Khor et al. Theoretical Biology and Medical Modelling (2015) 12:15
DOI 10.1186/s12976-015-0014-1

THEORETICAL BIOLOGY AND MEDICAL MODELLING

REVIEW Open Access

General overview on structure prediction of twilight-zone proteins

Bee Yin Khor, Gee Jun Tye, Theam Soon Lim and Yee Siew Choong*

* Correspondence: yeesiew@usm.my
Institute for Research in Molecular Medicine, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia

Abstract

Protein structure prediction from amino acid sequence has been one of the most challenging aspects in computational structural biology despite significant progress in recent years showed by critical assessment of protein structure prediction (CASP) experiments. When experimentally determined structures are unavailable, the predictive structures may serve as starting points to study a protein. If the target protein consists of homologous region, high-resolution (typically <1.5 Å) model can be built via comparative modelling. However, when confronted with low sequence similarity of the target protein (also known as twilight-zone protein, sequence identity with available templates is less than 30 %), the protein structure prediction has to be initiated from scratch. Traditionally, twilight-zone proteins can be predicted via threading or *ab initio* method. Based on the current trend, combination of different methods brings an improved success in the prediction of twilight-zone proteins. In this mini review, the methods, progresses and challenges for the prediction of twilight-zone proteins were discussed.

Introduction

Specific function and mechanism of a protein can be elucidated from the three dimensional (3D) structure of a protein. The most accurate way to determine a high resolution protein structure is through experimental methods such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy [1, 2]. As of January 2015, the Protein Data Bank (PDB) has over 100,000 deposited protein structures (www.rcsb.org) [3]. With the increasing number of deposited protein structure in PDB, the data is highly beneficial to the computational approach that utilized information from these experimentally-determined structures. Although the number of experimentally-determined protein structures is increasing at an accelerated rate, at the same time, numbers of known protein sequences from genome sequencing projects are increasing. To bridge the protein sequence-structure gap, computational protein 3D structure predictions from its amino acid sequence provide potential solution [4]. Computational protein structure prediction may not be as accurate as experimental method but they often reveal the molecular insight from the predicted structure and could generate hypotheses which are useful to complement the experimental approach and provide fundamental understanding of a protein [5, 6]. Therefore, when the experimentally-determined structures are unavailable, these predictive structures may serve as the starting points to study the protein.

Protein structure prediction is a method of translating the protein sequence into 3D structure through computational algorithms. Computational approaches for prediction protein 3D structures can be generally divided into three categories (comparative modelling, threading and *ab initio* approach). It can also be categorized into template-based (TBM) and template-free (FM) modelling [7, 8]. Comparative modelling and threading method are categorized into TBM as they depend on the availability of a template from solved protein structures [9]. FM (also known as *ab initio* or *de novo* method) is potentially able to predict protein structures without any template [8, 10]. To date, comparative modelling is the most successful and accurate method to produce a reliable structure. However, structure accuracy highly depends on how strong the relationship between target and template (sequence identity >30 %). For closely related protein sequence, sequence similarity usually falls above 30 % [4, 10, 11]. Over 95 % of protein chains with low sequence identity have different structures and this reduced the accuracy of the predictions [12]. As the sequence identity decreases, it leads to the probability of identifying incorrect templates and generating less accurate models with errors in predicted models, such as errors in side-chain packing, distortions and shifts in correctly aligned regions, errors in regions without a template and errors due to template misalignment [13, 14]. In addition, searching for homologous proteins is difficult when the sequence identity is low (also known as the "twilight-zone"), where the sequence identity falls between 10 and 30 % [15]. Thus, when the value is low, sequence identity is generally not a statistical reliable predictor to generate accurate model. Therefore, in such situation threading and *ab initio* method offer an alternative way for protein structure prediction. Previously, twilight zone protein structure prediction focused on the sequence alignment [16–20], the secondary structure prediction [21, 22] as well as the physiochemical properties of amino acids [23–25] to improve the quality of the built model. The scoring function e.g. position specific scoring matrices (PSSMs), Levitt-Gerstein (LG) score [26], LiveBench [27], MaxSub [28], S-score [29], *C*-score [30] where then used to rank the built models. Besides, obtaining an accurate structure for twilight-zone protein is challenging [31]. For this reason, this review will be emphasized on methods for prediction twilight-zone protein from scratch. Focus will be put on threading, *ab initio* and the current trend in protein structure prediction for twilight-zone proteins.

Threading method

Threading, also known as fold recognition, is used to identify protein templates in PDB bank for similar fold or similar structural motif to the target protein [32]. The concept for threading is similar to comparative modelling but comparative modelling only considers sequence similarity between target protein and template, while protein threading considers the structural information in the template [33]. The critical step of threading is to identify correct template proteins with similar folds to the target protein and make correct alignment [34]. Protein threading compares a target sequence against one or more protein structures to detect and obtain the best compatibility of sequence-structure template pair [1, 33]. They identify best fits of target sequence with the fold template based on the generated alignments and each template is calculated according to different scoring function. Commonly used alignment scores to identify precise target-template alignments include sequence profile-profile alignments (PPA), sequence-

structural profile alignments, secondary structure match, hidden-Markov models (HMM) and residue-residue contact [1]. The alignment algorithms are able to search for remotely homologous sequences in the databases. Therefore, even if sequence similarity is low (<30 %), threading method can be used to obtain similar folds or structural motifs for the target sequence. Traditionally, pair-wise comparison is used for matching of single sequences of target and template in the database. PPA, which can be used to detect weak similarities between protein families, is most often used and popular threading approach (successfully used in CASP7 for I-TASSER) [35, 36]. The new threading algorithm MUSTER (Multi-Source ThreadER) showed that accuracy of PPA can be further improved by incorporating various sequence and structure information (e.g. sequence profiles, secondary structure prediction, torsion angles, solvent accessibility and hydrophobic scoring matrix). MUSTER showed a better performance with TM-score 5–6 % higher than PPA in the testing proteins [34].

The overall procedure for I-TASSER is illustrated in Fig. 1. In general, I-TASSER divided the protein structure prediction into four steps: i) template identification, ii) structural reassembly, iii) model construction and, iv) final model selection. In the first step, the query sequence is threaded through PDB library to identify appropriate fragment using LOMETS algorithm [37]. This will be followed by continuous fragments

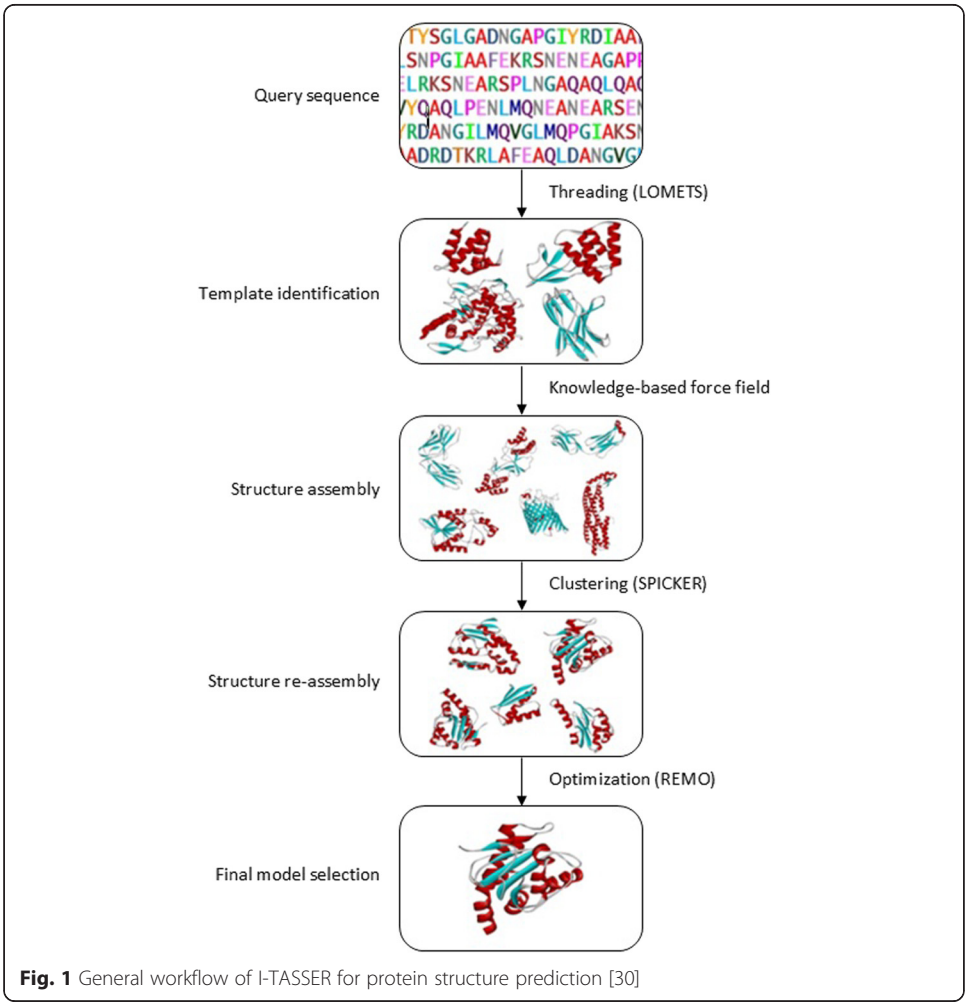

Fig. 1 General workflow of I-TASSER for protein structure prediction [30]

from the threading alignments are used to assemble full-length models that aligned well, with the unaligned regions (loops/tails) built by *ab initio* modelling [38]. The structure assembly simulations are guided by a knowledge-based force field, including: i) general knowledge-based statistics terms from the PDB, ii) spatial restraints from treading templates, iii) sequence-based contact predictions from SVMSEQ (a support vector machine based residue-residue contact predictor) [37]. After that, fragment assemble simulation is performed again and are clustered by SPICKER [39]. After superposition, all the clustered structures are averaged to obtain the cluster centroids. The final full atomic models are obtained by REMO which builds the full-atomic models from the selected I-TASSER decoys through the optimization of the hydrogen-bonding networks [40]. The forces in REMO protocol include H-bonding, clash/break-amendment, I-TASSER restraints and CHARMM22 potential [37]. For the final top 5 models selection, I-TASSER uses SPICKER to cluster and report up to five models corresponding to the five largest structure clusters. These steps are the essential advantage of TASSER for is its ability to drive the template structures closer to the native than the input templates by ~2–3 Å [41–43]. The confidence level of the predicted model was estimated by *C*-score (Eq. 1).

$$\text{C–score} = ln\left(\frac{M}{M_{tot}} \times \frac{1}{RMSD} \times \frac{1}{7}\sum_{i=1}^{7}\frac{Z(i)}{Z_0(i)}\right) \quad (\text{Eq.1})$$

TASSER has been tested in CASP6 experiment and emerged as one of the most successful structure prediction methods. It is however, TASSER failed to correctly predict the relative orientation of multiple domain proteins. TASSER's performance for free modelling targets is yet to be satisfactory as the success rate for non-homologous single-domain proteins is around two thirds [20, 44].

Since no single program has been reported to be outperformed others (within all threading approach), the consensus structure prediction method (meta-server approach) is therefore developed. With this approach, a number of models by multiple threading programs are generated. The idea behind this approach is the models that are generated by different programs are closest to native and less likely to make a common inaccurate prediction [31]. Available meta threading servers include 3D-Jury [45], and LOMETS [46]. 3D-Jury is a meta-server that collects and compares models from various remote protein structure prediction servers [45]. Therefore, the final performance is highly dependent on the inputs from the servers [46]. On the other hand, LOMETS locally installed all threading alignments programs, including PPA, HMM, structural profile and contact-based alignment. This will allow the users to obtain the predictions of all servers quickly compare with 3D-Jury [46]. The meta-server approaches have previously dominated the server prediction in CASP6 experiments. However, in CASP7 experiment, Zhang-Server (I-TASSER) showed better performance than all available meta-server (will be discussed in section 'Current trend in protein structure prediction') [47].

Ab initio method

When there is no homologous structure in PDB or the relationship is so distant until it could not be detected by threading, *ab initio* folding is the alternative way to generate protein structure from scratch [1]. This method is termed template-free modelling (FM)

(also known as *ab initio* or *de novo* modelling) as it originally referred to methods that based on the first principle laws of physics and chemistry. The idea is also based on Anfinsen's thermodynamic hypothesis [48]. As above-mentioned, Anfinsen's hypothesis stated that protein structure prediction depends solely on amino acid sequence [49]. The prerequisite of these modelling methods is that the native structure has the global minimum free energy among all available conformations [32]. Therefore, efficient and reliable algorithm is in need to limit the conformational space in order to minimize the energy function so that the protein is tend to be in its native state [50, 51].

There have been a variety of methods developed for *ab initio* protein structure generation. The leading approach is the fragment-based assembly method, an idea of Bowie and Eisenberg [11, 51, 52]. Based on this idea, Rosetta [53] was developed and was exceedingly successful in FM as Rosetta is able to produce accurate models nearer to its native structures [54, 55]. Fig. 2 shows the general workflow of Rosetta in protein structure prediction. The idea of fragment-based assembly is that the smaller fragments are restricted to the local structures by most closely related sequence in protein structure database [51, 54]. The lengths of the fragments vary by different programs and the fragment libraries comprise fragments from high-resolution known PDB structures. In Rosetta, fragment libraries of three- and nine- residue were exploited [53]. The original fragment insertion method by Rosetta showed consistent and accurate result compared to other *ab initio* structure predictions in CASP7 [53]. Generation of fragments is important in Rosetta after the completion of secondary structure prediction and it can be done through Robetta server [56, 57]. The program iterates over three- and nine-residue of the sequence and looks for similar sequences from the fragment libraries that Rosetta uses to guide the search of conformational space in predicting protein structures [58]. In Rosetta, method is done by Monte-Carlo algorithm to obtain native condition of protein conformations [53, 59]. Monte-Carlo algorithm generates a structure prediction by randomly inserting fragment predictions into the structure and the energy function is defined as the Bayesian probability of structure/sequence [54]. Bayes statistical theorem is exploited as a scoring function (Eq. 2) [59, 60]:

$$P(structure|sequence) = P(structure) \times \frac{P(sequence|structure)}{P(sequence)} \quad \text{(Eq.2)}$$

Rosetta energy functions are classified into two: knowledge-based centroid energy function that uses coarse-grained or low-resolution energy function to treat the side chains as centroids, and the knowledge-based all atom energy function that combines Lennard-Jones potential and a knowledge-based conformation-dependent amino acid internal free energy term [61]. The all atom energy function is more accurate but it is slower comparing with the centroid energy function as the side-chain atoms, van der Waals interaction, hydrogen bonds and pair wise solvation free energy is taking into consideration in all atom energy function. Both coarse-grained and all-atom energy function has been successfully used to predict high resolution protein structures from their sequences.

A newer method, QUARK by Yang Zhang group, successfully predicted models of correct folds for 8 out of 18 proteins with length less than 150 residues in CASP9 [62]. QUARK fragment assembly starts from random conformation that enable it to construct

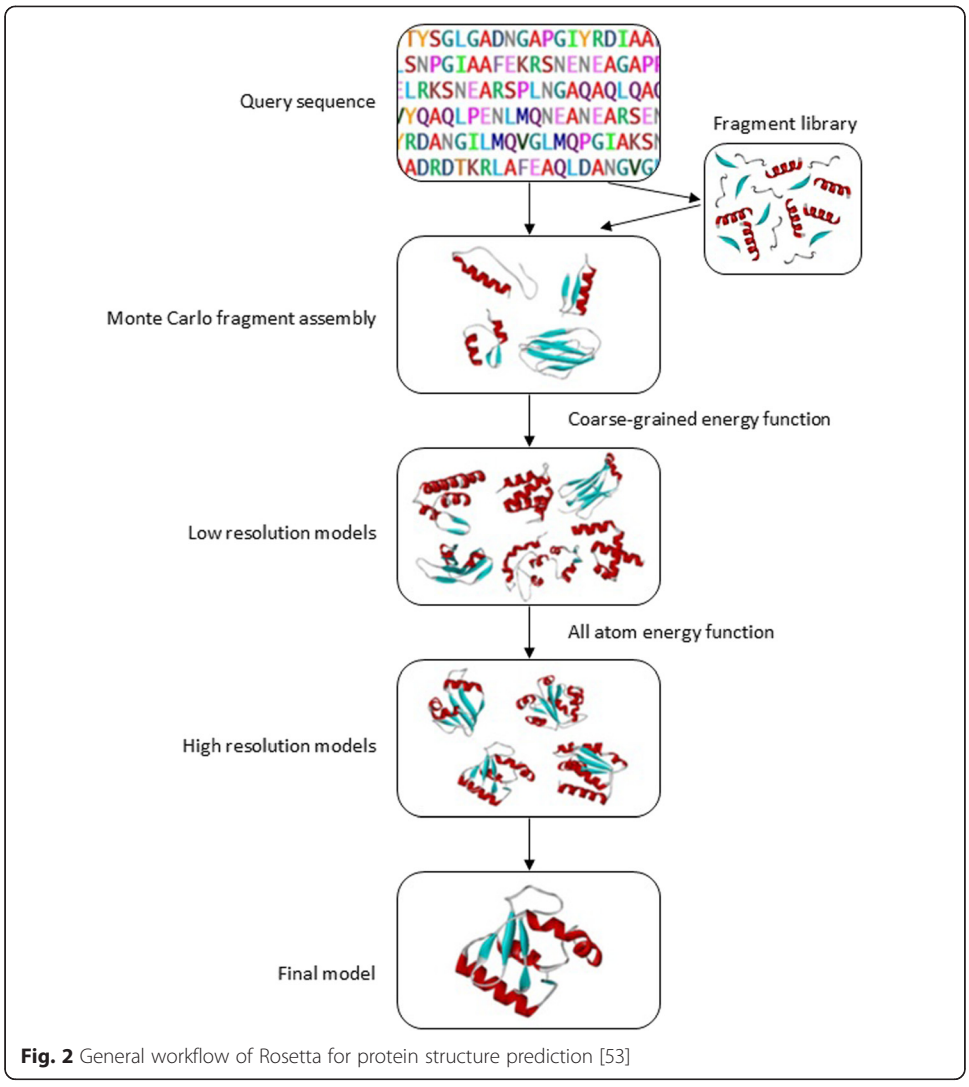

Fig. 2 General workflow of Rosetta for protein structure prediction [53]

new protein folds from scratch [63]. In QUARK, the models are assembled from small continuous fragments ranged from 1 to 20 residues excised from unrelated proteins by Monte-Carlo simulation [11, 63]. Both Rosetta and QUARK showed the importance of assembling structural models using small fragments by their significant performance in CASP9 [64]. In CASP10, QUARK successfully predicted model with larger size range in FM modelling (>150 residues) [62].

Current trend in protein structure prediction

In order to improve the performance of *in silico* approaches, the boundaries between the protein structure prediction methods have overlapped due to the integration of the strength of different approaches [31]. Recent CASP experiments demonstrated that composite approaches can achieve additional advantages in structure prediction. Since no single approach can perform better than others for all protein prediction, the emergence of new trend is the combination/hybrid of different protein structure prediction approaches [32, 63].

I-TASSER (Iterative Threading ASSEmbly Refinement) is one notable successful composite approach in the CASP experiments [30]. I-TASSER method is based on the secondary structure enhanced profile-profile threading alignment extended from TASSER algorithm for iterative structure assembly and refinement of protein molecules [43, 65]. I-TASSER retrieves structural template from PDB library through a meta-threading server, termed LOMETS. By year 2010, the online I-TASSER server has generated more than 30,000 full-length structure and function predictions for more than 6000 registered users [30]. I-TASSER can consistently predict correct folds and also high-resolution for small single-domain protein (<120 residues) with a lower computational time (5 CPU hours for I-TASSER and 150 CPU days per target for Rosetta). In CASP7, CASP8, CASP9 and CASP10, I-TASSER was ranked as the best server for protein structure prediction [66].

Butterfoss *et al.* presented blind-structure prediction for three peptoids using the hierarchical combination of Replica Exchange Molecular Dynamics (REMD) simulation and Quantum Mechanical (QM) refinement [67]. They have managed to predict a *N*-acryl peptoid trimer and a cyclic peptoid nonamer with backbone RMSD of only 0.2 and 1.0 Å, respectively. Their findings showed that physical modeling is able to performed *de novo* structure prediction for small peptoid molecules.

In 2013, *Bhageerath*-H Strgen, another homology/*ab initio* hybrid algorithm was developed. The method was tested in CASP9 experiments and showed 93 % of the targets were in the pool of decoys. The results showed that *Bhageerath*-H Strgen is capable of searching the protein fold for near-native conformation. Strategy in *Bhageerath*-H Strgen involved secondary structure prediction, database search for sequence based on the input amino acid sequence, fold recognition, template-target alignment, and template-based modelling by MODELLER [4]. The missing residues with no fragments are modelled using *Bhageerath ab initio* modelling. In their study, they showed that *Bhageerath*-H Strgen performs better than Rosetta and I-TASSER [68].

The Robetta server (http://robetta.bakerlab.org) is an automated server for protein structure and analysis. Protein structures can be generated in the presence or absence of similarity to homologous proteins of known structure. BLAST, PSI-BLAST, FFAS03 or 3D-Jury is used to search for a match to the solved protein structure. When there is a confident match, comparative modelling is used for protein structure prediction. If no match is found, *ab initio* Rosetta fragment insertion method will be used for prediction [58]. In CASP8 experiment, Robetta is ranked as the top 4 best performing groups [69].

Successes and challenges for twilight-zone protein modelling

The successful rates for twilight-zone protein modelling are increasing over the years with numerous successful examples have been reported. In year 2008, *Leucosporidium antarcticum* antifreeze protein was predicted by comparative modelling, threading and *ab initio* approaches due to low sequence identity. Their study suggests that I-TASSER (*ab initio* approach) is useful for low resolution protein structure prediction for twilight-zone protein.

In 2011, *Chlamydia trachomatis* protein CT296 was determined using both computational method (I-TASSER) and X-ray crystallography method. Despite having no homologs, the result showed that the structure of CT296 predicted by *ab initio* I-TASSER has overall structural similarity (root mean square, RMSD of 2.72 Å for 101/137 residues)

to the high-resolution X-ray crystallography structure (1.8 Å). The result showed that I-TASSER is effective to predict accurately twilight-zone protein structures that have no primary sequence homolog with any known proteins [70]. This is an encouraging study for the most challenging twilight-zone protein modelling in protein structure prediction.

Successes in the structure prediction for gas vesicle protein GvpA from haloarchaeon *Haloferax mediterranei* have also been reported. The protein structure was predicted by Strunk *et al.*, and Ezzeldin *et al.*, in year 2011 and 2012 respectively [71, 72]. The structure prediction was first carried out by Strunk and colleague via *ab initio* approach (Rosetta). The predicted structure suggested that GvpA possessed two α-helices and two β-strands. The secondary structure elements (α- β- β- α) is similar with the NMR structures obtained for GvpA protein from cyanobacterium *Anabaena flos-aquae* [73]. Mutation in α-helix and β-turn affected the ability to form gas vesicle. This *in vivo* data on GvpA mutants support the major structural features from the proposed structures.

In the subsequent year, Ezzeldin and colleagues predicted GvpA protein from *Halobacterium sp. NRC-1* with computational comparative modelling (by MODELLER and SCRATCH), threading (by I-TASSER) and *ab initio* modelling (by Rosetta) [72]. All the predicted structures were equilibrated through molecular dynamics (MD) simulation. Average MM-PBSA energy and standard deviation were calculated and ranked. From the comparison of the top ranked predicted structures and an earlier model proposed by Strunk *et al.*, it showed that two sequences possess 93 % identity despite of belonging to different organisms [71]. Furthermore, the structures possessed an α- β- β- α secondary structure, in agreement with previous experimental data and their secondary structure prediction [72]. The predicted model thus support the hypothesis that homologous sequences synthesized by different organisms should exhibit similar structures [72].

Another research in year 2014 was the structure prediction of *Bm*R1 protein from *Brugia malayi*. In the study, the *Bm*R1 protein (206 residues) was modelled via comparative modelling, threading and *ab initio* approaches. The predicted models were evaluated and compared. Based on the model evaluation, the *ab initio* approaches by Rosetta outperformed others method with a quality and reliable structure from structure validation and evaluations [74].

Despite the rapid progress in structure prediction, there are still significant challenges in the current method [32]. As demonstrated in the CASP experiments, the successful of twilight-zone protein modelling via FM is only limited to small protein below 100 residues [63]. With increasing protein size, the conformational space will also increase proportionally. As mention earlier, it is important to limit conformational space in order to obtain lowest free energy. In CASP 10, QUARK successfully predicted two FM targets with length >150 residues [62]. Although there are successful predictions for twilight-zone protein, there is still a need for a consistent successful rate. For example, in spite of the reported successful cases, the QUARK program has difficulty to consistently assemble the correct protein structures with length >100–120 residues from scratch [63, 75].

Another challenge in twilight-zone protein is to distinguish the correct distantly related proteins from unrelated proteins. The accuracy of comparative modelling is highly dependent on the sequence similarity between the target sequence and template. For closely related protein, sequence similarity usually above 30 % [4, 10]. When the sequence similarity decreases, probability of getting a reliable structure decreases. For this reason,

the algorithms and programs to identify correct templates from related proteins play a significant role. Although various template searching algorithms are available online, efficient and consistent template detection is still essential especially for distantly related protein sequences.

Conclusion and future direction

The elucidation of a protein structure is vital in order to aid the understanding of the biological roles of it in living cells. Comparative modelling can generate high resolution model when evolutionary related homologous templates are identified. The structure of a query protein from different evolutionary origin can be predicted by threading method to recognize folds similar to query. A query must be built from scratch by *ab initio* modelling when no structurally related proteins were found in the template database. Here, we have presented a general review on twilight-zone protein structure prediction from the point of view in both threading and *ab initio* approaches. Although each method reported successful predictions, the composite approaches from threading, *ab initio* and other various methods have showed marked improvement compared to the single method alone. The bottleneck of the twilight zone protein modelling is that the success/accuracy rate is decreased when the protein size is increased. Significant challenges remain in distant-homology identification and refinement. Compounded by the complexity of structure prediction is that about one tenth of proteins are disordered for their physiochemical roles. Therefore, the development of a reliable, efficient and consistent algorithm in fold-recognition and refinement would influence for accuracy in the prediction of twilight-zone proteins.

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

BYK drafted the manuscript. GJT, TSL and YSC revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by Science Fund (305/CIPPM/613232) from Malaysian Ministry of Science, Technology and Innovations.

Received: 8 June 2015 Accepted: 27 August 2015
Published online: 04 September 2015

References

1. Wu S, Zhang Y. Protein structure prediction. In: Edwards D, Stajich J, Hansen D, editors. Bioinformatics. New York: Springer; 2009. p. 225–42.
2. Nguyen MN, Madhusudhan MS. Biological insights from topology independent comparison of protein 3D structures. Nucleic Acids Res. 2011;39, e94.
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. Webb B, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol. 2014;1137:1–15.
5. Petrey D, Honig B. Protein structure prediction: inroads to biology. Mol Cell. 2005;20:811–9.
6. Wooley JC, Ye Y. A historical perspective and overview of protein structure prediction. In: Xu Y, Xu D, Liang J, editors. Computational methods for protein structure prediction and modeling. New York: Springer; 2007. p. 1–43.
7. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. Nat Protoc. 2012;7:1511–22.
8. Maurice KJ. SS Thread: template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. J Comput Chem. 2014;35:644–56.
9. Fiser A. Template-based protein structure modeling. Methods Mol Biol. 2010;673:73–94.
10. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins. 2011;79:1–5.
11. Xu D, Zhang Y. Toward optimal fragment generations for *ab initio* protein structure assembly. Proteins. 2013;81:229–39.

12. Mizianty M, Kurgan L. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics. 2009;10:414.
13. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291–325.
14. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen M-y, et al. Comparative protein structure modeling using modeller. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates III JR, editors. Current protocols in bioinformatics. New York: Wiley; 2006. p. 5.6.1–5.6.30.
15. Hansen SF, Bettler E, Wimmerová M, Imberty A, Lerouxel O, Breton C. Combination of several bioinformatics approaches for the identification of new putative glycosyltransferases in *Arabidopsis*. J Proteome Res. 2008;8:743–53.
16. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol. 2001;307:721–35.
17. Huang YM, Bystroff C. Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics. 2006;22:413–22.
18. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12:85–94.
19. Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol. 1995;249:816–31.
20. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins. 2005;61:91–8.
21. Homaeian L, Kurgan LA, Ruan J, Cios KJ, Chen K. Prediction of protein secondary structure content for the twilight zone sequences. Proteins. 2007;69:486–98.
22. Kurgan L, Chen K. Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun. 2007;357:453–60.
23. Gruber M, Soding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. J Struct Biol. 2006;155:140–5.
24. Szilagyi A, Gyorffy D, Zavodszky P. The twilight zone between protein order and disorder. Biophys J. 2008;95:1612–26.
25. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000;41:415–27.
26. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A. 1998;95:5913–20.
27. Rychlewski L, Fischer D, Elofsson A. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins. 2003;53 Suppl 6:542–7.
28. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2000;16:776–85.
29. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. BMC Bioinformatics. 2001;2:5.
30. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725–38.
31. Mihăşan M. Basic protein structure prediction for the biologist: a review. Arch Biol Sci. 2010;62:857–71.
32. Roy A, Zhang Y. Protein structure prediction. Chichester: Wiley; 2012.
33. Xu J, Jiao F, Yu L. Protein structure prediction using threading. Methods Mol Biol. 2008;413:91–121.
34. Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins. 2008;72:547–56.
35. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol. 2002;315:1257–75.
36. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep. 2013;3:2691.
37. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins. 2009;77 Suppl 9:100–13.
38. Wu S, Skolnick J, Zhang Y. *Ab initio* modeling of small proteins by iterative TASSER simulations. BMC Biol. 2007;5:17.
39. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A. 2004;101:7594–9.
40. Li Y, Zhang Y. REMO: a new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins. 2009;76:665–76.
41. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins. 2007;69:108–17.
42. Pandit SB, Zhou H, Skolnick J. Tasser-based protein structure prediction. In: Rangwala H, Karypis G, editors. Introduction to protein structure prediction. New Jersey: Wiley; 2010. p. 219–42.
43. Zhang Y, Skolnick J. Segment assembly, structure alignment and iterative simulation in protein structure prediction. BMC Biol. 2013;11:44.
44. Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, et al. Analysis of TASSER-based CASP7 protein structure prediction results. Proteins. 2007;69:90–7.
45. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 2003;19:1015–8.
46. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 2007;35:3375–82.
47. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008;18:342–8.
48. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181:223–30.
49. Hoque MT, Chetty M, Sattar A. Genetic Algorithm in *ab Initio* protein ptructure prediction using low resolution model: a review. In: Sidhu AS, Dillon TS, editors. Biomedical data and applications. Heidelberg: Springer; 2009. p. 317–42.
50. Bonneau R, Baker D. *Ab initio* protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct. 2001;30:173–89.
51. Ishida T, Nishimura T, Nozaki M, Inoue T, Terada T, Nakamura S, et al. Development of an *ab initio* protein structure prediction system ABLE. Genome Inform. 2003;14:228–37.
52. Bowie JU, Eisenberg D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci U S A. 1994;91:4436–40.

53. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, et al. Rosetta in CASP4: progress in *ab initio* protein structure prediction. Proteins. 2001;45:119–26.
54. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997;268:209–25.
55. Simoncini D, Zhang KYJ. Efficient sampling in fragment-based protein structure prediction using an estimation of distribution algorithm. PLoS ONE. 2013;8, e68954.
56. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P, et al. Automated prediction of CASP-5 structures using the Robetta server. Proteins. 2003;53:524–33.
57. Chivian D, Kim DE, Malmström L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated robetta protocols. Proteins. 2005;61:157–66.
58. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004;32:W526–W31.
59. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004;383:66–93.
60. Holzinger A, Dehmer M, Jurisica I. Knowledge discovery and interactive data mining in bioinformatics–state-of-the-art, future challenges and research directions. BMC Bioinformatics. 2014;15 Suppl 6:I1.
61. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J. Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry. 2010;49:2987–98.
62. Xu D, Zhang Y. *Ab Initio* structure prediction for *Escherichia coli*: towards genome-wide protein structure modeling and fold assignment. Sci Rep. 2013;3:1895.
63. Zhang Y. Interplay of I-TASSER and QUARK for template-based and *ab initio* protein structure prediction in CASP10. Proteins. 2013;82:175–87.
64. Kinch L, Shi SY, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins. 2011;79:59–73.
65. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.
66. Nahar N, Rahman A, Moś M, Warzecha T, Ghosh S, Hossain K, et al. *In silico* and *in vivo* studies of molecular structures and mechanisms of AtPCS1 protein involved in binding arsenite and/or cadmium in plant cells. J Mol Model. 2014;20:1–16.
67. Butterfoss GL, Yoo B, Jaworski JN, Chorny I, Dill KA, Zuckermann RN, et al. *De novo* structure prediction and experimental characterization of folded peptoid oligomers. Proc Natl Acad Sci U S A. 2012;109:14320–5.
68. Dhingra P, Jayaram B. A homology/*ab initio* hybrid algorithm for sampling near-native protein conformations. J Comput Chem. 2013;34:1925–36.
69. 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;77:50–65.
70. Kemege KE, Hickey JM, Lovell S, Battaile KP, Zhang Y, Hefty PS. *Ab initio* structural modeling of and experimental validation for *Chlamydia trachomatis* protein CT296 reveal structural similarity to Fe(II) 2-oxoglutarate-dependent enzymes. J Bacteriol. 2011;193:6517–28.
71. Strunk T, Hamacher K, Hoffgaard F, Engelhardt H, Zillig MD, Faist K, et al. Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*. Mol Microbiol. 2011;81:56–68.
72. Ezzeldin HM, Klauda JB, Solares SD. Modeling of the major gas vesicle protein, GvpA: from protein sequence to vesicle wall structure. J Struct Biol. 2012;179:18–28.
73. Sivertsen AC, Bayro MJ, Belenky M, Griffin RG, Herzfeld J. Solid-state NMR characterization of gas vesicle structure. Biophys J. 2010;99:1932–9.
74. Khor BY, Tye GJ, Lim TS, Noordin R, Choong YS. The structure and dynamics of BmR1 protein from *Brugia malayi*: *In silico* approaches. Int J Mol Sci. 2014;15:11082–99.
75. Xu D, Zhang Y. *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 2012;80:1715–35.