J. Mol. Biol. (1996) **257**, 457–469

JMB

Statistical Potentials Extracted From Protein Structures: How Accurate Are They?

Paul D. Thomas[1] and Ken A. Dill[2]*

[1]*Graduate Group in Biophysics, University of California, San Francisco CA 94143-0448, USA*

[2]*Department of Pharmaceutical Chemistry University of California San Francisco CA 94143-0446, USA*

**Corresponding author*

''Statistical potentials'' are energies widely used in computer algorithms to fold, dock, or recognize protein structures. They are derived from: (1) observed pairing frequencies of the 20 amino acids in databases of known protein structures, and (2) approximations and assumptions about the physical process that these quantities measure. Using exact lattice models, we construct a rigorous test of those assumptions and approximations. We find that statistical potentials often correctly rank-order the relative strengths of interresidue interactions, but they do not reflect the true underlying energies because of systematic errors arising from the neglect of excluded volume in proteins. We find that complex residue–residue distance dependences observed in statistical potentials, even those among charged groups, can be largely explained as an indirect consequence of the burial of non-polar groups. Our results suggest that current statistical potentials may have limited value in protein folding algorithms and wherever they are used to provide energy-like quantities.

Keywords: protein folding; knowledge-based potential; Boltzmann ensemble; residue partitioning; protein structure recognition

Introduction

Our purpose here is to evaluate ''statistical potentials'' and the premises that underlie them. Statistical potentials are widely used as empirical energy functions to judge the quality of proposed protein structure models (Lüthy *et al*., 1992; Wilmanns & Eisenberg, 1993), to identify the native fold or correct folding motif of an amino acid sequence among many incorrect alternatives (Hendlich *et al*., 1990; Jones *et al*., 1992; Bryant & Lawrence, 1993), to identify possible folds for a sequence of unknown structure (Bowie *et al*., 1991; Sippl & Weitckus, 1992), to predict docking of protein structures (Pellegrini & Doniach, 1993), to find amino acid sequences compatible with a desired structure (Godzik & Skolnick, 1992), and to simulate protein folding (Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Sun, 1993; Kolinski & Skolnick, 1994).

Statistical potentials are putative energies that are derived from amino acid pairing frequencies observed in known protein structures. The idea was first proposed by Tanaka & Scheraga (1976). Miyazawa & Jernigan (1985) took a major step forward in including terms to explicitly consider solvent effects. Sippl (1990) and others (Hendlich *et al*., 1990; Jones *et al*., 1992) extended these methods to include dependence on pairwise separation of residues in space and along the sequence. Bryant & Lawrence (1993) developed a log-linear statistical model to analyze protein structures separately, rather than using simple sums over distributions of residues in all proteins. More recently, statistical potentials have been refined by adding other statistical terms involving residue triplets (Godzik & Skolnick, 1992), dihedral angles (Nishikawa & Matsuo, 1993; Kocher *et al*., 1994), solvent accessibility and hydrogen-bonding (Nishikawa & Matsuo, 1993).

The basic idea behind statistical potentials is simple. We illustrate the idea using an idealized example. Suppose large numbers of the 20 amino acids were somehow to distribute themselves in a gas phase at temperature T. If the interactions are purely pairwise, the distributions can be described by the equilibrium pairwise density $\rho_{ij}(r)$ between any two amino acid types $i, j = 1, 2, \ldots, 20$ at

Abbreviations used: PDB, Protein Data Bank; 2D, two-dimensional; HP, hydrophobic-polar; PP, polar-polar; HH, hydrophobic-hydrophobic.

0022–2836/96/120457–13 $18.00/0

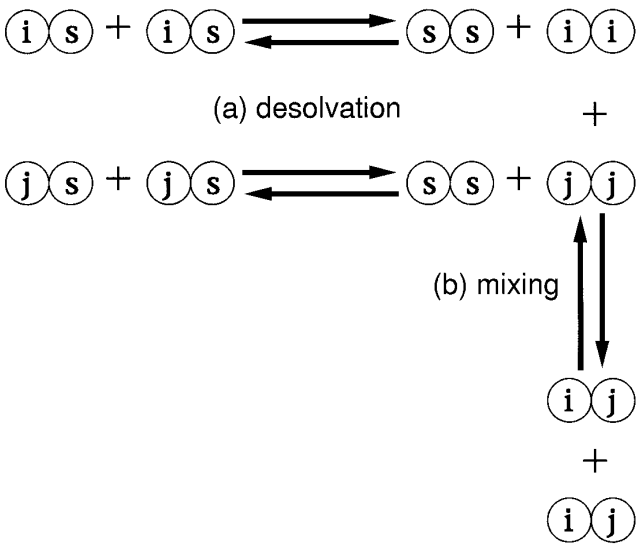

Figure 1. Hypothetical process for extracting contact energies between residues of type *i* and *j* from contact distributions in proteins. (a) ''Desolvation'' of two *i*-solvent contacts to form *i*-*i* and solvent-solvent contacts: extracted contact energy = $w_{ii} + w_{00} - 2w_{i0}$, where 0 denotes solvent and w_{xy} are defined by equation (2) using a random mixture of solvent and residues as the reference state. (b) ''Mixing'' of *i*-*i* and *j*-*j* contacts to form two *i*-*j* contacts: extracted contact energy = $2w_{ij} - w_{ii} - w_{jj}$, where w_{xy} are defined by equation (2) using a random mixture of residues, weighted according to average degree of burial in protein structures, as the reference state.

distance *r*. In this case, the interaction free energy, $w_{ij}(r)$, can be calculated from the observed densities by the Boltzmann relation:

$$w_{ij}(r) = -kT \ln\left(\frac{\rho_{ij}(r)}{\rho^*}\right) \quad (1)$$

where k = Boltzmann's constant and ρ^* is the reference state pair density at infinite separation where the particle interaction is zero. This example shows how to infer pairwise energies from the average spatial distributions of amino acids in this idealized gas phase example.

But protein crystal structures (and NMR structures) are not gas phases of amino acids in dynamic equilibrium. Certain assumptions and approximations are usually made to obtain energy-like quantities from protein structures. First, amino acid pair density functions $\rho_{ij}(r)$ are constructed by summing the static densities observed in different proteins from the Brookhaven Protein Data Bank (PDB, Bernstein *et al*., 1977) rather than averaging different states of the same protein. Second, it is necessary to choose a reference state (the pair density corresponding to zero-energy). Miyazawa & Jernigan (1985) introduced the use of the ''random-mixing approximation'', which assumes that in the absence of interactions, the amino acids and solvent molecules would be uniformly distributed throughout the available volume. In a random mixture the number of contacts between different monomers depends only on the relative concentrations of those monomer types. For example, because alanine residues are more common in proteins than methionine residues, a random mixture will have more Ala-Ala contacts than Met-Met. Finally, using the Boltzmann equation supposes an equilibrium between the observed pairing state and the reference state. Each amino acid pair is assumed to be independent of all the other pairs in the molecule. For weakly interacting particles in the gas phase, this is a reasonable approximation. However, one of the most remarkable features about proteins is the extremely close packing of residues (Richards, 1977). In addition, amino acids are covalently linked in specific sequences. These are the premises we test below. We do not test the assumption that interactions are pairwise additive, nor do we treat local interactions (e.g. dihedral angle potentials).

There are two main approaches to calculating amino acid pair potentials. In one approach, the interactions between amino acids are assumed to be short-ranged, and are approximated using a ''contact potential'' (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985). In the Miyazawa & Jernigan (1985) formulation, a contact energy w_{ij} is an average of amino acid pairings over distances shorter than some cutoff distance r_c:

$$w_{ij} = -kT \ln\left(\frac{\int_0^{r_c} \rho_{ij}(r)\, dr}{\int_0^{r_c} \rho_{ij}^*(r)\, dr}\right) \quad (2)$$

Miyazawa & Jernigan (1985) recognized that protein folding involves desolvating two monomers *i* and *j* before forming a contact between them. To account for this approximately, Miyazawa and Jernigan invented a hypothetical two-step process of contact formation (Figure 1). In the first step, monomers *i* and *j*, which are regarded as being solvated in the denatured state, go through a self-pairing (*i* with *i*, *j* with *j*, solvent with solvent). In the second step, the *i*-*i* and *j*-*j* pair ''bonds'' are broken and *i*-*j* bonds are made. These steps involve applying equation (2) four times, for breaking two contacts and forming two contacts.

In the Miyazawa and Jernigan approach, each of the two steps, desolvation and mixing, is based on a different random mixture reference state. For desolvation, the reference state is a uniform mixture of solvent molecules and amino acid residues. For mixing, which involves moving amino acids within a compact globule, the reference state weights residue positions in terms of their degree of burial.

The second class of statistical pair potentials allows for distance-dependence of the interactions.

For such potentials, equation (1) has been applied to individual small distance intervals. In this case, another normalization is also needed. Sippl (1990) solved the problem of how to calculate the expected ''uniform density'' reference state for distance-dependent potentials. The reference density of pair distances at each distance interval depends not only on the frequencies of the residue pairs in question, but also on the total number of pair distances observed at that distance. For example, more pairs of amino acids are separated by 10 Å than by 80 Å, because of the small sizes of proteins. Because dividing up frequencies both by pair type and distance interval results in many parameters with few proteins to define them, Sippl (1990) developed a ''sparse data correction'' which corrects for the energies calculated using equation (1) by an uncertainty factor.

Testing the premises of statistical potentials

Statistical potentials are arguably intended to mimic the natural energies that drive amino acids to form contacts. How well do statistical potentials, ''extracted'' from native structure databases, reflect the ''true'' underlying energies? It is not known, because there is no independent knowledge of nature's true underlying energies. Here we devise a test that circumvents that problem. We generate different ''model PDBs'' using an exact lattice model for which the underlying energies can be specified exactly. We extract putative energies from observing the monomer pairing frequencies in each model PDB, and compare the extracted energies to the true energies. We define the term ''true energy'' to mean the actual contact free energy that causes the protein to fold into its given native state, and the term ''extracted energy'' to mean the energy-like quantity that is obtained from observing the monomer pairing frequencies in the database of native structures and using the assumptions described above. It is not important that the lattice model is not a perfect mimic of real proteins. We are simply performing a consistency check of the methods that generate statistical potentials. We aim to learn how much error is introduced by their neglect of chain connectivity, amino acid sequence and excluded volume.

The ''AB-model'' consists of chains of two monomer types, A and B, having lengths $L = 11$ to 18 on two-dimensional (2D) square lattices. We specify a true contact potential, involving three interaction energies (for AA, AB, and BB contacts) which defines the total energy for any conformation of any AB sequence on the 2D lattice. Monomers are in contact if they are non-bonded nearest neighbors on the lattice. Since there are only three energies, we are able to explore all possible sets of unique contact potentials, and all the unique native structures for each given energy function.

Our consistency check runs as follows. (1) We select a set of true contact free energies, which we denote with capital letters E_{AA}, E_{AB} and E_{BB}. (2) For each chain length, we perform an exhaustive search of conformational space, and find the native states (lowest true energy) of all 2^L sequences. (3) We make a ''database'' of the unique native structures. (4) We use two representative statistical potential extraction methods, one contact-based and one distance-dependent, to extract statistical energies from this lattice model database. We denote the extracted energies with lowercase e_{AA}, e_{AB} and e_{BB}. The contact potential is extracted by the method of Miyazawa & Jernigan (1985) and the distance-dependent potential by a simplified version of the method of Sippl (1990) that considers all residues in the short chains to be of the same ''topological level'' (i.e. there is no dependence on sequence separation). We selected these two methods as representative of the methodology of statistical potentials because they are the most widely referenced in the current literature.

Our approach involves no sampling problems since we exhaustively enumerate the complete database of all sequences having a unique fold. The data are not sparse, as they are for real proteins. We have approximately 1200 to 60,000 contacts to define three parameters (400 to 20,000 observations per parameter). For comparison, Miyazawa and Jernigan used about 27,000 contacts to define 210 parameters, or about 130 observations per parameter. For the distance-dependent potentials, we include the sparse-data correction of Sippl (1990), though for most of the lattice model databases we consider there are enough pair distances to approach the uncorrected values. We do not address questions of database size or sampling problems. Rather, our purposes here are (1) to investigate the principles of statistical potentials, and (2) to assess how accurately current statistical potentials might reflect the real amino acid contact energies in proteins.

Results

First we explore the simplest version of the AB model, where the true potential involves only a single energy, namely the HP (Hydrophobic-Polar) model: $E_{HH}:E_{HP}:E_{PP} = -1:0:0$. In this model, contacts between H monomers are favorable relative to solvent contacts, while HP and PP contacts are energetically equivalent to solvent contacts. The folding properties of this model are known in some detail (Lau & Dill, 1989, 1990; Chan & Dill, 1991a,b; Lipman & Wilbur, 1991; Shortle *et al*., 1992; Miller *et al*., 1992; Unger & Moult, 1993; O'Toole & Panagiotopoulos, 1993; Camacho & Thirumalai, 1993a; Camacho & Thirumalai, 1993b; Chan & Dill, 1994; Chan *et al*., 1995; reviewed by Dill *et al*., 1995). Figure 2 shows that the Miyazawa and Jernigan procedure correctly determines that the HH contact interaction is dominant and attractive. However, the extracted energies are not equal to the true energies. Two of the main errors introduced by the extraction process are: (1) the extracted energies e_{HP} and e_{PP} are found to be non-zero, whereas the true energies E_{HP}

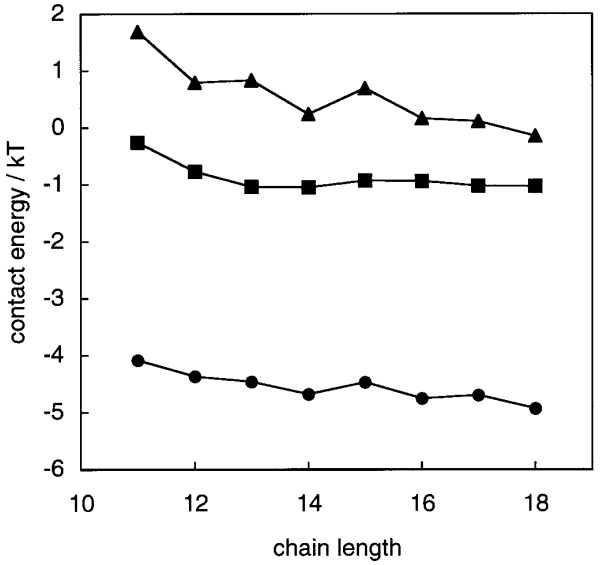

Figure 2. Extracted statistical potentials for the HP model *versus* chain length: e_{HH} (circles, true potential was $-\epsilon$, where $\epsilon > 0$), e_{HP} (squares, compare with the true potential of 0), e_{PP} (triangles, compare with the true potential of 0).

and E_{PP} are zero, and (2) all the extracted interactions depend on chain length, whereas the true energies do not. These errors arise from the approximations made in the extraction procedures for statistical potentials, as described below.

The problem: interactions are not independent

For the HP model, the extraction process infers that the HP interaction is more favorable than the PP interaction even though the true energies are zero for both HP and PP. The problem is not that HP contacts are more common than PP in the database; the problem is the assumption that the pairwise interactions are independent. The HH interaction, which dominates in this model, indirectly affects HP and PP pairing frequencies.

In the HP lattice model, favorable HH contacts are made preferentially, which has two effects on the observed frequencies of other contacts: (1) because each structure makes only a limited number of interresidue contacts, HH contacts deplete the total number of contacts available for any other interresidue contact (HP and PP contacts), and (2) HH contacts deplete the supply of H contact surfaces that are available for any other contacts with H monomers (HP and H-solvent contacts). But the random mixture reference state supposes that all contact types are uniformly available. Thus in the model database, PP, HP and H-solvent pairs are underrepresented relative to the reference state, so forming these contacts appears as an unfavorable contribution to the extracted contact energies (breaking them will appear favorable). The extracted HP contact energy therefore includes two large terms of opposite sign, an unfavorable HP contact forming term and a favorable H-solvent contact breaking term (Figure 1). The extracted PP contact energy, on the other hand, is dominated only by the unfavorable term for forming a PP contact; P residues are about equally solvated in both the native and reference states. The net result is that, as a result of the true HH interaction, the extracted HP contact energy appears more favorable than the extracted PP interaction. The HP and PP interactions are ''coupled'' to the HH interaction.

The coupling of interactions among different types of residue pairs is also illustrated by considering distance-dependent potentials (Figure 3). While the true potential in the HP model is just a first-neighbor HH contact interaction (i.e. a favorable ''spike'' at a distance of one lattice unit), the extracted potentials erroneously give a distance dependence. The extracted interactions are favorable over some distance ranges and unfavorable over others. The reason for this incorrect and complex distance dependence is the assumed uniform distributions in the extraction procedure reference states, as described below.

The incorrect extracted potentials come from two types of coupling. First, in both the distance dependent and contact potentials, the extracted energy of a monomer pair at a given distance is influenced by other pairs at the same distance. For the HP lattice model, the high density of HH pairs at short distances causes a correspondingly low density (relative to uniform) of HP and PP pairs at those distances. As a result, the extracted HP and PP potentials are erroneously unfavorable at short range. Second, for distance dependent extracted potentials, the energy of a monomer pair at a given distance of spatial separation is influenced by the same pair at different distances. For instance, there is a high density of HH pairs at short distances, due to the true HH attraction. The total density of distances between HH pairs is the same in both the database and the uniform distribution reference state, so the higher (than uniform) concentration of HH pairs at short distances causes a compensating depletion (relative to uniform) at longer distances. The concentrations at different distances are treated as independent by equation (1), but independence is a poor approximation.

In summary, the complex extracted distance dependence of HP and PP interactions follows from the HH interactions. H monomers cluster, giving many short HH distances and few long HH distances. The HH attraction drives P monomers to the protein surface, resulting in many long PP distances. The HH attraction similarly causes many intermediate HP distances, between interior Hs and exterior Ps. Thus the HP and PP pairing frequencies are not independent of the HH interaction.

This coupling of interactions is not an artifact of our model. A comparison of Figures 4 and 5 shows the same coupling in potentials extracted from the Protein Data Bank. We classified the residues in proteins into just two types, interior or exterior, and then used the method of Sippl (1990) to extract a potential between these two geometrically defined

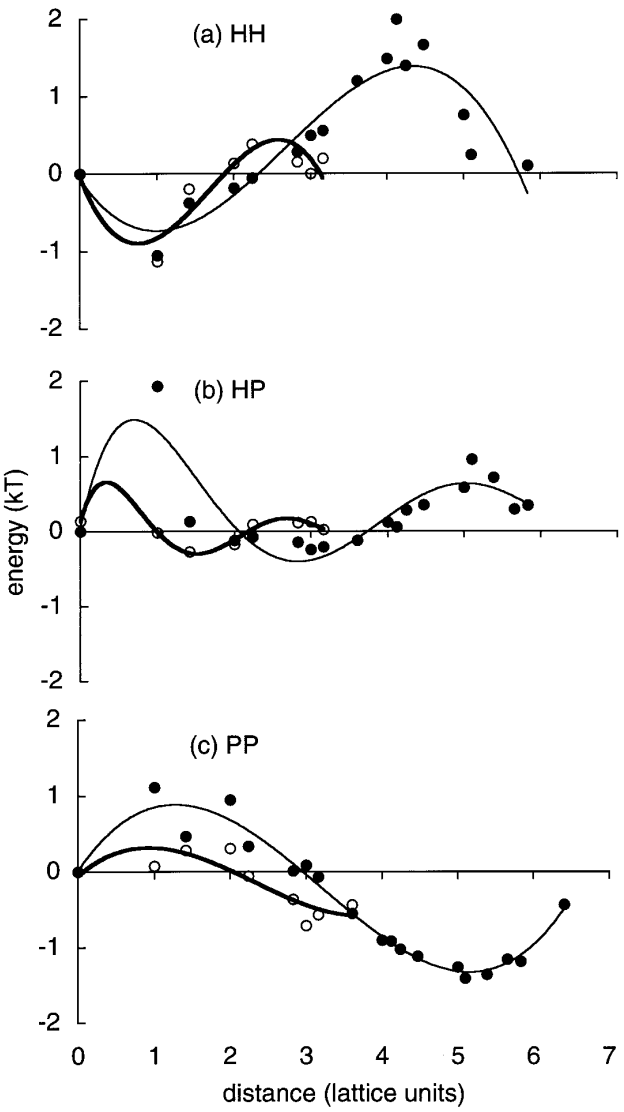

Figure 3. Distance-dependent statistical potentials extracted from 2D HP structure databases using the method of Sippl (1990), including sparse data correction ($\sigma = 1/50$). Lines are polynomial fits to the data to guide the eye. The extracted potentials for chain lengths of 11 (open circles) and 18 (filled circles) are shown. (a) Extracted HH interaction (compare with the true potential, a favorable ''spike'' at a distance of one lattice unit, and 0 for other distances). (b) Extracted HP interaction (the true potential is 0 for all distances). (c) Extracted PP interaction (the true potential is 0 for all distances).

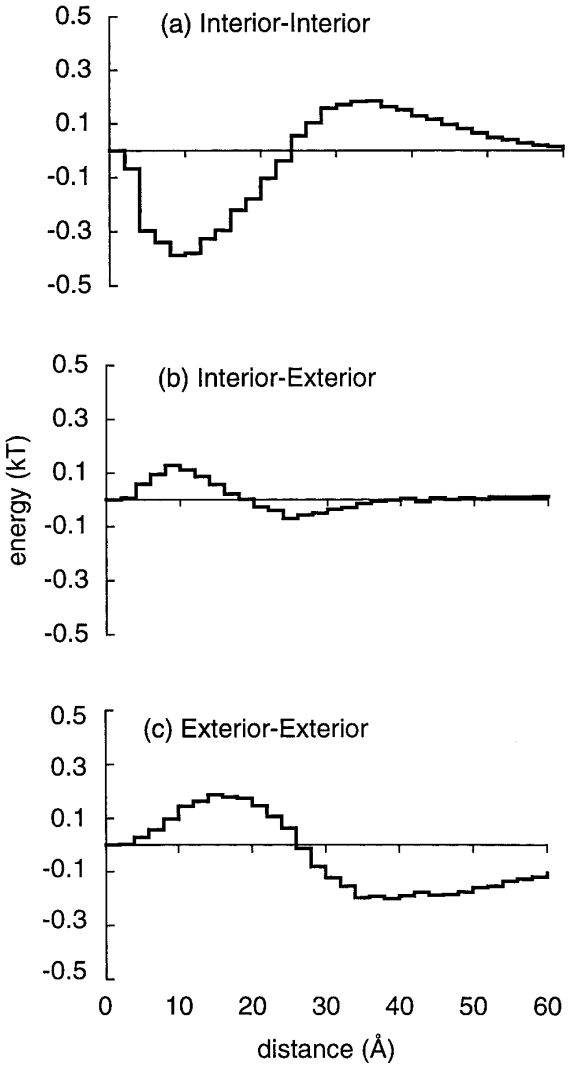

Figure 4. Distance-dependent potentials extracted from real protein structures using only two monomer types, interior or exterior: (a) Interior-interior (compare with Figure 5(a)), (b) interior-exterior, (c) exterior-exterior (compare with Figure 5(b) and (c)). Potentials are extracted for the same sequence separations in the same set of proteins, and using the same equations (including sparse data correction, scaling for three parameters rather than 210) as in Figure 5. Interior/exterior positions are determined using the program ACCESS (Lee & Richards, 1971). The full backbone and side-chain center of mass is used for the calculation, with a probe radius of 2.0 Å to compensate for the single-atom side-chain representation. If the accessible surface area of a carbon atom positioned at the side-chain center of mass (C^{α} for glycine) exceeds 30 $Å^2$, the residue is considered to be exposed.

''residue types.'' This test shows that the interior-interior pairing ''energy'' (Figure 4(a)) is nearly identical to that calculated by Hendlich *et al*. (1990) for the pair Ile-Val (Figure 5(a)), two hydrophobic residues. Apparently the 30 different extracted energy parameters for the different distance intervals are mainly just reflecting that isoleucine and valine residues tend to be in protein interiors. Moreover, Figure 3 shows that a simple nearest-neighbor HH attraction in the lattice model is sufficient to give the same functional form as that for Ile-Val pairs in proteins. Hence the apparently complex distance dependence among amino acids in proteins may reflect little more than that hydrophobic residues attract each other.

More strikingly, Figure 3 shows that the HH contact interaction is also sufficient to give an extracted PP potential (Figure 3(c)) that has roughly the same functional form as for both Lys-Asp (unlike charges) and Asp-Asp (like charges)

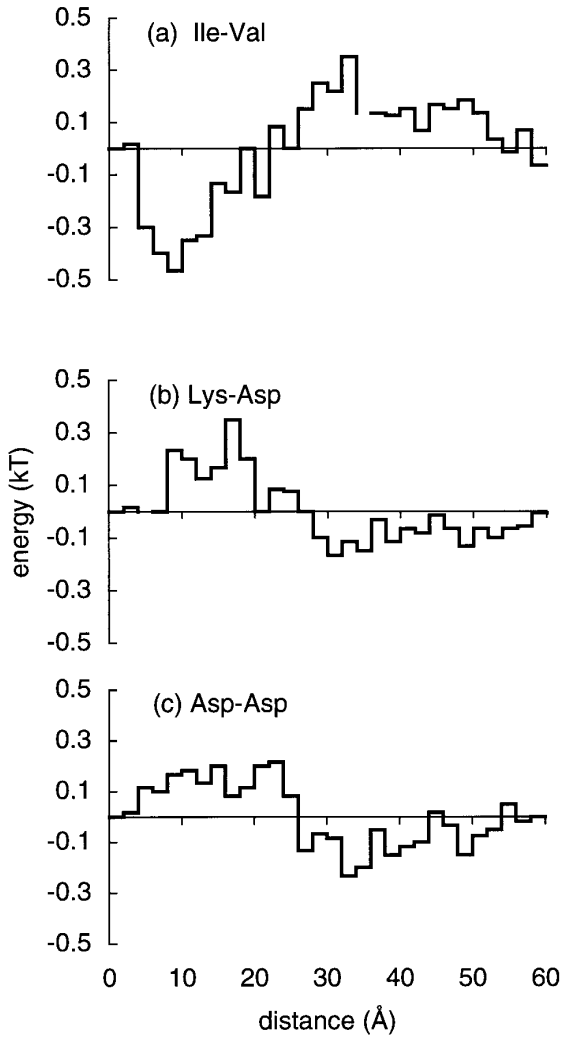

Figure 5. Distance-dependent potentials extracted by Hendlich *et al*. (1990) for residue pairs separated in primary sequence by 61 to 100: (a) valine-isoleucine, (b) lysine-aspartate, (c) aspartate-aspartate. Note the similarity between (b) and (c).

(Figure 5(b) and (c)) as extracted from the PDB by Hendlich *et al*. (1990). That is, the extracted charged residue interactions in proteins are not mainly due to electrostatics; it appears to be mainly because charged residues are driven to the protein surface by the non-polar attractions of other amino acids. This explains the observation that more Coulomb-like charge-charge interactions are extracted from the PDB when statistics are compiled only on surface residues (Kocher *et al*., 1994). Considering only surface residues is equivalent to using a different reference state for the potentials that takes into account the average effect of the hydrophobic residues on the observed charge-charge distributions.

Some statistical potentials have tens of thousands of parameters for pairing frequencies as a function of distance and sequence separation. But most of those parameters may be redundant, all reflecting mainly that non-polar amino acids form a core surrounded by polar residues. To test this possible redundancy, we performed the same threading test as Hendlich *et al*. (1990), but using only a single ''energy'' parameter that accounts for contacts between hydrophobic residues. Table 1 compares our single-parameter results with the threading potential of Hendlich *et al*. (1990). Table 1 shows that this single hydrophobic contact energy parameter correctly identifies the native conformation as having the lowest energy in nearly as many cases (37/65) as the much more complex potential of Hendlich *et al*. (41/65). Most of the failures in the simple model are also failures in the complex model. We conclude that: (1) most of the information about protein energetics contained in complex statistical potentials is simply hydrophobic clustering propensity, as has been noted before (Casari & Sippl, 1992; Bryant & Lawrence, 1993), and (2) the threading test (with no insertions and deletions) is not a particularly challenging test of energy functions. Our results for the threading test are consistent with those of Bryant & Amzel (1987), who found that counting hydrophobic contacts can distinguish between correctly and grossly misfolded proteins.

Interior-exterior partitioning: effects of protein size and composition

Whereas the true potentials between amino acids cannot depend on the chain length, the extracted potentials do (Figures 2 and 3). For the distance-dependent potentials extracted from our 2D HP lattice model (Figure 3), the functional form is similar for chain lengths of 11 or 18, but the scale of the distance dependence is different. The scale is a geometric description of the average location of monomers relative to the core or surface. For example, H residues are overrepresented in the cores of our model structures, which are larger for the 18-mers than for the 11-mers; accordingly, the extracted HH energy changes from attractive to repulsive between a distance of $\sqrt{5}$ and $\sqrt{8}$ lattice units for the 18-mers, but only between $\sqrt{2}$ and two lattice units for the 11-mers. Similarly, the PP interaction reaches a minimum at three lattice units for the 11-mers, but about five lattice units for the 18-mers; this corresponds to the difference in average conformational diameters. Analogous geometric properties may account for the results reported by Hendlich *et al*. (1990) that a potential extracted from a database of smaller proteins performs slightly better at recognizing the native conformations of other small proteins than a potential extracted from a database of proteins of all sizes.

In the case of the extracted contact potentials, the chain length dependence (Figure 2) is due primarily to the desolvation terms (Figure 1), which take a form similar to that of a transfer from a solvated to a buried state. As Janin (1979) first noted, apparent interior-exterior ''partition energies'' (extracted from frequencies of amino acids in buried and

Table 1. A count of hydrophobic contacts succeeds in the ''threading test'' nearly as often as a more complex potential for identifying native structures

Protein PDB code	Extracted potential Position	Hydrophobic contact count Position	N_{HH}	ΔN_{HH}
4SBV A	1	1	227	15
3ADK	1	1	167	24
2STV	5	1	154	2
1HMG B	71	803	88	–34
1GCR	1	1	188	43
2ALP	1	1	202	30
3WGA A	1	1	147	29
2SGA	6	1	181	45
2LZM	1	1	183	38
4DFR A	1	1	178	14
1LH4	1	1	189	10
1MBD	1	1	161	20
2SOD O	1	1	136	38
2LHB	1	1	183	17
2HHB B	1	1	172	26
2PKA B	5	3	116	–9
2HHB A	1	1	155	4
3FXN	1	1	173	46
1LZ1	1	1	174	42
2AZA A	1	1	120	8
2CCY A	2	2	140	–1
1RN3	1	1	99	8
1BP2	1	1	115	2
1PP2 R	1	1	118	0
155C	1	1	99	13
1PAZ	1	1	152	29
2PAB A	1	1	119	0
1HMQ A	1	115	81	–16
2C2C	1	16	92	–11
1CPV	1	1	121	2
1ACX	2	1	116	7
1REI A	1	1	95	9
2CDV	1	1548	41	–24
2SSI	1	7	119	6
4CYT R	1	53	59	–18
1WRP R	1	160	77	–16
1PCY	1	1	101	10
1HVP A	5	1	89	3
3FXC	2	1	102	13
2GN5	26	108	65	–13
1HIP	11	24	83	–9
2B5C	1	1	62	7
1CC5	235	7	86	–8
351C	1	11	73	–5
2PKA A	1	42	62	–10
3ICB	1	1	66	19
2ABX A	65	12	56	–9
1HOE	125	4	69	–3
1CTF	71	1	101	17
1SN3	1	1	70	8
1CSE I	1	3	60	–4
2EBX	1	2	32	–1
2MT2	1	1	71	10
4PTI	1	9	52	–5
1OVO A	1	35	39	–11
1FDX	64	1	71	1
5RXN	2571	1	44	0
1CRN	24	8	45	–6
1BDS	8	4	42	–2
2RHV 4	74	9690	9	–28
1PPT	10	360	17	–10
1INS B	301	186	25	–13
1GCN	4563	666	11	–10
1MLT A	107	340	20	–11
1INS A	483	686	14	–14

Position of the correct native conformation in a list of threaded conformations sorted by the total distance-dependent statistical energy of Hendlich *et al*. (1990) (column 2), or just by counting non-polar contacts (Ala, Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val; column 3). A contact is defined as a C^{β}–C^{β} distance of less than 8 Å. N_{HH} is the number of non-polar contacts in the native conformation, and ΔN_{HH} is the difference between the number of non-polar contacts in the native conformation and in the lowest energy non-native threaded conformation.

exposed positions in proteins) will depend on the surface-to-volume ratio of the protein because the greater volume inside larger proteins more readily accommodates non-polar monomers. This explains the systematic errors noted by Miyazawa & Jernigan (1985) in trying to predict surface-to-volume ratios using their extracted contact energies. Miyazawa and Jernigan assumed their extracted energies would not depend on the surface-to-volume ratios since the true energies do not depend on this quantity.

Extracted partition energies also depend on amino acid composition even though the true potentials cannot. For the lattice model, extracted energies become more positive with increasing content of H monomers (Figure 6). This is because there are more H monomers than needed to fill the core (2D lattice 18-mers can have at most five monomers completely sequestered from solvent), so additional H-monomers must be at least partially exposed to solvent. Because of coupling, the extracted HP and PP energies also increase with increasing non-polar content; with more H residues in the sequence, the chances increase that a given conformation will make only HH contacts. For example, out of all 164 structures in the database of 18-mers with 14 H monomers, 162 make only HH contacts and no HP or PP contacts.

To account approximately for the effects of surface-to-volume ratio and composition on extracted energies, we define the ''partition propensity'' (π) of a given protein or set of proteins as:

$$\pi = \frac{2n_c}{q_H n_H} \tag{3}$$

where n_c is the total number of contacts in a given protein, q_H is the average coordination number of a hydrophobic (H) residue (taken from Miyazawa & Jernigan, 1985) and n_H is the number of H residues in the protein. Physically, $2n_c$ is the number of coordination sites involved in all residue-residue contacts (roughly proportional to the total buried surface of the molecule), and $q_H n_H$ is the total number of H coordination sites (roughly proportional to the total hydrophobic surface).

The partition propensity is therefore a crude measure of how effectively hydrophobic surface can be buried in a given structure. If a protein has a low partition propensity, it has more hydrophobic residues than are needed to fill a core, so some will not be able to partition effectively into a core. A high partition propensity means there is a large buried core, and few hydrophobic residues in the sequence to fill it. Figure 7(b) shows how the extracted HH contact energy depends on the partition propensity in the lattice model.

The same dependence of extracted energies on partition propensity is found for real proteins (Figure 8). We sorted a group of 346 non-homologous protein chains (Hobohm & Sander, 1994) from the PDB by the partition propensity of each protein, labeling each residue type as H (hydrophobic) or P (others). We made 326 sets of 20 proteins by sliding

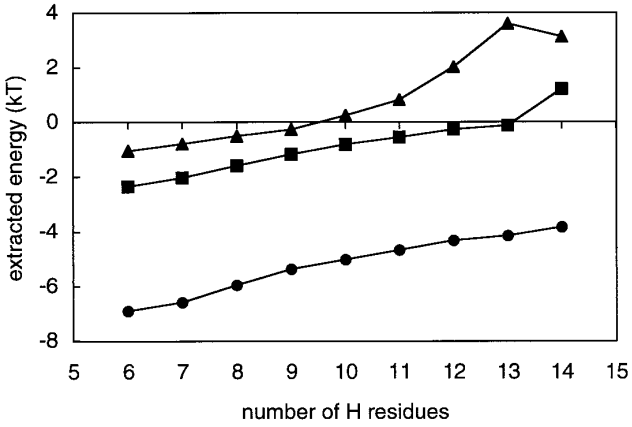

Figure 6. Extracted potentials depend on composition. The extracted HH (circles), HP (squares) and PP (triangles) energies for 2D lattice chains, $L = 18$, *versus* the number of H residues (n_H) in the sequences.

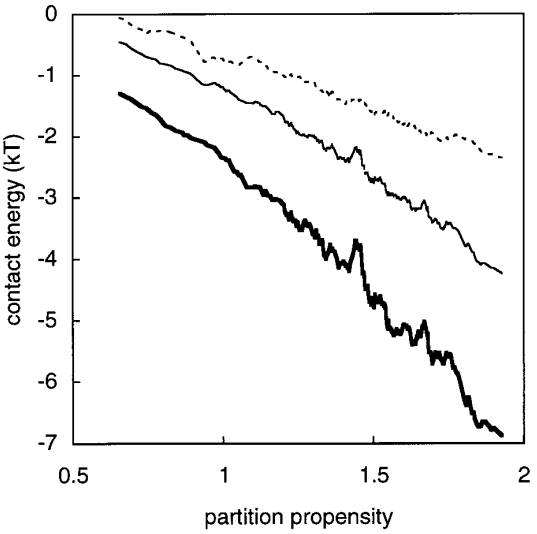

Figure 8. Extracted energies depend on partition propensity for proteins in the PDB. Hydrophobic residues (H) are Ala, Cys, Ile, Leu, Met, Phe, Tyr, Trp and Val; others are classed as P. The HH energy is shown by the bold line, HP by the continuous line and PP by the broken line.

a window along the sorted list of proteins. That is, the ith set contains the ith through $i + 19$th proteins in the sorted list. Figure 8 plots the extracted contact energies of each set against the average partition propensity of the set. The extracted energy between hydrophobic groups becomes significantly more favorable with increasing partition propensity, from about $-1.3kT$ to $-6.9kT$, while the extracted PP energy increases from about $0kT$ to $-2.4kT$. The statistical potentials depend systematically on the size and amino acid composition of the proteins in the given set.

Consider two different 69-protein databases: one containing proteins with an average partition propensity $\pi = 1.8$ and the other having $\pi = 1.0$. Figure 9 plots the 210 contact energies extracted from one database against the same ones from the other database. The correlation between the two energy sets is only modest ($R = 0.62$). Even the rank ordering of contact pairings can be affected by protein size and composition. For example, for $\pi = 1.8$ the extracted Phe-Lys contact energy is more favorable than Ile-Val, whereas for $\pi = 1.0$ Phe-Lys is less favorable than Ile-Val. Coupling effects are different in these two protein sets, because of the different degree of burial of the hydrophobic residues. Phe is usually completely buried in the proteins having higher partition propensity, so a Phe-solvent contact is very unfavorable by equation (2), and any contact with Phe appears very favorable because it breaks a Phe-solvent contact. However for proteins having more hydrophobic residues than their cores can accommodate, Phe is often sacrificed to the surface, and Phe-solvent contacts are less unfavorable. If the relative strengths of the extracted contact energies depend on average properties of the proteins in the database, such as protein size and composition, then how can we know which, if

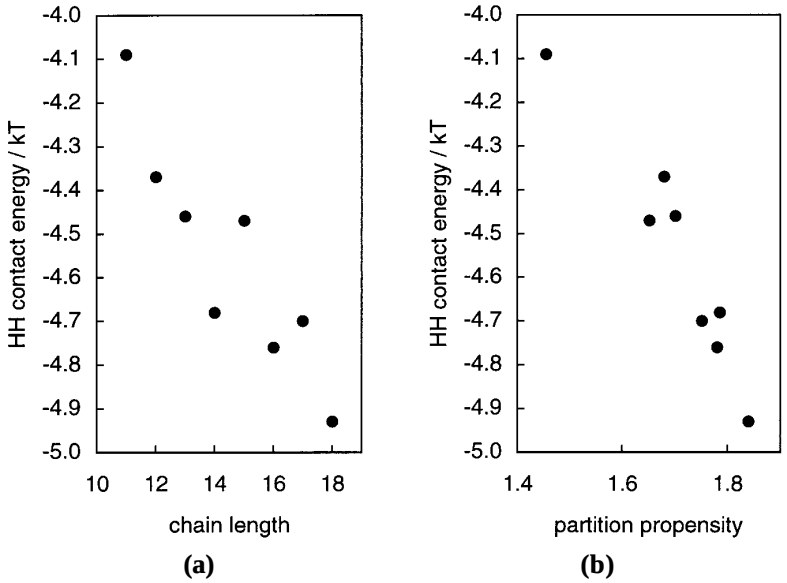

Figure 7. Extracted energies for the 2D HP model depend on (a) chain length, (b) average partition propensity of the structures in the database.

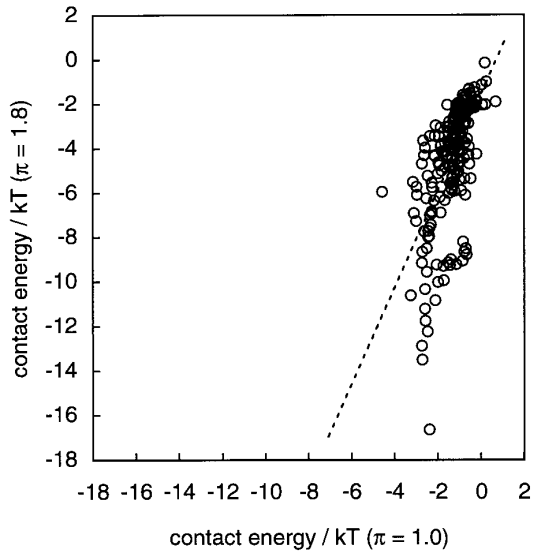

Figure 9. Contact energies extracted from two different sets of 69 proteins in the PDB. The energies calculated for the set of highest partition propensity ($\pi = 1.8$) are plotted against the energies for the same pairs calculated for the set of lowest partition propensity ($\pi = 1.0$). The slope of the linear best fit is 2.2, with an intercept of $-1.6kT$.

any, set of proteins gives extracted energies that are similar to nature's underlying energies? Clearly simply increasing the number of proteins in a database does not ''average out'' the coupling of the extracted energies. However, there may be ways to construct contact pair density functions or reference states that can take into account properties such as the partition propensity.

The Boltzmann distribution: does it apply?

Among the deepest questions underlying statistical potentials is whether the Boltzmann distribution law, equations (l) or (2), is appropriate for converting pair frequencies in proteins to energies. What is the meaning of temperature in these expressions? The Boltzmann distribution law applies to a single closed system held at fixed temperature that can populate different energy levels. In the gas-phase amino acid pairing example above, increasing the temperature would cause Ala-Trp contacts, for example, to resemble a random distribution. But the PDB is many proteins, not a single system. The database is fixed; each protein has no degrees of freedom that are affected by temperature changes. The amino acid pairings in lysozyme, for example, are the same at $T = 300$ K as at $T = 0$ K. Consequently, the extracted potentials contain no information about protein stability; extracted energies will be the same whether the native conformations of the proteins are stable by 10^{-5} or 10^{5} kcal/mol. Should we use $T = 0$ K, $T = 300$ K, or some other temperature?

Nevertheless, some evidence supports the use of a Boltzmann distribution. In particular, for some protein ''substructures'' (e.g. proline residues), frequencies of different ''states'' (e.g. *cis*- and *trans*-peptide conformations) observed in the PDB correlate with the frequencies expected from thermodynamic behavior. Observed frequencies in proteins are converted to energy differences using the Boltzmann relation, and then plotted against the energy differences (expected from either theory or experiment) between different states according to thermodynamic behavior at $T = 300$ K. If the correlation is linear, the slope can be adjusted by choosing the best-fit value for T in the Boltzmann relation (i.e. so the slope of the plot is 1). This yields an effective ''temperature'' for that substructure in proteins relative to the temperature of a true Boltzmann ensemble. This has been done for the following protein properties. Extracted partitioning energies correlate (Miyazawa & Jernigan, 1985; Rose *et al*., 1985; Lawrence *et al*., 1987; Miller *et al*., 1987) with oil/water transfer experiments (e.g. Nozaki & Tanford, 1971; Fauchàre & Pliska, 1983). Also well predicted are distributions of ϕ/ψ dihedral angles (Pohl, 1971; Kolaskar & Prashanth, 1979), charged residues (Bryant & Lawrence, 1991), *cis*- and *trans*-proline residues (MacArthur & Thornton, 1991), the sizes of empty cavities (Rashin *et al*., 1986) and residue stabilization of secondary structure elements (Serrano *et al*., 1992). Remarkably, for all of these different protein substructures, the apparent temperature is of the same order of magnitude, between about 150 K and 600 K.

Finkelstein *et al*. (1993) have proposed a theory to explain why substructures have about the same frequencies in proteins as they would have in thermodynamic equilibrium. They argue, based on a random heteropolymer model of proteins, that the number of random sequences having a native structure that contains any given substructure depends exponentially on the energy of that substructure. This model predicts that the temperature in the Boltzmann relation should be the same for all types of substructures, and roughly equal to room temperature.

We test here the Boltzmann distribution assumption for the interior-exterior partitioning of residues. We find that the apparent temperature for the extracted partition energies depends systematically on the average partition propensity (equation (3)) for the set of proteins. We divided the 346 PDB protein structures, sorted by partition propensity, into five sets. For each set, we define the extracted exterior-interior partition energy for each amino acid i:

$$\Delta G_i = G_{inside} - G_{outside} = -kT \ln\left(\frac{n_{ir}}{n_{i0}}\right) \qquad (4)$$

where n_{ir} is the number of contacts between residue type i and other residues (these are ''interior'' sites), and n_{i0} is the number of contacts with solvent (residue type 0). Both n_{ir} and n_{i0} are estimated as by Miyazawa & Jernigan (1985). Physically, n_{ir} corresponds roughly to the average fractional surface area of residues of type i that is buried in protein

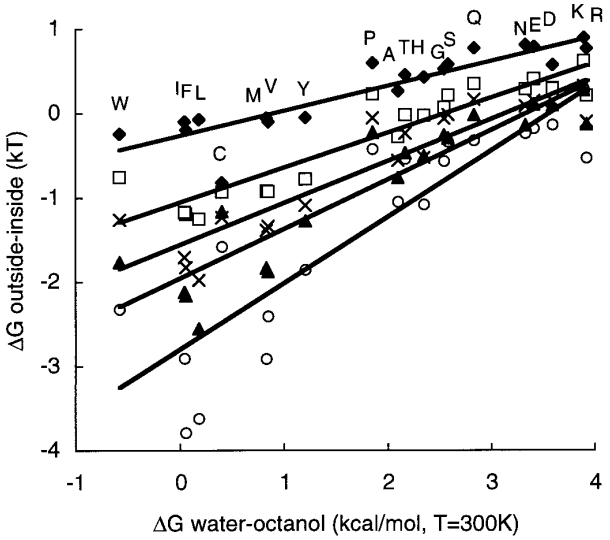

Figure 10. Extracted exterior-interior partition energies of the 20 amino acids, for protein sets having different propensities, *versus* experimental water-octanol transfer energies (Fauchère & Pliska, 1983). The lines are regression fits to (from top to bottom) $\pi = 1.0$ (filled diamonds, $R = 0.90$, $m = 0.28$), $\pi = 1.3$ (open squares, $R = 0.92$, $m = 0.42$), $\pi = 1.5$ (crosses, $R = 0.92$, $m = 0.50$), $\pi = 1.6$ (filled triangles, $R = 0.92$, $m = 0.58$), and $\pi = 1.8$ (open circles, $R = 0.88$, $m = 0.79$).

interiors, while n_{i0} corresponds roughly to the average fractional exposed surface area. Figure 10 plots the partition energies extracted from each protein set against the experimental energies for transferring each amino acid from water to octanol.

We find that (1) extracted energies correlate with experimental partition energies, consistent with the use of the Boltzmann expression, but (2) there is no single temperature that is relevant. The first point is supported by the high correlation ($R = 0.88$ to 0.92) for all five protein sets. The relevant temperature is determined by the slopes of these plots. In Figure 10 the temperature relevant to the set $\pi = 1.8$ is $T = 300\ \mathrm{K}/[(0.59)(0.79)] = 640\ \mathrm{K}$, and to the set $\pi = 1.0$ is $T = 300\ \mathrm{K}/[(0.59)(0.28)] = 1800\ \mathrm{K}$, based on slope factors of 0.79 and 0.28 relative to oil/water partitioning at 300 K. This result suggests that with respect to interior-exterior residue partitioning the proteins in the PDB may not be well-modeled by the random heteropolymer assumption of Finkelstein *et al*. (1993). Figure 10 shows that the effective temperature of interior-exterior partitioning depends on the length, composition and compactness of the proteins in the database, while the random heteropolymer model results are independent of protein length (due to cancellation of length-dependent terms) and composition (due to sequence averaging).

What is the physical meaning of the apparent ''temperature'' of a single protein structure or a database of structures? Here is an analogy, based on the buried/exposed partitioning of amino acids. Non-polar side-chain surface is buried in its ''ground state'' and exposed in its ''excited state.'' In a Boltzmann distribution, the amount of surface in the excited (exposed) state will increase with increasing temperature. But for a large protein having few enough hydrophobic monomers (high partition propensity) that it buries all its hydrophobic residues in the core (the ground state), the Boltzmann analogy gives an apparent temperature of 0 Kelvin with respect to hydrophobic residue partitioning. On the other hand, small proteins with many non-polar residues will be ''hotter'' because those proteins are ''forced,'' by surface-to-volume and composition constraints, to expose hydrophobic monomers. Hence proteins and databases can differ in their ''temperatures'' of interior-exterior partitioning.

Can statistical potentials correctly recognize native structures?

The results above indicate that statistical potentials may not quantitatively reflect the true energies that cause amino acid pairings in proteins. Here we ask if they succeed in a more modest test: do they correctly identify a native structure among a set of decoys? In this case, the value of the temperature is unimportant. The temperature just scales all interactions to the same degree, so while it affects the absolute stability, it does not affect the rank orderings of different structures. Therefore, for a statistical potential to succeed in correctly rank-ordering the energies of different structures, it only needs to be approximately correct to within an arbitrary scaling constant $C > 0$. In the AB lattice model we require only that: $e_{AA} \cong CE_{AA}$, $e_{AB} \cong CE_{AB}$ and $e_{BB} \cong CE_{BB}$.

To test the accuracy of the extracted energies in structure prediction, we now turn to the three-energy AB model. For chains of length $L = 14$ of two monomer types A and B, we create databases of minimum energy structures for different ''true'' contact potential functions, and extract statistical contact energies from each database. We then compare the extracted energies e_{AA}, e_{AB} and e_{BB} to the true energies E_{AA}, E_{AB} and E_{BB}. The most important test is whether, for all sequences in a given database, the true native conformation is identified as having the lowest value of the extracted energy over all possible conformations of that chain sequence. Unlike structure recognition tests that have been performed for real proteins, here we can exhaustively explore the conformational space for each AB sequence.

Table 2 shows the contact energies extracted from databases constructed using different true potentials. To facilitate comparison, both the true and extracted potentials are scaled relative to kT such that the strongest attractive contact interaction has an energy of –5 units. Table 2 shows that in all cases where the three true contact energies are different, the extracted contact energies have the same rank ordering as the true energies. For example, for the true potential $E_{AA}:E_{AB}:E_{BB} = -5:-4:-1$, the extracted potential $e_{AA}:e_{AB}:e_{BB} = -5:-3:+0.8$ correctly predicts

Table 2. AB model test

True $E_{HH}:E_{HP}:E_{PP}$	Extracted $e_{HH}:e_{HP}:e_{PP}$	Number of sequences	Prediction success (%)
–5 : –5 : –1	–5 : –3.7 : +1.4	173	84
–5 : –4 : –1	–5 : –3.0 : +0.8	1388	74
–5 : –3 : –1	–5 : –2.4 : 0.0	1374	64
–5 : –2 : –1	–5 : –2.1 : –0.5	1726	99
–5 : –1 : –1	–5 : –1.5 : –1.0	913	97
–4 : –5 : –1	–1.1 : –5 : +1.8	1059	93
–3 : –5 : –1	–0.8 : –5 : +1.8	1046	95
–2 : –5 : –1	–0.5 : –5 : +1.4	1060	95
–1 : –5 : –1	+1.3 : –5 : +1.3	918	90
–5 : –1 : –5	–5 : +1.2 : –5	915	90
–5 : –1 : –4	–5 : –0.2 : –3.6	2264	99
–5 : –1 : –3	–5 : –0.6 : –2.5	1922	92
–5 : –1 : –2	–5 : –1.1 : –2.1	2040	100
–5 : –3 : +1	–5 : –2.6 : +2.5	1514	96
–5 : –3 : +2	–5 : –2.6 : +4.7	1467	99
–5 : –3 : +3	–5 : –2.6 : +6.8	1465	99

Left column is the scaled set of true energies used to generate the lattice model database for L=14. Second column is the scaled set of energies found by the statistical potential extraction procedures. Column 3 is the number of unique folding sequences in the database for each true potential. Column 4 shows how often the extracted potential correctly identifies the native structure within the database (i.e. the structure having the lowest extracted energy is also the structure with the lowest true energy).

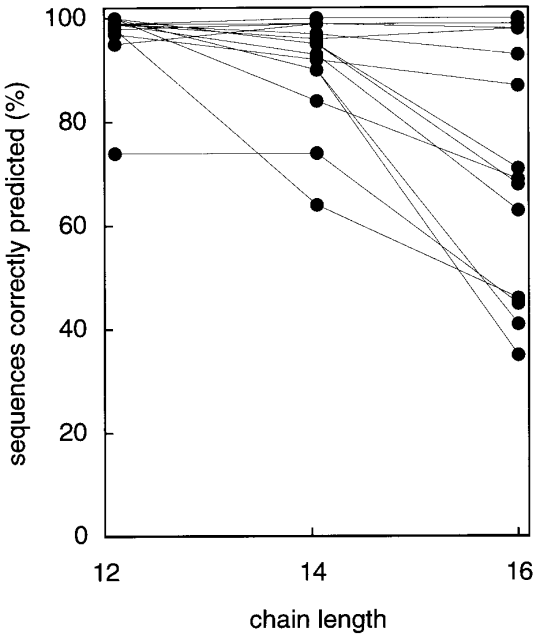

Figure 11. Percent correct structure prediction by the HP model statistical contact potential *versus* chain length for 2D model. The structure of a sequence is predicted correctly if the true native conformation is also lower in extracted energy than any other conformation. Each set of connected points represents a different potential in Table 2. For the 12-mer and 14-mer chains, sequence space searching is exhaustive; for the 16-mer chains, random AB sequences are sampled until 1000 sequences having unique native structures are found for each true potential.

that the AA interaction is the most favorable and the BB interaction is least favorable (although it incorrectly predicts a repulsive BB contact energy).

For the three-interaction model, the extracted potentials qualitatively reflect the true potential. But this is not always sufficient for structure prediction since the extracted contact energies are usually different from the true energies, and sometimes even opposite in sign. As a result, the total extracted contact energy does not always correctly predict the native conformations of AB sequences. For only one of the model databases are the native structures correctly predicted for 100% of the sequences. Sometimes statistical potentials are calculated using a compact reference state rather than the unfolded state of Miyazawa and Jernigan. But this reference state is still a random mixture, and statistical energies calculated using the compact reference are essentially just shifted by a constant positive amount from the values in Table 2. For our lattice model, we find that such a shift generally decreases the success of the potential in identifying correct native conformations.

The extraction procedures fail to reproduce the correct rank ordering of interactions for degenerate potentials, in which two interactions are equal (Figure 2, Table 2). In these cases, the interactions that are equal in the true potential are not equal in the extracted potential (except when the true AA and BB interactions are equal, since the three-interaction potential is symmetric and sequence space is symmetric). This incorrect ''splitting'' of degenerate energies results from the coupling between different interactions, as we noted for the HP model. For the true potential $E_{AA}:E_{AB}:E_{BB}$ = –5:–5:–1, the extracted energies are $e_{AA}:e_{AB}:e_{BB}$ = –5:–2:+1. The AA interaction appears stronger than the AB interaction because few A residues are exposed to solvent (so any A contact is very favorable). Forming an AA contact breaks two A-solvent contacts while forming an AB contact breaks only one, so the extracted AA energy is more favorable. In addition, the sign of the extracted BB interaction is wrong because BB contacts are less common than in a random mixture since Bs prefer to form AB contacts.

For the 14-monomer AB model, the correct native conformations are predicted for 64 to 100% of the sequences in a given database. But Figure 11 shows that the success in structure prediction generally decreases with chain length, even over very short chain lengths from 12 to 16 monomers. Since real proteins are much longer than our lattice chains (having hundreds to thousands of interresidue contacts), Figure 11 suggests that extracted statistical energies may have limited success in identifying the native states of protein sequences among a set of reasonable alternative structures.

Conclusions

We test some of the principles that underlie statistical potentials. Statistical potentials are energy-like quantities that are extracted from protein structure databases, based on certain assumptions. They have been used to model the true energies that cause proteins to fold, dock with ligands, and

recognize other proteins. We test the premises behind statistical potentials using exact lattice models, and we verify our conclusions, where possible, with tests on the PDB.

We conclude that the principal weakness in all the current statistical potentials is their assumption that the frequencies of each type of amino acid pair, such as Ala-Leu, are independent of other types of pairs, such as Phe-Gly. In a relatively small, compact object such as a protein, the space taken by other amino acids is a strong constraint on the possible positions of each given pair. The clustering of hydrophobic amino acids is probably a stronger determinant of the statistical potentials among charged groups than electrostatics are. We find that, whereas true potentials cannot depend on chain length or composition, extracted potentials do. This too appears to be mainly a consequence of the burial of hydrophobic surface in small compact objects of differing sizes and compositions.

There are a few caveats in interpreting our lattice model results. First, excluded volume is a more stringent constraint in two dimensions than in three, simply because of the reduced number of possible spatial neighbors of each residue. Furthermore, sequence effects may be more pronounced in our short-chain model. As a control, we have constructed a database of low-energy configurations of long (60-monomer) HP chains on a 3D lattice, and found results that are similar to those from the 2D HP model. Second, there are only two monomer types, so that the dominant interaction (e.g. the HH interaction) will tend to be the primary determinant of all the observed pair distributions. The real energetics of protein folding are undoubtedly more complex.

For real protein structures, we demonstrate that the use of the Boltzmann distribution law to convert interior-exterior residue partitioning frequencies to energies, which defines a database temperature relative to octanol-water partition energies, is not firmly grounded. The choice of a relevant temperature is strongly dependent on the choice of proteins in the database. We define a quantity we call the partition propensity of a given protein, which determines the relevant temperature in a Boltzmann equation. It may be possible to use this quantity to weight the burial frequencies observed in each protein to obtain database-independent partition energies.

Acknowledgements

P.D.T. is a Howard Hughes Predoctoral Fellow. We thank Kai Yue for suggesting this approach.

References

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

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.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Peptide Protein Res.* **29**, 46–52.

Bryant, S. H. & Lawrence, C. E. (1991). The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions. *Proteins: Struct. Funct. Genet.* **9**, 108–119.

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

Camacho, C. J. & Thirumalai, D. (1993a). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 6369–6372.

Camacho, C. J. & Thirumalai, D. (1993b). Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Letters*, **71**, 2505–2508.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Chan, H. S. & Dill, K. A. (1991a). Sequence-space soup of proteins and copolymers. *J. Chem. Phys.* **95**, 3775–3787.

Chan, H. S. & Dill, K. A. (1991b). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 447–449.

Chan, H. S. & Dill, K. A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238–9257.

Chan, H. S., Bromberg, S. & Dill, K. A. (1995). Models of cooperativity in protein folding. *Phil. Trans. Roy. Soc. Lond. ser. B*, **348**, 61–70.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding—a perspective from simple exact models. *Protein Sci.* **4**, 561–602.

Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic parameters II of amino-acid side-chains from the partitioning of N-acetyl-amino acid amides. *Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.

Finkelstein, A. V., Gutun, A. M. & Badretdinov, A. Ya. (1993). Why are the same protein folds used to perform different functions? *FEBS Letters*, **325**, 23–28.

Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Ya. (1995). Boltzmann-like statistics of protein architectures. Origins and consequences. *Sub-Cell. Biochem.* **24**, 1–26.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl Acad. Sci. USA*, **89**, 12098–12102.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.

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

Kocher, J. P., 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.

Kolaskar, A. S. & Prashanth, D. (1979). Empirical torsional potential functions from protein structure data. *Int. J. Peptide Protein Res.* **14**, 88–98.

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

Lau, K. F. & Dill, K. A. (1989). A lattice statistical mechanics model of the onformational and sequence spaces of proteins. *Macromolecules*, **22**, 3986–3997.

Lau, K. F. & Dill, K. A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl Acad. Sci. USA*, **87**, 638–642.

Lawrence, C., Auger, I. & Mannella, C. (1987). Distribution of accessible surfaces of amino acids in globular proteins. *Proteins: Struct. Funct. Genet.* **2**, 153–161.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Lipman, D. J. & Wilbur, W. J. (1991). Modelling neutral and selective evolution of potein folding. *Proc. Roy. Soc. London, ser. B*, **245**, 7–11.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

MacArthur, M. W. & Thornton, J. M. (1991). Influence of proline residues on protein formation. *J. Mol. Biol.* **218**, 397–412.

Miller, R., Danko, C. A., Fasolka, M. J., Balazs, A. C., Chan, H. S. & Dill, K. A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys.* **96**, 768–790.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.

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

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811–820.

Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine polypeptides in aqueous ethanol and dioxane solutions. *J. Biol. Chem.* **246**, 2211–2217.

O'Toole, E. M. & Panagiotopoulos, A. Z. (1993). Effect of sequence and intermolecular interactions on the number and nature of low-energy states for simple model proteins. *J. Chem. Phys.* **90**, 3185–3190.

Pellegrini, M., & Doniach, S. (1993). Computer simulation of antibody binding specificity. *Proteins: Struct. Funct. Genet.* **15**, 436–444.

Pohl, F. M. (1971). Empirical protein energy maps. *Nature New Biol.* **234**, 277–279.

Rashin, A. A., Ionif, M. & Honig, B. (1986). Internal cavities and buried waters in globular proteins. *Biochemistry*, **25**, 3619–3625.

Richards, F. M. (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.

Serrano, L., Sancho, J., Hirshberg, M. & Fersht, A. R. (1992). Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N- and C-caps and the replacement of alanine by glycine or serine at solvent-exposed interfaces. *J. Mol. Biol.* **227**, 544–559.

Shortle, D., Chan, H. S. & Dill, K. A. (1992). Modeling the effects of mutations on the denatured states of proteins. *Protein Sci.* **1**, 201–215.

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.

Sippl M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, **250**, 1121–1125.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.

Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75–81.

Wilmanns, M. & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proc. Natl Acad. Sci. USA*, **90**, 1379–1383.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct. Funct. Genet.* **6**, 193–209.

Edited by F. Cohen

(Received 6 September 1995; accepted 30 November 1995)