Protein Science (1993), *2*, 877–883. Cambridge University Press. Printed in the USA.

REVIEW

Catching a common fold

TOM L. BLUNDELL AND MARK S. JOHNSON

Imperial Cancer Research Fund Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, London WC1E 7HX, United Kingdom

(RECEIVED February 4, 1993; REVISED MANUSCRIPT RECEIVED March 22, 1993)

As the first proteins were sequenced in the 1950s, it was evident that they belonged to families. The determination of protein three-dimensional structures during the late 1960s and early 1970s (e.g., insulins, globins, and serine proteinases) confirmed that related proteins from different species adopt similar tertiary structures characteristic of each family. The sequence variations within a family reflected the restraints of the tertiary structures: apart from the catalytic or binding residues, invariant amino acids were most often in the protein core, inaccessible to solvent and with a key role in the protein architecture.

The fascination with families of proteins was deepened with the realization that many proteins, with quite unrelated sequences, could adopt a common fold. Rossmann, Matthews, Branden, Richardson, and many others recognized similarities between the tertiary structures or domains that occur in many quite different proteins (Richardson, 1981); these included $\alpha\beta$-nucleotide binding motifs (Rossmann fold), $\alpha\beta$-barrels (TIM barrel), β-jelly rolls, four α-helix bundles, and immunoglobulin domains (β-Ig fold). These protein topologies underlined the fact that tertiary structures could be considered as simple combinations of secondary structural elements packed together in a limited number of ways: $\alpha\beta\alpha\beta\alpha\beta$, $\alpha\alpha\alpha\alpha$, $\beta\beta\beta\beta$, and so on. It seemed that protein structures could be predicted from sequences by combinatorial assembly of the basic elements of secondary structure, following various rules about handedness of the loops connecting them and the avoidance of strands that were "cross-overs." However, such combinatorial approaches to the protein folding problem depend on correct assignment of α-helices, β-strands, and coils, and this remains a formidable challenge.

Attention was also distracted by an often fruitless argument on evolution. It seems likely that many protein structures have converged by the evolution of stable, common folds. Equally many proteins have evolved by swapping exons corresponding to structural, and sometimes functional, modules to give rise to complex multidomain structures. But it is difficult to be confident of divergent evolution, and in any case the knowledge is not very useful. Karl Popper reminds us that a hypothesis is of little scientific value unless an experiment can be devised that might falsify it; this is certainly difficult for hypotheses about divergent evolution of protein folds. A more useful line of enquiry ignores the question of convergent or divergent evolution and simply asks: "Can we recognize a sequence that will adopt a known protein fold?" A related question, often known as the inverse folding problem, is: "If we know the three-dimensional structure of one protein, can we predict the sequences that might adopt a similar fold?" The discussion of these questions provides the theme of this review.

Comparison and clustering of protein folds

A necessary first step to understanding common folds is the comparison of protein three-dimensional structures. We must establish which parts of the structure are topologically equivalent, how much they differ in space, and where there are insertions or deletions in one structure relative to the others. If such comparisons are automated and made quantitative, we can then estimate "distances" between related protein structures. This allows us to cluster proteins first into "nuclear families" with closely similar and probably homologous family members and then into "extended families" sharing a common fold (Fig. 1).

For homologous families with sequence identities of $\geq 50\%$, much can be achieved by alignment of their sequences using dynamic programming procedures based on the algorithm of Needleman and Wunsch (1970) for pairwise or multiple sequence alignments (Barton & Sternberg, 1987a; Feng & Doolittle, 1987). These methods usually consider the mutation rates of amino acid residues

Reprint requests to: Tom L. Blundell, Imperial Cancer Research Fund Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom.

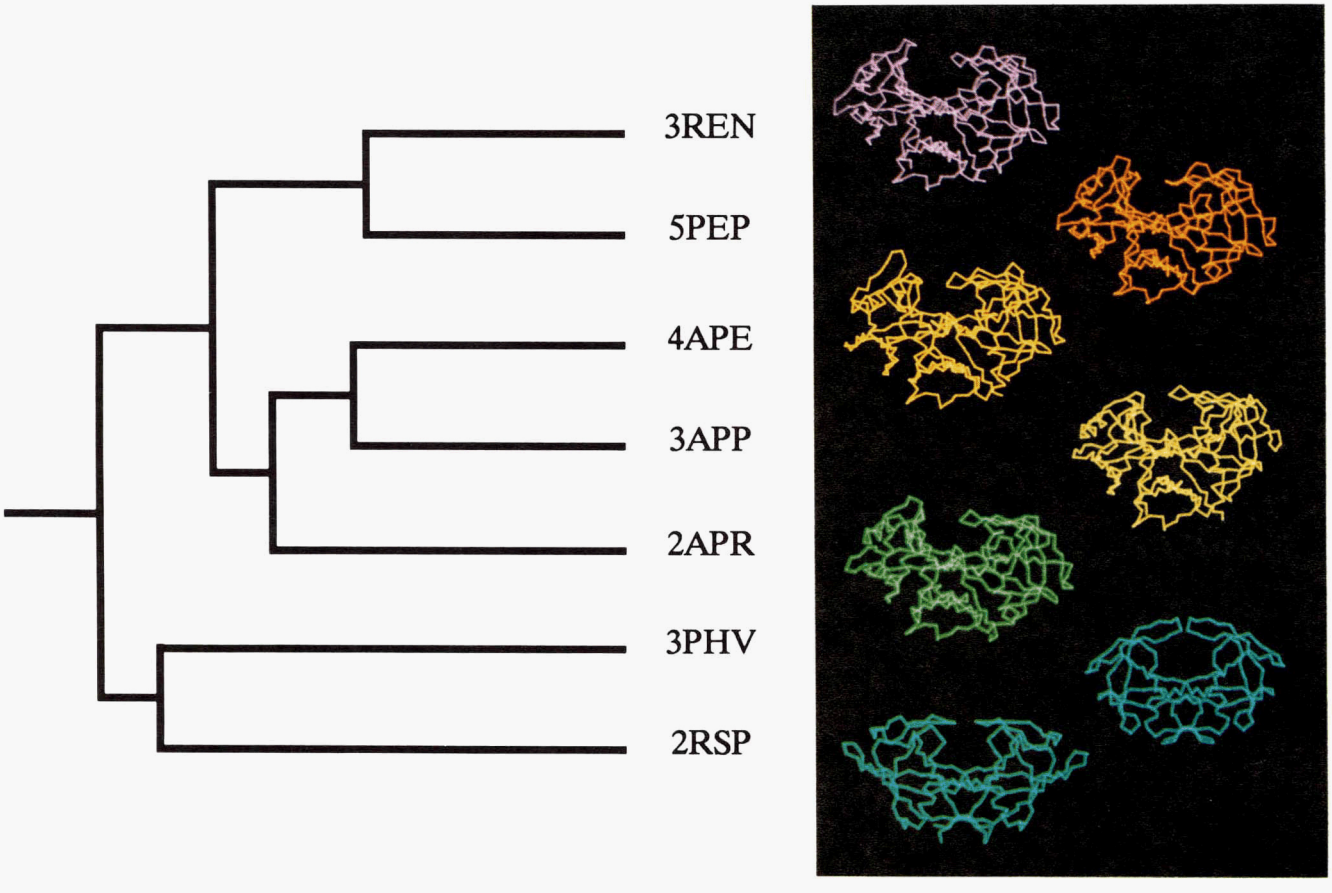

Fig. 1. Cladogram illustrating the relationships among several representative aspartic proteinases. From top to bottom: the mammalian proteinases: mouse renin (Brookhaven code [Bernstein et al., 1977]: 3REN) and porcine pepsin (5PEP); the fungal proteinases: endothiapepsin (4APE), penicillopepsin (3APP), and rhizopuspepsin (2APR); and the retroviral proteinases (dimers): the proteinase from the human immunodeficiency virus (3PHV) and the Rous sarcoma proteinase (2RSP). The tree was computed on the basis of the three-dimensional structures.

to derive optimal comparison scores and corresponding alignments, but other properties such as physicochemical parameters can also be included (Taylor, 1986; Argos, 1987).

In order to define topological equivalence we must compare protein tertiary structures. This often involves rigid-body least-squares superposition of the C_α positions. Several homologous structures can be aligned (Sutcliffe et al., 1987a; Russell & Barton, 1992) without preference to any one in the set in order to define a framework, which comprises a series of helices or strands that are conserved in the family. However, although dissimilar proteins usually retain the general arrangement of strands and helices, differences in orientation and position may preclude their direct superposition (Chothia & Lesk, 1986; Hubbard & Blundell, 1987; Johnson et al., 1990a,b).

The definition of topologically equivalent residues in polypeptides that have little sequence similarity but adopt similar folds was addressed more than a decade ago (see Matthews & Rossmann [1985] for a review). The methods included information about main-chain direction in the alignment or based their comparisons on rigid-body superposition of small parts of the whole structure. Others used relationships between secondary structure elements (Murthy, 1984; Richards & Kundrot, 1988) or compared segments of protein structures (Vriend & Sander, 1991). Taylor and Orengo (1989a,b) and Šali and Blundell (1990) have proposed the use of dynamic programming for comparisons of three-dimensional structures. In the work of Taylor and Orengo (1989a,b), interatomic vectors between residues are compared. In the computer program COMPARER of Šali and Blundell (1990), several features of protein sequences and structures such as local conformation, accessibility, and direction of chain are simultaneously compared. Because the method considers the most conserved protein features at a number of structural levels, it can be used to align distantly related protein structures, which is difficult to achieve by using least-squares superposition alone.

In most sequence alignments, a uniform gap penalty function implies that residues at any position of a protein have had the same chance of deletion or insertion during

evolution. However, insertions and deletions occur more often on the protein surface than in the core, and less often within a secondary structure element than in a loop. Such information relating to the mechanism of protein evolution was applied to the alignment of distantly related sequences by Lesk et al. (1986) and Barton and Sternberg (1987b). The approach has been implemented in the structural comparison program COMPARER (Zhu et al., 1992) by defining gap penalties in terms of structural parameters defined by analyses of families of proteins.

The aligned or equivalenced protein structures can then be clustered. A matrix of distances computed between all pairs of proteins can be used to construct a tree concisely describing relationships among them. Although extensive methodology for tree construction from protein sequences has been developed for the study of evolution (Doolittle, 1990), the clustering of protein three-dimensional structures has been less studied. Rossmann and colleagues (Rao & Rossmann, 1973; Eventoff & Rossmann, 1975) constructed cladograms based on structural features alone to describe distant phylogenetic relationships among the mononucleotide and dinucleotide binding proteins. Johnson et al. (1990a,b) have showed that a structural distance metric, defined from fractional topological equivalence and root mean square deviation between superposed members of the family, can give useful cladograms that correlate well with those derived from sequences (Fig. 2). We have extended this approach by reflecting additional structural and sequence features in the classification (Fig. 2) (Johnson et al., 1990a; Šali & Blundell, 1990). Moreover, because these features can include relationships such as hydrogen bonding patterns, which are known to be conserved in evolution, structures that bear little similarity in other respects can be compared and classified at statistically significant levels.

What is invariant in a common fold?

The question of conservation of amino acids in proteins with a common fold can be addressed by examining a da-

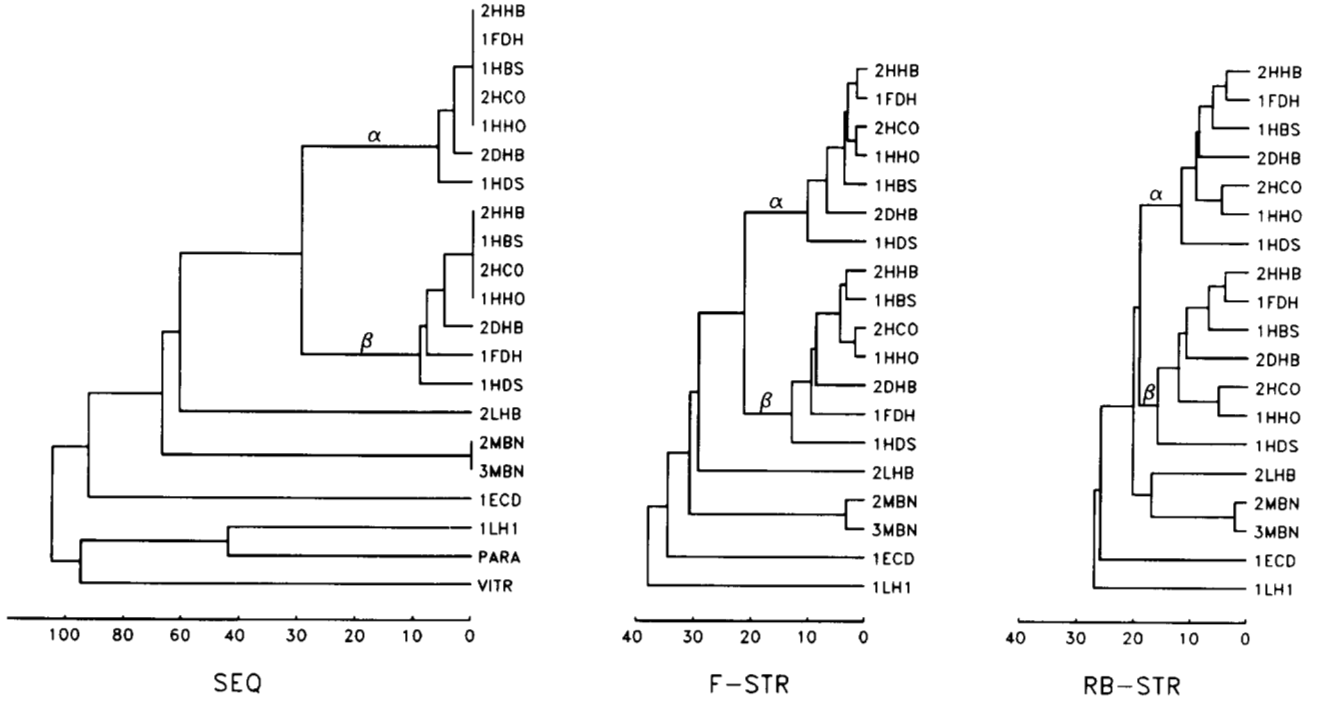

Fig. 2. Cladograms showing the relationships among 19 globin sequences (SEQ) and globin structures (F-STR and RB-STR). Sequences were multiply aligned and the tree (SEQ) constructed from distances based upon the alignment scores. The tree derived from the superposition of three-dimensional structures as rigid bodies (RB-STR) was based on both the root mean square deviations and the number of topologically equivalent main-chain C_α atoms for pairs of superposed proteins. The third tree (F-STR) was computed using a more flexible approach to structural alignments and considered the following features: the physical properties of amino acids (5%), residue main-chain solvent accessibilities (20%), residue identities (10%), and absolute main-chain directions in space (20%). *Labels:* The α and β-chains of human deoxyhemoglobin (Brookhaven code [Bernstein et al., 1977]: 2HHB), human carbonmonoxyhemoglobin (2HCO), human oxyhemoglobin (1HHO), human sickle cell hemoglobin (1HBS), human fetal deoxyhemoglobin (1FDH), horse deoxyhemoglobin (2DHB), deer sickle cell (1HDS) and sea lamprey hemoglobin V (2LHB), sperm whale metmyoglobin (2MBN), sperm whale deoxymyoglobin (3MBN), erythrocruorin of *Chironomus thummi thummi* (1ECD), and leghemoglobin of *Lupinus luteus* (1LH1). The sequences of the bacterial hemoglobin of *Vitreoscilla* (VITR) and the globin of the non-leguminous *Parasponia andersonii* (PARA) are also included within the sequence comparison. Figure used with permission (Johnson et al., 1990a).

tabase of alignments constructed by equivalencing topological features of family members (Overington et al., 1990, 1992; Zhu et al., 1992). Rules for the substitution of amino acids in three-dimensional structures are derived by counting how many times two residue types occur at structurally equivalent positions. We have constructed a number of specific substitution tables (Overington et al., 1990) in which we consider only a subset of residues that have a certain structural environment. Structural features included were solvent accessibility, local main-chain conformation (positive ϕ angle, helical, β-strand, or other), and side-chain hydrogen bonding to peptide groups or other side chains.

In general these analyses confirm that solvent-inaccessible residues are among the most conserved residues in a family, but they also underline the fact that the substitution patterns themselves vary significantly in different environments. Unique patterns characterize local structural features, especially those that involve positive ϕ values (selecting glycine, asparagine, and aspartic acid with highest probability) (Overington et al., 1990, 1992). Substitutions of aspartic acid, asparagine, glutamine, serine, and threonine are strongly influenced by side-chain accessibility and hydrogen bonding. For example, hydrogen-bonded and inaccessible aspartic acids are among the most highly conserved residues in families of proteins, whereas asparagines in the same environment are rarely invariant. Clearly, knowledge of the topology of one member of the family allows some prediction of the variation of sequence that can be tolerated with retention of the fold.

How many family folds?

Once we can compare and align protein sequences and structures, we can begin to address the question of the number of families with a common fold. However, it becomes apparent soon that if we are not interested primarily in the question of divergent evolution, the answer will be operationally defined. Do we wish to know how many families of proteins can be recognized by their sequences in a search of a sequence database? Do we wish to discover how many families have members that can be usefully superposed to define equivalence of $\geq$50% amino acid residues? Do we wish to define the number of families with a similar topology and, if so, how many insertions or deletions of helices and strands are allowed before we form another family?

Even with agreement on the operational definition of a family, there are many approaches to estimating the number of families. For a protein crystallographer, the most obvious approach is to consider the percentage of new structures defined by X-ray analysis and NMR that have structures that are similar to those defined previously. The answer is probably about 50% if we consider general topologies. One example in the past year has been the structure of the flexible multiple domain polymerase, porphobilinogen deaminase, which is involved in assembling the heme, chlorophyll, and B_{12} pyrrole precursor; this structure has two domains that are topologically identical to anion binding proteins such as phosphate and sulfate binding proteins (Louie et al., 1992). This may reflect a functional similarity, as porphobilinogen is also an anion.

At present there are rather more than 100 such families in the present database, indicating the existence of at least 200 families in nature if we have sampled half the existing protein folds already. However, protein crystallographers select proteins that are available in large quantities, either naturally or through cloning and expression, and crystallization further selects those that are soluble and do not comprise flexible regions of polypeptide or indeed flexible strings of modules. NMR spectroscopists select proteins that are small and usually relatively rigid. It is therefore optimistic to think that we have seen half of the protein folds existing in nature, and the estimate of 200 protein folds will need to be revised upward.

Dayhoff and coworkers (1983) have postulated the existence of nearly 1,000 different protein families. As Chothia (1992) has pointed out, about one-third of the sequences defined in genome studies are related to sequences previously determined, and about a quarter of the sequences determined are at least 25% identical with those of three-dimensional structures defined by X-ray analysis and NMR (Sander & Schneider, 1991; Pascarella & Argos, 1992). This gives an upper estimate of about 1,500 families. However, sequence searching is a most ineffective way of identifying sequences that clearly belong to a family like the globins (Johnson et al., 1993), and there will be many members with similar topologies but sequence identities less than 25%. Our own conclusion is therefore that the number of families with distinct topologies is much less, probably between 500 and 700.

This estimate implies that with the increasing number of new structures defined each year, we should move toward experimental definition of one example of each common fold by the end of the century. This implies that if methods to identify the folds from their sequences can be developed, and if comparative modeling can be extended to distantly related protein topologies, then we shall be able to provide at least rough indications of protein three-dimensional structures for most sequences defined by genome sequencing projects (Blundell et al., 1987).

Recognizing common folds from sequences

The major difficulty in using the common fold as a framework for protein modeling is the association of a new sequence with a characterized protein fold. This has been successfully addressed on many occasions by using structural information to identify key features in protein architecture and associating these with invariant or con-

served sequences. Many researchers have done this qualitatively. One example of many was our own proposal that relaxin adopted the insulin fold (Bedarkar et al., 1977). This involved alignment of the key half-cystines and conserved glycines, ensuring that the solvent-inaccessible residues were conserved as hydrophobic. The alignment was then used to model the relaxin structure on the basis of the known three-dimensional structure of insulin. The general features of this model have been confirmed recently by an X-ray analysis of relaxin crystals (Eigenbrot et al., 1991).

One of the first attempts to use structural information systematically was by Taylor (1986), who developed a method of generating templates for each part of the framework of a protein generated from superposition on the basis of known three-dimensional structures of proteins in a family. Eisenberg and coworkers developed an alternative approach, profile analysis, in which sequences were aligned for a family of proteins and the alignments were used to assess the probability of finding an amino acid at each position in the protein fold (Gribskov et al., 1987). Pickett et al. (1992) have exploited the flexible template procedure devised by Barton and Sternberg (1990) and have shown its usefulness in locating β/α-barrels. All three methods improved the success of finding related sequences when compared to using a single sequence in the search procedure. However, these methods require knowledge of several sequences for any protein family and benefit from the increased level of alignment accuracy (e.g., for the template or profile) when structural information is also available.

What can be achieved if only one sequence and three-dimensional structure comprises the total family membership? This has been addressed in several ways. Ponder and Richards (1987) used a library of side-chain rotamers and sought to find combinations of sequence and side-chain conformation that would allow retention of a known three-dimensional structure. This provides a powerful approach to identifying closely related sequences, but it is computationally very expensive. It is limited by the fact that distantly related proteins evolve through relative translations and reorientations of the elements of secondary structures in order to allow changes in side-chain shapes and volumes. This flexibility in structure is difficult to allow for in such analyses. Jones et al. (1992) have sought to overcome this problem by using knowledge-based potentials (Sippl, 1990). They thread a sequence through a known structure and ask, for each alignment: Can the sequence of interest adopt that three-dimensional fold? Maiorov and Crippen (1992) have derived a similar potential that recognizes the correct folding of globular proteins.

In order to escape from the limitations of the three-dimensional structure of any one member of the family, it may be necessary to "project" the restraints of the three-dimensional fold onto the one dimension of the sequence (Šali et al., 1990) and to work by comparing sequence templates or profiles. This can be approached by determining the propensity of an amino acid to occur in each class of local structural environment defined by solvent accessibility and secondary structure, as shown by Eisenberg and his colleagues (Bowie et al., 1991). Alternatively it can be achieved by calculating substitution tables as a function of local environment (Overington et al., 1990; Luthy et al., 1991). The method of Johnson and coworkers (Overington et al., 1992; Johnson et al., 1993) uses expanded amino acid substitution tables that take into account the local environment in the tertiary structure.

Each of these methods is able to detect distantly related sequences that adopt a particular protein fold. But none of the methods is successful in identifying all known topological relationships. Most methods seem able to identify actin on the basis of the ATPase fragment of the heat shock cognate protein even though the proteins have only 10% identity in sequence (Fig. 3). This is almost certainly because they are rather similar in length, at 375 and 386 amino acid residues, and there are not too many other proteins of this type and length. The major problem is in introducing useful gap penalties. Johnson et al. (1993) have done this by introducing structure-dependent gap penalties, as used in the comparisons of protein three-dimensional structures. However, when there are long insertions and deletions within the similar domains, or where one structure contains extra domains on either end, then the comparisons are much more problematic. Several of the methods have difficulty in identifying mammalian serine proteinases on the basis of the bacterial enzymes or vice versa; this is clearly a consequence of the insertions in the sequences of one group relative to the other.

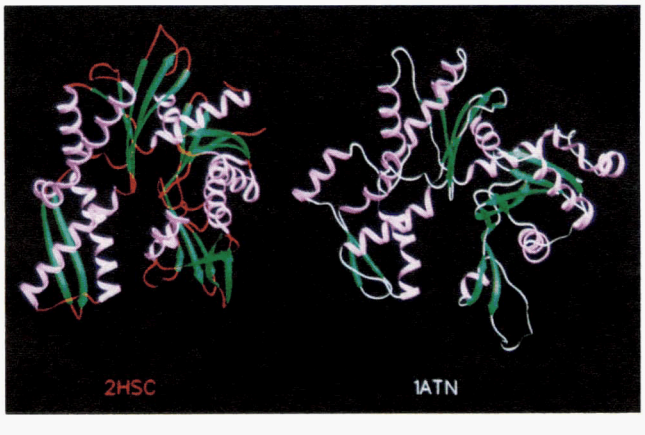

Fig. 3. Ribbon drawing for the 40-kDa fragment of the heat shock protein (Brookhaven code [Bernstein et al., 1977]: 2HSC) and actin (1ATN). Despite sharing this complex topology, the sequence identity based on structural comparisons is less than 10%. Helices have been colored lavender and strands have been colored green; loops are red in 2HSC and white in 1ATN.

Of the recent methods aimed toward catching a common fold, none is sufficiently good to recognize all members of a known protein class. For instance, consider the globins. A single globin structure challenged to find all globin sequences in a data bank will catch roughly 95% of the globins if the structure is a mammalian myoglobin or α or β chain of hemoglobin. One hundred percent of the globins will not be located before the first nonglobin, and the performance is much worse for globins such as erythrocruorin or leghemoglobin. Some of the difficulties lie with the method of comparison between a sequence and a structure and are independent of the information under comparison. Dynamic programming techniques such as that of Needleman and Wunsch (1970) suffer when proteins of different lengths are compared. Procedures that compare segments can have difficulty in reconstructing an alignment and are more time consuming. As mentioned above, the "gap problem" (Doolittle, 1981) is still a notorious problem. On the positive side, multiple alignments of proteins can help to provide templates or profiles that are much more sensitive to the detection of related folds.

Conclusions

Having identified the fold of a new sequence, we can use the information to predict the three-dimensional structure of the protein. This is often carried out subjectively, but more systematic approaches are now being developed. For example, we can use rules that relate the side-chain dihedral angle with the residue type at equivalent positions in homologous proteins (Summers et al., 1987; Sutcliffe et al., 1987b). Most methods depend on the assembly of rigid fragments (Jones & Thirup, 1986; Blundell et al., 1987, 1988; Claessens et al., 1989). In our approach (Blundell et al., 1988; Topham et al., 1993) we select three sets of fragments that define the framework, the structurally variable, mainly loop regions, and the side chains.

These modeling procedures are very successful when the known structures cluster around that to be predicted and when the percent sequence identity to the unknown is high (greater than 40%). There is still much to be done in improving these approaches. What is clear is that the identification of a sequence with a previously characterized "common fold" provides a very useful restriction of conformational space and a very helpful starting point for producing a useful model. Even without a precise model it will often provide clues about function in general and ligand binding in particular. These will increase the value of sequence information, which will surely increase in volume as the various genome sequencing projects get under way in the coming decade.

Acknowledgments

We thank Gordon Louie and Chris Thorpe for assistance in the preparation of Figure 3.

References

Argos, P. (1987). A sensitive procedure to compare amino acid sequences. *J. Mol. Biol. 193*, 385–396.

Barton, G.J. & Sternberg, M.J.E. (1987a). A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol. 198*, 327–337.

Barton, G.J. & Sternberg, M.J.E. (1987b). Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng. 1*, 89–94.

Barton, G.J. & Sternberg, M.J.E. (1990). Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *J. Mol. Biol. 212*, 389–402.

Bedarkar, S., Turnell, W.G., Blundell, T.L., & Schwabe, C. (1977). Relaxin has conformational homology with insulin. *Nature 270*, 449–451.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L., & Sutcliffe, M. (1988). Knowledge-based protein modelling and design. *Eur. J. Biochem. 172*, 513–520.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., & Thornton, J.M. (1987). Knowledge-based prediction of protein structure and the design of novel molecules. *Nature 326*, 347–352.

Bowie, J.U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into known three-dimensional structures. *Science 253*, 164–170.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature 357*, 543–544.

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823–826.

Claessens, M., van Cutsem, E., Lasters, I., & Wodak, S. (1989). Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng. 2*, 335–345.

Dayhoff, M.O., Barker, W.C., & Hunt, L.T. (1983). Establishing homologies in protein sequences. *Methods Enzymol. 91*, 524–545.

Doolittle, R.F. (1981). Similar amino acid sequences: Chance or common ancestry? *Science 214*, 149–159.

Doolittle, R.F., Ed. (1990). *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Methods in Enzymology*, Vol. 183. Academic Press, San Diego.

Eigenbrot, C., Randal, M., Quan, C., Burnier, J., O'Connell, L., Rinderknecht, E., & Kossiakoff, A.A. (1991). X-ray structure of human relaxin at 1.5 Å. Comparison to insulin and implications for receptor binding determinants. *J. Mol. Biol. 221*, 15–21.

Eventoff, W. & Rossmann, M.G. (1975). The evolution of dehydrogenases and kinases. *CRC Crit. Rev. Biochem. 3*, 112–140.

Feng, D.-F. & Doolittle, R.F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol. 25*, 351–360.

Hubbard, T.J.P. & Blundell, T.L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful in protein modelling. *Protein Eng. 1*, 159–171.

Gribskov, M., McLachlan, A., & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA 84*, 4355–4358.

Johnson, M.S., Overington, J.P., & Blundell, T.L. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol. 231*, in press.

Johnson, M.S., Šali, A., & Blundell, T.L. (1990a). Phylogenetic relationships from three-dimensional protein structures. *Methods Enzymol. 183*, 670–690.

Johnson, M.S., Sutcliffe, M.J., & Blundell, T.L. (1990b). Molecular anatomy: Phyletic relationships from three-dimensional structures of proteins. *J. Mol. Evol. 30*, 43–59.

Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature 358*, 86–89.

Jones, T.H. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J. 5*, 819–822.

Lesk, A.M., Levitt, M., & Chothia, C. (1986). Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng. 1*, 77–78.

Louie, G.V., Brownlie, P.D., Lambert, R., Cooper, J.B., Blundell, T.L., Wood, S.P., Warren, M.J., Woodcock, S.C., & Jordan, P.M. (1992). Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. *Nature 359*, 33–39.

Luthy, R., McLachlan, A.D., & Eisenberg, D. (1991). Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins Struct. Funct. Genet. 10*, 229–239.

Maiorov, V.N. & Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol. 227*, 876–888.

Matthews, B.W. & Rossmann, M.G. (1985). Comparison of protein structures. *Methods Enzymol. 115*, 397–420.

Murthy, M.R.N. (1984). A fast method of comparing protein structure. *FEBS Lett. 168*, 97–102.

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol. 48*, 443–453.

Overington, J.P., Donnelly, D., Johnson, M.S., Šali, A., & Blundell, T.L. (1992). Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci. 1*, 216–226.

Overington, J.P., Johnson, M.S., Šali, A., & Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc. R. Soc. Lond. B 241*, 132–145.

Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng. 5*, 121–137.

Pickett, S.D., Saqi, M.A.S., & Sternberg, M.J.E. (1992). Evaluation of the sequence template method for protein structure prediction. Discrimination of the $(\beta/\alpha)_8$-barrel fold. *J. Mol. Biol. 228*, 170–187.

Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol. 193*, 775–791.

Rao, S.T. & Rossmann, M.G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol. 76*, 241–256.

Richards, F.M. & Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: Secondary and first level supersecondary structure. *Proteins Struct. Funct. Genet. 3*, 71–84.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem. 34*, 167–339.

Russell, R.B. & Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct. Funct. Genet. 14*, 309–323.

Šali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol. 212*, 403–428.

Šali, A., Overington, J.P., Johnson, M.S., & Blundell, T.L. (1990). From comparison of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci. 15*, 235–240.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet. 9*, 56–68.

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol. 213*, 859–883.

Summers, N.L., Carlson, W.D., & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol. 196*, 175–198.

Sutcliffe, M.J., Haneef, I., Carney, D., & Blundell, T.L. (1987a). Knowledge-based modelling of homologous proteins: I. Three-dimensional frameworks. *Protein Eng. 1*, 377–384.

Sutcliffe, M.J., Hayes, F.R.F., & Blundell, T.L. (1987b). Knowledge-based modelling of homologous proteins: II. Rules for the conformations of substituted side chains. *Protein Eng. 1*, 385–392.

Taylor, W.R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol. 188*, 233–258.

Taylor, W.R. & Orengo, C.A. (1989a). Protein structure alignment. *J. Mol. Biol. 208*, 1–22.

Taylor, W.R. & Orengo, C.A. (1989b). A holistic approach to protein structure alignment. *Protein Eng. 2*, 505–519.

Topham, C., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S., & Blundell, T.L. (1993). Identification of key residues in structurally variable regions of proteins using conformationally-constrained environmental substitution tables: Applications to loop fragment ranking in modelling of protein structure. *J. Mol. Biol. 229*, 194–220.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins Struct. Funct. Genet. 11*, 52–58.

Zhu, Z.-Y., Šali, A., & Blundell, T.L. (1992). A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng. 5*, 43–51.