doi:10.1016/j.jmb.2005.07.054 J. Mol. Biol. (2005) 352, 986–1001

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

An Atomic Environment Potential for use in Protein Structure Prediction

Christopher M. Summa[1,2], Michael Levitt[2] and William F. DeGrado[1]*

[1]*Department of Biochemistry and Biophysics, The University of Pennsylvania Medical School Philadelphia, PA 19104-6059 USA*

[2]*Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305-5126, USA*

**Corresponding author*

We describe the derivation and testing of a knowledge-based atomic environment potential for the modeling of protein structural energetics. An analysis of the probabilities of atomic interactions in a dataset of high-resolution protein structures shows that the probabilities of non-bonded inter-atomic contacts are not statistically independent events, and that the multi-body contact frequencies are poorly predicted from pairwise contact potentials. A pseudo-energy function is defined that measures the preferences for protein atoms to be in a given microenvironment defined by the number of contacting atoms in the environment and its atomic composition. This functional form is tested for its ability to recognize native protein structures amongst an ensemble of decoy structures and a detailed relative performance comparison is made with a number of common functions used in protein structure prediction.

Keywords: protein structure prediction; statistical atomic potential; environment potential

Introduction

The development and testing of functions for the modeling of protein energetics is an important part of current research aimed at understanding protein structure and function. For research in the fields of computational protein design,[1–5] protein folding simulation,[6,7] and protein structure prediction,[8,9] the potential function is the keystone upon which such research bears its weight. Each of these applications has very specific criteria that dictate how useful a given potential function will be when used within a specific context. In general, a few criteria emerge that may be used for testing the utility of a new formulation: any useful potential should ideally recognize the native structure for a given protein as a global minimum and its calculation should be computationally efficient.

Much can be learned through statistical analyses of interacting groups in experimentally determined protein structures. Such analyses provide the basis for knowledge-based potentials of mean force,[10] as well as the derivation of the potential described here. We describe the derivation and testing of knowledge-based atomic environment potential for the modeling of protein structural energetics. Multi-body potential functions have shown great promise for threading, fold recognition, and crystal structure validation when derived both at the residue level[11–14] and the atomic level,[15–21] which led us to believe that this approach had a reasonable chance of success. A formulation has been chosen that focuses on the composition of the atomic neighborhood around each of the atoms in the protein. It is from this composition that energies are assigned to each of the proteinaceous atoms. We show that an atomic environment approach allows us to capture structural information that is absent in a pairwise atomic formulation (especially in the absence of explicit solvent). We will also show that this potential is able to recognize the native state of a protein amongst an ensemble of well-formed decoys[22,23] for a large database of native/decoy sets. In the following sections the functional form of the potential, and the methods used to derive statistics from a database of known protein structures are described in detail.

Abbreviations used: PAT, primary atom type; CI, chemical identity; ME, microenvironment; SASE, solvent accessible surface area.

E-mail address of the corresponding author: wdegrado@mail.med.upenn.edu

Results

A knowledge-based environment potential

Potential functions for protein energy modeling

0022-2836/$ - see front matter © 2005 Published by Elsevier Ltd.

applications fall into two general groups, those based on molecular mechanics force-fields,[24–26] and those derived from a dataset of high-resolution protein structures,[10,27,28] often called knowledge-based potentials. We have taken the latter approach because we believe that there is a significant amount of information in high-resolution crystal structures that may be gleaned with statistical analysis. A statistical analysis also affords us a tool with which to test certain hypotheses about the nature of interatomic interactions in proteins.

The potential that is derived and tested here is a non-bonded, atomic environment potential. Each atom in a protein structure is assigned an energy based on the number and composition of the other non-bonded atoms that fall within a sphere of a given radius around the atom in question. We generate the statistics used to derive this potential by analyzing a dataset of 1066 protein chains. The first task is to assign atom "types" to each of the protein heavy atoms.

Atom "type" designations

Two different methods were used for assignment of atom types to atoms in the protein database. The term "primary atom-type", or PAT, will be used for the first atom type designation. Each atom in the database was assigned to one of 20 different PAT groups, based on chemical identity, and degree of chemical substitution. The PAT assignments for the 20 naturally occurring amino acids are listed in Table 1. Each atom was also given another designation, which is used to define the microenvironment of a PAT. After examining a number of different clustering algorithms and schemes to reduce the 20 primary atom types to a more manageable set of environmental atoms, we ultimately settled for a scheme based purely on whether the atom is a carbon, oxygen, sulfur, or nitrogen, which we refer to as the "chemical identity" or CI. Each protein atom in the database, therefore, was assigned two different atom-type identifiers: the PAT, with a range of 20 possible values, and the CI, which has four possible values.

Microenvironment frequency calculation

Each microenvironment (or ME) is defined as the group of CI atoms that are within a radial distance r of a given PAT (Figure 1) minus those atoms that are ignored due to exclusion rules. Atoms that are within s residues along the chain relative to the residue that contains the reference PAT are excluded, effectively excluding all atoms bonded to the PAT atom. The value s will be referred to as the "residue skipping number". An ME is specified by: the number of each type of CI atoms that it contains, and the total number of atoms it contains, or the coordination number (Figure 2). For example, the microenvironment [2N,1C,1O,0S] has a coordination number of 4. Each atom in the database is assigned a PAT, and a ME. The observed frequency of each PAT/ME is calculated by counting over all the protein atoms in the high-resolution protein database.

Counting methods and derivation of pairwise probabilities

In an effort to understand how much "information" is carried by this construction relative to

Table 1. Atom type designations for atoms found in the 20 commonly occurring amino acids in proteins

Atoms in category	PAT	CI
All residues—N	1	N
All residues—C^{α}	2	C
All residues—C	3	C
All residues—O	4	O
Asp—$O^{\delta1}$, $O^{\delta2}$: Glu—$O^{\varepsilon1}$, $O^{\varepsilon2}$	5	O
Val—$C^{\gamma1}$, $C^{\gamma2}$: Leu—$C^{\delta1}$, $C^{\delta2}$: Ile—$C^{\gamma2}$, $C^{\delta1}$: Met—C^{ε}: Thr—$C^{\gamma2}$	6	C
Ala, Leu, Glu, Gln, Phe, Tyr, Trp, Met, Pro, Arg, Lys—C^{β}: Met, Pro, Arg, Lys—C^{γ}: Ile—$C^{\gamma1}$: Lys—C^{δ}	7	C
Ser, Cys, Asp, Asn, His—C^{β}: Glu, Gln—C^{γ}: Pro, Arg—C^{δ}: Lys—C^{ε}	8	C
Val, Ile—C^{β}: Leu—C^{γ}	9	C
Phe, Tyr—$C^{\delta1}$, $C^{\delta2}$, $C^{\varepsilon1}$, $C^{\varepsilon2}$: Phe—C^{ζ}: Trp—$C^{\varepsilon3}$, $C^{\zeta2}$, $C^{\zeta3}$, $C^{\eta2}$	10	C
Asp, Asn, His—C^{γ}: Glu, Gln—C^{δ}: Tyr, Arg—C^{ζ}: Trp—$C^{\varepsilon2}$	11	C
Ser, Thr—O^{γ}: Tyr—O^{η}	12	O
Asn—$N^{\delta2}$: Gln, His—$N^{\varepsilon2}$: Trp—$N^{\varepsilon1}$: His—$N^{\delta1}$	13	N
Cys—S^{γ}: Met—S^{δ}	14	S
Thr—C^{β}	15	C
Trp, His—$C^{\delta1}$: His—$C^{\varepsilon1}$	16	C
Phe, Tyr, Trp—C^{γ}: Trp—$C^{\delta2}$	17	C
Asn—$O^{\delta1}$: Gln—$O^{\varepsilon1}$	18	O
Arg—N^{ε}, $N^{\eta1}$, $N^{\eta2}$	19	N
Lys—N^{ζ}	20	N

The left-hand column lists the atoms that are grouped into subcategories, based on common traits such as bonding pattern, partial charge, and hydrophobicity. There are two atom types assigned to each atom, the primary atom type, or PAT, is a fine-grained designation and is used to describe the reference atom in a microenvironment cluster. The chemical identity type (or CI) is used to describe the composition of the microenvironment cluster in the vicinity of a given reference atom.

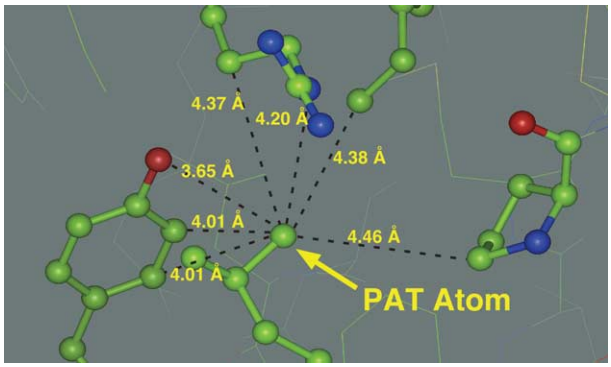


Figure 1. Depiction of a cluster of atoms forming a microenvironment around the methyl carbon PAT atom. All those atoms within a radius of 4.6 Å (the counting radius) of the PAT atom are counted with one exception: If the sequential distance between the residue containing the PAT atom and the residue containing the CI atom is smaller than the residue skipping number, the contact is ignored. The microenvironment shown here has two fundamental properties: the coordination number of the reference atom, and the chemical identities of the contacting atoms. The Figure was generated using InsightII from MSI, Inc.

what might be contained in a simple pairwise potential, we compared the frequency with which each PAT/ME is seen in our training database to the frequency expected if the probabilities of each interaction were statistically independent events (i.e. purely pairwise). Our null hypothesis, to which our measured frequencies are compared, can be stated as follows: the probability of finding an atom of type X in the environment of our reference atom is independent of the probability of finding any other atom of type Y in the same environment. The expected values for the null hypothesis were calculated as follows.

The expected conditional probability of a particular PAT/ME ($p_{(\mathrm{ME,coord|PAT})}$) where the probability of finding each atom in an ME is a statistically independent event is defined as the product of the conditional probability of finding a given coordination number once the PAT is known ($p_{(\mathrm{coord|PAT})}$) and the probability of finding an ME once both the PAT and coordination number are known ($p_{(\mathrm{ME,coord|PAT})}$) as follows:

$$p_{(\mathrm{ME,coord|PAT})} = p_{(\mathrm{coord|PAT})} \times p_{(\mathrm{ME|coord,PAT})} \quad (1)$$

where the expected probability of a given coordination x once the PAT (of type y) is known $p_{(\mathrm{coord}_y|\mathrm{PAT}_x)}$ is approximated as:

$$p_{(\mathrm{coord}_y|\mathrm{PAT}_x)} = \frac{f_{(\mathrm{coord}_y|\mathrm{PAT}_x)}}{\sum_{i=0}^{\max(i)} f_{(\mathrm{coord}_i|\mathrm{PAT}_x)}} \quad (2)$$

$f_{(\mathrm{coord}_y|\mathrm{PAT}_x)}$ is the database frequency of a coordination number y for a given PAT_x. Since we

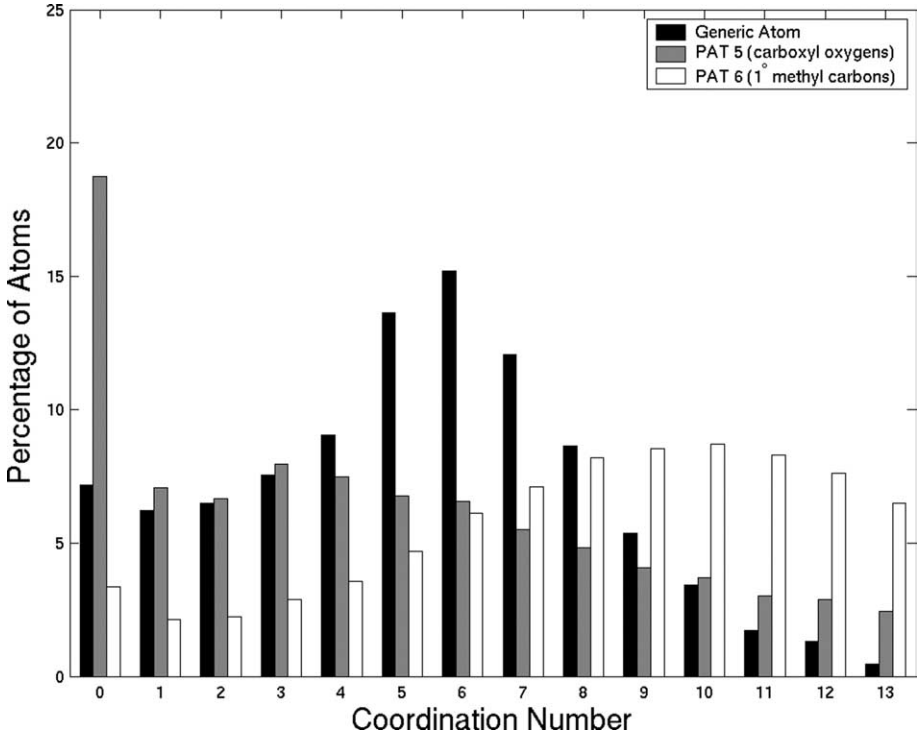


Figure 2. Coordination number histograms. Shown are the distributions of coordination numbers for two specific primary atom types (PATs) and the distribution over all atom types. An atom with a PAT of type 6 (white bars), which is a methyl carbon, is most likely to be buried and have a number of neighboring non-bonded atoms in its environment (most likely coordination is 9). In contrast, a carboxyl O (PAT 5, grey bars) is most likely to be fully solvent-exposed and have a coordination number of 0. These histograms were generated using a radial distance of 4.8 Å and a residue skipping number (s) of 3. The Figure was generated using MATLAB (The MathWorks, Inc., Natick, MA).

want to calculate expected frequencies for each PAT/ME we need to define the expected or standard state probability of finding a given coordination number for each atom type, which we define as $p_{(\mathrm{coord}_y|\mathrm{PAT}_{\mathrm{baseline}})}$. Thus, it is necessary to specify the coordination number distributions, using some standard state assumption. We tried several procedures, and chose the one that provided the best performance in practical applications. Perhaps the simplest is to use the coordination number probability distribution observed for all atoms in the database as the standard state distribution. The problem, however, is that some atom types occur more frequently than others, which skewed the distributions. We therefore first computed the frequency distributions for each atom type, and then took the unweighted mean of these distributions to approximate the standard state probability distribution of coordination numbers.

Specifically, $p_{(\mathrm{coord}_y|\mathrm{PAT}_{\mathrm{baseline}})}$ was calculated by normalizing the frequency distributions for the specific PATs such that each distribution contributed equally to the baseline distribution. The normalized coordination number frequency distributions of the frequencies were generated as follows:

$$f^{\mathrm{norm}}_{(\mathrm{coord}_y|\mathrm{PAT}_i)} = f_{(\mathrm{coord}_y|\mathrm{PAT}_i)} * \left[\frac{\frac{1}{20} * \sum_{j=0}^{\max(j)} \sum_{i=1}^{N_{\mathrm{PAT}}} f_{(\mathrm{coord}_j|\mathrm{PAT}_i)}}{\sum_{j=1}^{\max(j)} f_{(\mathrm{coord}_j|\mathrm{PAT}_i)}} \right] \tag{3}$$

where $f_{(\mathrm{coord}_x|\mathrm{PAT})}$ is the observed frequency of a given coordination number for a given PAT. The expression in brackets represents the normalization factor, with the numerator representing 1/20th of the total number of atoms in all distributions, and the denominator representing the number of atoms in the current distribution. max(j) is the total number of coordinations considered (15 in this case) and N_{pat} is the total number of PATs. We can then define the "baseline" distribution as the simple sum of the normalized frequency distributions:

$$f_{(\mathrm{coord}_y|\mathrm{PAT}_{\mathrm{baseline}})} = \sum_{i=1}^{N_{\mathrm{PAT}}} f^{\mathrm{norm}}_{(\mathrm{coord}_y|\mathrm{PAT}_i)} \tag{4}$$

where the sum is taken over all 20 PAT atom types. This leads to the calculation of the expected (baseline) probability of an atom in our database having a given coordination number:

$$p_{(\mathrm{coord}_y|\mathrm{PAT}_{\mathrm{baseline}})} = \frac{f_{(\mathrm{coord}_y|\mathrm{PAT}_{\mathrm{baseline}})}}{\sum_{j=0}^{\max(j)} f_{(\mathrm{coord}_j|\mathrm{PAT}_{\mathrm{baseline}})}} \tag{5}$$

We also examined other methods to create these expectation distributions that were based on the topology of a given atom within a residue or chemical characteristics. However, more fine-grained custom distributions for individual atom types failed to significantly improve performance.

The expected probability of finding a given ME about an atom of a given PAT and coordination is:

$$p_{(\mathrm{ME}|\mathrm{coord},\mathrm{PAT})} = W \prod_{i=1}^{y} p_{(\mathrm{CI}(i)|\mathrm{coord},\mathrm{PAT})}{}^{n_{(\mathrm{CI}(i))}} \tag{6}$$

where n is the number of atoms of type CI(i) in a given ME, and $p_{(\mathrm{CI}(i)\mathrm{coord},\mathrm{PAT})}$ is the fraction of atoms of type CI(i) found in the database about a given PAT with a given coordination, and W is the number of possible ways of producing the composition of the ME, and is defined as:

$$W = \frac{\left(\sum_{i=1}^{y} n_{(\mathrm{CI}(i))}\right)!}{\prod_{i=1}^{y} (n_{(\mathrm{CI}(i))}!)} \tag{7}$$

where y is the number of CI atom types, in this case, four.

Figure 3 shows the percentage of atoms in our protein database that have PAT/MEs whose frequency of occurrence in significantly (>95%) different from what we expect if each atom in the microenvironment is statistically independent from every other atom in the environment. Significance values were measured *via* a Poisson distribution for PAT/MEs in which the number of counts was lower than ten, and for PAT/MEs with counts of ten or higher, a chi-squared distribution was used. It is interesting to note that, for a counting radius of between 4.0 Å and 5.0 Å and a skipping number s of between 2 and 4, greater than 60% of all atoms in the database fall into PAT/MEs that differ significantly from the null hypothesis (i.e. statistical independence of probabilities).

Calculating a pseudo-energy

Satisfied that we were capturing information that by definition could not be present in a pairwise treatment of the atomic interactions in a database of known protein structures, we generated a scoring function based on the PAT/ME framework. In order to generate a score, however, we needed to compare the counts we see in the database to an expected frequency, though a slightly different one than that used above to measure statistical independence. The "standard state" or expected frequency of a particular PAT/ME, assuming that a given ME occurs *via* random chance with no specific interactions between atoms, is defined in a similar manner as above (equations (1)–(7)), except that the individual probabilities for finding a given CI (the $p_{(\mathrm{CI}(i))}$) are now assumed to be independent of the coordination number and PAT. In other words, the probability of finding a nitrogen in the ME is calculated based solely on the fraction of nitrogen atoms in the database as a whole, and has nothing to do with the choice of a PAT or coordination number. This has the effect of setting a standard

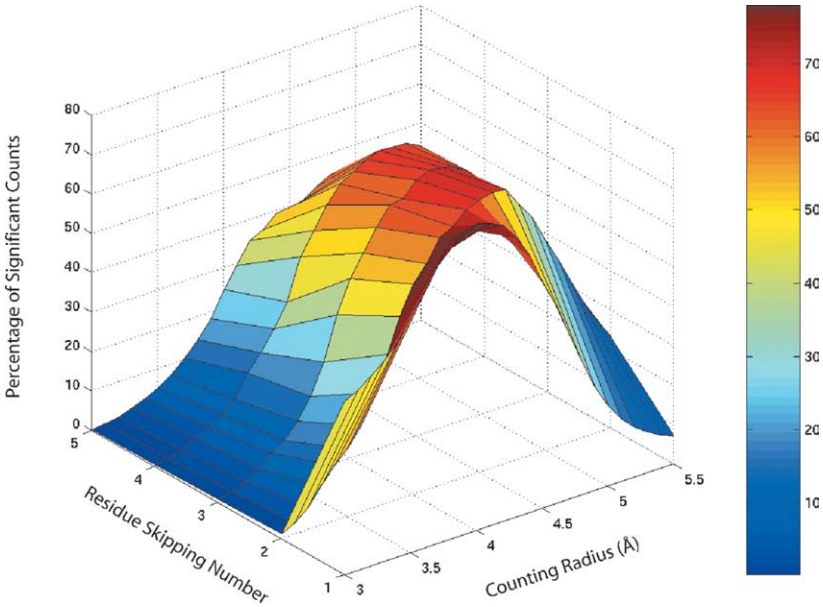


Figure 3. How many atoms have clusters that are significantly different from the pairwise expectation? Histogram showing the number of atoms in the training dataset that have a PAT/ME that occurs with a frequency significantly (>95%) different from what would be expected if all the atoms in the ME occurred with a probability that is statistically independent from the probabilities of the other atoms. For certain counting radii and skipping number, the percentage can achieve values as high as 77% (in the fringe region nearest the viewer at counting radius 4.0 Å and residue skip 1). This Figure was generated using MATLAB (The Mathwworks Inc., Natick, MA).

state in which each PAT sees the same mixture of atoms in its MEs, and true preferences for different compositions are more apparent.

The Boltzmann construction was used to express a pseudo-energy with a functional form defined as the natural log of the ratio of the number of times a given PAT/ME is found in the database ($f_{(\mathrm{ME,coord|PAT})}$) to the number of times it is expected to occur ($p_{(\mathrm{ME,coord|PAT})}$):

$$E_{(\mathrm{ME,coord|PAT})} = -\ln \frac{f_{(\mathrm{ME,coord|PAT})}}{p_{(\mathrm{ME,coord|PAT})}} \qquad (8)$$

The pseudo-energy of a particular configuration of a protein chain can then be defined as the sum over all N atoms in the protein:

$$E_{\mathrm{tot}} = \sum_{i=1}^{N} E_i \qquad (9)$$

The form of equation (9) leaves open the question of how to assign an energy to a PAT/ME for which the $(f_{(\mathrm{ME,coord|PAT})})=0$ (i.e. is never seen in the database). Since we do not *a priori* distinguish between those PAT/MEs that were underrepresented because of the size of our database, those that were underrepresented due to particularly unfavorable energetics, and those that were physically impossible *via* steric arguments, we apply the following statistical argument to estimate the counts: If the distribution of counts is assumed to be binomial, an estimate of counts expected given an infinite sample size can be calculated by taking the highest possible mean value consistent with seeing no counts (with a 95% confidence interval). While this is a best-case estimate, in practice, the estimated numbers are extremely small, resulting in an unfavorable energy when the Boltzmann formalism is applied.

Choosing a PAT/ME variant *via* a jackknife test

A stringent test of any potential function is its ability to detect the correct folded conformation for a given sequence from a library of thoughtfully constructed decoys. This type of test is notoriously difficult for even the most sophisticated potential functions, and no known potential function has yet shown 100% accuracy at decoy detection for every decoy set. We tested variants of our microenvironment potential function (each having a different critical distance from the PAT atom and/or residue skipping number) for its ability to discriminate between a native protein structure and a well-constructed decoy set. The aim was to choose a variant of the PAT/ME for future use. Potentials were generated with values of r between 3.5 Å to 5.1 Å in 0.1 Å increments (inclusive) with values of s between 1 and 4 (inclusive), resulting in 68 possible potentials.

For the jackknife test, the Decoys-R-Us database[23] was split up three different ways, creating three training sets, and three corresponding test sets. There was no overlap between the training set and its corresponding test set. Each training set was then scored with each of the 68 possible PAT/ME potentials. The measurement we used to judge success in the decoy scoring was the sum-of-log-ranks, or:

$$\mathrm{Score} = \sum -\log(\mathrm{Rank}_{\mathrm{native}}) \qquad (10)$$

where R_{native} is the rank of the score of the native when compared to all the decoys in the set. The smallest negative number is best using this metric, with a score of 0.0 being the best possible. The best scoring PAT/ME for each training set was then used to score each of the corresponding test sets, and these scores are compared (Table 2) with results obtained using a simple van der Waals potential

taken from the AMBER force field,[25] a pairwise electrostatic potential term also taken from AMBER, the sum of these two terms (representing the entire non-bonded contact energy of a typical molecular mechanics force field without either an explicit or implicit solvent model), CHARMM19[24] van der Waals and coulombic terms, and the ΔE and ΔE^{solv} of Delarue & Koehl,[18] the ΔG^{env} of Koehl & Delarue,[17] and distance-dependent atomic potentials RAPDF[29] and DFIRE.[30] In each case, the chosen variant of the PAT/ME showed performance exceeding (or nearly exceeding) the best performance of the other potentials tested in this work (Table 2, columns 2–3). The values for the sum-of-log-ranks shown in Table 2 represent the scores achieved on the Test set for each of the potentials tested. In choosing the best parameters for the PAT/ME function, the variant with the best score for the Training set (in bold, data not shown) was chosen for comparison to the other potentials. The sum-of-log-ranks metric consistently placed the PAT/ME among the top four potentials with RAPDF, the CHARMM19 pairwise electrostatic potential term, and the ΔE of Delarue & Koehl. For each training set, a slightly different variant of the PAT/ME potential was chosen as "best" *via* the jackknife procedure. For set 1, the best PAT/ME chosen from the training set had a 4.8 Å radius and a three residue skip, for set 2, a 4.5 Å radius and a three residue skip, and for set 3 a 4.8 Å radius and a one residue skip. The potentials were also compared to one another *via* another metric, the number of sets found where that energy function ranked the native structure as having the best (lowest) energy. A "perfect" function would have a score equal to the total number of decoy sets (132 in this case), where a "poor" one would have a score of zero. Using this metric, the PAT/MEs also scored well relative to the other potentials, with RAPDF, the CHARMM19 electrostatics term, and the ΔE of Delaure & Koehl and also scoring somewhat poorer (Figures 4–6).

Two representative data sets (for the decoy set Ig structal-1igc, and the decoy set 4-state-reduced-1ctf) are shown in Figure 6. Both an easy and a hard decoy set are shown. For this type of test, the ideal function would, of course, have the native structure as the lowest energy. One would hope, also, that the energy would decrease with r.m.s.d. as we approached the conformation of the native

Table 2. Results of jackknife test procedure

	Sum of log scores for test sets			# Sets in which native is ranked best
Scoring function used	Test set 1	Test set 2	Test set 3	Full set
United-atom vdW (AMBER)	−193.8	−340.2	−188.3	24
Coulombic (AMBER)	−149.9	−278.8	−146.0	34
United-atom vdW+coulombic (AMBER)	−181.5	−327.0	−180.4	25
United-atom vdW (CHARMM19)	−184.1	−331.4	−176.2	32
Coulombic (CHARMM19)	−78.6	−143.3	−69.9	78
ΔE (Delarue and Koehl)	−86.9	−132.7	−50.1	73
ΔE^{solv} (Delaure and Koehl)	−175.1	−320.2	−170.7	20
ΔG^{env} (Koehl and Delarue)	−133.9	−285.9	−156.8	34
RAPDF	−114.6	−218.3	−103.7	54
DFIRE	−145.8	−299.0	−154.8	39
4.8_3 (Best performance for training set 1)	**−68.7**	−91.3	−25.8	97
4.5_3 (Best performance for training set 2)	−87.4	**−139.1**	−54.0	96
4.8_1 (Best performance for training set 3)	−64.3	−118.8	**−70.8**	84

A jackknife test was performed to find the optimal values of the counting radius and residue skip parameters for the PAT/ME potential functions. In each jackknife set, values of the counting radius were varied from 3.5 Å to 5.1 Å in 0.1 Å increments and the residue skip was varied from one to four residues. Potentials for every possible combination of counting radius and residue skip were generated and used to score "training" sets (corresponding to some fraction of the total decoy database) and the "test" sets (corresponding to whatever decoy sets remained). The differences between the test sets lie in the manner in which the entire dataset was split into training set and test set. For set 1, every other decoy set of the full decoy data set was assigned to the training set, and the rest comprised the test set. For set 2, the training set consisted of only the 4-state-reduced decoy sets[22] and the remaining decoys were placed in the test set. Set number 3 is the converse of set 1 (i.e. test set 1 was used for training and training set 1 was used for testing). The full decoy set consisted of a total of 132 decoy sets. Values for sum of log scores for test sets indicated in bold represent the multi-body scoring function that scored best on the respective training set. Since this score is defined as $\text{Score} = \Sigma - \log(\text{Rank}_{\text{native}})$, the best possible score is zero, and scores with larger absolute values reflect poorer discrimination. The final column represents the number of times a given function had the best rank for its native structure over the entire test set. Set 1 had 67 subsets in its training set and 66 in its test set. Set 2 had seven subsets in its training set and 126 in its test set, and set 3 had 66 subsets in its training set and 67 in its test set. The united-atom vdW (AMBER) and united-atom Coulombic (AMBER) scores were generated from the united atom AMBER force field.[25] The united-atom vdW (CHARMM19) and united-atom coulombic (CHARMM19) potentials were extracted from the CHARMM19 force field.[24] ΔE and ΔE^{solv} are environment potentials characterized by the fractions of atomic surface accessible to polar or non-polar atoms.[18] ΔG^{env} is a free energy based on environment[17] and RAPDF[29] and DFIRE[30] are statistically derived pairwise atomic potentials.

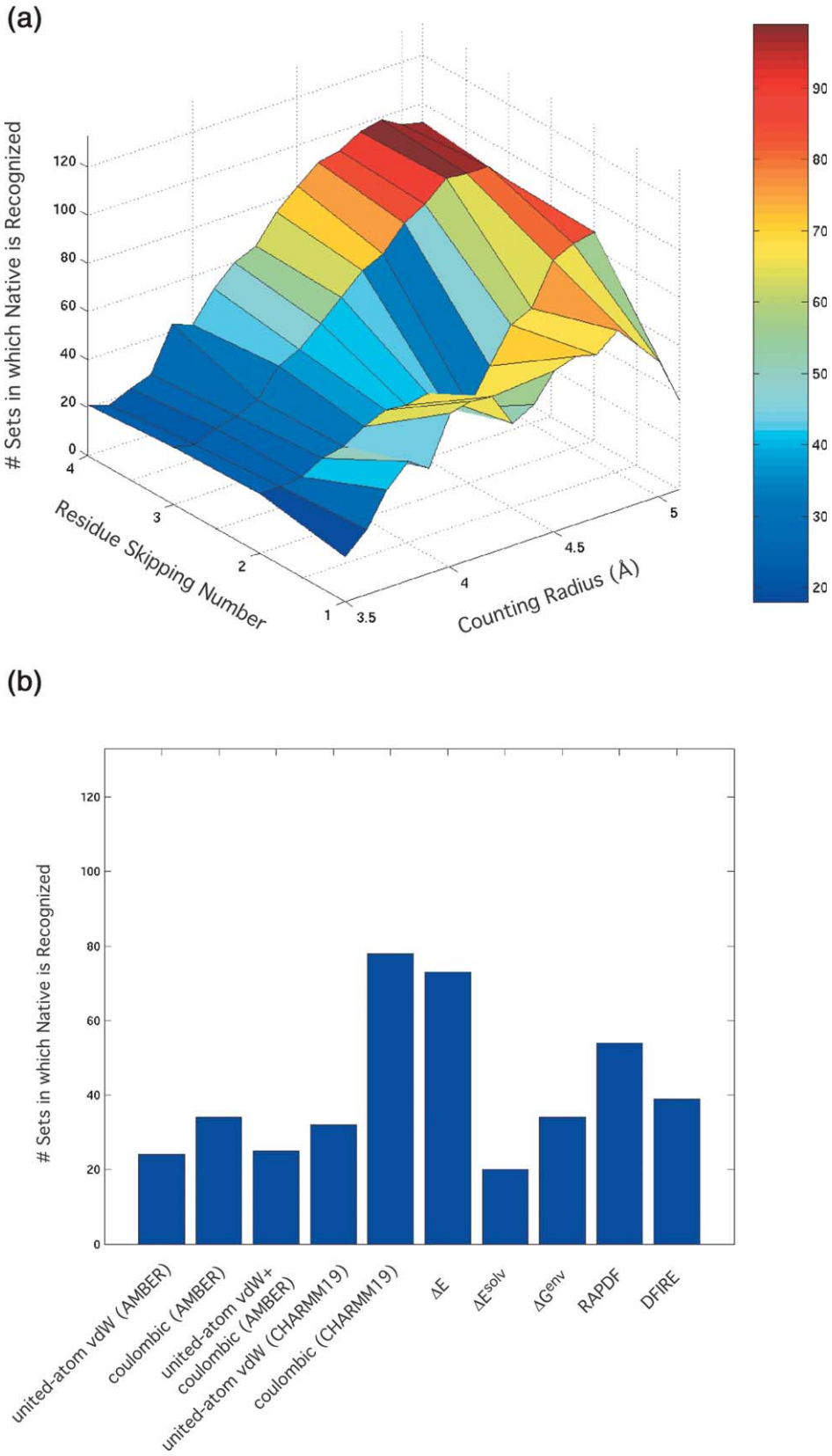
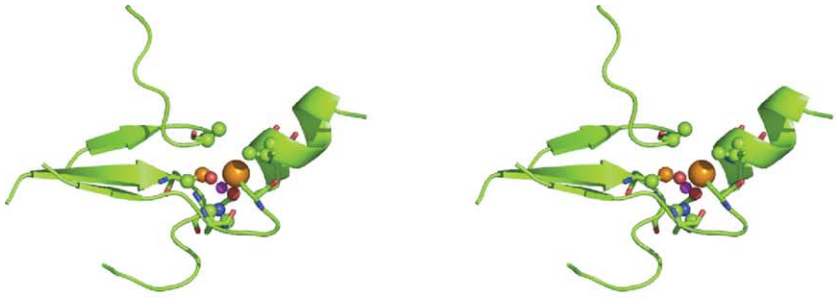
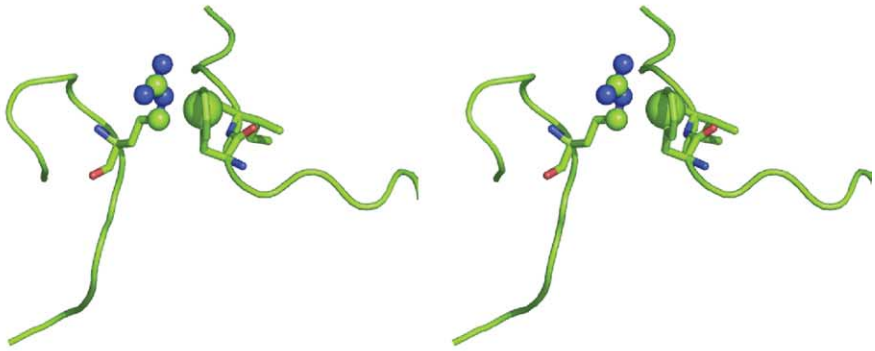


Figure 4. Number of decoy sets in which the native structure is assigned the lowest energy. For each of the energy functions tested, the total number of decoy sets (out of a possible 132) for which the native is given the best rank is shown. In (a), the microenvironment score is tested for its ability to recognize the native structure, and in (b), for comparison, are the other potentials tested in this work. This Figure is a graphical representation of the last column of Table 2, and was generated using MATLAB (The MathWorks Inc., Natick, MA).



(a) PAT=14, Rank=3, PDB ID=1i6p
(3N,8C,1O,1S)



(b) PAT=10, Rank=2, PDB ID=1arb
(3N,2C,0O,0S)

(c) PAT=19, Rank=1, PDB ID=1e6u
(3N,2C,0O,0S)

Figure 5. Examples of PAT/MEs from the database occurring more frequently than the pairwise expectation. These clusters (shown as stereo pairs) occur more often than would be expected if each atom occurrence were a statistically independent event. For each PAT, the 20 MEs that show the greatest deviation from expected counts were ranked (data not shown) For each example PAT/ME shown above, both the PAT and the atoms of the ME are highlighted, and the PAT is rendered slightly larger than the others to distinguish it from the atoms of the ME. Carbon atoms are shown in green, nitrogen atoms in blue, oxygen atoms in red, and sulfur atoms in orange. (a) An example of a PAT/ME around a Cys sulfur atom, in which a Zn binding site is recognized as occurring more often than predicted *via* the pairwise approximation. The Zn and water oxygen (shown) are not explicitly considered as part of the ME, yet this is recognized as a particularly good cluster around Cys sulfur. Examples (b) and (c) show identical MEs (3N,2C,0O,0S) but for different PAT atoms. In (b), the PAT is a Phe ring carbon, and the Figure shows it packed against the guanidino group of an Arg side-chain. In (c), a tightly packed Arg-Arg pair can be seen: the PAT in this case is the nitrogen of the guanidino group of the Arg side-chain. This Figure was generated using the program PyMOL.[58]

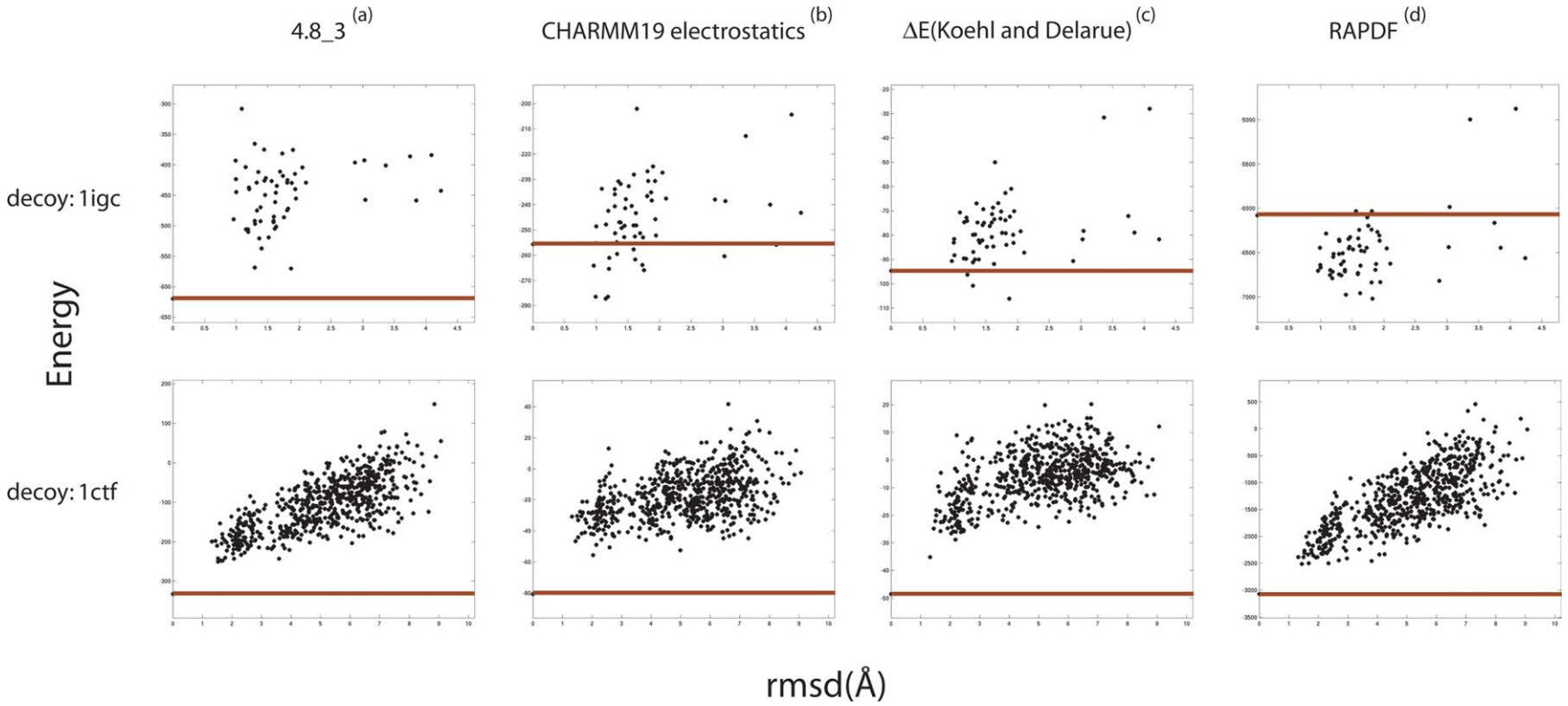


Figure 6. Representative scatter plot of the predictive capacities of different atomic scoring functions. A red horizontal line highlights the score of the native state; p, this represents the baseline score to which the scores of the decoys are compared. A perfect scoring function should not generate a score for a decoy that is below this basal level. It is also beneficial to see a trend in which the score decreases toward that of the native structure as the r.m.s.d values between the decoys and the native structure approach 0. For the PAT/ME scoring functions, the radial distance (r) and the residue skipping number (s) are shown in the title of the plot. This Figure was generated using the program MATLAB (The MathWorks Inc., Natick, MA).

structure. The Figure clearly shows that while this trend exists, none of the functions tested shows perfect behavior under both these criteria.

Detailed performance testing on an expanded decoy database

Once a PAT/ME variant had been chosen for testing, the criteria set for comparison and the database of decoys, were greatly expanded. The Rosetta all atom decoy set[31] and the Loops decoy set[32] were added to the analysis and the following analyses were performed for each decoy and potential. The details of this analysis have been compiled into tables and placed in the Supplementary Data in order to maintain the readability of this paper.

Native rank test

This test ranks the native structure amongst the ensemble of non-native decoys using each of the scores tested in the section above (results are presented in Supplementary Data, Table S2). Since the ability to discriminate the native structure from non-native structures is a stringent test of the performance of any potential used in protein structure prediction, reasonable performance on this test is important. Some decoy sets can be classified as "easy", since most of the functions tested were able to recognize the native state, while some are very difficult, with the native not recognized by any of the functions tested (as can be seen in Supplementary Data, Table S2, and in the compiled comparison data in Supplementary Data, Table S6).

Best decoy rank test

The r.m.s.d. rank of the best scoring decoy in each set was calculated (Supplementary Data, Table S2). This rank is a measure of the practical utility of a given function for near native decoy recognition.

Correlation of score versus C^{α}*-r.m.s.d.*

The linear correlation between scores and C^{α}-r.m.s.d. values is commonly employed[30,33] to show that a trend exists that might allow low C^{α}-r.m.s.d. decoys to be chosen from the decoy set by choosing low-scoring decoys. We also examined the rank order correlation, because there is not necessarily a linear relation between the r.m.s.d. of a series of decoy and its energy. The correlation coefficients determined by the linear and rank correlation methods are highly related ($r=0.98$). Therefore, although both values are reported (Supplementary Data, Table S6), only the linear correlation coefficients were used in the final combined analysis.

Table 3. Compiled relative performance data

	PAT/ME (4.8_3)	United-atom vdW (AMBER)	Coulombic (AMBER)	United-atom vdW+coulombic (AMBER)	United-atom vdW (CHARMM19)	Coulombic (CHARMM19)	ΔE (Delarue & Koehl)	ΔG^{env} (Koehl & Delarue)	ΔE^{solv} (Delarue & Koehl)	RAPDF	DFIRE
Native rank	117	81	41	80	94	128	95	70	22	83	89
Rank of best decoy	27	13	29	16	37	38	26	43	19	29	55
Linear correlation	8	7	2	8	10	10	1	117	4	30	47
Rank order correlation	10	5	3	5	12	16	7	102	4	43	41
log P_{B1}	27	14	29	16	37	38	27	43	19	30	54
log P_{B10}	62	61	72	66	75	80	68	88	50	75	92
Z-score	49	0	10	4	27	66	12	12	2	6	40
Enrichment	28	18	28	18	46	39	26	75	15	41	42

Shown is the sum over all sets tested in which each potential can be labeled as the best relative performer. Scores for each metric are weighted equally.

Table 4. Correlation analysis of tested metrics

	Native rank	Rank of best decoy	Rank order correlation	log P_{B1}	log P_{B10}	Z-score	Enrichment
Native rank	1.00	0.28	0.01	0.29	0.35	0.74	0.20
Rank of best decoy		1.00	0.61	0.99	0.87	0.53	0.80
Rank order correlation			1.00	0.62	0.60	0.02	0.88
log P_{B1}				1.00	0.87	0.54	0.79
log P_{B10}					1.00	0.38	0.76
Z-score						1.00	0.27
Enrichment							1.00

The correlation in performance of each pair of testing metrics is shown. Highly positively correlated metrics are assumed to be redundant measures of performance.

Fraction enrichment

The fraction enrichment metric (Supplementary Data, Table S4) is a measure of the overlap between: (1) the set of decoys consisting of the top 10% with the lowest C^{α}-r.m.s.d.; and (2) the set of decoys consisting of the top 10% of calculated scores.

Z-score

The Z score is defined as:

$$Z = -\frac{E_n - \langle E \rangle}{\sigma} \tag{11}$$

where E_n is the energy of the native protein, $\langle E \rangle$ is the mean energy over all decoys, and σ is the standard deviation of the distribution of decoy energies. A large positive number for this metric is preferable because it implies a large energy gap between the native structure and the mean of the decoy energy distribution (Supplementary Data, Table S4).

Probability of choosing the "best" decoy

A direct measure of a scoring function's utility in structure prediction is the probability that the best scoring decoy is also the lowest C^{α}-r.m.s.d. decoy in the set. This value, presented as the $\log_{10} P_{B1}$[33] (Supplementary Data, Table S5), is related to the rank of the best scoring decoy metric above, but in this construction, the overall size of the decoy set is taken into account as a normalization factor to differentiate between larger, more difficult decoy sets and smaller, easier ones.

Probability of choosing the best decoy among the top ten scoring decoys

A slightly less stringent criterion for comparison is the probability that the decoy with the lowest in a given set is among the ten best-scoring decoys. These values ($\log_{10} P_{B10}$ as per Wang *et al.*[33]) are also presented in Supplementary Data, Table S5.

Relative comparison of performance

In order to compare the overall performance of each of the potentials tested here, the potentials were ranked for each decoy/performance measure pair, and the best performer (or set of performers) was tabulated (Supplementary Data, Table S6). No one potential performs consistently best for all comparison metrics. While CHARMM19 united-atom Coulombic emerges as the best performer in the native structure rank test (with the PAT/ME coming in second), it does not perform as well as DFIRE in the best decoy rank test (Table 3). The clear "winner" in both the linear and rank order correlation tests is ΔG^{env}. The best scoring function in the log probability tests is DFIRE, followed closely by ΔG^{env}, RAPDF. In the Z score test, CHARMM19 Coulombic emerges with as the best performer, followed by PAT/ME and DFIRE and in the fraction enrichment test, ΔG^{env} again takes the top spot, followed, interestingly, by CHARMM19 van der Waals term.

To determine the overall best performance, we wished to use as different performance measures as possible, and to remove measures that were redundant to avoid biasing the conclusions by inclusion of multiple performance measures that test essentially the same thing. This was accomplished by conducting a correlation analysis of the performance metrics for all possible pairs of potential function (Table 4). The "rank of best decoy" measure correlates quite strongly with log P_{B1} and log P_{B10} ($r = 0.99$ and 0.87, respectively). To create a reduced set of more independent tests, we chose only one performance measure if a pair of measures had a correlation coefficient greater than 0.85. This left four measures, consisting of the native rank, rank of best decoy, Z-score, and enrichment measures. The scores were then normalized so their mean would be 1.0, and are reported in Table 5.

There is a large spread in the scores for the various potential functions, ranging from 0.36 for ΔE^{solv} to 1.79 for the Coulombic term of CHARMM 19. Of the knowledge-based potentials, PAT/ME scored second only to DFIRE. We consider this result particularly significant, given the course-grain nature of the PAT/ME method, and the fact that the method is partially limited by the number

Table 5. Comparison of reduced metric data

	PAT/ME (4.8_3)	United-atom vdW (AMBER)	Coulombic (AMBER)	United-atom vdW+coulombic (AMBER)	United-atom vdW (CHARMM19)	Coulombic (CHARMM19)	ΔE (Delarue & Koehl)	ΔG^{env} (Koehl & Delarue)	ΔE^{solv} (Delarue & Koehl)	RAPDF	DFIRE
Native rank	1.43	0.99	0.50	0.98	1.15	1.56	1.16	0.86	0.27	1.01	1.09
Rank of best decoy	0.89	0.43	0.96	0.53	1.23	1.26	0.86	1.42	0.63	0.96	1.82
Z-score	2.37	0.00	0.48	0.19	1.30	3.19	0.58	0.58	0.10	0.29	1.93
Enrichment	0.82	0.53	0.82	0.53	1.35	1.14	0.76	2.19	0.44	1.20	1.23
Mean	1.38	0.49	0.69	0.56	1.26	1.79	0.84	1.26	0.36	0.87	1.52

The relative performance over the entire dataset of decoys is compared for all independent comparison metrics. The scores have been normalized so that the mean over potentials for each metric is 1.0.

of counts, a limitation that should become less severe in coming years as the size of the crystallographic database is increased.

Discussion

The potential function discussed herein was derived in an effort to capture information relating to multi-body atomic interactions in static protein structures that would be impossible with a pairwise potential. There is evidence to suggest that, even at the residue level, higher order interactions play a crucial role in protein folding[34,35] and function.[13,14] We believe that this should be the case to an even greater extent for an atomic potential and have decided to explore an atomic environment potential for the following reasons.

First, in a residue-based potential, each residue is represented as a single point, effectively reducing the topological structure of the protein to that of a linear chain of beads. This simple connectivity can have the effect of adding errors to any potential that is derived from a database of known structures (see Thomas & Dill[36] for a detailed discussion of this phenomenon). No such simplification can be made at an atomic level, and the atomic interactions one sees in a folded protein are highly dependent on a much more complex topological landscape than one sees with the beads on a string model. This effect is illustrated in Figure 7, and is of particular importance when one is estimating atomic energy parameters from a database of folded protein structures. By looking at clusters explicitly, we try to capture as much information as possible about the nature of the atomic environment while, hopefully, minimizing the impact that this effect might have on the validity of our data.

Second, the major driving force in protein folding, the hydrophobic effect, is by its very nature an aggregate phenomenon of the hydrophobic protein atoms, the hydrophilic atoms, and the solvent.[37–40] The best-known measure of this effect is the empirical relationship between solvent accessible surface area (SASE) and the octanol-water transfer energy.[39] The SASE is an inherently multi-atom metric and cannot be generated without a full atomic representation of the protein. Since both the number of neighbors and the identity of those neighbors have been folded into a single function in the PAT/ME functional form, it contains information that is difficult to capture with a pairwise potential. Specifically, information about the burial propensity of a given atom type and the preferred atomic environment for a given atom type upon burial are difficult to capture with any pairwise atomic treatment. It is exactly this shortcoming of the pairwise energy functions that necessitates the calculation of the SASE. SASE calculations are computationally intensive relative to simple pairwise atomic distance calculations. The PAT/ME potential captures much of the information gained by SASE while retaining the

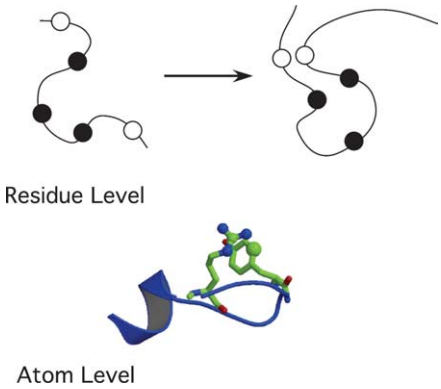


Figure 7. Topological effects of counting at different levels of complexity. At the residue level, bringing together two residues that are mutually attractive (in white) can create the erroneous appearance that other residues nearby in sequence are also mutually attractive (white–black interactions) when gleaning contact information from a database. This problem is exacerbated on the atomic level because of the more complex bonding topology that exists at the atomic level. When looking at clusters, however, no assumption is made about the individual interactions *per se*, only that a given cluster is more (or less) likely to occur than might be expected *via* random chance. This Figure was generated using the programs MOLSCRIPT[59] and Raster3D.[60]

relatively minor computational cost of a pairwise distance calculation. In recent work, Zhou & Zhou[41] have shown that burial propensity at the residue level can be captured by the atomic pairwise potential DFIRE. DFIRE (as well as RAPDF) has been formulated such that every proteinaceous atom type is unique, in contrast to the environment potential discussed here. It may be that the more fine-grained nature of these particular pair potentials, coupled with the residue-specific nature of the atom types, adds back information into the pairwise potential that might otherwise be lost when atom types are grouped according to their intrinsic chemical properties.

It is interesting to note that the variants of the PAT/ME potential with critical distances/skipping number combinations showing the greatest difference from the statistical independence assumption were the same variants that scored well in the jackknife procedure for decoy discrimination. This is likely a result of maximizing the number of atoms in the training dataset falling within the mid-range of the coordination number histogram. It is in the mid-range region of coordination number in which the most information can be captured: too many atoms in the high range limits the data coverage (due to the large number of possible ME compositions), and too many atoms in the low range prohibits the discrimination of unique atomic environments (due to the very small number of possible ME compositions).

We have undertaken the derivation of this potential with the hope of ultimately using it for both protein structure prediction and computational *de novo* protein design applications. Buried polar interactions are inherently difficult to design, because of the multibody nature of their environments. Figure 5 shows examples of interactions that were far more common than we would expect if each interaction were independent, leading to the conclusion that, especially in the case of buried polar interactions, an atomic environment potential might be able to capture information about the energetics of such interactions that a pairwise potential may not, by its very nature, be able to provide. One potential use of this type of information is as an aid in the design of buried metal-binding sites without the need to explicitly simulate the metal ligand. A simple van der Waals or Lennard-Jones potential with an electrostatic term, often used in *de novo* protein design for scoring the residues on the protein interior, is of limited utility when dealing with buried polar residues, although a strict stipulation that all buried hydrogen bonds be satisfied has recently met with success.[42]

As can be seen from the results, while the variant of the PAT/ME potential presented here is not the top performer for every decoy set and for every metric tested herein, it remains a powerful tool for protein structure comparison, and should prove to be a valuable addition to the arsenal of existing tools for this purpose. It performed on-par in our tests with many pairwise non-bonded atomic potential treatments, and has been successful at recognizing the native structure amongst challenging decoys at least as often as the more common non-bonded potentials such as the AMBER[25] non-bonded potential and pairwise potentials of mean force.

One caveat, however, with using the PAT/ME construction is that there is nothing currently built into this potential rendering it capable of preventing steric overlaps. Implicitly, overlaps should be discouraged, since they result in micro-environments that are too dense, but, in practice, this construction should always be used in conjunction with another term capable of preventing atomic clashes.

A final caveat of this approach is the incompatibility of any non-pairwise function with the DEE[5,43–47] algorithm, often used in homology modeling and *de novo* protein design applications. However, while the PAT/ME construction may not be used in initial pruning steps in which a DEE is employed, that does not prevent its subsequent use in a stochastic search step using Monte-Carlo or genetic algorithm based methods, especially when buried polar residues are desired.

Future directions

One potential problem with any energy derivation method that relies on a database of experimental protein structures is that the quality of the statistics that may be derived from atomic positions is only as good as the data on which it is

based. We have tried to filter low-resolution X-ray structures from our database, but it remains possible that errors occur in one or more of the structures used for this analysis. Future work on this potential will certainly use even larger datasets to improve sampling and reduce the effects of small errors in the raw data. Furthermore, increasing the dataset as new X-ray crystal structures of proteins become known is likely to improve the accuracy of our PAT/ME potential, because its accuracy is currently limited to a great extent by the number of counts in many of the microenvironments. Thus, we are now at the lower limit of adequate sampling using an atomic environment approach. By contrast, sufficient data are currently available for adequate sampling of most aspects of pairwise atomic interactions. Thus, we expect that the accuracy and use of atomic environment potentials will now expand more rapidly than can be expected for pairwise potentials.

It must also be remembered that, while the current PAT/ME construction is based on 20 total atom types, some of the potentials to which we compare its performance in this work contain as many as 167 different, residue-specific atom types. A future variant of the PAT/ME potential containing a larger number of atom types is expected to improve its performance.

In an effort to improve the accuracy of pairwise-residue and pairwise-atomic functions, much research has lately been devoted to optimization techniques. In these approaches, rather than deriving the energetic parameters from a set of known structures, the parameters are evolved by the computer with the stipulation that the native structure must always have the lowest energy among a large set of decoys. In this manner, the parameters are tuned such that the potential energy of the native state will be lower than any other, non-native state[48–51]. While our present method of derivation fulfils this criterion for most cases tested thus far, it may be possible to improve the parameters by using such a machine-learning approach. From an initial guess of the energetics of a given cluster (based on our statistics) the energies can be "tuned" in an iterative fashion in order to improve prediction accuracy.

Methods

The structure database

The database of proteins used for this analysis is a subset of the PISCES list[52] using a 35% sequence identity threshold, filtered as follows:

(1) No NMR structures were included.
(2) Crystallographic structures with resolutions beyond 2.0 Å were discarded.
(3) Structures with a significant number of missing residues were discarded.
(4) Structures consisting of only a C^{α} trace were discarded.
(5) Structures with less than 20 residues were discarded.
(6) Structures with large cofactors (such as a protein bound to DNA or heme) were discarded.

The final size of the database was 1066 chains (see Supplementary Data, Table 51), consisting of 1,918,682 total atoms. In cases where the asymmetric unit listed in the PDB file did not represent the biologically relevant oligomer, the full biomolecule representation was downloaded from the EBI macromolecular structures database.[53] The computer program "reduce"[54] was used on all PDB files to correct glutamine and asparagine side-chain amide geometries prior to collection of any statistics.

The potential function that we have derived is formally a united-atom potential, so hydrogen atoms in the database are not considered in this work.

The decoy database

The decoy folds tested in the jackknife procedure were downloaded from the Decoys-'R'-Us website†.[23] The decoy sets tested in this analysis were the "4 state reduced" sets,[22] the "lmds" sets,[55] the "fisa" sets,[31] the "globins" sets,[23] the "lattice_ssfit" sets.[56] A total of 132 decoy sets were scored from this database. For the comprehensive comparison of potentials, the Loops decoy set[32] from this website and the Rosetta all-atom[57] decoy sets (obtained from David Baker's website‡) were also used.

In each case, the native structure file was pared to correspond exactly to the number of residues represented in the decoy structures. If the native structure was solved by NMR, the first model in the file was chosen to represent the native. In cases were a small number of amino acid side-chain atoms were not present in the native structure, those side-chains were modeled in using InsightII and subjected to energy minimization using the CVFF forcefield to ensure reasonable coordinates. If the native structure and the decoys could not be made to correspond exactly (e.g. same number of residues, heavy atoms, etc.) then the set was discarded.

Energy calculations

RAPDF[29] scores were calculated with the program "potential_rapdf" from the RAMP suite of programs§ using the potential "astral_159_e4_allatoms_xray_scores." DFIRE scores were calculated with the program "dfire" obtained from the author.[30] The CHARMM19[24] van der Waals and coulombic terms, and the ΔE and ΔE^{solv} of Delarue & Koehl,[18] the ΔG^{env} of Koehl & Delarue[17] were calculated using software obtained from Patrice Koehl. The PAT/ME and AMBER all-atom scores were calculated using an in-house C++ program (ProtCAD) on a cluster of dual processor Itanium computers running the Linux operating system.

Acknowledgements

The authors thank Patrice Koehl for helpful discussion, use of computer code and resources,

† http://dd.stanford.edu
‡ http://depts.washington.edu/bakerpg/
§ http://www.ram.org

and critical reading of the manuscript. We thank Ram Samudrala for making the Decoys-R-Us database available for general use, as well as for general help and the use of his software. We thank Yaoqi Zhou for the use of his software. We also thank Kim Sharp, Roland Dunbrack, Jeff Saven and Mitchell Lewis for insightful discussions and critical analysis. C.M.S. acknowledges the support of the National Science Foundation Postdoctoral program in Biological Informatics; W.F.D. and M.L. thank and acknowledge grant support from the NIH (GM63718 to M.L. and GM54616 to W.F.D.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.07.054

References

1. DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F. & Lombardi, A. (1999). *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* **68**, 779–819.
2. Bolon, D. N., Voigt, C. A. & Mayo, S. L. (2002). *De novo* design of biocatalysts. *Curr. Opin. Chem. Biol.* **6**, 125–129.
3. Saven, J. G. (2002). Combinatorial protein design. *Curr. Opin. Struct. Biol.* **12**, 453–458.
4. Kortemme, T. & Baker, D. (2004). Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.* **8**, 91–97.
5. Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comp. Chem.* **19**, 1505–1514.
6. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* **66**, 549–579.
7. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396.
8. Bonneau, R. & Baker, D. (2001). *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189.
9. Laziridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139–145.
10. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
11. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
12. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
13. Karlin, S. & Zhu, Z. Y. (1996). Clusters of charged residues in protein 3-dimensional structures. *Proc. Natl Acad. Sci. USA*, **93**, 8350–8355.
14. Karlin, S. & Zhu, Z. Y. (1996). Characterizations of diverse residue clusters in protein 3-dimensional structures. *Proc. Natl Acad. Sci. USA*, **93**, 8344–8349.
15. Vriend, G. & Sander, C. (1993). Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallog.* **26**, 47–60.
16. Colovos, C. & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519.
17. Koehl, P. & Delaurue, M. (1994). Polar and non-polar atomic environment in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264–278.
18. Delarue, M. & Koehl, P. (1995). Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J. Mol. Biol.* **249**, 675–690.
19. Bagley, S. C. & Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci.* **4**, 622–635.
20. Wei, L., Altman, R. B. & Chang, J. T. (1997). Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. *Pacific Symp. Biocomput.* **2**, 465–476.
21. Karlin, S., Zhu, Z. Y. & Baud, F. (1999). Atom density in protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 12500–12505.
22. Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392.
23. Samudrala, R. & Levitt, M. (2000). Decoys 'R' Us: a database of incorrect protein conformations for evaluating scoring functions. *Protein Sci.* **9**, 1399–1401.
24. Brooks, B. R., Bruccoleri, B. D., Olafson, D. J., States, S., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.
25. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G. *et al*. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784.
26. Levitt, M., Hirschberg, M., Sharon, R. & Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Commun.* **91**, 215–231.
27. Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.
28. Melo, F. & Feytmans, E. (1997). Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **267**, 207–222.
29. Samudrala, R. & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895–916.
30. Zhou, H. & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726.
31. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
32. Samudrala, R. & Moult, J. (1998). A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol.* **279**, 287–302.
33. Wang, K., Fain, B., Levitt, M. & Samudrala, R. (2004).

Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.* **8**.
34. Munson, P. J. & Singh, R. K. (1997). Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein Sci.* **6**, 1467–1481.
35. Kannan, N. & Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **292**, 441–464.
36. Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457–469.
37. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–64.
38. Sippl, M. J. & Casari, G. (1992). Structure-derived hydrophobic potential. *J. Mol. Biol.* **224**, 725–732.
39. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
40. Huang, E. S., Subbiah, S., Tsai, J. & Levitt, M. (1996). Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol.* **257**, 716–725.
41. Zhou, H. & Zhou, Y. (2004). Quantifying the effect of burial of amino acid residues on protein stability. *Protein SFG*, **54**, 315–322.
42. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Compuational design of a biologically active enzyme. *Science*, **304**, 1967–1971.
43. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**, 539–542.
44. Gordon, D. B. & Mayo, S. L. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Struct. Fold. Des.* **7**, 1089–1098.
45. Lasters, I., De Maeyer, M. & Desmet, J. (1995). Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng.* **8**, 815–822.
46. Lasters, I., Desmet, J. & DeMaeyer, M. (1997). Dead-end based modeling tools to explore the sequence space that is compatible with a given scaffold. *J. Protein Chem.* **16**, 449–452.
47. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429–445.
48. Crippen, G. M. (1996). Easily searched protein folding potentials. *J. Mol. Biol.* **260**, 467–475.
49. Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.
50. Maiorov, V. N. & Crippen, G. M. (1994). Learning about protein folding *via* potential functions. *Proteins: Struct. Funct. Genet.* **20**, 167–173.
51. Mirny, L. A. & Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264**, 1164–1179.
52. Wang, G. & Dunbrack, R. L., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
53. Henrick, K. & Thornton, J. M. (1998). PQS: A protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
54. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747.
55. Keasar, C. & Levitt, M. (2003). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* **329**, 159–174.
56. Xia, Y., Huang, E. S., Levitt, M. & Samudrala, R. (2000). *Ab initio* construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300**, 171–185.
57. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **37**, 171–176.
58. Delano, W. L. (2002). *The PyMOL Molecular Graphics System*. DeLano Scientific LLC, San Carlos, CA, USA.
59. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
60. Merrit, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524.

Edited by B. Honig

(Received 27 October 2004; received in revised form 20 June 2005; accepted 20 July 2005)
Available online 3 August 2005