Article No. mb982092 *J. Mol. Biol.* (1998) **283**, 507–526

JMB

Protein Structure Prediction by Threading. Why it Works and Why it Does Not

Leonid A. Mirny and Eugene I. Shakhnovich*

Harvard University Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138 USA

We developed a novel Monte Carlo threading algorithm which allows gaps and insertions both in the template structure and threaded sequence. The algorithm is able to find the optimal sequence-structure alignment and sample suboptimal alignments. Using our algorithm we performed sequence-structure alignments for a number of examples for three protein folds (ubiquitin, immunoglobulin and globin) using both "ideal" set of potentials (optimized to provide the best *Z*-score for a given protein) and more realistic knowledge-based potentials. Two physically different scenarios emerged. If a template structure is similar to the native one (within 2 Å RMS), then (i) the optimal threading alignment is correct and robust with respect to deviations of the potential from the "ideal" one; (ii) suboptimal alignments are very similar to the optimal one; (iii) as Monte Carlo temperature decreases a sharp cooperative transition to the optimal alignment is observed. In contrast, if the template structure is only moderately close to the native structure (RMS greater than 3.5 Å), then (i) the optimal alignment changes dramatically when an "ideal" potential is substituted by the real one; (ii) the structures of suboptimal alignments are very different from the optimal one, reducing the reliability of the alignment; (iii) the transition to the apparently optimal alignment is non-cooperative. In the intermediate cases when the RMS between the template and the native conformations is in the range between 2 Å and 3.5 Å, the success of threading alignment may depend on the quality of potentials used.

These results are rationalized in terms of a threading free energy landscape. Possible ways to overcome the fundamental limitations of threading are discussed briefly.

© 1998 Academic Press

**Corresponding author*

Keywords: threading; Monte Carlo procedure; protein structure prediction

Introduction

The problem of predicting protein conformation from sequences is of great importance and has drawn a lot of attention recently (see e.g. Moult *et al*., 1997; Shakhnovich, 1997a; Finkelstein, 1997; Jones, 1997; Levitt, 1997) with hundreds of papers from dozens of groups.

A most desirable solution to the problem is to find a model and an algorithm that stimulate folding of a protein pretty much in a way that mimics natural protein folding and converges to the native conformation. While some success along these lines has been documented (Kolinski & Skolnick, 1994), this approach encounters a number of serious technical difficulties, making *ab initio* structure prediction hardly feasible now and perhaps in the foreseeable future (Finkelstein, 1997; Ortiz *et al*., 1998; Mirny & Shakhnovich, 1996). The main reason for such a reclusion was discussed by Shakhnovich (1997a), Finkelstein (1997) and Mirny & Shakhnovich (1996): a "good" folding model must be detailed enough to reproduce energetics faithfully and yet simple enough to be computationally feasible, a blend that has not been reached yet.

The energetics requirement is a very important one as far as folding is concerned: the energy function must be precise enough to single out the unique native structure as the global energy minimum among an astronomically large number of decoys, some of which have very low energy (Shakhnovich, 1994). The precision of potentials

Abbreviation used: MC, Monte Carlo.

E-mail address of the corresponding author: eugene@diamond.harvard.edu

0022–2836/98/420507–20 $30.00/0 © 1998 Academic Press

required to achieve this goal was analyzed by Bryngelson (1994) and Pande *et al*. (1995) for lattice model chains and by Mirny & Shakhnovich (1996) for real proteins. The analysis by Mirny & Shakhnovich (1996) suggests that at the level of a simple two-body approximation of energetics and structure-less amino acids there may not exist any potential which is able to fold real proteins into their native conformations. Adding more details into the model is probably the way to go. However, this complicates the search in conformational space, making computations far more demanding. Thus, *ab initio* folding success is contingent on finding a safe pathway between the Scilla of incorrect energetics and the Kharibda of a too complicated and thus computationally infeasible model.

The complications inherent in *ab initio* folding were realized early by a number of workers in the field, and alternative approaches were suggested; the most notable of them is threading (Finkelstein & Reva, 1991; Finkelstein, 1997). The key idea of the threading method is to decrease dramatically the number of decoys. This is achieved by constraining all protein conformations to a smaller subset of conformations obtained by threading through known protein structures that serve as a scaffold for the protein sequence in question and finding the energetically optimal alignment of the sequence to the scaffold structure. From the physical point of view the threading problem is somewhat equivalent to folding, because it also requires searching over a large set of possible alignments for the one that delivers minimum ''energy''. It was shown (Lathrop, 1994) that such a search is an NP complete problem (i.e. that there is an apparent ''Levinthal'' paradox in threading). As in folding, the search in threading is biased by the energy function, so that the related key issue is the precision of the energy function. The rationale for using threading rather than folding is the hope that a less precise energy function will suffice for the search over a more constrained conformational set: in this case the native state should be distinguished as having the lowest energy among the smaller number of alternatives. However, such simplification of the conformational space comes at a serious price. The reason is that the native structure itself may not belong to the constrained conformational set! In this case, threading seeks an approximate solution, i.e. the one that is closest to the native state in the conformational set of alignments. However, if this best solution is relatively distant structurally from the native state, its energy may be considerably higher than the energy of the native state (even with the ''ideal'' potential that strongly favors the native state). This factor clearly may decrease the energy gap, balancing on the negative side the gain achieved due to the restriction of conformational space. Obviously, the conformational space restriction, which is the basis of the threading approach, is not an ''innocent'' approximation that is guaranteed to work almost by definition. Clearly this issue requires a detailed study that aims to address a question of gains and losses made by threading approximations and add to our intuition about which factors are more important for particular models and when should we expect success in threading simulations and when we cannot.

As pointed out above, threading is very much like folding in terms of key questions and difficulties. In folding one asks basically two questions: are energy functions correct? and is a conformational search efficient enough to find the global minimum? An important advance in protein folding theory, which started from the seminal work of Go (Taketomi *et al*., 1975), is understanding that those two questions can be studied separately. The first approach proposed by Go (Taketomi *et al*., 1975) was to design an energy function that favors the native contacts and disfavors non-native ones. Such an energy function gives rise to fast folding (Gutin *et al*., 1996). However, the artificial penalties imposed on non-native interactions make the model somewhat unphysical for studying the physical principles of folding, because in real life strong non-native contacts cannot be excluded *a priori*. In fact they occur in some proteins (Lacroix *et al*., 1997). A more physical model for folding is based on sequence design, which generates, for any given potential function, special sequences for which the native structure is guaranteed to be the global minimum. Then the same potential is used for folding as the one used to design sequences (Shakhnovich, 1994; Shakhnovich *et al*., 1996a). In this case folding simulations quickly converge to the native state (Gutin *et al*., 1996; Shakhnovich, 1994). The properties of energy landscape and dynamics that lead to the native state can be studied in detail with implications for folding and evolution of real proteins (Shakhnovich *et al*., 1996a). An approach that is very similar in spirit is to design a potential which provides low energy to a natural sequence in its native structure (Goldstein *et al*., 1992; Mirny & Shakhnovich, 1996; Koretke *et al*., 1996; Hao & Scheraga, 1996; Ortiz *et al*., 1998). While this energy function may not be transferable to other proteins (Mirny & Shakhnovich, 1996) it serves its purpose by providing a free energy landscape with a large gap and hence reasonably fast folding. One conclusion from the analysis carried out for a number of folding models suggests that Monte Carlo (MC) simulation represents a powerful search strategy that is very efficient in finding the native state on a physically reasonable landscape.

Here we take a similar systematic approach to study threading. First we develop and present a Monte Carlo threading algorithm, which allows gaps and insertions both in structure and in sequence. The advantage of the Monte Carlo approach is that it converges to the Boltzmann distribution. This feature makes this method a valuable tool to map and characterize a free energy landscape and outline the physical requirements of convergence to the global minimum solution.

To test and rationalize the MC threading method we apply it to a number of problems of increasing complexity. The reason behind this ''gradual'' approach is the need to differentiate between limitations intrinsic to the threading approach, as suggested above, and the ones that originate from uncertainties in potential functions and the form of the Hamiltonian (scoring function) used.

Following this program, first we use our algorithm for the structure-structure alignment. It turns out that in the framework of our approach the structure-structure alignment is a counterpart of the Go model studied in folding: it provides the ''ideal'' potential function in which only the native interactions are favorable. This simple implementation of the method makes it straightforward to compare it with existing heuristic approaches such as the structure alignment algorithm *Dali* (Holm & Sander, 1993) used to build the FSSP database. The comparison shows that the MC procedure proposed in this work gives a more optimal alignment than *Dali* (with the same scoring function as used in *Dali*).

Next, we turn to a more realistic two-body energy function where interaction energy depends on amino acid types, and distances between them rather than on their location in the native structure. We design an ''ideal'' parameter set for this Hamiltonian scoring function and study how the approximate character of the potential function affects the results of threading at various degrees of similarity between the template and the native conformations.

Moving closer to the realm of structure prediction, we explore the accuracy of threading, using one of the knowledge based potentials that are currently available in the literature (Miyazawa & Jernigan, 1996).

In order to make our conclusions significant and general we carried out the analysis for three fold classes: α/β (ubiquitin and its structural homologues), all-β (class I immunoglobulin fold) and all-α (globin fold). The results are consistent between the fold classes studied. This allows us to arrive at quantitative conclusions concerning the degree of similarity between the native structure and the template that is required for successful sequence-structure alignment.

In what follows we provide a detailed discussion of the ubiquitin superfamily followed by the data (Tables 4 and 5) with comments for the immunoglobulin and globin fold.

Model

MC threading

To sample possible sequence-structure alignments and search for the alignment with the minimal energy we use the Monte Carlo (MC) procedure. The power of the MC procedure is that it allows us to find a global minimum on a variety of rough landscapes (Allen & Tildesley, 1987). In the search for a minimum, it samples possible alignments and allows us to study statistical properties of the energy landscape. This made the MC procedure extremely useful in the study of various disordered physical systems such as spin-glasses (Binder, 1995), liquids (Allen & Tildesley, 1987) and proteins (Shakhnovich, 1997b; Abkevich *et al*., 1995; Pande *et al*., 1997; Sali *et al*., 1994).

We develop the MC procedure to sample sequence-structure alignments. In contrast to previous work (Bryant, 1996; Lathrop & Smith, 1996), the whole space of possible alignments is sampled, allowing arbitrary gaps in both sequence and structure and arbitrary fragments of matching sequence and structure. Gap penalties are not used.

General scheme

To construct the MC procedure one needs to find a suitable representation of sequence-structure alignment and design a move set which could make sampling effective, i.e. provide a high acceptance ratio for the trial moves. In Methods we describe the alignment representation and the move set we designed.

Simulation starts from a random alignment. At each step of the procedure we make a trial move which slightly changes the alignment. Then we compute how energy changes, ΔE, upon this move. The move is accepted or rejected according to the Metropolis scheme (Metropolis *et al*., 1953). A move is always accepted if $\Delta E \leqslant 0$, and is accepted with probability $P = \exp(-\Delta E/T)$ if $\Delta E > 0$, where temperature T is a parameter of the procedure. Irrespective of whether the move is accepted or rejected another step is done and so on.

This procedure was proven to converge to the Boltzmann ensemble (Allen & Tildesley, 1987). In the simulation, the equilibrium ensemble is obtained by recording an alignment every 1000 steps. The interval is important to avoid correlation between sampled alignments (Binder, 1986). The typical length of the run is 10^6-10^7 steps which is 10-100 times longer than the typical time required to reach the alignment with the minimal energy (at the optimal temperature). An ensemble obtained in this way is used to compute average values of different quantities, such as energy and various distance measures (see below).

Energy function

The energy of an alignment is given by the sum of pairwise residue-residue interactions. If a structure of protein $\mathcal{I}$ is represented by a set of pairwise distances between residues $r^{\mathcal{I}}_{ii'}, i, i' = 1, \ldots, I$ and a sequence of protein $\mathcal{J}$ aligned to this structure is $a_j, j = 1, \ldots, J$ then:

$$E = \sum_{i<i'=1}^{I} V(p_i, p_{i'}, r^{\mathcal{I}}_{ii'}) \qquad (1)$$

where $V(j, j', R)$ is the energy of interaction between residues j and j' located at distance R in space, and p_i is the number of the residue in the sequence aligned with position i in the structure (see Methods: Alignment representations). If a_j and a'_j are identities of residues in positions j and j', then:

$$V(j, j', R) = U(a_j, a'_j, R), \quad (2)$$

where $U(\xi, \eta, R)$ is the potential which gives the energy of interaction between residue *types* ξ and η located at distance R in space. Particular forms of $V(j, j', R)$ and $U(\xi, \eta, R)$ used for threading are described below.

To write the energy function in this form we made the following assumptions.

(1) Energy is a sum of pairwise residue-residue interactions.

(2) Interaction between residues is a function of their identities only (not of the distance between them along the chain (Sippl, 1995) or their local environments).

(3) Energy of interaction between residues depends on the distance between some ''representative'' points of the residues (C^α, C^β, center of mass, etc.) and not on their relative orientation (Berriz *et al.*, 1997; Bahar & Jernigan, 1996).

In this study all the distances are measured between C^β atoms. We do not use penalties for gaps. Instead, similar to Holm & Sander (1993), we constrain the length of a fragment to be greater than or equal to $L_{\min} = 6$.

Note that threading with a pairwise potential is equivalent to alignment of two matrices: $r_{ii'}$ and $V(j', j, \cdot)$ under the constraint that diagonals of the matrices coincide.

Results

Testing MC search strategy with an ''ideal'' potential

Our first goal is to evaluate the proposed MC threading as a search strategy for threading as well as to probe a possible alignment energy landscape. The results of threading, however, always depend on both the potential and the search strategy. In order to test the search strategy and eliminate the problem of inaccurate potential we use an ''ideal'' potential.

''Ideal'' potential is the one which guarantees the lowest energy to the native conformation. Using *dRMS* (distance RMS, see Methods) as the energy function is an example of an ''ideal'' potential (Elofsson *et al.*, 1996). Clearly, the native conformation has $dRMS = 0$ and all other conformations have $dRMS > 0$. In the case of contact potential, the so-called Go-matrix (Go & Abe, 1981), which we described in the Introduction, can serve as an ''ideal'' potential. Namely, if i and i' are in contact (i.e. $r_{ii'} < R_{\text{cut}}$) in the native structure, then $V(j, j', r < R_{\text{cut}}) = -1$ and 0 otherwise, where R_{cut} is a contact cutoff. Then all conformations which do not have all the native contacts have higher energy than the native conformation. Below we use different ''ideal'' potentials to test the MC threading algorithm.

Since an ''ideal'' potential guarantees the lowest energy to the native conformation, success of threading and quality of alignment depend only on the search strategy (threading algorithm).

Threading is equivalent to structure comparison

When the energy $V(j, j', R)$ of interaction between residues j and j' is a function of $r^{\mathcal{J}}_{jj'}$ (distance between j and j' in the native structure $\mathcal{J}$), threading becomes equivalent to structure-structure comparison. For example, when:

$$V(j, j', R) = \frac{(r^{\mathcal{J}}_{jj'} - R)^2}{(N_{\text{alg}} - 1)(N_{\text{alg}} - 2)} \quad (3)$$

the energy function becomes equal to $dRMS^2$:

$$E = \frac{1}{(N_{\text{alg}} - 1)(N_{\text{alg}} - 2)} \sum_{i+2<i'=1}^{I} (r^{\mathcal{J}}_{jj'} - r^{\mathcal{I}}_{ii'})^2$$

$$= dRMS^2(\mathcal{I}, \mathcal{J}) \quad (4)$$

where N_{alg} is the length of alignment, i.e. number of matched residues. Hence, we can use our threading algorithm for structure-structure alignment as well as for threading. In general any threading algorithm which allows pairwise distance-dependent potential can be used for structure comparison.

Then the first test for a threading algorithm is to apply it to protein structure-structure alignments and compare the results with those obtained by a widely used structure comparison algorithm (Holm & Sander, 1993).

Testing alignment procedure: structure-structure alignment

Here we compare our results with the structure alignments made by Holm & Sander available in the FSSP database (Holm & Sander, 1993). FSSP was built using the *Dali* algorithm (Holm & Sander, 1993), which utilizes the following scoring function for structure comparison:

$$V(j, j', R) = \left(0.2 - 2 \times \frac{r^{\mathcal{J}}_{jj'} - R}{r^{\mathcal{J}}_{jj'} + R}\right) \exp(-(r^{\mathcal{J}}_{jj'} + R)/40) \quad (5)$$

We use the negative of this *Dali* scoring function as a potential for MC threading. In fact, this potential represents an ''ideal'' Lennard-Jones-like potential for threading. It has a deep minimum at distance $R = r^{\mathcal{J}}_{jj'}$, rapidly increases at small $R < r^{\mathcal{J}}_{jj'}$ and

slowly increases at $R > r^{\mathcal{J}}_{jj'}$, approaching a constant value. (The exponent leads to a very slow decrease of the potential at large $R \approx 40$ Å.) This shape makes the *Dali* potential equivalent to an ''ideal'' distance-dependent potential, which has a minimum when two residues are at the separation that they have in the native structure.

Finding the optimal alignment

To assess our Monte Carlo threading as a search strategy we apply it to structure-structure comparison. Optimal alignments obtained by our procedure are then compared with alignments provided in the FSSP database. Importantly, we use exactly the same measure of fold similarity as the one used by Holm & Sander to build the FSSP database.

Here we consider a few examples of structures which have similar folds and no significant sequence similarity:

ubiquitin-ubiquitin self-alignment (1*ubi*) : Ubi-Ubi

ubiquitin-cRaf1 (1*gua*) : Ubi-Gua

ubiquitin-G-protein (1*igd*) : Ubi-Igd

ubiquitin-ferrodoxin, chain A(1*frr*) : Ubi-Frr

PDB identifiers are taken for abbreviation of each pair. As a control we also use a pair of proteins with no structural similarity:

ubiquitin-plastocyanin (1*plc*) : Ubi-Plc

This control mimics the case when the wrong structure has been chosen as a template for threading.

For every pair, we compare the optimal alignment found by our MC threading and the alignment reported in the FSSP. Table 1 summarizes the results of this test.

For all three cases and for the self-alignment we find alignments that are very close to the FSSP alignments. Moreover, our procedure clearly ''outperforms'' the *Dali* algorithm used to build the FSSP. Using the same scoring (energy) as *Dali*, our procedure finds alignments which have better scores (lower energy) than alignments reported by FSSP. FSSP was built using a heuristic algorithm, which found alignments close to, but different from the optimal ones (Holm & Sander, 1993). In contrast, our procedure uses the Monte Carlo search strategy and is known to find the global optimum when simulated annealing is properly set up (Kirkpatrick, 1984).

From these results we conclude that our procedure successfully passed the first test. These results also demonstrate the superiority of the Monte Carlo procedure in comparison to a heuristic algorithm in the search for the global optimum. Clearly, Monte Carlo threading can easily find the right alignment when an exact distance-dependent potential of residue-residue interactions is provided. Can similar success be achieved when a contact potential (also an ''ideal'' one) is used instead?

Contact potential *versus* distance-dependent potential

A vast majority of protein models currently used for threading rely on contact approximation, i.e. two residues are said to interact with each other if the distance between them is less than a cutoff distance. The energy in contact approximation is given by:

$$E = \sum_{i<i'=1}^{I} B_{p_i,p'_i} \Delta^{\mathcal{I}}_{ii'} \tag{6}$$

where $B_{jj'}$ is the energy of a contact between residues in positions j and j' of protein $\mathcal{J}$, and $\Delta^{\mathcal{I}}_{ii'} = 1$ if positions i and i' are in contact in protein $\mathcal{I}$ and $\Delta^{\mathcal{I}}_{ii'} = 0$ otherwise. Two residues i and i' are said to be in contact if the distance between their C^{β} atoms $r_{ii'} < R_{\text{cut}} = 8.0$ Å. If a_j, $j = 1, \ldots, J$ is the sequence of the protein $\mathcal{J}$, then the energy of a contact $B_{jj'} = U(a_j, a_{j'})$, where $U(\zeta, \eta)$, $\zeta, \eta = 1, \ldots, 20$ is a potential of interactions between residues of types ζ and η. The long-standing question is how this approximation affects the accuracy of threading alignments.

To address this question we make threading using both contact and distance-dependent potentials. Comparing optimal alignments obtained with these models demonstrates the impact of contact approximation on the accuracy of alignments.

Table 1. Comparison of the optimal structure-structure alignments obtained by minimization of the *Dali* function by MC threading and by the *Dali* algorithm (Holm & Sander, 1993) (as reported in FSSP (Holm & Sander, 1997))

Proteins	FSSP			MC threading			
	Dali	*dRMS* (Å)	L_{alg}	*Dali*	*dRMS* (Å)	L_{alg}	Q_{alg}
Ubi-Ubi	–	0.00	76	−2832	0.00	76	1.00
Ubi-Gua	−729	2.05	68	−1047	2.03	58	0.82
Ubi-Igd	−344	2.34	46	−362	2.58	50	0.87
Ubi-Frr	−579	2.79	61	−648	2.40	60	0.62

The distance between alignments is given by Q_{alg} (see Methods).

Table 2. Comparison of the optimal structure-structure alignments which minimize distance-dependent function (columns 2 to 4) and overlap between contacts (columns 5 to 6)

Proteins	Contact			Distance-dependent			
	Dali	*dRMS* (Å)	L_{alg}	*Dali*	*dRMS* (Å)	L_{alg}	Q_{alg}
Ubi-Ubi	−2832	0.00	76	−2832	0.00	76	1.00
Ubi-Gua	−1029	2.16	69	−1047	2.03	58	0.96
Ubi-Igd	−239	3.15	52	−362	2.58	50	0.90
Ubi-Frr	−591	3.06	66	−648	2.40	60	0.84
Ubi-Plc(control)	592	5.54	65	−272	2.67	42	0.30

The distance between alignments is given by Q_{alg}.

An ''ideal'' potential for the contact approximation is given by:

$$B_{jj'} = -\Delta^{\mathcal{J}}_{jj'} \tag{7}$$

Threading with ''ideal'' contact potential is equivalent to structure-structure alignment, which maximizes the number of common contacts between the two structures. In other words, we align contact maps of the two proteins to maximize $Q = \Sigma^{\mathcal{I}}_{i<i'=1} \Delta^{\mathcal{J}}_{p_i,p'_i} \Delta^{\mathcal{I}}_{ii}$, which is the overlap between contacts in the matrices. Table 2 compares optimal alignments with contact and distance-dependent approximations for the same five cases.

Clearly, using contacts instead of a distance-dependent potential introduces very small changes in the alignments. These changes are predominantly shrinkage/expansion of fragments by one or two residues. We also observed that the closer the two proteins are, the smaller the effect on contact approximation on the accuracy of the alignment.

The major advantage of the contact approximation, compared to the distance-dependent one, is the small number of parameters required to define a potential. Since contact approximation does not change the optimal alignment very much it can be used efficiently for threading. Below we use the contact approximation for all threading experiments.

Thermodynamics of alignment

Here we study the generic properties of the threading energy landscape. Following the approach of ''ideal'' potential we make structural alignments of ubiquitin with cRaf1, G-protein, ferrodoxin, plastocyanin and ubiquitin itself. Note that these alignments are equivalent to threading of the ubiquitin sequence through corresponding structures when an ''ideal'' potential is used.

For every case we make several MC runs, each at different constant temperature T. Each run yields an equilibrium ensemble of alignments. Averaging over this ensemble we obtain $\langle E \rangle$, $C_v = (\langle E^2 \rangle - \langle E \rangle^2)/T^2$, and $\langle Q_{alg} \rangle$ as a function of T (see Figures 1 and 2). The temperature at which C_v has the main peak is the transition temperature T_f.

In order to study suboptimal alignments we perform a long MC run at transition temperature T_f. Alignments sampled every 1000 MC steps constitute an equilibrium ensemble of suboptimal alignments. For this ensemble of alignments we compute frequency w_{ij} of a match between every pair of residues i and j as:

$$w_{ij} = \frac{1}{M} \sum_{m=1}^{M} \delta_{j,p^m_i} \tag{8}$$

where p^m is the mth alignment out of M in the ensemble. Figures 3A, 4A and 5A present w_{ij} for three pairs of proteins considered above.

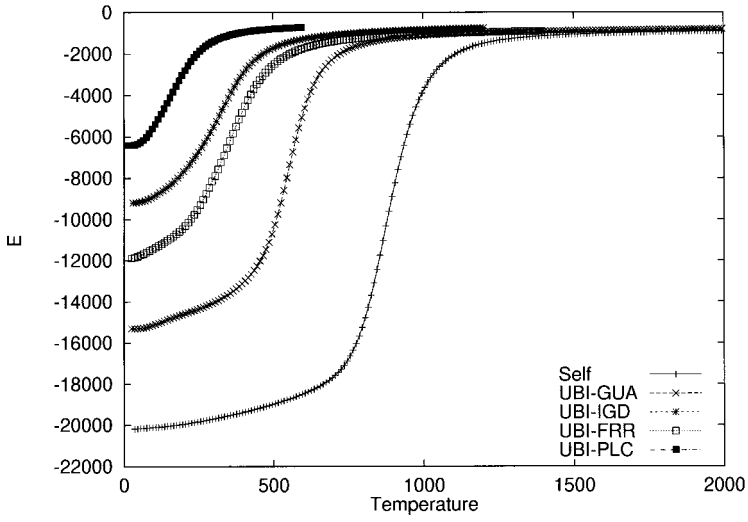

Figure 1. The average energy in equilibrium ensemble as a function of temperature for different pairs of aligned proteins.

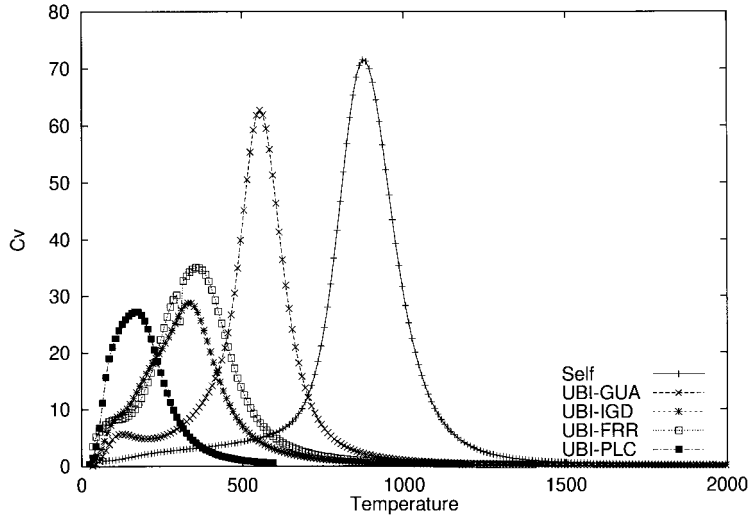

Figure 2. Heat capacity C_υ as a function of temperature. The main peak on C_υ corresponds to the transition temperature (T_f).

After normalization:

$$\tilde{w}_{ij} = \frac{w_{ij}}{\Sigma_{j=1}^{J} w_{ij}}$$

we compute the positional entropy of alignments S_i as:

$$S_i = -\sum_{j=1}^{J} \tilde{w}_{ij} \log \tilde{w}_{ij}. \quad (9)$$

Positional entropy S_i introduced in this way clearly measures the degree of uncertainty in matching the residue i. If i is always matched with j, then pos-

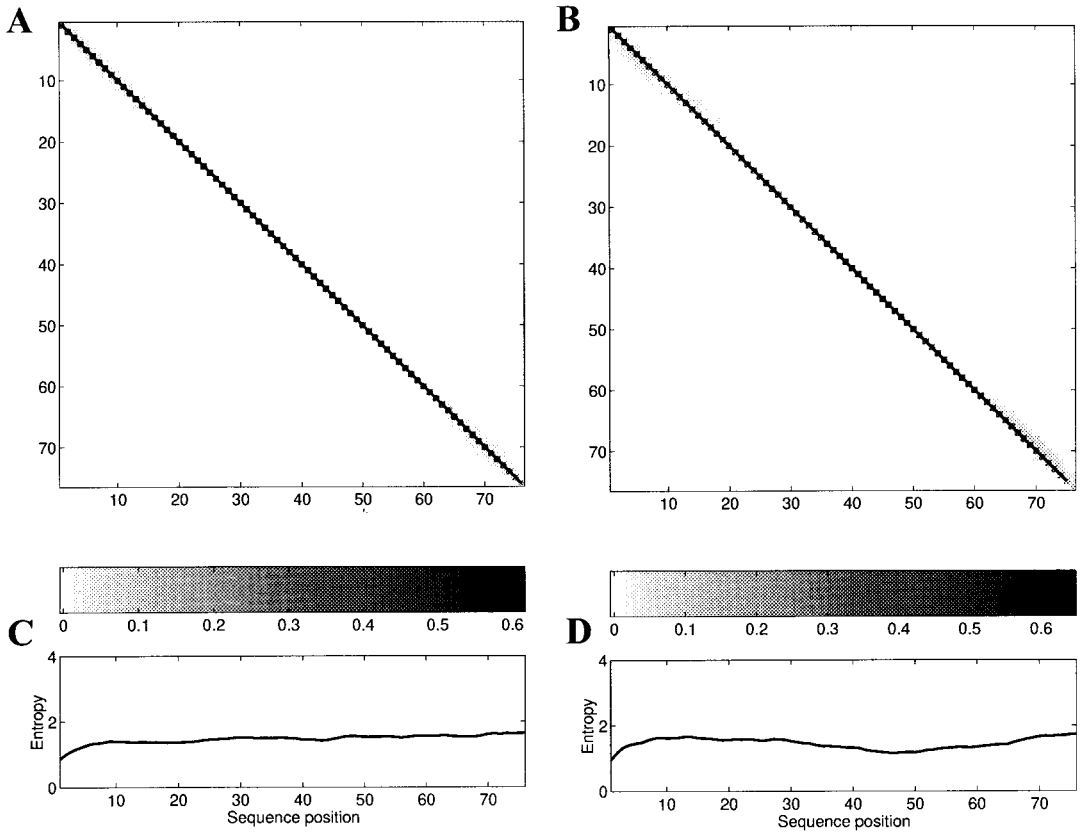

Figure 3. Self-alignment of ubiquitin. Equilibrium ensemble of alignments for structure-structure (A and C) and sequence-structure alignments (B and D) obtained at T_f. A and B, Distribution of matches w_{ij} in the alignments. The continuous line shows the optimal alignment. C and D, Positional entropy S_i, which is the measure of reliability of alignment.

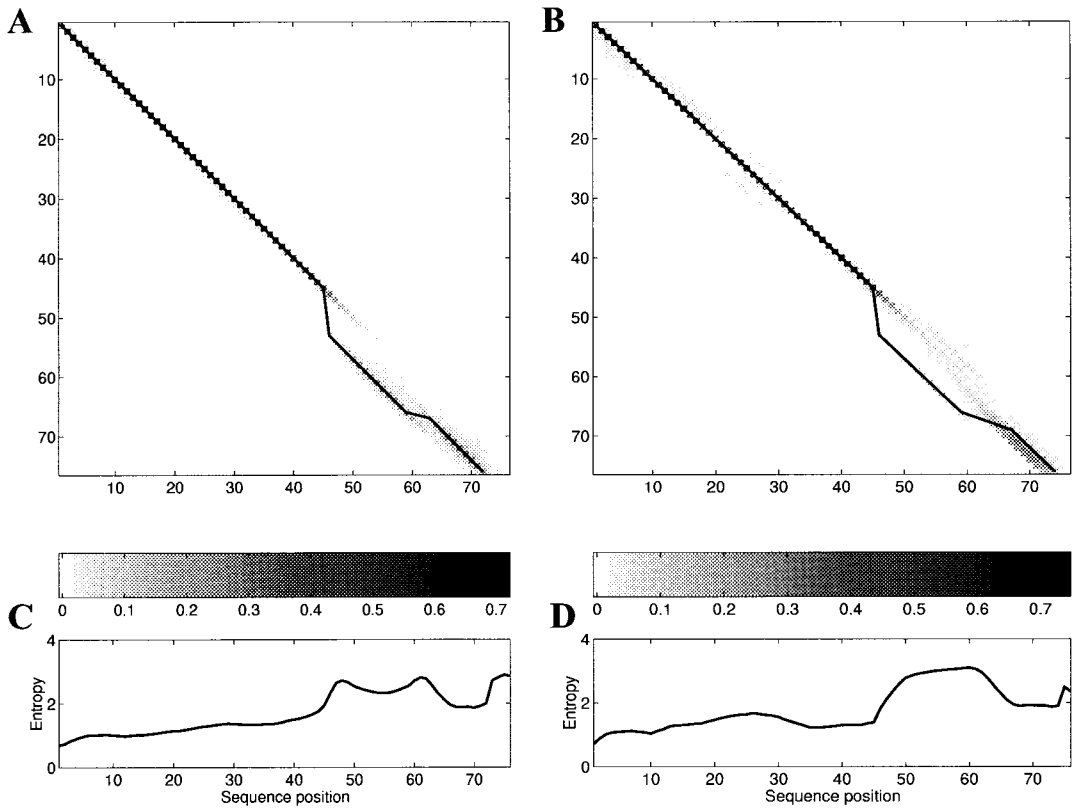

Figure 4. Ubiquitin-cRaf1 protein. Equilibrium ensemble of alignments for structure-structure (A and C) and sequence-structure alignments (B and D) obtained at T_f. A and B, Distribution of matches w_{ij} in the alignments. The continuous line shows the optimal alignment. C and D, Positional entropy S_i.

itional entropy $S_i = 0$. In contrast, S_i has its maximum if i is matched equally often with all j terms. Figures 3C, 4C and 5C present S_i for the three cases studied.

Self-threading

First we consider the case of self-alignment, i.e. the structural alignment of ubiquitin with itself (Ubi-Ubi). Figures 1 and 2 present average energy E and heat capacity $C_\upsilon = dE/dT$ as a function of T.

As we raise the temperature, T, a sharp transition in E is observed (see Figure 1). The jump in E shows that at high T the ensemble consists predominantly of incorrect high-energy alignments, whereas at low T the optimal (''native'') alignment dominates in the ensemble. What is more important is that this transition is very cooperative (it has a clear first order-like type). The cooperative nature of the transition can easily be seen from the sharpness of transition in E, the peak on C_υ (Figure 2) and the bimodal histogram (Figure 6A) obtained at transition temperature (T_f). Drawing a parallel with protein folding, we can say that the two peaks on the energy histogram correspond to two distinct coexisting states: ''folded'' (correct optimal alignment) and ''unfolded'' (random alignments). Both ''folded'' and ''unfolded'' states are free energy minima. As T decreases the ''unfolded'' state is destabilized. What is more important, as in the first-order transition scenario, the incorrect alignments with relatively low energy have high free energy and are not populated in the equilibrium ensemble obtained by the MC procedure, because equilibrium ensemble should be dominated by the correct alignment and a few suboptimal alignments with very similar structure.

Figure 3A presents alignments constituting the equilibrium ensemble. Clearly, the optimal alignment (main diagonal, for this case) dominates over all alternative alignments. Suboptimal alignments, which contain incorrect i,j matches, constitute a tiny fraction of the ensemble, since $w_{ii} \approx 0.6$ and $w_{ij \neq i} < 0.01$. This domination of the optimal alignment is also manifested in a uniformly low positional entropy S_i (see Figure 3C).

Analysis of the self-alignment test brings us to the following conclusions: (i) the optimal alignment can be easily found by the MC procedure; (ii) the transition for random alignments to the optimal alignment is cooperative; (iii) the optimal alignments dominates in the ensemble. Although self-alignment is an essential test for any threading procedure, it definitely represents the simplest possible case. Now we consider more realistic cases when the structure of ubiquitin is aligned against various similar folds.

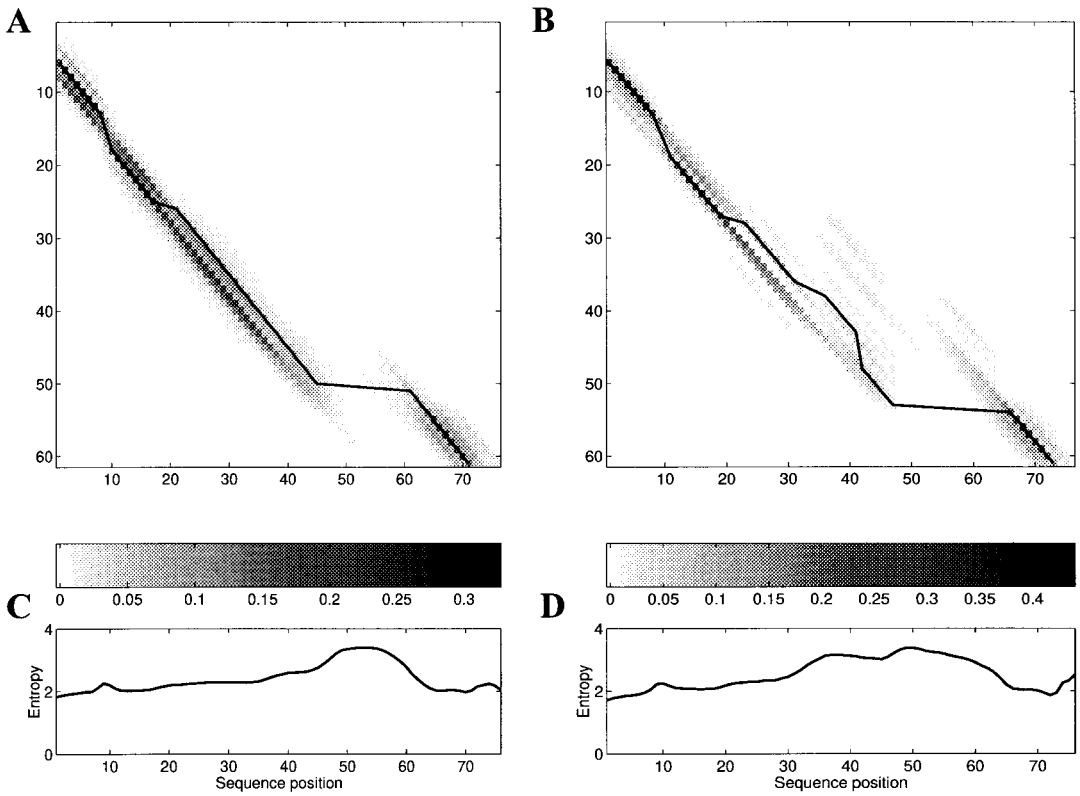

Figure 5. Ubiquitin-G-protein. Equilibrium ensemble of alignments for structure-structure (A and C) and sequence-structure alignments (B and D) obtained at T_f. A and B, Distribution of matches w_{ij} in the alignments. The continuous line shows the optimal alignment. C and D, Positional entropy S_i.

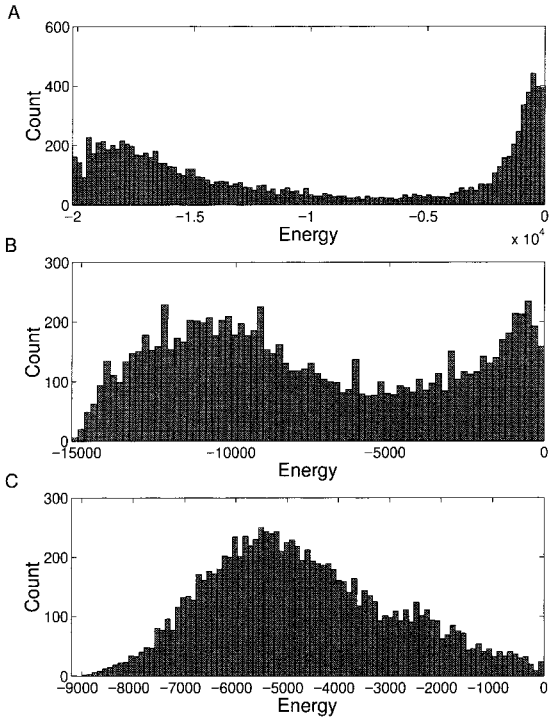

Figure 6. Distribution of alignment energies at T_f. A, Self-alignment of ubiquitin; B, ubiquitin-cRaf1; C, ubiquitin-cRaf1 protein. Bimodal distribution is a clear indicator of cooperative first-order-like transition.

High structural similarity: ubiquitin-cRaf1 protein

Now we make an alignment of ubiquitin and cRaf1 structures (Ubi-Gua). For this pair the FSSP database (Holm & Sander, 1997) reports $dRMS = 2.2$ Å for 58 atoms, providing $Z = 5.5$.

As in the case of self-alignment the transition is very sharp (see Figure 1) yielding a peak of C_v almost as high as for the self-alignment (see Figure 2). Another observation is that self-alignment has a transition temperature T_f higher than all other cases. The transition temperature for ubiquitin-cRaf1 $T_f^{\text{ubi-Gua}}$ is the next highest. To ensure that transition in this case retains its first-order character we make a histogram of the energy of the equilibrium ensemble at T_f (see Figure 6B). The observed bimodal distribution is a clear indication of the first-order transition.

As in the case of self-alignment, the optimal alignment dominates over the alternative alignments in the ensemble (see Figure 4A). Positional entropy, however, reveals a region of alignment (residues 50 to 60), which has a high degree of uncertainty in placing a gap and a fragment (see Figure 4C). The rest of the alignment has very low entropy and, hence, high certainty.

As with self-alignment, the ubiquitin-cRaf1 pair exhibits a first-order-like transition, which is associated with high accuracy and is manifested by a high peak in C_v and a bimodal energy distribution.

Does every pair of aligned proteins exhibit the first-order transition or is it an indicator of a good ''match'' between the two structures (or a sequence and a structure in the case of threading)? To address this question we consider two other pairs of proteins with a moderate structural similarity and a control case with no structural similarity.

Low structural similarity: ubiquitin-G-protein and ubiquitin-ferrodoxin

Both pairs of proteins share a common fold, but do not exhibit a great deal of similarity. According to the FSSP database (Holm & Sander, 1997) the ubiquitin-G-protein pair has $dRMS = 2.8$ Å for 46 atoms providing very moderate $Z = 2.8$, and the ubiquitin-ferrodoxin pair has $dRMS = 3.6$ Å for 63 atoms, providing a better $Z = 3.6$.

Both of these proteins exhibit no first-order transition in structural alignments with ubiquitin as can be seen for small C_υ peaks (Figure 2) and a monomodal distribution of energies in the equilibrium ensemble at T_f (see Figure 6C). As expected, moderate similarity between structures leads to the absence of cooperative transition.

Importantly, accuracy of the alignment also suffers a lot, as fluctuations in several positions of alignment increase. There is a substantial degree of uncertainty in placing every residue. A fuzzy pattern of w_{ij} shows that in most suboptimal alignments every residue is shifted two or three positions from its optimal match (see Figure 5A). These alternative alignments are very frequent in the equilibrium ensemble and have an energy only slightly above the optimal energy. A higher level of S_i is another indicator of this uncertainty (see Figure 5C).

No structural similarity: ubiquitin-plastocyanin

To consider the limiting case when proteins have almost no structural similarity we make an alignment of ubiquitin and plastocyanin structures. The transition in this case is very smooth and the peak in C_υ is smaller than for all other cases (see Figures 1 and 2). Alignment exhibits a great deal of uncertainty in several places (data not shown). This case represents a ''reference point'' opposite to that of self-alignment. For any pair of aligned proteins we can measure the peak of C_υ and compare it with two other peaks: the peak of C_υ for self-alignment and the peak for this reference, poor alignment.

Summary

Pairs of proteins, which are similar to each other, exhibit cooperative transition and provide the optimal alignment with a high reliability. In contrast, proteins which share a smaller degree of similarity have substantial uncertainties in alignments and exhibit no first-order transitions.

Absence of the first-order-like transition can be a sensitive indicator of a moderate similarity between the native structure of a protein and the structure used for threading. What is more important, it indicates that there is a huge number of suboptimal alignments which are different from the optimal alignments, but have the energy very close to the optimal one. This represents a very difficult case for threading. In fact, small (inevitable!) errors in the potential can decrease the energy of suboptimal alignments and increase the energy of the optimal alignment making it a non-optimal one. We consider this effect in more detail below.

Threading with a real potential

No residue-residue potential is able to provide an ''ideal'' pattern of interactions, where all native contacts are attractive and all non-native are ambivalent (see equation (7).) In fact, a potential assigns interaction energies to residue types, not positions! Here we study how threading alignment changes when we use a full residue-residue potential instead of an ''ideal'' one. In other words, instead of making structure-structure alignment we make a sequence-structure alignment with a real potential. We, however, chose a residue-residue potential $U(\xi\eta)$, $\xi, \eta = 1, \ldots, 20$ to provide a pattern of interactions $B_{jj'} = U(a_j, a_{j'})$, which resembles the ''ideal'' pattern as closely as possible. Essentially, we use the ''best'' potential in place of an ''ideal'' one and study changes in alignments associated with this minimal change in potential. In other words, we introduce a minimal inevitable ''noise'' in the ''ideal'' potential and study the stability of the optimal alignment to this noise.

The ''best'' potential is obtained by an optimization procedure described by Mirny & Shakhnovich (1996) and, briefly, in Methods. We optimize the residue-residue potential (20×20 matrix) for a single protein (ubiquitin in this case) in order to make as many native contacts attractive and non-native contacts repulsive as possible. This optimal potential provides the lowest energy to the ubiquitin sequence in the native structure of ubiquitin and maximizes the energy gap between this structure and the bulk of alternative ones (Mirny & Shakhnovich, 1996). Using this potential we make MC threading of the sequence of ubiquitin through the four alternative structures (ubiquitin itself, cRaf1, G-protein and ferrodoxin). Optimal alignments obtained in this way are then compared with the optimal alignments obtained by structure-structure comparison (see above).

Self-alignment

Self-alignment of ubiquitin using residue-residue potential shows that the native structure remains at the global energy minimum and the optimal alignment does not change. Although seven matches are lost from the optimal alignment (see Table 3) there are no displacements in the align-

Table 3. Comparison of the optimal structure-structure and optimal sequence-structure alignments for ubiquitin

Proteins	Structural alignment ("ideal" potential)			Threading the "best" potential				Threading MJ96 potential			
	L_{alg}	RMS	*dRMS*	L_{alg}	RMS[a]	*dRMS*[a]	Q_{alg}	L_{alg}	RMS[a]	*dRMS*[a]	Q_{alg}
Ubi-Ubi	76	0.00	0.00	69	0.00	0.00	0.92	74	1.83	1.79	0.82
Ubi-Gua	69	2.00	1.80	62	2.76	2.65	0.75	74	5.16	4.21	0.33
Ubi-Igd	52	3.01	2.49	46	4.17	3.40	0.53	55	6.92	4.98	0.36
Ubi-Frr	66	3.03	2.68	64	4.49	3.96	0.61	71	5.42	4.87	0.26
Ubi-Plc	65	5.62	6.07	54	10.23	10.62	0.09	–	–	–	–

RMS is computed over C^{α} atoms after optimal superposition of the matched residues. *dRMS* is computed over C^{β} atoms. $RMS_{C\alpha}$ is a better measure of the distance between the backbones, whereas $dRMS_{C\beta}$ is better in characterizing similarity between folds.

[a] For fair comparison of three alignments (structural, "best" and MJ96) *dRMS* and *RMS* were computed over the same length (the length of the shortest alignment). Q_{alg} is the distance between the threading and structural alignments. It measures the degree of success in threading.

ment. A low level of positional entropy S_i (see Figure 3B and D) demonstrates domination of the optimal alignment over the alternative ones.

Ubi-Gua

The optimal structure-structure alignment is no longer the lowest energy. Another alignment becomes the optimal one when we make threading with the "best" potential. This optimal threading alignment is slightly different from the optimal structure-structure alignment (see Table 3). Particularly, the alignment gets shorter as some matches are lost. More important, the last fragment of threading alignment (eight residues) shifts by one residue (compare continuous lines in Figure 4A and B). This shift leads to a slight increase of *dRMS*, but still yields absolutely accurate mounting of the ubiquitin sequence on the cRaf1 for the rest $87\% = (62 - 8)/62$ of the residues.

Comparing the density of matches w_{ij} for structure-structure (Figure 4A) and structure-sequence alignments (Figure 4B), we observe an increased contribution from wrong alignments in the sequence-structure case. The major part of the optimal alignment, however, sustains this competition brought about by deviation from "ideality" in the residue-residue potential.

Ubi-Igd

For this case, the best structure-structure alignment is by far not the lowest one in energy. Threading with the "best" residue-residue potential changes the optimal alignment drastically. Three out of six fragments are misplaced. All boundaries of the fragments are changed. Only 53% of the matches in the optimal threading alignment are present in the structural one. Alignment gets shorter and provides RMS = 4.17 Å for 46 C^{α} atoms, compared to RMS = 3.01 Å for 46 best matching C^{α} atoms for the structure-structure alignment.

Importantly, threading suboptimal alignments becomes very different from the optimal one and are present with substantial frequency in the equilibrium ensemble (Figure 5B). These lead to low reliability of the optimal alignment (Figure 5D). Note that this result is obtained using the "best" residue-residue potential, optimized for the sequence of ubiquitin.

Threading with a knowledge-based potential

We showed that when a minimal possible "noise" is introduced into potential, the optimal alignment with a distant template changes substantially, yielding RMS > 5 Å. In contrast, alignments with close templates sustain minimal possible "noise" in the "best" potential. How does an optimal alignment with different templates change when a realistic knowledge-based potential is used?

Table 3 presents the results of threading with the potential derived by Miyazawa & Jernigan (Table 4, upper half of Miyazawa & Jernigan, 1996) (MJ96). As expected, the accuracy of alignment decreased dramatically yielding RMS > 5 Å for all pairs. Even self-alignment of ubiquitin becomes inaccurate with $RMS(C^{\alpha}) = 2.36$ Å over 74 residues. Importantly, threading with the close template (Ubi-Gua) was rather accurate with the "best" potential but broke down with MJ96.

In other words, while threading through a distant template becomes inaccurate, even under a minimal possible "noise", threading through a close template can sustain minimal "noise", breaking down under a higher level of "noise" in potential.

However, there is a possibility that an accurate enough potential can provide rather good alignment on a very close template.

We conclude that further deviation of potential from the "ideal" one leads to the situation that low-energy decoys become energetically optimal rather than the structurally optimal alignment. To understand this phenomenon more deeply we study the energy landscape of threading.

Energy landscape of threading

The aim now is to study the general properties of the alignment energy landscape, i.e. how the energy of an alignment changes as it approaches the optimal one. The degree of similarity between two alignments is measured by Q_{alg} (see Methods).

In order to study the energy-similarity relationship for each case we perform a long MC run at the transition temperature. Alignments sampled every 1000 MC steps constitute the equilibrium ensemble. We compute the number of alignments in the ensemble which have the energy in the interval $[E, \in E + \delta E\}$ and similarity to the optimal alignment $Q_{alg} \in \{Q_{alg}, Q_{alg} + Q\delta_{alg}]$.

Structure-structure alignment

Figure 7A, B and C present logarithms of the number of alignments in each energy-similarity interval for different cases.

These results clearly demonstrate the fundamental difference between the cases which exhibit first-order transition (self-threading and Ubi-Gua) and those which do not (Ubi-Igd and Ubi-Frr). In the case of the first-order transition (see Figure 7A and B) we observe a well-focusing landscape, i.e. a pronounced correlation between the degree of similarity Q_{alg} and energy. Two states are predominantly populated: ''folded'' low E, high Q_{alg}; and ''unfolded'', high E, low Q_{alg}. In contrast, when no first-order transition is observed, the landscape is much less focusing, i.e. there are several low energy alignments different from the optimal one (low E, high Q_{alg}). These decoy low-energy alignments constitute a serious danger for the success of a threading procedure. As explained above structure-structure comparison is equivalent to threading with an ''ideal'' potential. Then, a potential different from an ''ideal'' one can decrease an already small energy difference between the optimal and the decoy alignments and, hence, make one of the decoys the optimal one. As shown above, that is exactly the case when the residue-residue potential is used for threading.

Sequence-structure alignment

Figure 7D, E and F present logarithms of the number of alignments in each energy-distance for threading with the ''best'' residue-residue potential. Comparing these with corresponding ones for structure-structure alignment (A, B and C) we observe the following. (i) Both Ubi-Ubi (self-alignment) and Ubi-Gua energy landscapes sustain the change of the potential and both have focusing landscape. (ii) The threading of Ubi-Igd, however, suffers a lot from the application of a non-ideal potential (compare Figure 7C and F). The optimal alignment becomes very different from the structural one. As expected, one of the former low-

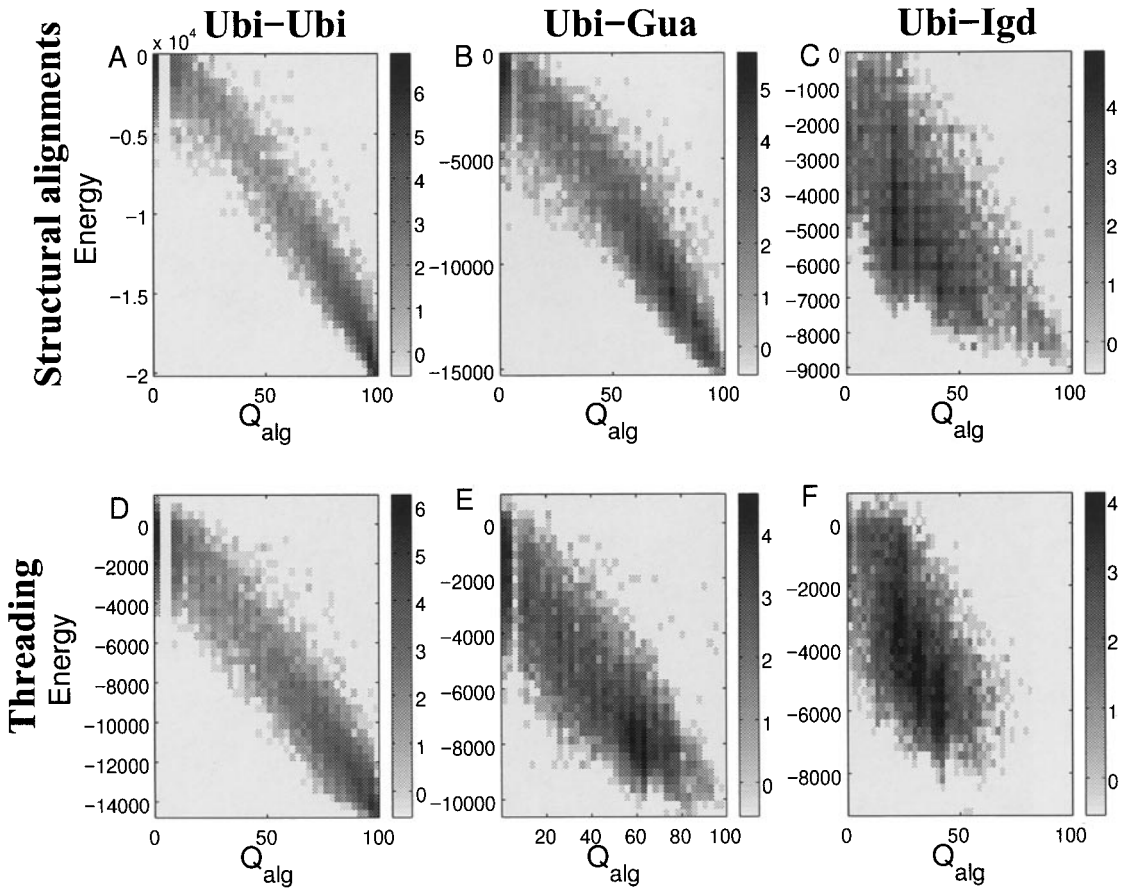

Figure 7. Energy landscapes of threading. The gray level shows the number of alignments in each E, Q_{alg} interval. A, B and C, Structure-structure alignments. D, E and F, Sequence-structure alignments. Pairs of proteins used are shown as the title of each column. All for the equilibrium ensembles of alignments obtained at T_f.

Table 4. Comparison of the optimal structure-structure and optimal sequence-structure alignments for the immunoglobulin fold

	Structural alignment (''ideal'' potential)			Threading the ''best'' potential				Threading MJ96 potential			
Proteins	L_{alg}	RMS	*dRMS*	L_{alg}	RMS	*dRMS*	Q_{alg}	L_{alg}	RMS	*dRMS*	Q_{alg}
Ten-Ten	89	0.00	0.00	84	0.00	0.00	0.94	84	3.27	3.24	0.60
Ten-Fnf	89	0.97	1.00	82	0.94	0.90	0.93	88	2.64	2.86	0.57
Ten-HhrB	80	4.41	3.25	82	2.74	2.26	0.77	87	4.26	3.98	0.47
Ten-Hnf	73	2.88	3.02	73	3.95	3.73	0.53	84	16.89	11.75	0.13
Ten-Cid	74	2.73	2.78	69	3.67	3.31	0.23	82	13.96	8.66	0.17
Ten-Tit	80	5.34	3.52	75	4.83	4.42	0.34	80	5.39	5.05	0.34

Notation as for Table 3. Proteins codes used in the Table and sequence identity with 1ten: Ten, tenascin (100%); Fnf3, fibronectin (3-d domain) (22%); HhrB, human growth hormone (19%); Hnf, CD2 (8%); Cid, CD4 (11%); Tit, titin (9%).

energy decoys becomes the optimal alignment, while the structural alignment obtains a rather high energy. (iii) The nicely focusing landscape of Ubi-Gua is deformed and several low-energy decoys with $Q_{alg} \approx 60\%$ appear (compare Figure 7B and E). We expect that further changes in the potentials can make one of the decoys the optimal alignment.

These results bring us to the conclusion that the optimal alignment may be sensitive to the noise in the potential. The degree of this sensitivity depends on the degree of structural similarity between the native structure and the template structure used for threading. The closer the template structure is to the native structure, the less sensitive the optimal alignment. It is crucial to establish the generality of this conclusion and make it as quantitative as possible. To achieve that, other threading examples should be considered. In the following subsection we briefly discuss threading experiments for two more proteins folds: immunoglobulins (all β) and globins (all α).

Examples of threading with other proteins

In order to generalize conclusions drawn from the study of proteins with a ubiquitin-like fold (αβ class) we repeated the analysis for proteins of immunoglobulin (β class) and globin (α class) folds. For each fold we make structural alignments and threading alignments with both ''best'' and MJ96 potentials. Tables 4 and 5 summarize the results of these experiments.

Consistent with our results for ubiquitin, the further the template is from the native structure, the more sensitive the alignments are to the noise in potential.

By choosing templates of increasing distance (RMS) from the native structure we can see how alignment accuracy depends on the distance. The structure of fibronectin (Fnf) is very close to the structure of tenascin with RMS < 2 Å. Threading of tenascin's sequence through the structure of fibronectin gives very accurate alignment for both ''best'' and MJ96 potential. A template that is close to the native structure can provide good structure predictions, even with an available knowledge-based contact potential (of course, if the threading algorithm is good enough to find alignments of the lowest energy!). A bit more distant structure of human growth hormone (HhrB) provides a good alignment with the ''best'' potential ($Q_{alg} > 0.75$, RMS < 3 Å), but it misses the right alignment with a more ''noisy'' MJ96 potential ($Q_{alg} < 0.5$, RMS > 4 Å). More distant from HhrB template structure of tenascin provide RMS > 3 Å even for the ''best'' potential and yields totally wrong alignments with MJ96 (see Table 5).

For a single domain of hemoglobin (Ash) even the closest available structure (Fal) gives a poor alignment with RMS > 4 Å when the ''best'' potential is used. Colicin A, a well-known structural analog of hemoglobin yields completely incorrect alignments even for the ''best'' potential. We see that as in the other cases, the templates that are distant from the native structure can not allow successful in threading with any potential.

Table 5. Comparison of the optimal structure-structure and optimal sequence-structure alignments for globin fold

	Structural alignment (''ideal'' potential)			Threading the ''best'' potential				Threading MJ96 potential			
Proteins	L_{alg}	RMS	*dRMS*	L_{alg}	RMS	*dRMS*	Q_{alg}	L_{alg}	RMS	*dRMS*	Q_{alg}
Ash-Ash	147	0.00	0.00	141	0.00	0.00	0.93	146	0.77	0.82	0.95
Ash-Fal	140	1.91	1.76	123	4.19	3.44	0.62	143	3.44	3.02	0.81
Ash-Hbg	137	3.16	2.60	115	5.77	4.36	0.51	138	5.87	4.62	0.31
Ash-ColA	114	4.24	3.43	102	7.44	6.05	0.42	146	15.30	8.68	0.00

Notation as for Table 3. Protein codes used in the table and sequence identity with 1ash; Ash, hemoglobin (domain one) (100%); Fal, myoglobin (12%); Hbg, hemoglobin(deoxy) (14%); ColA, colicin a (6%).

MC threading versus *"frozen" approximation*

Finally we compare the performance of Monte Carlo threading with the widely used ''frozen'' approximation (Flockner *et al.*, 1997). Under this approximation the energy of a target sequence a_j, $j = 1, \ldots, J$ of protein $\mathcal{J}$ threaded into a structure of protein $\mathcal{I}$ is assumed to be a sum of energies of single mutations needed to make sequence $\mathcal{J}$ out of the *sequence* b_i, $i = 1, \ldots, I$ of protein $\mathcal{I}$. If the energy of protein $\mathcal{I}$ after a single mutation $b_k \to a_k$ is:

$$e(b_k \to a_k) = \sum_i U(b_i, a_k)\Delta_{ik}$$

then the energy of the whole alignment p_i under ''frozen'' approximation is:

$$E = \sum_{i'=1}^{I} e(b_{i'} \to a_{p'_i}) = \sum_{i<i'=1}^{I} U(b_i, a_{p'_i})\Delta_{ii'}$$

The ''frozen'' approximation essentially turns a problem of sequence-structure alignment into a sequence-sequence alignment with a position-dependent substitution matrix, allowing application of fast and exact sequence alignment algorithms (Smith & Waterman, 1981; Needleman & Wunsch, 1970). This approximation, however, is very crude, since all interactions between residues of the target sequence $\mathcal{J}$ are replaced by an external field acting on the sequence from the residues of protein $\mathcal{I}$.

We implemented the ''frozen'' approximation using local sequence alignment (Smith & Waterman, 1981) and linear gap penalty function. To get the best results from the ''frozen'' approximation we varied the values of the two gap penalties trying 500 different values for each. An alignment which produced the lowest *dRMS* was selected for further comparison. The results for the ''frozen'' approximation applied to ubiquitin threaded through its three analogs with the MJ96 potential are presented in Table 6.

Table 6. Results of threading with ''frozen'' approximation using the MJ96 potential

Proteins	SeqID (%)	L_{alg}	*dRMS* C^{β}	Q_{alg}
Ubi-Ubi	100	76	0.00	1.00
Ubi-Gua	14	76	6.00	0.59
Ubi-Igd	4	61	7.11	0.04
Ubi-Frr	6	76	6.61	0.13
Ten-Ten	100	89	0.00	1.00
Ten-Fnf	22	89	1.79	0.90
Ten-HhrB	19	80	7.23	0.11
Ten-Hnf	9	89	18.21	0.08
Ten-Cid	11	89	8.87	0.00
Ten-Tit	9	89	6.06	0.07
Ash-Ash	100	147	0.00	1.00
Ash-Fal	12	146	4.87	0.01
Ash-Hbg	14	147	6.49	0.17
Ash-ColA	6	147	8.49	0.00

Gap penalties were chosen to minimize the *dRMS*. As expected, *dRMS* values are much higher for all proteins with no distinct sequence homology (compare *dRMS* with Tables 3 to 5). The only pair that yields an accurate alignment (Ten-Fnf) has a distinct sequence homology.

Clearly, the alignments obtained with the ''frozen'' approximation are very far from the structural ones. The only example where alignment was close to the optimal is tenascin threaded through the third domain of fibronectin. This result comes as no surprise, since the two proteins have extremely close structures (RMS < 1.0 Å) and some sequence and evolutionary homology (*SeqID* = 20%). For all other cases the alignments obtained by MC threading with the same potential were much more accurate than alignments obtained using the ''frozen'' approximation.

Discussion

Here we presented a systematic approach to the problem of protein structure prediction by threading. As in folding, the problem of threading has two components: (i) a search strategy, which is able to find the optimal alignment of sequence and a structure; and (ii) a potential, which provides the lowest energy to the native and similar structures of a protein. First, we developed a Monte Carlo algorithm to search through the space of alignments for the optimal one. To test the algorithm we applied it to structure-structure alignments using a *Dali* potential. For each case studied our algorithm successfully found the optimal alignments which another heuristic algorithm (Holm & Sander, 1993) failed to find. We also showed that for proteins with similar structures, the optimal alignment is not substantially changed when the distance-dependent potential is replaced by a much simpler contact potential. Then we considered structure-structure alignments as sequence-structure threading with an hypothesized ''ideal'' potential, which makes all native and only native contacts attractive (so-called ''Go-model'' in protein folding).

Although no residue-residue potential can provide this ''ideal'' pattern of interaction, one can build a potential which provides an interaction pattern as close to the ''ideal'' one as possible (Mirny & Shakhnovich, 1996). Using an optimization technique (Mirny & Shakhnovich, 1996) we built this potential for a single protein, ubiquitin, and then threaded its sequence through different protein structures. Comparison of the optimal alignments obtained by threading and corresponding structural alignments brought us to the following conclusions. (i) Although the used residue-residue potential reconstitutes an ''ideal'' pattern of interactions in the best possible way, optimal threading alignments are different from optimal structural alignments. (ii) The degree to which an optimal alignment changes depends strongly on the degree of similarity between the native structure and the template structure used for threading. Particularly, threading through the native structure

and a highly similar one yields an optimal alignment very similar to the structural one. In contrast, for proteins sharing a moderate similarity with the native one, the optimal threading alignment is very different from the structural one and provides a poor model of the native structure (RMS $C^{\alpha} > 4$ Å). These results and detailed analysis of the threading landscape brought us to the major conclusion that in order to get an accurate threading alignment one needs, most importantly, a very similar template structure. If this is not present, even an incredibly accurate residue-residue potential is useless.

Threading *versus* folding

In folding, the problem is to find the native conformation, whereas threading aims at finding a conformation similar to the native one. The fundamental difference is that in threading the native conformation is not present in the space of possible conformations and only an approximate (''native-like'') conformation can be found. The native conformation of a protein has an energy much lower than the energy of other (random) conformations, where a ''native-like'' conformations is less stable and not that much below other conformations in energy. This makes threading a much more challenging problem as one needs to find a conformation that is not as distinct in energy from others as the native. When one threads a sequence through its native structure, the optimal alignment is separated from dissimilar decoys by a large energy gap, whereas when one uses a structure similar to the native as a template for threading, the gap is much smaller. The further the template structure is from the native one, the smaller the energy gap between the optimal alignment and the random ones. Figure 8 presents a schematic of energy spectra for folding and threading.

From the physical point of view, threading using a template close to the native structure is similar to folding to the structure which has a moderate energy gap. This situation was studied extensively with model proteins (Mirny *et al*., 1996; Abkevich *et al*., 1996), where different energy gaps between the native conformation and the bulk of unfolded ones are responsible for different regimes of folding. When the energy gap is large (folding of designed, evolved sequences), the native conformation is stable relative to changes in temperature and to mutations (Pande *et al*., 1995; Vendruscolo *et al*., 1997), and folding is fast and exhibits a cooperative, first-order-like transition. When the energy gap is small (poorly designed, random sequences; Mirny *et al*., 1996; Abkevich *et al*., 1996) the native conformation is much less stable with respect to temperature and, importantly, is very unstable to mutations (Shakhnovich & Gutin, 1991) and to errors in the potential (Vendruscolo *et al*., 1997; Pande *et al*., 1995; Bryngelson, 1994). Folding in this case is slow because thermodynamic stability requires a low temperature at which a protein gets trapped in several misfolded low-energy conformations. Stability with respect to changes in potential and mutations have the same nature: both raise the energy of the ground state, decreasing the gap. If the gap is large enough it sustains destabilization of the native state and it remains the ground state. If the gap is small, the destabilizing decrease in the gap can eliminate the gap completely and make another (random!) conformation the state with the lowest energy (see the cartoon in Figure 8).

In this study we observed similar regimes for threading. When the template is close to the native structure (self-threading, Ubi-Igd) the optimal alignment is separated by a large energy gap and we observe: (i) cooperative transition; (ii) focusing energy landscape; (iii) high reliability of alignment (low S_i, similar suboptimal alignments); and, most importantly (iv) stability of the optimal alignment to the changes in potential. Alternatively, when the template is moderately close to the native structure (Ubi-Gua, Ubi-Frr) the gap is small and we

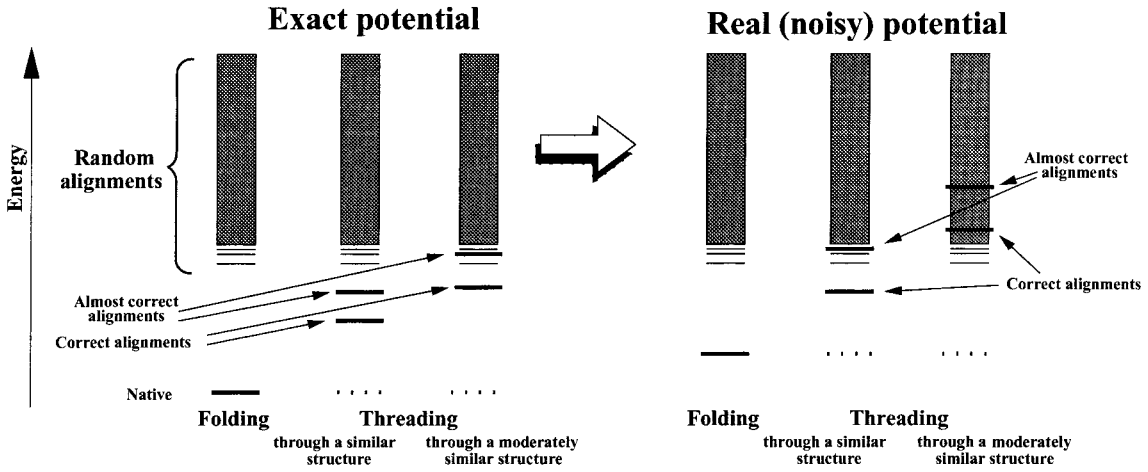

Figure 8. Schematic cartoon of the density of states for folding and threading. See Discussion.

observe: (i) smooth transition; (ii) poorly focusing energy landscape; (iii) low reliability of alignment (high S_i, dissimilar suboptimal alignments); and, most importantly (iv) low stability of the optimal alignment to the changes in potential.

Factors affecting the accuracy of alignment

These results suggest that there are two major factors affecting the accuracy of threading alignment: (i) the degree of similarity between the template structure and the native one; (ii) the accuracy of the potential. Both factors are responsible for providing a large energy gap between the right alignment and the bulk of decoys. The gap is essential to guarantee the lowest energy and high reliability for the right alignment.

The crucial role of the degree of similarity for the success of threading became evident after the recent experiment in ''blind'' structure prediction, CASP2 (Moult *et al*., 1997; Levitt, 1997; Marchler-Bauer & Bryant, 1997; Eisenberg, 1997; Shortle, 1997; Dunbrack *et al*., 1997). Several predictors Jones, 1997; Marchler-Bauer & Bryant, 1997) noted that success in threading could be attributed more to target proteins rather than methods. Only those proteins that have a globally similar (more than 60% of residues superimposable) structure in the database were modeled accurately (Marchler-Bauer & Bryant, 1997; Levitt, 1997). The degree of similarity between the two structures (*dRMS*, fraction of matched residues, fraction of common contacts, etc.) required to provide high accuracy of threading using currently available potentials is still to be estimated.

Earlier we showed (Mirny *et al*., 1996) than an ''individual'' potential optimized for a single protein is not transferable to others. On the other hand, a ''universal'' potential optimized simultaneously for all proteins in the database works worse for every single individual protein than the ''individual'' potential. Hence, alignments obtained with any realistic ''universal'' potential should be much worse than those obtained with the ''individual'' one used in our threading tests. This is indeed the case as seen in Tables 3 and 5.

While our analysis is carried out only on a limited set of examples (three folds) the consistency of the results between them makes it possible to make a preliminary quantitative conclusion of what degree of success in sequence-structure alignment can one expect from threading calculations.

When the RMS between the template and the native structure is less than 2 Å the sequence-structure alignment is likely to be accurate with any available and reasonably good potential function.

When the RMS between the template and the native structure is greater than 3.5 Å, the sequence-structure alignment is expected to be grossly inaccurate, independent of the potential function used.

The RMS between 2 Å and 3.5 Å represents a ''twilight zone'' where the results may depend strongly on the potential used as well as particular proteins.

The analysis of more examples using the MC threading procedure will allow us to make the quantitative estimates of expected threading accuracy more precise.

We are planning to carry our this research in the near future.

Fold recognition *versus* alignment recognition

How does the accuracy of threading alignment affect the accuracy of fold recognition? Why are current threading algorithms able to find the right template structure, yet fail to produce accurate alignments (Marchler-Bauer & Bryant, 1997; Levitt, 1997; Jones, 1997)?

The major difference between fold recognition and sequence-structure alignments is in the number of alternatives to choose from. Fold recognition is aimed at finding a structure in a representative fold database which contains about 1000 folds (Holm & Sander, 19970. In contrast, the threading algorithm applied to two proteins of length 100 needs to find the optimal in the space of $2^{100} \approx 10^{30}$ alignments (Waterman, 1995). Clearly, fold recognition is much less demanding of the accuracy of potential and the power of a search strategy. Fold recognition can tolerate errors in threading alignment. In fact, to distinguish a structure close to the native one from a distant structure one needs to find any low energy alignment of the close structure. Only a tiny fraction of alignments has the energy low enough to distinguish the close structure, but because of the huge alignment space this fraction numbers in the millions of alignments and any of them is equally good for the purpose of fold recognition. In contrast, here we show that alignment recognition requires much more accurate potential and a more powerful search strategy.

To be successful in fold recognition, one need not send the arrow right to the apple; one just needs to shoot in the right direction.

Future directions

One of the possible ways to overcome this fundamental problem in threading is to use some extra information about the query sequence. Evolutionary information can be of great importance for structure prediction and has already been taken into account (explicitly or implicitly) in different ways (Fischer & Eisenberg, 1996; Gerloff *et al*., 1997; Russell *et al*., 1996; Defay & Cohen, 1996; Goldman *et al*., 1996; Ortiz *et al*., 1998). A popular way of using evolutionary information is to predict a secondary structure or the degree of solvent accessibility from multiple sequence alignments (Rost & Sander, 1995) and then use this prediction to refine threading. Use of intrinsically inaccurate predictions, however, should not necessarily improve folding or threading, (Rost *et al*., 1997) but instead can derail them. Rational utilization of

evolutionary information clearly requires deeper structural understanding of the evolutionary traces observed in homologous proteins (Shakhnovich *et al*., 1996b; Mirny *et al*., 1998).

Another substantial improvement may come from building more complicated (but still computationally tractable!) models of proteins. Taking into account side-chains, their sizes and shapes, can eliminate from the search some threading alignments that produce unfeasible side-chain packing and thus focus the landscape to the native state.

Methods

Alignment representation

An alignment between two proteins of length I and J is represented by a matrix A_{ij}, where $i = 1,\ldots,I$ and $j = 1,\ldots,J$:

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ is aligned with } j \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Another way of presenting an alignment is by a pointer p_i:

$$p_i = \begin{cases} j & \text{if } i \text{ is aligned with } j \\ 0 & \text{if } i \text{ is not aligned to any residue} \end{cases} \quad (11)$$

In this study we do not allow double matches (i.e. $\Sigma_{i=1,\ldots,I} A_{ij} \leqslant 1$). The reverse of any fragment in the alignment is also forbidden, i.e. if $A_{ij} = 1$, then for any $i' > i$ and $j' < j$ $A_{i'j'} = 0$. (In general, reverse of protein fragments in sequence structure alignment might make sense and improve performance (Finkelstein, 1997).) Under these constraints, matrix A_{ij} should have the form shown in Figure 9A, i.e. an alignment composed of runs of aligned residues separated by gaps in either or in both proteins. These runs are referred to below as fragments of alignment. Each fragment is a set of matches

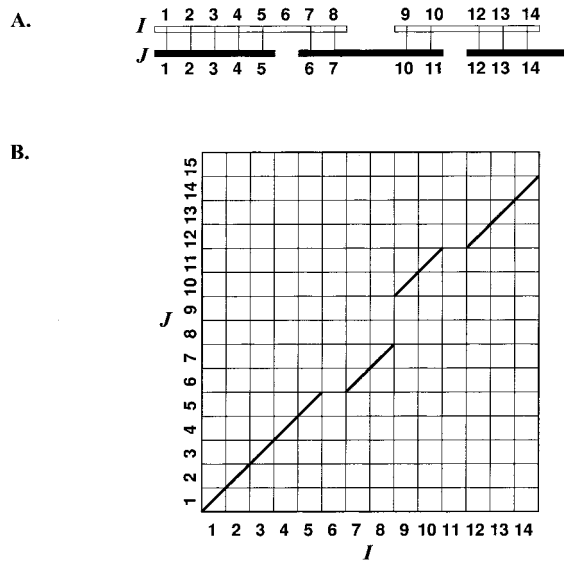

Figure 9. Representation of a protein-protein alignments. A, Example of the alignment; B, its matrix representation.

$(A_{ij} = 1 \text{ for } (ij) = (i', j'),\ (i'+1, j'+1), \ldots, (i'+L, j'+L))$ framed into gaps $A_{i'-1,j'-1} = A_{i+L+1,j+L+1} = 0$. These fragments are used in the move set as elementary building blocks.

Move set

Each move in the move set is designed to change the alignment preserving most of the matches and, hence, leading to a small change in energy. Another important feature of the proposed move set is that it allows easy introduction of constraints on the minimum length of a fragment or maximum length of a gap. This flexibility is achieved by making moves on fragments, rather than creating and destroying single matches. Here we constrain the fragment length to be greater or equal to $L_{\min}$ = six residues. Moves used in the current implementation are shown in Figure 10.

Shift

A whole fragment is shifted in any of four directions. The distance fragment shifted is chosen randomly and uniformly between 1 and M, where M is the space available to the next fragment in the direction of the shift.

Shrink/expand

A fragment is shrunk (expanded) on either end on n residues. The value of n is chosen randomly from the exponential distribution, i.e. $P(n) = a + b \times \exp(-n/5)$. Parameters a and b set the limits on n. Limits are chosen to provide the length of the obtained fragment $L \geqslant L_{\min}$ and not overlap with another fragment.

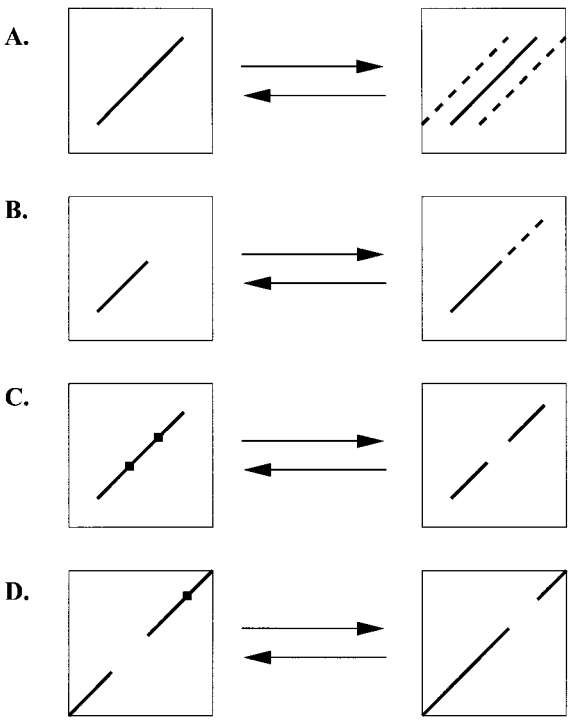

Figure 10. Move set. A, Shift; B, shrink/expand; C, split/merge; D, jump.

Split/merge

Fragments which happen to be located head-to-tail of each other are merged into a single fragment. Fragments are merged every ten steps (starting from step 5), a randomly chosen fragment is split into two fragments at a random point, providing that lengths of both fragments $L_1, L_2 \geqslant L_{\min}$.

These moves are sufficient for successful MC alignment. However, we use one extra move which makes sampling more efficient.

Jump

A piece of one fragment is deleted and joined to the next fragment. The length of the piece is chosen randomly and uniformly so that the length of the remaining part of the fragment $L \geqslant L_{\min}$.

Although we do not have an explicit move for the creation (destruction) of a fragment, new fragments can appear (disappear). In fact, a new fragment can appear by expansion of an existing one and further splitting. Similarly, a fragment can disappear by merging with an existing one followed by shrinking. At least one fragment is present in the alignment. These moves can efficiently sample all possible alignments between two proteins. Note that in contrast to other work (Lathrop & Smith, 1996) the lengths and number (Bryant, 1996) of fragments can vary.

This move set provides an acceptance ratio for Monte Carlo steps of about 0.6 at transition temperature T_f. A typical simulated annealing protocol requires 20-50 temperature levels with 10^6 steps on each. It takes about four minutes of CPU time on a Pentium 233 for each 10^6 MC steps for proteins of approximately 100 residues.

Measures of similarity

Distance between protein conformations

We use three different measure of distance between protein structures. Consider proteins $\mathcal{I}$ and $\mathcal{J}$ which have length I and J residues, respectively, and alignment between them given by p_i (see above). Then:

$$dRMS = \sqrt{\frac{1}{(N_{\text{alg}}-1)(N_{\text{alg}}-2)} \sum_{i,i'>i+2}^{I} (r^{\mathcal{J}}_{p_i,p_{i'}} - r^{\mathcal{I}}_{ii'})^2} \qquad (12)$$

where $r^{\mathcal{I}}_{ii'}$ and $r^{\mathcal{J}}_{jj'}$ are distances between residues in each protein. In this study we use distances between C^β atoms as distances between corresponding residues.

Another measure based on distance matrices $r_{ij'}$ was introduced by Holm & Sander, 1993) and is used in *Dali* structure comparison program:

$$Dali = \sum_{i,i'>i+2} \left(0.2 - \frac{r^{\mathcal{I}}_{ii'} - r^{\mathcal{J}}_{r_i,p_{i'}}}{d}\right) \exp(-d/20) \qquad (13)$$

where:

$$d = \frac{r^{\mathcal{I}}_{ii'} + r^{\mathcal{J}}_{p_i,p_{i'}}}{2}$$

The *Dali* measure is much more useful for structure comparison as it ''weights'' matches between closely located residues much more than between distant ones (see Model for details).

A much simpler measure of protein similarity is the overlap of contact maps to the two proteins:

$$Q = \frac{1}{\min(N_{\mathcal{I}}, N_{\mathcal{J}})} \sum_{i,i'>i+2}^{I} \Delta^{\mathcal{J}}_{p_i,p_{i'}} \times \Delta^{\mathcal{I}}_{ii'} \qquad (14)$$

where:

$$\Delta_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are in contact: } r_{ii'} < R_{\text{cut}} \\ 0 & \text{otherwise} \end{cases} \qquad (15)$$

N_χ is the number of contacts in protein χ. In this study we use contact cutoff $R_{\text{cut}} = 8$ Å between C^β atoms. Contact overlap Q, in contrast to *dRMS* and *Dali*, is focused on contacts between residues and, hence, it completely ignores residues that are far from each other and might not contribute to the energy of the conformation.

Distance between alignments

To measure distance between alignments we introduce overlap Q_{alg} between alignment p_i and q_i:

$$Q_{\text{alg}} = \frac{1}{N_{\text{alg}}} \sum_{i=1}^{I} \text{sign}(p_i \times q_i) \times |p_i - q_i| \qquad (16)$$

where $\text{sign}(0) = 0$ and $\text{sign}(x > 0) = 1$, ($q_i, p_i \geqslant 0$ for all i).

Random alignments

To generate a random alignment we make 10^4 steps of the MC procedure at $T = \infty$, i.e. accepting all moves. Alignments randomized in this way are used as starting points for optimization runs.

Optimization of potential

The aim of this procedure is to find a potential $U(\xi, \eta)$, $\xi, \eta = 1, \ldots, 20$ that provides the pattern of interactions between residues $B_{jj'} = U(a_j, a_{j'})$ as close to an ''ideal'' pattern $B^{\text{ideal}}_{jj'} = -\Delta_{jj'}$ as possible. This is achieved by maximizing the correlation between $B_{jj'}$ and $-\Delta_{jj'}$:

$$\rho(U) = \rho(B(U), -\Delta) = -\frac{\langle B\Delta\rangle - \langle B\rangle\langle\Delta\rangle}{\sqrt{\sigma^2(B)\sigma^2(\Delta)}} \qquad (17)$$

as a function of potential U. Averages $\langle\cdot\rangle$ are taken over all $j < j'$ elements. Optimization is performed by a Monte Carlo procedure, where at each step a randomly chosen element $U(\xi,\eta)$ is increased (decreased) by $\delta = 0.01$ and the increase is accepted or rejected according to the Metropolis scheme (Metropolis *et al.*, 1953). Since equation (17) does not change upon linear transformation of U, we set $\langle U\rangle = 0$ and $\sigma(U) = 1$ (see Mirny & Shakhnovich (1996) for details).

Importantly, $\rho(U)$ defined above is a linear function of the Z-score, which measures the energy gap between the native conformation and the bulk of alternative conformations:

$$Z = \frac{E_N - \langle E\rangle_{\text{conf}}}{\sigma(E)_{\text{conf}}} \qquad (18)$$

where E_N is the energy of the native conformation; $\langle E\rangle_{\text{conf}}$ and $\sigma(E)_{\text{conf}}$ are the mean and the variance of energy of alternative conformations. When $\langle E\rangle_{\text{conf}}$ and $\sigma_{\text{conf}}(E)$ are computed as described by Mirny & Shakhnovich (1996):

$$\rho = \frac{Z}{\sqrt{n_{\text{total}} - n}} \tag{19}$$

where n is the number of native contacts in $\mathcal{J}$ and n_{total} is the total number of possible contacts $\mathcal{J}(\mathcal{J}-2)/2$. Then, by maximizing ρ we maximize Z and make the native conformation of the proteins a pronounced energy minimum (Mirny & Shakhnovich, 1996).

Acknowledgments

We are grateful to Victor Abkevich and Cecilia Clementi for fruitful discussions. This work was supported by NIH grant GM52126.

References

Abkevich, V., Gutin, A. & Shakhnovich, E. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.

Abkevich, V., Gutin, A. & Shakhnovich, E. (1996). How the first biopolymers could have evolved. *Proc. Natl Acad. Sci, USA*, **93**, 839–844.

Allen, M. & Tildesley, D. (1987). *Computer Simulation of Liquids*, Oxford University Press, New York.

Bahar, I. & Jernigan, R. (1996). Coordination geometry of nonbonded residues in globular proteins. *Fold. Des.* **1**, 357–370.

Berriz, G., Gutin, A. & Shakhnovich, E. (1997). Cooperativity and stability in a langevin model of protein-like folding. *J. Chem. Phys.* **106**, 9276–9285.

Binder, K. (1986). *Monte Carlo Methods in Statistical Physics*, Springer-Verlag, Berlin and New York.

Binder, K. (1995). *The Monte Carlo Method in Condensed Matter Physics*, Springer-Verlag, Berlin and New York.

Bryant, S. (1996). Evaluation of threading specificity and accuracy. *Proteins: Struct. Funct. Genet.* **26**, 172–185.

Bryngelson, J. (1994). When is a potential accurate enough for structure prediction–theory and application to a random heteropolymer model of protein-folding. *J. Chem. Phys.* **100**, 6038–6045.

Defay, T. & Cohen, F. (1996). Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**, 314–323.

Dunbrack, R., Gerloff, D., Bower, M., Chen, X., Lichtarge, O. & Cohen, F. (1997). Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (casp2), asilomar. *Fold. Des.* **2**, R27–R42.

Eisenberg, D. (1997). Into the black of night. *Nature Struct. Biol.* **4**, 95–97.

Elofsson, A., Fischer, D., Rice, D., Le, G. & Eisenberg, D. (1996). A study of combined structure/sequence profiles. *Fold. Des.* **1**, 451–461.

Finkelstein, A. (1997). Protein structure: what is it possible to predict now? *Curr. Opin. Struct. Biol.* **7**, 60–71.

Finkelstein, A. & Reva, B. (1991). Search for the most stable folds of protein chains. *Nature*, **351**, 497–499.

Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Proteins Sci.* **5**, 947–955.

Flockner, H., Domingues, F. & Sippl, M. (1997). Protein folds from pair interactions: a blind test in fold recognition. *Proteins: Struct. Funct. Genet.* **1**, 129–133 (Suppl.).

Gerloff, D., Cohen, F., Korostensky, C., Turcotte, M., Gonnet, G. & Benner, S. (1997). A predicted consensus structure for the n-terminal fragment of the heat shock protein hsp90 family. *Proteins: Struct. Funct. Genet.* **27**, 450–458.

Go, N. & Abe, H. (1981). Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers*, **20**, 991–1011.

Goldman, N., Thorne, J. & Jones, D. (1996). Using evolutionary trees in proteins secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**, 196–208.

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

Gutin, A., Abkevich, V. & Shakhnovich, E. (1996). Chain length scaling of proteins folding time. *Phys. Rev. Letters*, **77**, 5433–5436.

Hao, M.-H. & Scheraga, H. (1996). How optimization of potential function affects protein folding. *Proc. Natl Acad, Sci. USA*, **93**, 4984–4989.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Holm, L. & Sander, C. (1997). Dali/fssp classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231–234.

Jones, D. (1997). Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7**, 377–387.

Kirkpatrick, S. (1984). Optimization by simulated annealing: quantitative studies. *J. Stat. Phys.* **34**, 975–986.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. ii. Application to protein. *Proteins: Struct. Funct. Genet.* **18**, 353–366.

Koretke, K., Luthey-Schulten, Z. & Wolynes, P. (1996). Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* **5**, 1043–1059.

Lacroix, E., Bruix, M., Lopez-Hernandez, E., Serrano, L. & Rico, M. (1997). Amide hydrogen exchange and internal dynamics in the chemotactic protein chey from *Escherichia coli*. *J. Mol. Biol.* **271**, 472–487.

Lathrop, R. (1994). The protein threading with sequence amino acid interaction preferences is np-complete. *Protein Eng.* **7**, 1059–1068.

Lathrop, R. & Smith, T. (1996). Global optimum proteins threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* **255**, 641–665.

Levitt, M. (1997). Competitive assessment of protein fold recognition and alignment accuracy. *Proteins: Struct. Funct. Genet.* **1**, 92–105 (Suppl.).

Marchler-Bauer, A. & Bryant, S. (1997). A measure of success in fold recognition. *Trends Biochem. Sci.* **22**, 236–240.

Metropolis, N., Rosenbluth, A., Rosenbluth, R., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Mirny, L. & Shakhnovich, E. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264**, 1164–1179.

Mirny, L., Abkevich, V. & Shakhnovich, E. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103–116.

Mirny, L., Abkevich, V. & Shakhnovich, E. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*. In the press.

Miyazawa, S. & Jernigan, R. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.

Moult, J., Hubbard, T., Bryant, S., Fidelis, K. & Pedersen, J. (1997). Critical assessment of methods of protein structure prediction(casp): Round ii. *Proteins: Struct. Funct. Genet.* **1**, 2–7(Suppl.).

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.

Ortiz, A., Kolinski, A. & Skolnick, J. (1998). Nativelike topology assembly of small proteins using predicted restraints in monte carlo folding simulations. *Proc. Natl Acad. Sci. USA*, **95**, 1020–1025.

Pande, V., Grosberg, A. & Tanaka, T. (1995). How accurate must potential be for successful modeling of protein-folding. *J. Chem. Phys.* **103**, 9482–9491.

Pande, V., Grosberg, A. & Tanaka, T. (1997). On the theory of folding kinetics for short proteins. *Fold. Des.* **2**, 109–114.

Rost, B. & Sander, C. (1995). Progress of 1d proteins structure prediction at last. *Proteins: Struct. Funct. Genet.* **23**, 295–300.

Rost, B., Schneider, R. & Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480.

Russell, R., Copley, R. & Barton, G. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365.

Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature*, **369**, 248–251.

Shakhnovich, E. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Letters*, **72**, 3907–3910.

Shakhnovich, E. (1997a). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.

Shakhnovich, E. (1997b). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.

Shakhnovich, E. & Gutin, A. (1991). Influence of point mutations on protein structure: probability of a neutral mutation. *J. Theoret. Biol.* **149**, 537–546.

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996a). Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–98.

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996a). Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–98.

Shortle, D. (1997). Structure prediction: folding proteins by pattern recognition. *Curr. Biol.* **7**, R151–R154.

Sippl, M. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.

Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Taketomi, H., Udea, Y. & Go, N. (1975). *Int. J. Pept. Protein Res.* 445–449.

Vendruscolo, M., Maritan, A. & Banavar, J. (1997). Stability threshold as a selection principle for protein design. *Phys. Rev. Letters*, **78**, 3967–3970.

Waterman, M. (1995). *Introduction to Computational Biology*, Chapman & Hall, New York.

Edited by F. Cohen

(Received 13 February 1998; received in revised form 15 July 1998; accepted 16 July 1998)