A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure

JAMES U. BOWIE, ROLAND LÜTHY, DAVID EISENBERG

The inverse protein folding problem, the problem of finding which amino acid sequences fold into a known three-dimensional (3D) structure, can be effectively attacked by finding sequences that are most compatible with the environments of the residues in the 3D structure. The environments are described by: (i) the area of the residue buried in the protein and inaccessible to solvent; (ii) the fraction of side-chain area that is covered by polar atoms (O and N); and (iii) the local secondary structure. Examples of this 3D profile method are presented for four families of proteins: the globins, cyclic AMP (adenosine 3′,5′-monophosphate) receptor-like proteins, the periplasmic binding proteins, and the actins. This method is able to detect the structural similarity of the actins and 70- kilodalton heat shock proteins, even though these protein families share no detectable sequence similarity.

AS A RESULT OF THE MOLECULAR BIOLOGY REVOLUTION, we now know 50 times the number of protein sequences as three-dimensional (3D) protein structures (Fig. 1). This disparity hinders progress in many areas of biochemistry because a protein sequence has little meaning outside the context of its 3D structure. The disparity is less severe than the numbers might suggest, however, because different proteins often adopt similar 3D folds (*1, 2*). As a result, each new protein structure can serve as a model for other protein structures. These structural similarities probably reflect the evolution of the current array of protein structures from a small number of primordial folds (*3–5*). If the number of folds is indeed limited, it is possible that crystallographers and nuclear magnetic resonance spectroscopists may eventually describe examples of essentially every fold. In that event, protein structure prediction would reduce, at least in crude form, to the inverse protein folding problem—the problem of identifying which fold in this limited repertoire a given sequence adopts.

The inverse protein folding problem is most often approached by seeking sequences that are similar to the sequence of a protein whose structure is known. If a sequence relation can be found, it can often be inferred that the protein of unknown structure adopts a fold similar to the protein of known structure. The strategy works well for closely related sequences, but structural similarities can go undetected as the level of sequence identity drops below 25 percent, the level Doolittle has called "the twilight zone" (*6, 7*).

The authors are in the Molecular Biology Institute and the Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024–1570.

A more direct attack on the inverse protein folding problem was taken by Ponder and Richards (*8*), who adopted quite literally the suggestion of Drexler (*9*) and Pabo (*10*) that one should search for sequences that are compatible with a given structure. In their "tertiary template" method, the backbone of a known protein structure was kept fixed and the side chains in the protein core were then replaced and tested combinatorially by a computer search to find which combination of new side chains could fit into the core. A set of core sequences was thereby enumerated that could in principle be tolerated in the protein structure. In this manner, the method of tertiary templates provides a direct link between 3D structure and sequence.

The rules used to relate 1D sequence and 3D structure in the tertiary template method may be excessively rigid. Proteins that fold into similar structures can have large differences in the size and shape of residues at equivalent positions (*11–22*). These changes are tolerated not only because of replacements or movements in nearby side chains, as explored by Ponder and Richards, but also as a result of shifts in the backbone (*13, 16, 17, 23, 24*). Moreover, insertions and deletions, which are commonly found in related protein structures, were not considered in the implementation of tertiary templates. In order to describe realistically the sequence requirements of a particular fold, the constraints of a rigid backbone and a fixed spacing between core residues must somehow be relaxed.

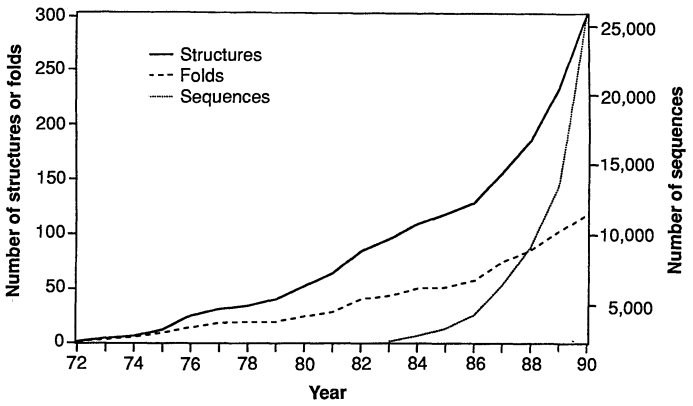

Fig. 1. The determination of amino acid sequences (right-hand scale) is outpacing the determination of 3D structures (left-hand scale) by a factor of 50. Also the number of structures is increasing faster than the number of folds: the cumulative number of structures deposited through 1990 is roughly twice the number of distinctly different protein folds. The number of sequences is the number deposited in the PIR database (*57*). The number of structures is the number of coordinate sets deposited in the Brookhaven Protein Data Bank (*58*), eliminating structures that differ only by a bound ligand, mutation, or space group. The number of folds is a subjective estimate of the number of "distinctly different structures," and should be regarded as having an uncertainty of at least ±20 in 1990.

Table 1. A comparison of a sequence homology search and a compatibility search with CRP. All proteins with Z scores greater than 6.0 in either the sequence homology search or the compatibility search are listed. Z score (1D) refers to the scores obtained from a sequence homology search with a sequence profile constructed with the *Escherichia coli* CRP sequence. Z score (3D) refers to the scores obtained from a structure compatibility search with a 3D profile constructed from the *E. coli* CRP structure (*38*). Percent identity refers to the percentage of identical amino acids in the sequences aligned with the program BESTFIT (*56*). For the sequence homology search, a gap-opening penalty of 4.5 and a gap-extension penalty of 0.05 was used. For the structure compatibility search, a gap-opening penalty of 5.0 and a gap-extension penalty of 0.05 was used. In the sequence homology search, the next highest scoring protein after fnr, Bam HI–ORF4 protein from Fowlpox virus, had an insignificant Z score of 4.90.

Protein	Z score (3D)	Z score (1D)	Percent identity
cAMP receptor protein—*E. coli* (CRP)	46.53	72.99	100.0
cAMP receptor protein—*Salmonella typhimurium* (CRP)	44.13	72.45	99.5
Hypothetical 24.1-kD protein—*Lactobacillus casei*	11.84	12.74	25.6
Regulatory protein fixK—*Rhizobium meliloti*	10.65	9.26	21.1
Regulatory protein fnr—*E. coli*	9.20	7.03	21.2
Protein kinase, cGMP-dependent—bovine	8.24	—	22.0
Protein kinase type III regulatory chain—fruit fly	6.62	—	20.9
DNA polymerase accessory protein 44—bacteriophage T4	6.58	—	19.7
Protein kinase type II regulatory chain—fruit fly	6.47	—	20.9
Protein kinase, cAMP-dependent, regulatory chain II-α—human	6.33	—	21.2
Protein kinase type I regulatory chain—fruit fly	6.15	—	20.9
Protein kinase, cAMP-dependent, type II regulatory chain—bovine	6.06	—	20.9

Overview of 3D compatibility searching with 3D structure profiles. Our method, outlined in Fig. 2, extends the link between 3D structures and sequences, but in a way that simulates the malleability of real proteins. We start with a known 3D structure and determine three features of each residue's environment: (i) the total area of the side chain that is buried by other protein atoms; (ii) the fraction of the side-chain area that is covered by polar atoms or water; and (iii) the local secondary structure. Based on these parameters, each residue position is categorized into an environment class. In this manner, a 3D protein structure is converted into a 1D string, like a sequence, which represents the environment class of each residue in the folded protein structure. We then seek the most favorable alignment of a protein sequence to the environment string.

How can this environment string be aligned to a protein sequence? The method relies on the clear preferences of each of the 20 amino acids for different environmental classes. For example, it is rare to find a charged residue buried in a nonpolar environment. Thus, by determining the environment class of a given position in a protein structure, it is possible to assign a score for finding each of the 20 amino acid types at that position in some related protein structure. We call these scores 3D-1D scores. The 3D-1D scores can then be used in a sequence alignment algorithm to find the best alignment of amino acid sequences to the environment string. The quality of alignment is taken as a measure of the compatibility of the sequence with the 3D structure. The method simulates the malleability of protein structures because no rigid tests for compatibility are applied. In particular, gaps are allowed in the alignment and unfavorable amino acids can be placed at any position, provided these low scores are overcome by enough favorable amino acid–environment pairings (high 3D-1D scores). Because the quality of the alignment to an environment string is not related to sequence similarity in any simple way, we call the sequence database searches

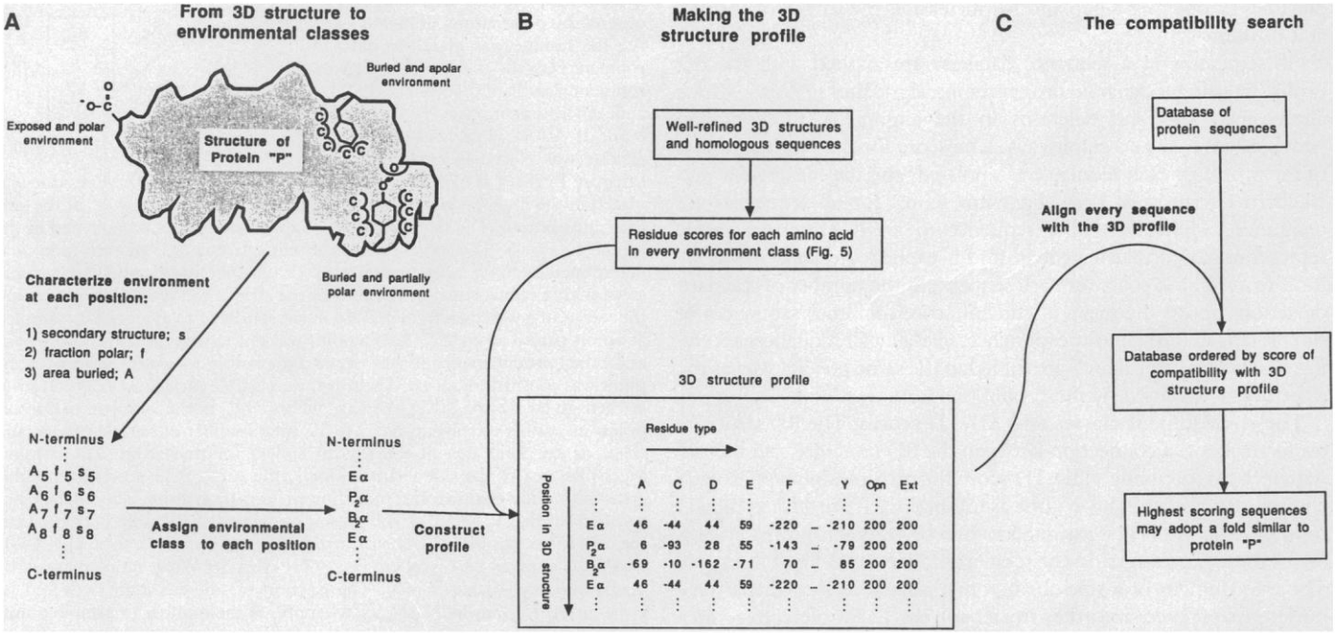

Fig. 2. Schematic description of the construction of a 3D structure profile (**A** and **B**) and of a 3D compatibility search of the sequence database (**C**). The 3D structure profile shown at the bottom of (B) is a portion of the profile for sperm whale myoglobin (Fig. 3), giving scores for only four positions of the structure (corresponding to residues 5, 6, 7, and 8) and for only 6 of the 20 amino acids.

Fig. 3. An example of a 3D profile. The example shows the first ten positions of the sperm whale myoglobin 3D profile (*59*). This profile was used in the compatibility search of Fig. 6. The environment group is listed for each position, followed by scores for placing each of the amino acids at that position. The actual profile is 153 positions long, the length of the sperm whale myoglobin sequence. The scores placed in each row are the 3D-1D scores of Fig. 5, multiplied by 100. The most effective gap penalties are determined empirically. In this case, gaps in helical regions were forbidden by setting very high gap penalties for the helical positions (positions 3 through 10 in the profile). In contrast, relatively low gap opening (Opn) and gap extension (Ext) penalties were used for the coil regions (positions 1 and 2).

		Amino acid type													Gap penalty	
Position in fold	Environment class	A	C	D	E	F	G	...	R	S	T	V	W	Y	Opn	Ext
1	E	12	-46	22	3	-190	113	...	-32	32	12	-91	-214	-94	2	0.02
2	B_2	-66	-5	-128	-135	105	-166	...	-80	-117	-76	60	102	112	2	0.02
3	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
4	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
5	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
6	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
7	B_2α	-69	-10	-162	-71	90	-149	...	6	-147	-150	68	50	85	200	200
8	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
9	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
10	B_1α	-66	-73	-197	-174	132	-253	...	-167	-273	-129	66	100	18	200	200
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.

using the environment strings 3D compatibility searches to distinguish them from homology searches.

3D structure profiles. In order to search a sequence database for the proteins most compatible with an environment string, we used the Profile method (*25, 26*), which was originally developed for detecting sequence homology but is sufficiently general to be expanded to our new purpose. A profile is a position-dependent scoring table in which each position is assigned 20 scores for the likelihood of finding any of the 20 amino acids at that position. In previous implementations of the Profile method, these scores were based on information from families of sequences (*27, 28*). What distinguishes the present 3D structure profiles from sequence profiles is that now the profile scores are the 3D-1D scores computed from the environments of residues in a 3D structure, not from sequences.

Part of the 3D structure profile for sperm whale myoglobin is shown in Fig. 3. Each row in the 3D structure profile represents an amino acid position in the 3D structure. The second column gives the environment class of that position in the folded protein (described below). The following 20 columns give the 3D-1D score for placing each of the 20 amino acid types in the environment found at that position in the structure. The last two columns give the penalties of opening a gap and for increasing the length of the gap at a position.

All sequences in a sequence database are aligned with the 3D profile by using a dynamic programming algorithm (*29, 30*), which allows insertions and deletions in the alignment. Optimal gap penalties were chosen empirically. The score for the best alignment of the profile to each sequence is tabulated, and the mean value and standard deviation of best alignment scores for all sequences are computed. The match of a sequence to a 3D structure profile representing a particular protein fold is expressed quantitatively by its Z score. The Z score for each sequence is the number of standard deviations above the mean alignment score for other sequences of similar length (*26*). In our experience, virtually all sequences receiving Z scores greater than 7 are folded in the same general way as the structure represented by the profile.

The environment classes and 3D-1D scores. The 3D structure profile makes the connection between the 3D structure and the 1D sequence by specifying a 3D-1D score for each residue type in each environmental class. This is done as follows. Each position in the 3D protein structure is first assigned to one of 18 environment classes. Six of these represent side-chain environments, as defined in Fig. 4. The environment of a side chain is first classed as buried, partially buried, or exposed according to its solvent-accessible surface area (*31, 32*). The buried and partially buried residue environments are further subdivided based on the fraction of the environment consisting of polar atoms (*33*). The buried class is subdivided into three classes, labeled B_1, B_2, and B_3 in order of increasing environmental polarity. Similarly, the residue positions in the partially buried class are subdivided into two types, labeled P_1 and P_2 in order of increasing polarity. Since we treat water as polar, exposed positions are necessarily in a polar environment. Consequently, the exposed side-chain category, labeled E, is not subdivided into polarity classes. To account for the slight preferences of certain residue types to be in particular secondary structures, residues in the side-chain environment classes are further distributed into three secondary structure types, α helix, β sheet, and other, to give a total of 18 environment classes.

The 3D-1D scores for matching the 20 amino acids with the 18 environment classes are given in Fig. 5. In general, residues with large hydrophobic side chains are found in the buried classes B_1, B_2, and B_3, whereas hydrophilic residues are favored in the exposed class E. If, however, a buried position has a polar environment (an

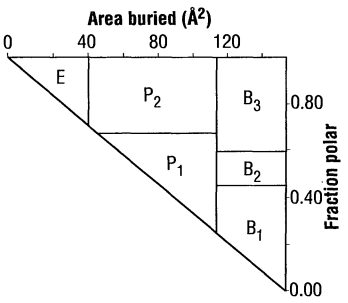

Fig. 4. The six side-chain environment categories. Two environmental characteristics were determined for each side chain: A, the total area buried in the protein structure; and f, the fraction of the side-chain area covered by polar atoms. If $A > 114$ Å^2, the residue was placed in environment class B_1 if $f < 0.45$, environment class B_2 if $0.45 \leq f < 0.58$, and environment class B_3 if $f \geq 0.58$. If $40 < A \leq 114$ Å^2, the residue was placed in environment category P_1 if $f < 0.67$ and environment class P_2 if $f \geq 0.67$. A residue was placed in the exposed environment category E if less than 40 Å^2 of the side chain was buried. The determination of the cutoff values is explained in the legend to Fig. 5. The solvent-accessible surface area (*31*) of each atom was determined by first placing imaginary "solvent spheres" around each protein atom with a radius equal to the sum of the atom's van der Waals radius and the radius of a water molecule. The solvent sphere of each atom was sampled at points placed every 0.75 Å. If a point was not within the solvent sphere of any other protein atom, it was deemed accessible to water, otherwise the point was considered buried. The solvent-accessible surface area of each atom is then given by $(N_{acc}/N_{total})Area_{ss}$, where N_{acc} is the number of sample points accessible to solvent, N_{total} is the total number of sample points, and $Area_{ss}$ is the total area of the solvent sphere for that atom. The solvent-accessible area of the side chain is simply the sum of the solvent-accessible areas of the side-chain atoms, including the α carbon atom. The total area of a side chain that is buried in the protein is defined as the difference between the solvent-accessible side-chain area in the protein and in a Gly-X-Gly tripeptide as given by Eisenberg *et al.* (*33*). Van der Waals radii are given by Richmond and Richards (*60*). The fraction of side-chain area covered by polar atoms is given by N_p/N_{total}, where N_p is the number of sample points covered by polar atoms or exposed to solvent. Sample points covered by atoms of the side chain itself were not counted. If a sample point was within the solvent sphere of both a polar and a nonpolar atom, the closer atom took precedence.

environment with potential hydrogen bond donors and acceptors), it should be less unfavorable to place polar side chains at that position. This trend is evident among the polar residues. For example, glutamine has an unfavorable 3D-1D score in the most nonpolar, buried environment B_1, but scores favorably in the polar, buried environment B_3. Within each environmental class, the preference for the secondary structure types generally follow the trends found in earlier studies. For example, according to the Chou and Fasman propensities (*34*), lysine has a higher propensity to be in a helix than in a sheet. A similar trend is seen in Fig. 5. In short, the table of 3D-1D scores provides the link of 3D structure to 1D sequence in the 3D structure profile method in the same way that the Dayhoff mutational matrix (*27, 35*) supplies the link between two sequences in the earlier sequence profile method (*25*).

3D compatibility search with a 3D structure profile for myoglobin. A demonstration that a 3D structure profile can actually detect sequences compatible with a known 3D structure is offered by the well-characterized globin family (*36*). In Fig. 6 the *Z* scores are shown for all sequences in the database aligned to a 3D structure profile constructed from the coordinates of sperm whale myoglobin (*37*). As shown, 511 of the 544 globin sequences score more highly than any nonglobin sequence. The results shown in Fig. 6 from the 3D structure profile are qualitatively similar to the results of a sequence profile (*25*) constructed from the myoglobin sequence, but differ in two significant aspects. First, because no specific-sequence information was used to construct the profile, sperm whale myoglobin is not the highest scoring protein sequence in the database. In a sequence homology search, the sperm whale myoglobin sequence must be the highest scoring sequence as it would produce a perfect match. Second, the 3D structure profile was somewhat more selective for globin sequences than is the sequence profile computed from the sperm whale myoglobin sequence. In general we find that a 3D structure profile is less sensitive to specific sequence relations and more sensitive to general structural similarity than a sequence homology search.

3D compatibility search with a 3D structure profile of cylic AMP receptor protein. The greater sensitivity of a 3D compatibility search over a sequence homology search in detecting distant structural relations is also seen in the case of the cyclic AMP (adenosine 3,5′-monophosphate) receptor protein (CRP). CRP is a DNA binding protein responsible for the activation of transcription when bound to the effector molecule cAMP. Its sequence is similar to those of a number of other DNA binding proteins as well as to the cAMP-dependent protein kinase family (*38–42*). In Table 1 the result of a sequence homology search in which a profile was constructed from the CRP sequence is compared with the result of a 3D compatibility search that made use of a 3D profile of the CRP structure. Both profiles detect significant relations between CRP and the fnr and FixK proteins, both known DNA binding proteins, as well as a hypothetical protein from *Lactobacillus casei*. The 3D profile, however, also detects a structural relation between CRP and the cAMP-dependent protein kinase family that the sequence profile does not. Clearly, the 3D compatibility search is able to detect distant relations, well below the level of 25 percent sequence

Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-1.16
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.59	-2.16
$B_2\alpha$	0.50	0.90	0.85	1.01	0.63	0.68	1.12	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\beta$	0.01	1.18	1.06	0.76	1.31	1.06	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.66	-1.66	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\alpha$	0.92	-0.03	0.58	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.96	-0.98	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.36	-1.05	1.01	0.10	0.66
$P_1\alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.95	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.52	-0.21	-0.28
$P_1\beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-2.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.31	-0.62	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.63	-0.13	-0.61	0.38	-1.12	-0.74	-1.29
$P_2\alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.26	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.61	0.50
$P_2\beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.98	-1.29	-0.57	0.84	0.59	-0.08	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.86	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
$E\alpha$	-1.35	-2.20	-2.10	-1.58	-2.76	-1.10	-0.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
$E\beta$	0.64	-0.90	0.30	-1.66	-1.47	-1.74	-0.68	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.03	0.22	-0.25	-0.14	-0.32

The 3D-1D scoring table. The scores for pairing a residue *i* with an nent *j* is given by the information value (*61*),

$$\text{3D-1D score } ij = \ln\left(\frac{P(i:j)}{Pi}\right)$$

where $P(i:j)$ is the probability of finding residue *i* in environment *j* and *Pi* is the overall probability of finding residue *i* in any environment. These probabilities were determined from a database of 16 known protein structures and sets of homologous sequences aligned to the sequence of known structure as described in Lüthy *et al.* (*28*). For each position in the aligned set of sequences, we determined the environment category of the position from the known structure and counted the number of each residue type found at the position within the set of aligned sequences. A residue type was counted only once per position. For example, if there were ten aspartates and one glycine found at a position in a set of aligned sequences, then both the Asp and Gly counters were both incremented by only one. The total number of residue replacements in our database was 8273. If the number of residues *i* in an environment *j* was found to be zero, the number was increased to one so that $P(i:j)$ was never zero. Boundaries for the environment categories (shown in Fig. 3) were adjusted iteratively to maximize the total 3D-1D score summed over all residues in our database:

$$\text{Total 3D-1D score} = \sum_{ij} N_{ij} \ln\left(\frac{P(i:j)}{Pi}\right)$$

where N_{ij} is the number of residues *i* in environment *j*. In this case, if N_{ij} was zero, the number was not increased to one. Instead, that term in the sum was treated as zero.

identity, that are often difficult to detect by sequence similarity.

3D compatibility search based on ribose binding protein (RBP) from *Escherichia coli*. The 3D structure profiles confirm and extend proposals that the *lac* and related repressors have structures similar to those of periplasmic sugar binding proteins (*43, 44*). RBP is a periplasmic protein involved in ribose transport. It is a member of a family of periplasmic binding proteins that have related folding patterns, yet little sequence similarity (*45*). Some sequence similarity has been noted between RBP, galactose binding protein (GBP), and arabinose binding protein (ABP), although ABP is the most dissimilar of the three (*45*). Müller-Hill also described sequence similarity between ABP and the *lac* and *gal* repressors (*43*). On the basis of this sequence similarity and the known structure of ABP, a model of the sugar binding site of *lac* repressor has been proposed (*44*).

A sequence search in which a sequence profile was constructed from the RBP sequence is shown in Fig. 7A. The highest scoring proteins in the sequence homology search are indeed RBP and GBP. The next highest scoring protein is *pur* repressor, which is a member of the *lac* repressor family. On the basis of sequence similarity, however, the case for overall structural similarity between RBP and *pur* repressor is relatively weak. The *Z* score for the sequence profile is in the range (less than 7) where spurious relations can occur.

The case for similar structures is greatly strengthened with a 3D compatibility search based on a 3D structure profile made from the RBP structure with the use of coordinates provided by S. Mowbray (Fig. 7B). The two highest scoring proteins are RBP and GBP, but the next highest scoring proteins are all members of the *lac* repressor family. We note that they all have quite significant *Z* scores greater than 8. This result suggests that the effector binding domains of these repressors indeed fold in a manner similar to RBP. ABP is not a high-scoring protein, suggesting that the structures of the *lac* repressor family and RBP are more similar than the structures of ABP and RBP. Moreover, a 3D compatibility search with a 3D profile constructed from the ABP structure did not reveal a significant structural relation between ABP and the repressor proteins.

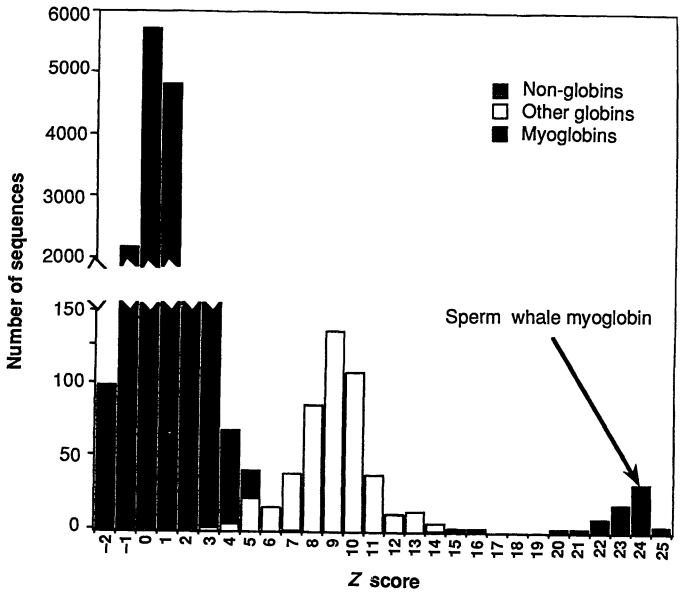

Fig. 6. Results of a compatibility search for the structure of sperm whale myoglobin. Myoglobin sequences are represented by black bars, other globin sequences are represented by white bars, and all other sequences are shown in gray bars. Sperm whale myoglobin is the eighth highest scoring protein (*Z* score = 23.7). Gaps were not allowed in helical regions (as defined in the protein data bank file). In nonhelical regions, a gap-opening penalty of 2.0 and a gap-extension penalty of 0.02 was used.

Thus, the RBP structure may prove to be a better model of the overall structure of the effector binding domains of the *lac* repressor family than the structure of ABP.

3D compatibility search with a 3D structure profile for actin. In 1990 3D structures were reported for the NH_2-terminal domain of the 70-kD bovine heat shock cognate protein (HSC 70) (*46*) and of muscle actin in a complex with deoxyribonuclease I (DNase I) (*47*). Kabsch *et al.* found "unexpected . . . almost perfect structural agreement" between the two structures, although there is virtually no sequence similarity (*47*). The similarity in structure in the absence of sequence similarity would seem to present a severe test of

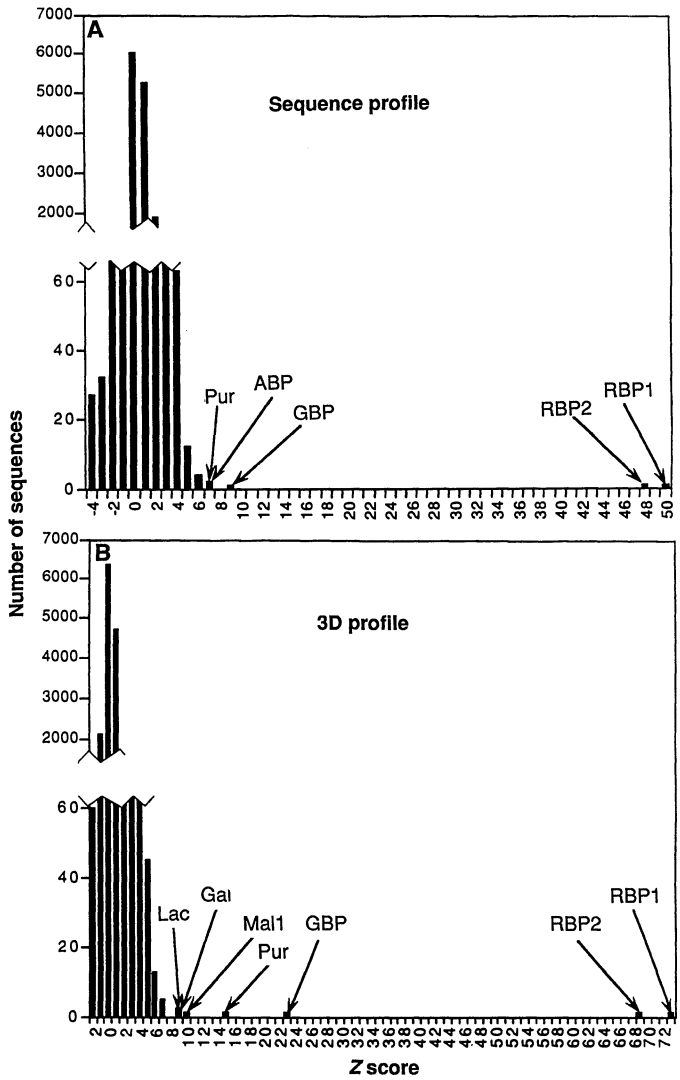

Fig. 7. Comparison of a sequence homology search and a structure compatibility search with ribose binding protein (RBP). (**A**) The results of a sequence homology search with a sequence profile constructed from the *E. coli* RBP sequence. The bar graph shows the number of sequences that give a particular *Z* score. A gap-opening penalty of 4.5 and a gap-extension penalty of 0.05 were used. The highest scoring proteins in (A) are RBP1 (*E. coli* RBP precursor, *Z* score = 49.0), RBP2 (*Salmonella typhimurium* RBP precursor, *Z* score = 47.9), GBP (*E. coli* galactose binding protein, *Z* score = 8.0), Pur (*E. coli* pur repressor, *Z* score = 6.1), and ABP (*E. coli* arabinose binding protein, *Z* score = 6.0). (**B**) The results of a structure compatibility search with a 3D profile constructed from the *E. coli* RBP structure. The bar graph shows the number of sequences that give a particular *Z* score. A gap-opening penalty of 5.0 and a gap-extension penalty of 0.2 were used. The highest scoring proteins labeled in (B) are RBP1 (*Z* score = 72.2), RBP2 (*Z* score = 68.9), GBP (*Z* score = 22.2), Pur (*Z* score = 14.2), Mal (*E. coli* Mal I protein, *Z* score = 9.0), Gal (*E. coli gal* repressor, *Z* score = 8.5), and Lac (*Klebsiella pneumoniae lac* repressor, *Z* score = 8.1).

3D structure profiles. Accordingly, we constructed a 3D structure profile from the actin coordinates and carried out a 3D compatibility search. The top scoring proteins are listed in Fig. 8. After the actin sequences (fgr is an actin-protein kinase fusion protein), the next four highest scoring protein sequences are all members of the 70-kD heat shock protein family, three of which have Z scores greater than 7. Thus, the 3D compatibility search clearly detects the structural correspondence between actin and members of the 70-kD heat shock protein family, a result unobtainable by a sequence homology search.

Relating 1D sequence and 3D structure. Prediction of protein structures from sequences requires a link between 3D structures and 1D sequences. In our method, this link is provided by the reduction of a 3D structure to a 1D string of environmental classes, that is, at the level of sequences. After this first step, the complexity of 3D space is eliminated, but the 3D-1D relation at the heart of the protein folding problem is preserved in the 3D structure profile. That related sequences can be detected by 3D profiles, which contain no direct information about amino acid type, might seem surprising. This result suggests that the environmental classes based on area and polarity are important parameters of folding.

In order to predict protein structures that are only distantly related to some known structure, some way of simulating the malleability of real proteins is required. Distantly related proteins differ in the majority of their side chains and also frequently differ in segments of backbone, particularly in loops that connect segments of secondary structures. The 3D profiles simulate this malleability of proteins by using a statistical approach embodied in the 3D-1D table (Fig. 5) and also in the dynamic programming algorithm. In particular, the tolerance of local unfavorable amino acid pairings and insertions and deletions in the alignments introduce considerable flexibility. The dynamic programming algorithms (*29, 30*) have long been used to align related sequences and more recently, have been applied to the alignment of similar 3D structures (*48, 49*). In our work, we have attempted to bridge the gap between sequence and structure. Thus our method merges two distinct lines in the study of proteins. One is the sequence comparison and database searching line (*50–52*), and the other is that of conformational energy calculations and consideration of stereochemistry and packing (*53, 54*).

Protein	Z score
↑	88.11
69 of 71 Actin Sequences	↕
↓	21.22
Kinase-related transforming protein (fgr)- feline sarcoma virus	17.47
Actin 5C - fruit fly	9.29
68-kD Heat shock protein - mouse	8.12
70-kD Heat shock protein - frog	7.95
70-kD Major heat shock - fruit fly	7.03
70-kD Heat shock cognate protein-bovine	**6.99**
HNRNP complex, protein C - frog	6.74
70-kD Heat shock cognate protein - human	6.31

Fig. 8. Sequence compatibility search with a 3D structure profile for actin (*47*). All sequences that received a Z score of 6.0 or greater are listed. A gap-opening penalty of 5.0 and a gap-extension penalty of 0.2 were used. The fgr protein is the result of a gene fusion between actin and a tyrosine-specific protein kinase (*63*). The bovine HSC70 protein, known to have a similar structure to actin, received a Z score of 6.99 and is shown in bold type.

In a 3D structure profile, stereochemistry and energetics enter implicitly into the assignment of the environmental class through the buried area of its residue and the polarity of atoms in the environment (*31, 55*). The end result is an alignment of a sequence to a 3D structure.

Although 3D profiles permit prediction of some protein structures from amino acid sequences, there are limitations to the predictive ability of the method. The most severe limitation is that no structure can be predicted for which no previous example is known. The reason is simply that each 3D profile is prepared from the atomic coordinates of a structure. Of course, the known "structure" could be a hypothetical or model structure, in which case a 3D compatibility search could reveal sequences consistent with the model. A second limitation arises because a 3D profile can detect only sequences that adopt a similar tertiary structure. Similar topology alone is not sufficient. For example, the 3D compatibility search with a 3D profile of the RBP structure detected only the closest structural relatives of RBP among the many periplasmic binding proteins of similar topology. As structures diverge, the pattern of residue environments that characterize a particular tertiary structure may change too greatly to be recognized. Finally, the structure predicted from a 3D profile is essentially the structure of the protein from which the profile is constructed. Obviously some procedure of energy refinement is necessary to adapt this crude, starting structure to a more accurate structure. Despite these limitations, 3D compatibility searches are clearly able to detect structural relations that may not be apparent by sequence similarity. Thus, compatibility searches should provide a useful complement to sequence homology searches in our attack on the inverse protein folding problem.

REFERENCES AND NOTES

1. M. Levitt and C. Chothia, *Nature* **261**, 552 (1976).
2. J. S. Richardson, *Adv. Prot. Chem.* **34**, 167 (1981).
3. R. L. Dorit, L. Schoenbach, W. Gilbert, *Science* **250**, 1377 (1990).
4. W. Gilbert, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901 (1987).
5. M. Go and M. Nosaka, *ibid.*, p. 915.
6. R. F. Doolittle, *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA 1986).
7. C. Sander and R. Schneider, *Proteins* **9**, 56 (1991).
8. J. W. Ponder and F. M. Richards, *J. Mol. Biol.* **193**, 775 (1987).
9. K. E. Drexler, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5275 (1981).
10. C. Pabo, *Nature* **301**, 200 (1983).
11. W. R. Taylor, *J. Mol. Biol.* **188**, 2333 (1988).
12. J. U. Bowie, and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
13. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
14. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989).
15. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988).
16. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **136**, 225 (1980).
17. _____, *ibid.* **160**, 325 (1982).
18. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
19. W. S. Sandberg and T. C. Terwilliger, *ibid.* **245**, 54 (1989).
20. J. A. Wells, *Biochemistry* **29**, 8509 (1990).
21. B. A. Katz and A. Kossiakoff, *J. Biol. Chem.* **261**, 15480 (1986).
22. T. Alber *et al.*, *Nature* **330**, 41 (1987).
23. T. Alber *et al.*, *Science* **239**, 631 (1988).
24. L. Weaver *et al.*, *Biochemistry* **28**, 3793 (1989).
25. M. Gribskov, A. D. McLachlan, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4355 (1987).
26. M. Gribskov, R. Lüthy, D. Eisenberg, *Methods Enzymol.* **183**, 146 (1990).
27. M. O. Dayhoff and R. V. Eck, *Atlas of Protein Sequence and Structure 1967–68* (National Biomedical Research Foundation, Silver Spring, MD, 1968).
28. R. Lüthy, A. McLachlan, D. Eisenberg, *Proteins* **10**, 229 (1991).
29. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
30. T. F. Smith and M. S. Waterman, *Adv. Appl. Math.* **2**, 482 (1981).
31. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
32. T. J. Richmond and F. M. Richards, *ibid.* **119**, 537 (1978).
33. D. Eisenberg, M. Wesson, M. Yamashita, *Chem. Scr.* **29A**, 217 (1989).
34. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47**, 45 (1978).
35. M. O. Dayhoff and R. M. Schwartz, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1979), p. 353.
36. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).

37. T. Takano, *ibid.* **110**, 537 (1977).
38. I. T. Weber and T. A. Steitz, *ibid.* **198**, 311 (1987).
39. S. Spiro and J. R. Guest, *FEMS Microbiol. Rev.* **75**, 399 (1990).
40. I. T. Weber, K. Takio, K. Titani, T. A. Steitz, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 7679 (1982).
41. I. T. Weber and T. A. Steitz, *Biochemistry* **26**, 343 (1987).
42. I. T. Weber, J. B. Shabb, J. D. Corbin, *ibid.* **28**, 6122 (1989).
43. B. Müller-Hill, *Nature* **302**, 163 (1983).
44. C. F. Sams, N. K. Vyas, F. A. Quiocho, K. S. Matthews, *ibid.* **310**, 429 (1984).
45. N. K. Vyas, M. N. Vyas, F. A. Quiocho, *J. Biol. Chem.* **266**, 5226 (1991).
46. K. M. Flaherty, C. DeLuca-Flaherty, D. B. McKay, *Nature* **346**, 623 (1990).
47. W. Kabsch, H. G. Mannherz, D. Suck, E. F. Pai, K. C. Holmes, *ibid.* **347**, 37 (1990).
48. W. Taylor and C. Orengo, *J. Mol. Biol.* **208**, 1 (1989).
49. A. Sali and T. L. Blundell, *ibid.* **212**, 403 (1990).
50. W. M. Fitch, *ibid.* **16**, (1966).
51. R. F. Doolittle, Ed., *Methods in Enzymology* (Academic Press, New York, 1990), vol. 183.
52. A. D. McLachlan, *J. Mol. Biol.* **62**, 409 (1972).
53. G. Nemethy and H. A. Scheraga, *Q. Rev. Biophys.* **10**, 239 (1977).
54. M. Levitt and A. Warshel, *Nature* **253**, 694 (1975).
55. D. Eisenberg and A. D. McLachlan, *ibid.* **319**, 199 (1986).
56. J. Devereux, P. Haeberli, O. Smithies, *Nucleic Acids Res.* **12**, 387 (1984).
57. D. G. George, W. C. Barker, L. T. Hunt, *ibid.* **14**, 11 (1986).
58. F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977).
59. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
60. T. J. Richmond and F. M. Richards, *J. Mol. Biol.* **119**, 537 (1978).
61. R. Fano, *Transmission of Information* (Wiley, New York, 1961).
62. G. Naharro, K. C. Robbins, E. P. Reddy, *Science* **223**, 63 (1984).
63. We thank W. Kabsch, K. C. Holmes, S. Mowbray, and F. Quiocho for permission to compute 3-D profiles from undeposited coordinates; D. George for information on the number of sequences deposited in each release of the PIR database; A. McLachlan, J. Miller, A. Olson, and J. Perry for discussion; and NIH and the Lita Annenberg Hazen Charitable Trust for support. J.U.B. is a DOE–Energy Biosciences Research Fellow of the Life Sciences Research Foundation.

29 March 1991; accepted 31 May 1991