Title of article: Scoring functions for *ab initio* protein structure prediction

Suggested running head: Scoring functions for *ab initio* methods

Authors: Enoch S. Huang, Ram Samudrala, Britt H. Park

Department of Structural Biology

Stanford University School of Medicine

Stanford, CA 94305

1. INTRODUCTION

The native conformation of a protein is generally assumed to be the one with the lowest free energy [1]. The successful prediction of protein structure depends on the surmounting of three sub-problems: (1) choosing a representation of protein conformation that includes structures similar to the correct conformation but limits the search space; (2) formulating a scoring function that relates a particular protein conformation to its free energy; and (3) devising a method to combine the first two elements in a search through conformational space for the state with the globally optimum score. These three requirements apply to the major classes of protein structure prediction: homology modeling, threading (fold recognition), and *ab initio* folding. In this chapter we focus on the second of the three sub-problems, that of developing energy functions, and place an emphasis on functions tailored for *ab initio* folding, although much of the discussion will also apply to threading.

The form of a scoring function is dependent on the particular type of problem to be tackled. For instance, in homology modeling, the backbone (or fold) of the target protein is assumed to be known, as it is derived from a related protein with known structure. A suitable function computes the total score for interactions between pairs of side-chains and side-chains with the backbone, to build side-chain conformations. However, in threading and *ab initio* folding, one is primarily concerned with capturing the overall fold, or topology, of the backbone. For example, consider an *ab initio* folding scenario in which one starts with a fully extended polypeptide backbone and attempts to fold it with respect to some scoring function. In order to make the search problem more tractable by reducing the degrees of freedom afforded to the protein, side-chain atoms are often reduced to a single coordinate [2], thereby decreasing the computational overhead; likewise, the applicable scoring functions are reduced in complexity. Threading techniques also use these simplified functions to score the alignment of probe sequences mounted upon structures and sub-structures found in the protein data bank (PDB) [3]. Such functions are suitable because the original side-chain conformations of the template are discarded when a probe sequence replaces the identity of the residues.

Obviously, simplified functions cannot be rooted in the same physical principles as the all-atom functions used for the molecular simulation of proteins which require the explicit positions of all the atoms in the protein [4, 5, 6, 7]. Parameters for these potential energy functions, or force fields, are obtained from experimental data and quantum mechanical calculations. In contrast, most of the scoring functions used in protein structure prediction fall into the category of knowledge-based potentials of mean force [8, 9]. The term "knowledge-based" refers to the statistical analysis of the properties found within the database of experimentally-determined protein structures. Knowledge-based functions mine the information-rich protein database by converting properties seen native proteins into "pseudo-energies" that reflect the compatibility of a given sequence with a structure. A wealth of properties of native structures is readily extracted, for instance the pairwise interaction of residues, the exposure of non-polar groups to solvent, the propensity of sequences to form secondary structure, and the close packing of protein atoms [10, 11, 12, 13]. The choice of the property is at the discretion of the modeller; hence, a knowledge-based function can be derived using a range of fold representations, from a string of secondary structure assignments to a full-atom representation. While simplified scoring functions are typically knowledge-based, the converse is not true.

Knowledge-based energy functions are not without problems in their theoretical justification [14, 15, 16, 17, 18, 19, 20]. Although the details of this discussion are beyond the scope of this chapter, the main points are presented here. First, knowledge-based functions derive their parameter sets from experimental data, typically by applying the inverse Boltzmann law to the observed properties in the protein database:

$$\Delta E = -kT\ln\left(\frac{f_1}{f_2}\right)$$

where the energy difference ΔE between two states is related to the ratio of their occupancies (f_1 and f_2); T is the temperature (°K) and k is the Boltzmann constant. f_1 is the frequency of observations of a certain type in the database, and f_2 is the number of observations expected by chance (defined by the chosen reference state, see below).

At least four assumptions underlie the application of the inverse Boltzmann law in this fashion: (1) the set of known stable folds of different proteins are representative of proteins in general; (2) the protein set represents a system at equilibrium; (3) the observed frequencies are independent of each other and their environment; and (4) the observed frequencies are distributed according to the Boltzmann law.

However, Thomas & Dill have shown that inter-residue interactions are not independent [17]. Rather, the result of a dominating hydrophobic effect is to influence the types of interactions that polar residues make, simply because each structure can only make a limited number of inter-residue contacts. For example, the extracted parameters for charged residues do not mainly reflect electrostatic interactions; charged residues are driven to the protein surface by the non-polar interactions, coupled by chain connectivity and excluded volume effects. Also, Kocher *et al.* argue that since protein folding is cooperative, inter-residue interactions cannot be independent [14]. Finally, Thomas & Dill have shown that the size of the proteins used to compile the parameters can also skew the extracted scores [17].

To circumvent the need for the assumptions surrounding the conversion of database statistics to true energies, some methods rely instead on the Bayesian formalism (i.e. conditional probabilities) to formulate a scoring function [21, 22]. The two formalisms are analogous and follow the same methodology in practice. We therefore refer to all the knowledge-based functions discussed in this chapter as "scoring" or "objective" functions.

Given that there are a multitude of scoring functions designed for protein structure prediction by threading and *ab initio* folding, it is important to understand how they work. In the next section, we will provide examples from work conducted in our laboratory, and in the literature. We will dissect out the essential components of scoring functions for *ab initio* folding, and compare and contrast the similarities and differences among them. Our intent is not to do an comprehensive review [8, 9, 16, 23, 24] but to stereotype the different components of the various scoring functions and explain their specific roles.

Ab initio folding methods can be largely placed into two main categories: fold generation by exhaustive enumeration or by minimization. Each of these classes can further be sub-divided into lattice and off-lattice (torsion-based) approaches. We will look at an example of each of these four sub-types of ab initio folding methods. Threading functions will not be discussed explicitly, but many knowledge-based functions used in *ab initio* folding can be directly applied to evaluating sequence-structure compatibility in a threading context [10, 12, 13, 25]. However, successful threading or fold recognition is by no means limited to the knowledge-based functions described in this chapter. Many excellent alternatives exist, including methods that use environmental profiles [11], predicted secondary structure [26, 27, 28], and multiple sequence alignment [29].

2. METHODS

2.1. General issues

Although all of the scoring functions discussed below were developed and tested for *ab initio* folding, some are exclusively knowledge-based. Some do not rely on the database of known structures, but model forces such as the hydrophobic effect, and hydrogen bonding. A few others combine the two approaches. For the knowledge-based functions, we will discuss some general procedural issues.

2.1.1. Selection of a database

The standard procedure for constructing a fold library to compile a scoring function is to choose a non-redundant set of proteins that reflect all known folds. One way to do this is to require that no two proteins in the set share more than 30% sequence identity. Undesired bias can arise from over-representing proteins of a certain size or topology (for instance, alpha-helical proteins), and thus a balanced mixture of proteins with different secondary structures must be used. The set should also be as large as possible to make the observed statistics robust.

2.1.2. Jack-knifing

Development and validation of a scoring function must proceed without specific knowledge of the target protein. A true threading or *ab initio* experiment would be carried out only in the absence of a template structure with suitably high sequence similarity (otherwise the problem shifts from fold recognition / generation to homology modelling!). Thus, validation of a given scoring function for use in threading or *ab initio* folding must ensure that no inadvertant use of information occurs. One commonly employed technique is that of "jack-knifing." Consider the case where parameters for a scoring function is extracted from a database of 300 proteins. Presumably the parameters reflect the tendencies of native proteins in general with respect to some property of interest (for instance, frequencies of pairwise contacts), but in reality the parameters will be biased in some degree towards the 300 proteins. In practice, this implies that one cannot validate the scoring function on a test set of proteins that includes any of the proteins used to compile the parameters (or any related proteins thereof). Furthermore, training or optimizing a scoring function with respect to performance on a fixed test set, whether the database was previously jack-knifed or not, is tantamount to introducing knowledge of the test set.

2.1.3. Correction for sparse data

If one is extracting many properties from the database, the problem of sparse data arises. Sippl [10] suggests the following correction for the observed frequency of sequence s in structural state *c:*

$$\rho'_{s,c} = \frac{1}{\sigma + m_s}\left(\sigma\rho_c + m_s\rho_{s,c}\right)$$

where m_s is the number of occurrences sequence s appears in the database and $\rho_{s,c}$ is the unadjusted frequency that sequence s appears in structural state c. The effective sequence-dependent frequency $\rho'_{s,c}$ is equal to a combination of the sequence-independent frequency ρ_c and the actual number of sequence-dependent occurrences of structural state *c.* The adjustable parameter σ sets the relative weight of the sequence-independent term (chosen as 50 in [30]). This correction for sparse data is most commonly employed when one is generating potentials of mean force in at various sequence separations (see section 2.2.2).

2.1.4. Choice of a reference state

Knowledge-based scoring functions express their "pseudo-energies" relative to a reference state. For example, a reference state might represent a system in which the actual interaction energy between residue pairs equals zero; i.e., a system exhibiting the contact frequencies of a randomly interacting system. This state may or may not include explicit solvent molecules, the presence of which dramatically affects the resulting effective energy of interaction between two residues. Since there are many ways to formulate a reference state [20], this issue will be individually addressed where applicable.

2.2. Exhaustive enumeration methods

2.2.1. A lattice model

In the studies by Hinds & Levitt [31, 32], all possible conformations of a sequence were generated, subject to the bounds, spacing, and geometry of the lattice. The knowledge-based scoring functions used by the study had the following functional form:

$$E = \sum_{contacts} e_{ij}$$

where e_{ij} is the contact score between residues types i and j and the total score E is the sum of all pairwise scores observed in the lattice structure. These so-called contact functions typically are square-welled, i.e. the interaction between a pair of residues is value e_{ij} if the residues are within a cutoff distance (6 to 8 Å is customary) and zero otherwise.

The parameters for the 210 values for e_{ij} (i.e., in a 20x20 symmetrical matrix) are calculated as

$$e_{ij} = \frac{N_{ij}^{obs}}{N_{ij}^{exp}}$$

where N_{ij}^{obs} is the number of observed contacts between residue types i and j. In the selected database and N_{ij}^{exp} is the number of contacts made in the reference state, or

$$N_{ij}^{exp} = \sum_{p} C_p \frac{T_{ijp}}{T_p}$$

where p is a protein in the database, T_p is the total number of possible tertiary contacts, and C_p is the number of actual tertiary contacts. The total number of contacts T_p is a simple function of the total number of residues in protein p, N_p:

$$T_p = \left(N_p - 4\right)\left(N_p - 5\right)$$

T_p is not exactly equal to N_p^2 because interactions between nearest neighbors along the sequence ($|i\text{-}j| < 5$) are disregarded. T_{ijp} is equal to the number of i and j pairs that are not nearest neighbors in the sequence. The ratio T_{ijp}/T_p is effectively the product of the concentrations of i and j. Contacts in the database are counted whenever a heavy atom of one residue is within 4.5 Å of a heavy atom of another residue.

This technique of recovering effective contact energies from the database is also referred to as the “quasi-chemical approach” [2, 20]. Briefly, this approximation treats the interacting centers (e.g. residues) as disconnected units that interact randomly and whose expected (or reference) contact frequency is proportional to their relative concentrations. This particular method uses a reference state with the compactness and packing patterns of native proteins.

The goal of exhaustive *ab initio* methods is achieved when the fold closest to the native structure corresponds to the global energy minimum. If there is more than one fold that resembles the native fold to within some RMS or DME cutoff, then ideally that subset of folds has better scores than all the other, non-native folds.

The tetrahedral lattice of Hinds & Levitt is a coarse lattice in that it is only able to generate walks suggestive of the overall native trace [31, 32]. On the other hand, this lattice can support exhaustive enumeration of most small proteins. The number of total walks is therefore very large (on the order of 10^7). Hinds & Levitt [31] did not report the rank of the nearest-native fold in the ensemble, but they note that out of the lowest-energy 10^3 to 10^4 folds, there are on the order of 10 native-like folds (4 to 5Å DME).

2.2.2. An off-lattice model

Next, we examine the four-state off-lattice model of Park & Levitt [30, 33]. By using only four states in Ramachandran space, one can reproduce the native fold to about 2 Å RMS error. Unfortunately, exhaustive enumeration of a small protein (100 residues) implies 4^{100}, or 10^{60} conformations, which is intractably large. However, if one enforces idealized native secondary structure (i.e., one state each to represent α and β states), allowing only 10 selected loop and turn residues to assume the four possible (ϕ, ψ) possibilities, then one only needs to contend with 4^{10} folds (about a million). After applying a generic compactness filter, only ~200,000 structures remain. One may think of this fold ensemble as the set of all possible arrangements of native secondary structure.

As in the case of Hinds and Levitt [31, 32], for a given set of conformations there were many (on the order of 10^2) near-native folds (≤ 4 Å RMS deviation from the native structure) present. Park et al. [34] evaluated a series of scoring functions by computing a Z-score (defined as the number of standard deviations a particular score departs from the mean score in the set) for each near-native fold. The best functions had the most negative average Z-scores for the near-native folds (a Z-score ≥ 0 means that the function did not discriminate better than random for that structure). Table 1 lists some of representative functions and their average Z-scores for 8 small proteins. Park et al. [34] also reported that for many functions, one of the near-native folds would rank very high in the score-sorted list. For instance, the Shell function placed a near-native fold within the top 100 of every fold ensemble for 8 different proteins (corresponding to the top 0.1 to 1% of a score-scored list). However, many non-native folds were also among the lowest-scoring conformations in each set, even though the near-native folds overall were favored. In other words, none of the simplified knowledge-based functions could identify near-native folds without also including some non-native folds.

Table 1: Performance of four selected energy functions.
Four energy functions described by Park et al. [34] are tested on eight semi-exhaustive off-lattice decoy sets. The average Z-score for the near-native folds (those within 4 Å RMS error of the native fold) is shown for each function.

Function	**<Z-score>**
Histogram	-1.27
Shell	-1.78
Contact(MJ)	0.03
HF	-1.51

The Shell function, the top performer out of our representative set of four functions is a simple contact function. Whenever a pair of residues that is more than one residue apart in the sequence is within 7 Å, a score e_{ij} is counted. Nearest neighbors in the sequence are ignored simply because they are always in close spatial proximity with each other, and hence should not contribute to the signal. Residues are reduced to a single "interacting center" (or virtual centroid) 3 Å from the Cα atom along the Cα-Cβ vector.

The 210 parameters e_{ij} are computed essentially in the same manner as described in 2.2.1, in that a compact, randomly mixed reference state is employed. Two subtle differences are:

1. The Shell function counts residues in contact (both in the database and in the set of ab initio folds) when their virtual centroid positions are within 7 Å of each other.
2. The value of T_p for the Shell function reflects the smaller sequence separation cutoff for interacting residues (only $|i-j| < 2$ are ignored).

The Histogram function is an implementation of the Sippl [10] potential of mean force (PMF). Unlike contact functions, which typically apply the quasi-chemical approximation in an explicit reference state, a PMF uses an implicit reference state (see below). The potential of mean force W between two interacting centers (e.g. residues) i and j is defined as:

$$W_{ij}(r) = -kT\ln\left(\frac{\rho_{ij}^{obs}(r)}{\rho(r)}\right)$$

where ρ_{ij} is the observed probability density that residues i and j are at distance r.

The reference state is a hypothetical state for the polypeptide that reflects the observed inter-residue distances of the database with sequence information removed. As inthe case for contact functions via the quasi-chemical approximation, the energy parameters are extracted from the observed amino acid distributions in a subset of the PDB. This function is named the Histogram function because it relates the energy of interaction as a function of observed inter-residue distances (calculated as the distance between the Cβ atoms). Hence, instead of recovering 210 pairwise contact parameters for the 20 amino acids, 210 histograms are generated. Each histogram reflects the relative frequency of inter-residue distances sampled at 20 uniformly spaced intervals. Furthermore, the classic implementation of the Sippl function [10] involves modeling the role of local and long-ranged pairwise interactions by generating separate histograms for pairs of residues at a given separation along the polypeptide chain (called a topological level).

In the Park & Levitt [30] implementation of the PMF described by Sippl [10]., ten histograms for each of the residue-pair interactions were generated in the following manner: 8 for local interactions with sequence separation 3 through 10, inclusive; 1 for medium range interactions (sequence separation 11 through 50, inclusive); and 1 for all other, long-range interactions.

Spatial distance bins were computed for each histogram by storing the minimum and maximum distances and dividing the range into 20 equal distance bins. The correction for sparse data was applied (section 2.1.3). If there is no sample in a particular bin, its occurrence was re-set to one to prevent the computed energy from going to infinity. Fortunately, these slots correspond to geometries that are very unlikely in real proteins.

The last of the three knowledge-based functions is called the Contact(MJ) function. This function differs from the Histogram and Shell functions in one key respect: the reference state is a random mixture of solvent and amino acids, which directly models the effect of desolvation in protein folding. A quasi-chemical reaction between two amino acids *i* and *j* in solvent can be expressed as:

$$i\text{-}0 + j\text{-}0 \Leftrightarrow i\text{-}j + 0\text{-}0$$

where 0 represents a solvent molecule. The effective energy of desolvation and contact formation e_{ij} is determined by separate terms for the effective energy of breaking the *i*-0 and *j*-0 interactions and forming *i-j* and 0-0:

$$e_{ij} = e'_{ij} + e'_{00} - e'_{i0} - e'_{j0}$$

Each energy parameter *e′* is determined by the same equation used in the Shell energy function.

The effects of introducing solvation-dependent energies e'_{i0}, e'_{j0}, and e'_{00} include:

1. Desolvation energies that are more favorable for polar and charged residues than hydrophobic residues.
2. The introduction of a favorable solvent interaction term, e'_{00}, which causes all the energy terms to be more favorable (more negative) by a constant.

Each of the e' terms is computed separately using the equation above [31, 32]. To extract the parameters, the following are required:

1. Each residue type i has an average coordination number q_i, estimated by scanning the database of known structures for buried residues of type i. When q_i is greater than the number of inter-residue contacts made by a particular residue i, then the difference is set to the number of contacts between residue i and solvent.
2. The number of solvent molecules is a free parameter equal to twice the number of residues in the protein [30].
3. The total number of contacts in the system is equal to T_p plus the number of solvent contacts, which is set to the number of solvent molecules times its coordination number. The coordination of water was set to the the average residue coordination number.
4. The total number of solvent-solvent contacts is equal to the number of solvent contacts minus the total number of residue-solvent contacts.

The hydrophobic fitness (HF) function [35] is unusual in that it derives no parameters from the PDB. Instead, it simply rewards favorable arrangements of hydrophobic and polar residues. A conformation is scored favorably if hydrophobic residues (of any type) make more contacts with other hydrophobic residues than would be expected on average. The overall score is weighted by a term that rewards the burial of hydrophobic residues. The form of the HF function is:

$$HF = -\left(\sum_i B_i\right)\left(\sum_i \left(H_i - H_i^\circ\right)\right)$$

where i is hydrophobic {C, F, I, L, M, V, W}; B$_i$ is the number of virtual side-chain centroids within 10 Å; H$_i$ is the number of hydrophobic residues (plus Y) with 7.3 Å. H_i° is the expected number of hydrophobic contacts based on a random distribution of the other residues surrounding residue i, disregarding the nearest neighbors in the sequence. H_i° is computed by multiplying the fraction of hydrophobic residues with the number of contacts residue i makes.

2.3. Minimization methods

Scoring functions take on different forms when structure prediction is attempted on a lattice by minimization protocols. When one is not concerned with exhaustive enumeration, a finer lattice may be used, thereby improving the accuracy to which a native fold may be represented. The trade-off is that one can never be sure that the best fold can be found by minimization, either because of imperfections in the energy function, search strategy, or both.

Unlike scoring functions used in exhaustive methods, a scoring function used in minimization must bear the additional burden of favoring generic features of native states, namely secondary structures and compactness. Exhaustive methods can enforce compactness simply by setting the bounds of a lattice or by simply discarding structures that do not satisfy a radius of gyration cutoff. In contrast, minimization starts with a random or extended state, and compactness must be monitored by at least one component of the scoring function. The problem of secondary structure formation may be surmounted by imposing native secondary structure assignments. Otherwise, a combinatorial explosion in the search process is averted by biased sampling of conformational space [21] or by *ad hoc* terms in the energy function favoring secondary structure formation (for instance, via hydrogen bonding). The implementation of these terms are typically specific to a given structural representation (e.g., lattice models with a certain geometry and spacing), so we will not discuss the functional forms or parameter derivation at length unless they are illustrative of general issues.

2.3.1. Minimization on a lattice

In the study by Kolinski & Skolnick [36], a dual lattice model was used for folding by optimization of a scoring function. In their scheme, a coarse lattice was used for the early stages of folding from an expanded state, and refinement of the initial structures was performed on a finer lattice. The entire scoring function is written as

$$E = E_{\mathrm{C}\alpha} + E_{H-bond} + E_{rot} + E_{sg-local} + E_{one} + E_{pair} + E_{tem}$$

and can be divided into three components: sequence-independent terms, sequence-dependent local and long-range terms, and multi-body side-chain interactions.

$E_{\mathrm{C}\alpha}$ and E_{H-bond} are the two sequence-independent terms. $E_{\mathrm{C}\alpha}$ acts as an effective Ramachandran potential. Every *i*, *i*+3 inter-Cα distance and its corresponding chirality (defined by the three intervening virtual Cα-Cα bonds) are compared with those extracted from the PDB. The resulting energy term enforces local geometries that favor secondary structure formation. The second generic term, E_{H-bond}, models H-bond formation based on pairs of Cα-Cα vertices that are 4 or more residues apart in the sequence. A hydrogen bond between Cα vertices *i* and *j* must satisfy the following geometrical restrictions:

$$\left|(\mathbf{b}_{i-1} - \mathbf{b}_i) \bullet \mathbf{r}_{ij}\right| \le a_{\max}$$
$$\left|(\mathbf{b}_{j-1} - \mathbf{b}_j) \bullet \mathbf{r}_{ij}\right| \le a_{\max}$$

where **b** is a backbone vector, $\mathbf{r}_{ij}$ is the vector between the Cα positions, and a_{max} is a parameter set by the lattice spacing. A H-bonding cooperativity term rewards the formation of hydrogen-bond networks by adding to E_{H-bond} subtotal a separate score when consecutive sets of residues *i,j* and *i*±1, *j*±1 are hydrogen-bonded.

For sequence specific energy terms, a simplified representation of side-chain rotamers (a single interaction center) was used. The energy of a given rotamer was simply tied to the frequency of that rotamer in the library (E_{rot}). The angle θ between two consecutive Cα-side-chain vectors was computed and scored as

$$E_{sg-local} = -\ln\left(\frac{\cos\theta^{obs}}{\cos\theta^{\exp}}\right)$$

where the expected occurrence assumes a uniform distribution of states. $E_{sg-local}$ refers to the local interaction of side-groups (or side-chains).

The long range interactions include a one-body term (E_{one}) and a pair potential (E_{pair}). The former is designed to drive hydrophobic residues into the interior of a folded chain. This term is designed to penalize extended states and addresses a central need of all minimization methods to drive the collapse of a polypeptide chain. This single-body term takes on two forms, one that is related to the position of a given residue from the center of mass of the polypeptide chain and a second that considers the number of contacts it makes relative to the average number for that residue in the database. The pair potential has a repulsive term that chases steric clash between side-chains and other side-chains and the main-chain and a statistically-derived scoring function similar to those described elsewhere in this chapter. The cutoff distances for repulsion and pairwise interaction are dependent on the residues involved. The strength of attraction is modulated by a factor f that reflects the average backbone orientation of the secondary structures:

$$f = 1.0 - \left[\cos^2(\mathbf{u}_i, \mathbf{u}_j) - \cos^2(20°)\right]^2$$

where $\mathbf{u}_i$=$\mathbf{r}_{i+2}$-$\mathbf{r}_{i-2}$ with $\mathbf{r}_i$ being the position of the ith Cα vertex. The maximum of this function occurs at 20° and the minimum at 90°.

Finally, the multibody term $E_{tem}$was added to simulate the cooperativity of side-chain packing from a molten state to a more native-like state. The authors note that in the absence of this term, the folds have the character of molten globules, i.e.

with well-defined secondary structure but more expanded than close-packed tertiary structures. The multibody term assumes the following form:

$$E_{tem} = (e_{ij} + e_{i+k,j+n})C_{ij} \times C_{i+k,j+n}$$

where C_{ij} = 1 when residues i and j are in contact and residue spacing $|k| = |n|$; k and n assume values of {±3, ±4}.

The relative strength of each of these contributions was set by requiring that secondary structure be more prevalent in the collapsed states than in unfolded conformations.

Starting from a random configuration on the coarse lattice, folding was attempted for three small proteins [37]. In the interest of conciseness, we focus on the folding of Protein A, in many ways the most successful experiment of the three. The 60-residue fragment of this protein adopts a three-helical bundle topology. Folding of this protein was carried out 45 times using a simulated annealing protocol on the coarser lattice. In 19 trials the correct three-helical conformation was seen, in another 11 trials, a three-helical bundle of incorrect topology persisted. Overall, the average conformational energy of the correct folds was lower than that of the incorrect folds, and the reproducibility of the non-native folds was much lower than for the native-like folds. Further refinement of 5 near-native folds upon the finer lattice yielded structures in the 2-3 Å RMS error range (excluding the residues at the N and C termini).

It should be noted that evaluation of a scoring function *per se* in minimization experiments is difficult because the observed performance is dependent on the search strategy as well as the function used. Generally speaking, the best methods available today can provide native-like folds in a significant fraction of the folding trials, as was the case for Protein A summarized above. However, successful convergence to a native-like fold is still limited to a handful of proteins.

2.3.2. Folding in torsion space

For our example for minimization in torsion space, we choose the work by Dill and co-workers [38]. As with many *ab initio* methods, the authors rely on the constraint of native secondary structure in order to overcome the vast conformational search problem. Unlike the exhaustive enumeration strategy of Park & Levitt [30], this minimization method has large dihedral library with which to place the rigid secondary structure elements. As a first step, the conformational search is powered by a genetic algorithm [39] that operates on a string of paired (ϕ, ψ) dihedral angles. A second step then refines the search by choosing a random adjustable residue and changing the torsion angles incrementally to probe the local energy surface for minima. The protein chain is reduced to a backbone with ideal bonds and angles and *trans* peptide conformations, and side-chains are represented by a virtual atom centered at the average rotamer observed in the PDB.

The scoring function of Sun, *et al.* [38] could afford to be much simpler than the one described above. Since native secondary structure was already in place, the energy terms favoring secondary structure formation were rendered unnecessary. In fact, their scoring function is surprisingly simple, as it relies mostly on hydrophobic interactions balanced by steric repulsion:

$$E_{Total} = E_{HH} + E_{ex}$$

where E_{HH} is an attractive interaction between hydrophobic residues {A, C, I, L, M, F, W, Y, V}. The magnitude of the attraction is distance-dependent, but the functional form is an analytical expression rather than a database derivation like the Histogram function (section 2.2.2). The expression is:

$$E_{HH} = \sum_{i} \sum_{j>i+1} e_{ij} f(d_{ij})$$

where e_{ij} is -1 if and only if i and j are hydrophobic residues and zero otherwise and d_{ij} is the distance between side-chain centroids of residues i and j. The coefficient f modulates the attraction by the following sigmoidal function:

$$f(d_{ij}) = \frac{1.0}{1.0 + e^{(d_{ij} - d_0)/d_t}}$$

d_t is a parameter that sets the sharpness of the sigmoidal function (set to 2.5 Å) and d_0 sets the interaction distance (6.5 Å) as the midpoint of the curve. The attraction is set to zero at 12 Å.

The excluded volume term is also a sigmoidal function between pairs of Cα atoms or side-chain centroids:

$$E_{ex} = C \times \sum_{ij} \frac{1.0}{1.0 + e^{(d_{ij} - d_{eff})/d_w}}$$

where d_w is 0.1 Å and d_{eff} is 3.6 Å for Cα atoms and 3.2 Å for side-chain centroids. The constant C sets the scale for the repulsive term higher relative to the attractive term.

To aid the formation of β-sheets during the folding process, a score of -1.0 for hydrogen bonding between beta-strands was added for every instance when certain geometical conditions were met (O-H distance < 2.5Å and N-H-O angle between 120° and 180°).

The method was tested on 10 small proteins. Of these, four of the lowest-scoring models were within 4 Å RMS error. The authors did not report the scores of the most native-like folds the representation could allow in terms of RMS error, but 8 of the native structures had scores less favorable than the structures found by the genetic algorithm.

2.4. Extending knowledge-based functions to the atomic level

Regardless of the initial fold representation used, protein structures are most useful when detailed atomic coordinates are known. While simplified scoring functions are capable of distinguishing near-natives from non-natives a significant fraction of the time, they will not work as well in situations where subtle differences between different conformations exist. To capture the finer details of atom-atom interactions in protein, such as interactions between side chain atoms and the rest of protein, a more detailed representation is necessary. For example, in a situation where two conformations are quite similar to the experimental structures (within 1-3 Å RMS error for the Cα atoms), we need all the information we can possibly obtain from the two conformations to determine which one is more accurate. A one-point-per-residue scoring function may not be able to discriminate as well as an all-atom discriminatory function, which takes into account the environment of all the atoms of the main and the side chain of each residue.

The all-atom probability discrimination function (PDF) as formulated by Samudrala & Moult [22] is similar to potential of mean force by Sippl [10], but the formulation is in Bayesian terms, and there is greater detail in the representation. 167 different atom types are used. Scores for interactions between pairs of atoms for all 167x167 possible pairs and for 18 distance ranges (0..3,3-4,4-5,...,19-20 Å) are compiled using the expression:

$$s(d_{ab}|C) = \frac{-\ln P(d_{ab}|C)}{P(d_{ab})}$$

$s(d_{ab}|C)$ is the conditional probability of observing two atoms *a* and *b* interacting at a distance *d* in a correct/native conformation C. $P(d_{ab}|C)$ is the probability of seeing atom types *a* and *b* in distance bin *d* in a correct conformation and is calculated by:

$$P(d_{ab}|C) = \frac{N(d_{ab})}{\sum_{d} N(d_{ab})}$$

$P(d_{ab})$ is the probability of seeing atom types *a* and *b* in the distance *d* in any conformation, correct or incorrect:

$$P(d_{ab}) = \frac{\sum_{ab} N(d_{ab})}{\sum_{d}\sum_{ab} N(d_{ab})}$$

$N(d_{ab})$ is the number of occurrences of *a,b* pairs in distance bin *d*.

A scoring function *S* proportional to the negative log conditional probability of conformation being correct is used to calculate the total score of a conformation, given a set of *i,j* interatomic distances:

$$S\left(\left\{d_{ab}^{ij}\right\}\right)=\sum_{ij} s\left(d_{ab}|C\right)$$

The PDF described above avoids sparse data problems by not separating local and non-local interactions. While this leads to an averaging of the two sorts of environments in the parameters for the scoring function, it does not appears to diminish predictive ability [22].

The reference state d_{ab} is in Bayesian terms referred to as a "prior distribution." In this case, the prior distribution is that found in the set of possible compact conformations, with the assumption that averaging over different atom types in experimental conformations is an adequate representation of the random arrangements of these atom types in any compact conformation.

Samudrala & Moult [22] have shown that discrimination between native and non-native folds deteriorates as the detail in fold representation is reduced. To illustrate that point here, we run a detailed all-atom scoring function that takes into account interactions between all 167 pairs of atoms, and another function that uses only Cβ-Cβ interactions, for two sets of protein structure conformations. The first is a set of 269 conformations of 434 repressor (PDB entry 1r69) ranging in RMS deviation (RMSD) from 0.95 to 14.95 Å. The second is a set of "deliberately misfolded structures" created by Holm and Sander [40]. In the latter case, 26 "misfolded conformations" are created by placing the sequences of the proteins on completely different structures of identical length, and then energy minimizing them to make them look as protein-like as possible. These misfolded conformations range from 8.66 to 22.43 Å RMSD with respect to the corresponding native structures.

Table 2 gives the results for the two types of scoring function for 1r69 set of conformations, and Table 3 gives the results for the two types of functions for the misfolded decoy set.

Table 2: Comparison of the all-atom and scoring Cβ-Cβ functions for a set of 269 conformations of 434 repressor (PDB entry 1r69). For each function, the root-mean-square

deviation (RMSD) of the best scoring conformation, the correlation between the scores and the RMSD of the conformation with that score, and Z score, using two different cutoffs to identify near-natives, is given. The detailed all-atom function performs slightly better than the Cβ-Cβ scoring function.

	RMSD of best scoring structure	correlation between score and RMSD	Z score (1 & 2 Å cutoff)
All-atom	1.67 Å	0.80	-1.75/-1.42
Cβ-Cβ	1.80 Å	0.63	-1.65/-1.32

Table 3: Comparison of the all-atom and Cβ-Cβ scoring functions for a set of 26 deliberately misfolded structures. For each function, the percentage of structures correctly discriminated and the average discrimination ratio (score of the incorrect conformation divided by the score of the correct conformation; the lower the ratio, the better the discrimination) is given. The all-atom function performs significantly better than the Cβ-Cβ function.

	% of structures correctly discriminated	Average discrimination ratio
All-atom	100%	0.38
Cβ-Cβ	77%	0.66

Even though in the case of the 1r69 decoy set, the Cβ-Cβ scoring function does quite well, the best scoring conformation selected by the all-atom function is slightly lower in RMS error, and there is a better correlation between the score of the conformation and the RMS error to the native conformation. The Z score for the single conformation below 1.0 Å and the 27 conformations below 2 Å is also slightly better in case of the all-atom function.

Given the results in Table 2, it might seem better to use a reduced representation to speed up the calculation of the fitness of a conformation, since the detailed representation is only slightly better. When we examine the results in Table 3, we see that for the 26 misfolded structures, the all-atom function is able to identify all the 26 misfolded conformations as being incorrect, with a significant degree of discrimination (the ratio is the score of the incorrect conformation divided by the score of the correct conformation; the lower the ratio, the greater the discrimination). However, the Cβ-Cβ scoring function is unable to correctly identify 6 of the 26 structures as being non-native, and the average discrimination ratio is poor relative to the ratio for the all-atom scoring function.

While a function should be able to do more than just discriminate native conformations from non-native ones, this results indicates that in an exhaustive or semi-exhaustive folding simulation, the simplified scoring function is more likely to fail since it is unable to tell a native structure from a conformation that is significantly different in this simple test.

From these, and other similar tests [41], it appears that taking into account as much information as is available in a protein conformation enables one to achieve better near native discrimination. Given that it is not too difficult to generate all-atom models from approximate representations [42, 43], the all-atom scoring function is an useful tool for protein structure prediction.

2.5. Summary

A well-suited scoring function for ab initio folding represents the most native-like conformation as more favorable than all other non-native ones. Current methods do not entirely succeed in this regard, as non-native folds have scores that are as good as the near-native candidates, thereby presenting false positives in exhaustive sampling or traps in minimization. In general, functions that employ compact reference states are more effective when selecting near-native folds from sets of compact folds.

The style of protein structure prediction largely dictates the functional forms and components necessary to compute the score of a conformation. A complete minimization without external constraints generally requires terms that enforce secondary structure and compactness along with pair-specific interactions. However, applying a biased conformational search based on sequence information [21] can greatly reduce the complexity of the energy function necessary to recover a significant number of native-like folds by minimization.

The success of the binary (hydrophobic and polar) functions suggests that most of the specificity of the knowledge-based functions, at least with respect to reduced representations, is due to the frequent occurrence of hydrophobic contacts in the interior of native proteins. However, this success was observed in the context of tertiary fold recognition; the native secondary structure was already in place.

The use of all-atom scoring functions for selecting near-native folds bears promise. To overcome the computational overhead involved in using an all-atom function, one approach could involve sampling large amounts of conformational space using a simplified fold representation and selecting the top scoring conformations using a simple and fast scoring function. All-atom coordinates for these conformations can then be built, and the best conformations selected using the all-atom function. This complementary method of structure prediction would reduce the number of false positives selected by the simplified function and help avoid local minima traps.

3. NOTES

3.1 Generic simplified energy functions

3.1.1. Interaction centers

Contact functions may vary with respect to their designated "interaction centers." Park et al. [34] test contact energy functions that use the Cα as a separate type of interaction center (in addition to the 20 amino acid centroids). It appears that the inclusion of the Cα is detrimental for threading methods as it crudely monitors the local backbone fitness. Since threading methods derive their backbone conformations directly from native structures, the Cα energy terms only add noise to the signal [34].

The placement of a virtual centroid is also arbitary. For instance, one might take the mean projection of side-chain centroids in the database onto the Cα-Cβ vector [13] or the average atomic coordinate centers of all side-chains of a given type [14]. However, the overall performance of a scoring function does not seem to be very sensitive to the placement of a single interaction center.

3.1.2. Distance-dependent energetics

Contact functions are step-functions; when residues are within an arbitrary cutoff distance an energy term is added to the total score. A single cut-off can be applied, as in the case of the Shell function described above. Alternatively, one could define different effective interaction distances depending on the pair of residues [30].

Any “on/off “contact approach may be considered as non-physical because Coulombic and van der Waals interactions smoothly increase and decrease as a function of spatial distance. To address this issue, Park et al. [34] tested a series of functions with pairwise energetics identical to the contact functions, but with Lennard-Jones style functional forms [44]:

$$E = \sum_{ij} \frac{A_{ij}}{r_{ij}^{8}} - \frac{B_{ij}}{r_{ij}^{4}}$$

where A_{ij} and B_{ij} are energy parameters dependent on the contact energy e_{ij} between residues i and j and the effective distance of interaction between i and j. However, the more complex distance-dependent functions did not perform any better than simple contact functions at discriminating near-native folds in the test set described earlier [34].

3.1.3. Multi-body interactions

Most statistical potentials are based on frequencies of pairwise interaction, but functions that include higher-order terms have been developed [25, 45]. A recent study on four-body interactions describes tendencies that cannot be captured by a pair-potential, such as the preference for certain side-chain size combinations in the hydrophobic core [45]. It would be interesting to test the performance of these potentials on the decoy sets described in this chapter.

3.1.4. Reference state

In the Park & Levitt [30] implementation of the solvent-exposed reference state (section 2.2.2), all 210 residue pairwise energies are negative, which means that the formation of new protein-protein contacts is always preferred. Practically speaking, if one were to use a solvent-exposed reference state to fold an polypeptide chain from an extended conformation, a function such as the Contact(MJ) would favor compact conformations and drive chain collapse. However, the drawback of using the solvent-exposed reference state in screening already compact conformations is that the discrimination between the states is weak. Thus, the Shell function, which uses a generic compact shape as a reference exhibits far better performance in the Park & Levitt ab initio test [30]. On the other hand, the Shell function is less adept at recognizing a native fold from an semi-folded, expanded decoy conformation generated by molecular dynamics at high temperature [34], suggesting that this function cannot be used in minimization methods without another term that monitors compactness.

3.2 Histogram function

Park et al. [34] observed that the distance-dependent energies extracted by this function can lead to undesirable results in certain situations. Because the database of proteins used to compile the parameters includes proteins of all sizes, the most-favored inter-residue distances for a given pair do not reflect those of the small proteins that serve as ab initio targets. This implies that if one tries to fold a small protein using only a PMF without an additional term to enforce compactness, then the most-favored structures will be more expanded than the native protein. For example, Simons et al. [21] used a scoring method related to the Histogram function to drive the folding of their small proteins, but also considered the radius of gyration as part of their final objective function.

3.3. Hydrophobic fitness function

This function, which does not require any parameters from the database, performed surprisingly well in most of our tests. However, because of its unusual functional form, is expected to be less amenable for minimization than screening discrete folds. Moreover, since it does not consider disulfide pairings, near-native fold recognition for small proteins that depend on disulfide bridges is noticeably worse than average [34].

3.4. All-atom scoring function

All the inter-atomic distances in the conformation are calculated given a set of coordinates. The number of occurrences of atom pairs at particular distances are stored. This process is repeated for all the coordinate files in the database. Once the raw counts are collated, a table of negative log conditional probability scores for all the 167x167 possible pairs of atoms for the 18 distance ranges [22] is computed (section 2.4).

The all-atom scoring function is susceptible to the problems that plague other knowledge-based functions: (1) the pairwise interactions observed are not independent of each other, (2) lack of sufficient observations to extract the "pseudo-energies" accurately, (3) the choice of a proper reference state, and (4) an averaging of environments. In practice (2) is not a severe problem in this implemention as the function does not use sequence separation, resulting in a greater number of observations in a given distance bin; (3) is chosen for the application at hand: to discriminate compact native conformations from non-native ones; (1) and (4) require taking into account higher-order interactions, which given the size of the current protein data bank [7] leads to sparse data. As a consequence, a compromise must be made between the number of parameters used and the size of the database. Based on our studies on various decoy sets (see below), we feel these compromises are justified.

3.5 Using decoy sets to evaluate scoring functions

Decoys (non-native or near-native conformations) are generally used to test whether a scoring function is useful. While the utility of a function lies in its use in exhaustive or minimization methods, a scoring function has to at least do well in decoy-based tests before it can be considered for simulation. Use of decoys has its pitfalls, the primary one being that there may be artifacts in a particular decoy set that are picked up by a scoring function, resulting in accurate discrimination for that decoy set but not for others. For example, the misfolded decoys described in section 2.4 are slightly expanded relative to the native structure. Thus a simple function that measures the amount of compactness does better than the Cβ-Cβ scoring function with a compact reference state. However, this simple function does not work as well as the Cβ-Cβ function for the 1r69 decoy set.

Thus an "ideal" function is one that discriminates well (100%) for a variety of decoy sets. Adding detail to the function appears to move us closer to this goal [22,41].

4. REFERENCES

1. Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. Science, 181, 223-230.

2. Miyazawa, S. & Jernigan, R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 18, 534–552.

3. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Jr, Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542.

4. Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp. Phys. Comm. 91, 215-231.

5. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. & Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187-217.

6. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Jr, Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. & Kollman, P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117, 5179-5197.

7. Jorgensen, W. and Tirado-Rives, J. (1988) The OPLS potential function for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J. Amer. Chem. Soc. 110, 1657-1666.

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

9. Jernigan, R.L. & Bahar, I. (1996) Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6, 195-209.

10. Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213, 859–883.

11. Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253, 164-170.

12. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. Nature, 358, 86-89.

13. Bryant, S.H. & Lawrence, C.E. (1993) An empirical energy function for threading protein sequence through folding motif. Proteins: Struct. Funct. Genet. 16, 92–112.

14. Kocher, J.-P.A., Rooman, M.J. & Wodak, S.J. (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol. 235, 1598–1613.

15. Godzik, A., Kolinski, A. & Skolnick, J. (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci. 4, 2107-2117.

16. Godzik, A. (1996) Knowledge-based potentials for protein folding: what can we learn from known protein sequences? Structure, 4, 363-366.

17. Thomas, P.D. & Dill, K.A. (1996) Statistical potentials extracted from protein structures: how accurate are they? J. Mol. Biol. 257, 457-469.

18. Ben Naim, A. (1997) Statistical potentials extracted from protein structures: are these meaningful potentials? J. Chem. Phys. 107, 3698-3706.

19. Rooman, M.J. & Wodak, S.J. (1995) Are database-derived potentials valid for scoring both forward and inverted protein folding? Protein Eng. 8, 849-858.

20. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997) Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci. 6, 676-688.

21. Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268, 209-225.

22. Samudrala, R. & Moult, J. (1997) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. (in press).

23. Wodak, S.J. & Rooman, M.J. (1993) Generating and testing protein folds. Curr. Opin. Struct. Biol. 3, 247-259.

24. Moult, J. (1997) Comparison of database potentials and molecular mechanics force field. Curr. Opin. Struct. Biol. 7, 194-199.

25. Godzik, A., Kolinski, A. & Skolnick, J. (1992) Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227, 227-238.

26. Rice, D.W. & Eisenberg, D. (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. 267, 1026-1038.

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

28. Di Francesco, V., Garnier, J. & Munson, P.J. (1997) Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. J. Mol. Biol. 267, 446-463.

29. Defay, T.R. & Cohen, F.E. (1996) Multiple sequence information for threading algorithms. J. Mol. Biol. 262, 314-323.

30. Park, B. & Levitt, M. (1996) Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J. Mol. Biol. 258, 267-392.

31. Hinds, D.A. & Levitt, M. (1992) A lattice model for protein structure prediction at low resolution. Proc. Natl Acad. Sci. USA, 89, 2536–2540.

32. Hinds, D.A. & Levitt, M. (1992) A lattice model for protein structure prediction at low resolution. Proc. Natl Acad. Sci. USA, 89, 2536–2540.

33. Park, B. & Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. 249, 493–507.

34. Park, B.H., Huang, E.S. & Levitt, M. (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J. Mol. Biol. 266, 831-846.

35. Huang, E.S., Subbiah, S. & Levitt, M. (1995) Recognizing native folds by the arrangement of hydrophobic and polar residues. J. Mol. Biol. 252, 709–720.

36. Kolinski, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins: Struct. Funct. Genet. 18, 338-352.

37. Kolinski, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. II. Application to Protein A, ROP, and crambin. Proteins: Struct. Funct. Genet. 18, 353-366.

38. Sun, S., Thomas, P.D. & Dill, K.A. (1995) A simple protein folding algorithm using a binary code and secondary structure constraints. Protein Eng. 8, 769-778.

39. Holland, J.H. (1975) Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor, MI.

40. Holm, L. & Sander, C. (1992) Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225, 93-105.

41. Samudrala, R., Huang, E.S. & Levitt, M. (in preparation)

42. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226, 507-533.

43. Holm, L. & Sander, C. (1991) Database algorithm for generating protein backbone and side-chain coordinates from a Cα trace: application to model building and detection of coordinate errors. J. Mol. Biol. 218, 183–194.

44. Wallqvist, A. & Ullner, M. (1994) A simplified amino acid potential for use in structure prediction of proteins. Proteins: Struct. Funct. Genet. 18, 267-280.

45. Munson, P.J. & Singh, R.K. (1997) Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci. 6, 1467-1481.

5. ACKNOWLEDGEMENTS

We thank Michael Levitt for his support. Part of this work were supported by NIH Grant GM45514.