A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection

Hui Lu and Jeffrey Skolnick*
Laboratory of Computational Genomics, Donald Danforth Plant Science Center, St. Louis, Missouri

ABSTRACT **A heavy atom distance-dependent knowledge-based pairwise potential has been developed. This statistical potential is first evaluated and optimized with the native structure *z*-scores from gapless threading. The potential is then used to recognize the native and near-native structures from both published decoy test sets, as well as decoys obtained from our group's protein structure prediction program. In the gapless threading test, there is an average *z*-score improvement of 4 units in the optimized atomic potential over the residue-based quasichemical potential. Examination of the *z*-scores for individual pairwise distance shells indicates that the specificity for the native protein structure is greatest at pairwise distances of 3.5–6.5 Å, i.e., in the first solvation shell. On applying the current atomic potential to test sets obtained from the web, composed of native protein and decoy structures, the current generation of the potential performs better than residue-based potentials as well as the other published atomic potentials in the task of selecting native and near-native structures. This newly developed potential is also applied to structures of varying quality generated by our group's protein structure prediction program. The current atomic potential tends to pick lower RMSD structures than do residue-based contact potentials. In particular, this atomic pairwise interaction potential has better selectivity especially for near-native structures. As such, it can be used to select near-native folds generated by structure prediction algorithms as well as for protein structure refinement. Proteins 2001;44:223–232.** © **2001 Wiley-Liss, Inc.**

Key words: potentials; decoy set; gapless threading *z*-scores

INTRODUCTON

Current prediction approaches to protein structure are based on the thermodynamic hypothesis that the native structure is at the lowest free energy state that lies in the lowest potential energy basin under physiological conditions. A potential that can discriminate between the native and misfolded structures is crucial for any protein structure prediction protocol to be fully successful. Toward this end, two different types of potential energy functions are currently in use.[1–6] The first class of potentials, the so-called physical-based potential, is based on the fundamental analysis of forces between atoms.[7,8] The second class, the so-called knowledge-based potentials, extracts parameters from experimentally solved protein structures.[9–14] The advantage of the first class of potentials is that, in principle, they can be derived from the laws of physics; the disadvantage is that the calculation of free energy is very difficult because this computation should include an atomic description of the protein and the surrounding solvent. Currently this type of computation is generally too expensive for protein folding.[15] On the other hand, with today's computer resources, knowledge-based potentials can be quite successful at fold recognition[16–19] and *ab initio* structure prediction[20,21] (see also *Proteins* special issue, 2000, on CASP3).

For knowledge-based potentials, quite often the distribution of pairwise distances is used to extract a set of effective potentials between residues or atoms. In most cases, the knowledge-based potential is built and then used on reduced protein models, i.e., using one or two points for each residue to represent a protein.[22–24] These points are usually located at the coordinates of each residue's C_α atoms, C_β atoms or at coordinates of the center of mass of each side chain. Simplified models are used to represent proteins because the computational cost of fold recognition or *ab initio* predictions using the all-atom representations of proteins would be prohibitive. Nevertheless, several potentials for higher-resolution models had been developed in the hope of providing better discriminatory power than are obtained with residue-based potentials.[25–28] For ranking structures near the native fold, and for protein structure refinement, the detailed interactions between side-chain atoms from different residues may be required to rank correctly the low root-mean-square deviation (RMSD) structures.[29]

One question that should be answered for every newly developed potential is how much of an improvement it is over previous potentials. For residue-based potentials, two kinds of measure for a potential are typically used: *z*-scores from gapless threading, and the ability to discriminate native structures from decoys. For the atomic resolution potentials developed previously,[25–28] the discrimination ability of native structure from decoys had been used for

Grant sponsor: National Institutes of Health; Grant number: GM RR-12255.

*Correspondence to: Jeffrey Skolnick, Donald Danforth Plant Science Center, 893 N. Warson Rd., St. Louis, MO 63141.
E-mail: skolnick@danforthcenter.org

Received 15 December 2000; Accepted 4 April 2001

evaluation. To the best of our knowledge, atomic potentials have not been tested using *z*-scores in gapless threading. This is probably due to the amount of computer time needed to reconstruct the atomic details of a large number of protein structures for a gapless threading test. Typically one needs to generate in the order of 10,000–100,000 structures for each target sequence. Also in question is the quality of the all-atom rebuilding programs required for such a calculation. On the other hand, the behavior in gapless threading is a quick indication of the quality of the potentials. The potentials that have better *z*-scores in gapless threading tests usually have better discriminative ability in *ab initio* folding (J. Skolnick, unpublished observations). A good *z*-score is a necessary, but not sufficient, condition for the potential to be useful for structure prediction.

Gapless threading consists of the evaluation of hundreds of thousands of structures. This is statistically more reliable than picking 10 or 20 native structures from a similar number of decoys. A recently developed atomic reconstruction program[30] provides a fast and reliable protocol for the all-atom rebuilding process. Thus, it is possible for us to evaluate our atomic potential by comparing it with other potentials. We then evaluate our potential on decoy sets taken from the PROSTAR website,[27] which has been tested against several residue-based and atom-based potentials. We also compare the performance of all atomic knowledge-based potentials and residue-contact potentials developed in our group,[13] as well as to an atomic potential (RAPDF) and a residue contact potential (CDF) posted on the PROSTAR website.[27] Another decoy set[31] is used to test the current potential's ability to pick native and near-native structures from decoys that have native secondary structures and compact self-avoiding conformations.

Furthermore, we test the current potential on decoys generated by generalized comparative modeling, that is, threading followed by a Monte Carlo protein structure prediction and refinement program,[32] as applied to the sequences in the Fischer database.[33] The goal is to pick the lowest possible RMSD structures from thousands of generated decoy structures. Since this test is against one of the best protein structure prediction programs currently in use, the results will allow us to evaluate the practical value of this atomic potential.

The organization of this article is as follows. First we provide the details of constructing the atomic potential. Then we present the comparison of the present potentials with several published potentials (both at residue-based and atom-based levels) on gapless threading tests and on decoy sets from the web and from our structure prediction program. Finally, we examine the significance of this work and discuss the future development and use of the atomic potential.

METHODS

Potential Construction

The potential is developed for an all-heavy-atom representation in which the heavy atoms in the protein are classified on a residue-specific and on an intra-residue-position-specific basis; that is, the C_α of LEU is different from the C_β of LEU and is also different from the C_α of ILE. Hydrogen atoms are ignored. In total, there are 167 different atom types in the current model. The distances between any two atoms are divided into 14 distance shells, starting from 1.5 Å, with each distance shell 1 Å in width. For example, distance shell number 3 covers the distance from 3.5 Å to 4.5 Å. The last distance shell, the fourteenth shell, covers all distances from 14.5 Å to infinity. The potential for any pair of heavy atoms *i* and *j* is calculated using the formula:

$$\varepsilon(i, j, d) = -RT \ln \left[\frac{N(i, j, d)_{\text{obs}}}{N(i, j, d)_{\text{exp}}}\right] \quad (1)$$

where d is the distance shell number. $N(i,j,d)_{\text{obs}}$ and $N(i,j,d)_{\text{exp}}$ are the observed number of contacts and the expected number of contacts in distance shell d, respectively. In the quasichemical approach, $N(i,j,d)_{\text{exp}}$ is defined as:

$$N(i, j, d)_{\text{exp}} = N(d)\chi_i\chi_j \quad (2)$$

where χ_k is the mole fraction of atom type k, and $N(d)$ is total observed number of pair contacts in distance shell d.

The essential question in the derivation of a knowledge-based potential is the choice of reference state, that is, how to calculate $N(i,j,d)_{\text{exp}}$ from a number of structures. As described by Skolnick et al.,[13] there are three ways to define the reference state: a composition-independent scale, a partial-composition-corrected scale, and a composition-corrected scale. The difference between these three scales is as follows: in the composition-independent scale, the observed distance-dependent contact numbers from different proteins are pooled together and the expected contact numbers are calculated from the overall mole fractions of the individual types of atoms in the entire database. In the partial-composition-corrected scale, the observed contact numbers are pooled together, but the expected contact numbers are the sum of expected contact numbers (calculated according to eq. 2) for each individual protein. In the composition-corrected scale the potential is calculated by averaging the pair potential eq. 1 for the proteins that have both atom type *i* and type *j* in their structures. In the current atomic potential calculation, all three scales have been calculated and tested using gapless threading.

The atomic potential calculation used in eqs. 1 and 2 is somewhat different from a previous atomic potential developed in our group.[25] Previously, the expected number of contacts in a certain distance shell was proportional to the number of contacts a certain type of atom made to all other atoms in that shell. In that way, the expected number of contacts for residues on the protein surface would be less than that of residues located inside the protein because the atoms on surface residues have a smaller total number of contacts. The current schemes in which the expected number of contacts is proportional to the mole fractions of the corresponding atoms reflects a better reference state, a random set of compact protein-like environments, for the

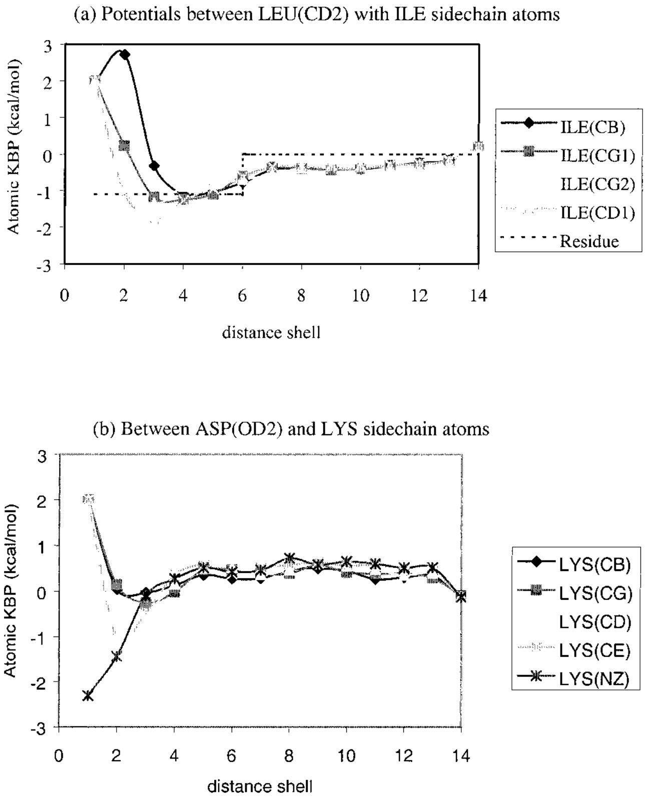


Fig. 1. Distance-dependent atomic potentials (**a**) between hydrophobic residues LEU and ILE, and (**b**) between hydrophilic residues LYS and ASP. Each distance shell is 1 Å wide. Distance shell number *l* corresponds to distances between *l*.5 Å to (*l*+1)0.5 Å.

potential. In the cases where certain atom pairs are not observed at short distances, the potentials in these distance shells are set to 2.

Figure 1 shows the pairwise potentials between the CD2 atom in LEU and different side-chain atoms in ILE, and between the OD1 atom in ASP and different side-chain atoms in LYS, using the composition-independent scale. The potentials between LEU and ILE atoms indicate attractive interactions between hydrophobic residues. These potentials have well-defined wells at short distances and maintain negative values at longer distances. For comparison, the residue-based contact potential described by Skolnick et al.[13] is also drawn in Figure 1(a), which has a value of -1.1 and a cutoff distance of 6.8 Å. The residue-based potential qualitatively agrees with the atomic potential. For interactions between atoms in the hydrophilic residues ASP and LYS, the potentials have negative wells at short distances, <3.5 Å, and then turn positive at longer distances. Hydrophilic residue pairs at distances longer than 4.5 Å, even for opposite-charged ones, are repulsive. One interesting observation is that when comparing potentials $\varepsilon(i,k,d)$ and $\varepsilon(j,k,d)$, where i and j are side-chain atoms from the same residue and k is a side-chain atom from a different residue, the differences in the potentials reside in the short distance shells (d <5). For distances longer than 5.5–6.5 Å, these potentials are almost the same.

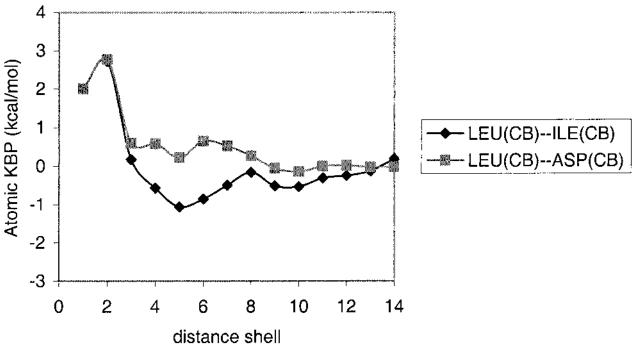


Fig. 2. Atomic potential between C_β atoms of LEU and of ILU, which is attractive, and the atomic potential between C_β atoms of LEU and of ASP, which is repulsive. Distance shell number *l* corresponds to distances from *l*.5 Å to (*l*+1)0.5 Å.

Because the residue-specific atomic model used in this case is the most detailed representation when one ignores the location in the sequence of the residues, the current potential can be used directly to check whether other quasichemical potentials based on reduced models by grouping certain atoms are reasonable. Grouping atoms that belong to the same residue is fine, as seen from Figure 1. The side-chain atoms from the same residue, when paired with a common third atom, have similar potentials. For example, when pairing different ILE side-chain atoms with the LEU(CD2) atom, these pairwise potentials have similar shape; thus, we can group the ILE side-chain atoms together for the construction of potentials for the reduced model. Of course, when using these kinds of potentials from reduced models, some atom-specific details and side-chain orientation information would be lost. When grouping atoms from different residues one needs to be more careful. For example, grouping GLU(OZ1) and ASP(OZ1) is possible because these two residues have similar characteristics, and the OZ1 occupies a similar position in these residues. The potential $\varepsilon(\text{GLU(OZ1)},k,d)$ and $\varepsilon(\text{ASP(OZ1)},k,d)$ have similar shapes, where k is any type of atom and d is the distance shell. But grouping the C_β of ILE and the C_β of ASP may not work very well. The potentials between ILE(C_β) and LEU(C_β) and between the ASP(C_β) and the LEU(C_β) are qualitatively different (Fig. 2) and reflect the difference in connectivity of the C_β in different residues. The potential between ILE(C_β) and LEU(C_β) has a well at 6 Å and is attractive from 3 to 14 Å. But the potential between ASP(C_β) and LEU(C_β) is repulsive at all the distances. As a result, the atoms ASP(C_β) and ILE(C_β) should not be grouped together when one tries to reduce the number of atom types in his or her protein models.

Database Selection

A set of 1,291 protein structures were randomly selected from the nonhomologous Protein Data Bank (PDB) select library[34] in order to generate the atomic potential. The sequence identity between any pair of sequences in the database is less than 25%. The gapless threading target set (16 sequences) and the template set (532 structures) were randomly selected from the same PDB library. The template set had only about one-half of the structures overlap with the subset used for generating the potential. The 16 sequences in the target set are not included in the 1,291 structures used for potential construction. The sequence sets are listed on our web page (http://bioinformatics.danforthcenter.org/).

Atomic Detail in Protein-Structure Rebuilding

In gapless threading and in ranking structure prediction results from the Fischer database sequences, the all-atom model of the protein structure has to be built from the reduced model. The building process uses the program described by Feig et al.[30] The program starts from a SICHO representation describing the protein by the center of mass of the C_α plus side-chain heavy atoms[35] of the protein and rebuilds backbone and side-chain heavy atoms consecutively. The accuracy of the rebuilt model is 1 Å RMSD for all atoms and 0.6 Å RMSD for C_α atoms.[30]

RESULTS

Evaluation of the Atomic Potential With Gapless Threading

Using gapless threading to evaluate the current atomic potential is done as follows. The target sequences are threaded through the side-chain centers of mass of the template structures; then using an all-atom rebuilding program[30] to reconstruct an all-atom representation of the decoy structures. Typically 10,000–100,000 decoy structures are rebuilt through this process for each of the target sequences. The object is to preserve the side-chain contact map. For native structures, the rebuilding program has been shown to be able to reconstruct the backbone to an RMSD of <0.6 Å for the C_α atoms compared with the native structures. The program typically takes 20 s to reconstruct a protein with about 100 residues on an SGI O2 machine.

As presented in Figure 3, the gapless threading results are compared between residue-based potentials and atom-based potentials. The atomic potentials with the composition-independent scale performed better than residue potentials in all 16 cases. On average, the atomic potential has a z-score of 4.1 units better than residue-based potentials, where the z-score of the native structure is defined as:

$$z = \frac{E_{\text{average}} - E_{\text{native}}}{\sigma} \tag{3}$$

where E_{average} and σ are the mean and standard deviation of the energies of all the structures generated by gapless threading.

In order to check the distance-related performance in the gapless threading, we have calculated the z-score of every distance shell. In 15 of the 16 cases, the shape of the z-scores is similar to that in Figure 4(a,b). The z-scores are low, lower than the z-score from the residue-based contact potential at short distances (first two distance shells) and at long distances (distance shell number 7 and up). The most significant z-scores come from distance shells between 3.5 and 6.5 Å, which is the first solvation shell for each atom. When we combined the energies from 3.5 to 6.5 Å, we obtained the best z-scores. As such, we will use atomic energy in this distance range in future structure predictions. In one case, the shape of the z-score-distance bin is different [Fig. 4(c)]. The z-score is very high at short distances, then decreases gradually. Nevertheless, the z-score of the combined energy from 3.5 to 6.5 Å is still better than that of residue-based potentials. So it is safe to say that atomic potentials using energies from 3.5 to 6.5 Å bins have the best selectivity in most cases and will likely outperform residue-based potentials.

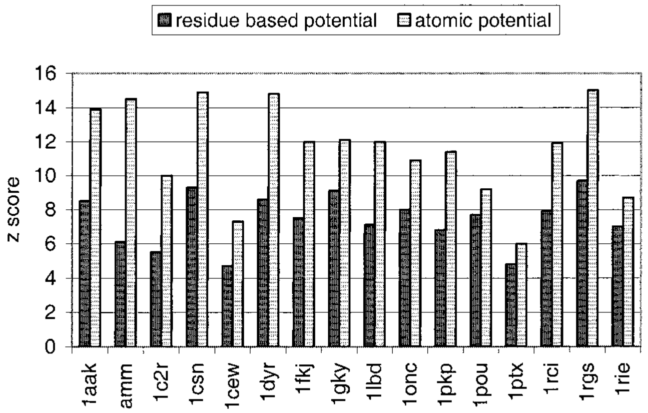
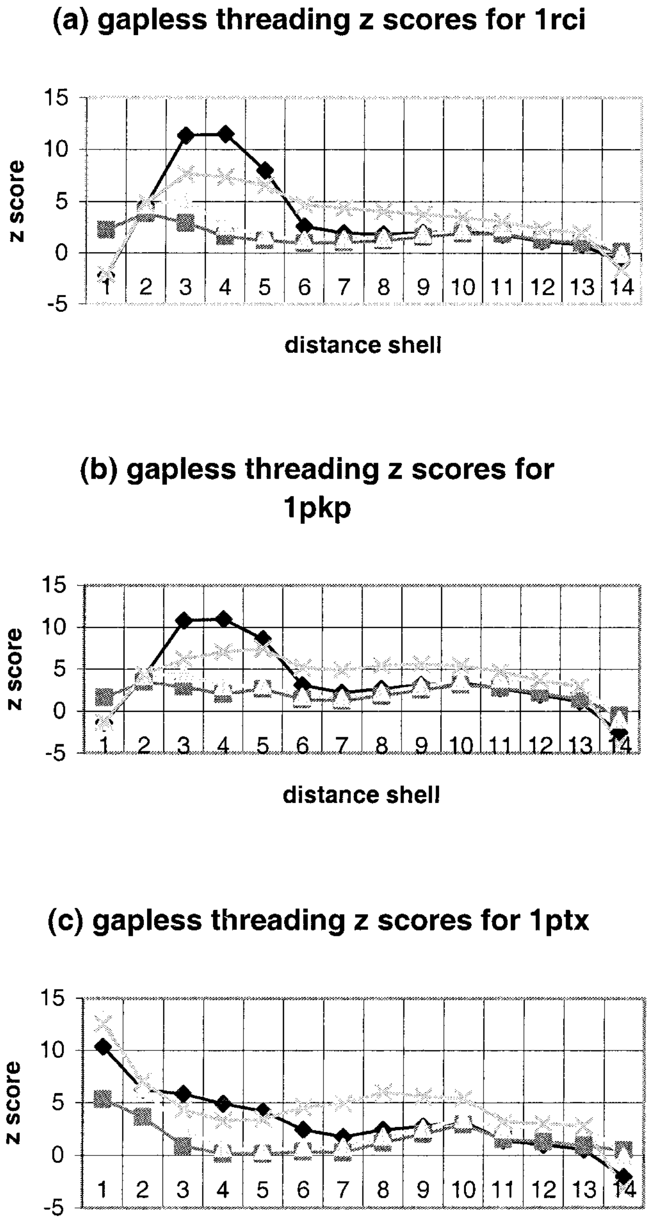


Fig. 3. Comparison of z-scores of gapless threading for atomic knowledge-based potential (striped) and residue contact potentials (solid) with a composition-corrected scale. In every case, the atomic knowledge-based potential has a better z-score than the residue-based potential. The average z-score for the atomic knowledge-based potential is 11.6. The average z-score for the residue contact potential with a composition-corrected scale quasi chemical potential is 7.5.

We have also evaluated the different reference states as proposed by Skolnick et al.[13] The potentials from the partially correct composition scale perform similarly to those from the composition-independent scale; i.e., the average z-score is ~12. But potentials from the composition-corrected scale perform much worse than do the above two scales, with an average z-score of only 4. This poor performance can be attributed to the way the potential of the composition-corrected scale is constructed, which is the weighted average of pairwise potentials from individual proteins. For 167 atomic types and 14 distance bins, individual proteins would not be able to provide enough statistics for the proper construction of the potentials. Thus, the overall averaging will not provide good results. This situation is somewhat different for residue potentials, where the composition-corrected scale can provide an average z-score that is 2 units better than the composition-independent scale.

In Figure 4, we also compared the performance of the atomic potentials computed by four different means: including or excluding the atoms from the first and second sequence of neighboring residues of each residue in the computation of $N(i,j,d)$ and $N(i,j,d)_{\text{exp}}$; and including or excluding the backbone atoms C, N, and O in the computation of the $N(i,j,d)$ and $N(i,j,d)_{\text{exp}}$. In the first solvation shell, 3.5–6.5 Å, the potentials calculated excluding the



(a) gapless threading z scores for 1rci

(b) gapless threading z scores for 1pkp

(c) gapless threading z scores for 1ptx

Fig. 4. *z*-scores from gapless threading for each distance shell. Results from four types of atomic knowledge-based potentials are plotted. ◆, including backbone atoms and excluding neighbor residues; X, excluding backbone atoms and excluding neighbor residues; △, including backbone atoms and including neighbor residues; shaded squares, excluding backbone atoms and including neighbor residues. See text for details.

TABLE I. Comparison of Different Potentials for the PROSTAR Decoy Sets

	Misfold	ifu	*Asilomar*	pdbrr and spga
Atomic KBP[a]	24/24[d]	32/43	37/41	5/5
RAPDF[b]	24/24	30/43	37/41	5/5
Residue contact potential[c]	24/24	22/43	35/41	4/5
CDF[b]	19/24	21/43	35/41	5/5

[a] The atomic KBP is the atomic potential developed in this article.
[b] RAPDF and CDF are atomic and residue-based potentials, respectively, from Samudrala and Moult.[27]
[c] This is a residue-based quasichemical potential from Skolnick et al.[13]
[d] The first number in each cell is the number of correctly identified decoys, and the second number is the total number of decoys. The first column lists the subsets of decoys.

TABLE II. Performance of Different Potentials When the Cutoff of the Discrimination Ratio Is 0.95*

	Misfold	ifu	*Asilomar*	pdbrr and spga
Atomic KBP	24/24[a]	32/43	36/41	5/5
RAPDF	24/24	29/43	35/41	4/5
Residue contact potential	24/24	20/43	31/41	4/5
CDF	17/24	10/43	15/41	4/5

*The discrimination ratio is defined as the ratio of energy of the decoy to the energy of the native structure.
[a] The first number in each cell is the number of correctly identified decoys, and the second number is the total number of decoys. The first column lists the subsets of decoys.

first and second neighbors performed much better than did those including the first and second neighbors. And the best potential is the one that includes the backbone atoms but excludes the neighbors from the protein sequence. For the rest of the distance shells, the four potentials have a similar performance, except for distances between 6.5 and 12.5 Å (the second and third solvation shells). The potential calculated excluding neighbors and excluding the backbone atoms C, N, and O performs a little bit better than the rest of the three potentials when including either or both of the neighbors or backbone C, N, and O in the potential constructions.

Performance on Decoy Set Structures

The atomic potential developed in this study has been tested on the decoy sets posted on the PROSTAR website, http://prostar.carb.nist.gov, and described in Samudrala and Moult.[27] The goal of this test is to discriminate between the native structure and one or more misfolded or low-resolution structures. The results are listed in Tables I and II. The atomic potential presented performs better than the residue-based potentials. When compared with the atomic potential RAPDF,[27] the current atomic potential has similar but better performance. Tables I and II also compare the contact potential CDF posted on the PROSTAR website with the composition-corrected quasichemical residue contact potential published previously (Skolnick et al.[13]).

In the *misfold* subset, both the atomic knowledge-based potential and atomic potential RAPDF select 100% of the cases correctly. Interestingly, the atomic potential RAPDF needs to use a long cutoff such as 15 Å or 20 Å to achieve 100% correct selectivity, and gets about 80% correct when using a cutoff of 5 or 10 Å.[27] By contrast, using the current atomic knowledge-based potential, a 100% correct selection can be achieved using any of the following cutoffs: from 3.5 to 6.5 Å, from 3.5 to 9.5 Å, and from 3.5 to 12.5 Å. As for the residue potentials, the residue contact potential with a composition-corrected scale from Skolnick et al.[13] picked the native structure 100% of the time, while the

PROSTAR contact potential picked the native structure 75% of the time. This subset had also been used to evaluate a previous pairwise atomic potential,[25] and it correctly picked 23 of the 24 native structures.

The current atomic knowledge-based potential has a much better differentiation ratio—defined in the PROSTAR website as the ratio of the energy of the decoy to that of the native structure—than the RAPDF potential.

In the test set *ifu*, our atomic knowledge-based potential failed to pick out 12 of the 42 native structures, while RAPDF missed 13. As in the first test set, the current atomic potential also has a better discrimination ratio. In this test set, neither of the residue-based potentials performed very well. This is probably because the targets in these subsets are protein pieces and it is difficult for residue potentials to evaluate structures based on a small number of pair interactions.

In the test set *Asilomar*, both the atomic knowledge-based potential and RAPDF missed 4 of 41 test cases. The residue potentials displayed similar performance to that of atomic potentials.

In the test sets *pdberr* and *sgpa*, where the decoys are low-resolution experimental structures, every potential correctly picked the high-resolution structures in all cases, except that the residue contact potential with a composition-corrected scale missed one case. In this case, the residue potential has a discrimination ratio of 1.01, while the contact potential from PROSTAR has a discrimination ratio of 0.99.

Note that, in some cases, the energies of the decoy structures and native structures are very close. That the native structure only has a slightly lower energy than that of the decoy may not be a reliable result. Thus, we used a somewhat stricter standard for claiming that the potential can distinguish between the native structure and the decoy. Instead of using the standard posted in PROSTAR, which shows that when the discrimination ratio is ≤1 for a correct pick, the qualifying line is moved to 0.95, to ensure that there is a recognizable energy gap between the native structure and the decoy. We should state that the discrimination ratio is not a good measure, but the better quantity, the *z*-score, is not meaningful for a small number of decoys in each sequence, where sometimes there is just one decoy structure. The results using the new standard are listed in Table II. For the atomic knowledge-based potential, the new standard does not affect many of the results. For those cases in which the native structure has lower energies, it shows that the energy gap between the native structure and the decoy is large. For RAPDF, in five cases the native and decoy energies are very close. The significant change of the results comes from evaluating the *Asilomar* subset using the contact potential CDF. CDF can only pick 15 out of the 42 native structures with large energy gaps.

Performance on Park and Levitt Decoy Set

The Park and Levitt decoy test set consists of 7 sequences, each with ~600–700 decoys that cover structures showing an RMSD ranging from 0 (native structure) to 10 Å from native structure. This test set was generated using a four-state off-lattice model together with a relaxation method.[31] The decoy structures all have the native secondary structures.

TABLE III. Evaluation of the Atomic Potential Using the Park and Levitt Decoy Set

PDB code	Native structure ranking	RMSD of lowest energy decoy (Å)	Correlation between energy and RMSD
1ctf	1	1.7	0.6
1r69	1	1.9	0.5
1sn3	1	2.0	0.5
2cro	1	2.8	0.7
3icb	1	1.9	0.8
4pti	1	1.4	0.5
4rxn	1	1.9	0.6

RMSD, root-mean-square deviation.

These decoys are evaluated using the current atomic potential. The results are listed in Table III. In all cases, the native structure has the lowest energy. In six of the seven cases, the decoy structures with the lowest energy are <2 Å away from native structures. In only one case, the lowest-energy decoy is 2.7 Å RMSD away from native structure, but the second lowest-energy decoy structure has an RMSD of 1.9 Å.

We have also calculated the correlation between the RMSD and the atomic potentials for these decoys. The results are also listed in Table III. In all cases, the correlations are >0.5. Figure 5 plots two of the energy vs. RMSD: 3icb and 4pti correspond to the highest and lowest correlation. The fact that there is a strong correlation between the RMSD and atomic energy in all cases suggests that structure refinement with the current atomic potential is possible if secondary structures are correctly predicted.

Performance on Structures Generated from GENECOMP

Our group has done extensive work on the refinement of predicted structures generated by threading.[35,36] As a test set, we have attempted to refine the set of sequences in the Fischer database[33] in standard threading benchmarks using program GENECOMP.[32] The goal of this test set is to pick low RMSD structures from the high RMSD structures. The structure constructed for each probe sequence from the Fischer database has been simulated with 100 trajectories, and each trajectory has 200 frames recorded. Using a residue-based potential energy,[35] which is actually the potential energy used to generate the trajectories, we picked the two lowest-energy structures from each trajectory. This gives us 200 decoy structures for each sequence. In 28 cases, the structures we selected include decoys that are <7 Å RMSD from the native structures. We will concentrate the evaluations of the atomic potential on these cases. This should be a difficult test, as the test structures are "smartly" misfolded, and the structures we picked from the trajectories are already the lowest or next lowest energies when evaluated with knowledge-based potentials on the residue level.

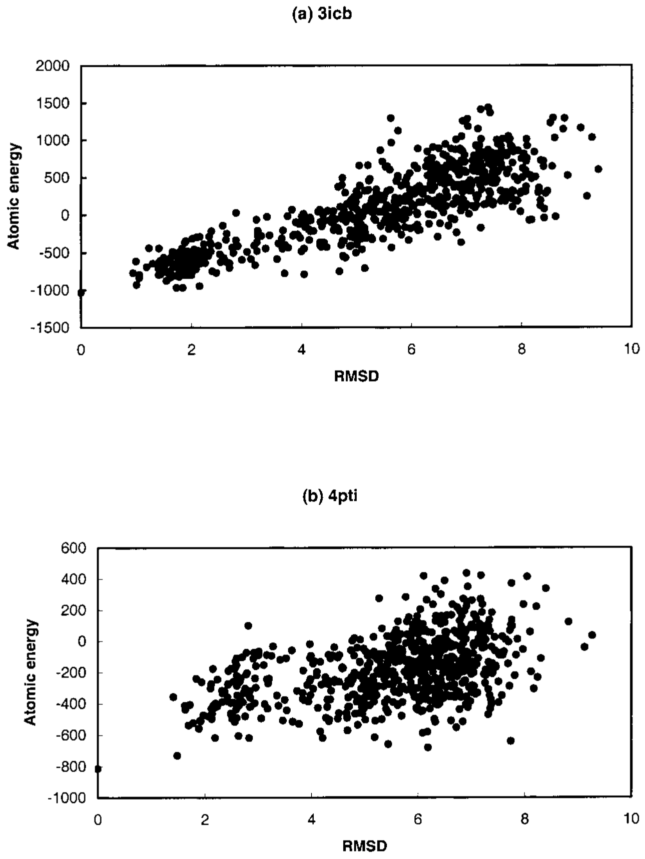


Fig. 5. Energy (in RT kcal/mol) vs. RMSD (in Å) for (**a**) 3icb and (**b**) 4pti for the decoy sets from Park and Levitt.[31]

Tables IV and V compare the atomic potential with two residue potentials: quasichemical and one including weak sequence similarity information. It has been shown that potential with weak sequence similarity information performed significantly better than normal quasichemical potential.[13] We performed two comparisons: (1) we compared the structures picked by the lowest energies from the atomic and residue potentials (Table IV), and (2) we compared the best RMSD structure from the top five structures selected by atomic and residue potentials (Table V). When compared with the quasichemical residue potential, the atomic potential selects lower RMSD structures in seven cases and higher RMSD structures in two cases when the lowest energy pick is used. The term "better" is defined as when the RMSD difference is larger than 10% of the RMSD value of the worse structure. A 0.5-Å difference would be significant for structures of ~2 Å RMSD, but it would not mean as much for structures of RMSD 7 Å away from the native ones. When the best structure in the top five selections is used, the atomic potential picks better structures in nine cases and worse in only one case. As presented by Skolnick et al.,[13] potentials, including weak sequence homologous information, perform better than do pure quasichemical potentials. When comparing the atomic potential with the residue contact potential, including weak sequence similarity information, the atomic potential did better for six cases but did worse for three cases of the lowest-energy structure and did better for five cases, while it did worse for three cases of the best structure in the top five selections.

We also checked the energy of native structures as compared with those of the decoys. In 25 of the 28 cases, the native structure has the lowest energy when compared with the 200 decoy structures selected in the previous section. In one case, the native structure has the second lowest energy and in another case, the native structure has the third lowest energy. In only one case, 1hrhA, does the native structure have the median energy of the decoys. This finding suggests that in most cases refinement near the native structure using the current atomic potential energy might be possible.

Performance on Structures from Unfolding Simulations

Unfolding simulations using molecular dynamics (MD) at high temperatures have been performed for proteins 1gb1, 2pcy, and 2zat. Starting from native structures, these high temperature MD simulations started from native structures and generated conformations of 0–4 Å RMSD from native ones in 100 picoseconds (ps). Figure 6 shows the atomic pairwise potential energy vs. the RMSD from these simulations. These plots show a good correlation of the pairwise atomic potential with the RMSD. In all three cases, the correlation coefficients between energy and RMSD are >0.6. Furthermore, in all three cases, the lowest-energy structure is only ~0.5 Å RMSD away from the native structure.

DISCUSSION

The present study shows that the atomic distance-dependent pair potential is consistently better than the residue-based pair potential in a wide range of test cases. The atomic potential is evaluated using gapless threading. The increment of the average z-score by 4 units quantitatively measures the improvement of the atomic potential over the residue-based potential with a composition-corrected scale. The higher specificity of the atomic potential energy for the native and near-native structures arises from amplification of the correct contact on the atomic level.

We have shown that the details of the potential construction are very important, particularly the ways of computing the N_{obs} and N_{exp}. Figure 4 demonstrates that if we included atoms from neighboring pairs in the potential construction and protein structure evaluation, the z-score would be even worse than if we used residue-based quasichemical potentials. The choice of reference states also has some impact on the performance of the potential. With the same quasichemical approach, these fine tunings of the potential are the main reason that the current potential outperformed those previously published.

The z-scores calculated from each distance shell (Fig. 4) imply that the native structure is distinguished from misfolded structures in the pairwise distances in the first solvation shell (3.5–6.5 Å). This conclusion is different from the one published earlier,[27] which shows that a longer cutoff, ≤20 Å is better.

TABLE IV. Comparison of Atomic Knowledge-Based Potential and Residue-Based Potentials for Sequences from the Fischer Database: Comparison of the Lowest-Energy Structures Picked by Different Potentials*

	Atomic knowledge-based potentials	Residue contact potential with correct composition scale	Residue-based potential with weak sequence similarity
Atomic knowledge-based potential	—	1hip_	1aba_ 1hip_ 1stfI
Residue contact potential with correct composition scale	1tlk_ 1onc_ 1rcb_ 1cauB 1stfI 2sas_ 1fc1A 1isuA[a]	—	1aba_ 1rcb_ 1stfI
Residue-based potential with weak sequence similarity	1hrhA 1rcb_ 1cauB 1stfI 2hpdA 2sas_	1hom_	—

*The atomic knowledge-based potential is the potential presented in the current study. The residue contact potentials with composition-corrected scales and with weak sequence similarity were presented by Skolnick et al.[13]

[a]The Protein Data Base (PDB) codes listed in each of the central cells indicate the cases in which the potential listed on the top of its column outperforms the potential listed on the left in its row. For example, the second column lists the cases for which the atomic knowledge-based potential outperforms the residue contact potentials. For example, in the third column and third row there are 8 cases in which the atomic potential is better than the residue contact potential.

TABLE V. Comparison of Atomic Knowledge-Based Potential and Residue-Based Potentials for Sequences from the Fischer Database: Comparison of the Best RMSD Structure from the Top Five Lowest-Energy Structures Picked by Different Potentials

	Atomic knowledge-based potential	Residue contact potential with correct composition scale	Residue-based potential with weak sequence similarity
Atomic knowledge-based potential	—	1aba_ 1hip_ 1cauB	1onc_ 1aba_ 3chy_ 1hip_
Residue contact potential with correct composition scale	1hom_ 1fc1A 1hrhA 3chy_ 1rcb_ 3hla_	—	1onc_ 1aba_ 3chy_ 1rcb_ 3hla_
Residue-based potential with weak sequence similarity	1hom_ 1fc1A 1isuA 1hrhA 1cgrA 1cauB 1c2rA 2sarA 2hpdA	1isuA 1cauB 1c2rA 2hpdA 2sarA	—

RMSD, root-mean-square deviation.

In Figure 1, one can see that for longer distances the different side-chain atoms tend to have similar potentials. And the potentials are typically <0.5 RT kcal/mol away from 0. This observation may explain why residue-based contact potentials work quite well as shown by the z-score in gapless threading. The residue-based contact potentials catch the most significant part of the pairwise interactions. We also showed that the composition-corrected scale cannot be used in atomic distance-dependent potential construction due to the lack of adequate statistics.

The tests on various decoy sets from the PROSTAR website demonstrated that the atomic knowledge-based potential performs consistently better than residue potentials and previous atomic potentials. Not only can the current atomic knowledge-based potential pick the native structures in more cases, but the discrimination ratio of the atomic knowledge-based potential is also better than that of the other potentials.

The Park and Levitt decoy set had been shown to be quite a challenge for a simple residue contact potential and van der Waals potential, where the lowest-energy structures typically were 6–10 Å RMSD away from native ones.[31] The improved residue-based potential[37] also cannot recognize the native and near-native structures in all cases. Using current atomic potential, there are good correlations between energy and RMSD in all cases, and the successful selection of native and near-native structures (2 Å RMSD) with the lowest energy in every test case were also very encouraging.

Extensive testing of the Fischer database sequences demonstrated that the atomic knowledge-based potential could select lower RMSD structures in more cases than could be achieved with the residue-based potentials. This finding implies that the atomic knowledge-based potential will have immediate practical value in protein structure prediction.

Recent work[13,38,39] suggests that using homologous sequence averaging computation of the potential energy would greatly enhance the specificity of the potentials. With the use of a gapless threading test, the average z-score improved by 3 units when using weak sequence-similarity potentials and homologous sequence averaging techniques. It would be natural to assume that similar ideas will also improve the performance of the current

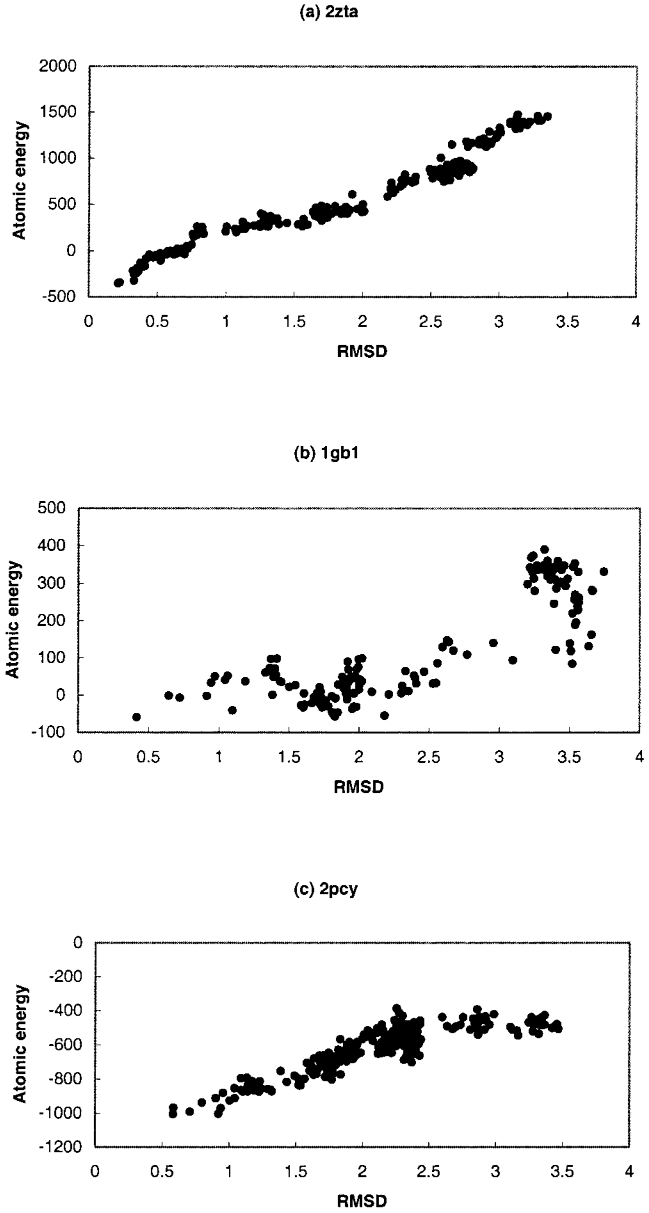


Fig. 6. Energy (in RT kcal/mol) vs. RMSD (in Å) for 2zta (**a**), 1gb1 (**b**), and 2pcy (**c**) for unfolding simulations using molecular dynamics.

atomic knowledge-based potential. Although it will be much more expensive for these protocols to be implemented because the sequence homologous average usually requires detailed atomic reconstruction of at least 10 times more structures than those of target sequences, it may be of use for a limited number of protein-structure predictions.

The results from the unfolding simulation (Fig. 6) show the possibility of refining the structure with atomic potentials because of the good correlation between the RMSD and the atomic potential for structures near the native state. Previously published results[40] showed that both the residue-based potential and molecular mechanics potential with a GB/SA term have good correlation with RMSD over a large range of distances (0–15 Å), but unfortunately for structures close to native ones (RMSD <5 Å), the energies and RMSD were not correlated. Using the current potential, structure refinements from medium-resolution prediction results (4 to 6Å RMSD) to near-native structures (2–3 Å RMSD) will be more likely to be successful.

REFERENCES

1. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
2. Vajda S, Sippl M., Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–238.
3. Mirny L, Shakhnovich E. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
4. Hao M, Scheraga H. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
5. Miyazawa S, Jernigan R. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 1999;36,357–369.
6. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
7. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMm: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 1983;4:187–193.
8. Cornell W, Ciepak P, Bayly C, Gould I, Merz K, Frguson D, Spelleyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. Biochemistry 1995;117:5179–5197.
9. Jernigan R, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
10. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized statitical mechanical energy functions for sequence structure alignment. Protein Sci 1996;5:1043–1059.
11. Thomas P, Dill K. Statistical potentials extracted from protein structure: how accurate are they? J Mol Biol 1996;257:457–469.
12. Betancourt M, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1996;8:361–369
13. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
14. Tobi T, Elber R. Distance-dependent, pair potential for protein folding: Results from linear optimization. Proteins 2000;41:40–46.
15. Duan Y, Kollman P. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
16. Jones D, Taylor W, Thornton J. A new approach to protein fold recognition. Nature 1992;358:86–89.
17. Bowie J, Luthy R, Eisenberg D. Method to identify protein sequences that fold into known three-dimensional structures. Science 1991;253:164–170
18. Godzik A, Skolnick J, Kolinski A. A topology fingerprint approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
19. Bryant S, Lawrence C. An empirical energy function for threading protein sequence through folding motif. Proteins 1993;16:92–112.
20. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268,209–225.
21. Skolnick J, Kolinski A, Ortiz A. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
22. Covell D. Folding protein α-carbon chains into compact forms by Monte Carlo methods. Proteins 1992;14:409–420.
23. Sun S. Reduced representation model of protein structure prediction: statistical potential and generic algorithms. Protein Sci 1993;2:762–785.
24. Kolinski A, Jaroszewski L, Rotkiewicz P, Skolnick J. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass. J Phys Chem 1998;102:4628–4637.

25. Debolt S, Skolnick J. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng 1996;9:637–655.
26. Sippl M, Ortner M, Jaritz M, Lackner P, Flockner H . Helmholtz free energies of atom pair interactions in proteins. Folding Design 1996;1:288–298.
27. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
28. Melo, F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
29. Monge A, Lathrop E, Gunn J, Shenkin P, Friesner R. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
30. Feig, M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks C. Accurate reconstruction of all-atom protein representation from side-chain based low resolution model. Proteins 2000;41:86–97.
31. Park B, Levitt M. Energy functions that discriminate X-ray and near-native foldes from well-constructed decoys. J Mol Biol 1996; 258:367–392.
32. Kolinski A, Betancourt M, Kihara D, Rotkiewicz P, Skolnick J. Generalized Comparative Modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;(in press).
33. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pacific Symp Biocomput 1996;300–318.
34. Hobohm U, Sander C Selection of a representative set of structures from Brookhaven Protein Databank. Protein Sci 1992;1:409–417.
35. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for the improvement of threading-based protein models. Proteins 1999; 37,592–610.
36. Skolnick J, Kihara D. Defrosting the frozen approximation, PROSPECTOR: A new approach to threading. Proteins 2001;42:319–331.
37. Simons K, Ruczinski I, Kooperberg C, Fox B, Bystroff, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
38. Finkelstein A. 3D protein folds: homologs against errors—an estimate based on the random energy model. Phys Rev Lett 1998;80:4823–4825.
39. Reva B, Skolnick J, Finkelstein A. Averaging interaction energies over homologs improves fold recognition in gapless threading. Proteins 1999;35:353–399.
40. Mohanty D, Dominy B, Kolinski A, Brooks C, Skolnick J. Correlation between knowledge-based and detailed atomic potentials. Proteins 1999;35:447–452.