Search for a new description of protein topology and local structure

Lukasz Jaroszewski & Adam Godzik

The Burnham Institute
10901 N.Torrey Pines Road, La Jolla, CA 92037
phone (858) 646 3168, e.mail {lukasz,adam}@burnham-inst.org

Abstract

A novel description of protein structure in terms of the generalized secondary structure elements (GSSE) is proposed. GSSE's are defined as fragments of the protein structure where the chain doesn't radically change its direction. In this new language, global protein topology becomes a particular arrangement of the relatively small number of large, rod like GSSE's. Protein topology can be described by an adjacency matrix giving information, which GSSE's are close in space to each other and defining a graph, where GSSE's are equivalent to vertices and interactions between them to edges. The information about the local structure is translated into the local density of pseudo-Cα atoms along the chain and the curvature of the chain. This new description has a number of interesting and useful features. For instance, enumeration theorems of graph theory can be used to estimate a number of possible topologies for a protein built from a given number of elements. Different topologies, including novel ones, can be generated from the known by various permutations of elements. Many new regularities in protein structures become suddenly visible in a new description. A new local structure description is more amenable to predictions and easier to use in fold predictions.

Introduction

The single most important decision in protein structure analysis is a choice of language used for protein structure description. Such choices define what we see as similar or different, and even on how we calculate energy. Helices or beta strands, usually used to depict protein structures, are based on local structural regularities, but at the same time they make a protein structure easier to analyze by defining a relatively small number of structural blocks. Other choices are focusing on interaction maps of proteins in form of contact maps (Godzik, Skolnick and Kolinski 1993), or their surfaces. Beautiful cartoon diagrams of proteins based on local secondary structure introduced by J.Richardson are by far the most popular way of depicting protein structures, to a very large extent shaping the way we think about proteins.

For instance, based on this representation, a complete classification of all known protein structures was attempted by several groups, and constantly updated lists of known protein folds are available over the Internet (CATH 1995, SCOP 1995). At the same time, defining blocks forming the global protein topology on the basis of local regularity and effectively mixing up two perspectives, one local, one global, has many drawbacks. For instance, various definitions of secondary structure, as well as local irregularities may lead to assigning various positions (or even various numbers) of secondary structure elements to otherwise very similar protein structures. Also, there are many proteins with only small amounts of secondary structure, that in cartoon diagrams come across as irregular, almost random structures.

In this contribution we attempt to solve these problems by completely separating the two perspectives (local and global). A new description of a global protein topology will be based only on the direction of the chain, and a local description only on the density of packing of amino acids close in sequence. Details of this approach will be explained in the Methods section of the paper, with several applications of the new description presented in the Results section, with even more suggested in the Discussion and Conclusions. In short, complete separation of the local and global structural information allows the global regularities to be easier to find and classify and local regularities easier to predict. For instance, it becomes obvious that constraints of packing together a relatively small number of rod-like objects play a fundamental role in limiting the number of possible protein topologies.

Methods

The crucial question which must be addressed before one attempts to build a simplified description of protein structures is a definition of an elementary building block used in the description, which preferably should be as large as possible. A continuous fragment of regular local secondary structure (SSE) seems to be an obvious choice here and indeed, it is used as a basis of most descriptions. There is however an important problem in using fragments of secondary structure, defined on the basis of the local regularities in secondary structure in a global fold description. There are many examples of locally regular structures, separated by a short elements of irregular structure, at least as classified their hydrogen bonded pattern (Kabsch and Sander 1983), but without any changes general direction. Should they count as one, or as two elements? Another potential problem is presented by stretches of irregular structure. These fragment, while devoid of any local regularity, often “connect” one side of the molecule with the other side, and, as such, in a global topology structure have a role similar to other regular α helices. Using secondary structure assignments, such as produced by the DSSP program (Kabsch and Sander 1983) about 50% of all protein residues are classified in states other than helix or extended. For some proteins, this percentage is much lower and their description in a standard ribbon diagram is difficult. These problems with the ambiguity of the local regularities as used in defining elements of protein structure can be solved by focusing on a global, instead of a local description of a protein chain.

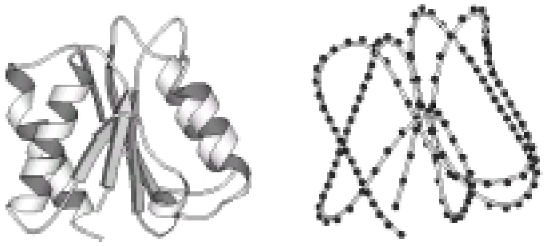

Figure 1. The ribbon diagram (left) and a smoothed backbone of flavodoxin (PDB code 1fla).

The process of averaging out the local characteristics of the protein chain is illustrated in Figure 1, where the backbone is “smoothed” out by using an averaging procedure, where a Cα position is replaced by an average calculated in a 5-residue window. After the smoothing the protein chain looses all its local “wiggles”, all that remains is the information about the global topology of a chain. We would call such a chain a “smoothed backbone” and the positions of Cα atoms after the smoothing procedure will be called pseudo-Cα atoms.

Now, a natural definition of a building block would be a fragment of a protein chain that is approximately straight, i.e. continues from one turn, where the protein chain changes its direction, to another. We would call such elements “generalized secondary structure elements” or GSSE. This definition includes all standard SSE’s, and at the same time solves many ambiguities: one GSSE might be built from several SSE’s (such as the first two helices in endonuclease) or from fragments devoid of any regular secondary structure (such as the first, third and a sixth element in endonuclease).

We propose a simple definition of a GSSE based on the straightforward application of the previous “naive” definition. We define an angular correlation matrix, with its [i,j] element equal to the angle between vectors [$C\alpha_i$,$C\alpha_{i+1}$] and [$C\alpha_j$, $C\alpha_{j+1}$]. An example of such a matrix is presented graphically in Figure 2 for myoglobin.

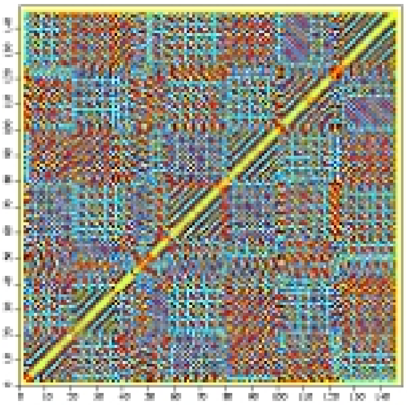

Figure 2a. The “angle map” of myoglobin : angles between two consecutive Cα atoms are shown in colors, from black (parallel) to red (anti-parallel).

In Figures 2a and 2b, the angle between two Cα (or pseudo-Cα) atoms is denoted by color, from black for 180 deg (parallel), to blue for 90 deg (perpendicular) to red for 0 deg (anti-parallel). For the native backbone structure (Figure 2a), the angular correlation matrix doesn’t reveal much regularity. But using a “smoothed” protein backbone (lower part of the Figure 2), such as illustrated in the Figure 1, presents a very different picture with very clear regularity, visible as well defined black boxes of angles close to 180 degrees along the diagonal.

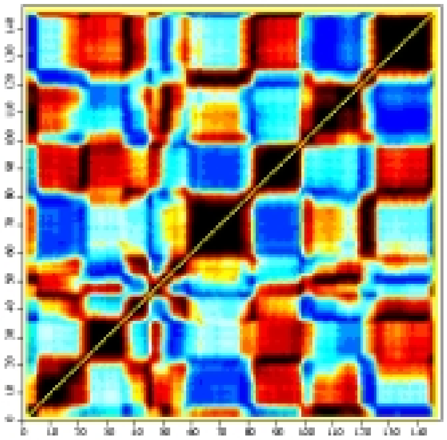

Figure 2b. The “angle map” of a smoothed backbone of myoglobin (see the text for details): angles between two consecutive pseudo-Cα atoms are shown in colors, from black (parallel) to red (anti-parallel).

These boxes are used to define GSSE boundaries, and this definition is surprisingly robust in respect to the number of smoothing procedures and threshold values for the segment definition. Three smoothing cycles and a threshold of 120 degrees for the lowest vector-vector angle within the segment were chosen arbitrarily and used throughout this paper. This is equivalent to a “smoothed” protein backbone, such as shown in Figure 1. A similar definition of a structural segment was used previously in a “block and U-turn” model of protein structure (Kolinski et al. 1997).

As seen in Figure 1, the pseudo-Cα chain obtained with a smoothing procedure described above can have different linear density of pseudo-Cα atoms along the chain. Helical segments correspond to a high linear density of about 0.7 pseudo-Cα atom per angstrom (1.5 Å between pseudo-Cα atoms), segments with extended (β) local conformation have a local density of about 0.3 (i.e. 3Å between atoms). Interestingly there are a lot of dense fragments that are not classified as helical by standard secondary structure classification programs. The links between GSSE can be characterized both by the linear density of pseudo-Cα atoms along the chain and by an angle between consecutive pseudo-Cα atoms.

Results:

Global description: protein topology as a graph

To complete a construction of a simplified description of protein topology, we now have to characterize a mutual orientation of segments in space. One of several possibilities is to provide a list of segments adjacent in space, another, a list of local angles describing how the chain turns. Within the former definition, there is a wide choice of possible definitions of “adjacency”. We have decided to use the definition based on distance in space between central fragments of segments, identified according to the procedure above. This is cross-correlated with the information of inter-residue interactions between segments.

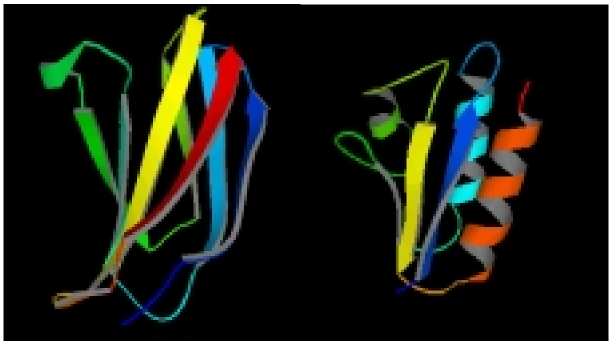

Figure 3. Ribbon diagrams of tenascin (1ten) and phosphotransferase (1poh) – proteins used in examples in the text.

Both steps are combined to create a completely automated procedure of assigning segments and deciding about the information about the segments proximity in space. This procedure provides a description in the form of an adjacency matrix, such as illustrated in Table I for examples of phosphotransferase (PDB code of 1pof) and tenascin (PDB code of 1ten). The same proteins will be also used in several examples later in the paper, their structures are shown in Figure 3 as ribbon diagrams and their smoothed backbones are shown in Figure 4.

Table I

						-
					-	1
				-	1	0
			-	1	0	0
		-	1	1	1	1
	-	1	1	1	0	1
-	1	0	0	0	1	1

						-
					-	1
				-	1	0
			-	1	0	0
		-	1	1	1	0
	-	1	0	1	1	1
-	1	1	0	0	1	1

Table I. Phosphotransferase (1pof) and tenascin (1ten) adjacency tables

The protein segments adjacency table resembles an adjacency table of a graph (Wilson and Watkins 1990), and can be treated as such. Following this analogy segments become equivalent to vertices of a graph and interactions between segments as edges of a graph. Because protein segments have a specified order, this is a labeled graph, with consecutive numbers of segments as vertices labels.

The idea of representing a protein structure as a graph is not new, and was used for identification of tertiary similarities between proteins (Mitchell et al. 1990; Grindley et al. 1993) and automated identification of certain motifs (Arteca and Metzey 1990; Koch, Kaden, and Selbig 1992; Flower 1994). However here we use a slightly different definition of the vertex (GSSE) and of the edge (interaction between segments), but more importantly we are going to use the graph representation of protein structures for much more general purposes.

The adjacency matrix, as shown in Table I is a two dimensional matrix, but it could be described as a vector, build from adjacency matrix cells according to the numbering scheme from Table II. Because the adjacency matrix is symmetric, this vector is completely equivalent to the whole matrix. On the other hand, this vector can be treated as a binary number (the index), which in turn can be represented as a decimal number. Thus the tenascin topology as described by its adjacency matrix can be represented as a binary number 110011101111110100101. This represents a unique (for a given segment and segment interaction definitions) representation of an arbitrary protein topology by a numerical index.

Table II

						-
					-	21
				-	19	20
			-	16	17	18
		-	12	13	14	15
	-	7	8	9	10	11
-	1	2	3	4	5	6

Table II. Cell numbering

Table III

						-
					-	1
				-	1	0
			-	1	0	0
		-	1	1	1	0
	-	1	1	1	0	1
-	1	1	0	0	1	1

Table III. Rearranged 1poh adjacency for the index calculation matrix

Now imagine a process, where we reconnect existing segments in a new order, by cutting some loops and introducing new ones, at the same time keeping constant the mutual arrangements of segments. For a labeled graph this is called a permutation of vertices, and on the level of the adjacency matrix it is equivalent to a simultaneous exchange of some columns and rows. For instance, by exchanging the first and the seventh row (and appropriate columns) the adjacency matrix of phosphotransferase from Table I becomes equivalent to that of Table III.

At the same time, the index of this new topology would change as well: from 100011111011111100101 it would become equal to 110011111011110100101. Coincidentally, permuted index of phosphotransferase would become very similar (with two differences) to that of tenascin. Indeed, analysis of the mutual positions of GSSE's 2-6 in both proteins reveals that both have a Greek key motif, even that one is a α/β protein, and the other an all beta protein. This similarity is illustrated below in Figure 4, where smoothed chains of corresponding fragments from both molecules are displayed. This similarity is almost impossible to notice in Figure 3, where both proteins are displayed as ribbon diagrams.

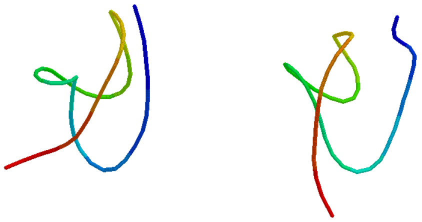

Figure 4. The structures of tenascin (1ten) and phosphotransferase (1poh), shown here as smoothed backbones (only equivalent elements are shown). The ribbon diagrams for both proteins were shown in Figure 3. Note the similarity between positions of GSSE's in both proteins.

Each of all possible permutations of vertices of a graph would have a different index. The one with the largest (in a numerical sense) index is called the canonical graph. It can be used to represent the group of graphs, which can be changed into each other by changing labels for various segments. In graph theory such graphs are called isomorphous and on the level of protein topology it represents identical arrangements of segments having different topology. Such an arrangement, which we will call a scaffold, is equivalent to an unlabeled graph.

At this point our analogy between protein topologies and graphs is complete. Every protein topology can be represented as a labeled graph, identified by its index. Groups of proteins sharing the same scaffold, i.e. proteins that differ only by different connectivity of segments can be represented as unlabelled graphs. Such groups can be identified by calculating canonical indices for each topology; topologies with the same canonical index are build on the same scaffold.

Treating a protein topology as a graph has many interesting consequences. Among them is the possibility of enumerating all existing labeled and unlabelled graphs with a given number of vertices (i.e. segments) and edges (i.e. interaction regions) (Harary and Palmer 1973).

Table IV

	4	5	6	7	8	9	10	11	12
4	2	1	1						
5	3	5	5	4	2	1	1		
6		6	13	19	22	20	14	9	5
7			11	33	67	107	132	138	126
8					23	89	236	486	1169

Table IV. Enumeration of the number of possible unlabelled graphs with x number of interaction regions and y number of segments

A complete enumeration for graph with 8 vertices and 10 edges lists over 700,000 different graphs divided into 236 significantly different (i.e. non-isomorphic) arrangements of linkers, which could be represented for instance by their canonical representation. Total number of graphs having up to 10 vertices is well over $3*10^7$.

The number of possible protein topologies is obviously much smaller. For instance, while we can build a graph for every protein topology, the reverse is not always true. There are also many empirical rules followed by protein topologies. For instance a single vertex (GSSE) must have at least 2 and no more than five edges (interactions). Accounting for these two factors lowers an upper estimate for possible protein topologies at around $3*10^5$.

This number is still significantly larger than estimates based on extrapolations of new folds appearing in newly determined protein structures. One factor not included in the analysis above, is that many structural families could be represented by graphs sharing a large common subgraph, but classified here as being different. Additional empirical rules, which if implemented could lower the estimate even further are currently under development.

Results

Local structure as a density of pseudo-Cα atoms

The type of secondary structure in GSSE can be identified by a density of pseudo-Cα atoms, or its inverse, a distance between consecutive pseudo-Cα atoms. In this description, helices correspond to short distances and strands to larger distances. Many irregular structures, that in the standard description are put together into a huge “unclassified” category, can be now differentiated and put into “helical-like” or extended-like category. This addresses an important problem in protein local structure description. The standard secondary structure definitions used to identify elements such as alpha helices or beta strands are based on arbitrary thresholds. After smoothing, these thresholds are almost eliminated, with distances falling smoothly into two distinct categories. Because of that, the distances between pseudo-Cα atoms in a smoothed trace are easier to predict than secondary structure. Figure 6 presents a correlation between a predicted and real secondary structure for a factor H protein (1hfi), as classified in a DSSP three state model (H,E,C) used in most secondary structure prediction programs. Prediction was performed with a nearest neighbor method developed in our group (Rychlewski and Godzik 1997) and the PHD algorithm (Rost and Sander 1993). Both methods achieve secondary structure prediction accuracy of over 70% in the three state prediction. On the latest CASP3 prediction contest our nearest neighbor method achieved accuracy of 76% on a large set of over 20 prediction targets (Orengo 1999). In this particular case, it is clear that both methods completely miss the series of beta strands at the N-terminal of the molecule

```
Seq EKIPCSQPPQIEHGTINSSRSSQESYAHGTKLSYTCEGGFRISEENETTCYMGKWSSPPQCE
sec ____EE____E__EEE________EE____EEEEE______E____EEEEE__EE_____E_
PHD _EE______________________EE___EEEEEE___EEEE___EEEEEE__________
NN  _______________________________EEEEE____EE___E_EEE____________
```

Figure 5a. Secondary structure in a DSSP classification for factor H protein (PDB code 1hfi) compared to two predictions (DSSP and NN methods).

It is difficult to compare secondary structure predictions if the prediction is done for a different type of secondary structure description. Also program such as neural network based PHD program would have to be retrained for a different description. However, the nearest neighbor method is based on a simple averaging procedure among fragments chosen for their sequence similarity to the prediction target (Rychlewski and Godzik 1997). Therefore it is simple to change the local structure description – one has to change in the database of the known protein structures and the modified structure description for the prediction target would automatically follow. We have prepared database of GSSE descriptions for all proteins from the PDB and applied the NN algorithm, exactly like described in (Rychlewski and Godzik 1997) and used to obtain data in Figure 5b. Figure 5b presents the comparison of predicted and real distances between pseudo-Cα atoms for a simplified model of 1hfi. It is interesting to compare the prediction in Figure 5b to the prediction in the lower line of Figure 5a. Both predictions being made using the same nearest neighbor method algorithm (only different properties were being averaged among neighboring fragments). Note a completely wrong discrete secondary structure prediction in the N-terminal half of the molecule.

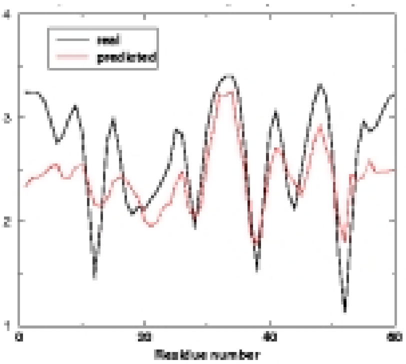

Figure 5b. An example of a prediction of distances between pseudo-Ca atoms for factor H protein (PDB code 1hfi). The predicted structure (red line) is compared to the native secondary structure (solid black line).

At the same time, the continuous description prediction correctly captured the pattern of beta/turn preferences, even if it exaggerated the level of departure from an ideal extended structure.

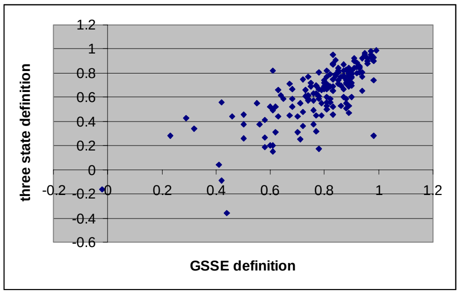

Figure 6. Comparison of accuracy of secondary structure predictions in the three state model (y-axis) and the pseudo-Cα distances model (x-axis).

Using the same nearest neighbor secondary structure prediction algorithm, the correlation for predicted and real pseudo-Cα atom distances is 0.8, as compared to 0.64 for H,E,C prediction over a large database of proteins.

This systematic difference is illustrated in Figure 6, which shows the correlation between the real and predicted pseudo-Cα atom distances vs. the same correlation for the standard three state (H,E,C) prediction. Almost all data points lay below the diagonal, i.e. suggesting that GSSE definition of secondary structure can be predicted more accurately for almost all proteins in the database.

At the same time the pseudo-Cα distances are much better description of a local structure. When both prediction target and the templates are described by their correct secondary structure in the H,E,C representation, only half of proteins in the benchmark (the UCLA 68 residue benchmark) can be correctly identified by their secondary structure alone (data not shown). This number increases by about 50% when the pseudo-Cα distances are used in a similar experiment. The important point here is that even exactly the same information (local protein structure) is being processed, the predictions are easier to make and the information is more useful, when alternative description of the local structure is being used. Application of such descriptions to protein fold recognition remains to be tested.

Discussion and Conclusions

A novel description of protein structure was introduced, based on a naïve picture of identifying fragments of protein chain where the chain doesn't change its direction. The new description allows separating the details of the local structure from the information about a global topology of the protein chain. With such separation it was possible to focus on new regularities on the global level, such as unexpected structural similarities between protein of different structural classes. At the same time, the new description of the local structure in terms of average distance between pseudo-Cα atoms after the smoothing procedure is much more amenable to local structure prediction and carries more information as judged by the fold prediction accuracy.

This new type of description remains to be tested in other applications, which could include:

1. Automated classification of protein structures
2. Very fast determination of structural similarities.
3. Conformational searches by complete enumeration of possible topologies. For smaller number of linkers it is possible to limit the search to 100-200 conformations.
4. Fold predictions based on the library of all possible topologies.
5. Fold predictions based on the new definition of secondary structure

The new definition of secondary structure has been recently tested in the context of the analysis of structure-structure alignments (Shindyalov and Bourne 1998). With the new definition, about 75% of an average protein can be classified as belonging to GSSEs, in contrast to only about 50% that can be assigned to secondary structure elements, as defined by a local secondary structure. Despite that, about 80% of all gaps in the alignments still fall between GSSEs, which is close to the similar percentage for standard secondary structure elements. This result strongly suggests that the locally irregular structural segments included in the GSSE definitions have many features previously associated with secondary structure elements.

Tests in specific applications such as mentioned above would eventually prove if this particular language for the description of protein structure is more useful that the traditional one. However, it is clear that to achieve progress in fields such as fold prediction and modeling, new ideas about how to describe a protein structure are necessary.

Acknowledgments

This research was supported by NIH grant GM60049.

References

Arteca, G. A., and Metzey P.G. 1990. A method for the characterization of foldings in protein ribbon models. *J. Mol. Graph*. 8:66-80.

CATH 1995. Protein Structure Classification. London, UC BSM. wwww.biochem.ucl.ac.uk/bsm/cath

Flower, D. R. 1994. Automating the identification and analysis of protein beta-barrels. *Prot. Engineering*. 7:1305-1310.

Godzik, A.; Skolnick, J.; and Kolinski, A. 1993. Regularities in interaction patterns of globular proteins. *Prot. Engineering* 6:801-810.

Grindley, H. M., et al. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol*. 229:707-721.

Harary, F., and Palmer E. M. 1973. Graphical Enumeration. New York, Academic Press.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

Koch, I., F.; Kaden F.; and Selbig J. 1992. Analysis of protein sheet topologies by graph-theoretical techniques. *Proteins* 12:314-323.

Kolinski, A., et al. 1997. A method for the prediction of surface U-turns and transglobular connections in small proteins. *Proteins* 27:290-308.

Mitchell, E. M., et al. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J.Mol.Biol*. 212:151-166.

Orengo, C. A.et al. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure and contacts prediction. *Proteins* 37 (S3):149-170.

Rost, B. and Sander C. 1993. Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol*. 232:584-599.

Rychlewski, L., and Godzik, A. 1997. Secondary structure prediction using segment similarity. *Prot. Engineering*. 10:1143-1153.

Shindyalov, I. N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Prot. Engineering* Sep;11:739-47

SCOP (1995). Structural classification of proteins. Oxford, MRC Cambridge. scop.mrc-lmb.cam.ac.uk/scop/

Wilson, R. J., and Watkins, J. J. 1990. Graphs. An introductory approach.: Wiley.