Protein Science (1996), *5*:1043–1059. Cambridge University Press. Printed in the USA.

Self-consistently optimized statistical mechanical energy functions for sequence structure alignment

K.K. KORETKE, Z. LUTHEY-SCHULTEN, AND P.G. WOLYNES
School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

(RECEIVED December 11, 1995; ACCEPTED March 25, 1996)

Abstract

A quantitative form of the principle of minimal frustration is used to obtain from a database analysis statistical mechanical energy functions and gap parameters for aligning sequences to three-dimensional structures. The analysis that partially takes into account correlations in the energy landscape improves upon the previous approximations of Goldstein et al. (1994, 1995) (Goldstein R, Luthey-Schulten Z, Wolynes P, 1994, *Proceedings of the 27th Hawaii International Conference on System Sciences.* Los Alamitos, California: IEEE Computer Society Press. pp 306–315; Goldstein R, Luthey-Schulten Z, Wolynes P, 1995, In: Elber R, ed. *New developments in theoretical studies of proteins.* Singapore: World Scientific). The energy function allows for ordering of alignments based on the compatibility of a sequence to be in a given structure (i.e., lowest energy) and therefore removes the necessity of using percent identity or similarity as scoring parameters. The alignments produced by the energy function on distant homologues with low percent identity (less than 21%) are generally better than those generated with evolutionary information. The lowest energy alignment generated with the energy function for sequences containing prosite signatures but unknown structures is a structure containing the same prosite signature, providing a check on the robustness of the algorithm. Finally, the energy function can make use of known experimental evidence as constraints within the alignment algorithm to aid in finding the correct structural alignment.

Keywords: homologous modeling; protein sequence alignment; protein structure prediction; statistical mechanical energy functions

Protein folding is a problem of discrimination. This is true for the physical process of folding a protein *in vitro*. It is even more valid for the practical problem of protein structure prediction. Many algorithms for protein structure prediction lead to an ensemble of structures that satisfy, to a modest extent, the *a priori* constraints on protein structure that can be inferred from a database and encoded in a semiempirical energy function (Bowie et al., 1991; Godzik et al., 1992; Goldstein et al., 1992a, 1994; Nishikawa & Matsuo, 1993; Sippl, 1993; Maiorov & Crippen, 1994). Recently, the problem of discrimination in protein folding has been highlighted using a statistical mechanical perspective on the problem of energy function-based structure prediction methods. Previous work along these lines has used only the simplest version of the statistical mechanical treatment of the energy landscape (Goldstein et al., 1992a, 1992b, 1994, 1995). In this paper, we show how more sophisticated approximations can be used that take into account the partial ordering of incorrectly folded or predicted structures. The resulting self-consistently optimized energy functions lead to more reliable matching of sequences to structures and to better alignments.

Discrimination enters the physical chemistry of protein folding through the competition between the driving forces that funnel the flow of protein configuration toward the native structure and the trapping of the molecule in misfolded configurations. Most random heteropolymers have an energy landscape in which trapping occurs, but in which there is no global guidance. Although random heteropolymers have a lowest or ground-state structure, there are traps nearly equivalent in energy that must be discriminated against. Obligate freezing into such a trap occurs during protein folding, when the system reaches a so-called glass transition temperature, T_G. The funneling process, on the other hand, allows folding to occur at a higher temperature, T_F. Only a fraction of the sequences will be thermodynamically foldable above the glass transition temperature. These sequences are said to satisfy the "principle of minimal frustration" (Bryngelson & Wolynes, 1987; Sasai & Wolynes, 1990; Goldstein et al., 1992a, 1992b). For these sequences, the ground-state structure is considerably more stable than the competing traps. Conceptually, for fast folding it is necessary to distinguish the ground state from individual traps in detail. The statistical energy land-

Reprint requests to: P.G. Wolynes, School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801; e-mail: wolynes@aries.scs.uiuc.edu.

scape theory gives estimates for the trap energy once the variance of the energies of generic misfolded structures is known. This estimate is based on the so-called random energy model approximation. Within the REM approximation, the ratio of T_F to T_G that determines (to a large extent) the nature of the folding kinetics can be determined. When the ratio is large, trapping is unimportant and molecular dynamics based on the energy functions can be quite efficient. The maximizing of T_F/T_G is a quantitative formulation of the principle of minimal frustration (Bryngelson & Wolynes, 1987). Indeed, within the REM approximation, the ratio is monotonically related to the difference in energy of the folded state and the typical collapsed states. It is easy to see then that maximizing that ratio not only improves folding kinetics, but also allows one to discriminate most efficiently between the folded state and the bulk of collapsed states by algorithms other than molecular dynamics.

The generality of the discrimination idea was made clear by Goldstein et al. (1994, 1995), who have shown how the quantitative physicochemical criterion based on the principle of minimal frustration also gives the best Bayesian discrimination for the probability of getting a correctly folded structure by sequence-structure alignment if the random energy model landscape is assumed to be correct. This statistical mechanical theory, however, makes clear that the random energy approximation is only partially correct. That approximation is based on a lack of correlations in the energies of states in the landscape, but, when some energy terms are more important than others, correlations that satisfactorily minimize the large energy terms but not the smaller contributions to the energy lead to partial ordering of the competing states. The following Gedankenexperiment makes clear the problem. Suppose we find that the energy function that discriminates correctly folded structures from the bulk of collapsed structures gives a very large interaction between two specific kinds of amino acids because folded structures always possess this particular kind of contact, but it occurs very rarely in randomly collapsed structures. If that interaction energy is too large, when the collapsed structures are allowed to readjust by alignment with insertions and deletions, or readjust by molecular dynamics, an anomalously large fraction of the competing energy states will satisfy this empirical correlation too. After this, in fact, this pair interaction would lead to very little discrimination between the folded structure and the *minimal energy* misfolded structures because those minima have already been partially ordered to satisfy this interaction. This is a consequence of the energy landscape being correlated, and is why the REM approximation would be poorer than expected.

Both the statistical mechanical and Bayesian theory suggest a way in which these correlations can be partially taken into account. When the folded energy is computed, it should not be based on the energy difference of folded structures and all collapsed structures, but should contain the energy difference between the folded structures and the thermally occupied minima in the ensemble of the collapsed structures. This is the appropriate stability gap (Bryngelson et al., 1995). Also, the variance of energies should be computed with the minima rather than all collapsed structures playing the crucial role, at least as a first approximation.

The recipe of maximizing T_F/T_G, taking into account the partial ordering of misfolded structures, is more complicated than the simple algorithm based on the REM because the minima themselves depend on the energy function. This leads to a self-consistent optimization problem, much like that done in the Hartree–Fock approximation of quantum chemistry. The iterative development of energy functions then proceeds along the following lines: A first approximation to the energy function is evaluated by solving the variational problem of maximizing the stability gap in units of the standard deviation of energy of collapsed structures. New structures are generated by alignment of a given set of trial sequences from the learning set against known structures. The new minima are then used to re-estimate the stability gap between correct folds and these minima, and the variance of energy of the minima and this is iterated until self-consistency or maximum degree of discrimination averaged over the training set is achieved. We see that this procedure for inferring energy functions is very much related to the issue of specific negative design in the problem of making foldable proteins *de novo*.

In this paper, we carry out such a self-consistent optimization of statistical mechanical energy functions based on a trial energy function that includes context terms, contact terms, and specific terms for hydrogen bonding. We show that the initial energy functions do lead to mildly correlated energy landscapes and that, upon initial alignment, there is a partial ordering of competing structures. This ordering is largely parallel to the microphase separation of two-letter code lattice models, which has been emphasized in the work of Dill and coworkers (Dill et al., 1995; Dill & Stigter, 1995). When the optimization procedure is self-consistently carried out, discrimination is greatly improved. Results are presented for a test set as well as a large number of predictions of structures where structural similarity might be inferred through functional relationships. We herein describe the energy function and how the progress of each iteration was monitored for higher discrimination. Then, the results of using the self-consistent energy function in threading sequences to putative structures are described. This is followed by a discussion of the effectiveness of the self-consistent energy function. Finally, we conclude with the methodology used in this study.

Energy function

We have designed an energy function to evaluate sequence-structure compatibility in terms of contributions from profile (E_p), pairwise contacts (E_{ct}), hydrogen bonding (E_{hb}), gap penalty (E_g), and satisfaction of experimental constraints (E_{cs}):

$$E_T = E_p + E_{ct} + E_{hb} + E_g + E_{cs}. \tag{1}$$

These energy terms can be expressed as a linear function of energy parameters, γ. The profile energy contribution (E_p), similar to the energy term described in previous works by Goldstein et al. (1992b) and Eisenberg and co-workers (Bowie et al., 1991), is a measure of the propensity of an amino acid to reside in a particular context of amino acids:

$$E_p = \sum_{i=1}^{N} \gamma^p(A_i, SS_i, SA_i), \tag{2}$$

where N denotes the number of residues in a protein and γ^p denotes the energy parameter, which is a function of amino acid identity A_i, secondary structure SS_i, and surface accessibility

SA_i. The secondary structure of the scaffold proteins has been predetermined with the DSSP algorithm (Kabsch & Sander, 1983), and each residue was assigned to be in either an α-helix, β-sheet, turn (β-turn, 3/10 helix), or random coil. The surface accessibility of each residue of the scaffold proteins has been calculated previously by using the algorithm of Richards (1977) as implemented in MidasPlus (Ferrin et al., 1988), and was assigned as outside if more than 18% of the side chain is exposed to the surface; otherwise, the residue was assigned as inside.

The pairwise contact energy contribution (E_{ct}), analogous to the energy term introduced in previous works by Goldstein et al. (1992b) and Miyazawa and Jernigan (1985), has been modified to handle multi-body interactions by monitoring for possible multiple cysteine bond formations to a single cysteine residue. This is conceptually similar to, but distinct from the three-body interactions introduced by Skolnick and coworkers (Godzik et al., 1992). This modification was needed due to the strong interaction between cysteines resulting from the covalent nature of disulfide bonds and the chemical saturation of this covalent interaction:

$$E_{ct} = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \sum_{k=1}^{2} \gamma_k^{ct}(A_i, A_j) u(r_k^{ct} - r_{ij}) + \sum_{cys} min[\gamma_1^{ct}(\mathrm{Cys},\mathrm{Cys}) u(r_1^{ct} - r_{ij})], \quad (3)$$

where γ^{ct} is an energy parameter that is a function of the identities of amino acids A_i and A_j in contact within a cut-off range; and u is a unit step function dependent upon the C_β distance between residues i and j, r_{ij}. This energy function has two cut-off ranges: a short range, where 0.0 Å $< r_1 <$ 5.0 Å, and a long range, where 5.0 Å $< r_2 <$ 12.0 Å. If the C_β distance of residues A_i and A_j are within one of these ranges, the corresponding γ^{ct} is added to the total energy. The γ_1^{ct}(Cys,Cys) is the largest energy parameter. To prevent multiple cysteine bond formation in the mean-field alignment, only unique cysteine-cysteine contacts within the 0.0–5.0 Å range are included in the overall energy.

The hydrogen bonding energy (E_{hb}) monitors a generic description for two types of backbone hydrogen bond patterns between N_i and O_j atoms, α-helix and β-sheet. Different types of hydrogen bonding energy terms have been described by other authors (Nishikawa & Matsuo, 1993; Srinivasan & Rose, 1995). Our form of the hydrogen bond term is:

$$E_{hb} = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \sum_{k=1}^{2} \gamma_k^{hb}(i,j), \quad (4)$$

where N denotes the number of residues in a protein; γ_1^{hb}, an energy parameter that is a function of any residues i and j involved in an α-helix hydrogen bond; and γ_2^{hb}, an energy parameter that is a function of any two residues i and j involved in a β-sheet hydrogen bond. This energy term depends on secondary structure assignments explicitly, and does not deal with hydrogen bonding in turns or with side chains.

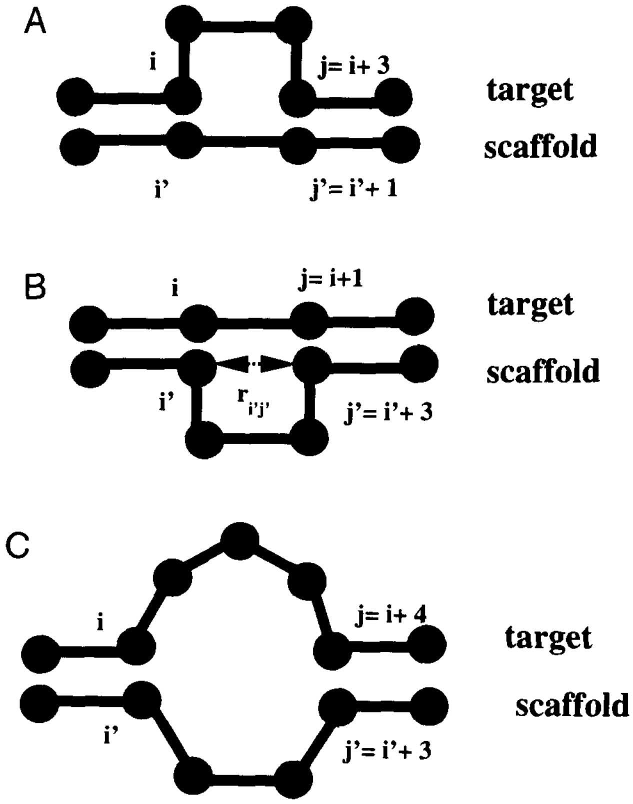

Fig. 1. Examples of the three categories of gaps. **A:** Insertions of the target residues where, if residue i on the target protein is pinned to residue i' on the scaffold protein, target residue j, the next pinned target protein residue, is pinned to the scaffold residue j', with $j > i + 1$, but $j' = i' + 1$. **B:** Deletions of the scaffold residues, where $j' > i'$, but $j = i + 1$ and $r_{i'j'} <$ 7.5 Å. **C:** Bulged gaps were both $j > i + 1$ and $j' > i' + 1$.

The gap energy (E_g) represents three different classes of gaps as shown in Figure 1: insertions ($j - i > 1$ with $r_{i'j'} <$ 3.9 Å, the distance between successive C_α atoms in the scaffold, illustrated by example A), deletions ($j - i = 1$, with 3.0 Å $< r_{i'j'} <$ 7.5 Å, illustrated by example B), and bulges ($j - i > 1$ and $r_{i'j'} >$ 4.0 Å, illustrated by example C). The use of three gap types was first introduced by Zuker (1991). In our work, the initial gap parameters for these three types of gaps were determined by a Bayesian statistical analysis of the distributions for each gap type in correct versus random alignments (Goldstein et al., 1994, 1995). A simple functional form was used to describe the ratio of the correct over the random gap distributions, resulting in a gap energy with the following form:

$$E_g = \sum_{i=k}^{n} \begin{cases} \gamma_1^g + \gamma_2^g(j - i) & \text{if } j - i > 1,\ r_{i'j'} \le 3.9\ \text{Å} \\ \gamma_3^g + \gamma_4^g r_{i'j'} + \gamma_5^g r_{i'j'}^2 & \text{if } j - i = 1,\ 3.0\ \text{Å} < r_{i'j'} \le 7.5\ \text{Å}\ , \\ \gamma_6^g + \gamma_7^g(j - i) + \gamma_8^g \dfrac{r_{i'j'}^2}{(j - i)} & \text{if } j - i > 1,\ r_{i'j'} \ge 4.0\ \text{Å} \end{cases} \quad (5)$$

where n denotes the total number of gaps in an alignment; γ^g, an energy parameter; i and j, the residue positions in the training protein; i' and j', the residue positions in the scaffold; and $r_{i'j'}$ refers to the three-dimensional distance between C_α coordinates of residues i' and j' in the scaffold. For all gap types, no gaps were allowed to be placed within the middle of a secondary structure unit nor was the length of any given gap allowed to exceed 15 residues. This last restriction is based on typical gaps appearing in multiple sequence alignments and can be varied easily.

The constraint energy term (E_{cs}) is used only in our mean-field alignment to aid in guiding an alignment to incorporate experimental data. Possible experimental constraints include distance constraints between certain residues determined from labeling or mutation experiments, as well as structural information from CD experiments.

A set of known structures from various folding classes is chosen to train the energy parameters. Each protein in this set has at least one structural analogue in the Protein Data Bank (PDB). A structural analogue is defined as a known structure that is similar to the native fold. This includes homologues as well as nonrelated sequences. The term structural analogue is used instead of homologue because the energy function aligns two sequences based on structural not genetic information. A training protein's native structure, along with its alignment to each of its known structural analogues, constitute the ensemble of "correct" folds. Misfolded structures are taken to be relatively compact structures with varying degree of correct secondary structure. The full ensemble (3,000 states) of misfolded structures of each training protein are generated by translating the training protein's sequence over protein structures with unrelated folds. Once a training set has been selected, an optimized energy function can then be obtained by maximizing the dimensionless ratio of T_F/T_G, or equivalently the ratio of the stability gap δE between the average energy of the correct folds and the mean energy of the full ensemble of misfolded structures to the standard deviation ΔE of the energies of the misfolded structures (Goldstein et al., 1992a, 1992b). This energy function can be expressed as a linear function of the energy parameters, $\boldsymbol{\gamma}$, in which $\delta E = \mathbf{A}\boldsymbol{\gamma}$ and $\Delta E^2 = \boldsymbol{\gamma}\mathbf{B}\boldsymbol{\gamma}$, where $\mathbf{A}$ and $\boldsymbol{\gamma}$ are vectors and $\mathbf{B}$ is a matrix given by:

$$A_i = \langle \lambda_i \rangle_{correct} - \langle \lambda_i \rangle_{misfolded} \tag{6}$$

and

$$B_{ij} = \langle \lambda_i \lambda_j \rangle_{misfolded} - \langle \lambda_i \rangle_{misfolded} \langle \lambda_j \rangle_{misfolded}, \tag{7}$$

respectively, where the index i denotes a specific energy contribution; λ_i is the frequency that the ith energy interaction occurs in a structure. There are 210 contact interactions for each cut-off range, 160 profile parameters, 8 gap parameters, and 2 hydrogen bond parameters. The $\langle \lambda_i \rangle_{correct}$ denotes the frequency of occurrence of that particular interaction averaged over the native structure of a training protein and its alignments to its structural analogues; and $\langle \lambda_i \rangle_{misfolded}$ is the average frequency of occurrence of that particular interaction in the ensemble of misfolded structures for the given training protein. The solution of the maximization problem for T_F/T_G leads to an explicit form for the optimal $\boldsymbol{\gamma}$; $\boldsymbol{\gamma} = \langle \mathbf{B}^{-1} \rangle \langle \mathbf{A} \rangle$, where $\langle \mathbf{B}^{-1} \rangle$ and $\langle \mathbf{A} \rangle$ are averaged over the set of training proteins.

The thermally occupied minima, defined as the alignments of the training proteins to unrelated folds that are produced with this energy function, are partially ordered to satisfy individual large interaction energy terms. Re-evaluating the energy function by maximizing the ratio of $\delta E/\Delta E$, where the new stability gap is defined as the difference between the mean energy of correct folds and the mean energy of the thermally occupied minima of misfolded structures, and the new standard deviation is over the energy distribution of these minima, will increase the discrimination between correct and misfolded structures. Determination of the energy function is achieved in a self-consistent fashion because the thermally occupied minima of the misfolded structures are dependent upon the energy function. Therefore, the optimal $\boldsymbol{\gamma}$ values for each iteration, n', are calculated as follows:

$$\boldsymbol{\gamma}_{n'} = \langle B_{n'}^{-1} \rangle \langle A_{n'} \rangle. \tag{8}$$

where the $\langle \lambda_i \rangle_{misfolded}$ values in Equations 6 and 7 now denote the average frequency of occurrence in the thermally occupied minima of misfolded structures rather than the full ensemble. In order to provide smoother convergence, a relaxation method was employed where a linear combination of the γ_{n-1} and the current $\gamma_{n'}$ values was performed to produce the final γ_n values:

$$\boldsymbol{\gamma}_n = (1 - \epsilon)\boldsymbol{\gamma}_{n-1} + \epsilon\gamma_{n'}. \tag{9}$$

In the work presented here, $\epsilon = 0.33$.

The progress of each successive optimization was evaluated with a discrimination score for each protein in the training set:

$$D_n = \frac{\delta E}{\Delta E_n}, \tag{10}$$

where

$$\delta E = \langle E_f \rangle_{correct} - \langle E_m \rangle_{minima}, \tag{11}$$

$$\Delta E^2 = \langle E_m^2 \rangle_{minima} - \langle E_m \rangle_{minima}^2. \tag{12}$$

Again the $\langle\,\rangle$ indicates an average over energy states. The energy of the folded structure, $\langle E_f \rangle_{correct}$, is evaluated as the average over the native fold and the training protein's alignments to each of its structural analogues. The energy of the misfolded structures, $\langle E_m \rangle_{minima}$, is evaluated as the energy of the training protein's alignments to a set of scaffolds with nonrelated folds produced by using the nth order iterated energy function. Figure 2 shows an example of the discrimination between two different energy functions, $n = 0$ (Fig. 2A) and $n = 1$ (Fig. 2B) for the training protein myoglobin (Takano, 1977, PDB code 5MBN).

The energy function used in the zeroth order ($n = 0$) approximation was based on previous work done by Goldstein et al. (1994, 1995), in which the energy parameters, excluding the gap energy terms, are calculated by solving the variational problem of maximizing the stability gap in units of the variance of the full ensemble of misfolded structures (Fig. 2). The penalty term for gaps was then evaluated separately using the Bayesian version of the analysis. Once the $\gamma_{n=0}$ values of the energy function have been calculated, the thermally occupied minima of the misfolded structures can be produced by aligning a set of training proteins against known structures. Unlike the distribution of all misfolded structures, the thermally occupied minima now

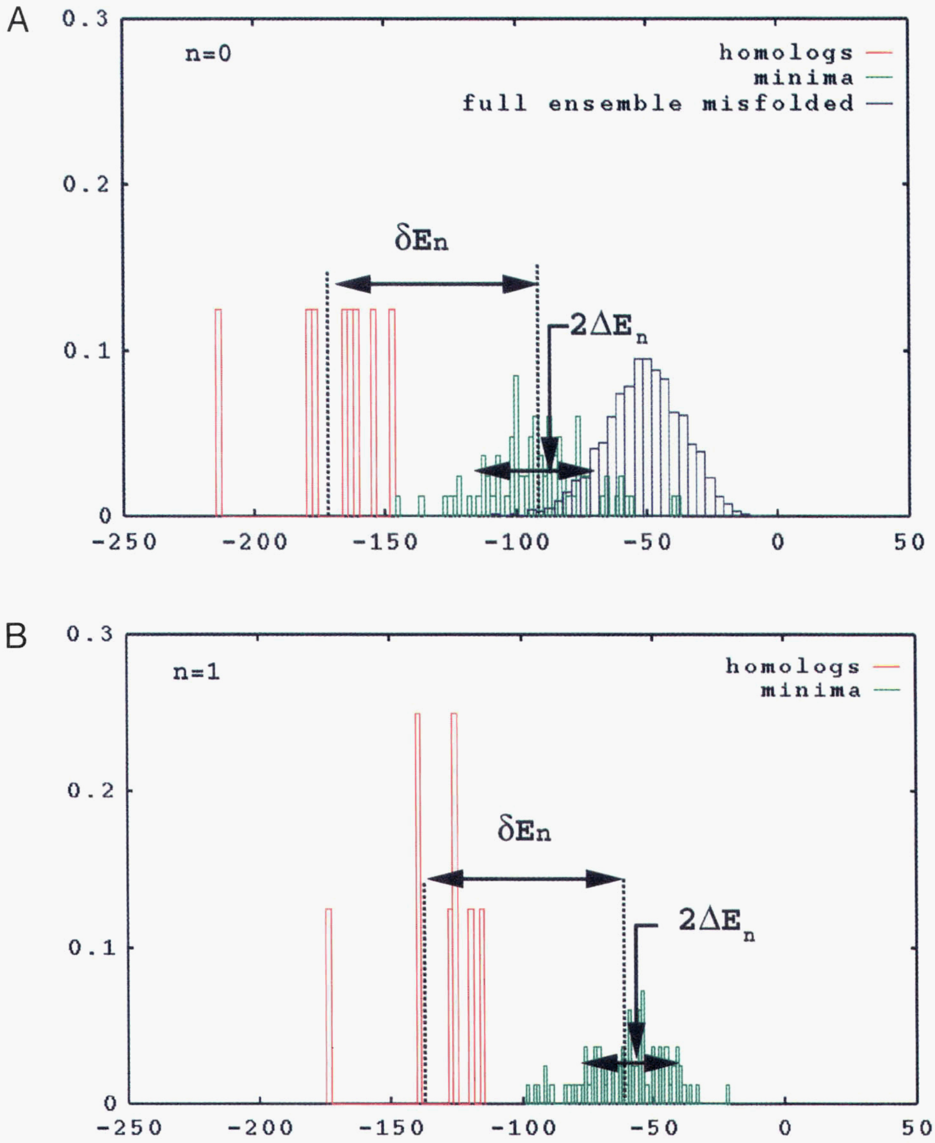

Fig. 2. Distributions of energetic states for the training protein myoglobin. The discrimination score D_n is evaluated by calculating the stability gap, δE, over the standard deviation, ΔE. **A:** Distributions of energetic states for the correct structures (native + alignment to homologues), the thermally occupied minima of misfolded structures (alignment to 83 unrelated scaffolds), and the full ensemble (3,000 states) of misfolded structures (5mbn sequence translated over unrelated structures) in that order, from left to right. All structures are evaluated with the zeroth order approximation energy function. **B:** Distributions of energetic states for the correct structures and thermally occupied minima of misfolded structures in that order, from left to right. All of the structures are evaluated with the first iterate energy function.

contain insertions and deletions. The introduction of insertions and deletions in the misfolded structures allows the gap penalty term to be evaluated directly with the rest of the energy terms. Once the minima have been obtained, the stability gap and variance can be re-estimated until self-consistency or maximum degree of discrimination averaged over the training set is achieved.

Results

Discrimination scores

The discrimination scores D_n (Table 1) for the 29 representative training proteins (see the Materials and methods) indicate that, in general, the $\boldsymbol{\gamma}_n$ values are better able to differentiate between correct and misfolded structures than the $\boldsymbol{\gamma}_{n-1}$ values. The maximum degree of discrimination for the majority of the training proteins is obtained with the 2nd-iterative energy parameters (see Supplementary material in Electronic Appendix). As seen in the last column of Table 1, the γ_2 parameters provide a mean increase in discrimination of 44% over the γ_0 parameters. The only two decreases in discrimination through self-consistent optimization are Bence-Jones immunoglobin (Epp et al., 1975, PDB code 1REI(A)) and hemerythrin (Stenkamp et al., 1978, PDB code 1HMQ(A)). Nevertheless, the self-consistent energy function generates a perfect alignment for the native scaffold for each protein, and produces alignments to their respective

Table 1. *Discrimination scores, Dn, for each successive iteration for the 29 training proteins*[a]

						Second
PDB	NRES	Name	Zeroth	First	Second	zeroth[b]
α Proteins						
2cro	63	434 Cro protein	3.70	3.60	4.40	1.19
1wrp(R)	102	TRP repressor (trigonal form)	2.41	3.80	4.44	1.84
1ccr	111	Cytochrome *c*	4.43	4.62	5.02	1.13
1hmq(A)	113	Hemerythrin (MET)	3.80	4.22	3.48	0.92
1bp2	123	Phospholipase	6.35	6.35	7.63	1.20
2hhb(A)	141	Hemoglobin (deoxy)	2.37	2.90	3.36	1.42
3cln	143	Calmodulin	5.37	5.95	6.99	1.30
1fdh(G)	146	Hemoglobin (deoxy, human fetal)	3.31	3.63	4.39	1.33
5mbn	153	Sperm whale myoglobin (deoxy)	3.79	4.62	5.49	1.45
β Proteins						
5pcy	99	Plastocyanin	5.35	6.28	6.82	1.28
1rei(A)	107	Bence-*Jones immunoglobulin	3.93	4.15	3.59	0.91
2paz	123	Pseudoazurin (cupredoxin)	6.55	8.69	8.85	1.35
2i1b	153	Interleukin-1*β	6.14	7.62	9.46	1.54
1f19(H)	215	R19.9(IG*G2B=K=) Fab fragment	2.14	2.79	2.55	1.20
3hfm(H)	215	IG*G1 Fab fragment	4.71	5.40	6.62	1.41
2plv(3)	235	Poliovirus (TYPE 1, Mahoney strain)	2.58	3.79	5.16	2.00
1r1a(2)	238	Rhinovirus serotype 1 (HRV1) coat protein	2.53	3.18	3.91	1.55
1cms	323	Chymosin B	3.36	4.19	5.56	1.65
α+β or α/β Proteins						
1fdx	54	Ferredoxin	1.97	2.53	2.21	1.12
1alc	122	α-*Lactalbumin	6.63	7.51	7.78	1.17
1rbb(A)	124	Ribonuclease B	5.68	6.49	6.91	1.22
1snc	135	Staphylococcal nuclease	6.31	8.15	9.61	1.52
2dhf(A)	182	Dihydrofolate reductase	3.36	4.28	5.01	1.49
2act	218	Actinidin (sulfhydryl proteinase)	5.06	6.59	7.60	1.50
2prk	279	Proteinase K	3.76	6.33	6.89	1.83
1pfk(A)	319	Phosphofructokinase (R-state)	3.85	5.56	7.10	1.84
2ldx	331	Apo-Lactate dehydrogenase	2.92	3.92	5.03	1.72
3gpd(G)	334	D-Glyceraldehyde-3-phosphate dehydrogenase	3.44	4.72	7.24	2.10
2liv	344	Leucine/isoleucine/valine-binding protein	5.34	6.90	8.83	1.65
Mean[c]			4.18	5.13	5.93	1.44

[a] The energy of the correct structure for each training protein is the average of itself and its structural analogues.

[b] Ratio of the second iterate discrimination score versus the zeroth-order discrimination score.

[c] The mean value is the average discrimination score over all the folding motifs. The greatest average discrimination is obtained with the second iterate values and these γ values are used in all the calculations presented in this paper.

structural analogues that are more energetically stable than alignments to unrelated structures.

The partial ordering of competing structures through the sequence–structure alignment minimization can be observed most directly through the surface-accessibility parameter. In the full distribution of misfolded structures, hydrophobic residues have equal probability of being placed on the surface or being buried (Fig. 3). On the other hand, in the thermally occupied minima, there are fewer hydrophobic residues on the surface, which is more comparable to the correct folds and thus more representative of a protein found in nature. The change in inside/outside placement does not vary significantly among the thermally occupied minima.

Alignments of known structures

To test our energy function, we created a test set of 16 known structures representing the various folding types of α, β, α/β, and $\alpha+\beta$. Each protein in this list had to have a known structural analogue with limited sequence similarity (less than 31%). In general it is difficult to find appropriate test proteins because we need the X-ray crystal structures of the test protein and at least one analogue with low percent identity. Most of the known X-ray structures with this requirement fall within similar folding classes (Murzin et al., 1995). Due to this constraint, only 4 of the 16 test proteins chosen are in folding classes (Murzin et al., 1995) different from those of the training set. Each of the 16 test proteins was aligned to 42 putative scaffolds, which included the native structure and one structural analogue with low percent identity (17.3–30.3%). A mean-field threading program was used to align the test proteins to each of the putative scaffolds using the self-consistently optimized energy function as the scoring matrix. The structure with the lowest energy was considered to be the predicted structure (Table 2). The q-scores listed in Table 2 are used as a measure of structure similarity (Goldstein et al., 1992b) between the X-ray structure and the

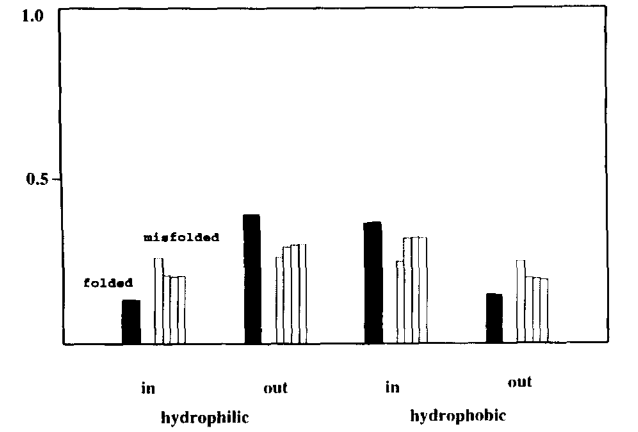

Fig. 3. Frequency of occurrence of hydrophobic and hydrophilic residues both in the interior and surface of molten globular structures. The shaded regions indicate the frequency for the correct (crystallographic) structures. The unshaded regions correspond to the misfolded structures generated by translations and after various stages of optimization (n = 0, 1, 2 iterate energy functions, respectively). Initially there is little correlation between inside/outside and hydrophobicity. Upon continued self-optimization, microphase separation in the minima generated from unrelated scaffolds becomes clear.

alignments derived using the present energy function (Equation 1) (q), as well as the standard (q-NW) and a modified (q-P-NW) Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) (see the Materials and methods) using a modified Dayhoff scoring matrix (Gribskov & Burgess, 1986). The q-score measures what fraction of the pairwise distances between corresponding residues in the aligned structure ($r_{i'j'}$) match the correct distances (r_{ij}^*) in the X-ray structure:

$$q = [N(N-1)]^{-1} \sum_{j>i} \exp[-(r_{ij}^* - r_{i'j'})^2/\sigma^2]. \quad (13)$$

The width of the gaussian function is less than 2 Å and has weak dependence on the sequence length, $\sigma^2 = 2|j-i|^{0.15}$, to allow a greater tolerance in the distances for residues further apart. If there is an error greater than 2 Å at large separation in the sequence ($j - i > 100$), no contribution to q exists and the comparison is poor. The q-scores emphasize compact regions in a protein (e.g., the core domain) and deemphasize discrepancies in isolated fragments, particularly at the ends of the aligned proteins. A q-score > 0.4 is interpreted as an indication of structural similarity and corresponds to an RMS value < 6 Å, as shown in Table 2. The average q-score of misfolded structures

Table 2. *The two most energetically stable alignments using the self-consistent energy function for a test set of 16 sequences with known structures*[a]

	Self-recognition		Predicted structure							
PDB	Target	q[b]	PDB	Analogue	RMS[c]	RMS Nw[d]	q[b]	q-NW[e]	q-P-NW[f]	% Ident.[g]
α-Proteins										
256b(A)	Cytochrome *b*562 (oxidized)	1.00	2ccy(A)	Cytochrome *c*′	5.01	11.62	0.44	0.23	0.49	17.3
3icb	Calcium binding protein	1.00	4tinc	Troponin C	3.30	3.37	0.52	0.51	0.51	29.3
1flpz	Crab hemoglobin I	1.00	2lbh	Hemoglobin V	2.74	4.39	0.64	0.48	0.41	22.5
2lh4	Leghemoglobin (deoxy)	1.00	5mbn	Sperm whale myoglobin	2.82	4.01	0.57	0.55	0.39	18.3
1mba	Sea hare myoglobin (MET)	1.00	5mbn	Sperm whale myoglobin	2.56	5.30	0.63	0.48	0.68	26.0
1gsq	Glutathione S-transferase	1.00	1gta	Glutathione S-transferase	4.79	3.57	0.46	0.63	0.68	21.8
1r69	434 Repressor (amino-terminal)	1.00	1lrd(4)	lambda Repressor-operator	3.27	2.97	0.55	0.61	0.61	28.6
β-Proteins										
1aaj	Amicyanin	1.00	6pcy	Plastocyanin	3.29	4.34	0.43	0.40	0.42	20.2
2rhe	Bence-*Jones protein	1.00	3hfm(H)	IG*G1 Fab fragment	4.96	4.92	0.46	0.43	0.47	27.2
1cd8	CD8 (T-cell CO-receptor)	1.00	2fb4(H)	Immunoglobin Fab fragment	5.00	9.24	0.40	0.30	0.36	21.3
3hfm(L)	IG*G1 Fab fragment	0.94	3hfm(H)	IG*G1 Fab fragment	7.64	7.06	0.31	0.38	0.32	25.2
2ifb	Intestinal fatty acid binding protein	1.00	1crb	Cellular retinol binding protein	3.02	2.89	0.64	0.63	0.67	30.53
α+β and α/β Proteins										
1fx1	Flavodoxin	1.00	3fxn	Flavodoxin	2.32	2.49	0.61	0.62	0.63	30.4
3dfr	Dihydrofolate reductase	1.00	8dfr	Dihydrofolate reductase	4.16	3.20	0.51	0.66	0.48	29.6
6ldh	Apo-lactate dehydrogenase	0.92	4mdh(A)	Malate dehydrogenase	3.93	9.20	0.47	0.36	0.40	20.2
2gbp	Galactose/glucose binding protein	0.93	1abp	Arabinose-binding protein	6.34	5.33	0.53	0.47	0.47	22.5

[a] For every target sequence, the structure with the most stable energy was the native one followed by the test protein's structural analogue.

[b] Structural similarity score between the test protein's X-ray structure and the alignment to its structural analogue produced by the self-consistent energy function.

[c] RMSD for Cα atoms only calculated with X-PLOR for alignment produced by the energy function.

[d] RMSD for Cα atoms only calculated with X-PLOR for alignment produced by NW.

[e] Structural similarities between the test protein's X-ray structure and the alignment to its structural analogue using NW.

[f] Structural similarities between the X-ray structure and the alignment to its structural analogue produced with P-NW.

[g] Percent identity based on the NW alignment. All sequences were aligned to a set of 42 putative structures containing the native structure and one structural analogue with limited percent identity (17.3–30.5%).

is roughly 0.14 (see Fig. 4), with an RMS value between 13 and 21 Å. The RMS values were calculated by X-PLOR (Brünger, 1987) using only the C_α coordinates. Self-recognition was achieved for all test cases generally with exactly correct alignment, although three had minor shifts resulting in *q*-scores slightly lower than 1.0. This is to be expected to some extent because the alignment algorithm uses the structural information of a scaffold, not sequence identity to set up a scoring matrix for an alignment between two sequences. The scaffold with the second most stable energy in all 16 test cases was the proteins' structural analogue.

One alignment of interest is the threading of cytochrome *b*562 subunit A (Mathews et al., 1975, PDB code 256B(A)), to the subunit A of cytochrome *c* prime's scaffold (Weber et al., 1980, PDB code 2CCY). The self-consistently optimized energy function produces an alignment with a *q*-score of 0.44, whereas the NW alignment only gives a *q*-score of 0.28. The distance plots of the energy function and NW alignments of 256B(A) → 2CCY(A) versus the actual 256B(A) X-ray structure are shown in Figure 5. The energy function alignment produces a structure that is more similar to the original *b*562 native form, as is also evident from the comparison of secondary structure given in Figure 6. This figure also illustrates the advantage of not depending solely on sequence identity to align two sequences. The present energy function alignment has a sequence identity of 13.2% compared to 20.1% for the NW alignment, but is structurally much more similar.

In the case of apo-lactase dehydrogenase (Abad-Zapatero et al., 1987, PDB code 6LDH), the alignment methods have comparable *q*-scores, but the energy function-based algorithm produces a structure with a decidedly better RMS value. The nearly equivalent *q*-scores indicate that the core structures are about the same and the distance plots (Fig. 7) indicate that the discrepancy in the NW alignments (both standard and physically based NW) arises from an incorrect structure assignment to the first 20 residues.

The alignment of a squid glutathione S-transferase (Ji et al., 1994, PDB code 1GSQ) to a blood fluke glutathione S-transferase (McTigue et al., 1995, PDB code 1GTA) produced from the self-consistent energy function has a much lower *q*-score than the alignments produced with evolutionary scoring matrices. The energy difference between the alignment produced with the self-consistent energy function compared to the alignment generated with the modified Dayhoff matrix is not that significant (−167 units for EF alignment and −154 units for P-NW). The lower *q*-score in the energy function alignment arises because the energy function does not allow the gap in the first half of the helix (residues 38–40 of 1GTA) that occurs in the P-NW alignment. It is energetically unfavorable because the structure loses three hydrogen bonds, has slightly lower profile contributions, and contact interactions in the longer cut-off range. Figure 8 shows the distance plots of the alignments produced by the energy function (Fig. 8A) and the P-NW (Fig. 8B). This figure illustrates the loss of tertiary interactions (i.e., the distance plot of the EF

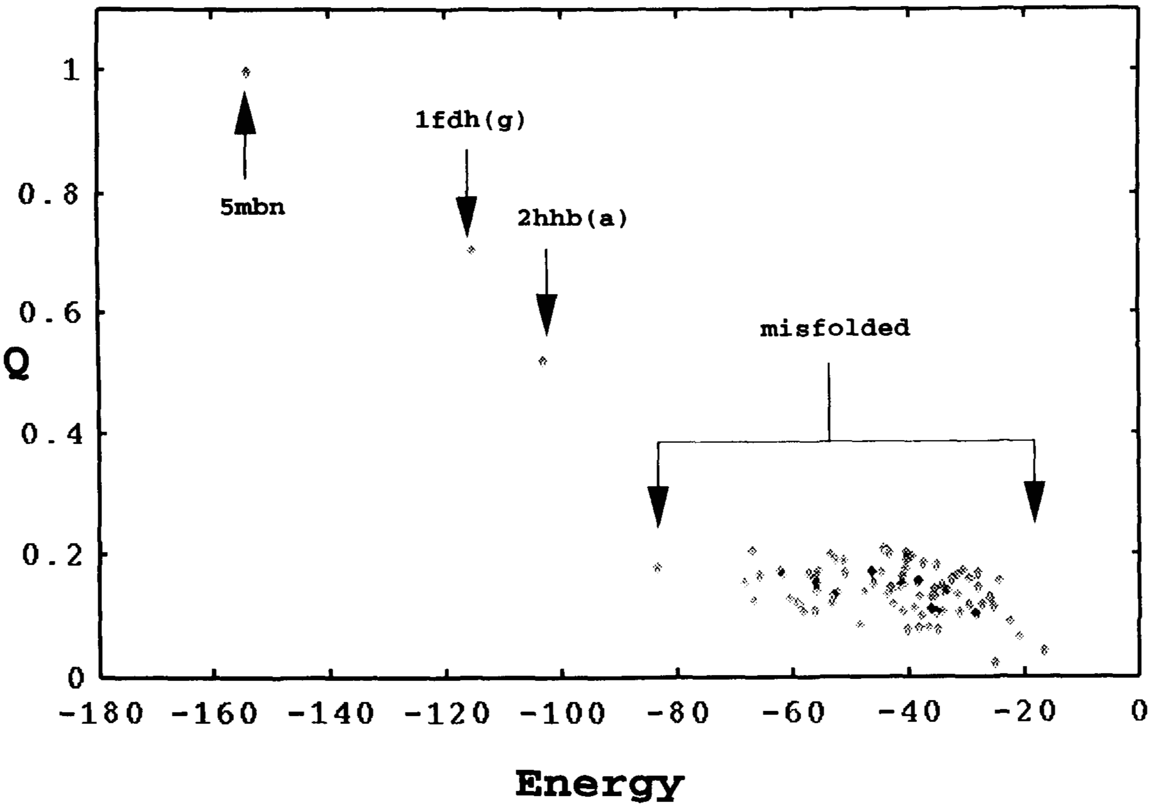

Fig. 4. Energy of alignment versus *q*-score for the sperm whale myoglobin sequence aligned to its native fold, two homologous scaffolds, and 83 nonrelated structures.

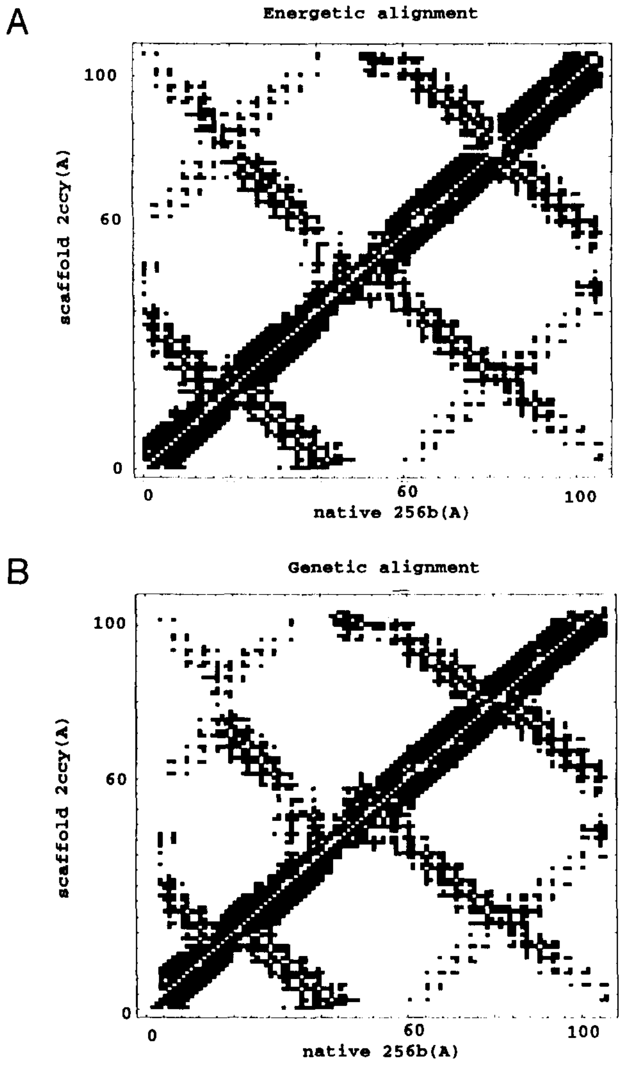

Fig. 5. Distance plots of 256b(A) comparing alignments generated from two different schemes, self-consistent energy functions and evolutionary scoring matrix with default parameters. **A:** Distance plot from the 256b(A) X-ray structure (lower half) and the alignment of 256b(A) to 2ccy(A) using the self-consistent energy function (upper half). All contacts indicated have an $r^{C}_{\beta ij} < 12$ Å. This plot illustrates the high structural similarity between the X-ray structure and our predicted structure. **B:** Distance plot of the 256b(A) X-ray structure (lower half) and the alignment of 256b(A) to 2ccy(A) using NW (upper half).

alignment has off-diagonal areas that are less dense than the distance plot of the NW alignment) due to the shift in the energy function alignment.

The alignment of the galactose/glucose binding protein (Vyas et al., 1988, PDB code 2GBP) to arabinose binding protein (Quiocho & Vyas, 1984, PDB code 1ABP) produced using the energy function has a higher similarity to the native fold between residues 200 and 265 than the alignment generated from the genetic scoring matrix. The tertiary interactions are more native-like in this area for the energy function alignment, as can be seen in Figure 9. However, the energy function does a poor job of aligning the last 30 residues. The native fold of 2GBP has an extented loop (residues 276–288) away from the core of the protein, which the 1ABP scaffold does not contain. The energy function does not place a gap in the target at this point, nor any place later in the sequence due to energetic reasons. This gives rise to RMS deviations (RMSDs) ranging from 12 to 20 Å for the final 30 residues. The NW alignment shows a higher degree of correct tertiary structure between the residue ranges of 255–265 and 299–309 (Fig. 9) arising from a gap in the target between residues 292 and 297. Therefore, the higher *q*-score for the energy function alignment is due to the higher similarity to the native fold in the core of the protein and the lower RMS score is due to the poor alignment of the final 30 residues.

Although the quality (defined here by *q*-score and RMS values) of the alignments produced by various methods are generally similar, it is important to mention that only the self-consistent energy function alignments can be ordered by energy, as shown in Table 1. Thus, these alignments reflect accurately the thermodynamic stability of the alignment, whereas the traditional percent identity can be misleading. In general, the energy function generates better alignments than the standard evolutionary scoring matrices when the percent identity between the two sequences is low (less than 21%).

Structure prediction for proteins with prosite signatures

Another type of test for this energy function is to create a set of proteins with unknown structures but whose sequences contain the same PROSITE signature. These sequences were then aligned to 43 putative scaffolds, which included two structures containing the corresponding PROSITE signature. Four different sets of proteins were chosen: a set containing the immunoglobin/major histocompatibility signature, the flavidoxin signature, the dihydrofolate reductase signature, and the thioredoxin signature.

Immunoglobin and major histocompatibility proteins are involved in the immune response against bacterial and viral antigens, respectively. Immunoglobin proteins are found on the surface of B cells and bind a specific antigen. Histocompatibility proteins, found on the surface of nearly all nucleated vertebrae cells, will bind an antigenetic fragment and then be recognized by T cell receptor proteins. Both types of proteins have a region containing an Ig_Mhc motif of (F,Y)xCx(V,A)xH, with four conserved residues, F or Y, C, V or A, and H. We chose the chain sequences of 10 immunoglobin and 9 histocompatibility proteins from the SWISS-PROT database (Bairoch & Boeckmann, 1994) due to their low sequence identity (less than 28%) with two known structures containing the Ig_Mhc motif (Padlan et al., 1989, PDB code 3HFM; Garrett et al., 1989, PDB code 3HLA). 3HLA(A) and 3HFM(H) are two-domain structures in which one domain is structurally similar. The structure with the lowest energy for each sequence was either the 3HFM(H) or 3HLA(A) (Table 3). In each case, the 19 SWISS-PROT sequences align their Ig_Mhc motif to the Ig_Mhc motif in both the 3HFM(H) and 3HLA(A) scaffolds.

Flavidoxin proteins are involved in the electron transfer in photosynthesis. These proteins contain a prosite motif of (L,I,V) (L,I,V,F,Y) (F,Y) x (S,T) xx (A,G) xTxxxAxx (L,I,V). We chose seven protein sequences from the SWISS-PROT database that had relatively low sequence identity with two flavidoxins of known structure (Smith et al., 1977, PDB code 3FXN; Watenpaugh et al., 1972, PDB code 1FX1). These seven flavidoxin sequences were then aligned to 42 putative structures, two of which were 3FXN and 1FX1. The scaffold with the lowest energy for

```
A   Alignment:           Secondary structure assignment

           1
X-ray:      ttHHHHHHHH  HHHHHHHHHt  ttHHHHHHHH  HHHHHHHHHH  tttttttttt
EF:         HHHHHHHHHH  HHHHHHHttt  ttHHHHHHHH  HHHHHHtttt  tttttttttt
NW:         HHHHtttttt  ttt--HHHHH  HHHHHHHHHt  tttttttttt  tttttBtHHH

          51
            tttttHHHHH  HHHHHHHHHH  HHHHHHHHHH  tttHHHHHHH  HtHHHHHHHH
            ttttHHTHHH  HHHHHHHHHH  HHHHHHHHHH  H-tHHHHHHH  HHHHHHHHHH
            HtttHHHHHH  HHHHHHHHHH  HHHHHHHHtH  HHHHHHHHHH  HHHHHHHHHH

         101
            HHHHHt
            HHHHHB          (85% secondary structure correct)
            HBtt--          (60% secondary structure correct)
```

```
B   Alignment:           Secondary structure assignment

         131
X-ray:      tttBBBBBtt  tHHHHHHHtt  tttttttttt  tttBBBtttt  BBBBBBBttt
5pcy:       ---BBBBBBB  ttBBBBBBt-  -----tBBt-  ---BBBBBtt  tBBBBBBBBt
2paz:       ---BBBBBBB  Btttttttt-  -----tBBt-  ---BBBBBtt  tBBBBBBBtt
1paz:       ---BBBBBBB  BBttttt---  -----BBBBB  BBBBBBtttt  BBBBBBBB--

         181
            ttBBBBBttt  tBBBBBtttt  tBBBBBtttt  BBBBBBBBBt  ttttttttBB
            tBBBBBB-tt  tttttttttt  BBBBBBBBtt  tBBBBBBBB-  tt---tttBB
            ttttBBBttt  tttttBBBtt  ttBBBBBBtt  tBBBBBBB-t  tt---tttBB
            ttttBBBttt  ttBBBBBttt  tBBBBB-ttt  tBBBBBBt-t  tt---ttttt

         231
            BBBBBBHHHH  HHHHHHHHHH  H
            BBBBBB----  ----------  -   (75% secondary structure)
            BBBBtttHHH  HHHttttHHH  H   (69% secondary structure)
            BBBBBBtttH  HHHHHHHttH  H   (73% secondary structure)
```

Fig. 6. Secondary structure assignments arising from two different alignments. **A:** Secondary structure assignment of each residue as defined by DSSP algorithm for 256b(A) X-ray structure, the alignment of 256b(A) to 2ccy(A) generated using the self-consistent energy functions, and the alignment of 256b(A) to 2ccy(A) obtained from using Needleman–Wunsch. The secondary structure of each residue is represented as either an α-helix (H), a β-sheet (B), or either a turn or random coil (t). The "-" symbol represents a gap in the alignment that is assigned as a random coil in calculating percent correct secondary structure. **B:** Secondary structure of each residue in the copper A domain of cytochrome *c* oxidases defined by the report of the X-ray structure and the alignments of copper A to 5pcy, 2paz, and 1paz generated using the self-consistent energy function with experimental constraints. Secondary structure assignment as in A.

each sequence is one of the two known flavidoxins (Table 3). In all cases, the flavidoxin sequences align their flavidoxin signature to the flavidoxin signature found in both known structures.

Dihydrofolate reductase proteins are cytosolic enzymes that help catalyze the reaction of NADPH + $H^+ \rightarrow NADP^+ + H_2$. These proteins contain a Dhfr signature of (L,I,F) Gxxxx (L,I,V,M,F) PW. A set of five sequences containing this motif were extracted from SWISS-PROT database. These sequences were then aligned to 43 random structures. Two of these structures were dihydrofolate reductase proteins (Bolin et al., 1982, PDB code 4DFR(B); McTigue et al., 1993, PDB code 8DFR). The lowest energy structure for all of the sequences was either 4DFR(B) or 8DFR (Table 3), and all sequences have their Dhfr signature aligned to the Dhfr signature of the 4DFR(B) or 8DFR.

Thioredoxins are small proteins of approximately 100 residues in length. They participate in various redox reactions via the reversible oxidation of an active center disulfide bond. These proteins contain a thioredoxin signature of (S,T,A)x(W,G) C(A,G,V)(P,H)C (T)x(W)C(G)(P)C. A set of six sequences containing this motif was extracted from the SWISS-PROT database. These sequences were then aligned to 43 random structures. One of the structures in this set was thioredoxin (reduced form) (Holmgren et al., 1990, PDB code 2TRX(A)). In each case, the lowest energy structure was 2TRX(A), and all sequences have their thioredoxin signature aligned to the signature in 2TRX(A).

Use of constraints in the prediction of the copper A domain of cytochrome c oxidase from Paracocuus denitrificans

Cytochrome *c* oxidase is a three-subunit complex that is found in the membrane of the mitochondria and is the last part of the respiratory chain. The copper A domain is part of the subunit II and is water soluble. This domain contains two copper ions

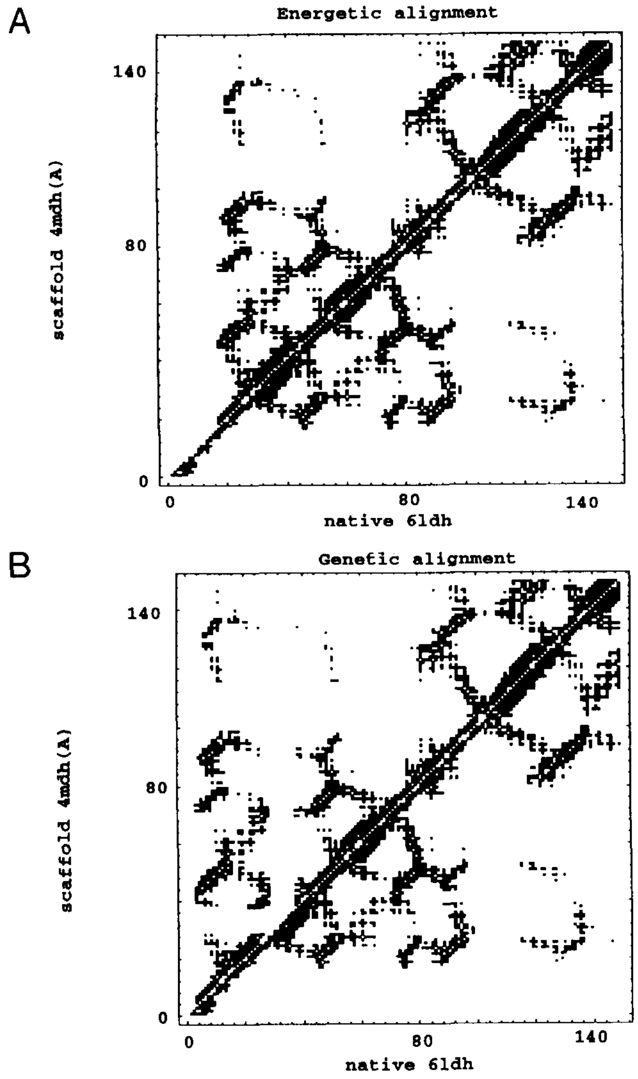

Fig. 7. Distance plots of the first 150 residues of 6ldh comparing the results of two alignment schemes, self-consistent energy function and evolutionary scoring matrix. **A:** Comparison of the X-ray structure of 6ldh (lower half) versus the alignment of 6ldh to 4mdh(A) using the self-consistent energy function (upper half). **B:** Alignment of 6ldh to 4mdh(A) scaffold (upper half) produced by Needleman-Wunsch alignment algorithm to the 6ldh X-ray structure. The evolutionary scoring matrix incorrectly aligns the first 20 residues of the 6ldh sequence to the 4mdh(A) scaffold, causing a discrepancy between the predicted and native structures. The alignment produced from the self-consistent energy function does not exhibit this discrepancy.

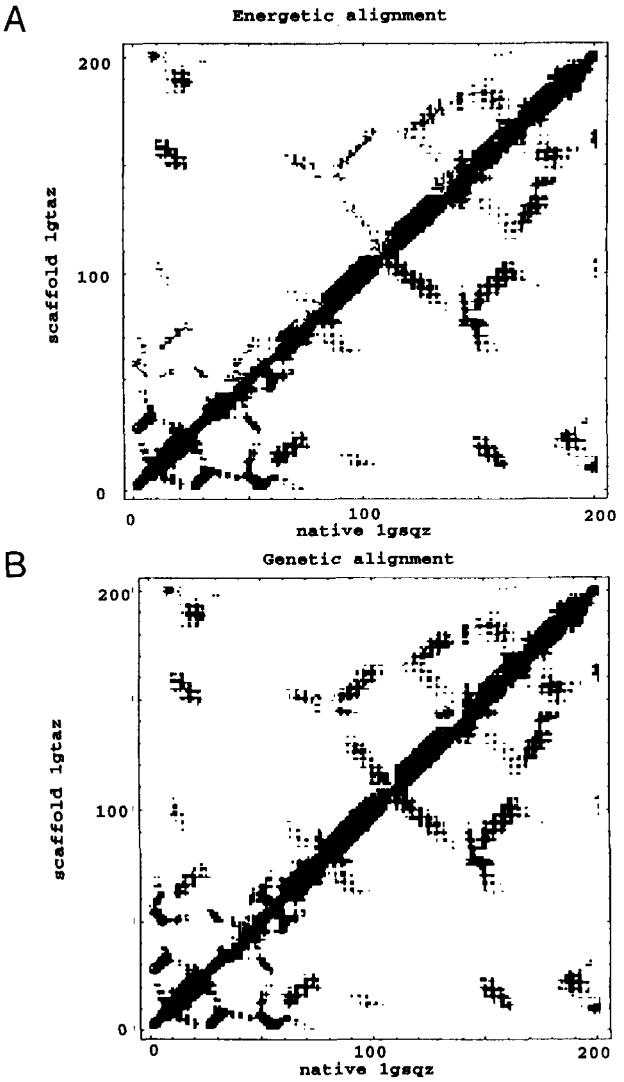

Fig. 8. Distance plots of 1gsq aligned to 1gta. **A:** Comparison of the native fold of 1gsq (lower half) to the alignment of 1gsq to the 1gta scaffold using the self-consistent energy function (upper half). **B:** Alignment of 1gsq to 1gta using the PNW alignment algorithm (upper half) compared to 1gsq's native fold (lower half). The alignment produced using our energy function has less tertiary structure than the one produced using the standard defaults with the modified Dayhoff matrix.

that accept an electron from cytochrome *c*. When this domain is separated from the membrane, it has an ER spectrum that is similar to blue copper proteins. Blue copper proteins have four conserved residues that are the ligands for their copper ion, H, C, H, M. That the ligands of the copper A sequence are H181, C216, C220, and M227 for one of the coppers, and Q218, C216, C220, and C224 for the other copper ion has been inferred by numerous bioinorganic and spectroscopic experiments (Han et al., 1991; Antholine et al., 1992; van der Oost et al., 1992; Kelly et al., 1993; Steffens et al., 1993; Farrar et al., 1995). The ligands for the first copper ion are the same as in the blue copper proteins, so constraints were included (see the Materials and methods) in the energy function for these four residues based on their distances found in the blue copper structures. The residues involved in the constraints, the distance ranges between these residues, as well as the stability and penalty energy units are given in Table 4. The top five most stable structures of the alignment of this sequence (residues 100–253 of subunit II of cytochrome *c* oxidase) to 60 putative β-sheet structures are all blue copper proteins. The analogy to blue copper proteins had been suggested earlier on other grounds (Han et al., 1991; Steffens et al., 1993).

The *P. denitrificans* cytochrome *c* oxidase structure was published after our predictions were made (Iwata et al., 1995). The chemical constraints inferred previously were confirmed to be correct. The copper A soluble domain of subunit II has 10 beta

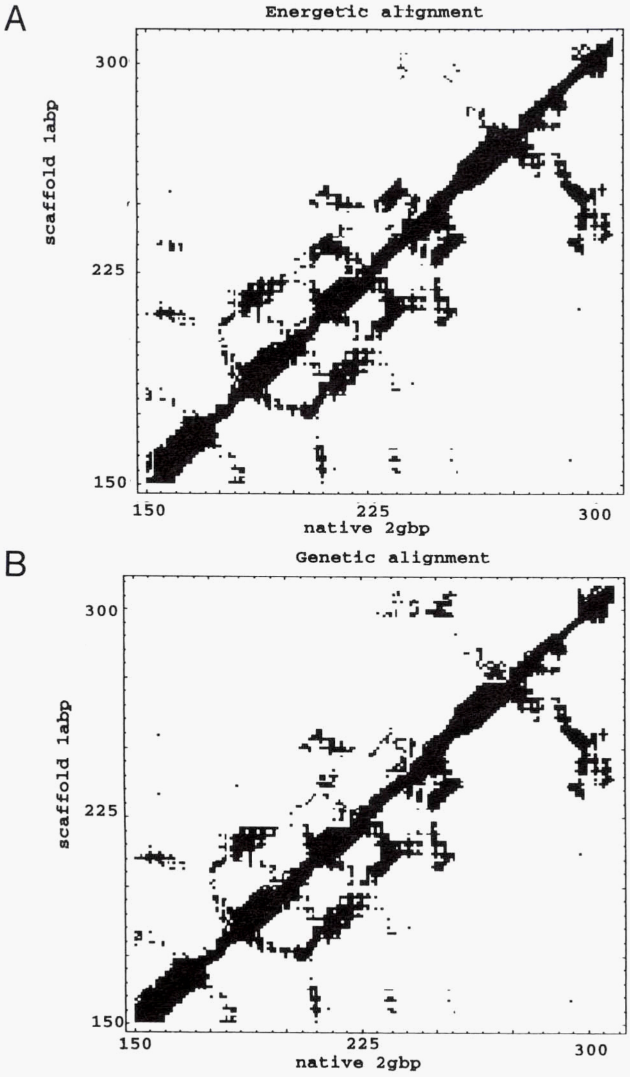

Fig. 9. Distance plots of 2gbp aligned to 1abp. **A:** Comparison between the X-ray structure of 2gbp (lower half) and the alignment of 2gbp to the 1abp scaffold using the self-consistent energy function (upper half). **B:** Alignment of 2gbp to the 1abp scaffold (upper half) using the default parameters with a modified Dayhoff matrix compared to the X-ray structure of 2gbp (lower half). This figure illustrates the stronger tertiary interactions between residues 200–260 of the energy function alignment versus the genetic alignment. This figure also shows the loss of tertiary interactions occurring with the last 20 residues for the energy function alignment due to the poor alignment of residues 280–309.

strands and 2 helices. Figure 10 shows the C_α trace of our most stable fold for this sequence, which is based on the plastocyanin scaffold (Church et al., 1986, PDB code 5PCY) where the residues known to be in beta strands are colored red and the binding ligands are colored green. The plastocyanin scaffold does not contain any helices, whereas the copper A domain contains two. However, the region between β-strands 3 and 4 in our alignment has a gap insertion of 14 residues that could possibly be modeled into a helix. The secondary structure assignments of our top three alignments of 5PCY, 1PAZ, and 2PAZ have a total of 75%, 69%, and 73% correct secondary structure assignment, respectively (Fig. 6). The percent of secondary structure assigned correctly to the copper A sequence is slightly lower than one would expect. However, this is mainly due to the inability to model the helix found between β-strands 3 and 4 of the copper A structure because this structural element is lacking in the blue copper protein scaffolds used in this analysis. This helps illustrate that the predicted structure can only be as good as the scaffolds available for alignment.

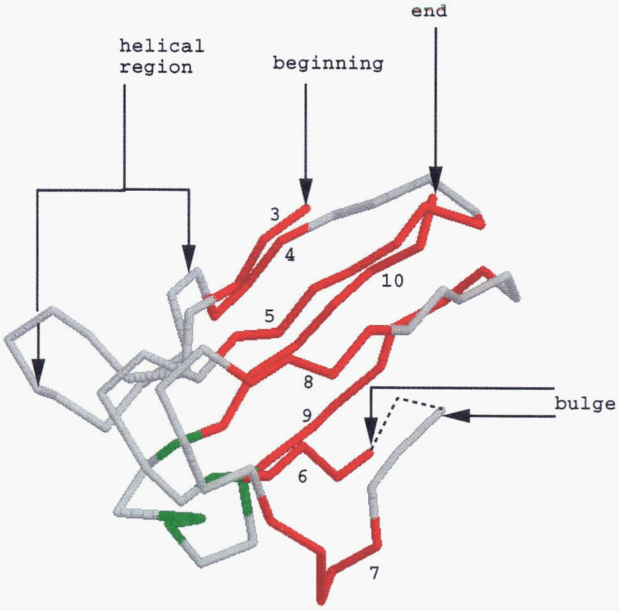

Fig. 10. Sequence of the copper A domain of the subunit II of cytochrome *c* oxidase from *P. denitrificans* threaded onto the 5pcy structure. Residues in red are known to form β-sheets 3–10 as defined by the report on the cytochrome *c* oxidase X-ray structure. Green depicts ligand-binding residues. A bulge gap is found between residues 185 and 187 with a C_α distance of 7.2 Å. The first helical region, indicated with arrows, is not assigned correct secondary structure. The first and last residues of the aligned copper A sequence are labeled with arrows.

Discussion

In the present work, a self-consistent procedure taking into account the partial ordering of misfolded structures was used to determine an energy function that gives improved sequence–structure alignments. The optimization of the energy function through self-consistency allowed for a higher discrimination between the correct and thermally occupied minima in the ensemble of misfolded states by incorporating some correlation between the folded and misfolded energy states. It also enabled us to optimize simultaneously the gap energy parameters along with all the other energy terms. In testing the degree of discrimination D_n of a given iterative optimization n (Table 1), only a few of the training proteins have slightly lower discrimination with the self-consistent energy function. Even in the case of 1REI(A), the present energy function is still able to find self-alignment and the structurally analogous scaffolds of 1REI(A) as the lowest energy states upon sequence threading. The problem of decrease in discrimination for 1REI(A) is probably enhanced by our definition of surface accessibility. Surface accessibility is defined for a complete protein, with all subunits included. Because we use subunit scaffolds as single subunit proteins, the surface accessibility for some of the residues in these

Table 3. *The most stable alignment using the self-consistent energy function for a set of sequences with unknown structure that contain a PROSITE signature*[a]

	Target name		Predicted homologue	%
SWISS-PROT	Name	PDB	Name	Ident.
Ig_Mhc Signature	(F,Y)xCx(V,A)xH			
B2MG_BOVIN	Bovin β-2 μ-globuline	3hfm(H)	IG*G1 Fab fragment	20.8
B2MG_PONPY	Orangutan β-2 μ-globuline	3hfm(H)	IG*G1 Fab fragment	19.6
B2MG_RAT	Rat β-2 μ-globuline	3hfm(H)	IG*G1 Fab fragment	18.6
HA2B_MOUSE	Mouse H-2 histocomp. antigen	3hla(A)	Human class I histocompatibility	16.9
HA2P_HUMAN	Human H-2 histocomp. antigen	3hla(A)	Human class I histocompatibility	25.3
HA2Q_MOUSE	Mouse H-2 histocomp. antigen	3hla(A)	Human class I histocompatibility	19.3
HA2S_MOUSE	Mouse H-2 histocomp. antigen	3hla(A)	Human class I histocompatibility	18.5
HA2Z_HUMAN	Human HLA histocomp. antigen	3hla(A)	Human class I histocompatibility	19.4
HB25_HUMAN	Human HLA histocomp. antigen	3hla(A)	Human class I histocompatibility	22.9
HB2B_HUMAN	Human HLA histocomp. antigen	3hla(A)	Human class I histocompatibility	21.2
HB2G_HUMAN	Human HLA histocomp. antigen	3hla(A)	Human class I histocompatibility	19.7
HB2S_HUMAN	Human HLA histocomp. antigen	3hla(A)	Human class I histocompatibility	23.2
KACA_RAT	Rat IG κ chain	3hfm(H)	IG*G1 Fab fragment	22.6
KACB_RAT	Rat IG κ chain	3hfm(H)	IG*G1 Fab fragment	21.7
KAC_MOUSE	Mouse IG λ chain	3hfm(H)	IG*G1 Fab fragment	19.8
LAC1_MOUSE	Mouse IG λ chain	3hla(A)	Human class I histocompatibility	24.2
LAC3_MOUSE	Mouse IG λ chain	3hfm(H)	IG*G1 Fab fragment	27.2
LAC5_MUSSP	Mouse IG λ chain	3hla(A)	Human class I histocompatibility	24.2
LAC_HUMAN	Human IG λ chain	3hla(A)	Human class I histocompatibility	22.2
Flavidoxin signature	(L,I,V)(L,I,V,F,Y)(F,Y)x(S,T)xx(A,G)...			
FLAV_ANASP	*Anabaena* sp. flavodoxin	1fx1	Flavodoxin	29.5
FLAV_AZOVI	*Azotobacter vinelandii* flavodoxin	1fx1	Flavodoxin	25.2
FLAV_CHOCR	*Chondrus crispus* flavodoxin	1fx1	Flavodoxin	29.3
FLAV_CLOAB	*Clostridium acetobutylicum* flavodoxin	1fx1	Flavodoxin	23.1
FLAV_DESVH	*Desulfovibro vulgaris* flavodoxin	3fxn	Flavodoxin	30.4
FLAV_KLEPN	*Klebsiella pneumoniae* flavodoxin	1fx1	Flavodoxin	23.1
FLAV_RHOCA	*Rhodobacter capsulatus* flavodoxin	1fx1	Flavodoxin	25.2
Dhfr signature	(L,I,F)Gxxxx(L,I,V,M,F)PW			
DYRB_ECOLI	*E. coli* DHFR	4dfr(B)	Dihydrofolate reductase	29.7
DYRC_ECOLI	*E. coli* DHFR	4dfr(B)	Dihydrofolate reductase	34.2
DYR_BPT4	Bacteriophage T4 DHFR	4dfr(B)	Dihydrofolate reductase	25.8
DYR_ENTFC	*Enterococcus faecium* DHFR	4dfr(B)	Dihydrofolate reductase	33.3
DYR_LACCA	*Lactobacillus casei* DHFR	4dfr(B)	Dihydrofolate reductase	27.7
Thioredoxin signature	(S,T,A)x(W,G)C(A,G,V)(P,H)C(A)...			
THI1_YEAST	*Saccharomyces cerevisiae* thioredoxin I	2trx(A)	*E. coli* thioredoxin	32.7
THIO_CHICK	Chick thioredoxin	2trx(A)	*E. coli* thioredoxin	24.0
THIO_HUMAN	Human thioredoxin	2trx(A)	*E. coli* thioredoxin	22.1
THIO_MOUSE	Mouse thioredoxin	2trx(A)	*E. coli* thioredoxin	22.1
THIO_RABIT	Rabit thioredoxin	2trx(A)	*E. coli* thioredoxin	22.1
THIO_RAT	Rat thioredoxin	2trx(A)	*E. coli* thioredoxin	22.1

[a] All sequences were aligned individually to 45 putative structures in which two of these structures contained the corresponding PROSITE signature. In all cases, the lowest energy structure was a protein with the appropriate PROSITE signature.

units is incorrect. Of the 14 structurally analogous scaffolds used for 1REI(A), only one was a single subunit protein. Also, all of the subunit proteins have two domains, whereas 1REI(A) is a single-domain protein. This creates differences in chemical properties between the analogues and 1REI(A), which affects the alignment. One way to improve the profile term in the energy function would be to redefine the surface accessibility for all the scaffolds that are subunits.

The compatibility of a sequence with a structure was evaluated with the total energy of the final alignment. The results of aligning known structure sequences to various scaffolds showed that the present energy function assigns self-recognition with the lowest energy and its structurally analogous scaffold with the second lowest energy. The alignments of these sequences to their structural analogues are similar to ones produced from NW with a few exceptions. One is the alignment of 256B(A) to 2CCY. For 256B(A), most standard alignment programs do not assign 2CCY as a homologue. Our energy function not only chooses this as the second most stable structure, but also produces an alignment with considerable structural overlap.

Table 4. *Constraints used in the alignment of the copper A domain of the subunit II of cytochrome c oxidase*

Stability 1[a]:	30.0
Stability 2[b]:	15.0
Penalty[c]:	−40.0

Residue 1	Residue 2	Experimental distance[d]
H181	C216	6.30 Å
H181	M227	5.10 Å
C216	M227	5.15 Å
H181	C220	8.50 Å[e]
H181	H224	8.50 Å[e]
C216	C220	5.50 Å[e]
C216	H224	5.50 Å[e]
C220	M227	5.50 Å[e]
C220	H224	5.50 Å[e]
H224	M227	5.50 Å[e]

[a] The energy contribution to stabilize a desired interaction within range 1.

[b] The energy contribution to stabilize a desired interaction within range 2.

[c] The coefficient of the energy contribution to destabilize an unfavorable interaction.

[d] The experimental distance assigned to each interaction was based on the largest distance of the corresponding interaction between ligand binding residues found in the blue copper structures.

[e] These constraints were applied to the second distance range only.

Alignments produced from the mean-field dynamic programming can be sometimes be improved by using experimental constraints. Adding experimental constraints to the alignment can help guide an alignment to its correct fold by forcing it out of any local minima in which it might be caught. This may imitate Nature in that it is very likely that chain topology in metalloproteins is partially fixed by interactions with the metal ion, not solely by the interactions in the simple alignment energy function. For illustration, this technique was used to improve the prediction of the soluble copper A domain of subunit II of cytochrome *c* oxidase. The constraints guided the alignment so that the copper ligand binding residues were appropriately configured.

There is an on-going survey of a subset of sequences from the Swiss-Prot database. This representative subset includes the 349 sequences from the *Escherichia coli* gene-protein database referenced in the Swiss-Prot database. These sequences are being aligned to a minimal representative set of unrelated scaffolds. The results of this survey will be published in a future article.

It is currently not computationally efficient to run a sequence against the complete PDB using this method. However, this is not necessary because the energy function is based on structural compatibility, and the alignments are only performed against a minimal representative set of unrelated folds. If the scaffold with the lowest energy has structural analogues, a further refinement can be performed by aligning the target sequence against the complete set of analogues to obtain the most energetically favorable structure.

The alignment procedure using the self-consistent energy function is not as trivial as standard sequence-sequence alignments. Our method of alignment uses only structural information from the scaffold and not the identity of its residues. Therefore, it is possible for a sequence to be threaded to its native structure and not be perfectly aligned. This self-consistent energy function is most efficient in recognizing sequence–structure compatibility when the sequence identity between two sequences is low (less than 21%). The energy function is dependent upon the structural information learned from known proteins. The greater the diversity of folds used in the training set, the better the energy function should be at recognizing distant sequence–structure compatibility. The quality of the sequence–structure alignment is dependent upon the availability of known scaffolds or structures. The predicted structure is only as good as the structural similarity between it and the available scaffolds.

Materials and methods

Pairwise contact energy term

The multiple-body term in the contact energy E_{ct} (Equation 3) monitors the cysteine–cysteine interactions during an alignment and prevents multiple bonds to any single cysteine. At the start of an alignment, the γ_1^{ct}(Cys,Cys) parameter is set to 0.0, the C_β distances between the various cysteines calculated, and the pair with the minimum distance found. The optimal γ_1^{ct}(Cys,Cys) value is assigned for the interaction between this minimal pair. The next shortest cysteine–cysteine distance is then found. If the two cysteines in this interaction are not one of the cysteines involved in the previous interaction, the optimal γ_1^{ct}(Cys,Cys) value is assigned to this interaction. This process is continued until all cysteine pair interactions have been evaluated. The cysteine–cysteine monitoring is performed for each successive refinement of alignment.

Hydrogen bonding energy term

The hydrogen bonding energy E_{hb} (Equation 4) allows for two types of backbone hydrogen bonds, α-helix and β-sheet. An α-helix E_{hb} energy contribution is assigned to residues A_i and A_j when $j = i + 4$ and both residues are in an α-helix, as defined previously by DSSP algorithm (Kabsch & Sander, 1983). The assignment of β-sheet hydrogen bonds is made through a complex algorithm based on oxygen distance and physiochemical information that can distinguish between parallel and antiparallel sheets (Fig. 11). The residues A_i and A_j are given one β-sheet E_{hb} energy contribution when assigned in a parallel sheet formation and two β-sheet E_{hb} energy contributions when assigned in an antiparallel sheet formation.

Experimental constraint energy term

The mean-field threading program uses the experimental constraint energy term (E_{cs}) to stabilize an alignment whenever the constraint is met in the scaffold and to destabilize the alignment whenever the constraint is not met. The stabilization parameter, *cvalue*, and penalty, *pvalue*, are determined empirically. For experimental constraints involving a distance between two residues, A_i and A_j, the experimental distance between two characteristic atoms (usually C_β) r_{ij}^{expt} must be known. To avoid conflicts with the cut-off values in the contact term, we considered two possible distance tolerances: range 1, Δr_1^{con} and range 2, Δr_2^{con}. The first range, Δr_1^{con}, spans r_{ij}^{expt} − 2 Å to r_{ij}^{expt}, and the second range, Δr_2^{con} spans r_{ij}^{expt} − 2 Å to r_{ij}^{expt} + 2 Å.

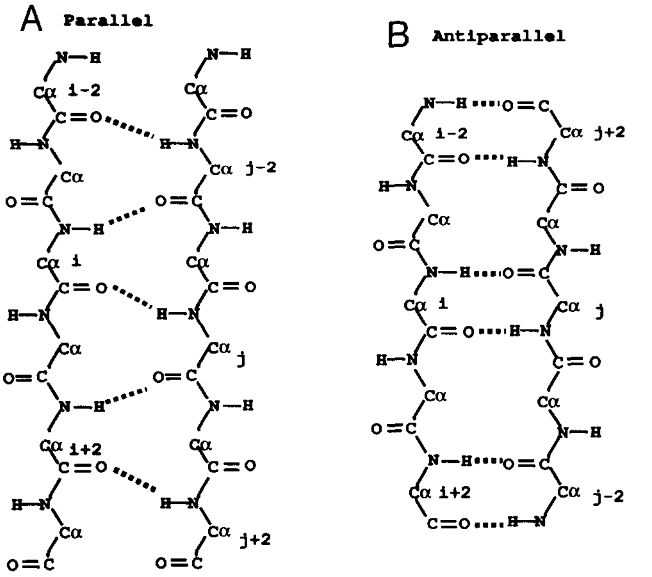

Fig. 11. A description of the β-sheet hydrogen bond assignment. Two general requirements for β-sheet hydrogen bond assignment are: the oxygen distance between residues i and j must be less than 4.0 Å and both residues must have been defined previously to be in a β-sheet by the DSSP algorithm. **A:** One hydrogen bond energy contribution will be assigned to the interaction between residues i and j for parallel sheet formation when both of the general requirements are met for the i, j pair and either $i + 2, j + 2$ or $i - 2, j - 2$ interactions also meet the two requirements. **B:** Method by which two hydrogen bond energy contributions will be assigned to the interaction of residues i and j for antiparallel sheet formation when both general requirements are met for the i, j pair, the C_α distance between residues i and j is less than the distance between $i, j - 2$ and $i, j + 2$, and neither $i + 2, j + 2$ or $i - 2, j - 2$ interactions meet the general requirements. We do not assign different energies for parallel or antiparallel sheets here, but this definition will allow such a distinction to be made.

These two constraints are applied during the alignment algorithm when the alignment score is calculated for residue A_i to be at various positions in the scaffold $A_{i'}$ (Goldstein et al., 1994, 1995). The program calculates the C_β distance between residues in the scaffold $r_{i'j'}$ in which A_i and A_j are pinned. If this distance is within range 1, the score for A_i to be pinned at $A_{i'}$ is given an added energy stability of *cvalue*(1). If not, the score is given a penalty that is a linear function of distance determined by:

$$penalty = pvalue(1)|r_{i'j'} - r^{con}|, \qquad (14)$$

where

$$r^{con} = \begin{cases} r_{ij}^{expt} - 2\text{ Å} & \text{if } r_{i'j'} < r_{ij}^{expt} - 2\text{ Å} \\ r_{ij}^{expt} & \text{if } r_{i'j'} > r_{ij}^{expt} \end{cases}$$

Next, the $r_{i'j'}$ distance is checked to see if it falls within range 2. If it does, the alignment score for A_i to be pinned to $A_{i'}$ is given an added energy stability of *cvalue*(2). If not, the penalty is applied (Equation 14). For experimental constraints that involve fixing segments of secondary structure, a similar stabilizing and destabilizing scheme is applied.

Self-consistent optimization

The zeroth-order approximation of the energy parameters, γ_0 was calculated in a similar fashion to the previous work of Goldstein et al. (1994, 1995). There the optimization of the energy parameters was a two-step process. The first step was to evaluate the energy function without the gap energy parameters and with the full ensemble of misfolded structures modeled by translating the set of training proteins' sequences over various unrelated scaffolds (Fig. 2). Once these energy parameters were calculated, the gap energy parameters were estimated.

To begin the work on self-consistency, a set of training proteins was chosen. The main requirement for the training set was that each one of the chosen proteins must have at least one structural analogue in the PDB (Bernstein et al., 1977; Abola et al., 1987). The alignments of the structural analogues to the training proteins were generated using a modified Dayhoff similarity score matrix (Gribskov & Burgess, 1986) in a modified Needleman-Wunsch algorithm (Needleman & Wunsch, 1970), which did not allow for gaps in a scaffold within α-helices or β-sheets nor which allow gaps with a C_α distance greater than 7.5 Å. We refer to alignments produced from this algorithm as physically constrained Needleman-Wunsch (P-NW) alignments. If an alignment had a q-score of 0.40 or greater, it was considered a structural analogue. A training set of 29 proteins consisting of 9 α-helical proteins, 9 β-sheet proteins, and 11 proteins with different percentages of both α/β and α+β folds was chosen (Table 1). The training proteins represent the following classes: lambda repressor-like DNA binding domains, Trp repressor, cytochrome c, 4 helix up-and-down bundle, globin-like, EF-hand, cupredoxin, immunoglobin-like β sandwiches, β-trefoil, viral coat proteins, acid protease, ferredoxin-like, lysozyme-like, ribonuclease A-like, OB-fold, dihydrofolate reductase, cysteine protease, subtilases, phosphofructokinase, NAD(P)-binding Rossmann fold domain, and periplasmic receptors. These training proteins were aligned to a set of 86 scaffolds with unrelated folds using the γ_0 energy parameters in a mean-field programming technique for sequence-structure alignment (Goldstein et al., 1994, 1995).

Once the initial alignments were produced for both the structural analogues and the 86 unrelated scaffolds, a first iterate optimization of the γ values was calculated. The γ_5^g parameter (Equation 5) was set equal to 0.0 in the optimization because it caused the $\mathbf{B}^{-1}$ matrix to be ill-conditioned due to lack of statistical information. The thermally occupied minima of misfolded structures were obtained by using the alignments of the training proteins to the 86 unrelated protein scaffolds, excluding the training protein and any structural analogues that might have been in the set of 86 proteins (Fig. 2). Therefore, at most, there were only 86 misfolded structures per training protein. Because these alignments had gaps, residues in a training sequence that were not aligned to any scaffold position were given an energy contribution of zero. Once these $\gamma_{n'=1}$ values were obtained, they were linearly combined with the γ_0 values to produce the final $\gamma_{n=1}$ values. The 29 training proteins were realigned to the same 86 unrelated protein scaffolds using the $\gamma_{n=1}$ values for the energy parameters, creating a new set of minima (Fig. 2) that were then used for a second iterate optimization according to Equations 6–9. The set of training proteins was not re-aligned to their structural analogues because the P-NW alignments were considered to be the optimal alignments.

This procedure of aligning the training proteins to the 86 unrelated scaffolds and then re-optimizing to get another iteration of γ values was repeated until a maximum degree of discrimination D_n averaged over the training set was achieved.

Alignment procedure

The alignments presented in this paper were produced using the current energy function with a mean-field dynamic programming algorithm described previously (Goldstein et al., 1994, 1995). Two different scoring matrices were used to generate two sets of initial alignments. The first set of initial alignments were generated by "frozen approximation," which constructs an initial scoring matrix S^0 based on the current energy function in which the identity $A^s_{i'}$ of the amino acids in the scaffold structure were used in evaluating the energy contributions of the contact term:

$$S^0_{ii'} = E_p + E_{hb} + \sum_{j'} \sum_{k} \gamma_k^{ct}(A_i, A_{j'}) u(r_k^{ct} - r_{i'j'}). \qquad (15)$$

Here $S^0_{ii'}$ is the initial scoring matrix for residue i of the target sequence to be pinned to residue i' of the scaffold. The contact between residues i and j' are based on the identity of residue i of the target sequence and residue j' of the scaffold sequence. The second set of initial alignments were generated using the P-NW program in which the scoring matrix is based on the modified Dayhoff similarity matrix (Gribskov & Burgess, 1986). Once initial alignments were produced, further refinements of the alignments were made with a mean-field iterated procedure. The alignments used to generate the thermally occupied minima of misfolded structures were based on only the first choice of initial alignments with three iterations of refinement. All other alignments were produced both with the self-consistently optimized energy function initial alignments and five iterations of refinement, and with the P-NW initial alignments followed by five iterations of refinement using the present energy function. The results of the two different types of alignments were compared, and the one with the most stable energy was considered the more stable fold. The initial alignment is the most difficult part of the threading program. Allowing for two different types of initial alignments helps prevent choosing an alignment that might have been trapped in a local minima.

Structural data

The 29 training proteins (shown in bold) and their corresponding structural analogues were selected from the PDB (Bernstein et al., 1977; Abola et al., 1987). Their PDB codes are: **2CRO**: 1R69; **1WRP(R)**: 2WRP(R); **1CCR**: 3C2C, 5CYT(R); **1HMQ(A)**: 2MHR; **1BP2**: 1P2P; **2HHB(A)**: 2HHB(B), 2LHB, 5MBN, 1MBA, 1HDS(B), 1HDS(A), 1FDH(G); **3CLN**: 4TNC; **1FDH(G)**: 1HDS(A), 1HDS(B), 1MBA, 2HHB(A), 2HHB(B), 2LH4, 2LHB, 5MBN; **5MBN**: 2LHB, 2HHB(B), 2HHB(A), 1MBA, 1HDS(B), 1HDS(A), 1FDH(G); **5PCY**: 1PAZ; **1REI(A)**: 2FB4(H), 2FB4(L), 2FBJ(H), 2FBJ(L), 2HFL(L), 2RHE, 3HFM(H), 3HFM(L), 4FAB(H), 4FAB(L), 1MCP(L), 1MCP(H), 1F19(L), 1F19(H); **2PAZ**: 1PAZ; **2I1B**: 4I1B; **1F19(H)**: 1F19(L), 1MCP(H), 1MCP(L), 2FB4(H), 2FB4(L), 2FBJ(H), 2HFL(L), 3HFM(H), 3HFM(L), 4FAB(H); **3HFM(H)**: 4FAB(H), 2FBJ(H), 2FB4(H), 1MCP(H), 1F19(H); **2PLV(3)**: 4RHV(3), 2MEV(3), 1R1A(3); **1R1A(2)**: 2MEV(2), 2PLV(2), 4RHV(2); **1CMS**: 2APR, 3APP, 4PEP; **1FDX**: 4FD1; **1ALC**: 1LZ1, 2LYZ; **1RBB(A)**: 1RNS, 1SRN(A); **1SNC**: 2SNS; **2DHF(A)**: 3DFR, 4DFR(B), 8DFR; **2ACT**: 9PAP; **2PRK**: 1TEC(E), 1SBT; **1PFK(A)**: 2PFK(A), 2PFK(D); **2LDX**: 6LDH; **3GPD(G)**: 4GPD(1); and **2LIV**: 2LBP. All proteins have a resolution of 2.5 Å or better.

The following 86 protein structures were selected from the PDB (Bernstein et al., 1977; Abola et al., 1987) for the set of unrelated scaffolds. Their PDB codes are: 1FDX, 2CRO, 3FXC, 5PCY, 1WRP(R), 3RNT, 256B(A), 1REI(A), 2CDV, 2SSI, 1TRX, 5CPV, 1CCR, 3C2C, 1HMQ(A), 2RHE, 1CY3, 2MHR, 1ALC, 2PAZ, 1BP2, 1RBB(A), 2CCY(A), 2AZA(A), 1LZ1, 1SNC, 3FXN, 2HHB(A), 2SNS, 3CLN, 1FDH(G), 1FX1, 2SOD(B), 1TNF(A), 5MBN, 2I1B, 2TMV(P), 4DFR(B), 4TNC, 3LZM, 1CD4, 1GCR, 2DHF(A), 2STV, 3ADK, 3GAP(A), 9PAP, 1F19(H), 3HFM(H), 2ACT, 1MCP(L), 4SBV(C), 2PLV(3), 3CNA, 1R1A(3), 2CGA(A), 1YPI(A), 1R1A(2), 2CAB,2PLV(2), 3HLA(A), 1SBT, 2PRK, 1RHD, 5CPA, 2GBP, 8ATC(A), 3TLN, 1PFK(A), 2TBV(C), 1CMS, 4PEP, 5HMG(A), 2LDX, 4MDH(A), 3GPD(G), 9API(A), 2LIV, 2LBP, 8ADH, 3XIA, 1WSY(B), 3ICD, 3PGK, 1CTS, and 3GRS. All proteins have a resolution of 2.5 Å or better.

Acknowledgments

We thank Professor Harry Gray and Ben Ramirez for inspirational discussions and input on the experimental chemical constraints for copper A domain of cytochrome *c* oxidase. This work was supported by the National Institutes of Health grant (PHS2 R02GM44557-06) and the Carver Grant. Computations were performed at the National Center for Supercomputing Applications in Urbana.

References

Abad-Zapatero C, Griffith J, Sussman J, Rossmann M. 1987. Refined crystal structure of dogfish m4 apo-lactace dehydrogenase. *J Mol Biol 198*:445.

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications*. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107-132.

Antholine W, Kastrau D, Steffens G, Buse G, Zumft W, Kroneck P. 1992. A comparative epr investigation of the multicopper proteins nitrous-oxide reductase and cytochrome-*c* oxidase. *Eur J Biochem 209*:875-881.

Bairoch A, Boeckmann B. 1994. *Nucleic Acid Res 22*:3578-3580.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535-542.

Bolin J, Filman D, Mathhews D, Hamlin R, Kraut J. 1982. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 angstroms resolution. *J Biol Chem 257*:13650.

Bowie J, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164-112.

Brünger AT. 1987. *X-PLOR version 3.1. A system for X-ray crystallography and NMR*. New Haven, Connecticut: Yale University Press.

Bryngelson J, Onuchic J, Socci N, Wolynes P. 1995. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins Struct Funct Genet 21*:167-195.

Bryngelson J, Wolynes P. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA 84*:7524-7528.

Church W, Guss J, Potter J, Freeman H. 1986. The crystal structure of mercury-substituted poplar plastocyanin at 1.9 angstroms resolution. *J Biol Chem 261*:234.

Dill K, Bromberg S, Yue K, Fiebig K, Yee D, Thomas P, Chan H. 1995. Principles of protein folding – A perspective from simple exact models. *Protein Sci 4*:561–602.

Dill K, Stigter D. 1995. *Modeling protein stability as heteropolymer collapse, vol 46*. New York: Academic Press, Inc. pp 59-104.

Epp O, Lattman E, Schiffer M, Huber R, Palm W. 1975. The molecular structure of a dimer composed of the variable portions of the Bence-Jones protein rei refined at 2.0 angstroms resolution. *Biochemistry 14*:4943.

Farrar J, Lappalainen P, Saraste M, Thomson A. 1995. Spectroscopic and mutagenesis studies on the cu-a centre from the cytochrome-*c* oxidase complex of *Paracoccus denitrificans*. *Eur J Biochem 232*:294–303.

Ferrin TE, Huang CC, Jarvis LE, Langridge R. 1988. The midas display system. *J Mol Graphics 6*:13–27.

Garrett T, Saper M, Bjorkman P, Strominger J, Wiley D. 1989. Specificity pockets for the side chains of peptide antigens in hla-aw68. *Nature 342*:692.

Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol 227*:227–238.

Goldstein R, Luthey-Schulten Z, Wolynes P. 1992a. Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA 89*(11):4918–4922.

Goldstein R, Luthey-Schulten Z, Wolynes P. 1992b. Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc Natl Acad Sci USA 89*:9029–9033.

Goldstein R, Luthey-Schulten Z, Wolynes P. 1994. A Bayesian approach to sequence structure alignment algorithms for protein structure recognition. In: *Proceedings of the 27th Hawaii International Conference on System Sciences*. Los Alamitos, California: IEEE Computer Society Press. pp 306–315.

Goldstein R, Luthey-Schulten Z, Wolynes P. 1995. The statistical mechanical basis of sequence alignment algorithms for protein structure recognition. In: Elber R, ed. *New developments in theoretical studies of proteins*. Singapore: World Scientific.

Gribskov, Burgess. 1986. Sigma factors from *E. coli*, *B. subtilis*, phage sp01 and phage t4 are homologous proteins. *Nucleic Acids Res 14*(16):6745–6763.

Han J, Adman E, Beppu T. 1991. Resonance ramon spectra of plastocyanin and pseudoazurin: Evidence for conserved cysteine ligand conformation in cupredoxins (blue copper proteins). *Biochemistry 30*:1-904-13.

Holmgren A, Soderberg BO, Eklund H, Branden CI. 1990. Crystal structure of thioredoxin from *Escherichia coli* at 1.68 angstroms resolution. *J Mol Biol 212*:167.

Iwata S, Ostermeier C, Ludwig B, Michel H. 1995. Structure at 2.8 angstroms resolution of cytochrome *c* oxidase from *Paracoccus denitrificans*. *Nature 376*:660–669.

Ji X, Sesay MA, Dickert L, Prassad SM, Johnson WW, Ammon HL, Armstrong RN, Gilliland GL. 1994. Structure and function of the xenobiotic substrate binding site of a glutathione S-transferase as revealed by X-ray crystallographic analysis of product complexes with the diastereomers of 9-(S-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene. *Biochemistry 33*:1043.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kelly M, Lappalainen P, Talbo G, Haltia T, van der Oost J, Saraste M. 1993. Two cysteines, two histidines and one methionine are ligands of a binuclear purple copper center. *J Biol Chem 268*:16781–16787.

Maiorov V, Crippen G. 1994. Learning about protein folding via potential functions. *Proteins Struct Funct Genet 20*:167–173.

Mathews F, Bethge P, Czerwinski E. 1975. The structure of cytochrome *b*562 from *Escherichia coli* at 2.5 angstroms resolution. *J Biol Chem 254*:1699.

McTigue M, Davies J, Kaufman B. 1993. Crystal structure of chicken liver dihydrofolate reductase complexes with nadp+ and bipterin. *Biochemistry 32*:6855–6862.

McTigue MA, Bernstein SL, Williams DR, Tainer JA. 1995. Purification and crystallization of a schistosomal glutathione S-transferase. *Proteins Struct Funct Genet 22*:55–57.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Murzin A, Brenner SE, Hubbard TJP, Chothia C. 1995. Scop – Structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536–540.

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

Nishikawa K, Matsuo Y. 1993. Development of pseudoenergy potentials for assessing protein 3-d-1-d compatibility and detecting weak homologies. *Protein Eng 6*(8):811–820.

Padlan E, Silverton E, Sheriff S, Cohen G, Smith-Gill S, Davies D. 1989. Structure of an antibody-antigen complex. Crystal structure of the hy/hel-10 fab-lysozyme complex. *Proc Natl Acad Sci USA 86*:5938.

Quiocho F, Vyas N. 1984. Novel stereospecificity of L-arabinose-binding protein. *Nature 310*:381–386.

Richards FM. 1977. Areas, volumes, packing, and protein structure. *Annu Rev Biophys Bioeng 6*:151–176.

Sasai M, Wolynes P. 1990. Molecular theory of associative memory hamiltonian models of protein folding. *Phys Rev Lett 65*:2740–2743.

Sippl M. 1993. Boltzmann principle, knowledge-based mean fields and protein folding – An approach to the computational determination of protein structures. *J Comput-Aided Mol Design 7*(4):473–501.

Smith W, Burnett R, Darling G, Ludwig M. 1977. Structure of the semiquinone form of flavodoxin from *Clostridium* sp. extension of 1.8 angstroms resolution and some comparisons with the oxidized state and reduced forms. *J Mol Biol 117*:195.

Srinivasan R, Rose G. 1995. Linus: A hierarchic procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.

Steffens G, Soulimane T, Wolff G, Buse G. 1993. Stoichiometry and redox behavior of metals in cytochrome *c* oxidase. *Eur J Biochem 213*:1149–1157.

Stenkamp R, Sieker L, Jensen L, McQueen J. 1978. Structure of methhemerythrin at 2.8 antgstroms resolution. Computer graphics fit of an averaged electron density map. *Biochemistry 17*:2499.

Takano T. 1977. Structure of myoglobin refined at 2.0 angstroms resolution. Structure of deoxymyoglobin from sperm whale. *J Mol Biol 110*:569.

van der Oost J et al. 1992. *EMBO J 11*:3209–3217.

Vyas NK, Vyas MN, Quiocho FA. 1988. Surgar and signal transducer binding sites of the *Escherichia coli* galactose chemoreceptor protein. *Science 242*:1290.

Watenpaugh K, Sieker L, Jensen L, Legall J, Dubourdieu M. 1972. Structure of the oxidized form of a flavodoxin at 2.5-angstroms resolution. Resolution of the phase ambiguity by anomalous scattering. *Proc Natl Acad Sci USA 69*:3185.

Weber P, Bartsch R, Cusanovich M, Hamlin R, Howard A, Jordan S, Kamen M, Meyer T, Weatherford D, Xuong N, Salemme F. 1980. Structure of cytochrome *c* prime. A dimeric, high-spin haem protein. *Nature 286*:302.

Zuker M. 1991. Suboptimal sequence alignment in molecular biology alignment with error analysis. *J Mol Biol 221*(2):403.