BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 2 2006, pages 129–133
doi:10.1093/bioinformatics/bti773

Structural bioinformatics

'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments

Jean-Christophe Gelly[1,2], Alexandre G. de Brevern[1,*] and Serge Hazout[1,†]

[1]INSERM U726, Equipe de Bioinformatique Génomique & Moléculaire (EBGM), Université Denis Diderot—Paris 7, case 7113, 75251 Paris Cedex 05, France and [2]UMR CNRS 7602, Institut de Minéralogie et Physique des Milieux Condensés (IMPMC), Université Pierre & Marie Curie, 4, Place Jussieu, 75252 Paris Cedex 05, France

Received on September 2, 2005; revised on October 19, 2005; accepted on November 8, 2005
Advance Access publication November 14, 2005
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The object of this study is to propose a new method to identify small compact units that compose protein three-dimensional structures. These fragments, called 'protein units (PU)', are a new level of description to well understand and analyze the organization of protein structures. The method only works from the contact probability matrix, i.e. the inter Cα-distances translated into probabilities. It uses the principle of conventional hierarchical clustering, leading to a series of nested partitions of the 3D structure. Every step aims at dividing optimally a unit into 2 or 3 subunits according to a criterion called 'partition index' assessing the structural independence of the subunits newly defined. Moreover, an entropy-derived squared correlation R is used for assessing globally the protein structure dissection. The method is compared to other splitting algorithms and shows relevant performance.

Availability: An Internet server with dedicated tools is available at http://www.ebgm.jussieu.fr/~gelly/

Contact: debrevern@ebgm.jussieu.fr.

1 INTRODUCTION

The organization of three-dimensional proteins structures can be represented as an assembly of different secondary structure elements in a particular arrangement (Michalopoulos *et al*., 2004; Richardson, 1981). This topology characterizes a unique and particular fold (Chothia and Finkelstein, 1990). Several distinct combinations of secondary structures, generally 2–4, form particular motifs that can be found in many different folds: the super-secondary structures. Many of them have been well characterized such as the simple β-hairpins (Sibanda and Thornton, 1993), to more complex associations such as helix–turn–helix, β–α–β, four-helix bundle or Greek key (Efimov, 1994; Richardson, 1981). Unfortunately, many folds contain very few or no super-secondary structures, e.g. the knottins (Gelly *et al*., 2004).

Since the eighties many authors have proposed different methods to hierarchically split protein structures into small compact units, with the aim of describing the different levels of protein structure organization (Go, 1981; Guo *et al*., 2003; Janin and Wodak, 1983; Lesk and Rose, 1981; Sowdhamini and Blundell, 1995; Tsai and Nussinov, 1997). The rules used by these methods are quite different. Gō determined structural units by visual inspection of three-dimensional structures and the Cα-Cα distance map. Janin and Wodak scanned polypeptide chains to find minimal interface area between putative compact globular units represented by pseudo atom groups. To identify compact units, Lesk & Rose described the protein fragments as inertial ellipsoids and selected the most compact ones using a progressive growing approach. The method proposed by Sowdhamini & Blundell to identify protein domains and super secondary elements was based on Cα-Cα distances between secondary structures. The algorithm developed by Tsaï & Nussinov used a complex scoring function, based on compactness, hydrophobicity and isolatedness, measures stability of a candidate building block. Only few servers are accessible to the scientific community at this time and mainly focus on protein domains (Alexandrov and Shindyalov, 2003; Pugalenthi *et al*., 2005).

In this paper we propose a description at an intermediate level of organization, between secondary structures elements and domains, the Protein Unit (PU). A PU is a compact subregion of the 3D structure corresponding to one sequence fragment. The basic principle is that each PU must have a high number of intra-PU contacts, and, a low number of inter-PU contacts. To identify theses PUs, we have developed a novel method called 'Protein Peeling'. This approach aims at cutting the 3D protein structure into a limited set of PUs. The an algorithm aims to define a series of successive nested partitions; this leads to the building of a tree (or a hierarchy) showing the successive splitting of the PUs into sub-PUs.

Thus, an organization of protein structures can be considered in a hierarchical manner: from secondary structures to structural domains with the PUs as intermediate elements.

2 METHODS

2.1 Principle of the 'Protein Peeling' algorithm

The 'Protein Peeling' method is summarized in the flowchart of Figure 1. It lies on an iterative principle similar to hierarchical clustering. The consecutive steps are: (1) at a given iteration, the protein sequence is cut into M PUs. For each PU, the optimal cutting into 2 or 3 sub-PUs is searched. A criterion, named partition index (PI) derived from the Matthews correlation coefficient (MCC), assesses the quality of every split by quantifying the subunits independence in terms of contacts. The resulting cutting maximizes the PI criterion. (2) A squared entropy-derived correlation R is used for assessing globally the quality of the protein partition into PUs. As this

*To whom correspondence should be addressed.
†Deceased.

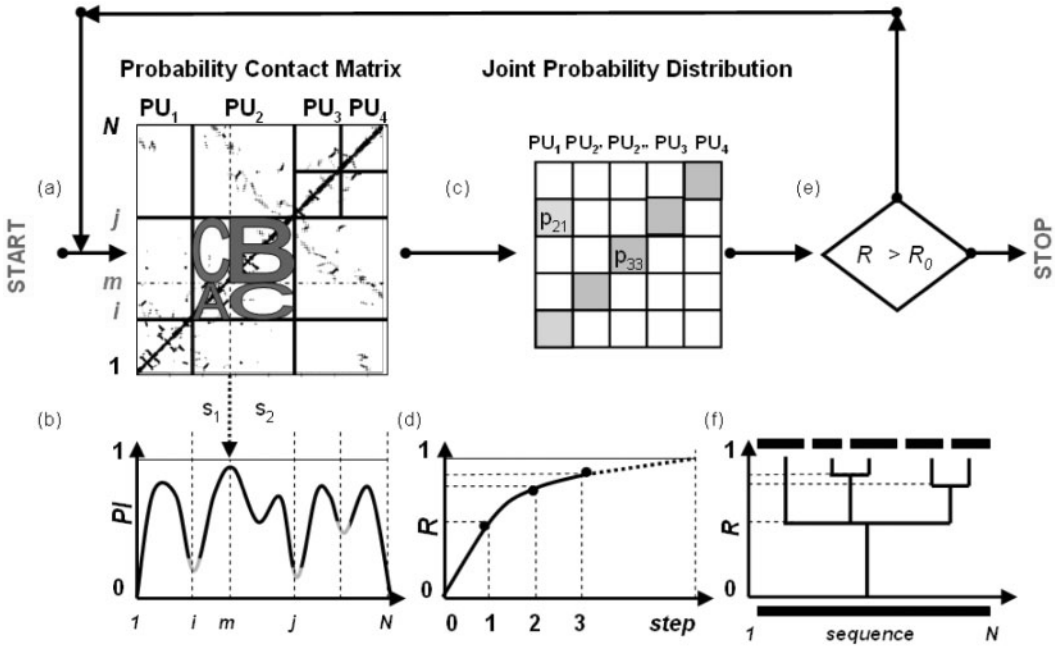

Fig. 1. Flowchart of protein peeling process. The protein has already been cut into 4 PUs (see Supplementary Data 1 for details). (**a**) The contact probability matrix with previous splitting event is represented. The Partition Index (PI) is computed for every PU, it uses the intra PUs contacts (shown by letters A and B) and the inter PUs contacts (C). The corresponding scores are shown in (**b**). The maximal PI score is found associated the position m of PU_2. So, PU_2. is split into $PU_{2'}$. and $PU_{2''}$, R value is estimated to quantify the quality of the splitting process. (**c**) R is based on joint probability distribution, i.e. the proportions of contact within and between every PUs. (**d**) R increases with the number of splitting events. (**e**) If R is more than R_0 a user-fixed threshold, the protein peeling is stopped, otherwise a new split is done. (**f**) A tree representation of PUs cutting and evolution of R could so be done to analyze the organization of PUs.

correlation increases after each new split, the process is reiterated while the R-value is less than an user-fixed threshold R_0.

2.2 A contact probability matrix for defining a continuous measure of the contacts

The contacts within a protein 3D structure can be characterized by the inter-Cα distance matrix **D**. The inter-Cα distance is simply an Euclidean distance between the Cα of the polypeptide chain. Classically, the contact matrix can be defined as a Boolean matrix **C** in which the element $C(i, j)$ equals 1 if the inter-Cα distance $d(i, j)$ between the ith and jth Cα of the protein backbone is less than a cut-off d_0. We used a logistic transformation to have continuous values instead of Boolean values (see Equation 1). Thus, the contact probability matrix **P** is derived from the matrix **D** of the inter-Cα distances; the distance $d(i, j)$ is translated into a probability $p(i, j)$ by a logistic function:

$$p(i,j) = \frac{1}{1 + \exp\left[\frac{d(i,j) - d_0}{\Delta}\right]} \quad (1)$$

with the parameters d_0 and Δ fixed to 8 and 1.5, Å, respectively in our study. For a distance $d(i, j)$ close to 0, the contact probability $p(i, j)$ is almost 1. For a large distance, $p(i, j) = 0$.

2.3 A partition index used for defining an optimal splitting of a protein unit into subunits

As previously defined, one PU is a protein fragment structurally compact, i.e. the contacts between the residues of the PU are numerous, compared to the number of contacts between the residues of this PU and the other residues of the protein. A measure must be defined to assess the relevance of the cutting of a PU into 2 or 3 subunits.

A PU (or at the beginning, the whole protein) is associated with a protein sequence $\boldsymbol{s}$ comprised between positions $[i, j]$, ($i < j$, $i = 1$ and $j = N$ at the beginning). The sequence is cut into two parts $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ associated with the positions $[i, m]$ and $[m + 1, j]$, respectively. The symmetric contact probability submatrix associated with the sequence $\boldsymbol{s}$ is shared into 3 sub-matrices corresponding to the sum of the contact probabilities between the residues of $\boldsymbol{s_1}$ with itself (noted A, Fig. 1), $\boldsymbol{s_2}$ with itself (B), and $\boldsymbol{s_1}$ with $\boldsymbol{s_2}$ (C).

To assess the presence of numerous contacts within the subunits $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ and a limited number of contacts between them, we have used Matthews' coefficient correlation (MCC) (Matthews, 1976). The MCC measure is translated into a partition index, $PI_{ij}(m)$:

$$PI_{i,j}(m) = \frac{AB - C^2}{(A + C)(B + C)} \quad (2)$$

Thus, the quality of the splitting of the PU into two subunits is quantified via a correlation. The complete absence of contacts between these two subunits (i.e. $C = 0$) leads to a maximal value of the partition index (i.e. 1). A large presence of contacts between subunits ($C > 0$) induces a low PI value. To define the optimal splitting of a given unit, we search for the position m ($i < m < j$) in the protein sequence such as where the partition index is maximal.

We also considered the possibility of splitting a given unit into three units rather than two (see Supplementary data 2a). Some constraints on the splitting may be added; for instance, the cutting cannot appear in repetitive secondary structures when their lengths are useless than $L_{\min}$ (only the long α-helices or β-strands can be cut).

2.4 Assessing the global 3D structure protein splitting into PUs

The 'protein peeling algorithm' carries out a series of nested partitions of the 3D protein structure. Nonetheless, it is necessary to assess globally the

protein splitting and to define the rule for stopping the process. For this purpose, we used a squared entropy-derived correlation, R (Hazout, submitted for publication), based on the contacts within and between PUs. This index quantifies globally the dependence of the PUs in terms of contact.

The contact probability matrix **P** is translated into a joint probability matrix p_{XY} of dimension $M \times M$ (M is the number of PUs). The matrix elements are the sums of the contact probabilities within PUs or between PUs, divided by the global contact probability sum (Fig. 1c).

The mutual entropy $M(X, Y)$ is computed (see Equation 3). It quantifies the dependence between two discrete random variables, X and Y, from the joint probability distribution, p_{XY}, and the marginal probability distributions, p_X, and p_Y.

$$M(X,Y) = \sum_{i=1,M} \sum_{j=1,M} p_{XY}(i,j) \ln \left[\frac{p_{XY}(i,j)}{p_X(i) p_Y(j)} \right] \quad (3)$$

The squared entropy-derived correlation R extends the classical concept of correlation defined for two continuous variables (Pearson correlation ρ) to two discrete variables. Its expression is:

$$R = \sqrt{1 - \exp[-2M(X,Y)]} \quad (4)$$

This squared correlation measures the independence level between two discrete variables. It varies between 0 and $R_{\max}$ (see Supplementary data 2b). Thus, in the absence of contacts between PUs, the matrix is diagonal (R close to 1).

2.5 An entropy-derived measure for assessing the PU compactness

To analyze individually every PU, we have computed an entropy-derived measure that allows assessing their compactness. This index named PU compaction index (CI) focuses on the non-local contact in the PUs.

We select the submatrix $\mathbf{P}_m$ from the matrix **P**, corresponding to the contact probabilities within the mth PU (noted PU_m), and we transform this matrix into a joint probability distribution $\mathbf{p}_m$ $\{p_m(i, j), i = 1, N_m, j = 1, N_m\}$, where N_m denotes the number of residues located in the studied PU (see Supplementary data 2c).

As a first step, the 'equivalent number of contacts within a given PU' is computed. This quantity is derived from the 'equivalent number of states' (de Brevern *et al*., 2000; de Brevern and Hazout, 2003; Etchebest *et al*., 2005) and is noted $Neq(PU_m)$.

$$Neq(PU_m) = e^{H(X,Y)} \quad (5)$$

with

$$H(X,Y) = -\sum_{i=1,Nm} \sum_{j=1,Nm} p_{XY}(i,j) \ln[p_{XY}(i,j)] \quad (6)$$

This index measures the diversity in terms of number of states, and it varies between 1 and ${N_m}^2$. Then, we eliminate the distant contacts (i.e. $|i - j| > 6$) from the conjoint probability distribution $\mathbf{p}_m$ and compute the equivalent number of 'closest' contacts $Neq(PU_m{}^*)$. The difference $\Delta Neq = [Neq\ (PU_m) - Neq(PU_m{}^*)]$ measures the equivalent number of distant contacts. So the ratio $\Delta Neq/N_m$, defines CI. It measures the equivalent number of distant contacts per residue within a given PU. A weak value specifies a unit of type extended, and a high value a unit of type compact.

3 RESULTS

3.1 A protein example

Figure 2 summarizes different results of the peeling process performed on globular actin (pdb code 1atnA) using our server (www.ebgm.jussieu.fr/~gelly).

Hierarchical tree splitting. A Progressive top-down splitting algorithm produces PUs of decreasing size. To inspect manually the organization of protein units at different levels of the peeling process, a tree is constructed (Fig. 2A). This tree shows the different levels of the hierarchical organization of various substructures into proteins. Starting from the entire protein at the root of the tree, nodes depict splitting events and are linked to their corresponding parent nodes. Leafs of the node represent the final PUs; they are presented following their relative positions in the polypeptide sequences. Three levels of PU splitting are displayed.

In the first stage, the structure is cut out in three areas [1–137] (CI = 1.3), [138–340] (CI = 1.0) and [341–374] (CI = 0.0). Two compact areas of large size (137 and 203 residues) and a C-terminal area in the form of a curved propeller of 38 residues appear. Then the second area is cut out in three segments corresponding to secondary superstructures at the second stage. As the last step, the first area is cleaved into three areas detailed in Figure 2A: a β strand, an $\alpha + \beta$ and an α/β. The end of the procedure shows the characterization of 7 PUs. All figures were dynamically generated on the web site with the Pymol program (DeLano, 2002 http://www.pymol.org).

Contacts map. Figure 2B shows the contact map. The probability contacts are drawn from black (corresponding to a probability of 1) to white (probability equal to 0). The final protein units produced at the last level of splitting are displayed with coloring area corresponding to the colors of Figure 2A along the diagonal, i.e. the sequence. This representation is a convenient way to understand the final splitting process and eventually manually controls the procedure. It appears that the strongest densities of contact are well joined together in central the square blocks and the contacts with the long range are well isolated by cutting.

3.2 Comparison to others methods and to experimental results

We have compared our own cutting algorithm to some methods that aim at cutting the protein structure into smaller units. Comparison for different proteins is shown in Figure 3.

In Figure 3A, final results of globular actin dissecting procedure with Protein Peeling and Tsaï and Nussinov methods are displayed (Tsai and Nussinov, 1997). In each case, three principal parts, corresponding to domains, have been identified. In the further steps, these three areas have been split into similar elements of comparable sizes and delimitations. In Figure 3B, lysosyme partitions of the protein Peeling and Foldons methods are detailed (Panchenko *et al*., 1996). The resulting units are quite comparable and compatible for all methods, in terms of size and limits along the sequence.

The intrinsic performance of the method and compatibility with a folding model were compared with results of the protein folding experiments carried out by Rumbley and co-authors for cytochrome c (Rumbley *et al*., 2001). It is comforting to note that the first levels of dissection correspond to experimental delimitation of different early folded elements (Fig. 3C). All the different elements (N- and C-terminal helix segments and connecting regions) observed experimentally are found with the Protein Peeling approach.

3.3 Peeling a protein: web server and specific tools

One of our goals is to provide an easy-to-use interface to splitting a protein and to offer a database of pre-cutting proteins for proposed analysis. Thanks to the PISCES server (Wang and Dunbrack, 2003), we obtained a representative set of structures extracted from the

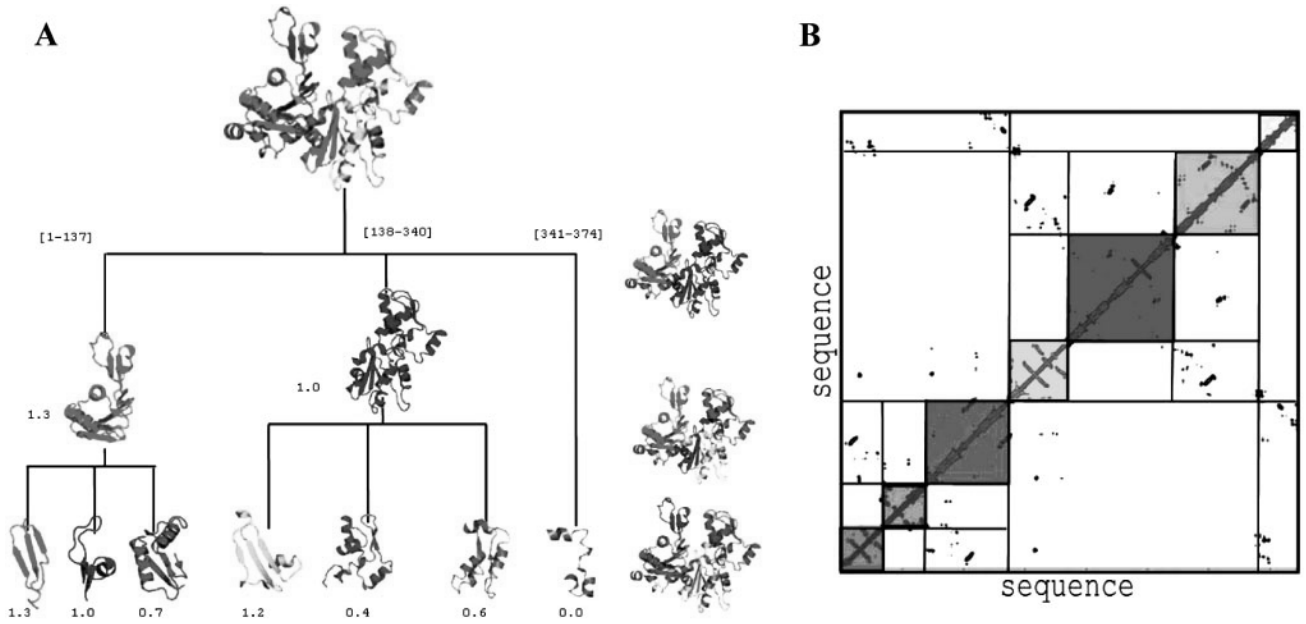

Fig. 2. Protein peeling example of actin protein (pdb code 1atn) with default parameters, shown with PyMol software (DeLano, 2002), $R_0 = 95$, minimal size of PU = 16 residues. (**A**) Hierarchical tree of peeling process with compaction indexes and limits of protein units. (**B**) Matrix of contact probability. The contact probabilities are represented from black ($p(i,j) = 1$) to white ($p(i,j) = 0$), the grey colors represent intermediate probabilities. Each splitting event is represented by vertical and horizontal lines. Contact area of each PU obtained at the last step is colored accordingly to protein unit color in Figure 2A.

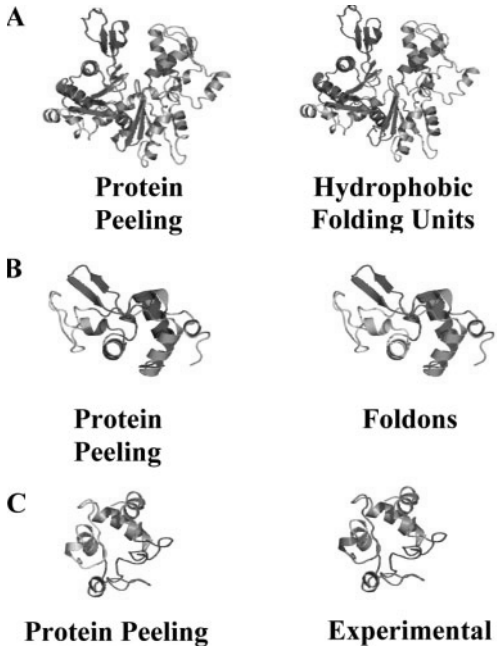

Fig. 3. Proteins splitting by peeling and other methods. (**A**) Globular actin (Tsai and Nussinov, 1997); (**B**) Lysosyme (1lys) (Panchenko *et al*., 1996); (**C**) Cytochrome **C** (1akk) (Rumbley *et al*., 2001).

Protein Data Bank (Berman *et al*., 2000). This non-redundant set of protein structures includes 2309 elements from crystallographic experiments with better than 2 Å of resolution. The proteins shared no >30% of sequence identity. All these structures have been dissected by the Protein Peeling procedure. Results were stocked in a flat file database and these pre-cut proteins could be easily accessed with a search engine displayed on the main page of the peeling protein Internet site.

4 PERSPECTIVES

Our database of pre-cutting proteins provides useful materials for further analysis on the structure, size, composition in amino acid and secondary structures of protein units. Such experiments open the way to other ambitious developments like construction of three-dimensional structures of proteins with protein units as it has been shown with similar approaches (Haspel *et al*., 2003; Inbar *et al*., 2003). Our research will focus on the comparisons with related works such as the structural trees of proteins (Efimov, 1997), DIAL (Pugalenthi *et al*., 2005) or the PDP approach (Alexandrov and Shindyalov, 2003). In the same way, other compaction indexes will be implemented in our server.

ACKNOWLEDGEMENTS

This paper is dedicated to the loving memory of Pr. Serge Hazout. The authors want to express their appreciation to Cristina Benros and Catherine Etchebest for their kind help in the course of this research. This work was supported by ACI Action Bioinformatique 2003–2004. We thank Université Paris 7, Institut National de la Recherche Médicale, Centre National de la Recherche Scientifique and Ministère de l'Education Nationale de l'Enseignement Supérieur et de la Recherche for partial support.

Conflict of Interest: none declared.

REFERENCES

Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429–430.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chothia,C. and Finkelstein,A.V. (1990) The classification and origins of protein folding patterns. *Annu. Rev. Biochem.*, **59**, 1007–1039.

de Brevern,A.G. *et al*. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**, 271–287.

de Brevern,A.G. and Hazout,S. (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics*, **19**, 345–353.

DeLano,W.L.T. (2002) , The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.

Efimov,A.V. (1994) Common structural motifs in small proteins and domains. *FEBS Let*., **355**, 213–219.

Efimov,A.V. (1997) Structural trees for protein superfamilies. *Proteins*, **28**, 241–260.

Etchebest,C. *et al*. (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins*, **59**, 810–827.

Gelly,J.C. *et al*. (2004) The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic Acids Res*., **32**, D156–159.

Go,M. (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*, **291**, 90–92.

Guo,J.T. *et al*. (2003) Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res*., **31**, 944–952.

Haspel,N. (2003) Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci*., **12**, 1177–1187.

Inbar,Y. *et al*. (2003) Protein structure prediction via combinatorial assembly of substructural units. *Bioinformatics*, **19**(Suppl. 1), i158–168.

Janin,J. and Wodak,S.J. (1983) Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol*., **42**, 21–78.

Lesk,A.M. and Rose,G.D. (1981) Folding units in globular proteins. *Proc. Natl Acad. Sci. USA*, **78**, 4304–4308.

Matthews,B.W. (1976) X-ray crystallographic studies of proteins. *Annu. Rev. Phys. Chem*., **27**, 493–523.

Michalopoulos,I. *et al*. (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res*., **32**, D251–254.

Panchenko,A.R. *et al*. (1996) Foldons, protein structural modules, and exons. *Proc. Natl Acad. Sci. USA*, **93**, 2008–2013.

Pugalenthi,G. *et al*. (2005) DIAL: a web-based server for the automatic identification of structural domains in proteins. *Nucleic Acids Res.*, **33**, W130–132.

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

Rumbley,J. *et al*. (2001) An amino acid code for protein folding. *Proc. Natl Acad. Sci. USA*, **98**, 105–112.

Sibanda,B.L. and Thornton,J.M. (1993) Accommodating sequence changes in beta-hairpins in proteins. *J. Mol. Biol*., **229**, 428–447.

Sowdhamini,R. and Blundell,T.L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein. Sci*., **4**, 506–520.

Tsai,C.J. and Nussinov,R. (1997) Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci*., **6**, 24–42.

Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.