PROTEINS: Structure, Function, and Genetics 45:241–261 (2001)

Linear Programming Optimization and a Double Statistical Filter for Protein Threading Protocols

Jaroslaw Meller[1,2] and Ron Elber[1]*
[1]*Department of Computer Science, Cornell University, Ithaca, New York*
[2]*Department of Computer Methods, Nicholas Copernicus University, Torun, Poland*

ABSTRACT **The design of scoring functions (or potentials) for threading, differentiating native-like from non-native structures with a limited computational cost, is an active field of research. We revisit two widely used families of threading potentials: the pairwise and profile models. To design optimal scoring functions we use linear programming (LP). The LP protocol makes it possible to measure the difficulty of a particular training set in conjunction with a specific form of the scoring function. Gapless threading demonstrates that pair potentials have larger prediction capacity compared with profile energies. However, alignments with gaps are easier to compute with profile potentials. We therefore search and propose a new profile model with comparable prediction capacity to contact potentials. A protocol to determine optimal energy parameters for gaps, using LP, is also presented. A statistical test, based on a combination of local and global *Z*-scores, is employed to filter out false-positives. Extensive tests of the new protocol are presented. The new model provides an efficient alternative for threading with pair energies, maintaining comparable accuracy. The code, databases, and a prediction server are available at http://www.tc.cornell.edu/CBIO/loopp. Proteins 2001;45:241–261.**

Key words: linear programming; potential optimization; decoy structures; threading; gaps

INTRODUCTION

The threading approach[1–8] to protein recognition is an alternative to the sequence-to-sequence alignment. Rather than matching the unknown sequence S_i to another sequence S_j (one-dimensional [1D] matching), we match the sequence S_i to a shape $\mathbf{X}_j$ (three-dimensional [3D] matching). Experiments found a limited set of folds compared with a large diversity of sequences, suggesting the use of structures to find remote similarities between proteins. Thus, the determination of overall folds is reduced to tests of sequence fitness into known and limited number of shapes.

Sequence-to-structure compatibility is commonly evaluated using reduced representations of protein structures. Points in 3D space represent amino acid residues, and an effective energy of a protein is defined as a sum of interresidue interactions. The effective pair energies can be derived from the analysis of contacts in known structures. Such knowledge-based pairwise potentials are widely used 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,[1,5,16] taking the form of a sum of individual site contributions, depending on the structural environment of a site. For example, the solvation/burial state or the secondary structure can be used to characterize different local environments. The advantage of profile models is the simplicity of finding optimal alignments with gaps (deletions and insertions into the aligned sequence) that permit identification of homologous proteins of different length. Using the dynamic programming (DP) algorithm,[17–20] optimal alignments with gaps in polynomial time can be computed, as compared with the exponential number of all possible alignments.

In contrast to profile models, the potentials based on pair energies do not lead to exact alignments with dynamic programming. A number of heuristic algorithms, providing approximate alignments, have been proposed.[21] However, they cannot guarantee an optimal solution with a less than exponential number of operations.[22] Another common approach is to approximate the energy by a profile model, the so-called frozen environment approximation (FEA), and to perform the alignment using DP.[23]

In this article, we evaluate several different scoring functions for sequence-to-structure alignments, with parameters optimized by linear programming (LP).[24–26] The capacity of the energies is explored in terms of a number of protein shapes that are recognized with a certain number of parameters. We propose a novel profile model, designed to mimic pair energies, which is shown to have accuracy comparable to that of other contact models. We discuss gap energies and introduce a double *Z*-score measure (from global and local alignments) to assess the results. The proposed threading protocol emphasizes structural fitness (as opposed to sequence similarity) and can be made a part of more complex fold recognition algorithms that use

Grant sponsor: National Institutes of Health; Grant sponsor: DARPA; Grant sponsor: Polish State Committee for Scientific Research; Grant number: 6-P04A-06614).

*Correspondence to: Ron Elber, Department of Computer Science, Cornell University, Upson Hall 4130, Ithaca, NY 14853. E-mail: ron@cs.cornell.edu

Received 11 December 2001; Accepted 4 June 2001

TABLE I. Definitions of Different Groups of Amino Acids Used in the Present Study*

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

*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. The amino acid types are HYD POL CHG CHN GLY ALA PRO TYR TRP 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.

family profiles, secondary structures, and other patterns relevant for protein recognition.

THEORY AND COMPUTATIONAL PROTOCOLS

Functional Form of the Energy

In this section, we formally define the families of pairwise and profile models. We also introduce a novel threading onion model (THOM), which is investigated in subsequent sections. In the widely used pairwise contact model, the score of the alignment of a sequence S into a structure $\mathbf{X}$ is a sum of all pairs of interacting amino acids:

$$E_{\text{pairs}} = \sum_{i<j} \phi_{ij}(\alpha_i, \beta_j, r_{ij}) \tag{1}$$

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

Let us consider the widely used contact potential. If the geometric centers of the side-chains are closer than 6.4 Å; the two amino acids are then considered in contact. The total energy is a sum of the individual contact energies:

$$\phi_{ij}(\alpha_i, \beta_j, r_{ij}) = \left\{ \begin{array}{cc} \varepsilon_{\alpha\beta} & 1.0 < r_{ij} < 6.4\,\text{Å} \\ 0 & \text{otherwise} \end{array} \right\} \tag{2}$$

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

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:

$$E_{\text{profile}} = \sum_i \phi_i(\alpha_i, \mathbf{X}) \tag{3}$$

As previously, α_i denotes the type of an amino acid that was placed at site i of $\mathbf{X}$. For example, if α_i is a hydrophobic residue, and x_i is characterized as a hydrophobic site, the energy $\phi_i(\alpha_i, \mathbf{X})$ will be low (high score). If α_i is charged, 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. In threading onion model 1 (THOM1), which was used in the past as an effective solvation potential,[1,2] the total energy of the protein is a direct sum of the contributions from structural sites and can be written as

$$E_{\text{THOM1}} = \sum_i \varepsilon_{\alpha_i}(n_i) \tag{4}$$

The energy of a site depends on two indices: (1) the number of neighbors to the site—n_i (a neighbor is defined by a cutoff distance—eq. 2); and (2) 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 comparable to that of the detailed pairwise model.

THOM1 provides a nonspecific interaction energy, which has relatively low prediction ability as compared with pairwise interaction models (see section, Application to Potential Design and Analysis). Threading onion model 2 (THOM2) attempts to improve the accuracy of the environment model, making it more similar to pairwise interactions.

We 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 neighbors to site j. The type of amino acid at site i is α_i. Only one of the amino acids in contact is known. The total contribution to the energy of site i is a sum over all contacts to this site

$$\phi_{i,\text{THOM2}}(\alpha_i, \mathbf{X}) = {\sum_j}' \varepsilon_{\alpha_i}(n_i, n_j)$$

The prime indicates 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:

$$E_{\text{THOM2}} = \sum_i {\sum_j}' \varepsilon_{\alpha_i}(n_i, n_j) \tag{5}$$

It is possible to define an effective contact energy using THOM2:

$$V_{ij}^{\text{eff}} = \varepsilon_{\alpha_i}(n_i, n_j) + \varepsilon_{\alpha_j}(n_j, n_i) \tag{6}$$

TABLE II. Definitions of Contact Types for the THOM2 Energy Model*

Type of site	$n' = 1,2;\bar{1}$	$n' = 3,4,5,6;\bar{5}$	$n' \geq 7;\bar{9}$
$n = 1,2;\bar{1}$	$(\bar{1},\bar{1})$	$(\bar{1},\bar{5})$	$(\bar{1},\bar{9})$
$n = 3,4;\bar{3}$	$(\bar{3},\bar{1})$	$(\bar{3},\bar{5})$	$(\bar{3},\bar{9})$
$n = 5,6;\bar{5}$	$(\bar{5},\bar{1})$	$(\bar{5},\bar{5})$	$(\bar{5},\bar{9})$
$n = 7,8;\bar{7}$	$(\bar{7},\bar{1})$	$(\bar{7},\bar{5})$	$(\bar{7},\bar{9})$
$n \geq 9;\bar{9}$	$(\bar{9},\bar{1})$	$(\bar{9},\bar{5})$	$(\bar{9},\bar{9})$

*There are 15 types of energy terms in THOM2 that are based on contacts in the first and the second contact layers. A contact between two amino acids is "on" if the distance is <6.4 Å. We consider five types of sites in the first layer and three in the second layer. Thus, there are $20 \times 15 = 300$ different energy terms for 20 different amino acids. 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.

Hence, we can formally express the THOM2 energy as a sum of pair energies

$$E_{\mathrm{THOM2}} = \sum_{i<j} V_{ij}^{\mathrm{eff}}$$

The effective energy mimics the formalism of pairwise interactions. However, in contrast to the usual pair potential, the optimal alignments with gaps can be computed efficiently with THOM2, as structural features alone determine the "identity" of the neighbor.

We use a coarse-grained model that leads 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 the section, Application to Potential Design and Analysis.

Linear Programming Protocol for Optimization of the Energy Parameters

Consider the alignment of a sequence S of length n, into a structure $\mathbf{X}$ of length m. In order to optimize the energy parameters for the amino acid interactions (the gap energies are discussed in the section, Protocol for Optimization of Gap Energies), we employ the so-called gapless threading, in which sequence S is fitted into the structure $\mathbf{X}$ with no deletions or insertions. Hence, the length of the sequence must be shorter than, or equal to, the length of the protein chain. If n is shorter than m, we may try $m - n + 1$ possible alignments varying the structural site of the first residue (and the following sequence).

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. The energy function $E(S, \mathbf{X}, \mathbf{p})$ depends on a vector $\mathbf{p}$ of q parameters (thus far undetermined).

Consider the sets of structures $\{\mathbf{X}_i\}$ and sequences $\{S_j\}$. There is an 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. Let $\mathbf{X}_n$ be the native structure. A condition for an exact recognition is

$$E(S_n, \mathbf{X}_j, \mathbf{p}) - E(S_n, \mathbf{X}_n, \mathbf{p}) > 0 \qquad \forall\ j \neq n \tag{7}$$

In the set of inequalities (Eq. 7), the coordinates and sequences are given, and the unknowns are the parameters we need to determine.

The LP protocol makes it possible to measure the difficulty of a training set. The number of parameters of the energy function necessary to satisfy all the inequalities is derived from the set of structures, as defined in eq. 7. Whereas the statistical potentials are based on the analysis of native structures only, the LP protocol is using sequences threaded through misfolded structures during the process of learning. As a result, the LP has the potential for accumulating more information, in an attempt to place the energies of the misfolded sequence as far as possible from the energy of the native state. In fact, the LP protocol can be used to optimize the Z-score of the distribution of energy gaps.[27] In the remainder of this section, we describe the technique to solve the inequalities of eq. 7.

The "profile" and pairwise interaction models considered in this work can be written as a scalar product:

$$E = \sum_{\gamma} n_{\gamma}(\mathbf{X}) p_{\gamma} \equiv \mathbf{n}(\mathbf{X}) \cdot \mathbf{p} \tag{8}$$

where $\mathbf{p}$ is the vector of parameters 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 denotes the types of the amino acid pairs, whereas in THOM1, it denotes the types of sites characterized by the identity of the amino acid at the site and the number of its neighbors. $n_{\gamma}(\mathbf{X})$ is the number of contacts, or sites of a specific type found in the structure $\mathbf{X}$.

Using the representation of eq. 8, we may rewrite eq. 7 as follows:

$$\mathbf{p} \cdot \Delta\mathbf{n}_j = \sum_{\gamma} p_{\gamma}[n_{\gamma}(\mathbf{X}_j) - n_{\gamma}(\mathbf{X}_n)] > 0 \qquad \forall\ j \neq n \tag{9}$$

Standard linear programming tools can solve Eq. 9 for $\mathbf{p}$. We use the BPMPD program of Meszaros,[28] which is based on the interior point algorithm. In the present computations, we seek a point in parameter space that satisfies the constraints, and we do not optimize a function in that space. Without a function to optimize the interior point, the algorithm places the solution at the "maximally feasible" point, which is at the analytic center of the feasible polyhedron that defines the "accessible" volume of parameters.[27,29]

The set of inequalities that we wish to solve includes tens of millions of constraints that could not be loaded directly into the computer memory (we have access to machines with 2–4 Gigabytes [Gb] of memory). Therefore,

the following heuristic approach was used. Only a subset of the constraints is considered:

$$\{\mathbf{p}\cdot\Delta\mathbf{n} < C\}_{j=1}^{J}$$

where the threshold of C is chosen to restrict the number of inequalities to a manageable size (~500,000 inequalities for 200 parameters). Hence, during a single iteration, we considered only the inequalities that are more likely to be relevant for further improvement by being smaller than cutoff C. This subset 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. 9). If no violations are detected, the process is stopped (a solution was found). If negative inner products are found in the remaining set, a new subset of inequalities below C is collected. The process is repeated, until it converges. Sometimes convergence is difficult to achieve, necessitating human intervention in the choices of the inequalities. For example, mixing subsets of inequalities from previous runs with the lowest inequalities obtained with the new parameters helps avoid the problem of "oscillating" between solutions.

Protocol for Optimization of Gap Energies

In this section, we discuss the derivation of the energy terms for gaps and deletions that enable the detection of homologs. 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 placed at a virtual structural site).

The gap residue, —, is considered another amino acid. We assign to it a score (or energy), $\varepsilon(\underline{\mathbf{X}})$, according to its environment. Gap training is similar to the training of other amino acid residues, in contrast to the usual ad hoc treatment of gap energies. The proposed treatment is also more symmetric than the different penalties for opening and extending gaps.

The database of "native" and decoy structures is different, however, for gapless and gap training. 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 sequence S_n from the coordinate training set and by replacing it with a homologous structure, $\mathbf{X}_h$. The best alignment of the native sequence, S_n, into the homologous structure, $\mathbf{X}_h$, with an initial guess of gap penalties, defines $\bar{S}_n$. The extended sequence, $\bar{S}_n$, with gap residues at certain (fixed from this point on) positions becomes our new (pseudo-) native sequence of the structure $\mathbf{X}_h$.

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

$$E(\bar{S}_n, \mathbf{X}_j, \mathbf{p}) - E(\bar{S}_n, \mathbf{X}_h, \mathbf{p}) > 0 \qquad \forall\ j \neq h, n \qquad (10)$$

Equation 10 is different from eq. 7 because we consider the extended set of amino acids, $\bar{S}$ instead of S, and the native-like structure is $\mathbf{X}_h$ instead of $\mathbf{X}_n$.

To limit the scope of the computations, we optimize the scores of the gaps only. Thus, we do not allow the amino acid energies optimized separately (see the section, Linear Programming Training of "Minimal" Models) to change while optimizing parameters for gaps. Moreover, the sequence $\bar{S}$, obtained by a certain prior (e.g., structure-to-structure) alignment, or from the experimental data, if available, is held fixed. In other words, threading of the extended sequence with fixed positions and number of gap residues (treated now as any other residue), $\bar{S}$, against all other structures in the training set is used, in order to generate a corresponding set of inequalities, (eq. 10). This optimization, although limited, and clearly not the final word on the topic, is still expected to be better than a guess. Further studies of gap penalties are in progress (T. Galor, J. Meller, and R. Elber, unpublished data). Optimization of gaps has been attempted in the past.[23,30]

In principle, one could optimize deletion penalties using a similar protocol. In this article, 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 deletion penalty is set equal to the cost of insertion averaged over the two nearest structural sites. No explicit dependence on the amino acid type is assumed.

Double *Z*-Score Filter for Gapped Alignments

In later sections on these assessments we 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[17]) and alignments of continuous fragments of sequences into continuous fragments of structures (local alignment[18]). 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.[26,31] Although it is our goal that the (free) energy function that we design is sufficiently accurate that the native state (the native sequence threaded through the native structure) is the lowest in energy, this is not always the case. Our exact training is for the training set and for gapless threading only (see the section, Application to Potential Design and Analysis). The results were not extended to include exact learning with gaps, or exact recognition of structures of related proteins that are not the native. Such extensions are difficult, as the number of inequalities for $\bar{S}$ is exponentially larger than the number of inequalities without gaps.

Other investigators use the Z-score as an additional filter or as the primary filter,[19,32,4,6] 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, in addition to the initial energy filter, provides improved discrimination as compared with a single Z-score test.

The Z-score is a dimensionless "normalized" score, defined as

$$Z = \frac{\langle E \rangle - E_p}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}} \quad (11)$$

The energy of the current "probe," i.e., 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. The Z-score measures the deviation of our "hits" from random alignments (alignments with scores far from the "random" average value are more significant). Following common practice,[32–34] we generate the distribution of random alignments numerically, employing sequence shuffling. That is, we consider the family of sequences obtained by permutations of the original sequence. The set of shuffled sequences has the same amino acid composition and length as the native sequence, and all the shuffled sequences have the same energy in the unfolded state (the energy of an amino acid with no contacts is set to zero).

In the section, Assessing the Distribution of Z-Scores for Gapped Alignments, we estimate numerically the probability $P(Z_p)$ of observing a Z-score of greater than Z_p by chance for local threading alignments. The relatively high likelihood of observing large Z-scores for false-positives makes predictions based on the Z-score test problematic. 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 a 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.[35] Nevertheless, using environment-dependent gap penalty, the Z-score of the global alignment proved a useful independent filter (see later sections on these assessments). We observe that good scores are obtained for length differences (between sequence and structure) that are on order of 10%. By contrast, when the differences in length are profound the global alignment fails. Large differences imply identification of domains and not a whole protein. This is a different problem, not addressed in the present work.

APPLICATION TO POTENTIAL DESIGN AND ANALYSIS

In this section, we analyze and compare several pairwise and profile potentials, optimized using the LP protocol. Given the training set, either we obtain a solution (exact 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). The larger the number of proteins that are recognized with the same number of parameters, the better the energy model. We find that, in general, the "profile" potentials have lower capacity than that of the pairwise interaction models.

Training and Test 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,[31] which 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. 7). 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.,[25] which we call the TE set. This set is considerably more demanding. It includes structures chosen according to diversity of protein folds, but also some homologous proteins ($\leq$60% sequence identity), and poses a significant challenge to the energy function. For example, the set is infeasible for threading onion model 1 (THOM1), even when using 20 types of amino acids (see the next section). 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 testing sets to evaluate the new potentials in terms of both gapped and gapless alignments. These test sets contain proteins that are structurally dissimilar to the proteins included in the training sets, specified by the root-mean-square deviation (RMS) between the structures. A structure-to-structure alignment algorithm, based on the overlap of the contact shells defined for the superimposed side-chain centers in analogy with THOM2 (disregarding however the identities of amino acids), was used (J. Meller and R. Elber, unpublished results).

The first testing set, referred to as S47, consists of 47 proteins: 25 proteins from the CASP3 competition[36] and 22 related structures, chosen randomly from the list of DALI[37] relatives of the CASP3 targets. Using CASP3-related structures is a convenient way of finding protein structures that are not sampled in the training. None of the 47 structures has homologous counterparts in the HL set, and only six (representing three different folds) have counterparts in the TE set, with a cutoff for structural (dis)-similarity of 12 Å RMS (between the superimposed side-chain centers).

The second test set, referred to as S1082, consists of 1,082 proteins that were not included in the TE set and that are different by $\geq$3 Å RMS (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 at $\leq$3 Å RMS) sample of protein families. The training and testing sets are available from the web.[38]

TABLE III. Comparing the Capacity of Different Threading Potentials*

Potential	Hinds–Levitt set	Tobi–Elber set
Pairwise, HP model, par. free	200	456
Pairwise, 10 aa, 55 par	246[a]	504
Pairwise, 20 aa, 210 par	246[a]	530
Pairwise, 20 aa, 210 par	237	594[a]
THOM1, 20 aa, 200 par	246[a]	474
THOM2, 10 aa, 150 par	246[a]	478
THOM2, 20 aa, 300 par	246[a]	428
THOM2, 20 aa, 300 par	236	594[a]

*Capacity for recognition of pairwise and profile threading potentials measured by gapless threading on Hinds–Levitt (HL) and Tobi–Elber (TE) representative sets of proteins (see the section, Training and Test Sets). Threading onion model 1 (THOM1) performs significantly worse than pairwise potentials. THOM2 shows a comparable performance and is able to learn the TE set (see also Table X). For each potential, the number of amino acid (aa) types used and the resulting number of parameters are reported. The number of correct predictions for structures in HL and TE sets is given in the second and third columns, respectively.
[a]The training set used (either HL or TE).

Linear Programming Training of "Minimal" Models

This section addresses the question: What is the minimal number of parameters required to obtain an exact solution for the HL and for the TE sets? 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. 7, for all sequences S_n and structures $\mathbf{X}_j$, are satisfied.

The pairwise model requires the smallest number of parameters (i.e., 55) to solve the HL set exactly (Table III). Only 10 types of amino acids were required: HYD POL CHG CHN GLY ALA PRO TYR TRP HIS (see also Table I). THOM1 and THOM2 require 200 and 150 parameters, respectively, to provide an exact solution on the same (HL) set (Table III). It is impossible to find an exact potential (of the functional forms we examined) for the HL set without (at least) 10 types of amino acids. The potentials optimized on the HL set are then used to predict the folds of the proteins of the TE set. Again, we find that the pairwise interaction model performs better than threading onion models.

An indication that THOM2 is a better choice than THOM1 is included in the next comparison. It is impossible to find parameters that will solve the TE set exactly 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 lead to feasible inequalities if the number of parameters is 300 for THOM2 and 210 for the pairwise potential. The set of parameters that solved the TE set exactly does not solve exactly the HL set, as the latter set includes proteins not included in the TE set.

We have also attempted to solve the TE set using pair energies and THOM2 with a smaller number of parameters. The problem proved 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.

Analysis of THOM2

As discussed earlier, in the section, Theory and Computational Models, the THOM1 potential provides a new set of parameters for an effective solvation model that was used in the past. Because in applying the LP protocol we can only solve the HL set, the solution for that set gives our optimal THOM1 energies, as included in Table IVA. In this section, we analyze THOM2 in detail, which has significantly higher capacity than THOM1. However, the double layer of neighbors makes the results more difficult to understand.

Figure 1 presents a contour plot of the total contributions of different types of contacts to the native energies of the native alignments in the TE set. The plots show the energy contributions as a function of the number of neighbors of the primary site (with known amino acid identity) and the number of contacts to a secondary site, n', respectively. The results for two types of residues, lysine and valine, are presented. The contribution of a given type of contact is defined as $f \cdot \varepsilon_\alpha(n, n')$, where $\varepsilon_\alpha(n, n')$ is the energy of a given type of contact, and f is the frequency of that contact, observed in the TE set.

It is possible to find a very attractive (or repulsive) site that makes only negligible contribution to the native energies, since it is extremely rare (i.e., f is small). Table V displays specific examples. 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.[26] It is of interest to check whether 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. 6 to define the energy of a contact. In the usual pairwise interaction, the energy of a valine–valine contact is a constant and is independent of other contacts that the valine may have.

Table VI lists 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 to very rare type of contact. The plausible interpretation is that these rare contacts are used to enhance recognition in some cases, due to specific "homologous features." Significant differences are observed also for the frequently occurring types of contacts that contribute in accord with the "general principle" of rewarding contacts between hydrophobic sites. For example, the effective energies of contacts between valine of five neighbors with

TABLE IV. Parameters of Some Threading Potentials Trained Using the LP Protocol*

A: THOM1[a]

	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	3.38	−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	10.00	10.00	−0.78	10.00	0.71

B. THOM2[b]

	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 threading onion model 1 (THOM1) potential trained on the Hinds–Levitt (HL) set of proteins.
[b]Numerical values of the energy parameters for threading onion model 2 (THOM2) potential trained on the Tobi–Elber (TE) set of proteins.
*Rows correspond to either different types of sites (THOM1) or contacts (THOM2). Columns correspond to different types of amino acids. See text for details.

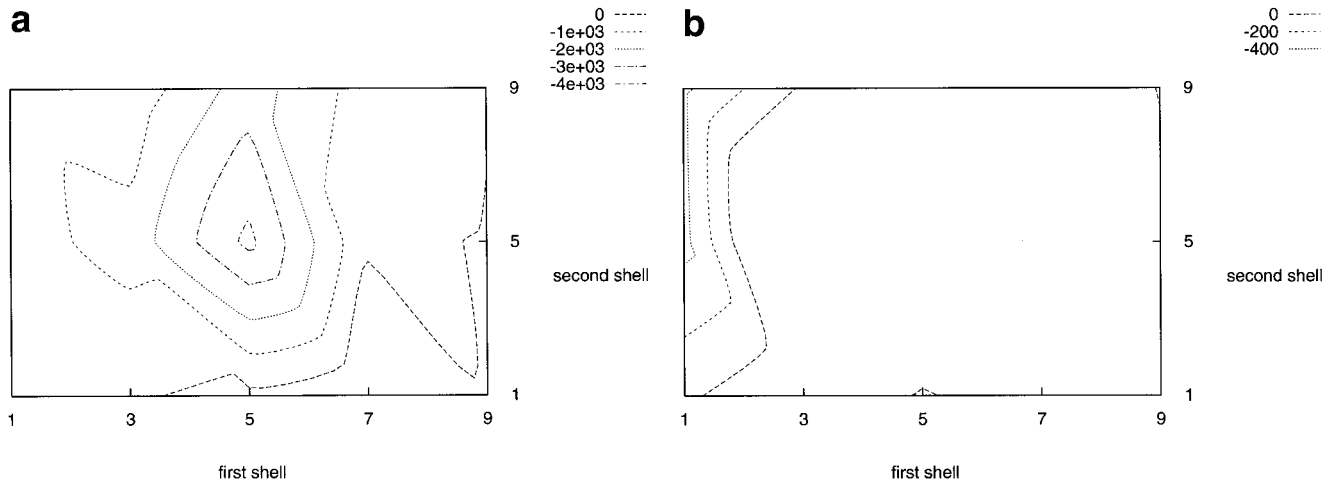

Fig. 1. Contour plots of the total energy contributions to the native alignments in the Tobi–Elber (TE) set for valine and lysine residues as a function of the number of neighbors in the first and second shells. **a:** Contacts involving valine residues with five to six neighbors with other residues of medium number of neighbors stabilize most of the native alignments. **b:** Only contacts involving lysine residues with a small number of neighbors stabilize native alignments.

another valine of three, five, or seven neighbors are equal to −0.44, −0.54, −0.61, respectively. Hence, THOM2 includes significant cooperativity effects. The optimal parameters for THOM2 potential are provided in Table IVB.

Training of Gap Energies

In this section, we apply the linear protocol for the optimization of gap energies described earlier. Training concerns the gap energies for THOM2 only, and it is limited to a small set of carefully chosen homologous pairs. Despite the limited scope of our training, we obtain satisfactory results in terms of recognition of remote homologues, as discussed subsequently.

Pairs of homologous proteins from the following families were considered: globins, trypsins, cytochromes, and lysozymes (Table VII). The families were selected to represent different folds. 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, $\bar{S}_n$ includes only one gap for

TABLE V. Characterization of Native and Decoy Structures*

A: THOM1[a]

Type of site[a]	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: THOM2[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)

*Overall site and contact distributions are split into distributions for hydrophobic and polar residues (as defined in Table I), given in the parentheses.

[a]Frequencies of different types of sites, relevant for training of threading model 1 (THOM1), found in the native structures of the Hinds–Levitt (HL) set, as opposed to decoy structures generated using the HL set. 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 the HL set and 556.14 millions of decoy sites generated using the HL set, respectively.

[b]Frequencies of different types of contacts, appropriate for training of threading onion model 2 (THOM2), found in the native structures of the Tobi–Elber (TE) set, as opposed to decoy structures generated using TE. 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 the TE set and 10089.19 millions of decoy contacts generated using the TE set, respectively.

TABLE VI. Cooperativity in Effective Pairwise Interactions of the THOM2 Potential*

A: VAL residues[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: LYS residues[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

*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) as 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 be in general asymmetric.

[a]Effective energies of contact between two VAL residues with a different number of neighbors.

[b]Effective energies of contacts between two LYS residues.

alignment of 1ccr (sequence) versus 1yea (structure), and 22 gaps for 1ntp versus 2gch. The structures of the lysozymes 1lz5 and 1lz6 include engineered insertions that allow us to sample experimentally observed gap locations.

For the remaining families, the process of generating pseudo-native sequences 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 using the THOM1 potential, with an initial guess for the gap energies, provided in Table VIIIA. The ad hoc gap penalties favor gaps at sites with few neighbors, and they satisfy the following constraints: (1) the gap penalty should increase with the number of neighbors; (2) the energy of a gap with n 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); and (3) the energy of amino acids without contacts is set to zero, and therefore the gap energy is greater than zero. Given these constraints, the initial gap penalties are tuned up to minimize the discrepancies with the DALI[37] structure-to-structure alignments (we choose not to use the DALI alignments directly, since they involve deletions that are not trained explicitly at this stage; see the section, Protocol for Optimization of Gap Energies.

The "pseudo-native" structures with extended sequences, obtained as described above, are added to the HL set (while removing the original native structures). The energy functional form we used for the gaps is the same as for other amino acids in THOM2. "Gapless" threading into other structures of the HL set generates about 200,000 constraints for the gap energies, which are solved using the LP solver. The resulting gap penalties for THOM2 are given in Table VIIIB. The value of 10 is the maximal penalty allowed by the optimization protocol we used. The maximal penalty is assigned to gaps found only in decoy states and that have no native states to bound the penalty at lower values. For example, using our initial guess for gap penalties, we do not observe gaps at the hydrophobic cores of pseudo-native structures. Gaps are usually found in loops with significant solvent exposure, and we have no information in our training set on "native" gaps in sites that are neighbor-rich.

Table IX presents the results of optimal threading with gaps (using dynamic programming) for myoglobin (1mba) against leghemoglobin (1lh2) structure. We show the

TABLE VII. Pairs of Homologous Structures Used for Training of Gap Penalties

Native[a]	Homologous[a]	Similarity[b]
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
1lzl (lysozyme, 130)	1lz5 (1lzl + 4 res insert, 134)	99%, 0.5 Å, 130 res
1lzl (lysozyme, 130)	1lz6 (1lzl + 8 res insert, 138)	99%, 0.3 Å, 129 res

[a]For each pair, the native and the homologous structures are specified by their Protein Data Bank (PDB) codes, names, and lengths, respectively.
[b]The similarity between the native and the homologous proteins is defined in terms of the sequence identity (%), root-mean-square (RMS) distance (Å), and length (number of residues), as defined by structure-to-structure alignments obtained by submitting the corresponding pairs to the DALI server.[37]

TABLE VIII. Gap Penalties for THOM2 Model as Trained by the LP Protocol*

A: THOM1[a]	
Type of site	Penalty
(0)	0.1
(1)	0.3
(2)	0.6
(3)	0.9
(4)	2.0
(5)	4.0
(6)	6.0
(7)	8.0
(8)	9.0
(9)	10.0
B: THOM2[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

*The limited set of homologous structures presented in Table VII is used.
[a]Initial guess of gap penalties for different types of sites in threading onion model 1 (THOM1).
[b]Optimized gap penalties for different types of contacts in threading onion model 2 (THOM2). Penalties that are not specified explicitly are equal to the maximum value of 10.0. Note that the training is limited and will be extended in a future work.

initial alignment (with the ad hoc gap parameters in Table VIIIA), defining the pseudo-native state, and the results for optimized gap penalties for THOM2. The location of gaps in the initial alignment is largely consistent with the DALI[37] structure-to-structure alignment. Four out of seven insertions coincide with the DALI superposition of the two structures, two insertions are shifted by three residues (see footnote to Table IX). The THOM2 alignment (different from the initial setup) is less consistent with the DALI alignment. Interestingly, however, it provides a better superposition of α-helices. The gaps appear (as expected) in loop regions (e.g., the CD, EF, and GH loops). An exception is the gap at position 9 (in 1lh2), located in the middle of the α-helix instead of position 2, as suggested by the DALI alignment. Further tests of alignments with gaps are presented in the section we present threading results for the pairwise TE potential (see the section, Tests of the Model).

To compute optimal alignments with the FEA, we need to set gap penalties for the TE potential. Pairwise models are not the focus of our study, and we do not attempt to optimize gap energies for the TE potential. Therefore, for the sake of fair comparison, we introduce ad hoc gap penalties based on a similar functional model, for both the TE and THOM2 potentials.

After some experimentation, the insertion penalties are chosen to be proportional to the number of neighbors to a site, $\varepsilon_{-}^{\mathrm{TE}}(n) = 0.2 \cdot (n + 1)$ and $\varepsilon_{-}^{\mathrm{THOM2}}(n) = 1.0 \cdot (\langle n \rangle + 1)$, for the TE and THOM2 potentials (the averaged number of neighbors, $\langle n \rangle$, in a class n belongs to, is used for THOM2; see Table II), respectively. This choice is consistent with the trained THOM2 gap energies, which also penalize sites of no neighbors. The proportionality coefficients were gauged using the same families 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 as described in the section, Protocol for Optimization of Gap Energies. For further comparisons with sequence-to-sequence alignments, we also introduce environment-dependent gap penalties that are used for family recognition in conjunction with the BLOSUM50[39] substitution matrix,

$$\varepsilon_{-}^{B50}(n) = (5 - n) - 8$$

(see the section, Assessing the Specificity of the Protocol).

Assessing the Distribution of *Z*-Scores for Gapped Alignments

In this section, we compute numerical distributions of the *Z*-scores for local and global threading alignments, using THOM2 and the gap penalties shown in Table VIIIB. On the basis of these distributions, we derive empirical cutoffs for the double *Z*-score test (discussed in the section, Double *Z*-Score Filter for Gapped Alignments) that filters out all the incorrect predictions observed in our benchmark. Further tests of the specificity, as well as sensitivity of the double *Z*-score filter, are included in the following sections.

TABLE IX. Comparison of Alignments of Myoglobin (1mba) Sequence into Leghemoglobin (1lhz) Structure*

A: THOM1[a]	
.....1.....2.....3.....4.....5.....	1–59
SLS**AAEADLAGKSWAPVFA**N**KNANGLDFLVALFEK**F**PDSANF**FADFKGKS**VADIK**ASP**K**	1mba
GAL**TESQAALVKSSWEEFNANIPKHTHRFFILVLEI**A**PAAKDL**FSFLKGTSEVPQN**NPE**	1lh2
.....1.....2.....3.....4.....5.....	1–59
6.....7.....8.....ii...9.....0.....1.....	60–116
LRDVSSRIFTRLNEFVNNAAN**AGKMSA--MLSQFAKEHVG**FGVG**SAQFENVRSMFPGFV**	1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTD**ATLKNLGSVHVSKGVADAHFPVVK**EA**ILKTI**	1lh2
6.....7.....8.....9.....0.....1.....	60–118
...2..i..i.....3.....4..i...i.i	117–146
ASVAAP-PA-G**ADAAWTKLFGLIIDALK-AA**G-A-	1mba
KEVVGAKW**SEELNSAWTIAYDELAIVIKKEMDDA**A	1lh2
.2.....3.....4.....5...	119–153
B: THOM2[b]	
.....i.1.....2.....3.....4.....i....i.i.	1–55
SLS**AAEAD-LAGKSWAPVFA**N**KNANGLDFLVALFEK**F**PDSANF**FADFKGK-S**VAD-I-K**	1mba
GAL**TESQAALVKSSWEEFNANIPKHTHRFFILVLEI**A**PAAKDL**FSFLKGTSEVPQN**NPE**	1lh2
.....1.....2.....3.....4.....5.....	1–59
....6.....7.....i.8.i.....9.....0.....1..	56–112
ASP**KLRDVSSRIFTRLNEFVN**NA-AN**A-GKMSAMLSQFAKEHVG**FGVG**SAQFENVRSMF**	1mba
LQAHAGKVFKLVYEAAIQLEVTGVVVTD**ATLKNLGSVHVSKGVADAHFPVVK**EA**ILKTI**	1lh2
6.....7.....8.....9.....0.....1.....	60–118
.....2.i.....3.....4.....	113–146
PGFVASVAA-PPAG**ADAAWTKLFGLIIDALKA**AGA	1mba
KEVVGAKW**SEELNSAWTIAYDELAIVIKKEMDDA**A	1lh2
.2.....3.....4.....5...	119–153

*The location of insertions in the initial alignment (which is used for training of gap energies) is largely consistent with the DALI structure-to-structure alignment,[37] which aligns: residues 2–50 of 1mba to 3–51 of 1lh2, residues 53–56 of 1mba to 52–55 of 1lh2 (implying deletions at positions 51 and 52 in 1mba), residues 59–80 of 1mba to 56–77 of 1lh2, residues 81–86 of 1mba to 82–87 of 1lh2, residues 87–121 of 1mba to 89–123 (with the implied insertion at position 88 in 1lh2), residues 122–139 of 1mba to 126–143 of 1lh2 (implying two insertions at positions 124 and 125 in 1lh2), and residues 140–145 of 1mba to 145–150 of 1lh2 (with an insertion at position 144 in 1lh2), respectively. α-Helices in both structures are marked in boldface. Note that F- and G-helices are shifted considerably in the DALI alignment (there is no counterpart for the D-helix in 1lh2). The initial THOM1 alignment 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 of table 9.B, is in perfect agreement with the DALI superposition for residues 10–50 of 1lh2 (including A-, B-, and C-helices) and then departs from the DALI alignment, overlapping E-, F-, and G-helices with a smaller shift.

[a]Threading onion model 1 (THOM1) alignment with the initial gap penalties.

[b]Threading onion model 2 (THOM2) alignment with trained gap penalties.

To establish a cutoff for the Z-score (and not the energy itself) that eliminates false-positives, we estimate numerically the probability $P(Z_p)$ of observing a Z-score larger than Z_p by chance. The distribution of Z-scores for random alignments is generated by threading sequences of the S47 set through structures included in the HL 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 the section, Training and Test Sets). Therefore, any significant hit in this set may be regarded as a false-positive. Z-scores of local alignments are employed to estimate $P(Z_p)$. The number of local alignments with "good" energies (significantly lower than zero) is large, underlying the need for an additional selection mechanism to eliminate false-positives.

In local alignments, a contribution due to a given contact should be only included if it belongs to the alignment (which is not known to start with). This implies a "structural" FEA (see also the section, Assessing Protein Family Signals and the Sensitivity of the Protocol). When counting contacts, we assume that the sites in contact in the original structure belong to the aligned part of the structure. This may result in spuriously low energies of local matches, making the Z-score of the local threading alignment an important filter.

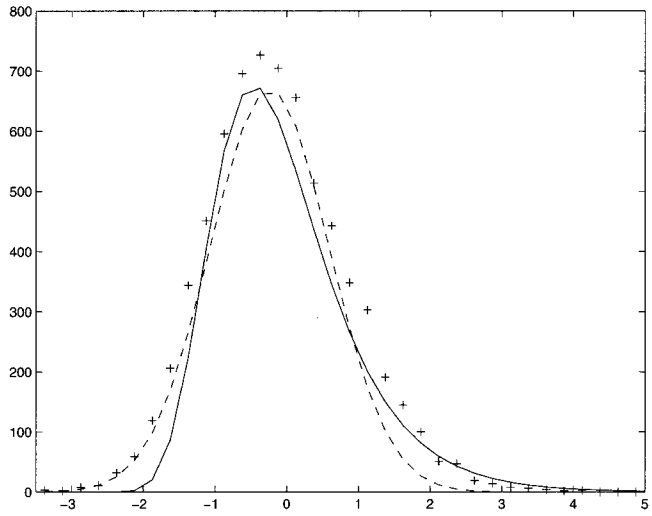

Fig. 2. Probability distribution function of the *Z*-scores computed with local threading alignments for the population of false-positives. A set of 47 sequences of proteins included in the S47 set is used to sample the distribution of the *Z*-scores for false-positives (proteins of the S47 set have no homologues in the Hinds–Levitt (HL) set; see text for details). Each of the sequences is aligned to all the structures included in the 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 for 6,813 local threading alignments is represented by +. Note the significant tail to the right, indicating a relatively high likelihood of observing false-positives with large *Z*-scores. The dotted line shows the attempted analytical fit to the gaussian distribution, whereas the solid line the attempted fit to the extreme value distribution (EVD). Note that actual distribution deviates significantly from both. According to the analytical fit to the EVD, the probability of observing a *Z*-score larger than Z_P by chance is equal to $P(Z_P) = 1 - \exp\{-\exp[-1.313 \cdot (Z_P + 0.466)]\}$ with the 98% confidence intervals: 1.313 ± 0.112 and 0.466 ± 0.079. For example, the probability of observing by chance a *Z*-score of >4 = 0.003. We emphasize, however, that the analytical fit to the extreme value distribution provides an upper bound for the observed number of observed false-positives.

As can be seen in Figure 2, the attempted analytical fit to the gaussian distribution underestimates the tail of the observed distribution. The analytical fit to the extreme value distribution,[40] in turn, provides an upper bound for the tail. In the realm of sequence comparison, the extreme value distribution has been used to model scores of random sequence alignments for both local ungapped alignments,[41] as well as local alignments with gaps.[42,43] However, we establish our thresholds on the basis of the numerical distribution.

The number of random alignments with a *Z*-score of >3, for example, is non-negligible (see the tail in Fig. 2 as well as the analytical estimate in the legend of Fig. 2). The expected number of false-positives observed in N trials is $N \cdot P(Z_p)$. Therefore, only relatively high *Z*-scores (that would miss, at the same time, many correct predictions) may result in significant predictions, when searching large databases. Restricting the *Z*-score test only to best matches (according to energy) is insufficient. We find that the double *Z*-score filter performs better, eliminating false-positives with a smaller number of correct predictions that are dismissed as insignificant.

Figure 3 displays 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 lower left corner are incorrect predictions (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. However, the *Z*-scores are computed using 1,000 shuffled sequences for both global and local alignments, which is sufficient for convergence. The converged results somewhat reduce the tails of the distribution. For example, the number of false-positives with a global *Z*-score greater than 2.5 and a local *Z*-score greater than 1.0 is equal to 3, as compared with 7 with only 100 shuffled sequences.

Figure 3 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 (see also the next section). However, this is the price we have to pay for high confidence 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 *Z*-score and a local *Z*-score greater than 3.0 and greater than 2.0 than that of the thresholds is <0.0001.

TESTS OF THE MODEL

We perform four tests in this section on the THOM2 potential. First, we compare the performance of the THOM2 and pair potentials from the literature, using gapless alignments and the S1082 set of proteins. Next, we consider alignments with gaps. We test the specificity and sensitivity of the double *Z*-score filter employed to assess the statistical significance of gapped alignments. Using the double *Z*-score filter, we analyze self-recognition for the S47 set of proteins that contains representatives of folds not sampled in the training. Next, tests of family recognition are presented, including comparison of THOM2 results with those of a pairwise model, using the FEA.

Evaluation of THOM2 and Pair Potentials by Gapless Threading

To make a comparison with pairwise potentials, and to test, at the same time, the generalization capacity of THOM2, we use the S1082 set. This set does not contain proteins included in the training set. However, as discussed in the section, Training and Test Sets, the threshold of 3 Å RMS for global structure-to-structure alignments (using side-chain centers) excludes only close structural homologues. Therefore, the S1082 set includes many structural variations of the folds used in the training. In general, it is difficult to find completely independent test sets when using training sets covering essen-

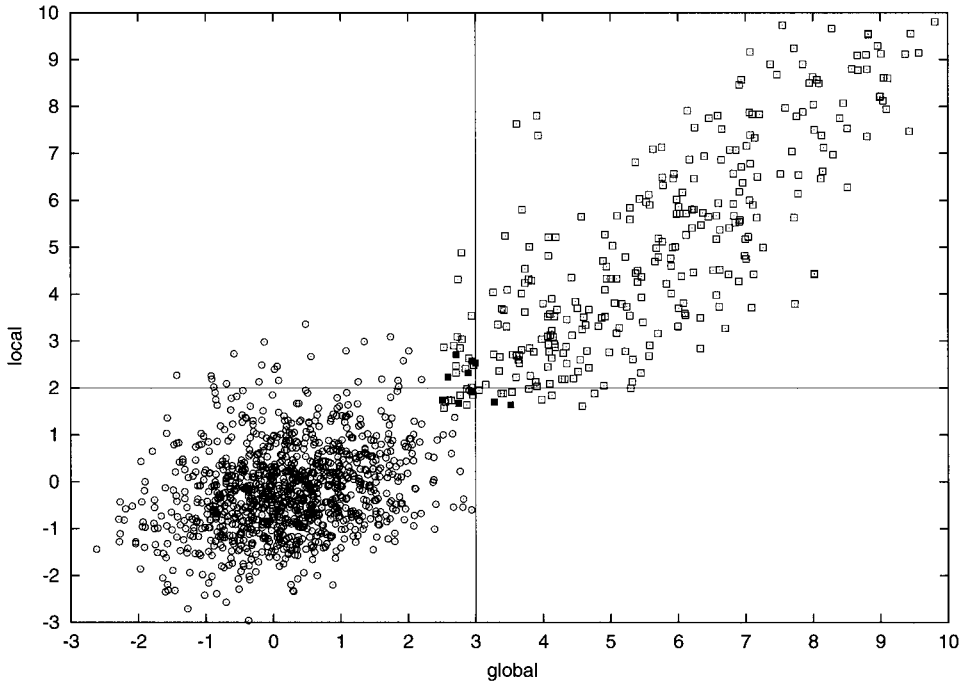

Fig. 3. The joint probability distribution for the *Z*-scores of global and local alignments. Circles at the lower left corner represent a population of 1,081 false-positives, resulting from the alignments of the S47 set sequences (see Fig. 4) against all structures in the Hinds–Levitt (HL) set (100 best global and 200 best local matches are considered, disregarding matches with positives energies of global alignments). The best pair scoring false-positive is slightly below the threshold (3,2). The population in the right upper corner represents (□) 331 pairs of HL sequences aligned to HL structures with global *Z*-scores of >2.5 and local *Z*-scores of >1. This set includes 236 native alignments and 95 non-native alignments; 10 matches are false-positives (■), and they are all below the threshold (3,2). Stiffer energy constraints were employed with only the 10 best global and 200 best local alignments considered. There is a population of true-positives below (2.5,1.0), which are not shown (including 10 native alignments). However, the number of false-positives below this threshold makes predictions within this range difficult.

tially all the known folds. This problem concerns all the knowledge-based potentials considered in this discussion.

Using gapless threading, we compare the performance of THOM2 with the performance of five knowledge-based pairwise potentials. As can be seen in Table X, the Godzik–Skolnick–Kolinski (GSK) potential[44] is the best in terms of the number of inequalities that are not satisfied, followed by the Betancourt–Thirumalai (BT),[45] Tobi–Elber (TE),[25] THOM2, Miyazawa–Jernigan (MJ),[46] and the Hinds–Levitt (HL)[47] potentials. However, in terms of the number of proteins recognized exactly (i.e., proteins with native energies lower than energies of all the decoys generated by gapless threading into all the structures in the S1082 set), the HL potential is the best, followed by TE, MJ, THOM2, BT, and GSK potentials.

The lack of correlation between the above two criteria is related to the fact that some of the above potentials, while recognizing very well many proteins, fare quite poorly for some of the proteins included in the S1082 set. Reducing the number of violated inequalities becomes important when applying some additional filters to select correct predictions from a small subset of energetically favorable matches (e.g., the *Z*-score test; see the section, Assessing the Distribution of *Z*-Scores for Gapped Alignments). Therefore, it would be desirable to satisfy both criteria at same time (also maximizing the *Z*-score of the distribution

TABLE X. Comparison of THOM2 and Knowledge-Based Pairwise Potentials, Using Gapless Threading*

Potential	Recog structs[a]	N_{sat} ineqs (M)[b]	*Z*-score[c]
HL	915	1.84	1.14
TE	914	0.20	1.45
MJ	902	0.28	1.23
THOM2	877	0.24	1.35
BT	861	0.17	1.26
GSK	819	0.08	1.35

*Results of gapless threading on the S1082 set (see text for the details). The results of threading onion model 2 (THOM2) potential are compared with five other knowledge-based pairwise potentials: Betancourt and Thirumalai (BT),[44] Hinds and Levitt (HL),[46] Miyazawa and Jerningan (MJ),[45] Godzik, Kolinski, and Skolnick (GKS),[43] and Tobi and Elber (TE).[25] The latter potential was trained using the linear programming (LP) protocol and the same (TE) training set. Note lack of correlation between the number of proteins that are missed and the number of inequalities, which are not satisfied. See text for further details.

[a]Potentials are ordered according to the number of proteins recognized exactly (out of 1,082).

[b]The number of inequalities that are not satisfied, out of approximately 95 million inequalities generated from the S1082 set (in units of millions).

[c]*Z*-scores (i.e., the ratios of the first and the square root of the second moments) for the distributions of energy differences between the native and misfolded structures.

of energy gaps). From this point of view, the TE, MJ, and THOM2 potentials seem to be somewhat better than the other four potentials. Gapless training of energies remains difficult problem, as reflected in Table X. None of the widely used potentials has a better than 90% success rate. In a set of 1,000 proteins, this translates into many errors.

The conclusion, which is important for the present work, is that the performance of the THOM2 potential is comparable to the performance of pairwise potentials, including the TE potential trained on the same set using a similar LP protocol. Since the proteins used in this test either were not included in the training or represent at least considerable variations of the structures included in the training, we conclude that the exact learning on the training set does not result in overfitting. This is further supported by the results (presented in the next section) for the S47 set of proteins that represent folds not sampled during the training.

Self-recognition by Gapped Alignments

We summarize first the performance of the THOM2 potential in terms of self-recognition of the HL set proteins by optimal alignments and *Z*-score filters. The HL set was partially learned (using gapless threading). However, our training did not include the *Z*-score or the possibility of gaps. Successful predictions based on the *Z*-score only are useful tests, even if performed on the training set of structures. Additionally, there are 40 proteins in the HL set that were not included in the learning (TE) set.

For each sequence, we generate all the global and local alignments into all the structures in the HL set. Energy and *Z*-score filters are considered. Of the total of 246 proteins, 234 native (global) alignments obtain the lowest energy and the highest *Z*-score. There are four native alignments resulting in weak *Z*-scores. The four failures are membrane proteins (from the photosynthetic reaction centers) that were not included in the training set. Only 5 of the remaining 242 native alignments obtain *Z*-scores of <3 (four alignments with *Z*-scores of >2.5 and one alignment with a *Z*-score of <2.5).

For the local alignments, we use the *Z*-score as the main filter, as there are many incorrect alignments with low energies. There are 226 local native alignments with *Z*-scores of >2 (177 of them of rank 1 and 35 of them of rank 2). Among the remaining 20 local native alignments, 9 result in very low *Z*-scores ($Z < 1.0$), including six structures from the training set. Using the double *Z*-score filter with the conservative threshold of 3 for global *Z*-scores and of 2 for local *Z*-scores results in dismissing 23 native alignments as insignificant.

In order to assess further the generalization capacity of THOM2 in terms of self-recognition by optimal alignments, we use the S47 set again. The structures of S47 proteins were embedded in the structures of the TE set, and the sequences of 25 proteins representing different folds in the S47 set were aligned into all the structures of this extended set. We observe that the native structures are found with high probability. A total of 20 of 25 structures result in native alignments with global *Z*-scores of >3 and local *Z*-scores of >2 (Table XI).

A less encouraging observation is the sensitivity of the results to structural fluctuations. THOM2 can identify related structures only if their distance is not too large. Seven out of 14 homologous structures with the DALI[37] *Z*-score for a structure-to-structure alignment of >10 are detected with high confidence. Only one homologous structure with the DALI *Z*-score of <10 is detected.

We would like to point out that only six structures (three pairs of structures representing three folds) of the S47 set had homologous counterparts in the training set. It is therefore reassuring that most of the native structures and significant fraction of the relatives are recognized in terms of both their energies and the *Z*-scores. Moreover, 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 exact learning (on a training set) preserves significant capacity for identification of new folds using optimal alignments with gaps.

Assessing the Specificity of the Protocol

We present examples of family recognition (i.e., identification of homologues) in terms of energy and double *Z*-score filter. Only a few homologues are identified in a large set of (decoy) structures. This allows us to assess the specificity of the protocol, providing a limited analysis of the sensitivity as well (see the next section for an extended assessment of the sensitivity). The test set S1082 is used. Eight families that have at least three representatives in the S1082 set are chosen to illustrate various aspects of THOM2 threading alignments, as compared with DALI[37] structure-to-structure alignments, as well as sequence-to-sequence alignments. The latter ones are generated using Smith–Waterman algorithm,[18] with the BLOSUM50[39] substitution matrix and structurally biased gap penalties (see the section, Training of Gap Energies). Since we do not incorporate family profiles in our threading protocol, we consider only pairwise sequence alignments for comparison in this discussion.

Similarly to threading, the confidence of sequence matches is estimated using *Z*-scores, defined 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 (for sequence alignment). If there is no clear evidence of sharing a common ancestor and a common function, the structural dissimilarity is used to define false-positives. Note that the distribution of *Z*-scores for sequence substitution matrices is different from that of threading potentials, with a very high *Z*-score for highly homologous sequences.

Regarding the specificity of threading results for the families considered in this discussion, we point out that there are only two energy-based predictions with relatively high global and local threading *Z*-scores that are false. They are still below our thresholds. The highest-

TABLE XI. Self-Recognition for Folds That Were Not Learned*

Name (len)[a]	DALI[b] *Z*-sc (RMS)	THOM2[c] Glob *Z*-sc	THOM2[c] Loc *Z*-sc
1hka (158)	33.0 (0.0)	**7.1**	**7.1**
1vhi (139)	4.3 (5.2)	0.2	0.3
2a2u (158)[d]	33.8 (0.0)	**2.5**	**4.0**
1bbp (173)[d]	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**
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**
1b10 (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**

*The S47 set of proteins is used in order to test the self-recognition. It is also a test of the sensitivity of the results to structural fluctuations for 25 different folds (of which 22 were not represented in the training set), using the double *Z*-score test.

[a]Pairs of homologous structures belonging to the S47 set are specified (three folds are represented by a single structure, for 2a2u its structural relative from the training set is included), using their Protein Data Bank (PDB) codes and lengths (specified in parentheses). If the domain is not specified and one refers to a multidomain protein, the A (or first) domain is used. High confidence predictions (global *Z*-score of >3.0 and local *Z*-score >2.0) are indicated in bold. Query sequences are indicated in italics (for each pair, the first line describes the native alignment and the second line an alignment into a homologous structure). Two of 25 native alignments gave weak signals (DNA binding protein 1blo and glycosidase 1bhe). Four other native alignments (2a2u, 1byf, 1jwe, and 1bkb) result in global *Z*-scores of somewhat <3.

[b]DALI[37] *Z*-scores and root-mean-square deviations (RMS) for structure-to-structure alignments into native and homologous structures. Low DALI *Z*-scores indicate that only short fragments of the respective structures are aligned and the resulting RMS may not be representative. Most of the homologous structures with a DALI *Z*-score of >10 are recognized with high confidence.

[c]Results of global and local THOM2 threading alignments of the 25 query sequences into an extended TE + S47 set.

[d]Alignment of the 2a2u sequence into the 1bbp structure was the only significant hit of any of the query sequences into the structures included in the training (Tobi–Elber [TE]) set. Thus, no false-positives with scores greater than our confidence cutoffs were observed.

scoring false-positive, namely the alignment of the aspartyl protease 1htrB into the xylanase 1clxA (*Z*-scores of 3.7 and 1.5, when converged using 1,000 shuffled sequences; see Table XIID), is still below our cutoffs. The alignment of the zinc-finger protein 1meyC into the Adrl DNA-binding domain 2adr is potentially the highest-scoring false-positive among the sequence-based matches. However, even though 1meyC and 2adr are structurally dissimilar according to DALI (RMS of 7.9 Å for 40 residues), they share very high sequence similarity (42% for 55 residues), have similar function, and are classified as related folds (zinc-finger design and classic zinc finger, respectively) by SCOP.[52] Other false-positives due to the sequence-to-sequence alignments obtain *Z*-scores of 5–7, which may cause difficulties in making predictions based on weak sequence similarity.

Regarding the sensitivity of the protocol, one finds first that all the native structures are with the lowest energies and are recognized with high confidence in terms of the double *Z*-score filter. We observe a varying degree of success in the recognition of family members and structural homologs, as illustrated in Table XIIA–H. Threading predictions are very robust for RAS, lactoglobulin, and glutathione transferase families. In the case of the RAS family (Table XIIA), a number of matches into remote structural relatives that share certain structural motifs with the RAS fold are observed. The structural similarity between lactoglobulins and bilin-binding proteins (that do not share detectable sequence similarity) is recognized (see alignment of 2blg into 2apd in Table XIIB). Glutathione transferases 1aw9 and 1axdA, with very weak signals from sequence alignments, are recognized as well.

By contrast, there are families for which threading performance is erratic, including phosphotransferase, cytochrome, and zinc-finger families that include matches recognizable by sequence alignment, of similar length and significant structural similarity, yet not recognized by threading (Table XIIC–F). The results for the pepsin-like acid proteases (Table XIIG) demonstrate missing matches attributable to significant differences in length, which are difficult to account for in global alignments. Local sequence and threading alignments for proteases 1pfzA and 1lyaB result in high *Z*-scores, but no signal from global threading alignment is observed. The family of small toxins is an example of relatively weak signals (both from threading and sequence alignment) that are below our universal cutoffs for false positives (Table XIIH).

Assessing Protein Family Signals and the Sensitivity of the Protocol

Three families are considered: 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 a simplified version of the double *Z*-score filter discussed earlier. Ideally, all the scores should be above the thresholds we presented. The

scores should also correlate with the RMS. The THOM2 results are compared with the results of the TE pairwise potential, which was trained on the same (TE) set using the LP protocol.[25]

The alignments due to the TE potential are computed using the first iteration of the FEA.[23] In THOM2, the number of neighbors to a secondary site determines 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.[23] Alternative to FEA are global optimization techniques[22] that are computationally expensive and difficult to use at the scale of testing presented here. Purely structural characterization of contact types in THOM2 avoids this problem, making the THOM2 potential amendable to dynamic programming, at least for global alignments (see the section, Self-recognition by Gapped Alignments).

Figures 4a–c shows the joint histograms of the sum of Z-scores for local and global THOM2 threading alignments (with trained gap penalties of Table VIIIB) versus the RMS between the superimposed side-chain centers (see the section, Training and Test Sets), for globins, immunoglobins, and POU-like domains, respectively. The vertical lines in the Figure 4 correspond to the sum of global and local Z-scores equal to 5, which approximately discriminates the high confidence matches (with the sum of local and global Z-scores of >5) and lower confidence matches that might be obscured by the false-positives. Nearly all pairs differing by <3 Å RMS can be identified by THOM2 threading alignments. Most of the matches within the range of 3–5 Å can still be identified with high confidence. Overall, 60%, 90%, and 95% of homologues with RMS <5 Å are recognized, for POU, globin and immunoglobin families, respectively. However, the number of matches with high confidence quickly decreases with the growing RMS.

The population of matches that are difficult to identify by pairwise sequence-to-sequence alignments, with structurally biased gap penalties (see the section, Training of Gap Energies and the section, Assessing the Specificity of the Protocol) is represented by the filled squares. All the matches represented by circles can be identified with high confidence by sequence-to-sequence alignments (i.e., they result in Z-scores of >8.0). Essentially all the pairs with RMS of <3 Å are identified by sequence alignments as well. Below this threshold, we observe many matches that can be still identified by threading, but not by sequence alignment (filled rectangles with the sum of threading Z-scores of >5). We also found examples of matches detected with high confidence by threading and not detected by PsiBLAST[48] (with default parameters and the PDB database) in many of the families considered: globins 1flp and 1ash, immunoglobin 2hfm and T-cell receptor 1cd8, toxins 1acw and 1pnh, lactoglobulin 2blg and bilin-binding protein 2apd, pheromones 2erl and 1erp, and POU-like proteins 1akh and 1mbg. By contrast, many sequence alignment matches are not detected by threading.

The performance of THOM2 and TE potentials is compared using 1D histograms for the sum of Z-scores for local and global threading alignments. For the sake of fair comparison, the ad hoc gap penalties, as defined in the section, Training of Gap Energies are used for both potentials. As can be seen in Figure 4d,e for globins and POU-like domains, the number of low Z-scores for THOM2 is smaller than the number of low Z-scores obtained with the TE potential and FEA. For example, the number of low confidence matches (which can still be roughly defined as matches below the cutoff of 5) for globins increases from 2,401 in the case of THOM2 to 3,350 (out of 8,558 matches) in the case of the TE potential. It can also be seen that the distribution of Z-scores is different. The TE potential yields many high Z-scores for alignments into very close homologues, as opposed to lower scores for more divergent pairs.

The somewhat worse performance of the pairwise model for these two families may result from the suboptimality of the alignments that we generate using the FEA. Interestingly, FEA with the TE potential also fails for a larger number of native alignments. For example, in the family of DNA binding proteins, the number of native alignments with very low Z-scores (<4) is equal to 7 for TE and only 2 for THOM2.

By contrast, there are families for which the TE potential works better. An example is the family of the immunoglobins (Fig. 4f). The FEA is expected to perform well when the sequence similarity is sufficiently high, since 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 an RMS of <4 Å can be detected by sequence alignments (Fig. 4c). Therefore, good performance of the FEA with the TE potential is expected in this case.

The above observation is further supported by the results of the FEA with the TE potential for eight families from the S1082 set, considered in the previous section. We do not include detailed results in this discussion. Instead, we summarize them. The threading results with the FEA and the TE potential are robust (and comparable to the THOM2 results) for RAS, SH3, and acid protease families that are represented by proteins of high sequence similarity. The results of the FEA are considerably worse for lactoglobulins and glutathione transferase families that are characterized by much lower success of sequence-based recognition (Table XII). At the same time, the FEA performs as poorly as THOM2 for cytochrome and zinc-finger families. An exception is observed for the toxin family, for which the FEA performs considerably better than THOM2, although there is no (or low) sequence similarity for some of the matches.

CONCLUSIONS AND FINAL REMARKS

We propose and apply an automated procedure for the design of threading models. 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

TABLE XII. Examples of Predictions for Eight Families of Homologous Proteins*

A: RAS family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
121p	**5.9**	**9.5**	**1**	**166**	**74.8**	**36.1/0.0/166/100**
1kao	6.1	5.9	2	167	46.4	28.5/1.4/166/49
3rabA	4.8	3.2	3	169	17.1	27.5/1.4/165/31
1ftn	3.7	3.8	10	193	21.7	22.9/1.8/161/35
1hurA	2.8	3.6	4	180	9.3	14.8/2.5/147/15
1kevA	—[k]	3.6	—[k]	351	—[k]	3.1/4.0/83/11
1mioB	—[k]	3.5	—[k]	458	—[k]	3.5/3.6/99/10[j]
1hdeA	—[k]	3.4	—[k]	310	—[k]	2.5/3.6/111/9
1ksaA	—[k]	*2.7*	—[k]	*366*	—[k]	—[k]
1cbf	—[k]	—	—[k]	*285*	*5.1*	*1.8/3.9/74/9*

B: Lactoglobulin family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2blg	**8.2**	**10.0**	**1**	**162**	**79.9**	**35.1/0.0/162/100**
1a3yA	4.7	3.8	5	149	10.6	17.6/2.4/140/17
1bj7	3.0	3.1	4	150	4.4	17.8/2.4/142/18
2apd	3.0	2.1	2	169	—[k]	11.8/3.0/138/15
1mup	1.7	2.5	3	157	8.9	19.2/2.2/146/16
2pcfB	—[k]	*3.0*	—[k]	*250*	—[k]	—[k]
1lgbC	—[k]	—[k]	—[k]	*159*	*6.0*	—[k]
1nglA	—[k]	—[k]	—[k]	179	5.8	12.0/3.5/136/14

C: Glutathione S-transferase family[a]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2gsq	**7.0**	**7.3**	**1**	**202**	**87.3**	**37.5/0.0/202/100**
1axdA	2.0	5.2	3	209	3.3	18.1/2.9/190/17
1gsdA	3.2	3.7	4	221	16.5	25.1/2.1/200/29
1aw9	4.3	2.5	2	216	4.9	18.4/3.1/194/19
1gnwA	—[k]	4.0	—[k]	211	5.0	17.1/3.1/187/17
1clxA	—[k]	*3.7*	—[k]	*347*	—[k]	*0.5/3.9/50/9*
1fhe	—[k]	—[k]	—[k]	217	11.5	20.9/2.3/195/25
2ljrA	—[k]	—[k]	—[k]	244	5.1	15.7/3.1/195/18
1ao7E	—[k]	—k	—[k]	*245*	*4.9*	—[k]

D: Phosphotransferase (SH3 domain) family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1aww	3.4	4.6	2	67	19.3	8.1/1.7/56/36[j]
2semA	3.9	2.6	3	58	13.8	10.2/1.5/56/31[j]
1fynA	**4.3**	**2.1**	**1**	**62**	**49.8**	**9.5/1.7/56/47[j]**
4hck	3.2	3.5	5	72	28.4	8.1/2.0/55/40[j]
1hsq	2.7	4.0	6	71	10.8	9.8/1.4/54/26[j]
1a3k	—[k]	*3.1*	—[k]	*137*	*3.2*	—[k]
1gbrA	—[k]	—[k]	—[k]	74	13.9	7.7/2.0/57/34[j]
1ark	—[k]	—[k]	—[k]	60	12.3	7.6/1.9/56/20[j]
lnksA	—[k]	—[k]	—[k]	*194*	*5.5*	—[k]

E: Cytochrome c family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
2cxbA	**6.8**	**6.0**	**1**	**124**	**57.4**	**28.1/0.0/123/100**
1co6	—[k]	—[k]	—[k]	107	15.9	14.4/1.7/99/36
1dsn	—[k]	*3.4*	—[k]	*333*	—[k]	—[k]
1crxA	—[k]	*3.1*	—[k]	*322*	—[k]	*0.9/3.3/50/8*
1ndoA	—[k]	—[k]	—[k]	*449*	*4.9*[k]	—[k]
451c	—[k]	—[k]	—[k]	82	3.9[k]	4.9/2.1/64/19
3cyr	—[k]	—[k]	—[k]	*107*	*2.9*[k]	—[k]

TABLE XII. (Continued)

F: Zinc-finger family[b]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1meyC	**5.5**	**3.6**	**1**	**87**	**36.8**	**8.9/1.8/82/51**[j]
1jhb	—[k]	*3.4*	—[k]	*106*	—[k]	—[k]
1iml	—[k]	*2.9*	—[k]	*76*	*6.5*	—[k]
2adr	—[k]	—[k]	—[k]	*60*	*11.4*	*1.2/7.9/40/35*[j]
2drpA	—[k]	—[k]	—[k]	66	9.9	4.9/2.6/58/33[j]

G: Aspartyl protease family[c]

Name[e]	GT[g]	LT[g]	Ene[h]	Len[f]	LS[g]	DALI[i]
1htrB	**9.6**	**8.3**	**1**	**329**	**95.0**	**56.8/0.0/329/100**
4cms	5.0	5.7	3	323	47.5	39.6/1.7/301/39[j]
2jxrA	5.6	3.7	2	329	43.9	37.0/2.1/307/41
1clxA	*3.7*	*1.5*	*4*	*347*	—[k]	—[k]
1egzA	—[k]	*3.6*	—[k]	*291*	—[k]	—[k]
1pfzA	—[k]	2.9	—[k]	380	32.2	31.7/2.4/298/29
1lyaB	—[k]	2.4	—[k]	241	32.0	9.3/2.4/83/59
2pia	*1.3*	—[k]	*6*	*321*	*6.0*	*0.5/4.3/49/6*

H: Scorpion toxin-like family[d]

Name[e]	GT[g]	LT[g]	Ene[b]	Len[f]	LS[g]
1pnh	2.8	2.9	2	31	—[k]
1acw	**2.8**	**2.1**	**1**	**29**	**15.6**
1mea	*2.5*	*1.3*	*4*	*28*	—[k]
1bh4	1.1	2.5	5	30	—[k]
1mtx	1.4	—[k]	10	39	6.0
2pta	—[k]	—[k]	—[k]	35	5.9
1ica	—[k]	—[k]	—[k]	40	4.8
1ilmA	—[k]	—[k]	—[k]	*61*	*3.3*

*Eight families, with a number of representatives included in the S1082 set, illustrating various degrees of success of our threading protocol in terms of sensitivity and specificity. Results are presented for global and local threading alignments using the threading onion model 2 (THOM2) potential, together with the results for (structurally biased) local sequence-to-sequence alignments and DALI structure-to-structure alignments. Representatives used as query sequences aligned to all the structures in the S1082 set are marked in boldface. Matches are ordered according to the sum of global and local threading *Z*-scores and according to *Z*-scores of the local sequence alignments if no threading signal is detected. False-positives (defined as matches with DALI *Z*-scores of <2.0) are indicated in italics. The highest-scoring false-positives for both: threading and sequence alignments are reported for each family.

[a]Example of family with successful threading predictions that do not share a detectable sequence similarity or that have a weak signal from sequence-to-sequence alignment (*Z*-score of <8.0).

[b]Example of family for which threading is less successful, missing a number of family members (of similar length) that can be detected by sequence-to-sequence alignment.

[c]Lack of detection when the difference in length is significant is expected, and it is one of the limitations of the present protocol.

[d]Example of family for which the DALI results could not be retrieved; therefore, the SCOP classification is used to define structural relatives (i.e., proteins that do share the knottins fold).

[e]Names of proteins (Protein Data Bank [PDB] codes).

[f]Lengths of proteins.

[g]*Z*-scores are computed using 50 shuffled sequences for a number of alignments with the lowest energies: 20 best matches in case of global threading (GT) alignments, 500 best matches in case of local threading (LT) alignments, and 50 best matches in case of local sequence (LS) alignments.

[h]Rank of the energy of global threading alignments is reported in the 4th column.

[i]DALI[37] alignments between the (known) structure of a query and the structure of a match are characterized in the last column: *Z*-score, RMS, length of the aligned fragment, and the identity for this fragment are provided.

[j]Comparisons with the FSSP representative of the query structure are used instead of a direct DALI alignment.

[k]Lack of a detectable (threading, sequence, or structural) similarity.

necessity for efficient threading algorithms with gaps, we selected the THOM2 as our best choice.

Statistical filters based on local and global *Z*-scores were outlined. We observe that, while using 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 between the superimposed side-chain centers of ≤ 5 Å

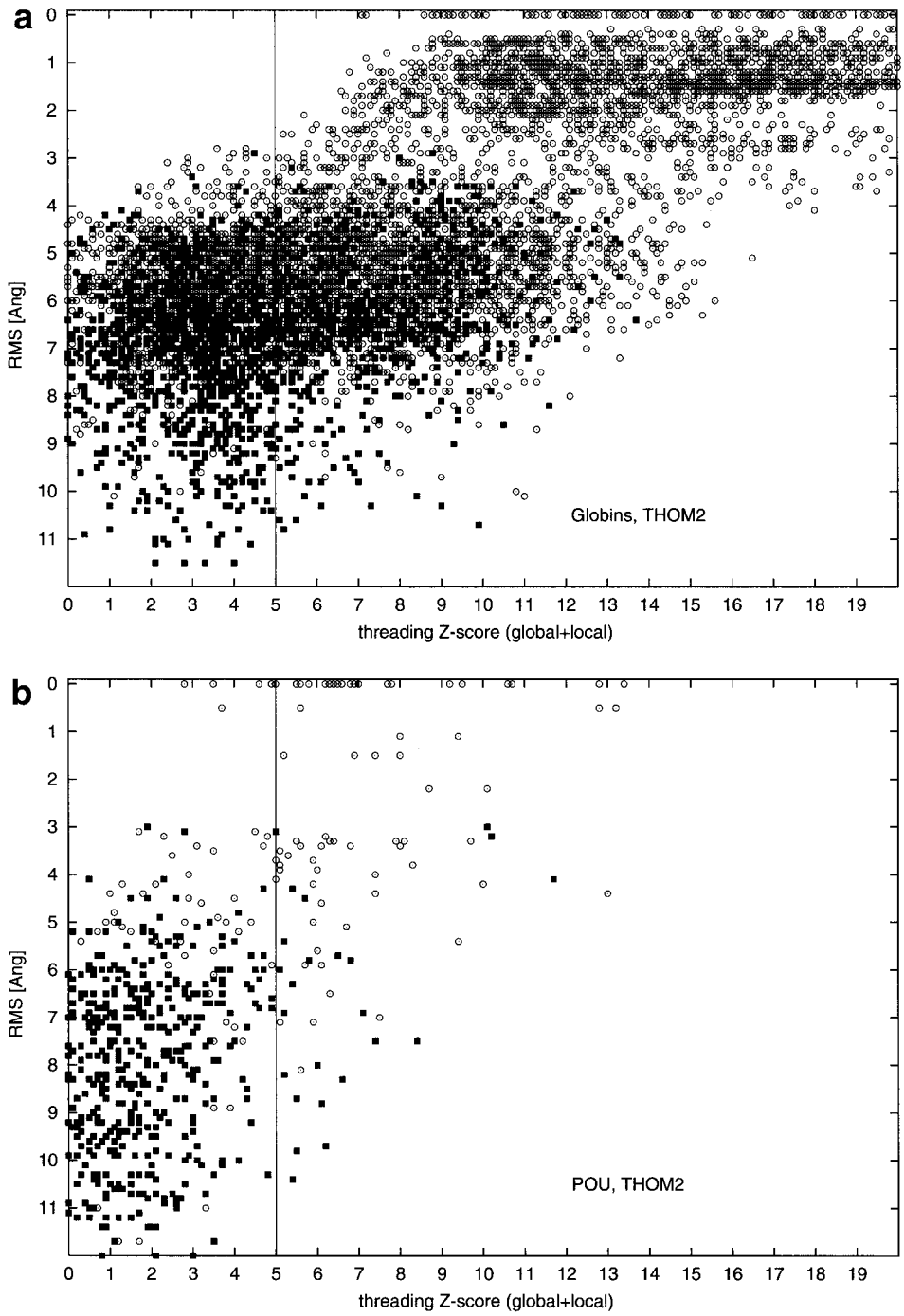

Fig. 4. Comparison of family recognition by THOM2 and pair energies. The results of THOM2 (with the trained gap penalties of Table VIIIB) for families of globins (**a**), POU-like domains (**b**), and immunoglobins (Fv fragments) (**c**). The joint histograms of the sum of *Z*-scores for local and global threading alignments versus the root-mean-square deviations (RMS) between the superimposed (according to structure-to-structure alignments) side-chain centers are presented. The population of matches that are difficult to identify by sequence-to-sequence alignments is represented by the filled squares. Next, the THOM2 results are compared to the results of Tobi–Elber (TE) pairwise potential,[25] using ad hoc gap penalties defined in text. The TE potential was optimized using the LP protocol and the same training set. The first iteration of the so-called frozen environment approximation (FEA)[23] is performed to obtain approximate alignments for the TE potential. One-dimensional histograms of the sum of *Z*-scores for local and global threading alignments for the globins (**d**), POU (**e**), and immunoglobin (**f**) families. Note, that the number of low THOM2 *Z*-scores (<5) is smaller for families of globins and POU-like proteins. By contrast, the TE potential and the FEA perform better for the family of immunoglobins, which is also easier for sequence alignment methods (see text for details).

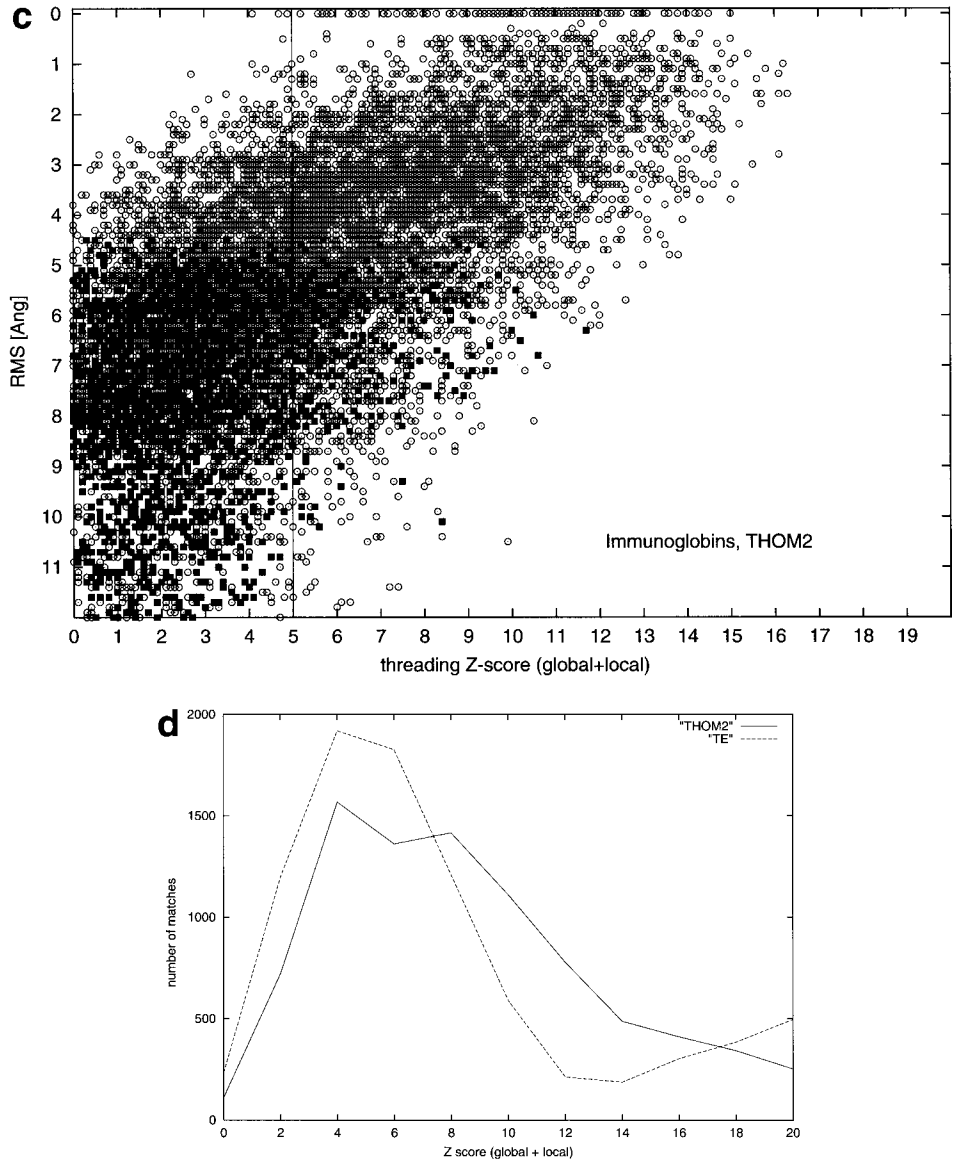

Figure 4. (Continued.)

and differences in length of ≤10%. We also observe many instances of successful recognition of family members that are not 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 sequence similarity, family profiles, or secondary structures as well. Such complementary "signals" are often employed in conjunction with pairwise potentials.[9–11,16] Threading protocols that are based exclusively on contact models were shown (consistent with our observations) to be quite sensitive to variations in structures.[49] THOM2 provides an alternative comparable in performance to pairwise potentials. Therefore, it can be used as a fast component of fold recognition methods employing pair energies, which is the target of a future work.

Despite the limitations of the threading protocol that is based on the THOM2 potential and the double *Z*-score filter (in terms of range of variations in structure and length that can be recognized), we found a number of useful predictions for remote homologues (e.g., ref. 50). Therefore, we decided to take part (group 280) in the recently held critical assessment of fully automated protein structure prediction methods (CAFASP),[51] even at the preliminary phase, without using additional information as secondary structures or family profiles. The performance of the LOOPP server[30] was about average for all fold recognition targets (e.g., LOOPP missed some targets recognizable by Psi-BLAST). How-

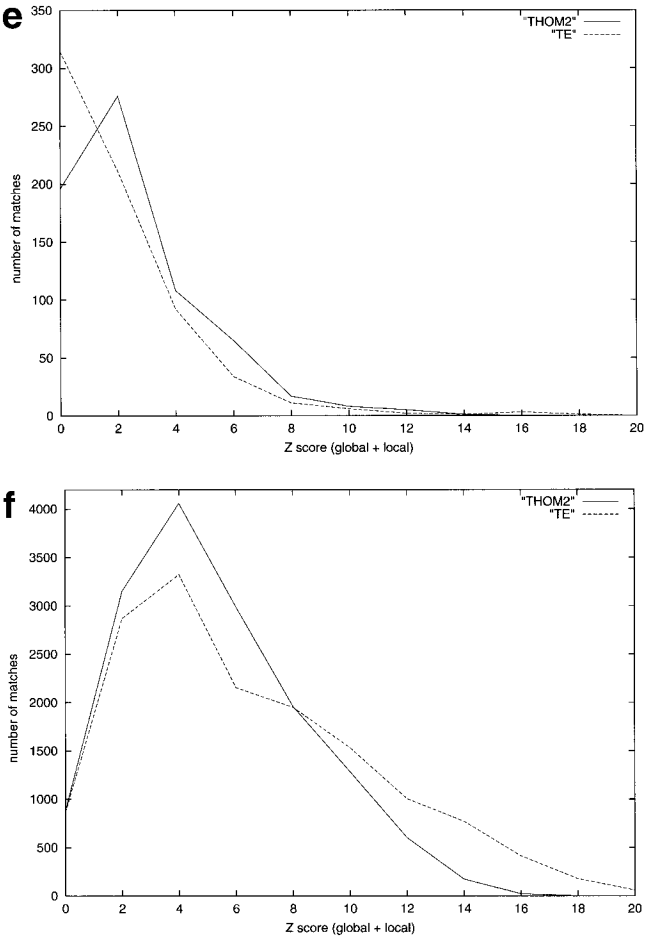

Figure 4. (Continued.)

ever, in the category of difficult-to-recognize targets, it was ranked among the best servers (rank 4 in the MaxSub 5.0 A evaluation), providing the best predictions among the servers for two difficult targets (T0097 and T0102).[51]

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institutes of Health National Center for Research Resources (NCRR) to the Cornell Theory Center (acting director Ron Elber) for the development of Computational Biology Tools. It was further supported by a seed grant from DARPA (to R.E.). Jaroslaw Meller acknowledges also partial support from the Polish State Committee for Scientific Research, grant 6 P04A-066-14.

REFERENCES

1. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Proteins 1992;13:258–271.
4. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
5. Ouzounis C, Sander C, Scharf M, Schneider R. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from 3D structures. J Mol Biol 1993;232:805–825.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993; 16:92–112.
7. Matsuo Y, Nishikawa K. Protein structural similarities predicted by a sequence–structure compatibility method. Protein Sci 1994;3: 2055–2063.
8. Mirny LA, Shakhnovich EI. Protein structure prediction by threading. Why it works and why it does not. J Mol Biol 1998;283:507–526.
9. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
10. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
11. Sternberg MJE, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
12. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. 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 1997;18:849–873.
13. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
14. Babajide A, Hofacker IL, Sippl MJ, Stadler PF. Neural networks in protein space: a computational study based on knowledge-based potentials of mean force. Folding Design 1997;2:261–269.
15. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF. Exploring protein sequence space using knowledge based potentials. J Compar Biol 1999;(in press).
16. Elofsson A, Fischer D, Rice DW, Le Grand S, Eisenberg D. A study of combined structure-sequence profiles. Folding Design 1998;1: 451–461.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–453.
18. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
19. Johnson MS, Overington JP, Blundell TL. Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol 1993;231:735–752.
20. Croman HT, Leiserson CE, Rivest RL. Introduction to algorithms. Cambridge, MA: MIT Press; 1985.
21. Lathrop RH, Smith TF. Global optimum protein threading with gapped alignment and empirical pair score functions. J Mol Biol 1996;255:641–665.
22. Lathrop RH. The protein threading problem with sequence amino-acid interaction preferences is NP-complete. Protein Eng 1994;7: 1059–1068.
23. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. The statistical mechanical basis of sequence alignment algorithms for protein structure prediction. In: Elber R, editor. Recent developments in theoretical studies of proteins. Singapore: World Scientific; 1996. pp 110–140.
24. Maiorov VN, Crippen GM, Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
25. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
26. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
27. Meller J, Wagner M, Elber R. Maximum feasibility guideline in the design and analysis of protein folding potentials. J Comp Chem 2001;(in press).
28. Meszaros CS. Fast Cholesky factorization for interior point methods for linear programming. Computer Math Applications 1996;31: 49–51.
29. Adler I, Monteiro RDC. Limiting behavior of the affine scaling continuous trajectories for linear programming problems. Math Program 1991;50:29–51.

30. Taylor WR, Munro RE. Multiple sequence threading: conditional gap placement. Folding Design 1997;2:S33–S39.
31. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
32. Bryant SH, Altschul SF. Statistics of sequence–structure threading. Curr Opin Struct Biol 1995;5:236–244.
33. Fitch WM. Random sequences. J Mol Biol 1983;163:171–176.
34. Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol 1985;2:526–538.
35. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In: Pacific Symposium on Biocomputing, Hawaii, 1996; p 300–318.
36. CASP3. Third community wide experiment on the critical assessment of techniques for protein structure prediction, Proteins 1999;Suppl 3; see also http://predictioncenter.lnl.gov/casp3.
37. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609; see also DALI server; http://www2.embl-ebi.ac.uk/dali.
38. Meller J, Elber R. Learning, Observing and Outputting Protein Patterns (LOOPP)—a program for protein recognition and design of folding potentials; http://www.tc.cornell.edu/CBIO/loopp.
39. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1989;89:10915–10919.
40. Gambel EJ. Statistics of extremes. New York: Columbia University Press; 1958.
41. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87:2264–2268.
42. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
43. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276:71–84.
44. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
45. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;2:361–369.
46. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256:623–644.
47. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
49. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
50. Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KP, Tanksley SD. Cloning transgenic expression and function of fw2.2: a quantitative trait locus key to the evolution of tomato fruit. Science 2000;289:85–88.
51. Fischer D, et al. CAFASP-2: the second critical assessment of fully automated structure prediction methods. Proteins CASP4 issue. 2001;(in press).
52. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of protein data base for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.