Searching Protein Structure Databases Has Come of Age

Lisa Holm
Chris Sander
Protein Design Group
European Molecular Biology Laboratory
D-69012 Heidelberg
Federal Republic of Germany

Abstract

The number of protein structures known in atomic detail has increased from one in 1960 [1] 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.

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 beta-barrel with identical topography to that in the SH3 domain (Figure 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 dihydrololate 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 (Figure 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 week, two or three 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 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] [12-19]. 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 unfolding 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 1.
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 that of individual residues in the other protein[21, 16, 22, 19]. The optimal alignment maximizes the sum of similarities of pairs of equivalent residues. An intrinsic limitation of the dynamic programming algorithms is that the order of such pairs along the two protein chains must be sequential, i.e., topological permutations are not detected.

In the Smith[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 an non-optimal initial superimposition, e.g., one generated according to sequence alignment. The Soap program[10] 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 Soap algorithm[9] was later improved by various filters to increase speed and reduce background noise[16].

Distance matrix algorithms

The use of intramolecular geometrical relationships, such as distances, to describe protein structures has the advantage of being independent of the coordinate frame[12, 13, 21, 23, 24, 14, 18]. The Protep program[13] reduces protein structure to secondary structure elements, i.e., helices and strands, and their geometrical 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 Cα-Cβ, Cβ-Cα distance matrices[23, 24, 18]. Instead of search patterns with sharp cutoffs, many of these algorithms use a continuous geometrical distance 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 partners 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[18] 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[15, 11]. The clustering method in Whatif[14] allows the detection of structural correspondences where the sequential order of equivalent fragments is not preserved, as can 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

The common structural core has been attacked at algorithm development are already beginning to yield interesting results. Using the new computer methods for scanning structural databases, a number of non-trivial similarities have been detected some time after the coordinates of the structures were made available (Table I). Non-trivial 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 more 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[25, 18, 26].

The two effects limit the total variety of observed protein folds: physical principles and the evolutionary history of natural proteins. Sequence database searches are a powerful tool in molecular biology because the sequence of a protein is an intrinsic property that is 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 more 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[25, 18, 26].

Two effects limit the total variety of observed protein folds: physical principles and the evolutionary history of natural proteins. Sequence database searches are a powerful tool in molecular biology because the sequence of a protein is an intrinsic property that is 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 more 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[25, 18, 26].

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[50]. The structures of three remotely related subclasses of lysozyme have been known for some time. Endochitinase shares with them a structural core of four helices and a small 3-stranded beta–sheet. These elements are arranged in topologically identical order in spite of massive peripheral insertions/deletions (Figure 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[51]. Much of the knowledge about lysozyme can be extended by analogy to these new members 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 beta–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, the other 10-stranded such as P2 myelin protein[15, 11]) have in common a conserved cluster of hydrophobic residues and an invariant tryptophan that ties together the first and last strands of the beta–barrel (Figure 4). Exactly the same packing pattern is seen in streptavidin, an argument in favor of an evolutionary relationship[30]. 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 (streptavidin = 10; retinol binding protein = 12). In other words, a helical path going around the barrel following the beta–sheet hydrogen bonds ends on the starting strand two residues farther in retinol binding protein. In addition, the beta–barrel of streptavidin has no central cavity and 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, 16–19] and generally accessible to the scientific community[4]. 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 biochemist’s laboratory. The unexpected similarities between the cold-shock protein and several oligonucleotide-binding proteins[37] and between the POLI-specific domain and lambda-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

http://biotech.dibms.unina.it/corsi/proteins_review.htm#RTFToC3

Pagina 2 di 7
of structure alignment algorithms. Financial support was from the EC Bridge program.

References


Figures

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].

2: Cystine-knot fold.

The cystine-knot fold lacks a pronounced hydrophobic core. The fold consists of two long, twisted beta--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 beta-2 (1TGI), nerve growth factor and transforming growth factors beta (1LG1). The cystine-knot fold lacks a pronounced hydrophobic core. The fold consists of two long, twisted beta--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 beta-2 (1TGI), nerve growth factor and transforming growth factors beta (1LG1).

Is streptavidin distantly related to calycins ?

Comparison of streptavidin (1PTS) and a member of the calycin superfamily (major urinary protein, 1MUP). The common core [18] consists of residues A13-A25, A27-A36, A38-A44, A51-A58, A76-A79, A88-A91, 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 Cα atoms of 2.9 Å over 76 residues. Side chains are shown for a conserved cluster of hydrophobic residues (bold in alignment below, beta--strands underlined). The ligand (black) is sequestered
inside the barrel in lipocalins and bound by loops in streptavidin. Drawn using MolScript [41].

22 NYHGKWKE 48 YTPEG-KSV 112 NYIG 131 FVVL 51 S bilin binding protein

1 AFUGTWK 40 IQQEG-NK 111 NELIQ 124 AKRIFKE fatty-acid binding protein

19 RFSGLTWY 45 FSVDETGQM 113 TYAVQ 133 YSFVFSRD retinol binding protein

16 GITGFWY 31 VTAGADGL 102 ARN 126 GHTDFIK streptavidin

### Table I: Recent examples of unexpected topological similarities

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

<table>
<thead>
<tr>
<th>structural class, superfamily,</th>
<th>new instance</th>
<th>new detec-</th>
<th>method of detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>or first instance of a fold</td>
<td>in PDB</td>
<td>ted im-</td>
<td>dihydrofolate reductase</td>
</tr>
<tr>
<td>C-terminal domain of biotin of myosin</td>
<td>yes</td>
<td>yes</td>
<td>eye[49]</td>
</tr>
<tr>
<td>C-terminal domain of biotin of myosin</td>
<td>yes</td>
<td>yes</td>
<td>eye[49]</td>
</tr>
<tr>
<td>C-terminal domain of biotin of myosin</td>
<td>yes</td>
<td>yes</td>
<td>eye[49]</td>
</tr>
<tr>
<td>C-terminal domain of biotin of myosin</td>
<td>yes</td>
<td>yes</td>
<td>eye[49]</td>
</tr>
<tr>
<td>N-terminal domain of arachidonic acid 15-lipoxygenase</td>
<td>no</td>
<td>yes</td>
<td>eye[66]</td>
</tr>
<tr>
<td>N-terminal domain of arachidonic acid 15-lipoxygenase</td>
<td>no</td>
<td>yes</td>
<td>eye[66]</td>
</tr>
<tr>
<td>N-terminal domain of arachidonic acid 15-lipoxygenase</td>
<td>no</td>
<td>yes</td>
<td>eye[66]</td>
</tr>
<tr>
<td>N-terminal domain of arachidonic acid 15-lipoxygenase</td>
<td>no</td>
<td>yes</td>
<td>eye[66]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>N-terminal domains of the S2 and S3 subunits of pertussis toxin</td>
<td>no</td>
<td>yes</td>
<td>eye[72]</td>
</tr>
<tr>
<td>Protein</td>
<td>Yes</td>
<td>No</td>
<td>Eye</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----</td>
<td>----</td>
<td>------</td>
</tr>
<tr>
<td>Oligonucleotide/oligosaccharide binding fold</td>
<td>Asp-RNA synthetase</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Staphylococcal nuclease</td>
<td>Yes</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>Phage M13 gene 5 protein</td>
<td>Yes</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>B-subunits of heat-labile enterotoxin and verotoxin-1</td>
<td>Yes</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>Cold-shock protein</td>
<td>No</td>
<td>Yes</td>
<td>76</td>
</tr>
<tr>
<td>Ribosomal protein S17</td>
<td>No</td>
<td>Yes</td>
<td>77</td>
</tr>
<tr>
<td>Toxic shock syndrome toxin-1</td>
<td>No</td>
<td>Yes</td>
<td>78</td>
</tr>
<tr>
<td>Subunits S2-S5 of pertussis toxin</td>
<td>No</td>
<td>Yes</td>
<td>72</td>
</tr>
<tr>
<td>Toxin-agglutinin fold</td>
<td>Complement regulatory protein CD59</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Astacin, N-terminal domain of thermolysin</td>
<td>Matrix metalloproteinases / collagenases</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sweet-tasting protein monellin</td>
<td>Cystatin family of thiol proteinase inhibitors</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Glutathione peroxidase, thioredoxin, glutaredoxin, glutathione S-transferase</td>
<td>Thioldisulphide interchange protein from E. coli</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DNA-binding domain of catabolite gene activator protein</td>
<td>DNA-binding domain of biotin repressor lexA repressor</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Histone H5 HNF-3/fork head DNA binding domain</td>
<td>Yes</td>
<td>Yes</td>
<td>85</td>
</tr>
<tr>
<td>DNA-binding domain of heat shock transcription factor</td>
<td>No</td>
<td>Yes</td>
<td>84</td>
</tr>
<tr>
<td>DNA polymerase I (Klenow fragment), HIV-1 reverse transcriptase</td>
<td>T7 RNA polymerase</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Thiamin</td>
<td>Transketolase pyruvate</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Diphosphate-dependent enzymes</td>
<td>Oxidase pyruvate</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Decarboxylase</td>
<td>No</td>
<td>Yes</td>
<td>88</td>
</tr>
<tr>
<td>Short-chain alcohol dehydrogenases (3alpha-,20beta-dihydroxy steroid dehydrogenase)</td>
<td>Dihydropteridine reductase</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lignin peroxidase, catalytic domain of prostaglandin H2 synthase-1</td>
<td>Yes</td>
<td>Yes</td>
<td>90</td>
</tr>
<tr>
<td>Methionine aminopeptidase</td>
<td>No</td>
<td>No</td>
<td>91</td>
</tr>
<tr>
<td>Animal and phage lysozymes</td>
<td>Plant endochitinase soluble lytic transglycosylase from E. coli</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Glutathione peroxidase</td>
<td>Yes</td>
<td>Yes</td>
<td>92</td>
</tr>
<tr>
<td>Thioredoxin</td>
<td>Yes</td>
<td>Yes</td>
<td>93</td>
</tr>
<tr>
<td>Glutaredoxin</td>
<td>Yes</td>
<td>Yes</td>
<td>94</td>
</tr>
<tr>
<td>Glutathione S-transferase</td>
<td>DNA-binding domain of catabolite gene activator protein</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DNA-binding domain of heat shock transcription factor</td>
<td>Yes</td>
<td>Yes</td>
<td>86</td>
</tr>
<tr>
<td>DNA polymerase I (Klenow fragment), HIV-1 reverse transcriptase</td>
<td>Yes</td>
<td>Yes</td>
<td>87</td>
</tr>
<tr>
<td>Thiamin</td>
<td>Yes</td>
<td>Yes</td>
<td>88</td>
</tr>
<tr>
<td>Diphosphate-dependent enzymes</td>
<td>Yes</td>
<td>Yes</td>
<td>88</td>
</tr>
<tr>
<td>Decarboxylase</td>
<td>Yes</td>
<td>Yes</td>
<td>88</td>
</tr>
<tr>
<td>Short-chain alcohol dehydrogenases (3alpha-,20beta-dihydroxy steroid dehydrogenase)</td>
<td>Yes</td>
<td>Yes</td>
<td>89</td>
</tr>
<tr>
<td>Lignin peroxidase, catalytic domain of prostaglandin H2 synthase-1</td>
<td>Yes</td>
<td>Yes</td>
<td>90</td>
</tr>
<tr>
<td>Methionine aminopeptidase</td>
<td>Yes</td>
<td>Yes</td>
<td>91</td>
</tr>
<tr>
<td>Animal and phage lysozymes</td>
<td>Yes</td>
<td>Yes</td>
<td>92</td>
</tr>
<tr>
<td>Glutathione peroxidase</td>
<td>Yes</td>
<td>Yes</td>
<td>93</td>
</tr>
<tr>
<td>Thioredoxin</td>
<td>Yes</td>
<td>Yes</td>
<td>94</td>
</tr>
<tr>
<td>Glutaredoxin</td>
<td>Yes</td>
<td>Yes</td>
<td>95</td>
</tr>
</tbody>
</table>