Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 9029–9033, October 1992
Biophysics

Protein tertiary structure recognition using optimized Hamiltonians with local interactions

(protein structure prediction/spin-glass theory)

RICHARD A. GOLDSTEIN*, ZAIDA A. LUTHEY-SCHULTEN*†, AND PETER G. WOLYNES*‡

*School of Chemical Sciences, †National Center for Supercomputing Applications, ‡Beckman Institute, University of Illinois, Urbana, IL 61801

Contributed by Peter G. Wolynes, June 26, 1992

ABSTRACT **Protein folding codes embodying local interactions including surface and secondary structure propensities and residue–residue contacts are optimized for a set of training proteins by using spin-glass theory. A screening method based on these codes correctly matches the structure of a set of test proteins with proteins of similar topology with 100% accuracy, even with limited sequence similarity between the test proteins and the structural homologs and the absence of any structurally similar proteins in the training set.**

The ability to predict the native tertiary structure of a protein based solely on its amino acid sequence has long been a goal of computational biophysics. The roughness of a realistic free-energy landscape with its attendant numerous local minima combined with the large number of conformational degrees of freedom for a protein chain have led to attempts to create alternative energy functions in terms of reduced descriptions of the protein configuration. In previous work, we explored the use of associative memory Hamiltonians, a particular type of folding code introduced by Friedrichs and Wolynes (1), which encodes correlations between the sequence of the target protein whose structure is to be determined and the sequences and structures of a set of "memory" proteins. Use of the associative memory formulation allowed us to apply the theory of spin glasses, whose relevance to protein folding has been explored (2–7), to create a nonslavishly realistic energy function for protein tertiary structure prediction optimized so as to facilitate rapid folding while avoiding local energy minima. We demonstrated the ability of an optimized associative memory Hamiltonian to correctly predict low-resolution structures of target proteins with low sequence similarity to the memory proteins by either a screening method or molecular dynamics with simulated annealing (8).

This optimization methodology can be extended to a wide range of different approaches that seek to create simplified Hamiltonians by taking advantage of the data base of known protein structures. One type of Hamiltonian introduced by Eisenberg and coworkers (9) seeks to determine what amino acid sequences are compatible with a particular protein fold

$$\mathscr{H}_{\mathrm{p}} = \sum_{i} \gamma^{\mathrm{p}}(A_i, C_i), \qquad [1]$$

where γ^{p} is a function of both the identity of residue i, A_i, and its context in the protein (e.g., surface accessibility, secondary structure, environmental polarizability), C_i. The values of $\{C_i\}$ are an explicit function of the protein configuration, providing the Hamiltonian's dependence on protein geometry. $\gamma^{\mathrm{p}}(A_i, C_i)$ was calculated based on the frequency of occurrence of particular residues in each possible context. A profile Hamiltonian of this form is capable of encompassing local propensities to particular backbone configurations and multibody potentials such as protein–solvent interactions in a simple way. A major advantage of such a Hamiltonian is that it is easily amenable to dynamic programming techniques, since it is manifestly invariant to insertions and deletions (10). It is not, however, able to generalize many of the specific two-body interactions such as disulfide and hydrogen bonds and salt bridges that give protein structures their specificity. Although their analysis was based on assumptions of statistical independence, which is problematic given the highly cooperative nature of protein folding, Lüthy *et al.* (11) were able to demonstrate the ability of a profile Hamiltonian to distinguish the correctly folded state of a protein from a set of possible configurations.

Another type of Hamiltonian is a contact-potential Hamiltonian such as that analyzed by Miyazawa and Jernigan (12)

$$\mathscr{H}_{\mathrm{c}} = \sum_{i<j} \gamma^{\mathrm{c}}(A_i, C_i, A_j, C_j) u(r_c - r_{ij}), \qquad [2]$$

where γ^{c} is a function of both the identity of residues i and j and their environment in the protein structure, and $u(r_c - r_{ij})$ is a unit step function equal to 1.0 when r_{ij} is less than some cutoff distance r_c. Miyazawa and Jernigan used a quasi-chemical approximation to derive the strength of the various interactions, neglecting many local interactions and propensities, and only roughly quantifying interactions of the protein with the solvent. Skolnick and Kolinski have shown (13) that, in spite of these limitations, with a modified version of Hamiltonian (Eq. 2) combined with exact knowledge of local backbone configuration it is possible to perform a lattice simulation that terminates with a correctly folded structure.

In this paper, we use spin-glass theory to optimize these two more traditional forms of Hamiltonians, both individually and in combination, for a set of training or example proteins. We then demonstrate the ability of the resulting Hamiltonians to predict the structures of a set of test proteins that exhibit only limited homology to the training set.

Spin Glasses and the Multiple-Minima Problem

For either molecular dynamics simulations or *in vivo* folding to yield the native state of a protein, it is necessary not only that the correct structure be stabilized by the various interactions but also that incorrect local minima be destabilized with respect to the native fold. Resolving, or even characterizing, this multiple-minima problem is difficult given the astronomically large number of possible conformations, even in proteins of small size.

The statistical mechanics of the multiple-minima problem in protein tertiary structure prediction can be understood heuristically by using theoretical methods originally applied to spin glasses (14). Spin glasses are spin systems with

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: GCG, Genetics Computer Group; PDB, Brookhaven National Laboratory Protein Data Bank.

random, frustrated interactions that, in the most general case, can compete with a simple nonfrustrated ordering interaction (e.g., ferromagnetism) (15). Reflecting this is a competition between two different phase transitions: an ordering transition (e.g., spin alignment) when the nonfrustrated interactions dominate and there is a gap in the energy spectrum between ordered and disordered states, and a glass transition to a state of frozen-in disorder when the random interactions dominate. For protein folding, these two transitions would correspond to the transition from a liquid-like state to the native folded state (at temperature T_f) or to a glassy state (at temperature T_g) (8, 14). The glass transition represents the transition to a state dominated by multiple minima; after the transition, the folded state is no longer kinetically accessible in the thermodynamic limit. Near the glass transition, the potential energy landscape becomes rough and folding kinetics become slow, nonexponential, and non-Arrhenius. Optimizing a particular protein folding Hamiltonian by maximizing the ratio of T_f/T_g allows structure prediction by using molecular dynamics simulations to be carried out at temperatures where trapping in local minima outside the folded state is unimportant.

Our previous analysis based on a random energy model indicated that T_f/T_g is maximized when $R^2 = \Delta E^2/\delta E^2$ is maximized, where δE is the width of the distribution of energy in the liquid-like states and ΔE is the average energy difference between these states and the correctly folded state (8). This provides a mathematical formulation of how optimality in protein folding can be achieved by stabilizing the correctly folded state with respect to all alternatively folded states.

In general, optimization of $\Delta E^2/\delta E^2$ represents a nonlinear problem, with possible multiple solutions. In the particular case when the Hamiltonian is linear with respect to a set of parameters $\{\gamma_i\}$, it is possible to express the energy of protein μ in its native state by $E^\mu_N = \Sigma_i\lambda^\mu_{N_i}\gamma_i$, and in liquid-like state k by $E^\mu_k = \Sigma_i\lambda^\mu_{k_i}\gamma_i$. One can then write ΔE and δE as $\Delta E = \mathbf{A}\gamma$ and $\delta E^2 = \gamma\mathbf{B}\gamma$, where $\mathbf{A}$ is a vector and $\mathbf{B}$ is a matrix given by

$$A_i = \lambda^\mu_{N_i} - \langle\lambda^\mu_{k_i}\rangle_k \qquad [3]$$

$$B_{ij} = \langle\lambda^\mu_{k_i}\lambda^\mu_{k_j}\rangle_k - \langle\lambda^\mu_{k_i}\rangle_k\,\langle\lambda^\mu_{k_j}\rangle_k, \qquad [4]$$

respectively, where the averages are over all of the liquid-like states k. Maximization of T_f/T_g leads to the explicit form for the optimal γ: $\gamma = \mathbf{B}^{-1}\mathbf{A}$.

γ can be optimized for a set of training proteins by averaging $\mathbf{A}$ and $\mathbf{B}$ over the training set. Once optimized, molecular dynamics can be used to generate the predicted structure of a given target protein. In this paper, however, we continue to use the screening method used by us and others (8, 9, 16–18). As has been pointed out, there is both theoretical and empirical evidence that there are a limited number of structural motifs found in globular proteins (9, 19–22). The tertiary structure prediction problem accordingly can be transformed from the problem of choosing the native state from among the astronomically large number of possible configurations to the problem of selecting between a much smaller set of motifs. Finding the structural motif of lowest energy for a given target protein can be done exhaustively or by using a mean-field self-consistent approach (17). This screening method can be related to lattice calculations, where the lattice is provided by the structures of known proteins (16). As a test of this methodology, we calculate the energy of the target sequence in the configuration of a set of trial structures representing structurally different proteins contained in the Brookhaven National Laboratory Protein Data Bank (PDB), using the Genetics Computer Group (GCG) BESTFIT alignments with default gap parameters to determine corresponding residues in the two proteins (23). The configuration of lowest energy or, alternatively, of largest $R = \mathbf{A}\gamma/(\gamma\mathbf{B}\gamma)^{1/2}$ as calculated in the trial structure configuration using the optimized values for γ is taken to represent the predicted structure of the target protein.

Structural similarities were characterized by q scores, where q measures what fraction of the pairwise distances between corresponding residues in the two proteins match within some tolerance (8)

$$q_{12} = [N(N-1)]^{-1}\sum_{i\neq j}\theta(r^1_{ij} - r^2_{i'j'}). \qquad [5]$$

Here θ is a gaussian function of width 1–2 Å depending on $j - i$, and r^1_{ij} and $r^2_{i'j'}$ are distances between residues i and j of protein 1 and corresponding residues i' and j' of protein 2, respectively. The q scores are related to rms distance deviations (drms values) used by Levitt (19), except q scores emphasize similar regions in the two proteins rather than the portions that are different. A q score > 0.4 was interpreted as indicating structural similarity.

Results

Profile Hamiltonian. Various forms of the profile Hamiltonian (Eq. 1) were optimized. Because the purpose of this work is more exploratory than definitive, a simpler representation of the amino acid environment than that in ref. 9 was used. As distinguished from their 18 categories of secondary structure, neighborhood polarity, and buried side-chain surface area, we encoded two parameters—secondary structure (helix, sheet, turn, or coil) and side-chain surface accessibility (inside or outside)—yielding eight different environments. Secondary structures were defined using the program DSSP (24), with 3–10 and π helices classified as turns. Side-chain accessible surface area was calculated by using the algorithm of Richards (25) as implemented in MIDASPLUS (26), with a residue side chain classified as outside if >15% of its side-chain surface area was accessible to solvent, compared with the surface area of the side chain in a Gly-Xaa-Gly tripeptide (27). With side-chain surface accessibility not defined, energy contributions for glycines depended on secondary structure only. Energy contributions were defined relative to the residue in an inside coil state, resulting in a Hamiltonian with 136 adjustable parameters.

A set of 42 proteins 50–270 residues long selected from the PDB were used as training proteins (28, 29). These proteins represented a range of tertiary folds, including α-helical proteins (e.g., sperm whale myoglobin; 5MBN), β-sheet proteins [e.g., mouse Fab fragment heavy chain; 3HFM(H)], and mixed proteins (e.g., chicken dehydrofolate reductase; 8DFR). We used the x-ray coordinates of a set of proteins 10–50% larger than the training protein, including all possible translations along the sequence, to model the liquid-like states (8, 16). $\mathbf{A}$ and $\mathbf{B}$ were averaged over the set of training proteins and used to calculate an optimal $\gamma^{\mathrm{P}}(A_i, C_i)$ listed at the top of Table 1. Results are qualitatively similar to propensities observed by others (9, 30), such as the correlation between surface propensity and hydrophobicity, the tendency of asparagines and glycines to be in turns, the α-helical propensity of arginine and methionine, and the β-sheet propensity of threonine and tyrosine.

Twenty-three test proteins were selected from the PDB data set, also 50–270 residues long, of various structural classes, so that a structurally homologous protein existed in the training set with maximum sequence similarity of <40% identity based on GCG BESTFIT alignments. For our data base, this cutoff represented the start of the region where sequence similarity did not necessarily imply structural ho-

Table 1. Context propensity vs. buried coil

	Buried			Surface			
aa	Helix	Sheet	Turn	Helix	Sheet	Turn	Coil
			Profile Hamiltonian				
Arg	0.26	0.07	0.18	1.89	0.95	0.85	1.38
Lys	−0.46	−0.20	−0.17	1.30	1.18	1.16	0.87
Asp	−0.13	−0.09	0.10	1.12	0.31	1.41	0.83
Gln	0.26	−0.05	0.23	1.20	1.01	1.28	0.86
Asn	−0.52	−0.66	−0.51	−0.12	0.02	0.94	0.42
Glu	0.20	−0.08	−0.16	1.58	0.87	1.23	0.52
His	0.20	−0.75	−0.66	−1.04	−0.29	−0.80	−0.84
Ser	−0.16	−0.53	−0.33	0.05	−0.23	0.61	0.49
Thr	−0.07	−0.03	−0.20	0.41	1.68	0.31	0.58
Pro	−1.37	−1.55	−0.93	−1.15	−1.35	0.03	−0.31
Tyr	−0.09	0.85	−0.43	−0.98	−0.43	−0.78	−1.08
Cys	−0.75	−0.67	−1.31	−1.03	−1.64	−1.35	−1.13
Gly	−0.53	−0.30	0.92	−0.53	−0.30	0.92	0.00
Ala	0.21	−0.37	−0.61	−0.27	−0.68	−0.57	−0.71
Met	1.83	−0.03	−0.19	−0.74	−0.32	−0.94	−0.58
Trp	−0.36	−0.48	−0.46	−1.78	−1.76	−1.38	−2.04
Leu	0.69	0.65	−0.30	−0.77	−0.67	−0.97	−0.96
Val	0.19	1.07	−0.55	−0.59	−0.43	−0.99	−0.82
Phe	0.72	0.70	0.24	−1.27	−0.09	−0.98	−0.68
Ile	0.91	1.18	−0.21	−0.79	−0.56	−0.96	−0.68
			Combination Hamiltonian				
Arg	0.19	0.41	0.13	1.82	1.00	0.67	1.64
Lys	−0.24	0.00	−0.40	1.36	1.39	1.06	1.16
Asp	0.04	0.27	−0.09	1.01	0.39	1.04	0.89
Gln	0.33	0.14	−0.05	1.25	0.82	0.72	0.71
Asn	−0.32	−0.48	−0.80	−0.24	0.10	0.58	0.36
Glu	0.35	0.29	−0.46	1.33	0.77	0.62	0.43
His	−0.20	−0.19	−0.50	−0.97	0.42	−0.54	−0.07
Ser	0.08	−0.45	−0.57	0.04	−0.48	0.14	0.39
Thr	0.29	0.35	−0.15	0.60	1.87	0.24	0.82
Pro	−1.51	−1.44	−1.21	−1.49	−1.50	−0.49	−0.41
Tyr	−0.08	0.87	−0.23	−0.28	0.21	−0.13	−0.16
Cys	−1.46	−1.23	−1.13	0.02	−0.76	−0.11	0.40
Gly	−0.54	−0.43	0.70	−0.54	−0.43	0.70	0.00
Ala	0.23	−0.18	−0.68	−0.06	−0.32	−0.43	−0.27
Met	1.74	−0.20	−0.29	−0.51	0.03	−0.60	0.28
Trp	−0.77	−0.25	−0.15	−1.25	−0.90	−0.70	−1.15
Leu	0.75	0.63	0.04	−0.03	−0.17	−0.18	−0.19
Val	0.32	1.07	−0.33	−0.12	−0.09	−0.34	−0.10
Phe	0.61	0.47	0.42	−0.72	0.34	−0.19	0.17
Ile	1.14	1.30	0.09	0.00	−0.01	−0.05	0.19

Optimized values $\gamma^{p}(A_i, C_i)$ for the profile Hamiltonian $\mathcal{H}_p$ and the combination Hamiltonian $\mathcal{H}_T = \mathcal{H}_p + \mathcal{H}_c$, showing the propensity of various amino acids (aa) to particular environments, relative to the buried coil state. Environments are as defined in the text. γ values for glycines are independent of side-chain solvent accessibility. Positive values indicate favorable interactions.

mology. Liquid-like states were constructed in the same way as for the training proteins. The set of trial structures for use with the screening method was generated by using the 42 training proteins as aligned to the test proteins using the GCG BESTFIT sequence alignment algorithm with default specifications. In contrast to our previous work, these alignments did not alter the form of the Hamiltonian, but only the set of trial structures. The profile Hamiltonian with optimized $\gamma^{p}(A_i, C_i)$ was able to correctly predict the structure of 20 of the 23 test proteins, failing with cro protein (2CRO), cytochrome *c*-551 (351C), and Bence Jones protein (2RHE). In two of these three cases, the incorrect prediction was lower in energy than the test protein in its native configuration, indicating that the problem was not in the use of the screening method. It is possible that more elaborate profile Hamiltonians, such as that described in ref. 9, would yield better results.

Contact Hamiltonian. There is a wide range of possible forms for the contact potential Hamiltonian (Eq. **2**). In general, to avoid overfitting of the data, simplified encodings of $\{A_i\}$ and $\{C_i\}$ were used. Allowing the value of $\gamma^{c}(A_i, C_i, A_j, C_j)$ to depend on contextual information as encoded in $\{C_i\}$, itself an explicit function of protein configuration, is one way of incorporating feature detection into the Hamiltonian, as proposed earlier (8, 10). Including environmental information deepened the minima corresponding to the native configuration of both the training and test proteins in their native configurations but tended to degrade results with the screening method due to variations in the environments of corresponding residues in similar structures. This variation is more serious for the case of contact potentials than for the profile Hamiltonian, as $\gamma^{c}(A_i, C_i, A_j, C_j)$ is a function of the environment of two different residues. A more exhaustive set of possible configurations used with the screening method might alleviate this problem. r_{ij} was defined as the distance between the C^{β} coordinates of the two residues, with the exception of glycine residues, where the C^{α} coordinate was used.

The best results were obtained with r_c = 10 Å. Grouping similar residues together did not significantly degrade the performance of the contact Hamiltonian. The top part of Table 2 shows the resulting optimized values of $\gamma^{c}(A_i, A_j)$, where residues were grouped into six categories on the basis of mutation rates as embodied in Dayhoff matrices (31), as defined in the legend to Table 2. This categorization itself could be optimized by using spin-glass theory. $\gamma^{c}(A_i, A_j)$ was assumed to be symmetric, resulting in only 21 adjustable parameters. Results are qualitatively what would be expected. The largest possible contribution comes from two cystines in close proximity. Hydrophobic and aromatic residues show a preference to cluster together, away from interactions with the solvent. Acidic and basic residues prefer to associate with each other, suggesting the effect of salt-bridge formation.

This contact Hamiltonian showed similar capabilities to the profile Hamiltonian, also correctly predicting 20 of the 23 test proteins, failing not only for 2CRO and 2RHE but also for leech Eglin-C [2TEC(I)]. In contrast with the results from the profile Hamiltonian, the energy minima of the natively folded test proteins were lower than the incorrect prediction in all three cases, indicating that a more comprehensive screening method might have yielded better results.

Table 2. Contact potentials

	C	S	N	H	V	F
			Contact Hamiltonian			
C	1.43	−0.03	−0.19	−0.11	0.68	0.87
S	−0.03	−0.01	−0.03	−0.02	−0.05	−0.07
N	−0.19	−0.03	−0.04	0.10	−0.29	−0.26
H	−0.11	−0.02	0.10	−0.13	−0.17	−0.07
V	0.68	−0.05	−0.27	−0.17	0.84	0.68
F	0.87	−0.07	−0.26	−0.07	0.68	0.74
			Combination Hamiltonian			
C	1.63	0.03	−0.09	−0.05	0.70	0.87
S	0.03	0.04	0.02	0.04	−0.08	−0.07
N	−0.09	0.02	0.04	0.19	−0.30	−0.24
H	−0.05	0.04	0.19	0.01	−0.17	−0.04
V	0.70	−0.08	−0.30	−0.17	0.75	0.62
F	0.87	−0.07	−0.24	−0.04	0.62	0.73

Optimized values of $\gamma^{c}(A_i, A_j)$ for the contact Hamiltonian $\mathcal{H}_c$ and the combination Hamiltonian $\mathcal{H}_T = \mathcal{H}_p + \mathcal{H}_c$, showing the propensity of amino acids of various categories to be within 10 Å of each other. The categories are as defined in ref. 31: **C** (sulfhydryl: Cys), **S** (small hydrophilic: Ser, Thr, Pro, Ala, Gly), **N** (acid, acid amide, hydrophilic: Asn, Asp, Glu, Gln), **H** (basic: His, Arg, Lys), **V** (small hydrophobic: Met, Ile, Leu, Val), and **F** (aromatic: Phe, Tyr, Trp). Positive values indicate favorable interactions.

Combined Hamiltonian. A complete Hamiltonian was constructed by combining the two Hamiltonians discussed above: $\mathcal{H}_T = \mathcal{H}_p + \mathcal{H}_c$. As in previous examples, $\gamma^p(A_i, C_i)$ depended on the identity of the respective residue and its environment, defined by surface accessibility and secondary structure, while $\gamma^c(A_i, A_j)$ depended on the category of the two residues, as defined above. This resulted in 136 + 21 = 157 adjustable parameters. Values of $\gamma^p(A_i, C_i)$ and $\gamma^c(A_i, A_j)$ are shown at the bottom of Tables 1 and 2, respectively; these values are comparable to those obtained by optimizing each part of the total Hamiltonian separately.

As illustrated in Table 3, the screening method resulted in correct predictions for all 23 test proteins, even when there was as little as 17% sequence identity between the test protein and any of the training proteins. Although the *R* values were lower than those obtained by using the associative memory Hamiltonian (8), this form of the Hamiltonian did not rely on sequence alignment techniques for its construction or on the ability to preassign structural class. The screening method was able to correctly predict the structure of the globin-like proteins even when the one globin protein in the training set used to determine γ, 5MBN, was deleted. Likewise, this Hamiltonian was able to correctly identify the structure of the four immunoglobulin and Bence Jones proteins, even when the one immunoglobulin-like protein [3HFM(H)] was omitted, and to identify the structure of the *L. casei* and *E. coli* dehydrofolate reductase [3DFR and 4DFR(B)] when chicken dihydrofolate reductase (8DFR) was omitted. The optimized Hamiltonian was also able to correctly correlate the sequences of the 13 leghemoglobins in the Swiss-Prot 21 data base based on the structure of sperm whale myoglobin, with only 15–20% sequence identity.

The energetic contributions from the profile and contact part of the combined Hamiltonian are highly correlated for the test proteins in their native configurations (linear correlation coefficient, 0.75), with the contact part of the Hamiltonian contributing about two-thirds of the energy. In contrast, the respective contributions of the two parts to the energy of the liquid-like states are almost completely uncorrelated (linear correlation coefficient, −0.01), offering evidence of the "consistency principle" of Gō (32) that the various types of interactions appear to be consistent with each other in the correctly folded state and that this state does represent a state of "minimum frustration" (3).

Conclusion

The results of this paper show that the optimization techniques developed to apply spin-glass theory to associative memory Hamiltonians may also be used to refine Hamiltonians outside the associative memory framework. Applying this technique to a particularly simple form of Hamiltonian involving local environmental preferences and residue contacts, it is possible to generate an optimized Hamiltonian that can correctly discriminate between similar and nonsimilar possible tertiary structures in all of the test examples tried. The Hamiltonian was able to select the correct fold for test proteins even when no examples of that fold were included in the training set, indicating that the Hamiltonian seems to be learning general principles of tertiary structure formation rather than details of specific structures.

The fact that there seems to be a consistency to different forms of interactions in the correctly folded state but not in the liquid-like states indicates that sensitivity in protein structure prediction can be augmented through the use of multiple forms of interactions. For instance, the predicted structures of 2CRO and 2RHE, based on either the profile Hamiltonian or the contact Hamiltonian, were incorrect; use of the combined Hamiltonian yielded the correct structural homolog for both of these examples. Combining the energy contributions from the Hamiltonians discussed in this paper with the associative memory Hamiltonians may yield even more discrimination between possible structures. Merging these interactions in a way to maximize predictive ability by

Table 3. Tertiary structure predictions using combination Hamiltonian

Target		Prediction			
PDB	Name	PDB	Name	% I	*R*
351C	*P. aeruginosa* cytochrome *c*-551	1CCR	Rice cytochrome *c*	22.20	4.20
1R69	434 repressor (α domain)	1LRD	λ repressor	28.60	5.77
2CRO	434 Cro	1LRD	λ repressor	16.40	3.31
1ALC	Baboon α-lactalbumin	2LYZ	Hen egg white lysozyme	37.20	9.93
1R1A	Human rhinovirus 1A coat (VP2)	2MEV	Monkey mengovirus coat (VP2)	31.80	7.03
2PLV	Human poliovirus coat (VP2)	2MEV	Monkey mengovirus coat (VP2)	37.10	6.32
4RHV	Human rhinovirus coat (VP2)	2MEV	Monkey mengovirus coat (VP2)	35.00	7.19
1TEC	Leech Eglin-C	2SNI	Barley chymotrypsin inhibitor II	37.10	4.93
1FX1	*D. vulgaris* flavodoxin	3FXN	*Clostridium MP* flavodoxin	31.10	9.00
IF19	Mouse Fab (L)	3HFM	Mouse Fab (H)	23.60	10.27
1REI	Human Bence Jones (variable)	3HFM	Mouse Fab (H)	29.80	7.99
2RHE	Human Bence Jones (variable)	3HFM	Mouse Fab (H)	28.70	7.65
3HFM	Mouse Fab (L)	3HFM	Mouse Fab (H)	26.20	8.73
4FAB	Mouse Fab (L)	3HFM	Mouse Fab (H)	21.30	7.11
3ICB	Bovine Ca-binding	5CPV	Carp Ca-binding parvalbumin B	31.30	5.90
1FDH	Human fetal hemoglobin (γ)	5MBN	Sperm whale myoglobin	22.90	8.05
1HDS	Deer sickle cell hemoglobin (β)	5MBN	Sperm whale myoglobin	24.80	6.58
1MBA	Sea hare myoglobin	5MBN	Sperm whale myoglobin	27.90	7.19
2HHB	Human hemoglobin (α)	5MBN	Sperm whale myoglobin	26.20	6.61
2LH4	Lupin leghemoglobin	5MBN	Sperm whale myoglobin	19.00	7.37
2LHB	Sea lamprey hemoglobin	5MBN	Sperm whale myoglobin	26.30	8.17
3DFR	*L. casei* DHFR	8DFR	Chicken DHFR	30.60	7.72
4DFR	*E. coli* DHFR	8DFR	Chicken DHFR	34.00	8.52

Results of screening method with a combination Hamiltonian, showing the test proteins, predicted structural homolog, sequence similarity of the two proteins as measured by percentage identity (% I) using GCG BESTFIT alignments, and the value of $R = \mathbf{A}\gamma/(\gamma\mathbf{B}\gamma)^{1/2}$ of the predicted structure. All of the test proteins were correctly paired with a protein of similar structure, as defined by $q > 0.4$. Organisms used were *Pseudomonas aeruginosa, Desulfovibrio vulgaris, Clostridium MP, Lactobacterium casei*, and *Escherichia coli*. DHFR, dihydrofolate reductase.

using earlier statistical treatments of known protein structures would be problematic, given their high degree of correlation. The techniques that we have developed based on spin-glass theory provide a simple way to optimize any particular combination of forms of interactions. Continued development of these techniques will allow even greater accuracy in structure prediction and give insight into the dominant interactions in folding.

Helpful interactions with Henrik Bohr, David Evensky, Pia Gratton, and Mark Sandrock, and critique of the text by Anthony Crofts, Hans Frauenfelder, and Klaus Schulten are gratefully acknowledged. Computations were carried out at the National Center for Supercomputing Applications in Urbana. This work was supported by National Institutes of Health Grant PHS 5 R01 GM 44557-03.

1. Friedrichs, M. S. & Wolynes, P. G. (1989) *Science* **246**; 371–373.
2. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
3. Bryngelson, J. D. & Wolynes, P. G. (1990) *Biopolymers* **30**, 171–188.
4. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6**, 597–601.
5. Shakhnovich, E. I. & Gutin, A. (1988) *Europhys. Lett.* **8**, 327–332.
6. Shakhnovich, E. I. & Gutin, A. (1989) *Stud. Biophys.* **132**, 47–56.
7. Wolynes, P. G. (1991) in *Spin Glasses and Biology*, ed. Stein, D. (World Sci., New York), in press.
8. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
9. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
10. Friedrichs, M. S., Goldstein, R. A. & Wolynes, P. G. (1991) *J. Mol. Biol.* **222**, 1013–1034.
11. Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992) *Nature (London)* **356**, 83–85.
12. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
13. Skolnick, J. & Kolinski, A. (1990) *Science* **250**, 1121–1125.
14. Sasai, M. & Wolynes, P. G. (1990) *Phys. Rev. Lett.* **65**, 2740–2743.
15. Mezard, M., Parisi, G. & Virasoro, M. A. (1987) *Spin Glass Theory and Beyond* (World Sci., Singapore).
16. Crippen, G. M. (1991) *Biochemistry* **30**, 4232–4237.
17. Finkelstein, A. & Reva, B. (1991) *Nature (London)* **351**, 497–499.
18. Sippl, M. J. & Weitckus, S. (1992) *Proteins* **13**, 258–271.
19. Hinds, D. A. & Levitt, M. (1992) *Proc. Natl. Acad. Sci USA* **89**, 2536–2540.
20. Richardson, J. (1981) *Adv. Protein Chem.* **34**, 167–339.
21. Finkelstein, A. V. & Ptitsyn, O. B. (1987) *Prog. Biophys. Mol. Biol.* **50**, 171–190.
22. Murzin, A. G. & Finkelstein, A. V. (1988) *J. Mol. Biol.* **204**, 749–770.
23. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
24. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577–2637.
25. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
26. Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988) *J. Mol. Graphics* **6**, 13–27.
27. Shrake, A. & Rupley, J. A. (1973) *J. Mol. Biol.* **79**, 351–371.
28. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
29. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Int. Union Crystallogr., Bonn), pp. 107–132.
30. Chou, P. & Fasman, G. (1974) *Biochemistry* **13**, 222–275.
31. George, D. G., Hunt, L. T. & Barker, W. C. (1988) in *Macromolecular Sequencing and Synthesis*, ed. Schlesinger, D. H. (Liss, New York), pp. 127–149.
32. Gō, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.