PROTEINS: Structure, Function, and Genetics 7:257–264 (1990)

Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets With Solvent Accessibility Patterns of Known Structures

James U. Bowie[1], Neil D. Clarke[2], Carl O. Pabo[3], and Robert T. Sauer[1]
[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; [2]Howard Hughes Medical Institute and Department of Molecular Biology and Genetics, and [3]Department of Biophysics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

ABSTRACT **Hydrophobic side chains often are buried in the interior of a protein, and evolutionarily related proteins usually maintain the hydrophobic character of buried positions. In this paper we show that a pattern of hydrophobicity values derived from a set of related protein sequences is well correlated with the linear pattern of side-chain solvent accessibility values, calculated from a known protein structure representative of the sequences. In several cases, information from aligned sequences can be used to select the correct tertiary fold from a large database of protein structures.**

Key words: protein families, structure prediction

INTRODUCTION

Residues buried in the interior of a protein are almost invariably hydrophobic, and removing these hydrophobic residues from water is a major factor in protein folding and stability.[1,2] The importance of these residues is clearly demonstrated by the conserved hydrophobic character of buried residues in related proteins.[3–7] In contrast, surface residues are more variable, accommodating both hydrophobic and hydrophilic side chains. Because of this difference, comparison of the sequences of related proteins allows one to identify residues that are likely to be buried and residues that are likely to be on the surface.[6,8] Such an analysis is most reliable when the sequences of many related proteins are known or when large numbers of neutral amino acid substitutions have been generated and analyzed.[8,9]

In this paper, we test whether patterns of conserved hydrophobic residues, inferred from a comparison of related protein sequences, are sufficiently characteristic of the three-dimensional fold to allow the identification of the correct structure in a database of known protein structures. If, in general, the preference for hydrophobic amino acids at particular sequence positions indicates that these residues are buried, then the pattern of hydrophobic residues should be reflected in a corresponding pattern of buried residues in the folded protein. We find that, in several tested cases, the similarity of these patterns is indeed sufficient to correctly identify the tertiary fold adopted by a set of protein sequences.

MATERIALS AND METHODS

Sequences, Sequence Alignments, and Protein Structures

The following sets of sequences were used in this study: 1) the sequences of nine globins, aligned by Lesk and Chothia[4] from the crystallographically determined structures; ii) nine M class cytochromes from an alignment based on the tuna cytochrome c structure[10] and one additional sequence, rice cytochrome c, whose structure has recently been found to closely resemble that of tuna cytochrome c[11]; iii) the CheY protein family, including the protein sequences of CheY, CheB, SfrA, OmpR, SpoA, SpoF, $NtrC_{Kp}$, $NtrC_{Bp}$, VirG, DctD, PhoB and ORF2 (J. Stock, personal communication); and iv) the EF-hand sequences from Szebenyi et al.[12]

The following 103 protein structures from the Brookhaven databank [13] were used in this study (Brookhaven designation); 156B, 1ABP, 1ACX, 1APR, 1BP2, 1CC5, 1CCR, 1CRN, 1CTF, 1CTX, 1ECD, 1FB4, 1FBJ, 1FC2, 1FDX, 1GCN, 1GCR, 1GP1, 1GPD, 1HIP, 1HMG, 1HMQ, 1INS, 1LZ1, 1MBD, 1MLT, 1NXB, 1PCY, 1PPD, 1PPT, 1PYP, 1RHD, 1RN3, 1RNS, 1SBT, 1SN3, 1TGN, 1TIM, 1TPA, 1UBQ, 2ABX, 2ACT, 2ADK, 2ALP, 2APP, 2AZA, 2B5C, 2CAB, 2CCY, 2CDV, 2CGA, 2CNA, 2CYP, 2EBX, 2EST, 2FD1, 2GCH, 2GN5, 2GRS, 2HHB, 2LH4, 2LZM, 2MDH, 2MT2, 2OVO, 2PAB, 2PKA, 2PTC, 2RHV, 2SGA, 2SNS, 2SOD, 2SSI, 2STV, 2TAA, 2TBV, 351C, 3C2C, 3CPV, 3CTS, 3FXC, 3ICB, 3PGK, 3PGM, 3RP2, 3SGB, 3TLN, 3WGA, 4ADH, 4APE, 4ATC, 4CYT, 4DFR, 4FXN, 4LDH, 4SBV, 5API, 5CHA, 5CPA, 5PTI, 5RXN,

Received April 26, 1989; accepted January 2, 1990.
Address reprint requests to Dr. Robert Sauer, Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139.

8CAT, 9PAP. Our database also includes Rop,[14] catabolite gene activating protein,[15] λ Cro repressor,[16] the N-terminal domain of bacteriophage 434 repressor,[17] and the N-terminal domain of bacteriophage λ repressor.[18] Our "database of known structures" consists of the unique subunits of these 108 proteins.

Hydrophobicity Patterns Derived From Sets *of Aligned Sequences*

Sets of aligned sequences were used to identify positions that require hydrophobic residues, and positions that can accommodate hydrophilic side chains. Each position in the set of aligned sequences was characterized by picking (with the steps described below) one of the most hydrophilic residues allowed at that position. Specifically 1) the amino acids found at each position in the aligned sequences were ranked by hydrophobicity, using the scale of Fauchere and Pliska[19]; 2) to account for possible alignment errors, sequence errors and occasional exceptions, the most hydrophilic residue found at each position was discarded (unless it was observed more than once); and 3) the most hydrophilic of the remaining residues was then picked as a measure of the extent to which non-hydrophobic residues are tolerated at this position. Arg and Lys residues were not included in this ranking unless they were the only residues found at a particular position. (Inspection of several globin structures showed that the long, aliphatic parts of the Arg and Lys side chains can effectively substitute for hydrophobic amino acids at some buried positions and still permit the charged groups at the ends of the side chains to reach the solvent.) This algorithm provides a rough measure of the "allowed hydrophilicity" at each residue position.

The steps listed above pick one amino acid to represent each position in the aligned sequences. This representation was further simplified by classifying these residues into three categories. Thus, each position in the aligned sequences is represented as H_1 (high hydrophobicity), H_2 (medium hydrophobicity), or H_3 (low hydrophobicity). H_1 is used if the amino acid which represents the allowed hydrophilicity at that position is Trp, Ile, Phe, Leu, Met, Val, or Cys. H_2 is used if the amino acid which represents the allowed hydrophilicity at that position is Tyr, Pro, Ala, Thr, His, Gly, or Ser. H_3 is used if the amino acid which represents the allowed hydrophilicity at that position is Gln, Asn, Glu, Asp, Lys, or Arg. The rationale for using these classifications is explained below.

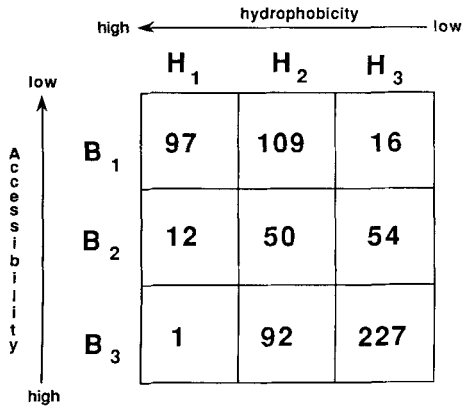
Solvent Accessibility Patterns Determined From Known Structures

For each unique subunit in our database of known structures, a string was created to represent the solvent accessibility at each of the consecutive residues along the polypeptide chain. The solvent accessibility of each position was designated as B_1 (buried), B_2 (partially buried), or B_3 (exposed). Solvent accessibilities were calculated using the Lee and Richards algorithm[20] as implemented in the program ACCESS by Handschucher and Richards. The fractional solvent exposure for each residue in the structures was determined by calculating the solvent exposed area of the Cα and side chain atoms, and then dividing by the solvent exposed area of the same atoms in an extended Ala-X-Ala tripeptide.[20] When determining which residues in a particular chain were buried, we ignored ions, all prosthetic groups except heme and the atoms of other subunits. We decided to use isolated subunits because our method of determining a single hydrophobicity pattern introduces a bias toward monomeric structures. Although a multimeric protein (such as hemoglobin) might have hydrophobic residues at the subunit interfaces, a related monomeric protein (such as myoglobin) probably would have hydrophilic residues at the corresponding positions. Since each of our sequence sets includes monomeric proteins, and since our hydrophobicity analyses are heavily biased by the most hydrophilic residues in the set of aligned sequences, the hydrophobicity pattern for the set of sequences should most closely match the accessibility pattern calculated for the monomeric proteins and for the isolated subunits of the multimeric proteins.

Aligning Hydrophobicity and Accessibility Strings

The hydrophobicity string determined from aligned sequences was compared to each of the accessibility strings for the known structures, and the optimal alignment with each was determined using the Needleman and Wunsch algorithm.[21] The quality of each alignment is obtained by summing the scores for all the paired hydrophobicity and accessibility groups along with the appropriate penalties for any gaps introduced in the alignment. The score for pairing a particular accessibility group, B_j, and a particular hydrophobicity group, H_i, can be read from Figure 1b. The development of this scoring table is discussed in the next section. Our alignment scheme also allowed gaps, although gaps in the alignment were prohibited within α-helices or β-sheets. (Helices and sheets were identified by the method of Kabsch and Sander[23].) Gaps in other areas were allowed, but a gap opening penalty of 3.0 and a gap extension penalty of 0.2 were applied. These gap penalties were determined empirically by maximizing the difference between the scores for the globin structures and the score for the highest scoring non-globin structure. In every case, the significance of the alignment scores was assessed by generating 50 random rearrangements of the solvent accessibility string, determining the best alignment

a)



b)

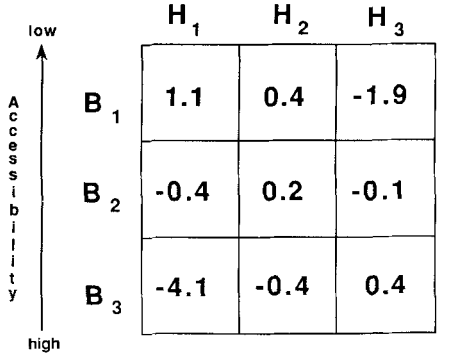


Fig. 1. **a:** Occurrence of each hydrophobicity/accessibility pairing in the globin alignments. As described in the text, the cutoffs that define each hydrophobicity and accessibility class were adjusted to maximize the information content in these globin comparisons. **b:** Scoring table used for evaluating alignments of hydrophobicity strings and solvent accessibility strings. These numbers represent the information values for the globin data (given above). These scores were used for all the alignments discussed in this paper. As described in the text, a gap opening penalty of 3.0 and a gap extension penalty of 0.2 were also applied.

score for each of these randomly rearranged sequences, and comparing the score of the correct solvent accessibility string with the mean and standard deviation of the scores for the randomized strings. Scores are reported as the number of standard deviations above or below the mean of the scores for the randomized strings.

Pair Scoring

To determine the values assigned to pairs of hydrophobicity groups and solvent accessibility groups, we compared the fractional solvent accessibility values of the residues in five globin structures (sperm whale myoglobin, human hemoglobin α and β chains, root nodule leghemoglobin, and larval insect erythrocrourin) with the hydrophobicity values obtained from a set of aligned globin sequences. Figure 1a shows the number of sequence positions of hydrophobicity group H_i which are properly aligned with structural positions of solvent accessibility group B_j. The scores assigned to each of the nine possible pairings are shown in Figure 1b and represent the information value, $I(H_i:B_j)$, for each pairing.[22] The information value is defined as $I(H_i:B_j) = \ln (P(H_i:B_j)/P(H_i))$, where $P(H_i:B_j)$ is the probability (i.e. the observed frequency) that the hydrophobicity group of a residue position is H_i if the accessibility group of that position is B_j, and $P(H_i)$ is the probability of any residue being a member of hydrophobicity group H_i. The variables that can be optimized in this scoring scheme are the cutoffs used to define the three fractional accessibility categories and the three hydrophobicity categories. The cutoffs used in defining each accessibility and hydrophobicity category were adjusted to maximize the total information, I_{total}, according to

$$\text{Max}\,(I_{total}) = \text{Max}\left(\sum_{i=1}^{3}\sum_{j=1}^{3} N(H_i,B_j)\,\ln\left(\frac{P(H_i:B_j)}{P(H_i)}\right)\right)$$

where H_i is the i^{th} hydrophobicity grouping, B_j is the j^{th} solvent accessibility grouping, and $N(H_i, B_j)$ is the total number of residues that are both in hydrophobicity group H_i and accessibility group B_j. The optimal accessibility groups were found to be 0–10%, 11–39%, and greater than 39% solvent exposure. The optimal hydrophobicity groupings were those listed above.

RESULTS

The goal of this work was to determine whether sequence information, from multiple aligned sequences, can be used to find a correct tertiary fold. We used the sequence information to determine a hydrophilicity/hydrophobicity pattern, and then tried to find a known structure with a similar pattern of solvent accessibility values along the polypeptide chain.

Figure 2 summarizes our approach. As a preliminary step, each protein subunit in the database is represented as a string of solvent accessibility values, with B_1 used to represent fully buried position, B_2 used to represent partially buried positions, and B_3 used to represent exposed positions. In some sense, this takes information from the full three-dimensional structures and "projects" it into a one-dimensional form that facilitates later comparison with sequence information. To study a particular set of sequences, a single string is generated to represent the extent to which hydrophilic residues are tolerated at each position of the aligned sequences (see Materials and Methods). The actual search procedure tries to align this hydrophobicity pattern with accessibility patterns from the structural database. Each structure in the database is used as a "trial structure" and the program finds the best alignment between the solvent accessibility pattern of the trial structure and the hydrophobicity pattern of the sequence set. The alignments are found using

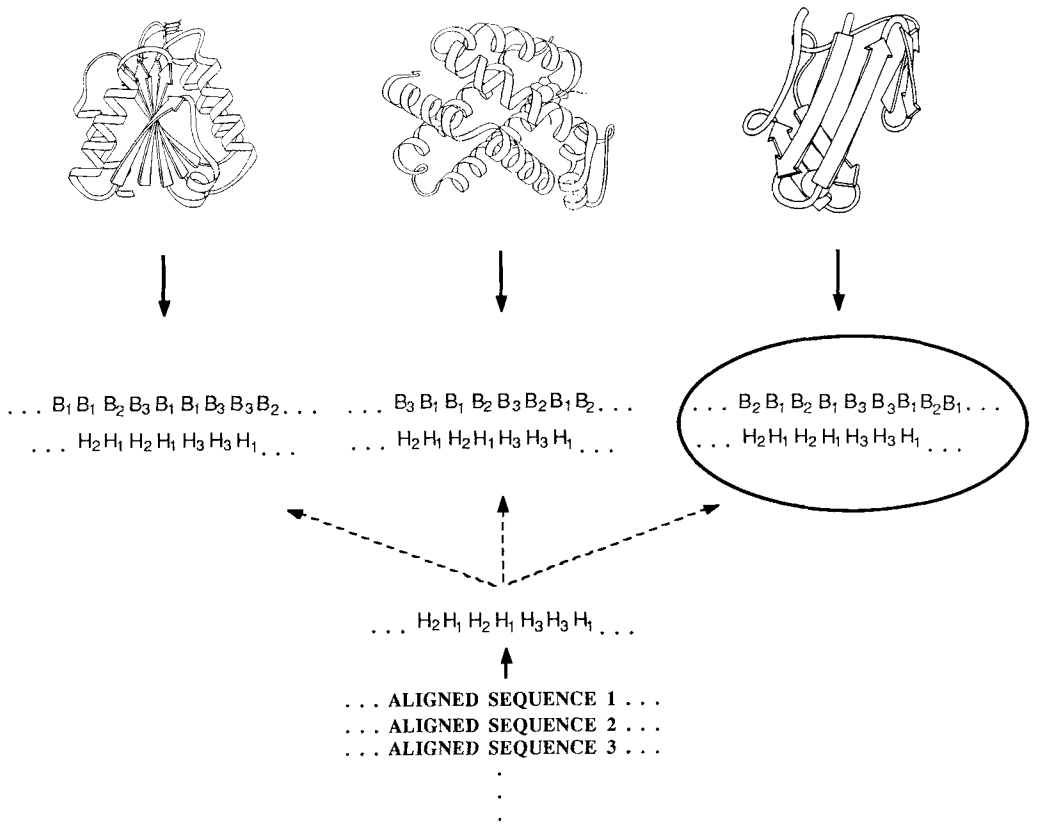


Fig. 2. Summary of pattern matching algorithm. Each protein in the structure database (represented here by a few of Jane Richardson's schematic drawings) is converted to a linear string representing the solvent accessibility of each residue in the folded protein. The arrows from the structures point to a part of the string derived for each of the structures. Shown at the bottom of the figure is a set of aligned protein sequences and part of the single hydrophobicity string generated from this set of sequences. This sequence string is then aligned as well as possible with each of the solvent accessibility strings. The structure which shows the highest quality alignment is circled. (The particular values shown here are for illustrative purposes only, and the position and length of the alignments will vary from one case to the next).

the algorithm of Needleman and Wunsch[21] and the scoring system uses weights that represent the information value for each possible pairing of a hydrophobicity group and an accessibility group. This procedure gives a score for the best alignment with each of the trial structures. The structure with the best overall score is then assumed to be the most plausible structure for the sequence family.

As described in Materials and Methods, we divided both the hydrophobicities and the solvent accessibilities into three groups. Each of the nine possible pairings between a solvent accessibility group and a hydrophobicity group was assigned a value based on how strongly the two groups were correlated in alignments of globin sequences and globin structures. Cutoffs for the hydrophobicity and accessibility groups were chosen by maximizing the ability of the program to distinguish the five globin structures from all other structures in the database.

Figure 1b shows the final scoring table used in evaluating the alignments. In the scoring table, positive numbers indicate that the two properties being compared are positively correlated in the globins; negative numbers indicate that the properties are negatively correlated. The most heavily weighted elements occur at the corners of the table. For example, a position that remains in the highly hydrophobic group, H_1, over the course of evolution is likely to be a member of the buried group, B_1, and is very unlikely to be in the exposed group, B_3. A position that accommodates hydrophilic residues (H_3) is fairly likely to be exposed (B_3) and is very unlikely to be buried (B_1). Because insertions or deletions in evolutionarily related proteins usually do not occur within alpha-helices or beta-sheets, gaps were not allowed in these regions.[3–5,7] Gap penalties for insertions or deletions in other regions were determined empirically to optimize scores for the globins.

Figure 3 shows the distribution of scores when the globin hydrophobicity pattern is matched with each protein structure in the database. The five highest-scoring proteins, which are clearly separated from all other structures, are the five globin structures that were present in the database. This shows that

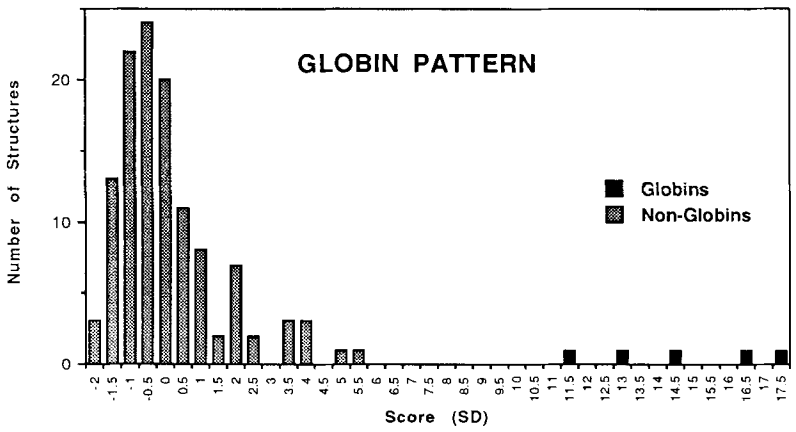


Fig. 3. Aligment scores for the globin hydrophobicity pattern with accessibility patterns calculated from the structure database. Scores are given in standard deviation units and refer to the alignment quality of a particular accessibility pattern, relative to the mean quality of 50 random strings of the same composition. The top five scoring protein structures (in order) are myoglobin, the α-subunit of human hemoglobin, leghemoglobin, erythrocrourin, and the β-subunit of human hemoglobin.

the pattern of hydrophobicity extracted from a set of aligned sequences can be used to identify the correct protein fold among a large collection of possible structures. To determine whether this method works on protein sequences and structures that were not used to adjust the scoring system, the algorithm was tested on several sequence sets, with the same scoring table and gap penalties used in the globin alignments. So that the method could be evaluated, sequence sets were chosen in which the crystal structure of at least one member had been solved.

Cytochrome C

The cytochrome c proteins are a relatively diverse family of heme proteins that can be grouped into four structural classes of different sizes. We considered sequences of the M class of cytochromes since two of these structures were in our database. Figure 4A summarizes the results of searching the database of known structures for the accessibility pattern that best matches the M class cytochrome hydrophobicity pattern. The two M class cytochromes in the structural database are the two highest scoring protein structures.

CheY Protein

CheY proteins are involved in chemotactic signal transduction in *Escherichia coli* and *Salmonella typhimurium* and show sequence similarity to a number of bacterial regulatory proteins.[24] As shown in Figure 4B, flavodoxin is identified by our search procedure as the structure in our database which is most consistent with the hydrophobicity pattern for CheY and its homologues. The crystal structure of the *S. typhimurium* CheY protein has recently been determined,[24] and Figure 5 shows a superimposition of the α-carbons of CheY and flavodoxin. The two structures are topologically related, both consisting of a five-membered β-sheet surrounded by α-helices which alternate with the β-stands along the sequence. This is an interesting test case for our algorithm since no statistically significant homology to flavodoxin is observed by conventional sequence alignment methods (application of the University of Wisconsin Genetics Computer Group program GAP to the alignment of the *E. coli* CheY sequence to the flavodoxin sequence and to 100 random rearrangements of the flavodoxin sequence showed that the alignment with the real sequence was somewhat worse than the mean of the alignments to the randomized sequences).

Calcium Binding Proteins

These proteins have a helix-loop-helix super-secondary structure, referred to as the EF-hand, and each domain of these proteins typically contains two such EF-hands. We used the combined hydrophobicity pattern of EF-hand sequences to search the structural database and the results are summarized in Figure 4C. One of the two calcium-binding proteins in the database, bovine intestinal calcium binding protein (ICBP), scores significantly higher than any other protein structure in the database. The other calcium-binding protein in the database, parvalbumin, did not score highly. Although these two calcium binding proteins contain qualitatively similar helix arrangements, they do differ substantially in both interhelix angles and helical lengths.[12] Clearly, ICBP and parvalbumin are sufficiently different that the solvent accessibility pattern for ICBP is well matched by the consensus hydrophobicity pattern while that of parvalbumin is not.

It is instructive to examine the two structures that score most highly after ICBP. Figure 6 shows the alignment of the EF-hand hydrophobicity pat-

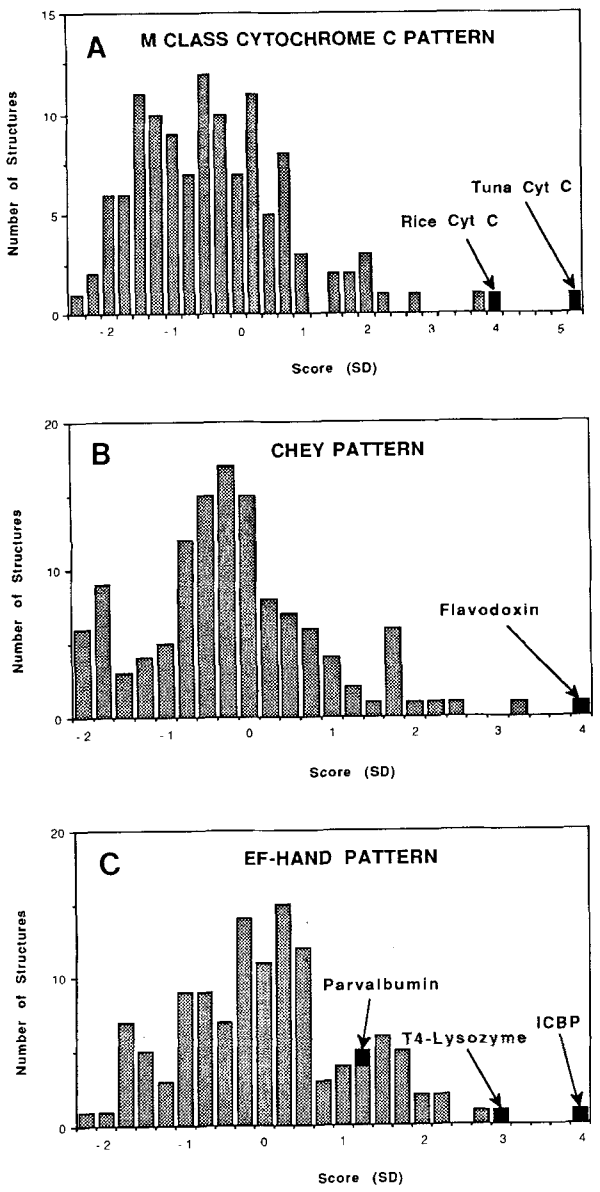


Fig. 4. Alignment scores (see the legend for Fig. 3) for the M class cytochromes c, the CheY protein family, and the EF-hand. The names, in order, of the top five scoring proteins for each histogram are: **A)** tuna cytochrome c, rice cytochrome c, cytochrome B562, superoxide dismutase, flavodoxin; **B)** flavodoxin, T4 lysozyme, cytochrome B562, yeast phosphoglycerate kinase, chymotrypsinogen A; **C)** the EF-hand: vitamin D dependent calcium binding protein, T4 lysozyme, catabolite gene activating protein, 434 repressor, arabinose binding protein.

tern with ICBP, T4 lysozyme, and catabolite gene activating protein (CAP). (This Figure also illustrates how a hydrophobicity pattern is generated.) As indicated in the Figure, the segment of T4 lysozyme which was identified by the EF-hand hydrophobicity pattern shares the same secondary structure features as the EF-hand. Furthermore, these two helices are oriented with respect to each other in a way reminiscent of the helices in the EF-hand.[25] Along with the CheY structure identification discussed above, this is a good example of the ability of our method to identify similar structural motifs in proteins that are not obviously related in sequence.

On the other hand, the structure identified in CAP is clearly unrelated to the EF-hand (there is no helix in the first 18 residues of the identified segment), even though the alignment score is similar to that in the other two proteins. Because there always is some chance of such spurious alignments, a high score cannot be considered proof of a correct structural identification. We also note that any structural model identified in this type of search must only be considered an approximate representation of the actual fold. For example, although each of the globin structures scores quite highly with the globin hydrophobicity pattern, individual helix packings in these structures can differ by up to 3 Å in interaxial distance and up to 30° in interhelix angle.[4]

DISCUSSION

Ultimately, one would like to be able to use sequence information to predict protein structures. In this paper, we have taken the preliminary step of testing whether we can use sequence information to pick out the correct structure from a database of known structures. We find that the correlation between the hydrophobicity of amino acid side-chains and their solvent accessibility in folded proteins is sufficiently good that a sequence pattern can be used to directly identify a structure.

The hydrophobicity information in a set of amino acid sequences was summarized by a single string of hydrophobicity values (high, medium, or low). Similarly, the information from the three-dimensional structures of proteins was summarized by representing the structures as strings of solvent accessibility values (buried, partially buried, or exposed). The results discussed in this paper show that these simple patterns can be surprisingly selective. The tertiary fold adopted by a set of sequences was correctly identified in several different test cases. Most notably, the flavodoxin structure was correctly identified as being most compatible with the hydrophobicity pattern of the CheY protein family, even though no significant sequence homology has been found.

It is striking that simple hydrophobicity patterns allow the identification of structural folds. This extends earlier observations that patterns of hydrophobicity can be a distinctive feature of the sequences of particular classes of protein. Sweet and Eisenberg have shown that the degree of correlation of hydrophobicity values between pairs of aligned sequences is a good criterion for the structural similarity of the two proteins.[26] In a similar vein, Bashford et al. have created a sequence template that uses residue hydrophobicity and size criteria to se-

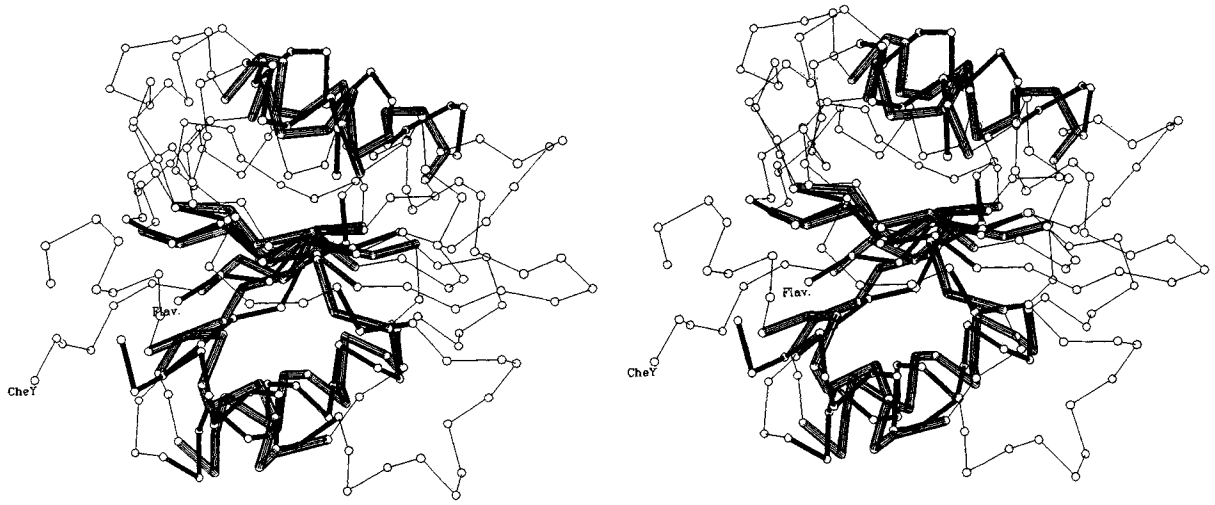


Fig. 5. Superimposition of the α-carbons of the CheY and flavodoxin. The bold segments identify the segments of the two proteins that were used for the alignment. Note that analogous secondary structures occupying similar positions in the folds of the two proteins. The aligned residues in flavodoxins are 1–6, 13–26, 29–35, 47–52, 62–74, 81–85, 100–106, 114–116, and 128–137. The corresponding residues in CheY are 8–13, 15–28, 31–37, 52–57, 60–72, 82–86, 95–101, 107–109, and 112–121. These 68 α-carbons (out of 138 for flavodoxin and 128 for CheY) are superimposed with an r.m.s. deviation of 3.0 Å. The alignment used for this figure differs in some respects from the best alignment predicted by the pattern matching program. In general, the precise alignment predicted by the program will depend on the gap penalties used. Analogous secondary structures will tend to be shifted with respect to each other if such shifts can reduce the number of gaps required to achieve a high score. For example, shifting the alignment of two helices by one helical turn may not make a large difference in the score of the aligned regions, but could save the insertion of an additional gap.

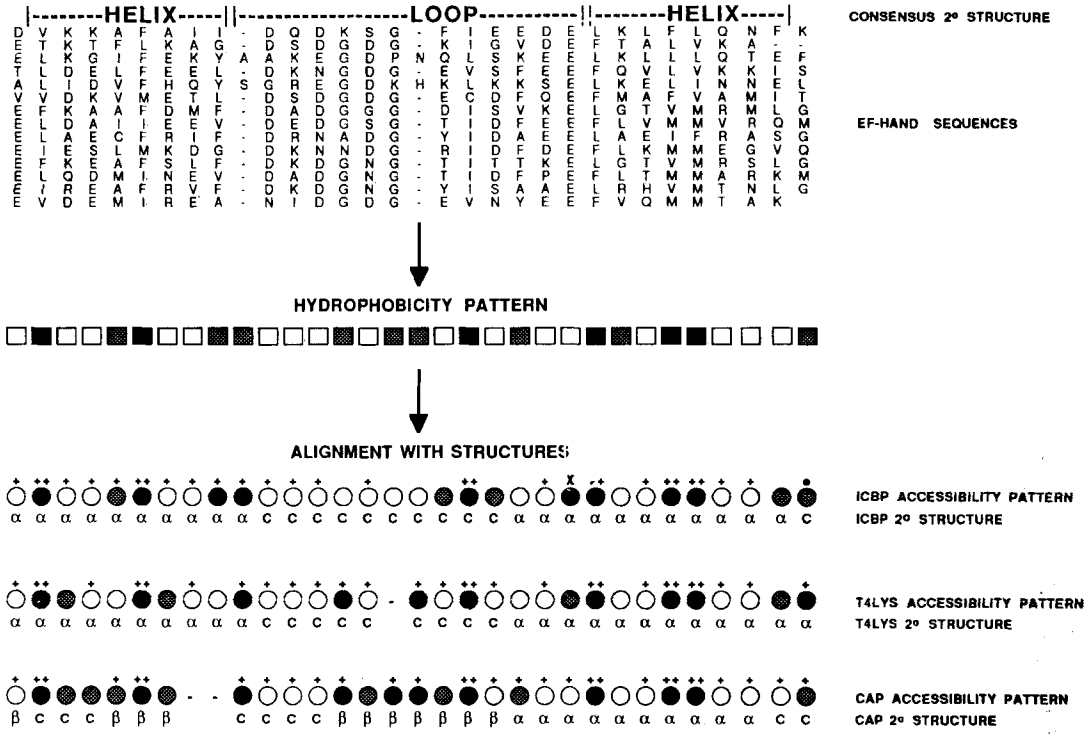


Fig. 6. Generation of the hydrophobicity pattern of EF-hand sequences and the alignments with the solvent accessibility patterns for the three proteins that gave the top scores. The amino acid sequences and consensus secondary structure are from Szebenyi et al.[12] The hydrophobicity pattern is shown below the sequences. Filled-in boxes correspond to members of the H_1 hydrophobicity group, shaded boxes to members of the H_2 group and open boxes to members of the H_3 group. The lower part of the figure shows the aligned solvent accessibility patterns of the proteins that gave the top three scores: vitamin D dependent calcium binding protein from bovine intestine [ICBP; residues 5 through 36], T4 lysozyme [T4-Lys; residues 41 through 71] and catabolite gene activating protein [CAP; residues 291 through 320]. Filled-in circles correspond to members of the B_1 group, shaded circles to members of the B_2 group, and open circles to members of the B_3 group. Gaps in the alignment are indicated by dashes. Symbols above the accessibility patterns indicate the quality of the matches. + + indicates a score of greater than 0.5; +, a score between 0 and 0.5; and an X, a score less than 0.5 The secondary structure of the three proteins is shown below each accessibility pattern. α denotes an α-helix; β, a β-strand; and C, a coil region.

lect globin sequences from a sequence database.[3] Here, we have shown that a hydrophobicity pattern can be correlated directly with structural features of the appropriate protein fold(s).

Our results are consistent with the idea that burying hydrophobic residues is of paramount importance in determining the conformation and stability of proteins. A recent mutagenic study of hydrophobic core residues of λ repressor suggests that the most important feature of interior residues is their hydrophobic character, and that the precise identity of the hydrophobic residue (i.e., its size and shape) is less important.[27] Similar conclusions have been reached from examination of homologous protein structures.[4,5] Apparently, protein structures are relatively tolerant of small changes in the shape and volume of amino acids in their interiors, but only in rare instances can they tolerate hydrophilic residues in the core.[5] This intolerance of hydrophilic residues in solvent-inaccessible positions is an important factor in the effectiveness of our alignment procedure.

The requirement that solvent-inaccessible residues be hydrophobic necessarily imposes sequence constraints on any polypeptide that adopts a particular fold. Our results suggest that the converse also is true: A given pattern of amino acid hydrophobicity imposes constraints on the number and kinds of conformations which can be adopted by a particular amino acid sequence. Theoretical studies using strings of "polar" and "non-polar" elements in a simple lattice model support the same conclusion.[28] Continued studies of how the pattern of polar and non-polar groups in an amino acid sequence can define a tertiary structure should help us to understand how protein sequence determines three-dimensional structure.

ACKNOWLEDGMENTS

We thank Jeff Stock for generously providing his aligned sequences for proteins in the CheY family and Anne Stock and Jeff Stock for providing coordinates for the α-carbons of CheY. We thank Will Gilbert and Upul Obeysekare for help with the computing. Program development was assisted by the University of Wisconsin Genetics Computer Group procedure library.[29] This work was supported by the Howard Hughes Medical Institute and by NIH grant AI-15706.

REFERENCES

1. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14:1–63, 1959.
2. Privalov, P.L. Stability of proteins. Small globular proteins. Adv. Protein Chem. 33:167–241, 1979.
3. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold. Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
4. 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.
5. Lesk, A.M., Chothia, C. Evolution of proteins formed by beta-sheets. The core of the immunoglobulin domains. J. Mol. Biol. 160:325–342, 1982.
6. Perutz, M.F., Kendrew, J.C., Watson, H.C. Structure and function of haemoglobin. Some relations between polypeptide chain configuration and amino acid sequence. J. Mol. Biol. 13:669–678, 1965.
7. Taylor, W.R. Identification of protein sequence homology by consensus template alignment. J. Mol. Biol. 188:233–258, 1986.
8. Bowie, J.U., Sauer, R.T. Identifying determinants of folding and activity for a protein of unknown structure. Proc. Natl. Acad. Sci. USA 86:2152–2156, 1989.
9. Reidhaar-Olson, J.F., Sauer, R.T. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241:53–57, 1988.
10. Dickerson, R.E. Cytochrome c and the evolution of energy metabolism. Sci. Am. 242:136–153, 1980.
11. Oochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S., Morita, Y. Structure of rice ferricytochrome c at 2.0 Å resolution. J. Mol. Biol. 166:407–418, 1983.
12. Szebenyi, D.M.E., Obendorf, S.K., Moffat, K. Structure of vitamin D-dependent calcium binding protein from bovine intestine. Nature 294:327–332, 1981.
13. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tsumi, M. The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
14. Banner, D.W., Kokkinidis, M., Tsernoglou, D. Structure of the colE1 rop protein at 1.7 Å resolution. J. Mol. Biol. 196:657–675, 1987.
15. McKay, D.B., Weber, T.A., Steitz, T.A. Structure of catabolite gene activator protein and 2.9 Å resolution. Incorporation of amino acid sequence and interactions with cyclic AMP. J. Biol. Chem. 257:9518–24, 1982.
16. Anderson, W.F., Ohlendorf, D.H., Takeda, Y., Matthews, B.W. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. Nature 290:754–758, 1981.
17. Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science 242:899–907, 1988.
18. Jordan, S.R., Pabo, C.O. Structure of the lambda complex at 2.5 Å resolution: details of the repressor-operator interactions. Science 424:893–899, 1988.
19. Fauchere, J., Pliska, V. Hydrophobic parameters π of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur. J. Med. Chem. Chim. Ther. 18: 369–375, 1983.
20. Lee, B. Richards, F.M. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.
21. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
22. Gibrat, J., Garnier, J., Robson, B. Further developments of protein secondary structure prediction using information theory. J. Mol. Biol. 198:425–443, 1987.
23. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
24. Stock, A.M., Mottenen, J.M., Stock, J.B., Schutt, C.E. Three-dimensional structure of CheY, the response regulator of bacterial chemotaxis. Nature 337:745–748, 1989.
25. Tufty, R.M., Kretsinger, R.H. Troponin and parvalbumin calcium binding regions predicted in myosin light chain and T4 lysozyme. Science 187:167–169, 1975.
26. Sweet, R.M., Eisenberg, D. Correlation of sequence hydrophobicities measures similarlity of three-dimensional protein structures. J. Mol. Biol. 171:479–488, 1983.
27. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature 339: 31–36, 1989.
28. Lau, K.F., Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22:3986–3997, 1989.
29. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.