Journal of
Data Mining in Genomics & Proteomics

Nahata and Runthala, J Data Mining Genomics Proteomics 2016, 7:4
DOI: 10.4172/2153-0602.1000206

Review Article OMICS International

Quick Reliable Exploration of the PDB Universe Seeks a New Template Search Algorithm

Sunil Nahata and Ashish Runthala*

Department of Biological Sciences, Birla Institute of Technology & Science, Pilani, Rajasthan, India

Abstract

Near-native protein structure prediction through Template Based Modelling (TBM) has been a major realistic goal of structural biology for several years. The TBM algorithms require the best-set of templates for a target protein sequence to maximally cover it and construct its correct topology. However, the accuracy of such prediction algorithms suffers from the algorithmic and logical problems of our template search measures which fail to quickly screen reliable structures for a target sequence. In this study, we employ the culled PDB95 dataset of 41,967 templates to predict the CASP10 target T0752 models for assessing the efficiency of the usually employ search engines PSI-BLAST and HHPred. Our analysis presents a detailed study in order to open new vistas for improving the accuracy of TBM prediction methodologies. It reveals weaknesses of most popular template search measures and thereby briefly provides a significant insight into the qualities of a foreseen template search algorithm to illustrate the need for a more reliable template search algorithm.

Keywords: PDB; PSI-BLAST; HHPred; Template; GDT; HHPred; MODELLER; TM_Score

Introduction

At molecular scale in life, proteins are the major functional molecules and are considered the building blocks of all life forms on earth. Smooth mapping of functional network of proteins in a cell so calls for the knowledge of their detailed structural conformation.

Currently (as of April 12th, 2016), 117,651 experimental protein structures have been released by the Protein Data Bank (PDB), while another 11,484 structures have been submitted and are awaiting release [1]. However, this number of structurally characterized proteins is way undersized when compared to 550,740 annotated sequences in UniProtKB/Swissprot database and 63,039,659 sequences in the entire UniProtKB/TrEMBL database. Even after removing the inferred homologous sequences from this database, the remaining 62,488,619 sequences also exceed the total count of their experimentally solved structures. So a major count of protein sequences does not have their experimental structures determined and this sequence-structure gap has been constantly increasing despite the development of dedicated X-ray crystallography pipelines. It has lastly inspired us to exploit the existing set of already solved protein structures as templates through TBM methodologies for predicting the reliable target models [2]. These algorithms utilize the target-template evolutionary relationship on the basis of fact that evolutionary related sequences share a similar structural topology, and are thus being constantly developed and redefined to reach experimental accuracy [3,4].

A TBM algorithm usually involve a fixed set of steps, viz., template search, template selection, construction of a target-template alignment, model building and lastly model assessment. The template search and selection step is the primary step for improvising the accuracy of the predicted model for a target sequence and thus several template search and selection algorithms including HHPred [5], COMA [6] and PSIPRED [7] have been developed. Further, an extensive improvisation of modelling has been performed through algorithms like MODELLER [8]. Knowledge based scoring functions such as TM_Score [9] have even been developed for selecting the highly accurate conformational decoy from the sampled set of models. It has been observed that bad models constructed through selection of inaccurate templates could rarely be structurally improved. Hence, template search and selection becomes the most important factor to decide the accuracy of target model prediction.

Template search and selection step aligns all the known protein structures to yield a well aligned and a reliable template hit. A daunting task it might seem to be, but being able to come across templates with a sequence identity more than 40% for the selected target can be counted upon as a chance to yield a well predicted model. But finding such a high sequence identity for most of the target sequences is a rarity. At times, the screened templates might also return false positives along with localized residue similarities to targets which might lead to the unreliable alignments [10]. Culling of the PDB database at various thresholds of sequence identity to select the correct template for a specific target is also performed [11].

Template search for a target sequence is normally done through position specific iterative - basic local alignment search tool (PSI-BLAST) [12] which is a sequence-profile based algorithm. Such template search algorithms are scored through the statistically significant level of most probable target-template aligned residue substitutions and usually employ E-value as the scoring measure for the template credibility. Despite this rigorous calculation, template sequences sometimes prove to be the false positive or spurious similarity at several local chunks against the target sequence. It is because the scoring matrices like BLOSUM [13] uniformly consider the same residue substation score at different locations of the target sequence and it poses a huge problem to search and select the distant although quite reliable templates for the considered target sequence. Several methods, like Context-Specific BLAST (CS-BLAST) [14], HMMER [15] and HHPred have thus been developed to solve such problems by considering the sequence and structural context of any residue substitution in a protein sequence. These methods hereby also consider mutation probabilities of residues

***Corresponding author:** Ashish Runthala, Department of Biological Sciences, Birla Institute of Technology & Science, Pilani, , Rajasthan, India, Tel: +911596 515-817; E-mail: ashish.runthala@gmail.com

Received October 10, 2016; **Accepted** October 27, 2016; **Published** October 31, 2016

Citation: Nahata S, Runthala A (2016) Quick Reliable Exploration of the PDB Universe Seeks a New Template Search Algorithm. J Data Mining Genomics Proteomics 7: 206. doi: 10.4172/2153-0602.1000206

Copyright: © 2016 Nahata S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

for insertions or deletions to make the template search further more effective. However, our template search algorithm is still not efficient. Thus solving the existing problem of structure prediction through computational means has become a major challenge for structural biologists. Despite our rigorous efforts, for algorithmic problems existing or unintentionally encoded, current template search and selection methods are insufficient to select correct hits consistently. Here in this article, the template search problem is experimentally proven for a Critical Assessment of Structure Prediction (CASP) target and the plausible reasons for our algorithmic failures are highlighted.

Methods

Selection of the target sequence

A CASP10 TBM-high accuracy (TBM-HA) target sequence T0752 is arbitrarily selected because of its short length of 156 residues and ease in computational processing.

Template set considered for modelling the target sequence

PDB database culled at a 95% sequence identity is recently downloaded on January 16th, 2014 from the MODELLER database (http://salilab.org/modeller/supplemental.html) for screening the best template(s) for T0752. It enables us to consider only a minimal number of 41,967 templates in comparison to the total count of 96,920 hits available in PDB, and saves our computational time and resource. Each of these 41,967 structures is then employed as an individual template to model the selected target and further assessing its accuracy against its experimentally solved native structure. These top-scoring template hits with TM_Score accuracy higher than 0.4 are lastly employed to screen all of their structural isoforms culled at the earlier step by the HHPred algorithm for finally selecting the best available set of templates for a target sequence.

Modelling the target sequence through MODELLER v9.9

The target sequence T0752 is modelled through MODELLER v9.9 with each of the selected 41,967 representative template structures through in-house automated scripts. These models are then evaluated through maxcluster tool based on Maxsub and TM_Score measures for assessing their topological similarity locally as well as globally through a sequence guided structural superimposition against their native structure parsed only for the domain(s) assessed during the CASP [16].

Extending the set of highly accurate templates to their structurally similar hits

For all the templates with even a bit insignificant TM_Score accuracy of 0.4, structurally similar protein structures culled by the MODELLER in the PDB95 dataset are also employed to model the target sequence for a more accurate selection of reliable hits and for predicting the near-native protein model for the considered sequence. Here the differential availability of all these top-scoring hits is further screened in the HHPred and PSI-BLAST results.

Results

Model accuracy of all selected templates

For the selected CASP10 156-residue target sequence, 41,967 representative templates out of 96,920 structures are employed. These templates resulted in quite a scattered pattern of model accuracy, in terms of TM_Score, with most of the hits proving to be futile for a single template based modelling, as expected. This modelling step yielded structures with TM_Score accuracy ranging from 0 to 1, as shown in Table 1. The models constructed through the selected representative templates are grouped at TM_Score intervals of 0.1 in this table.

Template set considered for modelling the target sequence

While majority of the considered templates prove to be insignificant ill-scoring hits with a TM_Score accuracy lesser than 0.3, only 42 templates show a TM_Score accuracy of 0.4, as enlisted in Table 1. Hence, the additional 141 structural isoforms shown for these hits by HHPred, although rejected in the earlier step, are also employed and it increased our considered set of plausibly reliable templates to 183. This set finally resulted in only 24 templates whose individual modelling accuracy surpassed 0.7 in terms of TM_Score, as shown in table 2. The table also represents the presence of these accurate templates in PSI-BLAST or HHPred results as "YES" along with the Global Distance Test-Total Accuracy (GDT-TS) score against the native structure. Table 3 summarizes the assessment scores of these top-24 templates through their pairwise alignments against the target sequence with four other scoring measures viz., the sequence identity percentage, count of the identical residues, BLOSUM62 score and the coverage span.

Discussion

The study is based on the TBM-HA target T0752 for the reason that this CASP10 sequence encodes a reasonably higher and usually encountered domain length of 156 residues, and this sequence is assessed as a single domain segment of 148 residues (2-149) that is the usual maximal length encountered as a single structural domain of a protein sequence. For T0752, the study tries to demonstrate that the existing best template search and selection tools such as PSI-BLAST and HHPred does not screen closely related templates efficiently.

Template search engines such as PSI-BLAST that are based on profile-sequence comparison methods and HHPred that are based on profile HMM algorithms are useful tools to search for a reliable template and to construct its heuristic alignment against a target protein sequence whose structure is yet to be solved experimentally. Template selection error by PSI-BLAST occurs due to the consideration of non-homologous portions at the ends with well-scoring aligned central portions [17]. The homologous over-extension issue is resolved to an extent by HHPred. However, its reliance on high-scoring domains in templates for target sequence results in an overall profile that is enhanced and a better reliable local alignment for only a short portion of the target sequence. Despite the fact that such profile HMMs in HHPred are adequately trained to detect even very distant evolutionary relationships for a target sequence against the templates and that it gives better results by being probabilistic in homolog detection along with an improved alignment [8], these servers are unable to deliver the

TM_Score	Total count of structures
0-0.1	692
0.1-0.2	39776
0.2-0.3	1430
0.3-0.4	27
0.4-0.5	28
0.5-0.6	8
0.6-0.7	3
0.7-0.8	2
0.8-0.9	0
0.9-1	1
Total	41,967

Table 1: TM_Score range of all the predicted models through all the 41967 representative templates.

complete or utmost reliable set of templates, as hereby shown.

Model accuracy of all the selected templates

As clearly enlisted in Table 1, there is a quite expected pattern of accuracy of all the selected representative templates. Here only 1 structure yielded a highly accurate model with TM_Score more than 0.9 and that was certainly the actual answer conformation for the selected target sequence. Excluding this answer conformation, only 14 templates showed TM_Score accuracy higher than a cutoff of 0.5 [18].

Encompassing and analyzing highly accurate templates to all structurally similar hits

The modelling of target sequence through all the 183 structurally similar hits yields only 24 models with a TM_Score accuracy of more than 0.7, as enlisted in Table 2. Out of these 42 templates and 141 structurally similar hits, only 2 templates 4GB5_A (Experimental structure of the target sequence) and 3B8L_A (Top ranked template by HHPred), are shown as representative structures by both the culled PDB95 dataset of MODELLER and PDB70 dataset of HHPred. It thus illustrates the loophole of our template search algorithms as the ones considered as redundant hits could actually prove to be the more accurate templates. Among 183 total structurally similar templates, 22 templates are available in the set of redundant structures and are discarded by our well-known supposedly reliable template search algorithms.

Where are we missing in our template search algorithms?

In search of a reliable template for a target sequence, we encounter several problems. Firstly, during the search for a reliable template to model the target sequence, a specific insertion in target sequence can be wrongly predicted to be in a coiled state confirmation by PSIPRED,

S No.	Template Model	TM_Score ≥ 0.7	GDT-TS	Resulted in	
				PSI-BLAST	HHPred
1.	4GB5A	0.999	100.000	YES	YES
2.	4STDC	0.836	72.774		
3.	4STDB	0.786	67.295		
4.	4STDA	0.782	65.753		
5.	7STDB	0.775	64.897		
6.	3STDA	0.774	63.699		
7.	3STDB	0.774	64.555		
8.	6STDA	0.774	64.384		
9.	5STDA	0.772	63.185		
10.	1STDA	0.771	65.068		
11.	5STDB	0.77	64.726		
12.	6STDB	0.77	64.726		
13.	7STDA	0.77	62.842		
14.	3STDC	0.768	62.842		
15.	5STDC	0.767	63.699		
16.	6STDC	0.767	63.014		
17.	7STDC	0.766	63.699		
18.	2STDA	0.763	63.356		
19.	3B8LD	0.735	64.384		
20.	3B8LA	0.733	63.87		YES
21.	3B8LC	0.733	63.87		
22.	3B8LB	0.73	62.5		
23.	3B8LE	0.73	63.87		
24.	3B8LF	0.73	64.384		

Table 2: Template Details For Top 24 Models With Tm_Score Accuracy Higher Than 0.7.

S.no	Template	Template Length	Sequence Identity	Identical Residues	BLOSUM Score	Cover Span
1	4gb5A	148	97.37	148	5.25	95.51
2	4stdC	164	18.12	29	0.34	98.8
3	4stdB	164	19.38	31	3.12	98.8
4	4stdA	164	19.38	31	3.12	98.8
5	7stdB	164	19.38	31	3.12	98.8
6	3stdA	162	19.5	31	3.14	100
7	3stdB	162	19.5	31	3.14	100
8	6stdA	164	19.38	31	3.12	98.8
9	5stdA	164	19.38	31	3.12	98.8
10	1stdA	162	19.5	31	3.14	100
11	5stdB	164	19.38	31	3.12	98.8
12	6stdB	164	19.38	31	3.12	98.8
13	7stdA	164	19.38	31	3.12	98.8
14	3stdC	162	19.5	31	3.14	100
15	5stdC	164	19.38	31	3.12	98.8
16	6stdC	164	19.38	31	3.12	98.8
17	7stdC	164	19.38	31	3.12	98.8
18	2stdA	162	20.75	33	3.77	100
19	3b8lD	144	25.33	38	0.83	100
20	3b8lA	147	25.08	38	0.83	98.14
21	3b8lC	144	25.33	38	0.8	100
22	3b8lB	144	25.33	38	0.83	100
23	3b8lE	144	25.33	38	0.8	100
24	3b8lF	143	24.75	37	0.83	100

Table 3: Pairwise alignment results of the top 24 templates against the target sequence.

if that insertion sequence data is unavailable in its repository of solved structures or is probably a novel fold. This is simply done to simply maximize the score of the prediction servers and this effortless trend normally comes at the cost of decreased biological significance of the evolutionary sequence insertions in the target sequence.

Secondly, errors on the basis of either misalignment or shifted alignment caused due to the addition or substitution of amino acid residues in the target or the template itself results in a bad model topology. That is why such alignments are either improved through other scoring measures and alignment algorithms like PRALINE [19] MUSCLE [20] and MAFFT [21] are manually curated.

Thirdly, PDB culling performed by HHPred, MODELLER or any other algorithm for that matter, compares structural similarity only of the templates without comparing their actual correlation against the target. So any template that is dissimilar or least similar to other templates is normally discarded. However, the discarded template could in fact share a biological relationship with the target [5,22].

Fourthly, while parsing the target template alignment, the start and the end chunks are normally pruned to simply improve the probability score of the alignment. Since considering the entire length could result in a lower TM_Score of the target model, only local alignments with high scoring conserved residues are normally considered for modelling a target. Such a practice should not be implemented as it is biologically important to predict the detailed target conformation, with a better topology closer to its actual native structure, and locally best alignment construction attempt through HHPred does not satisfy this constraint very well.

HMM based sequence comparison in a profile is yet another problem. HMM based residue substitution probability as such does not hold any biological significance. It is because the complete local

fold normally evolves together with an evolved alteration in some of its encoded residues and hence the complete fold should be scored together along with the individual residue probabilities. Other than this secondary structure constraint, loop residues are also important for a mutually correct orientation of secondary structure elements of the target sequence. Yet another problem here is the usage of scoring measures through alignment algorithms. These algorithms simply construct the best possible optimal target-template alignment without checking the biological validity and significance of the model constructed with the alignment results thus obtained. So, a set of sub-optimal alignments should also be included in the results by such algorithms as they could be structurally closer to the actual target conformation yielding a more accurate model topology.

Lastly as shown in the Table 3, the top 22 templates other than the HHPred resultant hits (4GB5_A and 3B8L_A) also show quite significant similarity against the target sequence. This table shows the template list in the order of their modelling accuracy, as shown earlier in the Table 2. The Table 3 further enlists the sequence length, identity, count of identical aligned residues, BLOSUM62 residue substitution score of the pairwise alignment (Constructed through MODELLER's default alignment script) and target coverage span scores for all the top 24 templates. It quite lucidly highlights the highly accurate score, certainly lesser than the answer structure 4GB5_A, although quite better than 3B8L_A. Hence, the need for an improved template search algorithm is strongly justified.

Conclusion

Because protein modelling accuracy is primarily rooted in the structural and functional homology of template against the target sequence, a thorough template search tool is mandatorily required to screen all the reliable hits for predicting the most accurate target models consistently. In this study, the template search measures PSI-BLAST and HHPred are scrutinized to assess the entirety of their results. Among a culled PDB95dataset of 41,967 structures 96,920 PDB entries in total and out of 24 templates with TM_Score more than 0.7, PSI-BLAST and HHPred protocols fail to search 23 and 22 hits, respectively. Therefore, this study attempts to highlight the logical problems prevailing in our normally employed template search measures and bridge the path for our further research methodologies.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
2. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15: 285-289.
3. Sali A (1998) 100,000 protein structures for the biologist. Nat Struct Biol 5: 1029-1032.
4. Runthala A, Chowdhury S (2016) Unsolved problems of ambient computationally intelligent TBM algorithms. In Hybrid Soft Computing Approaches: Research and Applications, Springer: Berlin 75-105.
5. Runthala A (2012) Protein structure prediction: challenging targets for CASP10. J Biomol Struct Dyn 30: 607-615.
6. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951-960.
7. Margelevicius M, Venclovas C (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics 11: 1-14.
8. Jones DT (1990) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.
9. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815.
10. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57: 702-10.
11. Wang G, Dunbrack RL (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589-1591.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
13. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89: 10915-10919.
14. Angermuller C, Biegert A, Soding J (2012) Discriminative modelling of context-specific amino acid substitution probabilities. Bioinformatics 28: 3240-3247.
15. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
16. Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000: 16.
17. Gonzalez MW, Pearson WR (2010) Homologous over-extension: a challenge for iterative similarity searches. Nucleic Acids Res 38: 2177-2189.
18. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26: 889-895.
19. Pirovano W, Feenstra KA, Heringa J (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 24: 492-497.
20. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
21. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.
22. Runthala A, Chowdhury S (2013) Protein Structure Prediction: Are we there yet?. In Innovations in Knowledge-based Systems in Biomedicine and Computational Life Sciences, Springer: Berlin 11-518.