Probabilistic divergence of a TBM methodology from the ideal protocol

Ashish Runthala
Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
Email: ashish.Runthala@gmail.com

Abstract: Protein structural information is essential for the detailed mapping of a functional protein network. For a higher modelling accuracy and quicker implementation, template based algorithms have been extensively deployed and redefined. The methods only assess the predicted structure against its native state/template, and do not estimate the accuracy for each modelling step. A divergence measure is postulated to estimate the modelling accuracy against its theoretical optimal benchmark. By freezing the domain boundaries, the divergence measures are predicted for the most crucial steps of a modelling algorithm. To precisely refine the score using weighting constants, big data analysis could further be deployed.

Keywords: Modelling, TBM, FM, Model, Accuracy

1. Introduction

Topological detail of a protein provides valuable resource for the routinely deployed life-science studies including the annotation of disulphide connectivity pattern [1,2], functional annotation [3-5], molecular docking and cross-talk based studies [6,7], enzyme design [8,9] and drug development [10]. For technical and resource limitations, the structure determination methodologies have failed to bridge the ever-increasing sequence-structure gap of the protein sequences [11], and have started deploying the predicted models for building more accurate solutions [12].

Protein structure prediction has thus been the prime topic of research for several years. The widespread interest is not only because of the incompetency of structure determination methods to quickly solve the protein structures, but it is also attributable to numerous tribulations impeding our modelling accuracy. As per the worldwide test critical assessment of structure prediction (CASP), the modelling algorithms have been classified into free modelling (FM) and template-based modelling (TBM) which encompasses threading and comparative modelling protocols [13,14]. While the FM methods deploy the knowledge based potentials and physiochemical interactions and attempts to maximally search the conformational space possible for a protein sequence (target) for constructing its lowest-energy model [15], the TBM algorithms deploy the experimentally solved protein structures (*templates*) to construct a protein model. The TBM strategy is developed on basis of observation that highly similar protein sequences usually encode the same topology [16], and structural folds are robustly conserved across the functional diversifications [17,18]. It uses a reliable set of conserved structural folds to model a protein sequence [19,20]. Although, a probabilistically estimated threading methodology has recently attempted to estimate modelling accuracy for the selected templates and algorithmic constraints, it could not succeed for the absence of a robust assessment protocol [21]. A more accurate protocol to estimate the accuracy of a modelling protocol is still the need of the hour to consistently build more accurate protein models for all protein sequences.

2. TBM algorithm

An accurate TBM algorithm holds the key to consistently model a reliable protein structure for every target. However, against the available set of templates, its practical utility is dependent on the level of attainable accuracy of each of its steps. Although its every step holds its own significance to build an accurate model, a few steps viz. template search, template selection and combination, target-template alignment and model assessment are the most crucial ones to fix up the modelling accuracy, as has also been recently observed [22,23].

2.1 Template search

As a correct template set maximally spans the conformational space for a target sequence and serves as the basis to construct a near-native structure, it is crucial to significantly cover a target through functionally similar templates for building its near-native model [23]. Many template search matrices including the target-template sequence identity are not applicable in every modelling instance [24]. It is also suggested that some templates are prone to yielding false positive and erroneous matches for many target sequences, and thus create inconsistencies when a novel fold is being predicted with conserved sequence chunks [25,26]. The probabilistic estimation of the credibility of templates is still not sufficient to detect only the meaningful hits, as resulted for HHPred [27,28], and its threshold needs to be correctly defined to discriminate the functionally and topologically similar templates from the spurious hits. Although difficult to consistently screen the correct template(s), the profile-alignment based algorithms including PSI-BLAST [29], combinatorial extension algorithms [30], profile-HMMs [31], Jackhmmer [32] and HHblits [28] are usually deployed, and their threshold values for various scoring parameters have a great impact on the accuracy. A stringent and weaker template search threshold orderly fails to sensitively span the appropriate distant templates and exclude the distantly related and unreliable hits from the constructed profile [33]. Though CASP13 has extensively deployed the deep learning strategies over CASP12, the average GDT-TS scores of difficult/template-free target models is only found to increase to 65.7 from 52.9 (http://predictioncenter.org/casp13/). Hence, selecting good templates is the still the major challenging step to build highly accurate protein models.

2.2 Template selection and combination

For a relatively simple target with one/two continuous/discontinuous structural domains, multiple templates are easily available, and it makes the selection of a few evolutionarily closest biological hits highly complicated. On the other hand, when there is no significantly detectable sequence relationship among templates for a target, selection of correct templates again becomes a major limiting factor for the accuracy of the a TBM algorithm. Recently, CASP13 (http://predictioncenter.org/casp13/index.cgi), has orderly tested 13, 46 and 37 easy, single domain and unsplit domain targets sharing an almost complete coverage against multiple templates. The template selection has been shown to be especially difficult when such targets encode a novel fold with conserved substructures [26,34]. The easy targets include the single domain of a monomer or the compact domain of a multimeric structure. Although the model topology constructed through incorrect templates can never be energetically relaxed to the native structure [35], the topologically close models have been

remarkably refined in CASP12 [36]. Hence, a correct selection of multiple templates, highly similar to each other for complementary augmentation, becomes essential to improve the overall target modelling accuracy [23]. The probability to predict such a set of correct, mutually complimentary and structurally similar templates becomes significant. Hence a good logically justified probability of such a crucial modelling step is needed.

2.3 Constructing a reliable target-template alignment

The modelling accuracy of a target is significantly dependent on the accuracy of its alignment against the template(s), which in turn depends on the proportion of correctly aligned and biologically significant residues. Profile-HMM-based algorithms deploy PFAM to avoid shifted alignment errors [31] and consider unaligned target chunks as single long insertions, and this hampers the overall modelling accuracy. As a reference guide, the structural information, inherent to the selected templates, has been also deployed by several protocols including combinatorial extension and HHPred [30,27]. However, most of these alignments need manual verification of the number, length and location of gaps, especially when a TBM-hard target is considered. The manually curated alignment becomes very useful to correct gaps incorporated in secondary structure elements, on the basis of sequence similarity and the secondary structure probability of the aligned residue chunks. Many modelling algorithms like MULTICOM [37,38] iteratively revisit the template selection and alignment steps to remove extraneous templates, or to overcome gaps localized in conserved regions.

A threading alignment usually attempts to be as accurate as a curated alignment, and through the maximum accuracy alignment algorithm, it has been shown to improve the accuracy of HMM-profile over the usual Viterbi scoring [39,28]. As the protein structures are robustly conserved over the sequence variations [40], many algorithms like MULTICOM [37,38] consider alignment ensemble to remove the irrelevant templates for overcoming the gaps bisecting the secondary structure/conserved regions. Hence, a probabilistic formulation is required to estimate the likelihood of a target-template alignment.

2.4 Model assessment

To select the most accurate model, the constructed/sampled decoy(s) are usually assessed against template/native structure on the basis of topological measures including root mean square deviation [41], template modeling score [42], global distance test score [43,44], contact area difference score [45], local distance difference test [46], SphereGrinder [47] and angular differences between successive atomic planes [48], and through solvent accessible surface area, hydrogen bonding network, pair-wise atomic interactions and molecular packing [49]. The global [50] and local [46,45] assessment scores are orderly used to select a topologically more accurate structure and find the inaccurately modelled regions for guiding the conformational sampling, usually deployed to reduce the non-physical atomic clashes [51]. However, even within an inadequately sampled landscape, multiple sampled models are usually close to the native topology [52,53], and for consistently selecting the most accurate model, a probabilistic estimation is needed to predict the likelihood of a protein model to be the correct near-native conformation.

Although a model-sampling-cum-assessment protocol has been developed through a probabilistic search [54], it could not prove to be successful. Although it shows that a soft-energetic bias steers the model sampling towards a diverse decoy dataset less prone to

energetic artifact structures, it also demonstrates the algorithmic limitations of the usually deployed energy functions Hamiltonian and Rosetta. Recently in CASP13, for an incompetent assessment protocol, D-Haven could only construct a model with a GDT_TS score of 0.8293 for the target T0955 in contrast to the respective score of 0.9512 of the top-ranked structure predicted by MESHI's lab (http://predictioncenter.org/casp13/).
As an attempt to deploy several diverse protocols more reliably, numerous meta-server algorithms have been developed for various modelling steps [55-57] as a hopeful attempt to deploy the majority of algorithms for building more accurate models. However, for consistently predicting the reliable models for all targets, the TBM algorithm still needs more accurate measures for evaluating the credibility of the modelling steps, in terms of the divergence score. It would empower our algorithms to quickly proofread the modelling strategy and rebuild an improved model. The further empowered TBM methodologies would provide credible models as the scaffolds to determine more accurate structures from the low-resolution experimental methodologies.

3. Probabilistic assessment of modelling steps

The standard practice of selecting the top-ranked templates is not consistently correct with a single unique algorithm. As TBM algorithms usually deploy a common set of independently implemented steps, it is possible to define an independent probabilistic assessment score for each step. Since a modelling strategy does not have an absolute assessment measure for every modelling step, we resort to estimating a measure of divergence for a modelling protocol from the results of computationally-feasible and the best-possible protocol with the highest modelling accuracy. Evaluating the modelling strategy of a predicted model this way against the one for the best possible theoretically feasible algorithmic structure, the credibility of a modelling methodology/step could be estimated. The score would be useful to find the most accurate protocol for each modelling step and improve the modelling accuracy.

3.1. Template search

It is well proven that the single-domain protein sequences, encoding utmost 200 residues, could have an average coverage of 70% against the templates and the predicted models could be within an RMSD of 5Å from their native state [58,18]. A target sequence usually shares a notable similarity against several templates/substructures, although the perfect set(s) of the non-redundant template(s) could maximally include a subset of N templates, including all the colocalized set of structural isoforms/topologically close structures. It is also logical to interpret that the likelihood of correctly screening such ideal entries is inversely proportional to the profile diversity [59].
Through profile-based strategies [27,60], integrated with the contact-map network [61] and deep-learning strategies [57], the templates are usually searched through the expanded sequence databases. An effective measure based on the colocalized alignment gaps, structural compactness and interactions with domain parsing need to be additionally assessed to evaluate the quality of an alignment for prioritizing the relevant templates only. As the effective template-scoring function/strategy is still missing, the protocol that screens the best set of templates as the top-ranked hits is actually the ideal one for a target sequence. While an ideal template is the experimentally determined target structure, the best set should be the union of all functionally related structures through which a near-native target model could be

reliably constructed. However, the modelling methodologies fail to consistently screen this complete set, and as a tradeoff, heuristically fix a different set (K) to span the target to result into incorrect model topology. Hence, the logical divergence (Div_{search}) for this step should be an absolute difference of the two sets and it should drastically increase when the additional/lesser templates are screened by a template-search algorithm, as formulated in the following equation I.

$$Div_{search} = \frac{|K - N|}{N \bigcup K} \qquad \text{- (I)}$$

3.2. Template selection and combination

It clusters the screened templates as per the alignment location over the target sequence and selects minimal number of top-ranking and structurally complementary hits to gain the maximum coverage. An ideal template-set specifically differs in size and for entries. As each template-set, spanning a specific target segment, is independent of another group, the divergence measure of a modelling protocol can be reasonably approximated. If C_n builds the set of n groups each having a total of a_1, a_2, a_3… a_n templates, and for every i^{th} cluster, k_i and n_i orderly defines the set of templates selected by a modelling protocol against the best possible set that could have built more accurate model, the logical divergence of this step ($Div_{selection}$) can be approximated as equation II.

$$Div_{selection} = 1 - \prod_{i=1}^{Cn} \frac{|k_i - n_i|}{(n_i \cup k_i)} \qquad \text{- (II)}$$

3.3. Constructing a reliable target-template alignment

For building a biologically meaningful target-template alignment and constructing a near-native target structure, the sequence similarity measure is usually deployed. However, when sequence similarity decreases, the evolutionary relationship among the related proteins cannot be reliably inferred [62,63]. Although the low-homology sequence alignments are curated through protocols like SFESA [64], the algorithm consistently constructing the most-optimal alignment is still not developed, i.e. the same protocol seldom builds the high accuracy model for the usual/CASP target sequence and its domains [23].

To estimate the divergence of the selected set of templates, a target-template alignment could be parsed into domains on the basis of Grishin plots [65] or DeepmetaPSICOV [66] or the alignment position against the number of gaps. As it could be accessed through a per-residue scoring difference for a common set of domains, the credibility of an alignment could be estimated for a template-dataset through the following equation, and it would be able to robustly assess even the partial accuracy of an alignment. If k_j and n_j orderly builds the set of templates selected by an algorithm against the ideal methodology for the jth domain encoding l_j residues, and s_{aci} and s_{opti} is the score of the ith residue for the jth domain of the selected alignment against the ideal structural alignment, the divergence (Div_{ali}) of a d-domain alignment can be estimated as:

$$Div_{ali} = 1 - \prod_{j=1}^{d} \{1 - \sum_{i=1}^{l_j} \frac{|(s_{aci} - s_{opti})|}{l_j}\} \quad \text{-(III)}$$

3.4. Model assessment

A global assessment of the constructed decoys helps to select the best model and guide the conformational sampling. It is still a major problem to select the most-accurate predicted model from the generated ones, especially if is a non-CASP target. All available measures are not consistently and unanimously correct to select the most accurate model for a target sequence. When domain orientation of a prediction is substantially variant than the closest template(s), measures like lDDT have proven to be more accurate, as recently observed for 11/65 targets in CASP13 (www.predictioncenter.org/CASP13/). However, for 43/65 oligomeric CASP13 targets, the assessment methodology is only found fairly competent for the monomer models and overall topology is not reliably predicted.

As has been repeatedly observed in all the CASPs, the best human predictions are found to be more accurate than the best TS servers which have only constructed structures of more consensus topology. The methodologies, building such an overall average topology for the top-ranked CASP server models, are not desirable for the advancement in the field. Despite the fact that for CASP13 target T0984, HHSearch could screen many templates with atleast 75% template-coverage, the best predicted model has only showed a GDT-TS and LDDT score of 57.94 (*monomeric*) and 64 (*oligomeric*). Furthermore, for these TBM-easy targets T0984-D1 and T0984-D2, the top-ranked models of KIAS-Gdansk and McGuffin has shown a GDT-TS score of 67.32 and 76.31 respectively, and it has again signified the requirement of a more robust assessment protocol.

Even when the models possess a global topological accuracy, the local stereochemistry should be correct and hence, the assessment team should work out at two levels. The conformational space of the models should be extensively sampled and it requires extensive data from both the modelling and assessment protocols. Therefore, for ignoring the near-native decoys (sd) and selecting the structurally divergent models from the constructed ones (cgd), the divergence of this step (Div_{ma}) could be defined as

$$Div_{ma} = 1 - \frac{sd}{cgd} \quad \text{-(IV)}$$

It has been recently quoted in CASP13 meeting that "Should the evaluation focus much more explicitly on the details of the model?" and it is implicated that the modelling protocol should primarily evaluate the topology, and if it is satisfactorily closer to the near-native conformation for almost the functionally important/conserved residues, the physicochemical details should be assessed.

As a linear scoring function fails to accurately assess the interdependency among the deployed sequence/structural parameters, the natural logarithm of the individual divergence scores should be considered to estimate the overall divergence of a modelling protocol. Considering α1-α4 as the four orderly defined normalization coefficients for the key parameters, the overall divergence can be postulated to be Div_{tot}. Strategically refining our recently published modelling protocol [23] along with the divergence scores (*In progress*),

with no extensive sampling, it shows a significantly higher modelling accuracy for a small randomly selected test-dataset of 5 CASP13 targets
(Table1, http://predictioncenter.org/casp13/).

$$Div_{tot} = 1 - [\alpha_1(1 - Div_{search}) x \alpha_2(1 - Div_{selection}) x \alpha 3(1 - Div_{ali}) x \alpha 4(1 - Div_{ma})]$$

Target (length)	Top-ranked CASP13 model			Our model		Modelling accuracy improvement over the best CASP model
	Algorithm	Templates	TM-Score/ GDT-TS/ RMSD	Templates	TM-Score/GDT-TS/ RMSD	
T0951_D1 (277)	IntFOLD5	3W04A, 5CBKA, 3W06A,5DNUA, 1WOMA, 3QVMA	0.971/93.284/ 0.952	5CBKA	**0.977/94.683/ 0.984**	0.006/1.399/ 0.032
T0961_D1 (505)	BAKER	4Y9JB	0.973/89.761/ 1.426	4Y9JA	0.963/86.282/ 1.544	-0.010/-3.479/ 0.118
T0967_D1 (81)	BAKER-ROSETTA SERVER	3W8G_B	0.905/90.190/ 1.031	5HOKD	**0.910/90.506/ 0.980**	0.005/0.316/ -0.051
T0971 (186)	BAKER-ROSETTA SERVER	4XBX_B	0.972/95.769/ 0.729	3EBTA	**0.978/97.663/ 0.642**	0.006/1.894/ -0.087
T0976 (252)	DELClab	2HHG_A, 1YT8_A	0.687/41.224/ 3.681	1YT8A	**0.768/51.531/ 2.798**	0.081/10.307/ -0.883
T0976_D1 (120)	DELClab	2HHG_A, 1YT8_A	0.916/87.607/ 1.206		0.780/68.376/ 1.728	-0.136/-19.231/ 0.522
T0976_D2 (124)	A7D	N/A	0.892/83.130/ 1.316		0.819/72.967/ 1.793	-0.073/-10.163/ 0.477

Table1: Model assessment results for the 5 randomly selected CASP13 targets
A more accurate prediction is highlighted with boldcase letters. The developed strategy has shown a promisingly higher modelling accuracy than the best CASP13 models, and a negative score indicates superior CASP13 model.

3. Discussion

As observed in various studies [50,67], a more-robust assessment protocol seems a necessity for predicting the near-native protein structure. The earlier studies have sought to screen the topological/physicochemical errors in the predicted models for selecting the accurate structures [49,11,23]. To the best of our knowledge, no previous assessment protocol has probabilistically evaluated the modelling accuracy of a protocol. For estimating the accuracy of a modelling protocol, our postulated equations unanimously rank the complete modelling strategy and do not simply assess a lastly predicted model against its native state/template.
The modelling steps are analyzed under the given framework of quantifying internal consistency with respect to a scoring function. For instance, |K- N| and |N- K| variables within the template search equation define the number of excessive/trivial and omitted templates, as it has recently been shown that the accurate templates are often missed by even the top-ranked template search protocols [23]. Thus, the over-/under-selection will be

appropriately dealt with this formula. Lastly, an equivalent assessment of all scoring terms is considered for an initial testing.

The best possible or most optimal algorithm is expected to deploy the best possible set of structures through the biologically meaningful alignment to yield the highest modelling accuracy, and in order that our analysis remains general, we compare this optimality at every step to another algorithm to evaluate the likelihood of its credibility. All assessments between procedures are made strictly under the premise of the same scoring function for both procedures, and without this assumption, there cannot be any kind of parity between the results of the two procedures. However, the divergence of a modelling methodology from the theoretical optimal protocol does not indicate the possibility of a more accurate algorithm. This is because the optimal procedure is computationally intractable for all practical purposes and therefore divergence of modelling results from the most optimal protocol is a measure to define the ranking methodology for each step so that the ill-performing steps could be curated. The lower the overall divergence score, closer is the prediction methodology to the ideally-best applicable algorithm. For the reason that some of the steps can affect the modelling accuracy much more significantly than the others, a well-trained set of different normalization coefficients is hereby required in the derived probabilistic equation. A major limitation of the method is that the coefficients ($\alpha 1$- $\alpha 4$) are strictly based on the deployed protocol. To add greater significance to these constants, big data analysis techniques such as machine learning should be further deployed to gather CASP information from the existing results of structure prediction algorithms and assign suitable values to the constants.

A combinatorial divergence score is parameterized to select the more accurate models and is logically shown to be more promising than the current assessment measures. Evaluating the modelling results for 5 randomly selected CASP13 targets against their native structures (*Parsed only for the officially assessed domain boundaries*), it indicates that our template-ranking strategy [23] has substantially improved the modelling accuracy. However, failure for some targets suggests the requirement of a threading protocol. Our article builds a chassis for developing a more accurate as well as automated modelling protocol that iteratively fixes a specific modelling step for each target sequence through the available sequence/structural constraints.

Conclusion

As it is well understood that 51 tokens can be best delivered through two tokens: 50 and 1, and attempting any other combination is simply sub-optimal. We have described a standard decomposition of any protein structure prediction procedure and have constructed a divergence procedure for every modelling step. The proposed function works through comparison of a deployed strategy against the best theoretical protocol. As it is difficult to trace the algorithmic errors at every step of a modelling protocol, the strategy lays a logical chassis to develop a more accurate assessment measure that heuristically refine the modelling algorithm and consistently build a near-native model for every target.

Acknowledgements The author acknowledges the University and department for providing the required resources/support.

In compliance with the ethical standards

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Funding: It is not a funded research.

Availability of data and material The constructed files/datasets analyzed in this study are available from the corresponding author on reasonable request.

References

1. Becker J, Maes F, Wehenkel L (2013) On the relevance of sophisticated structural annotations for disulfide connectivity pattern prediction. PLoS One 8 (2):e56621. doi:10.1371/journal.pone.0056621
2. Tam JP, Nguyen GKT, Loo S, Wang S, Yang D, Kam A (2018) Ginsentides: Cysteine and Glycine-rich Peptides from the Ginseng Family with Unusual Disulfide Connectivity. Scientific Reports 8 (1):16201. doi:10.1038/s41598-018-33894-x
3. Torng W, Altman RB (2019) High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics 35 (9):1503-1512. doi:10.1093/bioinformatics/bty813
4. Manohar Lal KUS, Neelam Mahala, Ashish Runthala, and Uma S. Dubey (2020) Camel milk α-lactalbumin as a potential anticancer molecule: A Bioinformatics analysis. Research Journal of Pharmaceutical, Biological and Chemical Sciences 11 (3):38-52. doi:10.33887/rjpbcs/2020.11.3.5
5. Kamjula V, Kanneganti A, Metla R, Nidamanuri K, Idupulapati S, Runthala A (2020) Decoding the vital segments in human ATP-dependent RNA helicase. Bioinformation 16 (2):160-170. doi:10.6026/97320630016160
6. Lapillo M, Tuccinardi T, Martinelli A, Macchia M, Giordano A, Poli G (2019) Extensive Reliability Evaluation of Docking-Based Target-Fishing Strategies. International Journal of Molecular Sciences 20 (5). doi:10.3390/ijms20051023
7. Jamithireddy AK, Runthala A, B G (2020) Evaluation of specificity determinants in Mycobacterium tuberculosis /anti- factor interactions. Biochemical and Biophysical Research Communications 521 (4):900-906. doi:10.1016/j.bbrc.2019.10.198
8. Amrein BA, Runthala A, Kamerlin SCL (2019) In-Silico-Directed Evolution Using CADEE. Computational Methods in Protein Evolution, Methods in Molecular Biology 1851:381-415. doi:10.1007/978-1-4939-8736-8_22
9. Phulara SC, Rajput VS, Mazumdar B, Runthala A (2020) Metabolic and Enzyme Engineering for the Microbial Production of Anticancer Terpenoids. In: 'Essentials of Cancer Genomic, Computational Approaches and Precision Medicine. pp 237-259. doi:10.1007/978-981-15-1067-0_10
10. Raghavan V, Agrahari M, Gowda DK (2019) Virtual screening of p53 mutants reveals Y220S as an additional rescue drug target for PhiKan083 with higher binding characteristics. Computational Biology and Chemistry 80:398-408. doi:10.1016/j.compbiolchem.2019.05.005
11. Runthala A (2012) Protein structure prediction: challenging targets for CASP10. Journal of Biomolecular Structure and Dynamics 30 (5):607-615. doi:10.1080/07391102.2012.687526
12. Ogorzalek TL, Hura GL, Belsom A, Burnett KH, Kryshtafovych A, Tainer JA, Rappsilber J, Tsutakawa SE, Fidelis K (2018) Small angle X-ray scattering and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy. PROTEINS: Structure, Function, and Bioinformatics 86 Suppl 1:202-214. doi:10.1002/prot.25452
13. Cheng J (2008) A multi-template combination algorithm for protein comparative modeling. BMC Structural Biology 8:18. doi:10.1186/1472-6807-8-18
14. Ma J, Wang S, Zhao F, Xu J (2013) Protein threading using context-specific alignment potential. Bioinformatics 29 (13):i257-265. doi:10.1093/bioinformatics/btt210
15. Corey Hardin TVP, Zaida Luthey-Schulten (2002) Ab initio protein structure prediction. Current Opinion in Structural Biology 12:176–181
16. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cellular and Molecular Life Sciences 60 (12):2637-2650. doi:10.1007/s00018-003-3114-8
17. Lisa N Kinch NVG (2002) Evolution of protein structures and functions. Current Opinion in Structural Biology 12:400–408
18. Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102 (4):1029-1034. doi:10.1073/pnas.0407152101

19. Kryshtafovych A, Venclovas C, Fidelis K, Moult J (2005) Progress over the first decade of CASP experiments. PROTEINS: Structure, Function, and Bioinformatics 61 Suppl 7:225-236. doi:10.1002/prot.20740
20. Kolodny R, Pereyaslavets L, Samson AO, Levitt M (2013) On the universe of protein folds. Annual Review of Biophysics 42:559-582. doi:10.1146/annurev-biophys-083012-130432
21. Simoncini D, Berenger F, Shrestha R, Zhang KY (2012) A probabilistic fragment-based protein structure prediction algorithm. PLoS One 7 (7):e38799. doi:10.1371/journal.pone.0038799
22. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A (2018) Evaluation of the template-based modeling in CASP12. Proteins: Structure, Function, and Bioinformatics 86:321-334. doi:10.1002/prot.25425
23. Runthala A, Chowdhury S (2019) Refined template selection and combination algorithm significantly improves template-based modeling accuracy. Journal of Bioinformatics and Computational Biology 17 (02):1950006. doi:10.1142/s0219720019500069
24. Anna Tramontano VM (2003) Assessment of Homology-Based Predictions in CASP5. PROTEINS: Structure, Function, and Genetics 53:352–368
25. Dunbrack RL, Jr. (2006) Sequence comparison and protein structure prediction. Current Opinion in Structural Biology 16 (3):374-384. doi:10.1016/j.sbi.2006.05.006
26. Ralf Jauch HCY, Prasanna R. Kolatkar, Neil D. Clarke (2007) Assessment of CASP7 structure predictions for template free targets. PROTEINS: Structure, Function, and Bioinformatics 69(Suppl 8):57–67. doi:10.1002/prot.21771
27. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951-960. doi:10.1093/bioinformatics/bti125
28. Remmert M, Biegert A, Hauser A, Soding J (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9 (2):173-175. doi:10.1038/nmeth.1818
29. Stephen F. Altschul TLM, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller and David J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25 (17):3389–3402
30. Ilya N.Shindyalov PEB (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11 (9):739–747
31. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Research 32 (Database issue):D138-141. doi:10.1093/nar/gkh121
32. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Computational Biology 7 (10):e1002195. doi:10.1371/journal.pcbi.1002195
33. Pagnuco IA, Revuelta MV, Bondino HG, Brun M, Ten Have A (2018) HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS One 13 (3):e0193757. doi:10.1371/journal.pone.0193757
34. Wang Y, Wang J, Li R, Shi Q, Xue Z, Zhang Y (2017) ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. Nucleic Acids Research 45 (W1):W400-W407. doi:10.1093/nar/gkx410
35. Mirjalili V, Feig M (2013) Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. Journal of Chemical Theory and Computation 9 (2):1294-1303. doi:10.1021/ct300962x
36. Hovan L, Oleinikovas V, Yalinca H, Kryshtafovych A, Saladino G, Gervasio FL (2018) Assessment of the model refinement category in CASP12. PROTEINS: Structure, Function, and Bioinformatics 86 Suppl 1:152-167. doi:10.1002/prot.25409
37. Wang Z, Eickholt J, Cheng J (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 26 (7):882-888. doi:10.1093/bioinformatics/btq058
38. Jianlin Cheng JL, Zheng Wang, Jesse Eickholt and Xin Deng (2012) The MULTICOM toolbox for protein structure prediction. BMC Bioinformatics 13:65
39. Biegert A, Soding J (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. Bioinformatics 24 (6):807-814. doi:10.1093/bioinformatics/btn039
40. Friedberg I, Margalit H (2002) Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function. Protein Science 11 (2):350-360. doi:10.1110/ps.18602

41. KABSCH W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A A32:922-923. doi:10.1107/S0567739476001873
42. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. PROTEINS: Structure, Function, and Bioinformatics 57 (4):702-710. doi:10.1002/prot.20264
43. Adam Zemla CV, John Moult, Krzysztof Fidelis (1999) Processing and Analysis ofCASP3 Protein Structure Predictions. PROTEINS: Structure, Function, and Genetics 3:22–29. doi:10.1002/(sici)1097-0134(1999)37:3+<22::aid-prot5>3.3.co;2-n
44. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Research 31 (13):3370-3374. doi:10.1093/nar/gkg571
45. Olechnovic K, Kulberkyte E, Venclovas C (2013) CAD-score: a new contact area difference-based function for evaluation of protein structural models. PROTEINS: Structure, Function, and Bioinformatics 81 (1):149-162. doi:10.1002/prot.24172
46. Mariani V, Biasini M, Barbato A, Schwede T (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29 (21):2722-2728. doi:10.1093/bioinformatics/btt473
47. Kryshtafovych A, Monastyrskyy B, Fidelis K (2014) CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. PROTEINS: Structure, Function, and Bioinformatics 82 Suppl 2:7-13. doi:10.1002/prot.24399
48. Shikhin Garg SK, Ashish Runthala (2016) Improved protein model ranking through topological assessment. In: Wong K-C (ed) Computational Biology and Bioinformatics: Gene Regulation. CRC Press, pp 406-424
49. Kryshtafovych A, Fidelis K (2009) Protein structure prediction and model quality assessment. Drug Discovery Today 14 (7-8):386-393. doi:10.1016/j.drudis.2008.11.010
50. Jing X, Dong Q (2017) MQAPRank: improved global protein model quality assessment by learning-to-rank. BMC Bioinformatics 18 (1):275. doi:10.1186/s12859-017-1691-z
51. Ashish Runthala SC Iterative optimal TM-score and Z-score guided sampling significantly improves model topology. In: International MultiConference of Engineers and Computer Scientists, Hong Kong, March 12 - 14 2014. IAENG, pp 123-128
52. Ashish Runthala SC (2013) Protein Structure Prediction: Are We There Yet? In: Tuan Pham LJ (ed) Knowledge-Based Systems in Biomedicine. vol SCI450. Springer, Berlin, Heidelberg, pp 79–115. doi:10.1007/978-3-642-33015-5_4
53. Runthala A, Chowdhury S (2016) Unsolved Problems of Ambient Computationally Intelligent TBM Algorithms. Hybrid Soft Computing Approaches, Studies in Computational Intelligence 611:75-105. doi:10.1007/978-81-322-2544-7_3
54. Kevin Molloy SS, and Amarda Shehu (2013) Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab-initio Protein Structure Prediction. IEEE/ACM transactions on computational biology and bioinformatics 10 (5):1162-1175. doi:10.1109/TCBB.2013.29
55. Wallner B, Larsson P, Elofsson A (2007) Pcons.net: protein structure prediction meta server. Nucleic Acids Research 35 (Web Server issue):W369-374. doi:10.1093/nar/gkm319
56. Buchan DWA, Jones DT (2018) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. PROTEINS: Structure, Function, and Bioinformatics 86 Suppl 1:78-83. doi:10.1002/prot.25379
57. Zheng W, Zhang C, Wuyun Q, Pearce R, Li Y, Zhang Y (2019) LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. Nucleic Acids Research 47 (W1):W429-W436. doi:10.1093/nar/gkz384
58. Kihara D, Skolnick J (2003) The PDB is a covering set of small protein structures. Journal of Molecular Biology 334 (4):793-802. doi:10.1016/j.jmb.2003.10.027
59. Panchenko AR (2003) Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Research 31 (2):683-689. doi:10.1093/nar/gkg154
60. Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. PROTEINS: Structure, Function, and Bioinformatics 72 (2):547-556. doi:10.1002/prot.21945
61. Buchan DWA, Jones DT (2017) EigenTHREADER: analogous protein fold recognition by efficient contact map threading. Bioinformatics 33 (17):2684-2690. doi:10.1093/bioinformatics/btx217
62. Rost B (1999) Twilight zone of protein sequence alignments. Protein Engineering 12 (2):85–94. doi:10.1093/protein/12.2.85

63. Rost B (2002) Enzyme Function Less Conserved than Anticipated. Journal of Molecular Biology 318 (2):595-608. doi:10.1016/s0022-2836(02)00016-5
64. Tong J, Pei J, Grishin NV (2015) SFESA: a web server for pairwise alignment refinement by secondary structure shifts. BMC Bioinformatics 16:282. doi:10.1186/s12859-015-0711-0
65. Kinch LN, Shi S, Cheng H, Cong Q, Pei J, Mariani V, Schwede T, Grishin NV (2011) CASP9 target classification. PROTEINS: Structure, Function, and Bioinformatics 79 Suppl 10:21-36. doi:10.1002/prot.23190
66. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Computational Biology 13 (1):e1005324. doi:10.1371/journal.pcbi.1005324
67. West CE, de Oliveira SHP, Deane CM (2019) RFQAmodel: Random Forest Quality Assessment to identify a predicted protein structure in the correct fold. PLoS One 14 (10):e0218149. doi:10.1371/journal.pone.0218149