

Influence of Protein Structure Databases on the Predictive Power of Statistical Pair Potentials

Emiko Furuichi and Patrice Koehl*
CNRS, Illkirch Graffenstaden, France

ABSTRACT A long standing goal in protein structure studies is the development of reliable energy functions that can be used both to verify protein models derived from experimental constraints as well as for theoretical protein folding and inverse folding computer experiments. In that respect, knowledge-based statistical pair potentials have attracted considerable interests recently mainly because they include the essential features of protein structures as well as solvent effects at a low computing cost. However, the basis on which statistical potentials are derived have been questioned. In this paper, we investigate statistical pair potentials derived from protein three-dimensional structures, addressing in particular questions related to the form of these potentials, as well as to the content of the database from which they are derived. We have shown that statistical pair potentials depend on the size of the proteins included in the database, and that this dependence can be reduced by considering only pairs of residue close in space (i.e., with a cutoff of 8 Å). We have shown also that statistical potentials carry a memory of the quality of the database in terms of the amount and diversity of secondary structure it contains. We find, for example, that potentials derived from a database containing α -proteins will only perform best on α -proteins in fold recognition computer experiments. We believe that this is an overall weakness of these potentials, which must be kept in mind when constructing a database. *Proteins* 31:139–149, 1998. © 1998 Wiley-Liss, Inc.

Key words: protein structure; statistical potentials; protein structure database; assessing protein models

INTRODUCTION

It is the ability of proteins to fold into unique three-dimensional (3-D) structures that allows them to function. Hence comprehension of the way in which the amino acid sequence determines the 3-D conformation of the native protein is essential to an understanding of biological processes. Should the laws of folding be known, protein structure prediction, de novo design of complicated protein folds, and

even the prediction of a protein's interactions with its environment would become tractable. The properties of proteins are directly related to their potential energy surfaces, with the native structure corresponding to the minimum of this surface. The challenge for theoretical biology is to obtain a good approximation of the true potential energy function and to derive methods for reaching the minimum of this function. Considerable efforts and progress have been made in recent years on both problems^{1–6}; this paper is concerned with the derivation of energy functions.

When energy is a critical quantity, biomolecular simulations rely essentially on accurate potential energy functions, or force fields, in which the energy parameters are traditionally optimized with respect to the properties of small model molecules (though data are now often extracted from quantum chemical calculations⁷). These force fields, based on full-atom representations of the protein structures, are commonly used in molecular mechanics and molecular dynamics studies, but remain impractical, for computing reasons, for studying issues implying long-time scales, such as protein folding. For such purposes, simplified models of protein structure seem more practical. These low-resolution methods require specific potential energy functions, most often derived from known protein structures. Databanks of protein structures are growing in size; it is therefore expected that potentials derived from them will become more and more accurate. Several methods have been proposed to extract information from databanks in the form of statistical (or knowledge-based) potentials, which can then be used for studying proteins whose structure is unknown.^{3,8,9} Among these methods, two approaches prevail. In the first, knowledge-based potentials are optimized to fit to known protein sequence and structure data. Maiorov and Crippen¹⁰ defined energy parameters that solve the threading problem, i.e., such that the native structure for a given sequence is clearly recognized from among a large set of decoys. Goldstein et al.¹¹

Emiko Furuichi's permanent address is Fukuoka Women's Junior College, 4-16-1 Gojo, Dazaifu, Fukuoka 818-01, Japan.

*Correspondence to: Patrice Koehl, UPR 9003 du CNRS, Pole API, Boulevard Sebastien Brant, 67400 Illkirch Graffenstaden, France. E-mail: koehl@bali.u-strasbg.fr

Received 3 July 1997; Accepted 24 October 1997

have developed an analytical method based on spin glass theory for determining the parameters that maximize the stability of the native protein structure relative to an average alternative structure, i.e., the foldability of the protein sequence in their terminology. This approach has been extended recently, with various definitions of the foldability.¹²⁻¹⁷

In the second approach, putative energy terms have been derived from amino acid pairing frequencies observed in the known protein structures, as initially proposed by Tanaka and Scheraga,¹⁸ and subsequently extended by Miyazawa and Jernigan¹⁹ explicitly to account for solvent effects, and also by Sippl et al.,^{20,21} who included the dependence on separation of residues in both space and sequence. These potentials are referred to as either log-odd potentials (if statistics alone are considered) or potentials of mean force (if a physical model based on statistical mechanics is considered). Further developments have been concerned with the introduction of new energy functions based on other statistical terms, such as residue triplets,^{22,23} solvent accessibility,²⁴ atomic environment,²⁵⁻²⁷ dihedral angle preferences,²⁸ ion pairs,²⁹ and hydrogen bonding,^{27,30} as well as on studies that try to define the best possible combination of these terms.³¹ Most of these potentials (derived from both methods) have been used for error recognition in protein models derived from X-ray data, NMR data, or modelization,^{32,33} for ab initio folding calculations,^{23,34,35} as well as for the threading or fold recognition problem,³ with the latter being commonly recognized as the first test to be performed to assess the quality of a given set of potential.

The basis on which statistical potentials are derived have been questioned. Recently, using simple lattice models, Thomas and Dill³⁶ have investigated both the principles of statistical potentials and the extent to which these potentials reflect real amino acid contact energies in proteins. Their results clearly identify problems with these potentials. In this article, we describe an extension of their studies in which we investigate potentials derived from protein 3-D structures directly, and in particular we address questions of database size, database content, and sampling problems.

FORMALISM

Let X be a state variable of a physical system in equilibrium. The probability that X takes the value x is given by the Boltzmann law:

$$P(X=x) = \frac{\exp\left[-\frac{E(x)}{kT}\right]}{Z} \quad (1)$$

where $E(x)$ is the energy of the system, k the Boltzmann's constant, T the temperature, and Z the

partition function. If x^* is the value X such that the interaction energy of the system is zero, then

$$P(X=x^*) = \frac{1}{Z}. \quad (2)$$

Combining Equations (1) and (2), we obtain

$$P(X=x) = P(X=x^*) \exp\left[-\frac{E(x)}{kT}\right]. \quad (3)$$

Conversely, if the probability density functions P can be measured, we can derive the energy of the system described by x from

$$E(x) = -kT \ln \left[\frac{P(X=x)}{P(X=x^*)} \right]. \quad (4)$$

$E(x)$ given in Equation (4) is the so-called potential of mean force.

In the case of proteins, potentials of mean force are mainly concerned with pairwise interaction energies between any two amino acid types a and b (from 1 to 20); in this case X is a distance functional, and $P_{ab}(X=r)$ is the marginal probability of observing two amino acids of type a and b at a distance r . If the amino acids are considered to float in a dilute idealized gas phase in which pairwise interactions dominate, Equation (4) is valid and becomes

$$E_{ab}(r) = kT \ln \left[\frac{P_{ab}(X=r)}{P_{ab}(X=+\infty)} \right] \quad (5)$$

since the interaction energy between a and b is zero at infinite separation only.

In a first approximation, the marginal probabilities P_{ab} are derived from the discrete frequencies observed in different proteins from the PDB databank. The definition of the reference state is not straightforward; this is sometimes referred to as the partition function problem (as seen from Eq. (2)). A zero interaction between two residue type a and b cannot be defined as such since two residues will always be located at a finite distance from each other. Miyazawa and Jernigan¹⁹ defined a statistical potential that only considered local interaction, i.e., such that the probabilities of amino acid pair occurrence are integrated over a range of distance $[0, r_c]$, where r_c is a small cutoff value (6.5 Å). The main strength of their potential was that they included an explicit treatment of solvent. In order to define the corresponding reference state, they introduced the 'random-mixing approximation,' in which amino acids and solvent distribute uniformly in the absence of interactions. In this random mixture, contacts only depend on the concentration of the amino acids, hence their normalization scheme basically cor-

rected for amino acid pair concentration. Sippl²⁰ constructed distance-dependent pair potentials, which included correction for amino acid pair concentrations in terms of amino acid types (e.g., ALA–GLY versus MET–TRP), as well as in terms of the distribution of distances observed (i.e., number of pairs observed at 10 Å vs. those observed at 50 Å). Sippl's formalism may be summarized as follows. Let $P(X = r)$ be the probability of observing any type of amino acid pair at distance r , and T the state variable that defines the amino acid pair type, then

$$P_{ab}(X = r) = \frac{P(X = r/T = (a, b))}{P(T = (a, b))}. \quad (6)$$

Using Equation (6), Equation (5) in Sippl's formalism becomes

$$E_{ab}(r) = -kT \ln \left[\frac{P(X = r/T = (a, b))}{P_{ab}(X = +\infty) P(T = (a, b))} \right]. \quad (7)$$

A similar expression prevails for the potential for an unspecific amino acid type, noted $E(r)$. The distance-dependent pair potential of Sippl is the net potential $\Delta E_{ab}(r)$, defined as

$$\begin{aligned} \Delta E_{ab}(r) &= E_{ab}(r) - E(r) \\ &= -kT \ln \left[\frac{P(X = r/T = (a, b))}{P(X = r) P(T = (a, b))} \right] \\ &\quad - kT \ln \left[\frac{P(X = +\infty)}{P_{ab}(X = +\infty)} \right]. \end{aligned} \quad (8)$$

The second term of the right-hand side of Equation (8) is independent of conformation, and only depends on the nature of a and b . For a given sequence, the sum of these terms is constant, and hence can be ignored in a fold recognition computer experiment (in which the same sequence is threaded through a set of protein folds), yielding a modified net potential:

$$\Delta E_{ab}(r) = -kT \ln \left[\frac{P(X = r/T = (a, b))}{P(X = r) P(T = (a, b))} \right]. \quad (9)$$

The net potential energy $\Delta E'(S,C)$ of a protein of sequence S that adopts a conformation C is then assumed to be the sum of the individual residue pair contributions, based on Equation (9). Rooman and Wodak³⁷ have shown that $\Delta E'(S,C)$ directly approximates the difference between the free energy of S in conformation C and that of S in a denatured-like state, and can thus be used even if the amino acid composition is not fixed.

But protein structures may not be considered as amino acids in the gas phase, and the assumptions

and approximations required to apply Equation (5) (or the modified forms proposed by Miyazawa and Jernigan¹⁹ as well as Sippl²⁰) may be too crude to maintain a physical meaning to this formalism. The physical premises of these potentials have been tested in detail in the study by Thomas and Dill.³⁶ Briefly, they have shown that these potentials do not reflect the true underlying energy of proteins, mainly because the assumption that frequencies of amino acid pairs are not independent of each other is not valid (amino acids are covalently linked in specific sequences). As a consequence of this study, we will refer to potentials calculated from Equation (9) as statistical potentials rather than potentials of mean force.

Thomas and Dill's research³⁶ was based on 2-D lattice models for which both configurational and sequence space could be explored systematically, hence they did not address questions of database size. The sampling problem is twofold. First, certain amino acid pairs are underrepresented in natural sequences, yielding poor statistics for estimating the corresponding probability density functions. This is certainly the case for any pairs involving tryptophan or methionine. This problem may not be crucial if the sparse data correction scheme of Sippl²⁰ is used. Second, the composition of the database itself may play an important role. Amino acids are covalently linked and form specific secondary structures, providing coherence in the relative positions of the amino acid pairs, which is again in opposition to the assumption that amino acid pairs are independent of each other (this assumption is required to apply the Boltzmann equation). The quantity as well as the quality (i.e., the types) of secondary structures that form the proteins contained in a specific databank may then have an effect on the quality of the potentials derived from this databank. These are the issues we address below. Since we are looking at structural effects, we have selected the distance-dependent statistical potentials originally described by Sippl,²⁰ to which we added a surface statistical potential based on surface accessibility^{24,38} to account for hydrophobic effects.

METHOD

We aim to learn how much bias is introduced in statistical potentials when they are extracted from databases with poor sampling in some structural aspects such as secondary-structure elements. Our procedure is defined as follows. First, we define four different databases, each containing one type only of three-dimensional fold (α , β , α/β , or $\alpha+\beta$). Second, we extract distance-dependent pairwise potentials as well as a surface potential to include solvent effect from all four databases. Each of these potentials are tested in fold recognition computer experiments, and the results are classified according to the folding type of the tested protein.

Protein Structure Database

The complete database of protein structures used in the present study consists of 125 nonhomologous proteins (in fact 125 chains) extracted from release 70 (October 1994) of the Brookhaven PDB.³⁹ This database has been subdivided into four subdatabases which we denote as A, B, AB, and A+B for α , β , α/β and $\alpha+\beta$ proteins, respectively, according to the protein classification proposed by Orengo et al.⁴⁰ All four databases, A, B, AB, and A+B, have approximately the same total number of residues. For reference, a fifth database which we denote as F was prepared randomly, which contains (approximately) the same number of residues as A, B, AB, and A+B, but with equal representation of the four folding type. The five databases are described in Table I.

The set of proteins chosen is not complete in that it does not include all protein folds known to date. This is not important since we are not testing sampling problems related to the size of the database, but those related to the quality of the database.

Extracting Statistical Potentials

Distance-dependent statistical potentials were extracted by the method of Sippl,²⁰ based on Equation (9). For two residues i and j in a protein,

$$\Delta E_{a_i a_j}^k(r) = -RT \ln \left[\frac{P_{a_i a_j}^k(r)}{P^k(r)} \right] \quad (10)$$

where a_i and a_j correspond to the amino acids at position i and j in the protein, respectively; $k = j - i$ is the topological level, or separation of these residues along the sequence; and r is the spatial distance between the C_α atoms of i and j . RT is a constant, taken to be $0.582 \text{ kcal mol}^{-1}$ (R is used rather than k so as to work with mol rather than individual molecule, while T is set to 293 K). The pair interactions are represented by several variants of potentials, depending on the topological level k . In the short range ($1 \leq k \leq 6$), individual potentials are compiled for each value of k . For medium- ($7 \leq k \leq 9$), and long-range ($10 \leq k$) separations, the pair potentials are condensed to a single type of potential.

Scanning the database yields density distributions F rather than probabilities P , as required in Equation (10). For each amino acid pair and each topological level, the shortest (R_{\min}) and largest (R_{\max}) distances between the corresponding C_α in the database are recorded. R_{\max} is then set to $\min(R_{\max}, R_c)$, where R_c is a cutoff distance. The complete distance range $[R_{\min}, R_{\max}]$ is then divided into 20 intervals of equal size. Because of the large number of possible potential bins (i.e., 20 distance intervals, 8 topological levels, and 400 amino acid pairs, giving a total of 64,000 different potential values to be considered), the number of observations for each bin is small, hence these densities cannot be used as such. To

TABLE I. Protein Structure Databases*

Database	Fold type	Proteins [†]
A	α -proteins	1aca, 1aep, 1lbbh, 1c5a, 1cc5, 1acol, 1ecd, 1gmf, 1hbg, 1lhdd, 1lith, 1le2, 31lmb, 1mba, 1mbc, 1lprc, 1r69, 1rcb, 1lrop, 1ycc, a256b, a2ccy, 2cy3, a2hmz, 2lh7, 2lhb, 2sas, a2scp, p2tmv, a2utg, r2wrp, a3sdp, 451c, 4cpv, 4icb, a4sdh
B	β -proteins	1acx, 1lbbp, 1cd8, 1lcob, 1epg, 1f3g, 1hcc, 1lfc, 1lnsb, 1paz, 1rbp, 1ltnf, e2er7, h2fb4, 2gcr, b2hla, a2ltm, a2pab, 2por, 2rhe, a2tlv, 3cd4, 3cna, 4fgf, a5hvp, 5rxn, e5sga, 7pcy
AB	α/β -proteins	1ald, 1ego, o1gd1, 1gky, 1trb, 2fcr, 2gbp, a2trx, 3adk, 3chy, 3dfr, 3pgm, 3trx, a4dfr, 4fxn, 5p21, a5tim, 6cpa, 6ldh, 8abp, 8dfr
A + B	$\alpha + \beta$ -proteins	1aak, 1ab2, 1aps, 1lbov, 1bp2, 1ctf, 1crn, 1lscse, 1eaf, 1fdx, 1lff, 1fxd, a1fxi, a1gat, a1il8, a1lts, 1lz1, a1msb, 1ppy, a1rnb, a1rve, 1sn3, 1snc, i1tgs, a1tpk, 1ubq, a2bop, 2gb1, 2ovo, a2sar, i2sic, a2tsc, 3b5c, 3cla, 3lzm, 4pti, 7rsa, b8atc, 9rnt, a9wga
F	all	1aak, 1aps, 1cd8, 1acol, 1fxd, a1gat, 1gky, 1lhdd, 31lmb, a1nsb, a1rnb, 1sn3, 1ycc, 2gbp, 2lhb, a2ltm, 2ovo, a2pab, 2sas, a2tlv, 3cla, 3lzm, 451c, a4dfr, a4sdh, e5sga, a5tim, 8abp, b8atc

*The full database contains all protein chains from the four databases A, B, AB and A + B; F is a random subset of the full database, in which all four fold types are represented.

[†]Brookhaven databank codes. If available, chain identifiers are given first.

correct for this sparse data problem, Sippl²⁰ proposed the following approximations:

$$P^k(r) \approx F^k(r) \quad (11)$$

and

$$P_{ab}^k(r) \approx \frac{1}{1 + m\sigma} P^k(r) + \frac{m\sigma}{1 + m\sigma} F_{ab}^k(r) \quad (12)$$

where the F values are the observed frequencies in the database of protein structures, m is the number of pairs (a,b) observed at distance r and topological level k , and σ is the weight given to each observation.

In all subsequent calculations, σ is set to 1/50, i.e., such that for $m = 50$, $F^k(r)$ and $F_{ab}^k(r)$ have the same weight.

The distance cutoff R_c was set to account for intramolecular interactions. To account for protein surface-solvent interactions, Sippl³⁸ proposed a simple model based on neighborhood calculation. In this study, a sphere of radius 12 Å is centered at the C_α atom of a particular residue and the number of residues N within this sphere is calculated. Following Equation (9), the statistical potential for solvent interaction for amino acid type a is defined as³⁸

$$\Delta O_a(n) = -RT \ln \left[\frac{P(\text{TYPE} = a, N = n)}{P(\text{TYPE} = a) P(N = n)} \right] \quad (13)$$

where TYPE is the variable that describes the amino acid type, and N the variable for the number of residues in the sphere of interaction. The probabilities P are derived directly from the observed frequencies in the protein structure database, without statistical correction.

The pair interaction energy $E_{\text{pair}}(S, C)$ for a sequence S in conformation C is obtained by summing Equation (10) over all interacting pairs with a spatial distance less than R_c . Similarly, the surface energy $E_{\text{surf}}(S, C)$ is obtained by summing Equation (13) over all amino acids. A combined energy $E_{\text{tot}}(S, C)$ is defined as

$$E_{\text{tot}}(S, C) = E_{\text{pair}}(S, C) + \frac{\sigma_{\text{pair}}}{\sigma_{\text{surf}}} E_{\text{surf}}(S, C) \quad (14)$$

where σ_{pair} and σ_{surf} are the standard deviation of the corresponding energies.

Testing the Potentials: The Hide-and-Seek Procedure

In the hide-and-seek computer experiment, the native fold X for a sequence S is hidden among a large number of nonnative folds, C , and the task is to retrieve X using the potential function to be tested as a guiding criterion. The task is successfully solved if $E(S, X)$ is lower than any of the $E(S, C)$. This procedure is in spirit similar to the so-called threading problem, though it does not allow for gaps. In this study, the hide-and-seek experiment is performed on a polyprotein⁴¹ constructed from the set of 125 protein structures in our full database (see above). The structures are joined using short fragments from protein structures, with the requirement that there are no close contacts between modules. The total length of our polyprotein is in the order of $L \approx 20,000$ residues. The sequence S of a given protein of length N is shifted along the polyprotein, yielding a set of decoys for S . The set D of conformations C obtained represents the conformational space accessible to S . This set contains $L - N + 1$ conformation,

and since $N \ll L$, the total number of conformations is essentially independent of N . The potential energies $E_{\text{pair}}(S, C)$, $E_{\text{surf}}(S, C)$, and $E_{\text{tot}}(S, C)$ are calculated for each C in D , as well as for the native conformation X . The quality of the statistical potentials derived from various databanks can then be estimated by a z-score:

$$z = \frac{E(S, X) - \langle E(S, C) \rangle}{\sigma} \quad (15)$$

where $\langle \rangle$ represents the average over all conformations C in D , and σ the corresponding standard deviation. If the discrimination is significant, z is expected to be large and negative (in other works, the signs in the definition of z are permuted, in which case z is expected to be large and positive).

The predictive power of a given statistical potential is expressed in terms of the average z-score, $\langle z \rangle$, calculated for a subset of the known native folds in a given database by

$$\langle z \rangle = \frac{1}{M} \sum_{i=1}^M z_i \quad (16)$$

where z_i is the individual z-score for protein i obtained from Equation (15), and M is the number of proteins included in the subset considered.

Kocher et al.⁴² have shown that the database of known structures, even assembled in a polyprotein, is a poor challenge for many empirical energy functions. As discussed by Park and Levitt,³¹ decoy structures should in fact include structures that are close to the native structure, be native-like in all properties but the overall fold (otherwise they may be distinguished trivially), and be diverse and numerous. All these criteria apply when an absolute test of a new potential is performed, in order to compare its power with other potentials. In this study, however, the quality of the reference set of decoy structures is not crucial, since we are performing relative tests on similar statistical potentials that differ only in the database from which they were extracted.

In the subsequent calculations, we will refer to two different types of 'database,' i.e., the set of proteins used to derive the statistical potentials, and the set of test proteins used in the hide-and-seek experiments to derive the average z-score. To avoid confusion, we will refer to the first set as a *database* of protein structures and to the second as a *pool* of known native folds.

RESULTS

Setting Up the Potential: Definition of the Cutoff Distance R_c

All pairwise potentials similar to those described by Sippl²⁰ ignore interactions beyond a certain distance cutoff value R_c , on the basis that the interac-

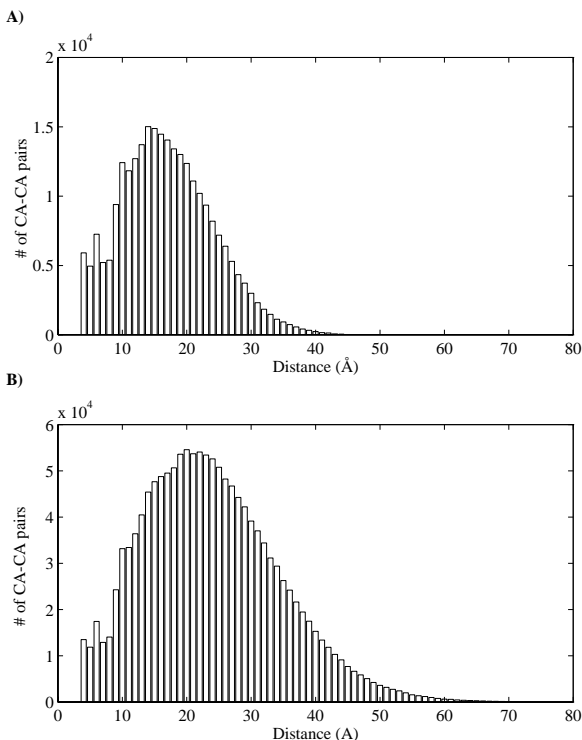


Fig. 1. Distributions of interresidue C_{α} - C_{α} distances in (A) database S of short proteins (i.e. smaller than 130 residues) and (B) database L of long proteins (i.e. longer than 130 residues). Database S contains 1ab2, 1aca, 1acx, 1aps, a1bov, 1bp2, 1c5a, 1cc5, 1cd8, 1cm, 1ctf, 1ego, 1epg, 1fdx, 1fkf, 1fxd, a1fxi, a1gat, a1gmf, 1hcc, c1hdd, a1il8, 311mb, d1lts, a1msb, 1paz, 1r69, a1rnb, a1rop, 1sn3, i1tgs, a1tpk, 1ubq, 1ycc, a256b, a2bop, a2ccy, 2cse, 2cy3, 2gb1, b2hla, a2hmz, 2ovo, a2pab, 2rhe, a2sar, i2sic, a2trx, a2utg, r2wrp, 3b5c, 3chy, 3trx, 451c, 4cpv, 4fgf, 4icb, 4pti, a5hvp, 5rxn, 7pcy, 7rsa, and 9rnt. Database L contains 1aak, 1aep, 1ald, a1bbh, a1bbp, a1cob, a1col, 1eaf, 1ecd, 1f3g, o1gd1, 1gky, 1hbg, 1lfc, a1ith, 1le2, 1lz1, 1mba, 1mbc, a1nsb, c1prc, 1pyp, 1rbp, 1rcb, a1rve, 1snc, a1tnf, 1trb, e2er7, h2fb4, 2fcr, 2gbb, 2gcr, 2lh7, 2lhb, a2ltn, 2por, 2sas, a2scp, a2tbv, p2tmv, a2tsc, 3adk, 3cd4, 3cla, 3cna, 3dfr, 3lzm, 3pgm, a3sdp, a4dfr, 4fxn, a4sdh, 5p21, 5sga, a5tim, 6cpa, 6ldh, 8abp, b8atc, 8dfr, and a9wga.

tions are not residue-specific and are determined simply by solvation effects. Sippl and Jaritz⁴¹ have studied the dependence of the performance of their pairwise statistical potentials on the distance cutoff. For a given database containing 68 proteins, they have shown that the predictive power of their potentials, as measured by the average z-score, was low for short truncation distances ($R_c < 10 \text{ \AA}$), and increased slowly but steadily for cutoff distances up to 30 Å, where it reached a maximum. For R_c values larger than 30 Å, (z) remained rather flat. In terms of successful identification of native folds, they have shown that truncation at a distance of 20 Å yields the best result; this is the value Sippl et al. have used.^{38,41} Interestingly Jones et al. used a value of 10 Å.²⁴

The cutoff dependence is directly related to the quality of the database used, more specifically to the

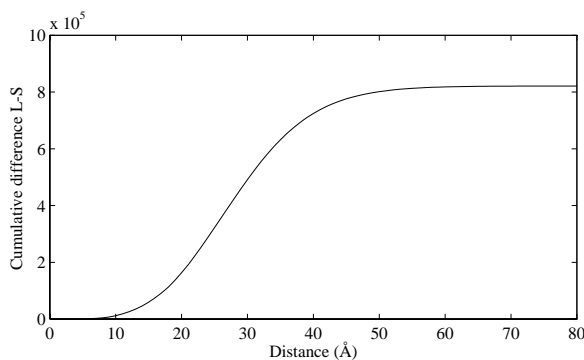


Fig. 2. Integral difference D between the two distributions S and L shown on Figure 1A and B, respectively. Distribution S was first scaled by a constant (yielding S') such that the first nonzero bins of both distributions are equal. The i^{th} bin of D is then defined as $D_i = D_{i-1} + L_i - S'_i$.

amount of small and large proteins it contains. Small proteins have few distances greater than 35 Å compared to large proteins. Also, the number of pair interaction increases with the square of the sequence length, hence large proteins bias the potentials. To investigate this problem, we divided our full database into two subsets, S and L, where S contains all protein chains with less than 130 residues, and L all remaining proteins. The distribution of C_{α} - C_{α} distances for both S and L are shown in Figure 1A and B, respectively. These two distributions are clearly different, with much more long-range C_{α} - C_{α} distances for L than for S. Interestingly, the two distributions are reasonably similar up to approximately 10 Å, as illustrated on Figure 2, indicating that the short-range interactions in protein are independent of protein size, while obviously very-long-range interactions are only present in large proteins. In order to be independent of the quality of the databases, pairwise potentials should then be considered only for short cutoff values.

To test this hypothesis further, pairwise potentials were calculated from each database S and L for various cutoff values, and used in hide-and-seek computer experiments with a pool of 40 test proteins (using the jack-knife procedure, i.e., the protein under test in the hide-and-seek procedure is removed from the database prior to the calculation of the potentials). The dependence of the average z-score (z) on distance cutoff values for databases S and L are shown on Figure 3. The results for both S and L are in agreement with the observations of Sippl and Jaritz, i.e., a significant improvement in predictive power is observed for cutoff values varying from 5 to 10 Å, followed by a steady but slow increase for $R_c > 10 \text{ \AA}$. A difference between S and L is observed, however. For small cutoff values ($R_c < 7-8 \text{ \AA}$), the predictive power of the statistical potentials derived from the database of small proteins (S) and large proteins (L) are nearly identical. A significant differ-

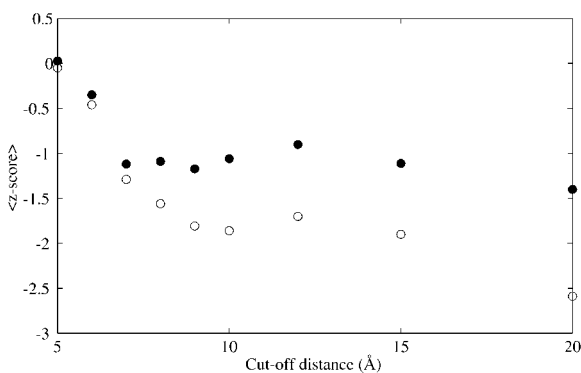


Fig. 3. Dependence of the predictive power of statistical pair potentials on the C_{α} - C_{α} distance cutoff: (○), statistical potential derived from large proteins only (database L), and (●), statistical potentials derived from small proteins only (database S; see legends of Fig. 1).

ence is detected, however, for larger values of R_c , in which case the average z-scores for (L) are better than those derived from (S). By increasing the distance cutoff, the corresponding amount of amino acid pairwise information introduced in the potentials is greater for large proteins than for short proteins, since small proteins have very few long interresidue distances. Though these pairwise potentials are not true mean force potentials as described by Thomas and Dill,³⁶ we expect them to be reasonably independent of the quality of the database in terms of the number of large proteins it contains. A compromise must be considered, since R_c should be large enough to increase the predictive power, but not too large as to be independent of the quality of the database. Based on Figure 3, we will rely on a value for R_c of 8 Å.

Illustration of the Problem: The Jack-Knife Procedure

The statistical potentials used in this study provide a measure of the correlation between pairs of amino acids and their spatial organization (Eq. (10)), or between an amino acid type and its solvent environment (Eq. (13)). It is expected that these correlations will suffice to identify the correct fold of a protein out of a large series of decoys, and this is tested in the hide-and-see experiment. Correct application of this procedure requires that the fold of the protein under test not be included in the database of proteins used to derive the potentials that evaluate the structure-sequence fitness; this is the so-called jack-knife procedure.

Theoretically, potentials of mean force are derived from the hypothesis of independence of all types of amino acid pairs (for pair potential), or of all types of amino acids (for surface potential), and thus should not include any memory of a specific fold. To assess this hypothesis, hide-and-see experiments were performed on two proteins (PDB code 2pab and 1gd1,

containing 114 and 334 residues, respectively), using the combined statistical potentials given in Equation (15), removing or not the corresponding folds from the database used to derive the statistical potentials. The size of the database may play a role here, in that the information corresponding to a given fold might be diluted if a large database is considered, making the jack-knife procedure less critical than in the case of a small database. Results are presented in Figure 4. Using the jack-knife procedure, the z-score decreases, i.e., the discriminative power of the potential increases as the size of the database increases. This has already been described by Sippl and Jaritz.⁴¹ When no precautions are taken, i.e., the native fold is included in the database, results differ significantly. First, z-scores obtained from this scheme are always significantly better than those obtained using the jack-knife procedure. This improvement in the discriminative power of the potentials is artificial in the sense that it only provides evidence of the structural memory of the potentials. Second, the apparent discriminative power decreases as the size of the database is increased, as expected from the dilution effect.

To eliminate this bias while checking other possible sources of structural memory in the potentials, all subsequent calculations were performed using the jack-knife procedure.

Effects of the Protein Database Content

The jack-knife tests presented above have shown that the statistical potentials described here retain a memory of the complete fold of a protein. We decided to investigate whether the amount of smaller structural elements of the proteins in the database, such as helices and strands, also influence the information content of the statistical potentials. Secondary-structure-specific statistical potentials derived from the five databases A, B, AB, A+B, and F, which contain α proteins, β proteins, α/β proteins, $\alpha+\beta$ proteins, and all types of proteins equally represented, respectively (see Methods above), were tested in hide-and-see experiments on pools of proteins also distinguished in terms of their overall fold. Both the effects of the content of the database from which the statistical potentials are derived, and the importance of the nature of the protein tested in the hide-and-see experiment, are investigated here. All results are summarized in Table II.

The results are astonishingly clear cut in the case of α and β proteins. First, statistical potentials derived from an all- α protein database perform better on α -protein than any other potentials. This is illustrated on Figure 5A, in which databases A and B are compared. Similarly, potentials derived from database B perform better on β -proteins (Fig. 5B). Secondly, the predictive power of the statistical potentials tested on α -proteins and β -proteins are

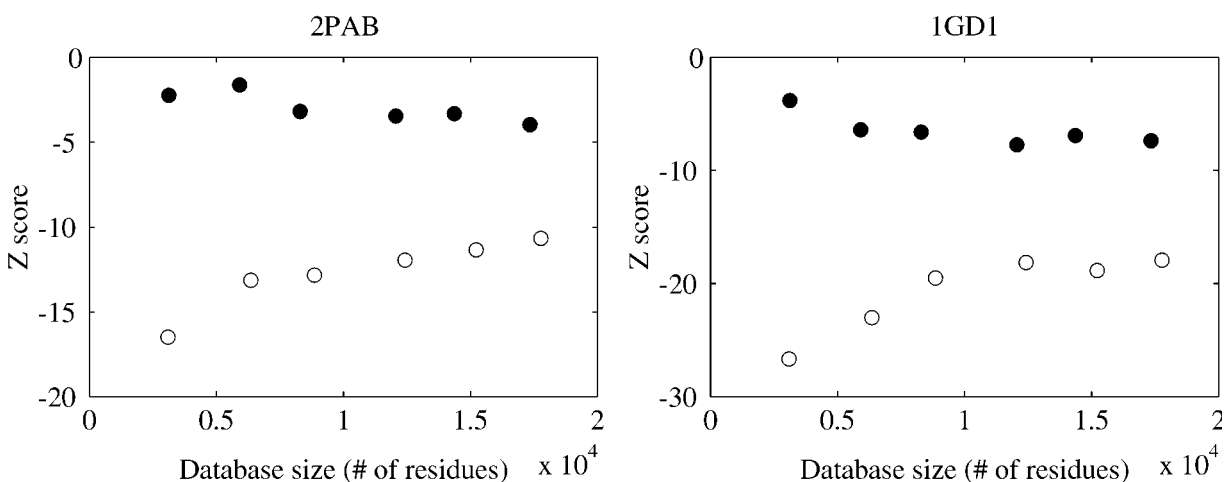


Fig. 4. Two examples of the dependence of the predictive power of statistical potentials in hide-and-seek experiments on the size of the database, using (●) or not (○) the jack-knife procedure (which consists of removing the protein to be tested from the database used to derive the potentials; see text for details).

TABLE II. Dependence of the Predictive Power (Measured by $\langle z \rangle$) of Statistical Pair Potentials on the Database They Are Derived From, and on the Test Proteins Considered

/Test database/proteins	α - proteins	β - proteins	α/β - proteins	$\alpha + \beta$ - proteins
Database A	-3.03	-0.79	-3.13	-1.55
Database B	-1.69	-3.03	-2.99	-2.12
Database AB	-3.03	-2.35	-4.32	-2.09
Database A + B	-2.86	-2.37	-3.74	-2.27
Database F	-2.95	-2.42	-3.90	-2.46

directly correlated to the helix or strand content of the database from which they are derived (Fig. 6).

Statistical potentials derived from the database AB also perform slightly better than the other potentials on a pool of α/β proteins, but the bias is much less than that observed for statistical potentials derived from database A and tested on a pool of α -proteins, for example. Interestingly, the predictive power of the statistical potentials tested on α/β proteins decreases for low and high contents of both α -helices (Fig. 7A) and β strands (results not shown). This suggests that the presence of both types of secondary structures in the database used to derive the potentials is required for a good identification of α/β protein.

A similar behavior is observed when the statistical potentials are tested on $\alpha + \beta$ proteins, though the requirement of the presence of α -helices in the database is less clear (Fig. 6). The positions of helices and strands in $\alpha + \beta$ proteins are not fold-specific, hence databases with similar contents of α -helices and β -strands should yield potentials with similar predictive power on $\alpha + \beta$ proteins. This was observed here (Fig. 7B) to the extent that the best predictive power is observed for database F.

DISCUSSION

Statistical potentials are now widely used as empirical energy functions to assess protein structure models, for protein fold recognition (i.e., the 'sequence recognizes structure,' or threading, problem), and for ab initio protein folding experiments. Many parameters have been considered as state variables to define these potentials, including local geometry (dihedral angles or pseudodihedral angles between C_α), short-range amino acid contacts, contacts with the solvent, radial distribution of amino acid pairs, charge distribution, and atomic environment. There is a general belief that since the number of known protein structures has increased greatly in recent years, and is expected to grow even faster in the near future, these potentials will become more and more accurate. Miyazawa and Jernigan⁴³ recently reevaluated their potentials using a database of 1,661 protein structures (compared to 41 in their original study). Interestingly, they showed that the larger database did not provoke substantial changes in the statistical potentials themselves, but rather confirmed their validity.

The advantages of statistical potentials are clear.^{2,8} They include the essential features of protein structures as well as solvent effects at a low computing cost. Because they are fast to compute, they allow better sampling of the conformational space or sequence space in computer folding or inverse folding experiments. However, the way in which these statistical potentials are derived have been questioned,^{36,37,42} raising doubts as to their value as energy-like quantities. Thomas and Dill³⁶ have drawn attention to systematic errors arising from the neglect of chain connectivity and excluded volume. In particular, they have shown that statistical poten-

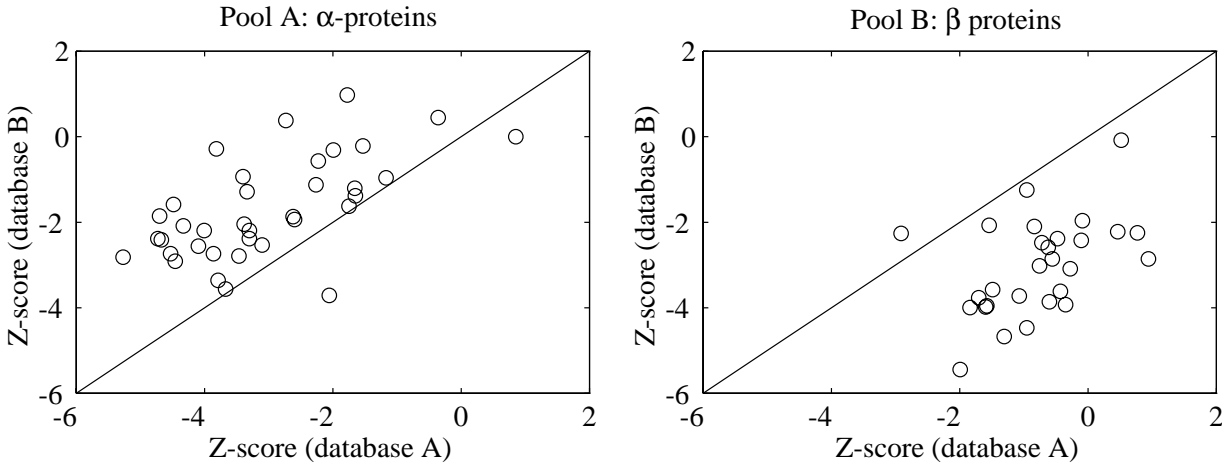


Fig. 5. Comparison of the predictive powers of statistical potentials derived from database A (containing mainly α -proteins) and from database B (β -proteins) in hide-and-seek experiments applied on a pool of α -proteins and a pool of β -proteins. The first diagonal is shown for sake of clarity.

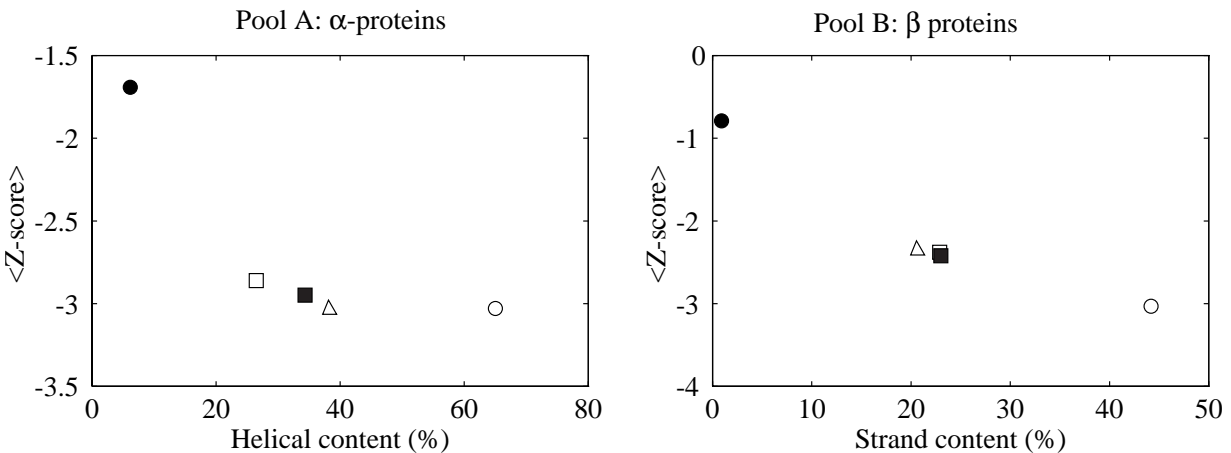


Fig. 6. Average predictive power of statistical pair potentials for recognizing α -fold (respectively β -fold), versus the helical (respectively strand) content of the database from which they are derived. (●) denotes database B; (□), database A+B; (■), database F; (△), database AB; and (○), database A.

tials derived from a protein structure database in the manners of Miyazawa and Jernigan, as well as Sippl, are strongly dependent on the length of the corresponding proteins. The validity for proteins of these results obtained from 2-D lattice models was recently questioned by Bahar and Jernigan,⁹ who repeated these calculations for the Miyazawa-Jernigan contact potentials using subsets of protein structures of different sizes. They found that the dependence of the potentials on the protein size is negligibly small, and argued that the results of Thomas and Dill were biased by the fact that they only used 2-D lattices, on which excluded volume is a much more stringent constraint than in regular 3-D structures. Using databases of true protein structures, we have shown here that there is a protein size

dependence for Sippl-like potentials (see Fig. 1), which is minimized if short contacts only are considered (i.e., for a small cutoff distance R_c). Instead of using short contacts only, Bahar and Jernigan⁹ have recently proposed another option by including a second normalization factor in Equation (10), which takes in account the volume $4\pi r^2 \Delta r$ of a spherical shell associated with a given distance range $r \pm \Delta r$.

We have shown in this study that, in addition to biases due to sequence composition and connectivity, statistical potentials carry a memory of the quality of the database used in terms of the amount and diversity of secondary structure it contains. We find, for example, that potentials derived from a database containing α -proteins only will perform better on α -proteins in fold recognition computer experiments.

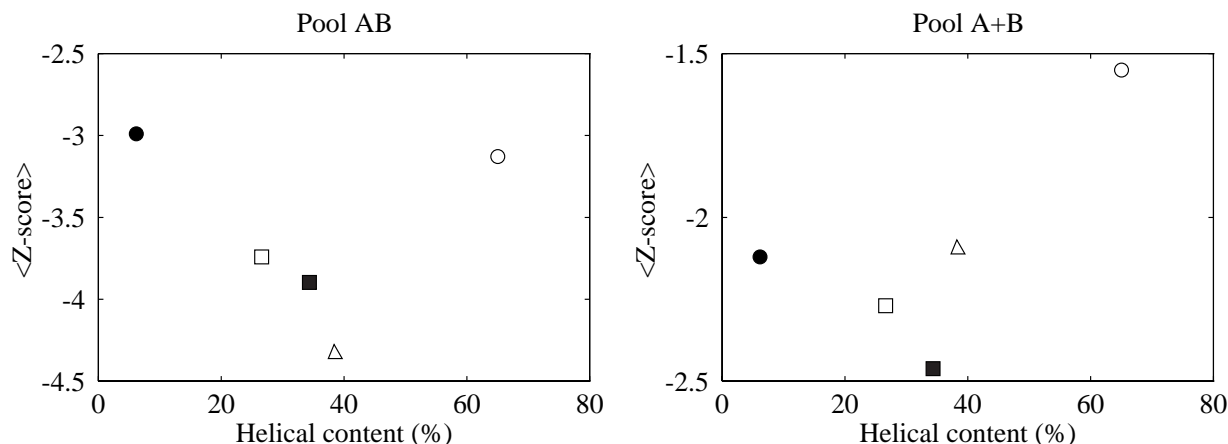


Fig. 7. Average predictive power of statistical pair potentials for recognizing α/β and $\alpha+\beta$ fold versus the helical content of the database from which they are derived. (●) denotes database B; (□), database A+B; (■), database F; (△), database AB; and (○), database A.

We believe that this is an overall weakness of these potentials, which must be kept in mind when constructing a database for deriving such potentials.

REFERENCES

- Dill, K.A., Bromberg, S., Yue, K.Z., et al. Principles of protein folding: A perspective from simple exact models. *Protein Sci.* 4:561–602, 1995.
- Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* 6:195–209, 1996.
- Jones, D.T., Thornton, J.M. Potential energy functions for threading. *Curr. Opin. Struct. Biol.* 6:210–216, 1996.
- Koehl, P., Delarue, M. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* 6:222–226, 1996.
- Straub, J.E. Optimisation techniques with applications to proteins: Recent developments in theoretical studies of proteins. 137–196, 1996.
- Berne, B.J., Straub, J.E. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.* 7:181–189, 1997.
- Halgren, T.A. Potential energy functions. *Curr. Opin. Struct. Biol.* 5:205–210, 1995.
- Sippl, M.J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235, 1995.
- Bahar, I., Jernigan, R.L. Interresidue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* 266:195–214, 1997.
- Maierov, V.N., Crippen, G.M. Contact potential that recognises the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888, 1992.
- Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Optimal protein folding codes from spin glass theory. *Proc. Natl. Acad. Sci. U.S.A.* 89:4918–4922, 1992.
- Sasai, M. Conformation, energy, and folding ability of selected amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* 92:8438–8442, 1995.
- Koretke, K.K., Luthey-Schulten, Z., Wolynes, P.G. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* 5:1043–1059, 1996.
- Hao, M.H., Scheraga, H.A. Optimizing potential functions for protein folding. *J. Phys. Chem.* 100:14540–14548, 1996.
- Thomas, P.D., Dill, K.A. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 93:11628–11633, 1996.
- Crippen, G.M. Easily searched protein folding potentials. *J. Mol. Biol.* 260:467–475, 1996.
- Mirny, L.A., Shakhnovich, E.I. How to derive a protein folding potential: A new approach to an old problem. *J. Mol. Biol.* 264:1164–1179, 1996.
- Tanaka, S., Scheraga, H.A. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 9:1976.
- Miyazawa, S., Jernigan, R.L. Estimation of effective inter-residue contact energies from protein crystal structures: Quasichemical approximation. *Macromolecules* 18:534–552, 1985.
- Sippl, M.J. 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.* 1990:859–883, 1990.
- Hendlich, M., Lackner, P., Weitckus, S., et al. Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* 216:167–180, 1990.
- Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.* 89:9029–9033, 1992.
- Kolinski, A., Godzik, A., Skolnick, J. A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J. Chem. Phys.* 98:7420–7433, 1993.
- Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. *Nature (London)* 358:86–89, 1992.
- Delarue, M., Koehl, P. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J. Mol. Biol.* 249:675–690, 1995.
- Eisenhaber, F. Hydrophobic regions on protein surfaces: Derivation of the solvation energy from their area distribution in crystallographic protein structures. *Protein Sci.* 5:1676–1686, 1996.
- Sippl, M.J., Ortner, M., Jaritz, M., Lackner, P., Flökner, H. Helmholtz free energies of atom pair interactions in proteins. *Folding Design* 1:289–298, 1996.
- DeWitte, R.S., Shakhnovich, E.I. Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy. *Prot. Sci.* 3:1570–1581, 1994.
- Bryant, S.H., Lawrence, C.E. The frequency of ion pair substructures is quantitatively related to the electrostatic potential: A statistical model for nonbonded interactions. *Proteins* 9:108–119, 1991.

30. Sippl, M.J. Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biol.* 260:644–648, 1996.
31. Park, B.H., Levitt, M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392, 1996.
32. Sippl, M., Weitckus, S. Detection of native-like models for amino-acid sequences of unknown three-dimensional structure in a database of known protein conformation. *Proteins* 13:258–271, 1992.
33. Sippl, M.J. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355–362, 1993.
34. Rabow, A.A., Scheraga, H.A. Lattice neural network minimization: Application of neural network optimization for locating the global-minimum conformations of proteins. *J. Mol. Biol.* 232:1157–1168, 1993.
35. Shakhnovich, E.I., Gutin, A.M. A new approach to the design of stable proteins. *Protein Eng.* 6:793–800, 1993.
36. Thomas, P.D., Dill, K.A. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* 257:457–469, 1996.
37. Rooman, M.J., Wodak, S.J. Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* 8:849–858, 1995.
38. Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding: An approach to the computational determination of protein structures. *J. Comput. Aided. Mol. Des.* 7:473–501, 1993.
39. Bernstein, F.C., Koetzle, T.F., Williams, G., et al. The protein databank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542, 1977.
40. Orengo, C.A., Flores, T.P., Jones, D.T., Taylor, W.R., Thornton, J.M. Recurring structural motifs in proteins with different structures. *Curr. Biol.* 3:131–139, 1993.
41. Sippl, M.J., Jaritz, M. Predictive power of mean force pair potentials. In: 'Proteins Structure by Distance Analysis.' Bohr, H., Brunak, S. (eds.). Amsterdam: IOS Press, 1994: 113–134.
42. Kocher, J.-P.A., Rooman, M.J., Wodak, S. Factors influencing the ability of knowledge-based potentials to identify native sequence–structure matches. *J. Mol. Biol.* 235:1598–1613, 1994.
43. Miyazawa, S., Jernigan, R.L. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644, 1996.