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

Alessandro Genoni, Giulia Morra, Giorgio Colombo

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

Alessandro Genoni [(1,2)*], Giulia Morra [(1)], Giorgio Colombo [(1)*]

(1) Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco 9, 20131 Milano, Italy.

(2) Equipe de Chimie et Biochimie Théoriques, UMR 7565 SRSMC – CNRS, Université de Lorraine, Boulevard des Aiguillettes BP 70239, 54506 Vandoeuvre-lès-Nancy Cedex, France.

* Correspondence to:

Giorgio Colombo, Istituto di Chimica del Riconoscimento Molecolare – CNR, Via Mario Bianco 9, 20131 Milano, Italy. E-mail: giorgio.colombo@icrm.cnr.it. Phone: +39-02-28500031; Fax: +39-02-28901239;

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

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[1].

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[2]. 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[3].

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

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

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

Considering the algorithmic strategies, they are mainly compactness-based approaches that rely on protein structural features (e.g., the contacts density or the C_α-C_α distance maps) and whose underlying principle consists in the fact that, under a correct domain assignment, the number of inter-domain interactions is lower than the intra-domain interactions[7]. In this context we can consider the Domain Parser method

where a graphic-theoretic approach is used to perform the domain decomposition[8, 9]. In particular, after associating each residue with a node of a connected network and all the significant residue-residue contacts with an edge characterized by a capacity value that depends on the type of the interaction between the two involved aminoacids, this technique aims at dividing the network into two or more connected parts in such a way that the total edge capacities across the divisions are minimized[8]. Furthermore, in order to improve the accuracy of the decompositions when there is not an exact *a priori* knowledge of the number of domains, the previous method has been afterwards improved including the hydrophobic moment profile and neural networks to check the quality of the identified domains[9]. Along these lines, the PDP (Protein Domain Parser) and DDomain approaches are based on the assumptions that each structural domain is associated with a set of continuous protein fragments and identify the domains using a normalized domain-domain contact-interaction profile[10, 11]. Other two interesting techniques that exploit the residue-residue contacts as basic tool to identify structural domains in proteins are PUU[12] and Domak[13]. The former is a fast algorithm that is able to deal with multiple-segment domains and that tries to determine potential folding units through an eigenvalue analysis of the contacts map, while the latter, starting from the already mentioned fundamental principle that a domain has more residue-residue contacts within than without, is a strategy that introduces secondary structure information and other factors in order to improve the agreement with a training set of protein structures with well-established domains definitions. In the spirit of the techniques briefly described above, Dodis[14] takes into account only the C_α atoms and the matrix of the C_α- C_α distances is clustered using an agglomerative hierarchical procedure that separates the C_α atoms into distinct domains. It is worthwhile to observe that this method does not aim at providing the

best partition, but it produces many possible decompositions of a protein structure into one or more domains. Finally, in the framework of the compactness-based strategies, it is interesting to mention the approach proposed by Taylor[15], a technique that basically consists in an Ising model and in which the protein residues (namely, the structural elements of the model) change their state (in this case, a multi-value label) in function of the state of the close residues (i.e., the state of the neighbors).

All the methods described so far are based on the concept of compactness, but they ignore recurrence criteria, namely the recognition that many identified domains are common to different proteins at a high level of statistical significance. Therefore, in order to overcome this drawback, Holm and Sander have proposed DALI[16], an automated method that is able to identify recurrent domains in proteins computing the recurrence in terms of statistical significance of structural similarity for many pairs of substructures.

A completely different way to automatically identify structural domains is based on the observation that the internal dynamics of proteins can be accurately described in terms of movements of approximately rigid subunits. In this context, examples are the translation-libration-screwlike (TLS) motion analysis developed by Schomaker and Trueblood[17], the DynDom technique introduced by Hayward *et al.*[18, 19], the strategy devised by Hinsen and coworkers[20, 21], where clusters of aminoacids are identified as quasi-rigid blocks taking into account only low-frequency modes of fluctuations, and the approach proposed by Yesylevskyy *et al.*[22], which perform a hierarchical clustering of the correlation patterns. Furthermore, it is worthwhile to mention the method recently introduced by Micheletti and coworkers that, considering only the essential dynamic spaces of a protein, cluster the aminoacids minimizing a phenomenological strain-energy function by means of a variational scheme, so

ensuring that the fluctuations within each single group are as small as possible compared to those across different groups[23].

Finally, for the sake of completeness, it is important to consider energy-based techniques and, in particular, the one developed by Berezovsky and coworkers[24, 25]. Relying on the importance of the van der Waals forces in the folding process, this approach allows obtaining an energy hierarchy of domains for globular proteins simply analyzing van der Waals energy profiles.

In this paper, mainly taking into account the definition of domain as a compact and independent folding unit, we propose an alternative energy-based strategy for identifying structural domains and subunits. In particular, extending the recently developed Energy Decomposition Method[26-30], we compute the non-bonded interaction energy matrix of the structure to be examined. After diagonalizing this matrix, we select an optimal subset of its eigenvectors that allow identifying the most stabilizing residues and the essential non-bonded interactions that define each folding unit. The optimally selected eigenvectors enable to reconstruct a filtered non-bonded interaction energy matrix that is characterized by a significant quenching of the inter-domain interactions and that is afterwards subject to a clustering procedure to identify the correspondences between the interaction energy clusters and the structural domains of the protein in exam. In other words, our new method detects the most relevant residue-residue interactions that implicitly define the nuclei which are necessary to stabilize the three-dimensional fold of the different domains (or subunits). Therefore, in analogy with the previously adopted terminology[26-30], the groups of fundamental pair-interactions define the folding nuclei of each domain. This approach has been implemented in a computer program and it has been tested using the comprehensive benchmark datasets assembled by Bourne and coworkers

(i.e., Benchamrk_2 and Benchamrk_3) to assess the capabilities of novel or already existing algorithmic methods[31]. The new technique has proved to be accurate in determining the exact number of domains and to provide information on the domains boundaries. It is worth observing that, unlike most of the automated approaches, the new strategy is not trained using a set of proteins with domains assigned by experts and, above all, it is a completely *ab initio* technique, namely, it relies only on the physico-chemical analysis of the energetic properties of the different protein fragments and no preliminary information is given as input. In this respect, the proposed method may be useful in the domain partitioning of novel folds or novel structures for which no previous information is available. Moreover, our method is deeply connected to the energy-properties of protein folding and, therefore, it provides further and different information compared to the analysis of contacts density maps or the analysis of structural and geometrical properties. In particular, we envisage that the energy-based domain identification proposed here may be a useful complement to the findings of other strategies and, for its nature, it may be used to provide experimentalists with immediately testable hypotheses (see the discussion of the results) on domains limits, boundaries, stability and isolation. Finally, as it will be discussed below, careful analysis of specific cases indicate that there is room for significantly improving the technique.

Theory

Underlying philosophy and general scheme. In the Energy Decomposition Method (EDM) recently developed by Colombo and coworkers[26-30], attention is focused on the protein non-bonded interaction energy matrix $\mathbf{E}^{nb}$ whose elements represent the non-bonded (namely, van der Waals and electrostatic) interaction energies between

residues. According to the EDM approach, the eigenvector ($\mathbf{v}_1$, also indicated as "first eigenvector") associated with the lowest eigenvalue (λ_1) of $\mathbf{E}^{nb}$ allows to identify most of the crucial aminoacids necessary for the stabilization of the protein fold[26-30]. In fact, using that eigenvector, it is possible to obtain a filtered non-bonded interaction energy matrix ($\mathbf{E}^{fold}$) that accounts for many of the essential residue-residue interactions responsible for the protein folding, namely we have:

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

where $\mathbf{v}_1^T$ is the transpose of the vector $\mathbf{v}_1$. This approximation usually holds for one-domain proteins, but in more complicated cases, such as multi-domain proteins, it has been observed that more eigenvectors are generally needed to completely represent all the interactions that are fundamental for the protein folding/stability. To better explain the reason why only one eigenvector is not sufficient for multi-domain proteins, in Figure 1 we have shown the eigenvector associated with the lowest eigenvalue of $\mathbf{E}^{nb}$ for proteins constituted by one, two, three and four domains. It is easy to observe that, except for the one-domain protein case (see Figure 1A), where almost all the components of the first eigenvector are significant, in all the other situations many components of this eigenvector are equal to zero (see Figures 1B, 1C and 1D) and, therefore, the eigenvector associated with the lowest eigenvalue does not contain by itself sufficient information in regards to the stabilization energy of the whole protein. 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

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

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

where N_e, namely the number of essential eigenvectors, is generally greater than N_D. The matrix $\mathbf{E}^{fold}$ is afterwards further filtered through a symbolization process to emphasize the significant non-bonded interactions and, finally, it is subject to a proper clustering procedure leading to the domains identification.

At this point it is worth noting an interesting parallelism. As the well-known essential dynamics strategy[32, 33] selects the most suitable eigenvectors of the atomic displacement covariance matrix to identify the correlated motions of large protein subunits, which can be also defined as dynamic domains[23], our novel method detects proper eigenvectors of the non-bonded interaction energy matrix to extract the fundamental interactions within each folding unit, so allowing another possible definition for the structural domains. It is due to this analogy that the selected eigenvectors and the matrix $\mathbf{E}^{fold}$ have been previously referred to as essential eigenvectors and essential folding matrix, respectively.

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

Before concluding, it is important to note that in this subsection we have described the general philosophy and scheme of our new strategy. All the details about the selection of the essential eigenvectors and about the symbolization of the essential folding matrix will be given in the following two subsections, while the clustering procedure will be examined in the “Methods” section.

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 **a**, **b** and **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 **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" **F**, namely a matrix with identical elements given by the following equation:

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

where N is the total number of residues and λ_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:

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

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

$$\mathbf{e} = \left\lfloor \frac{1}{\sqrt{\mathrm{N}}}, \frac{1}{\sqrt{\mathrm{N}}}, \mathrm{K}, \frac{1}{\sqrt{\mathrm{N}}} \right\rfloor \quad (6)$$

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[26-30] 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 (λ_1) of $\mathbf{E}^{nb}$ to obtain the matrix **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

$\mathbf{E}^{symb}$, different algorithms can be used and, therefore, it is not a distinctive aspect of the proposed technique. In the preliminary version of the program, whose results will be discussed in one of the next paragraphs, we have devised a non-optimized in-house clustering algorithm that will be briefly described in the "Methods" section.

Finally, for the sake of clarity, it is important to observe that our symbolized essential folding matrix $\mathbf{E}^{symb}$ is completely different from a classical protein contacts matrix (see Figure 4). In particular, if on the one hand it is true that the non-bonded energy interactions drop off quite fast with distance, on the other hand it is easy to observe that they survive for a longer range compared to the simple residue-residue contacts (see Figure 4). This indicates that $\mathbf{E}^{symb}$ does not only contain information about the density of interactions, but it also comprises important information about the interactions intensity and physico-chemical properties. Therefore, in this respect, our analysis provides complementary information that can enrich the one obtained by means of classical structure-based techniques.

Methods

The method described in the Theory section has been implemented in a computer program using the FORTRAN77 language and it has been afterwards tested considering the benchmark datasets recently constructed by Bourne and coworkers to assess the capabilities of algorithmic techniques[31].

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

the first step, all the possible small (user-defined in the input file) diagonal blocks in $\mathbf{E}^{symb}$ are examined and if their density is greater than or equal to the overall density (namely, the density of the whole matrix), they are taken into account (Figure 5A and Figure 5B). The second step consists in grouping the maximum number of consecutive diagonal blocks previously selected. To accomplish this task a starting diagonal block must be considered and the interaction density with its consecutive diagonal block (if it exists) is calculated (Figure 5C). If the interaction density is greater than or equal to the overall density, the two consecutive *tesserae* are merged in a temporary cluster (Figure 5D), otherwise the second diagonal block becomes the new starting unit. If a temporary diagonal cluster is obtained, the procedure is repeated, but now considering the interaction density between the current diagonal cluster and its possible consecutive diagonal block (Figure 5E). Therefore, at the end of the second step, diagonal clusters are obtained (Figure 5F), while the third step considers the possibility to merge them even if they are not consecutive (Figure 6A and Figure 6B). In particular, two generic clusters X and Y are merged if and only if the following three requirements are satisfied*: i*) the interaction density between X and Y is greater than or equal to the overall density; *ii*) X is the cluster that interacts most with Y (i.e., the interaction density between X and Y is greater than the interaction densities between any other cluster and Y); *iii*) Y is the cluster that interacts most with X. If two or more couples satisfy the previous conditions, we consider the couple with the highest interaction density. After each new merging the previous analysis is performed again and the procedure stops only when there is no other couples to be joined. In the fourth step, the clusters with dimension smaller than the minimum required size (user-defined in the input file) are discarded (Figure 6C and Figure 6D). Finally, in case of multi-fragment clusters, the gaps between the

fragments are analyzed (Figure 6E) and, if these gaps do not overlap with other clusters, they are merged with the multi-fragment cluster in exam (Figure 6F). Therefore, through the five steps just described, it is possible to obtain a first-level subdivision in domains, but it is important to note that each first-level domain is afterwards subject to the same five-step procedure with the only difference that the overall interaction density of the domain in exam replaces the overall interaction density of the whole protein. Of course, the clustering algorithm stops when all the *n*-level domains cannot be further subdivided into *(n+1)*-level substructures. Finally, it is worthwhile to observe that, since it is sometimes difficult to automatically decide which is the best partition among different possibilities, following the philosophy already adopted by Carugo[14] and by Berozovsky *et al.*[24, 25], the implemented algorithm provides all the levels of domain subdivision and let the user choose the proper level of resolution according to the particular circumstances.

Preparation of the Bourne benchmark datasets. The Bourne benchmarks considered in this paper are Benchmark_2 and Benchmark_3[31]. The former, which contains 315 chains and which has been assembled considering the consensus among three different expert methods in assigning the number of domains, is one of the most comprehensive consensus-based dataset of multi-domain proteins currently available to evaluate new or already existing automated strategies. The latter consists of 271 structures and it has been obtained removing from Benchmark_2 44 chains for which the expert methods disagree in assigning the domain boundaries. Half of each benchmark dataset is available on-line (http://pdomains.sdsc.edu) and they have been downloaded to perform the preliminary tests. Furthermore, 14 of the obtained structures have been discarded from Benchmark_2 (11 in the case of Benchmark_3)

mainly for two reasons: sequences of missing residues longer than 25 aminoacids or the presence of non-conventional residues in the crystallographic structures (see Table S1 in the Supporting Information for the complete list of discarded chains). Furthermore, in case of gaps shorter than or consisting of 25 residues, the missing aminoacids have been added by means of the following protocol: a) the sequences of missing residues have been designed manually by means of the molecular modeling environment Maestro[34] and they have been afterwards modeled as loops using MODELLER[35-38]; b) keeping frozen all the atoms non-belonging to the loops with the exception of the C_α atoms of the boundary residues (which have been constrained to their positions with force constants set equal to 100.0 $kJ \cdot mol^{-1} \cdot Å^{-2}$), the protein structure has been minimized through the package MacroModel[39] in implicit solvent (water), using the AMBER* force field and the default optimization parameters (conjugate gradient algorithm, van der Waals and electrostatic cutoffs set equal to 8.0 Å and 20.0 Å, respectively). Furthermore, during the optimization, the peptide bond dihedral angles of the modeled loops have been constrained to 180° (also in these cases the force constants have been set equal to 100.0 $kJ \cdot mol^{-1} \cdot Å^{-2}$).

Calculation of the non-bonded interaction energy matrices. In order to relax the starting crystal structures and any possible modeled loops, we have optimized the geometry of all the proteins in exam exploiting the AMBER9 package[40]. In particular, the ff03 force field and the steepest descent algorithm have been used and, furthermore, the solvation effects have been taken into account by means of the Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method. Therefore, the generic elements $\mathbf{E}_{ij}^{nb}$ of the protein non-bonded interaction energy matrices have been computed as the sum of three different terms:

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

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"

with respect to the consensus of expert methods); b) the fragmentation of domains; c) the boundary consistency, namely the measure of the overlap between the consensus domain boundaries detected by expert methods and the domain boundaries determined by our new approach.

About the first criterion, it is important to note that the errors can be classified as over-cuts (i.e., assigning more domains than the benchmark) and under-cuts (i.e. assigning fewer domains that the benchmark). Considering the available chains of Benchmark_2 as reference, we can observe (see Table 1) that the new method shows a very good accuracy (about 75%) in assigning the correct number of domains and a tendency to under-cut a few structures (17.5% of under-cuts). Furthermore, in order to fully characterize the capability of the new strategy in correctly determining the number of domains, Benchmark_2 has been afterwards subdivided into four subsets containing structures constituted by one, two, three and at least four domains, respectively. It is worth to note that we decided to group together the benchmark chains characterized by four, five and six domains in order to have a representative sample, even if the total number of structures in the obtained set still remains low (namely, six structures overall). As it can be seen in Table 1, the accuracy is quite high for one- and two-domain chains, but, as the number of domains in the examined structures rises, a significant number of under-cuts is observed. These results have urged us on investigating the reasons why the proposed technique has the tendency to underestimate the number of domains. One possible cause is the adopted clustering algorithm that is not able to properly “subdivide” the obtained symbolic essential folding matrix, as it can be seen, for instance, in the cases depicted in Figure 7. However, it is important to note that this drawback can be simply overcome improving the current non-optimized in-house clustering procedure or using an

already optimized algorithm, such as one of those associated with the other automated techniques for the domains identification. Another plausible reason for the under-cutting trend is connected to the specific physico-chemical bases of our approach that aims to identify all the possible significant non-bonded interactions. Therefore, if two (or more) domains significantly interact (especially in the frequent cases of compact protein structures or of compact interfaces between substructures that could be manually assigned to different domains), a single block is found to characterize the symbolic essential folding matrix and it becomes difficult to separate all the clusters corresponding to structural domains (see the examples represented in Figure 8). However, in these cases, our method indicates that certain interfaces take part in fundamental interactions and, hence, the proposed strategy might also provide additional information about the role of the interfaces in the structural stabilization of the protein. In fact, in these circumstances, the interface residues are responsible for a relevant amount of the stabilization energy necessary to define the specific 3D arrangement of the protein. Hence, it is conceivable to hypothesize that even if two protein fragments that share a large interface with extensive contacts have been assigned to different domains according to a certain set of rules, it may be possible that they are properly folded and structured *only* in the presence of those contacts. These situations are reminiscent of the case of the "obligate dimers", which have been properly studied in the context of protein-protein interactions[41]. In this case, our strategy, while providing different results in the assignment and partitioning compared to methods based on different criteria, gives direct insights into the importance of interface interactions in the stabilization of the 3D organization of a multi-domain protein. Such information may be used, for instance, to guide site-directed mutagenesis experiments aimed at disrupting or reinforcing specific interface

interactions in order to reveal their relative importance in domain and protein stabilization.

In Table 2 we have shown the performances of faster automated techniques (Domain Parser[8, 9], PDP[10] and PUU[12]) evaluated by Bourne and coworkers using the complete Benchmark_2 as reference[31]. For the sake of completeness, it is important to stress that Bourne and coworkers have obtained these results considering the complete Benchmark_2 and not its downloadable half that has been considered for our test calculations (see the "Methods" section). This is one of the reasons why we do not have a representative sample for four-, five- and six-domain proteins and, therefore, we have limited our comparisons with the other algorithmic approaches up to the three-domain case. From Tables 1 and 2 it is easy to observe that PDP is by far the strategy that overall provides the best agreement with the benchmark, while our method can be considered at the same level of Domain Parser and PUU. From a more detailed analysis that considers the one-, two- and three-domain subsets of Benchmark_2, it is possible to note that the new proposed approach is less reliable than the other automated techniques when it deals with one- and, especially, three-domain proteins, but it shows a very good accuracy for two-domain chains, resulting second only to the PDP strategy. However, it is also worth to observe that for three-domain proteins there is not a consensus among the other algorithmic methods taken into account since the agreement with the benchmark ranges from 53.7% (PUU) to 72.2% (PDP). This can be considered as an evidence of the uncertainty associated with the automated techniques for the domains identification.

As mentioned above, another important criterion to evaluate the capabilities of the new proposed strategy is the domains fragmentation[31]. It provides a measure of the balance between compact domains, which often correspond to the presence of many

fragments, and biologically meaningful domains, which are usually characterized by few fragments. Fragmentation is related to the three-dimensional organization of the primary sequence. In fact, a domain may consist of one single contiguous stretch of aminoacids, but, in many cases, polypeptide regions that are distant in sequence are physically close in the tertiary structure due to the specific folding of the protein. In the former case, isolating the domain from the whole protein would consist in cutting a continuous polypeptide sequence and the stabilization would be provided by intramolecular contacts among residues, which are kept together (namely, constrained in space) by the covalent bonds of the polypeptide chain. A stable single domain protein represents an extreme case of this situation. On the contrary, if many fragments were present, stabilizing the domain in isolation would require keeping together separated peptides in a specific geometric arrangement through non-covalent interactions. This would represent an entropically unfavorable condition compared to the previous case. As a consequence, a balance between a low number of fragments and their lengths with respect to the domain dimensions is required to properly stabilize the substructure.

It has been observed that the average fragmentation of domains generally correlates well with the average number of domains assigned by an algorithmic method and, therefore, this means that if an automated strategy assigns on average more domains, it also assigns on average more fragments per domain. In our case the average number of domains per chain and the average number of fragments per domain are 1.76 and 1.56, respectively. Comparing these values with the corresponding ones obtained for the expert consensus and for other algorithmic techniques[31], the former is generally lower while the latter is higher in all the situations. This result confirms the under-

cutting trend of the proposed strategy that tends to group many segments in the same cluster instead of identifying new domains.

Finally, in order to completely characterize the accuracy of the new method, we have taken into account the very stringent boundary consistency criterion that has enabled us to determine whether our domain assignments were identical or not to the corresponding consensus assignments provided by expert techniques. In particular, we have chosen an 80% boundary consistency according to which two assignments are considered identical only if they overlap for at least 80% of the chain length. Considering 124 structures of Benchmark_3 (see the subsection about the preparation of the Bourne benchmark datasets), it is possible to observe that the agreement with the expert consensus is at 56.1% and this increases to 65.9% taking into account a 70% boundary consistency (see Table 3). This trend can be observed much more clearly in Fig. 9 where, taking into account only the structures of Benchmark_3 for which our method assigns the correct number of domains, the fraction of chains with inconsistent boundaries is represented in function of the boundary consistency threshold: when the boundary consistency is set equal to 80% the fraction of inconsistent boundaries is 23.3% and it reduces to 14.4% and 10% for threshold values corresponding to 75% and 70%, respectively. These data show that in most of the cases in which the new technique does not reach the desired agreement with the benchmark dataset, the obtained results are not very far from the exact domain assignments (see examples in Fig. 10). Hence, we believe that the proposed strategy is well founded and that its accuracy and performance can be sensitively increased through targeted improvements of the algorithm.

The fact that a substantial fraction of protein domain decompositions has been reproduced using our new *ab initio* physical-chemistry based method, which takes

into account only the intramolecular non-bonded interaction energies of proteins, suggests that the energetic domain-decomposition can be profitably used in several applicative contexts. In particular, the energy-based domain definition could not only provide important insights into the determinants of the modular organization of protein structures and functional regions, but it may also be used as a comparative tool to highlight common features in protein families and super-families.

Moreover, in an applicative context, the knowledge of the relevant interactions for stabilization provided by our novel approach could be used to design experiments aimed at perturbing the stability of the domains and, therefore, to report on the structural or functional consequences of such perturbations. For example, in protein design the strategy could be exploited to suggest possible regions for the modification of protein sequences as well as to pinpoint locations where the sequence could be split into autonomously folding and structurally stable units. The rational generation/modification of stable protein constructs with specific functions may have important applications in several fields ranging from biotechnology to structural vaccinology.

Finally, the domains identified with our technique can be used to add biases in molecular simulations and in the coarse graining of protein structures for aiding the exploration of conformational spaces.

Conclusions

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

eigenvectors of the non-bonded interaction energy matrix, it is possible to highlight only those blocks of interactions that characterize each folding unit and, consequently, to detect the corresponding domains by means of a clustering procedure. This approach has been afterwards tested against benchmark datasets that were constructed to assess the capabilities and the features of automated methods. The collected results show that the novel strategy is quite accurate in determining the number of domains, despite its tendency to under-cut a few structures. Furthermore, a detailed analysis of the obtained results has suggested that there are two possible reasons to explain this shortcoming of the new method. One consists in the fact that the adopted clustering algorithm has the tendency to group together as many protein fragments as possible instead of identifying new clusters. This drawback can be overcome improving our non-optimized clustering procedure or implementing one of those optimized procedures already adopted by other algorithmic strategies. The second reason is that the novel technique tries to detect all the possible significant non-bonded interactions between aminoacids and, if the structure in exam is very compact, it becomes difficult to distinguish the different domains of the protein. If on one hand this can be obviously considered as an intrinsic bias of the proposed strategy, on the other hand, this can be also seen as the strength of the method since it provides relevant insights into the effective role played by specific interface residues and substructures in the stabilizing interactions required to obtain a specific 3D organization.

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

able to provide domain assignments not really different from the consensus ones and, therefore, only few algorithm refinements can significantly improve the results. These observations are important also considering the fact that our approach has not been previously trained against a suitable set of protein structures and it does not need any preliminary information to run, namely it can be considered as a real *ab initio* method. Furthermore, since the proposed strategy basically consists in analyzing the essential components of the residue-residue non-bonded interaction energy matrix, it is possible to get additional energy-information for the system under investigation. In particular, we can obtain insights into the energetic stability of the detected domains and a general picture of the folding energetics with the identification of all the possible crucial residues for the protein fold.

A limitation of the new approach might be represented by the associated computational expense since it entails a structure minimization to remove bad contacts from the starting crystal structure and a matrix diagonalization. However, it is worthwhile to point out that this can be considered a serious problem only for very large proteins and that this can be also considered as an additional cost to get more information compared to the manual or the usual compactness-based algorithms. Furthermore, nowadays, the fast pace in methodological developments and the continuous increase in computer power allow to perform long time scale all-atom Molecular Dynamics simulations of large systems[42-44] and, therefore, the use of only slightly computational demanding computational techniques does not represent a limiting drawback.

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

algorithmic technique is the most reliable and, therefore, we believe that our method may be used in combinations with other strategies either to provide a consensus view or to enrich existing assignment techniques with physico-chemical and energetic insights. Moreover, the new approach might be very useful in aiding research in several applicative and experimental contexts.

Acknowledgments

We thank A.I.R.C. (Project MFAG 11775) and the Cariplo VACCINI GtA and LOMBARDY ASTIL projects for financial support. We would like to thank Guido Scarabelli, Rubben Torella and Massimiliano Meli for helpful discussions.

Supporting Information Available

Table S1: complete list of discarded chains of Benchmark_2 and Benchmark_3; Table S2: identified domains and number of selected eigenvectors for all the examined structures; Complete references: 5, 40 and 42. This material is available free of charge via the Internet at http://pubs.acs.org.

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.

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.

38. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudham, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U.; Sali, A. *Curr. Protoc. Bioinformatics* **2006**, *5*, Unit 5.6.

39. *Macromodel, Version 9.5*; Schrödinger LLC: New York, 2007.

40. Case, D. A.; Darden, T. A.; Cheatman, T. E. III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M. Jr.; Pearlman, D. A.; Crowley, M. et al. *AMBER, Version 9*; University of California – San Francisco: San Francisco, 2009.

41. Tuncbag, N.; Gursoy, A.; Guney, E.; Nussinov, R.; Keskin, O. *J. Mol. Biol.* **2008**, *381*, 785-802.

42. Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B. et al. *Science* **2010**, *330*, 341-346.

43. Shan, Y. B.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181-9183.

44. Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhami, D. W.; Maragakis, P.; Shan, Y. B.; Xu, H.; Shaw, D. E. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13118-13123.

Table 1 – Performance of the method in assigning the correct number of domains (reference: available Benchmark_2). The results are shown for all the possible structures (*Overall*) and for subsets containing chains constituted only by one, two, three or at least four domains.

	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

Table 3 – Performance of the method in assigning the correct domain boundaries (Reference: available Benchmark_3) considering different values for the Boundary Consistency threshold.

	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)).

Figure 4. 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).

Figure 5. 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 *tesseare* 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.

Figure 6. 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.

Figure 7. 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).

Figure 8. Examples of cases in which the under-cutting trend is due to significant interactions in compact structures. (A) 1tuia: 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).

Figure 9. 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.

Figure 10. Examples of non-accurate domain boundaries assignment: (A) 1bc5a (boundary consistency: 71%): benchmark assignment (15-75 (blue) // 76-269 (red)), symbolized essential folding matrix, new strategy assignment (6-75; 131-150 (blue) // 81-125; 186-265 (red)); (B) 1c0ma (boundary consistency: 78%): benchmark assignment (1-167 (blue) // 168-221 (red)), symbolized essential folding matrix, new strategy assignment (11-130 (blue) // 146-221 (red)). Note that our residues numeration is different from the one in the Bourne database (see Supporting Information to reconstruct it).

Figure 1-

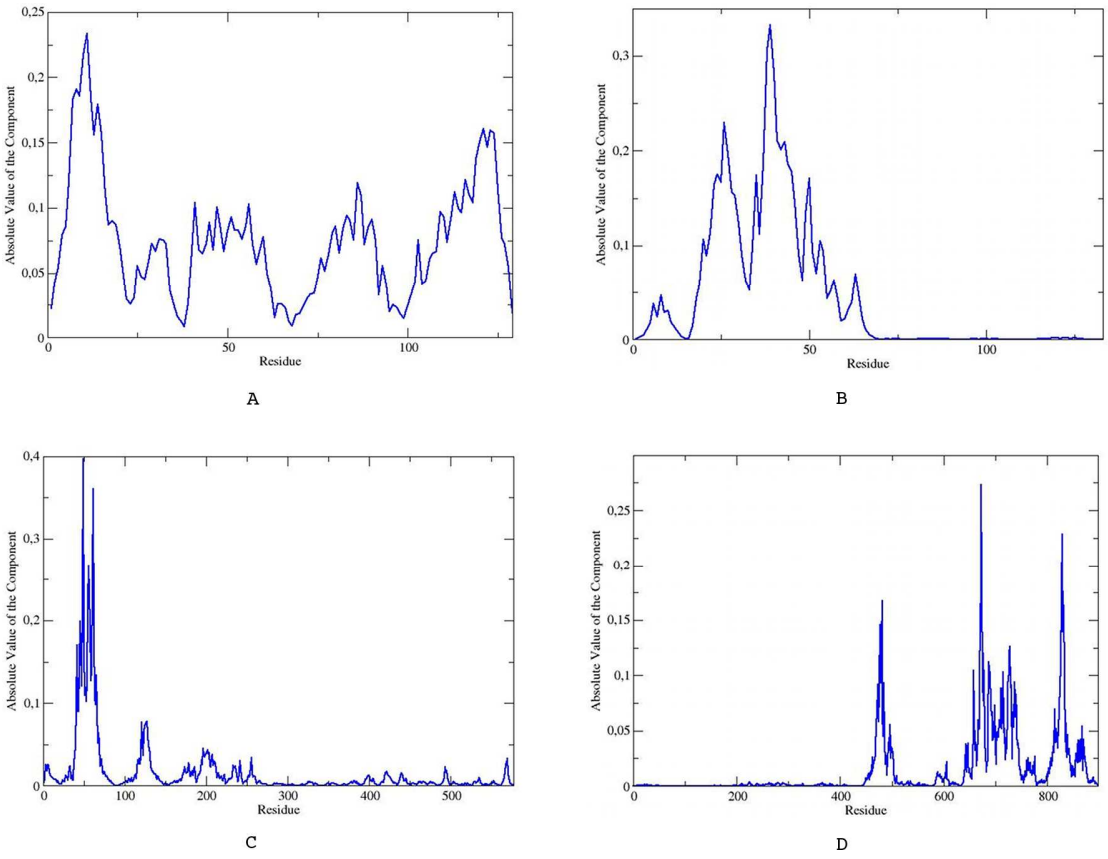


Figure 2 –

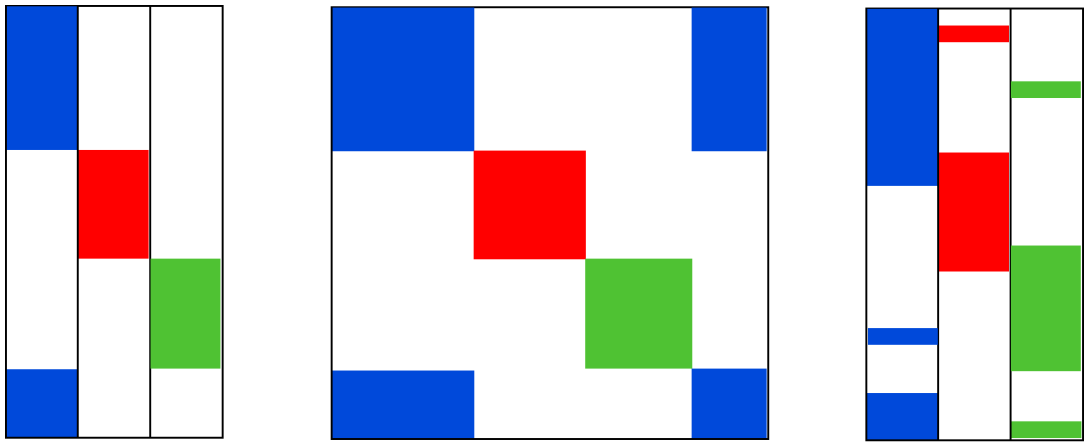


Figure 3 –

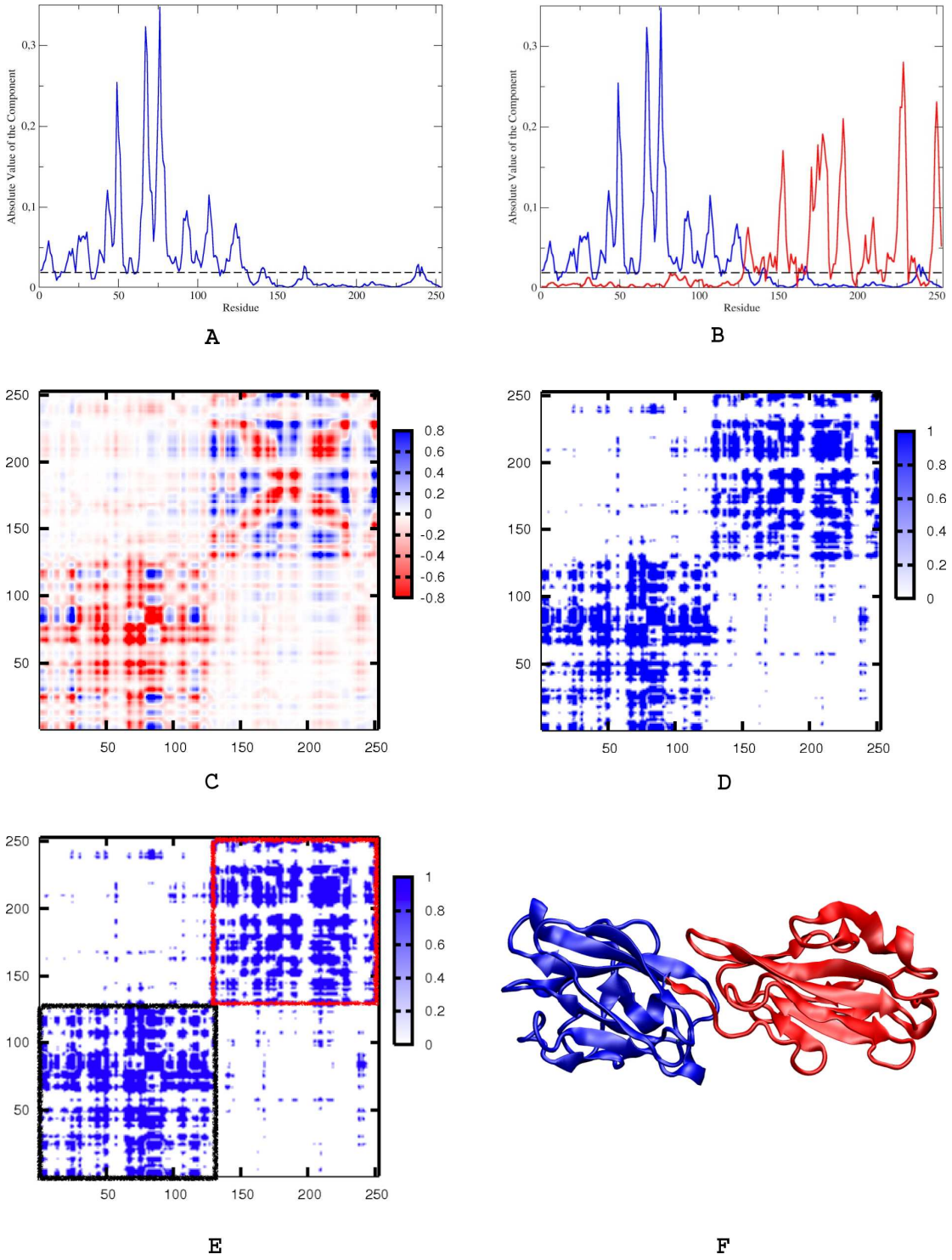


Figure 4-

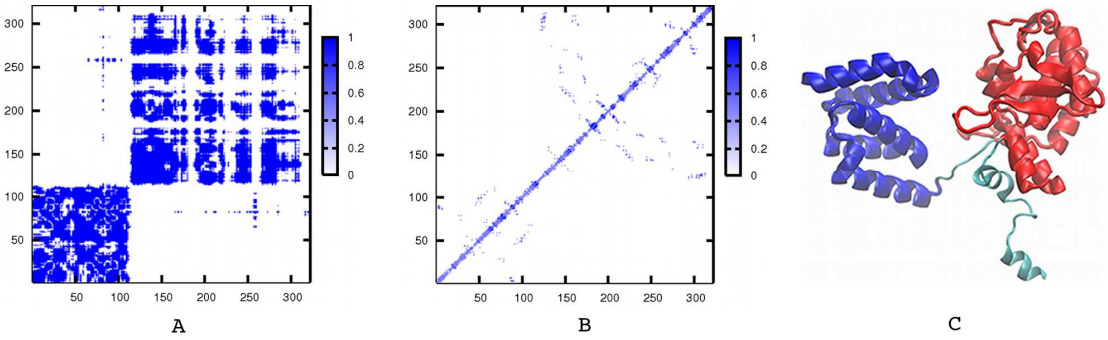


Figure 5 –

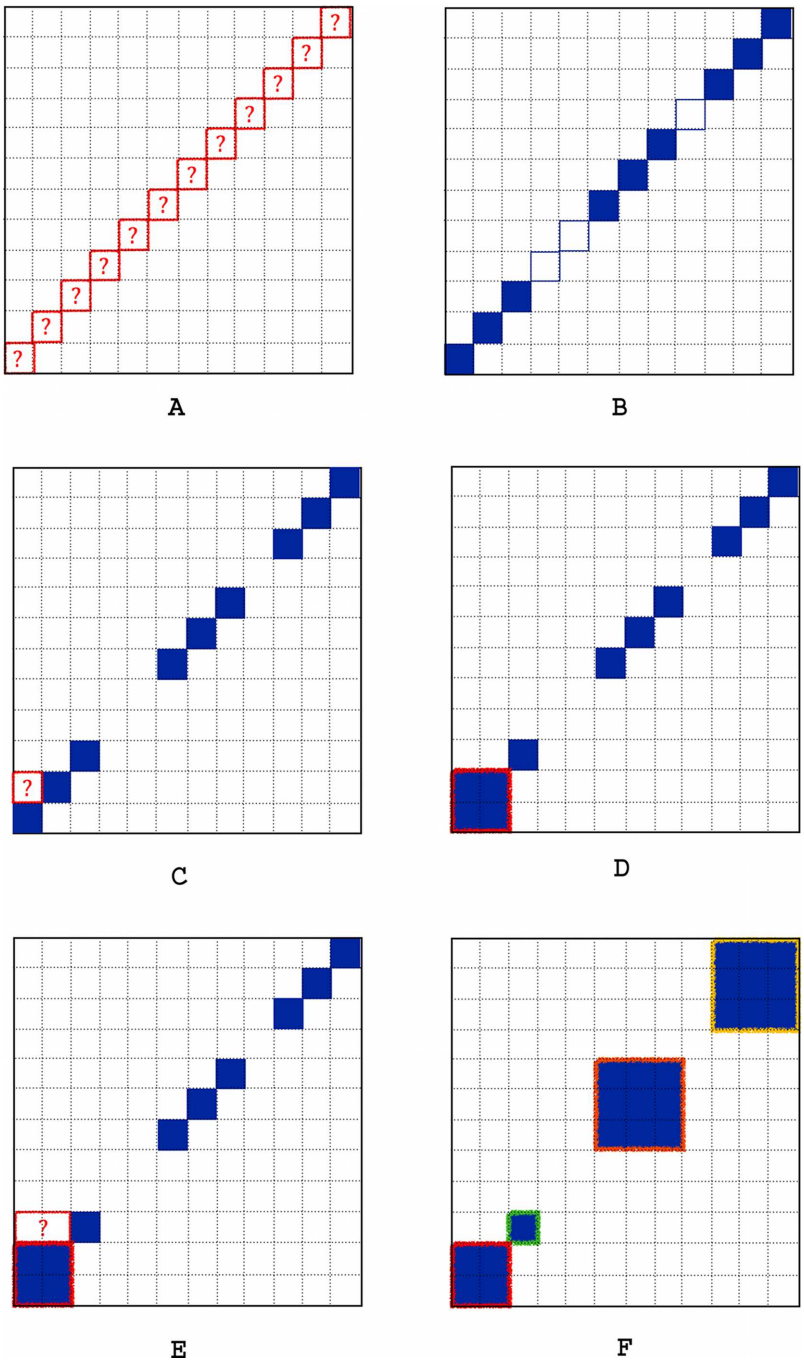


Figure 6 –

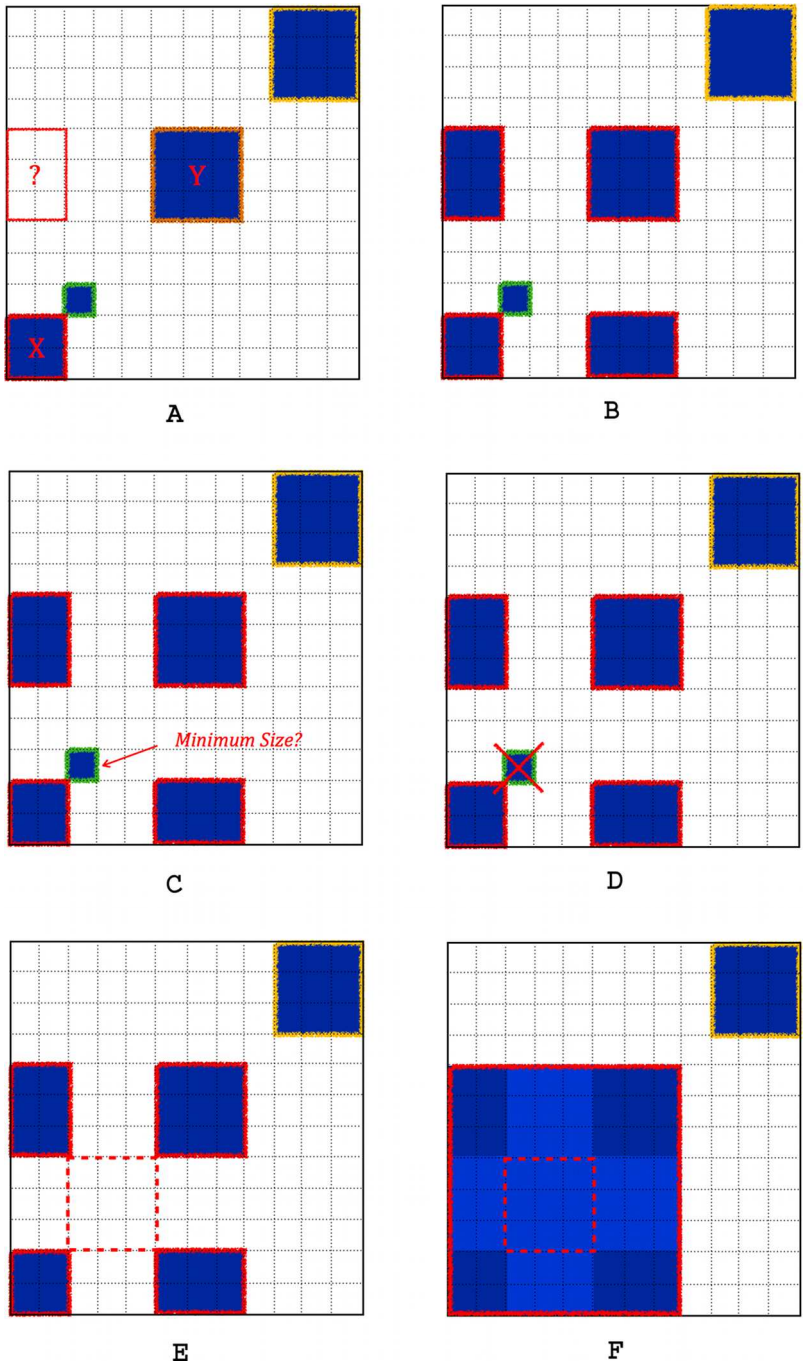


Figure 7 –

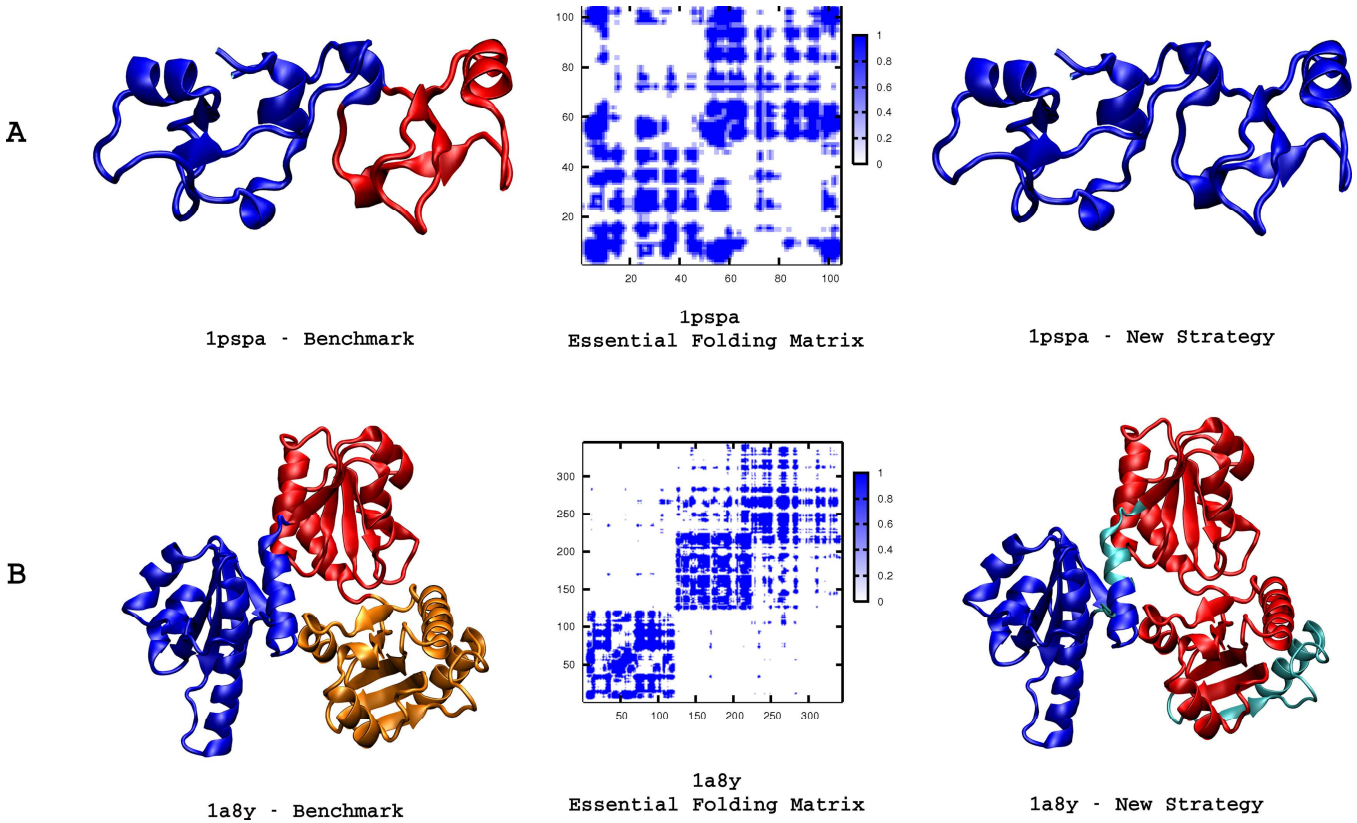

A
1pspa - Benchmark
1pspa
Essential Folding Matrix
1pspa - New Strategy
B
1a8y - Benchmark
1a8y
Essential Folding Matrix
1a8y - New Strategy

Figure 8 –

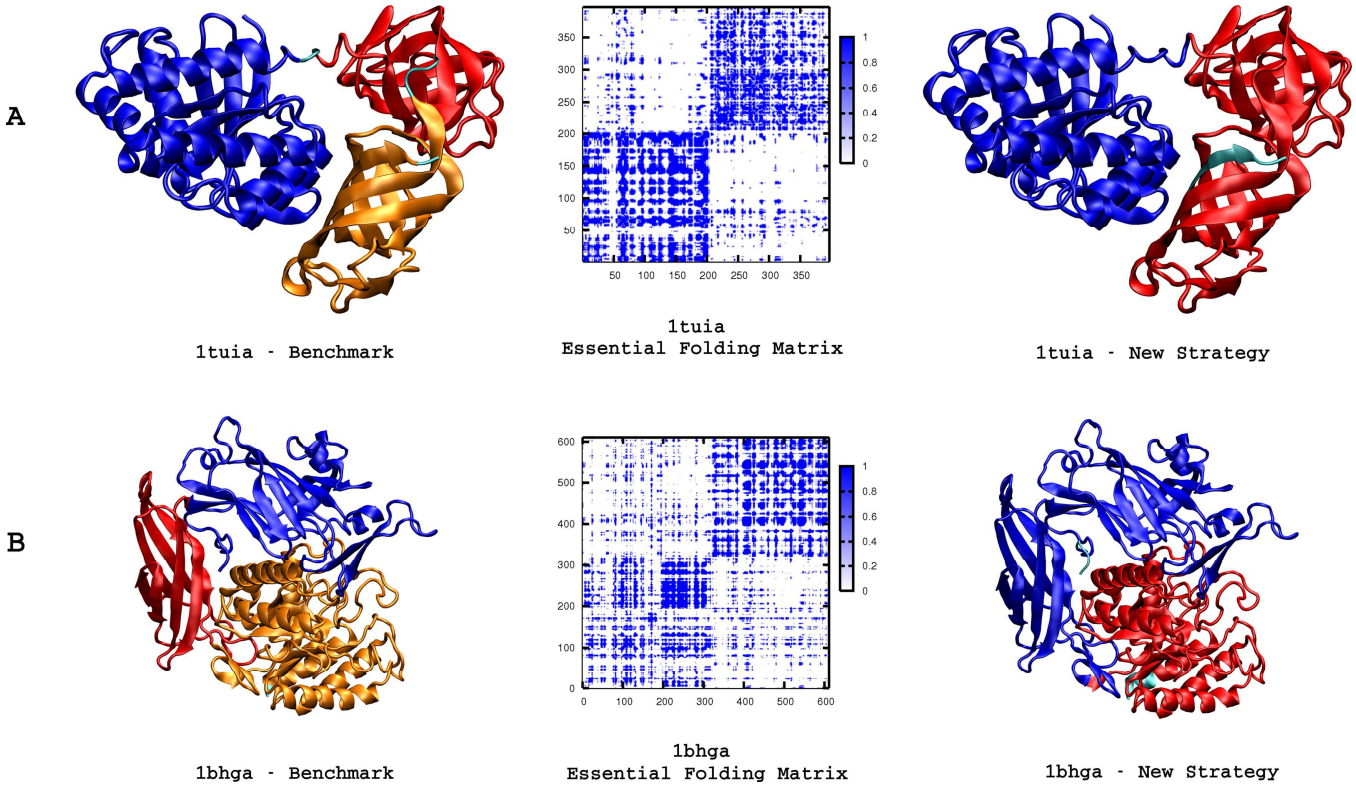

A
1tuia - Benchmark
1tuia
Essential Folding Matrix
1tuia - New Strategy
B
1bhga - Benchmark
1bhga
Essential Folding Matrix
1bhga - New Strategy

Figure 9 –

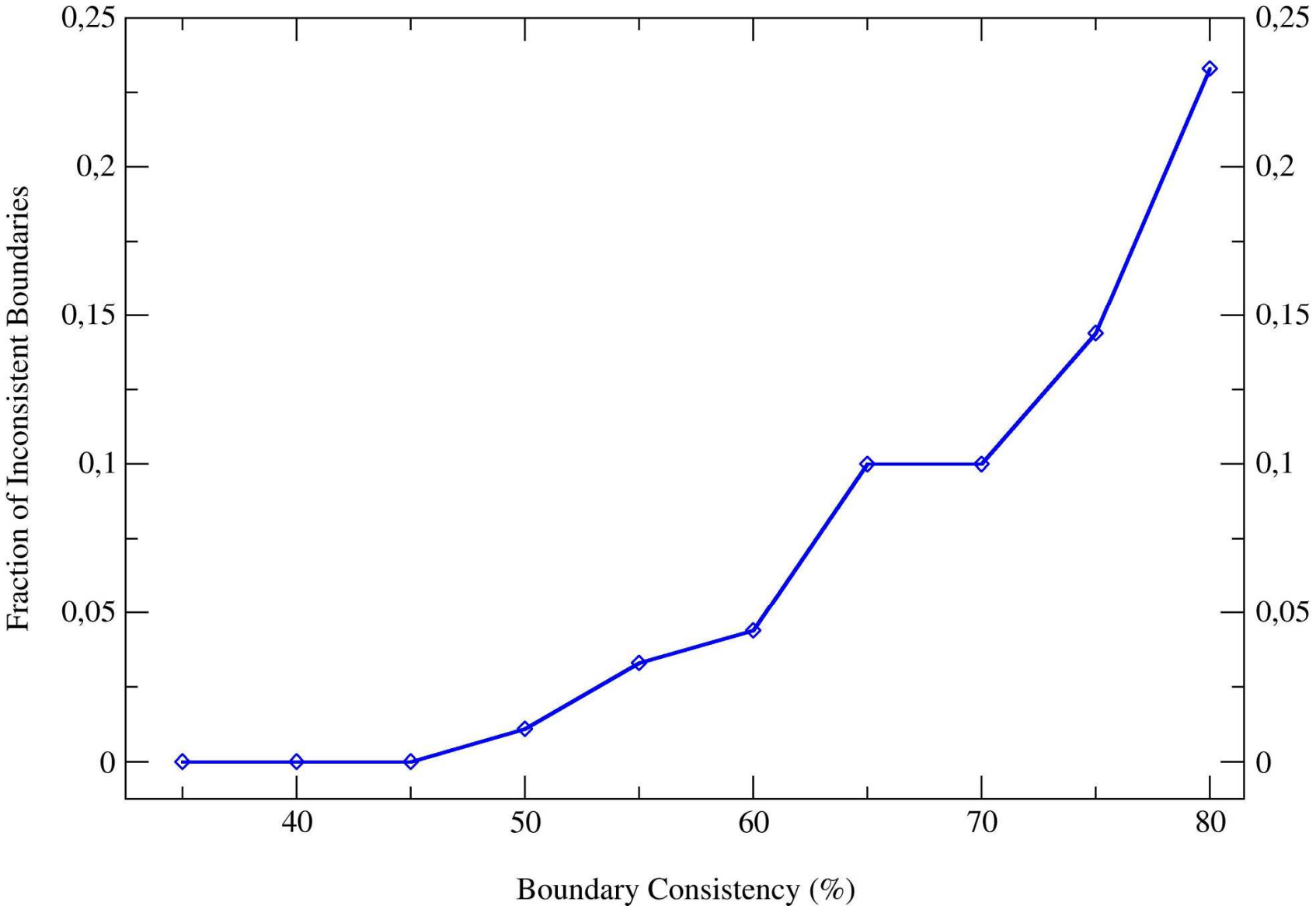

0,25
0,2
0,15
0,1
0,05
0
Fraction of Inconsistent Boundaries
40
50
60
70
80
Boundary Consistency (%)

Figure 10 –

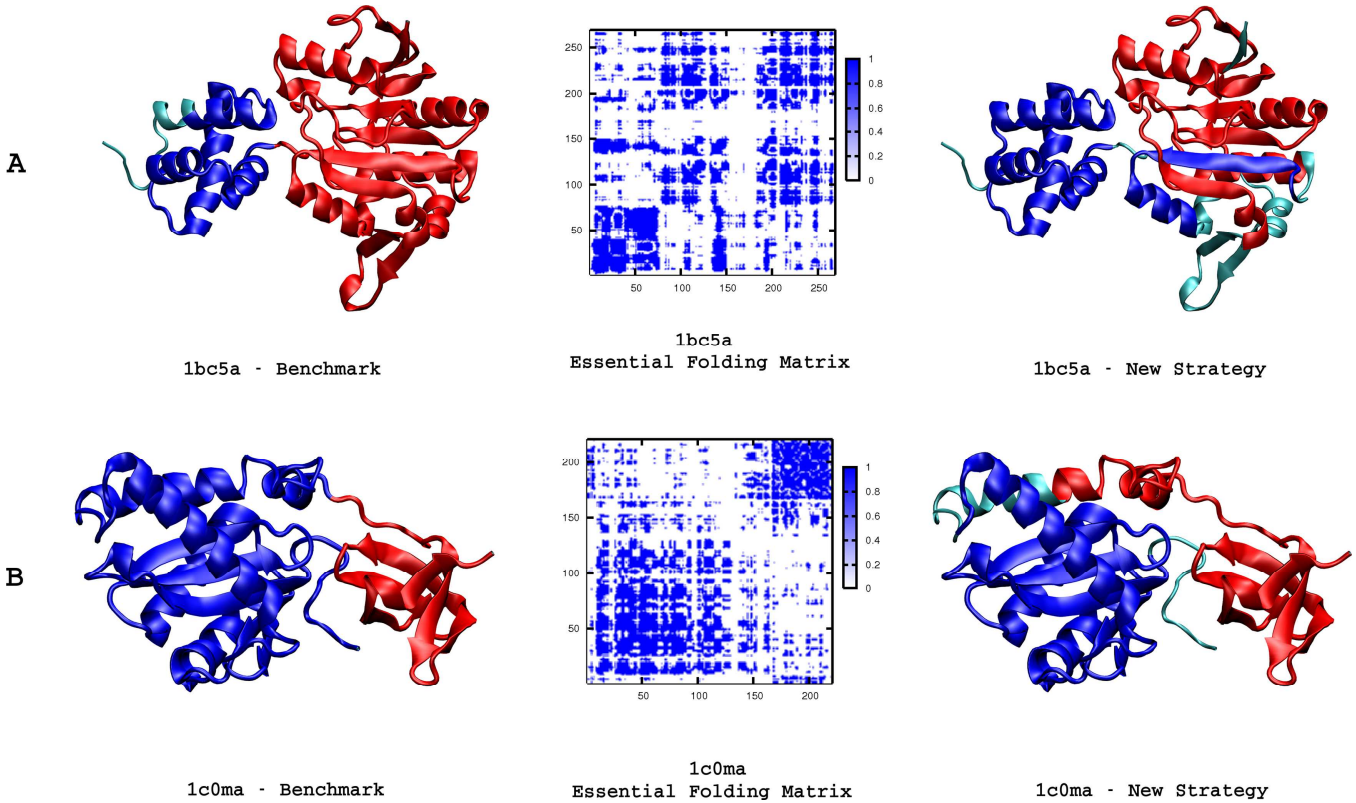
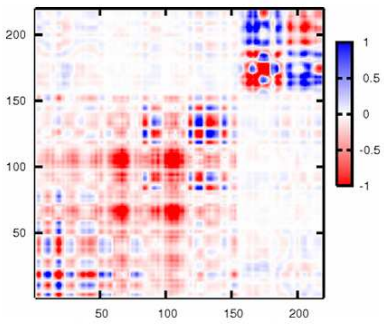
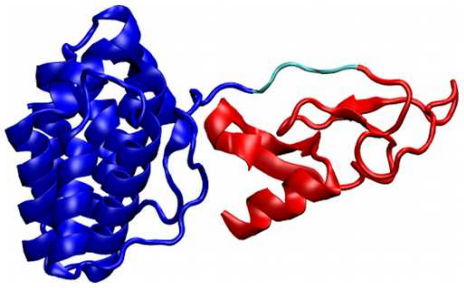

A
1bc5a - Benchmark
1bc5a
Essential Folding Matrix
1bc5a - New Strategy
B
1c0ma - Benchmark
1c0ma
Essential Folding Matrix
1c0ma - New Strategy

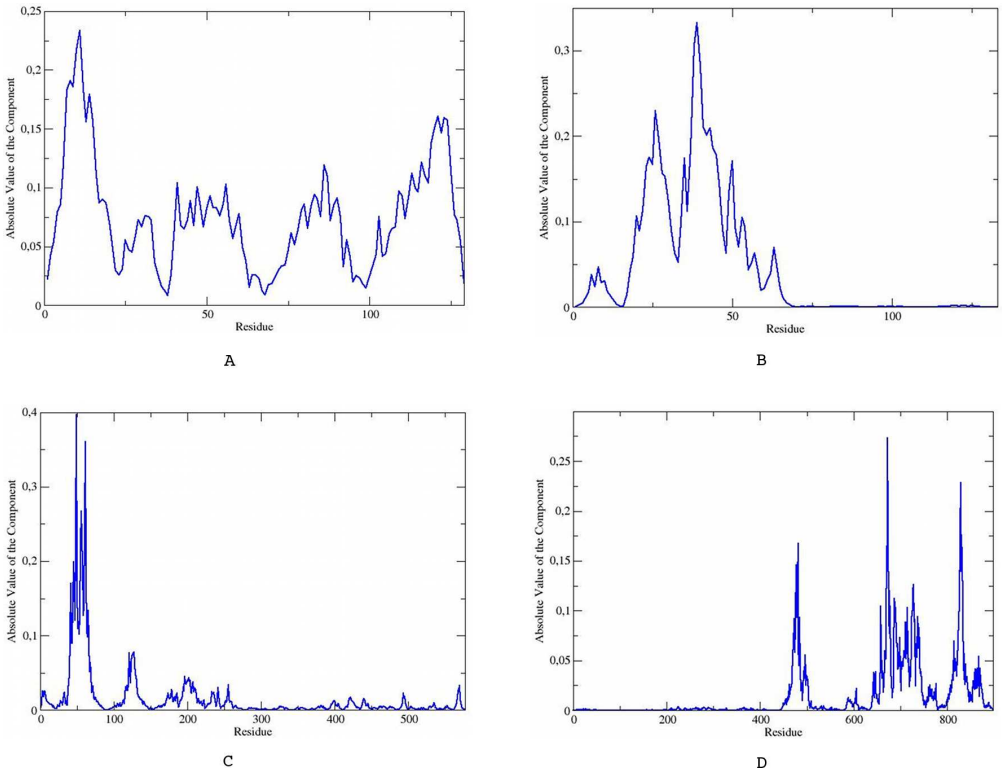
Table of Contents Image –



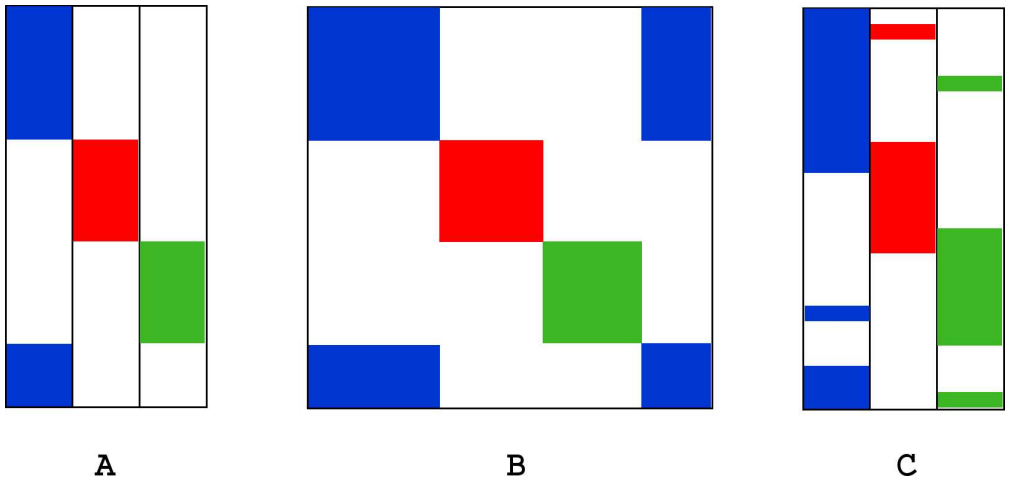
Essential Non-Bonded Interactions



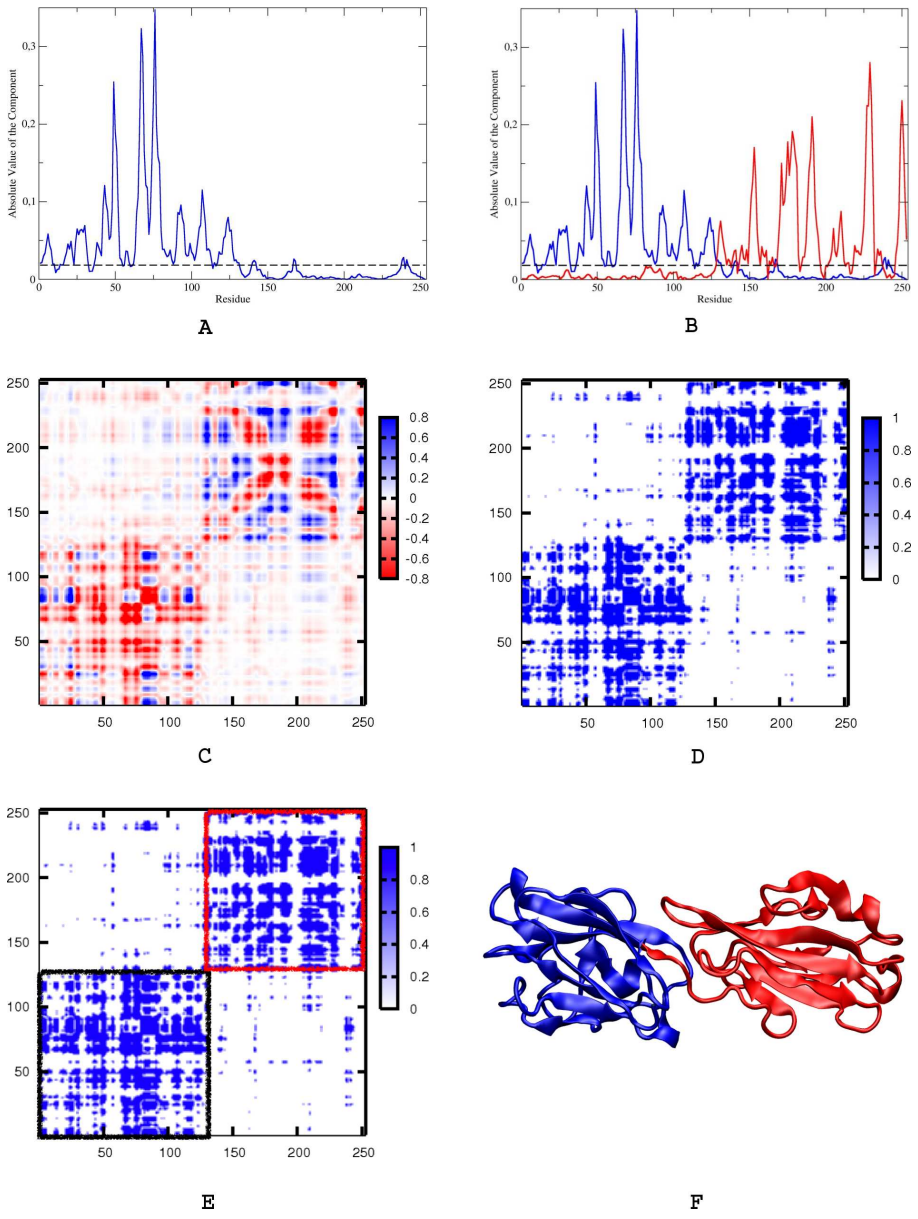
Protein Domains



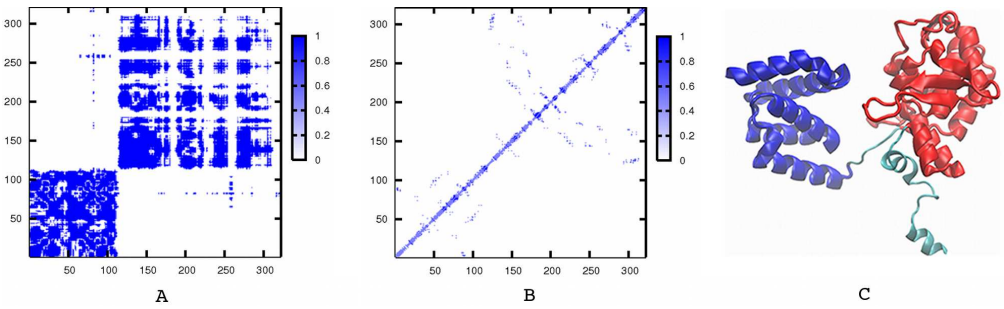
Components (in absolute value) of the eigenvector associated with the lowest eigenvalue of the non-bonded interaction energy matrix for (A) the one-domain chain 1rcb, (B) the two-domain chain 6paxa, (C) the three-domain chain 1ciy and (D) the four-domain chain 1dgkn.
150x114mm (300 x 300 DPI)



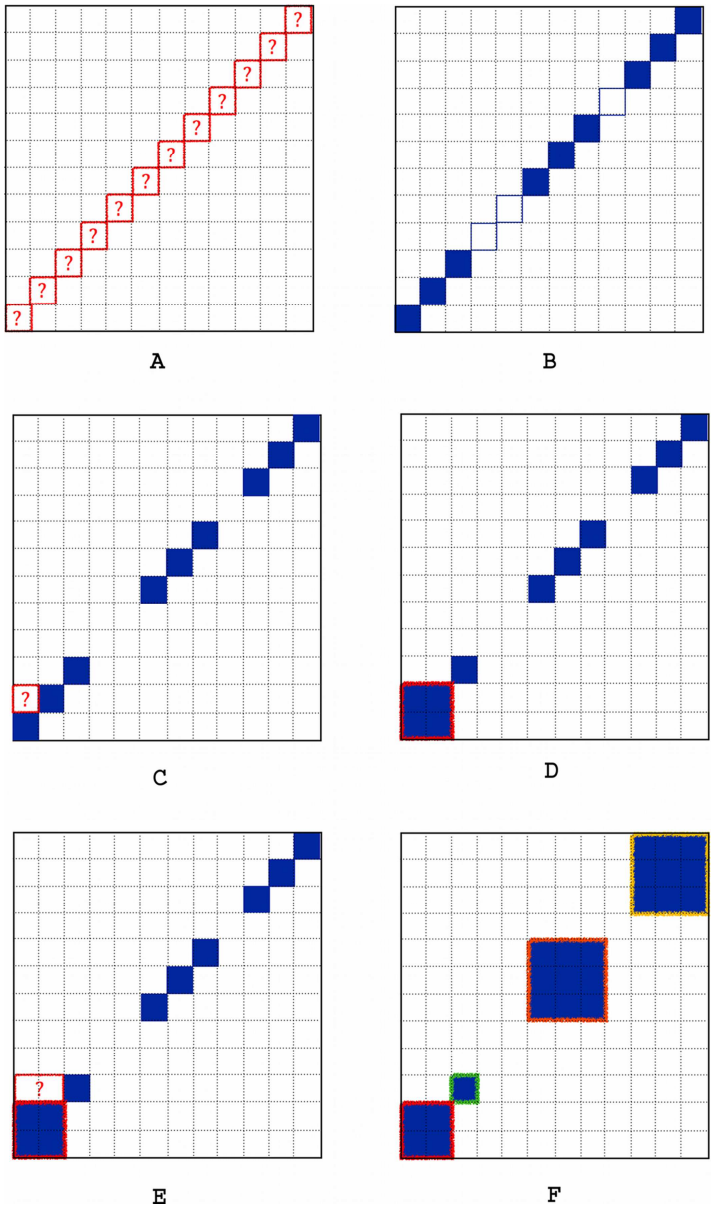
Working hypotheses. (A) Ideal situation for a three-domain protein: each domain is associated with an eigenvector with a block structure of significant components (here depicted in blue, red and green for the first, the second and the third eigenvector, respectively). The blocks do not overlap and their 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)



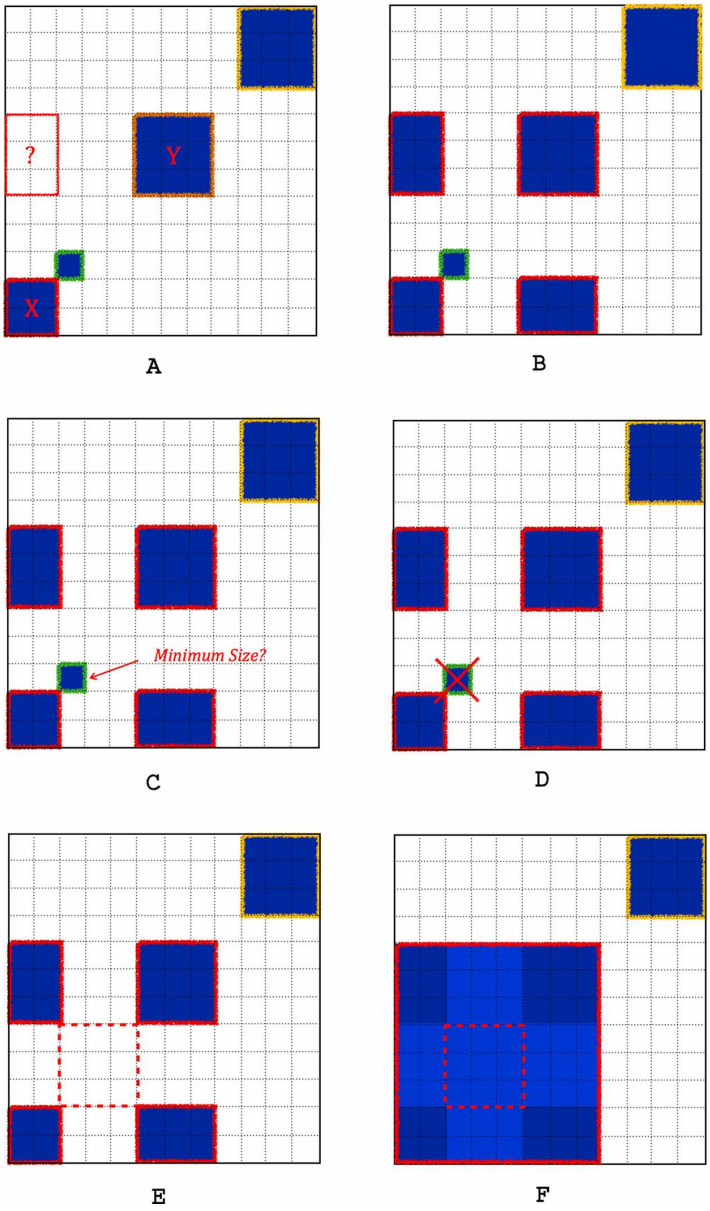
Selection of the essential eigenvectors, construction of the essential folding matrix and domains identification for the 1kzq protein: (A) components (in absolute value) of the eigenvector associated with the lowest eigenvalue (blue line) and significance threshold for the eigenvectors components (dotted 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)



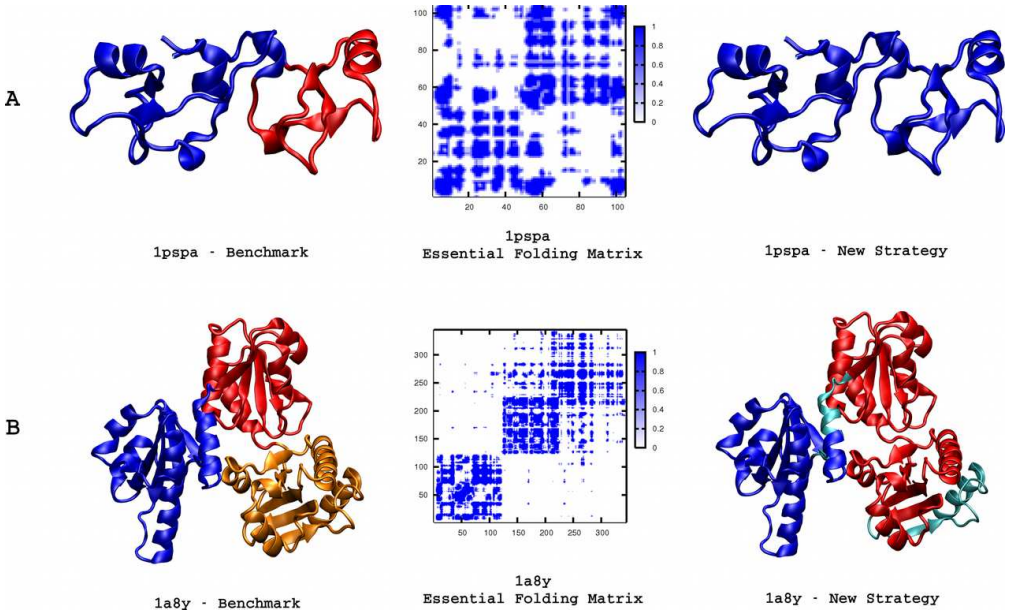
Comparison between the symbolized essential folding matrix (A) and the contacts matrix (D) of the 1crxa protein (C). The domains identified by the new technique are depicted in blue (residues 1-110) and red (residues 116-295). Note that our residues numeration is different from the one in the Bourne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



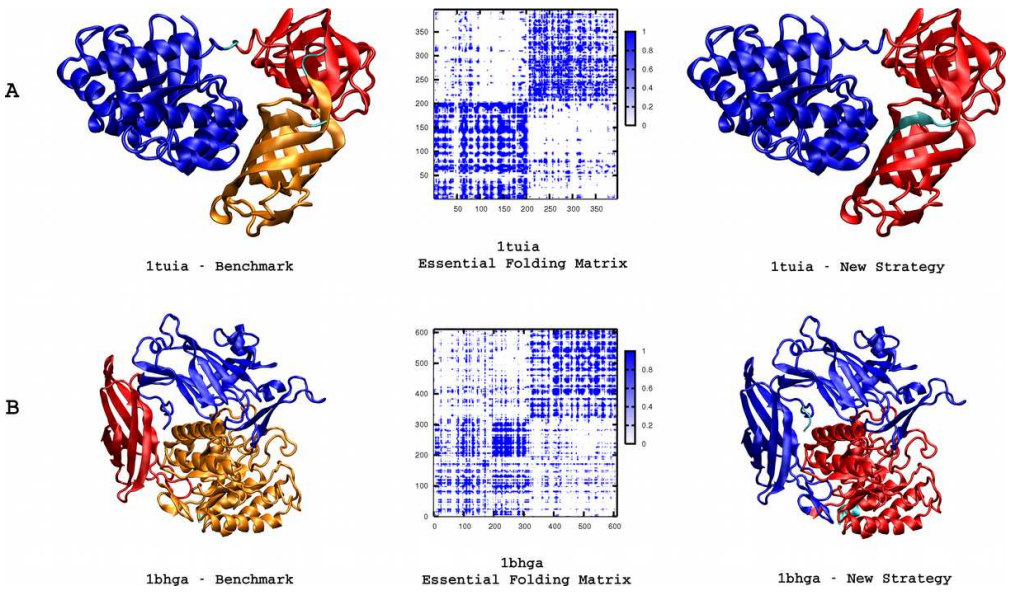
Clustering algorithm (early stage): (A) the densities of the small diagonal blocks are computed; (B) the small diagonal tesserae whose density is greater than or equal to the overall density are considered; (C) the interaction density between two consecutive diagonal blocks is examined; (D) two consecutive diagonal tesseare 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)



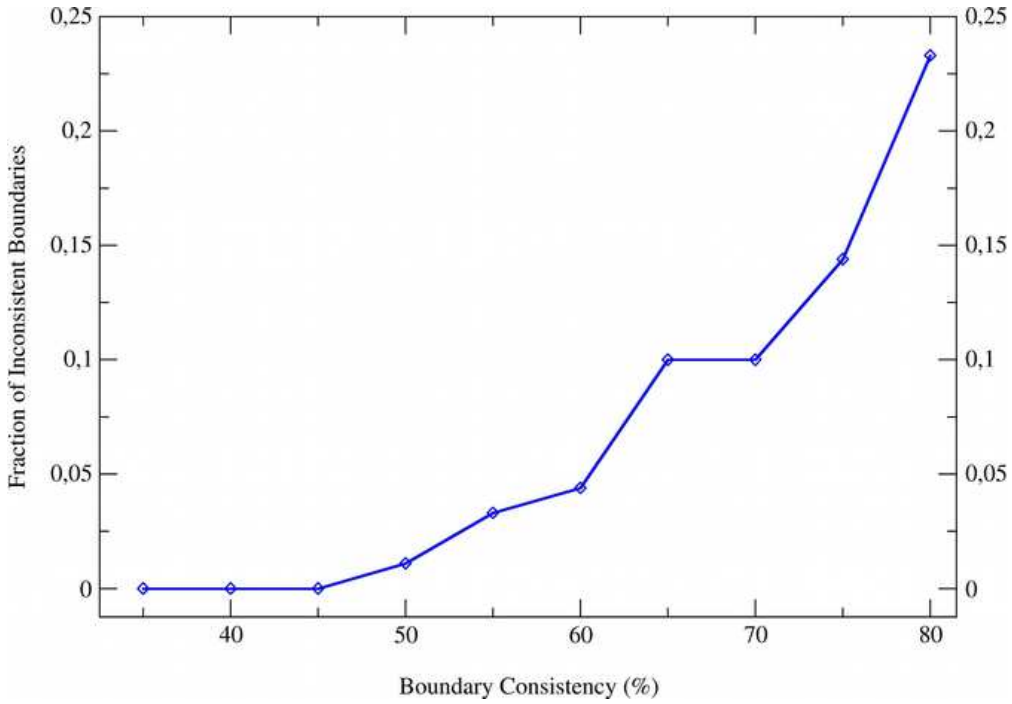
Clustering algorithm (late stage): (A) the possibility to join non-consecutive diagonal clusters is taken into account; (B) clusters obtained after merging all the possible non-consecutive diagonal blocks; (C) the clusters size is evaluated; (D) clusters with dimension smaller than the minimum required 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)



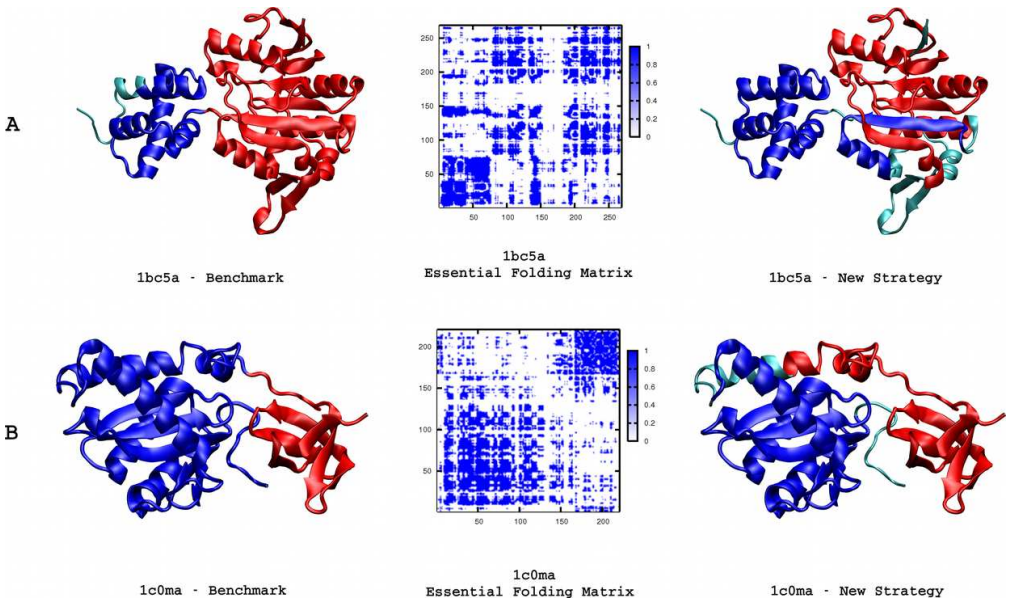
Examples of cases in which the under-cutting trend is due to the clustering algorithm. (A) 1pspa: benchmark assignment (1-98; 99-105 (blue) // 59-98 (red)), symbolized essential folding matrix, new strategy assignment (1-105 (blue)). (B) 1a8y: benchmark assignment (1-124 (blue) // 125-226 (red) // 227-345 (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)



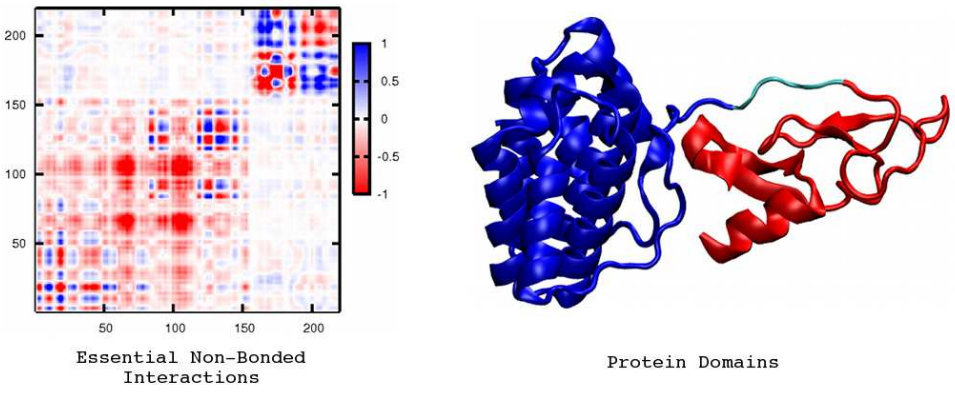
Examples of cases in which the under-cutting trend is due to significant interactions in compact structures. (A) 1tuia: benchmark assignment (1-205 (blue) // 208-299 (red) // 303-396 (orange)), symbolized essential folding matrix, new strategy assignment (1-210 (blue) // 211-390 (red)); (B) 1bhga: benchmark 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)



Examples of non-accurate domain boundaries assignment: (A) 1bc5a (boundary consistency: 71%): benchmark assignment (15-75 (blue) // 76-269 (red)), symbolized essential folding matrix, new strategy assignment (6-75; 131-150 (blue) // 81-125; 186-265 (red)); (B) 1c0ma (boundary consistency: 78%): benchmark assignment (1-167 (blue) // 168-221 (red)), symbolized essential folding matrix, new strategy assignment (11-130 (blue) // 146-221 (red)). Note that our residues numeration is different from the one in the Bourne database (see Supporting Information to reconstruct it).
104x61mm (300 x 300 DPI)



88x35mm (300 x 300 DPI)