

A Graph Representation for Protein β -Sheets and its Applications

Marc Parisien and François Major

*Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal, CP 6128 Succ. Centre-Ville,
Montréal, Québec, Canada H3C 3J7*

Abstract

A graph representation for protein β -sheets is here presented. The residue-level graph encodes all the topological features of β -sheets; peptidic bonds between residues of the same strand, β -sheet partnership and inter-strand H-bonding. The β -sheet topological graph has thus the expressiveness to accurately model all β -sheet features such as standard parallel/anti-parallel H-bond motifs, β -bulges and β -barrels. A database of atomic coordinates of β -sheets and their corresponding β -sheet topology graph has been compiled. A sub-graph isomorphism algorithm has been adapted from Ullman's original form to compare β -sheet topological graphs of appreciable sizes, therefore the β -sheet database can be scrupulously examined for particular motif searches. Applications vary from *De Novo* protein design to β -bulges and β -barrels analysis.

Key words: graph theory, β -sheet, β -bulge, β -barrel, subgraph isomorphism.

INTRODUCTION

Graphs are powerful mathematical abstractions which enable us to express relations, either quantitative, like distance costs, or qualitative, like pedigrees in genealogical trees, between connected objects. Literature shows that there are several graph representations of proteins, which allow for various tasks such as protein sketching [1], Nuclear Magnetic Resonance (NMR) structure determination [2,3] as well as motifs comparison, similarity search, protein fold identification and classification [4–8].

β -sheets are present in more than 80% of globular proteins. Proteins can be classified with the help of their β -sheet topology [9–11], i.e. number of strands, relative orientations of strands within the β -sheet. Also, similar β -sheet structures show to have conserved amino-acid type at specific key positions, which qualifies them as folding nucleus residues [12–14].

β -sheet topology graphs can be inferred from low resolution NMR spectroscopy where inter-strand H-bonds can be deduced from slowly exchanging amide resonances. Other cross-strands interactions can be calculated with help of strong and weak Nuclear Overhauser Effect (NOE [15–17]) data [18–28].

Secondary structure prediction or elucidation in conjunction to strand pairs alignment prediction [29–39] can also lead to sheet topology graphs. In this case, two hypothesis on the inter-strand H-bonds (and correspondingly, on the orientation up/down of a side-chain for a given residue) must be explored.

Once a graph representation of protein β -sheet can be calculated one can build a database of such graphs, and their corresponding amino-acid atomic 3-D coordinates, in order to compare them with each other or to perform a particular β -sheet motif search within this database. Indeed, several applications of the β -sheet topology graph can be thought of, notably:

- (1) β -sheet modeling. Given a β -sheet topology graph, each graph matching in the database can supply atomic 3-D coordinates that can be held as models for this β -sheet topology, and will conform in all points to the specified graph. This can be seen as going from a 2-D representation, i.e. the β -sheet topology graph, to actual 3-D models of β -sheets. It goes further than homology modeling since we do not restrict ourselves to β -sheet coordinates from proteins having similar sequences to the target, thus addressing the conformational search space spanned by β -sheets found in the PDB [40,41].
- (2) Sequence analysis. When a particular β -sheet motif is searched for in the database, all the graph-matching solutions can be superimposed, and thus the sequence entropy at each vertex, or residue position in the β -sheet topology graph, can be calculated to reveal key amino-acid types

for this particular motif. This can be done for β -bulges, β -barrels and for β -sheets from the same structural family but with distant sequences.

METHOD

β -Sheet Topology Graph

A graph $G = (V, E)$ is composed of a finite set of vertices $V = \{x_i\}$ and a finite set of edges $E = \{(x_i \in V, y_i \in V)\}$. The graph is said to be oriented if the edges set has ordered pairs. The in degree of a vertex is the number of oriented edges that is incident to that vertex, while the out degree is the number of edges leaving the vertex. A β -sheet topology graph is a graph G in which the set of vertices V is all the residues contained in the β -sheet, while the set of edges E describes all the various topological relations between the residues of V . This graph is said to be at residue level, because vertices of the graph are residues, compared to other graph representations of proteins, like secondary structure level graphs in which vertices are now secondary structure units. The β -sheet topology graph is oriented; edges start at the lowest residue number to terminate at the highest residue number of the connected residue pair, thus outlining the β -strand progression from N terminal to C terminal, as well as the relative β -strand positions within the β -sheet, and is weakly connected, that is, there is an undirected path between any pair of residues (however it is not strongly connected because some residues may have in degrees of zero), therefore each residue is attached to the β -sheet graph with either of the topological relations. The topological relations that are censured are:

Type C This relation expresses the backbone $C_i \rightarrow N_{i+1}$ peptidic covalent bond between residues R_i and R_{i+1} .

Type H This relation expresses the presence of at least an H-bond between residues R_i and R_j . H-bonds are defined by a classical Coulomb electrostatic interaction energy as calculated in the DSSP program [42].

Type P This relation expresses a β -sheet partnership between residues R_i and R_j , which are in the same register, thus side-by-side in the β -sheet. This partnership relation is also calculated by DSSP [42].

Type HP This relation is used when types **H** and **P** are simultaneously exhibited.

From there, we can say that the edge set E is composed of the specific edge type sets; $E = E^C \cup E^H \cup E^P$ (with $E^{HP} = E^H \cap E^P$).

Database

The culled PDB Select 25 database [43], a subset of PDB [44] whose sequences share no more than 25% identity, is used to provide atomic coordinates for the backbone of β -sheets. As of the 7th of February 2004, this database had 1966 chains. The cutoff values are 25% for sequence identity, 2.0 Å for resolution and 0.25 for R-Factor. Residues that are flagged by DSSP [42] in the 'E' state are considered as part of a β -sheet. The sharing of the same β -sheet identification number by such residues ensures that the corresponding graph is weakly connected. From there, a graph representation of the β -sheet is calculated, in which the vertices of the graph are the residues found within the β -sheet, while the edges encode for the β -sheet topological relations. Every β -sheet feature can be represented in a β -sheet topology graph; from standard parallel or anti-parallel β -strand pairings to non-canonical inter-strand H-bonding motifs, β -bulges and β -barrels. Figure 1 shows a hypothetical mixed parallel/anti-parallel β -sheet with all its topological relations explicated.

Subgraph Isomorphism

Once a β -sheet topology graph is defined we can now tackle the problem of comparing these graphs, or more specifically, to answer the question of whether a graph is included in another one. This problem is known as the subgraph isomorphism problem; is there a subgraph of one graph which is isomorphic to another graph, and is considered NP-complete [45]. Although the fact that β -sheet topology graphs are planar (even for β -barrels), and a polynomial-time algorithm exists [46] for solving such problems when the considered graphs are planar, we adapted Ullman's subgraph isomorphism algorithm [47] to the particular structure of β -sheet topology graphs, thus making possible the comparison of graphs with several hundred residues in few seconds of computation time.

Ullman's subgraph isomorphism algorithm between two graphs, $G^1 = (V^1, E^1)$ and $G^2 = (V^2, E^2)$ with $|V^1| \leq |V^2|$, starts by filling a matrix M^0 of size $|V^1| \times |V^2|$, in which an element in M_{ij}^0 is equal to 1 if it is possible, *a priori*, to map the i^{th} vertex of G^1 , namely G_i^1 , on the j^{th} vertex of G^2 , G_j^2 , by considering the number and type of incoming and outgoing edges of G_i^1 and G_j^2 . Let $In(G_i, \mathbf{T})$ be a function that counts the number of incoming edges of a given type \mathbf{T} in a graph G for the i^{th} vertex. Similarly, let $Out(G_i, \mathbf{T})$ count

the number of outgoing edges of type **T** for the i^{th} vertex in G . Then:

$$M_{ij}^0 = \begin{cases} \begin{array}{l} \text{In}(G_j^2, \mathbf{C}) \geq \text{In}(G_i^1, \mathbf{C}) \text{ and} \\ \text{In}(G_j^2, \mathbf{H}) + \text{In}(G_j^2, \mathbf{HP}) \geq \text{In}(G_i^1, \mathbf{H}) \text{ and} \\ \text{In}(G_j^2, \mathbf{P}) + \text{In}(G_j^2, \mathbf{HP}) \geq \text{In}(G_i^1, \mathbf{P}) \text{ and} \\ \text{In}(G_j^2, \mathbf{HP}) \geq \text{In}(G_i^1, \mathbf{HP}) \text{ and} \\ \text{Out}(G_j^2, \mathbf{C}) \geq \text{Out}(G_i^1, \mathbf{C}) \text{ and} \\ \text{Out}(G_j^2, \mathbf{H}) + \text{Out}(G_j^2, \mathbf{HP}) \geq \text{Out}(G_i^1, \mathbf{H}) \text{ and} \\ \text{Out}(G_j^2, \mathbf{P}) + \text{Out}(G_j^2, \mathbf{HP}) \geq \text{Out}(G_i^1, \mathbf{P}) \text{ and} \\ \text{Out}(G_j^2, \mathbf{HP}) \geq \text{Out}(G_i^1, \mathbf{HP}) \end{array} & 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

As an example, consider the graphs G^1 and G^2 in Figure 2. The resulting M^0 matrix is shown in Table 1(a).

Because of the special nature of the β -sheet topology graph, when a residue R_i^1 of graph G^1 is mapped onto a residue R_j^2 of graph G^2 we want also that the entire strand S_i^1 which includes R_i^1 be also mapped to the corresponding strand S_j^2 in G^2 which contains R_j^2 . That requirement makes us consider sub-matrices of M^0 , $S_i^1 \times S_j^2$, and zero-out any diagonals of these sub-matrices which admit a null entry. Now, the M^0 matrix contains much less one's, thus reducing the number of possible mappings to check. To pursue our example, the new matrix M^0 is shown in Table 1(b). Notice that the surviving diagonals have a residue mapping for each residue of each strand of G^1 , which happens to be the smallest graph in terms of number of residues.

Ullman's algorithm now generates permutation vectors $V[i] = j$ that tell which residue R_i^1 of G^1 is mapped on which residue R_j^2 of G^2 . The permutation vectors have $|V^1|$ entries. Since edges in β -sheet topology graphs are directed from lowest residue number to highest not all generated permutations are valid; only those that are strictly increasing are taken into account. This requirement has a huge impact on the pruning of the backtrack tree that is used to generate the permutation vectors.

The isomorphism test is then applied for each valid permutation vectors. For each labeled edge $e_{kl}^1 \in V^1 \times V^1$ (we suppose that if $e_{kl}^1 \notin E^1$ then it's label is \emptyset instead of one of **C**, **H**, **P**, **HP**) consider the mapped edge $e_{V[k]V[l]}^2$ linking the two mapped residues $R_{V[k]}^2$ and $R_{V[l]}^2$ in G^2 . For the isomorphism test between graphs G^1 and a subgraph of G^2 be found true, each individual label comparison $e_{kl}^1 \leftrightarrow e_{V[k]V[l]}^2$ must be held true, or else G^1 is not an isomorphic subgraph of G^2 relative to the given permutation vector V . The following truth table can be found in Table 2. Briefly, this table says that a covalent **C** edge in G^1 must be mapped to a covalent **C** edge in the isomorphic subgraph in G^2 , that a partner **P** edge can be mapped to either a partner **P** or an **HP** edge, that an H-bond edge **H** can be also mapped to either a partner **P** or

an **HP** edge, that an **HP** edge must map onto an **HP** edge, and finally that an unspecified relation \emptyset can be mapped to any types except the covalent **C** edge.

In our example, a total of four permutation vectors are generated but only two will pass successfully the isomorphism test. The first solution maps residues $\{111, 112, 113, 121, 122, 123\}$ of G^1 onto $\{211, 212, 213, 222, 223, 224\}$ of G^2 , while the second maps onto $\{221, 222, 223, 232, 233, 234\}$. Notice that the unspecified relation between residues R_{112}^1 and R_{122}^1 in G^1 does not prevent a more specific mapping in G^2 in which type **P** is found.

Distances

The sequence distance between two isomorphic graphs is the sum on all residue positions of the amino-acid substitution similarity score given in the PAM250 matrix [48], as if the β -sheet topological graph would serve as the alignment template. Higher values indicate higher similarity between the sequences of the isomorphic graphs. The RMSD, or root mean square deviation, is calculated in the standard way. At a given β -sheet residue position, one can calculate the sequence variation or entropy using Shannon's [49] celebrated equation $\sum_C p_i \log(p_i)$. The Kullback-Leibler distance [50] between two probability distributions, p and q , is $\sum_C p_i \log(p_i/q_i)$. The last two measures are dependent of the amino-acid partitioning C used; from the two class hydrophobic/polar split (HP model) to the full-fledged 20 symbols break up [51,52]. A better amino-acid partition could be generated using the methodology of Wang and Wang [53] not on the context-independent MJ matrix [54] but rather on the β - β environment-dependent one [55]. Furthermore, the β - β MJ matrix could be made specific for parallel and anti-parallel strands pairing [56,57].

RESULTS and DISCUSSION

Protein Design

Although the graph matching algorithm is not able to generate new β -sheet topologies, it can assess whether or not if a given topology is unheard of. In a recent article published by Baker et al. [58], the group claims to have engineered a novel protein fold expressed as Top7. The protein's β -sheet topology graph has been used to scan the graph database and was found to be no isomorphic subgraph of any β -sheet in the culled PDB Select 25, it is thus a unique β -sheet topology.

On the other hand, the β -sheet topology of human monocyte chemoattractant protein MCP-1 (PDB code 1DOK [59,60]) (See Figure 3(a)) has 71 isomorphic siblings in the culled PDB Select 25, not only in the Interleukin 8-like chemokines SCOP [10] family, but also in various other unrelated protein classes, as shown in Table 3. The MCP-1 β -sheet has a bulged residue at position ALA26, thus making the corresponding topology graph depart from the ubiquitous canonical anti-parallel 3-stranded sheets (we obtain 411 isomorphic β -sheet if we eliminate residue 25 and make residue 26 a partner **P** to residue 45). Since this β -sheet topology, including the β -bulge, seems to be used in a wide variety of proteins of different functions it can be speculated about the specific role of the β -bulge as to prevent amyloidosis [61], since the bulged strand is exposed to solvent. Figure 4 shows a superposition of the heavy backbone atoms of the 71 isomorphic graphs. The models are superimposed on their first strand, which comprises residues 25 to 31. The two closest models have an RMSD of 0.39 Å, while the two farthest have 6.99 Å. These models can serve as template for either *De Novo* protein design or homology modeling.

β -Bulge

Sequence analysis has been performed on the β -bulge region of MCP-1, residues 25,26,27,28,43,44,55. This β -bulge pattern is the most common occurring in anti-parallel strand pairs, and is of type C+ (classic), as defined in [62]. A total of 711 isomorphic siblings can be found in the culled PDB Select 25 for this particular residue motif shown in Figure 3(b). The results are summarized in Table 4. Four amino-acid classes were used to partition the findings of the subgraph isomorphism algorithm, and are those taken from [52], Table II. This β -bulge motif has particular amino-acid preferences at key positions. It is noteworthy to mention that our results differ slightly from those published in [62]. In particular, we find that residues in position **1** are not necessarily large hydrophobic residues, but instead, large aromatic residues and cysteine {CFYW} (only 6% occurrence) seem to be forbidden. Also, not only small residues {GPATS} (39%) are favored at position **2**, but also large polar residues {NHQEDRK} (41%), which account for the larger observed percentage, although large hydrophobic residues {MLIV} (only 7%) seem to be denied. Position **X** does not seem to have a particular preference, at least in our amino-acid subdivisions. It is also interesting to note some deviance from the reference amino-acid composition, especially at position 25 where large hydrophobic residues predominate (68%), and at position 45 for which aromatic residues are observed more than normally (26%). Even though the Shannon entropy measures in Table 4 doesn't reveal conserved amino-acid positions, exception made at residue position 25 where the entropy is very low due to the high presence of large hydrophobic amino-acids, the Kullback-Leibler

distance, on the other hand, is more sensible to departures to the reference amino-acid distribution (from β -sheets of PDB Select 25), and shows that positions 25 (highly large hydrophobic), 27 (predominantly polar or charged residues) and 45 (important aromatic occupancy) are of extremely specific amino-acid distribution, and could prove to be necessary for the C+ β -bulge type fold.

β -Barrel

A β -barrel is a β -sheet in which the first strand is H-bonded to the last one, and makes a barrel-like 3-D structure. Two parameters fully describe the topological features of regular β -barrels; the number of strands, n , taking part in the barrel and the shear number, S , that is the distance separating the starting residue on the first strand and the terminal residue on the same strand after a walk around the barrel in a direction perpendicular to the strand orientations [63]. Figure 5 shows two alternatives for ($n=8, S=8$) β -barrels called β -rings. Here, the relative strands orientation is parallel. The difference between the two graphs is the handedness of the partnership \mathbf{P} connections. It is interesting to note that these two β -ring motifs yield quite different results as for isomorphic siblings in the culled PDB Select 25 database. Figure 5(a) has 85 solutions while Figure 5(b) has none. This is due to the handedness of the crossover connections between two consecutive parallel strands [64]. The β -barrel encoded in the β -sheet topological graph of Figure 5(a) would put the α -helices *outside* of the barrel, whereas the one if Figure 5(b) would put them *inside* the barrel to satisfy the connection handedness (almost all examples are found to be right-handed). Within this theoretical framework it is possible to address various values of n and S .

CONCLUSION

A graph representation for protein β -sheets is here presented. The residue-level graph encodes all the topological features of β -sheets; peptidic bonds between residues of the same strand, β -sheet partnership and inter-strand H-bonding. The β -sheet topological graph has thus the expressiveness to accurately model all β -sheet features such as standard parallel/anti-parallel H-bond motifs, β -bulges and β -barrels. A database of atomic coordinates of β -sheets and their corresponding β -sheet topology graph has been compiled. A sub-graph isomorphism algorithm has been adapted from Ullman's original form to compare β -sheet topological graphs of appreciable sizes, therefore the β -sheet database can be scrupulously examined for particular motif searches. Applications vary from *De Novo* protein design to β -bulges and β -barrels analysis. The LINUX

version of the program and associated database can be found at the following address: <http://www-lbit.iro.umontreal.ca/bSheet/index.html>.

ACKNOWLEDGMENTS

We would like to thank Vincent Devloo for reviewing this manuscript. This work was jointly funded by the Canadian Institutes of Health Research (CIHR MT-14604), Genome Québec and Genome Canada. FM is a CIHR investigator.

References

- [1] M. Parisien, M. C. Peitsch, F. Major, A protein conformational search space defined by secondary structure contacts, *Pac. Symp. Biocomput.* 243 (1998) 425–436.
- [2] E. C. van Geerestein-Ujah, M. Mariani, H. Vis, R. Boelens, R. Kaptein, Use of graph theory for secondary structure recognition and sequential assignment in heteronuclear (^{13}C , ^{15}N) NMR spectra: application to HU protein from *Bacillus stearothermophilus*, *Biopolymers* 7 (1996) 691–707.
- [3] C. Bailey-Kellogg, A. Widge, J. J. Kelley, M. J. Berardi, J. H. Bushweller, B. R. Donald, The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data, *J Comput. Biol.* 7 (2000) 537–558.
- [4] E. M. Mitchell, P. J. Artymiuk, D. W. Rice, P. Willett, Use of techniques derived from graph theory to compare secondary structure motifs in proteins, *J. Mol. Biol.* 212 (1990) 151–166.
- [5] I. Koch, F. Kaden, J. Selbig, Analysis of protein sheet topologies by graph theoretical methods, *Proteins* 12 (1992) 314–323.
- [6] H. M. Grindley, P. J. Artymiuk, D. W. R. DW, W. P, Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm, *J. Mol. Biol.* 229 (1993) 707–721.
- [7] P. J. Artymiuk, A. R. Poirrette, H. M. G. HM, D. W. Rice, P. Willett, A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures, *J. Mol. Biol.* 243 (1994) 327–344.
- [8] A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton, C. Orengo, Recognizing the fold of a protein structure, *Bioinformatics* 19 (2002) 1748–1759.

- [9] J. S. Richardson, beta-sheet topology and the relatedness of proteins, *Nature* 268 (1977) 495–500.
- [10] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247 (1995) 536–540.
- [11] A. V. Efimov, A structural tree for proteins containing s-like beta-sheets, *FEBS Lett.* 437 (1998) 246–250.
- [12] S. W. Michnick, E. Shakhnovich, A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies, *Fold. Des.* 3 (1998) 239–251.
- [13] N. Kannan, S. Selvaraj, M. M. Gromiha, S. Vishveshwara, Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach, *Proteins* 43 (2001) 103–112.
- [14] R. Qamra, B. Taneja, S. C. Mande, Identification of conserved residue patterns in small beta-barrel proteins, *Protein Eng.* 15 (2002) 967–977.
- [15] A. W. Overhauser, Paramagnetic relaxation in metals, *Phys. Rev.* 89 (1953) 689–700.
- [16] A. W. Overhauser, Polarization of nuclei in metals, *Phys. Rev.* 92 (1953) 411–415.
- [17] I. Solomon, Relaxation processes in a system of two spins, *Phys. Rev.* 99 (1955) 559–565.
- [18] M. Delepierre, C. Dobson, F. Poulsen, Studies of beta-sheet structure in lysozyme by proton nuclear magnetic resonance. assignments and analysis of spin-spin coupling constants, *Biochemistry* 21 (1982) 4756–4761.
- [19] F. Inagaki, N. Clayden, N. Tamiya, R. Williams, Individual assignments of the amide proton resonances involved in the triple-stranded antiparallel pleated beta-sheet structure of a long neurotoxin, *Laticauda semifasciata* III from *Laticauda semifasciata*, *Eur. J. Biochem.* 123 (1982) 99–104.
- [20] K. Mayo, Epidermal growth factor from the mouse. physical evidence for a tiered beta-sheet domain: two-dimensional NMR correlated spectroscopy and nuclear Overhauser experiments on backbone amide protons, *Biochemistry* 24 (1985) 3783–3794.
- [21] G. Montelione, K. Wuthrich, E. Nice, A. Burgess, H. Scheraga, Identification of two anti-parallel beta-sheet conformations in the solution structure of murine epidermal growth factor by proton magnetic resonance, *Proc. Natl. Acad. Sci. (USA)* 83 (1986) 8594–8598.
- [22] P. Weber, S. Brown, L. Mueller, Sequential ^1H NMR assignments and secondary structure identification of human ubiquitin, *Biochemistry* 26 (1987) 7282–7290.

- [23] B. Stockman, A. Krezel, J. Markley, K. Leonhardt, N. Straus, Hydrogen-1, carbon-13, and nitrogen-15 NMR spectroscopy of *Anabaena* 7120 flavodoxin: assignment of beta-sheet and flavin binding site resonances and analysis of protein-flavin interactions, *Biochemistry* 29 (1990) 9600–9609.
- [24] T. Pochapsky, X. Ye, ¹H NMR identification of a beta-sheet structure and description of folding topology in putidaredoxin, *Biochemistry* 30 (1991) 3850–3856.
- [25] C. Ullman, P. Haris, K. Smith, R. Sim, V. Emery, S. Perkins, Beta-sheet secondary structure of an LDL receptor domain from complement factor I by consensus structure predictions and spectroscopy, *FEBS Lett.* 371 (1995) 199–203.
- [26] C. Ullman, P. Harris, D. Galloway, V. Emery, S. Perkins, Predicted alpha-helix/beta-sheet secondary structures for the zinc-binding motifs of human papillomavirus E7 and E6 proteins by consensus prediction averaging and spectroscopic studies of E7, *Biochem J.* 319 (1996) 229–239.
- [27] W. Zhang, T. Smithgall, W. Gmeiner, Sequential assignment and secondary structure determination for the Src homology2 domain of hematopoietic cellular kinase, *FEBS Lett.* 406 (1997) 131–135.
- [28] H. Ponstingl, G. Otting, NMR assignments, secondary structure and hydration of oxidized *escherichia coli* flavodoxin, *Eur. J. Biochem* 244 (1997) 384–399.
- [29] K. Maruyama, Y. Itoh, F. Arisaka, Circular dichroism spectra show abundance of beta-sheet structure in connectin, a muscle elastic protein, *FEBS Lett.* 202 (1986) 353–355.
- [30] L. H. Holley, M. Karplus, Protein secondary structure prediction with a neural network, *Proc. Natl. Acad. Sci. (USA)* 86 (1989) 152–156.
- [31] A. Perczel, K. Park, G. Fasman, Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel beta-sheet in proteins, *Proteins* 13 (1992) 57–69.
- [32] K. Asai, S. Hayamizu, K. Handa, Prediction of protein secondary structure by the hidden markov model, *CABIOS* 9 (1993) 141–146.
- [33] D. Frishman, P. Argos, Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence, *Protein Eng.* 9 (1996) 133–142.
- [34] J. Yang, Prediction of protein secondary structure from amino acid sequence, *J. Protein Chem.* 15 (1996) 185–191.
- [35] A. Galat, A note on circular-dichroic-constrained prediction of protein secondary structure, *Eur. J. Biochem.* 236 (1996) 428–435.
- [36] M. Asogawa, Beta-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network, *ISMB* 5 (1997) 48–51.

- [37] V. D. Francesco, P. McQueen, J. Garnier, P. Munson, Incorporating global information into secondary structure prediction with hidden markov models of protein folds, *ISMB* 5 (1997) 100–103.
- [38] M. Clementi, S. Clementi, G. Cruciani, M. Pastor, A. Davis, D. Flower, Robust multivariate statistics and the prediction of protein secondary structure content, *Protein Eng.* 10 (1997) 747–749.
- [39] M. Ito, Y. Matsuo, K. Nishikawa, Prediction of protein secondary structure using the 3D-1D compatibility algorithm, *CABIOS* 4 (1997) 415–424.
- [40] B. K. Ho, P. M. Curmi, Twist and shear in beta-sheets and beta-ribbons, *J. Mol. Biol.* 317 (2002) 291–308.
- [41] E. G. Eldon, R. Mukhopadhyay, C. Tang, N. S. Wingreen, Flexibility of beta-sheets: Principal-component analysis of database protein structures, preprint cond-mat/0309119 (2003).
- [42] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* 22 (1983) 2577–2637.
- [43] G. Wang, R. L. Dunbrack Jr, PISCES: a protein sequence culling server, *Bioinformatics* 19 (2003) 1589–1591.
- [44] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank, *Nucl. Acids Res.* 28 (2000) 235–242.
- [45] M. Garey, D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman & co., New York, 1979.
- [46] J. Hopcroft, J. Wong, Linear time algorithm for isomorphism of planar graphs, in: 6th Annual ACM Symposium on Theory of Computing, 1974, pp. 172–184.
- [47] J. Ullmann, An algorithm for subgraph isomorphism, *Journal of the ACM* 23 (1976) 31–42.
- [48] M. Dayhoff, R. Schwartz, B. Orcutt, A model of evolutionary change in proteins, in: M. Dayhoff (Ed.), *Atlas of Protein Science and Structure*, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Silver Spring, MD, 1978, pp. 345–352.
- [49] C. E. Shannon, W. Weaver, *The mathematical theory of communication*, University of Illinois Press, Urbana, IL, 1949.
- [50] S. Kullback, R. A. Leibler, On information and sufficiency, *Annals of Mathematical Statistics* 22 (1951) 79–86.
- [51] K. Fan, W. Wang, What is the minimum number of letters required to fold a protein?, *J. Mol. Biol.* 328 (2003) 921–926.
- [52] T. Li, K. Fan, J. Wang, W. Wang, Reduction of protein sequence complexity by residue grouping, *Protein Eng.* 16 (2003) 323–330.

- [53] J. Wang, W. Wang, Grouping of residues based on their contact interactions, *Phys. Rev. E* 65 (2002) 041911.
- [54] S. Miyazawa, R. L. Jernigan, Residue - residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, *J. Mol. Biol.* 256 (1996) 623–644.
- [55] C. Zhang, S. H. Kim, Environment-dependent residue contact energies for proteins, *Proc. Natl. Acad. Sci. (USA)* 97 (2000) 2550–2555.
- [56] S. Lifson, C. Sander, Antiparallel and parallel β -strands differ in amino acid residue preferences, *Nature* 282 (1979) 109–111.
- [57] S. Lifson, C. Sander, Specific recognition in the tertiary structure of beta-sheets of proteins, *J. Mol. Biol.* 139 (1980) 627–639.
- [58] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy, *Science* 302 (2003) 1364–1368.
- [59] T. M. Handel, P. J. Domaille, Heteronuclear (^1H , ^{13}C , ^{15}N) NMR assignments and solution structure of the monocyte chemoattractant protein-1 (MCP-1) dimer, *Biochemistry* 35 (1996) 6569–6584.
- [60] J. Lubkowski, G. Bujacz, L. Boque, P. J. Domaille, T. M. Handel, A. Wlodawer, The structure of MCP-1 in two crystal forms provides a rare example of variable quaternary interactions, *Nat. Struct. Biol.* 4 (1997) 64–69.
- [61] J. S. Richardson, D. C. Richardson, Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation, *Proc. Natl. Acad. Sci. (USA)* 99 (2002) 2754–2759.
- [62] A. W. Chan, E. G. Hutchinson, D. Harris, J. M. Thornton, Identification, classification, and analysis of beta-bulges in proteins, *Protein Sci.* 2 (1993) 1574–1590.
- [63] A. D. McLachlan, Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.* 128 (1979) 49–79.
- [64] J. F. Richardson, D. C. Richardson, Principles and patterns of protein conformation, Plenum Press, New York, 1989, Ch. 1, pp. 1–98.
- [65] R. A. Sayle, E. J. Milner-White, Rasmol: Biomolecular graphics for all, *Trends Biol. Sci.* 20 (1995) 374–376.

Legends to Tables

Table 1. Optimization of the M^0 matrix of the subgraph isomorphism problem between the two graphs depicted in Figure 2. S_l^k refers to the l^{th} strand of graph k .

(a) The M^0 matrix without optimization as computed by the original Ullman subgraph isomorphism algorithm.

(b) The M^0 matrix after the zero-out diagonal walk optimization.

Table 2. Truth table used in the subgraph isomorphism algorithm. e_{kl}^1 is an edge in G^1 between residues R_k^1 and R_l^1 , while $e_{V[k]V[l]}^2$ is an edge in G^2 between residues $R_{V[k]}^2$ and $R_{V[l]}^2$. The vector V is one of the mapping vectors generated by the algorithm such that the i^{th} entry in V , $V[i]$, is the residue in G^2 on which i of G^1 is mapped onto. The various topological relations **C**, **P**, **H** and **HP** as well as \emptyset are as defined in the text. Only true entries in the truth table are signaled; all other entries are to be taken as false.

Table 3. Subgraph isomorphic β -sheets of the monocyte chemoattractant protein 1 (MCP-1) [59,60]. The source β -sheet topological graph has been extracted from the PDB code 1DOK structure as annotated by the DSSP secondary structure assignment algorithm [42]. In the table, the PDB ID column refers to the PDB code of the isomorphic β -sheet. The S_1 column identifies which residues are part of the first strand mapping of the β -sheet, S_2 for the second strand and S_3 for the third strand. Residues are identified by their PDB residue number in addition to the optional chain identifier. The D_{SEQ} column is the sequence distance between the reference amino-acid sequence on the β -sheet of 1DOK with the isomorphic sibling, as calculated by the PAM250 substitution matrix [48], and higher values signify higher homology. The maximum D_{SEQ} value is 88 for the comparison of the β -sheet sequence of 1DOK with itself. The R_{SEQ} column is the sorted sequence distance rank. The D_{RMSD} column shows the RMSD distance between the crystal structure of the β -sheet of 1DOK compared to the isomorphic sibling. The RMSD is taken after strands S_1 of both structures have been aligned and is calculated from all heavy backbone atoms. The R_{RMSD} column is the sorted structure distance rank. The Molecule column is extracted from the PDB file under the ‘‘COMPND MOLECULE’’ keyword. The Header column is also taken from the PDB file but under the ‘‘HEADER’’ keyword.

Table 4. C+ class [62] β -bulge sequence analysis. This β -bulge is found in the monocyte chemoattractant protein 1 (MCP-1) [59,60]. There are 711 isomorphic siblings to this β -bulge. The ‘#’ column refers to the residue sequence number as found in PDB code 1DOK. The T column is the residue tags of the C+ β -bulge motif found in [62], Figure 2. The SE column is the

calculated Shannon [49] entropy for a given residue position. The KL column is the Kullback-Leibler distance [50] of the observed amino-acid distribution at a particular residue position compare to the reference observed amino-acid distribution found in β -sheets in general. The following columns are the 4-class amino-acid partitioning used for the sequence analysis, and comes from the works of [52], Table II. Amino-acids are mentioned by their 1-letter code. Classes are cysteine and aromatics {CFYW}, large hydrophobic {MLIV}, small {GPATS} and large polar or charged {NHQEDRK}. Each amino-acid class has an absolute occurrence count, N, as well as a relative occurrence count, %. The E row is the reference amino-acid distribution as found in β -sheets of the culled PDB Select 25 [43]. Entries which differ significantly from the reference distribution are underlined and boldfaced.

Legends to Figures

Figure 1. Hypothetical β -sheet topological graph. Strands are S_1 [95, 99], S_2 [108, 112] and S_3 [123, 127]. Strand S_1 is parallel to strand S_3 , while S_2 is anti-parallel to S_3 . Notice the difference in the H-bonding patterns between the parallel strand pair and the anti-parallel one. The various topological relations **C**, **P**, **H** and **HP** are displayed in the picture.

Figure 2. The two β -sheet topology graphs, G^1 and G^2 , used in Ullman’s subgraph isomorphism algorithm. In S_j^i , the superscript i refers to the graph number while the subscript j refers to the strand number within the graph.

(a) The graph G^1 which serves as the β -sheet topology to search for in G^2 . It is a two-stranded anti-parallel β -sheet with strands of length three. The partnership relation, **P**, between residues R_{112}^1 and R_{122}^1 has been omitted from the graph to show that it does not prevent a mapping in G^2 .

(b) The graph G^2 , a canonical three-stranded anti-parallel β -sheet with strands of length four. The two isomorphic solutions of G^1 in G^2 , I^1 and I^2 are highlighted in Grey boxes.

Figure 3. Observed β -sheet topology graph of monocyte chemoattractant protein 1 (MCP-1) (PDB code 1DOK) [59,60]. Topological features, namely covalent link **C**, β -sheet partnership **P** and H-bonding **H**, are calculated by the DSSP algorithm [42].

(a) The complete β -sheet topology graph of MCP-1.

(b) The β -sheet topology graph of the C+ type β -bulge [62] used for sequence analysis of the C+ motif, and found in MCP-1. Subscripts i to residue sequence number R_i refer to Thornton’s bulged residue nomenclature found in [62], Figure 2.

Figure 4. The 71 superimposed models of the β -sheet topological graph of monocyte chemoattractant protein 1 (MCP-1) (PDB code 1DOK) [59,60]. The input β -sheet topology graph is as Figure 3(a). These models are those of Table 3. The models have been superimposed on their first strand, which spawns residues 25 to 31. The two closest models have an RMSD of 0.39 Å, while the two farthest have 6.99 Å. Colors are blue for strand S_1 , green for strand S_2 and red for strand S_3 . The bulged residues, 26 and 27, are colored in yellow. The images were produced by RasMol [65].

(a) Side view.

(b) Bottom view.

(c) Front view.

(d) Top view.

Figure 5. β -barrel rings for β -strands with ($n=8, S=8$). In that configuration only two residues per strands are needed to express the ring in a topological graph. The dashed residues are used to show the connection between the

first and last strands.

(a) A β -ring which would result in right-handed crossover connections for them to pass outside the β -barrel, which are the prevalent connection types observed in solved protein structures [64].

(b) A β -ring which would result in left-handed crossover connections for them to pass outside the β -barrel. A right-handed connection would have to go by the inside of the barrel, which is impossible given the number of strands.

(a)(b)

Table 1

Table 2

Table 3

Table 4

Fig. 1.

(a)(b)

Fig. 2.

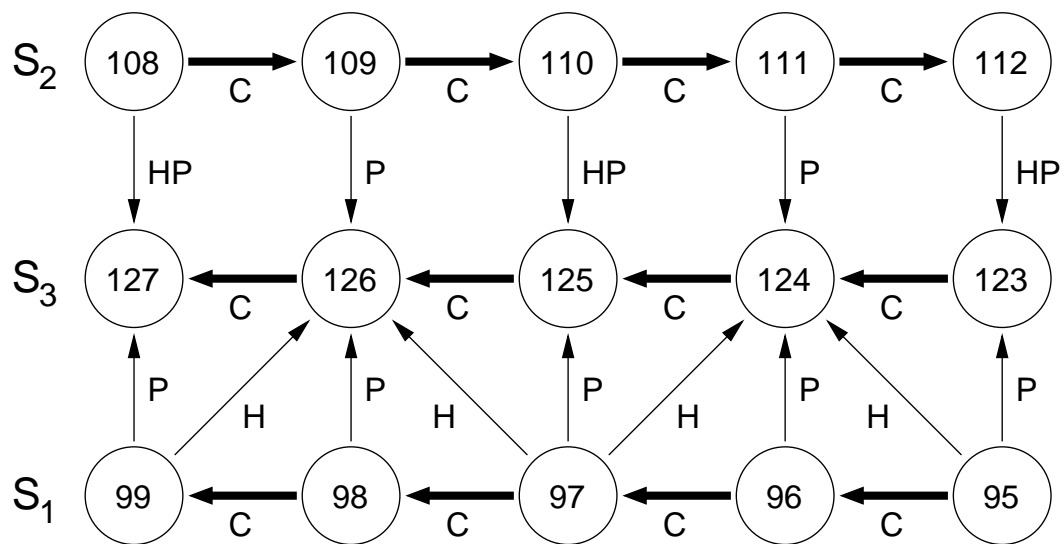
(a)(b)

Fig. 3.

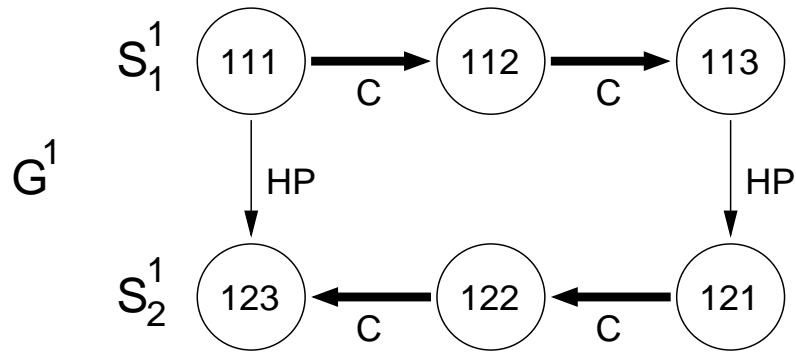
Fig. 4.

(a)(b)

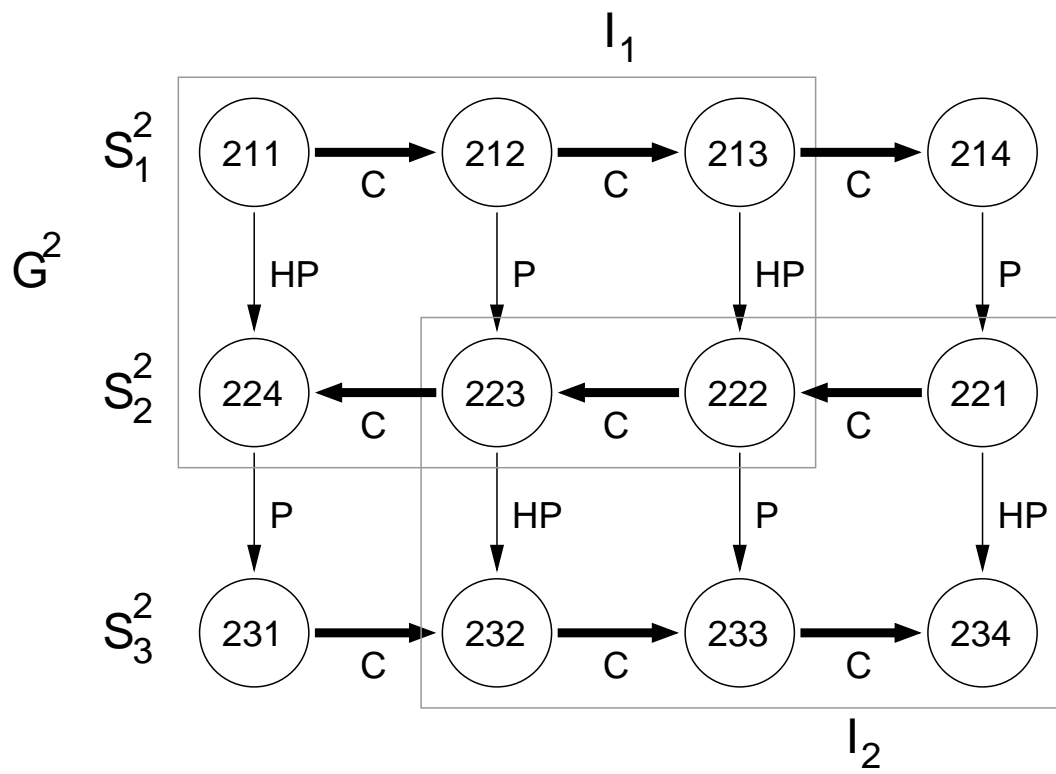
Fig. 5.



Parisien and Major, Figure 1.

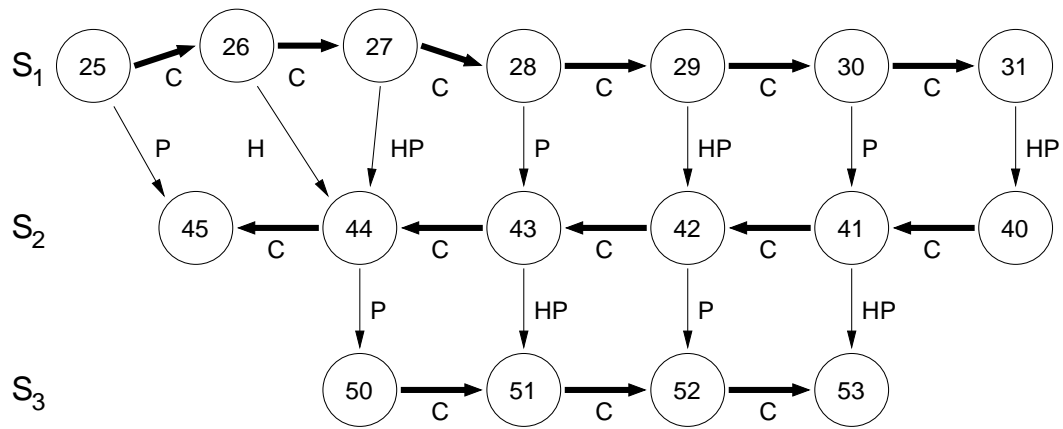


(a)

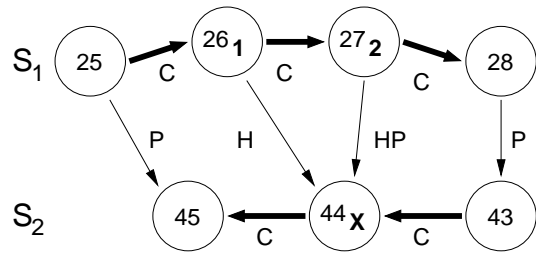


(b)

Parisien and Major, Figure 2.

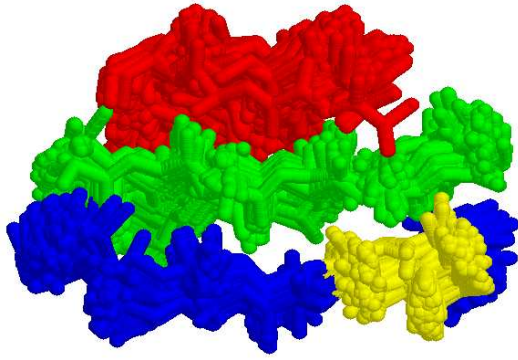


(a)

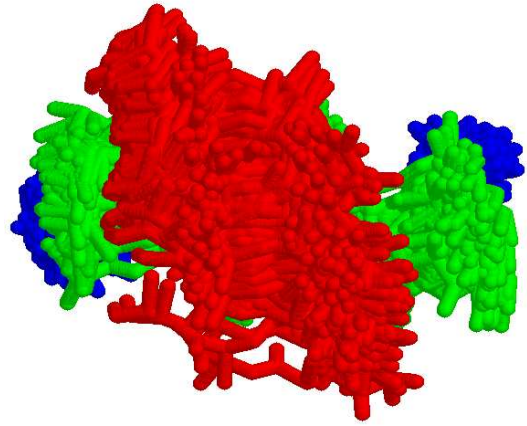


(b)

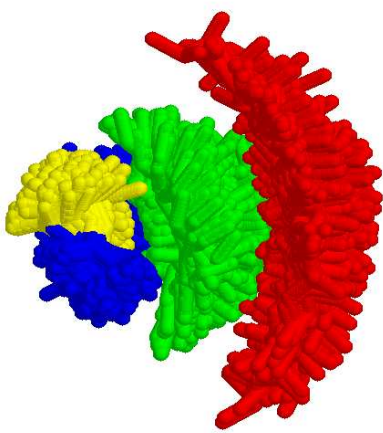
Parisien and Major, Figure 3.



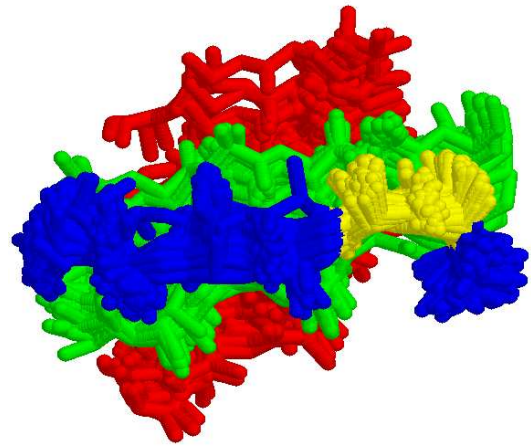
(a)



(b)

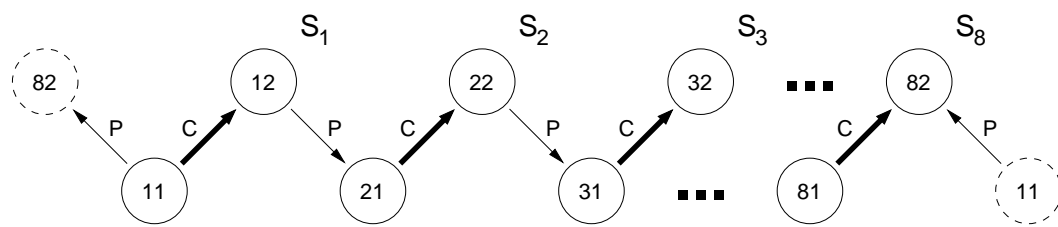


(c)

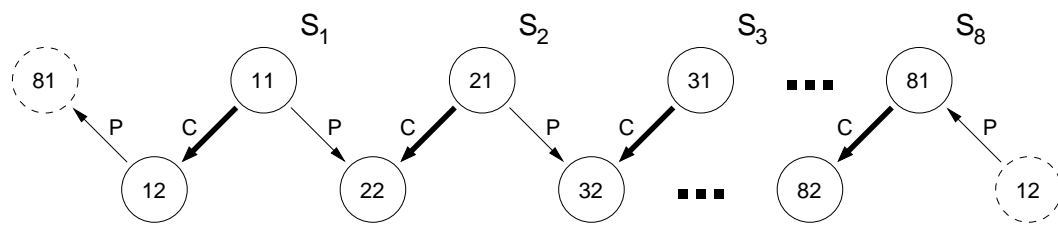


(d)

Parisien and Major, Figure 4.



(a)



(b)

Parisien and Major, Figure 5.

	S_1^2				S_2^2				S_3^2			
	1	0	1	0	1	0	1	0	0	0	0	0
S_1^1	0	1	1	0	0	1	1	0	0	1	1	0
	0	0	1	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1	0	0
S_2^1	0	1	1	0	0	1	1	0	0	1	1	0
	0	0	0	0	0	1	0	1	0	1	0	1

(a)

	S_1^2				S_2^2				S_3^2			
	1	0	0	0	1	0	0	0	0	0	0	0
S_1^1	0	1	0	0	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	1	0	0
S_2^1	0	0	0	0	0	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	1	0	0	0	1

(b)

Parisien and Major, Table 1.

		$e_{V[k]V[l]}^2$				
		C	H	P	HP	\emptyset
e_{kl}^1	C	T				
	H		T		T	
	P			T	T	
	HP				T	
	\emptyset		T	T	T	T

Parisien and Major, Table 2.

PDB ID	S ₁	S ₂	S ₃	D _{SEQ}	R _{SEQ}	D _{RMST}	R _{RMST}	Molecule	Header
1DOK	25A-31A	40A-45A	50A-53A	88	-	0.00	-	MONOCYTE CHEMOATTRACTANT PROTEIN 1	CHEMOATTRACTANT
1Q42	85A-91A	100A-105A	147A-150A	26	0	2.48	61	MRNA TRANSPORT REGULATOR MTR2	TRANSLATION
1TVX	39A-45A	55A-60A	65A-68A	24	1	0.80	4	NEUTROPHIL ACTIVATING PEPTIDE 2 VARIANT	CYTOKINE
1EC7	10A-16A	40A-45A	50A-53A	15	2	1.20	13	GLUCARATE DEHYDRATASE	LYASE
1EE8	44A-50A	53A-58A	63A-66A	14	3	2.24	57	MUTM (FPG) PROTEIN	DNA BINDING PROTEIN
1NH2	66D-72D	75D-80D	112D-115D	13	4	2.11	53	TRANSCRIPTION INITIATION FACTOR TFIID	TRANSCRIPTION/DNA
1C8C	20A-26A	29A-34A	43A-46A	13	5	0.89	5	DNA-BINDING PROTEIN 7A	DNA BINDING PROTEIN/DNA
2HFT	113-119	122-127	174-177	13	6	1.24	16	HUMAN TISSUE FACTOR	COAGULATION FACTOR
1QJ5	356A-362A	365A-370A	396A-399A	12	7	1.61	27	7,8-DIAMINOPELAGONIC ACID SYNTHASE	AMINOTRANSFERASE
1CEW	47I-53I	59I-64I	96I-99I	11	8	2.12	54	CYSTATIN	PROTEINASE INHIBITOR(CYSTEINE)
1JOV	98A-104A	107A-112A	126A-129A	6	9	2.96	64	H11317	UNKNOWN FUNCTION
1G4M	228A-234A	324A-329A	342A-345A	6	10	1.45	21	BETA-ARRESTIN1	SIGNALING PROTEIN
1GUI	39A-45A	48A-53A	146A-149A	5	11	2.05	45	LAMINARINASE 16A	CARBOHYDRATE BINDING MODULE
1JZ8	222A-228A	241A-246A	288A-291A	5	12	2.14	55	BETA-GALACTOSIDASE	HYDROLASE
1RVK	3A-9A	37A-42A	47A-50A	4	13	2.07	48	ISOMERASE/LACTONIZING ENZYME	UNKNOWN FUNCTION
1LSH	801A-807A	819A-824A	869A-872A	3	14	4.01	70	LIPOVITELLIN (LV-1N, LV-1C)	LIPID BINDING PROTEIN
1OBF	172O-178O	245O-250O	307O-310O	3	15	1.78	36	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE	GLYCOLYTIC PATHWAY
1MDL	6-12	36-41	46-49	3	16	1.07	8	MANDELATE RACEMASE	ISOMERASE
1MUC	6A-12A	36A-41A	46A-49A	3	17	1.22	15	MUCONATE LACTONIZING ENZYME	ISOMERASE
1LUG	90A-96A	117A-122A	142A-145A	2	18	2.67	63	CARBONIC ANHYDRASE II	LYASE
1B8A	25A-31A	34A-39A	47A-50A	2	19	3.14	65	ASPARTYL-TRNA SYNTHETASE	LIGASE
1E42	875A-881A	884A-889A	902A-905A	2	20	1.14	11	BETA2-ADAPTIN	ENDOCYTOSIS
1NJH	54A-60A	66A-71A	75A-78A	2	21	1.65	30	PROTEIN YOJF	UNKNOWN FUNCTION
1MB4	152A-158A	277A-282A	341A-344A	1	22	1.96	41	ASPARTATE-SEMIALDEHYDE DEHYDROGENASE	OXIDOREDUCTASE
1ES5	209A-215A	223A-228A	236A-239A	1	23	3.98	69	DD-TRANSEPTIDASE	HYDROLASE
1DZK	39A-45A	50A-55A	67A-70A	-1	24	1.62	28	ODORANT-BINDING PROTEIN	ODORANT BINDING PROTEIN
1N9E	377A-383A	529A-534A	689A-692A	-1	25	2.03	44	LYSYL OXIDASE	OXIDOREDUCTASE
1BKB	32-38	47-52	62-65	-1	26	1.37	19	TRANSLATION INITIATION FACTOR 5A	TRANSLATION
1K3X	44A-50A	53A-58A	63A-66A	-1	27	2.09	51	ENDONUCLEASE VIII	HYDROLASE/DNA
1JY1	196A-202A	266A-271A	275A-278A	-2	28	2.17	56	TYROSYL-DNA PHOSPHODIESTERASE	HYDROLASE
1NVM	150B-156B	210B-215B	270B-273B	-3	29	1.87	38	4-HYDROXY-2-OXOVALERATE ALDOLASE	LYASE/OXIDOREDUCTASE
1PJ1	343A-349A	415A-420A	429A-432A	-3	30	2.35	60	INOSITOL-3-PHOSPHATE SYNTHASE	ISOMERASE
1J5P	123A-129A	194A-199A	203A-206A	-3	31	1.66	31	HYPOTHETICAL PROTEIN TM1643	UNKNOWN FUNCTION
1WHO	28-34	65-70	75-78	-3	32	2.03	43	ALLERGEN PHL P 2	ALLERGEN
1GY7	66A-72A	81A-86A	100A-103A	-4	33	2.07	49	NUCLEAR TRANSPORT FACTOR 2	NUCLEAR TRANSPORT
1UMW	411A-417A	422A-427A	432A-435A	-4	34	1.73	33	SERINE/THREONINE-PROTEIN KINASE PLK	KINASE
1IJQ	408A-414A	419A-424A	430A-433A	-4	35	1.28	17	LOW-DENSITY LIPOPROTEIN RECEPTOR	LIPID TRANSPORT
1JC4	35A-41A	46A-51A	65A-68A	-5	36	1.45	22	METHYLMALONYL-COA EPIMERASE	ISOMERASE
1F2T	3A-9A	70A-75A	83A-86A	-5	37	3.15	66	RAD50 ABC-ATPASE	REPLICATION
1IWL	124A-130A	139A-144A	149A-152A	-6	38	3.45	68	OUTER-MEMBRANE LIPOPROTEINS CARRIER	PROTEIN TRANSPORT
1DPG	207A-213A	346A-351A	368A-371A	-6	39	1.50	24	GLUCOSE 6-PHOSPHATE DEHYDROGENASE	OXIDOREDUCTASE (CHOH(D) - NAD(P))
1OLZ	190A-196A	209A-214A	232A-235A	-7	40	1.93	40	SEMAPHORIN 4D	DEVELOPMENTAL PROTEIN
1NEP	15A-21A	37A-42A	89A-92A	-7	41	2.09	52	EPIDIDYMAL SECRETORY PROTEIN E1	LIPID BINDING PROTEIN
1JET	61A-67A	71A-76A	143A-146A	-8	42	1.01	7	OLIGO-PEPTIDE BINDING PROTEIN	PEPTIDE TRANSPORT
1D7U	351A-357A	360A-365A	404A-407A	-8	43	1.66	32	2,2-DIALKYLGLYCINE DECARBOXYLASE (PYRUVATE)	LYASE
1N9E	95A-101A	122A-127A	139A-142A	-9	44	1.87	37	LYSYL OXIDASE	OXIDOREDUCTASE
1HS6	18A-24A	37A-42A	101A-104A	-9	45	1.59	26	LEUKOTRIENE A-4 HYDROLASE	HYDROLASE
1MW9	268X-274X	410X-415X	423X-426X	-9	46	1.77	35	DNA TOPOISOMERASE I	ISOMERASE
1K8K	55C-61C	66C-71C	76C-79C	-9	47	1.31	18	ACTIN-LIKE PROTEIN 3	STRUCTURAL PROTEIN
1EYQ	30A-36A	48A-53A	93A-96A	-11	48	2.02	42	CHALCONE-FLAVONONE ISOMERASE 1	ISOMERASE
1NH2	248C-254C	257C-262C	280C-283C	-11	49	1.65	29	TRANSCRIPTION INITIATION FACTOR TFIID	TRANSCRIPTION/DNA
1OLZ	399A-405A	417A-422A	427A-430A	-11	50	0.67	0	SEMAPHORIN 4D	DEVELOPMENTAL PROTEIN
1F9Z	28A-34A	39A-44A	56A-59A	-11	51	1.41	20	GLYOXALASE I	LYASE
7AHL	81A-87A	248A-253A	275A-278A	-12	52	1.46	23	ALPHA-HEMOLYSIN	CYTOLYTIC PROTEIN
1KAF	128A-134A	137A-142A	151A-154A	-12	53	3.31	67	TRANSCRIPTION REGULATORY PROTEIN MOTA	TRANSCRIPTION
1Q7F	1004A-1010A	1014A-1019A	1024A-1027A	-12	54	0.77	2	BRAIN TUMOR CG10719-PA	TRANSLATION
1JZ8	822A-828A	835A-840A	853A-856A	-13	55	2.32	59	BETA-GALACTOSIDASE	HYDROLASE
1JKG	72A-78A	92A-97A	110A-113A	-13	56	1.91	39	P15	TRANSPORT PROTEIN
1KKO	3A-9A	54A-59A	64A-67A	-13	57	0.97	6	3-METHYLASPARTATE AMMONIA-LYASE	LYASE
1KSH	17B-23B	60B-65B	104B-107B	-15	58	1.56	25	ARF-LIKE PROTEIN 2	SIGNALING PROTEIN/HYDROLASE
1H2W	338A-344A	348A-353A	361A-364A	-15	59	0.78	3	PROLYL ENDOPEPTIDASE	HYDROLASE
1R17	285A-291A	310A-315A	386A-389A	-16	60	2.06	46	FIBRINOGEN-BINDING PROTEIN SDRG	CELL ADHESION
1I8D	109A-115A	118A-123A	158A-161A	-17	61	2.06	47	RIBOFLAVIN SYNTHASE	TRANSFERASE
1N2E	247A-253A	267A-272A	280A-283A	-17	62	2.08	50	PANTOTHENATE SYNTHETASE	LIGASE
1FW9	49A-55A	73A-78A	91A-94A	-18	63	2.50	62	CHORISMATE LYASE	LYASE
1OXX	301K-307K	310K-315K	325K-328K	-18	64	1.09	9	ABC TRANSPORTER, ATP BINDING PROTEIN	TRANSPORT PROTEIN
1KCM	42A-48A	56A-61A	87A-90A	-19	65	2.32	58	PHOSPHATIDYLINOSITOL TRANSFER PROTEIN ALPHA	LIPID BINDING PROTEIN
1FW9	126A-132A	139A-144A	152A-155A	-21	66	1.73	34	CHORISMATE LYASE	LYASE
1OLZ	272A-278A	287A-292A	304A-307A	-22	67	0.71	1	SEMAPHORIN 4D	DEVELOPMENTAL PROTEIN
1NWW	85A-91A	94A-99A	117A-120A	-25	68	1.16	12	LIMONENE-1,2-EPOXIDE HYDROLASE	HYDROLASE
1A12	242A-248A	251A-256A	261A-264A	-27	69	1.10	10	REGULATOR OF CHROMOSOME CONDENSATION 1	NUCLEOTIDE EXCHANGE FACTOR
1DQI	49A-55A	106A-111A	116A-119A	-32	70	1.21	14	SUPEROXIDE REDUCTASE	OXIDOREDUCTASE

Parisien and Major, Table 3.

#	T	SE	KL	CFYW		MLIV		GPATS		NHQEDRK	
				N	%	N	%	N	%	N	%
E		1.34e+00		13253	14	33231	36	23431	25	22993	25
25		9.63e-01	2.50e-01	61	9	482	<u>68</u>	114	16	54	<u>8</u>
26	1	1.21e+00	5.62e-02	45	<u>6</u>	331	47	145	20	190	27
27	2	1.19e+00	3.50e-01	90	13	52	<u>7</u>	277	<u>39</u>	292	<u>41</u>
28		1.28e+00	1.48e-02	82	12	313	44	164	23	152	21
43		1.26e+00	5.35e-02	118	17	331	47	166	23	96	<u>14</u>
44	X	1.29e+00	1.10e-02	79	11	299	42	158	22	175	25
45		1.29e+00	1.20e-01	188	<u>26</u>	272	38	184	26	67	<u>9</u>

Parisien and Major, Table 4.