New insights into template-based protein modeling techniques

David A. Tiberi[1], Minseon Kim*[2]

[1]Department of Anatomy and Cell Biology, McGill University, 3640 University Street, Montréal, Québec, H3A 2B2
[2]Department of Microbiology and Immunology, McGill University, 3775 University Street, Montréal, Québec, H3A 2B4

ABSTRACT

Introduction: While the development of genomic sequencing methods has greatly improved the efficiency of collecting sequence data, experimental methods to obtain structure information have been lagging significantly. In order to elucidate protein structures, researchers have developed computational structural modeling techniques such as homology modeling and fold recognition (threading). The general consensus is that homology modeling is a superior approach with templates of high sequence similarity to the desired target (>30%), whereas threading is better suited for lower (<30%) sequence similarity templates. We compared recently improved threading algorithms with homology modeling to test the validity of this consensus. **Methods**: The most current versions of MODELLER and I-TASSER were used for model generation. We then used common assessment criteria (N-Dope, Q-mean and PROCHECK) to verify the validity of the models. Structure comparisons were also made using Chimera's Cα root-mean-square deviation. **Results**: Contrary to our prior expectations, the model determined by threading showed similar or even better assessment results in some criteria compared to the model generated from homology modeling. Furthermore, the structure analysis showed that homology modeling and threading protocols yield models with root-mean-square deviations of under 2 Å when used on protein sequences that share sequence identities of at least 30% to the experimentally determined protein template. **Discussion**: We believe that recent improvements in threading algorithms will allow for broader applications of this methodology in large-scale modeling efforts. The fully automated steps could provide time efficacy. In contrast to popular belief in the modeling community, we have shown that threading could be a competitive means of modeling rather than a mere backup method.

KEYWORDS

Homology modelingfold recognition, template-based modeling, MODELLER, I-TASSER, Alpha-fetoprotein (AFP), Human serum albumin (HSA)

*Corresponding author:
minseon.kim@mail.mcgill.ca

Received: 8 January 2010
Revised: 7 March 2010

INTRODUCTION

Determining the structure of a novel protein from its primary sequence is vital to many aspects of modern biology. Its applications range from drug discovery in the pharmaceutical industry to enzyme optimization for biotechnological uses in industry (*1*). Due to efforts like the Human Genome Project and improved computing capabilities, the potential for molecular modeling to produce new biological insights has greatly increased (*2*). Currently, there are about 10.5 million protein sequences available in Swissprot and TrEMBL, of which the protein structures of barely 62,000 have been determined (*3*, *4*). Protein structural

genomics aims to solve one protein structure for each protein family. As long as one protein structure is derived experimentally, structures of proteins in the same family can be solved using computational means (*5*).

Computational modeling methods are separated into three broad approaches: homology modeling, *ab-initio*, and fold recognition (threading). Template-based modeling methods, such as fold recognition and homology modeling, are the most reliable for predicting the structure of a target protein (i.e., the protein sequence under study). However, their use is limited by the availability of an optimal template, a homologous protein (similarity due to common ancestry) with an experimentally determined structure (*6*). Homology modeling predicts protein structures based on their sequence similarity to homologous proteins with experimentally derived structures. This approach stems from the idea that evolutionarily related proteins tend to share structural similarities, which enables researchers to predict the structure of homologous proteins. Regions of conserved structure are computationally transferred from the template to target model, while the non-conserved regions are usually calculated with respect to favored energy states. *Ab-initio*, or free modeling, relies on basic thermodynamic assumptions but is not currently a practical modeling option. Lastly, fold recognition depends on limited number of protein structural folds is limited in nature. Thus, remote homologues can be identified through the shared folds between proteins even if sequence similarity is insufficient to identify potential template proteins. Fold recognition consists of placing and aligning the sequence of amino acids against a template structure. The software first searches the fold database, and the best-fitting fragments are selected. If no suitable fold is found in the database, ab-initio is used to build that section of the model (*7*). Consensus in the field of structural bioinformatics holds that homology modeling generates models that are closer to the native protein structure than fold recognition. That is, it produces models with lower root-mean-square deviation (RMSD) to the native protein structure, and it is the preferred approach when sequence similarity to a known template ranges from 30-50%. Fold recognition is mainly used when sequence similarity drops below 30%, since it can identify targets with only fold-level homology (*8*).

The modeling efforts described in this paper focus on human alpha-fetoprotein (AFP), a 590-amino acid serum protein with three domains (stable and autonomously folding regions) (*9*). AFP belongs to the blood plasma protein family, which also consists of human serum albumin (HSA), afamin and vitamin D-binding protein. It is produced at a high level by the fetal liver and yolk sac, but only trace amounts are found in normal adults. These background levels of AFP are normally maintained throughout the life of an individual except for a transient elevation in pregnant females. AFP selectively suppresses cell-mediated immunity and promotes cell proliferation (*10*). Blood levels of AFP are also used in pregnant women to detect fetal abnormalities such as Down syndrome and neural tube defects (*11*).

Determining the three dimensional structure of AFP poses several challenges for the structural community. The size and complexity of the molecule makes it difficult to obtain via recombinant DNA methods the amounts needed for x-ray crystallography (*12*).

However, the availability of experimentally determined HSA structures, that are similar in size (585 amino acids) and share high sequence identity to AFP, allows for computational modeling of AFP. We used template-based methods of homology modeling and fold recognition to build models of AFP based on HSA. While the overall sequence of AFP is 40% identical to HSA, the actual sequence identities for domains I, II and III are 29%, 41% and 48%, respectively (*9*). Based on the current consensus in the field, we expect that homology modeling will be most successful for domains II and III since they have the highest sequence identity to the HSA template. When sequence identity to the template drops below 30%, as is the case for domain I of AFP, homology-derived models become inaccurate due to sequence misalignment. Consequently, we predict fold recognition will yield a better model for domain I of AFP (13).

METHODS

Template-based modeling techniques were applied on a domain-by-domain basis using the following domain ranges: domain-I (amino acids 2-192), domain-II (amino acids 193-384) and domain-III (amino acids 385-591) (9).

HOMOLOGY MODELING

Homology modeling depends highly on template identification and the quality of the initial alignment. These crucial steps are followed by multiple-template modeling using MODELLER and subsequent loop-refinement.

TEMPLATE IDENTIFICATION

We searched for potential templates in the Protein Data Bank (PDB) using the MODELLER script build_profile.py. The script identified an HSA template (1N5U) and a vitamin D-binding protein template (1KXP). Due to its high sequence similarity to AFP, we used the 1N5U template exclusively. For the multiple-template modeling process, we searched the PDB for another HSA template. The difference between the two templates is suggested to be over 2 Å RMSD (*14*). We therefore selected the HSA structure 1AO6, which differs from 1N5U by 4.59 Å.

MODELLER

MODELLER is a homology modeling program, which creates target models by satisfying spatial restraints. Based on the alignment information, spatial restraints are derived and target models are generated with minimal violation to such restraints. MODELLER was chosen due to its reputation as one of the best performing modeling software available (*13*, *15*). We used Version 9v7 in this experimental protocol. MODELLER uses python scripts for each step of the process, including the manual refinements. It can be run on both Windows and Mac and is available at http://salilab.org/modeller/.

MULTIPLE-TEMPLATE MODELING

We aligned template structures using the MODELLER script salign_iterative.py. This script incorporates automatic iteration of the alignment procedure, rendering the parameter values unnecessary. The best alignment result based on a scoring function is displayed as an output file. We used the script align2d_mult.py to align the target sequence onto the template structures, incorporating both sequence and structure information. We then used the model_mult.py script to generate a set of five different models for each domain of AFP, resulting in a total of fifteen different models.

LOOP-REFINEMENT PROCESS

From the pool of generated models, we selected for each domain the model with the lowest N-Dope score, and hence the highest accuracy. Using DOPE-profiles, which visualize DOPE scores per residue as a graph, we chose residues with higher DOPE scores for the loop-refinement process. Loop regions occur where no conservation is found in the target-template sequence alignment; no conserved structures can be adopted from the template protein structure. For such regions, MODELLER enables *ab-initio* refinement using the script loop_refine.py, which gives a number of independently generated alternative loop conformations. The loop conformation with the lowest energy state is selected based on the n-DOPE scores.

FOLD RECOGNITION

AUTOMATED SERVER SELECTION

The most recent (2008) Critical Assessment of Techniques for Protein Structure Prediction (CASP) study formed the basis for the selection of automated threading servers. Using a double-blind approach, organizers make available to the structure prediction community sequences for which the crystallographic structure will be solved in the next few months, and they are challenged to make predictions of these targets. The study selected I-TASSER as the best automated prediction server. In addition, the large repository of published material on this server and its widespread use by the structural community ultimately led us to select it for this experiment.

I-TASSER

The target protein sequence is submitted along with an e-mail address to which the results will be sent. I-TASSER performs profile-profile searching of the PDB using the statistical profiles for sequences based on their tendency to mutate at each position. This enables broader detection of remote homologues that cannot be identified through mere sequence based searches (*16*). Aligned fragments are then assembled with unaligned fragments, which are built by means of *ab-initio*. The simulation built from this first round is then used by the program in an iterative step that further refines the model and chooses the model with lowest energy conformation as the final output. I-TASSER is available at http://zhang.bioinformatics.ku.edu/I-TASSER/.

STRUCTURE COMPARISON AND VISUALIZATION

To compare the models that were built using homology modeling and fold recognition, we created structural alignments using the MODELLER script salign_iterative.py. We then imported the output alignment file in .ali format into Chimera in order to match the alignment onto three dimensional protein structures. Chimera was chosen for visualization since it allows for simple importing of the alignment file, rendering manual adjustments unnecessary. From the imported alignment information, Chimera calculated the Cα root-mean-square deviation (RMSD) between the two structures.

FINAL ASSESSMENT STEP

Following model generation, we used N-DOPE, Q-mean and PROCHECK as quality assessment criteria. To simplify the procedure, we used N-DOPE scores to choose the best models for each domain. We then performed Q-mean and PROCHECK assessments upon this selection of top models in order to confirm their quality (*17*).

N-DOPE

N-Dope is derived from the original DOPE score, which is a statistical potential means used to quantify model accuracy. DOPE scores are not normalized with respect to protein size and have an abstract scale, so they cannot be used to make comparisons between different models. To allow for comparisons, normalized N-Dope scores are used. Lower values are indicative of higher accuracy (*14*).

Q-MEAN

Q-mean is a combination of five different statistical potentials enabling both global and local structural quality assessment. It is a relatively new assessment web server, which stresses combining several independent quality measures into one score. The web server is available at http://swissmodel.expasy.org/qmean.

Table 1. Human serum albumin structures that have been incorporated in template-based modeling

ID	RESOLUTION (Å)	NAME
1AO6	2.50	CRYSTAL STRUCTURE OF HUMAN SERUM ALBUMIN
1N5U	1.90	STUDY OF HUMAN SERUM ALBUMIN COMPLEXED WITH HEME
1GNI	2.40	SERUM ALBUMIN COMPLEXED WITH OLEIC ACID

The summary of each structure including resolution and its identification within the PDB database has been recorded. Homology modeling was performed using 1AO6 and 1N5U while fold recognition used 1N5U.

Table 2. Model assessment control results

	N-DOPE	QMEAN	PROCHECK
DOMAIN-I	-1.824	0.721	91.9%
DOMAIN-II	-1.759	0.698	95.0%
DOMAIN-III	-1.517	0.666	92.6%

As a control, the 1N5U serum albumin structure was analyzed using the three assessment criteria; namely, N-DOPE, Q-mean, and PROCHECK. All domains of 1N5U obtained scores that confirm their stable tertiary structure, which is expected since each domain was experimentally determined.

Table 3. Improvement of the AFP domain models from homology modeling after refinement step

BEFORE	REFINEMENT	AFTER REFINEMENT
DOMAIN-I	-0.838	-1.059
DOMAIN-II	-1.481	-1.611
DOMAIN-III	-0.745	-0.873

The N-DOPE scores of the domain models before and after the refinement step have been recorded. The scores confirm that the refinement has improved the relative accuracy of the models, which was quantified using N-DOPE. The corresponding changes in the DOPE-profile can be seen in FIGURE 1.

Table 4. Assessment results for AFP domain models

	N-DOPE		QMEAN		PROCHECK	
	Homology Modeling	Fold Recognition	Homology Modeling	Fold Recognition	Homology Modeling	Fold Recognition
DOMAIN-I	-1.059	-1.039	0.603	0.608	93.1%	93.0%
DOMAIN-II	-1.611	-1.671	0.611	0.599	91.0%	91.0%
DOMAIN-III	-0.873	-1.083	0.630	95.2%	95.2%	96.0%

The results for fold recognition generated models show slightly higher accuracy compared to the models from homology modeling although they are not as accurate as the experimentally determined structures. These results show that fold recognition was able to obtain models with similar or slightly higher accuracy to those of homology modeling.

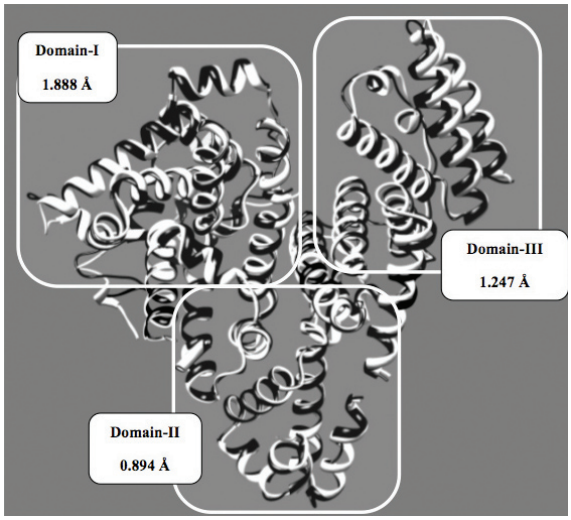

Fig. 1. RMSD measurements between the domain models made by homology modeling and fold recognition. The RMSD measurements have been utilized in order to quantify the similarity between the domain structures made in two different approaches. Yellow frame indicates homology-modeling model, while green indicates fold recognition model.

PROCHECK

PROCHECK assesses a protein model's stereochemistry, including its symmetry, geometry and packing quality (*13*). Among the many outputs that are given by PROCHECK, Ramachandran plots were the most utilized for our purposes. Each residue is arranged according to their stability; stable and accurate models are expected to have over 90% of their residues fall under the most favored region of the plot. PROCHECK scripts can be obtained by downloading PROCHECK-NT from http://ruppweb.dyndns.org/ftp_warning.html.

RESULTS

The assessment methods allowed us to obtain reference scores to which the generated models are compared. Table 1 describes the templates used. The assessment results performed on 1N5U are listed in Table 2. As expected, the 1N5U domain structures determined by x-ray crystallography yielded scores indicative of a native protein. This was confirmed by the three quality measurements.

Using homology modeling, AFP domain models were generated based on the 1AO6 and 1N5U templates. After the loop-refinement step, the energy profile was rendered more favorable, as evidenced by the N-Dope scores recorded in Table 3. This supports the idea that refinement steps, albeit requiring manual intervention, can improve the model quality significantly. Although all

three domains had N-DOPE scores less than zero, the domain II model has the highest relative accuracy based on N-DOPE scores. Q-mean and PROCHECK values, summarized in Table 4, also confirm that the generated models are within the acceptable range (see Methods for explanation), although the scores are shown to be less accurate than those of the control.

Table 4 summarizes the results obtained using fold recognition. I-TASSER selected 1GNI as the most suitable template, and the resulting models were subjected to the three assessment methods. Like in homology modeling, the N-DOPE scores obtained with I-TASSER were best for domain II, suggesting that domain II models were more accurate than those of domain I or domain III. Also, the N-DOPE scores obtained from fold recognition for domains II and III were more accurate than those obtained via homology modeling.

The overlay between homology and fold recognition models is shown in Fig. 1 along with the images of overlaid domain structures. All RMSD measures are below 2 Å, with domain II structures being the most similar.

DISCUSSION

Generating models of AFP allows for a comparative analysis of fold recognition and homology modeling. Our results challenge several key assumptions about these two techniques. The current consensus in structural bioinformatics is that homology modeling yields more accurate models than fold recognition. However, the assessment data we generated for these models indicate that both are capable of developing highly accurate models in the range of low to medium resolution x-ray crystallographic structures (Table 4).

The prevailing assumptions regarding the accuracy of both homology modeling and fold recognition must be revisited. The consensus is that fold recognition models often have a RMSD of 2-6 Å, with errors mainly occurring in the loop regions (*8*), while those of homology modeling often approach RMSD of 1-2 Å (*18*). This dogma was established some fifteen years ago, at a time when both of these techniques were still quite basic and unrefined. We do not dispute that for much of the last decade homology modeling has been the more accurate and preferred method of computational modeling when a template exists with sequence similarities above 30%. However, new algorithms have enabled the latest generation of fold recognition servers to generate models with accuracies that rival or surpass those of homology modeling. This result runs counter to the basic view in much of the published literature, and suggests a change in the assumptions regarding the accuracy of certain computational modeling techniques (*19*).

This paradigm shift first became apparent when I-TASSER generated the best 3D structure in CASP 7 in the automated server section (*6*, *20*). Two main factors contributed to the success of I-TASSER in this and subsequent competitions. First, an improved template refinement process that uses iteration was introduced, which reduces the RMSD by approximately 1 Å in the aligned regions (*20*). Furthermore, incorporating the refinement step with iteration skips the manual refinement step usually required during homology modeling. Unlike manual refinement, which is performed during homology modeling, I-TASSER's automated refinement ensures the same strict calculations and algorithms will be applied every time, making the process homogeneous. Secondly, the use of consensus target-template alignments (meta-server approach) by fold recognition software, including I-TASSER, greatly improves model generating capabilities. Consequently, the line between fold recognition and homology modeling has begun to blur.

The most recent CASP studies as well as our modeling work on AFP clearly provide evidence for fold recognition's ability to serve as a viable modeling method, even when sequences share over 30% similarity to known templates. Ideally, when attempting to model proteins with sequences that are 30-50% similar to known templates, researchers are encouraged to utilize both homology modeling and fold recognition approaches. This way, the results generated may be compared to each other, and the most reliable models can be selected. However, if time constraints must be considered, we feel that fold recognition, due to its rapid and user-friendly nature, may be used exclusively to generate models within the same range of accuracy as those made using a homology approach.

While researchers have greatly improved fold recognition servers, it is important to note that due to their automated protocols, bioinformaticians are unable to modulate the level of refinement, and thus the quality, of the resulting model. A novel and potentially timesaving approach would be to generate initial models using fold recognition with subsequent manual refinements using MODELLER. In order to achieve the daunting task of modeling all of the deposited sequences, we believe large-scale structural genomics projects should focus mainly on fold recognition approaches due to their rapid, automated protocols. Although homology modeling can provide viable models, the manual refinement steps require human intervention, thus making this technique impractical for large-scale structural genomic projects. New experimental protocols based on initial fold recognition and subsequent model refinements with MODELLER may allow structural genomic projects to elucidate the vast number of protein sequences yet to be determined.

ACKNOWLEDGMENTS

We would like to thank Dr. Murgita for his guidance and supervision of this project.

REFERENCES

1. K. A. Thiel, *Nature biotechnology*. **22**, 513 (2004).
2. T. Schwede *et al., Structure*. **17**, 151 (2009).
3. Swissprot and TrEMBL (http://us.expasy.org/sprot/).
4. RCSB Protein Data Bank (http://www.rcsb.org/odb/).
5. M. Tekeda-Shitaka, D. Takaya, C. Chiba, H. Tanaka, H. Umeyama, *Current Medicinal Chemistry*. **11**, 551 (2004).
6. A. Kryshtafovych, K. Fidelis, *Drug Discovery Today*. **14**, 386 (2009).
7. K. Mizuguchi, *Drug Discovery Today: Targets*. **3**, 18 (2004).
8. Y. Zhang, *Current Opinion in Structural Biology*. **19**, 145 (2009).
9. T. Morinaga, M. Sakai, T. G. Webmann, T. Tamaoki, *Proc. Natl. Acad. Sci. USA* **80**, 4604 (August, 1983).
10. L. Olding, R.A. Murgita R.A. *Reproductive Immunology-Current Topics in Microbiology and Immunology*. pp. 159-187. (1997)
11. C. De Mees *et al., Molecular and Cellular Biology*. **26**, 2012 (2006).
12. R. Boismenu *et al. Protein Expression Purification*. **10**:10-26. (2007)
13. Z. Xiang, *Current Protein and Peptide Science*. **7**, 217 (2006).
14. N. Eswar, D. Eramian, B. Webb, M. Shen, A. Sali, *Methods In Molecular Biology-Clifton Then Totowa*. **426**, 145 (2008).
15. C. Sansom, *Biochemist*. **30**, 34 (2008).
16. L. A. Kelley, M. J. Sternberg, *Nat Protoc*. **4**, 363 (2009).
17. D. Eramian *et al., Protein Science*. **15** (2006).
18. R. J. Read, G. Chavali, *Proteins: Struct Funct Bioinfo*. **69**, 27 (2007).
19. Y. Zhang, *Current Opinion in Structural Biology*. **18**, 342 (2008).
20. Y. Zhang, *Proteins: Structure, Function, and Bioinformatics*. **69**, 108 (2007).