Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.

ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

PROTEIN RECOGNITION BY SEQUENCE-TO-STRUCTURE FITNESS: BRIDGING EFFICIENCY AND CAPACITY OF THREADING MODELS

JAROSLAW MELLER

Department of Computer Science, Cornell University, Ithaca, NY, U.S.A.; and Department of Computer Methods, Nicholas Copernicus University, Torun, Poland

RON ELBER

Department of Computer Science, Cornell University, Ithaca, NY, U.S.A.

CONTENTS

I. INTRODUCTION

The threading approach [1–8] to protein recognition is a generalization of the sequence-to-sequence alignment. Rather than matching the unknown sequence S_i to another sequence S_i (one-dimensional matching), we match the sequence S_i to a shape $\mathbf{X}_j$ (three-dimensional matching). Experiments found a limited set of folds compared to a large diversity of sequences. A shape has (in principle) more detectable ''family members'' compared to a sequence, suggesting the use of structures to find remote similarities between proteins. Hence, the determination of overall folds is reduced to tests of sequence fitness into known and limited number of shapes.

The sequence–structure compatibility is commonly evaluated using reduced representations of protein structures. Assuming that each amino acid residue is represented by a point in three-dimensional space, one may define an effective energy of a protein as a sum of inter-residue interactions. The effective pair energies can be derived from the analysis of contacts in known structures. Knowledge-based pairwise potentials proved to be very successful in fold recognition [2,3,6,9–11], *ab initio* folding [11–13], and sequence design [14–15].

Alternatively, one may define the so-called ''profile'' energy [1,5] taking the form of a sum of individual site contributions, depending on the structural environment (e.g., the solvation/burial state or the secondary structure) of a site. The above distinction is motivated by computational difficulties of finding optimal alignments with gaps when employing pairwise models.

Consider the alignment of a sequence $S = a_1 a_2 \ldots a_n$ of length, n, where a_i is one of the 20 amino acids, into a structure $\mathbf{X} = (x_1, x_2, \ldots, x_m)$ with m sites, where x_j is an approximate spatial location of an amino acid (taken here to be the geometric center of the side chain). We wish to place each of the amino acids in a corresponding structural site $\{a_i \rightarrow x_j\}$. No permutations are allowed. In order to identify homologous proteins of different length, we need to consider deletions and insertions into the aligned sequence. For that purpose we introduce an ''extended'' sequence, $\bar{S}$ which may include gap ''residues'' (spaces, or empty structural sites) and deletions (removal of an amino acid, or an amino acid corresponding to a virtual structural site).

Our goal is to identify the matching structure $\mathbf{X}_j$ with the extended sequence $\bar{S}_i$. The process of aligning a sequence S into a structure $\mathbf{X}$ provides an optimal

score and the extended sequence $\bar{S}$. This double achievement can be obtained using dynamic programming (DP) algorithm [16–19]. In DP the computational effort to find the optimal alignment (with gaps and deletions) is proportional to $n \times m$, as compared to exponential number ($\approx 2^{n+m}$) of all possible alignments.

In contrast to profile models, the potentials based on pair interactions do not lead to optimal alignments with dynamic programming. A number of heuristic algorithms that provide approximate alignments have been proposed [20]. These algorithms cannot guarantee an optimal solution with less than exponential number of operations [21]. Another common approach is to approximate the energy by a profile model (the so-called frozen environment approximation) and to perform the alignment using DP [22]. In this work, we are aiming at deriving systematic approximations to pair energies that would preserve the computational simplicity of profile models.

Threading protocols that are based exclusively on pairwise models were shown to be too sensitive to variations in shapes [23]. Therefore, pairwise potentials are often employed in conjunction with various complementary ''signals,'' such as sequence similarity, secondary structures, or family profiles [9–11,24–28]. Such additional signals enhance the recognition when the tertiary contacts are significantly altered. In GenTHREADER [9], for example, sequence alignment methods are employed as the primary detection tools. A pairwise threading potential is then used to evaluate the consistency of the sequence alignments with the underlying structures. Bryant and co-workers use, in turn, an energy function which is a weighted sum of a pairwise threading potential and a sequence substitution matrix [10].

Distant-dependent pair energies are expected to be less sensitive to variations in shapes than simple contact models, in which inter-residues interactions are assumed to be constant up to a certain cutoff distance and are set to zero at larger distances. A number of distance-dependent pairwise potentials have been proposed in the past [29,30]. We consider both simple contact models and distance-dependent power law potentials and compare their performance with that of novel profile models.

We compute the energy parameters by linear programming (LP) [31–33]. There are a number of alternative approaches to derive the energy parameters. For example, statistical analysis of known protein structures makes it possible to extract ''mean-force'' potentials [34–38]. Another approach is the optimization of a single target function that depends on the vector of parameters such as T_f/T_g [39], the Z score [1], or the σ parameter [40]. We note also that optimization of the gap energies has been attempted in the past [22,41]. The statistical analysis is the least expensive computationally. The optimization approaches have the advantage that misfolded structures can be made part of the optimization, providing a more complete training. The LP approach is

computationally more demanding compared to other protocols. However, it has important advantages, as discussed below.

In LP training we impose a set of linear constraints (for energy models linear in their parameters) of the general form

$$\Delta E_{\text{dec, nat}} \equiv E_{\text{decoy}} - E_{\text{native}} > 0 \tag{1}$$

where E_{native} is the energy of the native alignment (of a sequence into its native structure) and E_{decoy} represents the energies of the alignments into non-native (decoy) structures. In other words, we require that the energies of native alignments be lower than the energies of alignments into misfolded (decoy) structures.

While optimization of the Z, T_f/T_g, and σ scores led to remarkably successful potentials [1,39,40], it focuses at the center of the distribution of the $\Delta E_{\text{dec, nat}}$'s and does not solve exactly the conditions of Eq. (1). For example, the tail of the distribution of the $\Delta E_{\text{dec, nat}}$ may be slightly wrong, and a fraction f of the $\Delta E_{\text{dec, nat}}$'s may "leak" to negative values. If f is small, it may not leave a significant impression on the first and second moments of the distribution; that is, the value of the Z score remains essentially unchanged. "Tail misses" is not a serious problem if we select a native shape from a small set of structures. However, when examining a large number of constraints, even if f is small, the number of inequalities that are not satisfied can be very large, making the selection of the native structure difficult if not impossible.

In contrast to the optimization of average quantities, the LP approach guarantees that all the inequalities in Eq. (1) are satisfied. If the LP cannot find a solution, we get an indication that it is impossible to find a set of parameters that solve all the inequalities in Eq. (1). For example, we may obtain the impossible condition that the contact energy between two ALA residues must be smaller than 5 and at the same time must be larger than 7. Such an infeasible solution is an indicator that the current model is not satisfactory, and more parameters or changes in the functional form are required [31–33]. Hence, the LP approach, which focuses on the tail of the distribution near the native shape, allows us to learn continuously from new constraints and improve further the energy functions, guiding the choice of their functional form.

In the present chapter we evaluate several different scoring functions for sequence-to-structure alignments, with parameters optimized by LP. Based on a novel profile model, designed to mimic pair energies, we propose an efficient threading protocol of accuracy comparable to that of other contact models. The new protocol is complementary to sequence alignments and can be made a part of more complex fold recognition algorithms that use family profiles, secondary structures, and other patterns relevant for protein recognition.

The first half of the chapter is devoted to the design of scoring functions. Two topics are discussed: the choice of the functional form (Section II) and the

choice of the parameters (Section III). The capacity of the energies is explored and optimal parameters are determined (Section IV). High capacity indicates that a large number of protein shapes are recognized with a small number of parameters.

The second part of the manuscript deals with optimal alignments. We design gap energies (Section V) and introduce a double Z-score measure (from global and local alignments) to assess the results (Section VI). Presentation of extensive tests of the algorithm (Section VII) is followed by the conclusions and closing remarks.

II. FUNCTIONAL FORM OF THE ENERGY

In a nutshell there are two ''families'' of energy functions that are used in threading computations, namely the pairwise models (with ''identifiable'' pair interactions) and the profile models. In this section we formally define both families and we also introduce a novel THreading Onion Model (THOM), which is investigated in the subsequent sections of the chapter.

A. Pairwise Models

The first family of energy functions is of pairwise interactions. The score of the alignment of a sequence S into a structure **X** is a sum of all pairs of interacting amino acids,

$$E_{\text{pairs}} = \sum_{i<j} \phi_{ij}(\alpha_i, \beta_j, r_{ij}) \tag{2}$$

The pair interaction model, ϕ_{ij}, depends on the distance between sites i and j and also depends on the types of amino acids, α_i and β_j. The latter are defined by the alignment, because certain amino acid residues $a_k, a_l \in S$ are placed in sites i and j, respectively.

We consider two types of pairwise interaction energies. The first is the widely used contact potential. If the geometric centers of the side chains are closer than 6.4 Å, then the two amino acids are considered in contact. The total energy is a sum of the individual contact energies:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \left\{ \begin{array}{ll} \varepsilon_{\alpha\beta}, & 1.0 < r_{ij} < 6.4\,\text{Å} \\ 0, & \text{otherwise} \end{array} \right\} \tag{3}$$

where i, j are the structure site indices (contacts due to sites in sequential vicinity are excluded, $i = 3 < j$), α, β are indices of the amino acid types (we drop the subscripts i and j for convenience), and $\varepsilon_{\alpha\beta}$ is a matrix of all the possible contact types. For example, it can be a 20×20 matrix for the twenty amino acids.

TABLE I
The Definitions of Different Groups of Amino Acids That Are Used in the Present Study[a]

Hydrophobic (HYD)	ALA CYS HIS ILE LEU MET PHE PRO TRP TYR VAL
Polar (POL)	ARG ASN ASP GLN GLY LYS SER THR
Charged (CHG)	ARG ASP GLU LYS
Negatively charged (CHN)	ASP GLU

[a]Note that 10 types of amino acids are found to be sufficient to solve the Hinds–Levitt set either by pairwise interaction models or by THOM2 (in the case of continuous LJ models, HIS was replaced by CYS). The amino acid types are HYD, POL, CHG, CHN, GLY, ALA, PRO, TYR, TRP, and HIS. The list implies that when an amino acid appears explicitly, it is excluded from other groups that may contain it. For example, HYD includes in this case CYS, ILE, LEU, MET, and VAL, while CHG includes ARG and LYS only, since the negatively charged residues form a separate group.

Alternatively, it can be a smaller matrix if the amino acids are grouped together to fewer classes. Different groups that are used in the present study are summarized in Table I. The entries of $\varepsilon_{\alpha\beta}$ are the target of parameter optimization.

The advantage of the single-step potential is its simplicity. This is also its weakness. From a chemical physics perspective the interaction model is oversimplified and does not include the (expected) distance-dependent interaction between pairs of amino acids. To investigate a potential with more "realistic" shape we also consider a "distance power" potential:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \frac{A_{\alpha\beta}}{r_{ij}^m} + \frac{B_{\alpha\beta}}{r_{ij}^n} \tag{4}$$

Here two matrices of parameters are determined: one for the m power, $A_{\alpha\beta}$, and one for the n power, $B_{\alpha\beta}$ $(m > n)$. The signs of the matrix elements are determined by the optimization. In "physical" potentials like the Lennard-Jones model we expect $A_{\alpha\beta}$ to be positive (repulsive) and $B_{\alpha\beta}$ to be negative (attractive). The indices m and n cannot be determined by LP techniques and have to be decided on in advance. A suggestive choice is the widely used Lennard-Jones [LJ(12,6)] model $(m = 12,\ n = 6)$. In contrast to the square well, the LJ(12,6) form does not require a prespecification of the arbitrary cutoff distance, which is determined by the optimization. It also presents a continuous and differentiable function that is more realistic than the square well model.

We show in Section IV that the LJ(12,6), commonly employed in atomistic simulations, performs poorly when applied to inter-residue interactions. Therefore other continuous potentials of the type described in Eq. (5) were investigated. We propose a shifted LJ potential (SLJ) that has significantly higher capacity compared to LJ and is closer in performance to that of the square well potential.

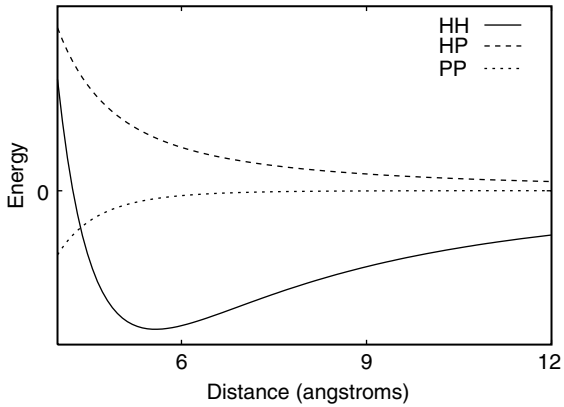


Figure 1. A sample plot of the Lennard-Jones-like potential that we developed. The functional form is $A_{\alpha\beta}/r_{ij}^6 + B_{\alpha\beta}/r_{ij}^2$ (LJ(6,2)), where the indices α and β denote the amino acid types and the indices i and j are the positions along the chain. $A_{\alpha\beta}$ and $B_{\alpha\beta}$ are optimized using the LP approach. The plot includes interactions of the types HH, HP, and PP, where H stands for hydrophobic and P stands for polar residues, respectively. The coefficients A and B are given in Table 7a. Note that the usual Lennard-Jones potential (LJ(12,6)) has a poor recognition capacity.

The SLJ is based on the replacement of $A_{\alpha\beta}/r_{ij}^{12}$ by $A_{\alpha\beta}/(r_{ij}+a)^{12}$, where a is a constant that we set to 1 Å.

The SLJ is a smoother potential with a broader minimum. An alternative potential that also creates a smoother and wider minimum is obtained by changing the distance powers. We also optimized a potential with the (unusual) $(m = 6,\ n = 2)$ pair. This choice was proven most effective and with the largest capacity of all the continuous potentials that we tried (Fig. 1).

B. Profile Models

The second type of energy function assigns "environment" or a profile to each of the structural sites [1]. The total energy $E_{profile}$ is written as a sum of the energies of the sites:

$$E_{profile} = \sum_i \phi_i(\alpha_i,\ \mathbf{X}) \tag{5}$$

As previously, α_i denotes the type of an amino acid a_k of S that was placed at site i of $\mathbf{X}$. For example, if a_k is a hydrophobic residue and x_i is characterized as a hydrophobic site, the energy $\phi_i(\alpha_i, \mathbf{X})$ will be low (score will be high). If a_k is charged, then the energy will be high (low score). The total score is given by a sum of the individual site contributions.

We consider two profile models. The first, which is very simple, was used in the past as an effective solvation potential [1,2,42]. We call it THOM1 (THreading Onion Model 1), and it suggests a clear path to an extension (which is our prime model), namely, THOM2. The "onion" level denotes the number of contact shells used to describe the environment of the amino acid. The THOM1 model uses one "contact" shell of amino acids. The more detailed THOM2 energy model (to be discussed below) is based on two layers of contacts.

In the "profile" potential THOM1, the total energy of the protein is a direct sum of the contributions from m structural sites and can be written as

$$E_{\mathrm{THOM1}} = \sum_i \varepsilon_{\alpha_i}(n_i) \tag{6}$$

The energy of a site depends on two indices: (a) the number of neighbors to the site, n_i [a neighbor is defined as for pairwise interaction—Eq. (2)], and (b) the type of the amino acid at site i, α_i. For 20 amino acids and a maximum of 10 neighbors we have 200 parameters to optimize, a number that is comparable to the detailed pairwise model.

THOM1 provides a nonspecific interaction energy, which, as we show in Section IV, has relatively low prediction ability when compared to pairwise interaction models. THOM2 is an attempt to improve the accuracy of the environment model, making it more similar to pairwise interactions. In order to mimic pair energies, we first define the energy $\varepsilon_{\alpha_i}(n_i, n_j)$ of a contact between structural sites i and j, where n_i is the number of neighbors to site i and n_j is the number of contacts to site j (see Fig. 2). The type of amino acid at site i is α_i. Only one of the amino acids in contact is "identifiable." The total contribution due to a site i is then defined as a sum over all contacts to this site $\phi_{i,\mathrm{THOM2}}(\alpha_i, \mathbf{X}) = \sum_j{}' \varepsilon_{\alpha_i}(n_i, n_j)$, with the prime indicating that we sum only over sites j that are in contact with i (i.e., over sites j satisfying the condition $1.0 < r_{ij} < 6.4$ Å and $|i-j| \geq 4$). The total energy is finally given by a double sum over i and j:

$$E_{\mathrm{THOM2}} = \sum_i \sum_j{}' \varepsilon_{\alpha_i}(n_i, n_j) \tag{7}$$

Consider a pair of sites (i, j) which are in contact and occupied by amino acids of types α_i and α_j. Let the number of neighbors of site i be n_i, and let for site j be n_j. The effective energy contribution of the (i, j) contact is

$$V_{ij}^{eff} = \varepsilon_{\alpha_i}(n_i, n_j) + \varepsilon_{\alpha_j}(n_j, n_i) \tag{8}$$

Hence, we can formally express the THOM2 energy as a sum of approximate pair energies $E_{\mathrm{THOM2}} = \sum_{i<j} V_{ij}^{eff}$.

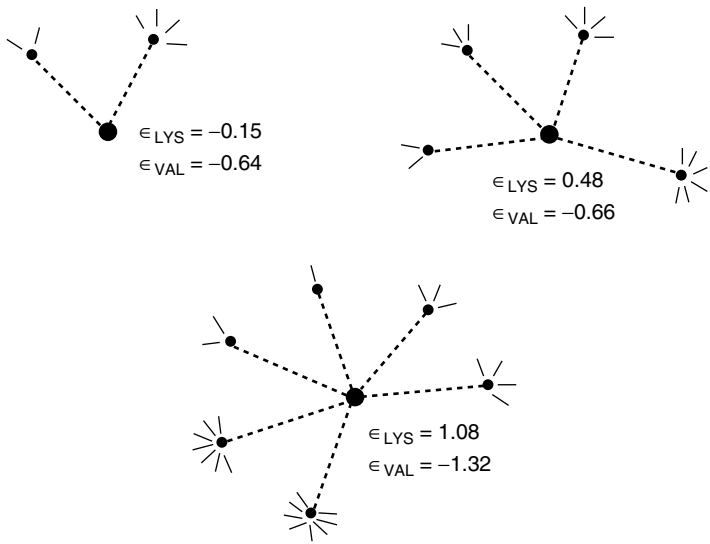


Figure 2. A schematic representation of the interactions with the THOM2 potential. THOM2 assigns scores according to two contact shells. As an example we show a sample of contacts to a site and the associated energies for valine and lysine. As expected, the hydrophobic residue (valine) strongly prefers to be at a site with a large number of neighbors in the first and second shells. Lysine is the extreme case on the polar side.

The effective energy mimics the formalism of pairwise interactions. However, in contrast to the usual pair potential the alignments with THOM2 can be done efficiently. Structural features alone (the number of the contacts) determine the ''identity'' of the neighbor. The structural features are fixed during the computations, making it possible to use dynamic programming. This is in contrast to pairwise interactions for which the identity of the neighbor may vary during the alignment. For 20 amino acids, the number of parameters for this model can be quite large. Assuming a maximum of 10 neighbors, we have $20 \times 10 \times 10 = 2000$ entries to the parameter array. In practice we use a coarse-grained model leading to a reduced set of structural environments (types of contacts) as outlined in Table II.

The use of a reduced set makes the number of parameters (300 when all 20 types of amino acids are considered) comparable to that of the contact potential. Further analysis of the new model is included in Section IV.

III. OPTIMIZATION OF THE ENERGY PARAMETERS

Here we consider the amino acid interactions (the gap energies are discussed in Section V). In order to optimize the energy parameters, we employ the so-called

TABLE II
Definitions of Contact Types for the THOM2 Energy Model[a]

Type of Site[b]	$n=1,2;\ \bar{1}$	$n=3,4,5,6;\ \bar{5}$	$n\geq 7;\ \bar{9}$
$n=1,2;\ \bar{1}$	$(\bar{1},\bar{1})$	$(\bar{1},\bar{5})$	$(\bar{1},\bar{9})$
$n=3,4;\ \bar{3}$	$(\bar{3},\bar{1})$	$(\bar{3},\bar{5})$	$(\bar{3},\bar{9})$
$n=5,6;\ \bar{5}$	$(\bar{5},\bar{1})$	$(\bar{5},\bar{5})$	$(\bar{5},\bar{9})$
$n=7,8;\ \bar{7}$	$(\bar{7},\bar{1})$	$(\bar{7},\bar{5})$	$(\bar{7},\bar{9})$
$n\geq 9;\ \bar{9};$	$(\bar{9},\bar{1})$	$(\bar{9},\bar{5})$	$(\bar{9},\bar{9})$

[a]The THOM2 model defines an energy of a site as a sum of contributions due to contacts to this site. A contact between two amino acids is "on" if their distance is smaller than 6.4 Å. Different types of contacts are defined by the number of neighbors to the two sites involved in contact i.e., the information about the first and second contact layer of a site is used (see Fig. 2). We consider five types of sites in the first layer (primary site i occupied by an amino acid of known type) and three types of sites in the second layer (secondary site j with no amino acid type assigned). Therefore, there are $5 \times 3 = 15$ types of contacts. The primary site i may be occupied by any of the 20 amino acids, leading to $20 \times 15 = 300$ different energy terms. A reduced set of amino acids is associated with a smaller number of parameters to optimize (for 10 types of amino acids, the number of parameters is $10 \times 15 = 150$). The notation we used for each type of site is based on a representative number of neighbors. The number of neighbors n in a given class and its representative are given in the first column (for different classes of sites in the first layer) and in the first row (for different classes of sites in the second layer). The intersections between columns and rows correspond to contacts of different types: a contact between two sites of medium number of neighbors is denoted by $(\bar{5},\bar{5})$, for example.

gapless threading in which the sequence S_i is fitted into the structure $\mathbf{X}_j$ with no deletions or insertions. Hence, the length of the sequence (n) must be shorter or equal to the length of the protein chain (m). If n is shorter than m, we may try $m-n+1$ possible alignments varying the structural site of the first residue $\{a_1 \rightarrow x_1, x_2, \ldots, x_{m-n+1}\}$.

The energy (score) of the alignment of S into $\mathbf{X}$ is denoted by $E(S,\mathbf{X},\mathbf{p})$, where $\mathbf{X}$ stands (depending on the context) either for the whole structure or only for a substructure of length n, relevant for a given gapless alignment. The energy function, $E(S,\mathbf{X},\mathbf{p})$, depends on a vector $\mathbf{p}$ of q parameters (so far undetermined). A proper choice of the parameters will get the most from a specific functional form, where we restrict the discussion below to knowledge-based potentials.

Consider the sets of structures $\{\mathbf{X}_i\}$ and sequences $\{S_j\}$. There is a corresponding energy value for each of the alignments of the sequences $\{S_j\}$ into the structures $\{\mathbf{X}_i\}$. A good potential will make the alignment of the "native" sequence into its "native" structure the lowest in energy. If the exact structure is not in the set, alignments into homologous proteins are also considered "native." Let $\mathbf{X}_n$ be the native structure. A condition for an exact recognition potential is

$$E(S_n, \mathbf{X}_j, \mathbf{p}) - E(S_n, \mathbf{X}_n, \mathbf{p}) > 0, \qquad \forall\, j \neq n \tag{9}$$

In the set of inequalities (9) the coordinates and sequences are given and the unknowns are the parameters that we need to determine. We first describe the sets used to train the potential and then describe the technique to solve the above inequalities.

A. Learning and Control Sets

Two sets of protein structures and sequences are used for the training of parameters in the present study. Hinds and Levitt developed the first set [43] that we call the HL set. It consists of 246 protein structures and sequences. Gapless threading of all sequences into all structures generated the 4,003,727 constraints [i.e., the inequalities of Eq. (8)]. The gapless constraints were used to determine the potential parameters for the 20 amino acids. Because the number of parameters does not exceed a few hundred, the number of inequalities is larger than the number of unknowns by many orders of magnitude.

The second set of structures consists of 594 proteins and was developed by Tobi et al. [32]. It is called the TE set and is considerably more demanding. It includes some highly homologous proteins (up to 60% sequence identity) and poses a significant challenge to the energy function. For example, the set is infeasible for the THOM1 model, even when using 20 types of amino acids (see Section IV). The total number of inequalities that were obtained from the TE set using gapless threading was 30,211,442. The TE set includes 206 proteins from the HL set.

We developed two other sets that are used as control sets to evaluate the new potentials in terms of both gapless and optimal alignments. These control sets contain proteins that are structurally dissimilar to the proteins included in the training sets. The degree of dissimilarity is specified in terms of the RMS distance between the structures. The structure-to-structure alignments (necessary for RMS calculations) were computed according to a novel algorithm [45].

The new structural alignment is based on dynamic programming and provides for closely related structures results that are comparable to the DALI program [44]. Contrary to DALI, we employ (consistently with our threading potentials) the side-chain coordinates, and not the backbone (C_α) atoms, while overlapping two structures (in fact, in analogy with THOM2, we overlap the contact shells, disregarding however the identities of amino acids). Thus, the results of our structure-to-structure alignments refer to superimposed side-chain centers. Our cutoff for structural dissimilarity is 12 Å RMSD.

The first control set, which is referred to as S47, consists of 47 proteins representing families not included in the training. This includes 25 structures used in the CASP3 competition [46] and 22 related structures chosen randomly from the list of VAST [47] and DALI [44] relatives of CASP3 targets. None of the 47 structures has homologous counterparts in the HL set, and only three have counterparts in the TE set. As measured by our novel (both global and local) structure-to-structure alignments, the remaining proteins differ from those

in the training sets by at least 12 Å with respect to HL set and 9.3 Å with respect to TE set (the RMS distance is larger than 12 Å for all but seven shorter proteins), respectively.

The second control set, referred to as S1082, consists of 1082 proteins that were not included in the TE set and which are different by at least 3 Å RMSD (measured, as previously, between the superimposed side chain centers) with respect to any protein from the TE set and with respect to each other. Thus, the S1082 set is a relatively dense (but nonredundant up to 3 Å RMSD) sample of protein families. The training and control sets are available from the web [48].

B. Linear Programming Protocol

The "profile" energies and the pairwise interaction models that were discussed in Section II can be written as a scalar product:

$$E = \sum_{\gamma} n_\gamma p_\gamma \equiv \mathbf{n} \cdot \mathbf{p} \tag{10}$$

where $\mathbf{p}$ is the vector of parameters that we wish to determine. The index of the vector, γ, is running over the types of contacts or sites. For example, in the pairwise interaction model the index γ is running over the identities of the amino acid pairs (e.g., a contact between alanine and arginine). In the THOM1 model it is running over the types of sites characterized by the identity of the amino acid at the site and the number of its neighbors. n_γ is the number of contacts, or sites of a specific type found in a fold. The "number" may include additional weight. For example, the number of alanine–alanine contacts in a protein is (of course) an integer. However, in the Lennard-Jones model, the contact type $A_{\alpha,\beta} \equiv p_\gamma$ is associated with additional geometric weight hidden in a continuous "number" function, $n_\gamma \propto 1/r^m$.

In the pairwise contact model, there are 210 types of contacts for the 20 amino acids. We have experimented with different representations and different numbers of amino acid types. While the Hinds–Levitt set can be solved with a reduced number of parameters, the more demanding requirements of the larger set necessitates (for all models presented here) the use of at least 210 parameters.

We wish to emphasize that the linear dependence of the potential energies on their parameters is not a major formal restriction. Any potential energy $E(\mathbf{X})$ can be expanded in terms of a basis set (say $\{n_\gamma(\mathbf{X})\}_{\gamma=1}^{\infty}$) in which the coefficients are unknown parameters:

$$E(S, \mathbf{X}, \mathbf{p}) = \sum_{\gamma=1}^{\infty} p_\gamma n_\gamma(\mathbf{X}) \tag{11}$$

Note that we deliberately used a similar notation to Eq. (11) and that the information on $\mathbf{X}$ and S is "buried" in $n_\gamma(\mathbf{X})$. A good choice of the basis set will converge the sum to the right solution with only a few terms. Of course, such a choice is not trivial to find, and one of the goals of the present chapter is to explore different possibilities.

The linear representation of the energy simplifies Eq. (9) as follows:

$$E(S_n, \mathbf{X}_j, \mathbf{p}) - E(S_n, \mathbf{X}_n, \mathbf{p}) = \sum_\gamma p_\gamma (n_\gamma(\mathbf{X}_j) - n_\gamma(\mathbf{X}_n)) = \mathbf{p} \cdot \Delta\mathbf{n}_j > 0 \qquad \forall j \neq n \tag{12}$$

Hence, the problem is reduced to the condition that a set of inner vector products will be positive. Standard linear programming tools can solve Eq. (12). We use the BPMPD program of C. S. Meszaros [49], which is based on the interior point algorithm. We seek a point in parameter space that satisfies the constraints, and we do not optimize a function in that space. In this case, the interior point algorithm places the solution at the "maximally feasible" point, which is at the center of the accessible volume of parameters [50].

The set of inequalities that we wish to solve includes tens of millions of constraints that could not be loaded into the computer memory directly (we have access to machines with two to four gigabytes of memory). Therefore, the following heuristic approach was used. Only a subset of the constraints is considered, namely, $\{\mathbf{p} \cdot \Delta\mathbf{n} < C\}_{j=1}^{J}$, with a threshold C chosen to restrict the number of inequalities to a manageable size (which is about 500,000 inequalities for 200 parameters). Hence, during a single iteration, we considered only the inequalities that are more likely to be significant for further improvement by being smaller than the cutoff C.

The subset $\{\mathbf{p} \cdot \Delta\mathbf{n} < C\}_{j=1}^{J}$ is sent to the LP solver "as is." If proven infeasible, the calculation stops (no solution possible). Otherwise, the result is used to test the remaining inequalities for violations of the constraints [Eq. (12)]. If no violations are detected, the process was stopped (a solution was found). If negative inner products were found in the remaining set, a new subset of inequalities below C was collected and sent to the LP solver. The process was repeated, until it converged. Sometimes convergence was difficult to achieve, and human intervention in the choices of the inequalities was necessary. Nevertheless, all the results reported in the present chapter were iterated to a final conclusion. Either a solution was found or infeasibility was detected.

IV. EVALUATION OF PAIR AND PROFILE ENERGIES

In this section we analyze and compare several pairwise and profile potentials, optimized using the LP protocol. As described in the previous section, given the

training set (HL or TE) and the sampling of misfolded (decoy) structures generated by gapless threading, either we obtain a solution (perfect recognition on the training set) or the LP problem proves infeasible.

We use the infeasibility of a set to test the capacity of an energy model. We compare the capacity of alternative energy models by inquiring how many native folds they can recognize (before hitting an infeasible solution). Next, using the control sets, we further test the capacity of the models in terms of generalization and the number of inequalities in Eq. (9) that can be still satisfied, although they were not included in the training. We use the same sets of proteins and about the same number of optimal parameters. The larger the number of proteins that are recognized with the same number of parameters, the better the energy model. We focus on the capacity of four models: the square well and the distance power-law pairwise potentials, as well as THOM1 and THOM2 models. We find that the ''profile'' potentials have in general lower capacity than the pairwise interaction models.

A. Parameter-Free Models

Perhaps the simplest comparison that we can make is for zero-parameter models, and this is where we start. Zero-parameter models have nothing to optimize. They suggest an immediate and convenient framework for comparison, independent of successful (or unsuccessful) optimization of parameters.

An example of pair interaction energy with no parameters is the famous H/P model [51]. In H/P the interactions of pairs of amino acids of the type HP and PP are set to zero and the HH interaction is $-\lambda$. The total energy of a structure is the number of HH contacts (n_i) of structure i times $-\lambda$; that is, $E_i = -n_i\lambda$. The positive parameter λ determines the scale of the energy, however, it does not affect the ordering of the energies of different structures. The difference $E_i - E_n = -\lambda(n_i - n_n)$ is positive or negative, regardless of the magnitude of $|\lambda|$. The existence of a solution of the inequalities in (9) is therefore independent of λ.

For the HL protein set with 246 structures, the HP model predicts the correct fold of 200 proteins. For the larger TE set, the HP recognizes correctly 456 of the 594 proteins. This result is quite remarkable considering the simplicity of the model used, and it raises hopes for even more remarkable performance of the pairwise interaction model once more types of pair interactions are introduced. It is therefore disappointing that the addition of many more parameters to the pairwise interaction model did not increase its capacity as significantly as one may hope, though gradual increase is still observed.

A simple, parameter-free THOM1 model can be defined as follows. As in the pairwise interaction, we consider two types of amino acids: H and P. The energy of a hydrophobic site is defined as $\varepsilon_H(n) = -\lambda n$. For a polar site it is $\varepsilon_P = 0$. It is evident from the above definitions that the parameter-free THOM1 cannot

possibly do better than the HP model, because neighbors of the type HH and HP are counted on equal footing. Indeed the parameter-free THOM1 is doing poorly in both HL and TE sets (only 118 of 246 proteins were solved for HL and 211 of 594 for TE).

B. "Minimal" Models

The parameter-free models are insufficient to solve exactly even the HL set. By "exact" we mean that each of the sequences picks the native fold as the lowest in energy using a gapless threading procedure. Hence, all the inequalities in Eq. (12), for all sequences S_n and structures $\mathbf{X}_j$, are satisfied and the LP problem of Eq. (12) is feasible. This section addresses the question; What is the minimal number of parameters that is required to obtain an exact solution for the HL and for the TE sets? The feasibility of the corresponding sets of inequalities [Eq. (12)] is correlated with the number of model parameters, as listed in Table III.

Consider first the training on the HL set (the solution of the TE set will be discussed in Section IV.D). For the square well potential we require the smallest number of parameters (i.e., 55) to solve the HL set exactly. Only 10 types of

TABLE III
Comparing the Capacity of Different Threading Potentials[a]

Potential	Hinds–Levitt Set	Tobi–Elber Set
SWP, HP model, par-free	200	456
SWP, 10 aa, 55 par	246*	504
SWP, 20 aa, 210 par	246*	530
SWP, 20 aa, 210 par	237	594*
LJ 12-6, 10 aa, 110 par	246*	125
SLJ 12-6, 10 aa, 110 par	246*	488
LJ 6-2, 10 aa, 110 par	246*	530
THOM1, HP model, par-free	118	221
THOM1, 20 aa, 200 par	246*	474
THOM2, 10 aa, 150 par	246*	478
THOM2, 20 aa, 300 par	246*	428
THOM2, 20 aa, 300 par	236	594*

[a]Capacity for recognition of pairwise and profile threading potentials is measured by gapless threading on Hinds–Levitt and Tobi–Elber representative sets of proteins. We compare the capacity of "parameter-free" models (such as the HP and the HP variant of THOM1), demonstrating the superiority of pair potential on profile model in the simplest possible case. We also show that the square well potential and the LJ(6,2) potential are significantly better than THOM1. THOM2, however, is showing comparable performance and is able to learn the TE set (see also Table IV). SWP stands for square well pairwise potential, and SLJ stands for shifted Lennard-Jones potential. For each potential the number of amino acids types used and the resulting number of parameters are reported. The training set used (either HL or TE) is indicated by an asterisk in the second or third column, respectively. The number of correct predictions for structures in HL and TE sets is given in the second and third columns as well.

amino acids were required: HYD, POL, CHG, CHN, GLY, ALA, PRO, TYR, TRP, HIS (see also Table I). The above notation implies that an explicit mentioning of an amino acid excludes it from other, broader subsets. For example, HYD includes now only CYS ILE LEU MET PHE and VAL, whereas CHG includes ARG and LYS only because the negatively charged residues form a separate group, CHN. The LJ, THOM1, and THOM2 models require 110, 200, and 150 parameters, respectively, to provide an exact solution of the same (HL) set (see table IV). It is impossible to find an exact potential for the HL set without (at least) 10 types of amino acids.

Smaller number of parameters led to infeasibility. The optimized models are then used "as is" to predict the folds of the proteins at the TE set. Again, we find that the pairwise interaction model is doing the best and is followed by THOM2 and THOM1, with LJ(12,6) closing.

The above test of the models optimized on the HL set gives an "unfair" advantage to the THOM models that are using more parameters. Nevertheless, even this head start did not change the conclusion that the pairwise square well model better captures the characteristics of sequence fitness into structures. Without the need for efficient treatments of gaps (see Section V), the pairwise interaction model should have been our best choice. Moreover, so far THOM2 is not significantly better than THOM1.

C. Evaluation of the Distance Power-Law Potentials

The LJ(12,6) model, which is a continuous representation of the pairwise interaction, performs poorly. The model trained exactly on the HL set predicts correctly only 125 structures from the 594 structures of the TE set. This result is surprising because the LJ is continuous and differentiable (and more realistic), and has more parameters.

A possible explanation for the failure of LJ(12,6) is the following. The LJ(12,6) is describing successfully atomic interactions. The shape of atoms is much better defined than the shape of amino acid side chains. Amino acids may have flexible side chains and alternative conformations, making the range of acceptable distances significantly larger. To represent alternative configurations of the same type of side chains, potentials with wide minima are required.

To test the above explanation and in a search for a better model, we also tried a shifted LJ function (SLJ) as well as an LJ-like potential with different powers ($m = 6,\ n = 2$, LJ(6,2); see also Fig. 1). As can be seen from Table IV, the "softer" potentials are performing better than the steep LJ(12,6) potential. For example, a LJ(6,2) potential trained on the HL set with 110 parameters (only 10 types of amino acids were used) recognizes correctly 530 proteins of the TE set. Thus, LJ(6,2) has a similar capacity to a square well potential, trained on the same set with 210 parameters.

TABLE IV
Comparison of Performance of THOM2 and Knowledge-Based Pairwise Potentials Using Gapless Threading[a]

Potential	Recognized Structures	Nonsatisfied Inequalities [mln]
BT	1447 (87.3%)	0.28
HL	1412 (85.2%)	3.53
MJ	1410 (85.1%)	0.48
THOM2	1396 (84.3%)	0.38
TE	1353 (81.7%)	0.33
SK	1293 (78.0%)	0.16

[a]The results of gapless threading on the TE set with 20 redundant structures excluded and extended by the S1082 set (see text for details) are reported. The resulting set of 1656 proteins generates about 226 million inequalities. The results of THOM2 potential are compared to five other knowledge-based pairwise potentials by Betancourt and Thirumalai (BT) [37], Hinds and Levitt (HL) [36], Myazawa and Jerningan (MJ) [34], Godzik, Kolinski, and Skolnick (GKS) [38] and Tobi and Elber (TE) [32]. The latter potential was trained using LP protocol and the same (TE) training set. Potentials are ordered according to the number of proteins recognized exactly (out of 1656), given in the second column (values in parentheses indicate the percentage of proteins recognized exactly). The third column contains the number of inequalities (out of 226 mln) that are not satisfied. Note lack of correlation between the number of proteins that are missed and the number of inequalities that are not satisfied.

This suggests that in *ab initio* off-lattice simulations of protein folding, which employ ''residue''-based potentials, LJ(6,2) may be more successful than commonly used LJ(12,6) [12]. Finally, we comment that the training of the LJ type potential was numerically more difficult than the training of the square well potential.

D. Capacity of the New Profile Models

We turn our attention below to further analysis of the new profile models. An indication that THOM2 is a better choice than THOM1 is included in the next comparison: the number of parameters that is required to solve exactly the TE set (see Table III). It is impossible to find parameters that will solve exactly the TE set using THOM1 (the inequalities form an infeasible set). The infeasibility is obtained even if 20 types of amino acids are considered. In contrast, both THOM2 and the pairwise interaction model led to feasible inequalities if the number of parameters is 300 for THOM2 and 210 for the square well potential (SWP). Note that the set of parameters that solved exactly the TE set does not solve exactly the HL set because the latter set includes proteins not included in the TE set.

We have also attempted to solve the TE set using SWP and THOM2 with a smaller number of parameters. For square well potential the problem was proven infeasible even for 17 different types of amino acids and only very similar amino acids grouped together (Leu and Ile, Arg and Lys, Glu and Asp).

Similarly, we failed to reduce the number of parameters by grouping together structurally determined types of contacts in THOM2. Enhancing the range of a ''dense'' site to be a site of seven neighbors or more also results in infeasibility.

Although the rare ''crowded'' sites need to be considered explicitly to solve the TE set with THOM2, a reduced form of the full THOM2 potential trained on the TE set is doing quite well. Consider the contacts $(\bar{9}, \bar{1})$, $(\bar{9}, \bar{5})$, and $(\bar{9}, \bar{9})$. These contacts are very rare and are therefore merged with the contact types $(\bar{7}, \bar{1})$, $(\bar{7}, \bar{5})$, and $(\bar{7}, \bar{9})$. After the merging the number of parameters drops to 200 (instead of 300). The ''new'' potential recognizes 540 proteins out of 594 of the TE set. Only 324 inequalities are not satisfied. Hence, adding 100 parameters increases the capacity of the potential only by a minute amount.

To make a comparison to potentials not designed by the LP approach and to test at the same time the generalization capacity of THOM2, we consider the set of 1656 proteins obtained by adding the S1082 set to the TE set (with 20 redundant structures i.e., structures differing by less than 3 Å with respect to other structures in the TE set removed. This is a demanding test because it contains many homologous pairs and many short proteins that may be similar to fragments of larger proteins. Using the gapless threading protocol, we evaluate the performance of five knowledge-based pairwise potentials. As can be seen from Table IV, the Betancourt–Thirumalai (BT) potential [37] recognized exactly the largest number of proteins, followed by the Hinds–Levitt (HL) [36], Miyazawa–Jernigan (MJ) [34], THOM2, Tobi–Elber (TE) [32], and Godzik–Skolnick–Kolinski (GSK) [38] potentials. However, in terms of the number of inequalities that are not satisfied, the GSK potential is the best, followed by BT, TE, THOM2, MJ, and HL potentials.

The performance of THOM2 potential (84.3% accuracy) is comparable to the performance of other square well potentials (including the TE potential trained on the same set). Because most of the proteins used in this test were not included in the training, we conclude that the perfect learning on the training set avoids overfitting the data.

E. Dissecting the New Profile Models

The THOM1 potential is the easiest to understand and we therefore start with it. In Fig. 3 we examined the statistics of THOM1 contacts from the HL learning set. The number of contacts to a given residue is accumulated over the whole set and is presented by a continuous line. We expect that polar residues have a smaller number of neighbors compared to hydrophobic residues, which is indeed the case. The distributions for hydrophobic and polar residues are shown in Figs. 3a and 3b, respectively. The distributions make the essence of statistical potentials that are defined by the logs of the distribution (appropriately normalized).

The statistical analysis employs only native structures, whereas our LP protocol is using sequences threaded through wrong structures (misthreaded)

during the process of learning. As a result, the LP has the potential for accumulating more information, attempting to put the energies of the mis-threaded sequence as far as possible from the correct thread. In Fig. 4 we show the results of the LP training for valine, alanine, and leucine that are in general agreement with the statistical data above. Nevertheless, some interesting and

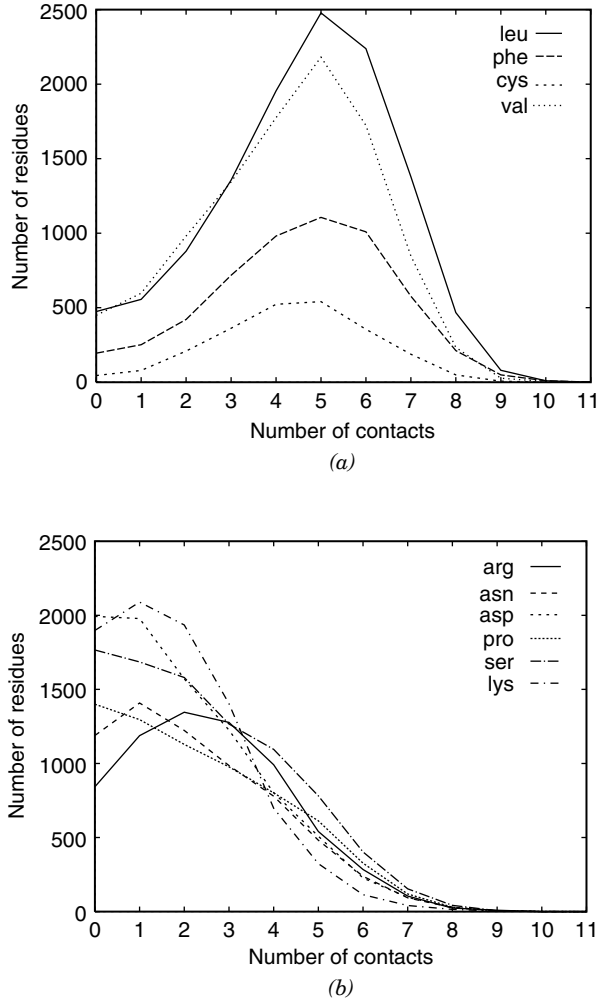
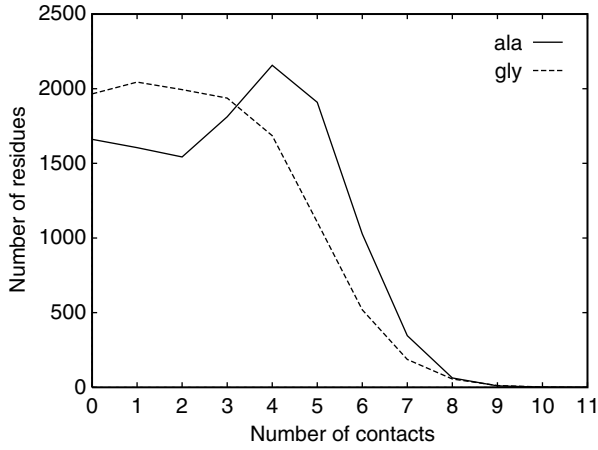


Figure 3. Statistical analysis of contacts for the THOM1 model. (a) Distribution of the number of contacts for hydrophobic residues. (b) Distribution of the number of contacts for (c) Data for alanine and glycine.



(c)

Figure 3 (*Continued*)

significant differences remain. For example, very rare valine residues with 10 neighbors obtain positive energies.

A plausible interpretation of this result is that these rare sites are used to enhance recognition in some cases, due to specific "homologous features." In Table Va we examined the type of contacts (in terms of the number of neighbors) for native and decoy structures.

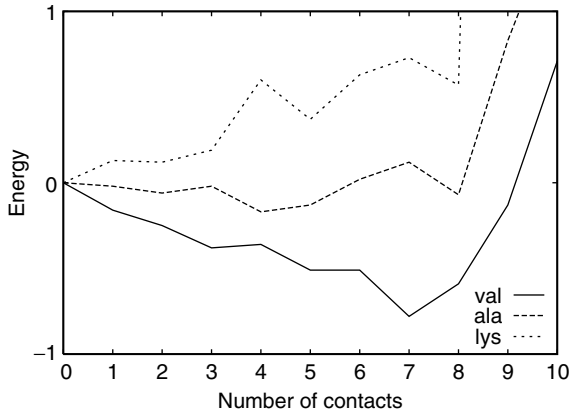


Figure 4. Potentials for THOM1 energy as extracted from LP training. Three residues are shown: alanine, lysine, and valine. Note that the minimum of the potential for valine is at seven neighbors. Note also that lysine has a minimum at zero neighbors.

TABLE V
Characterization of Native and Decoy Structures[a]

(**a**)

Type of Site[b]	Native (HYD/POL)	Decoys (HYD/POL)
(1)	16.97 (4.89/12.09)	24.20 (11.72/12.48)
(2)	17.30 (6.06/11.24)	21.72 (10.52/11.20)
(3)	17.72 (8.29/9.43)	18.70 (9.06/9.64)
(4)	16.60 (9.68/6.92)	15.00 (7.28/7.73)
(5)	14.62 (10.16/4.47)	10.79 (5.24/5.55)
(6)	9.96 (7.66/2.30)	6.04 (2.94/3.10)
(7)	4.95 (4.02/0.92)	2.63 (1.28/1.35)
(8)	1.57 (1.32/0.25)	0.77 (0.38/0.40)
(9)	0.26 (0.21/0.05)	0.12 (0.06/0.06)
(10)	0.04 (0.04/0.00)	0.02 (0.01/0.01)

(**b**)

Type of Contact	Native (HYD/POL)	Decoys (HYD/POL)
$(\bar{1},\bar{1})$	5.09 (1.59/3.50)	11.34 (5.48/5.85)
$(\bar{1},\bar{5})$	9.02 (2.99/6.04)	12.69 (6.15/6.54)
$(\bar{1},\bar{9})$	0.41 (0.15/0.26)	0.35 (0.17/0.18)
$(\bar{3},\bar{1})$	6.25 (2.88/3.37)	9.51 (4.60/4.91)
$(\bar{3},\bar{5})$	24.09 (13.01/11.08)	26.59 (12.91/13.68)
$(\bar{3},\bar{9})$	3.23 (1.88/1.35)	2.29 (1.12/1.18)
$(\bar{5},\bar{1})$	2.77 (1.81/0.96)	3.18 (1.54/1.64)
$(\bar{5},\bar{5})$	28.36 (20.96/7.40)	22.09 (10.75/11.34)
$(\bar{5},\bar{9})$	6.85 (5.11/1.74)	3.84 (1.87/1.96)
$(\bar{7},\bar{1})$	0.40 (0.31/0.09)	0.34 (0.16/0.17)
$(\bar{7},\bar{5})$	9.56 (8.00/1.56)	5.84 (2.85/3.00)
$(\bar{7},\bar{9})$	3.21 (2.60/0.61)	1.54 (0.75/0.79)
$(\bar{9},\bar{1})$	0.01 (0.01/0.00)	0.01 (0.01/0.01)
$(\bar{9},\bar{5})$	0.52 (0.44/0.08)	0.29 (0.15/0.14)
$(\bar{9},\bar{9})$	0.23 (0.19/0.04)	0.09 (0.05/0.05)

[a]Frequencies of different types of sites (relevant for the training of THOM1) found in the native structures of HL set as opposed to decoy structures generated using the HL set are presented in part a. In THOM1 the type of site is defined by number of its neighbors (n). Frequencies are defined by the percentage from the total number of 53,012 native sites in HL set and 556.14 millions of decoy sites generated using HL set, respectively. Frequencies of different types of contacts (appropriate for the training of THOM2) found in the native structures of TE set as opposed to decoy structures generated using TE are given in Table Vb. Different classes of contacts are specified in Table II. Frequencies are defined by the percentage from the total number of 439,364 native contacts in TE set and 10,089.19 millions of decoy contacts generated using TE set, respectively. The comparable site and contact distributions separated for hydrophobic and polar residues (as defined in Table I) are given in parentheses.

It is evident that native structures tend to have more contacts but that the difference is not profound. The deviations are the result of threading short sequences through longer structures (we have more threading of this kind). Such threading suggests a small number of contacts for the set of decoy structures. A sharper difference between native and decoy structures is observed when the contacts are separated to hydrophobic and polar (Table Vb). The difference in hydrophobic and polar contacts is very small at the decoy structures and much more significant for the native shapes.

Another reflection of the same phenomenon is the statistics of pair contacts. For the native structures we find that 42.6% of the contacts are of HH type, 38.2% are HP, and 19.3% are PP. This statistics is of the HL set that has a total of 93,823 contacts. For the decoy structures the statistics of pair contacts is vastly different. Only 23.5% of the contacts are HH, HP contacts are 50% of the total, and 26.5% are PP. The number of contacts that were used is 833.79 million. More details can be found in Tables Va and Vb.

THOM2 has significantly higher capacity, however the double layer of neighbors makes the results more difficult to understand. In Fig. 2 we showed the energy contributions of a few typical structural sites as defined by the THOM2 model. For example, the ''lowest'' picture in Fig. 2 is a site with six neighbors in the first contact shell and a wide range of neighbors in the second shell. The second shell includes a site with just two neighbors as well as a site with nine neighbors. The overall large number of neighbors suggests that this site is hydrophobic, and the corresponding energies of lysine and valine indeed support this expectation.

In Fig. 5 we present a contour plot of the total contributions to the energies of the native alignments in the TE set, as a function of the number of contacts in the first shell, n, and the number of secondary contacts to a primary contact, n', respectively. The results for two types of residues, lysine and valine, are presented. The contribution of a type of site to the native alignment is twofold: its energy $\varepsilon_\alpha(n, n')$ and the frequency of that site f. It is possible to find a very attractive (or repulsive) site that makes only negligible contribution to the native energies because it is extremely rare (i.e., f is small). For specific examples see Table VI. By plotting $f \cdot \varepsilon_\alpha(n, n')$ we emphasize the important contributions. Hydrophobic residues with a large number of contacts stabilize the native alignment, as opposed to polar residues that stabilize the native state only with a small number of neighbors.

It has been suggested that pairwise interactions are insufficient to fold proteins and higher-order terms are necessary [30]. It is of interest to check if the environment models that we use catch cooperative, many-body effects. As an example we consider the cases of valine–valine and lysine–lysine interactions. We use Eq. (8) to define the energy of a contact. In the usual pairwise

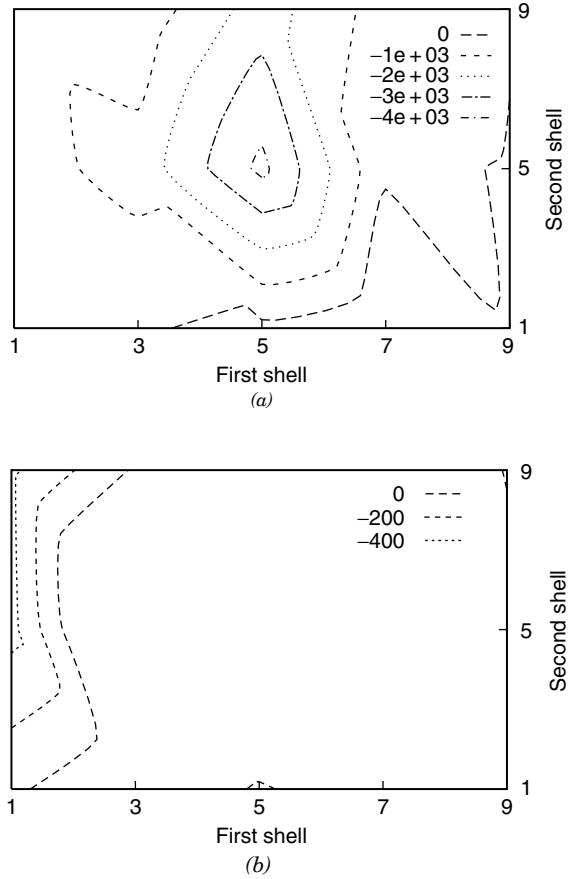


Figure 5. Contour plots of the total energy contributions to the native alignments in the TE set for valine and lysine residues as a function of the number of neighbors in the first and second shells. Part a shows that contacts involving valine residues with five to six neighbors with other residues of medium number of neighbors stabilize most the native alignments. On the other hand, as can be seen from part b, only contacts involving lysine residues with a small number of neighbors stabilize native alignments.

interaction the energy of a valine–valine contact is a constant and independent of other contacts that the valine may have.

In Table VI we list the effective energies of contacts between valine residues as a function of the number of neighbors in the primary and secondary sites. The energies differ widely from -1.46 to $+3.01$. The positive contributions refer,

TABLE VI
Cooperativity in Effective Pairwise Interactions of the THOM2 Potential[a]

(**a**)

	V($\bar{1}$)	V($\bar{3}$)	V($\bar{5}$)	V($\bar{7}$)	V($\bar{9}$)
V($\bar{1}$)	−0.56	−0.41	−0.17	−1.46	3.01
V($\bar{3}$)	−0.41	−0.34	−0.44	−0.30	−0.07
V($\bar{5}$)	−0.17	−0.44	−0.54	−0.61	−0.38
V($\bar{7}$)	−1.46	−0.30	−0.61	−0.49	−0.76
V($\bar{9}$)	3.01	−0.07	−0.38	−0.76	−1.03

(**b**)

	K($\bar{1}$)	K($\bar{3}$)	K($\bar{5}$)	K($\bar{7}$)	K($\bar{9}$)
K($\bar{1}$)	−0.03	−0.03	−0.19	1.18	0.69
K($\bar{3}$)	−0.03	0.28	0.40	0.58	0.61
K($\bar{5}$)	−0.19	0.40	0.52	0.83	0.86
K($\bar{7}$)	1.18	0.58	0.83	1.34	0.38
K($\bar{9}$)	0.69	0.61	0.86	0.38	−0.59

[a]For a pair of two amino acids α and β in contact, we have 25 different possible types of contacts (and consequently 25 different effective energy contributions) because α and β may occupy sites that belong to one of the five different types characterized by the increasing number of contacts in the first contact shell (see Table II). Moreover, the 5×5 interaction matrix will, in general, be asymmetric. The effective energies of contact between two VAL residues with a different number of neighbors are given in part a, whereas the energies of contacts between two LYS residues are given in part b.

however, to very rare types of contacts, and the energies of the probable contacts are negative as expected. Hence, the THOM2 model is compensating for missing information on neighbor identities by taking into account significant cooperativity effects.

To summarize the study of the potentials we provide, in Table VII, the optimal parameters for LJ(6,2), THOM1, and THOM2 potentials.

V. THE ENERGIES OF GAPS AND DELETIONS

In the present section we discuss the derivation of the energy for gaps (insertions in the sequence) and deletions. A gap residue is denoted by a $-$, and a deletion is denoted by a v. For example, the extended sequence $\bar{S} = a_1 - va_3 \ldots a_n$ has a gap at the second structural position (x_2) and a deletion at the second amino (a_2).

A. Protocol for Optimization of Gap Energies

The gap (an unoccupied structural site) is considered to be an (almost) normal amino acid. We assigned to it a score (or energy) according to its environment, like any other amino acid. Here we describe how the energy function of the gap was determined. The parameters were optimized for THOM1 and THOM2, because these are the models accessible to efficient alignment with gaps.

Gap training is similar to the training of other amino acid residues. Only the database of ''native'' and decoy structures is different. To optimize the gap parameters we need ''pseudo-native'' structures that include gaps. We construct such ''pseudo-native'' conformations by removing the true native shape $\mathbf{X}_n$ of the sequence S_n from the coordinate training set and by putting instead a homologous structure, $\mathbf{X}_h$. The best alignment of the native sequence into the homologous structure is $\bar{S}_n$ into $\mathbf{X}_h$, and it includes gaps. We require that the

TABLE VII
Parameters of Some of the Threading Potentials Trained Using the LP Protocol[a]

(**a**)

	HYD	POL	CHG	CHN	GLY	ALA	PRO	TYR	TRP	CYS
HYD	9.32	1.45	−0.44	−0.4	7.35	−1.09	2.17	−0.54	2.29	9.93
POL	1.45	−1.19	−1.07	−0.95	−1.55	−0.75	−1.12	1.41	2.7	0.49
CHG	−0.44	−1.07	2.62	−0.44	−0.35	−1.23	−0.67	0.21	−2.47	−2.51
CHN	−0.4	−0.95	−0.44	1.89	−0.01	3.58	1.32	6.73	8.92	−1.61
GLY	7.35	−1.55	−0.35	−0.01	−1.15	−1.11	2.23	−1.39	−1.17	−1.52
ALA	−1.09	−0.75	−1.23	3.58	−1.11	2.9	−1.53	5.64	−2.43	3.59
PRO	2.17	−1.12	−0.67	1.32	2.23	−1.53	6.51	8.86	8.64	−2.68
TYR	−0.54	1.41	0.21	6.73	−1.39	5.64	8.86	4.98	7.19	−2.55
TRP	2.29	2.7	−2.47	8.92	−1.17	−2.43	8.64	7.19	9.95	−3.74
CYS	9.93	0.49	−2.51	−1.61	−1.52	3.59	−2.68	−2.55	−3.74	−0.12

	HYD	POL	CHG	CHN	GLY	ALA	PRO	TYR	TRP	CYS
HYD	−2.34	0.47	1.71	1.11	−0.21	−0.35	1.22	−1.33	−0.98	−5.11
POL	0.47	0.01	−0.02	0.48	−0.07	−0.7	2.38	−0.81	−0.87	0.57
CHG	1.71	−0.02	0.23	−1.65	−0.51	1.13	0.05	−1.93	1.29	3.73
CHN	1.11	0.48	−1.65	0.12	0	1.58	−2.26	0.33	4.91	3.35
GLY	−0.21	−0.07	0.51	0	1.35	0.41	−0.82	0.47	−1.93	−3.59
ALA	−0.35	−0.7	1.13	1.58	0.41	−1.59	1.3	−2.38	2.12	1.19
PRO	1.22	2.38	0.05	−2.26	−0.82	1.3	−4.08	−3.2	−7.25	−1.37
TYR	−1.33	−0.81	−1.93	0.33	0.47	−2.38	−3.2	−2.9	−5.13	1.67
TRP	−0.98	−0.87	1.29	4.91	−1.93	2.12	−7.25	−5.13	−2.73	−0.2
CYS	−5.11	0.57	3.73	3.35	−3.59	1.19	−1.37	1.67	−0.2	−7.87

TABLE VII (*Continued*)

(**b**)

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
(1)	−0.02	0.10	−0.22	0.02	−0.13	0.02	0.05	−0.05	−0.15	−0.17	−0.04	0.13	−0.40	−0.52	0.29	−0.02	0.02	−0.20	−0.23	−0.16
(2)	−0.06	−0.23	−0.07	0.20	−0.37	0.21	−0.03	0.06	−0.05	−0.30	−0.22	0.12	−0.20	−0.25	0.24	−0.01	−0.10	−0.57	−0.27	−0.25
(3)	−0.02	−0.01	−0.01	−0.43	−0.72	0.09	0.10	0.05	−0.25	−0.48	−0.37	0.19	−0.66	−0.58	0.06	0.05	−0.12	−0.77	−0.37	−0.38
(4)	−0.17	0.12	0.29	0.37	−0.70	0.22	0.40	0.14	−0.31	−0.64	−0.41	0.60	−0.50	−0.68	0.22	0.00	0.21	−0.36	−0.39	−0.36
(5)	−0.13	0.22	0.20	0.68	−1.13	0.33	0.45	0.38	0.24	−0.53	−0.50	0.37	−0.39	−0.65	0.31	0.31	0.02	−0.65	−0.78	−0.51
(6)	0.02	0.32	0.17	0.43	−1.16	0.02	0.70	0.42	0.36	−0.57	−0.58	0.63	−0.80	−0.82	0.75	0.27	0.24	−0.46	−0.72	−0.51
(7)	0.12	−0.10	0.30	0.43	−1.27	0.46	0.39	0.20	0.27	−0.76	−0.54	0.73	−0.44	−0.40	0.42	0.09	0.36	0.12	−0.39	−0.78
(8)	−0.07	0.91	−0.12	−0.01	−1.60	0.51	0.83	0.29	−0.71	−1.37	−0.72	0.57	−0.66	0.25	0.02	0.36	0.15	−0.26	−0.74	−0.59
(9)	0.83	1.36	0.11	0.35	−1.71	0.82	10.00	2.12	0.38	−0.33	1.03	10.00	1.66	−1.03	1.13	2.23	−0.57	10.00	−0.38	−0.13
(10)	1.57	10.00	10.00	10.00	10.00	10.00	10.00	0.83	10.00	−0.93	−0.47	10.00	10.00	0.40	10.00	0.00	10.00	−0.78	10.00	0.71

(**c**)

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
(1,1)	0.23	−0.03	−0.03	−0.08	−0.82	−0.26	0.09	0.29	0.07	−0.12	−0.16	−0.02	0.21	−0.20	0.03	0.05	−0.07	−0.50	−0.64	−0.28
(1,5)	−0.21	−0.26	−0.10	0.20	−1.11	0.00	−0.08	0.00	0.03	0.31	−0.23	−0.13	−0.15	−0.29	−0.23	0.07	−0.09	−0.60	−0.40	−0.36
(1,9)	−6.01	−4.09	−5.42	−6.14	−7.27	−5.88	−5.80	−5.81	−4.75	−5.46	−5.85	−4.91	−4.97	−5.83	−6.17	−5.89	−5.89	−5.25	−6.79	−6.99
(3,1)	−0.01	−0.10	−0.17	0.02	−0.50	−0.09	0.11	0.31	0.04	−0.10	−0.10	0.11	−0.20	−0.17	−0.02	0.40	0.06	−0.31	−0.29	−0.05
(3,5)	−0.08	0.18	0.15	0.13	−0.69	0.12	0.24	0.04	−0.03	−0.29	−0.21	0.14	0.08	−0.32	−0.05	0.06	0.08	−0.36	−0.28	−0.17
(3,9)	−0.29	0.06	−0.33	0.08	−0.78	0.18	0.02	−0.13	−0.47	−0.60	−0.49	0.09	−0.85	−0.07	0.19	0.23	−0.15	−0.15	0.03	−0.27
(5,1)	0.13	−0.21	0.04	0.22	−0.15	−0.11	0.08	0.48	0.19	−0.15	−0.32	−0.06	−0.15	−0.27	0.17	0.19	0.34	−0.07	0.02	0.19
(5,5)	0.06	0.16	0.20	0.17	−0.60	0.04	0.13	0.18	−0.04	−0.25	−0.19	0.26	−0.26	−0.28	0.09	0.11	0.02	−0.36	−0.30	−0.27
(5,9)	−0.65	0.68	−0.26	−0.19	−0.82	−0.09	0.43	−0.36	−0.19	−0.47	−0.42	0.34	0.32	0.07	0.55	0.22	0.01	0.04	−0.46	−0.58
(7,1)	6.29	5.50	5.56	6.02	5.09	5.55	5.68	6.10	5.70	5.59	5.26	6.08	5.64	5.80	5.82	5.23	5.48	6.42	5.17	5.53
(7,5)	0.17	0.29	0.36	0.39	−0.28	0.28	0.45	0.33	0.28	−0.08	−0.01	0.50	0.24	−0.16	0.42	0.13	0.34	0.04	−0.08	−0.03
(7,9)	0.08	0.41	0.00	−0.15	−0.30	0.04	−0.27	0.05	0.69	0.04	−0.17	0.67	0.06	0.03	−0.71	0.82	0.24	−0.36	0.14	−0.25
(9,1)	10.00	4.50	6.05	5.21	4.00	5.94	10.00	10.00	10.00	10.00	6.22	5.59	4.91	6.02	9.61	10.00	10.00	5.88	10.00	10.00
(9,5)	0.26	0.30	0.26	0.71	0.41	−0.02	0.32	0.83	−0.09	1.26	−0.15	0.52	−0.19	0.43	3.07	0.43	0.52	−0.08	0.08	0.21
(9,9)	0.20	0.04	−0.37	−1.34	−1.19	0.47	1.37	−1.36	1.06	−1.99	−0.25	−0.29	1.41	−1.33	6.94	3.22	−0.54	0.81	−0.53	−0.52

[a]Numerical values of the energy parameters for three potentials are given: LJ(6,2) (part a—note that the "repulsive" coefficients *A* are given first, followed by the "attractive" coefficients *B*; the unit distance is 3 Å), THOM1 trained on an HL set of proteins (part b), and THOM2 trained on TE set of proteins (part c; see text for details). The rows in the tables correspond to either different types of amino acids (LJ) or to different types of sites (THOM1) or contacts (THOM2). The columns correspond to different types of amino acids.

alignment $\bar{S}_n$ into the homologous protein will yield the lowest energy compared to all other alignments of the set. Hence, our constraints are

$$E(\bar{S}_n, \mathbf{X}_j, \mathbf{p}) - E(\bar{S}_n, \mathbf{X}_h, \mathbf{p}) = \sum_{\gamma} p_\gamma (n_\gamma(\mathbf{X}_j) - n_\gamma(\mathbf{X}_h)) > 0 \qquad \forall j \neq h, n \tag{13}$$

Equation (13) is different from Eq. (12) in two ways. First, we consider the ''extended'' set of ''amino acids''—$\bar{S}$ instead of S. Second, the native-like structure is $\mathbf{X}_h$—a coordinate set of a homologous protein and not $\mathbf{X}_n$.

The number of inequalities that we may generate (alignments with gaps inserted into a structure and deletions of amino acids) is exponentially large in the length of the sequence, making the exact training more difficult. Some compromises on the size of samples for inequalities with gaps have to be made. To limit the scope of the computations, we optimize here the scores of the gaps only. Thus, we do not allow the amino acid energies (computed previously by gapless threading; see Section III) to change while optimizing parameters for gaps. Moreover, the sequence $\bar{S}$ (obtained by prior alignment of the native sequence against a homologous structure) is held fixed, and gapless threading against all other structures in the set is used to generate a corresponding set of inequalities [Eq. (13)]. By performing gapless threading of $\bar{S}_n$ into different structures, we consider only a small subset of all possible alignments of $\bar{S}_n$, because we fixed the number and the position of the gaps that we added to the native sequence S_n.

Pairs of homologous proteins from the following families were considered in the training of the gaps: globins, trypsins, cytochromes and lysozymes (see Table VIII). The families were selected to represent vastly different folds with a

TABLE VIII
Pairs of Homologous Structures Used for the Training of Gap Penalties[a]

Native	Homologous	Similarity
1mba (myoglobin, 146)	1lh2 (leghemoglobin, 153)	20%, 2.8 Å, 140 res
1mba (myoglobin, 146)	1babB (hemoglobin, chain B, 146)	17%, 2.3 Å, 138 res
1ntp (β-trypsin, 223)	2gch (γ-chymotrypsin, 245)	45%, 1.2 Å, 216 res
1ccr (cytochrome c, 111)	1yea (cytochrome c, 112)	53%, 1.2 Å, 110 res
1lz1 (lysozyme, 130)	1lz5 (1lz1 + 4 res insert, 134)	99%, 0.5 Å, 130 res
1lz1 (lysozyme, 130)	1lz6 (1lz1 + 8 res insert, 138)	99%, 0.3 Å, 129 res

[a]For each pair the native and the homologous structures are specified by their PDB codes, names, and lengths in the first and second column, respectively. In the third column the similarity between the native and the homologous proteins is defined in terms of sequence identity (%), RMS distance (angstroms), and length (number of residues) of the FSSP structure-to-structure alignment, obtained by submitting the corresponding pairs to the DALI server [44].

significant number of homologous proteins in the database. The globins are helical, trypsins are mostly β-sheets, and lysozymes are α/β proteins. Note also that the number of gaps differs appreciably from a protein to a protein. For example, $\overline{S}_n$ includes only one gap for the alignment of 1ccr (sequence) versus 1yea (structure), and 22 gaps for 1ntp versus 2gch.

The energy functional form that we used for the gaps is the same as for other amino acids. The "pseudo-native" structures with extended sequences are added to the HL set (while removing the original native structures). Gapless threading into other structures of the HL set results in about 200,000 constraints for the gap energies. Because we did not consider all the permutations of the gaps within a given sequence and our sampling of protein families is limited, our training for the gaps is incomplete. Nevertheless, even with this limited set we obtain satisfactory results. A representative set of homologous pairs that we used allows us to arrive at scores that can detect very similar proteins (e.g., the cytochromes 1ccr and 1yea) and also related proteins that are quite different (e.g., the globins 1lh2 and 1mba); see Table VIII.

The process of generating pseudo-native is as follows: For each pair of native and homologous proteins the alignment of the native sequence $\bar{S}_n$ into the homologous structure $\mathbf{X}_h$ is constructed. This alignment uses an initial guess for the gap energy, which is based on the THOM1 potential and was based on the following observations.

- The gap penalty should increase with the number of neighbors. For example, we require that $\varepsilon\_(n+1) > \varepsilon\_(n)$ for the THOM1 gap energy.
- The energy of a gap with contacts must be larger than the energy of an amino acid with the same number of contacts. The gap energy must be higher; otherwise, gaps will be preferred to real amino acids. For example, the THOM1 energy of the proline residue with one neighbor is 0.29. Therefore the gap energy must be larger than 0.29; or in general, $\varepsilon\_(n) > \varepsilon_k(n)$, where $k = 1, \ldots, 20$ (types of amino acids) and $n = 1, \ldots, 10$ (number of neighbors).
- The energy of amino acids without contacts is set to zero. The gap energy is therefore greater than zero.

In Table IX we provide the initial guess for the gaps (used to determine pseudo-native states) and the final optimal gap values for THOM1 and THOM2. The value of 10 is the maximal penalty allowed by the optimization protocol that we used. However, this value is not a significant restriction. A solution vector $\mathbf{p}$ can be used to generate another scaled solution $\lambda\mathbf{p}$, where λ is a positive constant.

Nevertheless, note that the maximal value is reached rather quickly. This may indicate that our sampling of inequalities is still insufficient from the perspective of native alignment. The values of gaps that are found only in decoy states are increasing without limit in the LP protocol. For example, it is so rare

TABLE IX
The Gap Penalties for THOM1 and THOM2 Models as Trained by the LP Protocol with the Limited Set of Homologous Structures from Table VIII[a]

(**a**)

Type of Site	Initial Penalty	Optimized Penalty
(0)	0.1	2.7
(1)	0.3	3.9
(2)	0.6	9.0
(3)	0.9	10.0
(4)	2.0	10.0
(5)	4.0	10.0
(6)	6.0	10.0
(7)	8.0	10.0
(8)	9.0	10.0
(9)	10.0	10.0

(**b**)

Type of Contact	Penalty
(0)	1.0
($\bar{1}$,$\bar{1}$)	8.9
($\bar{1}$,$\bar{5}$)	5.7
($\bar{1}$,$\bar{9}$)	10.0

[a]Initial and optimized gap penalties for different types of sites in the THOM1 model are given in part a. Optimized gap penalties for different types of contacts in the THOM2 model are given in part b. Penalties that are not specified explicitly are equal to the maximum value of 10.0.

to find a gap at the hydrophobic core of a protein that our protocol assigns to it the maximal penalty.

The gaps are favored in sites with a small number of contacts. This observation is expected, because gaps are usually found in loops with significant solvent exposure. Note that THOM2 is penalized for a gap for each individual contact.

In Table X we show the results of optimal threading with gaps (using dynamic programming) for myoglobin (1mba) against leghemoglobin (1lh2) structure. We show the initial alignment (with the *adhoc* gap parameters from Table IXa) defining the pseudo-native state, and we also show the results for optimized gap penalties for THOM1 and THOM2. These alignments are largely consistent with the DALI [44] structure–structure alignment (see Table X). Note that the gaps appear (as expected) in loop domains (e.g., the CD, EF, and GH loops). The only ''surprising'' gap is at position 9. Further tests of alignments with gaps for proteins that we did not learn are given in Section VI.

TABLE X
An Example of Output from the Program LOOPP for Sequence-to-Structure Alignments [48][a]

(**a**)

```
. . . . . 1. . . . . 2. . . . . 3. . . . . 4. . . . . 5. . . . .    1–59
SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKSVADIKASPK       1mba
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPE       1lh2
. . . . . 1. . . . . 2. . . . . 3. . . . . 4. . . . . 5. . . . .    1–59

6. . . . . 7. . . . . 8. . . . . ii. . . 9. . . . . 0. . . . . 1. . . . . .   60–116
LRDVSSRIFTRLNEFVNNAANAGKMSA–MLSQFAKEHVGFGVGSAQFENVRSMFPGFV       1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSKGVADAHFPVVKEAILKTI       1lh2
6. . . . . 7. . . . . 8. . . . . 9. . . . . 0. . . . . 1. . . . . .   60–118

. . . 2. . i. . i. . . . . 3. . . . . 4. . i. . . i. i      117–146
ASVAAP-PA-GADAAWTKLFGLIIDALK-AAG-A-                        1mba
KEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA                        1lh2
. 2. . . . . 3. . . . . 4. . . . . 5. . .                  119–153
```

(**b**)

```
. . . . . i. 1. . . . . i. 2. i. . . . . 3. . . . . i. . . . 4. . . . . 5. . . . .   1–55
SLSAAEAD-LAGKSWAPVF-ANK-NANGLDFLVALFEK-FPDSANFFADFKGKSVADIK       1mba
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPE       1lh2
. . . . . 1. . . . . 2. . . . . 3. . . . . 4. . . . . 5. . . . . .   1–59

. . . . 6. . . . . 7. . . . . i. . . . 8. . i. . . . . 9. . . . . 0. . . . . 1. .   56–112
ASPKLRDVSSRIFTRLNEFV-NNAANAG-KMSAMLSQFAKEHVGFGVGSAQFENVRSMF       1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSKGVADAHFPVVKEAILKTI       1lh2
6. . . . . 7. . . . . 8. . . . . 9. . . . . 0. . . . . 1. . . . . .   60–118

. . . . i. . . 2. . . . . 3. . . . . 4. . . . . .      113–146
PGFV-ASVAAPPAGADAAWTKLFGLIIDALKAAGA                    1mba
KEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA                    1lh2
. 2. . . . . 3. . . . . 4. . . . . 5. . .              119–153
```

(**c**)

```
. . . . . i. 1. . . . . 2. . . . . 3. . . . . 4. . . . . i. . . . i. i.   1–55
SLSAAEAD-LAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGK-SVAD-I-K       1mba
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPE       1lh2
. . . . . 1. . . . . 2. . . . . 3. . . . . 4. . . . . 5. . . . . .   1–59

. . . . 6. . . . . 7. . . . . i. 8. i. . . . . 9. . . . . 0. . . . . 1. .   56–112
ASPKLRDVSSRIFTRLNEFVNNA-ANA-GKMSAMLSQFAKEHVGFGVGSAQFENVRSMF       1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSKGVADAHFPVVKEAILKTI       1lh2
6. . . . . 7. . . . . 8. . . . . 9. . . . . 0. . . . . 1. . . . . .   60–118
```

TABLE X (*Continued*)

. . . . . 2. i. . . . . 3. . . . . 4.	113–146
PGFVASVAA-PPAGADAAWTKLFGLIIDALKAAGA	1mba
KEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA	1lh2
. 2. . . . . 3. . . . . 4. . . . . 5. . .	119–153

[a]We compare alignments of myoglobin (1 mba) sequence into leghemoglobin (1lh2) structure using the intial (part a) and trained gap penalties (part b for THOM1 and part c for THOM2). Note that the location of insertions in the initial alignment (which is used for training of gap energies) is to a large extent consistent with the DALI structure to structure alignment [44], which aligns: residues 2–50 of 1 mba to 3–51 of 1lh2 (helices A, B, and C), residues 53–56 of 1 mba to 52–55 of 1lh2 (implying deletions at positions 51 and 52 in 1 mba), residues 59–80 of 1 mba to 56–77 of 1lh2 (E helices), residues 81–86 of 1 mba to 82–87 of 1lh2, residues 87–121 of 1 mba to 89–123 (with the implied insertion at position 88 in 1 lh2), residues 122–139 of 1 mba to 126–143 of 1lh2 (implying two insertions at positions 124 and 125 in 1lh2) and residues 140–145 of 1 mba to 145–150 of 1lh2 (with an insertion at position 144 in 1lh2), respectively. Note also that F and G helices are shifted considerably in the DALI alignment (there is no counterpart of the D helix in 1lh2). The initial THOM1 alignment (part a) is in perfect agreement with the DALI superposition between residues 88 and 150 of 1lh2, except for two insertions at positions 128 and 147 (shifted by three residues with respect to the DALI alignment). The insertions at positions 88, 125, 151, and 153 coincide with the DALI alignment. The THOM2 alignment, with trained gap penalties (part c), is in perfect agreement with the DALI superposition for residues 10 to 50 of 1lh2 and then departs from the DALI alignment, overlapping E, F, and G helices with a smaller shift.

B. Deletions

Yet another technical comment is concerned with ''deletions'' that were mentioned above. A single deletion makes the native sequence shorter by one amino acid, leaving the structure unchanged. In sequence–sequence alignment, deletions can be made equivalent to insertion of gaps. In threading, however, the sequence and the structure are asymmetric. Deleting of residues (amino acids with no corresponding structural sites) or the insertion of gap residues (empty structural sites) is not the same operation.

Nevertheless, in the present chapter we exploit an assumed symmetry between insertion of a gap residue to a sequence and the placement of a ''delete'' residue in a ''virtual'' structural site. The deletions are assigned an environment dependent value that is equal to the averaged gap insertion penalty for the mirror image problem (shorter sequence instead of longer). The deletion penalty is set equal to the cost of insertion averaged over two nearest structural sites. No explicit dependence on the amino acid type is assumed.

While optimization for deletions is not performed in the present chapter, such an optimization is similar to the optimization of gaps. Consider a partial alignment of the sequence $\bar{S}_n = \ldots a_{j'-1} v_{j'} a_{j'+1} \ldots$ into a homologous structure, $\mathbf{X}_h = (\ldots, x_j, x_{j+1}, \ldots)$, in which $a_{j'-1}$ is placed into x_j, $a_{j'+1}$ is placed into x_{j+1}, and $v_{j'}$ is a deletion. What is the energetic cost associated with deleting $v_{j'}$? An estimate would be based on an analogous formulation to the gap residue:

$$\varepsilon_{v_j}(\bar{S}_n, \mathbf{X}_h) = \varepsilon_v(x_j, x_{j+1}) \tag{14}$$

We denoted the ''deletion'' residue by ''v'' because it corresponds to a virtual site inserted into the structure. The deletion is designed as a special energy term that depends on the nearest structural sites: x_j and x_{j+1}. The optimization of the new energy function is the target of a future work.

VI. TESTING STATISTICAL SIGNIFICANCE OF THE RESULTS

In the following we will consider optimal alignments of an extended sequence $\bar{S}$ with gaps into the library structures $\mathbf{X}_j$. We focus on the alignments of complete sequences to complete structures (global alignments [16]) and alignments of continuous fragments of sequences into continuous fragments of structures (local alignment [17]). In global alignments, opening and closing gaps (gaps before the first residue and after the last amino acid) reduce the score. In local alignments, gaps or deletions at the C and N terminals of the highest scoring segment are ignored. Only one local segment, with the highest score, is considered.

Threading experiments that are based on a single criterion (the energy) are usually unsatisfactory. While we do hope that the (free) energy function that we design is sufficiently accurate so that the native state (the native sequence threaded through the native structure) is the lowest in energy, this is not always the case. Our perfect training is for the training set and for gapless threading only. The results were not extended to include (a) perfect learning with gaps or (b) perfect recognition of shapes of related proteins that are not the native.

Despite significant efforts to eliminate all ''false-positive'' signals, the present authors are not aware of any energy function that can achieve this goal. Tobi and Elber [30] conjectured, based on significant numerical evidence, that it is impossible to use a general pair interaction model and to make the native structure the lowest in energy from a set of protein-like structures. The evidence was given for the (simpler) problem of gapless threading. In the present chapter we discuss the more complex problem of threading with gaps that makes the robust detection of the native state even more difficult.

Other investigators use the Z score as an additional filter or as the primary filter [18,52,4,6], and we follow their steps. The novelty in the present protocol is the combined use of global and local Z scores to assess the accuracy of the prediction. This filtering mechanism was found to provide improved discrimination as compared with a single Z score test.

A. The Z-Score Filter

The Z score, which may be regarded as a dimensionless, ''normalized'' score, is defined as

$$Z = \frac{\langle E \rangle - E_p}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}} \tag{15}$$

The energy of the current ''probe''—that is, the energy of the optimal alignment of a query sequence into a target structure—is denoted by E_p. The averages, $\langle \ldots \rangle$, are over ''random'' alignments (that still need to be defined). The Z score is designed as measure of the deviation of our ''hits'' from random alignments. The larger the value of Z, the more significant the alignment. This is because the score is far from the ''random'' average value.

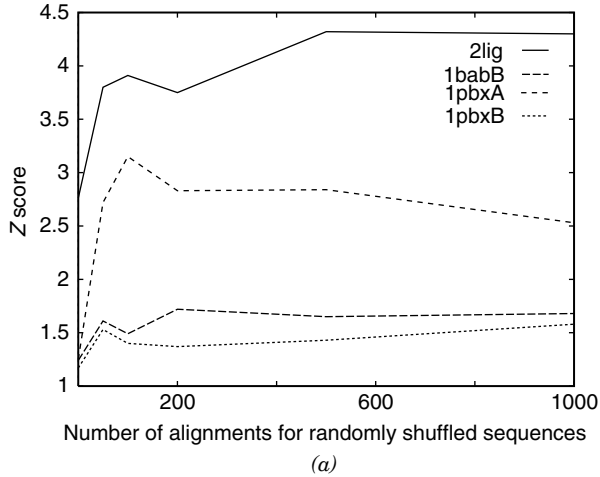
A nontrivial question is how we define a random alignment. The randomness can come from two sources: random structure or random sequence. It is common in *ab initio* folding to assess the correctness of a given structure by comparing its energy to the energies of other structures assumed random. This approach is useful if the number of structures is much larger than the number of sequences (typical of *ab initio* computations). However, in threading protocols the number of structures is relatively small and the number of sequences (with gaps) is significantly larger.

It is therefore suggestive to use a measure, which is based on random sequences instead of random structures. Following the common practice [52–54] we generate this distribution numerically, employing sequence shuffling of the probe sequence. Let $S_p = a_1 a_2 \ldots a_n$ be the probe sequence. We consider the family of sequences that is obtained by permutations of the original sequence.

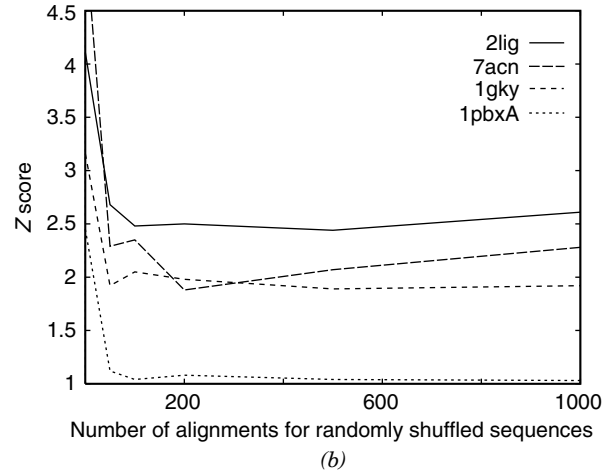
The set of shuffled sequences has the same amino acid composition and length as the native sequence. This leads to a deviation from ''true'' randomness (no constraints) that is used in analytical models. Nevertheless, the constraints are convenient to ''solve'' the problem of the energy of the unfolded state. In the unfolded state all amino acids are assumed to have no contacts with other amino acids. Therefore all the shuffled sequences have the same energy in the unfolded state.

We address the convergence of the Z score in Fig. 6. How many shuffled sequences do we need before we get a reliable estimate? For example, after 100 shuffles the Z score of the global alignment of 1pbxA into 2lig (two different families) suggests that the result is significant. However, enlarging the sample to include 1000 random probes significantly reduces the Z score below the ''cutoff'' of 3. Hence, especially when the signal is not very strong, it is important to fully converge the value of the Z score. The large number of alignments that are performed for the shuffled sequences (between 50 and 1000) makes the process computationally demanding and underlines the need of an efficient algorithm for genomics scale threading experiments.

An essential decision needed is what is a ''good'' score and what is a ''bad'' score. Intuitively, negative energies are assumed ''good.'' Negative energies are lower than the state with no contacts—that is, contacts with water molecules as in the unfolded state. However, no such intuition is obvious for the Z score. To establish a cutoff for the Z score that eliminates false positives, we consider the probability $P(Z_p)$ of observing a Z score larger than Z_p by chance. Clearly our



(a)



(b)

Figure 6. The convergence of the *Z* scores as a function of the number of shuffled sequences. The results for global and local alignments are presented in the parts a and b, respectively. The sequence of the aspartate receptor protein 2lig (not included in the training set) is aligned to all the structures of the HL set, and the best matches are shown. Note that hemoglobin 1pbxA is found among the good matches (false positive) with a global *Z* score of about 3 when using only 100 shuffled sequences to estimate the distribution for random sequences. Converging the *Z* scores makes it possible to better separate the native alignment with respect to incorrect alternatives. The *Z* score for local alignment of 2lig into 1pbxA is small (about 1) and suggests that this match is indeed a false positive. The initial values in the figure correspond to scaled energies of the alignments.

results will be statistically significant only if $P(Z_p)$ is very small. The expectation value of the number of occurrences of false positives in N alignments with a Z score larger than Z_p is $N \cdot P(Z_p)$.

To estimate $P(Z_p)$, we thread sequences of the S47 set through structures included in the Hinds–Levitt set. The probe sequences of known structures were selected to ensure no structural similarity between the HL set and the structures of the probe sequences (see Section III.A). Therefore any significant hit in this set may be regarded as a false positive.

Z scores of local alignments are employed to estimate $P(Z_p)$. In local alignments the number of ''good'' energies (significantly lower than zero) is large, underlining the need for an additional selection mechanism to eliminate false positives. It also makes it possible for us to estimate $P(Z_p)$ for a population of alignments with ''good'' scores. For each probe sequence, Z scores are calculated for 200 structures with the best energies. Only alignments with matching segments of at least 60% of the total sequence length are considered. One hundred shuffled sequences are used to compute the averages required for a single Z-score evaluation. A histogram of the resulting 6813 pairwise alignments is presented in Fig. 7.

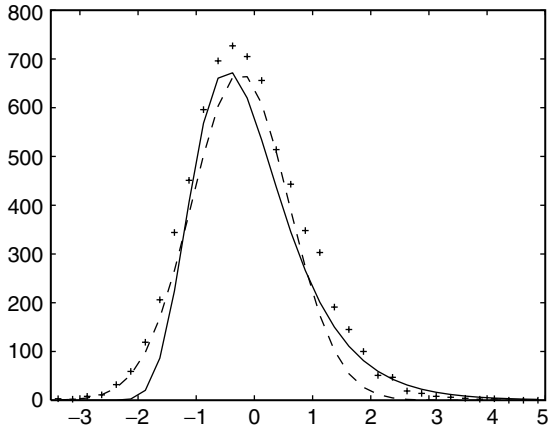


Figure 7. The probability distribution function of the Z scores computed for the population of false positives. A set of 47 sequences from the 547 set of proteins with known structures without homologs in the HL set is used to sample the distribution of Z scores for false positives. Each of the sequences is aligned to all the structures included in HL set. The Z scores are calculated for the 200 best matches (according to energy) using 100 shuffled sequences. The observed distribution of Z scores is represented by $+$. The dashed line shows the attempted analytical fit to a Gaussian distribution, whereas the solid line the analytical fit to the expected extreme value (double exponential) distribution. Note the significant tail to the right, which is the probability of obtaining a relatively large Z score by chance. See text for more details.

Let us denote by $\hat{p}(Z)$ the probability density of finding a Z-score value between Z and $Z + dZ$. Hence, $P(Z_p)$ is given by $P(Z_p) = \int_{-\infty}^{Z_p} \hat{p}(Z)\, dZ$. We approximate the observed distribution ('+') by an analytical fit to the extreme value distribution (represented by a continuous line in Fig. 6), which is defined by [55]

$$\hat{p}(Z) = 1/\sigma \cdot \exp[-(Z - a)/\sigma - e^{(Z-a)/\sigma}] \tag{16}$$

In the realm of sequence comparison, the extreme value distribution has been used to model scores of random sequence alignments for local, ungapped alignments [56] as well as for local alignments with gaps [57].

The observed distribution is asymmetric and has a long tail toward high Z-score values (which is the tail that we are mostly interested in). Note, however, that there are significant differences between the numerical data and the analytical fit (and of course from the symmetric Gaussian distribution; dotted line in Fig. 7). Some deviations are expected because the distribution we extracted numerically differs from a random distribution. As discussed above, we use, for example, only alignments with negative energies. Hence, the energy filter was already employed.

Using analytical fit, we find that $P(Z_p) = 1 - \exp[-\exp(-1.313 \cdot (Z_p + 0.466))]$ with the 98% confidence intervals 1.313 ± 0.112 and 0.466 ± 0.079. For example, we estimate that the probability of observing a random Z score that is larger than 4 is 0.003. We emphasize, however, that the analytical fit is an upper bound as is shown in Fig. 6. For example, the observed number of Z scores larger than 4.0 is equal to 3—as opposed to the expected number of finding a Z score larger than 4.0, which is equal to (according to the analytical fit) $6813 \cdot 0.003 = 20.4$.

We observe similar discrepancy for global threading alignments of all the sequences from the HL set into all the structures in the HL set. For each probe sequence we select the 10 best matches (with lowest energies) that are subsequently subject to the statistical significance test, resulting in a sample of 2460 Z scores. Only five of the calculated Z scores, which are larger than 3.0, correspond to false positives. Using the analytical fit from Fig. 7 the expected number of observing by chance Z scores larger than 3.0 is equal to 24.6. Thus, it seems that the conservative estimate of the tail of the extreme value distribution indeed provides an upper bound for the probability of observing a false positive with a low energy and a high Z score.

B. Double Z-Score Filter

When searching large databases, the probability of observing false positives is growing, because the expected number of false positives is $N \cdot P(Z_p)$, where N is the number of structures in the database. Therefore, only relatively high Z

scores may result in significant predictions. Unfortunately, there are many correct predictions with low Z scores that overlap with the population of false positives. A high cutoff will therefore miss many true positives. Restricting the Z score test to only best matches (according to energy) is still insufficient. Therefore we propose an additional filtering mechanism, based on a combination of Z scores for global and local alignments. The double Z-score filter eliminates false positives, missing much smaller number of correct predictions.

Global alignments (in contrast to local alignments) are influenced significantly by a difference in the lengths of the structure and the threaded sequence. The matching of lengths was considered too restricted in previous studies [58]. However, at our hands and using environment-dependent gap penalty, the Z

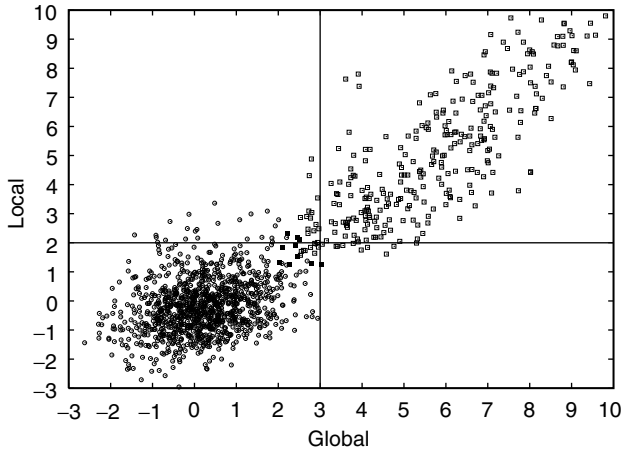


Figure 8. The joint probability distribution for the Z scores of global and local alignments. The distribution at the lower left corner (circles) is the result of the alignments of the 547 set sequences against all structures in HL set. The Z scores for the false positives are computed using 1000 shuffled sequences for both global and local alignments to ensure convergence. Only weak energy constraint are used; that is, 100 best global and 200 best local matches are subject to a Z score test, and then a given pair (global Z score, local Z score) is included if the energy of the global alignment is negative. The resulting 1081 pairs are included in the figure. The best pair in this population is slightly below the threshold (3.0,2.0). The population in the right upper corner represents (square boxes) 331 pairs of HL sequences aligned to HL structures with global Z scores larger than 2.5 and local Z scores larger than 1 [some of the Z scores fall beyond the (10,10) range). This set includes 236 native alignments and 95 non-native alignments. There are 10 matches that are false positives (filled squares), and they are all below the threshold (3, 2). Four of them are marginally so. The Z scores of this distribution were generated using 1000 shuffled sequences for global alignments, but only 50 for local alignments. Stiffer energy constraints were employed in which only the 10 best matches (according to energy) for global alignments and with 200 best matches for local alignments were considered. Of course, there is still a population of matches below (2.5,1.0) threshold (including 10 native alignments). However, the number of false positives below this threshold grows quickly, making predictions with Z scores in this range difficult.

score of the global alignment was proven a useful independent filter. This filter is an addition to the use of energy (of local and global alignments) and of the Z score of local alignments.

In Fig. 8 we present the joint probability distribution for global and local Z scores for a population of false positives versus a population of correct predictions. The squares at the upper right corner represent correct predictions, resulting from 331 native alignments (of a sequence into its native structure) and homologous alignments (of a sequence into a homologous structure) of the HL set proteins. The circles at the left lower corner are false positives obtained from the alignments of the sequences of the S47 set against all structures in the HL. The procedure is the same as the one used previously to generate the probability density function for the Z scores of local alignments (see Fig. 7). However, the Z scores are computed using 1000 shuffled sequences for both global and local alignments, which is sufficient to converge the values of the Z scores. The converged results reduce somewhat the tails of the distribution. For example, the number of false positives with a global Z score larger than 2.5 and a local Z score larger than 1.0 is equal to 3, as compared to 7 with only 100 shuffled sequences.

Figure 8 shows that the thresholds of 3.0 for global Z scores and of 2.0 for local Z scores are sufficient to eliminate all the false predictions. These cutoffs result in a number of misses, for example, 23 native alignments are dismissed as insignificant (see also the next section). However, this is a price we have to pay for high confidence levels in our predictions. The total number of pairwise alignments for which we compute the global and the local Z scores, and subsequently test for the presence of false positives, is about 10,000. Hence, we estimate that the probability of observing a single false positive with a global and a local Z score larger than the 3.0 and 2.0 thresholds is smaller than 0.0001.

VII. TESTS OF THE MODEL

There are three tests that we perform in this section on the THOM2 potential. We use optimal alignments and the double Z-score test proposed in Section VI. First, we analyze the results of threading the sequences of the HL set into all the structures of the HL set. Self-recognition and family recognition are discussed. Next, threading of the CASP3 sequences into an extended TE set is used to test the performance of the new threading protocol on the set of folds that were not included in the training. Finally, further tests of family recognition are presented, including the comparison of THOM2 results with those of a pairwise model using the frozen environment approximation.

A. The HL Test

The HL set was partially learned (using gapless threading). The first test verifies that the additional flexibility of gaps and deletion maintain good prediction

ability such as self- and family recognition. We note that our training did not include the *Z* score, so successful predictions based on only the *Z* score are useful tests even if performed on the training set of structures. The second test is a prediction experiment on proteins not included in the learning set. There are 40 new proteins that are included in Table XIa.

TABLE XI
A Summary of the THOM2 Threading Alignments of All the Sequences of the HL Set Into All the Structures of the HL Set[a]

(**a**)

1bbt1, 1gp1A, 1grcA, 1ipd, 1lap, 1lpe, 1phd, 1prcL, 1prcM, 1rbp, 1rhd, 1rnh, 1stp, 1wsyB, 2cna, 2cts, 2gbp, 2snv, 2wrpR, 3sicE, 4dfrA, 4gcr, 4rcrH, 4rcrL, 4rcrM, 7acn, 8adh, 4cms, 4i1b, 5fd1, 1atnA, 1tfd, 2aaiA, 2aaiB, 2bbkA, 2bbkB, 2lig, 2mnr, 2plv1, 2sas

(**b**)

Energy	*Z* Score	*N*
First	First	234
First	Second	4
First	Fourth	1
Second	Second	3
Weak	Weak	4

(**c**)

Z Score	*N*
First	177
Second or Third	35
Fourth and lower	14
Weak	11
Very Weak	9

[a]A list of proteins of the HL set that were not included in the training (TE) set is given in part a. A summary of the native global alignments is included in part b. Part c contains a summary for the native local alignments. The number of native alignments *N*, with ranks specified in terms of energies (first column in part b) and *Z* scores (second column in part b and the first column in part c), is given in the last column. For global alignments, ''weak'' is used to mark alignments with a weak energy or *Z*-score signals. There are four weak alignments corresponding to the photosynthetic centers membrane domains that were not included in the training set. Only five out of the remaining 242 native alignments obtain *Z* scores smaller than 3.0 (four alignments with *Z* scores larger than 2.5 and one alignment with a *Z* score smaller than 2.5). For local alignments, ''very weak'' denotes native alignments with *Z* scores smaller than 1.0, whereas ''weak'' marks alignments having *Z* scores larger than 1.0 and smaller than 2.0. There are 226 local native alignments with *Z* scores larger than 2.0. Note also that energy is not used to filter local alignments (beyond the initial restriction to 200 best candidates).

The self-recognition of the HL set proteins in terms of optimal alignments and *Z*-score filters is summarized in Tables XIb and XIc (see also Fig. 8). In Table XIb we provide the data for the global alignment. Energy and *Z*-score filters are considered. Of the total of 246 proteins, 234 are clear-cut cases (the energy and the *Z* scores of the native alignment are at the top). The four failures are membrane proteins (photosynthetic reaction centers) that were not included in the training set. In Table XIc the data for the local alignments are provided. We use only the *Z* score as a filter because there are many incorrect alignments with good (negative) energies. Among nine native alignments that are clear failures ($Z < 1.0$), six refer to structures that were included in the training set.

As examples of protein families, represented in the HL set, we discuss cytochromes, dehydrogenases, and acid proteases. Cytochromes were included in the training of the gaps, so we might expect that identification of cytochromes will be easy. Yet, this is not the case and we report a ''bad'' case scenario for some of the members of the family in Table XIIa. The *Z*-score values are below what we usually consider as a significant hit. Even though the correct proteins make it to the top, the global *Z* scores are too low (1.3–1.4) to confirm the prediction. The successful recognition of dehydrogenases and acid proteases families is shown in Tables XIIb and XIIc. We comment that most of the family members of the HL set are recognized irrespective of the choice of the probe sequence, as long as it belongs to a given family. More extensive tests of family recognition are discussed in Section VII.C.

Global *Z* scores reported in Tables XI and XII are converged using 1000 shuffled sequences. Local *Z* scores are, however, computed using only 50 shuffled sequences. The constraint here is of computational resources. Global *Z* scores are computed only for 10 energy-best structures and can be done

TABLE XII
Examples of Predictions for Families of Homologous Proteins[a]

(**a**)

Query srequence: 5cytR	Structure	Energy	*Z* score	RMS
Global alignments	5cytR	−22.1	4.1	0.0
	1ccr	−10.4	1.4	6.9
	3c2c	−10.4	1.4	4.9
	1rro	−11.2	1.3	—
	256bA	−12.0	1.0	—
Local alignments	5cytR	−31.0	3.9	0.0
	1ccr	−35.6	3.2	1.9
	1yea	−23.9	3.2	1.9
	2ccyA	−22.8	3.0	—
	2fox	−27.6	2.3	—

TABLE XII (*Continued*)

(**b**)

Query Sequence: 1llc	Structure	Energy	*Z* Score	RMS
Global alignments	1llc	−80.0	7.0	0.0
	1lldA	−60.7	4.4	5.3
	1ldnA	−52.9	4.2	4.6
	4mdhA	−47.4	2.1	6.7
	6ldh	−45.8	1.6	4.6
Local alignments	1ldnA	−73.4	5.2	4.1
	1llc	−89.8	5.2	0.0
	1lldA	−74.1	4.4	5.0
	6ldh	−73.4	4.3	4.4
	1ipd	−82.7	2.8	—

(**c**)

Query Sequence: 1pplE	Structure	Energy	*Z* Score	RMS
Global alignments	1pplE	−77.3	9.5	0.0
	2er7E	−61.4	7.3	2.9
	3aprE	−51.9	4.3	3.9
	4cms	−45.0	4.2	5.4
	4pep	−43.1	3.6	5.7
Local alignments	1pplE	−79.2	12.9	0.0
	2er7E	−68.6	8.3	2.9
	3aprE	−59.6	4.5	5.2
	4pep	−55.4	3.3	5.7
	1prcH	−46.6	2.2	—

[a]The results of global and local threading alignments for representatives of three families in the HL set are reported. The families are cytochromes (part a), lactate and malate dehydrogenases (part b), and pepsin-like acid proteases (part c). Five best alignments, ordered according to their *Z* scores (fourth column), are reported. The names of the query sequences are specified in the first column, target structures in the second, and the energy of the alignment in the fourth column, respectively. In the last column the RMS distance between the (known) structure of the probe (query) and the target structure, according to a novel structure-to-structure alignment (Meller and Elber [45]), is provided. RMS distances larger than 12 Å are indicated by a dash. Note that in a ''bad'' case scenario a distance of about 5 Å between the superimposed side-chain centers of 5cytR and 3c2c is sufficient to make threading identification virtually impossible because the *Z* score is too low (see part a). The local alignment provides a significantly improved *Z* score in this case. On the other hand, there are homologous structures that are not detected by the local alignments, although their global *Z* scores are high. Examples are malate dehydrogenase 4mdh (see part b) and acid protease 4cms (see part c). The structures with the PDB codes 1rro and 2fox (part a), 1ipd (part b) and 1prcH (part c) do not belong to the families of interest.

accurately. Local Z scores are computed for 200 alignments. The number of alignments with negative energies, which needs to be probed by an additional filter, is much larger for local alignments. With limited computational resources and/or a large-scale alignment project, it may be necessary to use Z scores that are not fully converged. For example, when aligning a 1pplE sequence into a 1prcH structure, a Z score of 1.8 with 1000 shuffled sequences is obtained, as opposed to 2.2 with only 100 shuffles sequences.

Finally, we remark that we were able to find alignments (with gaps) that have energies lower that the energy of the native state. Moreover, even aligning a sequence into its own structure may result in lower energy than the native if the addition of gaps and deletions is favorable. One such example is the alignment (with gaps) of 1llc onto its native shape.

B. Recognition of Folds Not Included in the Training

In order to assess the generalization capacity of THOM2 in terms of optimal alignments, we use the S47 set again. Let us recall that the S47 set is composed of CASP3 [46] targets and their relatives. Using CASP3-related structures is a convenient way of finding protein shapes that are not sampled in the training. The experiment we perform is for self-recognition and is not aimed at finding remote relatives (as in CASP). The results are summarized in Table XIII. The native and

TABLE XIII
Self-Recognition for Folds That Were Not Learned[a]

PDB Code (len)	FSSP Z-score (RMS)	THOM2 Global Z score	THOM2 Local Z score
1HKA (158)	33.0 (0.0)	**7.1**	**7.1**
1VHI (139)	4.3 (5.2)	0.2	0.3
2A2U (158)	33.8 (0.0)	**2.5**	**4.0**
1BBP (173)	11.6 (3.3)	**3.5**	**3.0**
2EZM (101)	55.3 (0.0)	**3.7**	**3.2**
1QGO (257)	46.0 (0.0)	**5.6**	**7.6**
1ABE (305)	6.4 (3.4)	0.5	0.4
1BYF (123)	29.5 (0.0)	1.8	2.8
1YTT (115)	16.4 (2.2)	−0.1	1.4
1JWE (114)	26.9 (0.0)	2.6	2.3
1B79 (102)	18.7 (1.3)	0.3	1.3
1B7G (340)	61.5 (0.0)	**8.7**	**8.8**
1A7K (358)	25.1 (2.9)	−0.4	−0.9
1EUG (225)	43.0 (0.0)	**3.4**	**3.0**
1UDH (244)	30.8 (1.7)	−1.0	2.9
1D3B (72)	18.4 (0.0)	**3.5**	**2.8**
1B34 (118)	13.4 (1.1)	1.9	2.0
1DPT (114)	24.8 (0.0)	**6.2**	**6.0**
1CA7 (114)	18.7 (1.2)	**4.0**	**2.5**
1BG8 (76)	19.1 (0.0)	**3.4**	**3.5**

TABLE XIII (*Continued*)

PDB Code (len)	FSSP *Z*-score (RMS)	THOM2 Global *Z* score	THOM2 Local *Z* score
1DJ8 (79)	16.2 (0.7)	**5.1**	**3.9**
1QFJ (226)	42.7 (0.0)	**8.1**	**8.4**
1VID (214)	7.1 (3.1)	−2.0	0.5
1BKB (132)	25.1 (0.0)	2.7	1.5
1EIF (130)	17.4 (1.6)	**3.5**	**2.0**
1B0N (103)	19.5 (0.0)	**4.7**	**5.0**
1LMB (87)	8.0 (5.3)	0.3	0.1
1BD9 (180)	38.8 (0.0)	**4.5**	**5.8**
1BEH (180)	36.0 (0.3)	**7.4**	**5.8**
1BHE (376)	70.2 (0.0)	6.7	0.6
1RMG (422)	36.9 (2.2)	0.9	—
1B9K (237)	39.7 (0.0)	**8.1**	**8.2**
1QTS (247)	36.1 (0.7)	**3.5**	**6.4**
1EH2 (95)	24.3 (0.0)	**6.0**	**6.5**
1QJT (99)	7.6 (2.5)	**3.6**	**3.7**
1BQV (110)	20.9 (0.0)	**3.5**	**2.3**
1B4F (82)	3.2 (3.3)	0.0	1.7
1CK2 (104)	26.0 (0.0)	**5.2**	**4.3**
1CN8 (104)	14.3 (2.2)	**5.3**	**2.0**
1BL0 (116)	24.9 (0.0)	0.5	0.5
1JHG (101)	3.4 (6.6)	1.1	1.0
1BNK (100)	24.9 (0.0)	**5.4**	**6.3**
1B93 (148)	31.4 (0.0)	**4.0**	**3.2**
1MJH (143)	6.1 (3.4)	0.3	1.3
1BK7 (190)	37.2 (0.0)	**7.7**	**9.0**
1BOL (222)	19.7 (2.3)	0.1	−1.0
1BVB (211)	37.3 (0.0)	**5.3**	**4.3**

[a]Twenty-two pairs of CASP3 targets and their structural relatives, as well as an additional three singleton targets, are added to the TE set. Their PDB codes are given in the first column (with lengths in parentheses). The actual CASP3 targets are given as the first structure of each pair (e.g., 1HKA from the pair 1HKA, 1VHI). If the domain is not specified and one refers to a multidomain protein, then the A (or first) domain is used. The results of global and local THOM2 threading of the 25 CASP3 sequences into an extended TE set (594 + 47 structures) are reported in the third and fourth column, respectively. Two of 25 native alignments gave weak signals (DNA-binding protein 1BLO and glycosidase 1BHE). Four other native alignments (2A2U, 1BYF, 1JWE, and 1BKB) provide global *Z* scores somewhat smaller than 3. The DALI *Z* scores and RMS deviations for structure-to-structure alignments into native and homologous structures are reported in the second column (the native structures have RMS distances of zero). Note that low *Z* scores indicate that only short fragments of the respective structures are aligned and the resulting RMS deviation may not be representative. Nine related structures, among the 14 pairs with the DALI Z score larger than 10, obtain *Z* scores larger than 3.0 and 2.0 for the global and local THOM2 threading alignments, respectively. The alignment of 2A2U sequence into the 1BBP structure was the only significant hit of any of the target sequences into the structures included in the training (TE) set. Thus, no false positives with scores above our confidence cutoffs were observed. All the predictions that can be made with a high degree of confidence are indicated by *Z* scores printed using boldface type.

homologous shapes were embedded in the structures of the TE set, and the sequences of CASP targets were aligned into all the structures of such extended set. We provide in the table the results of the native alignments and the alignments into related homologous structure, irrespective of their rank.

One encouraging observation is that the native structures are found with high probability. Twenty of 25 structures would have been found if the native structure was included in the set. A less encouraging observation is the sensitivity of the results to structural fluctuations. The THOM2 model can identify related structures only if their distance is not too large. Nine out of 14 homologous structures with the DALI [44] Z score for structure-to-structure alignment larger than 10 are detected with high confidence. Only one homologous structure with the DALI Z score lower than 10 is detected.

Only three among the 25 structures of the CASP3 targets included in Table XIII had homologous counterparts in the training set. These are 2a2u, 1byf, and 1eug with their respective homologous proteins 1bbp, 2msb, and 1akz. It is therefore reassuring that most of the native structure and a significant fraction of relatives are recognized in terms of both their energies and the Z scores. Also, there are no further significant hits into other structures from the TE set. Hence, no false positives above our confidence thresholds are observed in this test. We conclude that our nearly perfect learning (on a training set) preserves significant capacity for identification of new folds using optimal alignments with gaps.

Note also that good scores with the global alignment are obtained for length differences (between sequence and structure) that are on the order of 10%. This was made possible by using environment-dependent gaps. When the differences in length are profound (e.g., 1bqv versus 1b4f), it is obvious and expected that the global alignment will fail. Large differences are clearly focused on identification of domains and not a whole protein. This is a different problem, which the present chapter does not address.

C. Recognition of Protein Families: THOM2 Versus Pair Energies

Three families are considered here: globins (92 proteins), immunoglobins (Fv fragments, 137 proteins), and the DNA-binding, POU-like domains (26 proteins). Sequences of all family members are aligned optimally to all the structures in the family. Both the local and global alignments are generated for each sequence–structure pair, and the results are compared in terms of the sum of Z scores for global and local alignments. Thus we employ here a simplified version of the double Z-score filter discussed before. The THOM2 results are compared to the results of the TE pairwise potential, which was trained on the same set of 594 proteins using the LP protocol. The difference in the LP protocol was that an objective function was optimized.

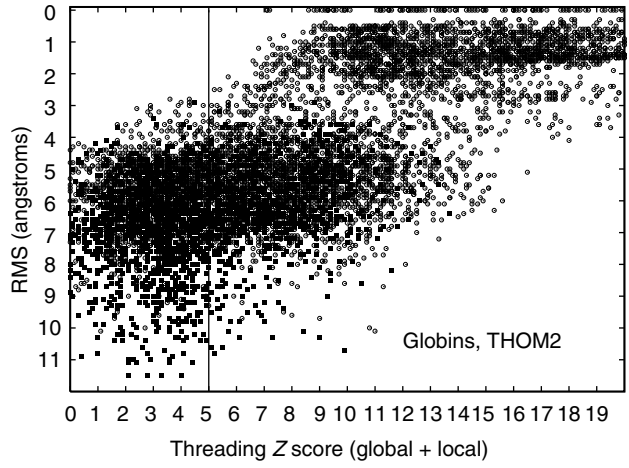
The alignments due to the pairwise potential are computed using the first iteration of the frozen environment approximation (FEA) [22]. That is, when

evaluating fitness of a query sequence into a structure, we assume that types of contacts are fixed according to the native identities of sites making contacts to a primary site occupied by a query residue. Such an approach is in fact a different profile approximation to the "true" pair energies. In THOM2, the number of neighbors to a secondary site approximates its identity, whereas in FEA it is approximated by the identity of the native residue at that site. In principle, the FEA should be iterated until self-consistency is achieved [22]. Purely structural characterization of contact types in THOM2 avoids this problem.

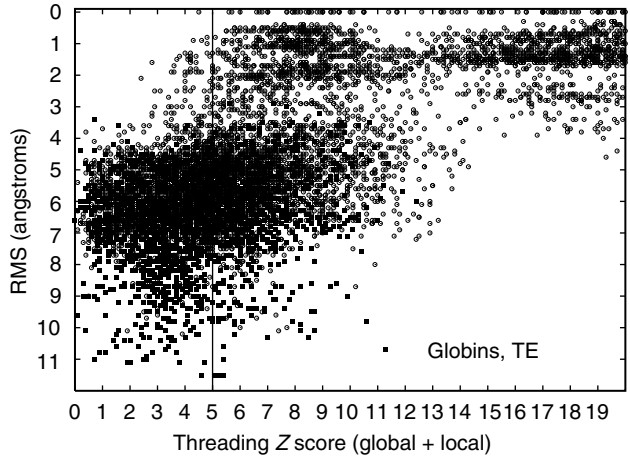
In order to compute optimal alignments with the FEA, we need to set the gap penalties for the TE potential. After some experimentation the insertion penalties are chosen to be proportional to the number of neighbors to a site, $\varepsilon_{-}(n) = 0.2 \cdot (n + 1)$. This choice is consistent with the THOM2 gap energies, which also penalize sites of no neighbors. The proportionality coefficient was gauged using the same families that were used to train THOM2 gap energies. However, no LP training was attempted. The deletion penalties are also consistent with the THOM2 model, and they are defined in the way described in Section V.

Figures 9a to 9f show the joint histograms of the sum of Z scores for local and global threading alignments versus the RMS deviations between superimposed (according to our novel structure-to-structure alignments; see Section III.A) side-chain centers. Figures 9a, 9c, and 9e show the results for THOM2 (for globins, immunoglobins, and POU-like domains, respectively), whereas Figs. 9b, 9d, and 9f show the corresponding results for TE potential with FEA. The vertical lines in the figures correspond to the sum of global and local Z scores equal to 5, which roughly discriminates the high confidence matches (with higher Z scores) and lower confidence matches that might be obscured by the false positives.

The population of matches that are difficult to identify by pairwise sequence-to-sequence alignments is represented by the filled squares. Sequence alignments are generated using Smith–Waterman algorithm with the BLOSUM50 substitution matrix (with the signs inverted) and structurally biased gap penalties [$\varepsilon_{-}(n) = 8 + (n - 5)$, where n is the number of neighbors to a site]. Confidence of matches is estimated using Z scores defined, analogously to threading alignments, by the distribution of scores for shuffled sequences. We find that structurally biased gap penalties improve the recognition in case of weak sequence similarity. We do not observe false positives with more than 50% of the query sequence aligned and with a Z score larger than 8 (the distribution of Z scores for sequence substitution matrices is vastly different from that of threading potentials, with very high Z score for homologous sequences). All the matches represented by circles can be identified with high confidence by pairwise sequence-to-sequence alignments.

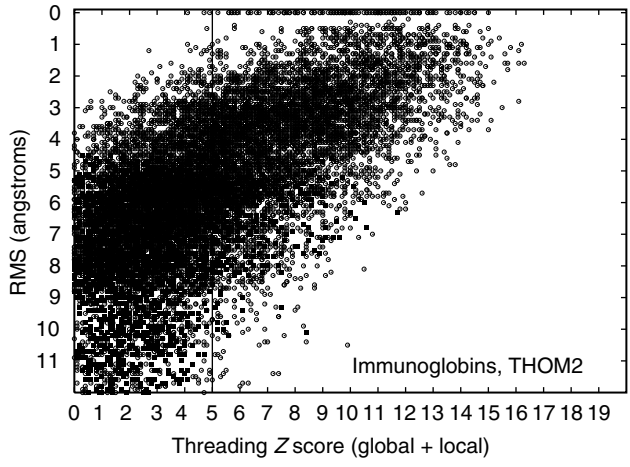


(a)

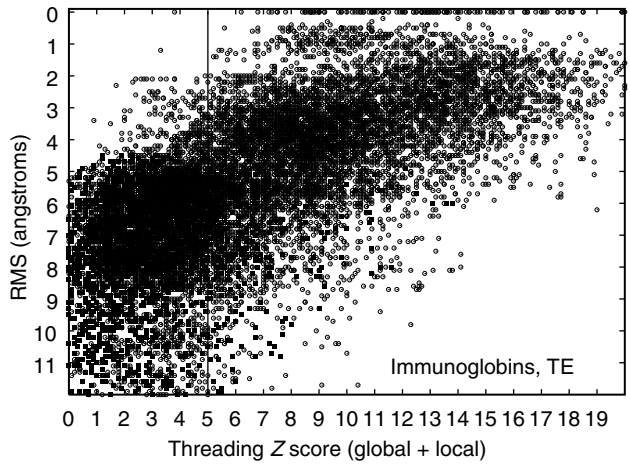


(b)

Figure 9. Comparison of family recognition by THOM2 and pair energies. The results of THOM2 for families of globins, immunoglobins (Fv fragments), and POU-like domains are compared to the results of Tobi–Elber (TE) pairwise potential. TE potential was optimized using LP protocol (with different target function) and the same training set. The first iteration of the so-called frozen environment approximation is performed to obtain approximate alignments for the TE potential. Parts a–f show the joint histograms of the sum of Z scores for local and global threading alignments versus the RMS deviations between superimposed (according to structure-to-structure alignments; see text for details) side-chain centers. Parts a, c, and e show the results for THOM2 (for globins, immunoglobins, and POU-like domains, respectively), whereas parts b, d, and f show the corresponding results for TE potential and the frozen environment approximation. The population of matches that are difficult to identify by pairwise sequence-to-sequence alignments is represented by the filled squares (see text for details). Note that the number of low THOM2 Z scores (for example, smaller than 5) is, on the average, smaller for families of globins and POU-like proteins. This is further highlighted in parts g and h, which show one-dimensional histograms of the sum of Z scores for local and global threading alignments for globins and POU-like domains. On the other hand, the TE potential and FEA perform better for immunoglobins family, which is also easier for sequence alignment methods (see text for details).



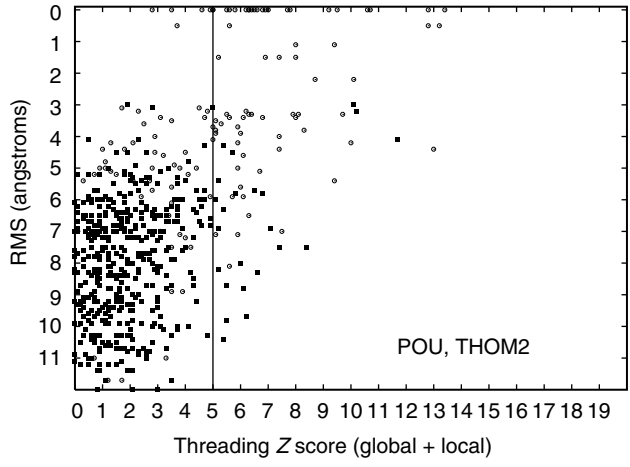
(c)



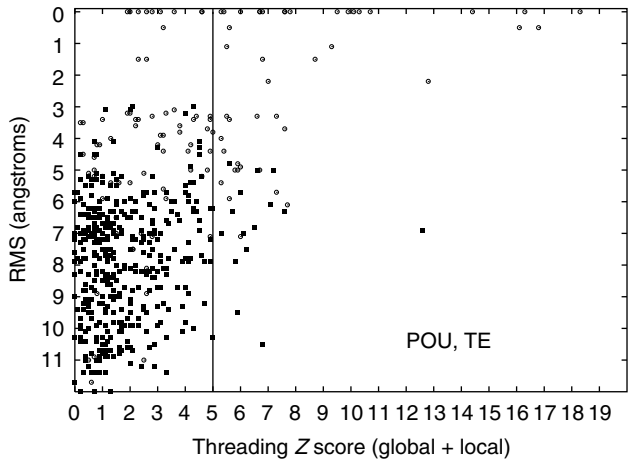
(d)

Figure 9 (*Continued*)

Nearly all pairs differing by less than 3 Å RMSD can be identified by THOM2 threading alignments. Most of the matches in the range between 3 and 5 Å can be still identified with high confidence. However, the number of confident matches (to the right with respect to vertical lines representing our cutoff of 5 in terms of sum of local and global *Z* scores) quickly decreases with the growing RMS distance. Essentially all the pairs with RMSD smaller than 3 Å can be also identified by pairwise sequence alignments. Below this threshold,



(e)

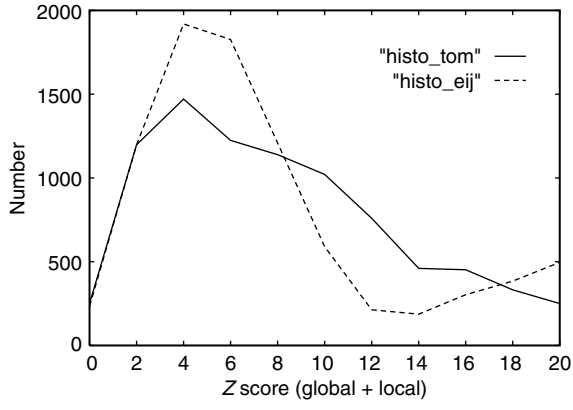


(f)

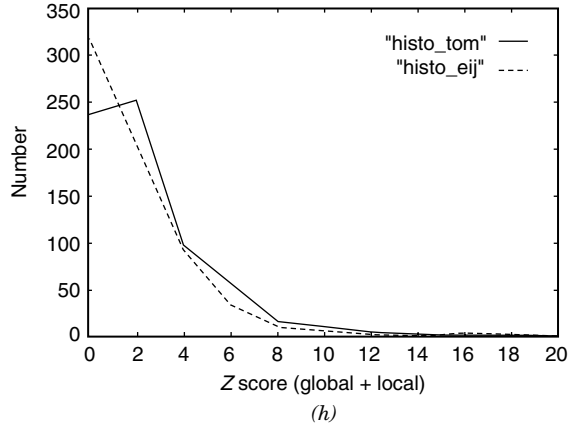
Figure 9 (*Continued*)

however, we observe many matches that can be still identified by threading but not by sequence alignment (filled rectangles corresponding to threading Z score higher than 5).

On the other hand, there are many matches due to the sequence alignment that are not detected by threading. Because we do not incorporate family profiles in our threading protocol, we do not include here a systematic comparison with the results of PsiBLAST [59]. However, we found examples



(g)



(h)

Figure 9 (*Continued*)

of matches detected with high confidence by threading and not detected by PsiBLAST in each of the families considered here (e.g, globins 1flp and 1ash or POU-like proteins 1akh and 1mbg).

Note that for the families of globins and POU-like domains the number of low THOM2 Z scores (for example smaller than 5) is, on the average, smaller than the number of low Z scores obtained with the TE potential and FEA. This is further highlighted in Figs. 9g and 9h showing one-dimensional histograms for the sum of Z scores for local and global threading alignments for globins and POU-like domains. For example, the number of low confidence matches ($Z < 5$) for globins increases from 2851 in case of THOM2 to 3265 in case of the TE potential. One can also notice that the distribution of Z

scores is different, with many very high Z scores for alignments into very close homologs as opposed to lower scores for more divergent pairs, in case of the TE potential.

Interestingly, FEA with the TE potential fails also for a larger number of native alignments. This is especially clear for the family of DNA binding proteins (see Figs. 9e and 9f). The number of native alignments with very low Z scores (smaller than 4) is equal to 7 in case of pairwise model and only 2 in case of THOM2. Because DNA binding proteins may be stabilized by contacts that are not included in our model, the energies of native alignments are quite poor. One striking example is the 1hdp. According to TE potential, 1hdp has the native energy equal to -0.42. An alternative alignment into its native structure with one insertion and one deletion in the sequence improves the energy to -0.63 despite the cost of gaps. On the other hand, THOM2 model seems to be capable of compensating for that using the information about the shape of the protein as encoded in the contact (solvation) shell characterization of each contact. The THOM2 native alignment for 1hdp is the lowest in energy and leads to higher Z scores.

The relatively worse performance of the pairwise model may result from the suboptimality of alignments that we generate using FEA, especially that our gap penalties for the TE potential were not optimized by LP protocol and we did not attempt to converge the FEA until self-consistency is achieved. However, as discussed above, in many instances it is clear that even better gap penalties will not be able to improve the observed scores. The specific functional form of our new profile model contributes to the relatively better performance too.

On the other hand, there are families for which the pairwise model works better. As can be seen from Figs. 9c and 9d, one such example is the family of immunoglobins. The FEA is expected to perform well when the sequence similarity is sufficiently high, because the information about the native sequences is used to generate optimal alignments. The divergence in terms of what can be detected by sequence similarity is larger for globins and POU-like proteins than for immunoglobins. For example, contrary to other families considered here, all the immunoglobins with RMSD smaller than 4 Å can be detected by sequence alignments. Therefore, good performance of the FEA with the TE potential is expected in this case.

VIII. CONCLUSIONS AND FINAL REMARKS

In the present chapter we proposed and applied an automated procedure for the design of threading potentials. The strength of the procedure, which is based on linear programming tools, is the automation and the ability of continuous exact learning. The LP protocol was used to evaluate different energy functions for accuracy and recognition capacity. Keeping in mind the necessity for efficient

threading algorithms with gaps, we selected the THOM2 model as our best choice.

Statistical filters based on local and global Z scores were outlined. We observe that, while using very conservative Z scores that essentially exclude false positives, the new protocol recognizes correctly (without any information about sequences) most of the family members with the RMS distance between the superimposed side chain centers of up to 4 Å. We also observe many instances of successful recognition of family members that cannot be confidently recognized by pair energies with the so-called frozen environment approximation.

The present approach is based on fitness of sequences into structures. Nevertheless, it is easily extendable to include also sequence similarity, family profiles, secondary structures, and other relevant signals. Because the THOM2 model provides an effective and comparable in performance alternative to pairwise potentials, it can be used as a fast component of fold recognition methods employing pair energies. It is the target of a future work.

The algorithms and threading potentials presented in this chapter are available in the program LOOPP (Learning, Observing, and Outputting Protein Patterns). The program (including the source code and sets of proteins for training and recognition) is available from the web [48]. It is also possible to submit sequences directly to our server.

Acknowledgments

This research was supported by an NIH NCRR grant to the Cornell Theory Center (acting director Ron Elber) for the developments of Computational Biology Tools. It was further supported by a seed grant from DARPA to Ron Elber. Jaroslaw Meller acknowledges also partial support from the Polish State Committee for Scientific Research (Grant 6 P04A 066 14).

References

1. J. U. Bowie, R. Luthy, and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).
2. D. T. Jones, W. R. Taylor, and J. M. Thornton, A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).
3. M. J. Sippl and S. Weitckus, Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins* **13**, 258–271 (1992).
4. A. Godzik, A. Kolinski, and J. Skolnick, Topology fingerprint approach to the inverse folding problem. *J. Mol. Biol.* **227**, 227–238 (1992).
5. C. Ouzounis, C. Sander, M. Scharf, and R. Schneider, Prediction of protein structure by evaluation of sequence–structure fitness. Aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.* **232**, 805–825 (1993).
6. S. H. Bryant and C. E. Lawrence, An empirical energy function for threading protein sequence through folding motif. *Proteins* **16**, 92–112 (1993).
7. Y. Matsuo and K. Nishikawa, Protein structural similarities predicted by a sequence–structure compatibility method. *Protein Sci.* **3**, 2055–2063 (1994).

8. L. A. Mirny and E. I. Shakhnovich, Protein structure prediction by threading. Why it works and why it does not. *J. Mol. Biol*. **283**, 507–526 (1998).
9. D. T. Jones, GenTHREADER An efficient and reliable protein fold recognition method for genomic sequenecs. *J. Mol. Biol*. **287**, 797–815 (1999).
10. A. R. Panchenko, A. Marchler-Bauer, and S. H. Bryant, Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol*. **296**, 1319–1331 (2000).
11. M. J. E. Sternberg, P. A. Bates, L. A. Kelley, and R. M. MacCallum, Progress in protein structure prediction: Assessment of CASP3. *Curr. Opin. Struct. Biol*. **9**, 368–373 (1999).
12. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, A united-residue force field for off-lattice protein structure simulations: Functional forms and parameters of long range side chain interaction potentials from protein crystal data. *J. Comp. Chem*. **18**, 849–873 (1997).
13. Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, *Ab initio* construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol*. **300**, 171–185 (2000).
14. A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler, Neural networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold. Des.* **2**, 261–269 (1997).
15. A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler, Exploring protein sequence space using knowledge based potentials. *J. Comp. Biol*. submitted, Santa Fe Institute preprint 98-11–103 (1999).
16. S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol*. **48**, 443–453 (1970).
17. T. F. Smith and M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol*. **147**, 195–197 (1981).
18. M. S. Johnson, J. P. Overington, and T. L. Blundell, Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol*. **231**, 735–752 (1993).
19. H. T. Croman, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 1985, Chapter 16.
20. R. H. Lathrop and T. F. Smith, Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol*. **255**, 641–665 (1996).
21. R. H. Lathrop, The protein threading problem with sequence amino-acid interaction preferences is NP-complete. *Protein Eng*. **7**, 1059–1068 (1994).
22. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, The statistical mechanical basis of sequence alignment algorithms for protein structure prediction, in *Recent Developments in Theoretical Studies of Proteins*, Ron Elber, ed., World Scientific, Singapore, 1996, Chapter 6.
23. S. H. Bryant, Evaluation of threading specificity and accuracy. *Proteins* **26**, 172–185 (1996).
24. A. Elofsson, D. Fischer, D. W. Rice, S. Le Grand, and D. Eisenberg, A study of combined structure–sequence profiles. *Fold. Des.* **1**, 451–461 (1998).
25. B. Rost, R. Schneider, and C. Sander, Protein fold recognition by prediction-based threading. *J. Mol. Biol*. **270**, 471–480 (1997).
26. K. T. Simons, I. Ruczinski, E. Huang, and D. Baker, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82–95 (1999).
27. A. R. Ortiz, A. Kolinski, and J. Skolnick, Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol*. **277**, 419–448 (1998).

28. J. Park, K. Karplus, C. Barret, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia, Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*. **284**, 1201–1210 (1998).
29. I. Bahar and R. L. Jerningan, Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol*. **266**, 195–214 (1997).
30. D. Tobi and R. Elber, Distance-dependent pair potential for protein folding: Results from linear optimization. *Proteins: Struct. Funct. Genet*. **41**, 40–46 (2000).
31. V. N. Maiorov and G. M. Crippen, Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol*. **227**, 876–888 (1992).
32. D. Tobi, G. Shafran, N. Linial, and R. Elber, On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40**, 71–85 (2000).
33. M. Vendruscolo and E. Domany, Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys*. **109**, 11101–11108 (1998).
34. S. Miyazawa and R. L. Jernigan, 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 (1996).
35. A. Godzik, A. Kolinski, and J. Skolnick, Knowledge-based potentials for protein folding: What can we learn from protein structures? *Proteins: Struct Funct Genet* **4**, 363–366 (1996).
36. D. A. Hinds and M. Levitt, A lattice model for protein structure prediction at low resolution. *Proc Natl. Acad. Sci. USA* **89**, 2536–2540 (1992).
37. M. R. Betancourt and D. Thirumalai, Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*. **2**, 361–369 (1999).
38. A. Godzik, A. Kolinski, and J. Skolnick, Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci*. **4**, 2107–2117 (1995).
39. J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J. Phys. Chem.* **93**, 2902–6915 (1989).
40. D. K. Klimov and D. Thirumalai, Linking rates of folding in lattice models of proteins with underlining thermodynamic characteristics. *J. Chem. Phys*. **109**, 4119–4125 (1998).
41. W. R. Taylor and R. E. Munro, Multiple sequence threading: conditional gap placement. *Fold. Des.* **2**, S33–S39 (1997).
42. G. N'emethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G.. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, Energy parameters in polypeptides. *J. Phys. Chem.* **96**, 6472–6484 (1992).
43. D. A. Hinds and M. Levitt, Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol*. **243**, 668–682 (1994).
44. L. Holm and C. Sander, The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res*. **22**, 3600–3609 (1994).
45. J. Meller and R. Elber, to be published.
46. CASP3. Third community wide experiment on the critical assessment of techniques for protein structure prediction. Asilomar, USA, 1998, http://Predictioncenter.lnl.gov/casp3
47. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne: The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
48. J. Meller and R. Elber, Learning, Observing and Outputting Protein Patterns (LOOPP)—a program for protein recognition and design of folding potentials, http://www.tc.cornell.edu/CBIO/loopp

49. C. S. Meszaros, Fast Cholesky factorization for interior point methods for linear programming. *Comput. Math. Appl*. **31**, 49–51 (1996).
50. I. Adler and R. D. C. Monteiro, Limiting behavior of the affine scaling continuous trajectories for linear programming problems. *Math. Program*. **50**, 29–51 (1991).
51. H. S. Chan and K. A. Dill, Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins: Struct. Funct. Genet.* **30**, 2–33 (1998).
52. S. H. Bryant and S. F. Altschul, Statistics of sequence–structure threading. *Curr. Opin. Struct. Biol*. **5**, 236–244 (1995).
53. W. M. Fitch, Random sequences. *J. Mol. Biol*. **163**, 171–176 (1983).
54. S. F. Altschul and B. W. Erickson, Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol*. **2**, 526–538 (1985).
55. E. J. Gambel, *Statistics of Extremes*, Columbia University Press, New York, 1958.
56. S. Karlin and S. F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268 (1990).
57. W. R. Pearson and D. J. Lipman, Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448; W. R. Pearson, Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol*. **276**, 71–84 (1998).
58. D. Fischer, A. Elofsson, D. Rice, and D. Eisenberg, Assesing the performance of fold recognition methods by means of a comprehensive benchmark, in *Pacific Symposium on Biocomputing, Hawaii*, 1996, pp. 300–318.
59. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acid Res*. **25**, 3389–3402 (1997).
60. J. F. Gibrat, T. Madej, and S. H. Bryant, Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*. **6**, 377–385 (1996).
61. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247**, 536–540 (1995).