Empirical limits for template-based protein structure prediction: the CASP5 example

B. Contreras-Moreira*,[1], I. Ezkurdia, M.L. Tress, A. Valencia*

Protein Design Group, Centro Nacional de Biotecnología, Madrid, Spain

Received 29 November 2004; revised 17 December 2004; accepted 5 January 2005

Available online 19 January 2005

Edited by Robert B. Russell

Abstract Most protein structure prediction methods use templates to assist in the construction of protein models. In this paper, we analyse the current state of template-based modelling approaches and reach an estimate of the empirical limits of these methods. Our analysis show that current prediction methods are already reaching these empirical accuracy limits in the easier cases, where finding a close homologue to the native target structure is not a problem. However, we find that even in the absence of alignment errors and using optimal templates, template-based methods have intrinsic limitations, suggesting that other methodologies, such as ab initio procedures, must be used if accuracy is ultimately to be improved.

Keywords: Template-based protein structure prediction; Comparative modelling; Fold-recognition; Fragment reconstruction; Accuracy limit

1. Introduction

Methods for protein structure prediction can be classified into two basic classes: those which use physical principles to fold a protein and those which use experimentally determined structures to help reconstruct the protein of interest. The first class is usually known as ab initio approaches [1]; the second includes related techniques such as comparative modelling, fold recognition and threading [2–7]. These generally use sequence alignments to map the sequence to be modelled onto protein templates of known structure and are guided by criteria such as sequence similarity or secondary-structure compatibility.

This paper deals mainly with the second class of methods, template-based methods. The empirical basis for these approaches comes from the observation by Chothia and Lesk [8] that protein sequence identity and structural similarity are correlated. According to their original results there are clear empirical limits for protein structure predictions based on single templates: for proteins sequences around 95% identical backbone deviations are expected to be under 1 Å RMS; when the sequence identity drops to 30%, deviations grow to around 4 Å RMS. These limits broadly agree with the observed performance of comparative modelling servers as measured by continuous benchmarks such as EVA [9] (see [10] for a review), and ultimately affect the quality and therefore the applicability of template-based predictions [11].

In addition to these natural restrictions, methods for template-based prediction of protein structure must solve two technical problems: the choice of the template closer to the target structure, and the derivation of the sequence alignment between the query and template protein closer to the optimal structural alignment. The lack of satisfactory solutions for these two problems has been identified as negatively affecting the performance of fold recognition and comparative modelling methods in previous "Critical Assessment of Techniques for Protein Structure Prediction" experiments (CASP [12]) [13,14].

However, choosing the correct template and alignment are not the only problems facing predictors. Even those models built from the correct template and alignment often require substantial refinement in order to be sufficiently close to the native target structure. This paper seeks to estimate the limits of current template-based structure prediction techniques under ideal conditions, that is building a model a posteriori using multiple optimal templates and in the absence of alignment errors.

We do that by allowing models to be built by combining aligned fragments from several templates, selected by structural similarity. We then measure, using the CASP GDT_TS score [15], how the best fragment-based predictions compare to the native target structure.

Additionally, we ask how far the predictions are from these best possible models. This gives us a better idea of how successful the current modelling methods are, how good they could be in the absence of the sequence alignment problem, and can implicitly tell us to what extent ab initio methods would be needed to improve the current performance of template-based methods.

2. Datasets, methods and algorithms

A collection of 68 targets, as split in domains by the CASP5 organisers, was taken as our test set (see http://predictioncenter.llnl.gov/casp5). These targets are proteins whose experimental structures were about to be released at the time CASP5 started (May, 2002). To model

*Corresponding authors. Fax: +34 91 585 45 06.
E-mail addresses: contrera@ccg.unam.mx (B. Contreras-Moreira), valencia@cnb.unam.es (A. Valencia).

[1] Present address: Programa de Genómica Computacional, Centro de Ciencias Genómicas, Av. Universidad s/n, Colonia Chamilpa, 62210 Cuernavaca, Morelos, México.

0014-5793/$30.00
doi:10.1016/j.febslet.2005.01.014

these targets we needed a library of PDB templates and for that we used a 90% non-redundant set of PDB chains from April 2002, obtained from PDB-SELECT [16]. Since CASP5 started in May 2002 we were sure that we did not have access to templates not available to the predictors at that time. This library included 6182 PDB chains.

Then, we designed our procedure with five steps to be applied to every CASP5 target:

(1) Search the template library for a list of significantly similar PDB structures.
(2) Superimpose all found templates in the target's frame of reference.
(3) Calculate a large number of fragment-based models from the ensemble of templates and evaluate them using typical CASP evaluation parameters.
(4) Compare the best fragment-based model to the target structures.
(5) Compare the best fragment-based model to the best model produced by CASP5 predictors.

Note that no fragment readjustment is performed, since we felt this was an ab initio technique.

Each of these steps was implemented as follows:

1. To search the template library we used the program MAMMOTH [17] and took only those templates that yielded a $-\ln E$ score over 4.5, to avoid using marginally similar structures. Only the top 40 hits were considered.
2. To superimpose the selected templates in the frame of reference of the target we used two programs, MAMMOTH and LGA [15], to generate alternative sequence-independent superpositions. Other superimposition protocols could be added in this step, but for demonstration purposes we felt that two were enough. In the case of MAMMOTH, the coordinates of the superimposed template needed to be transformed using the rotation matrices and translations provided in the output.
3. This was the most important step in our protocol, the generation of a collection of models for our target by fragment reconstruction from the superimposed templates. This step included several substeps, as shown in Fig. 1 and was based on a previous work [18]:

3.1. Construct a multiple alignment from the pairwise structural alignments between target and templates, with the target as the frame of reference. This multiple alignment can be regarded as a matrix with each row in the matrix corresponding to a template, each column to an aligned set of residues and their backbone coordinates.
3.2. In this sub-step, we define "fragment" as a contiguous set of template residues that have been aligned without gaps by either MAMMOTH or LGA. As suggested by related work [19,20], fragment length is an important parameter and here we tried values of 5 and 9 residues. To score fragments we used a score similar to GDT-TS [15,21], the main evaluator used in CASP experiments. GDT-TS score measures similarity between two structures based on a combination of the fractions of matching residues within distance cutoffs of 1, 2, 4, and 8 Å. MyGDT is similar to GDT-TS but calculated on just one superimposition. It is calculated as P1 + P2 + P4 + P8/4, where P.n. is the % of residues in the template closer than n Å to the corresponding residues in the target. In this sub-step, fragments of the chosen length in the matrix are labelled and their myGDT scores are calculated and stored.
3.3. For each labelled fragment a new fragment-based model by growing it towards both N and C termini within the matrix applying iteratively these greedy rules:
  3.3.1. If one or more fragments are available in the matrix to grow the model, choose the best one and add it. Fragments are scored according to their local myGDT score with respect to the target's coordinates.
  3.3.2. Otherwise, if possible, extend the solution model by one residue.
3.4. Rank all obtained fragment-based solutions in terms of global myGDT scores with respect to the target length and select the best ones within a given tolerance (set to 1 myGDT unit in this experiment).

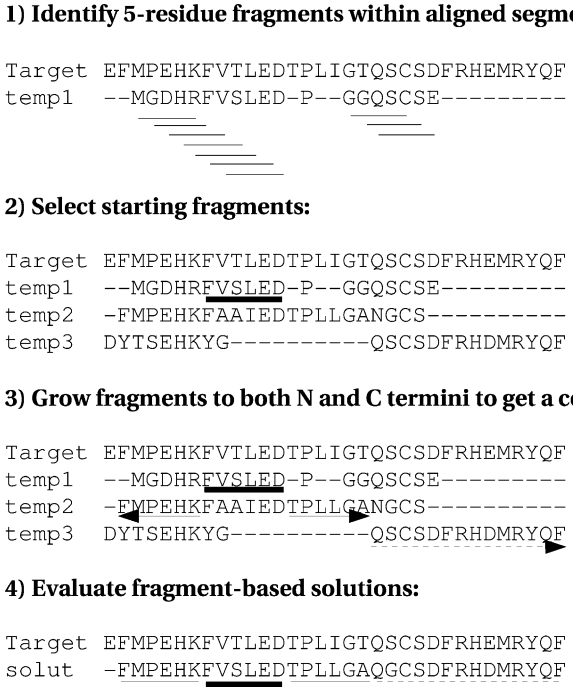

Fig. 1. Graphical example of the construction of fragment-based models from a set of superimposed templates. A possible fragment-based solution is built starting from a fragment of 5 residues in template 1. This fragment is then grown in both left and right (N and C-terminal) directions. Going left there are three options, three possible fragments and the one from template 2 was taken for having the best local myGDT score with respect to the target's coordinates. Towards the right of the starting fragment, initially only the fragment from template 2 was available. Then another fragment was extracted from template 3 and this was actually extended since no other fragments were available (see rules 3.3.1 and 3.3.2).

3.5. Use the sequence-dependent mode of LGA to calculate final GDT_TS scores for these selected models and store the best obtained GDT_TS, our estimation for the best accuracy achievable from these templates and these alignments.

Finally, we compare our fragment-based model to the corresponding best prediction in CASP5 and also to the experimental structure.

This protocol was implemented as a set of three Perl programs, available from the authors:

(1) *update_nrpdb.pl*: to create a local copy of the non-redundant set of PDB chains Pubs.
(2) *search_templates_mammoth.pl*: to scan the target against the template library using MAMMOTH and to finally generate a list of suitable templates.
(3) *fragbench.pl*: to superimpose the templates and to create a collection of fragment-based solutions evaluated in terms of GDT_TS. The initial template superpositions and the final selected models are printed to files in PDB format.

3. Results

After creating the non-redundant set of PDB chains of April, 2002, we ran *search_templates_mammoth.pl* for every target. In the case of the small target T0186_3 our search for templates produced no hits with $-\ln E$ scores over the threshold, so we added templates 1gkp_A, 1ie7_C, 1gkr_A, 4ubp_C, 1k1d_A, templates that were used by CASP5 predictors for this target. For the rest of targets our procedure was successful and the best template for each of them is shown in Table 1, together with the % of sequence identity after superposing with MAMMOTH.

We ran *fragbench.pl* for all 68 CASP5 targets, building models from five residue fragments and from nine residue fragments. The GDT-TS scores for both sets of ideal fragment-based models are shown in Fig. 2. In the figure, the native target structure would have a GDT-TS of 100.00, so the results show that even with the aid of error-free alignments and the optimum choice of fragments, the information from the fragment library is not sufficient to recreate the target structure. In general, the higher the sequence identity the higher the GDT-TS of the fragment-based model, though the models have a wide range of GDT-TS scores. For the 9 residue fragment models, for example, GDT-TS ranges from 37.98 for target T0132 to 96.81 for target T0137. One reason for the wide range of scores is that the structural information required to recreate the models is not always available, reflecting non-homogeneous distribution of structural space within the PDB.

In Fig. 3A, the GDT-TS scores for the best predicted models from CASP5 are superimposed on the scores for both sets of ideal fragment-based models. Here the targets are sorted according to sequence identity between the target and the closest template (from 11.6% for T0181 to 49% for T0154_1). In Fig. 3B, all three sets of models for all targets are replotted with the target-template sequence identity on the horizontal axis. Linear regression lines are also shown.

These two figures suggest that above a imaginary line of ~35% of sequence identity the best predictors are often reaching the limits of what can be done solely with the available structures in the database. It is in these cases where improvements on the performances of the template-based methods can only come from some form of ab initio method.

Table 1
List of CASP5 targets split in domains with the best template found by MAMMOTH to model them, the % of sequence identity and the $-\ln E$ value, the significance of the structural similarity

Target_domain	Template	% seq. id.	Aligned length	$-\ln E$
T0130	1lou_A	18.2	79/100	8.894
T0132	1bvq_A	17.8	123/147	14.36
T0136_1	1ef8_A	17.1	167/256	16.77
T0136_2	1ey3_A	16.5	202/264	17.12
T0137	1bwy_A	43.1	131/133	18.54
T0138	1dc8_A	15.3	121/135	15.82
T0141	1lba_#	23.5	141/187	9.996
T0142	1i9z_A	22.9	268/280	19.88
T0143_1	1agj_A	28.0	114/121	12.72
T0143_2	1qtf_A	34.8	94/95	13.13
T0146_1	1f9f_C	17.8	72/107	7.61
T0146_2	1ais_A	13	81/89	9.74
T0147	1f74_A	20.0	226/234	16.06
T0148_1	1aps_#	19.7	70/71	10.78
T0148_2	1ap8_#	16.7	86/91	9.660
T0149_1	1fts_#	25.0	187/201	15.63
T0149_2	1h7s_B	15.8	109/116	9.336
T0150	1ck9_A	33.7	96/97	14.11
T0151	3ull_B	33.8	94/106	12.58
T0153	1euw_A	31.9	127/134	16.98
T0154_1	1iho_A	49.1	175/185	22.99
T0154_2	1iho_A	34.0	97/103	14.17
T0155	1dhn_#	33.6	116/117	16.77
T0156	2bnh_#	17.2	155/156	7.896
T0157	1hjr_A	16.2	118/120	11.91
T0159_1	1ii5_A	19.7	155/167	9.517
T0159_2	1atg_#	14.3	135/142	6.744
T0161	1du0_B	17.9	55/154	6.206
T0162_1	1hu3_A	16.3	55/56	8.554
T0162_2	1flt_X	17.0	50/51	6.008
T0162_3	2mpr_A	17.9	167/168	10.10
T0165	1fj2_A	22.9	218/318	21.96
T0167	1jeo_A	37.8	170/180	20.47
T0168_1	3nul_#	16.0	125/170	5.959
T0168_2	4blm_A	17.4	134/141	6.763
T0169	1qsm_B	18.3	133/156	14.96
T0170	1hlr_A	17.4	68/69	8.295
T0172_1	1ej0_A	20.6	167/192	19.05
T0172_2	1f4i_A	23.8	44/101	5.868
T0173	1jil_A	19.1	255/287	9.329
T0174_1	1fi4_A	22.0	193/197	8.380
T0174_2	1fi4_A	16.0	154/155	8.382
T0176	1jrm_A	23.1	90/100	8.615
T0177_1	1gu9_F	40.4	52/57	7.435
T0177_2	1e8g_A	34.2	87/88	8.482
T0177_3	1gpj_A	29.0	73/75	9.984
T0178	1jcl_A	25.8	216/219	26.48
T0179_1	1inl_B	32.7	55/56	8.554
T0179_2	1inl_B	44.4	213/218	25.96
T0181	1xyp_A	11.6	110/111	5.008
T0182	1c24_A	42.7	248/249	30.61
T0183	1jcl_A	28.8	226/247	26.25
T0184_1	1jfz_A	33.8	145/165	19.57
T0184_2	1qu6_A	20.5	70/72	10.53
T0185_1	4uag_A	20.6	99/101	14.27
T0185_2	4uag_A	30.5	185/197	21.36
T0185_3	4uag_A	18.6	121/130	15.03
T0186_1	1gou_A	25.0	74/77	5.828
T0186_2	1ejr_C	15.0	231/250	18.78
T0186_3	1gkp_A	28.6	–	–
T0187_1	1jji_A	24.1	176/187	9.130
T0187_2	1ej0_A	16.7	172/227	12.23
T0188	1eo1_A	24.5	105/107	14.85
T0189	1bx4_A	20.5	299/319	27.97
T0191_1	1di6_A	17.9	134/139	12.59
T0191_2	2pgd_#	26.3	120/143	15.65
T0193_1	1f8r_A	16.0	73/74	10.23
T0193_2	1e3j_A	16.8	126/130	12.89

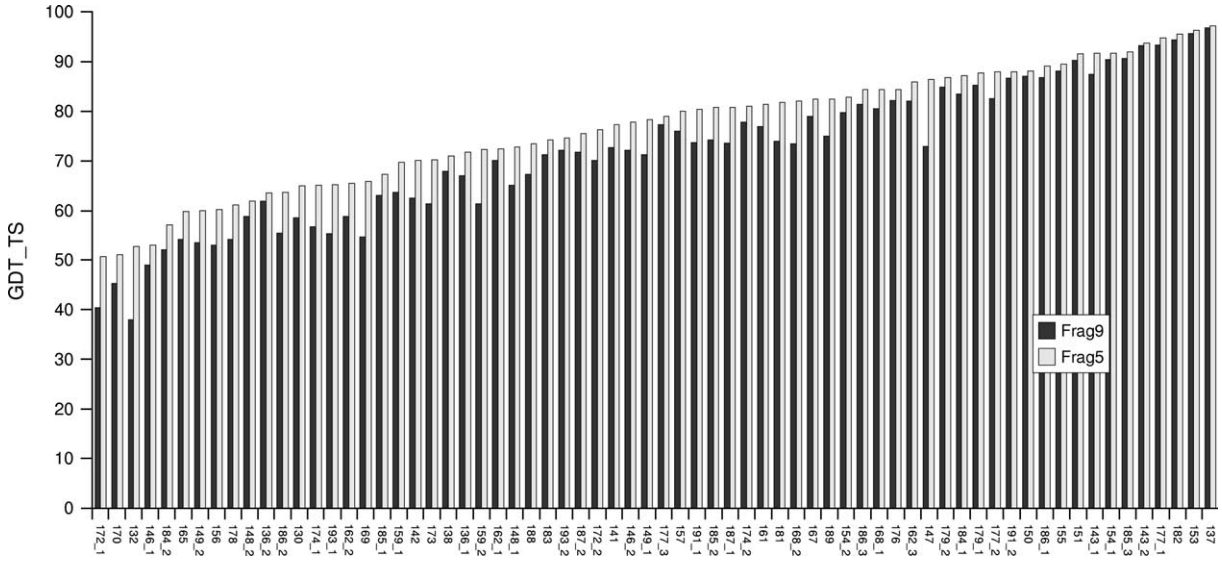

Fig. 2. Here the GDT-TS scores of the models built by FRAGBENCH using fragments of size 5 and 9 are plotted against each of the CASP5 targets. The scores are ordered by the 5 residue fragment model GDT-TS score.

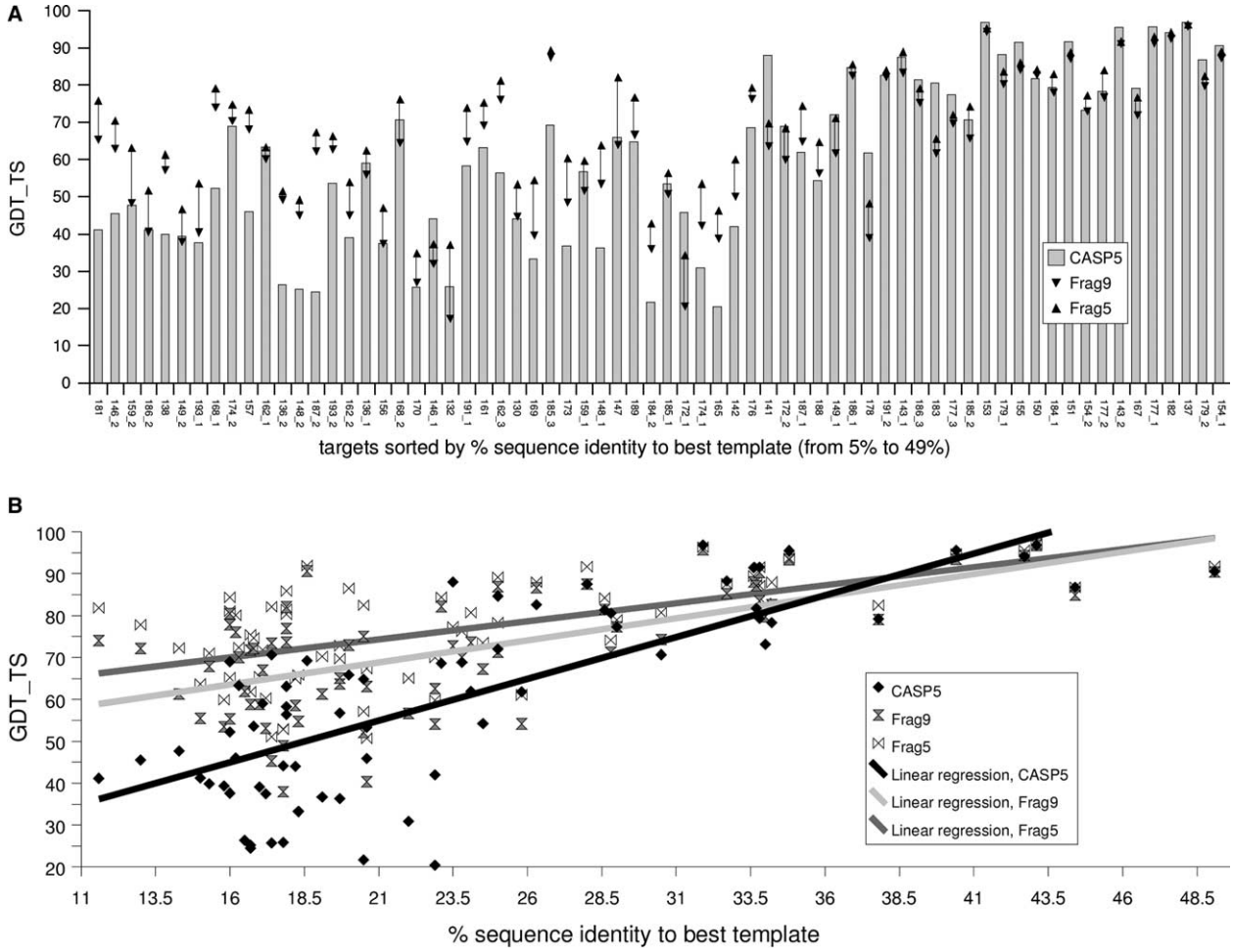

Fig. 3. CASP5 performance compared to empirical template-based limits. (A) The best CASP5 predictions are shown together with the best template-based solutions produced by FRAGBENCH, using fragments of size 5 and 9. (B) Same results plotted as a function of the % of sequence identity to the best template, used to draw linear-correlations that meet near the 35% imaginary line. This data shows that CASP5 methods cannot reach the fragment-based performance below the 35–40% threshold. It also shows that even minimizing alignment problems, models built from fragments have limited quality. Note: these figures show the same trends when excluding targets labelled as New Fold and Fold Recognition (analogy) in CASP5.

These results also show that below this line template-based approaches still have room for improvement with regards to alignment accuracy and template selection.

4. Discussion

CASP blind trials for protein structure prediction methods have identified two major sources of errors that affect template based prediction methods: selection of incorrect templates and errors in alignments. This paper shows the limits of template-based modelling methods in the absence of these problems. Even with optimal selection of templates (combining various chains) and perfect quality structural alignments, most CASP5 targets could not be predicted with GDT_TS scores better than 70%. This result shows that the current set of known structures is rather limited and does not contain the complete information necessary for building template-based models, even when the best combination of this structural information is known. In this sense the results suggest that only additional information can overcome the limitations of the available structural information, and ab initio methods would have to be combined with fragment-based methods to improve current performance of homology modelling techniques.

On the positive side, the results of comparing CASP5 models with the optimal multi-template models show that for the so called easy cases with templates over 30% of sequence identity with the query proteins, the best models are now almost as good as the optimal multi-templates, and we are reaching the limits of what the template-based methodology can do for the modelling of protein main chain. On the other hand, in the region below 30% identity the best models are far from even the optimal model that could be obtained with the best available structural information. It is in this region where the problems of template identification, combination and sequence alignments are still playing a major role.

The recent increase in growth of the structural databases ought not only to provide templates closer to the target structures in many cases, but also ought to provide a more diverse library of fragments for fragment-based modelling. It will be interesting in future years to see to what extent fragment-based modelling will benefit from the effects of this increased pool of template information and to see whether structure prediction groups can improve on the best possible template-based model.

Acknowledgement: Thanks to Osvaldo Graña for his help in using CASP5 data and to Adam Zemla for his assistance with LGA.

References

[1] Bonneau, R. and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects. Annual Review of Biophysics and Biomolecular Structure 30, 173–189.
[2] Greer, J. (1981) Comparative model-building of the mammalian serine proteases. Journal of Molecular Biology 153 (4), 1027–1042.
[3] Jones, T.A. and Thirup, S. (1986) Using known substructures in protein model building and crystallography. EMBO Journal 5 (4), 819–822.
[4] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. Nature 358 (6381), 86–89.
[5] Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. Journal of Molecular Biology 213 (4), 859–883.
[6] Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253 (5016), 164–170.
[7] Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 234, 779–815.
[8] Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. EMBO Journal 5 (4), 823–826.
[9] Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17 (12), 1242–1243.
[10] Contreras-Moreira, B., Fitzjohn, P.W. and Bates, P.A. (2002) Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. Applied Bioinformatics 1 (4), 177–190.
[11] Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. Science 294, 93–96.
[12] Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) Critical assessment of methods of protein structure prediction (CASP) – round V. Proteins 53 (Suppl. 6), 334–339.
[13] Tramontano, A., Leplae, R. and Morea, V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. Proteins (Suppl. 5), 22–38.
[14] Sippl, M.J., Lackner, P., Domingues, F.S., Prlic, A., Malik, R., Andreeva, A. and Wiederstein, M. (2001) Assessment of the CASP4 fold recognition category. Proteins (Suppl. 5), 55–67.
[15] Zemla, A. (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 31, 3370–3374.
[16] Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Science 1, 409–417.
[17] Ortiz, A.R., Strauss, C.E.M. and Olmea, O. (2002) MAMMOTH (Matching molecular models obtained from theory): an automated method for model comparison. Protein Science 11, 2606–2621.
[18] Contreras-Moreira, B., Fitzjohn, P.W. and Bates, P.A. (2003) In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. Journal of Molecular Biology 328 (3), 593–608.
[19] Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. Journal of Molecular Biology 281, 565–577.
[20] Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. (2002) Small libraries of protein fragments models native structures accurately. Journal of Molecular Biology 323, 297–307.
[21] Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. (1999) Processing and analysis of CASP3 protein structure predictions. Proteins Structure, Function and Genetics (Suppl. 3), 22–29.