Published in final edited form as:
Proteins. 2011 March ; 79(3): 853–866. doi:10.1002/prot.22923.

Protein domain assignment from the recurrence of locally similar structures

Chin-Hsien Tai[1], **Vichetra Sam**[2], **Jean-Francois Gibrat**[3], **Jean Garnier**[2,3], **Peter J. Munson**[2], and **Byungkook Lee**[1]

Chin-Hsien Tai: taic@mail.nih.gov; Vichetra Sam: vichet.sam@gmail.com; Jean-Francois Gibrat: Jean-Francois.Gibrat@jouy.inra.fr; Jean Garnier: jean.garnier@jouy.inra.fr; Peter J. Munson: munson@helix.nih.gov; Byungkook Lee: BKLee@mail.nih.gov

[1] Laboratory of Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

[2] Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA

[3] INRA, Unité Mathématique Informatique et Génome UR1077, Jouy-en-Josas 78350, France

Abstract

Domains are basic units of protein structure and essential for exploring protein fold space and structure evolution. With the structural genomics initiative, the number of protein structures in the Protein Databank (PDB) is increasing dramatically and domain assignments need to be done automatically. Most existing structural domain assignment programs define domains using the compactness of the domains and/or the number and strength of intra-domain versus inter-domain contacts. Here we present a different approach based on the recurrence of locally similar structural pieces (LSSPs) found by one-against-all structure comparisons with a dataset of 6,373 protein chains from the PDB. Residues of the query protein are clustered using LSSPs via three different procedures to define domains. This approach gives results that are comparable to several existing programs that use geometrical and other structural information explicitly. Remarkably, most of the proteins that contribute the LSSPs defining a domain do not themselves contain the domain of interest. This study shows that domains can be defined by a collection of relatively small locally similar structural pieces containing, on average, four secondary structure elements. In addition, it indicates that domains are indeed made of recurrent small structural pieces that are used to build protein structures of many different folds as suggested by recent studies.

Keywords

domain parsing; structure comparisons; clustering; LSSP; secondary structures; symmetric matrix factorization; singular value decomposition; pairwise correlation method

Correspondence to: Byungkook Lee, `BKLee@mail.nih.gov`.

Program availability

The source codes of the programs are available by writing to one of the authors. They are also available for download at the MIG web site: http://mig.jouy.inra.fr/mig2en/. A server (DOMIRE for DOMain Identification from REcurrence) is in preparation for running VAST (see: http://genome.jouy.inra.fr/vast/) and then any of the three procedures: SVD, SMF, or PCM. In addition to domain boundaries, this server will also provide for each query a list of the Locally Similar Structural Pieces (LSSPs) and a list of filtered targets according to their similarity to the query {R. Shrager et al., in preparation).

INTRODUCTION

More than 30 years ago, Wetlaufer [1] reported the existence of distinct domains in protein three-dimensional (3D) structures. Based on kinetic arguments, he suggested that the presence of these domains could be the consequence of independent nucleation processes in the early steps of the 3D structure formation. Since then, structural domains have been considered the building blocks of protein 3D structure and the basic units of folding, function and evolution. They are the cornerstone on which different protein structure classifications [2–4] are built and play a key role in proteomics initiatives [5] and various fields of bioinformatics such as fold recognition techniques or protein docking methodologies. Protein domains are also defined from the conservation of amino acid sequence [6–8], but we will be concerned with structural domains in this study.

Despite their fundamental role, there is a conspicuous lack of agreement in the definition of a domain, as pointed out by different authors [9–12]. According to the viewpoint adopted, domain definitions can be divided broadly into 3 categories:

- Geometric: a domain consists of one or several stretches of the polypeptide chain that form distinct and compact substructures within the protein that potentially could fold independently if cleaved from the protein;
- Functional: a domain is a unit of the structure that has a defined biochemical function;
- Recurrence: a domain is a recurrent substructure found in different proteins that has been preserved through evolution.

The vast majority of automatic or semi-automatic methods that partition 3D structures into domains are based on geometric criteria. Examples of geometric criteria that have been implemented are, for instance, the maximization of intra-domain and minimization of inter-domain contacts [13]or, alternatively, the definition of a minimal surface area for the interface between domains. Physicochemical considerations that have additionally been employed in several works include: domains should be compact with a hydrophobic core [14]; they should not consist of too many discontinuous pieces; domain boundaries should not cut secondary structure elements, in particular beta sheets; they should have a minimum size (fixed to about 30–40 residues in most methods).

A number of methods have implemented these different criteria, using either a top-down [15–18] or a bottom-up [14,19–21] approach. In the former, the method starts with the complete protein structure and recursively divides it into substructures until some stopping criterion is met. In the latter, residues are hierarchically clustered into larger groups. The list of methods given above is by no means exhaustive and the interested reader is referred to a recent review [22] and to chapter 6 of Todd Taylor's PhD dissertation [23] for a more comprehensive list of domain assignment methods and a concise description of their principles.

SCOP domains are defined manually and it is difficult to know exactly what factors are taken into consideration. However, numerous examples of SCOP domains suggest that both function and recurrence were considered. To the best of our knowledge, no automatic method depends solely on such an approach. Only the Dali Domain Dictionary (DDD) [4] makes use of whole domain recurrences to complement their otherwise geometry-based domain definition procedure, PUU [18].

Here we present methods that rely on a variant of the recurrence category, wherein domain assignments are made from a large collection of structurally similar pieces found in other protein structures. For this purpose, VAST [24,25] was used to compare each query protein to a representative set of 6,373 target protein chains from the PDB. A VAST comparison yields *cliques*, which are sub-structures of three or more secondary structure elements (SSE) that match between two protein structures (see Methods and Appendix to Sam et al. [26]). We collected a large number of cliques using very generous selection criteria from the VAST comparisons (as described in Methods) between each query and the whole of the target set. The locally similar structural pieces (LSSP) are the query sub-structures represented in these cliques (see Fig. 1). This collection of LSSPs was used to define domains in the query structure. We do not consider biochemical functions, thus departing from the domain definition used by SCOP. The objective was to explore the connection between collections of LSSPs and the notion of structural domains. We present three methods for this exploration, which employ different ideas on how best to translate the collection of LSSPs into concrete domain identification.

We tested these methods on a benchmark dataset of 128 proteins in the Balanced_Domain_Benchmak_3 of pDomain [27]. Although these methods are based on a simple principle, we find that they can produce domain definitions that are comparable to those of DomainParser [15], PDP [17] and PUU [18], which are all based essentially on geometric criteria. In many individual cases examined, the procedures gave results that were quite similar to the domains assigned manually. At first sight, it appears rather surprising that small structural pieces can collectively define full domains, but a number of works on protein fold evolution [28–31]have recently drawn attention to the importance of such small locally similar structural motifs.

METHODS

Target set

Protein chains that contribute to ASTRAL SCOP 1.71 domains with less than 40% sequence identity to each other (ASTRAL40 set) were collected. Forty-nine of these from obsolete PDB structures were excluded. When a domain consisted of more than one chain, all the chains of the domain were included. This gave us a total of 6373 chains as target set of structures. The list of the chain names is in the Supplementary Material (6373ChainNames.pdf).

15-chain set

A set of 15 protein chains was used during the development of our algorithms. The set includes both one-domain and multi-domain chains and chains with segmented domains, which are domains made of sequences that are not all contiguous in the primary sequence. Their sizes range from less than 70 to more than 1200 residues. Four are one-domain chains with less than 250 residues in the structure: 1aluA, 1avgI, 1b67A and 1fcyA. Another two one-domain chains are 1a0tP and 1qbkB, both of which are made of repeating elements: 1a0tP is a transmembrane beta-barrel, which is made of repeating beta-hairpins; 1qbkB is an alpha-alpha superhelical structure, which is made of alpha-alpha hairpin repeats. 9rubB is a two-domain protein according to both SCOP and CATH, one of which is a $(\beta/\alpha)_8$ TIM barrel domain. TIM barrels are one of the most abundant fold types in the protein structure universe but they are sometimes difficult to recognize as a single domain by an automatic method. SCOP treats 2minB and 1avhA as one-domain proteins but CATH cuts them into 4 domains, of which 3 (2minB) or 4 (1avhA) are repetitions of the same basic structural types. The inter-domain contacts are tight in these structures and nearly indistinguishable from the intra-domain contacts. SCOP breaks 8acnA and 1atnA into two domains but CATH treats both as 4-domain proteins. 1atnA is the structure used by Holm and Sander to illustrate their PUU procedure [18]. 8acnA is a chain for which correct domain assignments were difficult in an earlier attempt by one of us (Jongsun Jung and BL, unpublished work). 1kekA and 1oy8A are large chains with more than 1200 and 1000 residues, respectively. SCOP treats 1kekA as a 5-domain protein but CATH cuts it into 7 domains. SCOP and CATH agree on the number of domains in 1jjcB and 1oy8A, 6 and 8 domains respectively. Three chains, 3grsA, 1jjcB, and 1oy8A, contain segmented domains according to both SCOP and CATH.

128-chain set

The Balanced_Domain_Benchmark_3 of Benchmark Dataset in pDomains [27] was chosen for testing. Protein chains in this set have more than 90% overlap of the domain boundaries assigned by SCOP, CATH and the authors. There are 135 chains available for download, but VAST did not generate any cliques for 6 of them, presumably because these chains are short (only 43 to 72 residues) and have few secondary structures. Another one, 2pf2, was removed because SCOP and CATH did not agree on the number of domains for this chain. That left us with 128 chains for testing: 48 one-domain chains, 53 two-domain chains, 21 three-domain chains, 3 four-domain chains and 3 five-domain chains. This set did not include any chains in the 15-chain set described above.

Locally Similar Structural Pieces (LSSPs)

VAST was used for one against all structure comparison of each query chain to all chains in the target set. A clique, for the purposes of this paper, is a set of aligned pairs of residues that VAST generates after comparing a pair of structures. We used almost all the cliques that VAST generated by employing a rather liberal selection criteria: Pcli greater than −10 (see Appendix of Sam et al. [26] for a description of VAST Pcli) and RMSD less than 4 Å. An LSSP (locally similar structural pieces) is the set of aligned residues of either structure (see Fig. 1). Thus, a clique is made of two LSSPs, a query LSSP and the target LSSP. When the word is used without a qualifier, we shall mean the query LSSP.

Most LSSPs are made of discontinuous segments in the primary sequence of the query or the target that belong to secondary structural elements (SSEs). We included the residues in the gap between two segments among the aligned residues if the gap was less than 40 residues long. The idea behind this practice is that, even though the gap region is not aligned, it is likely to be a loop between SSEs and therefore a part of the aligned domain, although, obviously, such a simple length criterion will not exclude cases wherein the padded region includes secondary structural elements. Gaps longer than 40 residues were not included because such a long sequence could contain another domain. Such gap-filled or ‘padded’ LSSPs are referred to as pLSSPs. The query and target pLSSPs generally have different lengths because the gap lengths are not necessarily the same in the query and target sequences (see Fig. 1b).

Recognizing domains from the collection of pLSSPs

The set of pLSSPs is stored as an n-by-m binary matrix, the A-matrix, where n is the number of pLSSPs and m is the number of residues in the query chain (see Fig. 2b). The element A(i,j) is 1 if pLSSP i includes the query residue j or 0 otherwise.

The co-occurrence matrix, N, is an m-by-m square matrix (see Fig. 2c) generated from the pLSSP matrix by the simple operation, $N=A^T *A$, where A^T is the transpose of A. $N(i, j)$ is the number of pLSSPs that contain both residues i and j of the query protein.

We developed three different methods for deriving the number of domains and domain boundaries from the A- and N-matrices, which are described below. A fuller, mathematical description of each of these methods is given as a Supplementary Material (mathematics_of_the_methods.pdf).

The method based on SMF (Symmetric Matrix Factorization)—For a solution with *nd* number of domains, SMF partitions the *N*-matrix into *nd* by *nd* blocks and computes the average value, or density, of the matrix in each block. The density of the diagonal block measures the probability that two residues in the same domain are found in the same pLSSP, while the off-diagonal density measures the probability that two residues in different domains are found in the same pLSSP. The former probability favors the current domain definition while the latter disfavors it. Initially, the SMF procedure defines 12 domains by clustering residues with similar N matrix row vectors. SMF then generates solutions with successively fewer domains following a bottom-up approach in which the least well separated pair of domains is combined to generate the solution with one less domain. The least well separated domains are the pair for which the off-diagonal density relative to the smaller of the two diagonal densities is the largest. After 12 putative domain definitions are generated in this manner, the final solution is chosen using a score function that (1) penalizes a solution when the maximum off-diagonal density is large, (2) penalizes a solution when the minimum diagonal element is small, and (3) rewards a solution by how much the maximum off-diagonal density rises when the number of domains is increased by one.

Fig. 3a shows the scores for the 12 putative partitions for 1jjcB. It shows a clear maximum for the 6-domain solution. Figs. 3b–3d show three calculated *N*-matrices (see mathematics_of_the_methods.pdf in the Supplementary Material) corresponding to the 4, 6 and 8-domain solutions for 1jjcB. The 6-domain solution in Fig. 3c is most similar to the original *N*-matrix in Fig. 2c.

The method based on SVD (Singular Vector Decomposition)—In this method, the *N*-matrix, viewed as a 3D object (see Fig. 4a), is “sliced” with planes at different levels. The results are represented by the binary matrices, *B* (see mathematics_of_the_methods.pdf in the Supplementary Material). These are shown in Figs. 4e, 4f, 4g and 4h for threshold *T* values of 10, 100, 300, and 1000, respectively, for 1atnA. Increasing the threshold gradually reveals the domain organization of the protein. For a medium-sized, compact protein such as 1atnA, multiple pLSSPs can span the whole chain and low thresholds (Fig. 4e, *T*=10) show only one domain. Fig. 4f, with T=100, shows three, weakly resolved, domains (D1=[1–80]; D2=[81–190], [251–374]; D3=[191–250], approximately) while Fig. 4g (*T*=300) shows a somewhat noisy picture of the four domains. In particular, the presence of cross peaks indicates that the first domain consists of three segments. For larger thresholds (such as *T*=1000 in Fig. 4h), the picture of the domains slowly fades away until only motifs along the principal diagonal remain visible (the red peaks in Fig. 4a). At each threshold value, the precise domain boundaries are obtained by SVD of the binary matrix B (see mathematics_of_the_methods.pdf in the Supplementary Material). The principal difficulty with this approach is to select the appropriate threshold: a too low threshold will disclose only some of the domains (merging others into larger entities), a too high threshold will divide the structure into numerous motif-like substructures.

The Pairwise Correlation Method (PCM)—In PCM, we treat the alignment of pLSSPs to the query structure as a statistical process. For a particular query structure residue, we define a random variable taking on the value 1 if the particular pLSSP includes that residue and 0 otherwise. A similar random variable is defined for the other query residues. Thus, each pLSSP is an independent, multivariate observation for this set of random variables, and the collection of pLSSPs are the available data from which the domains are to be derived. In this setting, it is natural to investigate the pairwise correlations among the variables, and then group the query residues into domains according to the correlation structure, grouping positively correlated pairs into the same domain while excluding negatively correlated pairs, as much as possible. We establish an objective function to measure the quality of possible groupings, by simply summing the correlation values for residue pairs contained within one domain, and then summing over all domains. Hierarchical clustering provides a practical means to maximize this objective function and establishes the domain partition. Even though it is not guaranteed to find the absolute best possible partition, clustering provides a satisfactory approximation in most cases.

Analyzing the correlation matrix *R* (see mathematics_of_the_methods.pdf in the Supplementary Material) has several advantages over analyzing the *N*-matrix directly. As the correlations must fall in the range [−1,1], there is reduced dependence on the number of available data (number of pLSSPs). The *R*-matrix effectively scales the *N*-matrix so that rare domains (those that are less frequently represented in the reference database) have a similar chance of being recognized, as do the popular domains (which would likely produce many more pLSSPs than rare domains). We illustrate this with protein 1avhA, a 4-domain protein of 320 residues (Fig. 5). Even though the 4 domains appear nearly identical (Fig. 5d), the third domain does not recruit as many pLSSPs as do the others (Fig. 5a), making it harder to detect in the *N*-matrix. By contrast, the *R*-matrix shows a more uniform representation of all four domains (Fig. 5b). The progress of the hierarchical clustering is illustrated in Fig. 5c, showing the maximum at 4 domains. The 3-domain assignment shows only a slightly lower Q-score, indicating the ambiguity in choosing the correct number of domains. However, the 1- and 2- domain solutions are clearly poorer choices, using this method.

Another advantage of the PCM approach is that it naturally generates a stopping rule for the clustering algorithm, and hence a natural choice for the number of domains; PCM simply stops when the negatively correlated pairs begin to outweigh the positively correlated residue pairs.

Comparison of different domain definitions

The results of our algorithms were compared with the domain definitions used in SCOP and CATH and with the results from three other domain decomposition programs, DomainParser, PDP and PUU. For a quantitative comparison, we used the Net Domain Overlap (NDO) score [32] and SCOP or CATH, whichever gave the higher score, as the standard of truth for each query structure. A mathematical description of the NDO scoring scheme is given as a Supplementary Material (NDOscoring.pdf). The estimated standard error of the mean was calculated as $s/\sqrt{n}$ where s is the sample standard deviation about the mean and n is the number of query chains. MATLAB (http://www.mathworks.com/) was used to develop our programs and to generate the ribbon bar charts to indicate the domain organization of each chain.

RESULTS

In this section we first describe the graphical representation of the LSSPs that lies at the heart of our methodology. The next sub-section presents two interesting examples among the 15-chain set that we used for program development (see Methods) and the last sub-section compares domain assignment results obtained by our methods and other automatic methods on a known benchmark. The purpose of these comparisons is to make sure that the domains we propose are comparable to domains given by other methods based on different principles.

Characteristics of the LSSPs, pLSSPs and the N-matrix

Fig. 2a shows the position of LSSPs in one of the query proteins in the 15-chain set, 1jjcB. Each horizontal set of line segments represents one LSSP. The majority of them are short. The lengths of LSSPs for 1jjcB range from 12 to 175 residues with an average of 24. Most have only 3 to 5 secondary structure elements (SSEs). One query-target pair may produce more than one LSSP when more than one part of either the query or the target structure can be aligned. Therefore, the number of LSSPs may be, and usually is, much larger than the number of target structures. In the case of 1jjcB, there are 12,282 LSSPs from 2,924 target chains. The map of padded LSSPs (pLSSPs) of 1jjcB is shown in Fig. 2b.

Table 1 shows the average length of LSSPs and pLSSPs, as well as the number of aligned SSEs in LSSPs, for each chain in the 15-chain set. The histograms of the lengths of LSSPs and pLSSPs and of the number of SSEs in LSSPs for the 15-chain set are given as a Supplementary Material (15histos.pdf); they are similar across the 15-chain set. Overall, the LSSPs are small; on average, they are 26 residues long and contain 4 secondary structure elements. The padding process doubles the length of the LSSP (to 56 residues on average). The *N*-matrix of 1jjcB is shown in Fig. 2c. $N(i, j)$ is the number of pLSSPs that contain both residues *i* and *j* of the query protein. Frequent co-occurrence is an indication that the two residues (*i*,*j*) are in the same domain since they are likely to be structurally related. On the other hand, two residues in two different domains in the query structure are less likely to be observed in other structures unless the two domains pack closely. The *N*-matrix has a block diagonal appearance, from which the domain boundaries are readily discernible. Occasionally, a diagonal block is split by another diagonal block, in which case the split block represents a segmented domain. The off-diagonal blocks circled in dotted red along with the two diagonal blocks circled in solid red in Fig. 2c indicate such a segmented domain, made of residue 1 – 40 and 150–185, shown as two red bars in Fig. 2e and as red ribbons in Fig. 2f.

The N-matrices for all 15 chains are given as a Supplementary Material (15charts.pdf). These matrices were processed by the three different mathematical procedures described in the Methods section to define structural domains.

Examples

Here we describe the results for two examples, 1jjcB and 1oy8A. The domain boundaries are presented in different colors in a ribbon bar chart and in corresponding colors in a 3D structure drawing.

Fig. 2e shows the ribbon bar chart for 1jjcB (β-chain of the phenylalanyl tRNA synthetase). SCOP, CATH and PDP domains are similar and the 3D structure drawing shown in Fig. 2f is colored using the CATH definitions. SMF, SVD and PCM all gave six domains as did SCOP and the domain boundaries that agree with those of SCOP and CATH and with visual inspection to similar degrees of accuracy as other established methods (Domain Parser, PDP and PUU). It can be noted, however, that one helix in SCOP red domain (the domain colored red in Figs. 2e and 2f) partly belongs to the blue domain in SMF and SVD (see Fig. 2e). The *N*-matrix of 1jjcB does show some weak off-diagonal density corresponding to these two domains (dotted blue circles in Fig. 2c), indicating that there are pLSSPs that bridge the red and blue domains.

1oy8A, a multidrug efflux pump, has 1006 residues and the structure has an approximate two-fold symmetry, clearly visible in the *N*-matrix (Fig. 6c). According to SCOP, CATH, and our visual inspection, it is made of 8 domains, 4 of which are segmented (red, green, purple, and orange in Figs. 6a and 6b). None of the three published programs (Domain parser, PDP, and PUU) produce the same set of domains. But SMF defines the same 8 domains as SCOP and CATH (Fig. 6a). On the other hand, SVD and PCM combine the red and purple segmented domains into one and the blue and magenta domains into another (Figs. 6a and 6d). These latter two algorithms apparently did not ignore the off-diagonal densities that are visible in the *N*-matrix, indicated by red and blue dotted circles in Fig. 6c.

Agreement with other methods and among our methods

The domain boundaries for the 15-chain set are given in the Supplementary Material (15chainDB.pdf for numerical boundary residue numbers calculated by the three methods and 15charts.pdf for ribbon bar chart representation of domain boundaries by all methods). Table II shows the number of domains in this set of protein chains according to SCOP, CATH, visual inspection, and by different automatic methods. There are 5 chains for which SCOP and CATH do not agree on the number of domains. In each case, CATH cuts the chain into more domains than SCOP.

For a quantitative comparison of different domain assignment schemes, we use the NDO scores that we developed earlier for such purpose [32]. The NDO score essentially measures the fraction of residues correctly assigned minus those incorrectly assigned to a reference set of domains. We used SCOP and CATH domains as the reference. (See Methods.) The NDO scores of each domain assignment program for each query are given in Table III. These were calculated using either SCOP or CATH as the standard of truth, whichever gave the higher score. The average for each method with the estimated standard error of the mean are given in the last row of the table. The domain assignments made by visual inspection by one of us (BL) are closer to SCOP or CATH than by any of the automatic methods.

1qbkB is a superhelical structure with 890 residues. Although SCOP and CATH treat the whole chain as one domain, most programs cut it into 3 or more pieces. For the other 5 single domain chains, most programs recognize them well except PCM, which fails to recognize two of them as single domains.

Among the 9 multi-domain chains, for 5 of which SCOP and CATH disagree on the number of domains (Table II), our algorithms gave results that were comparable to those from the existing programs tested (Table III). The average NDO scores for these 9 chains are 86, 78 and 83% for SMF, SVD and PCM, respectively, compared to 81, 83, and 65% for DomainParser, PDP and PUU, respectively.

We then used a larger, well-known dataset for a more objective testing. For this purpose, we chose a 128 chain set obtained from pDomain Benchmark dataset [27]. SCOP and CATH agree on the number of domains for all of these chains. The domain boundaries for all chains by all methods are given in the Supplementary Material (128chainDB.pdf). The average NDO scores of each program are shown in Table IV. SMF and SVD are comparable with other programs in recognizing one-domain chains. For chains with three or more domains, PCM method gives results closer to SCOP and CATH than SMF and SVD. On the contrary SMF and SVD perform better for 1 and 2 domain proteins. This result gives some basis for the choice of the method for a domain analysis when the three are used together to analyze the *N*-matrix.

DISCUSSION

Collection of local structural similarities can define domains

The set of LSSPs for a protein chain, such as those shown in Fig. 2a for 1jjcB, represents a collection of local structural similarities found in a database of other protein structures. It is not clear from Fig. 2a that such a collection can define domains in the protein chain. However, when the LSSPs are padded with intervening query sequences and the collection written in the form of a binary matrix, which is then transformed to the co-occurrence *N*-matrix, the domain structure becomes evident. (See Fig. 2c.) Deriving the precise number and boundaries of domains from the *N*-matrix is not trivial. We devised three different algorithms for this process, but the process can probably still be improved. Nevertheless, the results reported here clearly indicate that such a collection of local structural similarities can be used to assign domains that largely agree with accepted domains. This is evident from visual inspection of the *N*-matrices and of the assigned domains of many individual proteins, as well as from the relatively high average NDO scores achieved for all proteins tested. It should be noted that this method of assigning domains does not use any assumptions, or use only minimal assumptions (see SVD, determination of the threshold value T in mathematics_of_the_methods.pdf of the Supplementary Material), on the geometric or folding properties of domains.

The apparent success of our methods is all the more remarkable because a vast majority of the pLSSPs arise from target structures that do not contain the relevant domain in the query chain. The large number of pLSSPs that define a domain indicates this to be the case. For example, Fig. 2d shows the pLSSPs for 1jjcB sorted by their mean (query) residue number and colored according to the SMF domain to which the pLSSP contributes the most. For the purple domain defined by SCOP, for example, there are 3,633 pLSSPs from 2,244 targets that contribute more to this domain than to any of the other five SCOP domains. Clearly, a representative set of 6,373 target structures will not contain over 2,000 structures that all contain one or more domains that resemble the purple domain of 1jjcB. This is not an exceptional example. The number of pLSSPs per domain is 6,030 and 6,263 on average for the 15- and 128-chain sets, respectively, using the CATH domain definitions. If SCOP definitions are used, the numbers are even larger.

The large number of pLSSPs arises from the use of generous criteria in selecting VAST cliques. Many pLSSPs contain only a small number of aligned residues and about equal number of padded residues. For example, the purple domain of 1jjcB is 207 residues long (residues 475 to 681) according to SCOP. The average length of the 3,633 pLSSPs assignable to this domain (each contributes more to this domain than to any other domain) is 56 including the padded residues, or 27% of the number of residues of the domain. The proportion of padded residues in these pLSSPs is 49% on average. Only 17 pLSSPs span more than 80% of the 1jjcB purple domain. Others represent a small part of a domain in the target structure that match a small part of this domain in the query structure. Many are from target domains that visually do not resemble the 1jjcB purple domain (The figure on the right-hand side of Fig. 1a shows a target structure that generated pLSSPs with the average number of aligned and padded residues for this domain). Table I and the histograms given in the Supplementary Material (15histos.pdf) show that the lengths of LSSPs and pLSSPs of all chains in the 15-chain set are similarly short and that they contain only 4 SSEs on average.

To further verify the role of pLSSPs that fall short of covering the whole of query domains, we did the following experiments. SMF was run for the 15 chains using only those pLSSP whose padded length was less than 80% of any domain defined by SCOP. It turns out that, on average, only 7% of the pLSSPs were removed by this criterion. The *N*-matrix looked quite similar and the average NDO score was 78 compared to 83 using all pLSSPs. Similar experiment was done using SVD on 1atnA with an even lower cut-off. 1atnA has 4 domains according to CATH and SVD, with a chosen threshold, also assigns 4 domains using all pLSSPs. For the experiment, 1,865 pLSSPs (14%) covering more than 70% of any CATH domain were removed. The change in the *N*-matrix was slight; the most notable feature was that the second domain became a little less visible. The number of domains remained the same and the domain boundaries varied by less than 10 residues except that the second domain picked up a 5-residue segment from the fourth domain, which became thereby segmented. These results indicate that short pLSSPs that do not cover the whole domain can nevertheless define the boundaries of the whole domain.

The important feature is that there are many more pLSSPs that span some region within one single domain than those that span between two or more domains. This is the feature that ultimately makes it possible to recognize domains in the *N*-matrix. The SMF algorithm, for example, defines domains by explicitly rewarding for the number and length of pLSSPs within a domain, which tends to break up domains into smaller, better-defined pieces, and penalizing for the number of pLSSPs that span between domains, which tends to fuse domains into larger ones. Padding LSSPs helps in this process since it brings in many query residues that are not aligned but which are in the domain because they are between SSEs that do align.

Characteristics of using collections of pLSSPs

All three methods (SMF, SVD, and PCM) use matrices and procedures that do not depend on the order of the columns or rows of the matrices. Therefore, segmented domains are treated like any other domain and do not require any special handling.

Structures that are made of repeating units without a central core are difficult to recognize as one-domain proteins. This type of structure is represented by 1a0tP and 1qbkB in the 15-chain set. 1qbkB is made of alpha-alpha helical hairpin repeats arranged in essentially one-dimensional, superhelical array. Both SCOP and CATH consider this to be a one-domain protein. However, most pLSSPs arise from structures that contain a small number of anti-parallel helices, which match at many different places along the superhelix. Therefore, the N-matrix is not a solid square but looks like a chain of overlapping smaller squares along the diagonal. (See the N-matrix for 1qbkB given in 15charts.pdf in the Supplementary Material.) The density outside of these smaller squares is low because the structure has no long-range interactions and few other structures have the helices arranged in the same superhelical manner over the long range. The low density away from diagonal makes it difficult to recognize this chain as a one-domain protein (see the ribbon bar chart for this protein given in 15charts.pdf in the Supplementary Material).

1a0tP is one unit of the homotrimeric porin. It is a large barrel made of 18 up-and-down beta strands. Most would agree with SCOP and CATH that this is a one-domain protein. However, owing to the large size of the barrel, there are only near-neighbor interactions, with no interaction across the barrel. Most pLSSPs arise from structures that contain a β-sheet with a few up-and-down β-strands, which match the query structure at many places around the barrel. Again, the N-matrix is not a solid square but has a look of a chain of overlapping small squares along the diagonal. Although both SMF and SVD managed to consider this as a one-domain protein (see the ribbon bar chart for this protein given in 15charts.pdf in the Supplementary Material), other parameter choices can easily break up this chain into many domains.

As described in the previous section, pLSSPs can collectively delineate domain boundaries in the query 3D structure. However pLSSPs can be a double-edged sword. When two domains are close in space, either because of the intimate domain interface or because they bind a common ligand, such as a DNA, a number of these pLSSPs can span both domains, thus blurring the boundary between them. For instance, in the *N*-matrix representation this causes “overlapping squares” for domains that follow each other in the sequence, or “off diagonal peaks” when they are separated in the sequence. This kind of problem happens for 1avhA, the 320-residue human annexin V, which is made of 4 helical annexin domains with intimate domain interfaces (Fig. 5d). Although there is a single target, 1bo9A (the 73-residue human annexin I), that matches each of the 4 domains individually, there are also many pLSSPs that span two domains. This obscures the boundaries between the domains and results in an *N*-matrix that only PCM, with its special statistics-based scaling feature, could interpret as representing 4 domains. A more judicious choice or filtering of the VAST cliques, LSSPs, and/or pLSSPs could help in such cases. Overall, the contribution of the pLSSPs is positive since, domains being evolutionarily stable structures, there are generally more pLSSPs matching within a single domain than those that span two separate domains.

None of the three procedures is yet a full-fledged tool for domain partition and we see at least two clear avenues for improving the methods to take full advantage of the LSSPs. One possibility is to filter the cliques to reduce the background as already suggested above. Another is to combine the three procedures in such a fashion as to preserve the positive features of each method and suppress or avoid negative features. Only after such refinements have been implemented would it be possible to make a reliable assessment of the effectiveness of the LSSP-based domain partition method compared to the largely geometry-based, and highly refined, methods such as PDP, Domain Parser, and PUU.

Implications

It is increasingly recognized [34,35] that geometric similarities exist between proteins that have different folds. These relationships between seemingly unrelated proteins are found at different scales, from small sets of SSEs to larger fragments. This new perception of existing relationships between proteins challenges usual hierarchical classification schemes [28] and lends support to a continuous view of the fold space [26,36]. Protein structures result both from evolutionary and physicochemical principles. Recent works [37–39] have described a number of genetic mechanisms (duplication, swapping and deletion, circular permutation, accretion of SSEs or “embellishment” around a core substructure) that explain how proteins derived from a common ancestor can nevertheless possess different folds. Other groups have focused their attention on the physical and chemical principles of domain formation. Zhang and colleagues [29] have shown that the observed repertoire of single-domain folds found in the PDB can be the result of geometric effects due to the packing of compact H-bonded secondary structures. This is in agreement with the work of Szustakowski et al. [30] who have built a dictionary of super-secondary structures that can be used as basic protein parts. These ‘protein legos’ can be assembled in a variety of ways to build different protein folds. To explain the presence of similar structural “motifs” in different folds, Lupas and collaborators [31] proposed a hypothesis according to which these motifs are the relics of a predomain world when protein structures consisted of conglomerations of short polypeptide chains (antecedent domain segments, ADSs). Single-chain domains arose later from the fusion of the corresponding ADS genes. This hypothesis is substantiated by the existence of modern repeat proteins, for instance ankyrin or HEAT proteins, where repetitive substructures are associated with obvious sequence similarities.

The methodology we propose in this article, based on the definition of domains from sets of pLSSPs found in many “unrelated” target proteins, is fully consistent with the above view of the evolutionary and physicochemical principles underlying single-fold (domain) structures. In this framework, defining domains in terms of recurrence of structural “pieces” in many structures is, in principle, a more natural approach than geometrically decomposing single structures.

CONCLUSIONS

We developed three different procedures for defining domains in a protein from a collection of relatively small, locally similar structural pieces (LSSP) that are found in other protein structures. Although these procedures have yet to be perfected, it is clear that one can obtain domains that largely agree with those of existing domain databases by these procedures. Therefore, LSSPs offer a new way of defining domains. In addition, the fact that this is possible indicates that domains are indeed made of recurrent, small structural pieces, in accordance with the conclusions of many recent studies by others.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Richard Shrager for preparing tables of filtered target structures for the 15-chain set, Dr. Woei-Jyh Lee for writing the mathematical description of the NDO scoring scheme, given as the Appendix to this article, and Dr. Todd Taylor for his criticism and valuable suggestions. Molecular graphics images were produced using the UCSF Chimera[40] package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081). This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research; by the Intramural Research Program of the NIH, Center for Information Technology; and by the National Institute for Agricultural Research (INRA), France, Mathématique, Informatique et Génome laboratory.

References

1. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A 1973;70(3):697–701. [PubMed: 4351801]
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093–1108. [PubMed: 9309224]
3. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536–540. [PubMed: 7723011]
4. Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins 1998;33(1):88–96. [PubMed: 9741847]
5. Fields S. Proteomics. Proteomics in genomeland. Science 2001;291(5507):1221–1224. [PubMed: 11233445]
6. Sonnhammer EL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci 1994;3(3):482–492. [PubMed: 8019419]
7. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997;28(3):405–420. [PubMed: 9223186]
8. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 1998;95(11):5857–5864. [PubMed: 9600884]
9. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci 1998;7(2): 233–242. [PubMed: 9521098]
10. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339. [PubMed: 7020376]
11. Tress M, Tai CH, Wang G, Ezkurdia I, Lopez G, Valencia A, Lee B, Dunbrack RL Jr. Domain definition and target classification for CASP6. Proteins 2005;61 (Suppl 7):8–18. [PubMed: 16187342]
12. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward consistent assignment of structural domains in proteins. J Mol Biol 2004;339(3):647–678. [PubMed: 15147847]
13. Islam SA, Luo J, Sternberg MJ. Identification and analysis of domains in proteins. Protein Eng 1995;8(6):513–525. [PubMed: 8532675]
14. Swindells MB. A procedure for detecting structural domains in proteins. Protein Sci 1995;4(1): 103–112. [PubMed: 7773168]
15. Xu Y, Xu D, Gabow HN. Protein domain decomposition using a graph-theoretic approach. Bioinformatics 2000;16(12):1091–1104. [PubMed: 11159328]
16. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci 1995;4(5):872–884. [PubMed: 7663343]

17. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics 2003;19(3):429–430. [PubMed: 12584135]
18. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19(3):256–268. [PubMed: 7937738]
19. Sowdhamini R, Blundell TL. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. Protein Sci 1995;4(3):506–520. [PubMed: 7795532]
20. Taylor, T.; Vaisman, I. Protein structural domain assignment with a Delaunay tessellation derived lattice. The 3rd International Symposium on Voronoi Diagrams in Science and Engineering; 2006; 2006.
21. Taylor WR. Protein structural domain identification. Protein Eng 1999;12(3):203–216. [PubMed: 10235621]
22. Veretnik, S.; Gu, J.; Wodak, SJ. Identifying structural domains in proteins. In: Gu, J.; Bourne, PE., editors. Structural Bioinformatics. 2. Hoboken, N.J: John Wiley & Sons, Inc; 2009. p. 485-513.
23. Taylor, TJ. Ph D. Fairfax, VA: George Mason University; 2006. Analysis of the structure and topology of real and model proteins using Delaunay tessellation.
24. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6(3):377–385. [PubMed: 8804824]
25. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23(3):356–369. [PubMed: 8710828]
26. Sam V, Tai CH, Garnier J, Gibrat JF, Lee B, Munson PJ. ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. BMC bioinformatics 2006;7:206. [PubMed: 16613604]
27. Holland TA, Veretnik S, Shindyalov IN, Bourne PE. Partitioning protein structures into domains: why is it so difficult? J Mol Biol 2006;361(3):562–590. [PubMed: 16863650]
28. Taylor WR. Evolutionary transitions in protein fold space. Curr Opin Struct Biol 2007;17(3):354–361. [PubMed: 17580115]
29. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A 2006;103(8):2605–2610. [PubMed: 16478803]
30. Szustakowski JD, Kasif S, Weng Z. Less is more: towards an optimal universal description of protein folds. Bioinformatics 2005;21(Suppl 2):ii66–71. [PubMed: 16204127]
31. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 2001;134(2–3):191–203. [PubMed: 11551179]
32. Tai CH, Lee WJ, Vincent JJ, Lee B. Evaluation of domain prediction in CASP6. Proteins 2005;61 (Suppl 7):183–192. [PubMed: 16187361]
33. Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 2005;5:17. [PubMed: 16164759]
34. Petrey D, Fischer M, Honig B. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. Proc Natl Acad Sci U S A 2009;106(41):17377–17382. [PubMed: 19805138]
35. Petrey D, Honig B. Is protein classification necessary?. Toward alternative approaches to function annotation. Curr Opin Struct Biol 2009;19(3):363–368. [PubMed: 19269161]
36. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins 2000;38(3):247–260. [PubMed: 10713986]
37. Andreeva A, Murzin AG. Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 2006;16(3):399–408. [PubMed: 16650981]
38. Kinch LN, Grishin NV. Evolution of protein structures and functions. Curr Opin Struct Biol 2002;12(3):400–408. [PubMed: 12127461]
39. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. Bioinformatics 2005;21(8):1308–1310. [PubMed: 15604105]

40. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13):1605–1612. [PubMed: 15264254]

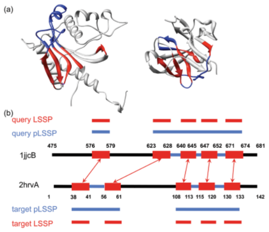

Figure 1. A typical LSSP

Panel (a) shows an LSSP between the query protein 1jjcB (only residues 475–681, corresponding to the purple domain in Figs. 2d, 2e and 2f, are shown) and the target protein 2hrvA. The structural similarity found by VAST corresponds to 4 anti-parallel strands of a sheet and a piece of another strand further up. Structurally the two domains have nothing in common but that sheet. Aligned segments are shown in red in both structures. Regions between two consecutive aligned segments in the query or the target sequences are included in the alignment when their length is less than 40 residues (this is shown in blue in the two structures). This generates “padded” LSSPs (pLSSPs).

Panel (b) presents the same LSSPs and pLSSPs “projected” along the query and target sequences. The red rectangles correspond to the VAST clique segments with their numbering along their respective sequences. Double-headed arrows indicate the structural correspondence between segments. Padded regions between the segments are displayed as thick blue lines. Notice that the query and target pLSSPs (and LSSPs) have gaps of different lengths and positions. The query pLSSP (respectively LSSP) displayed on top of panel (b) is one of the 12,282 pLSSPs (resp. LSSPs) shown in Fig. 2b (resp. Fig. 2a). The query pLSSP with 5 SSEs has a length of 56 residues, the mean of all pLSSPs for this domain, and a number of aligned residues of 26, the average for this domain.

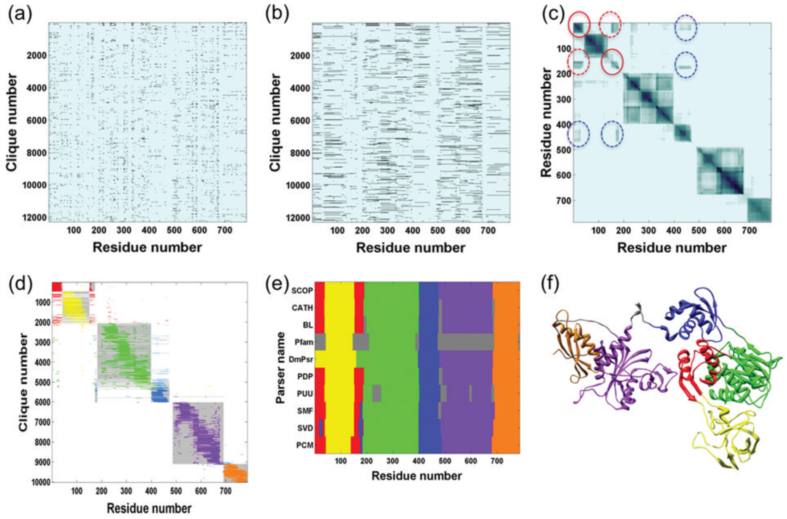

Figure 2. LSSPs and domains of 1jjcB

The six panels of this figure all pertain to 1jjcB. (a) Location of LSSPs (Locally Similar Structural Pieces) on the sequence of 1jjcB. The Y-axis is the LSSP serial number and the X-axis is the residue number. Black lines indicate the residues in 1jjcB that are aligned in the VAST cliques produced by the target chains. There are 12,282 cliques for which the RMSD is less than 4 Å. This figure can be misleading; because of the great compression of data along the Y-axis of the figure, many cliques are not visible. (b) Location of pLSSPs (gap-filled or "padded" LSSP). This is also the *A*-matrix, which has the value 1 on the line segments and 0 outside. This figure can be misleading; because of the great compression of data along the Y-axis of the figure, many cliques are not visible. (c) The heat map of the co-occurrence *N*-matrix of 1jjcB. The X and Y-axes are the residue numbers of 1jjcB. The pixel intensity indicates the value of the matrix element, which is the number of pLSSPs that contain both of the residues represented by the pixel position. The two squares along the diagonal, circled in red, indicate two segments of a segmented domain, which also produces the off-diagonal intensities, indicated by the red dotted circles. This domain is made of segments [1–39] and [151–186] of the polypeptide chain and is "interrupted" by the second domain [40–150] that appears as the second square along the diagonal of the heat map. This segmented domain is shown in red in panels d, e, and f. The off-diagonal intensities dot-circled in blue indicate that there are pLSSPs that span the red domain and another domain, which is colored blue in panels d, e, and f. (d) A sorted map of pLSSPs, colored according to the domain assignments made by SMF. The pLSSPs were sorted in ascending order of the mean of the serial numbers of the residues in the pLSSP. The gray shading indicates the boundaries of the domains. This figure can be misleading; because of the great compression of data along the Y-axis of the figure, many cliques are not visible. (e) The ribbon bar chart shows the domain boundaries of 1jjcB according to SCOP, CATH, visual inspection by one of the authors (BL), Pfam, and different domain partition programs indicated by their name (DmPsr for DomainParser). The grey areas indicate the residues that do not belong to any domain or, for Pfam, any protein family. (f) The structure of 1jjcB colored according to the CATH domains.

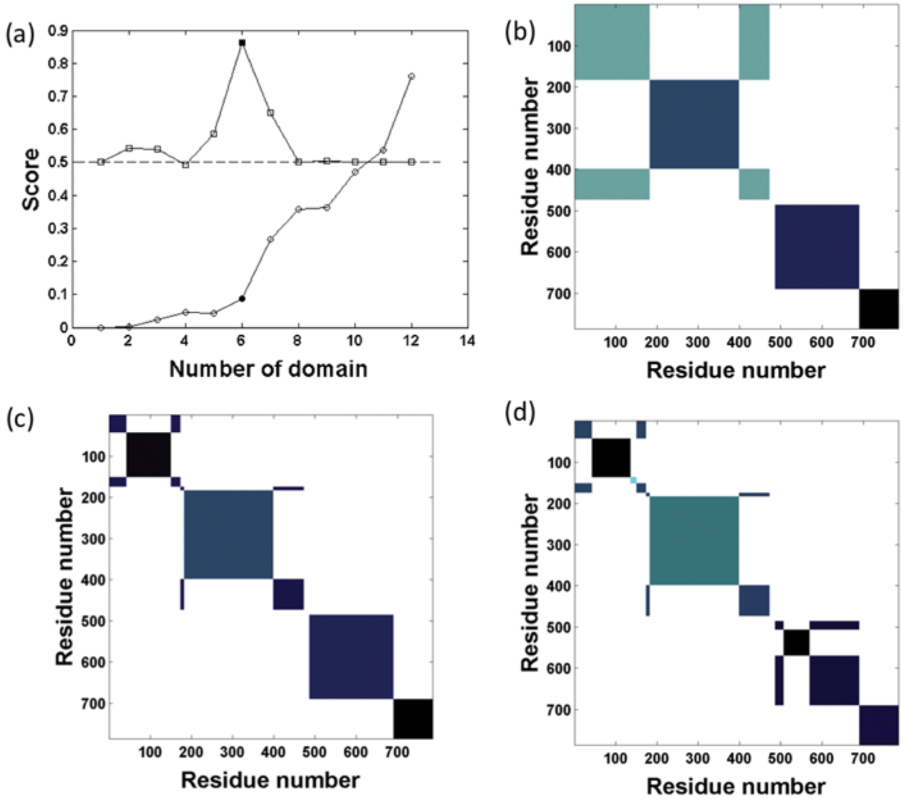

Figure 3. Partition of 1jjcB by the SMF algorithm

(a) The scores for partitions into different numbers of domains. The circles and squares are for the *Q1* and *2*Q+0.5*, respectively. (See mathematics_of_the_methods.pdf of the Supplementary Material.) The symbols are filled at the position of maximum *Q*. The remaining three panels show the calculated N_c-matrices for the 4-domain (b), 6-domain (c), and 8-domain (d) solutions, respectively.

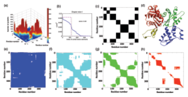

Figure 4. Partition of 1atnA by the SVD algorithm

(a) A 3D plot of the *N*-matrix. The values of the matrix elements are plotted along the z-axis, using colors according to the color scheme shown to the right of the graph. (b) The blue curve is the plot of the squares of the elements (sorted by decreasing order of magnitude) of the singular vector associated with the largest of the 4 singular values from the singular value decomposition of the binary matrix shown in panel g (see mathematics_of_the_methods.pdf of the Supplementary Material). The X-axis is for the elements of the singular vector, which correspond to the (scrambled) residues of 1atnA. The vertical dashed pink line is at the position of the singular value. In the ideal case, the elements to the left of the vertical line have the same non-zero value and correspond to the residues of the domain; all others are zero. The green dotted vertical line next to the pink one corresponds to a slight optimization of the singular value to find a position in the vicinity where the blue curve has the maximal slope. The score is the difference between the areas under the blue and the pink curves. (c) The *M*-matrix (matrix representation of an equivalence relation) using the CATH domain definition. (d) The structure of 1atnA colored according to the CATH domains. (e) to (h) Binarized *N*-matrices “sliced” at thresholds of 10, 100, 300 and 1000, respectively. The color of the matrix approximately corresponds to the color of the 3D plot of panel (a) at the slicing level.

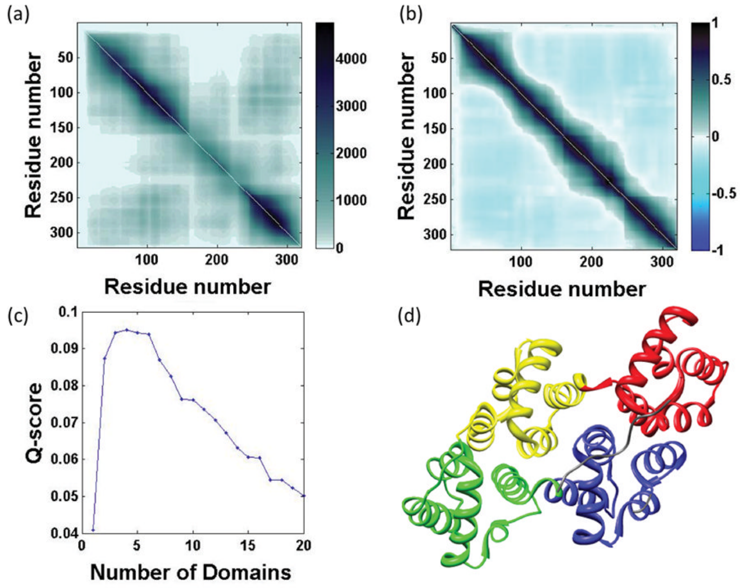

Figure 5. Partition of 1avhA by the PCM algorithm

(a) A heat map of the *N*-matrix. Residues 160 to 230 are clearly under-represented in the pLSSP collection. (b) A heat map of the pairwise residue correlation matrix, *R*. Now residues 160–230 receive a more nearly equal representation. Boundaries between the 4 domains are evidenced by the "pinching in" of the positive area along the diagonal at roughly residue 75, 150, and 240. (c) PCM *Q*-score vs. number of domains, showing optimum at 4 domains (see mathematics_of_the_methods.pdf of the Supplementary Material). Scores for 1 or 2 domains are strongly reduced compared to 4 domains, while the 3 and 5 domain results were nearly as good as the 4 domain result. Clustering proceeded from right to left, starting with an assignment having 320 domains, one for each residue. (d) The structure of 1avhA colored according to the CATH domain definition. PCM gave a 4-domain solution similar to CATH.

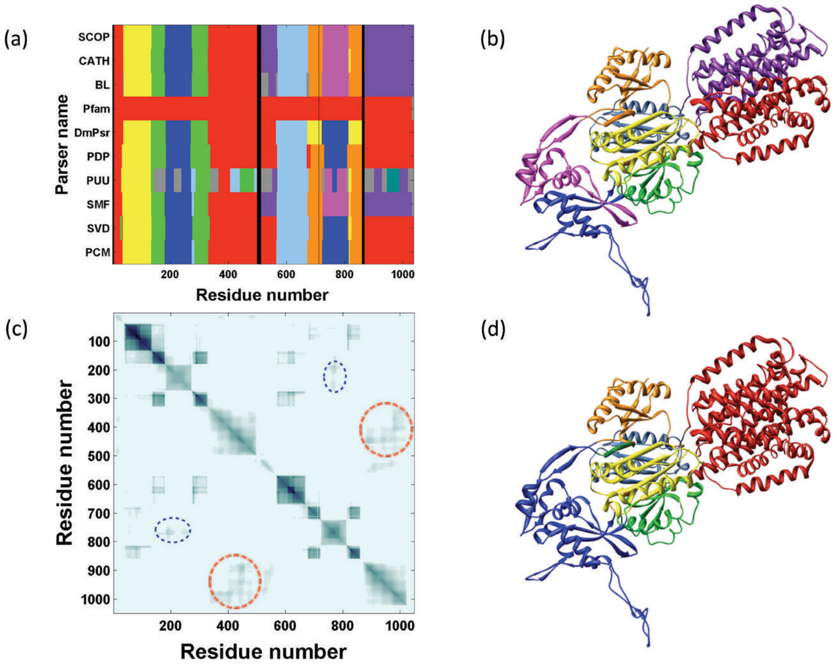

Figure 6. Domain organization of 1oy8A

(a) The ribbon bar chart shows the domain boundaries of 1oy8A according to the indicated databases and programs. The black areas are residues with missing coordinates in the PDB file. (b) The structure colored according to the CATH domain definitions. SCOP definitions are similar. SMF gave similar results. (c) The *N*-matrix of 1oy8A. The dotted circles indicate the off-diagonal signals, which made PCM and SVD to combine the red and purple domains together (red circle) and the blue and magenta domains together (blue circle). (d) The structure colored according to the PCM domain definitions. SVD gave similar results. Both combined the red and purple SCOP/CATH domains into one, and the magenta and blue domains into one.

Table I

The average length of LSSPs and pLSSPs and the average number of SSEs in LSSPs for the 15-chain set. The numbers in parenthesis are the standard deviations.

	Number of pLSSPs	LSSPs	pLSSPs	SSEs
1a0tP	16558	24 (11)	70 (31)	4 (1)
1aluA	4493	27 (15)	46 (24)	3 (0)
1avgI	9151	21 (8)	58 (23)	4 (1)
1b67A	1070	23 (9)	58 (6)	3 (0)
1fcyA	12584	26 (12)	50 (23)	4 (1)
1qbkB	17023	23 (12)	50 (26)	4 (1)
9rubB	14254	29 (23)	63 (40)	5 (2)
3grsA	10931	28 (19)	57 (30)	5 (2)
2minB	13114	31 (22)	66 (34)	5 (2)
1avhA	18239	25 (10)	52 (23)	4 (1)
8acnA	10510	25 (14)	53 (28)	4 (1)
1atnA	12885	27 (14)	55 (24)	5 (1)
1kekA	11614	27 (18)	59 (32)	4 (2)
1jjcB	12282	24 (11)	54 (25)	4 (1)
1oy8A	11280	27 (12)	53 (23)	4 (1)
Average	11733 (4467)	26	56	4

Table II

Number of domains for the 15-chain set

	SCOP	CATH	Visual (BL)	Domain Parser	PDP	PUU	SMF	SVD	PCM
1a0tP	1	1	1	1	1	1	1	1	3
1aluA	1	1	1	1	1	1	1	1	1
1avgI	1	1	1	1	1	1	1	1	1
1b67A	1	1	1	1	1	1	1	1	1
1fcyA	1	1	1	1	1	1	1	1	2
1qbkB	1	1	1	3	2	*	3	5	4
9rubB	2	2	2	2	2	2	2	3	3
3grsA	3	3	3	3	4	4	4	3	3
2minB	1	4	3	2	3	4	4	3	4
1avhA	1	4	4	3	2	3	1	1	4
8acnA	2	4	3	2	2	7	3	4	4
1atnA	2	4	4	2	4	4	4	2	3
1kekA	5	7	6	4	6	7	7	8	6
1jjcB	6	6	6	5	6	6	6	6	6
1oy8A	8	8	8	5	6	11	8	6	6
Sum	36	48	45	36	42	53*	47	46	51

*PUU did not give any result for 1qbkB. The sum is without 1qbkB.

Number of domains in each chain of the 15-chain set by SCOP, CATH, visual inspection (by BL), and different programs.

Table III

NDO scores for the 15-chain set

	Visual (BL)	Domain Parser	PDP	PUU	SMF	SVD	PCM
1a0tP	100.0	100.0	100.0	100.0	100.0	100.0	37.3
1aluA	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1avgI	100.0	100.0	100.0	100.0	97.2	100.0	100.0
1b67A	100.0	100.0	51.5	100.0	100.0	100.0	100.0
1fcyA	100.0	100.0	100.0	100.0	100.0	100.0	60.2
1qbkB	100.0	54.4	68.8	*	40.2	22.4	33.2
9rubB	98.3	99.6	99.6	73.8	96.7	90.1	89.2
3grsA	98.3	97.8	89.8	79.6	87.2	80.0	97.8
2minB	76.2	57.2	73.9	46.6	75.1	62.6	72.8
1avhA	94.8	64.3	66.7	64.3	99.4	100.0	87.9
8acnA	83.5	99.5	99.5	27.8	80.3	62.8	90.4
1atnA	98.3	80.1	91.1	93.9	65.9	78.5	74.5
1kekA	79.1	82.5	68.5	68.6	80.6	68.2	80.8
1jjcB	97.4	86.2	97.6	90.5	93.7	90.8	91.1
1oy8A	94.5	58.8	64.3	41.4	94.6	66.6	66.5
Average (esem†)	**94.7 (2.1)**	**85.4 (4.7)**	**84.8 (4.4)**	**77.6* (6.6)**	**87.4 (4.4)**	**81.5 (5.7)**	**78.8 (5.6)**

[*] PUU did not give any result for 1qbkB. The average was calculated excluding 1qbkB.

[†] Estimated standard error of the mean is in ().

NDO score for each chain of the 15-chain set by visual inspection and each domain parsing programs. SCOP or CATH, whichever gave the higher score, was used as the standard of truth for the NDO score calculation.

Table IV

Average NDO scores for the 128-chain set

	Number of chains	Domain Parser	PDP	PUU	SMF	SVD	PCM
All chains	128	92.9 (1.3)	93.1 (1.3)	86.7 (1.7)	87.5 (1.7)	87.3 (1.6)	83.2 (1.8)
1-domain chains	48	99.0 (1.0)	97.4 (1.2)	93.6 (1.8)	99.5 (0.1)	97.2 (1.3)	89.5 (3.1)
2-domain chains	53	90.2 (2.1)	92.1 (2.0)	80.9 (3.1)	82.6 (2.6)	87.7 (2.2)	79.0 (2.5)
Chains with 3 or more domains	27	87.3 (3.6)	87.4 (3.7)	86.0 (3.4)	75.6 (5.0)	68.7 (4.1)	80.1 (4.2)

NDO scores averaged over all chains of the 128-chain set and separately over the 1-domain chains, 2-domain chains, and chains with 3 or more domains. SCOP or CATH, whichever gave the higher score, was used as the standard of truth for the NDO score calculation. Estimated standard error of the mean is in (). The domain boundaries given for DomainParser, PDP and PUU were downloaded from the pDomains website (http://pdomains.sdsc.edu).