BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 7 2005, pages 981–987
doi:10.1093/bioinformatics/bti080

Structural bioinformatics

A new approach to prediction of short-range conformational propensities in proteins

Dominik Gront* and Andrzej Kolinski

Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

Received on June 8, 2004; revised on September 14, 2004; accepted on October 4, 2004
Advance Access publication October 27, 2004

ABSTRACT

Motivation: Knowledge-based potentials are valuable tools for protein structure modeling and evaluation of the quality of the structure prediction obtained by a variety of methods. Potentials of such type could be significantly enhanced by a proper exploitation of the evolutionary information encoded in related protein sequences. The new potentials could be valuable components of threading algorithms, *ab-initio* protein structure prediction, comparative modeling and structure modeling based on fragmentary experimental data.

Results: A new potential for scoring local protein geometry is designed and evaluated. The approach is based on the similarity of short protein fragments measured by an alignment of their sequence profiles. Sequence specificity of the resulting energy function has been compared with the specificity of simpler potentials using gapless threading and the ability to predict specific geometry of protein fragments. Significant improvement in threading sensitivity and in the ability to generate sequence-specific protein-like conformations has been achieved.

Availability: see: http://www.biocomp.chem.uw.edu.pl

Contact: dgront@chem.uw.edu.pl

INTRODUCTION

A short-range potential means an energy function that evaluates the probability of a certain local conformation of a protein with a given sequence of amino acids. Potentials proposed here depend on the amino acid identities and their sequence context, on the distance between two Ca atoms and, in some cases, on the predicted secondary structure. The idea to use the local sequence similarity for scoring protein structures is not new and has been used in different applications. Details vary between particular applications. For instance, the local sequence similarity could be used as a criterion for selection of short fragments of structures as building blocks (Simons *et al*., 1997) in a fold assembly procedure. It could also be used in derivation of short-range distance restraints (Skolnick *et al*., 2003) to support subsequent threading refinements or restrained *ab-initio* folding (Kolinski *et al*., 2001).

In the present work we provide a systematic derivation of a set of short-range potentials for protein threading, fold evaluation and *ab-initio* algorithms of structure assembly. Results obtained from various databases (various levels of sequence similarity), with and without a support of a given or predicted secondary structure are compared and the ability of the designed potentials to predict a precise geometry of short protein fragments is evaluated. The method is relatively simple and is based on careful analysis of the sequence–structure relationship that employs profile-to-profile alignments (Gribskov *et al*., 1987). The potentials have a form of energy histograms and could be easily implemented in various applications. Of course such potentials are protein dependent. Thus, the detailed prescription for their derivation is provided and for a number of example cases the full datasets were made available via our homepage (http://www.biocomp.chem.uw.edu.pl).

*To whom correspondence should be addressed.

MATERIALS AND METHODS

Input databases

In order to perform the computations described in this work, a custom-designed database has been prepared. The database contains protein structures combined with their sequence profiles. Each residue in a protein is described by its Ca coordinates and by a sequence profile column composed of 20 numbers, defining the probability of each amino acid type occurrence at the corresponding position in a multiple sequence alignment. The profiles were generated with PSIBLAST (Altschul *et al*., 1997) (number of iterations: 7, cutoff *e*-value for including the hit into the profile: 1e-7). Two non-redundant protein structure databases were used. The first one contained proteins with sequence similarity below 30%, the second one of proteins with sequence similarity below 90%. Both sets were extracted from PISCES database set (Wang and Dunbrack, 2003).

The backbone coordinates of some proteins included in the databases are gapped, i.e. some amino acids are missing. Therefore, the local structural properties, such as the $r_{i,i+k}$ distances between alpha carbons, were calculated only for the continuous fragments, i.e. when $r_{i,i+1} = 3.8 \pm 0.5$ Å for all residues included in a fragment. In the structural sense this non-broken subchain has been treated as a separate protein chain. However, the gaps mentioned above were not taken into account in the PSIBLAST search—entire sequences (possibly with some amino acids missing) were used for generating the sequence profile. Because for short (20 amino acids and shorter) fragments the PSIBLAST results are not significant, we decided not to use each separate subchain as an input for PSIBLAST. Gaps existing in the query sequence may result in gaps in the multiple sequence alignment. The resulting profile has been cut into fragments to match structural subchains.

Comparison of profiles

Only short fragments of sequence profiles were compared. Their lengths were fixed and equal L. Depending on the distance range for particular potentials, the optimal values of L have been found to be equal 17, 18 and 19 for $r_{i,i+2}$, $r_{i,i+3}$ and $r_{i,i+4}$ distances, respectively. In the simple case, the profile comparison score can be written as a sum of scores for aligning the related

columns from profiles 1 and 2:

$$S_{\mathrm{P}} = \sum_{i=1}^{L} \mathrm{score}(C_{1,i}, C_{2,i}) \tag{1a}$$

where $C_{j,i}$ is the column corresponding to i-th sequence position in j-th profile. The similarity score for two columns score$(C_{1,i}, C_{2,i})$ from two profiles is defined as follows:

$$\mathrm{score}(C_{1,i}, C_{2,i}) = \sum_{l=1}^{20}\sum_{k=1}^{20} M(k,l) \cdot C_{1,i,l} \cdot C_{2,i,k} \tag{1b}$$

where k and l are amino acid types ($k, l \in \{\mathrm{ALA}, \mathrm{GLY}, \mathrm{etc.}\}$), $C_{p,i,k}$ is the probability of the k-th amino acid type occurrence on the i-th position in the p-th profile. $M(k,l)$ denotes the similarity score for amino acids k and l. We used the BLOSUM62 similarity matrix.

It has to be noted that the raw score given by formulas (1a) and (1b) is highly dependent on the fragment length and its amino acid composition. Therefore, it was normalized in the form of z-score (Panchenko, 2003):

$$z - S_{\mathrm{P}} = \frac{S_{\mathrm{P}} - <S_{\mathrm{P}}>}{\sigma(S_{\mathrm{P}})} \tag{2}$$

The mean value of the score $<S_{\mathrm{P}}>$ and the standard deviation of the score $\sigma(S_{\mathrm{P}})$ has to be estimated for all the permutations of the columns in both profiles. Thus, $<S_{\mathrm{P}}>$ stands for an average alignment score for two profiles with a given length and amino acid composition, no matter what the amino acid order (column order in profiles) is. Consequently, $<S_{\mathrm{P}}>$ was calculated as follows:

$$<S_{\mathrm{P}}> = \frac{1}{L \cdot L}\sum_{j=1}^{L}\sum_{k=1}^{L} \mathrm{score}(C_{1,j}, C_{2,k}) \tag{3}$$

$\sigma(S_{\mathrm{P}})$ was calculated in a similar manner. Due to the non-local scoring we could not use any of the standard alignment tools such as the local sequence alignment (Smith and Waterman, 1981). Indeed, the values $<S_{\mathrm{P}}>$ and $\sigma(S_{\mathrm{P}})$ depend not only on the entire aligned fragment, but also on the amino acids pair being aligned in a given step of the dynamic programming algorithm. The present approach required the assumption that L is a constant. As a result the computational cost was greatly reduced.

Comparison of pairs of sequences

The new short-range potentials proposed in this work heavily rely on the sequence profiles. However, in order to evaluate the effect of evolutionary information on the specificity of the designed potentials, the same calculations (for prediction of the local distances in proteins) have been conducted with single protein sequences. The scoring formulas were very similar to those for scoring profiles:

$$S_{\mathrm{S}} = \sum_{i=1}^{L} M(s_{1,i}, s_{2,i}) \tag{4a}$$

$$z - S_{\mathrm{S}} = \frac{S_{\mathrm{S}} - <S_{\mathrm{S}}>}{\sigma(S_{\mathrm{S}})} \tag{4b}$$

were $s_{j,i}$ denotes the i-th amino acid in the j-th sequence. We did not derive potentials based on the (single) sequence similarity.

Short-range statistical potentials

The general idea of the design of the short-range potential follows our previous work (Kolinski *et al*., 1999; Kolinski and Skolnick, 1998). These potentials were extensively tested in various applications of the reduced protein models, from comparative modeling to *ab-initio* folding (Kolinski *et al*., 1999; Kolinski and Skolnick, 1998; Boniecki *et al*., 2003; Kolinski, 2004). For the reader's convenience it is briefly outlined below.

Potential functions R13, R14 and R15 have been derived for three types of short-range distances: between the i-th and $(i+2)$-th alpha carbons (called r13), the i-th and the $(i+3)$-th (called r14), and the i-th and the $(i+4)$-th alpha carbons (called r15). (According to the convention assumed in this work rij denotes distance and Rij potential corresponding to the rij distance). Denoting $\mathbf{r}_i$ as the coordinate vector of the i-th Ca and $\mathbf{v}_i$ as the unit vector along the virtual Ca–Ca bond, the distances mentioned above are defined as follows:

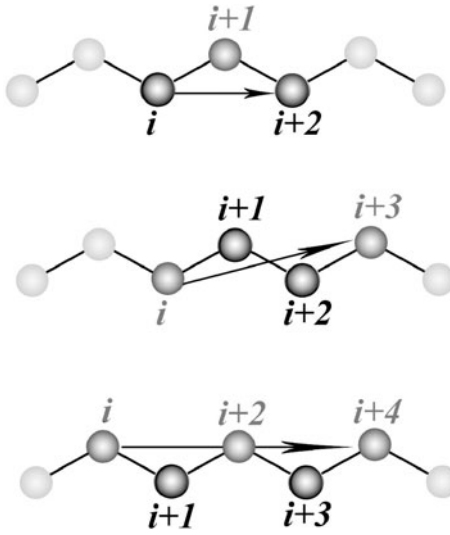

Fig. 1. Definitions of the r13 (at the top), r14 and r15 distances (at the bottom of the picture).

$$\begin{aligned} \mathrm{r13}_i &= |\mathbf{r}_i - \mathbf{r}_{i+2}| \\ \mathrm{r14}^*_i &= |\mathbf{r}_i - \mathbf{r}_{i+3}| \cdot \mathrm{sign}([\mathbf{v}_i \times \mathbf{v}_{i+1}] \cdot \mathbf{v}_{i+2}) \\ \mathrm{r15}_i &= |\mathbf{r}_i - \mathbf{r}_{i+4}| \end{aligned} \tag{5}$$

Statistics for each potential has been generated from a non-redundant structural database, described above. For the r13 statistics, the histograms contained 8 bins from 0 to 8 Å, for r14—24 bins from −12 to 12 Å and for r15 16 bins from 0 to 16 Å. The negative values of the r14 distances denote the left-handed conformations, while the positive ones stand for the right-handed conformations of the three successive Ca backbone vectors. The geometry of the 1–3 (two consecutive virtual Ca bonds) fragments has no chirality and the definition of the chirality of the 1–5 fragments is somewhat ambiguous. Thus, only the chirality of the 1–4 fragments is treated in the explicit way and is denoted by the symbol '*' in the abbreviation $\mathrm{r14}^*_i$. Then, the potentials have been calculated from the histograms as follows:

$$E_k = -\ln\left(\frac{n_i}{n_0} + t\right) \tag{6}$$

where E_k denotes the value of the potential for the k-th bin of the histogram, n_i is the number of observations for the k-th bin and n_0 is the expected number of observation for the k-th bin. The expected values of the histograms are easy to calculate:

$$n_0 = \frac{N_0}{k} = \frac{1}{k}\sum_{i=1}^{k} n_i \tag{7}$$

where N_0 is the total number of observations for the histogram and k is the number of bins in the histogram (8, 24 or 16—depending on the potential). In order to make all the potentials complete, the maximum for all the short-range interactions has been set equal to an arbitrary value of 2.0 and ascribed to all empty bins of the distance histogram.

All the potentials depend on the identity of the two amino acids (see Fig. 1):

(i) The R13 potential for the i-th residue depends on the identity of the i-th and the $(i+2)$-th amino acid.

(ii) The R14 potential for the i-th residue depends on the identity of the $(i+1)$-th and the $(i+2)$-th amino acid.

(iii) The R15 potential for the i-th residue depends on the identity of the $(i+2)$-th and the $(i+4)$-th amino acid.

The potentials could be also made specific to the local secondary structure:

(i) for R13—when amino acids $i, i+1$ and $i+2$ are helical, then the fragment is assigned as a helix. When all the three are in a beta sheet, the fragment is assigned as a beta-type. In all the other cases it is treated as a coil.

(ii) for R14—when amino acids $i, i+1, i+2$ and $i+3$ are helical, then the fragment is assigned as a helix. When all the four are in a beta sheet, the fragment is assigned as a beta-type. In all other cases it is treated as a coil.

(iii) for R15—when amino acids $i, i+1, i+2, i+3$ and $i+4$ are helical, then the fragment is assigned as a helix When all the five are in a beta sheet, the fragment is assigned as a beta-type. In all the other cases it is treated as a coil.

The secondary structure has been assigned by DSSP (Kabsch and Sander, 1983) program assuming the reduced three-letter code. For every type of secondary structure a separate set of potentials has been derived. Thus, for each distance type 1200 ($20 \times 20 \times 3$) different possibilities exist. Statistics for all the cases were collected separately and subsequently transformed into potentials. Consequently, the effects of known or predicted secondary structure can be incorporated into algorithms employing these potentials.

Short-range, protein-dependent (sequence similarity-based) potentials

The use of sequence profiles instead of sequences greatly improves the sensitivity of sequence comparisons. For instance, the assumption that local structural similarity follows local sequence similarity is employed in several secondary structure prediction methods, such as PSIPRED (Jones, 1999) or PHD (Rost and Sander, 1993).

In the present work the short-range potentials, which have to be derived separately for each protein sequence, are designed and evaluated. Statistics accounts only for profiles (of known protein structures) that are locally similar to the sequence profile of the query protein (for which the potential is calculated). Each observation is weighted by the local similarity score. The details of the procedure for calculation of the R13 potential are given below as an example.

Let us consider r13 distance between the i-th and the $(i+2)$-th residues. A protein profile fragment of length L, containing the $(i+2)$-th amino acid at its center is compared to all the profile's fragments in a database. In our case $L = 17$, therefore the i-th, $(i+1)$-th and the $(i+2)$-th residues were at positions 8, 9 and 10 in the fragment of interest.

To further improve the potentials, the secondary structure information can be used. A term scoring similarity between the predicted secondary structure for a query protein and the secondary structure of a protein in the structural database is added.

$$z - S_{\mathrm{PT}} = z - S_{\mathrm{P}} + \sum_{i=1}^{L} S_{\mathrm{T}}^{i} \tag{8}$$

S_{T}^{i} is the similarity score between the secondary structures of two residues:

$$S_{\mathrm{T}}^{i} = \begin{cases} -\varepsilon & \text{when both residues are in E or in H} \\ +\varepsilon & \text{when one of the residues is in E and the other is in H} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

For the central part of the fragment, i.e. $i \in [6, 10]$, $\varepsilon = 0.16$ and 0.08 for the remaining positions.

For each residue in the query protein (except the nine amino acids fragments at the N-terminus and the C-terminus of the sequence) separate histograms were generated. Only the observations of r13 distance with $z - S_{\mathrm{PT}}$ (or $z - S_{\mathrm{P}}$) bigger than a threshold value S_{MIN} were included. For each bin in the histograms, average score ($z - S_{\mathrm{PT}}$ or $z - S_{\mathrm{P}}$) was also calculated. Then, the homology potentials were calculated in a similar fashion as it was done for the simple statistical potentials. The main difference was that the number of hits for a bin in a given histogram was weighted by the average profile similarity score S_i for the bin:

$$E_k = -\ln\left(\frac{n_i \cdot S_i}{n_0} + t\right) \tag{10}$$

$$n_0 = \frac{N_0}{k} = \frac{1}{k}\sum_{i=1}^{k} n_i \cdot S_i \tag{10a}$$

Sometimes there are none, or very little locally similar profiles for some regions of the query sequence. In such cases the profile-based potential cannot be calculated, or the result would be irrelevant. Therefore, the profile-based potentials have been determined only in these cases where the number of hits in the histogram (denoted in the formulas written above as N_0) was higher than a certain threshold value N_{MIN}. Proper entries from the corresponding simple statistical potentials were used for the remaining positions along the sequence. Proper means the same database (PDB90 or PDB30) and the same level of the secondary structure information used in the derivation process.

Potential for scoring r14* and r15 distances have been derived in an analogous fashion. After optimization based on the gapless threading test described below, the best values of the cut-off parameters were found. In case where the secondary structure-based scoring was not applied $S_{\mathrm{MIN}} = 1.0$, otherwise $S_{\mathrm{MIN}} = 1.5$. In all cases $N_{\mathrm{MIN}} = 5.0$. Five observations per histogram may appear to be too small. However, it should be noted that in case of our profile-based potentials all observations fall into one or at most into two bins.

Gapless threading procedure

In order to calculate the gapless threading (Sippl and Weitckus, 1992) score for a pair of proteins, the shorter one has been thread within the longer. For each relative position of the first protein in the second protein the short-range energy of the first sequence in the structure of the second protein has been calculated, as well as the energy of the second sequence in the first structure. A minimum energy has been reported for each 'sequence–structure' pair. Let $E_{i,j}$ denote the energy for the i-th sequence and the j-th structure. Then the mean z-score for threading of the sequence through all the structures from the test set is calculated as follows:

$$z = \frac{1}{N}\sum_{k=1}^{N} \frac{E_{k,k} - <E_{k,i}>_i}{\sigma(E_{k,i})_i} \tag{11}$$

where $< E_{k,i} >_i$ is the average energy calculated for the k-th sequence in all the structures. The mean z-score for threading of all the sequences through a structure is calculated in a similar manner:

$$z = \frac{1}{N}\sum_{k=1}^{N} \frac{E_{k,k} - <E_{i,k}>_i}{\sigma(E_{i,k})_i} \tag{12}$$

The set of proteins used for the gapless-threading test contained only the continuous-chain proteins from the PDB30 database ($N = 1308$ structures). Each protein from the set has been thread through all the remaining proteins except of those longer (or shorter) by 80 amino acids or more than the query sequence.

RESULTS

Dependence between the local sequence similarity and the local structure similarity

For a pair of proteins from the database PDB30 every sequence fragment from the first protein has been compared with every fragment from the second protein. Due to the low-sequence similarity in the database none of the proteins was compared to its close homologues. For all possible pairs of sequence windows, $z - S_{\mathrm{S}}$, $z - S_{\mathrm{P}}$, and difference between r13 distances in the two structural fragments were calculated.

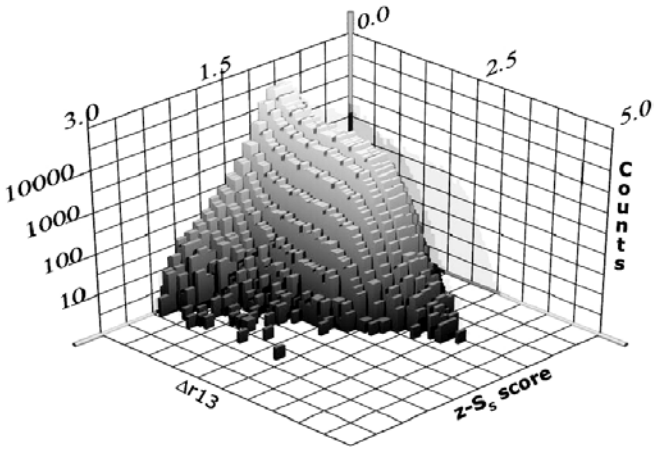

Fig. 2. Correlation between the $z - S_S$ score and the r13 error for the sequence comparisons.

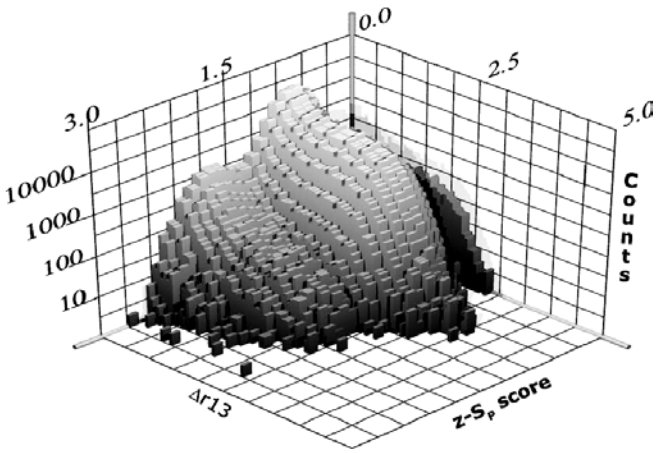

Fig. 3. Correlation between the $z - S_S$ score and the r13 error for the profile–profile comparison.

The collected statistics can be illustrated in a form of a two-dimensional histogram. Figure 2 shows the number of counts for a given $z - S_S$ score and the absolute value of the difference between the r13 values in the compared sequences. The difference between the r13 measures the error of the prediction. In Figure 3 a similar statistics for the case of local profile-based comparison is presented. In both cases a score below 1.4 is not statistically significant—any values of the r13 error in the range of 0–2.5 Å are almost equally probable. Comparison of Figure 2 and Figure 3 shows that the use of the sequence profiles (in contrast to sequences alone) leads to a significant fraction of the high scoring hits, with a very low error of the r13 predictions. Histograms for r14 and r15 distances look very similar to those for r13. It is clear that the value of the local profile similarity score higher than a certain threshold value usually implies a significant local structure similarity. Much larger number of the high scoring (and structurally very similar) fragments was detected using the profile-based approach. Thus, it is expected that the homology potentials should be much more specific.

Table 1. Homology potentials do not cover the whole protein

	% All	% H	% E	% C
PDB30				
Without secondary structure	14.6	14.3	16.3	13.9
With secondary structure	68	91	85.1	22.2
PDB90				
Without secondary structure	37.9	36.6	42.1	36.8
With secondary structure	76.7	92.8	89.4	54.8

The data indicate the percentage of residues, for which homology R13 potential were successfully derived: as an average on the entire sequences (column % All), on the helical residues (column % H), on the residues in beta-sheets (column % E) and on the coil residues (% C).

Table 2. Average percentages of the residues having the native distances in the global minimum of the R13 potential

	% Correctly assigned bins for All residues	Helical residues	Residues in beta-sheet	Residues in coil
PDB30				
Without secondary structure	74.7	88.0	66.5	67.6
With secondary structure	82.3	95.5	71.3	70.4
PDB90				
Without secondary structure	85.6	93.7	80.6	81.4
With secondary structure	84.5	95.4	73.8	77.8

Values were calculated for entire proteins (column % All) and for each type of the secondary structure: helix, beta-sheet and coil, in % H, % E and % C table columns, respectively.

Summary of the profile-based potentials

In order to asses the influence of different factors on the quality of the derived potentials several variants of R13, R14 and R15 potentials were calculated. The set of test proteins contained only non-broken protein chains from the PDB30 and PDB90 databases.

In order to asses the quality of potentials computed from the data which lack homology relationship with the query sequence, we used the PDB30 set (non-broken proteins and fragments, see 'Input databases' for details). The query protein was always removed from the source set. These (low-homology) potentials are addressed to *ab-initio* simulations.

On the contrary, for many sequences, there are many homologous proteins with already known structures. To model this case, we used the PDB90 as a source dataset.

When the secondary structure information is ignored, the profile-based potentials cover 14.6% of all the residues (the fraction of statistically significant hits), when derived from the PDB30 database, and 37.9% when derived from the PDB90 database. When the secondary structure information is included, this ratio raises to 68 and 76.7%, respectively. Detailed comparison of the results is given in Tables 1 and 2.

All calculations described above were conducted with known secondary structure assigned by DSSP. In order to check how the predicted secondary structure influences our potentials we repeated the threading test (threading trough all structures from the PDB30

set) for a set of 37 proteins randomly selected from the PDB30 database. PSIPRED (Jones, 1999) was used as a tool for the secondary structure prediction.

Simple statistical potentials derived from both PDB30 and PDB90 datasets were used as the reference baseline in the evaluation of the profile-based potentials. Gapless threading tests were used for this purpose. Prior the proper tests threading was also used as a tool for optimization of the algorithm's parameters L, N_{MIN} and S_{MIN}. Table 3 contains a summary of the evaluation of the relative performance of various potentials.

The z-scores for the simple statistical potentials appear to be very low. Nevertheless, these potentials (and similar potentials) perform very well in *ab initio* folding simulations and in threading calculations. The explanation is that the local conformational stiffness of polypeptide chains and formation of the secondary structure are cooperative phenomena. Thus, the specificity for larger fragments of the sequence could be significantly higher than it might be expected from the separated entries of the potential. The z-scores for the profile-based potentials are much higher implying a higher specificity. The predicted secondary structure information is almost as good as the exact secondary structure in augmenting the quality of the potentials. Simple tests based on the prediction of the local distances (see Fig. 4) show also a qualitative superiority of the profile-based approach. We expect that the potentials provided here will become a valuable tool for protein structure prediction.

Test of the profile-based short-range potential in the *ab-initio* loop modeling

The simple statistical potentials (employed here as a baseline for evaluation of the quality of the profile-based potentials) were used previously in various applications of protein modeling with a reduced representation of conformational space. It has been shown that the reduced models with knowledge-based force field (where the statistical potentials of the short-range interactions were their essential components) allow much more accurate modeling of protein fragments (or loops) than it is possible with more standard tools of molecular modeling (Boniecki *et al*., 2003; Kolinski, 2004). This way a range of applicability of comparative modeling could be significantly expanded. Recently, we used these statistical potentials for a well-controlled test of applicability of reduced models in loop modeling (Kolinski, 2004), demonstrating very good geometrical fidelity of the resulting models. The results were compared with a very similar test of comparative modeling done recently by Fiser and Sali (2003) for the new version of MODELLER. Here the same experiment is repeated using the profile-based potentials instead the simple statistical ones. All other conditions of the computational experiment are exactly the same as in our previous work, i.e. the same are: the test set of proteins, simulation technique and the remaining components of the force field.

The test set contains five small globular proteins of various structural classes. Using the PDB structures we made the DSSP assignments of their secondary structure. Regular elements of the secondary structure (helices—H and the extended fragments of β-sheets—E) were assumed to be a template for the loop modeling. Remaining portions of the structures were treated as unknown. Random starting conformations of the loops have been generated in the same fashion as in the previously performed experiment with the simple statistical potential of the short-range interactions. The loop optimization has been done using the CABS-reduced representation (Boniecki

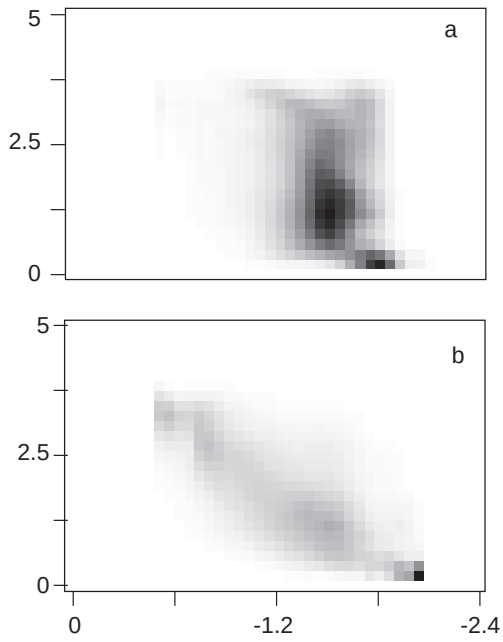

Fig. 4. Dependence between the local energy, calculated for five-residue fragments and dRMSD from native: (**a**) the statistical potential, (**b**) the profile-based potential. In both cases the potentials were derived with known secondary structure from PDB90 database. The gray scale is proportional to the number of counts.

et al., 2003; Kolinski, 2004) and the Replica Exchange Monte Carlo sampling protocol. Entire structures were optimized, although the core (or template) part was kept near the starting conformation by a set of strong native-like distance restraints. The lowest energy structures were selected for the final evaluation. The details of the CABS model, its force field and the sampling details could be found in our recent publications (Boniecki *et al*., 2003; Kolinski, 2004).

The results of the loop modeling are compared in Table 4. The data from the previous work are given in the parentheses. Clearly, the structures generated with the help of the profile-based short-range potentials are consistently more accurate. In two cases (2fdx and 2gb1) the improvement in the model quality was of a qualitative character. Thus it has been demonstrated that the new potentials have higher predictive power not only for the regular fragments of protein structures but they also improve the loop predictions. In all cases the conservative PDB30 versions of the potentials were used. Obviously, the PDB90 potentials can only be more accurate.

CONCLUSION

We derived and compared various short-range interaction potentials using multiple sequence alignments to identify related proteins and then profile–profile local alignments to score sequence to structure compatibility of short fragments. Geometry of these fragments was described as a set of short-range distances between the alpha carbons and the resulting statistical potentials were stored in a form of energy histograms. The potentials were tested in the context of gapless threading, in the ability to predict geometry of short fragments of protein backbone and in a conservative test of its

Table 3. Comparison between the profile-based and the simple statistical potentials in the gapless threading test

Potential type	PDB set	Secondary structure	Z-score[a] Sequence	Structure	Top 1[b] Sequence	Structure	Top 10[c] Sequence	Structure
(a)	30	N	2.05	0.32	0.31	0.02	0.49	0.13
Simple statistical	90	N	2.04	0.33	0.31	0.02	0.49	0.13
	30	P	2.96	2.22	0.68	0.50	0.87	0.77
	90	P	2.96	2.23	0.67	0.51	0.87	0.77
	30	Y	3.57	2.71	0.80	0.66	0.95	0.84
	90	Y	3.57	2.72	0.80	0.66	0.95	0.84
Homology (profile-based)	30	N	1.63	0.82	0.46	0.12	0.61	0.28
	90	N	2.82	2.46	0.79	0.56	0.90	0.73
	30	P	3.85	3.48	0.85	0.75	0.95	0.90
	90	P	3.90	3.76	0.91	0.86	0.95	0.91
	30	Y	4.08	3.30	0.85	0.74	0.97	0.89
	90	Y	3.91	3.87	0.80	0.74	0.88	0.89
(b)	30	N	2.30	0.28	0.35	0.01	0.53	0.10
Simple statistical	90	N	2.27	0.28	0.34	0.01	0.52	0.10
	30	P	3.50	1.53	0.67	0.24	0.89	0.52
	90	P	3.33	1.54	0.68	0.26	0.89	0.53
	30	Y	4.05	1.96	0.77	0.37	0.93	0.63
	90	Y	4.04	1.96	0.76	0.37	0.93	0.63
Homology (profile-based)	30	N	1.29	0.50	0.36	0.04	0.55	0.16
	90	N	1.88	1.83	0.66	0.31	0.83	0.50
	30	P	3.86	2.69	0.79	0.56	0.93	0.80
	90	P	3.77	3.00	0.80	0.63	0.94	0.84
	30	Y	4.12	2.79	0.79	0.55	0.93	0.77
	90	Y	4.64	3.86	0.86	0.82	0.92	0.86
(c)	30	N	3.92	0.97	0.59	0.03	0.71	0.14
Simple statistical	90	N	3.88	0.99	0.58	0.04	0.77	0.26
	30	P	3.50	3.13	0.76	0.60	0.91	0.84
	90	P	3.49	3.14	0.76	0.60	0.91	0.84
	30	Y	3.94	3.60	0.83	0.70	0.96	0.88
	90	Y	3.93	3.60	0.82	0.70	0.96	0.89
Homology (profile-based)	30	N	1.93	1.64	0.62	0.13	0.80	0.40
	90	N	5.43	2.57	0.89	0.61	0.95	0.81
	30	P	5.38	4.88	0.94	0.83	0.98	0.96
	90	P	5.32	6.30	0.94	0.85	0.98	0.96
	30	Y	4.55	4.08	0.84	0.72	0.96	0.87
	90	Y	6.75	6.53	0.94	0.84	0.96	0.92
(d)	30	N	3.64	0.53	0.55	0.02	0.14	0.71
Simple statistical	90	N	3.60	0.53	0.55	0.02	0.15	0.71
	30	P	3.53	2.37	0.76	0.50	0.93	0.78
	90	P	3.54	2.38	0.76	0.51	0.93	0.79
	30	Y	4.07	2.91	0.85	0.66	0.86	0.97
	90	Y	4.06	2.91	0.85	0.66	0.86	0.97
Homology (profile-based)	30	N	1.86	0.97	0.60	0.08	0.27	0.75
	90	N	3.42	2.42	0.87	0.52	0.73	0.95
	30	P	4.65	3.34	0.85	0.75	0.95	0.90
	90	P	4.58	4.72	0.91	0.81	0.97	0.95
	30	Y	4.50	3.62	0.87	0.74	0.96	0.88
	90	Y	5.62	5.17	0.92	0.82	0.95	0.90

The data show the results for R13 (a), for R14 (b), for R15 (c), and for all potentials combined together (d). Eight kinds of potentials have been tested: profile-based and simple-statistical, with and without secondary structure information, derived from PDB90 or PDB30 database.

[a] z-score calculated in the gapless threading of all structures through a sequence (column 'sequence'), all sequences through a structure (column 'structure').

[b] 'Top 1' shows the ratio of native sequences selected as the highest scoring from all the sequences used in threading sequences through a structure (column 'sequence') and the ratio of native structures selected as the best structure (column 'structure').

[c] 'Top 10' is analogous to 'Top 1', but shows the ratio of native structures or sequences found among ten best scoring.

N, secondary structure not used; P, predicted secondary structure; Y, known secondary structure (DSSP).

Table 4. Comparison of the performance of the sequence similarity-based potential with the simple statistical potentials in the loop modeling of globular proteins

Name	Type	N	N_L	N_{max}	cRMSD (Å) All	Core	Loops
1ten	β	89	41	7	1.62 (1.67)	0.53 (0.54)	2.18 (2.28)
256B	α	106	22	7	1.19 (1.28)	0.40 (0.42)	2.13 (2.32)
2fdx	α/β	138	50	6	1.12 (1.58)	0.44 (0.49)	1.60 (2.17)
2gb1	$\alpha+\beta$	56	21	6	0.88 (1.21)	0.53 (0.57)	1.26 (1.69)
4mba	α	146	34	8	1.25 (1.34)	0.60 (0.60)	2.17 (2.45)

N, protein length (number of residues); N_L, total number of the loop residues, N_{max}, the length of the longest loop; cRMSD, coordinate root mean square deviation from the native structure after the best superimposition; all, cRMSD for entire model after the best superimposition with the crystallographic structure; core, cRMSD for the core part of the model after best superimposition of the core; loops, cRMSD for the all loop residues of the model after best superimposition of the core structure. The data from the previous experiments with the simple statistical potentials are given in parentheses.

application to comparative modeling. It has been demonstrated that a higher level of sequence similarity in the structural database as well as known (or predicted) secondary structure increase the specificity and sensitivity of the potentials. Interestingly, the new potentials work well also in the loop regions of protein structures. Example data for a set of proteins are available on our homepage (http://www.biocomp.chem.uw.edu.pl). The algorithms for derivation of the potentials for large sets of proteins are available upon request. Future applications of the new potentials include refinement of the threading alignments, homology modeling with reduced representation of the protein conformational space and *ab initio* structure prediction for globular proteins.

ACKNOWLEDGEMENT

This work was partially supported by grant no PBZ-KBN-088/P04/2003. We would like to express our thanks to Anna Oleksy for critical reading of this manuscript.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Boniecki,M., Rotkiewicz,P., Skolnick,J. and Kolinski,A. (2003) Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.*, **17**, 725–738.

Fiser,A. and Sali,A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, **19**, 2500–2501.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci., USA*, **84**, 4355–4358.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.

Kolinski,A., Betancourt,M.R., Kihara,D., Rotkiewicz,P. and Skolnick,J. (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins*, **44**, 133–149.

Kolinski,A., Rotkiewicz,P., Ilkowski,B. and Skolnick,J. (1999) A method for the improvement of threading-based protein models. *Proteins*, **37**, 592–610.

Kolinski,A. and Skolnick,J. (1998) Assembly of protein structure from sparse experimental data: an efficient Monte Carlo Model. *Proteins*, **32**, 475–494.

Panchenko,A.R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.*, **31**, 683–689.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Simons,K.T., Kooperberg,C., Huang,E. and 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.

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

Skolnick,J., Zhang,Y., Arakaki,A.K., Kolinski,A., Boniecki,M., Szilagyi,A. and Kihara,D. (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **53** (Suppl. 6), 469–479.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.