Research Article

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*).

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.

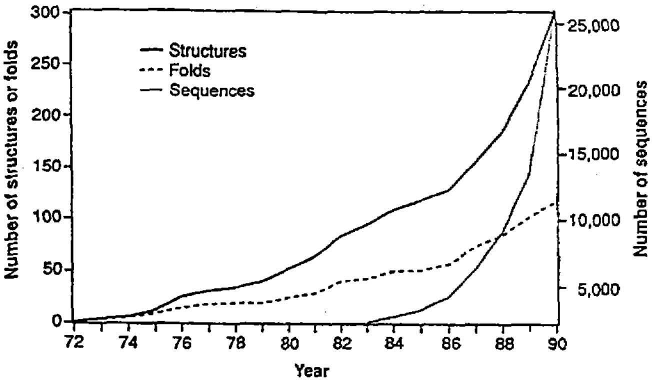


Fig. 1. The determination of amino acid sequences (right-hand scale) is outpacing the determination of 3D structures (left-hand scale) by a factor of 50. Also the number of structures is increasing faster than the number of folds: the cumulative number of structures deposited through 1990 is roughly 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.

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

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

rare to find a charged residue buried in a nonpolar environment. Thus, by determining the environment class of a given position in a

[redacted]

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

Position in fold	Environment class	Amino acid type													Gap penalty	
		A	C	D	E	F	G	...	R	S	T	V	W	Y	Opn	Ext

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

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.

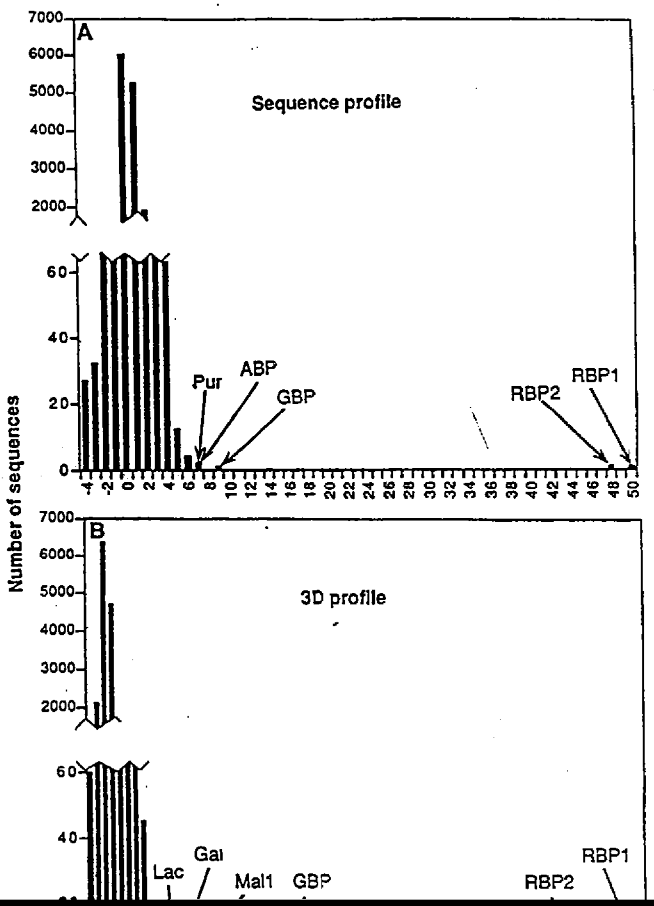
3D compatibility search with a 3D structure profile of cylic AMP receptor protein. The greater sensitivity of a 3D [illegible] identity, that are often difficult to detect by sequence similarity.

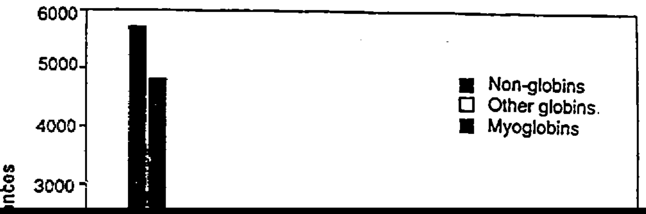
3D compatibility search based on ribose binding protein (RBP) from *Escherichia coli*. The 3D structure profiles confirm and extend proposals that the *lac* and related repressors have structures similar to those of periplasmic sugar binding proteins (*43, 44*). RBP is a periplasmic protein involved 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. 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





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

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

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