Protein Structure Prediction

Ambrish Roy, *University of Michigan, Ann Arbor, Michigan, USA*

Yang Zhang, *University of Michigan, Ann Arbor, Michigan, USA*

Advanced article

Article Contents

- Introduction
- Pipeline of the Composite Protein Structure Prediction: A Case Study
- Applications of Protein Structure Predictions
- Challenges of Modelling Transmembrane and Disordered Proteins
- Conclusion
- Acknowledgements

Online posting date: 15th August 2012

The goal of protein structure prediction is to estimate the spatial position of every atom of protein molecules from the amino acid sequence by computational methods. Depending on the availability of homologous templates in the PDB library, structure prediction approaches are categorised into template-based modelling (TBM) and free modelling (FM). While TBM is by far the only reliable method for high-resolution structure prediction, challenges in the field include constructing the correct folds without using template structures and refining the template models closer to the native state when templates are available. Nevertheless, the usefulness of various levels of protein structure predictions have been convincingly demonstrated in biological and medical applications.

Introduction

The ideal solution to the protein structure prediction problem is based on the physicochemical principles, that is, to find the native structure of proteins by identifying the lowest free-energy states. This dream was motivated by the Anfinsen's finding that the native structure is determined only by the protein's amino acid sequence which represents a unique, stable and kinetically accessible minimum of the free energy (Anfinsen, 1973). However, no success has been demonstrated along the line of first principle-based methods. This is mainly due to the lack of strategy that can precisely describe the subtle atomic interactions of intra-protein and protein–solvent interactions. Second, searching for the correct state through the giant number of possible conformations is a major challenge to current computing power.

On the other hand, the bioinformatics based approaches, that is, predicting target structures using information collected from solved structures of other related proteins which are deposited in the Protein Data Bank (PDB) (Berman *et al*., 2000), have enjoyed considerable success. The key procedures of bioinformatics-based approaches include query-template sequence alignments, fold-recognition, fragment-based structural assembly and multiple template-based structural refinements. For practical reasons, this note focuses mainly on the review of the successes and challenges of the bioinformatics-based methods. We also briefly introduce the current state of biological application of protein structure predictions. In **Table 1**, we list a summary of the commonly used web-servers for automated protein structure predictions.

eLS subject area: Structural Biology

How to cite:
Roy, Ambrish; and Zhang, Yang (August 2012) Protein Structure Prediction. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0003031.pub2

Review of protein structure prediction approaches

Protein structure prediction methods, depending on the extent to which they exploit the known experimental structures in the Protein Data Bank (PDB), have been broadly classified into three categories: *ab initio* folding, comparative modelling and threading. By definition, *ab initio* (or *de novo*) modelling originally referred to the methods that are based on the first principle laws of physics and chemistry. The guiding principle is that the native state of the protein lies at the global free-energy minimum (Anfinsen, 1973). Therefore, *ab initio* methods try to fold a given protein from the query sequence using various force fields and extensive conformational search algorithms. However, little success has been demonstrated by using the physicochemical principle-based approaches. The most successful methods in this category still use evolutionary and knowledge-based information to collect spatial restraints and short structural fragments to assist structural assembly procedure (Simons *et al*., 1997; Xu and Zhang, 2012). This category is now called 'free modelling' (FM) in the CASP experiments since many of the methods do not purely rely on the first principles (Moult *et al*., 2009). Despite of the progress in *ab initio* protein structure predictions, predicting 3D structure of proteins with >150 residues is still a major challenge. This size dependence is due to the higher number of secondary structure elements in the large proteins, which results in a much higher number of possible fold arrangements. The force field of most *ab*

Table 1 List of publicly available protein structure prediction tools

Name	Web address	Methods[a]
On-line protein structure prediction servers		
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	TBM + FM
Robetta	http://robetta.bakerlab.org/	FM
ModWeb	https://modbase.compbio.ucsf.edu/scgi/modweb.cgi	TBM
SwissModel	http://swissmodel.expasy.org/	TBM
HHpred	http://hhpred.tuebingen.mpg.de/hhpred	TBM
chunk-TASSER	http://cssb.biology.gatech.edu/skolnick/webservice/chunk-TASSER/index.html	TBM + FM
QUARK	http://zhanglab.ccmb.med.umich.edu/QUARK/	FM
Phyre	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id = index	TBM
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html	TBM
3D-Jury	http://meta.bioinfo.pl	TBM (meta–server)
LOMETS	http://zhanglab.ccmb.med.umich.edu/LOMETS/	TBM (meta–server)
PSIpred	http://bioinf.cs.ucl.ac.uk/psipred/	TBM + SS
Freely downloadable software for protein structure prediction		
Modeller	http://salilab.org/modeller/	TBM
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/download/	TBM + FM
Rosetta	http://www.rosettacommons.org/software/	FM
HHsearch	ftp://toolkit.lmb.uni-muenchen.de/HHsearch/	TBM
Scwrl4	http://dunbrack.fccc.edu/scwrl4/	SC

[a]TBM, template-based modelling; FM, free modelling; SS, secondary structure prediction; SC, Side-chain structure modelling.

initio approaches has no specificity to recognise the correct fold among the numerous fold candidates. Therefore, increasing the accuracy of the force field and the power of conformational search is essential to the solution of the problem.

In comparative modelling (CM, also called homology modelling), protein structure is constructed by matching the sequence of the protein of interest (target) to an evolutionarily related protein with a known structure (template) in the PDB. Thus, a prerequisite for CM technique is the presence of a homologous protein in the PDB library. For the protein targets where templates with a sequence identity $> 50\%$ are available in the PDB, the homologous templates can be easily identified with the sequence-template alignments precisely conducted. The backbone models generated using CM techniques can have a modelling accuracy of up to 1–2 Å RMSD from the native structure. For protein targets which have templates with sequence identity ranging from 30 to 50%, the target-template alignment is less accurate but the models often have $\sim 85\%$ of their core regions within an RMSD of 2–4 Å from the native structure, with errors mainly occurring in the loop regions (Jauch *et al*., 2007). However, when the target-template sequence identity drops below 30%, modelling accuracy by CM sharply decreases because of substantial alignment errors and the lack of significant template hits. Because CM builds models by copying the aligned structures of the templates or by satisfying distance/contact restraints from the templates (Marti-Renom *et al*., 2000), an essential limit of the approach is that the CM models usually have a strong bias and are closer to the template structure rather than to the native structure of the target protein (Read and Chavali, 2007; Tramontano and Morea, 2003). Accordingly, one of the important challenges to CM (and to all template-based methods) is how to refine the models closer to the native structure than the initial templates.

Threading (or fold recognition) refers to a bioinformatics procedure that identifies protein templates in the PDB library, which have a similar fold or similar structural motif to the target protein. It is similar to CM in the sense that both approaches try to build a structural model by using the experimentally solved structures as template. However, since many proteins with low sequence identity can have similar folds, threading aims to detect the target-template alignments regardless of the evolutionary relationship. The identification of precise target-to-template alignments is a significantly nontrivial problem when the sequence identity is low. Here, the design of accurate alignment scoring function is essential to the efficiency of the approaches. The commonly used alignment scores include secondary structure match, sequence-structural profile match (Bowie *et al*., 1991), sequence profile–profile alignments (Rychlewski *et al*., 2000; Soding, 2005) and residue–residue contacts (Skolnick *et al*., 2004; Xu *et al*., 1999), with the best scoring alignments usually detected by dynamic programming (Needleman and Wunsch, 1970) or hidden-Markov modelling (Eddy, 1998). Recently, it has been demonstrated that the methods of composite scoring functions including multiple structural features (e.g. solvent accessibility, torsion angles etc) can achieve additional gains in the protein template identifications (Wu and Zhang, 2008; Yang *et al*., 2011).

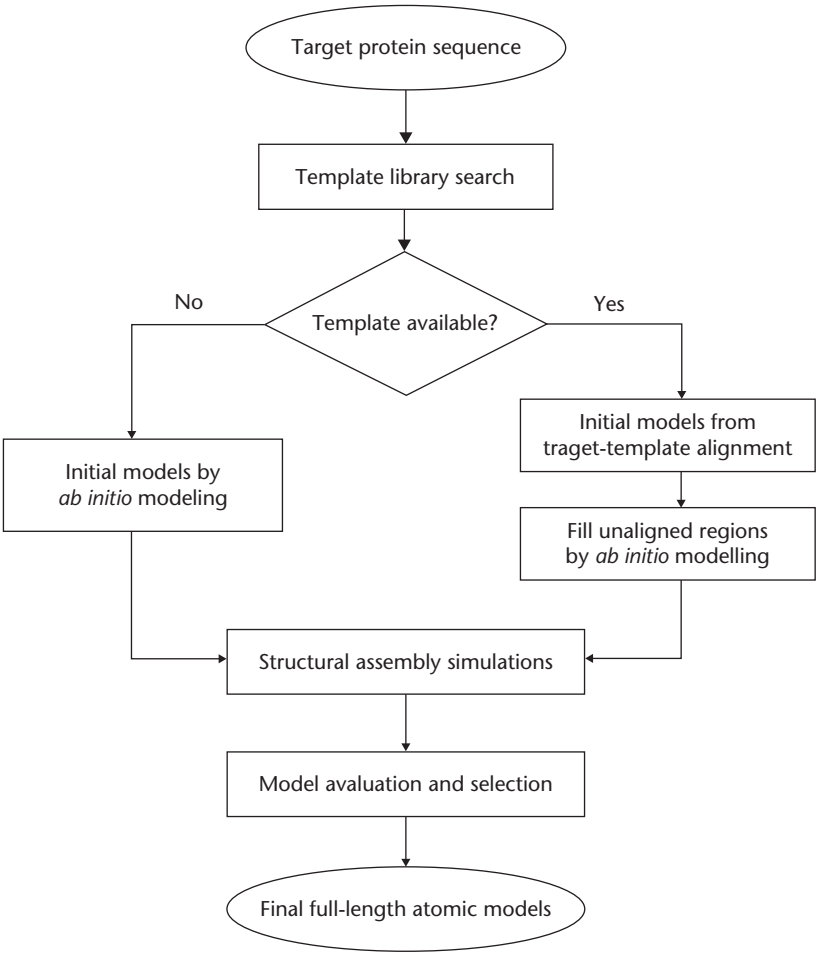

Figure 1 Pipeline of a typical composite protein structure prediction approach.

As a general trend in the field of protein structure prediction, the borders between the conventional categories of methods have become blurred. Many *ab initio* approaches use spatial restraints or structural fragments detected by threading (Bradley *et al*., 2005; Xu *et al*., 2011; Zhang *et al*., 2003); both threading and comparative modelling approaches rely on multiple sequence alignments. Meanwhile, since no single approach can outperform others for all protein targets, the second trend of the field is the prevalence of the so-called meta-server approaches (Fischer, 2006). A common meta-server approach is to generate a number of models by multiple programs which are developed by different laboratories, with the final models then selected from the best ranking ones (Ginalski *et al*., 2003; Wu and Zhang, 2007). Although different approaches have been attempted in protein template and model selections, the most efficient model selection approach appears to be the consensus selection, that is, the models that are most often generated by different methods are usually the one that is the closest to the native (Wallner and Elofsson, 2007; Zhang *et al*., 2010).

Rather than model ranking and selection, another efficient meta-server based approach is to reconstruct protein models using multiple template information, for example, exploiting the spatial restraints and structural fragments extracted from the multiple templates to guide the physics-based structural assembly simulations. The final models can thus have a refined quality compared to any of the individual templates. This method represents the most efficient and successful approach, as demonstrated by community-wide benchmark results of the recent CASP experiments (Das *et al*., 2007; Zhang, 2007; Zhou *et al*., 2007). As a case study, we dissect in detail the pipeline of such composite approach in the next section.

Pipeline of the Composite Protein Structure Prediction: A Case Study

A typical composite protein structure prediction pipeline involves five main steps (**Figure 1**): (a) identification of

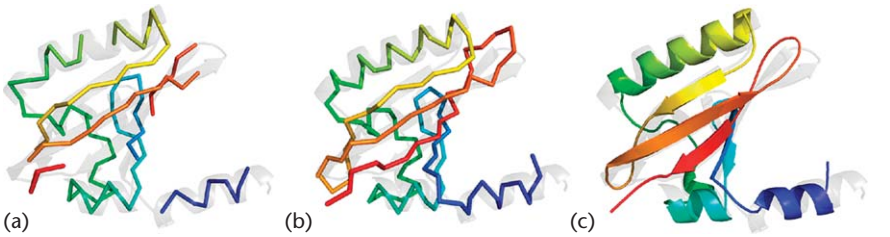

Figure 2 An example of the template-based modelling by I-TASSER server for PAS domain from *Burkholderia thailandensis* (PDBID: 3mqo). (a) Initial target model built by copying Cα coordinates from a nonhomology template (PDBID: 3lyx) identified by MUSTER, which contains multiple gaps; (b) full-length model constructed by the I-TASSER Monte Carlo assembly simulations; (c) final atomic structural model after atomic structural refinement. The grey background cartoon shows the X-ray structure.

template structures and construction of target-template alignments; (b) construction of initial model from template alignments with the unaligned regions constructed by *ab initio* modelling; (c) reduced-level structural assembly and refinement simulations; (d) atomic-level model construction and refinement; and (e) model selection. Below is a detailed explanation of the procedures.

Identification of templates and target-template alignment: Identification of suitable template structures is invariably the first step in the protein structure prediction pipeline. A variety of approaches have been developed for the identification of suitable templates. The simplest approach is the sequence level alignment based on the BLOSUM or PAM mutation scales with the highest scoring alignment identified by the Needleman–Wunsch or Smith–Waterman dynamic programming algorithms. When close homologous template proteins with sequence identities 30–40% or higher are available in the structure library, models generated using this approach is generally of high accuracy. However, as the target-template sequence identity falls below the threshold, alignment and modelling errors rapidly grows (**Figure 2a**).

For most proteins, it is difficult to identify close homologue in the existing structure library using the naïve sequence-to-sequence method. Most of the state-of-the-art structure modelling methods therefore use sophisticated threading algorithms to generate and score target-template alignments. Construction of this sequence-to-structure alignment is nontrivial and many algorithms use a combination of both sequence and structure information. In the threading program MUSTER (Wu and Zhang, 2008), for example, a composite scoring function (for aligning '*i*'th residue of the query to '*j*'th residue of the template) is defined as:

$$\begin{aligned}\mathrm{Score}(i,j) = {} & \mathrm{E}_{\mathrm{seq\_prof}} + \mathrm{E}_{\mathrm{sec}} + \mathrm{E}_{\mathrm{struc\_prof}} \\ & + \mathrm{E}_{\mathrm{sa}} + \mathrm{E}_{\mathrm{phi}} + \mathrm{E}_{\mathrm{hydro}} + \mathrm{E}_{\mathrm{shift}}\end{aligned} \quad [1]$$

where the first term $\mathrm{E}_{\mathrm{seq\_prof}}$ represents the profile–profile alignment score with the sequence profiles generated from multiple sequence alignments of the target and template sequences. The second term $\mathrm{E}_{\mathrm{sec}}$ computes the match between the predicted secondary structure of query and secondary structure of templates. The third term $\mathrm{E}_{\mathrm{struc\_prof}}$ calculates the score of aligning the structured-derived profiles of templates to the sequence profile of query. The fourth term E_{sa} counts for the difference between the predicted solvent accessibility of query and solvent accessibility of templates. The fifth and sixth terms ($\mathrm{E}_{\mathrm{phi}}$ and $\mathrm{E}_{\mathrm{psi}}$) count for the difference between the predicted torsion angles (phi and psi) of query and those of templates. The seventh term $\mathrm{E}_{\mathrm{hydro}}$ is an element of hydrophobic scoring matrix that encourages the match of hydrophobic residue (V, I, L, F, Y, W, M) in the query and the templates. And the last term $\mathrm{E}_{\mathrm{shift}}$ is introduced to avoid alignment of unrelated residues in local regions. Although the first term is sequence-based information, the second to seventh terms all correspond to the use of structural information. The sequence and structural information are then combined into a single-body energy term, which can be conveniently used in the dynamic programming algorithm for identifying the best alignment between the query and the template.

Since a single threading program often fails in identifying the best template, it is wise to collect template alignments generated from multiple threading programs, which can increase the coverage of different templates. Meanwhile, the consensus information of the template alignments from multiple programs can be used to identify better templates and structurally conserved residue regions. One of the first meta threading servers is 3D-Jury by Ginalski *et al*. (2003), which collects threading alignments from on-line servers of different laboratories through internet queries. But due to the availability and connection issues, the internet-based meta-servers suffer from speed and completeness of result collections. Recently, Wu and Zhang constructed the LOMETS meta-server that has all threading programs locally installed, which significantly improved the reliability and speed of the meta-server approaches. LOMETS currently consists of nine state-of-the-art threading programs based on a number of different alignment methods, including profile–profile alignments (MUSTER, PPA, SP3, Sparks), hidden-Markov models (HHsearch, SAM-T02), structural profile (FUGUE) and contact-based alignment (PROSPECT2, PAINT). A list of the most often used template identification servers can be seen in **Table 1**. Further explanation on the sequence alignment and template threading can also

be found in. **See also**: Protein Tertiary Structures: Prediction from Amino Acid Sequences

Construction of initial model using target-template alignment: Once the template proteins are selected based on the alignment scores, the next step is to generate initial target models by copying the Cα XYZ co-ordinates of the template residues to the target residues, for the aligned residue pairs. Since the chain connectivity is required for most methods, various methods are designed to construct the structural models for the unaligned regions (**Figure 2a**). For instance, I-TASSER builds the initial full-length models by filling the gaps using a self-avoiding random walk of C_α–C_α bond vectors of variable lengths from 3.26 to 4.35 Å. To guarantee that the last step of the random walk can quickly arrive at the first C_α of the next template fragment, the distance l between the current C_α and the first C_α of the next template fragment is checked at each step of the random walk; only the walks with $l < 3.54n$ are allowed, where n is the number of remaining C_α–C_α bonds in the walk. If the template gap is too big to span by a specified number of unaligned residues, a big C_α–C_α bond is kept at the end of the random walk and a spring-like force that acts to draw sequential fragments closer will be applied in subsequent structural assembly step.

Reduced-level structure assembly and refinement simulations: Once an initial model is generated, refinement simulations are conducted to reassemble the global topology and the local structures of the protein chains (**Figure 2b**).

The success of the refinement simulations depend on the accuracy of the force field and the efficient of the search engine. Although the physics-based atomic force fields can provide a reasonable description of protein–atom interactions in many aspects, the implementation requests atomic-level representation which are often too slow to refine proteins of a reasonable size. The knowledge-based potentials, which are often in reduced forms and derived from the statistical regularities of the structures in the PDB, have shown power in both protein structure recognition and fold assembly simulations (MacCallum *et al*., 2009; Summa and Levitt, 2007), where appropriate selections of reference states and structural features are proven to be of critical importance (Skolnick, 2006).

Second, given the force fields, efficient identification of the global energy minimum is nontrivial since most of the composite force fields are characterised with numerous local energy minima, which can easily trap the folding simulations. One way of speeding up the computational search process is to reduce the conformational entropy. For example, in TOUCHSTONE-II (Zhang *et al*., 2003), the authors constrained the conformational change of protein structure on a lattice system. In Rosetta (Simons *et al*., 1997) and I-TASSER (Zhang, 2007), fragment structures copied from the PDB templates are kept rigid during the simulation. These techniques can help significantly reduce the entropy of search because of constraint on conformational movements.

Another way of increasing the conformational search efficiency, which is also associated with the entropy reduction, is to reduce the level of protein structure representation. For example, in UNRES (Liwo *et al*., 2007), a protein residue is represented by three units of Cα atom, side-chain ellipsoid and peptide group. In I-TASSER (Roy *et al*., 2010; Zhang, 2007), the residue is specified by two units of Cα atom and the side-chain centre of mass. These reductions of structure representation can dramatically reduce the total number of conformations in the search space. However, although the reduced models have advantage of better conformational search, they may suffer from a lower accuracy of energy force field design.

Finally, a central theme in protein conformational search is the appropriate design of conformational updating and optimisation algorithms, with examples including Monte Carlo and molecular dynamics simulations, which will essentially decide the efficiency of the overall conformational searches. The detailed discussions on the various conformational searches can be found in many structure modelling literatures (Liwo *et al*., 2007; Simons *et al*., 1997; Xu and Zhang, 2012; Zhang *et al*., 2002).

Atomic-level model construction and refinement: Since most structural reassembly methods represent the protein chain by a reduced model, the detailed backbone and side-chain atoms need to be added for full-length atomic model construction, which is also to increase the practical usability of the structural models (**Figure 2c**). Here, basic rules of physical realism as observed in the experimentally solved structures, including bond length and angle constraints, steric overlaps and hydrogen-bonding network, should be satisfied. A number of algorithms have been developed to construct the full-atomic models from the reduced models, for example, Maxsprout (Holm and Sander, 1991), Pulchra (Rotkiewicz and Skolnick, 2008), REMO (Li and Zhang, 2009) and so on. Several programs also attempt to refine the topology of the structural models while adding the missed atoms (Zhang *et al*., 2011).

Model evaluation and selection: A number of structural conformations (also called structural decoys) will be resulted from the structural assembly simulations. The objective of this last step is to select the high quality 3D model of correct fold from all the possible alternative conformations that are closest to the native structure. A naïve approach is to perform a stereo-chemical check and determine how the model deviates from the basic regularities of known experimental structures. However, the structural models with the best local feature and physical realism do not necessarily correspond to that with the topology closest to the native state. Variants of all-atom physics-based and statistical potentials are often used for model quality estimation. Due to the importance, this effort of model ranking and selection has resulted in a new category called Model Quality Assessment Programs (MQAP) in the CASP experiments (Cozzetto *et al*., 2009; Fischer, 2006; Kryshtafovych *et al*., 2011) which aims to identify the best models from all the structures generated by the community of predictors. So far, the best method for model selection in MQAP is that based on consensus, that is, supposing that the models generated by most programs

have on average the best quality. Accordingly, structural clustering is a popular tool for model selection in many structural prediction pipelines (Roy *et al*., 2010; Simons *et al*., 1997). One such clustering tool is SPICKER which was designed to find the structural decoys which have the most number of neighbours in a hierarchical way (Zhang and Skolnick, 2004a, b).

Applications of Protein Structure Predictions

The biological usefulness of the models relies on the accuracy of the structure prediction (**Figure 3**). For example, models generated by CM using close homologues, usually meet the highest structural requirement and can be successfully used for studying the effect of SNP or mutations, designing new proteins using site-directed mutagenesis and screening compound libraries for structure based drug-discovery (Zhang, 2009). Of note, Sali and colleagues (Schlessinger *et al*., 2011) recently screened the KEGG Drug library against the model of the norepinephrine transporter (NET) generated by CM, leading to the discovery and experimental validation of a novel ligands for NET. Another important application of these high resolution CM models, has been shown by Tramontano and co-workers (Giorgetti *et al*., 2005), where models with GDT-score >0.84 were successfully used to obtain the phase information of the X-ray diffraction data by molecular replacement (MR). The authors found that the MR performance depends on the overall quality of the models, rather than on the local structures. Moreover, the best available structural templates identified by threading were much less successful in MR than the complete models, which buttress the importance of structural refinement in protein structure prediction.

Medium-resolution models, roughly in the range of 2.5–5 Å RMSD to native structure, are typically generated by threading and CM using distantly homologous templates. These models can be used for identifying the spatial locations of functionally important residues, such as active/binding sites and sites of disease-associated mutations. For example, Arakaki *et al*. (2004) assessed the possibility of assigning enzyme commission (EC) number by matching the active site motifs gleaned from structure decoys of various resolutions, and found that models of 3–4 Å resolution can be used to assign the first three digits of the EC number with an accuracy of 35%, whereas the accuracy drops down to 22% when models of 4–5 Å are used.

Even models with lowest resolution from otherwise meaningful predictions, that is, models with an approximately correct topology, predicted using either *ab initio*

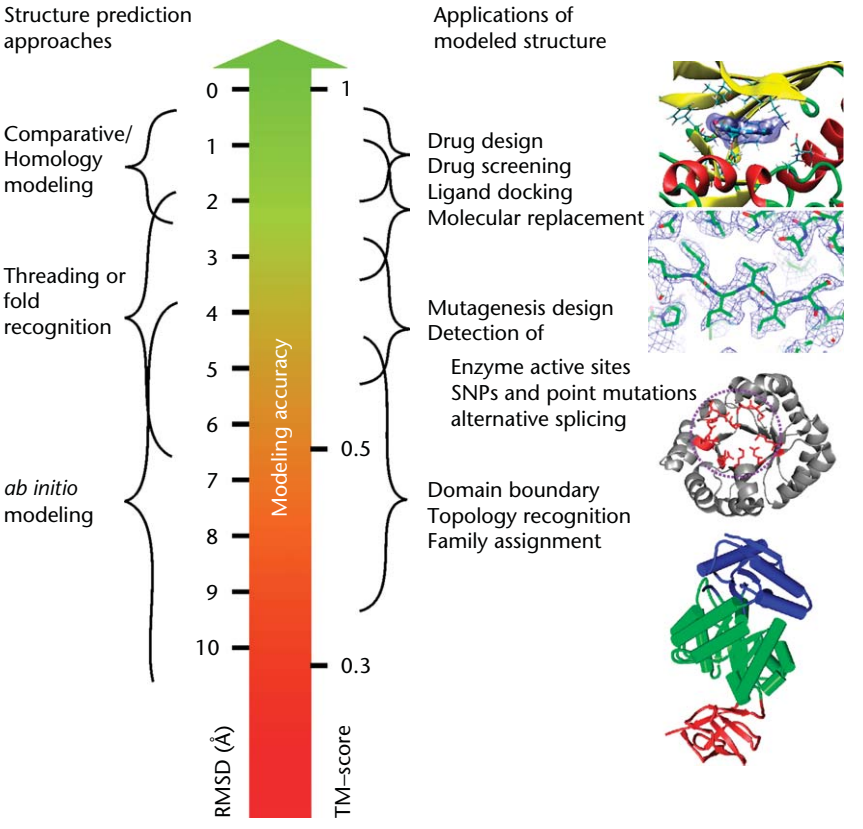

Figure 3 Approximate correspondence of the structure prediction algorithms, model accuracy, and the biological usefulness.

approaches or based on weak threading hits have a number of uses, including protein domain boundary identification, topology recognition and family/superfamily assignment. For instance, Malmstrom *et al*. (2007) modelled 3338 small protein domain (<150 residues) proteins in the yeast *Saccharomyces cerevisiae* genome using ROSETTA; the SCOP superfamily assignment was generated for 404 domains with trustable confidence scores, based on the structural comparison between the predicted models and the SCOP structures. An additional 177 assignments were made after integrating the Gene Ontology (GO) annotations.

One of the main impetuses for predicting protein structures is to use them for structure-based functional annotation. A convenient approach to the structure-based functional assignment involves global structural comparison of protein pairs for fold recognition and family assignment (Malmstrom *et al*., 2007; Zhang *et al*., 2006), which in many cases can be used to correctly infer the function. However, it is also recognised that the relationship between structure and function is not always straightforward, as many protein folds/families are known to be functionally promiscuous (Roy *et al*., 2009), and different folds can perform the same function. When the global structures are not similar, functional similarity may arise due to the conserved local structural motifs, which perform the same biochemical function, although in different global structural frameworks. In a recent development, Roy and Zhang (Roy and Zhang, 2012) showed that using low to medium resolution receptor structures and a combination of local and global structural similarities, ligand binding pockets can be identified in 65% of cases with an average error of 2 Å. Without knowing the ligand a priori, the ligand interacting residues assignment can be made with an average Matthews correlation coefficient of 0.60 and precision of 0.73.

Challenges of Modelling Transmembrane and Disordered Proteins

Despite the progress in protein structure prediction, many serious challenges need to be addressed. For instance, in the *ab initio* structure prediction category, we have no success in modelling medium-to-large size proteins with >150 residues. Similarly, in the template-based modelling (TBM) category, the refinement of predicted models beyond the best available templates is a major limitation. In the following, we review the relatively new emerging challenges in modelling the transmembrane proteins and the intrinsically disordered proteins.

Structure modelling of membrane proteins

Approximately 20–25% of proteins coded in the sequenced genomes are transmembrane proteins (TMP) (Krogh *et al*., 2001). Transmembrane proteins have diverse functional roles, which include the involvement in nutrient and metabolite transport, information flow, as well as energy production. Not surprisingly, TMPs are the most important targets for developing new pharmacological agents. However, despite of the rapid growth in the PDB library, TMPs represent only <2% of all known structures in the PDB, because they are both hard to crystallise and intractable by NMR.

Most of the existing bioinformatics tools for studying membrane protein structure are focused on predicting either (a) the location of trans-membrane domain, or (b) their topology, that is, the cellular location of *N*- and *C*- terminus of the polypeptide chain. These predictions are useful for the designing of further experiments to unravel the location of loops and number of trans-membrane segments. But the structural details are still missing in these 2D models, and 3D structure modelling of TMPs is usually required in the experiments.

The structure prediction of transmembrane proteins is generally considered more difficult than the globular ones, because of the lack of homologous protein structure in the template structure library. Moreover, the spatial profile of hydrophobic residues in TMPs are inverted compared to the globular proteins, which contributes negatively to the scoring functions of most protein structure prediction algorithms designed for globular proteins, as they promote hydrophobic residues towards the interior of the protein. Meanwhile, TMPs are on average >200 residues, which poses a formidable challenge to *ab initio* structure prediction because of the combinatorial complexity. The limitations have been partially alleviated with the development of hybrid methods that combine *ab initio* fragment assembly and sparse experimental restraints. For example, FILM by Pellegrini-Calace *et al*. (2003) developed a membrane specified potential which was combined with the *ab initio* structural assembly approach and able to fold a handful of small transmembrane proteins in 3–7 Å. Zhang and Skolnick (2004a, b) used TOUCHSTONE-II to reassemble the template fragments as identified by threading and successfully fold 6 out of 18 nonhomologous TMPs <300 residues with a RMSD below 6.5 Å. Recently, Barth *et al*. (2009) showed that models within 4 Å can be generated for transmembrane proteins of 190–300 residues, by constraining the helix–helix packing arrangements at particular positions as predicted from sequence or identified by experiments.

Although promising results have been demonstrated by the hybrid *ab initio* fragment assembly approaches, the accurate modelling of large-scale transmembrane proteins still rely on the presence of experimental template proteins. Fortunately, for G protein-coupled receptors (GPCRs) which comprise the largest family of TMPs and are considered as the most dominant drug targets, several proteins have been recently solved with the structures deposited in the PDB library (Cherezov *et al*., 2007; Palczewski *et al*., 2000). These structures provide important insights about the spatial arrangements of functional residues, which can

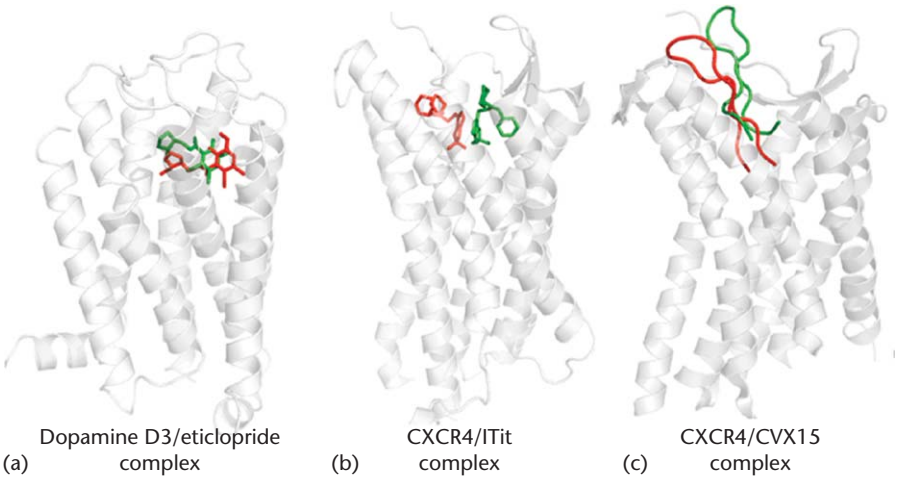

Figure 4 Predicted protein–ligand complexes using I-TASSER and BSP-SLIM in GPCR-Dock 2010. (a) Dopamine D3/eticlopride complex; (b) CXCR4 chemokine receptor with compound IT1t; and (c) CXCR receptor with peptide CVX15. The native ligand binding pose is shown in green and the predicted ligand pose in red.

be used for mutational analysis and ligand-docking experiments of other homologous GPCR proteins (Zhang *et al*., 2006). For example, in the recent community wide GPCR Dock 2010 experiment, the structures of three human GPCRs complexes: (1) Dopamine D3/eticlopride complex, (2) CXCR4/IT1t complex and (3) CXCR4 with CVX15, were solved and the participing predictors were asked to model the receptor structure and dock the co-crystallised ligand before the structures were released (Kufareva *et al*., 2011). This experiment provided a unique opportunity to objectively examine both the quality of the receptor structure modelling and ligand docking on modelled GPCR structures. The predicted receptor structure for the three complexes using I-TASSER had RMSDs of 1.6 Å, 2.27 Å and 2.82 Å to the crystal structures in the transmembrane region (**Figure 4**). Despite the high quality of the receptor models, the docked ligand conformation in these receptors had an RMSD of 3.42 Å, 9.78 Å and 8.88 Å for D3/eticlopride, CXCR4/IT1t and CXCR4/CVX15, respectively, which highlight the difficulty of low-resolution structure-based ligand docking. In D3/eticlopride complex, the binding pocket mainly lies in the transmembrane region, and therefore the ligand has a lower RMSD. Contrarily, the ligands in CXCR4/IT1t and CXCR4/CVX15 interacts extensively with ECL2, which shows large structure variability in templates and was therefore predicted incorrectly. Improvements in the accuracy of loop modelling of GPCRs will therefore have a profound impact on the accuracy of ligand-binding mode predictions and drug design.

Modelling of disordered proteins

The centre of the sequence-to-structure-to-function paradigm is the hypothesis that the amino acid sequence encodes for a structurally ordered 3D structure, which determines the functions of the protein molecule. However, it is increasingly realized that the intrinsic disordered proteins (Dunker *et al*., 2001) are common in many genomes, where either a part of protein (Intrinsic Disordered Regions) or sometimes the entire protein (Intrinsic Disordered Protein) exists as an ensemble of variable structures, which helps them to perform multiple functions by utilising their large intermolecular interfaces. Due to the inherent flexibility, it is difficult to characterize the structure of the disordered regions using the traditional X-ray crystallography or NMR techniques. Very often, the disordered residues can become ordered upon binding.

Structure predictions of disordered proteins are mainly focused on the identification of the disordered/ordered regions along the sequences (Monastyrskyy *et al*., 2011). The first method for computational disorder prediction was developed 15 years ago (Romero and Dunker, 1997), which was based on neural network training on a verity of sequence-based features to recognise the disordered regions. Since then, more than 50 methods to identify disorder have been developed. The major differences among the methods are in the training algorithms (neural network or support vector machine) and feature selections which includes residue-level and window-level information calculated from amino acid sequences, sequence profile from multiple sequence alignments, secondary structure, solvent accessibility, torsion angle, etc. Although the performance of different disorder prediction algorithms has become converged due to the similarity in methodology development, it was recently demonstrated that the information from multiple structure comparisons can help improve the accuracy of disorder prediction, that is, the regions that are structurally conserved within multiple models tend to be ordered, whereas the structurally varied regions in the models are most associated with disorder (McGuffin, 2008). The recent progress in the disorder structure predictions can be found in recent review (He *et al*., 2009) and the CASP assessment (Monastyrskyy *et al*., 2011). A

conceptual explanation of the intrinsically disordered proteins was given in. **See also**: Intrinsically Disordered Proteins

Conclusion

We have presented a general overview of the three categories of approaches used for protein structure prediction, including comparative modelling, threading and *ab initio* folding. Despite the significant efforts made in the field, template-based modelling, which construct models based on other solved protein structures, is the only reliable method for high-resolution predicted protein structure. Nevertheless, composite approaches that combine tools of threading, fragment assembly, *ab initio* modelling and structural refinements have demonstrated powers in modelling proteins of different homology level of protein sequences. A commonly used pipeline for the composite protein structure prediction was dissected, for illustrating the procedures.

Predicted protein structures have been extensively used for ligand screening and structure based drug-design, detecting functional site residues and designing mutagenesis experiments, helping molecular replacements, or identifying the impact of disease-associated SNPs and point mutations. Even for the models built from weakly homologous templates or by *ab initio* modelling, correctly predicted folds have been used for assigning protein families or identifying approximate domain boundaries. Although experiment structures are undoubtedly the most desirable, predicted models span many needs of most biologists.

Membrane proteins and intrinsically disordered proteins are often ignored by the mainstream protein structure prediction community. However, they play a crucial role in many cellular processes. Developing efficient algorithms for modelling these proteins will enlarge the scope of protein structure predictions, which will help us better understand the molecular basis of their functions.

Acknowledgements

This work is supported in part by the NSF Career Award (DBI 0746198) and the National Institute of General Medical Sciences (GM083107, GM084222).

References

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181**: 223–230.

Arakaki AK, Zhang Y and Skolnick J (2004) Large scale assessment of the utility of low resolution protein structures for biochemical function assignment. *Bioinformatics* **20**: 1087–1096.

Barth P, Wallner B and Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proceedings of the National Academy of Sciences of the USA* **106**: 1409–1414.

Berman HM, Westbrook J, Feng Z *et al*. (2000) The Protein Data Bank. *Nucleic Acids Research* **28**: 235–242.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.

Bradley P, Misura KM and Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.

Cherezov V, Rosenbaum DM, Hanson MA *et al*. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318**: 1258–1265.

Cozzetto D, Kryshtafovych A and Tramontano A (2009) Evaluation of CASP8 model quality predictions. *Proteins* **77**(suppl. 9): 157–166.

Das R, Qian B, Raman S *et al*. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**: 118–128.

Dunker AK, Lawson JD, Brown CJ *et al*. (2001) Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling* **19**: 26–59.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Fischer D (2006) Servers for protein structure prediction. *Current Opinion in Structural Biology* **16**: 178–182.

Ginalski K, Elofsson A, Fischer D and Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**: 1015–1018.

Giorgetti A, Raimondo D, Miele AE and Tramontano A (2005) Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics* **21**(suppl. 2): ii72–ii76.

He B, Wang K, Liu Y *et al*. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Research* **19**: 929–949.

Holm L and Sander C (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *Journal of Molecular Biology* **218**: 183–194.

Jauch R, Yeo HC, Kolatkar PR and Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**: 57–67.

Krogh A, Larsson B, von Heijne G and Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**: 567–580.

Kryshtafovych A, Fidelis K and Tramontano A (2011) Evaluation of model quality predictions in CASP9. *Proteins* **79**(suppl. 10): 91–106.

Kufareva I, Rueda M, Katritch V, Stevens RC and Abagyan R (2011) Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure* **19**: 1108–1126.

Li Y and Zhang Y (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* **76**: 665–676.

Liwo A, Khalili M, Czaplewski C *et al*. (2007) Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization

method with single training proteins. *Journal of Physical Chemistry B* **111**: 260–285.

MacCallum JL, Hua L, Schnieders MJ *et al*. (2009) Assessment of the protein-structure refinement category in CASP8. *Proteins* **77**(suppl. 9): 66–80.

Malmstrom L, Riffle M, Strauss Charlie EM *et al*. (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biology* **5**: e76.

Marti-Renom MA, Stuart AC, Fiser A *et al*. (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure* **29**: 291–325.

McGuffin LJ (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* **24**: 1798–1804.

Monastyrskyy B, Fidelis K, Moult J, Tramontano A and Kryshtafovych A (2011) Evaluation of disorder predictions in CASP9. *Proteins* **79**(suppl. 10): 107–118.

Moult J, Fidelis K, Kryshtafovych A, Rost B and Tramontano A (2009) Critical assessment of methods of protein structure prediction-Round VIII. *Proteins: Structure, Function, and Bioinformatics* **77**: 1–4.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443–453.

Palczewski K, Kumasaka T, Hori T *et al*. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**: 739–745.

Pellegrini-Calace M, Carotti A and Jones DT (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins* **50**: 537–545.

Read RJ and Chavali G (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins* **69**(suppl. 8): 27–37.

Romero Obradovic and Dunker K (1997) Sequence data analysis for long disordered regions prediction in the Calcineurin family. *Genome Inform Ser Workshop Genome Inform* **8**: 110–124.

Rotkiewicz P and Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry* **29**: 1460–1465.

Roy A and Zhang Y (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure* **20**: 987–997.

Roy A, Kucukural A and Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**: 725–738.

Roy A, Srinivasan N and Gowri VS (2009) Molecular and structural basis of drift in the functions of closely related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biology* **9**: S41–S55.

Rychlewski L, Jaroszewski L, Li W and Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science* **9**: 232–241.

Schlessinger A, Geier E, Fan H *et al*. (2011) Structure-based discovery of prescription drugs that interact with the norepinephrine transporter, NET. *Proceedings of the National Academy of Sciences of the USA* **108**: 15810–15815.

Simons KT, Kooperberg C, Huang E and Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology* **268**: 209–225.

Skolnick J (2006) In quest of an empirical potential for protein structure prediction. *Current Opinion in Structural Biology* **16**: 166–171.

Skolnick J, Kihara D and Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Protein* **56**: 502–518.

Soding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951–960.

Summa CM and Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences of the USA* **104**: 3177–3182.

Tramontano A and Morea V (2003) Assesment of homology based predictions in CASP 5. *Proteins* **53**: 352–368.

Wallner B and Elofsson A (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins* **69**: 184–193.

Wu S and Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**: 547–556.

Wu ST and Zhang Y (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research* **35**: 3375–3382.

Xu D and Zhang Y (2012) Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-based Force Field, *Proteins* **80**: 1715–1735.

Xu D, Zhang J, Roy A and Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79**(suppl. 10): 147–160.

Xu Y, Xu D, Crawford OH *et al*. (1999) Protein threading by PROSPECT: a prediction experiment in CASP3. *Protein Engineering* **12**: 899–907.

Yang Y, Faraggi E, Zhao H and Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**: 2076–2082.

Zhang J, Liang Y and Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**: 1784–1795.

Zhang J, Wang Q, Barz B *et al*. (2010) MUFOLD: A new solution for protein 3D structure prediction. *Proteins* **78**: 1137–1152.

Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**: 108–117.

Zhang Y (2009) Protein structure prediction: when is it useful? *Current Opinion in Structural Biology* **19**: 145–155.

Zhang Y and Skolnick J (2004a) SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* **25**: 865–871.

Zhang Y and Skolnick J (2004b) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal* **87**: 2647–2655.

Zhang Y, Devries ME and Skolnick J (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Computational Biology* **2**: e13.

Zhang Y, Kihara D and Skolnick J (2002) Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48**: 192–201.

Zhang Y, Kolinski A and Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal* **85**: 1145–1164.

Zhou H, Pandit SB, Lee SY *et al*. (2007) Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins* **69**(suppl. 8): 90–97.

Further Reading

Baker D and Sali A (2001) Protein structure prediction and structural genomics. *Science* **294**: 93–96.

Elofsson A and von Heijne G (2007) Membrane protein structure: prediction versus reality. *Annual Review of Biochemistry* **76**: 125–140.

Fink AL (2005) Natively unfolded proteins. *Current Opinion in Structural Biology* **15**(1): 35–41.

Zhang Y and Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the USA* **102**(4): 1029–1034.

Zhang Y (2008) Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology* **18**: 342–348.