PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles

Katrin Wolff,[1] Michele Vendruscolo,[2] and Markus Porto[1]*

[1] Institut für Festkörperphysik, Technische Universität Darmstadt, 64289 Darmstadt, Germany
[2] Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

ABSTRACT

One of the major bottlenecks in many ab initio protein structure prediction methods is currently the selection of a small number of candidate structures for high-resolution refinement from large sets of low-resolution decoys. This step often includes a scoring by low-resolution energy functions and a clustering of conformations by their pairwise root mean square deviations (RMSDs). As an efficient selection is crucial to reduce the overall computational cost of the predictions, any improvement in this direction can increase the overall performance of the predictions and the range of protein structures that can be predicted. We show here that the use of structural profiles, which can be predicted with good accuracy from the amino acid sequences of proteins, provides an efficient means to identify good candidate structures.

Proteins 2010; 78:249–258.

Key words: protein structure prediction; protein structural profiles; artificial neural network; position specific scoring matrices (PSSMs); critical assessment of structure prediction (CASP).

INTRODUCTION

The quality of methods to predict the structures of proteins from their amino acid sequences has advanced considerably over the last few years, as demonstrated clearly by the steadily improving quality of the results of the periodic community-wide critical assessment of structure prediction (CASP) exercise.[1–5] If sequence similarity to proteins of known structure can be detected, homology modeling can provide high-accuracy predictions.[6] When this strategy is not readily applicable, the structure should be predicted ab initio, a task that is often much more difficult.[7] Currently, the most widely used methods to perform this type of prediction use a molecular fragment replacement approach.[3,8,9] The first step in this approach consists of the homology-based generation of structural fragments, which are then assembled into low-resolution candidate structures.[10] In the subsequent step, a high-resolution structural refinement in many cases achieves predictions of great accuracy.[3] The best predictions, however, require the high-resolution refinement of a large fraction of the low-resolution decoys, a procedure that demands very considerable computational effort. To reduce this effort, reliable methods for the selection of the best structures are needed. One option to improve the selection is to define better energy functions for scoring of low-resolution candidate structures.[11] A second option, which appears to be very promising, is to perform a clustering of the decoys by their pairwise root mean square distances (RMSDs) and then consider the largest clusters[8] or the clusters of lowest energy.[12] Clustering by distance matrices and identifying the cluster of lowest energy has also proved successful in the reconstruction of protein structures from highly approximate backbone torsion angles.[13] A problem related to the selection of decoy structures is the ranking of predicted protein structures[14] where scoring functions that are weakly funneled toward the native state can be improved considerably by taking correlations between decoy structures[15] into account. The inclusion of sparse experimental data such as NMR chemical shifts also substantially improves protein structure determination.[16,17]

Here, we show that the use of structural profiles[18,19] provides an effective way for selecting candidate structures. Structural profiles corresponding to the native states of proteins can be computed from structures and used to efficiently compare them[20] or to analyze protein folding dynamics.[21] Most importantly in the context of this study, however, the structural profile of the native state of a protein can also be predicted with good accuracy from its amino acid sequence, for example using an artificial neural network trained on other protein structures and sequences, similarly to secondary

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Deutscher Akademischer Austauschsdienst; Grant number: D/08/08872; Grant sponsor: The British Council; Grant number: ARC 1319.
*Correspondence to: Markus Porto, Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstraße 6, 64289 Darmstadt, Germany. E-mail: porto@fkp.tu-darmstadt.de
Received 30 March 2009; Revised 26 June 2009; Accepted 30 June 2009
Published online 14 July 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22533

Table I
Number of Structures with RMSD Values Smaller Than One Standard Deviation Below the Mean N ($z_{RMSD} \leq -1$) for Various Selection Methods and Minimum RMSD in Å (in brackets)

PDBid	EC	predEC	Rosetta score	C_1	C^*	minRMSD
1pv0	72 (2.9)	26 (3.1)	10 (3.5)	0 (6.1, 199)	26 (3.7, 58)	2.6
1gb1	151 (1.6)	9 (1.7)	186 (1.5)	0 (3.8, 196)	91 (1.5, 91)	1.5
1shg	58 (3.6)	38 (4.0)	7 (4.9)	0 (6.9, 209)	0 (7.1, 49)	3.3
1jic	50 (3.8)	23 (5.8)	4 (6.4)	0 (10.9, 200)	0 (10.7, 164)	3.1
1r69	37 (1.6)	15 (2.2)	34 (2.3)	0 (3.0, 191)	1 (3.0, 70)	1.6
1c9oA	143 (2.9)	89 (2.9)	53 (3.2)	201 (2.8, 201)	0 (5.6, 104)	2.8
1mjc	127 (2.8)	80 (2.8)	138 (2.9)	166 (4.3, 201)	166 (4.3, 201)	2.8
1fgp	102 (5.9)	25 (9.4)	28 (8.7)	0 (10.7, 196)	21 (9.8, 185)	5.9
1ubq	119 (2.4)	97 (2.7)	78 (2.7)	26 (3.5, 196)	0 (4.6, 42)	2.3
1oqp	76 (4.1)	29 (4.2)	51 (4.2)	84 (4.5, 200)	0 (5.4, 70)	3.7
1btb	70 (3.7)	87 (3.4)	109 (3.1)	186 (6.2, 200)	174 (3.1, 174)	3.1
1p9yA	40 (5.7)	30 (6.1)	17 (5.4)	27 (8.1, 198)	25 (6.2, 198)	4.7

For the clustering method, the second number in brackets is the cluster size. The largest cluster is denoted as C_1, the cluster of lowest average Rosetta score as C^*. The last column gives the overall minimum RMSD in Å. For 1mjc C_1 and C^* coincide, for 1p9yA there are two equally large clusters C_1 one of which is also C^*. In every line the best method, disregarding the case of the exact EC, is underlined, which is usually the method achieving highest N ($z_{RMSD} \leq -1$). For 1mjc and 1oqp, the Rosetta score provides the best results over C_1 and for 1btb C^* over C_1 because the overall RMSD distributions are better, for 1r69 the case is left undecided between predicted EC and Rosetta score. The proteins are sorted by length.

structure prediction.[22] Prediction of structural profiles is significantly easier than the prediction of contact maps which is a notoriously difficult task.[23] The profile computed from each candidate structure in the decoy set can then be compared to the predicted target profile and structures with similar profiles are selected for refinement. By investigating both exact and predicted target profiles for structure selection, we demonstrate that the use of structural profiles constitutes a powerful method to reduce the computational effort in structure prediction.

METHODS

Generation of decoys

For each protein whose structure is to be predicted, we constructed 10,000 candidate structures using a library of fragments in a reduced representation, which included backbone and C_β atoms, using the standard Rosetta ab initio protocol[8] and excluding sequence homologs for the generation of fragments. These decoy generations run for 1–3 days per protein on a modern desktop computer. Although we discuss specifically the case of Rosetta, the results that we present in terms of decoy selection are expected to apply also to other methods for the generation of decoys.

Benchmark structures

The protein structures used for testing are listed in Table I (Supporting Information), and their lengths range from 46 to 209 amino acids. These lengths are taken from the corresponding PDB files, and, for 1shg, 1r69, 1p9yA, and 1gk9, differ from those in the FASTA files obtained from the PDB website. For fragment prediction and assembly, sequences were shortened at the termini to match the lengths and sequences of PDB structures. The distance of decoys to native structures was measured by C_α-RMSD and TM-score[24] and the distribution of RMSD values was also used to assess the quality of different selection methods.

CASP8 server predictions

To assess both longer protein structures and structures generated by different methods, we also applied our filtering scheme to models (server-only ab initio predictions) submitted to CASP8.[25] We discarded targets for which the experimental PDB structure was incomplete or ruptured (but allowed for possibly missing end termini) to simplify comparison to predictions. Of the 69 target proteins we thus considered only 29. A list of target proteins is given in Table III (Supporting Information) with lengths ranging from 69 to 533.

Prediction of the structural profiles

The structural profile used in this study is the effective connectivity (EC), which is defined as a linear combination of the eigenvectors of the contact map of the native state.[19] This definition requires the knowledge of the native structure and cannot be used for ab initio structure prediction. The structural profile can, however, be predicted from the amino acid sequence with good accuracy by using feed-forward artificial neural networks (ANN). To improve the predictions, the ANN does not use as input the single sequences whose profiles are to be predicted, but rather follows an approach developed in the context of secondary structure prediction[22] and takes evolutionary information into account by first obtaining position specific scoring matrices (PSSMs) using PSI-

BLAST.[26] The ANN consists of an input layer of 15 × 21 neurons, a hidden layer of 40 neurons, and a single output neuron. For each amino acid in the sequence, information from a sequence window of 15 amino acids centered at our residue in question enters the ANN and each place within this window defines probabilities for the 20 amino acid kinds that might occur at this position (PSSM). An extra neuron is set to 1 (and all others to 0) if the window extends the sequence (hence 15 × 21 input neurons). The activation function of the hidden layer is the hyperbolic tangent and the output is linear. The output neuron then gives the predicted profile entry for the central amino acid of the window. The ANN was trained on a representative subset of the PDB of 300,000 residues in total to minimize the squared differences between exact EC and prediction with early stopping to avoid overfitting. In this training, no difference in prediction quality could be observed depending on the inclusion or omission of sequence homologs.

Filtering of decoys

One way to narrow down the set of decoy structures to be considered for high-resolution refinement is to use the same score as in the decoy generation—here the standard low-resolution score of Rosetta[8]

$$E_{\text{score}} = E_{\text{env}} + E_{\text{pair}} + E_{\text{vdw}} + E_{\text{hs}} + E_{\text{ss}} + E_{\text{sheet}} + E_{\text{r-sigma}}. \quad (1)$$

This equation contains terms for residue interaction with environment (solvation) E_{env}, residue pair interaction E_{pair}, and van der Waals-interaction for steric repulsion E_{vdw}. The remaining terms, E_{hs}, E_{ss}, E_{sheet}, and $E_{\text{r-sigma}}$, account for packing of secondary structure elements. Using this filter the x% structures of lowest score are selected ($x \leq 2$).

The filtering score based on the structural profile for decoy j is defined as

$$\Delta_{\text{EC}}^{(j)} = \sum_{i=1}^{N} \left| t_i - c_i^{(j)} \right|^{\alpha}, \quad (2)$$

where the index i runs over all protein residues, $\mathbf{c}^{(j)}$ denotes the EC profile computed from candidate structure j in the decoy set and $\mathbf{t}$ the target EC (either predicted or computed from native structure) and $c_i^{(j)}$ and t_i are their respective vector entries. The exponent α is set to 2, but varying it between 0.5 and 4 makes hardly any difference to the filtering. Again, those structures are selected that score among the top x% of the entire set ($x \leq 2$).

The results depend on the choice of the contact threshold r_c for the contact map—either directly when computing structural profile from contact map or indirectly in the training of the ANN. Both prediction quality of ECs and filtering quality depend on this parameter but the dependence varies for different proteins. For the results reported here, r_c = 8.5 Å has been used as distance threshold between C_α atoms.

The quality of the prediction of the EC is measured by the Pearson's correlation coefficient ρ_c between predicted and exact ECs and may vary for different proteins. To obtain comparable data for different proteins and to assess the quality of the selection procedure as a function of the quality of the predictions, we simulated different sets of fixed correlations by linearly interpolating between predicted profiles and profiles obtained from native structures.

Clustering of decoys

Clustering by RMSD is used to find representatives of clusters of similar configurations and thus identify highly populated energy minima. This method is based on the idea that while the low-resolution energy score may not well preserve the depth of the native basin it may still preserve its width. Thus, configurations would be more densely sampled around the native basin which could be made visible by extracting large clusters of low pairwise RMSDs.

In order to compare this method to the filtering procedures described earlier, we aim at a largest cluster of 200 structures (2% of total set of 10,000). The clustering procedure used in this work involves five steps: (1) compute pairwise RMSDs between all configurations; (2) choose a RMSD threshold; (3) find the configuration with most neighbors (the largest cluster), i.e., with most other configurations within the RMSD threshold; (4) remove it together with its neighbors; (5) continue at step 3 considering the surviving configurations. Using this algorithm, the 10 largest clusters are extracted. The RMSD threshold is determined in a binary search to return a largest cluster of ~200 structures. The RMSD distribution of the largest cluster is then compared to those obtained by filtering. In a combination of two methods, the cluster of the lowest average energy score [Eq. (1)] is also determined among the 10 largest clusters. Additionally, the centers of all 10 largest clusters are compared to the native structure.

Assessment of selection quality

It is possible to consider several different measures to compare the performance of the various filtering methods. We tested two different distance measures, C_α-RMSD and TM-score (also based on C_α atoms).[24] The root mean square deviation (RMSD) is perhaps the most intuitive measure of protein similarity and adequate for closely related structures but carries less information for

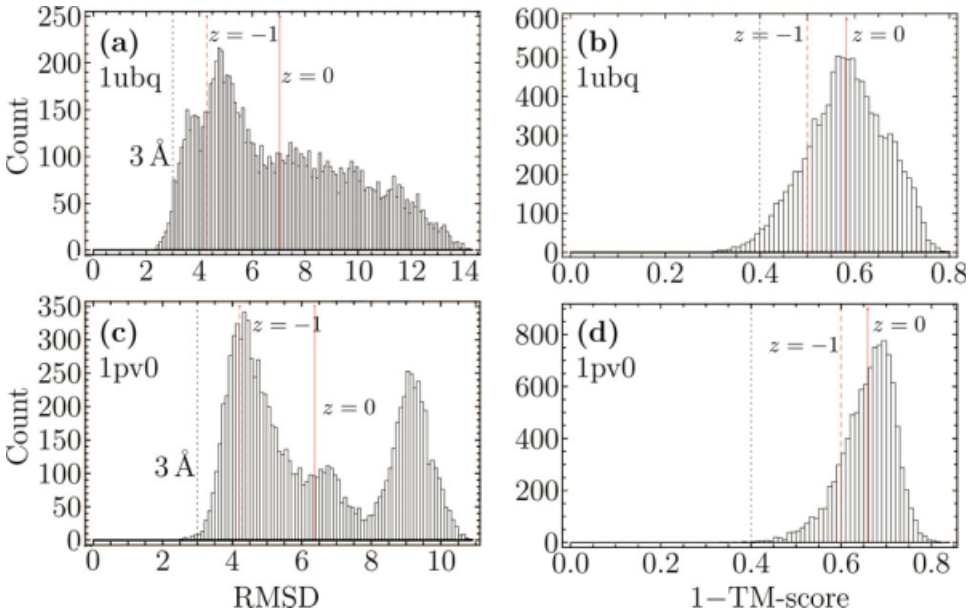

Figure 1

Distribution of RMSD **(a,c)** and 1 – TM-score **(b,d)** of decoys compared to target structure for (a,b) ubiquitin (PDBid 1ubq, length 76) and (c,d) Sda antikinase (PDBid 1pv0, length 46). Solid lines indicate the mean of the RMSD distributions, and dashed lines one standard deviation below the mean and dotted line 3 Å or TM-score 0.6. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

dissimilar structures.[27] The TM-score does not suffer from this disadvantage and can detect even quite weak similarities between structures. Another question is how to compare distributions (of either RMSD or TM-score) as selected by different filtering methods. Although in the following we refer in most places to the RMSD values, the results obtained by using the TM-score are essentially equivalent.

The entire distribution of the selected structures carries most information but is also difficult to compare quantitatively. The average RMSD (or average TM-score) to the native structure is also not very suitable as it is most important to find the few structures of very low RMSD (high TM-scores) values even if some bad structures are contained in the selection. The structure with the lowest RMSD value (highest TM-score) in the selection and its position in the internal ranking of structures by the scoring function may be more suited but is strongly afflicted by chance. We suggest that the number of structures in the selection below a certain RMSD (or above a certain TM-score) threshold is most informative; in our experience this threshold should be at about 3–4 Å (or a TM-score threshold of 0.6). However, distributions differ for different proteins and for some the decoy set do not contain such structures. Therefore, we choose a protein-dependent threshold and all structures with an RMSD one standard deviation lower than the mean RMSD ($z_{\mathrm{RMSD}} < -1$) (or mean TM-score, $z_{\text{TM-score}} < -1$) are considered as good structures.

RESULTS AND DISCUSSION

RMSD and TM-score distributions

For most small (i.e., less than 200 amino acids) proteins, the RMSD values of the decoys to target structures were found to range from 2.5 to 14 Å and TM-score from 0.2 to 0.8; we show two typical cases in Figure 1, 1ubq and 1pv0 (for the remaining structures see Supporting Information). Note that the *x*-axes of Figure 1(b,d) show 1 – TM-score such that, in both RMSD and TM-score plots, "good" decoys are sorted toward the left and that *x*-axes are to different scales. The dotted line indicates an RMSD of 3 Å and TM-score of 0.6 for comparison of different proteins, the dashed (red) line denotes one standard deviation below mean as a measure of relatively good decoys within a given decoy set.

Notable exceptions in decoy distributions are found for 1p9yA with some structures of very high RMSD (ca. 25 Å) and 1r69 where the entire RMSD distribution is fairly good and TM-scores even reach values of 0.9. In the latter case, a random selection might be suitable for refinement to high-resolution structures. It has to be noted that 1p9yA is a notoriously difficult case where two distant β-hairpins have to be in proximity, 1r69 on the other hand was used for calibration of the Rosetta score[28] explaining the very good decoy distribution.

The larger proteins of about 200 amino acids have RMSD distributions ranging from 7 to 30 Å, some (1ix9 and 1gk9) even without any decoys below an RMSD of

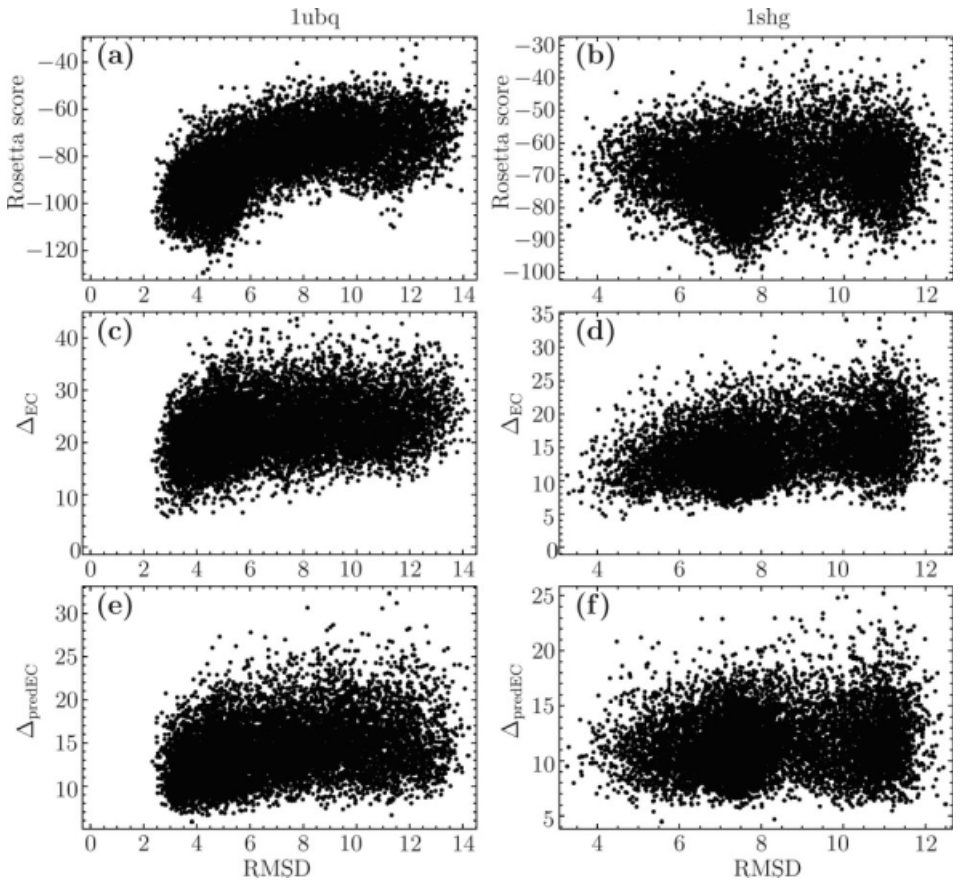

Figure 2

(**a,b**) Scatter plots of the Rosetta score, Eq. (1), as a function of RMSD values for (a) ubiquitin (1ubq, length 76) and (b) SH3 domain (1shg, length 57). (**c,d**) Scatter plots of Δ_{EC}, Eq. (2), as a function of RMSD values for (c) ubiquitin (1ubq) and (d) SH3 domain (1shg). (**e,f**) Scatter plots of Δ_{predEC}, also Eq. (2) but for predicted target profile, as a function of RMSD values for (e) ubiquitin (1ubq) and (f) SH3 domain (1shg).

10 Å, and TM-score distributions between 0.1 and 0.5, for some proteins even only up to 0.4. For these larger proteins, decoys were too distant from native structures to allow any meaningful comparison of filtering methods. Even when using TM-score as a distance measure, which is more sensitive to distant similarities than RMSD, filtering methods were unable to detect these weak signals. We therefore estimated the number of decoys necessary to expect one structure with RMSD $\leq$ 5 Å by approximating distributions as Gaussian (see Supporting Information) with mean and variance calculated from decoy sets. The results (see Supporting Information) show that up to 10^{12} structures would be required for 1ix9. As this number is too large, we chose to additionally test our filtering method on CASP8 models where good predictions of longer proteins are on-hand (see below).

Selection of structures by scoring functions

The correlation between the scoring function and the distance from the native structure and the funneling toward the native structure are all important features in a scoring function. The low-resolution energy scoring function of Rosetta is usually correlated to the RMSD values [see Fig. 2(a) for example protein 1ubq] but funneling toward the lowest RMSDs would require very extensive sampling and the decoy set to contain very near native structures. Instead, in Figure 2(a) there appears to be a small funnel toward structures of about 4 to 5 Å even though structures of 2 to 3 Å RMSD to the native structure exist in the decoy set. In other cases [see Fig. 2(b) for example protein 1shg], the scoring function fails more dramatically and leads to very different structures. The scoring function Δ_{EC} on the other hand appears more reliable in the scatter plots and always selects structures of very low RMSD among the best scoring structures [see Fig. 2(c) for 1ubq and 2(d) for 1shg]. Scatter plots for predicted structural profiles, Figure 2(e, f), again show less funneling than plots for exact profiles but do not lead to false minima either. Scatter plots with TM-score instead of RMSD are qualitatively similar (data not shown).

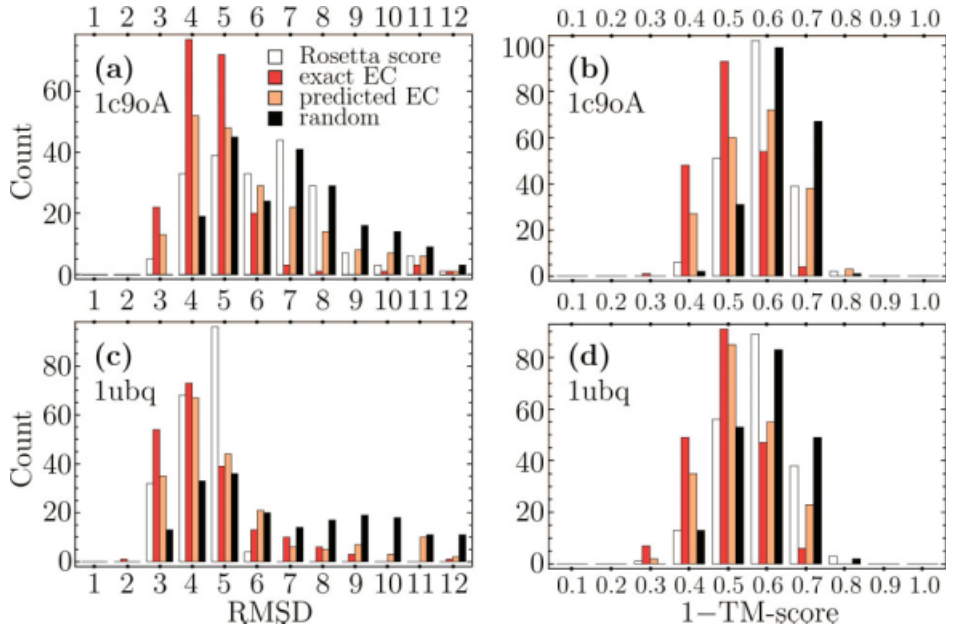

Figure 3

Comparison of RMSD **(a,c)** and 1 – TM-score **(b,d)** distributions of structures selected by EC, predicted EC, and Rosetta score for (a,b) chain A of cold shock protein (1c9oA, length 66) and (c,d) ubiquitin (1ubq, length 76). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 3 shows two examples of RMSD and TM-score distributions after selection of 200 structures that have the best ranking in terms of energy [Eq. (1)] or profile comparison [Eq. (2)]. Filtering by using the exact profile provides better results than filtering by Rosetta score in all but two cases (see Table I)—1btb and 1gb1. The performance of the two filtering methods is close for 1gb1 (especially if less than 200 structures are selected, see Supporting Information) and also for 1r69 and 1mjc where the outcome depends on the number of structures selected. Even though here filtering has been performed using the exact profile, which is not available for unknown structures, the results show the great potential of this filtering method. Filtering by predicted EC is comparable to filtering by energy score [e.g., in Fig. 3(c,d), even better in Fig. 3(a,b)]. Although the shape of distributions may vary between RMSD and TM-score, ordering of filtering methods remains unchanged. In the following plots and tables, discussing small protein structures, we therefore restrict ourselves to results for RMSD.

In eight of twelve cases, the exact EC is better than the energy score and in six of these cases the predicted EC is also better (as measured by number of good structures $N(z_{\mathrm{RMSD}} \leq -1)$, see Table I). Importantly, the great strength of filtering by EC lies with ranking structures of very low RMSD near the top (see Table I, numbers in brackets). Figure 4 shows for two examples how the number of good structures increases with the number of top structures as selected by the different filters. It appears that the curves for filtering by profile flatten with increasing number of structures while the curve for the energy score catches up [especially Fig. 4(b)]. This is important if, due to limited resources, only a small percentage of structures is to be selected for refinement. The black triangles denote the number of good structures expected in a random sample, $N(\mathrm{selected})\frac{|G|}{|A|}$, where G is the set of all good structures in the decoy set and A the entire decoy set. It holds $N(z_{\mathrm{RMSD}} \leq -1) = |G \cap S_f|$ with S_f the set of structures selected by filter f (EC, predicted EC, and Rosetta score).

Table II reports the correlation of predicted and exact EC to investigate the dependance of filtering quality on prediction quality. We note that the correlation coefficients observed for our test set are quite representative for the prediction quality in general and, while varying considerably, do not systematically deteriorate with increasing length of proteins. (The correlation coefficients for the protein structures of lengths larger than 200 amino acids vary from 0.58 to 0.78.) Columns 3 and 4 show the performance (number of good structures and lowest RMSD) for simulated "predictions," that is linear interpolations between exact and predicted EC to give arbitrary prediction qualities. Filtering by profile is performed using Δ_{EC} from Eq. (2), not ρ_c so, arguably this might be a better measure of prediction quality in our case. However, the correlation coefficient ρ_c is what is typically reported. We adhere to this convention and show filtering performance for fixed correlation to exact

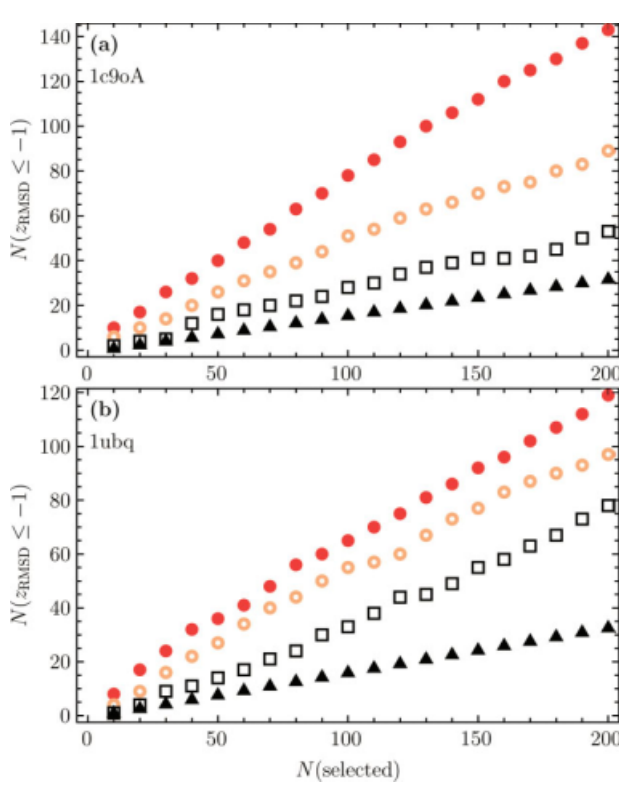

Figure 4

Comparison of good structures among those selected by EC (filled circles), predicted EC (empty circles), and Rosetta score (squares) **(a)** for chain A of cold shock protein (1c9oA, length 66) and **(b)** ubiquitin (1ubq, length 76). Triangles denote expectation values for random sample. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

profiles. In a separate test we used the normalized distance $\delta_{EC} = \Delta_{EC}/N$ between exact and predicted profile, where N is the length of the protein (data not shown). These results do not differ qualitatively from the results at fixed ρ_c shown here.

Table II shows that the dependence of filtering on prediction quality is not the same for different proteins. Protein 1gb1, for which filtering by predicted EC exhibited the worst performance, also has an exceptionally low correlation coefficient—but so does 1ubq for which filtering was quite good. A higher correlation to EC improves filtering performance for each protein, see Figure 5, (except for 1btb where filtering by predicted EC was better than by EC) and a correlation coefficient of $\rho_c = 0.8$ would put 1gb1 back among the others. Also for fixed ρ_c filtering performance varies considerably (even when measured in relation to exact EC) showing that prediction quality cannot be the only determinant of good filtering, protein-specific structure features or the structure distribution in the decoy set (and thus the sampling method) may also play a role. In particular, filtering by structural profiles compares favorably for decoy sets of lower quality. This is important as decoy sets deteriorate with increasing lengths of proteins whereas prediction quality of structural profiles does not systematically do so. If, however, the decoy sets are of poor quality (no structure below an RMSD of approximately 5 Å, as mentioned earlier), none of the filtering methods investigated is capable of extracting meaningful subsets.

Taken together these results indicate that improving structural profile predictors will help to further improve this method of filtering even if some structures are inherently more difficult than others.

Selection of structures by clustering

In the calculations that we performed, we could not find any systematic improvement when using the clustering method over the selection by Rosetta score or by the predicted profiles. Frequently, the largest cluster does not contain any good structures. Even in two of the cases where it apparently provides good results by higher $N(z_{RMSD} \leq -1)$ (1mjc and 1oqp) this may rather show the limits of the measure of selection quality used than actually mean better selection (see Table I). Visual inspection of the RMSD distribution (see Supporting Information) reveals that very good structures are missing in the largest cluster and high $N(z_{RMSD} \leq -1)$ is caused by structures just below the threshold of $z_{RMSD} = -1$. Only for protein 1c9oA [Fig. 6(b)] the largest cluster unexpectedly outperforms the other selection methods.

Using the cluster of lowest average Rosetta score instead of the largest one can give good results for those cases where the direct selection by score has also been successful (1btb, 1gb1). It is interesting to note that for 1btb clustering even improves the selection as compared to direct selection by the Rosetta score [Fig. 6(a)]. Both examples in Figure 6, however, are not representative for the performance of structure selection by clustering. While the RMSD threshold for clustering has been

Table II

Pearson's Correlation Coefficient ρ_c Between Predicted and Exact EC (Column 2), Interpolated EC (mEC) (Columns 3 and 4) with Fixed Correlation to Exact EC (0.8 resp. 0.9)

PDBid	ρ_c	mEC, $\rho_c = 0.8$	mEC, $\rho_c = 0.9$
1pv0	0.64	25 (3.1)	34 (3.1)
1gb1	0.30	67 (1.7)	102 (1.7)
1shg	0.62	41 (4.0)	43 (4.0)
1jic	0.72	22 (6.2)	27 (6.2)
1r69	0.54	21 (2.2)	25 (1.6)
1c9oA	0.80	90 (2.9)	116 (2.9)
1mjc	0.86	–	89 (2.8)
1fgp	0.35	37 (8.5)	49 (8.0)
1ubq	0.30	122 (2.5)	126 (2.5)
1oqp	0.69	32 (4.1)	39 (4.1)
1btb	0.48	87 (3.4)	82 (3.2)
1p9yA	0.69	30 (6.1)	33 (5.7)

Data reported are number of good structures $N(z_{RMSD} \leq -1)$ and lowest RMSD in the selection (in brackets). The proteins are sorted by length.

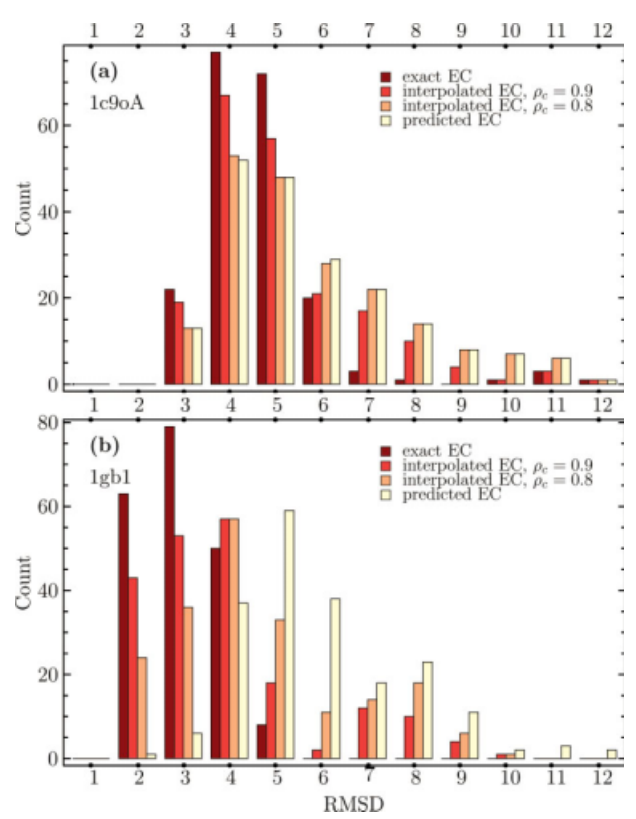

Figure 5

Comparison of RMSD distributions of structures selected by EC, predicted EC, and interpolations of the two profiles **(a)** for chain A of cold shock protein (1c9oA, length 66) and **(b)** immunoglobulin binding domain of protein G (1gb1, length 56). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

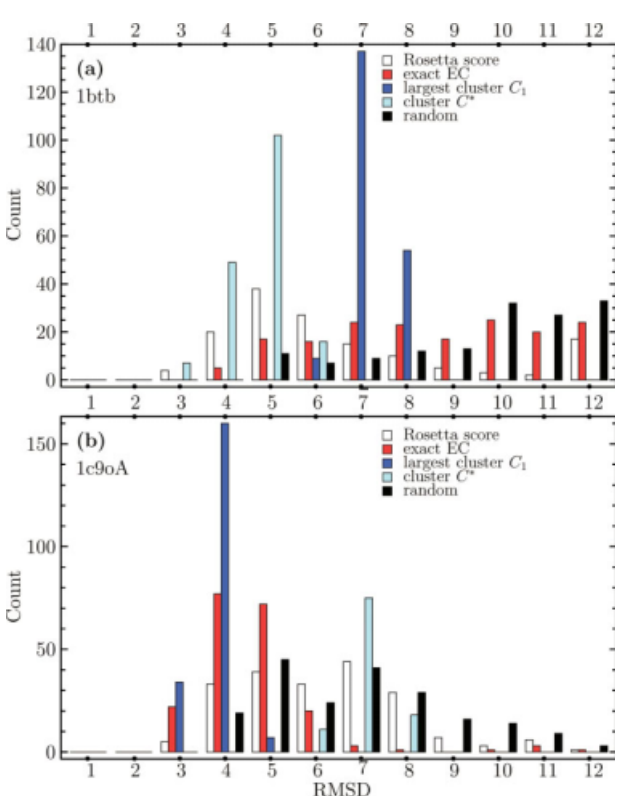

Figure 6

Comparison of RMSD distributions of structures selected by EC, Rosetta score, or clustering for **(a)** for barstar (1btb, length 89) and **(b)** chain A of cold shock protein (1c9oA, length 66).

chosen such that the largest cluster has ∼200 members and numbers $N(z_{\mathrm{RMSD}} \leq -1)$ are therefore comparable, clusters of lowest energy are usually smaller which has to be kept in mind when comparing absolute numbers.

We also compare the centers of the 10 largest clusters to the 10 top structures as ranked by EC or predicted EC, see Table III. Here, the number of good structures $N(z_{\mathrm{RMSD}} \leq -1)$ is not a good measure as only one cluster is expected to be really near-native. It is however significant information if no cluster center is among the good structures as is the case for four proteins. When using the structure of minimum RMSD as an alternative measure, selection by exact EC also outperforms the cluster centers in 10 cases and the predicted EC still wins in seven cases. Especially for those proteins where initial low-resolution sampling is poor (1p9yA, 1fgp) clustering fails whereas EC filtering may still identify relatively good conformations. Changing the clustering RMSD threshold such that the largest cluster contained 500 instead of 200 structures brought no significant changes. Hence, the hierarchical clustering approach is useful if the decoy set is known in advance to be very good but both Rosetta score and (especially) the predicted EC are more versatile in the sense that they are successful for good decoy sets but also tolerate decoy sets of only moderate prediction quality. Additionally, filtering by scoring function has the

Table III

Number of Good Structures N ($z_{\mathrm{RMSD}} \leq -1$) and Minimum RMSD in Å for Centers of the 10 Largest Clusters (Column 2) and 10 Structures of Lowest Δ_{EC} for Exact EC (Column 3) and Predicted EC (Column 4)

PDBid	C_{centers}	EC	predEC
1pv0	4 (3.5)	5 (2.9)	1 (3.8)
1gb1	3 (2.0)	10 (1.7)	0 (4.1)
1shg	0 (6.7)	9 (4.1)	3 (5.0)
1jic	0 (10.0)	6 (6.3)	2 (8.0)
1r69	1 (2.9)	1 (2.3)	2 (2.8)
1c9oA	5 (3.2)	10 (3.2)	6 (3.1)
1mjc	6 (3.8)	9 (3.2)	3 (4.6)
1fgp	0 (10.7)	8 (8.3)	1 (10.1)
1ubq	6 (2.5)	8 (2.7)	4 (3.4)
1oqp	3 (5.0)	6 (4.1)	2 (5.3)
1btb	3 (4.2)	6 (3.7)	6 (3.4)
1p9yA	0 (13.9)	3 (7.8)	2 (9.0)

The best method (not considering exact EC) is underlined in each row. The proteins are sorted by length.

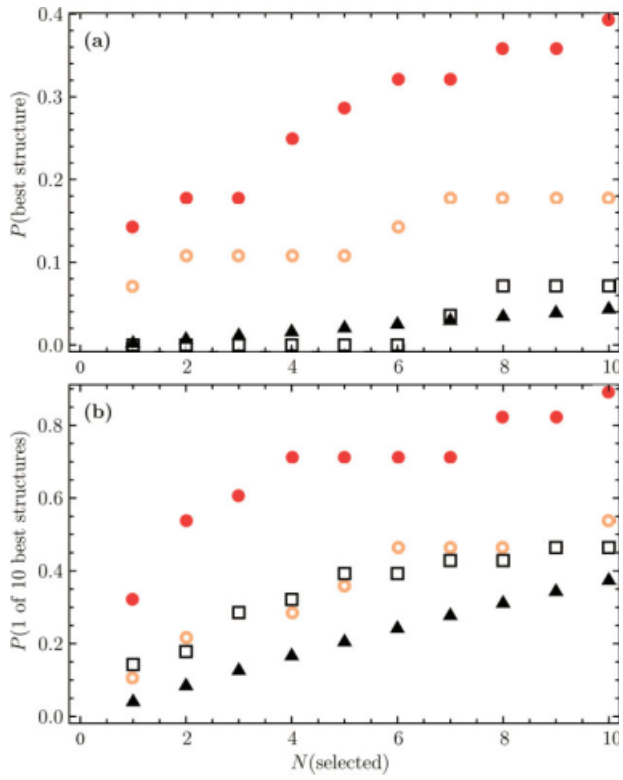

Figure 7

Comparison of relative frequency of selecting good structures by EC (filled circles), predicted EC (empty circles), and Rosetta score (squares) **(a)** if only single best structure counts as good or **(b)** any of 10 best structures. Triangles denote the probability of a random sample containing at least one good structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

advantage of being easily scalable—simply by varying the percentage of selected structures as in Figure 4.

CASP8 server models

In addition to the decoy sets generated by Rosetta, we also studied the server-only ab initio predictions of CASP8 to test our method on longer protein structures. For the 29 targets considered, 100 to 300 models per target are available, which we again ranked by Rosetta score, exact and predicted EC (the correlation coefficient between the latter two varies from 0.34 to 0.82). Although the CASP models were generated by various methods, we used the Rosetta energy as a scoring function to perform a consistent comparison.

Figure 7 shows the relative frequency of choosing at least one "good" structure for different filters and selecting 1 to 10 models per target, averaged over all 29 targets. The "random" curve gives the probability to do so by randomly sampling the models, taking into account the different numbers of models per target and subsequently averaging over all 29 targets. Clustering models by RMSD proved ineffective because of too small structure sets. In Figure 7(a) only the structure of very best TM-score is considered as good, in Figure 7(b) the 10 structures of highest TM-score. For this plot, TM-score was used instead of RMSD to measure model similarity to target as model quality differed considerably. Although, with lengths ranging from 69 to 533, not all of the CASP8 targets are longer than the structures considered before, we average over the entire set to get better statistics. Taking into account only the 20 structures longer than 120 amino acids yields a similar picture.

The results are consistent with those for smaller structures: The exact EC shows the best performance, and the predicted EC is comparable to Rosetta score but better if a stricter criterion of "good" structures is applied. It is however remarkable how well Rosetta score performed considering that models were not optimized using this score.

CONCLUSION

We have shown that the use of structural profiles provides an effective way to select near-native candidate structures in ab initio prediction methods. We have first provided a proof of principle by performing the selection by using exact profiles, and then demonstrated that the use of predicted profiles also offers results of very good quality. Furthermore, the results that we have presented concerning the selection by interpolated structural profiles indicate that the quality could be further improved by increasing the accuracy in the prediction of the structural profiles. We also wish to point out that predictions of structural profiles do not systematically deteriorate with increasing lengths of proteins whereas decoy sets do. If decoy sets contain at least some good structures, our approach is better than other filtering methods in coping with decoys of less quality while still being able to extract top structures from good sets. As shown by considering the structures from server-only ab initio predictions of CASP8, using predicted profiles for longer structures constitutes no bottleneck. As strategies for selecting conformations that exploit predicted structural profiles, such as those discussed in this work, provide results of good quality, it is worth considering them in standard ab initio prediction methods.

ACKNOWLEDGMENTS

The authors thank Jonas Minning for providing the predicted structural profiles for the proteins studied in this work and Andrea Cavalli for useful discussions.

REFERENCES

1. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61:225–236.

2. Kolinski A, Bujnicki JM. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins 2005;61:84–90.
3. Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
4. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins 2007;69:194–207.
5. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69:108–117.
6. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69:38–56.
7. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69:57–67.
8. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66.
9. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.
10. 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.
11. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 2001;30: 173–189.
12. Betancourt MR, Skolnick J. Finding the needle in a haystack: educing native folds from ambiguous ab initio protein structure predictions. J Comput Chem 2001;22:339–353.
13. Gong H, Fleming PJ, Rose GD. Building native protein conformation from highly approximate backbone torsion angles. Proc Natl Acad Sci USA 2005;102:16227–16232.
14. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins 2008;71:1175–1182.
15. Stumpff-Kane AW, Feig M. A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes. Proteins 2006;63:155–164.
16. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 2007;104:9615–9620.
17. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 2008;105:4685–4690.
18. Porto M, Bastolla U, Roman HE, Vendruscolo M. Reconstruction of protein structures from a vectorial representation. Phys Rev Lett 2004;92:218101.
19. Bastolla U, Ortiz AR, Porto M, Teichert F. Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. Proteins 2008;73:872–888.
20. Teichert F, Bastolla U, Porto M. SABERTOOTH: protein structural alignment based on a vectorial structure representation. BMC Bioinformatics 2007;8:425.
21. Wolff K, Vendruscolo M, Porto M. Stochastic reconstruction of protein structures from effective connectivity profiles. PMC Biophys 2008;1:5.
22. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
23. Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 2005; 7:180.
24. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
25. CASP8 website. Available at: http://predictioncenter.org/download_area/CASP8/server_predictions (accessed on June 15, 2009).
26. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
27. Wallin S, Farwer J, Bastolla U. Testing similarity measures with continuous and discrete protein models. Proteins 2002;50:144–157.
28. Rosetta forum. Available at: http://boinc.bakerlab.org/rosetta/forum_thread.php?id=1453#14335 (accessed on June 15, 2009).