PROTEINS: Structure, Function, and Genetics 42:378–382 (2001)

Protein Structural Alignments and Functional Genomics

James A. Irving,[1] James C. Whisstock,[1] and Arthur M. Lesk[2]*
[1]*Department of Biochemistry and Molecular Biology, Monash University, Clayton Campus, Melbourne, Victoria, Australia*
[2]*Department of Haematology, Wellcome Trust Centre for Molecular Mechanisms in Disease, Cambridge Institute for Medical Research, University of Cambridge Clinical School, Cambridge, United Kingdom*

ABSTRACT **Structural genomics—the systematic solution of structures of the proteins of an organism—will increasingly often produce molecules of unknown function with no close relative of known function. Prediction of protein function from structure has thereby become a challenging problem of computational molecular biology. The strong conservation of active site conformations in homologous proteins suggests a method for identifying them. This depends on the relationship between size and goodness-of-fit of aligned substructures in homologous proteins. For all pairs of proteins studied, the root-mean-square deviation (RMSD) as a function of the number of residues aligned varies exponentially for large common substructures and linearly for small common substructures. The exponent of the dependence at large common substructures is well correlated with the RMSD of the core as originally calculated by Chothia and Lesk (EMBO J 1986;5:823–826), affording the possibility of reconciling different structural alignment procedures. In the region of small common substructures, reduced aligned subsets define active sites and can be used to suggest the locations of active sites in homologous proteins. Proteins 2001;42:378–382.**

Key words: protein structure—function relationships; structural genomics; alignment

INTRODUCTION

Biological sequence alignment is the assignment of residue–residue correspondences. Structural alignment of proteins matches amino acids that occupy corresponding regions in molecular space. It is the basis of classifications of protein folds and underlies threading approaches to protein fold recognition. Computationally, structural alignment is related to the problem of extracting a maximal common substructure from point sets.[1–10]

The very large number of structural alignment methods creates the difficulty that different techniques often select different sets of residues from two protein structures.[9,11,12] The ambiguity is intrinsic to the problem: In principle, one method may align relatively small substructures that "fit" relatively well; another may choose larger substructures, with a poorer fit.

For closely related proteins, the definition of the core by Chothia and Lesk[13] provides a standardized set of alignable residues. In that work, results from all pairs of homologous proteins studied fell on a single curve relating the number of identical amino acids in aligned positions of the core to the root-mean-square deviation (RMSD) of the main-chain atoms.[1] This provides a quantitative measure of structural similarity. However, the observed simple relationship deteriorates somewhat at large divergence.[14] Also, the definition of the core cannot be applied to the set of proteins with little or no secondary structure.

To reconcile the results of different alignment methods that select substructures of different size, we studied the variation of RMSD against the number of residues, n. This variation has been presented previously by us[9] and others (e.g., Hubbard[15]), but it has not been analyzed in quantitative detail before.

Ideally, for each n less than the size of maximal common substructure, we would like to find the best-fitting substructure containing the $C\alpha$ atoms of n residues, and report the RMSD in atomic position. However, there is no guarantee that the maximal common substructure of any given size is unique,[9,11,12] and it is a difficult combinatorial problem to generate the complete tree of common substructures.[8,9] We therefore used an approximation, the "sieving" procedure.[16] Given two homologous proteins, we perform the following steps:

1. Find an approximation to the maximal common substructure of the sets of Cα atoms[17]
2. Report n and the RMSD in position of the corresponding atoms
3. Find and delete the worst-fitting pair
4. Recompute the superposition of the remaining pairs
5. Until the RMSD <0.2 Å, go to step 2
6. Graph the RMSD against n

RESULTS

The curves solve one of the problems that motivated this study: how to compare structural alignments of different pairs of proteins. Figure 1 shows fits of common substructures of the Papaya cysteine protease papain and a series

Grant sponsor: National Health and Medical Reseach Council of Australia; Grant number: 997144; Grant sponsor: Wellcome Trust

*Correspondence to: A.M. Lesk, Department of Haematology, Wellcome Trust Centre for Molecular Mechanisms in Disease, Cambridge Institute for Medical Research, University of Cambridge Clinical School, Wellcome/MRC Building, Hills Road, Cambridge CB2 2QH, United Kingdom. E-mail: aml2@mrc-lmb.cam.ac.uk

Received 18 August 2000; Accepted 29 September 2000

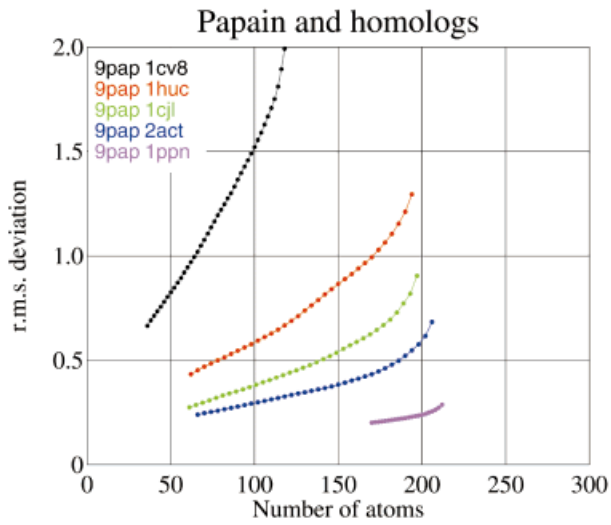

Fig. 1. Comparisons of Papaya papain [9pap] with an independent structure determination of the same molecule [1ppn] and with three increasingly diverged homologues: actinidin [2act], procathepsin L [1cjl], and cathepsin B [1huc]. The percentage of residues identical to 9pap is as follows: 1ppn, 100%; 2act, 49.0%; 1cjl, 42.6%; 1huc, 32.5%; and 1cv8, 16.1%.

TABLE I. Structural Alignment of Papaya Papain With an Independent Structure Determination of the Same Molecule and With Three Increasingly Diverged Homologues

Protein	PDB code	Source	Initial no. of residues aligned	RMSD of Cαs (Å)
Papain	9pap	Papaya	212	—
Papain	1ppn	Papaya	212	0.29
Actinidin	2act	Kiwi fruit	206	0.68
Procathepsin L	1cjl	Human	197	0.90
Cathepsin B	1huc	Human	194	1.29

PDB, Protein Data Bank; RMSD, root-mean-square deviation.

of homologues, including an alternative crystal form of papain itself (Table I). These graphs define the order of similarity much more perspicuously than single (n, RMSD) pairs: more distant homologues yield steeper, more elevated curves.

In order to produce results comparable to those achieved with established procedures, we applied this procedure to the set of homologous proteins treated by Chothia and Lesk,[13] comprising 32 pairs, including globins, cytochromes c, immunoglobulins, serine and cysteinyl proteinases, lysozymes, and dihydrofolate reductases; all coordinate sets are available from the Protein Data Bank (PDB).[18] Proteins from all-α, all-β, α-β, and α/β classes of proteins are represented.

Different protein pairs generate curves similar in general shape (Figs. 1, 2). For all 32 pairs of proteins, the graphs of RMSD against n have a linear region for small values of n and a region of exponential growth at larger values of n. Figure 2 shows the typical quality of the fits. In the 32 pairs of proteins studied, the range of the linear region varies within a range of 3–85 residues, with a mean of 27.2.

There is a good correlation between the RMSD of the core[13] and the exponent describing the behavior at high n (Fig. 3). This correlation has an important consequence in that it reconciles different structural alignment methods: The exponential parameter depends on the shape of the curve, rather than on a single point produced by a particular structure alignment method. From different structure alignment methods producing different aligned substructures from a given pair of proteins, one could derive the same exponential parameter by generating the curve of RMSD against n, and fitting it numerically. Using the correlation established in Figure 3, different methods initially producing aligned substructures of different sizes can yield the same core RMSD value.

APPLICATION TO STRUCTURAL GENOMICS: PREDICTION OF BINDING SITES IN PROTEINS OF UNKNOWN FUNCTION

After extensive "sieving" to produce a well-fitting reduced aligned substructure, the surviving regions of the structure, near the lower end of the curve, often include the active sites plus additional core secondary structural elements that appear to lend structural support to the binding site. Chothia and Lesk,[13] and others have pointed out that active sites are often the best-preserved regions in a divergent family of proteins. Such regions conserve the stereospecific interactions involved in ligand binding and catalysis. This procedure described therefore has potential application in identifying the active site of a family of homologous proteins of known structure but unknown function. This problem will grow in importance with the development of "structural genomics."

The pair of structures, YabJ from *Bactilis subtilis*[19] (PDB entry 1qd9), and YjgF from *Escherichia coli,*[20] (1qu9) are proteins of unknown mechanism of function related to the enzyme of known structure and function, chlorismate mutase.[21] All three structures are trimeric. Supposing that no homologue of known function were available—what could our methods reveal about the function of YabJ and YjgF? The initial structural alignment of YabJ and YjgF matched 358 residues, the $C\alpha$ of which fit with RMSD 0.94 Å. 52% of the aligned residues are identical. Sieving reduced this to a common subset of 107 aligned residues (67% identical; note the enrichment) that fit with RMSD 0.24 Å (shown in blue and black, Fig. 4). (Because the molecules are trimeric, this corresponds to 36 residues per monomer.) Structural superposition of chlorismate mutase[22] [2cht] and YabJ allowed us to position the transition-state-analogue inhibitors of chlorismate mutase within the YabJ and YjgF structures (magenta in Fig. 4; only the inhibitors appear in Fig. 4; the chlorismate mutase protein is not shown). The well-fitting minimal substructure of YabJ and YjgF includes the surroundings of the inhibitor, suggesting that YabJ and YjgF are enzymes, sharing a common active site with chlorismate mutase. This is consistent with inferences from patterns of conservation in the amino acid sequences.[19] Note that if we knew that

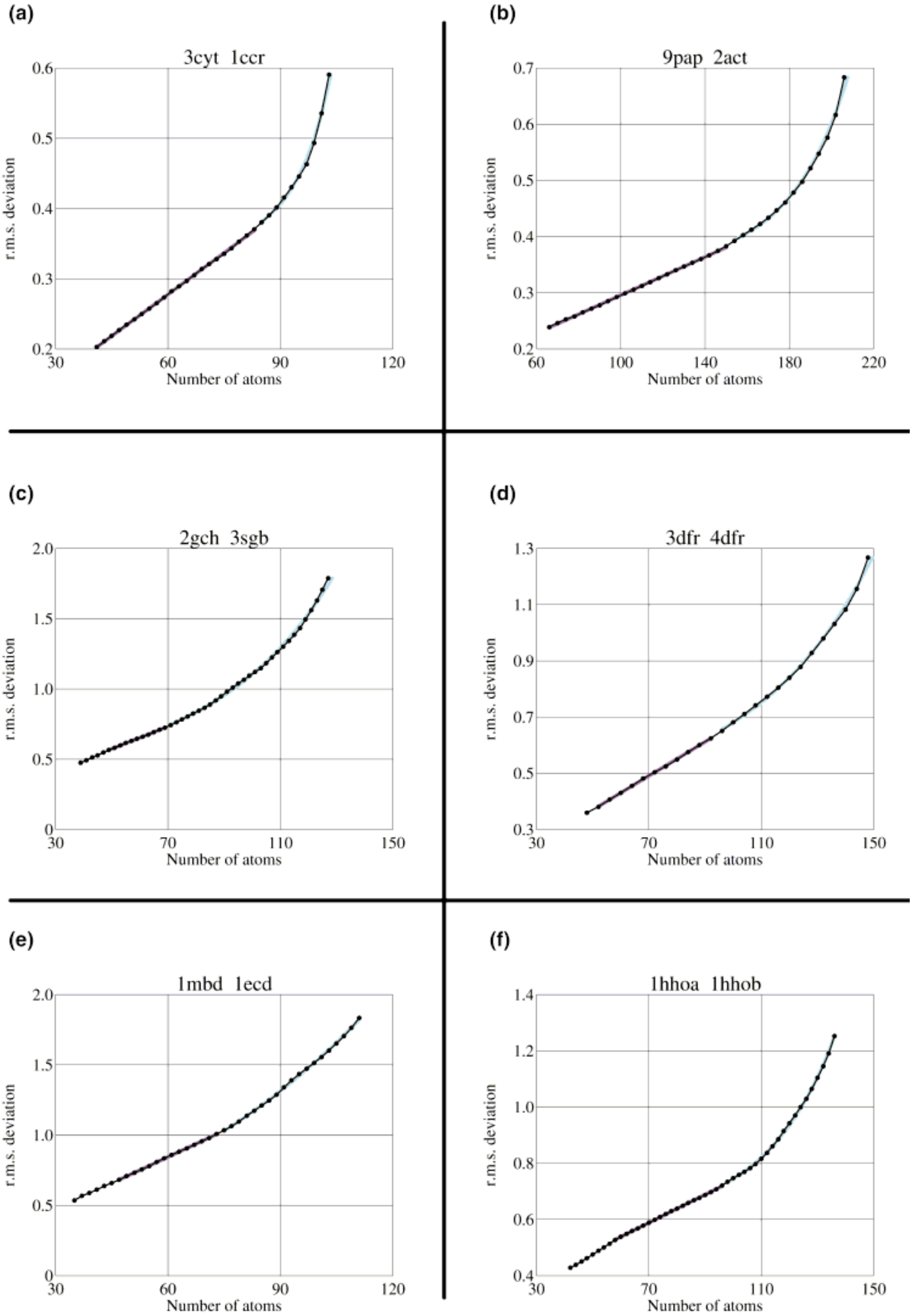
(a)
3cyt 1ccr
r.m.s. deviation
Number of atoms
(b)
9pap 2act
r.m.s. deviation
Number of atoms
(c)
2gch 3sgb
r.m.s. deviation
Number of atoms
(d)
3dfr 4dfr
r.m.s. deviation
Number of atoms
(e)
1mbd 1ecd
r.m.s. deviation
Number of atoms
(f)
1hhoa 1hhob
r.m.s. deviation
Number of atoms

Figure 2.

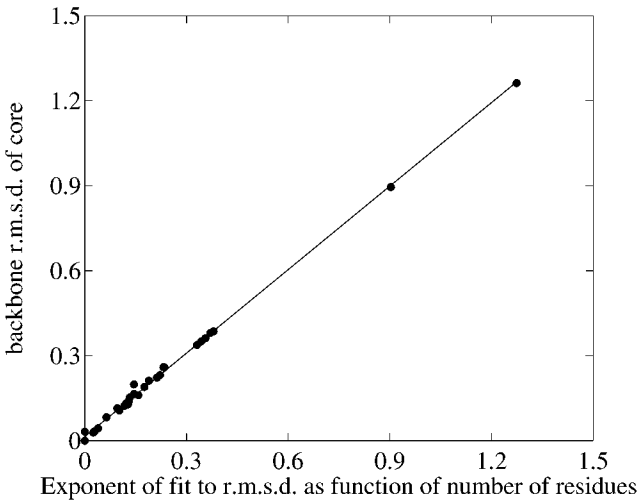

Fig. 3. Correlation of the root-mean-square deviation (RMSD) of the core as computed by Chothia and Lesk[13] with the exponent derived from fitting the curve of the variation of RMSD with number of residues in the region of larger substructures.

YabJ and YjgF were enzymes, the active site would be *expected* to appear at an interface between domains or subunits. The fact that the minimal substructure includes a crevice between domains is suggestive that the proteins of unknown function are in fact enzymes.

We emphasize that (1) this prediction of the binding sites of YabJ and YjgF is entirely independent of, and made no use of, the chlorismate mutase structures; and (2) the chlorismate mutase structures do not confirm our prediction—they merely make it reasonable. Experimental studies of YabJ and/or YjgF are required to test the prediction, which we wish to record before experiments reveal the answer.

Of course, many protein families do not have conserved binding sites and would give different results. In antibody variable domains, for instance, the binding sites are quite variable, but the VL–VH domain interface is well conserved,[23] and we would expect this to be retained as the best-preserved substructure. We are currently investigating protein families with different structural features to

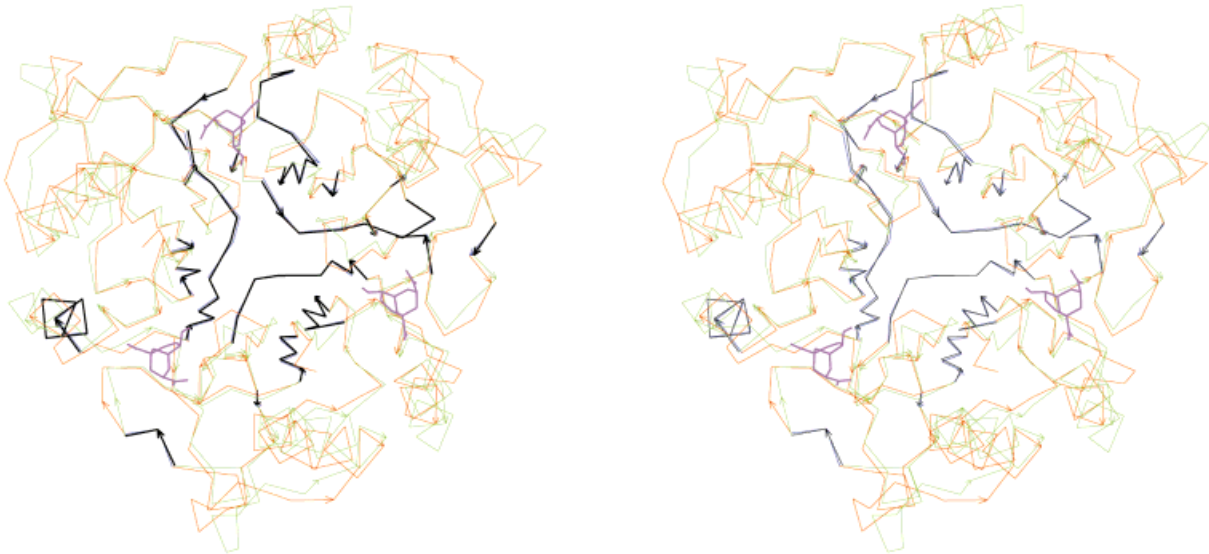

Fig. 4. Structures of related trimeric proteins YabJ from *Bacilis subtilis* [1qd9] and YjgF from *Escherichia coli* [1qu9] superposed on a reduced common substructure. Blue, reduced common substructure of YabJ; green, rest of YabJ; black, reduced common substructure of YjgF; red, rest of YjgF; magenta, inhibitors of chlorismate mutase [2cht]. Chlorismate mutase itself does not appear. The appearance of the inhibitors of chlorismate mutase in crevices in the YabJ and YgjF structures suggests that these crevices form the active sites of YabJ and YgjF and that they share this active site with chlorismate mutase. The prediction is based solely on the structures of YabJ and YgjF and could equally have been made if no structure of any relative with known function had been available.

Fig. 2. The variation of root-mean-square deviation (RMSD) with number of residues selected has two regimes: exponential behavior for large substructures and linear behavior for smaller substructures. The typical quality of the fits of these functions to the curves is shown: black, raw data; magenta, fit in linear region; cyan, fit in exponential region. **a:** Tuna and rice cytochromes c [3cyt, 1ccr]. **b:** papain and actinidin [9pap, 2act]. **c:** γ-Chymotrypsin and *S. griseus* proteinase b [2gch, 3sgb]; **d:** dihydrofolate reductases from *L. casei* and *Escherichia coli* [3dfr, 4dfr]. **e:** Sperm whale myoglobin and *Chironomus* erythrocruorin [1mbd, 1ecd]. **f:** α- and β-chains of human hemoglobin [1hho]. The method of data analysis involved first determining the extent of the linear region as follows. If N is the size of the maximal common substructure, for each interval $1 \leq I < j \leq N$, with $j - I \geq \max\{10, 0.4N\}$, we fitted the points of that interval to a straight line by standard least-squares formulae and computed $s\Delta$ = the standard error.[24] We recorded the minimum value sΔ (min) and determined the longest interval $n_1 \ldots n_2$, for which $s_\Delta \leq 1.5 \times s_\Delta$ (min). Suppose that the interval of linearity extends from numbers of residues n_1 to n_2. We then fitted the region $n_2 \leq n \leq N$ to an exponential function of the form: $A + Be^{\alpha(n - n_1)}$.

survey the types of regions of greatest structural similarity.

CONCLUSIONS

1. Graphs of RMSD against the number of residues aligned allow much more confident assignment of the degree of divergence than was achieved with single points.
2. It is possible to reconcile the results of different structural alignment methods by generating similar graphs from different starting points.
3. The graphs form a family of curves with exponential and linear regions. The exponential parameter is well correlated with the main-chain RMSD of the core.[13]
4. Examining the residues remaining after extensive sieving suggests the location of active sites in proteins of unknown function, a procedure with application to the results of structural genomics.

ACKNOWLEDGMENTS

The authors thank A.G. Murzin for discussion. J.C.W. thanks the National Health and Medical Research Council of Australia (997144) for support. A.M.L. is supported by the Wellcome Trust.

REFERENCES

1. Levine M, Stuart D, Williams J. A method for systematic comparison of the three-dimensional structures of proteins and some results. Acta Crystallogr A 1984; 40:600–610.
2. Karpen ME, de Haseth PL, Neet KE. Comparing short protein substructures by a method based on backbone torsion angles. Proteins 1989;6:155–167.
3. Liebman MN, Venanzi CA, Weinstein H. Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modelling enzyme recognition and specificity. Biopolymers 1985;24:1721–1758.
4. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
5. Vriend G, Sander C. Detection of common three-dimensional substructures in proteins. Proteins 1991;11:52–58.
6. Alexandrov NN, Takahashi K, Go N. Common spatial arrangement of backbone fragments in homologous and non-homologous proteins. J Mol Biol 1992;225:5–9.
7. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
8. Nichols WL, Rose GD, Ten Eyck LF, Zimm BH. Rigid domains in proteins: an algorithmic approach to their identification. Proteins 1995;23:38–48.
9. Lesk AM. Three-dimensional pattern matching in protein structure analysis. In: Galil Z Ukkonen E, editors. Combinatorial pattern matching. Lecture notes in computer science 937. Berlin: Springer-Verlag; 1995. p 248–260.
10. Bachar O, Fischer D, Nussinov R, Wolfson HJ. A computer vision based technique for 3-D sequence independent structural comparison of proteins. Protein Eng 1993;6:279–288.
11. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Folding Design 1996;1:123–132.
12. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
13. Chothia C, Lesk AM. Relationship between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
14. Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 1993;2:1811–1826.
15. Hubbard TJP. RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins Suppl 1999;3:15–21.
16. McPhalen CA, Vincent MG, Picot D, Jansonius JN, Lesk AM, Chothia C. Domain closure in mitochondrial aspartate aminotransferase. J Mol Biol 1992;227:197–213.
17. Lesk AM. Integrated access to sequence and structural data. In: Saccone C, editor. Biosequences: perspectives and user services in Europe. Brussels, Belgium: EEC; 1986. p 23–28 and references contained therein.
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
19. Sinha S, Rappu P, Lange SC, Mantsala P, Zalkin H, Smith J. Crystal structure of bacillus subtilis YabJ, a purine regulatory protein and member of the highly conserved YjgF family. Proc Natl Acad Sci USA 1999;96:13074–13079.
20. Volz K. A test case for structure-based functional assignment: the 1.2 Å crystal structure of the yjgf gene product from *Escherichia coli*. Protein Sci 1999;8:2428–2437.
21. 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.
22. Chook KM, Ke H, Lipscomb WN. Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. Proc Natl Acad Sci USA 1993;90:8600–8603.
23. Chothia C, Novotny J, Bruccoleri RE, Karplus M. Domain association in immunoglobulin molecules: the packing of variable domains. J Mol Biol 1985;186:651–663.
24. Snedecor GW. Statistical methods. Ames, IA: Iowa State University Press; 1965.