PROTEINS: Structure, Function, and Genetics 19:165–173 (1994)

MINI-REVIEW ARTICLE

Searching Protein Structure Databases Has Come of Age

Liisa Holm and Chris Sander
Protein Design Group, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

ABSTRACT **The number of protein structures known in atomic detail has increased from one in 1960 (Kendrew, J.C., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., Shore, V.C. Nature (London) 185:422–427, 1960) to more than 1000 in 1994. The rate at which new structures are being published exceeds one a day as a result of recent advances in protein engineering, crystallography, and spectroscopy. More and more frequently, a newly determined structure is similar in fold to a known one, even when no sequence similarity is detectable. A new generation of computer algorithms has now been developed that allows routine comparison of a protein structure with the database of all known structures. Such structure database searches are already used daily and they are beginning to rival sequence database searches as a tool for discovering biologically interesting relationships.**

Key words: algoriths, structure alignment, Protein Data Bank, protein superfamilies, structural homology

INTRODUCTION

Attending scientific conferences can lead to exciting revelations. A poster, presenting the three-dimensional structure of the recently solved biotin repressor, immediately caught the eye of the crystallographer R. Wierenga, whose group had just solved the structure of an unrelated small protein domain with homology to protein kinases, called SH3 (*src*-homology 3). Surprisingly, the C-terminus of biotin repressor folds into a six-stranded antiparallel β-barrel with identical topography to that in the SH3 domain (Fig. 1). At about the same time, A.G. Murzin,[2] using his encyclopedic memory of protein structures, recognized an SH3-like fold in the plasmid-encoded R67 dihydrofolate reductase,[3] also an unexpected relationship. On the other side of the globe, M.B. Swindells noted the similarity of the unusual elongated folds of the very distantly related tumor and nerve growth factors (Fig. 2), which appeared in print within some months.[4] These are anecdotal illustrations of the dramatic increase in the number of unexpected topological similarities between proteins with little or no sequence similarity over the past year or two, but they also are typical examples of the haphazard fashion in which most of these similarities were discovered.

BOOM OF NEW STRUCTURES

Soon, no single scientist will be able to hold all known protein structures in her memory and scan them for similarities. The world's journals publish experimentally determined protein structures at a rate that exceeds one paper per day (estimate for 1994). Every two or three weeks new and important structures are published. At the end of 1993, there were already 339 different structures in the Protein Data Bank (size of a representative set of protein structures,[5] using a 30% sequence identity cutoff), compared to only 155 a year earlier. In light of these developments, one may expect that structural analysis by visual inspection will soon be inadequate, in speed and/or precision, as the number of structures rapidly increases. How can we cope with this challenge?

We do not have to look far for a solution. The field of protein sequence analysis faced a similar situation many years ago, when sequence alignment using pencil and paper for all new sequences began to exceed human capacity. Efficient computer alignment algorithms were developed and turned into

Abbreviations: 1D, one-dimensional; 3D, three-dimensional; EGF, epidermal growth factor; HIV, human immunodeficiency virus; PDB, Protein Data Bank; SH2, *src*-homology 2; SH3, *src*-homology 3.

Received February 14, 1994; revision accepted March 25, 1994.
Address reprint requests to Liisa Holm and Chris Sander, Protein Design Group, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.

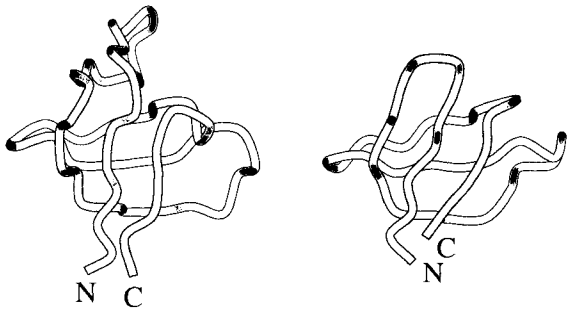

Fig. 1. SH3-fold. Small domains with an SH3-like topology have been identified in protein kinases (1SHG[39]), myosin subfragment-1, and in bacterial enzymes such as biotin repressor (1BIA[40]) and plasmid-encoded dihydrofolate reductase. Drawn using MolScript.[41]

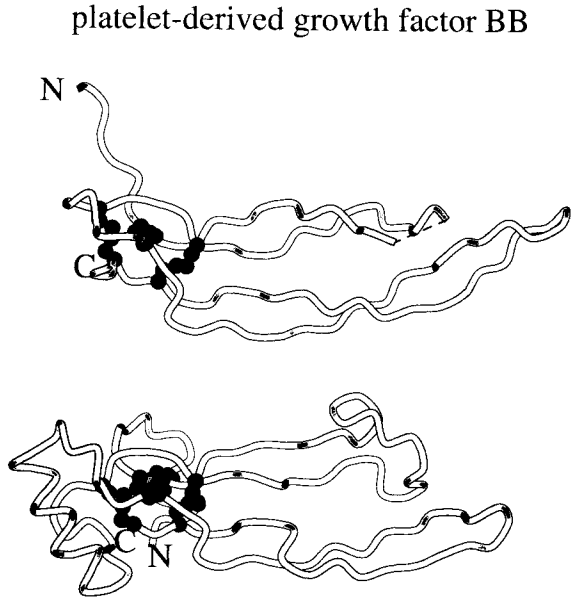

Fig. 2. Cystine-knot fold. The cystine-knot fold lacks a pronounced hydrophobic core. The fold consists of two long, twisted β-hairpins which are held together by three disulphide bridges at the base (black). Conserved half-cystine residues are the principal common sequence feature between known structures with this fold: transforming growth factors β2 (1TGI[42]), nerve growth factor, and platelet-derived growth factor BB (1PDG[43]). Drawn using MolScript.[41]

powerful database search tools that are now applied routinely to any newly deduced protein sequence.[6] This same scenario is now repeating for three-dimensional structures and the solution has the same three essential ingredients: larger databases, more powerful computers, and novel algorithms.

STRUCTURAL ALIGNMENT

Algorithms for structure comparison have to address the full complexity of similarity of shape in 3D space. Early computer methods[7–10] required manual initial alignment or were very slow or limited to close homologues. The last few years have seen the arrival of a new generation of search algorithms that are general, elegant, and/or fast.[11–19] The most efficient of these allow fully automated and rapid similarity searches through the entire database of more than a thousand three-dimensional structures.

The notion of structural equivalence becomes increasingly complex with increasing evolutionary distance. The conformation of a point mutant differs from that of the wildtype protein only locally and only by a few tenths of an Ångström. Much larger deviations are observed in pairs of homologous proteins: with increasing sequence dissimilarity, small shifts in the relative orientations of secondary structure elements accumulate and reach several Ångströms and tens of degrees, as described, e.g., for the globins.[20] At the largest evolutionary distances, only the topology of the fold or folding motif is conserved, i.e., the relative location of helices and strands and the loop connections between these. Deviations can be even larger and qualitatively different when structural similarity is the result of convergent rather than divergent evolution. In particular, convergent evolution may result in similar 3D folds that differ in the topology of loop connections.

Just as there is much latitude in the basic formulation of the structure comparison problem, many different types of optimization algorithm have been employed. These are briefly reviewed here with explicit references to programs that appear in Table I.

DYNAMIC PROGRAMMING ALGORITHMS

Dynamic programming algorithms are a standard tool for finding the global optimum in sequence alignment, a 1D problem. These algorithms have been applied to structure alignment, a 3D problem, by imposing the simplifying assumption that the environment of each individual residue can be adequately summarized in a number or vector or matrix that can then be compared with those of individual residues in the other protein.[16,19,21,22] The optimal alignment maximizes the sum of similarities of pairs of equivalent residues. An intrinsic limitation of the dynamic programming algorithm is that the order of equivalent pairs along the two protein chains must be sequential, i.e., topological permutations are not detected.

In the Stamp[22] and Align[19] programs, comparison is based on the spatial displacement of residue pairs in a 3D superimposition of the structures. The displacements are transformed to similarity scores and an iterative procedure successively refines a nonoptimal initial superimposition, e.g., one generated

TABLE I. Recent Examples of Unexpected Topological Similarities*

Structural class, superfamily, or first instance of a fold	New instance	Available in PDB	Detected immediately	Method of detection
G-protein like fold	cheY	Yes	No	Protep[48]
β-Grasp fold (ubiquitin)	Ferredoxin	Yes	No	Whatlf,[14] Ssap[16]
SH3 domains, R67 dihydrofolate reductase	C-Terminal domain of biotin, repressor	Yes	Yes	Eye[49]
	N-Terminal domain of myosin	No	Yes	Eye[50]
β-Trefoil fold	Hisactophilin	No	Yes	Whatlf[51]
Globins, phycocyanins	Membrane-insertion domain of colicin A	Yes	No	Dali,[52] Ssap[53]
Bacteriophage λ repressor	POU-specific domain	No	Yes	Dali,[38] eye[54]
	Globins	Yes	No	Align,[19] Ssap[53]
βαββαβ motif (muconolactone isomerase, ribosomal protein L7/L12, acylphosphatase etc.)	Histidine-containing phosphocarrier protein	Yes	No	Eye, Ssap[55]
	Ribosomal protein L9	No	Yes	Eye[56]
2nd domain of biotin repressor	SH2-like folding motif	Yes	No	Stamp[57]
	Seryl-tRNA synthetase-like folding motif	Yes	No	Dali
Ribonuclease H-like fold	ATPase subdomain la of actin, hexokinase, hsp70 and glycerol kinase, and connecting domain of HIV reverse transcriptase	Yes	No	Protep[58]
Propeller fold	Bacterial sialidase	Yes	Yes	Eye[59]
Calycins (small hydrophobic ligand binding proteins)	Streptavidin	Yes	No	Eye[36]
Barwin	Endoglucanase V	No	Yes	Eye[60]
Four-helical bundles	C-Terminal domain of kanamycin nucleotidyltransferase	No	Yes	Eye[61]
Transthyretin-like fold	Protocatechuate 3,4-dioxygenase	Yes	No	Dali
TIM (triose phosphate isomerase) barrel	β-Amylase	Yes	Yes	Eye[62]
EGF-like modules	N-Terminal domain of prostaglandin H2 synthase-1	No	Yes	Eye[63]
Fibronectin type III domains	Chloroplast cytochrome f	No	Yes	Eye[64]
Domain D of cyclodextrin glycosyltransferase	3rd domain of galactose oxidase	Yes	No	Dali[65]
Colipase-binding domain of pancreatic lipase	Domain I of arachidonic acid 15-lipoxygenase	No	Yes	Eye[66]
Phosphoglycerate mutase	Acid phosphatase	No	Yes	Eye[67]
Carboxypeptidase A	Leucine aminopeptidase	Yes	No	Protep[68]
Plant lectins	β-Endoglucanase	Yes	Yes	Eye[69]
	Human serum amyloid component P	Yes	Yes	Eye[70]
	Animal S-type lectins	Yes	Yes	Eye[71]
Rat mannose-binding protein (C-type eukaryotic lectin)	N-Terminal domains of the S2 and S3 subunits of pertussis toxin	No	Yes	Eye[72]
Protein R2 of ribonucleotide reductase	Methane monooxygenase	No	Yes	Eye[73]
Hematopoietic cytokines	Interleukin-5	No	Yes	Eye[74]
Oligonucleotide/oligosaccharide binding fold	Staphylococcal nuclease	Yes	No	Eye[75]
	Asp-tRNA synthetase	Yes	No	Eye[75]
	Phage M13 gene 5 protein	Yes	No	Eye[75]
	β-subunits of heat-labile enterotoxin and verotoxin-1	Yes	No	Eye[75]
	Cold-shock protein	No	Yes	Whatlf,[37] eye[76]
	Ribosomal protein S17	No	Yes	Eye[77]
	Toxic shock syndrome toxin-1	No	Yes	Eye[78]
	Subunits S2-S5 of pertussis toxin	No	Yes	Eye[72]

(continued)

TABLE I. Recent Examples of Unexpected Topological Similarities* (Continued)

Structural class, superfamily, *or first instance* of a fold	New instance	Available in PDB	Detected immediately	Method of detection
Toxin-agglutinin fold	Complement regulatory protein CD59	No	Yes	Eye[79]
Astacin, N-terminal domain of thermolysin	Matrix metalloproteinases/collagenases	No	Yes	Eye[80,81]
Sweet-tasting protein monellin	Cystatin family of thiol proteinase inhibitors	Yes	No	Eye[82]
Glutathione peroxidase, thioredoxin, glutaredoxin, glutathione *S*-transferase	Thiol:disulfide interchange protein from *E. coli*	Yes	Yes	Eye[83]
DNA-binding domain of catabolite gene activator protein	DNA-binding domain of biotin repressor	Yes	No	Dali,[28] eye[84]
	LexA repressor	No	Yes	Dali[28]
	Histone H5	Yes	Yes	Eye[85]
	HNF-3/*fork head* DNA binding domain	No	Yes	Eye[84]
	DNA-binding domain of heat shock transcription factor	No	Yes	Eye[86]
DNA polymerase I (Klenow fragment), HIV-1 reverse transcriptase	T7 RNA polymerase	No	Yes	Eye[87]
Thiamin diphosphate-dependent enzymes	Transketolase	Yes	Yes	Eye[88]
	Pyruvate oxidase	Yes	Yes	Eye[88]
	Pyruvate decarboxylase	Yes	Yes	Eye[88]
Short-chain alcohol dehydrogenases (3α,20β-dihydroxysteroid dehydrogenase)	Dihydropteridine reductase	Yes	No	Eye[89]
	UDP-galactose 4-epimerase	Yes	No	Dali, eye[29]
Cytochrome *c* peroxidase, myeloperoxidase	Lignin peroxidase	Yes	Yes	Eye[90]
	Catalytic domain of prostaglandin H2 synthase-1	No	Yes	Eye[63]
Methionine aminopeptidase	Creatine amidinohydrolase	No	No	Eye[91]
Animal and phage lysozymes	Plant endochitinase	Yes	No	Dali[30]
	Soluble lytic transglycosylase from *E. coli*	No	Yes	Eye[31]

*The compilation includes similarities detected by eye (published between March 1993 and March 1994) or by automated database searches.

according to sequence alignment. The Ssap program[16] elegantly eliminates the need for an explicit initial alignment by using dynamic programming at two levels. First, the structural environments of each residue pair are compared in a coordinate frame defined by the trial superimposition of their backbone atoms. The optimal alignment traces are accumulated in a master matrix from which the final alignment is derived. The idea is that only residue pairs with high scores in many trial superimpositions have a chance of being part of the best global match. The original Ssap algorithm[9] was later improved by various filters to increase speed and reduce background noise.[16]

DISTANCE MATRIX ALGORITHMS

The use of intramolecular geometric relationships, such as distances, to describe protein structures has the advantage of being independent of the coordinate frame.[12,13,14,18,21,23,24] The Protep program[13] reduces protein structure to secondary structure elements, i.e., helices and strands, and their geometric relations. The user constructs a search pattern that defines the allowed range of angles and distances between the elements of a structural motif. The complete set of perfect matches to the search pattern can be rapidly retrieved from a prestored database (a single violation is sufficient to discard a potential match).

Several other methods use a more detailed description of internal geometry, in the form of C^{α}–C^{α} distance matrices.[23,24,18] Instead of search patterns with sharp cutoffs, many of these algorithms use a continuous geometric similarity score summed over all equivalenced intramolecular distances. Optimization of this score as a function of residue equivalences cannot be solved by dynamic programming, as residues as well as all their intramolecular part-

ners have to be equivalenced in an overall consistent fashion. One way out is provided by Monte Carlo algorithms,[21,23] which efficiently explore the complex search space—although they are not guaranteed to locate the global optimum. The Dali program[14] builds up an optimal alignment using Monte Carlo optimization to combine pairs of matching fragments (matching submatrices of the distance matrices) into larger consistent sets of pairs. In our experience, the method is both sensitive and accurate, although, in its current implementation, not as fast as some pattern matching or 3D clustering algorithms.

3D CLUSTERING ALGORITHMS

Historically, the common structural core is often defined as a self-consistent set of pair equivalences with positional deviations below a user-defined threshold. Finding the largest possible common core that satisfies this condition is a complicated search problem. The problem can be partitioned into subproblems of finding pairs of non-gapped matching fragments, e.g., strands and helices. Consistent sets of fragment pairs can then be assembled using clustering methods.[11,15] The clustering method in Whatif[14] allows the detection of structural correspondences where the sequential order of equivalent fragments is not preserved (also true for distance matrix methods). In practice, however, several adjustable cutoffs are needed to maintain speed while allowing flexibility in the relative orientations and positions of equivalent fragments.

RECURRENT FOLDS

These efforts at algorithm development are already beginning to yield interesting results. Using the new computer methods for scanning structural databases, a number of nontrivial similarities have been detected some time after the coordinates of the structures were made available (Table I). Nontrivial in this context means similarities that are not easily found by sequence comparison alone. As more and more resemblances and remote evolutionary connections between new and old protein structures are discovered, the growth curve for unique folds rises much less steeply than that for the total number of known structures. The computer methods for structural alignment will be especially useful in creating and developing a classification of the building blocks of protein structure.[18,25,26]

Two effects limit the total variety of observed protein folds: physical principles and evolutionary history of natural proteins. Sequence database searches are a powerful tool in molecular biology because inferences concerning protein structure and function exploit the biological fact that these can be retained over large evolutionary distances. Structure database searches expand the realm of comparative analysis, and occasionally simplify biochemical classifications by unifying two or more protein families into one superfamily.[27–29] We illustrate this point with two examples of bio(techno)logical interest.

FROM STRUCTURAL SIMILARITY TO BIOLOGICAL FUNCTION

The first example sheds some light on the structure of barley endochitinase, an enzyme involved in plant defense reactions against chitin-containing pests (fungi and insects). Several sequences of the barley endochitinase family are known, but site-directed mutagenesis has so far failed to identify the active site. As the result of a database search with the structure of endochitinase, we identified a subtle but unambiguous similarity to lysozymes from animals and phage.[30] The structures of three remotely related sublcasses of lysozyme have been known for some time. Endochitinase shares with them a structural core of four helices and a small 3-stranded β-sheet. These elements are arranged in topologically identical order in spite of massive peripheral insertions/deletions (Fig. 3). Significantly, the location and composition of the active site and key structural residues, as seen in lysozymes, are also conserved in endochitinase. Recently, the structure determination of a bacterial muramidase has revealed a fifth subclass of lysozymes.[31] Much of the knowledge about lysozyme can be extended by analogy to these new members and used to guide experiment.

DESIGN OF A VERSATILE BINDING MOTIF

Sometimes biological intuition is ambivalent on the question of evolutionary divergence vs. physical convergence. The second example illustrates this conceptual difficulty in protein taxonomy. Streptavidin, known for its very high affinity for biotin,[32] has a similar antiparallel, up-and-down β-barrel topology as a diverse set of tissue-specific carrier proteins that selectively bind a variety of small hydrophobic ligands, called calycins.[33] The two subclasses of calycins (one 8-stranded such as retinol binding protein,[34] the other 10-stranded such as P2 myelin protein[35]) have in common a conserved cluster of hydrophobic residues around an invariant tryptophan that ties together the first and last strands of the β-barrel (Fig. 4). Exactly the same packing pattern is seen in streptavidin, an argument in favor of an evolutionary relationship.[36] Two other proteins with similar topologies, catalase and photoactive yellow protein, do not have this packing pattern. The key argument against an evolutionary relationship is based on architectural differences, notably in shear number s (streptavidin $s = 10$; retinol binding protein $s = 12$). In other words, a helical path going around the barrel following the β-sheet hydrogen bonds ends on the starting strand two residues farther in retinol binding protein. In addition, the β-barrel of streptavidin has no central cavity and

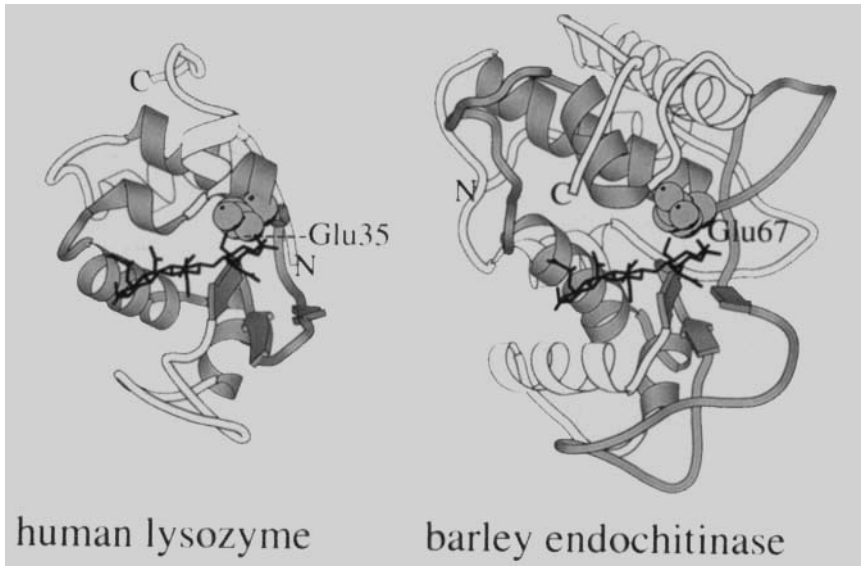

Fig. 3. Active site revealed by structural similarity between plant endochitinase and animal and phage lysozymes. The ribbon diagram[41] highlights structurally equivalent elements of the common core (shaded) between animal and phage lysozymes and barley endochitinase. Regions outside the common structural core are white. The active site of lysozymes is known but that of endochitinase is yet biochemically uncharacterized. A trisaccharide inhibitor of lysozyme (black) was transferred from the cocrystal with human lysozyme (9LYZ) to an equivalent position in the structure of endochitinase. Glu-35 is the principal catalytic residue in human lysozyme (1LHM[44]). By analogy, the structurally equivalent Glu-67 is proposed to be the catalytic residue of barley endochitinase (1BAA[45]).

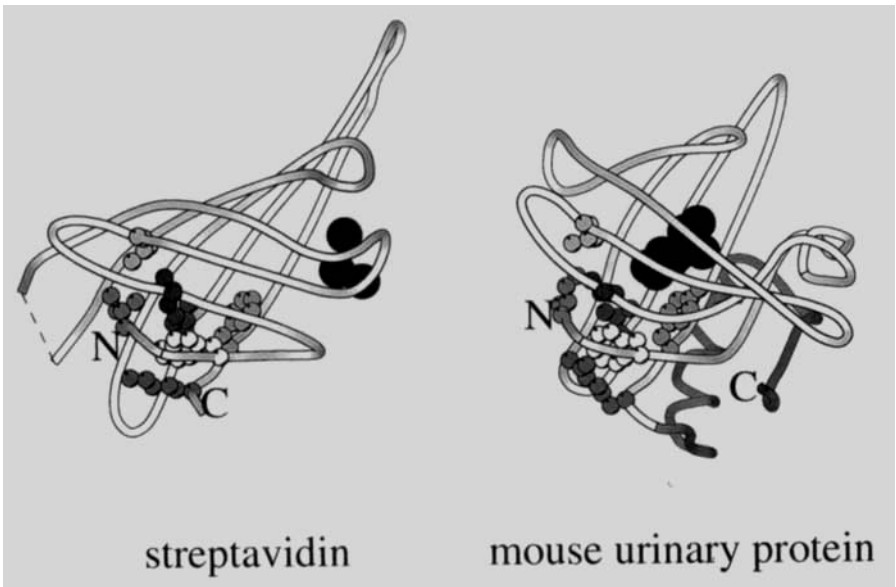

Fig. 4. Is streptavidin distantly related to calycins? Comparison of streptavidin (1PTS[46]) and a member of the calycin superfamily (major urinary protein, 1MUP[47]). The common core[18] consists of residues A13–A25, A27–A36, A38–A44, A51–A58, A76–A79, A88–A91, A93–A98, A101–A111, A121–A133 in 1PTS and 15–27, 45–61, 68–75, 86–93, 96–112, 115–127 in 1MUP, with a positional root mean square deviation of C^{α} atoms of 2.9 Å over 76 residues. Side chains are shown for a conserved cluster of hydrophobic residues (bold in alignment below, β-strands underlined). The ligand (black) is sequestered inside the barrel in lipocalins and bound by loops in streptavidin. Drawn using MolScript.[41]

```
18 KINGEWHT  49 IHVLE-NSL  103 NFLMA  120 LMGLYGRE  major urinary protein
22 NYHGKWWE  48 YTPEG-KSV  112 NYIIG  131 FVWVLSRS  bilin binding protein
 1 AFDGTWKV  40 ITQEG-NKF  111 NELIQ  124 AKRIFKKE  fatty-acid binding protein
19 RFSGTWYA  45 FSVDETGQM  113 TYAVQ  133 YSFVFSRD  retinol binding protein
16 GITGTWYN  31 VTAGADGAL  102 ARINT  126 GHDTFTKV  streptavidin
```

binds the ligand by loops, whereas the barrel of retinol binding protein is wider and more ellipsoidal and sequesters the ligand deep inside the barrel. Classification issues aside, the packing motif of calycin–streptavidin indisputably has been an evolutionary success.

MORE TO COME

The importance of structure comparison will continually increase with the rapid growth of the pool of known structures. At some time in the future, an important part of the protein folding problem will effectively evaporate as structure and sequence comparison tools classify any new protein into an existing family.

We propose that any newly solved protein structure be routinely compared with those in the Protein Data Bank, for the detection of possible topological resemblances. The software required to perform rapid and rigorous searches of structural databases is now sufficiently mature[14,16–18] and generally accessible to the scientific community.* The experience of the past year (Table I) shows that the chances of finding interesting similarities in the database are greater than those of a fold being unique. The recognition of a distant relationship to a well-characterized protein may reveal possible functional analogies that in turn may lead to considerable time savings in the biochemists' laboratory. The unexpected similarities between the cold-shock protein and several oligonucleotide-binding proteins[37] and between the POU-specific domain and λ repressor,[38] both identified by automated database searches upon completion of the structure determinations, are motivating examples. Searching protein structure databases has come of age.

ACKNOWLEDGMENTS

We thank M.B. Swindells and A.G. Murzin for interesting discussions on the importance of protein structure database searches, and G. Vriend, G. Tuparev, and W.R. Taylor for the comparison of structure alignment algorithms. Financial support was from the EC Bridge program.

REFERENCES

1. Kendrew, J.C., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., Shore, V.C. Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London) 185:422–427, 1960.
2. Murzin, A.G. Familiar strangers. Nature (London) 360: 635, 1992.
3. Matthews, D.A., Smith, S.L., Baccanari, D.P., Burchall, J.J., Oatley, S.J., Kraut, J. Crystal structure of a novel trimethoprim-resistant dihydrofolate reductase specified in *Escherichia coli* by R-plasmid R67. Biochemistry 25: 4194–4204, 1986.
4. Swindells, M.B., Daopin, S., Cohen, G.H., Davies, D. Structural similarity between transforming growth factor beta-2 and nerve growth factor. Science 258:1160–1162, 1992.
5. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
6. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. Protein Sci. 1:1677–1690, 1992.
7. Rossmann, M.G., Argos, P. Exploring structural homology of proteins. J. Mol. Biol. 105:75–95, 1976.
8. Matthews, B.W., Rossmann, M.G. Comparison of protein structures. Meth. Enzymol. 115:397–420, 1985.
9. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
10. Zuker, M., Somorjai, R.L. The alignment of protein structures in three dimensions. Bull. Math. Biol. 51:55–78, 1989.
11. 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. 9:769–789, 1992.
12. Kleywegt, G., Jones, T. 3rd European Workshop on Crystallography of Biological Molecules, p. M2, Como, Italy, May 1993.
13. Mitchell, E.M., Artymiuk, P.J., Rice, D.W., Willett, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J. Mol. Biol. 212:151–166, 1990.
14. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. Proteins 11:52–58, 1991.
15. Alexandrov, N.N., Takahashi, K., Go, N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. J. Mol. Biol. 225:5–9, 1992.
16. Orengo, C.A., Brown, N.P., Taylor, W.T. Fast structure alignment for protein databank searching. Proteins 14: 139–167, 1992.
17. Grindley, H.M., Artymiuk, P.J., Rice, D.W., Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J. Mol. Biol. 229:707–721, 1993.
18. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
19. Subbiah, S., Laurents, D.V., Levitt, M. Structural similarity of DNA-binding domains of bacterophage repressors and the globin core. Current Biol. 3:141–148, 1993.
20. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225–270, 1980.
21. Sali, A., Blundell, T.L. Definition of general topological equivalence in protein structures. J. Mol. Biol. 212:403–428, 1990.
22. Russell, R.B., Barton, G.J. Multiple protein sequence alignment from tertiary structure: Assigment of global and residue confidence levels. Proteins 14:309–323, 1992.
23. Barakat, M.T., Dean, P.M. Molecular structure matching by simulated annealing. III. The incorporation of null correspondences into the matching problem. J. Computer-Aided Mol. Design 5:107–117, 1991.
24. Subbarao, N., Haneef, I. Defining topological equivalences in macromolecules. Protein Engin. 4:877–884, 1991.
25. Holm, L., Ouzounis, C., Sander, C., Tuparev, G., Vriend, G. A database of protein structure families with common folding motifs. Prot. Sci. 1:1691–1698, 1992.
26. Orengo, C.A., Flores, T.P., Taylor, W.R., Thornton, J.M. Identification and classification of protein fold families. Prot. Eng. 6:485–500, 1993.
27. Bork, P., Sander, C., Valencia, A. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc. Natl. Acad. Sci. U.S.A. 89:7290–7294, 1992.

*In particular, the authors offer to assist those with a new structure in attempts at discovering similarities. Send C^{α} coordinates by e-mail to holm@embl-heidelberg.de. A list of structurally similar proteins, aligned in 3D using Dali, will be returned.

28. Holm, L., Sander, C., Cox, M., Fogh, R., Boelens, R., Vliet, P.C. v.d., Schnarr, M., Rüterjans, H., Kaptein, R. In preparation.
29. Holm, L., Murzin, A.G., Sander, C. Three sisters, different names: 3α,20β-Dihydroxysteroid dehydrogenase, dihydropteridine reductase and UDP-galactose 4-epimerase. Nature Struct. Biol. 1:146–147, 1994.
30. Holm, L., Sander, C. Structural similarity between plant endochitinase and lysozymes from animals and phage: An evolutionary connection. FEBS Lett. 340:129–132, 1994.
31. Thunnissen, A.-M.W.H., Dijkstra, A., Kalk, K., Rozeboom, H., Engel, H., Keck, W., Dijkstra, B. Doughnut-shaped structure of a bacterial muramidase revealed by X-ray crystallography. Nature (London) 367:750–753, 1994.
32. Bauer, A.J., Rayment, I., Frey, P.A., Holden, H.M. The molecular structure of UDP-galactose 4-epimerase from *Escherichia coli* determined at 2.5 Å resolution. Proteins 12:372–381, 1992.
33. Flower, D.R., North, A.C.T., Attwood, T.K. Structure and sequence relationships in the lipocalins and related proteins. Prot. Sci. 2:753–761, 1993.
34. Cowan, S.W., Newcomer, M.E., Jones, T.A. Crystallographic refinement of human serum retinol binding protein at 2 Å resolution. Proteins 8:44–61, 1990.
35. Cowan, S.W., Newcomer, M.E., Jones, T.A. Crystallographic studies on a family of cellular lipophilic transport proteins. Refinement of P2 myelin protein and the structure determination and refinement of cellular retinool-binding protein in complex with all-trans retinol. J. Mol. Biol. 230:1225–1246, 1993.
36. Flower, D.R. Structural relationship of streptavidin to the calycin protein superfamily. FEBS Lett. 333:99–102, 1993.
37. Schnuchel, A., Wiltscheck, R., Czisch, M., Herrier, M., Willimsky, G., Graumann, P., Marahiel, M.A., Holak, T.A. Structure in solution of the major cold-shock protein from *Bacillus subtilis*. Nature (London) 364:169–171, 1993.
38. Dekker, N., Cox, M., Boelens, R., Verrijzer, C.P., Vliet, P.C. v.d., Kaptein, R. Solution structure of the POU-specific DNA-binding domain of Oct-1. Nature (London) 362: 852–855, 1993.
39. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., Saraste, M. Crystal structure of a Src-homology 3 (SH3) domain. Nature (London) 359:851–855, 1992.
40. Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J., Matthews, B.M. The E. coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin and DNA-binding domains. Proc. Natl. Acad. Sci. U.S.A. 89:9257–9261, 1992.
41. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
42. Kaopin, S., Piez, K., Ogawa, Y., Davies, D. Crystal structure of transforming growth factor-beta2: an unusual fold for the superfamily. Science 257:369–372, 1992.
43. Oefner, C., D'Arcy, A., Winkler, F., Eggimann, B., Hosang, M. Crystal structure of human platelet-derived growth factor BB. EMBO J. 11:3921–3926, 1992.
44. Inaka, K., Taniyama, Y., Kikuchi, M., Morikawa, K., Matsushima, M. The crystal structure of a mutant lysozyme C77/95A with increased secretion efficiency in yeast. J. Biol. Chem. 266:12599–12603, 1991.
45. Hart, P., Monzingo, A., Ready, M., Ernst, S., Robertus, J. Crystal structure of an endochitinase from *Hordeum vulgare L.* seeds. J. Mol. Biol. 229:189–193, 1993.
46. Weber, P.C., Pantoliano, M.W., Thompson, L.D. Crystal structure and ligand-binding studies of a screened peptide complexed with streptavidin. Biochem. 31:9350, 1992.
47. Bocskei, Z., Groom, C.R., Flower, D.R., Wright, C.E., Phillips, S.E., Cavaggioni, A., Findlay, J.B., North, A.C. Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. Nature (London) 360:186–188, 1992.
48. Artymiuk, P.J., Rice, D.W., Mitchell, E.M., Willett, P. Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph theoretical techniques. Prot. Engin. 4:39–43, 1990.
49. Noble, M., Musacchio, A., Saraste, M., Courtneidge, S., Wierenga, R. Crystal structure of the SH3 domain in human Fyn: comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectin. EMBO J. 12:2617–2624, 1993.
50. Rayment, I., Rypniewski, W.R., Schmidt-Bäse, K., Smith, R., Tomchik, D.R., Benning, M.M., Winkelmann, D.A., Wesenberg, G., Holden, H.M. Three-dimensional structure of myosin subfragment-1: A molecular motor. Science 261: 50–58, 1993.
51. Hazabettl, J., Gondol, D., Wiltscheck, R., Otlewski, J., Schleicher, M., Holak, T.A. Structure of hisactophilin is similar to interleukin-1beta and fibroblast growth factor. Nature (London) 359:855–858, 1992.
52. Holm, L., Sander, C. Globin fold in a bacterial toxin. Nature (London) 361:309, 1993.
53. Orengo, C.A., Flores, T.P., Jones, D.T., Taylor, W.R., Thornton, J.M. Recurring structural motifs in proteins with different functions. Current Biol. 3:131–139, 1993.
54. Assa-Munt, N., Mortimore-Smith, R.J., Aurora, R., Herr, W., Wright, P.E. The solution structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage lambda repressor DNA-binding domain. Cell 73: 193–205, 1993.
55. Swindells, M.B., Orengo, C.A., Jones, D.T., Pearl, L.H., Thornton, J.M. Recurrence of a binding motif? Nature (London) 362:299, 1993.
56. Hoffman, D., Davies, C., Gerchman, S., Kycia, J., Porter, S., White, S., Ramakrishnan, V. Crystal structure of prokaryotic ribosomal protein L9: A bilobal RNA-binding protein. EMBO J. 13:205–212, 1994.
57. Russell, R.B., Barton, G.J. An SH2-SH3 domain hybrid. Nature (London) 364:765, 1993.
58. Artymiuk, P.J., Grindley, H.M., Kumar, K., Rice, D.W., Willett, D.W. Three-dimensional structural resemblance between the ribonuclease H and connection domain of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. FEBS Lett. 324:15–21, 1993.
59. Crennell, S.J., Farman, E.F., Laver, W.G., Vimr, E.R., Taylor, G.L. Crystal structure of a bacterial sialidase (from *Salmonella typhimurium* LT2) shows the same fold as an influenza virus neuraminidase. Proc. Natl. Acad. Sci. U.S.A. 90:9852–9856, 1993.
60. Davies, G.J., Dodson, G.G., Hubbard, R.E., Tolley, S.P., Dauter, Z., Wilson, K.S., Hjort, C., Mikkelsen, J.M., Rasmussen, G., Schülein, M. Structure and function of endoglucanase V. Nature (London) 365:362–364, 1993.
61. Sakon, J., Liao, H., Kanikula, A., Benning, M., Rayment, I., Holden, H. Molecular structure of kanamycin nucleotidyltransferase determined to 3.0 Å resolution. Biochemistry 32:11977–11984, 1993.
62. Mikami, B., Hehre, E., Sato, M., Katsube, Y., Hirose, M., Morita, Y., Sacchettini, J. The 2.0-A resolution structure of soybean beta-amylase complexes with alpha-cyclodextrin. Biochemistry 32:6836–6845, 1993.
63. Picot, D., Loll, P., Garavito, M. The X-ray structure of the membrane protein prostaglandin H2 synthase-1. Nature (London) 367:243–249, 1994.
64. Martinez, S., Huang, D., Szczepaniak, A., Cramer, W., Smith, J. Crystal structure of chloroplast cytochrome f reveals a novel cytochrome fold and unexpected heme ligation. Structure 2:95–105, 1994.
65. Bork, P., Holm, L., Sander, C. Sequence and structural comparison of immunoglobulin-like domains. In preparation.
66. Boyington, J.C., Gaffney, B.J., Amzel, L.M. The three-dimensional structure of an arachidonic acid 15-lipoxygenase. Science 260:1482–1486, 1993.
67. Schneider, G., Lindqvist, Y., Vihko, P. Three-dimensional structure of rat acid phosphatase. EMBO J. 12:2609–2615, 1993.
68. Artymiuk, P., Grindley, H., Park, J., Rice, D., Willett, P. Three-dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph-theoretical techniques. FEBS Lett. 303:48–52, 1992.
69. Keitel, T., Simon, O., Borriss, R., Heinemann, U. Molecular and active-site structure of a *Bacillus* 1,3-1,4-betaglucanase. Proc. Natl. Acad. Sci. U.S.A. 90:5287–5291, 1993.
70. Emsley, J., White, H., O'Hara, B., Oliva, G., Srinivasan, N., Tickle, I., Blundell, T., Pepys, M., Wood, S. Structure of pentameric human serum amyloid P component. Nature (London) 367:338–345, 1994.

71. Liao, D.-I., Kapadia, G., Ahmed, H., Vasta, G.R., Herzberg, O. Structure of S-lectin, a developmentally regulated vertebrate beta-galactoside binding protein. Proc. Natl. Acad. Sci. U.S.A. 91:1428–1432, 1994.
72. Stein, P., Boodhoo, A., Armstrong, G., Cockle, S., Klein, M., Read, R. The crystal structure of pertussis toxin. Structure 2:45–57, 1994.
73. Rosenzweig, A., Frederick, C., Lippard, S., Nordlund, P. Crystal structure of a bacterial non-haem iron hydroxylase that catalyzes the biological oxidation of methane. Nature (London) 366:537–543, 1993.
74. Milburn, M., Hassel, A.M., Lambert, M.H., Jordan, S.R., Proudfood, A.E.I., Graber, P., Wells, T.N.S. A novel dimer configuration revealed by the crystal structure at 2.4 A resolution of human interleukin-5. Nature (London) 363: 172–176, 1993.
75. Murzin, A.G. OB(oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. EMBO J. 12:861–867, 1993.
76. Schindelin, H., Marahiel, M.A., Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein. Nature (London) 364:164–168, 1993.
77. Golden, B., Hoffmann, D., Ramakrishnan, V., White, W. Ribosomal protein S17: Characterization of the three-dimensional structure by 1H and 15N NMR. Biochemistry 32:12812–12820, 1993.
78. Acharya, K., Passalacqua, E., Jones, E., Harlos, K., Stuart, D., Brehm, R., Tranter, H. Structural basis of superantigen action inferred from crystal structure of toxic-shock syndrome toxin-1. Nature (London) 367:94–97, 1994.
79. Fletcher, C., Harrison, R., Lachmann, P., Neuhaus, D. Structure of a soluble, glycosylated form of the human complement regulatory protein CD59. Structure 2:185–199, 1994.
80. Gomis-Rüth, F.-X., Kress, L.F., Bode, W. First structure of a snake venom metalloproteinase: A prototype for matrix metalloproteinases/collagenases. EMBO J. 12:4151–4157, 1993.
81. Lovejoy, B., Cleasby, A., Hassell, A., Longley, K., Luther, M., Weigl, D., McGeehan, G., McElroy, A., Drewry, D., Lambert, M. et al. Structure of the catalytic domain of fibroblast collagenase complexed with an inhibitor. Science 263:375–377, 1994.
82. Murzin, A.G. Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors. J. Mol. Biol. 230:689–694, 1993.
83. Martin, J.L., Bardwell, J.C.A., Kuriyan, J. Crystal structure of the DsbA protein required for disulphide bond formation *in vivo*. Nature (London) 365:464–467, 1993.
84. Clark, K.L., Halay, E.D., Lai, E., Burley, S.K. Co-crystal structure of the HNF-3/*fork head* DNA-recognition motif resembles histone H5. Nature (London) 364:412–420, 1993.
85. Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L., Sweet, R.M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature (London) 362:219–223, 1993.
86. Harrison, C.J., Bohm, A.A., Nelson, H.C.M. Crystal structure of the DNA binding domain of the heat shock transcription factor. Science 263:224–227, 1994.
87. Sousa, R., Chung, Y.J., Rose, J.P., Wang, B.-C. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. Nature (London) 364:593–399, 1993.
88. Muller, Y.A., Lindqvist, Y., Furey, W., Schulz, G.E., Jordan, F., Schneider, G. A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. Structure 1:95–103, 1993.
89. Krook, M., Ghosh, D., Strömberg, R., Carlquist, M., Jörnvall, H. Carboxyethyllysine in a protein: Native carbonyl reductase/NADP+-dependent prostaglandin dehydrogenase. Proc. Natl. Acad. Sci. U.S.A. 90:502–506, 1993.
90. Poulos, T., Edwards, S., Wariishi, H., Gold, M. Crystallographic refinement of lignin peroxidase at 2 Angstroms. To be published 1994.
91. Murzin, A.G. Can homologous proteins evolve different enzymatic activities? TIBS 18:403–405, 1993.