Protein Engineering, Design & Selection vol. 19 no. 5 pp. 187–193, 2006
Published online March 13, 2006 doi:10.1093/protein/gzj018

A knowledge-based scoring function based on residue triplets for protein structure prediction

Shing-Chung Ngan, Michael T.Inouye and Ram Samudrala[1]

Computational Genomics Group, Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195, USA

[1]To whom correspondence should be addressed.
E-mail: ram@compbio.washington.edu

One of the general paradigms for *ab initio* protein structure prediction involves sampling the conformational space such that a large set of decoy (candidate) structures are generated and then selecting native-like conformations from those decoys using various scoring functions. In this study, based on a physical/geometric approach first suggested by Banavar and colleagues, we formulate a knowledge-based scoring function, which uses the radii of curvature formed among triplets of residues in a protein conformation. By analyzing its performance on various decoy sets, we determine a good set of parameters—the distance cutoff and the number of distance bins—to use for configuring such a function. Furthermore, we investigate the effect of using various approaches for compiling the prior distribution on the performance of the knowledge-based function. Possible extensions to the current form of the residue triplet scoring function are discussed.

Keywords: *ab initio* prediction/Bayesian/protein structure

Introduction

In protein structure prediction, a given sequence with one or more known homologs whose conformations have been experimentally determined can be modeled with comparative modeling techniques (Blundell *et al*., 1987; Bajorath *et al*., 1994; Johnson *et al*., 1994; Sali 1995; Sanchez and Sali, 1997). On the other hand, a sequence with no obvious homologs is often modeled using *ab initio* methods (Friesner and Gunn, 1996; Jones 1997; Levitt *et al*., 1999). One of the general paradigms for *ab initio* structure prediction involves sampling the conformational space such that a large set of 'decoy' structures are generated and then selecting native-like conformations from those decoys using various scoring functions (Samudrala *et al*., 1999; Samudrala and Levitt, 2002). Since the first papers on protein structure prediction appeared some 30 years ago, both conformational space sampling and scoring function design have remained as major challenges in *ab initio* structure prediction to this day (Moult *et al*., 1997, 1999, 2001, 2003).

There are two broad categories of scoring functions. The first category of functions are largely based on some aspects of the known physics of molecular interaction, such as the van der Waals force, electrostatics, and the bending and torsional forces, to determine the energy of a particular conformation (Brooks *et al*., 1983; Weiner *et al*., 1986; Jorgensen and Tirado-Rives, 1988; Nemethy *et al*., 1992; Cornell *et al*., 1995; MacKerell *et al*., 1998). The second category of functions are knowledge-based. Each of these knowledge-based functions tries to capture some aspects of the protein native conformations, such as the tendency of a certain amino acid to be exposed or buried relative to the solvent, or to be part of the helix, strand or coil local structure and so on. These knowledge-based functions are compiled based on the statistics of a database of experimentally determined protein structures (Wodak and Rooman, 1993; Sippl 1995; DeBolt and Skolnick, 1996; Gilis and Rooman, 1996; Jernigan and Bahar, 1996; Zhang *et al*., 1997; Samudrala and Moult, 1998). Interaction between these two categories of functions has resulted in a fertile ground for the experimentation and construction of new scoring functions. In this study, based on a physical/geometric approach first suggested by Banavar and colleagues (Maritan *et al*., 2000; Banavar *et al*., 2002, 2003a, b), we formulate and analyze an analogous knowledge-based scoring function (denoted as the residue triplet scoring function), which involves the radii of curvatures formed among triplets of residues in a protein conformation. We also investigate the effect of using various approaches for compiling the prior distribution on the performance of the knowledge-based function.

The paper is organized as follows. We first briefly review the physical/geometric approach of Banavar and colleagues. We then describe the construction of a knowledge-based scoring function which incorporates some key features from that of Banavar *et al*. The performance of the knowledge-based function in structure prediction is evaluated through its application to 41 decoy sets of various quality. Finally, we propose some possible extensions to the current form of the scoring function.

Theoretical background and methods

The three-body potential of Banavar et al.

In Maritan *et al*. (2000) and Banavar *et al*. (2002, 2003a, b), Banavar and colleagues viewed a protein chain as a system of discrete particles and considered interactions among any three particles through a three-body potential. By drawing a circle through any given three particles, the radius of curvature could be determined and was used as the input variable to the potential function. In their Monte-Carlo simulation of protein chain folding, a Lennard-Jones type function was chosen as

the potential. It was demonstrated that protein-like structures, such as short segments of helices with special pitch-to-radius ratio, sheets and hairpins, were naturally obtained as ground states in their simulations.

A knowledge-based formulation of the three-body potential

Our formulation of the knowledge-based residue triplet potential is analogous to the standard pairwise residue distance-dependent scoring function, with two main modifications. First, the two-body potential in the pairwise case is replaced by a three-body potential. Second, the pairwise residue distances, which form inputs to the score calculation for a given conformation, are replaced by the radii of curvature of residue triplets. It should be noted that a residue triplet does not necessarily consist of three residues consecutive in sequence, just as a residue pair does not necessarily correspond to a pair of neighboring residues in the two-body potential. Precisely, in terms of the Bayesian statistics formalism as described in Samudrala and Moult (1998), we view a given set of conformations for a protein sequence as comprising of two subsets: a subset of correct conformations {C} and a subset of incorrect conformations {I}. For a given conformation, we calculate the probability that it belongs to the subset of correct structures {C}, given some properties of the conformation. In our present case, these properties are the set of distances $\{r_{abc}^{ijk}\}$, where r_{abc}^{ijk} is the radius of curvature formed by residues i, j and k of residue types a, b and c. The probability is denoted as $P(C|\{r_{abc}^{ijk}\})$. Using Bayes' theorem, one obtains

$$P(C)P(\{r_{abc}^{ijk}\}|C) = P(\{r_{abc}^{ijk}\})P(C|\{r_{abc}^{ijk}\}) \quad (1)$$

where $P(\{r_{abc}^{ijk}\}|C)$ is the (posterior) probability of observing the set of radii of curvature $\{r_{abc}^{ijk}\}$ in a correct structure, $P(\{r_{abc}^{ijk}\})$ is the (prior) probability of observing such a set of radii in any correct or incorrect structure and $P(C)$ is the probability that any structure picked at random is a member of the correct set. To ensure computational feasibility, we make a simplifying assumption that the radii are independent of one another:

$$P(\{r_{abc}^{ijk}\}|C) = \prod_{i,j,k} P(r_{abc}^{ijk}|C); \; P(\{r_{abc}^{ijk}\}) = \prod_{i,j,k} P(r_{abc}^{ijk}) \quad (2)$$

Combining Equations (1) and (2) gives

$$P(C|\{r_{abc}^{ijk}\}) = P(C)\prod_{i,j,k} \frac{P(r_{abc}^{ijk}|C)}{P(r_{abc}^{ijk})} \quad (3)$$

Equation (3) suggests a scoring function S, which is proportional to the negative log conditional probability that the given structure is correct, given a set of radii of curvature:

$$S(\{r_{abc}^{ijk}\}) = -\sum_{i,j,k} \log\left(\frac{P(r_{abc}^{ijk}|C)}{P(r_{abc}^{ijk})}\right) \quad (4)$$

Before one can use Equation (4) as a scoring function, the statistics for the posterior probability $P(r_{abc}^{ijk}|C)$ and the prior probability $P(r_{abc}^{ijk})$ need to be compiled. Specifically, to compute the statistics for $P(r_{abc}^{ijk}|C)$, we tabulate the radii of curvature generated by residue triplets in a set of experimentally determined conformations available from the Protein Data Bank (PDB) (Westbrook *et al*., 2003; Bourne *et al*., 2004). This set of conformations was created by first selecting all proteins that appear in the *e*-value filtered ASTRAL SCOP genetic domain sequence subset list with the threshold *e*-value set at 10^{-4} (Chandonia *et al.*, 2004). Subsequently, we retained proteins whose lengths are less than 300 residues (primarily for computational efficiency) and removed proteins whose PSI-BLAST *e*-values are less than 2 with respect to a set of 41 protein sequences we later use for test decoy set generation and scoring function testing. This results in a total of 3150 structures (hereafter denoted as the database of solved protein structures). We then evaluate the quantity

$$P(r_{abc}^{ijk}|C) \equiv \frac{N(r_{abc})}{\sum_r N(r_{abc})} \quad (5)$$

where $N(r_{abc})$ is the number of occurrences of triplets with residue types a, b and c whose radius of curvature is in the distance bin r. For compilation of the statistics of $P(r_{abc}^{ijk})$, we attempt three approaches in this study. In the first approach, for each protein sequence in the database of the solved protein structures, we use an *ab initio* conformational space sampling protocol to generate 10 decoy structures, as a result yielding a total of $3150 \times 10 = 31\,500$ decoy structures (hereafter denoted as the database of decoy structures). The *ab initio* conformational space sampling protocol consists of a Monte-Carlo method with simulated annealing procedure, with move set based on the standard fragment replacement scheme, namely, the existing conformation of three consecutive residues at a random position is replaced by the torsion values of three residues with identical sequence from an experimentally determined structure (Simons *et al*., 1997; Hung and Samudrala, 2003). The energy function used to generate the decoys is a combination of the all-atom distance-dependent function, a hydrophobic compactness function and a bad contacts function (Samudrala *et al*., 1999; Samudrala and Levitt, 2002). We use the database of the 31 500 decoy structures to determine the prior distribution $P(r_{abc})$ analogous to the way the database of the solved structures is used in Equation (5) for the posterior distribution:

$$P(r_{abc}) = \frac{N(r_{abc})}{\sum_r N(r_{abc})} \quad (6)$$

As a second approach, we apply the mixture method described in Samudrala and Moult (1998), i.e. instead of using the database of the 31 500 decoy structures, the database of the 3150 solved structures is employed and averaging is done across the various residue types when determining the prior distribution. Specifically, $P(r_{abc})$ is calculated by

$$P(r_{abc}) = P(r) = \frac{\sum_{abc} N(r_{abc})}{\sum_r \sum_{abc} N(r_{abc})} \quad (7)$$

where $\sum_{abc} N(r_{abc})$ is the number of contacts among all residue triplets in a particular distance bin r in the database of the solved structures, regardless of residue types. Finally, as a third approach, Equation (7) is again employed to compile the statistics of the prior distribution, i.e. averaging is again performed across the various residue types. However, the compilation is done on the database of the 31 500 decoy structures, instead of the 3150 solved structures.

Generation of test decoy sets and evaluation of the residue triplet scoring function

To evaluate the performance of the residue triplet scoring function in distinguishing native-like from non-native

conformations, we apply it to 41 test decoy sets of various quality. The 41 test decoy sets correspond to 41 protein sequences, some of them taken from the second through fifth Community Wide Experiments on the Critical Assessment of Techniques for Protein Structure Prediction (Moult *et al*., 1997, 1999, 2001, 2003) and the rest randomly picked from the PDB. Each decoy is generated using the same conformational space sampling protocol described in the preceding sub-section. Each run consists of 100 000 iterations using the fragment replacement move set and yields 10 decoys at the end of the run. One thousand seeds are used to generate 10 000 decoys for each test decoy set.

Table I gives the PDB identifiers and the SCOP classifications of the 41 protein sequences used in generating the test decoy sets. Also included is the C_α root mean squared deviation (RMSD) of the best decoy relative to the corresponding native structure in each test set. Among them, 15 test decoy sets have their best structures below 6 Å C_α RMSD relative to their native conformations. (Twenty-four decoy sets have their best structures below 7 Å C_α RMSD relative to their native conformations.) We denote those 15 sets as the high quality test decoy sets.

Table I. List of the protein sequences used in generating the test decoy sets

Protein	SCOP classifications	Length	Min. RMSD
d1b0n-A2	a.35.1.3 (A:1–68)	68	2.729
d1b33-N	d.30.1.1 (N:)	67	7.349
d1b34-A	b.38.1.1 (A:)	80	7.943
d1b4b-A	d.74.2.1 (A:)	71	5.506
d1b79-A	a.81.1.1 (A:)	102	5.29
d1ck9-A	d.79.3.1 (A:)	104	7.661
d1ctf	d.45.1.1 (-)	68	4.37
d1dgn-A	a.77.1.1 (A:)	89	4.482
d1dj8-A	a.57.1.1 (A:)	79	5.092
d1dtj-A	d.51.1.1 (A:)	74	4.902
d1e68-A	a.64.2.1 (A:)	70	3.794
d1eai-C	g.22.1.1 (C:)	61	6.914
d1edz-A2	c.58.1.2 (A:3–148)	146	9.277
d1efu-B3	a.5.2.2 (B:1–54)	54	5.247
d1ev0-A	d.71.1.1 (A:)	58	6.641
d1f53-A	b.11.1.4 (A:)	84	9.123
d1fc3-A	a.4.6.3 (A:)	119	8.184
d1fmt-A1	b.46.1.1 (A:207–314)	108	7.385
d1g6e-A	b.11.1.6 (A:)	87	7.891
d1g7d-A	a.71.1.1 (A:)	106	5.867
d1goi-A1	b.72.2.1 (A:447–498)	52	6.111
d1gut-A	b.40.6.1 (A:)	67	6.459
d1h5p-A	b.99.1.1 (A:)	95	8.223
d1h8a-C1	a.4.1.3 (C:87–143)	57	2.941
d1ijy-A	a.141.1.1 (A:)	122	7.916
d1ira-Y1	b.1.1.4 (Y:1-101)	101	8.317
d1iwg-A1	d.58.44.1 (A:38–134)	97	5.7
d1jju-A3	b.1.18.14 (A:274–351)	78	6.614
d1jos-A	d.52.7.1 (A:)	100	5.302
d1jyg-A	a.60.11.1 (A:)	69	3.471
d1k2y-X2	c.84.1.1 (X:155–258)	104	6.889
d1ktz-B	g.7.1.3 (B:)	106	8.586
d1l9l-A	a.64.1.1 (A:)	74	4.041
d1msp-A	b.1.11.2 (A:)	124	9.932
d1n69-A	a.64.1.3 (A:)	78	6.753
d1qu6-A1	d.50.1.1 (A:1–90)	90	8.597
d1rie	b.33.1.1 (-)	127	9.548
d1sra	a.39.1.3 (-)	151	8.781
d1sro	b.40.4.5 (-)	76	6.031
d2igd	d.15.7.1 (-)	61	6.508
d7gat-A	g.39.1.1 (A:)	66	7.248

Each row lists the PDB identifier of the sequence, the SCOP classification, the length of the protein sequence and the C_α RMSD of the best decoy structure relative to the native conformation in the test decoy set. Fifteen test decoy sets have their best structures below 6 Å C_α RMSD relative to their corresponding native conformations. Twenty-four test decoy sets have their best structures below 7 Å C_α RMSD relative to their corresponding native conformations.

We use two measures to evaluate the quality of the residue triplet scoring function. This first measure is the enrichment ratio. After the scoring function is applied to a test decoy set, we count the number of decoys (denoted as a) which are in the top 10% both in terms of their residue triplet scores and their C_α RMSD relative to the native structures. The expected number in a random distribution is $10\% \times 10\% \times \{\text{number of decoys in the set}\}$ (denoted as b). The enrichment ratio is a/b. A value above 1 indicates enrichment over the random distribution. The second measure is obtained via the receiver-operating characteristic (ROC) analysis. A decoy structure is a priori classified as true positive if its C_α RMSD relative to the native structure is in the top 10% among all the decoys in the test set. The lower 90% decoy structures are classified as true negative. After the residue triplet score has been computed for each decoy in a test set, we start with the best scoring decoy and expand the collection of the 'native-like' decoys by adding one decoy at a time. The true positive fraction and the false positive fraction (FPF) are determined for each successive step and plotted against each other to generate the ROC curve. The area under a truncated ROC curve (with $0 \leqslant \text{FPF} \leqslant 0.1$ in this study) generated by the residue triplet scoring function (denoted as A_s), divided by the expected area under a truncated ROC curve corresponding to the random distribution (denoted as A_r), indicates the improvement of the scoring function over the random distribution. The percentage improvement is simply $100\% \times (A_s - A_r)/A_r$.

Selection of the distance cutoff and the number of distance bins

Before one can compile the statistics for the posterior probability $P(r_{abc}|C)$ and the prior probability $P(r_{abc})$ using Equations (5–7), the distance cutoff, the number of distance bins and the bin sizes have to be fixed. It is not clear a priori what the best values for these parameters are. Thus, we try a number of possibilities in this study. Distance cutoffs from 12 to 16 Å and numbers of bins ranging from 4 to 11 are tested. Bin widths are determined in the following manner: Figure 1 depicts the distribution of the radius of curvature for triplets (regardless of residue types) observed in the database of the solved structures. If, for example, we fix a cutoff distance of 15 Å and the number of bins to be five, then we choose the bin widths in such a way that each bin will have approximately equal area underneath the distribution curve, holding roughly the same number of observed radii. There are of course other ways to sub-divide the bin sizes. We perform the subdivision in this particular manner mainly to restrict the search space for finding reasonably good parameter values.

Results and discussion

A good parameter set for configuring the residue triplet scoring function

Figures 2a, 3a and 4a illustrate the various enrichment ratios that the residue triplet scoring functions produce and Figures 2b, 3b and 4b show the corresponding percentage improvement in the truncated ROC measure.

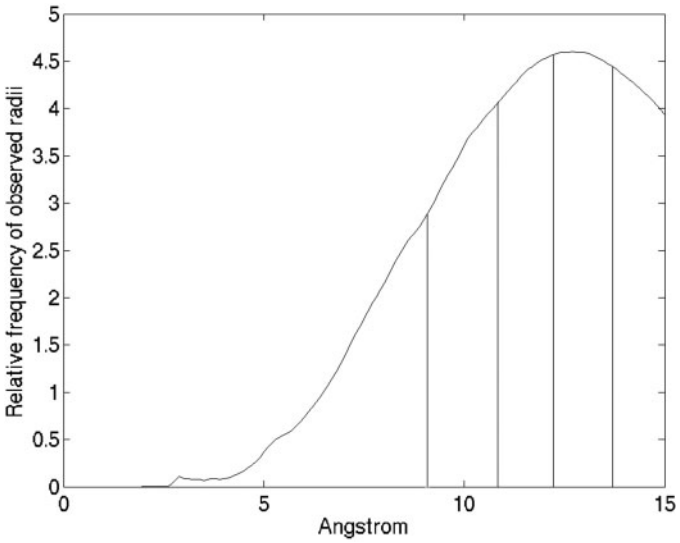

Fig. 1. Distribution of the radii of curvature for all triplets. Triplets are obtained from a database of solved protein structures and are considered regardless of residue types. In this example, we sub-divide the area under the distribution curve into 5 bins, with the distance cutoff at 15 Å. Each bin has approximately equal area, which means that they hold roughly the same number of observed radii.

(Figures 2–4 extract and summarize data in Supplementary Tables I–III, respectively available at *PEDS* online.) Figure 2 illustrates the performance of the scoring functions [haereafter denoted as the residue specific decoy structure based triplet (RSDT) functions] that employ a residue type specific compilation of the prior distribution $P(r_{abc})$ derived from the database of the 31 500 decoy structures. In Figure 3, the scoring functions [hereafter denoted as the residue non-specific solved structure based triplet (RNST) functions] use a residue type non-specific compilation of the prior distribution derived from the database of the 3150 solved structures. In Figure 4, a residue type non-specific compilation of the prior distribution derived from the database of the 31 500 decoy structures is employed in constructing the scoring functions [hereafter denoted as the residue non-specific decoy structure based triplet (RNDT) functions].

Overall, focusing on the performances of the scoring functions on the high quality test decoy sets (i.e. the 15 test decoy sets that contain structures of less than 6 Å C_α RMSD relative to the native conformations), by comparing Figures 2(a and b), 3(a and b) and 4(a and b), we see that a good set of parameters for the residue triplet scoring function is a distance cutoff of 14 Å with 7 distance bins (alternatively, a distance cutoff of 14 Å with 8 bins also gives similar performance) and with the prior distribution $P(r_{abc})$ generated with a residue type specific compilation of the database of the 31 500 decoy structures. This produces an enrichment ratio of ~1.33 and an ROC improvement of ~45%. Analysis based on the standard leave-one-out cross-validation yields similar results, with an average enrichment ratio of 1.32 and an average ROC improvement of 42%. For test decoy sets of lesser quality, this particular configuration of the scoring function maintains the overall enrichment ratio above 1.21 and the ROC improvement above 30% [numerical values detailed in Supplementary Tables Ia(ii–v) and Ib(ii–v) available at *PEDS* online].

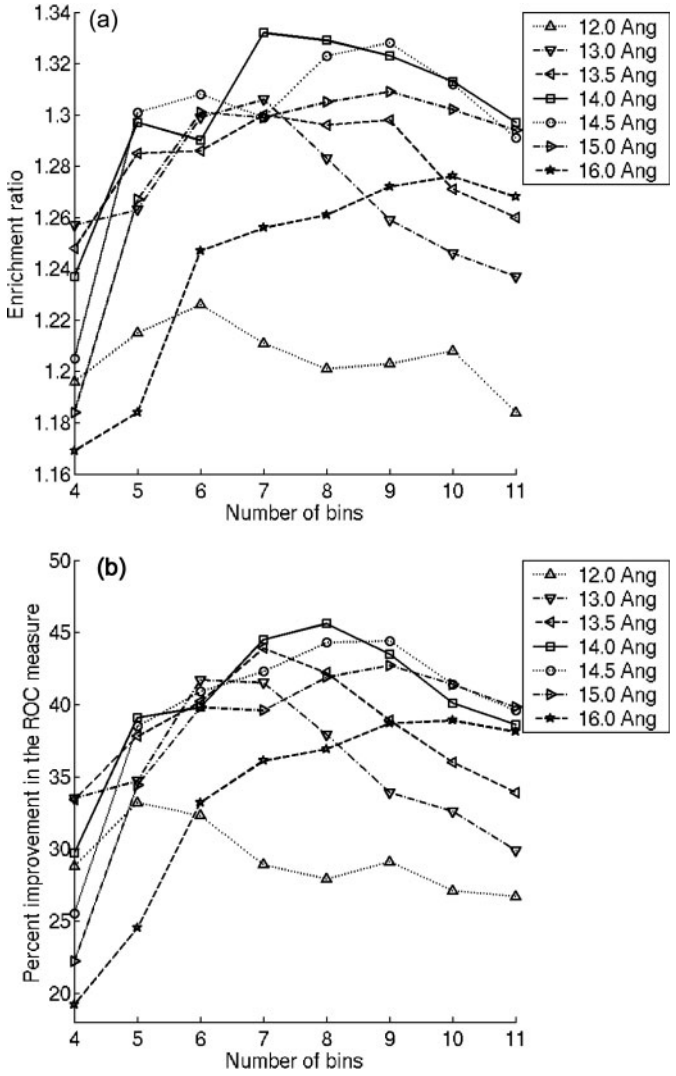

Fig. 2. Performance of the RSDT functions. Shown are (**a**) the average enrichment ratios and (**b**) the percentage improvement in the ROC measure achieved by the RSDT functions when they are applied to the high quality test decoy sets. The RSDT functions are constructed with a residue type specific compilation of the prior distribution derived from the database of 31 500 decoy structures. Distance cutoff ranging from 12 to 16 Å and the number of bins ranging from 4 to 11 are examined. Configurations with a distance cutoff of 14 Å with 7 distance bins and with a distance cutoff 14 Å with 8 distance bins give the best results.

Choice of the prior distribution

By inspecting Figures 2–4, we observe that switching from using a prior distribution $P(r_{abc})$ generated with a residue type specific compilation of the database of decoy structures, to the one generated with a residue non-specific compilation of the database of solved structures and the one generated with a residue non-specific compilation of the database of decoy structures, depresses the performance of the residue triplet scoring function in general. For example, for the high quality decoy sets, the best enrichment ratios are ~1.15 (Figure 3a) and ~1.18 (Figure 4a) and the best ROC improvements are ~17% (Figure 3b) and ~23% (Figure 4b) for the functions configured with the latter two prior distributions. These values are lower than the enrichment ratio of 1.33 and the ROC improvement of 45% for the RSDT function.

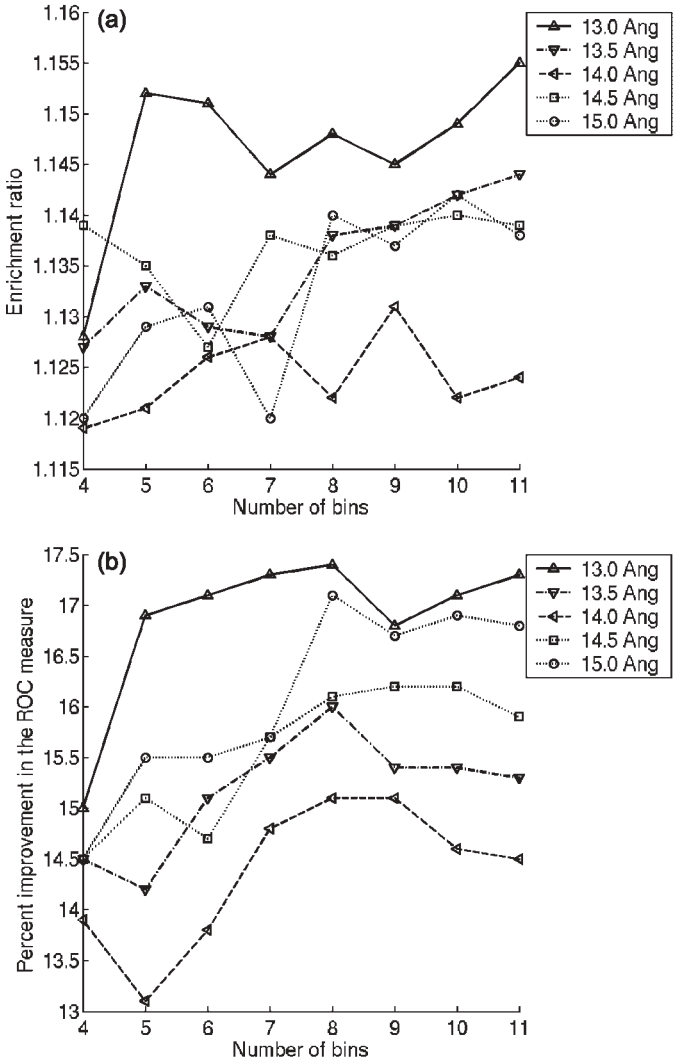

Fig. 3. Performance of the RNST functions. Shown are (**a**) the average enrichment ratios and (**b**) the percentage improvement in the ROC measure achieved by the RNST functions when they are applied to the high quality test decoy sets. The RNST functions are constructed with a residue type non-specific compilation of the prior distribution derived from the database of 3150 solved structures. Distance cutoff ranging from 13 to 15 Å and the number of bins ranging from 4 to 11 are examined. Comparing with Figure 2, we observe that the RNST scoring functions generally have lower performances.

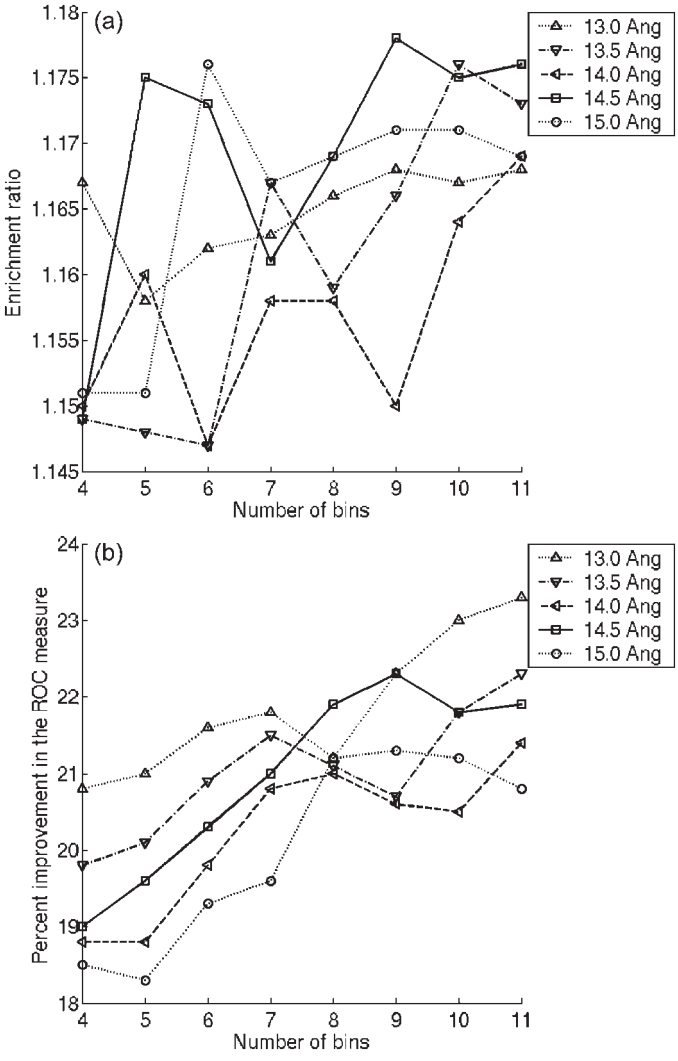

Fig. 4. Performance of the RNDT functions. Shown are (**a**) the average enrichment ratios and (**b**) the percentage improvement in the ROC measure achieved by the RNDT functions when they are applied to the high quality test decoy sets. The RNDT functions are constructed with a residue type non-specific compilation of the prior distribution derived from the database of 31 500 decoy structures. Distance cutoff ranging from 13 to 15 Å and the number of bins ranging from 4 to 11 are examined. Comparing with Figure 2, we observe that the RNDT scoring functions generally have lower performances.

The best performing RSDT, RNST and RNDT scoring functions are selected from Figures 2–4 and their enrichment ratios and ROC percentage improvements are plotted in Figures 5 and 6 across test decoy sets of various quality. The performance differences among the RSDT, RNST and RNDT functions depicted in these two figures indicate that a residue type specific derivation of the prior distribution can boost the accuracy of the scoring function over one based on a residue type non-specific derivation. Furthermore, according to the figures, the performance of the RNDT function seems to be slightly better than that of the RNST function. This observation suggests the importance of using the same conformational space sampling protocol for creating test decoy sets as well as for generating the database of decoy structures for prior distribution derivation, at least in the context of constructing the residue triplet scoring function. Despite the above-mentioned disadvantage, the RNST scoring function is still useful in instances where a priori information about the conformational space sampling protocol used in generating the test decoy sets is either not known or not utilized, since only the database of solved structures is needed in compiling the statistics of the prior distribution. A good way to further explore and understand the comparative effectiveness of the three approaches for prior distribution estimation is to study them in the context of other knowledge-based functions (for example, in the construction of the pairwise residue distance-dependent scoring function).

Comparing the performance of the residue triplet scoring function to other established functions

To provide a rough yardstick for measuring the performance of the residue triplet scoring function, we apply the all-atom distance-dependent conditional probability discriminatory function [denoted as the RAPDF function in Samudrala and Moult (1998)] to the 41 test decoy sets. The RAPDF

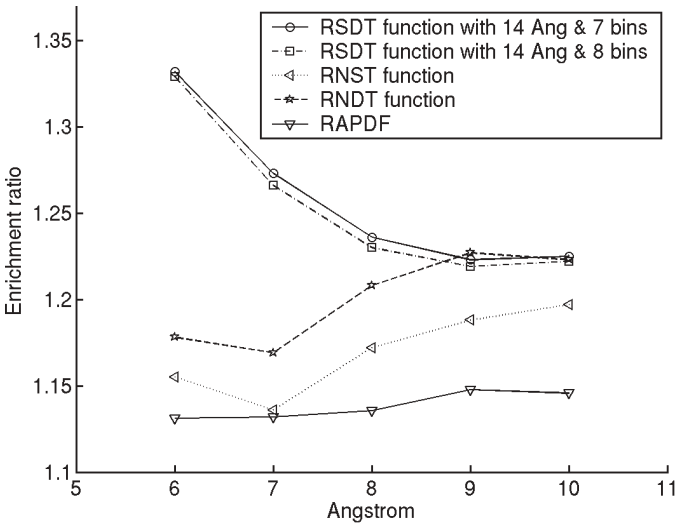

Fig. 5. Performance of the various types of residue triplet scoring functions. Triplet functions are evaluated using the average enrichment ratios on test decoy sets of various quality. For example, the circle at coordinate (6 Å,1.332) indicates that the RSDT function configured with a distance cutoff of 14 Å and 7 distance bins achieves an average enrichment ratios of 1.332 for the test decoy sets that contain structures of less than 6 Å C_α RMSD relative to the native conformations. From Figure 3a, we select the best performing RNST scoring function. The left-pointing triangles in the current figure indicate the average enrichment ratios achieved by that function. The best performing RNDT scoring function is analogously chosen from Figure 4a, represented by the stars in current figure. We also include the performance of one other scoring function in the figure. The downward pointing triangles correspond to the all-atom distance-dependent conditional probability discriminatory function, a two-body potential. Overall, the RSDT functions give the best performances.

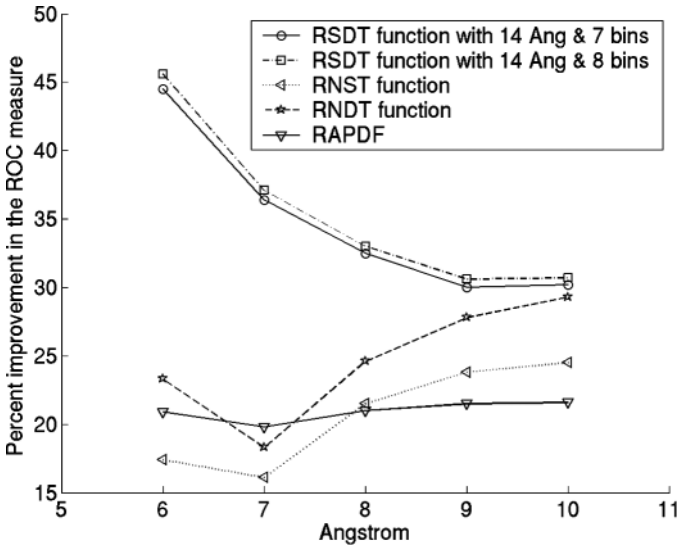

Fig. 6. Performance of the various types of residue triplet scoring functions. Triplet functions are evaluated using the average ROC percent improvements on test decoy sets of various quality. For example, the circle at coordinate (6 Å,44.5%) indicates that the RSDT function configured with a distance cutoff of 14 Å and 7 distance bins achieves an average percent improvement of 44.5% for the test decoy sets that contain structures of less than 6 Å C_α RMSD relative to the native conformations. From Figure 3b, we select the best performing RNST scoring function. The left-pointing triangles in the current figure indicate the average percent improvement achieved by that function. The best performing RNDT function is analogously chosen from Figure 4b, represented by the stars in the current figure. We also include the performance of one other scoring function in the figure. The downward pointing triangles correspond to the all-atom distance-dependent conditional probability discriminatory function, a two-body potential. Overall, the RSDT functions give the best performances.

function has been studied and compared with other functions in the literature [e.g. see Lu and Skolnick (2001), de Bakker *et al*. (2003) and Zhang *et al*. (2004)]. In the present study, this function is compiled with the database of the 3150 solved structures. The resulting enrichment ratios and percentage improvements in the ROC measure for the RAPDF function are shown in Figures 5 and 6, respectively. These figures show that the residue triplet functions with the configuration of a distance cutoff of 14 Å with 7 bins and of a distance cutoff of 14 Å with 8 bins both perform reasonably well in comparison.

In addition, we also apply a local-triplet (LT) scoring function described in Lezon *et al*. (2004) to the test decoy sets. The LT function uses a specially designed five-letter alphabet to represent the Ramachandran angles and evaluates a given decoy with a two-step process, in which a sequence–structure and a structure–structure mapping of the LTs are performed. It has been shown to have produced good results in the fold recognition of coarse-grained protein tertiary structures. In the present study, for the high quality test decoy sets, this function yields an average enrichment ratio of 1.10 and an average ROC percent improvement of 18.1%. Comparing these results with Figures 5 and 6 again confirms that the residue triplet functions with the configuration of a distance cutoff of 14 Å with 7 bins and of a distance cutoff of 14 Å with 8 bins perform well.

Examination of low counts

In order for the posterior probabilities $P(r_{abc}^{ijk}|C)$ estimated with Equation (5) and the prior probabilities $P(r_{abc})$ estimated with Equation (6) to be statistically meaningful, there needs to be sufficient counts for the denominator $\sum_r N(r_{abc})$ for each residue triplet type (a,b,c). Our results indicate that for the RSDT function with a distance cutoff of 14 Å and 7 distance bins, in the posterior probabilities estimation based on the database of the solved structures, the triplet type tryptophan–tryptophan–tryptophan has the count of 4717, the lowest among all triplet types. With 7 distance bins, this gives an average of ∼674 counts per bin. For the prior probabilities estimation based on the database of the decoy structures, the triplet type tryptophan–tryptophan–tryptophan has the count of 48 177, again the lowest among all triplet types. With 7 distance bins, this gives an average of ∼6882 counts per bin. Thus, in both cases, the counts are sufficiently high for Equations (5) and (6) to provide statistically valid estimates of the respective probabilities. Similar low count results are also obtained for the RNST and RNDT functions.

Conclusion

In this study, we construct and analyze a residue triplet knowledge-based scoring function. The scoring function is inspired by the previous work of Banavar and colleagues, who studied chain folding using a physical/geometric approach in which the inputs to their Lennard-Jones type potential were the radii of curvature of residue triplets. Their computer simulations showed a number of interesting results, e.g. naturally obtaining ground states with protein-like local structures, such as helices with specific pitch-to-turn ratio, sheets and hairpins.

Our formulation of the residue triplet scoring function follows the standard approach used in constructing the pairwise residue distance-dependent potential, with two modifications: (i) the two-body potential is replaced by a three-body one and (ii) the pairwise distances are replaced by the radii of curvature corresponding to residue triplets. Three different approaches for estimating the prior distribution of the radius of curvature are tested. Also tested are the use of various distance cutoffs and numbers of bins in constructing the knowledge-based potential. To evaluate the performances of the various possible configurations, we generate 41 test decoy sets of different quality and apply the various configurations of the scoring function on the test decoy sets. Our numerical experiments show that a distance cutoff of 14 Å, with either 7 or 8 distance bins and with the statistics of the prior distribution of the radius of curvature derived from a database of decoy structures in a residue type specific manner, produces good results.

We discuss briefly some possible modifications and extensions to the current form of the residue triplet scoring function. First, instead of using a straight 14 Å distance cutoff across the different residue types, the distance cutoff can be chosen in a residue type specific manner. That is, for given residue triplets of specific residue types a, b and c, one can compile the statistics and observe the log-odd score $S(r_{abc})$ of such a triplet type as a function of the radius of curvature r_{abc}. A good cutoff value for the triplet type will correspond to the radius of curvature at which this function decays to zero. Second, the residue-based function can be augmented to an all-atom form. Using a detailed atomic description for protein confirmations may yield a more accurate scoring function for discriminating native-like from non-native conformations. Third, as suggested in Banavar *et al*. (2003b), residue quadruplets instead of triplets can be used to construct an analogous scoring function. In such a case, the radius of curvature will be replaced by the radius of the sphere formed by four residues. Both the second and the third extensions require increased computing power, but they are still computationally tractable for small proteins with sizes <120 residues. Finally, we note that in the residue triplet formulation, a large radius of curvature can be generated either by three neighboring residues subtending an angle close to 180°, or by three residues distant from one another and forming an equilateral triangle. The fact that the two configurations are not distinguishable in the triplet formulation suggests that it is beneficial to combine the residue triplet scoring function with a two-body distance-based scoring function to further enhance the decoy discrimination ability. A detailed study of how to combine the residue triplet function with other potentials will be presented elsewhere.

Acknowledgements

The authors thank all the members of the Samudrala Group and the anonymous reviewers for their insightful suggestions for improving the content of the manuscript. The authors also thank Mr Tim Lezon and Prof. Jayanth Banavar for their helpful discussions and for providing data files and programs for implementing their scoring function. This work is supported in part by a Searle Scholar Award, a NSF CAREER award, a NSF grant DBI-0217241 and a NIH grant GM068152-01 to R.S., as well as the University of Washington's Advanced Technology Initiative in Infectious Diseases. Funding to pay the Open Access publication charges for this article was provided by the Searle Award.

References

Bajorath,J., Stenkamp,R., Aruffo,A. (1994) *Protein Sci.*, **2**, 1798–1810.
Banavar,J.R., Maritan,A., Micheletti,C. and Trovato,A. (2002) *Proteins*, **47**, 315–322.
Banavar,J.R., Flammini,A., Marenduzzo,D., Maritan,A. and Trovato,A. (2003a) *ComPlexUs*, **1**, 4–13.
Banavar,J.R., Gonzalez,O., Maddocks,J.H. and Maritan,A. (2003b) *J. Stat. Phys.*, **110**, 35–50.
Blundell,T.L., Sibanda,B.L., Sternberg,M.J.E. and Thornton,J.M. (1987) *Nature*, **326**, 347–352.
Bourne,P.E. *et al.* (2004) *Nucleic Acids Res.*, **32**, D223–D225.
Brooks,B., Bruccoleri,R., Olafson,B., States,D., Swaminathan,S. and Karplus,M. (1983) *J. Comput. Chem.*, **4**, 187–217.
Chandonia,J.M., Hon,G., Walker,N.S., LoConte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) *Nucleic Acids Res.*, **32**, D189–D192.
Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M.Jr, Fergusson,D.M., Spellmeyer,D.C., Fox,D.C., Caldwell,J.W. and Kollman,P.A. (1995) *J. Am. Chem. Soc.*, **117**, 5179–5197.
de Bakker,P.I.W., DePristo,M.A., Burke,D.F. and Blundell,T.L. (2003) *Proteins*, **51**, 21–40.
DeBolt,S.E. and Skolnick,J. (1996) *Protein Eng.*, **8**, 637–655.
Friesner,R.A. and Gunn,J.R. (1996) *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 315–342.
Gilis,D. and Rooman,M. (1996) *J. Mol. Biol.*, **257**, 1112–1126.
Hung,L.H. and Samudrala,R. (2003) *Nucleic Acids Res.*, **31**, 3296–3299.
Jernigan,R.L. and Bahar I. (1996) *Curr. Opin. Struct. Biol.*, **6**, 195–209.
Johnson,M.S., Srinivasan, N., Sowdhamini,R. and Blundell,T.L. (1994) *Crit. Rev. Biochem. Mol. Biol.*, **29**, 1–68.
Jones,D.T. (1997) *Curr. Opin. Struct. Biol.*, **7**, 377–387.
Jorgensen,W. and Tirado-Rives,J. (1988) *J. Am. Chem. Soc.*, **110**, 1657–1666.
Levitt,M., Gerstein,M., Huang,E., Subbiah,S. and Tsai,J. (1999) *Annu. Rev. Biochem.*, **66**, 1368–1372.
Lezon,T., Banavar,J.R. and Maritan,A. (2004) *Proteins*, **55**, 536–547.
Lu,H. and Skolnick,J. (2001) *Proteins*, **44**, 223–232.
MacKerell,A.D. Jr *et al*. (1998) *J. Phys. Chem. B*, **102**, 3586–3616.
Maritan,A., Micheletti,C., Trovato,A. and Banavar,J. (2000) *Nature*, **406**, 287–290.
Moult,J., Hubbard,T., Bryant,S.H., Fidelis,K. and Pedersen,J.T. (1997) *Proteins*, **29**, 2–6.
Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) *Proteins*, **37**, 2–6.
Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2001) *Proteins*, **45**, 2–7.
Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) *Proteins*, **53**, 334–339.
Nemethy,G., Gibson,K.D., Palmer,K.A., Yoon,C.N., Paterlini,G., Zagari,A., Rumsey,S. and Scheraga,H.A. (1992) *J. Phys. Chem.*, **96**, 6472–6484.
Sali,A. (1995) *Curr. Opin. Biotech.*, **6**, 437–451.
Samudrala,R. and Levitt,M. (2002) *BMC Struct. Biol.*, **2**, 3–18.
Samudrala,R. and Moult,J. (1998) *J. Mol. Biol.*, **275**, 895–916.
Samudrala,R., Xia,Y., Levitt,M. and Huang E.S. (1999) In Altman,R., Dunker,K., Hunter,L., Klein,T. and Lauderdale,K. (eds), *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific Press, Singapore, pp. 505–516.
Sanchez,R. and Sali,A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 206–214.
Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) *J. Mol. Biol.*, **268**, 209–225.
Sippl,M. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.
Weiner,S., Kollman, P., Nguyen,D. and Case,D. (1986) *J. Comput. Chem.*, **7**, 230–252.
Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) *Nucleic Acids Res.*, **31**, 489–491.
Wodak,S. and Rooman,M. (1993) *Curr. Opin. Struct. Biol.*, **3**, 247–259.
Zhang,C., Liu,S. and Zhou,Y. (2004) *Protein Sci.*, **13**, 391–399.
Zhang,C., Vasmatzis,G., Cornette,J.L. and DeLisi,C. (1997) *J. Mol. Biol.*, **267**, 707–726.

Received August 23, 2005; revised December 30, 2005; accepted January 9, 2006

Edited by Janet Thornton