Template-Based Modeling of Protein Structure

David Constant
Biochemistry 218
December 11, 2011

Introduction.

Much can be learned about the biology of a protein from its structure. Simply put, structure determines function. Protein structure prediction has become more and more important as the number of available non-redundant protein sequences grows while experimental structure determination methods remain expensive and time consuming. The field of structural prediction continues to advance, and has made some progress on filling the gap between the number of available sequences and structures, but there is much ground to be made up as the known proteome expands.

There are two general approaches to protein structure prediction: *ab initio* modeling and template-based modeling (sometimes called comparative or homology modeling). In *ab initio* modeling, chemical and physical constraints are used to predict the most favorable structure for a given sequence. The main limitation for this technique is its enormous computational cost. Comparative techniques, however, use previously determined structures of proteins similar to the target as templates for building a model of the query. These techniques rely on the fact that proteins with similar functions (especially those evolutionarily related) often have similar sequences, which adopt specific structural conformations. The overall similarity as well as the alignment of the query and template sequences will obviously have an effect on the quality of the ultimate model that is predicted. In recent years, the line between *ab initio* and template-based modeling has become increasingly blurred, and many modeling programs feature some elements of both approaches for an overall increase in the accuracy of predicted protein structures. In this review, I will focus on template-based structural modeling techniques. I will describe several publicly available prediction servers that participated in the most recent Critical Assessment of Structural protein Prediction (CASP) experiment, CASP9, held in 2010.

Finally, I will outline some of the main challenges to further advancement the field currently faces.

Template-based modeling

In brief, the process of template-based protein structure prediction consists of identifying proteins with solved structures that are homologous to the query, aligning the query to the template, and refining the model. In practice, the process is much more complex, and the actual methods employed by various prediction programs vary significantly. Notably, the best-ranked servers that participated in the CASP 9 experiment used distinct methodology to generate high-quality models. While some techniques for performing each step are clearly superior, others are more or less suitable for a particular experimental goal. The CASP 9 results show that there are multiple ways to generate accurate models and one method or server does not dominate all others [1].

Template based modeling relies on the principle that proteins with similar 3D structures have similar primary sequences. The accurate identification of evolutionarily or functionally related proteins based on sequence identity is essential to the generation of a near-native model using this method. Simple BLAST searches comparing sequences to sequences can be sufficient for very easy queries (those with very high sequence identity to good template proteins). More sophisticated techniques greatly improve the models generated for intermediate and difficult queries, where the best available templates are 30-50% and <30% identity to query, respectively. In these cases, using alignment techniques comparing the query to templates using sequence profiles generated by PSI-BLAST or HMM can result in the identification of many more true-positive homologous proteins and consequently much better templates [1,2]. In the past, the use of PSI-BLAST greatly increased the accuracy of homology modeling, and the CASP 9 organizers performed a rough analysis to determine the degree to which template identification has advanced past this technique. They created PSI-BLAST sequence profiles for each query in CASP 9, and found that in most cases, the servers were able to find better templates than those found by these profiles for up to 96% of the queries [1]. While this analysis is

far from rigorous, what is clear is that methods for identifying templates have continued to improve.

When the query has multiple types of domains, treating these domains separately can also significantly increase the model quality. Some automated prediction servers do this as a matter of course. This approach, however, increases the need for accurate domain identification and loop modeling and may not generate a good full-chain model. In cases where the sequence identity between query and template is >50%, predictions can be as close as ~1 Å C_α RMSD from the native structure. When the identity is between 30-50%, the models are rarely more than ~4 angstroms C_α RMSD from native (usually ~2-3 Å). Below 30% identity, template-based modeling is much less effective [2,3]. Generally, the more accurate the alignment is and the greater its coverage of the query, the more accurate the eventual model will be.

After identification of homologous template proteins, alignment of the query can progress. The template used for alignment can either be assembled from multiple solved structures, taking the best template for each domain or region and then building the full-length template from these fragments (called "threading"), or it can be simply the single best template found. Particularly in cases where the query has multiple domains or low homology to any known structures, threading is often preferable, but its success depends in large part on the accuracy and sensitivity of the initial template search. Multiple-templated techniques have been gaining popularity and, on average, increase the accuracy of the models generated with them over single-templated techniques [4].

Once the query has been aligned to a template, a 3D model can be constructed. The backbone of the model is built based on the template, and residue side-chain conformations (side-chain packing) are determined based on allowable rotamer conformation and sometimes optimization of free-energy states. Unaligned loops are modeled by either *ab initio* methods or using structural information from database searches. The accuracy of predicted loop regions is highly correlated with their length; if the region is less than ~6 residues, it is usually highly accurate, but becomes much more variable with increasing length [2]. Side-chain packing is usually very accurate, if the backbone alignment between the query and template is good, and is usually best within

densely packed regions. The importance of peripheral side-chain conformations depends on the biological question being asked, and only for certain applications is this critical information.

Refinement of the models generated from the template alignments is a necessary step, but can be prone to significant error and is one of the major bottlenecks in the field [5]. The most promising methods employ molecular dynamics to explore the conformational space of the sequence and find the most favored arrangement, but this requires a significant amount of computational power. A quicker but more deterministic method is to calculate the position of atoms as a probability density function based on the template sequence backbone and allowable bond angles and distances of the query. In either case, the model refinement process can easily reduce the overall accuracy of the model by either altering conserved and well-templated regions or failing to find optimal energy minima for a given set of atoms. It is hoped that further progress will be made to reduce the computational cost and and facilitate the accurate application of molecular dynamics approaches.

Publicly Available Prediction Servers

I-TASSER

Initially developed and tested in 2004 during the CASP6 experiment by Yang Zhang and Jeffrey Skolnick [6], Iterative-Threading/ASSEmbly/Refinement (I-TASSER, URL http://zhanglab.ccmb.med.umich.edu/I-TASSER/) is one of the best fully automated and publicly available template-based structure prediction servers, and was ranked first in the "server only" category of CASP7 and 8, and second in CASP9 [1]. I-TASSER has undergone much development since its debut in CASP6, and has become increasingly composite in methodology, combining sophisticated techniques for each stage of the model building process [6,7,8,9].

The first step I-TASSER performs is to create a sequence profile for the query using PSI-BLAST. The secondary structure of this sequence is then predicted using PSIPRED, a highly accurate secondary structure prediction server developed at the University College

London. Using the constraints provided by PSI-BLAST and PSIPRED, the query is then threaded through the PDB structure library using the Local Meta-Threading-Server (LOMETS, URL http://zhanglab.ccmb.med.umich.edu/LOMETS/), which uses eight component servers (FUGUE, HHsearch, MUSTER, PROSPECT2, PPA-I, SAM, SP3, and SPARKS) to find the best possible templates for the query. The component servers generate 20 models each ranked by Z-score, for 160 total models from which the ten best are selected by a function that weights models based their Z-score and the component servers average TM-score (a global comparison score measuring how similar a model is to an experimental structure).

The continuous fragments from the threading alignments are then excised from their respective template structures and assembled into a full-length model with un-templated regions built by *ab initio* modeling; the model is initially optimized by a replica-exchange Monte-Carlo simulation, guided by a knowledge-based force field combining spatial restraints, contact predictions, and backbone/side-chain correlations. The conformations from low-temperature replicas are clustered and averaged with the SPICKER algorithm [10] (URL http://zhanglab.ccmb.med.umich.edu/SPICKER/) and the cluster centroids are then submitted to a second round of threading and refinement. Finally, the lowest-energy structures from this second round are selected and all-atom models from the C_α backbone are built primarily by optimizing H-bonding networks, and then subjected to a further refinement step called fragment-guided molecular dynamics (FG-MD) to further refine and optimize local geometry, H-bonding, and steric clashes of each atom in the model. As a final step, the I-TASSER pipeline annotates the finished model by matching it against protein function databases and provides gene ontology terms and enzyme classification numbers.

HHPred/MODELLER

The HHpred server, in 2005, was the first to use profile HMMs to identify templates for modeling [11]. Developed by the Söding group, HHpred in its current form (URL http://toolkit.lmb.uni-muenchen.de/hhpred) was ranked highest in the "server only" category in CASP 9 [1]; aside from this, the main advantage of HHpred is its vast increase in speed relative to other methods without reducing accuracy – where I-TASSER

(and others) takes on the order of days, whereas HHpred returns structures in mere minutes [12]. From the point of view of attaining the goal of generating a structure for every known protein, this method is clearly superior; however, HHpred does not include any alternative alignments, and relies on the separately developed MODELLER program to actually generate the model. The HHpred alignment, once generated, is used to build several models using the fully automated MODELLER program, and the best is chosen as the final model.

Several different versions of HHpred were run in CASP9, but their results were identical and their methods very similar [1,12], so the method will be discussed in general. First, a query sequence alignment is performed by iterative HMM searches against a non-redundant database, and an HMM profile is built. A database containing HMMs for a subset of PDB sequences with known structures is then searched with the query HMM using HHsearch, an algorithm for pair-wise alignment of HMMs and ranking of alignments. A trained neural network is then used to predict the TM-score of the model that would be built by each HMM alignment, and the alignments are re-ranked based on this. Then, from the top-ranked alignments for each segment of the query, a full-length alignment is built for 3D structure prediction using MODELLER.

The MODELLER program (available stand-alone at http://salilab.org/modeller/) is used by HHpred to build 3D models based on the profile-profile HMM alignments generated in the first phase. The 3D structures are built by satisfying the spatial restraints of the C_α-C_α bond lengths and angles, the dihedral angles of the side-chains, and van der Waals interactions. These restraints are calculated from the template structures, and forms a model that represents the most probable conformation of the query based on homologous proteins [13]. MODELLER is generally considered a very good structural prediction program, but it does not model side-chain conformations very well given their inherently higher degree of positional uncertainty.

Robetta

The ROBETTA server was ranked fifth in the “server only” category of CASP9 (as Baker-Rosetta server) behind HHpred, I-TASSER (as Zhang-server), QUARK (also out

of the Zhang lab) and Seok-server, a server that is not publicly available based out of Korea. ROBETTA combines the ROSETTA method of all-atom model refinement to their lowest free-energy state with a meta-server for template identification and alignment [14]. The query sequence is first parsed into domains using the Ginzu protocol, developed by David Baker's group [15], which uses BLAST, pairwise HMM searching, and PSI-BLAST to identify domains and their discrete boundaries. Unlike I-TASSER and HHpred, if ROBETTA finds multiple domains in a query, it will model them separately and then assemble those models into a full chain later, which is a major difference in the techniques. Once the domains have been parsed and suitable templates have been identified, multiple alignments are generated by several methods (HHSEARCH, Compass, and Promals) and the best are selected for modeling. Loop regions are assembled from fragment libraries and optimized to fit the template.

The core of ROBETTA is in the model refinement process, which uses Monte Carlo simulations to find low-energy conformations of a given model. Each attempted move in these simulations consists of random backbone torsion angles, optimized side-chain rotamer conformation and minimization of the disagreeable backbone and side-chain angles. For easy queries, a single best template is often easy to find using this method, which employs BLAST for just this reason. But when the queries are more difficult, there are often multiple similarly ranked templates. In this case, ROBETTA performs multiple randomly seeded energy-minimizing Monte Carlo simulations to determine the template that results in the lowest energy conformation of the query. This adds a large computational burden to harder targets, which is the most significant drawback of the ROBETTA server.

Progress and Obstacles

Template selection

Significant advances have been made in template selection and building using sophisticated techniques such as PSI-BLAST and HMM profile-profile alignments, but little progress has been made since they became common. In the analysis of the CASP 9 results, it was found that even the best methods (mufold and HHpred) were only able to

produce a model better or even as good as the "single best available template" in the PDB for at most ~30% of queries, and the majority of prediction servers performed wors than this [1]. However, there was at least one model submitted for each target that showed improvement over the best available template for roughly two thirds of the CASP 9 targets [1]. Additionally, the overall performance of servers in CASP 9 exceeded that of CASP8 only on targets of intermediate difficulty [16]. Optimistically, this can be seen as an increase in the difficulty of the targets in CASP9, but this is hard to show. Despite recent advances in template selection, though, there is still further progress to be made in this critical first step of homology modeling.

Model building

While the quality of models produced in the CASP experiments has consistently improved between rounds, progress has slowed recently [16]. In most cases, the improvement over the best available template is due to correct modeling of unaligned regions, most often by *ab initio* methods, which are quite accurate for short targets. It is clear that the potential for improvement over the best template is greater for those queries classified as "hard" targets, that is, those with low homology to available templates and more unaligned portions. Overall, for those queries with at least 15 unaligned residues 35% of these residues were correctly modeled (within 3.8 Å of the experimental structure) regardless of target difficulty [16]. The overall quality of template-based models is most heavily influenced by the degree of alignment of the query with the best available template. Directed development and expansion of the templates available in the PDB could have an enormous impact on the overall applicability of homology modeling.

Model refinement

One major obstacle of current homology modeling techniques lies in the choice of model refinement process. Often, the refinement process decreases the accuracy of the model by altering highly conserved or well-templated regions. This could also be contributing to the finding that for each server, only the minority of queries were modeled more accurately than if the best template was simply copied. It may be that the servers are in fact finding the best template, but then refining "away" from the native state of the query.

Given the organization of the CASP experiments, this hypothesis has thus far been untestable, but could easily be tested if the initial templates found by each server could be compared with the final model and the experimental structure. It would be expected that, if the refinement process were increasing accuracy, the final model would be closer to the experimental structure than the template. Clearly, further testing and development is necessary for the molecular dynamics techniques to meet their potential, not least because HHpred, which uses primarily quickly calculated positional likelihoods in refinement, out-performed other techniques which used very time-consuming and costly simulations to refine their models.

Quality/error assessment

For several rounds now, CASP experiments have included an explicit category for model quality assessment. In this experiment, the global and per-residue error values for each model as determined by quality assessment methods are compared with the values as determined by superposition of each model with the experimental structure. In CASP 9, there were several methods for which the correlation between these values approached perfection; however, there are several important caveats to this apparently very encouraging result. First, these are separate from actual structure prediction methods, the vast majority of which are not able to provide realistic confidence measures. Almost none are able to provide accurate models and good confidence measures at the same time. Second, the winning quality assessment methods all used clustering techniques, looking at multiple models from different servers, to calculate the quality of each individual model. In practice, it may not always be useful or feasible to obtain many models in order to calculate their quality. The performance of single-model quality assessments was rather poor relative to the clustering methods. An additional consideration is that if the low-quality models are removed from the assessment of the clustering methods, then they perform significantly worse, indicating that when the available models are of similar quality it becomes much more difficult to determine their absolute quality as compared to the experimental structure [17]. Most structural prediction methods do not provide a realistic quality assessment relative to native, and even the best stand-alone quality assessment programs need significant improvement for some situations [17,18].

Currently, even when provided with a predicted model the confidence measures should not be entirely trusted, but rather one should rely on more absolute measures of realistic bond angles, lengths, etc., to judge model quality. Additionally, use of more than one prediction method is recommended, and at all times the biology of the particular protein should always be kept in mind.

Conclusion

Template-based modeling techniques have become highly sophisticated, and in some cases can generate extremely accurate models from the primary sequence of a protein. The recent advances can largely be attributed to a more integrative approach to the problem of structure prediction, particularly with respect to template selection and threading as can be seen by the dominance of meta-servers in this step. The further integration of the most sophisticated and effective methods for each step of the process is necessary for the field to grow further. In particular, one of the most critical areas for improvement is in assessment of the accuracy of a model. If the models generated are to be useful for biologists, they must be trustworthy, and currently each model must be carefully scrutinized before it can be deemed reliable. The directed development of the available structural templates could increase the number of queries that can be modeled with confidence, and while this would be good for biology as a whole it would not advance the field of protein structure prediction explicitly, unless the entire goal is to generate a structure for every sequence known. I would argue that while this is a useful and important goal, the issue of how to "build" a protein from first principles is a more interesting and ultimately more rewarding – and yet, even with good templates it is still no trivial matter to model a folded protein. As the lines between *ab initio* and template-based modeling grow, it will be interesting to see where the field leads next. With CASP10 just around the corner in 2012, we may not have to wait long for at least part of an answer.

Bibliography

[1] V. Mariani, F. Kiefer, T. Schmidt, J. Haas, T. Schwede. Assessment of template-based protein structure predictions in CASP9. *Proteins* (2011) **79** Suppl 10:37-58

[2] L. Jaroszewski, V. Protein Structure Prediction Based on Sequence Similarity. *Methods in Molecular Biology* (2009) **569**:129-156

[3] X. Qu, R. Swanson, R. Day, J. Tsai. A Guide to Template Based Structure Prediction. *Current Protein and Peptide Science*, (2009) **10**:270-285

[4] J. Cheng. A multi-template combination algorithm for protein comparative modeling *BMC Structural Biology* (2008) **8**:18

[5] K. Ginalski. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology* (2006) **16**:172-177

[6] Y. Zhang, A. Arakaki, J. Skolnick. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* (2005) Suppl 7:91–98

[7] Y. Zhang. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* (2007) **69** Suppl 8:108-117

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

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

[10] Y. Zhang, J. Skolnick. SPICKER: a clustering approach to identify near-native protein folds. *Journal of Computational Chemisty* (2004) **25:**6

[11] J. Söding, A. Biegert, A. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* (2005) **33**:W244-W248 (Web Server issue

[12] Söding J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics **21**, 951-960. doi:10.1093/bioinformatics/bti125.

[13] A. Sali, T. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* (1993) **234**, 779-815

[14] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev,J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B. Kim, R. Das, N. Grishin, D. Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77**(Suppl 9):89-99.

[15] D. Chivian, D. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. Strauss, R. Bonneau, C. Rohl, D. Baker. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* (2003) **53**(Suppl 6):524-33

[16] A. Kryshtafovych, K. Fidelis, J. Moult. CASP9 results compared to those of previous CASP experiments. *Proteins* (2011) **79**(Suppl 10):196-207

[17] A Kryshtafovych, K. Fidelis, A. Tramontano. Evaluation of model quality predictions in CASP9. *Proteins* (2011) **79**(Suppl 10):91-106