

Protein structure similarities

Patrice Koehl

Comparison of protein structures can reveal distant evolutionary relationships that would not be detected by sequence information alone. This helps to infer functional properties. In recent years, many methods for pairwise protein structure alignment have been proposed and are now available on the World Wide Web. Although these methods have made it possible to compare all available protein structures, they also highlight the remaining difficulties in defining a reliable score for protein structure similarities.

Addresses

Department of Structural Biology, Fairchild Building,
Stanford University, Stanford, California 94305, USA;
e-mail: koehl@csb.stanford.edu

Current Opinion in Structural Biology 2001, 11:348–353

0959-440X/01/\$ – see front matter

© 2001 Elsevier Science Ltd. All rights reserved.

Abbreviations

3D	three-dimensional
PDB	Protein Data Bank
PSD	protein structural distance
RMS	root mean square
SSE	secondary structure element

Introduction: understanding protein structure is central to the post-genomic era

Only 10 years ago, sequencing the whole genome of even a simple organism appeared to be a formidable task that would require several decades. Major progress in molecular biology, as well as the strong determination of individuals, and both private and public organizations, has made this process a reality. As of the end of the year 2000, more than 30 genomes have been fully sequenced, including yeast, *Drosophila melanogaster* and *Caenorhabditis elegans*. All the sequences of the corresponding genes are publicly accessible (see, for example, the NCBI site at <ftp://ncbi.nlm.nih.gov/genbank/genomes>). The full value of these sequence data will only be realized when all gene sequences are assigned their roles in the cell. As it is not feasible to study experimentally every protein in all genomes, the function and biological role of a newly sequenced protein is usually inferred from a characterized protein using sequence and/or structure comparison methods. Functional inference based on sequence only is limited by the so-called twilight zone, where similarities can no longer be reliably detected (around 25% sequence identity). To improve and complement this method, a major effort aimed at increasing our knowledge of protein structure has been undertaken, under the name ‘structural genomics’ [1]. The success of this approach is strongly linked to our ability to organize in databases the wealth of information that results from the structural genomics effort, as well as depending on the development of data mining techniques that extract the relevant information contained in these databases.

All protein structures determined experimentally, either by X-ray crystallography or NMR spectroscopy, are deposited in a centralized resource, the Protein Data Bank (PDB) [2]. As of 5th December 2000, the PDB contains 13,861 structures of proteins, nucleic acids and protein–nucleic acid complexes. A striking feature derived from this wealth of data is that nearly all proteins have structural similarities to other proteins. Although these similarities may arise from general principles of physics and chemistry that limit the number of protein folds, they may also result from evolutionary relationships. Approaches that identify and examine these structural relationships have relied on the classification of proteins, using either structural information alone (CATH [3] and FSSP [4]) or a combination of structural and evolution information with a good deal of human expertise (SCOP [5]). In this paper, I will review recent progress in how structural similarities are identified.

Protein structure alignment

Any classification of a set of objects into clusters of similar objects requires a definition of similarity and dissimilarity. In the case of protein molecules, such a measure is provided by structural alignment. A structural comparison program needs to be automatic and fast; the latter criterion is crucial for large-scale all-against-all computer experiments required for clustering the protein structure space [6]. Though significant progress has been made over the past decade, a fast, reliable and convergent method for protein structural alignment is not yet available. Recent developments have focused both on the search algorithm and on defining the target function to be minimized, that is, a quantitative measure of the quality of an alignment.

The most direct approach to the comparison of two protein structures is to move the set of points representing one structure as a rigid body over the other, and look for equivalent residues. This can only be achieved for relatively similar structures and will fail to detect local similarities of structures sharing common substructures. To avoid this problem, the structures can be broken into fragments (usually secondary structure elements [SSEs]), but this can lead to situations in which the global alignment can be missed. Recent work has focused on combining the local and global criteria in a hierarchical approach. These methods proceed by first defining a list of equivalent positions in the two structures, from which a structural alignment can be derived. This initial equivalence set is defined by methods such as dynamic programming [7,8^{*}], comparison of distance matrices [9], fragment matching [10,11], geometric hashing [12,13], maximal common subgraph detection [14,15] and local geometry matching [16]. Optimization of this equivalence set is performed using dynamic programming [8^{*},17,18,19^{*}], Monte Carlo algorithms or simulated

annealing [9], a genetic algorithm [20*] and incremental combinatorial extension of the optimal path [21].

Most of the methods for protein structure alignment quantify the quality of the alignment on the basis of geometric properties of the set of points representing the structures. Some of these methods compare the respective distance matrices of each structure, trying to match the corresponding intramolecular distances for selected aligned substructures [7,9,21,22]. Other methods compare the structures directly after superposition of aligned substructures, trying to match the positions of corresponding atoms [10,11,16,17,23]. Interestingly, there is no consensus on the definition of a match of distances or of atomic positions needed for either of these two schemes. When comparing two pairs of atoms between two structures, Taylor and Orengo [7] defined a distance or similarity score in the form $a/(D+b)$, where D is the difference between the two intramolecular distances, and a and b are arbitrarily defined constant values. Holm and Sander [9] defined a similarity score as $(a-[D/\langle D \rangle])\exp(-[\langle D \rangle/b]^2)$, where $\langle D \rangle$ is the average of the two intramolecular distances. Rossmann and Argos [24], and Russell and Barton [25] used a score $\exp(-[D/a]^2)\exp(-[S/a]^2)$, where S takes into account local neighbors for each pair of atoms. As another example of a scoring scheme for minimizing intermolecular distances, Levitt and co-workers [17,18] defined a score $a/(1+[R/b]^2)$, where R is the distance between a pair of corresponding atoms in the two structures. At this stage, there is no clear evidence as to which score performs best.

All the techniques cited above use geometry for the comparison, ignoring similarities in the environment of the residues. Suyama *et al.* [26] proposed another approach in which they ignored the 3D geometry altogether and compared structures on the basis of 3D profiles [27] alone, using dynamic programming. These profiles include information on solvent accessibility, hydrogen bonds, local secondary structure states and sidechain packing. Although this method is able to align two-domain proteins with different relative orientations of the two domains, it often generates inaccurate alignments [26]. Jung and Lee [28*] recently improved upon this method by iteratively refining the initial profile alignment using dynamic programming and 3D superposition. Their method, referred to as SHEBA, was found to be fast and as reliable as other alignment techniques (though it was only tested on a small number of protein pairs).

Kawabata and Nishikawa [29*] derived a novel scoring scheme for generating structural alignments based on the Markov transition model of evolution. The similarity score between two structures i and j is defined as $\log P(j \rightarrow i)/P(i)$, where $P(j \rightarrow i)$ is the probability that structure j changes to structure i during evolution, and $P(i)$ is the probability that structure i appears by chance. The probabilities are estimated using a Markov transition model that is equivalent to the Dayhoff's substitution model for amino acids. Three types

of scores were considered: a score based on accessibility to solvent; a residue-residue distance score; and an SSE score. They show that their method recognizes more similarities between proteins known to be homologous than FSSP [4].

Root mean square as a measure of protein similarity

Though most of the algorithms for protein structure alignments use scoring schemes that differ significantly from simply taking into account interatomic distances (see above), the root mean square (RMS) deviation remains the measure reported to describe the similarity between two proteins. Two different RMS values have been proposed, $cRMS$ and $dRMS$. Given two sets of coordinates, the $cRMS$ is the norm of the distance vector between the two sets, provided that they have been optimally superposed:

$$cRMS = \sqrt{\frac{1}{N} \sum_{i=1}^N (\|\mathbf{x}(i) - \mathbf{y}(i)\|^2)} \quad (1)$$

where N is the number of atoms in the list of equivalence, and \mathbf{x} and \mathbf{y} are the coordinates of atom indexed i in protein A and protein B, respectively.

The $dRMS$ measures the difference between the respective distance matrices of each structure:

$$dRMS = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N (d_{ij}^A - d_{ij}^B)^2} \quad (2)$$

where d_{ij}^A and d_{ij}^B are the distances between atoms i and j in molecules A and B, respectively. Both $cRMS$ and $dRMS$ are based on the L2-norm (i.e. the Euclidian norm) and, as such, they suffer from the same drawback as the residual, χ^2 , in least-squares minimization: the presence of outliers introduces a bias in the search for an optimal fit and the final measure of the quality of the fit may be artificially poor because of the sole presence of these outliers. As a result, RMS is a useful measure of structural similarity only for closely related proteins [30]. Several other measures have therefore been proposed to circumvent these problems. In an attempt to provide a unified statistical framework for sequence comparison and structure comparison, Levitt and Gerstein [31] defined a structural similarity score S_2 that sums their scoring scheme for structural alignment (see above):

$$S_2 = A \left(\sum_{i=1}^N \frac{1}{1 + \left(\frac{\|\mathbf{x}(i) - \mathbf{y}(i)\|}{B} \right)^2} - Ngap/2 \right) \quad (3)$$

S_2 was defined as a more reliable indicator of structure similarity than RMS because it depends most strongly on the best-fitting pairs of atoms (thereby removing the weights of outliers), whereas RMS gives equal weight to all pairs of

atoms. Interestingly, Lesk [32] recently proposed replacing the L₂-norm in the RMS definition by the L_∞ norm, also called the Chebyshev norm, yielding a new score:

$$S_{\infty} = \max_{i \in [L, N]} \{ \| \mathbf{x}(i) - \mathbf{y}(i) \| \} \quad (4)$$

S_{∞} reports the worst-fitting pair of atoms (after optimal superposition of the two structures) and, as such, is even more sensitive to outliers than the RMS.

Yang and Honig [19[•]] defined a new protein structure similarity measure, the protein structural distance (PSD). PSD combines a secondary structural alignment score and the RMS deviation of topologically equivalent residue pairs. It thus incorporates the resolution power of both RMS for closely related structures and the secondary structure score for proteins that can be very different. By analyzing the PSD scores obtained from more than one and a half million pairs of proteins, Yang and Honig proposed that there is a continuous aspect of protein conformation space, in apparent disagreement with structural classification databases such as SCOP (Structural Classification Of Proteins [5]) and CATH (Class, Architecture, Topology and Homologous Superfamilies [3]).

May [33[•]] assessed 37 different protein structure similarity measures in terms of their robustness in generating robust and accurate clusters in a hierarchical classification of 24 protein families. Interestingly, it was found in this study that the sum of ranks of distances at aligned positions was a better measure than the direct sum of distances and that RMS computed over the subset of core-aligned positions performs better than normal RMS. Variations in the hierarchical classification of protein structures raise the question of the validity not only of the measure used for the clustering, but also of the hierarchical clustering itself.

The difficulty of defining a similarity score between protein structures is most probably a reflection of the fact that the problem of structure comparison does not have a unique answer [34–36]. This could also reflect the fact that the problem is ill posed and that additional information is required to characterize a problem with a well-defined solution. For example, in fold recognition applications, predictors will focus on the well-conserved core region of the protein and pay less attention to the loop geometry. In such cases, it makes sense to define a similarity score that only includes atoms in the core.

Similarity measures for protein structure prediction

A quantitative measure of the similarities of protein structures is essential for a critical assessment of the quality of protein structure predictions, such as those generated for CASP (a community-wide experiment on the Critical Assessment of techniques for protein Structure Prediction, organized in the form of a meeting held in alternating years at Asilomar, California). In the special case of comparing a

predicted structure with the corresponding experimental structure, the equivalence list is known because the two sequences are identical, which reduces the complexity of the problem. On the other hand, each prediction may omit different residues and different parts of the structure may have different accuracies.

Hubbard [37[•]] solved the problem by generating a large number of superpositions and calculating the best RMS for each number of equivalent residues (not necessarily contiguous). The result is the RMS/coverage graph, which was used for the evaluation of predictions at CASP3. This plot can also be interpreted as defining the number of equivalent residues for a given RMS value (the Adam Zemla's global distance test, GDT, used in CASP4).

Defining recurrent local motifs in protein conformations

Steric and chemical constraints reduce the number of viable conformations of amino acid residues within a protein [38]. Interestingly, these spatial limitations are not independent in consecutive residues along the protein sequence. For example, amino acids within a secondary structure element (α helix or β strand) have nearly identical local geometry. Even in the general case, the correlations between the geometries of consecutive residues are so strong that it should be possible to construct a small data bank of protein fragments that can be used as elementary building blocks to reconstruct virtually all native protein structures [39]. Databases of protein fragments have proved useful for protein structure reconstruction based on experimental data [40] and for homology modeling [41–45], as well as for *ab initio* protein structure prediction [46].

Several methods of classification of protein loops have been described [47–49]. Some of these procedures have been extended and applied to the problem of the automatic definition of recurrent local structural motifs [43,50–53,54[•],55]. Basically, these methods first extract a large database of protein fragments (overlapping or not) and this database is subjected to a clustering algorithm. Though simple in concept, these procedures raise some challenges that require special attention. Firstly, similarities between overlapping fragments generate noise for the clustering algorithm. Secondly, clustering algorithms require a measure of similarity (S) that satisfies the triangular inequality (i.e. for three fragments A, B and C, $S[A,B] \leq S[A,C] + S[B,C]$). Although this is the case for RMS, the latter might not be a good measure of protein structure similarity (see above). Finally, the question of defining the best fragment size to consider has not yet been solved. Despite these problems, Unger *et al.* [43] have shown that about 100 representative hexamers can be combined to cover 99% of the structures. Micheletti *et al.* [54[•]] have recently improved upon this initial study, showing that, with nonredundant libraries containing fragments of 4, 5 or 6 residues, they can fit a set of 10 proteins to within 1 Å.

Table 1

Web sites for protein structure alignment servers and programs.

Program	Server interface	Program download	Method
CE	http://cl.sdsc.edu	ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/src	Extension of the optimal path [21]
DALI	http://www2.ebi.ac.uk/dali	http://jura.ebi.ac.uk:8765/~holm/DaliLite	Distance matrix alignment. This is the most widely used program [9].
KENOBI	http://sullivan.bu.edu/kenobi	http://sullivan.bu.edu/kenobi	Genetic algorithm [20•]
PRISM		http://www.columbia.edu/~ay1	SSE alignment followed by iterative refinement of the equivalence list [19•]
PROSUP	http://anna.came.sbg.ac.at/prosup/main.html	ftp://ftp.came.sbg.ac.at/pub/Prosup	Hierarchical alignment (to build initial equivalence list), followed by dynamic programming refinement [35].
SAP		http://mathbio.nimr.mrc.ac.uk/tools	Double dynamic programming [8•]
SHEBA	http://lily.nci.nih.gov/~jung/index.html	http://lily.nci.nih.gov/~jung/sheba_program.html	Hierarchical alignment with profiles [28•]
TOP	http://bioinfo1.mbfys.lu.se/TOP	ftp://bioinfo1.mbfys.lu.se/pub/guoguang(topv6)	SSE alignment [56•]
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html		Vector alignment [57]
STRUCTAL	http://bioinfo.mbb.yale.edu/align/server.cgi		Double dynamic programming [17]

The methods described above define contiguous structural motifs in proteins. Wako and Yamato [53] recently proposed a novel method to detect motifs that is free of this limitation. In their approach, the Delaunay tessellation is applied to the set of C α atoms of the proteins. Each tetrahedron of the tessellation is given a code (i.e. a string of digits) based on properties of the tetrahedron and its neighbors. Tetrahedra with the same code are grouped into sets. The local structures in each set were found to be similar enough to represent a motif. Some of these motifs are parts of secondary structures and others are irregular.

Conclusions

Comparing two protein structures and giving a quantitative measure of their similarities remains an active area of development in structural biology, as demonstrated by the number and diversity of new methods for protein comparison that have recently been published. Most of these methods are fast enough to make full database searches possible. Furthermore, many groups involved in this research have generously made their programs available for use over the Internet and the World Wide Web. In some cases, the program itself is accessible for download, either as an executable or as a full source package (Table 1). These are wonderful tools and I do encourage the reader to test several of these sites.

Defining the similarities between two protein structures remains a difficult problem. The exponential increase of the size of the structural databases introduces new constraints in that methods developed for measuring structural similarities must be fast enough to allow full database searches. On the other hand, this is what makes this whole field both fascinating and essential for structural genomics, in that these databases contain a wealth of information that still needs to be unraveled.

The problem of comparing two protein structures can be reformalized as the problem of comparing two sets of points in 3D space. As such, it can be seen as a classical problem of computational geometry, and it is expected that collaboration between structural biologists well versed in deciphering protein structures and computer scientists who focus on geometric problems should provide the synergy required for significant progress.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Kim SH: **Shining a light on structural genomics.** *Nat Struct Biol* 1998, 5:643-645.
 2. Bernstein FC, Koetzle TF, Williams G, Meyer DJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based archival file for macromolecular structures.** *J Mol Biol* 1977, 112:535-542.
 3. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH: a hierarchic classification of protein domain structures.** *Structure* 1997, 5:1093-1108.
 4. Holm L, Sander C: **The FSSP database of structurally aligned protein fold families.** *Nucleic Acids Res* 1994, 22:3600-3609.
 5. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, 247:536-540.
 6. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, 6:377-385.
 7. Taylor WR, Orengo CA: **Protein structure alignment.** *J Mol Biol* 1989, 208:1-22.
 8. Taylor WR: **Protein structure comparison using iterated double dynamic programming.** *Protein Sci* 1999, 8:654-665.
This paper describes a protein structure comparison method that allows the generation of large populations of high scoring alternate alignments. This method is an improvement upon Taylor's initial method, described in [7].
 9. Holm L, Sander C: **Protein-structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, 233:123-138.

10. Vriend G, Sander C: **Detection of common three-dimensional substructures in proteins.** *Proteins* 1991, 11:52-58.
11. Alexandrov NN, Takahashi K, Go N: **Common spatial arrangements of backbone fragments in homologous and nonhomologous proteins.** *J Mol Biol* 1992, 225:5-9.
12. Fischer D, Bachar O, Nussinov R, Wolfson H: **An efficient automated computer vision based technique for detection of three-dimensional structural motifs in proteins.** *J Biomol Struct Dyn* 1992, 9:769-789.
13. Nussinov R, Fischer D, Wolfson H: **A computer vision based three-dimensional approach for the comparison of protein structures.** *FASEB J* 1992, 6:A349.
14. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P: **A graph-theoretic approach to the identification of three-dimensional patterns of amino-acid side-chains in protein structures.** *J Mol Biol* 1994, 243:327-344.
15. Artymiuk PJ, Poirrette AR, Rice DW, Willett P: **The use of graph-theoretical methods for the comparison of the structures of biological macromolecules.** *Topics Curr Chem* 1995, 174:73-103.
16. Wu TD, Schmidler SC, Hastie T, Brutlag DL: **Regression analysis of multiple protein structures.** *J Comput Biol* 1998, 5:585-595.
17. Subbiah S, Laurents DV, Levitt M: **Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core.** *Curr Biol* 1993, 3:141-148.
18. Gerstein M, Levitt M: **Comprehensive assessment of automatic structural alignment against a manual standard; the SCOP classification of proteins.** *Protein Sci* 1998, 7:445-456.
19. Yang AS, Honig B: **An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance.** *J Mol Biol* 2000, 301:665-678.
 This paper describes a new measure of protein structural similarity, the protein structural distance (PSD). PSD includes both a secondary structure alignment score and RMS. Using PSD scores computed over more than one and a half million pairs of protein structures, Yang and Honig show that there is a continuous aspect of protein conformation space.
20. Szustakowski JD, Weng ZP: **Protein structure alignment using a genetic algorithm.** *Proteins* 2000, 38:428-440.
 This paper describes a new method for protein structure alignment based on a genetic algorithm.
21. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, 11:739-747.
22. Mizuguchi K, Go N: **Comparison of spatial arrangements of secondary structural elements in proteins.** *Protein Eng* 1995, 8:353-362.
23. Madej T, Gibrat JF, Bryant SH: **Threading a database of protein cores.** *Proteins* 1995, 23:356-369.
24. Rossmann MG, Argos P: **Exploring structural homology of proteins.** *J Mol Biol* 1976, 105:75-95.
25. Russell RB, Barton GJ: **Multiple protein-sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.** *Proteins* 1992, 14:309-323.
26. Suyama M, Matsuo Y, Nishikawa K: **Comparison of protein structures using 3D-profile alignment.** *J Mol Evol* 1997, 44:S163-S173.
27. Bowie JU, Lüthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, 253:164-170.
28. Jung J, Lee B: **Protein structure alignment using environmental profiles.** *Protein Eng* 2000, 13:535-543.
 A new hierarchical protein structure alignment method is described. An initial alignment is derived by comparing the environmental profiles of the two proteins, without consideration of their 3D structures. This alignment is then iteratively refined, in which new alignments are found by 3D superposition of the structures.
29. Kawabata T, Nishikawa K: **Protein structure comparison using the Markov transition model of evolution.** *Proteins* 2000, 41:108-122.
 This paper describes a new score to evaluate protein structure similarity. Transition probabilities $P(i \rightarrow j)$ between two structures i and j are evaluated using the Markov transition model, which is similar to the Dayhoff's substitution model. These probabilities are used to derive a similarity score between two structures i and j as $\log P(i \rightarrow j)/P(i)$, where $P(i)$ is the probability that structure i appears by chance. A structure comparison program was developed based on this score and this program was found to recognize more homologous similarity than DALI. Unfortunately, this program is not publicly available.
30. Mizuguchi K, Go N: **Seeking significance in three-dimensional protein-structure comparisons.** *Curr Opin Struct Biol* 1995, 5:377-382.
31. Levitt M, Gerstein M: **A unified statistical framework for sequence comparison and structure comparison.** *Proc Natl Acad Sci USA* 1998, 95:5913-5920.
32. Lesk AM: **Extraction of geometrically similar substructures: least-squares and Chebyshev fitting and the difference distance matrix.** *Proteins* 1998, 33:320-328.
33. May ACW: **Towards more meaningful hierarchical classification of amino acid scoring matrices.** *Protein Eng* 1999, 12:707-712.
 Protein structure similarity measures are assessed in terms of the robustness of the resulting trees generated by hierarchical clustering of 24 known protein families. This paper emphasizes the problems of RMS as a measure of similarities and, more generally, the need to assess the applicability of hierarchical clustering to structural data.
34. Orengo CA, Swindells MB, Michie AD, Zvelebil MJ, Driscoll PC, Waterfield MD, Thornton JM: **Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity.** *Protein Sci* 1995, 4:1977-1983.
35. Feng ZK, Sippl MJ: **Optimum superimposition of protein structures: ambiguities and implications.** *Fold Des* 1996, 1:123-132.
36. Godzik A: **The structural alignment between two proteins: is there a unique answer?** *Protein Sci* 1996, 5:1325-1338.
37. Hubbard TJP: **RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions.** *Proteins* 1999, 37:15-21.
 Structure comparison is an essential part of the assessment of protein structure prediction. This paper describes a new method for that specific problem, in which the results of a large number of superpositions of sets of residues in the prediction and in the experimental structures (not necessarily contiguous) are presented graphically, in the so-called RMS/coverage graphs. These graphs have proven very valuable for assessing the results of CASP3.
38. Ramachandran G, Sasisekharan V: **Conformation of polypeptides and proteins.** *Adv Protein Chem* 1968, 23:283-437.
39. Jones TA, Thirup S: **Using known substructures in protein model-building and crystallography.** *EMBO J* 1986, 5:819-822.
40. Kleywegt GJ, Jones TA: **Databases in protein crystallography.** *Acta Crystallogr D* 1998, 54:1119-1131.
41. Claessens M, Vancutsem E, Lasters I, Wodak S: **Modeling the polypeptide backbone with spare parts from known protein structures.** *Protein Eng* 1989, 2:335-345.
42. Blundell T, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibanda BL, Sutcliffe M: **18th Krebs Hans lecture: knowledge-based protein modeling and design.** *Eur J Biochem* 1988, 172:513-520.
43. Unger R, Harel D, Wherland S, Sussman JL: **A 3D building-blocks approach to analyzing and predicting structure of proteins.** *Proteins* 1989, 5:355-373.
44. Summers N, Karplus M: **Modelling of globular proteins. A distance-based search procedure for the construction of insertion/deletion regions and Pro-nonPro mutations.** *J Mol Biol* 1990, 216:991-1016.
45. Levitt M: **Accurate modelling of protein conformation by automatic segment matching.** *J Mol Biol* 1992, 226:507-533.
46. Simons KT, Kooperberg C, Huang ES, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J Mol Biol* 1997, 268:209-225.
47. vanVlijmen HWT, Karplus M: **PDB-based protein loop prediction: parameters for selection and methods for optimization.** *J Mol Biol* 1997, 267:975-1001.
48. Rufino SD, Donate LE, Canard LHJ, Blundell TL: **Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling.** *J Mol Biol* 1997, 267:352-367.
49. Li WZ, Liu ZJ, Lai LH: **Protein loops on structurally similar scaffolds: database and conformational analysis.** *Biopolymers* 1999, 49:481-495.

50. Rooman MJ, Rodriguez J, Wodak SJ: **Automatic definition of recurrent local-structure motifs in proteins.** *J Mol Biol* 1990, **213**:327-336.
51. Conklin D: **Machine discovery of protein motifs.** *Machine Learning* 1995, **21**:125-150.
52. Lessel U, Schomburg D: **Similarities between protein 3D structures.** *Protein Eng* 1994, **7**:1175-1187.
53. Wako H, Yamato T: **Novel method to detect a motif of local structures in different protein conformations.** *Protein Eng* 1998, **11**:981-990.
54. Micheletti C, Seno F, Maritan A: **Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies.** *Proteins* 2000, **40**:662-674.

The goal of the study described in this paper is to provide minimal sets of protein oligomers (termed 'oligons') that are able to represent any protein.

It is shown that meaningful classifications of protein fragments cannot be done for lengths greater than six or less than four residues. On the other hand, a few dozen oligons of four, five or six residues can be used to reproduce any protein. The libraries of oligons are available at <http://www.sissa.it/~michelet/prot/repset>.

55. Diamond R: **On the multiple simultaneous superposition of molecular-structures by rigid body transformations.** *Protein Sci* 1992, **1**:1279-1287.
56. Lu GG: **TOP: a new method for protein structure comparisons and similarity searches.** *J Appl Crystallogr* 2000, **33**:176-183.
This paper describes a suite of programs available on the World Wide Web for protein structure comparison. The comparison is performed by fragment matching between the two proteins. These programs include options for database processing via Internet-based and Web-based servers.
57. Gibrat JF, Madej T, Spouge JL, Bryant SH: **The vast protein structure comparison method.** *Biophys J* 1997, **72**:MP298.