 Finally, for the sake of completeness, it is important to consider energy-based  
9  
10 techniques and, in particular, the one developed by Berezovsky and coworkers<sup>24, 25</sup>.  
11  
12 Relying on the importance of the van der Waals forces in the folding process, this  
13  
14 approach allows obtaining an energy hierarchy of domains for globular proteins  
15  
16 simply analyzing van der Waals energy profiles.  
17

18  
19 In this paper, mainly taking into account the definition of domain as a compact and  
20  
21 independent folding unit, we propose an alternative energy-based strategy for  
22  
23 identifying structural domains and subunits. In particular, extending the recently  
24  
25 developed Energy Decomposition Method<sup>26-30</sup>, we compute the non-bonded  
26  
27 interaction energy matrix of the structure to be examined. After diagonalizing this  
28  
29 matrix, we select an optimal subset of its eigenvectors that allow identifying the most  
30  
31 stabilizing residues and the essential non-bonded interactions that define each folding  
32  
33 unit. The optimally selected eigenvectors enable to reconstruct a filtered non-bonded  
34  
35 interaction energy matrix that is characterized by a significant quenching of the inter-  
36  
37 domain interactions and that is afterwards subject to a clustering procedure to identify  
38  
39 the correspondences between the interaction energy clusters and the structural  
40  
41 domains of the protein in exam. In other words, our new method detects the most  
42  
43 relevant residue-residue interactions that implicitly define the nuclei which are  
44  
45 necessary to stabilize the three-dimensional fold of the different domains (or  
46  
47 subunits). Therefore, in analogy with the previously adopted terminology<sup>26-30</sup>, the  
48  
49 groups of fundamental pair-interactions define the folding nuclei of each domain.  
50  
51 This approach has been implemented in a computer program and it has been tested  
52  
53 using the comprehensive benchmark datasets assembled by Bourne and coworkers  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 (i.e., Benchamrk\_2 and Benchamrk\_3) to assess the capabilities of novel or already  
4 existing algorithmic methods<sup>31</sup>. The new technique has proved to be accurate in  
5 determining the exact number of domains and to provide information on the domains  
6 boundaries. It is worth observing that, unlike most of the automated approaches, the  
7 new strategy is not trained using a set of proteins with domains assigned by experts  
8 and, above all, it is a completely *ab initio* technique, namely, it relies only on the  
9 physico-chemical analysis of the energetic properties of the different protein  
10 fragments and no preliminary information is given as input. In this respect, the  
11 proposed method may be useful in the domain partitioning of novel folds or novel  
12 structures for which no previous information is available. Moreover, our method is  
13 deeply connected to the energy-properties of protein folding and, therefore, it  
14 provides further and different information compared to the analysis of contacts  
15 density maps or the analysis of structural and geometrical properties. In particular, we  
16 envisage that the energy-based domain identification proposed here may be a useful  
17 complement to the findings of other strategies and, for its nature, it may be used to  
18 provide experimentalists with immediately testable hypotheses (see the discussion of  
19 the results) on domains limits, boundaries, stability and isolation. Finally, as it will be  
20 discussed below, careful analysis of specific cases indicate that there is room for  
21 significantly improving the technique.  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

## 50 Theory

51  
52 *Underlying philosophy and general scheme.* In the Energy Decomposition Method  
53 (EDM) recently developed by Colombo and coworkers<sup>26-30</sup>, attention is focused on  
54 the protein non-bonded interaction energy matrix  $\mathbf{E}^{nb}$  whose elements represent the  
55 non-bonded (namely, van der Waals and electrostatic) interaction energies between  
56  
57  
58  
59  
60

1  
2  
3 residues. According to the EDM approach, the eigenvector ( $\mathbf{v}_1$ , also indicated as “first  
4 eigenvector”) associated with the lowest eigenvalue ( $\lambda_1$ ) of  $\mathbf{E}^{nb}$  allows to identify  
5 most of the crucial aminoacids necessary for the stabilization of the protein fold<sup>26-30</sup>.  
6  
7  
8  
9  
10  
11 In fact, using that eigenvector, it is possible to obtain a filtered non-bonded interaction  
12 energy matrix ( $\mathbf{E}^{fold}$ ) that accounts for many of the essential residue-residue  
13 interactions responsible for the protein folding, namely we have:  
14  
15  
16

$$\mathbf{E}^{fold} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T \quad (1)$$

17  
18  
19  
20  
21 where  $\mathbf{v}_1^T$  is the transpose of the vector  $\mathbf{v}_1$ . This approximation usually holds for one-  
22 domain proteins, but in more complicated cases, such as multi-domain proteins, it has  
23 been observed that more eigenvectors are generally needed to completely represent all  
24 the interactions that are fundamental for the protein folding/stability. To better explain  
25 the reason why only one eigenvector is not sufficient for multi-domain proteins, in  
26 Figure 1 we have shown the eigenvector associated with the lowest eigenvalue of  $\mathbf{E}^{nb}$   
27 for proteins constituted by one, two, three and four domains. It is easy to observe that,  
28 except for the one-domain protein case (see Figure 1A), where almost all the  
29 components of the first eigenvector are significant, in all the other situations many  
30 components of this eigenvector are equal to zero (see Figures 1B, 1C and 1D) and,  
31 therefore, the eigenvector associated with the lowest eigenvalue does not contain by  
32 itself sufficient information in regards to the stabilization energy of the whole protein.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
In order to overcome this drawback, in this paper we propose a new strategy based on  
the following working hypotheses (see also Figure 2A): each domain may represent  
an independent folding unit and it should be autonomously stable if excised from the  
original protein. Therefore, in analogy with the case of one-domain proteins, we  
postulate that: (a) for each domain there should exist only one associated eigenvector  
recapitulating the most significant interactions of the domain, in the same way as the

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

eigenvector associated with the lowest eigenvalue recapitulates the most significant protein interactions for single-domain proteins; (b) each “domain eigenvector” should have a block structure, namely the significant components of each “domain eigenvector” should correspond to the residues belonging to the domain; (c) the union of all the significant blocks should cover all the protein residues. In an ideal case the previous hypotheses would be satisfied and, after selecting the proper set of eigenvectors, the “filtered” non-bonded interaction energy matrix would be simply obtained as

$$\mathbf{E}^{fold} = \sum_{i=1}^{N_D} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \quad (2),$$

with  $N_D$  as the number of protein domains, and it would have a block structure that enables to immediately identify the protein subunits. In fact, in a situation like the one represented in Fig. 2B, the matrix  $\mathbf{E}^{fold}$  (from now on, also referred to as “essential folding matrix”) completely neglects the inter-domain interactions whereas each of its blocks represents the essential folding interactions of the corresponding domain.

In a real situation, the eigenvectors of the non-bonded interaction energy matrix  $\mathbf{E}^{nb}$  are characterized by overlapping blocks of significant components (see Fig. 2C) and more than one eigenvector for each domain is generally needed. In order to satisfy as much as possible the working hypotheses introduced above, it is necessary to select the smallest set of eigenvectors that cover the largest part of residues (i.e., components) with the minimum redundancy. Hence, the corresponding essential folding matrix is obtained as

$$\mathbf{E}^{fold} \approx \mathbf{E}_{approx}^{fold} = \sum_{i=1}^{N_e} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \quad (3),$$

1  
2  
3 where  $N_e$ , namely the number of essential eigenvectors, is generally greater than  $N_D$ .

4  
5  
6 The matrix  $\mathbf{E}^{fold}$  is afterwards further filtered through a symbolization process to  
7  
8  
9 emphasize the significant non-bonded interactions and, finally, it is subject to a proper  
10  
11 clustering procedure leading to the domains identification.

12  
13 At this point it is worth noting an interesting parallelism. As the well-known essential  
14  
15 dynamics strategy<sup>32, 33</sup> selects the most suitable eigenvectors of the atomic  
16  
17 displacement covariance matrix to identify the correlated motions of large protein  
18  
19 subunits, which can be also defined as dynamic domains<sup>23</sup>, our novel method detects  
20  
21 proper eigenvectors of the non-bonded interaction energy matrix to extract the  
22  
23 fundamental interactions within each folding unit, so allowing another possible  
24  
25 definition for the structural domains. It is due to this analogy that the selected  
26  
27 eigenvectors and the matrix  $\mathbf{E}^{fold}$  have been previously referred to as essential  
28  
29 eigenvectors and essential folding matrix, respectively.

30  
31  
32 Summarizing, the object of our analysis is a matrix whose relevant elements  
33  
34  
35 correspond to the essential components of the protein non-bonded interaction energy  
36  
37 matrix, components that allow to identify the most important residues in defining and  
38  
39 stabilizing the three dimensional organization of the different sequences. In turn, this  
40  
41 gives insights into the determinants of the structural stability of the detected domains.

42  
43 Before concluding, it is important to note that in this subsection we have described the  
44  
45 general philosophy and scheme of our new strategy. All the details about the selection  
46  
47 of the essential eigenvectors and about the symbolization of the essential folding  
48  
49 matrix will be given in the following two subsections, while the clustering procedure  
50  
51 will be examined in the “Methods” section.  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

*Selection of the eigenvectors.* As explained above, each essential eigenvector should be characterized by a block of significant components that generally correspond to the residues of the associated domain (as already mentioned, we have exact correspondence only in an ideal case). Therefore, a threshold to discriminate the significant components from the non-significant ones is crucial for the selection of the proper eigenvectors and, in our new approach, it has been chosen as the median of the distribution of the absolute values of all the eigenvectors components. In other words, a component is considered significant if and only if its absolute value is greater than that threshold (see Figure 3A).

After determining the significant components for all the eigenvectors of  $\mathbf{E}^{nb}$ , it is possible to start the selection. The first pick is always the eigenvector associated with the lowest eigenvalue, while the other ones are selected bearing in mind that it is necessary to identify the smallest set of eigenvectors that cover the largest part of the residues with the minimum redundancy. To accomplish this task, we always look for the eigenvector whose significant components overlap as little as possible with the set of residues significantly covered by the already chosen eigenvectors (see Figure 3B) and, if two or more eigenvectors are characterized by the same overlap value, the one associated with the lower eigenvalue is considered. For a complete understanding of this selection procedure, let us consider a simple example where  $\mathbf{e}_1$  and  $\mathbf{e}_2$  constitute a set of already selected eigenvectors. For the sake of simplicity, we indicate the significant components with 1 and the non-significant components with 0:

$$\mathbf{e}_1 = [0, 0, 0, 0, 1, 1, 0, 1]$$

$$\mathbf{e}_2 = [1, 0, 0, 0, 1, 1, 0, 0]$$

It is easy to observe that we have four significantly covered residues: the residues 1, 5, 6 and 8. Now, let us imagine to pick the third eigenvector from this set:

$$\mathbf{a} = [1, 0, 0, 0, 1, 0, 1, 1]$$

$$\mathbf{b} = [1, 1, 1, 1, 0, 0, 0, 0]$$

$$\mathbf{c} = [1, 1, 0, 0, 1, 0, 1, 0]$$

The significant components of  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  respectively overlap with three, one and two of the residues that are already significantly covered by  $\mathbf{e}_1$  and  $\mathbf{e}_2$ . Hence, to guarantee the minimum redundancy, eigenvector  $\mathbf{b}$  will be selected.

Finally, it is important to note that an eigenvector is subject to two further check-tests to be definitely accepted. The first one computes the amount of information that the new eigenvector introduces, namely the number of new significant components divided by the total number of residues. If this quantity is lower than or equal to 0.01 the eigenvector is discarded, otherwise is run through the second test that checks if the redundancy threshold is exceeded. In particular, if at least 50% of residues are significantly covered at least three times (i.e., by at least three selected eigenvectors), the eigenvector in exam is rejected and the selection stops.

*Construction of the essential folding matrix and its symbolization.* Exploiting equation (3), the set of selected eigenvectors is used to construct the essential folding matrix (see Figure 3C). However, in order to highlight the really essential non-bonded interactions, it is necessary to further filter the resulting matrix and this goal can be accomplished considering the “flat threshold-matrix”  $\mathbf{F}$ , namely a matrix with identical elements given by the following equation:

$$\mathbf{F}_{ij} = f = \frac{\lambda_1}{N} \quad (4),$$

where  $N$  is the total number of residues and  $\lambda_1$  is the lowest eigenvalue of the non-bonded interaction energy matrix. This matrix corresponds to the case in which all the non-bonded interactions are identical and its “flat element”  $f$  can be used as threshold

to extract the significant non-bonded interactions in  $\mathbf{E}^{fold}$ . Therefore, the essential folding matrix is transformed into the symbolized essential folding matrix  $\mathbf{E}^{symb}$  (see Figure 3D) following this simple rule:

$$\mathbf{E}_{ij}^{symb} = \begin{cases} 1 & \text{if } |\mathbf{E}_{ij}^{fold}| > |f| \\ 0 & \text{if } |\mathbf{E}_{ij}^{fold}| \leq |f| \end{cases} \quad (5)$$

For the sake of completeness, it is important to observe that the “flat threshold-matrix” is obtained considering the following normalized “flat eigenvector”:

$$\mathbf{e} = \left[ \frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}} \right] \quad (6)$$

From a physical point of view, this vector represents a situation in which each residue equally contributes to the structural stabilization of the protein and this is the reason why its “flat component” has been successfully used in the Energy Decomposition Method<sup>26-30</sup> to identify the crucial aminoacids for the protein fold. Therefore, it is obvious that the “flat eigenvector” can be also exploited to detect the really significant non-bonded interactions in the essential folding matrix  $\mathbf{E}^{fold}$  and, in particular, in our new approach it is associated with the lowest eigenvalue ( $\lambda_1$ ) of  $\mathbf{E}^{nb}$  to obtain the matrix  $\mathbf{F}$  of identical elements given by equation (4).

After the symbolization, exploiting the concept that for a domain the number of significant intra-domain non-bonded interactions is greater than the number of significant inter-domain non-bonded interactions,  $\mathbf{E}^{symb}$  is subject to a clustering procedure to identify the possible protein domains (see Figure 3E and Figure 3F). Considering this last step, it is worthwhile to point out that the essence of our novel strategy consists in the proper selection of the eigenvectors and in the construction/symbolization of the essential folding matrix. The clustering step is important to detect the domains of the protein in exam, but after obtaining the matrix



1  
2  
3  $\mathbf{E}^{symb}$ , different algorithms can be used and, therefore, it is not a distinctive aspect of  
4  
5  
6 the proposed technique. In the preliminary version of the program, whose results will  
7  
8 be discussed in one of the next paragraphs, we have devised a non-optimized in-house  
9  
10 clustering algorithm that will be briefly described in the “Methods” section.  
11  
12 Finally, for the sake of clarity, it is important to observe that our symbolized essential  
13  
14 folding matrix  $\mathbf{E}^{symb}$  is completely different from a classical protein contacts matrix  
15  
16 (see Figure 4). In particular, if on the one hand it is true that the non-bonded energy  
17  
18 interactions drop off quite fast with distance, on the other hand it is easy to observe  
19  
20 that they survive for a longer range compared to the simple residue-residue contacts  
21  
22 (see Figure 4). This indicates that  $\mathbf{E}^{symb}$  does not only contain information about the  
23  
24 density of interactions, but it also comprises important information about the  
25  
26 interactions intensity and physico-chemical properties. Therefore, in this respect, our  
27  
28 analysis provides complementary information that can enrich the one obtained by  
29  
30 means of classical structure-based techniques.  
31  
32  
33  
34  
35  
36  
37  
38  
39

## 40 **Methods**

41  
42 The method described in the Theory section has been implemented in a computer  
43  
44 program using the FORTRAN77 language and it has been afterwards tested  
45  
46 considering the benchmark datasets recently constructed by Bourne and coworkers to  
47  
48 assess the capabilities of algorithmic techniques<sup>31</sup>.  
49  
50  
51  
52  
53

54 *The clustering algorithm.* As already mentioned, the clustering of the symbolized  
55  
56 essential folding matrix  $\mathbf{E}^{symb}$  has been performed exploiting a non-optimized in-  
57  
58 house algorithm. This technique consists in five main steps and relies on the density  
59  
60 of significant non-bonded interactions, from now on simply referred to as density. In

1  
2  
3 the first step, all the possible small (user-defined in the input file) diagonal blocks in  
4  
5  $E^{symb}$  are examined and if their density is greater than or equal to the overall density  
6  
7  
8 (namely, the density of the whole matrix), they are taken into account (Figure 5A and  
9  
10 Figure 5B). The second step consists in grouping the maximum number of  
11  
12 consecutive diagonal blocks previously selected. To accomplish this task a starting  
13  
14 diagonal block must be considered and the interaction density with its consecutive  
15  
16 diagonal block (if it exists) is calculated (Figure 5C). If the interaction density is  
17  
18 greater than or equal to the overall density, the two consecutive *tesserae* are merged  
19  
20 in a temporary cluster (Figure 5D), otherwise the second diagonal block becomes the  
21  
22 new starting unit. If a temporary diagonal cluster is obtained, the procedure is  
23  
24 repeated, but now considering the interaction density between the current diagonal  
25  
26 cluster and its possible consecutive diagonal block (Figure 5E). Therefore, at the end  
27  
28 of the second step, diagonal clusters are obtained (Figure 5F), while the third step  
29  
30 considers the possibility to merge them even if they are not consecutive (Figure 6A  
31  
32 and Figure 6B). In particular, two generic clusters X and Y are merged if and only if  
33  
34 the following three requirements are satisfied: *i*) the interaction density between X  
35  
36 and Y is greater than or equal to the overall density; *ii*) X is the cluster that interacts  
37  
38 most with Y (i.e., the interaction density between X and Y is greater than the  
39  
40 interaction densities between any other cluster and Y); *iii*) Y is the cluster that  
41  
42 interacts most with X. If two or more couples satisfy the previous conditions, we  
43  
44 consider the couple with the highest interaction density. After each new merging the  
45  
46 previous analysis is performed again and the procedure stops only when there is no  
47  
48 other couples to be joined. In the fourth step, the clusters with dimension smaller than  
49  
50 the minimum required size (user-defined in the input file) are discarded (Figure 6C  
51  
52 and Figure 6D). Finally, in case of multi-fragment clusters, the gaps between the  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 fragments are analyzed (Figure 6E) and, if these gaps do not overlap with other  
4 clusters, they are merged with the multi-fragment cluster in exam (Figure 6F).  
5  
6 Therefore, through the five steps just described, it is possible to obtain a first-level  
7 subdivision in domains, but it is important to note that each first-level domain is  
8 afterwards subject to the same five-step procedure with the only difference that the  
9 overall interaction density of the domain in exam replaces the overall interaction  
10 density of the whole protein. Of course, the clustering algorithm stops when all the  $n$ -  
11 level domains cannot be further subdivided into  $(n+1)$ -level substructures. Finally, it  
12 is worthwhile to observe that, since it is sometimes difficult to automatically decide  
13 which is the best partition among different possibilities, following the philosophy  
14 already adopted by Carugo<sup>14</sup> and by Berozovsky *et al.*<sup>24, 25</sup>, the implemented  
15 algorithm provides all the levels of domain subdivision and let the user choose the  
16 proper level of resolution according to the particular circumstances.  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33

34  
35  
36 *Preparation of the Bourne benchmark datasets.* The Bourne benchmarks considered  
37 in this paper are Benchmark\_2 and Benchmark\_3<sup>31</sup>. The former, which contains 315  
38 chains and which has been assembled considering the consensus among three  
39 different expert methods in assigning the number of domains, is one of the most  
40 comprehensive consensus-based dataset of multi-domain proteins currently available  
41 to evaluate new or already existing automated strategies. The latter consists of 271  
42 structures and it has been obtained removing from Benchmark\_2 44 chains for which  
43 the expert methods disagree in assigning the domain boundaries. Half of each  
44 benchmark dataset is available on-line (<http://pdomains.sdsc.edu>) and they have been  
45 downloaded to perform the preliminary tests. Furthermore, 14 of the obtained  
46 structures have been discarded from Benchmark\_2 (11 in the case of Benchmark\_3)  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 mainly for two reasons: sequences of missing residues longer than 25 aminoacids or  
4 the presence of non-conventional residues in the crystallographic structures (see Table  
5 S1 in the Supporting Information for the complete list of discarded chains).  
6  
7 Furthermore, in case of gaps shorter than or consisting of 25 residues, the missing  
8 aminoacids have been added by means of the following protocol: a) the sequences of  
9 missing residues have been designed manually by means of the molecular modeling  
10 environment Maestro<sup>34</sup> and they have been afterwards modeled as loops using  
11 MODELLER<sup>35-38</sup>; b) keeping frozen all the atoms non-belonging to the loops with the  
12 exception of the C<sub>α</sub> atoms of the boundary residues (which have been constrained to  
13 their positions with force constants set equal to 100.0 kJ·mol<sup>-1</sup>·Å<sup>-2</sup>), the protein  
14 structure has been minimized through the package MacroModel<sup>39</sup> in implicit solvent  
15 (water), using the AMBER\* force field and the default optimization parameters  
16 (conjugate gradient algorithm, van der Waals and electrostatic cutoffs set equal to 8.0  
17 Å and 20.0 Å, respectively). Furthermore, during the optimization, the peptide bond  
18 dihedral angles of the modeled loops have been constrained to 180° (also in these  
19 cases the force constants have been set equal to 100.0 kJ·mol<sup>-1</sup>·Å<sup>-2</sup>).

20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44 *Calculation of the non-bonded interaction energy matrices.* In order to relax the  
45 starting crystal structures and any possible modeled loops, we have optimized the  
46 geometry of all the proteins in exam exploiting the AMBER9 package<sup>40</sup>. In particular,  
47 the ff03 force field and the steepest descent algorithm have been used and,  
48 furthermore, the solvation effects have been taken into account by means of the  
49 Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method.  
50  
51 Therefore, the generic elements  $E_{ij}^{nb}$  of the protein non-bonded interaction energy  
52 matrices have been computed as the sum of three different terms:  
53  
54  
55  
56  
57  
58  
59  
60

$$\mathbf{E}_{ij}^{nb} = \mathbf{E}_{ij}^{el} + \mathbf{E}_{ij}^{vdW} + \mathbf{G}_{ij}^{solv} \quad (7),$$

with  $\mathbf{E}_{ij}^{el}$ ,  $\mathbf{E}_{ij}^{vdW}$  and  $\mathbf{G}_{ij}^{solv}$  as the electrostatic, van der Waals and solvent contributions, respectively. The non-bonded interaction energy matrices have been afterwards diagonalized and the obtained eigenvalues and eigenvectors have been used for the domains identification as described in the Theory section.

## Results and Discussion.

Exploiting the program that implements the strategy introduced in this paper, we have determined the domains associated with all the examined structures (i.e., 143 chains for Benchamrk\_2 and 124 chains for Benchamk\_3) using on the average 3.8, 5.0, 5.6 and 6.0 eigenvectors when one, two, three and four domains have been detected, respectively. It could be surprising that we have generally selected more than one eigenvector also in the case of one-domain proteins, but this is simply due to the adopted algorithm that, as seen above, completes the selection of the eigenvectors when at least 50% of residues are significantly covered at least three times (see the “Selection of the eigenvectors” subsection). This criterion has been chosen as a comprehensive consensus to deal with multi-domain chains in order to identify the smallest set of eigenvectors that cover the largest part of the aminoacids with the minimum redundancy. Of course, this introduces a non-necessary redundancy when one-domain proteins are examined, but this does not generally invalidate the results. Afterwards, we have considered several criteria to assess the capabilities of the proposed strategy against the benchmark data (see Table S2 in the Supporting Information for the list of identified domains and the number of selected eigenvectors for all the examined structures). In particular, we have analyzed: a) the performance of the new technique in assigning the exact number of domains (of course, “exact”

1  
2  
3 with respect to the consensus of expert methods); b) the fragmentation of domains; c)  
4  
5 the boundary consistency, namely the measure of the overlap between the consensus  
6  
7 domain boundaries detected by expert methods and the domain boundaries  
8  
9 determined by our new approach.  
10  
11

12 About the first criterion, it is important to note that the errors can be classified as  
13  
14 over-cuts (i.e., assigning more domains than the benchmark) and under-cuts (i.e.  
15  
16 assigning fewer domains than the benchmark). Considering the available chains of  
17  
18 Benchmark\_2 as reference, we can observe (see Table 1) that the new method shows  
19  
20 a very good accuracy (about 75%) in assigning the correct number of domains and a  
21  
22 tendency to under-cut a few structures (17.5% of under-cuts). Furthermore, in order to  
23  
24 fully characterize the capability of the new strategy in correctly determining the  
25  
26 number of domains, Benchmark\_2 has been afterwards subdivided into four subsets  
27  
28 containing structures constituted by one, two, three and at least four domains,  
29  
30 respectively. It is worth to note that we decided to group together the benchmark  
31  
32 chains characterized by four, five and six domains in order to have a representative  
33  
34 sample, even if the total number of structures in the obtained set still remains low  
35  
36 (namely, six structures overall). As it can be seen in Table 1, the accuracy is quite  
37  
38 high for one- and two-domain chains, but, as the number of domains in the examined  
39  
40 structures rises, a significant number of under-cuts is observed. These results have  
41  
42 urged us on investigating the reasons why the proposed technique has the tendency to  
43  
44 underestimate the number of domains. One possible cause is the adopted clustering  
45  
46 algorithm that is not able to properly “subdivide” the obtained symbolic essential  
47  
48 folding matrix, as it can be seen, for instance, in the cases depicted in Figure 7.  
49  
50 However, it is important to note that this drawback can be simply overcome  
51  
52 improving the current non-optimized in-house clustering procedure or using an  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 already optimized algorithm, such as one of those associated with the other automated  
4 techniques for the domains identification. Another plausible reason for the under-  
5 cutting trend is connected to the specific physico-chemical bases of our approach that  
6  
7  
8  
9  
10 aims to identify all the possible significant non-bonded interactions. Therefore, if two  
11  
12  
13 (or more) domains significantly interact (especially in the frequent cases of compact  
14  
15  
16 protein structures or of compact interfaces between substructures that could be  
17  
18 manually assigned to different domains), a single block is found to characterize the  
19  
20 symbolic essential folding matrix and it becomes difficult to separate all the clusters  
21  
22 corresponding to structural domains (see the examples represented in Figure 8).  
23  
24 However, in these cases, our method indicates that certain interfaces take part in  
25  
26 fundamental interactions and, hence, the proposed strategy might also provide  
27  
28 additional information about the role of the interfaces in the structural stabilization of  
29  
30 the protein. In fact, in these circumstances, the interface residues are responsible for a  
31  
32 relevant amount of the stabilization energy necessary to define the specific 3D  
33  
34 arrangement of the protein. Hence, it is conceivable to hypothesize that even if two  
35  
36 protein fragments that share a large interface with extensive contacts have been  
37  
38 assigned to different domains according to a certain set of rules, it may be possible  
39  
40 that they are properly folded and structured *only* in the presence of those contacts.  
41  
42 These situations are reminiscent of the case of the “obligate dimers”, which have been  
43  
44 properly studied in the context of protein-protein interactions<sup>41</sup>. In this case, our  
45  
46 strategy, while providing different results in the assignment and partitioning  
47  
48 compared to methods based on different criteria, gives direct insights into the  
49  
50 importance of interface interactions in the stabilization of the 3D organization of a  
51  
52 multi-domain protein. Such information may be used, for instance, to guide site-  
53  
54 directed mutagenesis experiments aimed at disrupting or reinforcing specific interface  
55  
56  
57  
58  
59  
60

1  
2  
3 interactions in order to reveal their relative importance in domain and protein  
4 stabilization.  
5  
6

7  
8 In Table 2 we have shown the performances of faster automated techniques (Domain  
9 Parser<sup>8, 9</sup>, PDP<sup>10</sup> and PUU<sup>12</sup>) evaluated by Bourne and coworkers using the complete  
10 Benchmark\_2 as reference<sup>31</sup>. For the sake of completeness, it is important to stress  
11 that Bourne and coworkers have obtained these results considering the complete  
12 Benchmark\_2 and not its downloadable half that has been considered for our test  
13 calculations (see the “Methods” section). This is one of the reasons why we do not  
14 have a representative sample for four-, five- and six-domain proteins and, therefore,  
15 we have limited our comparisons with the other algorithmic approaches up to the  
16 three-domain case. From Tables 1 and 2 it is easy to observe that PDP is by far the  
17 strategy that overall provides the best agreement with the benchmark, while our  
18 method can be considered at the same level of Domain Parser and PUU. From a more  
19 detailed analysis that considers the one-, two- and three-domain subsets of  
20 Benchmark\_2, it is possible to note that the new proposed approach is less reliable  
21 than the other automated techniques when it deals with one- and, especially, three-  
22 domain proteins, but it shows a very good accuracy for two-domain chains, resulting  
23 second only to the PDP strategy. However, it is also worth to observe that for three-  
24 domain proteins there is not a consensus among the other algorithmic methods taken  
25 into account since the agreement with the benchmark ranges from 53.7% (PUU) to  
26 72.2% (PDP). This can be considered as an evidence of the uncertainty associated  
27 with the automated techniques for the domains identification.  
28  
29

30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
As mentioned above, another important criterion to evaluate the capabilities of the  
new proposed strategy is the domains fragmentation<sup>31</sup>. It provides a measure of the  
balance between compact domains, which often correspond to the presence of many



1  
2  
3 fragments, and biologically meaningful domains, which are usually characterized by  
4 few fragments. Fragmentation is related to the three-dimensional organization of the  
5  
6 primary sequence. In fact, a domain may consist of one single contiguous stretch of  
7  
8 aminoacids, but, in many cases, polypeptide regions that are distant in sequence are  
9  
10 physically close in the tertiary structure due to the specific folding of the protein. In  
11  
12 the former case, isolating the domain from the whole protein would consist in cutting  
13  
14 a continuous polypeptide sequence and the stabilization would be provided by  
15  
16 intramolecular contacts among residues, which are kept together (namely, constrained  
17  
18 in space) by the covalent bonds of the polypeptide chain. A stable single domain  
19  
20 protein represents an extreme case of this situation. On the contrary, if many  
21  
22 fragments were present, stabilizing the domain in isolation would require keeping  
23  
24 together separated peptides in a specific geometric arrangement through non-covalent  
25  
26 interactions. This would represent an entropically unfavorable condition compared to  
27  
28 the previous case. As a consequence, a balance between a low number of fragments  
29  
30 and their lengths with respect to the domain dimensions is required to properly  
31  
32 stabilize the substructure.  
33  
34

35  
36 It has been observed that the average fragmentation of domains generally correlates  
37  
38 well with the average number of domains assigned by an algorithmic method and,  
39  
40 therefore, this means that if an automated strategy assigns on average more domains,  
41  
42 it also assigns on average more fragments per domain. In our case the average number  
43  
44 of domains per chain and the average number of fragments per domain are 1.76 and  
45  
46 1.56, respectively. Comparing these values with the corresponding ones obtained for  
47  
48 the expert consensus and for other algorithmic techniques<sup>31</sup>, the former is generally  
49  
50 lower while the latter is higher in all the situations. This result confirms the under-  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 cutting trend of the proposed strategy that tends to group many segments in the same  
4 cluster instead of identifying new domains.  
5  
6

7  
8 Finally, in order to completely characterize the accuracy of the new method, we have  
9 taken into account the very stringent boundary consistency criterion that has enabled  
10 us to determine whether our domain assignments were identical or not to the  
11 corresponding consensus assignments provided by expert techniques. In particular, we  
12 have chosen an 80% boundary consistency according to which two assignments are  
13 considered identical only if they overlap for at least 80% of the chain length.  
14  
15 Considering 124 structures of Benchmark\_3 (see the subsection about the preparation  
16 of the Bourne benchmark datasets), it is possible to observe that the agreement with  
17 the expert consensus is at 56.1% and this increases to 65.9% taking into account a  
18 70% boundary consistency (see Table 3). This trend can be observed much more  
19 clearly in Fig. 9 where, taking into account only the structures of Benchmark\_3 for  
20 which our method assigns the correct number of domains, the fraction of chains with  
21 inconsistent boundaries is represented in function of the boundary consistency  
22 threshold: when the boundary consistency is set equal to 80% the fraction of  
23 inconsistent boundaries is 23.3% and it reduces to 14.4% and 10% for threshold  
24 values corresponding to 75% and 70%, respectively. These data show that in most of  
25 the cases in which the new technique does not reach the desired agreement with the  
26 benchmark dataset, the obtained results are not very far from the exact domain  
27 assignments (see examples in Fig. 10). Hence, we believe that the proposed strategy is  
28 well founded and that its accuracy and performance can be sensitively increased  
29 through targeted improvements of the algorithm.  
30  
31

32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58 The fact that a substantial fraction of protein domain decompositions has been  
59 reproduced using our new *ab initio* physical-chemistry based method, which takes  
60

1  
2  
3 into account only the intramolecular non-bonded interaction energies of proteins,  
4  
5 suggests that the energetic domain-decomposition can be profitably used in several  
6  
7 applicative contexts. In particular, the energy-based domain definition could not only  
8  
9 provide important insights into the determinants of the modular organization of  
10  
11 protein structures and functional regions, but it may also be used as a comparative  
12  
13 tool to highlight common features in protein families and super-families.  
14  
15

16  
17 Moreover, in an applicative context, the knowledge of the relevant interactions for  
18  
19 stabilization provided by our novel approach could be used to design experiments  
20  
21 aimed at perturbing the stability of the domains and, therefore, to report on the  
22  
23 structural or functional consequences of such perturbations. For example, in protein  
24  
25 design the strategy could be exploited to suggest possible regions for the modification  
26  
27 of protein sequences as well as to pinpoint locations where the sequence could be split  
28  
29 into autonomously folding and structurally stable units. The rational  
30  
31 generation/modification of stable protein constructs with specific functions may have  
32  
33 important applications in several fields ranging from biotechnology to structural  
34  
35 vaccinology.  
36  
37

38  
39 Finally, the domains identified with our technique can be used to add biases in  
40  
41 molecular simulations and in the coarse graining of protein structures for aiding the  
42  
43 exploration of conformational spaces.  
44  
45  
46  
47  
48  
49

## 50 **Conclusions**

51  
52 In this paper we have proposed a novel algorithmic technique aimed at identifying  
53  
54 structural protein domains as independent protein folding units. In order to  
55  
56 accomplish this task we have developed a method for the proper analysis of the non-  
57  
58 bonded interaction energies. In particular, we have shown that, optimally selecting the  
59  
60

1  
2  
3 eigenvectors of the non-bonded interaction energy matrix, it is possible to highlight  
4  
5 only those blocks of interactions that characterize each folding unit and, consequently,  
6  
7 to detect the corresponding domains by means of a clustering procedure. This  
8  
9 approach has been afterwards tested against benchmark datasets that were constructed  
10  
11 to assess the capabilities and the features of automated methods. The collected results  
12  
13 show that the novel strategy is quite accurate in determining the number of domains,  
14  
15 despite its tendency to under-cut a few structures. Furthermore, a detailed analysis of  
16  
17 the obtained results has suggested that there are two possible reasons to explain this  
18  
19 shortcoming of the new method. One consists in the fact that the adopted clustering  
20  
21 algorithm has the tendency to group together as many protein fragments as possible  
22  
23 instead of identifying new clusters. This drawback can be overcome improving our  
24  
25 non-optimized clustering procedure or implementing one of those optimized  
26  
27 procedures already adopted by other algorithmic strategies. The second reason is that  
28  
29 the novel technique tries to detect all the possible significant non-bonded interactions  
30  
31 between aminoacids and, if the structure in exam is very compact, it becomes difficult  
32  
33 to distinguish the different domains of the protein. If on one hand this can be  
34  
35 obviously considered as an intrinsic bias of the proposed strategy, on the other hand,  
36  
37 this can be also seen as the strength of the method since it provides relevant insights  
38  
39 into the effective role played by specific interface residues and substructures in the  
40  
41 stabilizing interactions required to obtain a specific 3D organization.  
42  
43  
44  
45  
46  
47  
48  
49

50 Overall, analyzing the outcomes provided by the current version of the program, it is  
51  
52 worthwhile to observe that the proposed method is well grounded. For instance,  
53  
54 considering the consistency of the assigned domain boundaries, it is possible to obtain  
55  
56 a very good agreement with the proper benchmark dataset only slightly reducing the  
57  
58 consistency threshold (e.g., 75% or 70%). This means that the new strategy is already  
59  
60

1  
2  
3 able to provide domain assignments not really different from the consensus ones and,  
4  
5 therefore, only few algorithm refinements can significantly improve the results. These  
6  
7 observations are important also considering the fact that our approach has not been  
8  
9 previously trained against a suitable set of protein structures and it does not need any  
10  
11 preliminary information to run, namely it can be considered as a real *ab initio* method.  
12  
13 Furthermore, since the proposed strategy basically consists in analyzing the essential  
14  
15 components of the residue-residue non-bonded interaction energy matrix, it is  
16  
17 possible to get additional energy-information for the system under investigation. In  
18  
19 particular, we can obtain insights into the energetic stability of the detected domains  
20  
21 and a general picture of the folding energetics with the identification of all the  
22  
23 possible crucial residues for the protein fold.  
24  
25  
26  
27

28  
29 A limitation of the new approach might be represented by the associated  
30  
31 computational expense since it entails a structure minimization to remove bad  
32  
33 contacts from the starting crystal structure and a matrix diagonalization. However, it  
34  
35 is worthwhile to point out that this can be considered a serious problem only for very  
36  
37 large proteins and that this can be also considered as an additional cost to get more  
38  
39 information compared to the manual or the usual compactness-based algorithms.  
40  
41 Furthermore, nowadays, the fast pace in methodological developments and the  
42  
43 continuous increase in computer power allow to perform long time scale all-atom  
44  
45 Molecular Dynamics simulations of large systems<sup>42-44</sup> and, therefore, the use of only  
46  
47 slightly computational demanding computational techniques does not represent a  
48  
49 limiting drawback.  
50  
51  
52  
53

54  
55 Finally, we want to stress that our new strategy is another possible automated tool to  
56  
57 perform domain assignments. As already pointed out in the introduction, the quite  
58  
59 ambiguous definition of protein structural domain does not allow determining which  
60

1  
2  
3 algorithmic technique is the most reliable and, therefore, we believe that our method  
4  
5 may be used in combinations with other strategies either to provide a consensus view  
6  
7 or to enrich existing assignment techniques with physico-chemical and energetic  
8  
9 insights. Moreover, the new approach might be very useful in aiding research in  
10  
11 several applicative and experimental contexts.  
12  
13  
14  
15

### 16 17 **Acknowledgments**

18  
19 We thank A.I.R.C. (Project MFAG 11775) and the Cariplo VACCINI GtA and  
20  
21 LOMBARDY ASTIL projects for financial support. We would like to thank Guido  
22  
23 Scarabelli, Rubben Torella and Massimiliano Meli for helpful discussions.  
24  
25  
26  
27

### 28 29 **Supporting Information Available**

30  
31 Table S1: complete list of discarded chains of Benchmark\_2 and Benchmark\_3; Table  
32  
33 S2: identified domains and number of selected eigenvectors for all the examined  
34  
35 structures; Complete references: 5, 40 and 42. This material is available free of charge  
36  
37 via the Internet at <http://pubs.acs.org>.  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## References

1. Wetlaufer, D. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 697-701.
2. Veretnik, S.; Bourne, P. E.; Alexandrov, N. A.; Shindyalov, I. N. *J. Mol. Biol.* **2004**, *339*, 647-678.
3. Alden, K.; Veretnik, S.; Bourne, P. E. *BMC Bioinformatics* **2010**, *11*, 310.
4. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536-540.
5. Greene, L. H.; Lewis, T. E.; Addou, S.; Cuff, A.; Dallman, T.; Dibley, M.; Redfern, O.; Pearl, F.; Nambudiry, R.; Reid, A. et al. *Nucleic Acid Res.* **2007**, *35*, D291-D297.
6. Majumdar, I.; Kinch, L. N.; Grishin, N. V. *PLoS ONE* **2009**, *4*, e5084.
7. Wodak, S. J.; Janin, J. *Biochemistry* **1981**, *20*, 6544-6552.
8. Xu, Y.; Xu, D.; Gabow, H. N. *Bioinformatics* **2000**, *16*, 1091-1104.
9. Guo, J.-T.; Xu, D.; Kim, D.; Xu, Y. *Nucleic Acid Res.* **2003**, *31*, 944-952.
10. Alexandrov, N.; Shindyalov, I. *Bioinformatics* **2003**, *19*, 429-430.
11. Zhou, H.; Xue, B.; Zhou, Y. *Protein Sci.* **2007**, *16*, 947-955.
12. Holm, L.; Sander, C. *Proteins* **1994**, *19*, 256-268.
13. Siddiqui, A. S.; Barton, G. J. *Protein Sci.* **1995**, *4*, 872-884.
14. Carugo, O. *J. Appl. Cryst.* **2007**, *40*, 778-781.
15. Taylor, W. R. *Protein Eng.* **1999**, *12*, 203-216.
16. Holm, L.; Sander, C. *Proteins* **1998**, *33*, 88-96.
17. Shomaker, V.; Trueblood, K. N. *Acta Cryst. B* **1968**, *24*, 63-76.
18. Hayward, S.; Kitao, A.; Berendsen, H. J. C. *Proteins* **1997**, *27*, 425-437.
19. Hayward, S.; Berendsen, H. J. C. *Proteins* **1998**, *30*, 144-154.
20. Hinsen, K. *Proteins* **1998**, *33*, 417-429.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
21. Hinsen, K.; Thomas, A.; Field, M. J. *Proteins* **1999**, *34*, 369-382.
  22. Yesylevskyy, S. O.; Kharkyanen, V. N.; Demchenko, A. P. *Biophys. Chem.* **2006**, *119*, 84-93.
  23. Potestio, R.; Pontiggia, F.; Micheletti, C. *Biophys. J.* **2009**, *96*, 4993-5002.
  24. Berezovsky, I. N. *Prot. Eng.* **2003**, *16*, 161-167.
  25. Koczyk, G.; Berezovsky, I. N. *Nucleic Acid Res.* **2008**, *36*, W239-W245.
  26. Tiana, G.; Simona, F.; De Mori, G. M. S.; Broglia, R. A.; Colombo, G. *Protein Sci.* **2004**, *13*, 113-124.
  27. Ragona, L.; Colombo, G.; Catalano, M.; Molinari, H. *Proteins* **2005**, *61*, 366-376.
  28. Colacino, S.; Tiana, G.; Colombo, G. *BMC Struct. Biol.* **2006**, *6*, 17.
  29. Morra, G.; Colombo, G. *Proteins* **2008**, *72*, 660-672.
  30. Genoni, A.; Morra, G.; Merz, K. M. Jr.; Colombo, G. *Biochemistry* **2010**, *49*, 4283-4295.
  31. Holland, T. A.; Veretnik, S.; Shindyalov, I. N.; Bourne, P. E. *J. Mol. Biol.* **2006**, *361*, 562-590.
  32. Amadei, A.; Linnsen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412-425.
  33. van Alten, D. M. F.; Amadei, A.; Linnsen, A. B. M.; Eijsink, V. G. H.; Vriend, G.; Berendsen, H. J. C. *Proteins*, **1995**, *22*, 45-54.
  34. *Maestro, Version 8.0*; Schrödinger LLC: New York, 2007.
  35. Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779-815.
  36. Fiser, A.; Do, R. K.; Sali, A. *Protein Sci.* **2000**, *9*, 1753-1773.
  37. Marti-Renom, M. A.; Stuart, A.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325.



- 1  
2  
3 38. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudham, M. S.; Eramian,  
4 D.; Shen, M.-Y.; Pieper, U.; Sali, A. *Curr. Protoc. Bioinformatics* **2006**, *5*,  
5 Unit 5.6.  
6  
7  
8  
9  
10 39. *Macromodel, Version 9.5*; Schrödinger LLC: New York, 2007.  
11  
12 40. Case, D. A.; Darden, T. A.; Cheatham, T. E. III; Simmerling, C. L.; Wang, J.;  
13 Duke, R. E.; Luo, R.; Merz, K. M. Jr.; Pearlman, D. A.; Crowley, M. et al.  
14 *AMBER, Version 9*; University of California – San Francisco: San Francisco,  
15 2009.  
16  
17  
18  
19  
20  
21 41. Tuncbag, N.; Gursoy, A.; Guney, E.; Nussinov, R.; Keskin, O. *J. Mol. Biol.*  
22 **2008**, *381*, 785-802.  
23  
24  
25  
26 42. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.;  
27 Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B. et al.  
28 *Science* **2010**, *330*, 341-346.  
29  
30  
31  
32 43. Shan, Y. B.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw,  
33 D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181-9183.  
34  
35  
36  
37 44. Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhami, D. W.; Maragakis, P.; Shan,  
38 Y. B.; Xu, H.; Shaw, D. E. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13118-  
39 13123.  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Table 1 – Performance of the method in assigning the correct number of domains  
4 (reference: available Benchmark\_2). The results are shown for all the possible  
5  
6 structures (*Overall*) and for subsets containing chains constituted only by one, two,  
7  
8  
9 three or at least four domains.  
10  
11  
12  
13  
14

|                   | Correct (%) | Under-cut (%) | Over-cut (%) |
|-------------------|-------------|---------------|--------------|
| Overall           | 74.8        | 17.5          | 7.7          |
| 1-Domain          | 88.5        | 0.0           | 11.5         |
| 2-Domain          | 79.7        | 15.6          | 4.7          |
| 3-Domain          | 42.9        | 47.6          | 9.5          |
| At least 4-Domain | 16.7        | 83.3          | 0.0          |

Table 2 – Performances of the Domain Parser, PDP and PUU methods in assigning the correct number of domains (reference: complete Benchmark\_2). The results are shown for all the possible structures (*Overall*) and for subsets containing chains constituted only by one, two, or three domains.

|          | Domain Parser (%) |           |          | PDP (%) |           |          | PUU (%) |           |          |
|----------|-------------------|-----------|----------|---------|-----------|----------|---------|-----------|----------|
|          | Correct           | Under-cut | Over-cut | Correct | Under-cut | Over-cut | Correct | Under-cut | Over-cut |
| Overall  | 77.1              | 18.5      | 4.5      | 85.0    | 3.8       | 11.2     | 74.2    | 7.3       | 18.5     |
| 1-Domain | 97.2              | 0.0       | 2.8      | 96.3    | 0.0       | 3.7      | 94.4    | 0.0       | 5.6      |
| 2-Domain | 72.5              | 21.0      | 6.5      | 83.3    | 2.2       | 14.5     | 69.6    | 10.1      | 20.3     |
| 3-Domain | 64.8              | 31.5      | 3.7      | 72.2    | 11.1      | 16.7     | 53.7    | 9.3       | 37.0     |

1  
2  
3 Table 3 – Performance of the method in assigning the correct domain boundaries  
4  
5  
6 (Reference: available Benchmark\_3) considering different values for the Boundary  
7  
8 Consistency threshold.  
9  
10

|                            | Boundary Consistency (%) |      |      |      |      |
|----------------------------|--------------------------|------|------|------|------|
|                            | 80%                      | 75%  | 70%  | 65%  | 60%  |
| Accurate domain assignment | 56.1                     | 62.6 | 65.9 | 65.9 | 69.9 |
| Failed domain assignment   | 17.1                     | 10.6 | 7.3  | 7.3  | 3.3  |
| Under-cut                  | 17.9                     | 17.9 | 17.9 | 17.9 | 17.9 |
| Over-cut                   | 8.9                      | 8.9  | 8.9  | 8.9  | 8.9  |

## Figure Captions

Figure 1. Components (in absolute value) of the eigenvector associated with the lowest eigenvalue of the non-bonded interaction energy matrix for (A) the one-domain chain 1rcb, (B) the two-domain chain 6paxa, (C) the three-domain chain 1ciy and (D) the four-domain chain 1dgkn.

Figure 2. Working hypotheses. (A) Ideal situation for a three-domain protein: each domain is associated with an eigenvector with a block structure of significant components (here depicted in blue, red and green for the first, the second and the third eigenvector, respectively). The blocks do not overlap and their union covers all the residues. (B) Essential folding matrix in an ideal situation for a three-domain protein: the blue, red and green blocks immediately identify the protein domains. (C) Real situation: the eigenvectors associated with the domains have a block structure, but their blocks overlap and do not cover all the residues.

Figure 3. Selection of the essential eigenvectors, construction of the essential folding matrix and domains identification for the 1kzq protein: (A) components (in absolute value) of the eigenvector associated with the lowest eigenvalue (blue line) and significance threshold for the eigenvectors components (dotted horizontal line); (B) components (in absolute value) of the second essential eigenvector (red line); (C) essential folding matrix constructed with the set of essential eigenvectors; (D) symbolized essential folding matrix; (E) clustering of the symbolized essential folding matrix; (F) identified domains for the 1kzq protein (1-125 (blue) // 126-253 (red)).

1  
2  
3 Figure 4. Comparison between the symbolized essential folding matrix (A) and the  
4 contacts matrix (D) of the 1crxa protein (C). The domains identified by the new  
5 technique are depicted in blue (residues 1-110) and red (residues 116-295). Note that  
6 our residues numeration is different from the one in the Bourne database (see  
7 Supporting Information to reconstruct it).  
8  
9  
10  
11  
12  
13  
14  
15  
16

17 Figure 5. Clustering algorithm (early stage): (A) the densities of the small diagonal  
18 blocks are computed; (B) the small diagonal *tesserae* whose density is greater than or  
19 equal to the overall density are considered; (C) the interaction density between two  
20 consecutive diagonal blocks is examined; (D) two consecutive diagonal *tesseare* are  
21 merged in a new temporary diagonal cluster; (E) the clustering procedure along the  
22 diagonal continues computing the interaction density between the temporary diagonal  
23 cluster and its consecutive diagonal block; (F) diagonal clusters obtained at the end of  
24 the early stage of the clustering algorithm.  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37

38 Figure 6. Clustering algorithm (late stage): (A) the possibility to join non-consecutive  
39 diagonal clusters is taken into account; (B) clusters obtained after merging all the  
40 possible non-consecutive diagonal blocks; (C) the clusters size is evaluated; (D)  
41 clusters with dimension smaller than the minimum required size are discarded; (E) the  
42 existence of non-overlapping gaps between fragments of multi-fragment clusters is  
43 checked; (F) the non-overlapping gaps are merged with the corresponding clusters.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4 Figure 7. Examples of cases in which the under-cutting trend is due to the clustering  
5 algorithm. (A) 1pspa: benchmark assignment (1-98; 99-105 (blue) // 59-98 (red)),  
6 symbolized essential folding matrix, new strategy assignment (1-105 (blue)). (B)  
7  
8 1a8y: benchmark assignment (1-124 (blue) // 125-226 (red) // 227-345 (orange)),  
9 symbolized essential folding matrix, new strategy assignment (6-115 (blue) // 126-315  
10 (red)). Note that our residues numeration is different from the one in the Bourne  
11  
12 database (see Supporting Information to reconstruct it).  
13  
14  
15  
16  
17  
18  
19  
20  
21

22 Figure 8. Examples of cases in which the under-cutting trend is due to significant  
23 interactions in compact structures. (A) 1tuia: benchmark assignment (1-205 (blue) //  
24 208-299 (red) // 303-396 (orange)), symbolized essential folding matrix, new strategy  
25 assignment (1-210 (blue) // 211-390 (red)); (B) 1bhga: benchmark assignment (1-203  
26 (blue) // 204-307 (red) // 308-610 (orange)), symbolized essential folding matrix, new  
27 strategy assignment (6-320 (blue) // 321-605 (red)). Note that our residues numeration  
28 is different from the one in the Bourne database (see Supporting Information to  
29 reconstruct it).  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42

43 Figure 9. Fraction of inconsistent boundaries in function of the boundary consistency  
44 threshold. Only structures for which the new strategy assigns the correct number of  
45 domains are considered.  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Figure 10. Examples of non-accurate domain boundaries assignment: (A) 1bc5a  
4 (boundary consistency: 71%): benchmark assignment (15-75 (blue) // 76-269 (red)),  
5 symbolized essential folding matrix, new strategy assignment (6-75; 131-150 (blue) //  
6 81-125; 186-265 (red)); (B) 1c0ma (boundary consistency: 78%): benchmark  
7 assignment (1-167 (blue) // 168-221 (red)), symbolized essential folding matrix, new  
8 strategy assignment (11-130 (blue) // 146-221 (red)). Note that our residues  
9 numeration is different from the one in the Bourne database (see Supporting  
10 Information to reconstruct it).  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Figure 1-

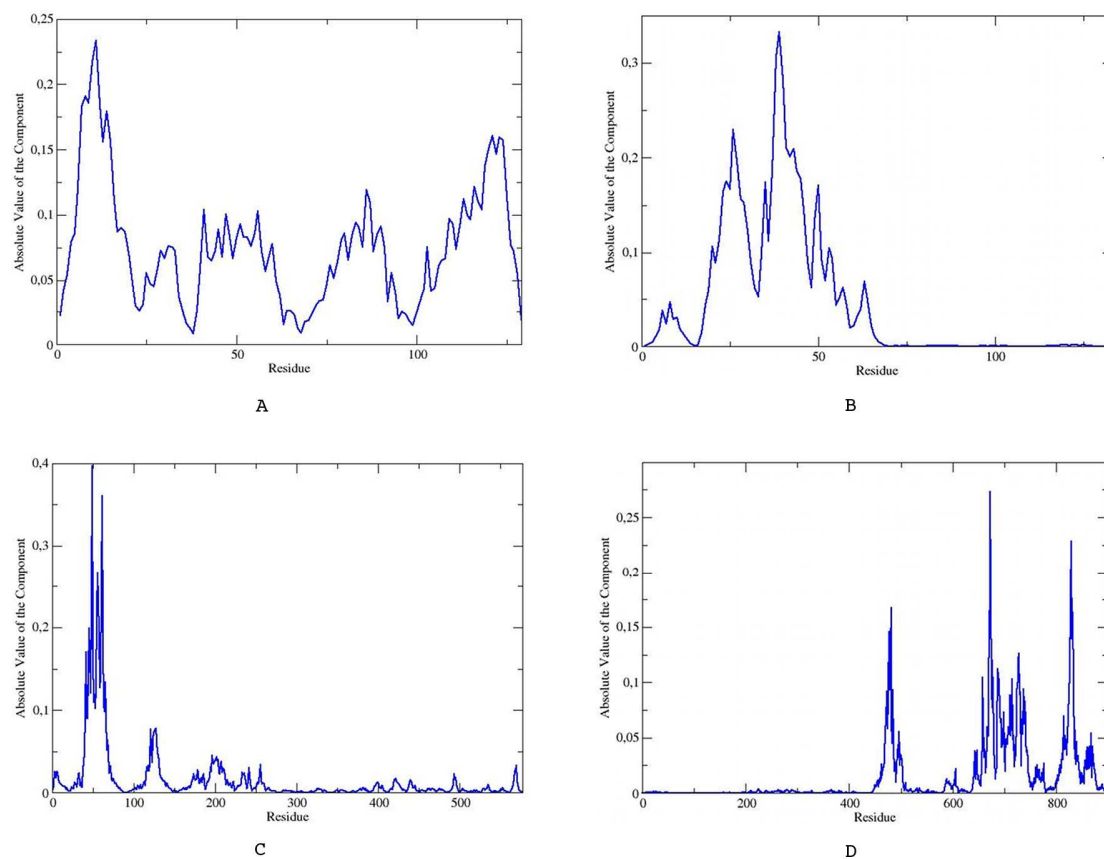
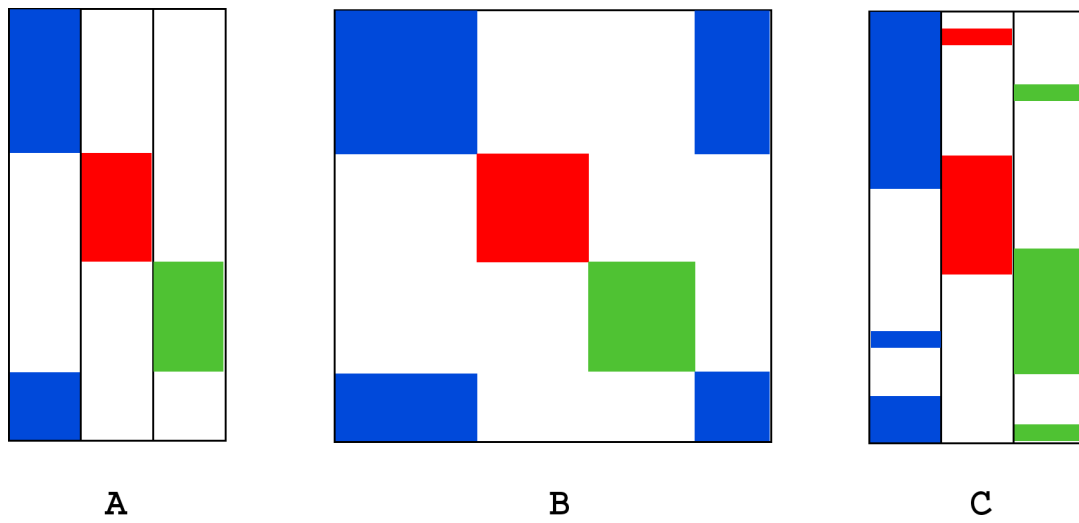


Figure 2 –



A

B

C

Figure 3 –

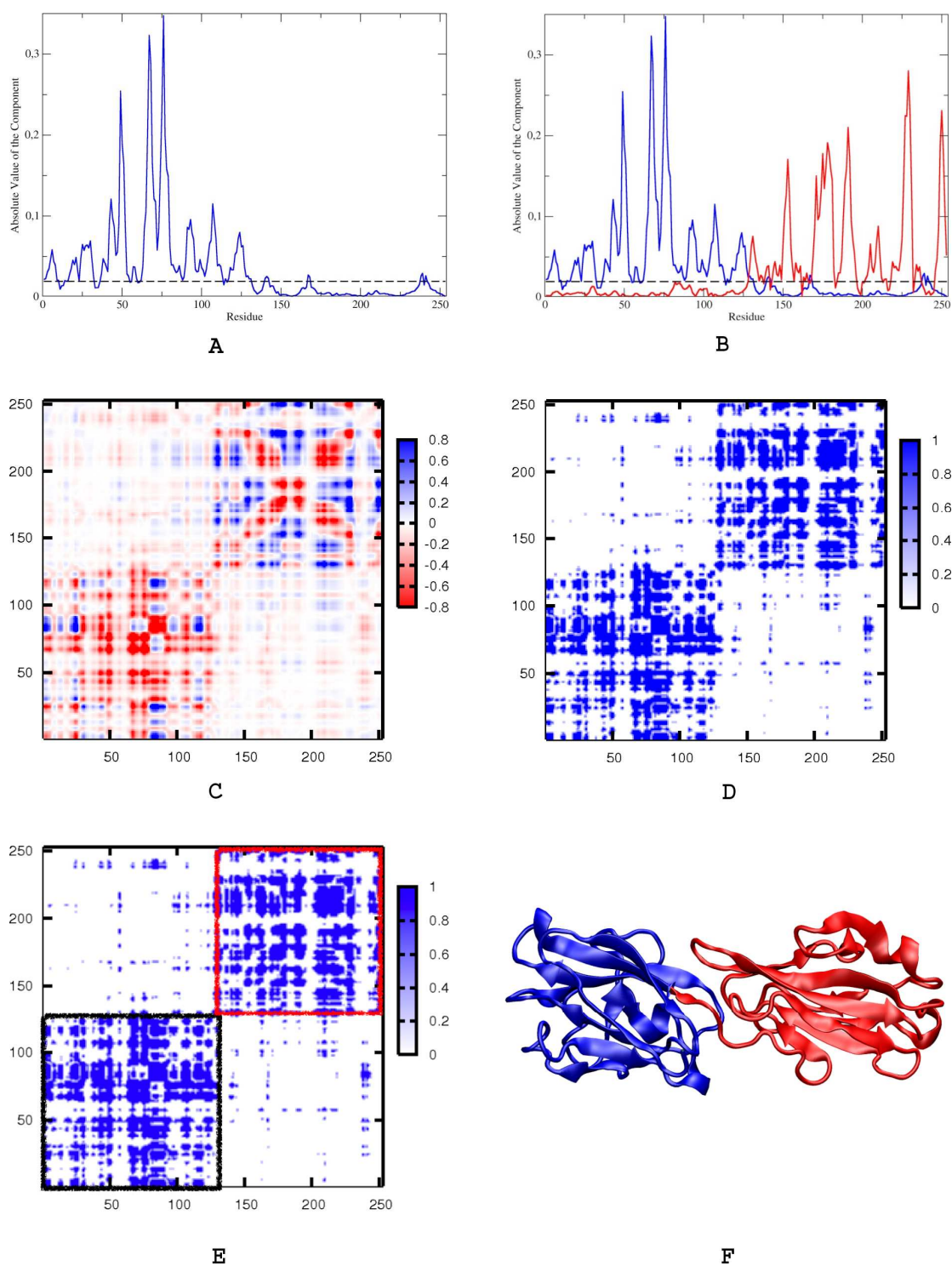


Figure 4-

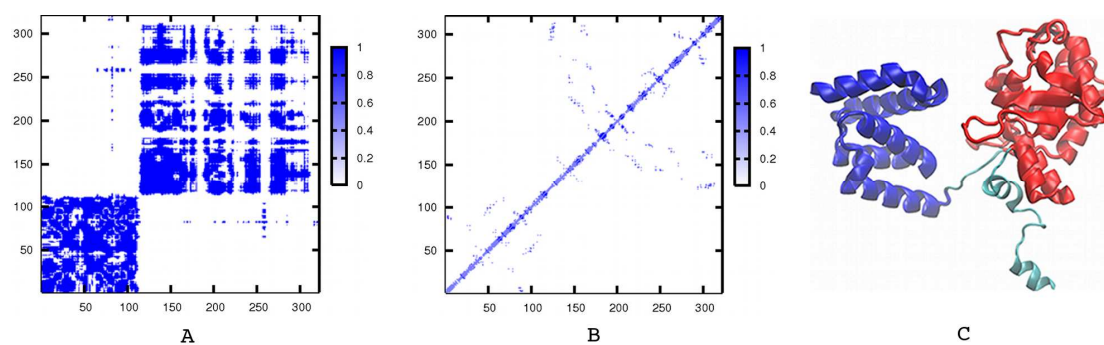
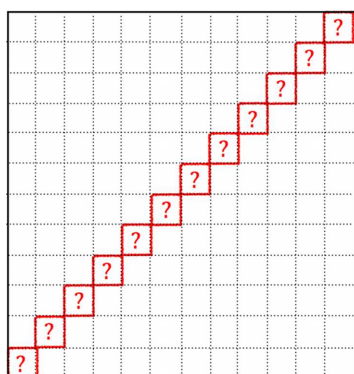
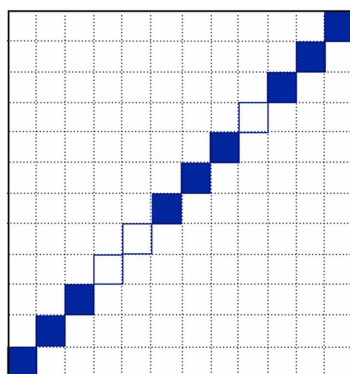


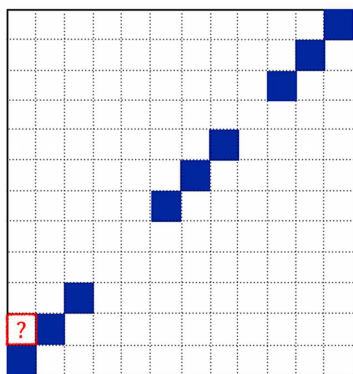
Figure 5 –



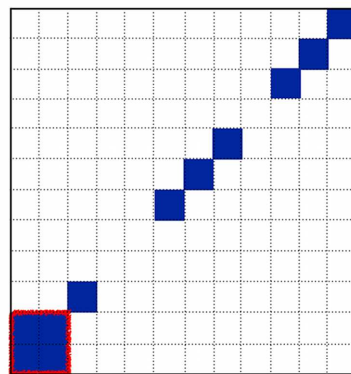
A



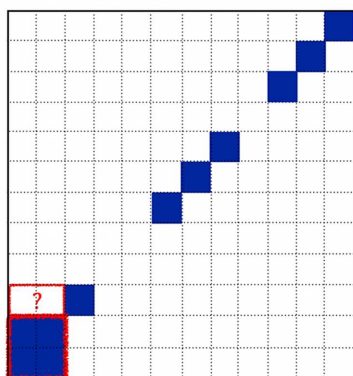
B



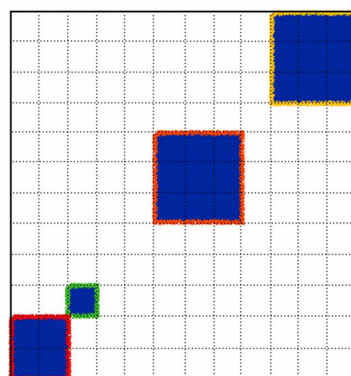
C



D



E



F

Figure 6 –

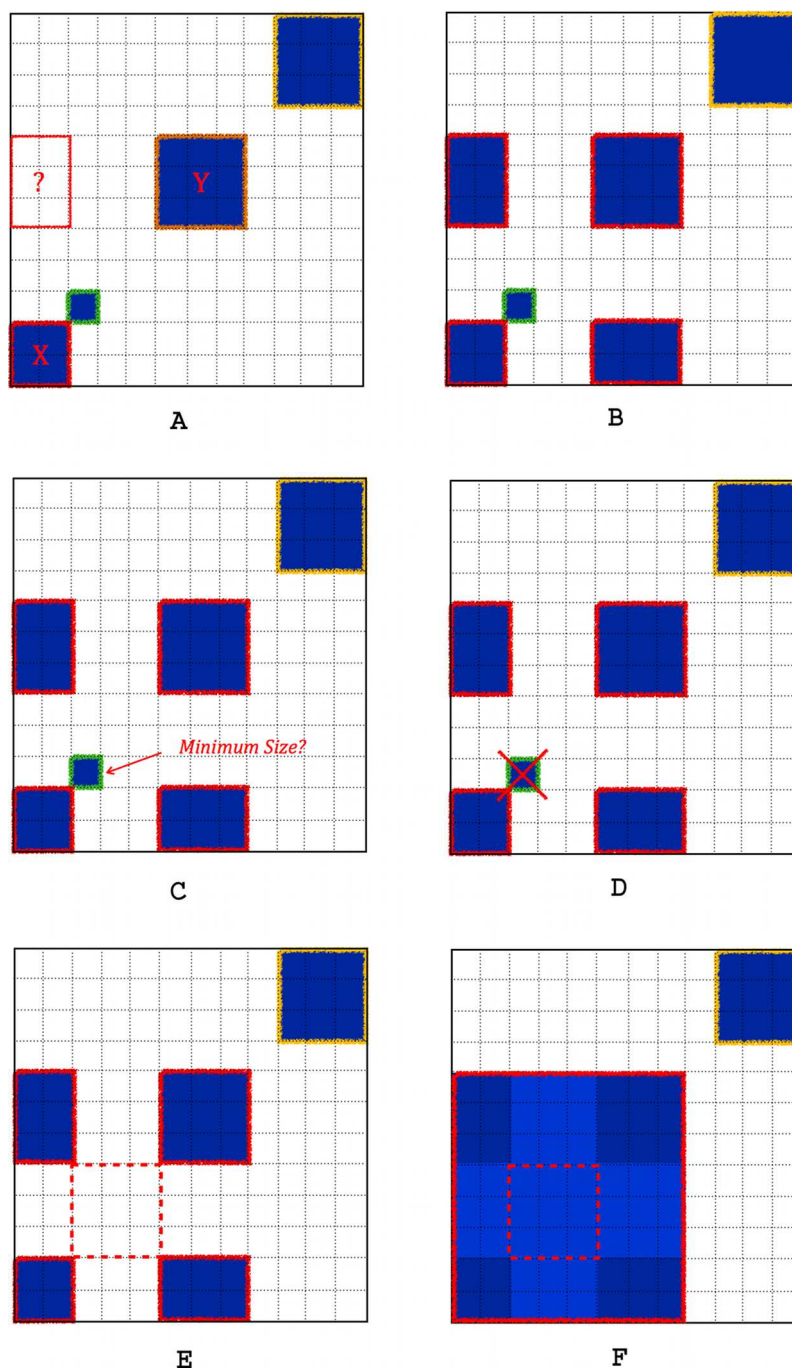




Figure 7 –

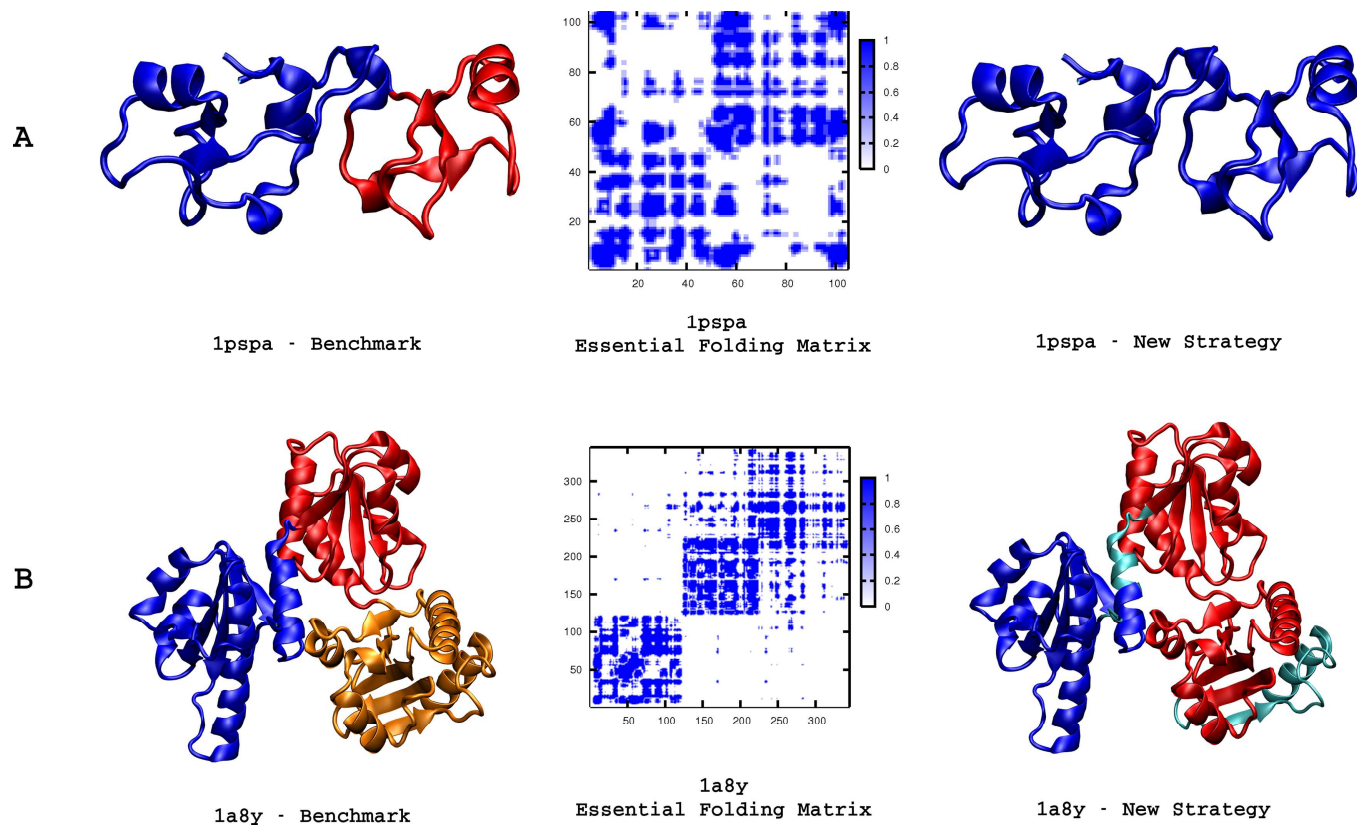


Figure 8 –

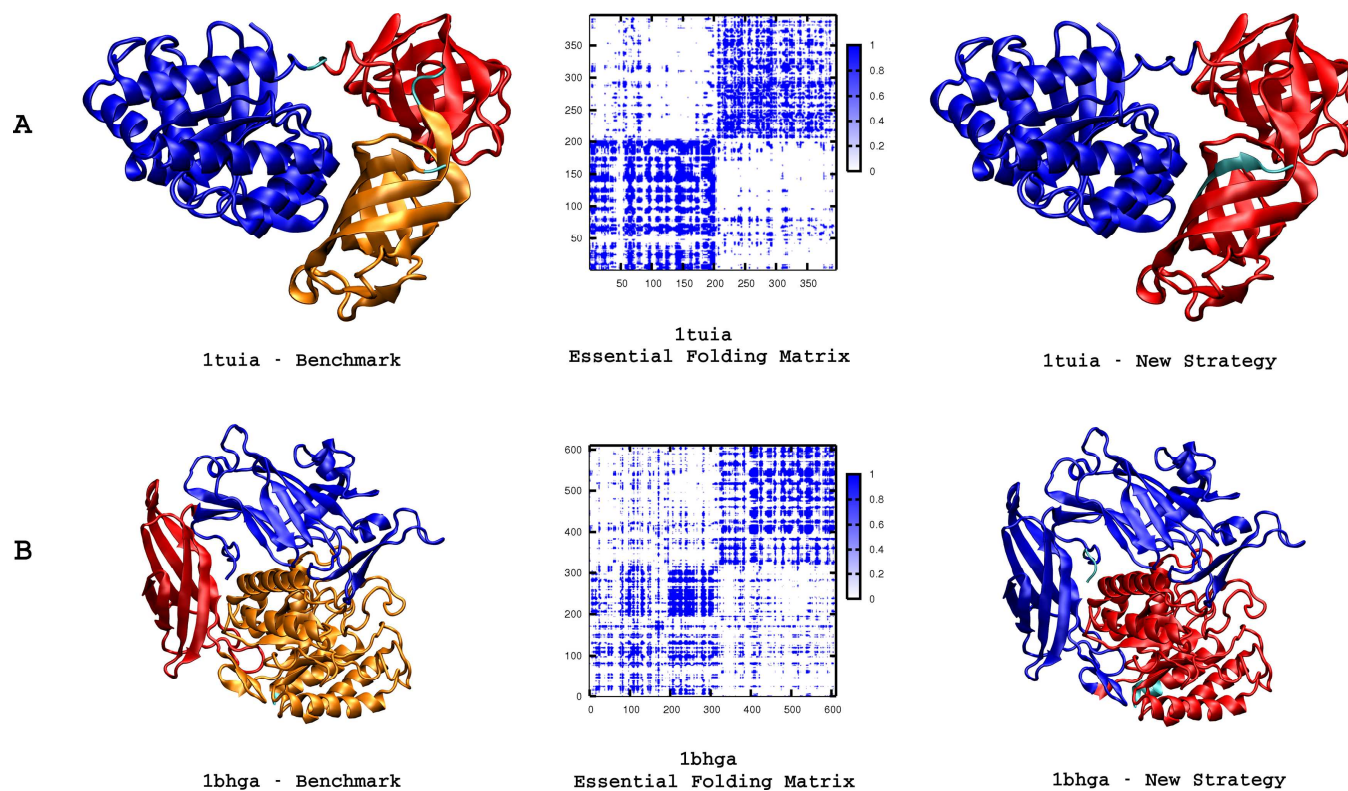




Figure 9 –

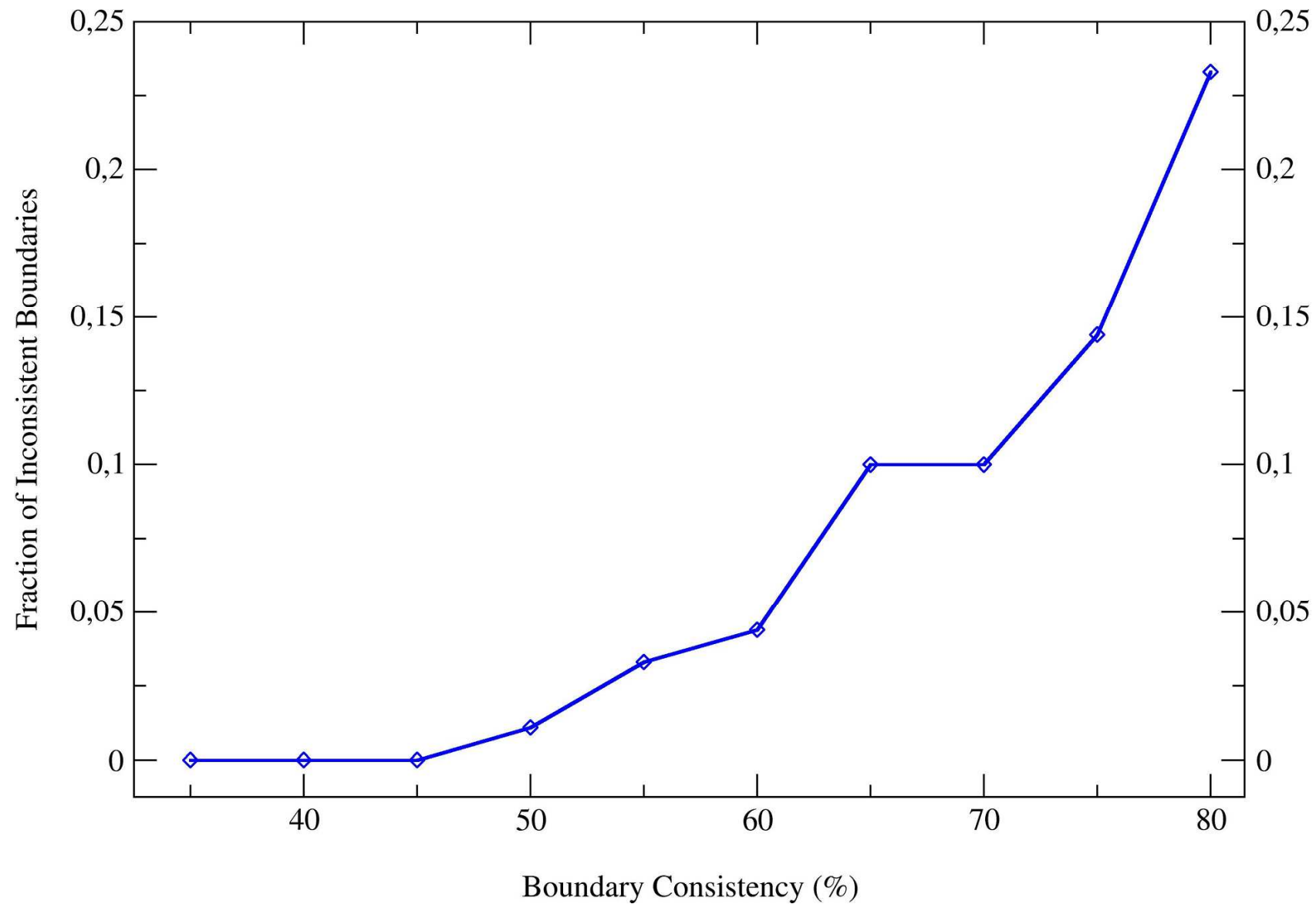
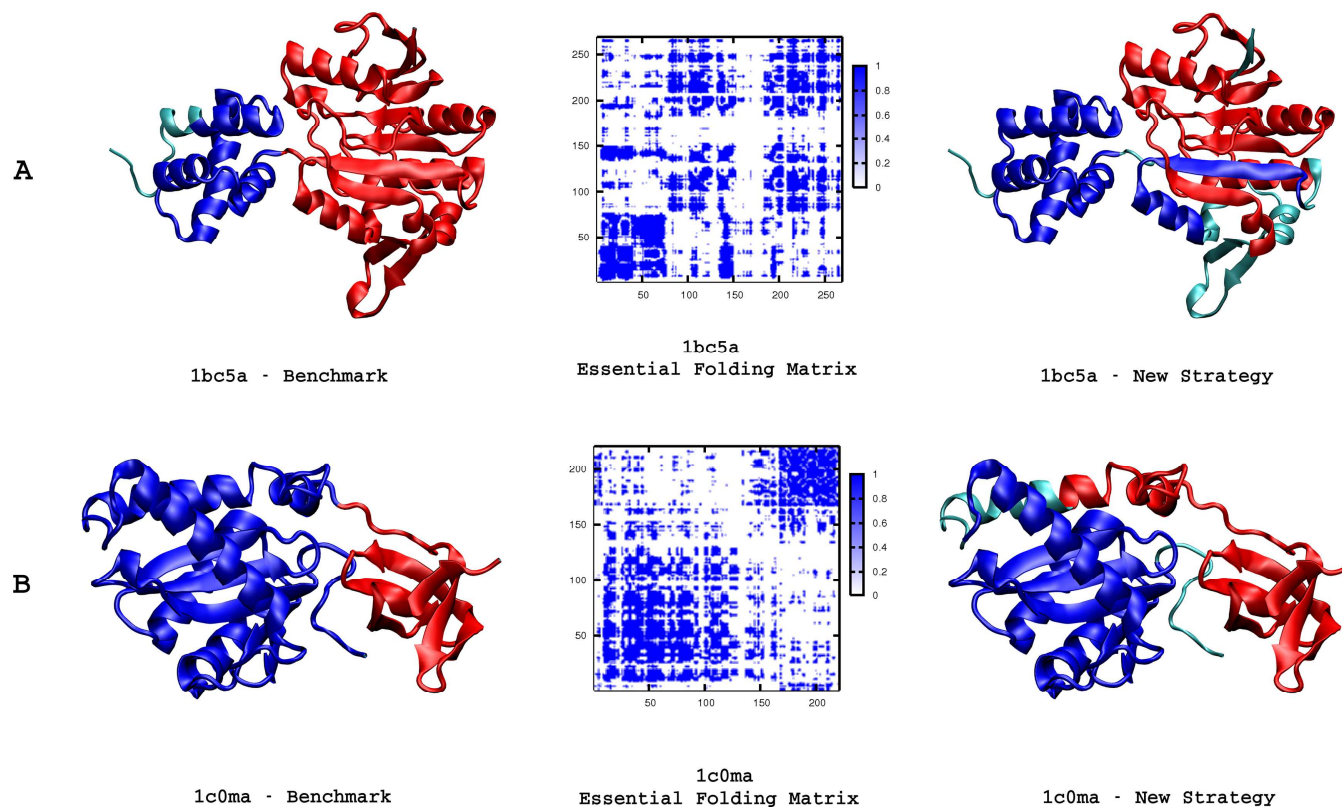
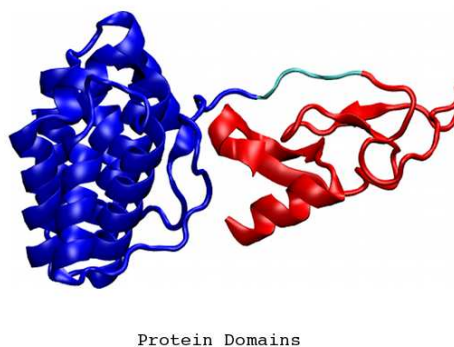
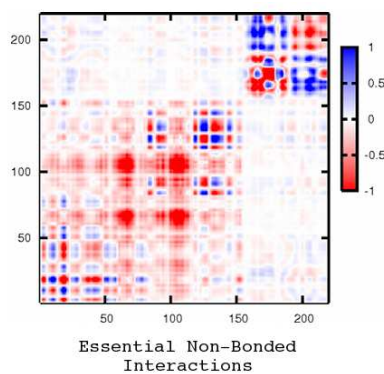
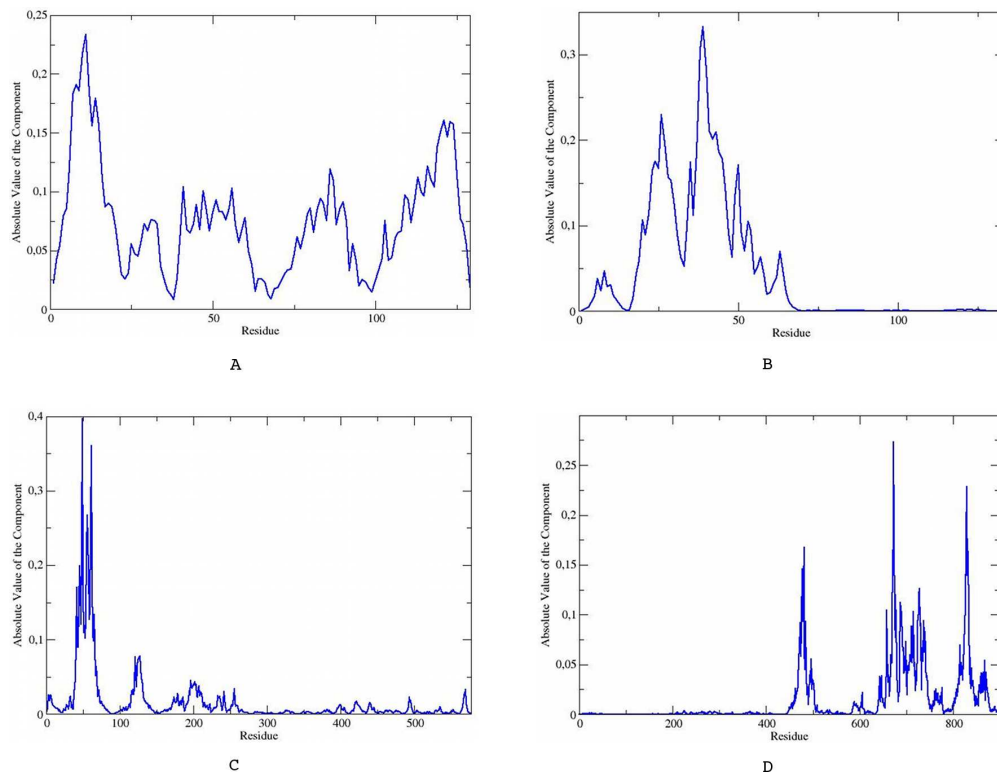


Figure 10 –

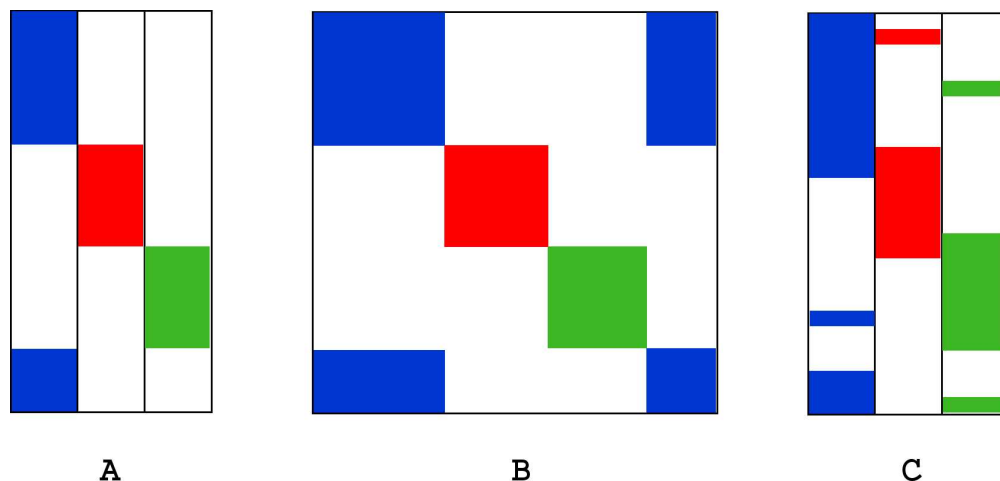


## Table of Contents Image –



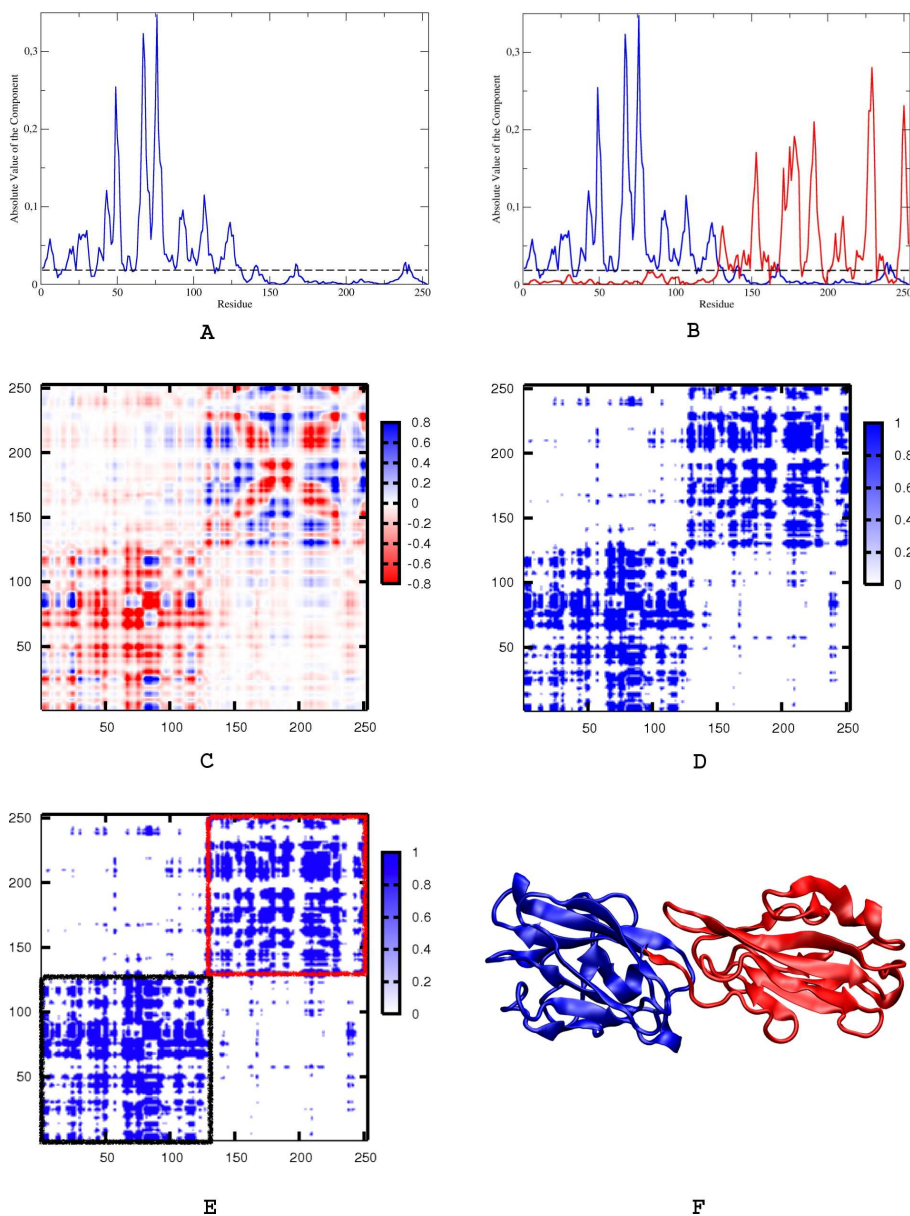


Components (in absolute value) of the eigenvector associated with the lowest eigenvalue of the non-bonded interaction energy matrix for (A) the one-domain chain 1rcb, (B) the two-domain chain 6paxa, (C) the three-domain chain 1ciy and (D) the four-domain chain 1dgnk.  
150x114mm (300 x 300 DPI)

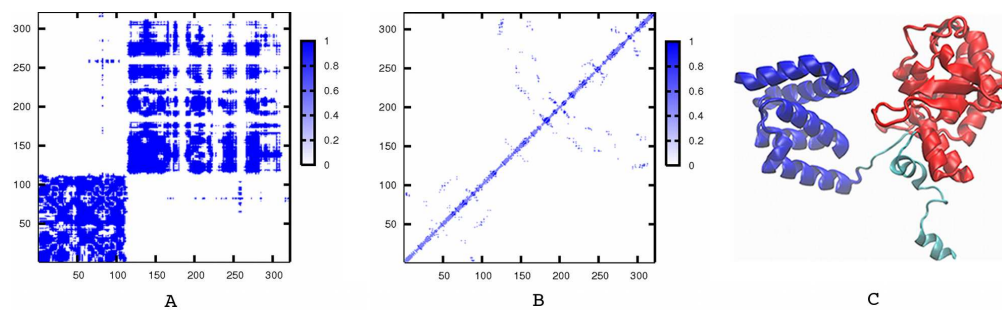


Working hypotheses. (A) Ideal situation for a three-domain protein: each domain is associated with an eigenvector with a block structure of significant components (here depicted in blue, red and green for the first, the second and the third eigenvector, respectively). The blocks do not overlap and their union covers all the residues. (B) Essential folding matrix in an ideal situation for a three-domain protein: the blue, red and green blocks immediately identify the protein domains. (C) Real situation: the eigenvectors associated with the domains have a block structure, but their blocks overlap and do not cover all the residues.

142x66mm (300 x 300 DPI)

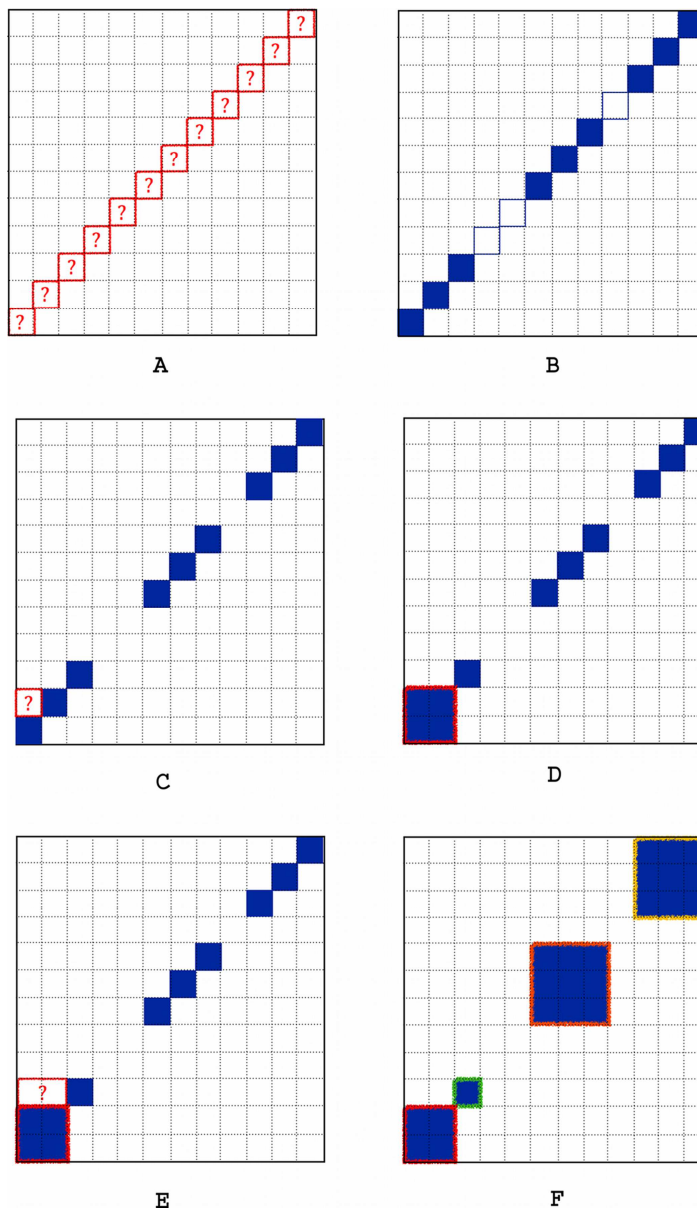


Selection of the essential eigenvectors, construction of the essential folding matrix and domains identification for the 1kzq protein: (A) components (in absolute value) of the eigenvector associated with the lowest eigenvalue (blue line) and significance threshold for the eigenvectors components (dotted horizontal line); (B) components (in absolute value) of the second essential eigenvector (red line); (C) essential folding matrix constructed with the set of essential eigenvectors; (D) symbolized essential folding matrix; (E) clustering of the symbolized essential folding matrix; (F) identified domains for the 1kzq protein (1-125 (blue) // 126-253 (red)).  
160x213mm (300 x 300 DPI)



Comparison between the symbolized essential folding matrix (A) and the contacts matrix (D) of the 1crxa protein (C). The domains identified by the new technique are depicted in blue (residues 1-110) and red (residues 116-295). Note that our residues numeration is different from the one in the Bourne database (see Supporting Information to reconstruct it).  
167x50mm (300 x 300 DPI)

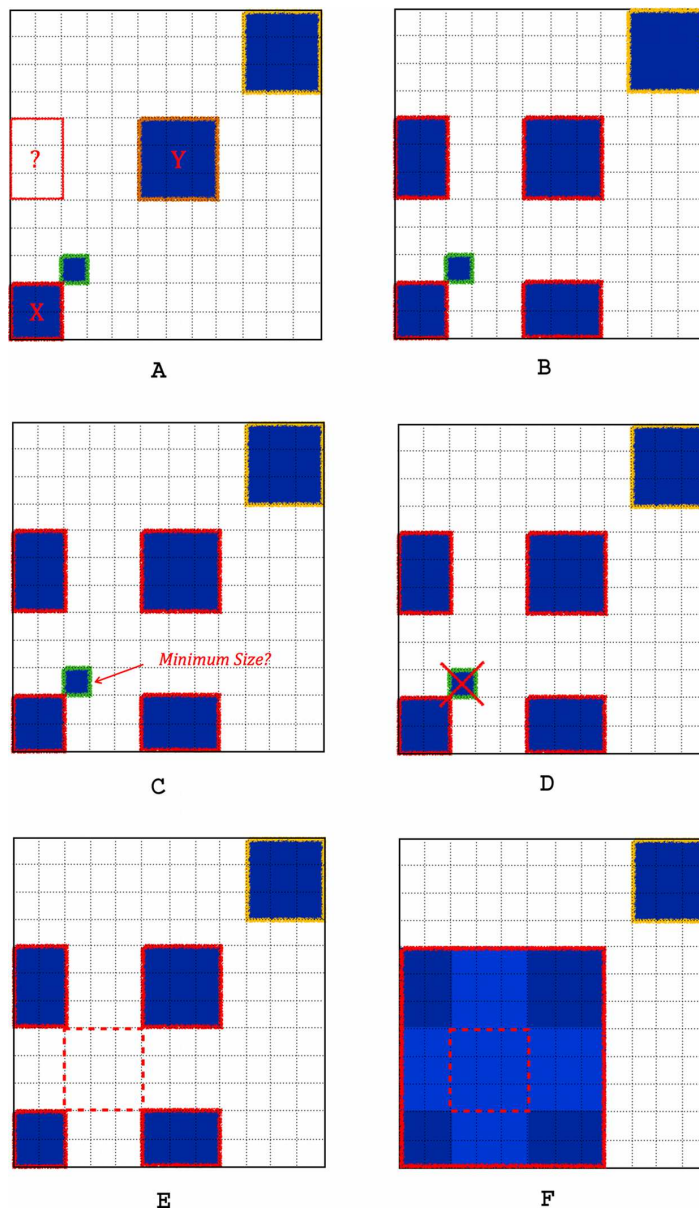




Clustering algorithm (early stage): (A) the densities of the small diagonal blocks are computed; (B) the small diagonal tesserae whose density is greater than or equal to the overall density are considered; (C) the interaction density between two consecutive diagonal blocks is examined; (D) two consecutive diagonal tesserae are merged in a new temporary diagonal cluster; (E) the clustering procedure along the diagonal continues computing the interaction density between the temporary diagonal cluster and its consecutive diagonal block; (F) diagonal clusters obtained at the end of the early stage of the clustering algorithm.

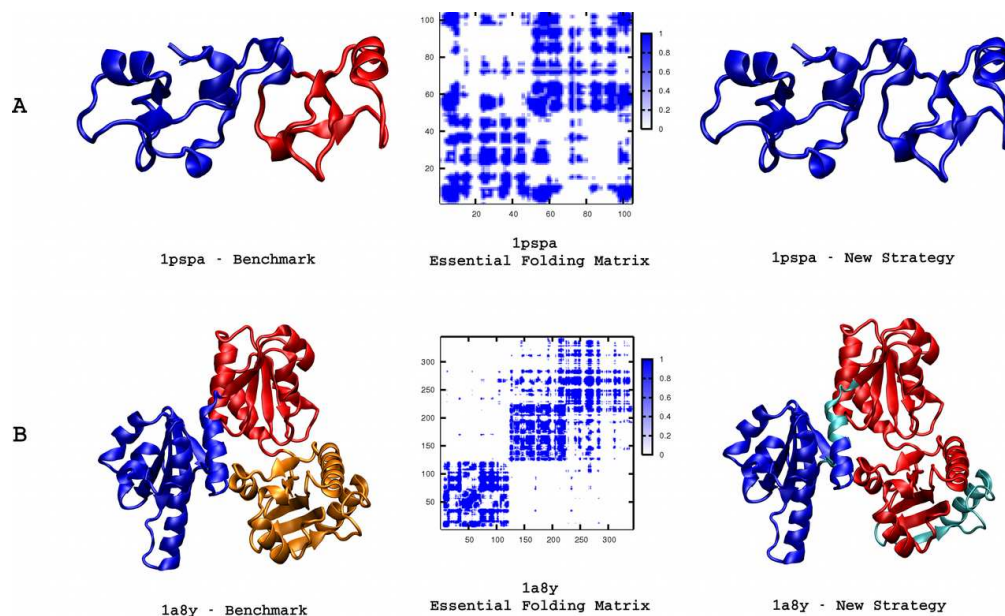
178x303mm (300 x 300 DPI)



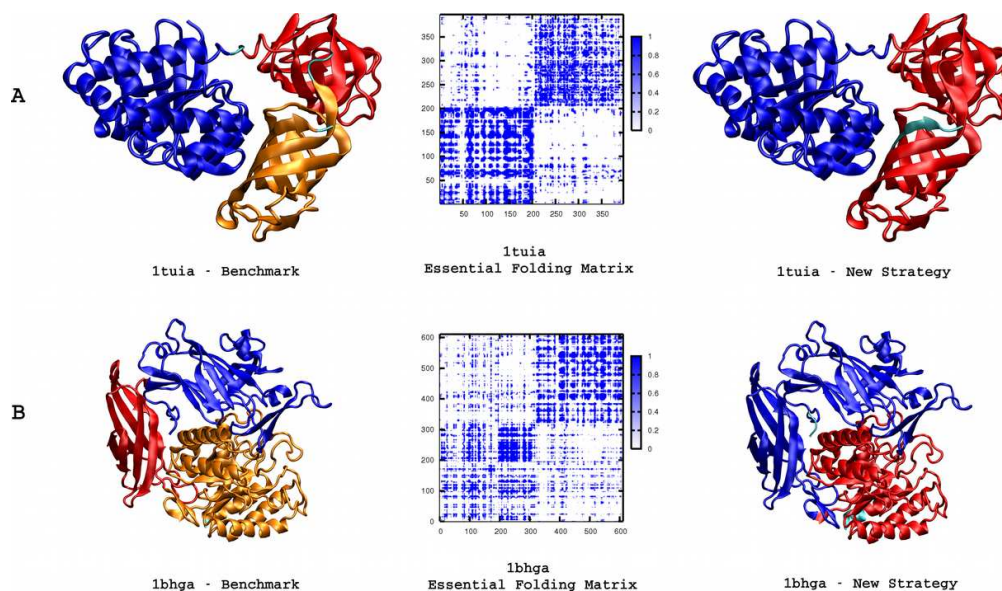


Clustering algorithm (late stage): (A) the possibility to join non-consecutive diagonal clusters is taken into account; (B) clusters obtained after merging all the possible non-consecutive diagonal blocks; (C) the clusters size is evaluated; (D) clusters with dimension smaller than the minimum required size are discarded; (E) the existence of non-overlapping gaps between fragments of multi-fragment clusters is checked; (F) the non-overlapping gaps are merged with the corresponding clusters.

105x179mm (300 x 300 DPI)

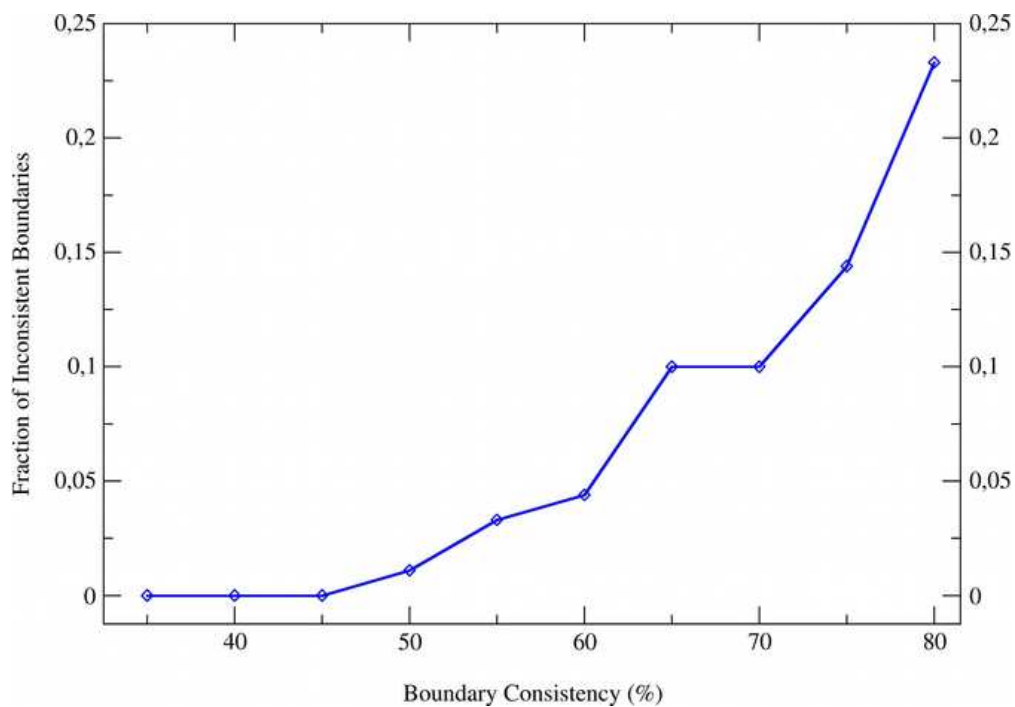


Examples of cases in which the under-cutting trend is due to the clustering algorithm. (A) 1pspa: benchmark assignment (1-98; 99-105 (blue) // 59-98 (red)), symbolized essential folding matrix, new strategy assignment (1-105 (blue)). (B) 1a8y: benchmark assignment (1-124 (blue) // 125-226 (red) // 227-345 (orange)), symbolized essential folding matrix, new strategy assignment (6-115 (blue) // 126-315 (red)). Note that our residues numeration is different from the one in the Bourne database (see Supporting Information to reconstruct it).  
107x64mm (300 x 300 DPI)

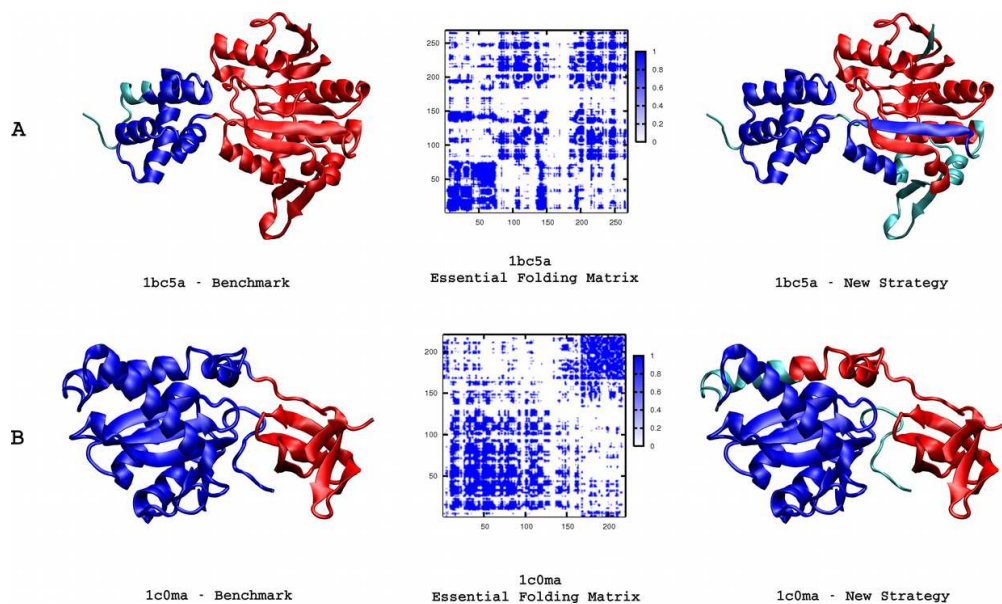


Examples of cases in which the under-cutting trend is due to significant interactions in compact structures. (A) 1tua: benchmark assignment (1-205 (blue) // 208-299 (red) // 303-396 (orange)), symbolized essential folding matrix, new strategy assignment (1-210 (blue) // 211-390 (red)); (B) 1bhga: benchmark assignment (1-203 (blue) // 204-307 (red) // 308-610 (orange)), symbolized essential folding matrix, new strategy assignment (6-320 (blue) // 321-605 (red)). Note that our residues numeration is different from the one in the Bourne database (see Supporting Information to reconstruct it).

102x59mm (300 x 300 DPI)

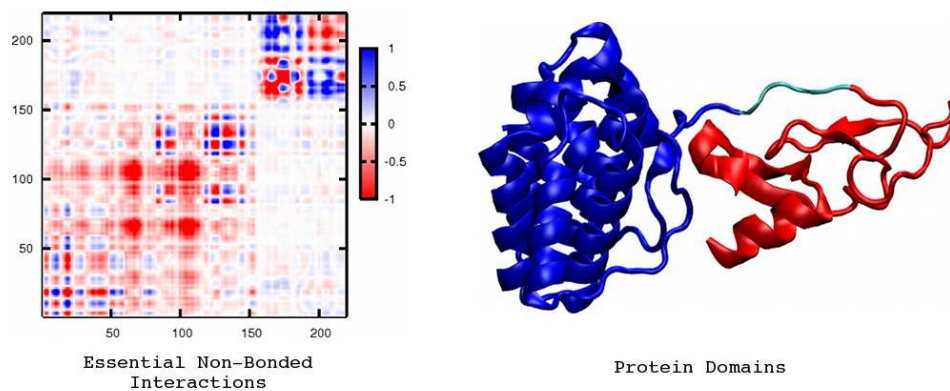


Fraction of inconsistent boundaries in function of the boundary consistency threshold. Only structures for which the new strategy assigns the correct number of domains are considered.  
56x39mm (300 x 300 DPI)



27 Examples of non-accurate domain boundaries assignment: (A) 1bc5a (boundary consistency: 71%);  
28 benchmark assignment (15-75 (blue) // 76-269 (red)), symbolized essential folding matrix, new  
29 strategy assignment (6-75; 131-150 (blue) // 81-125; 186-265 (red)); (B) 1c0ma (boundary  
30 consistency: 78%): benchmark assignment (1-167 (blue) // 168-221 (red)), symbolized essential  
31 folding matrix, new strategy assignment (11-130 (blue) // 146-221 (red)). Note that our residues  
32 numeration is different from the one in the Bourne database (see Supporting Information to  
33 reconstruct it).

104x61mm (300 x 300 DPI)



88x35mm (300 x 300 DPI)