Review

Michal Brylinski*

Is the growth rate of Protein Data Bank sufficient to solve the protein structure prediction problem using template-based modeling?

Abstract: The Protein Data Bank (PDB) undergoes an exponential expansion in terms of the number of macromolecular structures deposited every year. A pivotal question is how this rapid growth of structural information improves the quality of three-dimensional models constructed by contemporary bioinformatics approaches. To address this problem, we performed a retrospective analysis of the structural coverage of a representative set of proteins using remote homology detected by COMPASS and HHpred. We show that the number of proteins whose structures can be confidently predicted increased during a 9-year period between 2005 and 2014 on account of the PDB growth alone. Nevertheless, this encouraging trend slowed down noticeably around the year 2008 and has yielded insignificant improvements ever since. At the current pace, it is unlikely that the protein structure prediction problem will be solved in the near future using existing template-based modeling techniques. Therefore, further advances in experimental structure determination, qualitatively better approaches in fold recognition, and more accurate template-free structure prediction methods are desperately needed.

Keywords: comparative modeling; COMPASS; HHpred; Protein Data Bank; protein fold recognition; protein structure prediction; protein threading; template-based modeling.

DOI 10.1515/bams-2014-0024
Received December 22, 2014; accepted January 8, 2015; previously published online February 7, 2015

Introduction

Linus Pauling, a chemist, peace activist, educator, and the only person to be awarded two unshared Nobel Prizes (Nobel Prize in Chemistry in 1954 and Nobel Peace Prize in 1962), concluded his Nobel Lecture with the following remark: "We may, I believe, anticipate that the chemist of the future who is interested in the structure of proteins, nucleic acids, polysaccharides, and other complex substances with high molecular weight will come to rely upon a new structural chemistry, involving precise geometrical relationships among the atoms in the molecules and the rigorous application of the new structural principles, and that great progress will be made, through this technique, in the attack, by chemical methods, on the problems of biology and medicine" [1]. Indeed, structural biology has been rapidly developing over the past half-century contributing to many major breakthroughs in life sciences.

The specific and unique three-dimensional structures give macromolecules the ability to perform various cellular functions. Consequently, macromolecular structure, folding, and structural alterations affecting the function of proteins and nucleic acids are of prime importance to biologists. Biological structures are typically resolved experimentally by X-ray crystallography and nuclear magnetic resonance spectroscopy; nonetheless, these techniques are often expensive and time consuming. As a result, known biological sequences greatly outnumber available structures; as of December 2014, the National Center for Biotechnology Information Reference Sequence Database [2] contains 46,968,574 sequences, whereas the Protein Data Bank [3] (PDB) features 105,097 structures. A high demand for protein structures stimulates the development of computational structure prediction methods, which in fact represent the only strategy to keep up with the rapidly growing volume of sequence data.

***Corresponding author: Michal Brylinski,** Department of Biological Sciences, 202 Life Sciences Bldg., Louisiana State University, Baton Rouge, LA 70803, USA; and Center for Computation and Technology, 2054 Digital Media Center, Louisiana State University, Baton Rouge, LA 70803, USA, E-mail: michal@brylinski.org.
http://orcid.org/0000-0002-6204-2869

Prediction of protein tertiary structures

Current techniques for protein structure prediction fall into two major categories: template-free and template-based methods [4, 5]. The first group comprises various algorithms that simulate the folding of polypeptide chains into their native conformations [6–8]. These approaches build upon fundamental physical principles, that is, they do not directly use structural information extracted from related proteins. Despite the promising progress in our understanding of folding mechanisms and the development of template-free methods [9], the quality of protein models constructed by template-free modeling is generally insufficient for subsequent functional annotation and drug discovery compared to experimental structures [10–12]. In contrast, template-based methods routinely generate models whose accuracy is often comparable to that of low-resolution experimental structures [13–15]. In comparative protein structure prediction, the three-dimensional model of a target protein is constructed using a template that is typically an evolutionarily related protein whose structure has been solved experimentally [16–18]. In a nutshell, a template protein is first identified in the PDB and the target-to-template alignment is calculated. Next, according to this alignment, an initial model is generated using the coordinates of equivalent residues in the template. Missing residues and loops are added to the initial model, which is subsequently subjected to a refinement procedure to optimize intramolecular contacts and the packing of side chains.

Certainly, the recognition of quality templates and the construction of correct alignments are critical to the success of comparative modeling. This can be achieved using purely sequence-based methods to identify evolutionarily closely related homologs [19–22]; however, detecting remotely related templates in the "twilight zone" of sequence similarity [23] requires more sensitive algorithms such as protein threading and fold recognition [24]. Many of these methods employ sequence profile alignments [25, 26] to compare a profile of the target protein constructed from homologous sequences against profiles generated for a nonredundant subset of proteins from the PDB. In addition, scoring functions used in threading commonly incorporate structural information in the form of pair potentials, solvent accessibility, secondary structure profiles, backbone dihedral angles, and hydrophobic interactions [27–31]. Finally, the accuracy of protein fold recognition can be further improved by combining several threading algorithms into meta-threading pipelines [32, 33].

Cheshire Cat in structure bioinformatics

Unquestionably, the past two decades have seen an encouraging progress in protein structure prediction. The quality of modeled structures is constantly improving, making them suitable for a wide range of applications, including molecular function inference, the prediction of the effect of mutations, and rational drug design [34–39]. This progress can be attributed to two major factors: the advances in algorithms and methods for comparative protein structure modeling and the continuous growth of structure databases. It has been already established that the PDB is likely complete at the level of compact, single domain protein structures, viz. suitable structure templates are present in the PDB to reliably model any protein sequence [40, 41]. Therefore, developing efficient fold recognition algorithms that are capable of detecting these templates can, in principle, solve the protein folding problem [42].

On the other hand, perhaps the growth rate of the PDB is also sufficient for the current algorithms to be able to effectively detect structure templates and build accurate models for any biological sequence in the near future. If so, one can contentedly wait as a Cheshire Cat, a famous character in Lewis Carroll's novel, for the protein folding problem to be solved just by having enough structures for a near-complete mapping to the sequence space using existing bioinformatics tools. For instance, it was estimated that solving 16,000 carefully selected structures would provide structural models for approximately 90% of 300,000 sequences available in the Swiss-Prot and TrEMBL [43] databases back in 2000 [44]. Considering the exponential growth of the PDB, these structures were anticipated to become available in about a decade [44, 45]. As expected, the structural coverage of Swiss-Prot during that period has increased [46]; nonetheless, the near-complete structural mapping of the sequence space has yet to be attained.

In this communication, we perform a retrospective analysis of the structural coverage of a representative set of proteins using remote homology. Two state-of-the-art sequence profile-based algorithms for fold recognition are tested: COMPASS [47] and HHpred [48]. We analyze 12 time-snapshots of the PDB downloaded from the archive of the Research Collaboratory for Structural Bioinformatics that cover a period from 2005 to 2014. The results are presented in terms of the success detecting structurally related proteins, the quality of target-to-template alignments, and the overall prediction confidence. Although the structural coverage is constantly increasing, we demonstrate that the protein folding problem is unlikely to be solved in the near future using a Cheshire Cat approach without further advances in the methods for protein structure prediction.

Growth of the PDB

The exponential growth rate of the PDB often found in literature is calculated from the total number of entries deposited on a regular basis [49–51]. Our analysis reveals a similar trend (Figure 1, black triangles). This exponential growth results from the increasing number of structures released in the PDB every year. Nevertheless, a nonredundant content at both sequence and structure levels is more meaningful from a point of view of protein structure prediction. On that account, we first clustered each of the 12 PDB snapshots at 40% sequence identity using CD-HIT [52] and counted the number of representative families. Figure 1 (gray squares) shows that the number of homologous clusters increases linearly at a rate of 1400 per year on average. Next, we partitioned each snapshot into a set of structurally related proteins by using the Fr-TM-align program [53] to calculate the pairwise TM-score [54] between individual chains. The resulting structure similarity matrix was subsequently clustered using a greedy algorithm [55] at the TM-score threshold of 0.5 [56]. Similar to sequence clusters, the number of structurally related groups of proteins increases linearly at a rate of 650 per year on average (Figure 1, crossed circles). Thus, despite the exponential growth of the PDB, its nonredundant sequence and structure components increase linearly at much slower pace.

Structural coverage of the protein sequence space

A pivotal question is how this growth of structural information improves the quality of three-dimensional models constructed by contemporary bioinformatics approaches. To address this problem, we compiled a representative dataset of 7818 proteins 100–200 amino acids in length, whose experimental structures were available in the PDB as of April 2014. Here, the redundancy was removed at 40% sequence identity using CD-HIT [52]. Next, we used two state-of-the-art fold recognition algorithms, COMPASS [47] and HHpred [48], to identify weakly homologous templates for the target proteins in 12 time-snapshots of the PDB covering a period from 2005 to 2014. By excluding those template proteins that share more than 40% sequence identity with their targets, we focus on the capability of sequence profile-based methods to identify

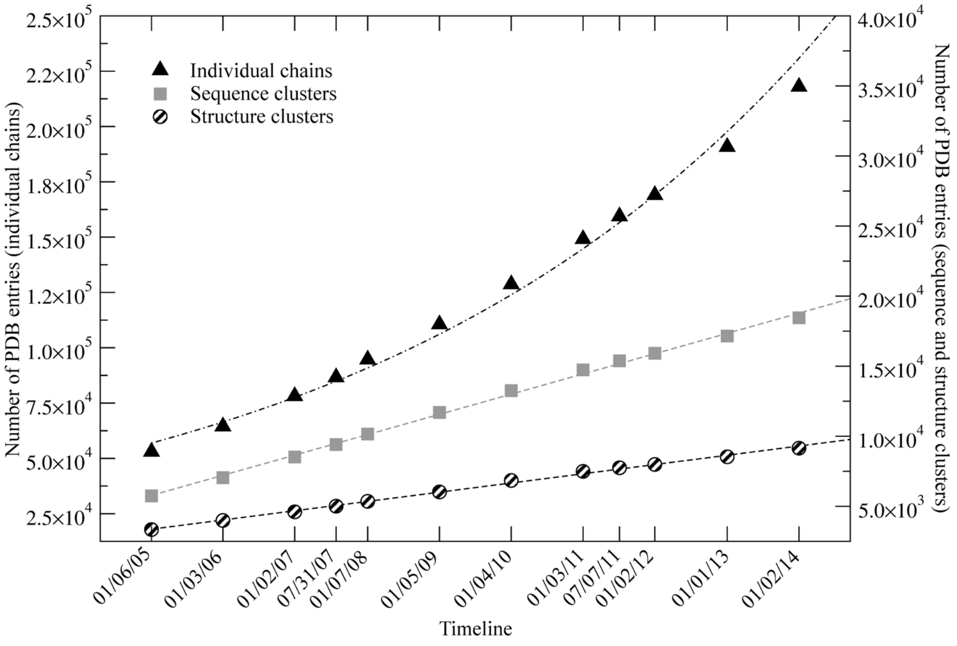

Figure 1: Growth rate of the PDB.
At any given time, we counted the number of protein chains (black triangles), the number of sequence clusters obtained by clustering individual chains at 40% sequence identity (gray squares), and the number of structure clusters calculated by clustering individual chains at a TM-score of 0.5 (crossed circles).

structure templates in the “twilight zone” of sequence similarity. Each top-ranked hit is compared to its target using sequence and structure similarity, the quality of threading alignment, and the estimated prediction confidence.

Figure 2A shows that the overall sequence identity between target proteins and the top hits detected by both programs increased over almost a decade by only a couple of percentage points within a narrow range of 24%–28%. We note that closely related templates with more than 40% sequence identity to their targets are excluded in this study. Interestingly, the prediction confidence increased from the average *E*-value of approximately 0.18 in 2005 and 2006 to 0.12 to 0.14 in 2010 to 2014 with a significant boost in the confidence scores during a period from 2007 to 2009 (see Figure 2B). Confidence estimates provided by both fold recognition algorithms are well correlated with the actual prediction accuracy; therefore, as expected, the quality of detected templates improved as well. This is shown in Figure 2C for TM-score values calculated from structure alignments constructed by Fr-TM-align and in Figure 2D for TM-score values calculated over threading alignments reported by COMPASS and HHpred. For instance, using HHpred, the average TM-score of structure alignments increased in the first 3 years from 0.60±0.20 in 2005 to 0.64±0.20 in 2008. This fast initial growth rate shown as dashed lines in Figure 2C and D slowed down afterward and the corresponding average TM-score values increased during the last 3 years from 0.66±0.19 in 2011 to only 0.67±0.19 in 2014. Of course, threading algorithms should also generate correct alignments that would allow for the construction of high-quality models of the target proteins. Therefore, in Figure 2D, we plot TM-score values calculated over threading alignments. In general, the quality of threading alignments is approximately 10% lower than those reported by Fr-TM-align; however, it has been increasing at fairly similar rates. For example,

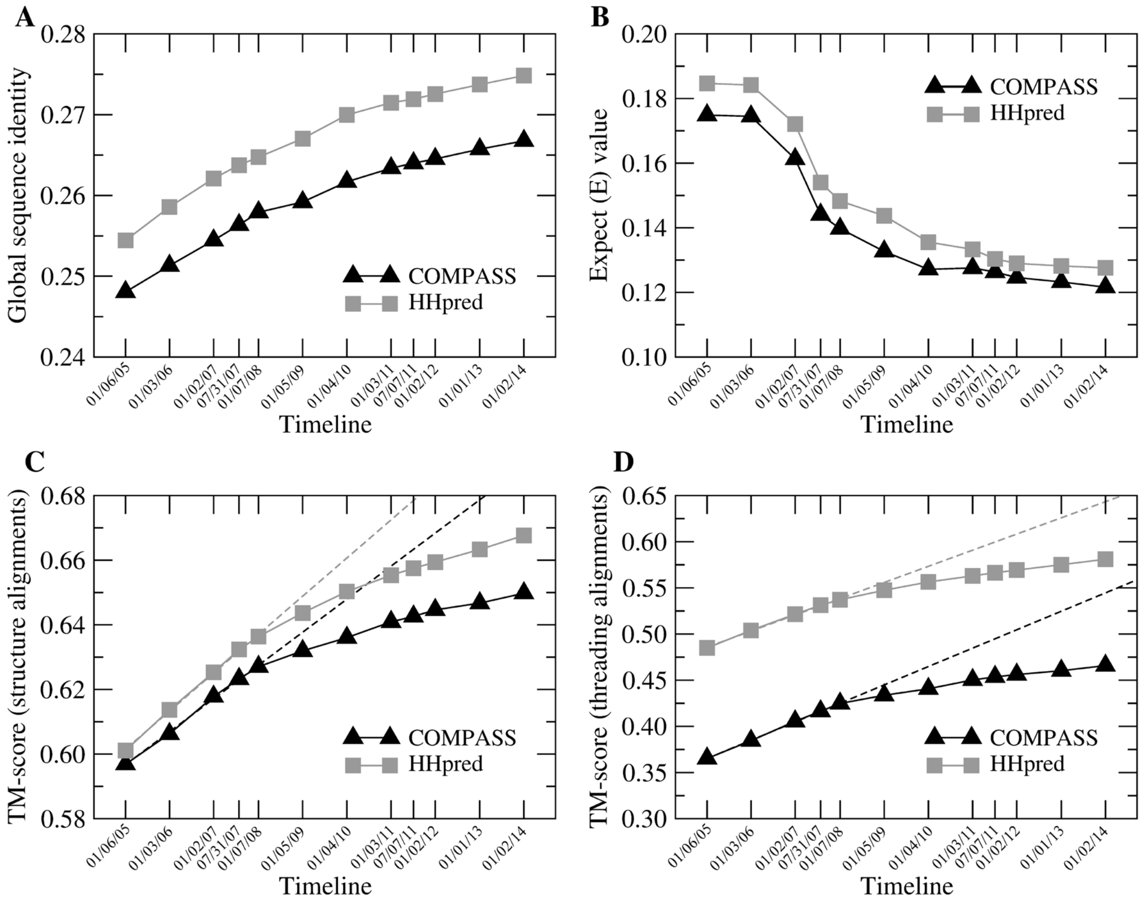

Figure 2: Quality of template structures detected by COMPASS and HHpred over time.
(A) Target-template global sequence identity, (B) expect values that correspond to the prediction confidence, and the TM-score calculated using (C) structure and (D) threading alignments. Only the top-ranked templates are considered; each data point represents the average value calculated over the benchmarking proteins. The dashed regression lines in C and D are calculated by fitting a linear equation to the first five data points (2005–2008).

Table 1: Time-dependent improvement of the quality of template structures detected by COMPASS and HHpred. Each row represents one snapshot of the PDB.

Date	Structure alignments		Threading alignments	
	COMPASS	HHpred	COMPASS	HHpred
01/06/05				
01/03/06	0.001[a]	<0.001[a]	<0.001[a]	<0.001[a]
01/02/07	<0.001[a]	<0.001[a]	<0.001[a]	<0.001[a]
01/07/08	0.002[a]	0.001[a]	<0.001[a]	<0.001[a]
01/05/09	0.099	0.020[b]	0.035[b]	0.014[b]
01/04/10	0.169	0.032[b]	0.106	0.029[b]
01/03/11	0.193	0.098	0.123	0.105
01/02/12	0.203	0.183	0.168	0.117
01/01/13	0.472	0.207	0.302	0.146
01/02/14	0.292	0.151	0.189	0.135

First, we calculated the average TM-score and the corresponding SD using structure alignments and threading alignments of the identified templates against experimental structures. Then, the statistical significance was evaluated using the *t* statistic and a p-value calculated versus the preceding snapshot with those values <0.05 (significant) and 0.01 (highly significant) marked by superscripts [b] and [a], respectively.

the average TM-score for threading alignments reported by HHpred increased from 0.48±0.28 in 2005 to 0.54±0.27 in 2008 and from 0.56±0.25 to 0.58±0.25 during the last 3-year period. Note that these trends are independent of the fold recognition algorithm and qualitatively similar results are obtained for COMPASS, although with a somewhat lower accuracy compared to HHpred.

Finally, using the *t* statistic, we calculated p-values for the TM-score distribution for each time-snapshot with respect to the preceding release of the PDB. The results reported in Table 1 indicate that highly significant improvements with p<0.01 for both the template quality (structure alignments) and the accuracy of the corresponding threading alignments occur between 2005 and 2008. Still, significant improvements with p<0.05 are observed during a period from 2009 to 2010; however, improvements after 2010 are statistically insignificant.

Conclusions

Current approaches to protein structure modeling routinely construct high-quality models for biological sequences, provided that reliable template structures can be identified in the PDB. Nonetheless, fold recognition fails to detect structurally related proteins for many targets despite that these templates are present in threading libraries. Because the PDB undergoes an exponential expansion in terms of the number of macromolecular structures deposited every year, one could expect that there will be enough structures at some point to solve the protein folding problem using contemporary structural bioinformatics tools. In this communication, we investigate this issue by performing fold recognition for a representative set of proteins using state-of-the-art algorithms against a dozen of PDB snapshots covering a period from 2005 to 2014. We show that the number of proteins whose structures can be confidently predicted indeed continuously increases on account of the growth of the PDB; however, this encouraging trend noticeably slowed down around the year 2008. At the current pace, it is unlikely that the protein structure prediction problem will be solved in the near future using existing modeling techniques. Therefore, in addition to advances in experimental structure determination, qualitatively better approaches to fold recognition as well as more accurate template-free structure prediction techniques would be required to achieve a complete structural coverage of the protein sequence space.

Acknowledgments: Portions of this research were conducted with high-performance computational resources provided by the Louisiana State University (HPC@LSU; http://www.hpc.lsu.edu) and the Louisiana Optical Network Institute (LONI; http://www.loni.org). We thank Dr. Wei Feinstein who read the manuscript and provided critical comments.
Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This work was supported by the Louisiana Board of Regents through the Board of Regents Support Fund [contract LEQSF(2012-15)-RD-A-05].
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Pauling L. Modern structural chemistry. Nobel Lecture: December 11, 1954.
2. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014;42:D756–763.

3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
4. Guo JT, Ellrott K, Xu Y. A historical perspective of template-based protein structure prediction. Methods Mol Biol 2008;413:3–42.
5. Dorn M, E Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: methods and computational strategies. Comput Biol Chem 2014;53PB:251–76.
6. Honig B. Protein folding: from the levinthal paradox to structure prediction. J Mol Biol 1999;293:283–93.
7. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–5.
8. Zhang J, Li W, Wang J, Qin M, Wu L, Yan Z, et al. Protein folding simulations: from coarse-grained model to all-atom model. IUBMB Life 2009;61:627–43.
9. Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. Proteins 2014;82:Suppl 2:164–74.
10. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y. Assessment of CASP8 structure predictions for template free targets. Proteins 2009;77:Suppl 9:50–65.
11. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins 2011;79:Suppl 10:59–73.
12. Tai CH, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. Proteins 2014;82:Suppl 2:57–83.
13. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009;77:Suppl 9:18–28.
14. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins 2014;82:Suppl 2:43–56.
15. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins 2011;79:Suppl 10:37–58.
16. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16:172–7.
17. Lushington GH. Comparative modeling of proteins. Methods Mol Biol 2015;1215:309–30.
18. Qu X, Swanson R, Day R, Tsai J. A guide to template based structure prediction. Curr Protein Pept Sci 2009;10:270–85.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
20. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. Biol Direct 2012;7:12.
21. Biegert A, Soding J. Sequence context-specific profiles for homology searching. Proc Natl Acad Sci USA 2009;106:3770–5.
22. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–8.
23. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
24. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–9.
25. Joseph AP, de Brevern AG. From local structure to a global framework: recognition of protein folds. J R Soc Interface 2014;11:20131147.
26. Koonin EV, Wolf YI, Aravind L. Protein fold recognition using sequence profiles and its application in structural genomics. Adv Protein Chem 2000;54:245–75.
27. Bennett-Lovsey RM, Herbert AD, Sternberg MJ, Kelley LA. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins 2008;70:611–25.
28. Peng J, Xu J. Low-homology protein threading. Bioinformatics 2010;26:i294–300.
29. Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72:547–56.
30. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 2003;1:95–117.
31. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27:2076–2082.
32. Brylinski M, Lingam D. eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. PLoS One 2012;7:e50200.
33. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35:3375–82.
34. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today 2004;9:659–69.
35. Liu T, Tang GW, Capriotti E. Comparative modeling: the state of the art and protein drug target structure prediction. Comb Chem High Throughput Screen 2011;14:532–47.
36. Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H. Protein structure prediction in structure based drug design. Curr Med Chem 2004;11:551–8.
37. Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol 2009;19:145–55.
38. Brylinski M. Nonlinear scoring functions for similarity-based ligand docking and binding affinity prediction. J Chem Inf Model 2013;53:3097–112.
39. Brylinski M. eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models. PLoS Comput Biol 2014;10:e1003829.
40. Skolnick J, Zhou H, Brylinski M. Further evidence for the likely completeness of the library of solved single domain protein structures. J Phys Chem B 2012;116:6654–64.
41. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA 2006;103:2605–10.
42. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 2005;102:1029–34.
43. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform 2002;3:275–84.
44. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–66.
45. Yan Y, Moult J. Protein family clustering for structural genomics. J Mol Biol 2005;353:744–59.
46. Grabowski M, Joachimiak A, Otwinowski Z, Minor W. Structural genomics: keeping up with expanding knowledge of the protein universe. Curr Opin Struct Biol 2007;17:347–53.
47. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–36.
48. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–60.

49. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907.
50. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. How community has shaped the Protein Data Bank. Structure 2013;21:1485–91.
51. Campbell ID. Timeline: the march of structural biology. Nat Rev Mol Cell Biol 2002;3:377–81.
52. Li W, Godzik A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
53. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 2008;9:531.
54. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–10.
55. Cormen TH, Leiserson CE, Rivest RL, Stein C. Greedy algorithms. Introduction to algorithms. MIT Press, 1990:414.
56. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score=0.5? Bioinformatics 2010;26:889–95.