A comparison of different functions for predicted protein model quality assessment

Juan Li[1] · Huisheng Fang[2]

Received: 20 February 2016 / Accepted: 8 July 2016 / Published online: 3 August 2016

Abstract In protein structure prediction, a considerable number of models are usually produced by either the Template-Based Method (TBM) or the ab initio prediction. The purpose of this study is to find the critical parameter in assessing the quality of the predicted models. A non-redundant template library was developed and 138 target sequences were modeled. The target sequences were all distant from the proteins in the template library and were aligned with template library proteins on the basis of the transformation matrix. The quality of each model was first assessed with QMEAN and its six parameters, which are C_β interaction energy (C_beta), all-atom pairwise energy (PE), solvation energy (SE), torsion angle energy (TAE), secondary structure agreement (SSA), and solvent accessibility agreement (SAE). Finally, the alignment score (score) was also used to assess the quality of model. Hence, a total of eight parameters (*i.e.*, QMEAN, C_beta, PE, SE, TAE, SSA, SAE, score) were independently used to assess the quality of each model. The results indicate that SSA is the best parameter to estimate the quality of the model.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9924-1) contains supplementary material, which is available to authorized users.

✉ Huisheng Fang
hsfang889@163.com

[1] Department of Hematology, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu 210008, People's Republic of China

[2] School of Life Science and Technology, China Pharmaceutical University, Nanjing, Jiangsu 210009, People's Republic of China

Keywords Protein structure prediction · QMEAN · Secondary structure agreement

List of abbreviations

C_beta	C_β interaction energy
PE	All-atom pairwise energy
SE	Solvation energy
TAE	Torsion angle energy
SSA	Secondary structure agreement
SAA	Solvent accessibility agreement
TBM	Template-based method
CASP	Critical assessment of protein structure prediction

Introduction

Predicting the three-dimensional structure of a protein based on its amino acid sequence is important for understanding protein function in several biological activities. The Critical Assessment of protein Structure Prediction (CASP) is the current gold standard for predictive protein modeling and classifies the prediction methods into three main approaches [1]: homology modeling [2, 3], threading [4–6], and ab initio predictions [7–9]. All three approaches generate a large number of candidate models, which are then assessed for quality.

Homology modeling and threading share many similarities and are known as Template-Based Modeling (TBM). These predictive methods use a protein structure library to identify a "best fit" target sequence to an unknown protein. In contrast, ab initio modeling conducts a conformational search under the guidance of a designed energy function. This energy function first calculates the energy of the searched conformation according to a certain force field (e.g., amber). If the calculated energy is low, the conformation is conserved; if the calculated energy is high, the conformation is abandoned. This procedure usually generates a number of possible conformations (structure decoys), and final models are selected among these [7–9].

Determining a predictive structure by TBM consists of four steps. First, known structures (templates) are searched for sequences closely related to the target sequence. Second, the target sequence is aligned to the template structure. These two steps are critical for TBM [10].Third, the structural framework is built by copying the structure of the template protein in the aligned regions or by satisfying the spatial restraints from templates (for example, the distance between two atoms cannot be less than 0.1 nm). Finally, the unaligned loop regions are constructed and side-chain atoms are added.

The quality of the generated protein model is dependent on the selected template [11, 12]. Usually, the final choice of template is determined by the statistical significance of the alignment score between the query (target protein) and the template sequences [13]. There are four categories that determine a good match between structure and sequence: (1) a significant sequence match with significant structural match, (2) a significant sequence match with non-significant structural match, (3) a non-significant sequence match with significant structural match, and (4) a non-significant sequence match with non-significant structural match [14]. This suggests the fact that a significant sequence match (*e.g.*, small E-value) does not necessarily result in structures similarities; on the other hand, non-significant sequence alignments might result in similarities in the protein structure. These facts tell us that the alignment score may not correlate with the chosen template's quality, especially for a distant homolog and, therefore, it is important to identify reliable scoring functions to determine the quality [15].

The disconnect between alignment score and template quality suggests that the current protein modeling methods may not necessary select the best model when searching a template library. To test this hypothesis, we searched the protein "d1bmta1" (SCOP database, http://scop.mrc-lmb.cam.ac.uk/scop/) against a non-redundant database in which one protein was selected for each family in the SCOP database. The MaxSub score (determined using Eq. 1) from the best model was 0.603, while the score for the best alignment was only 0.173 (Li J and Fang H, unpublished observations). MaxSub is defined as:

$$\mathrm{MaxSub} = \frac{\sum_{i=1}^{n} \frac{1}{1+\left(\frac{d_i}{d}\right)^2}}{q} \qquad (1)$$

and the full description of the formula is provided in [14].

The above data suggest that identifying the model closest to the native conformation (*i.e.*, its MaxSub score) will result in an accurate protein structure prediction. This should be the case for both the template-free approach [16, 17] and the template-alignment approach [18, 19]. There have been several studies that targeted the development of algorithms to address this issue (*i.e.*, estimating the quality of the predicted model) and these have been summarized by Benkert et al. [20, 21].

QMEAN [20], which uses a composite score consisting of a linear combination of six parameters, is one of the best prediction methods currently in use. The six parameters are C_β interaction energy (C_beta), all-atom pairwise energy (PE), solvation energy (SE), torsion angle energy (TAE), secondary structure agreement (SSA), and solvent accessibility agreement (SAA). A higher quality model can be derived from the comprehensive evaluation of the existing score parameters in QMEAN when compared to prediction methods using simply an alignment score or its derived statistics (*i.e.*, E value only).

In this study, we used QMEAN to assess the quality of the protein models generated by the method in our lab [22]. Since the six QMEAN parameters represent different features of a protein, we also used the individual parameter scores to separately assess the quality of the models. Interestingly, we found that SSA has a higher probability of predicting the quality of the protein model than the other five parameters.

Materials and methods

Datasets

Based on SCOP-1.67, we built a non-redundant template database composed of globular proteins. Each template protein in the database derived from different protein families present in the SCOP database. We selected one protein from each family; since there are 2160 families in SCOP-1.67, the database was composed of 2160 template proteins. Due to a large difference between any two selected proteins (generally, the sequence identities are <30 %) and the lack of redundancy, the database constructed was defined as non-redundant. One-hundred-thirty-eight query proteins distantly related to the template database were selected from distinct families for analysis. Moreover, the 138 proteins selected belong to the four classes of globular proteins (*i.e.*, all α, all β, α/β, α + β structures) [23].

Alignment and assessment of the models

Each query protein was aligned with the above template database on the basis of the BLOSUM62 transformation matrix with the dynamic program method; models were then generated according to the alignment. First, each query protein sequence was aligned with the sequence of the templated protein. The sequence identity was calculated as the number of identical residues (between the target sequence and the templated protein sequence) divided by the number of total residues of the shorter of two sequences.

Afterward, the backbone structure of the template protein, obtained after alignment with the target protein, was copied as the predicted structure of target protein. Since there are 2160 proteins in template database and, in theory, each alignment between the query sequence and template protein sequence can generate one model, 2160 models were generated. The alignment quality was calculated using the TMscore algorithm (http://bioinformatics.buffalo.edu/TMscore) [6], which compared the predicted model structure with its experimental structure and then produced MaxSub value.

Details of the TMscore algorithm are described in Li and Fang [22]. The alignment between template protein sequence and target protein sequence includes aligned region and gap (unaligned) region. The backbone structure of the target protein is achieved by copying the backbone structure of template protein into corresponding aligned regions. If a loop of the template protein is exactly in the alignment region (the sequence alignment region between the template protein and the predicted protein), that loop of the template protein could be the loop structure of the predicted protein.

Here, MaxSub is used to estimate the quality of the model because it is independent of the protein size. The quality of each model in this study was also assessed using QMEAN (http://swissmodel.expasy.org/qmean/cgi/index.cgi). For each model, the MaxSub score, the QMEAN score, and the scores from each of the QMEAN's six parameters were collected.

Results

Relationship between MAXSUB and C_beta, PE, SE, TAE, SSA, and SAA parameters; QMEAN and alignment score

Figure 1 shows the relationship between MaxSub and eight parameters (C_beta, PE, SE, TAE, SSA, SAA, QMEAN, and alignment score) for the randomly selected protein "d1bmta1". We determined that SSA had the strongest correlation with MAXSUB compared to the other parameters (Fig. 1e). Table 1 summarizes the results from the analysis of the eight parameters using a *t* test. The *t* value of SSA was significantly higher than that of the other seven parameters tested. These results indicate that SSA significantly correlates with MaxSub. Moreover, as shown in Table 1, the alignment score was not significantly different than that according to QMEAN. However, the correlation between these two parameters (alignment score and QMEAN) was significantly higher than of the other five QMEAN parameters (C_beta, PE, SE, TAE, and SAA). From the correlation coefficient results, SSA was significantly higher than the others.

MaxSub score of selected models

We then identified the near-native model for each target protein based on the MaxSub score of selected models. We investigated the performance of the aforementioned eight parameters using three criteria: the sum of MaxSub, the number of correct predictions [21], and the *t* value (Table 2). Table 2 shows that the sum of MaxSub scores based on SSA is the highest; followed by C_beta, PE, and QMEAN. Next, we investigated the number of correct models, which required a MaxSub score >0.1 [21, 24]. From higher to lower score, the parameters can be listed as SSA, PE, TAE, C_beta, and QMEAN. Using the *t*-test, we found the *t*-values for quality of those models selected by SSA compared with those selected by SE (2.473), TAE (2.905), SAA (4.488), QMEAN (2.215), and alignment score (3.050) were all larger than *t* (137, 0.005). This indicates that SSA performs best in identifying the near-native protein model when compared to SE, TAE, SAA, and QMEAN. The *t* value between QMEAN and SAA is 2.944, indicating that QMEAN is significantly better than SAA in identifying the near-native structure. Finally, the *t*-value between alignment score and C_beta (−2.059) and PE (−2.005) indicates that the performance of C_beta and PE is superior to the alignment score. Moreover, the *t*-value between alignment score and SSA was 2.132, which was larger than *t* (137, $p = 0.05$), indicating that the alignment score is not better than SAA in assessing the quality of the predicted model. Hence, SSA is a best parameter to estimate the quality of a predicted model.

Discussion

TBM generates the model for the alignment of the query sequence with the template protein sequence on the basis of the score matrix or the position specific score matrix (*i.e.*, profile). In general, the number of models obtained is equivalent to the number of alignments. The best models are selected according to their alignment score or E value [2–6].

Unlike TBM, a successful ab initio modeling procedure could help answering the basic questions of how and why a protein adopts a specific structure out of many possibilities. Of note, most current template-free modeling methods still

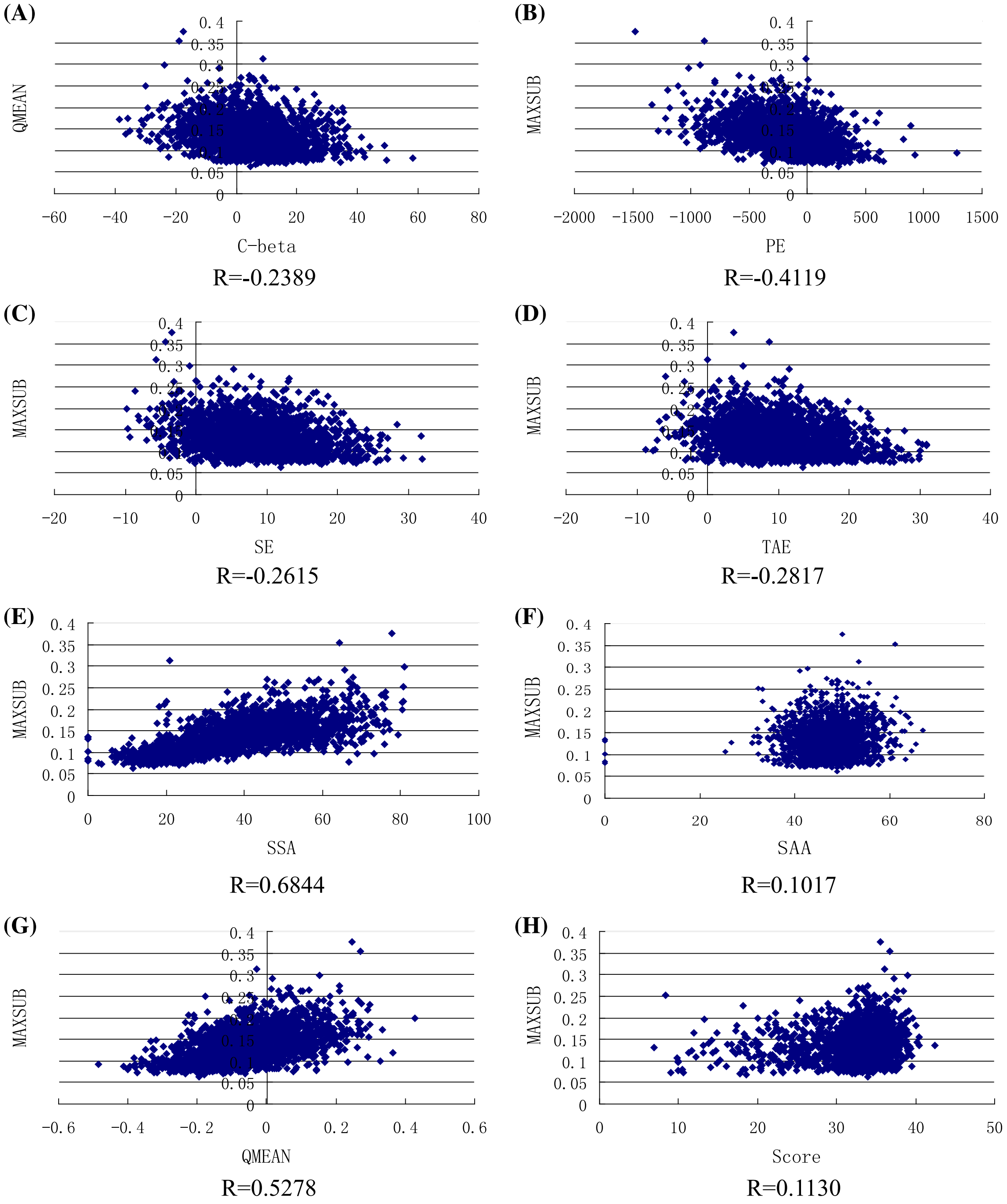

Fig. 1 Relationship between 8 parameters and MAXSUB for protein d1bmta1. **a** C_beta interaction energy (C_beta); **b** all-atom pairwise energy (PE); **c** solvation energy (SE); **d** torsion angle energy (TAE); **e** secondary structure agreement (SSA); **f** solvent accessibility agreement (SAA); **g** QMEAN; **h** alignment score (Score)

rely on template-based methods by using local sequence matching of fragments present in the database [25]. Similar to TBM, the ab initio prediction usually produces a considerable amount of candidate models. To identify the high quality models among the large number of candidate models is very important, and requires accurate evaluation of model quality.

The secondary structure of a protein is very close to its tertiary structure (*i.e.*, its native structure). A number of researchers have produced hypotheses in order to understand the relationship between secondary structure prediction and tertiary structure. Rost [26] summarized the areas of concern for secondary structure prediction as: regions likely to undergo structural change, proteins that need to be

Table 1 *t* test value between the eight parameters evaluated

	C_beta	PE	SE	TAE	SSA	SAA	QMEAN	Score
C_beta	0.000	−15.374	−0.800	−4.593	−22.815	3.790	−22.574	−12.389
PE	15.374	0.000	14.403	9.895	−7.560	14.658	−4.470	−1.546
SE	0.800	−14.403	0.000	−4.068	−20.175	4.277	−18.106	−11.618
TAE	4.593	−9.895	4.068	0.000	−20.535	6.856	−15.473	−9.598
SSA	22.815	7.560	20.175	20.535	0.000	22.254	7.701	4.056
SAA	−3.790	−14.658	−4.277	−6.856	−22.254	0.000	−21.784	−16.756
QMEAN	22.574	4.470	18.106	15.473	−7.701	21.784	0.000	0.608
Score	12.389	1.546	11.618	9.598	−4.056	16.756	−0.608	0.000

Table 2 Performance comparison of identification of near native models between the eight parameters

	C_beta	PE	SE	TAE	SSA	SAA	QMEAN	Score
MAXSUB[a]	15.858	15.282	14.769	13.746	17.252	11.874	15.040	13.863
Correct[b]	49	57	44	50	61	38	48	47
t(C_beta)[c]	0.000	0.624	1.907	2.023	−1.569	3.451	1.056	2.059
t(PE)	−0.624	0.000	0.651	1.982	−1.670	3.919	0.240	2.005
t(SE)	−1.907	−0.651	0.000	1.159	−2.473	2.790	−0.427	1.167
t(TAE)	−2.023	−1.982	−1.159	0.000	−2.905	1.839	−1.306	−0.205
t(SSA)	1.569	1.670	2.473	2.905	0.000	4.488	2.215	3.050
t(SAA)	−3.451	−3.919	−2.790	−1.839	−4.488	0.000	−2.944	−2.131
t(QMEN)	−1.056	−0.240	0.427	1.306	−2.215	2.944	0.000	1.216
t(Score)	−2.059	−2.005	−1.167	0.205	−3.050	2.131	−1.216	0.000

[a] Sum of MAXSUB

[b] The number of query proteins, whose MAXSUB of identified models by these eight parameters is larger than 0.1 (A47, A48)

[c] t value 1.960{t (∞, 0.05)} $<$ t (137, 0.05) $<$ 1.980{t (120, 0.05)}

classified based on secondary structure predictions from genome analysis, and predictions from the sequence to the 2D and 3D structures. Ortiz et al. [27] successfully used secondary structure predictions as one component of their 3D structure prediction method. Eyrich et al. [28] minimized the energy of arranging the predicted rigid secondary structure segments. Lomize et al. [29] also started protein structure predictions from secondary structure segments. Chen et al. [30] suggested using secondary structure predictions to reduce the complexity of molecular dynamics simulations. Samudrala et al. [31] combined secondary structure-based simplified presentations with a particular lattice simulation, attempting to enumerate all possible folds [31, 32].

The work summarized above demonstrates that predicting the protein secondary structure is essential for the prediction of the protein tertiary structure (topology). In this study, our results revealed that SSA is better suited for evaluating the predictive model quality (Fig. 1e). This algorithm may play an indispensable role in protein structure prediction. However, there is some bias in our experiments, which may overestimate the performance of SSA. One source of bias is the specified class of protein secondary structure (helix, sheet, and coil). Helix and sheet are considered regular structures; however, the protein fold consists mainly of coil. If the coil structural detail is predicted well, then measurement of SSA will exhibit better performance and identify a high quality model. Therefore, the drawback of using SSA is the requirement of structural details of the coil region.

Conclusions

Our analysis shows that SSA is the best parameter in assessing the quality of structural models, followed by C_beta, PE, and QMEAN. SSA is an essential parameter when choosing a predictive protein model template.

Acknowledgments This work was supported by Grants from the National Natural Science Foundation of China (No. 31171386 and 31371399), the Jiangsu Province Natural Science Foundation (No. BK2011093).

References

1. Murzin AG (2001) Progress in protein structure prediction. Nat Struct Biol 8:110–112

2. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
3. Sánchez R, Sali A (1997) Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7:206–214
4. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170
5. Panchenko AR, Marchler-Bauer A, Bryant SH (2000) Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 296:1319–1331
6. Xu D, Zhang J, Roy A, Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins: Struct Funct Bioinform 79:147–160
7. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kaźmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci USA 98:2329–2333
8. Simons KT, Strauss C, Baker D (2001) Prospects for ab initio protein structural genomics. J Mol Biol 306:1191–1199
9. Kolinski A, Skolnick J (1998) Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 32:475–494
10. Elofsson A (2002) A study on protein sequence alignment quality. Proteins 46:330–339
11. Pandit SB, Zhang Y, Skolnick J (2006) TASSER-Lite: an automated tool for protein comparative modeling. Biophys J 91:4180–4190
12. Kryshtafovych A, Fidelis K (2009) Protein structure prediction and model quality assess. Drug Discovery Today 14:386–393
13. Pearson WR, Sierk ML (2005) The limits of protein sequence comparison. Curr Opin Struct Biol 15:254–260
14. Levitt M, Gerstein M (1998) A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 95:5913–5920
15. Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16:776–785
16. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
17. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 101:7594–7599
18. Contreras-Moreira B, Fitzjohn PW, Bates PA (2003) In silico protein recombination:enhancing template and sequence alignment selection for comparative protein modelling. J Mol Biol 328:593–608
19. John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 31:3982–3992
20. Benkert P et al (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277
21. Wallner B, Fang H, Elofsson A (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins 53:534–541
22. Li J, Fang H (2012) Substitution transformation of score matrix for improving alignment quality of local sequence of distantly related proteins. Current Bioinformtics 7:35–42
23. Murzin AG, Brenner SE, Hubbard T (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
24. Fang H, Wallner B, Lundström J, von Wowern C, Elofsson A (2001) Improved fold recognition by using the Pcons consensus approach. Protein Struct Prediction: Bioinform Approach 1:397–416
25. Chen H, Kihara D (2008) Estimating quality of template-based protein models by alignment stability. Proteins 71:1255–1274
26. Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134:204–218
27. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 37:177–185
28. Eyrich VA, Standley DM, Felts AK, Friesner RA (1999) Protein tertiary structure prediction using a branch and bound algorithm. Proteins 35:41–57
29. Lomize AL, Pogozheva ID, Mosberg HI (1999) Prediction of protein structure: the problem of fold multiplicity. Proteins 37:199–203
30. Chen CC, Singh JP, Altman RB (1999) Using imperfect secondary structure predictions to improve molecular structure computations. Bioinformatics 15:53–65
31. Samudrala R, Xia Y, Huang E, Levitt M (1999) Ab initio protein structure prediction using a combined hierarchical approach. Proteins 37:194–198
32. Samudrala R, Huang ES, Koehl P, Levitt M (2000) Constructing side chains on near-native main chains for ab initio protein structure prediction. Protein Eng 13:453–457