

A SOFTWARE PIPELINE FOR PROTEIN STRUCTURE PREDICTION

Michael S. Lee
Computational Sciences and Engineering Branch
U. S. Army Research Laboratory
Department of Cell Biology and Biochemistry
U. S. Army Medical Research Institute of Infectious Diseases
Frederick, MD 21702

In-Chul Yeh, Nela Zavaljevski, Paul Wilson, and Jaques Reifman
DoD Biotechnology High Performance Computing Software Applications Institute
U. S. Army Medical Research and Materiel Command
Frederick, MD 21702

ABSTRACT

We have developed a software suite to predict protein structures from sequence through the integration of multiple non-commercial programs. The Army and DoD medical and scientific communities will be able to use this software to annotate structures of sequenced pathogenic and host genomes. Such structural predictions can be used in therapeutic and vaccine design as well as many areas of basic biological research. In this work, initial assessments of the software are made. Most importantly, these tests include evaluation of the quality of predicted structural models as a function of sequence similarity to known protein structures.

1. INTRODUCTION

Protein structure prediction is an integral tool for the current proteomic and systems biology efforts taking place in the DoD and Army medical research communities. Many genomes of pathogenic organisms have been sequenced in recent years. However, biologists must uncover the structural and functional nature of the corresponding translated proteins. Accurate structural models of proteins can be used in computational drug and vaccine design. Protein structure models at lower resolutions are still useful for functional annotation. Furthermore, knowledge gleaned from structural models can help biologists determine which proteins are critical in metabolic pathways and should be targets for drug design and other experimental studies (Bonneau, Baliga et al. 2004).

Protein structure prediction algorithms generally fall into two categories: comparative modeling (Madhusudhan 2005) and *de novo*. Comparative modeling approaches aim to find similarities between the input (or query) sequence and one or more sequences of known protein structure templates. After identification of likely matches, the query sequence is aligned onto each candidate template to produce a model structure. This procedure is

surprisingly accurate for cases when the query and template sequences are very similar (e.g., sequence homology greater than 40%) (Kryshtafovych, Venclovas et al. 2005).

When the sequence homology to known protein structures is more limited, i.e., less than 30%, fold recognition/threading methods are employed. In threading methods, a sequence of unknown structure is threaded through templates from the structure database and alternative sequence-structure alignments are scored using various empirical conformational energy calculations. The best-performing threading programs, such as GenTHREADER (Jones 1999), FUGUE (Shi, Blundell et al. 2001), and 3D-PSSM (Kelley, MacCallum et al. 2000), use hybrid approaches that combine sophisticated energy functions with sequence and structure homology. We chose to use the fold recognition software PROSPECT II, which is also based on a hybrid approach. The authors of PROSPECT II reported excellent performance on the standard fold recognition benchmarks (Kim, Xu et al. 2003).

Given templates obtained from either sequence similarity or fold recognition, protein models are built using comparative modeling software, such as MODELLER (Fiser and Sali 2003) and Nest (Petrey, Xiang et al. 2003). Besides building models from alignments and templates, these programs try to find optimal side chain placement and provide putative structures where there are gaps in the alignment.

As the homologies between the query sequence and known structures become more remote, two problems arise. First, finding the best templates with threading becomes difficult. Second, even with the optimal structural template in hand, alignment of the query sequence onto this template becomes challenging when the query and template sequences differ substantially (Kryshtafovych, Venclovas et al. 2005).

Failure of a fold recognition algorithm to find analogous templates or lack of confidence in sequence alignments leads one to consider *de novo* programs to build model structures. Generally speaking, *de novo* algorithms aim to fold up a protein based on its sequence using an energetic function. Small templates (less than 10 residues apiece) are often used as building blocks to enhance the sampling of an exponentially large conformational space (Simons, Kooperberg et al. 1997). Research in *de novo* algorithms is still in its infancy and reflects the fact that the grand challenge protein-folding problem remains unsolved. However, limited success on small proteins (i.e., less than 100 residues) has recently been achieved (Bradley, Misura et al. 2005). Also, structural models derived by *de novo* algorithms can be used to predict fold types and, consequently, protein function (Bonneau, Baliga et al. 2004).

The Biotechnology High Performance Computing Software Application Institute (BHS AI) has developed a software package that integrates several state-of-the-art structure prediction methods. The program accepts as input a protein sequence, and as output provides domain boundary information, protein-fold identification, and three-dimensional atomic models. The resolution of the structural models is often related to the similarity of the query sequence with experimentally known protein structures. In this study, we describe the software that we have developed and assess the accuracy of each component.

2. METHODS

Our structure prediction pipeline is a software program that integrates several stand-alone structure prediction methods as seen in Figure 1. Other notable examples of pipelines developed in recent years include TASSER (Zhang, Arakaki et al. 2005) and Robetta (Kim, Chivian et al. 2004).

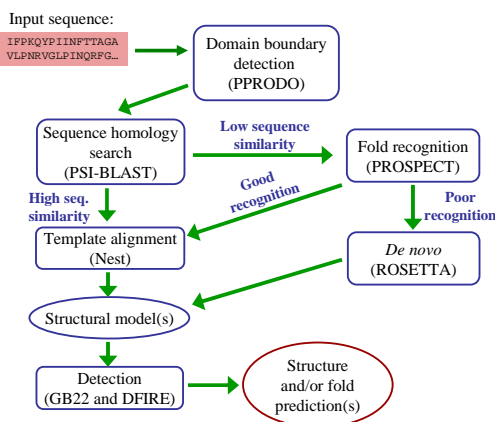


Figure 1. Schematic diagram of the protein structure prediction pipeline developed by the BHS AI.

First, the program PPRODO (Sim, Kim et al. 2005) is used to determine if the query sequence can be broken up into independently folding units or domains. This program requires output from the PSI-BLAST (Altschul, Madden et al. 1997) sequence alignment program and the PSIPRED (Jones 1999) secondary structure prediction tool. PPRODO uses a neural network algorithm to score each residue with the likelihood of being a linker region between domains. Currently, the query sequence can be broken into two domains if the maximum PPRODO residue score is above some cutoff value (see the Results section for values). Future work will consider the prediction of two or more domain boundaries for a given chain.

Next, the PSI-BLAST program is performed on each designated domain to determine the sequence similarity of this sequence segment with a database of the sequences of all of the known protein structures from the protein data bank (PDB) (Berman, Battistuz et al. 2002). First, PSI-BLAST is run for three iterations on a non-redundant database of multiple genomes. Then, the profile generated from this search is used in a single BLAST run on PDB sequences only. This protocol is called PDB-BLAST (Bujnicki, Elofsson et al. 2001).

If the best found sequence similarity is below some threshold (e.g., 20%), or no matches are obtained, a fold recognition program, PROSPECT (Kim, Xu et al. 2003), is used to deduce more distant relationships between the domain sequence and a library of thousands of protein fold templates derived from the SCOP 1.69 database (Andreeva, Howorth et al. 2004). In the present study, for the purposes of benchmarking, PROSPECT is run regardless of the highest PDB-BLAST sequence similarity.

We use PROSPECT in two stages. In the first stage, the complete template database is screened using PROSPECT with a simplified scoring function, which neglects pairwise interactions. In the second stage, a user-specified number of best hits is evaluated using a refined PROSPECT scoring function, which includes pairwise interactions. Fold recognition reliability is computed by random shuffling of the query sequence, which increases computation time.

With templates and sequence alignments from either PDB-BLAST and/or PROSPECT, a molecular modeling program, Nest (Petrey, Xiang et al. 2003), is used to build three-dimensional atomic models of the domain sequence using the obtained alignments and templates. Future work will consider the concatenation of multiple domains to produce a single protein structure (Kim, Chivian et al. 2004).

When no templates can be found for a query sequence, the *de novo* folding program Rosetta (Simons, Kooperberg et al. 1997) is used. Rosetta assembles three and nine-residue peptide fragments in random configurations to generate thousands of atomic models. This program is the most computationally intensive component of the pipeline and is limited to 150 residue segments given its computational complexity. From a large set of generated models ($N = 10,000$), the best ones are detected using a scoring function. We utilize two scoring functions, which are described below.

Given several template-based and/or *de novo* models, it is not possible to know which one is the closest to the actual native structure. For this reason, we assessed three criteria for their ability to detect the best model structures. The first criterion, most commonly used, is the percentage sequence similarity of the template to the query sequence. The assumption is that the higher the sequence similarity is, the more accurate the model is. This descriptor is, of course, not available for *de novo* models.

The second and third criteria are scoring functions. One scoring function is a knowledge-based potential known as DFIRE (distance-scaled finite ideal-gas reference) which has shown good success in detecting the native protein structure from a set of incorrect models (Zhou and Zhou 2002). The performance of DFIRE for detecting the nearest-to-native structure has not been tested as extensively (Summa, Levitt et al. 2005). The DFIRE model is outlined in detail elsewhere (Zhou and Zhou 2002). Generally speaking, DFIRE incorporates atomistic structural data from nearly 2000 single-chain X-ray protein structures. The DFIRE function incorporates the logarithm of the probability of interactions at different distances between all pairs of non-hydrogen atoms.

The other scoring function we tested is an all-atom force-field potential plus an implicit solvent model, PARAM22 / GBMV-SA, abbreviated here as GB22. This function consists of the CHARMM PARAM22 force field (Mackerell, Bashford et al. 1998), a generalized Born molecular volume implicit solvent model (Lee, Feig et al. 2003), and a surface area-based hydrophobicity term (Feig and Brooks 2002). This energy function is normally used in molecular mechanics simulations (Lee and Olson 2006), but has also been shown to be a good scoring function for structural model detection (Feig and Brooks 2002). Before scoring a structure with GB22, steric clashes are ameliorated with 200 steps of minimization using PARAM22 and a simple distance-based dielectric function (Feig and Brooks 2002).

There are many ways to numerically assess model structures against the native structure (Kryshtafovych, Venclovas et al. 2005). Two standard methods were used

in this work. The root-mean-squared-deviation (RMSD) of the backbone alpha-carbon positions,

$$RMSD = \frac{1}{N_{res}} \sqrt{\sum_{i=1}^N |\vec{x}_i^{model} - \vec{x}_i^{native}|^2}, \quad (1)$$

where $\{\vec{x}_i\}$ are the coordinates of the alpha carbon on residue i , and N_{res} is the total number of residues. The coordinates are obtained following best-fit superposition of the model onto the native. Values of 2.5 Å or less for RMSD indicate model structures that might be useful in drug and vaccine design. To account for models with fewer (or more) residues than the native, another measure, the global distance test (GDT) averages over local and global accuracy:

$$GDT = \frac{1}{4}(P_1 + P_2 + P_4 + P_8), \quad (2)$$

where P_m is the percentage of residues of the native that fit within an RMSD of m (Å) to the model. For example, a GDT score of 100 indicates an RMSD of 1 Å or less between model and native. Models with GDT scores greater than 80 would likely be useful for drug-design efforts. It is important to remember, however, that the user does not know beforehand the quality of the model structure as RMSD and GDT scores can only be obtained with the correct native structure available.

3. RESULTS

A wide variety of query sequences have been tested and the output information has been compared with known protein structures. First, the accuracy of domain boundary predictions for known two-domain proteins is assessed. Second, the quality of template-based models versus the various criteria functions is analyzed. Finally, the quality and utility of the *de novo* based models are evaluated.

3.1 Domain boundary prediction

We were able to deduce the optimal cutoff score for which to accept predictions from the domain boundary-recognition program, PPRODO, based on the evaluation of 5172 multi-domain proteins. We designate a prediction is correct if it is within 15 residues of the domain boundary assigned by the Molecular Modeling Database MMDB (Chen, Anderson et al. 2003). Overall, PPRODO made 3176 correct predictions of the domain boundary within 15 residues. Using the score as a forecaster of whether the boundary prediction was right or wrong, the true positive rate, false positive rate, and accuracy are illustrated in Fig. 2. The receiver operating characteristic (Centor 1991) of these data is illustrated in Fig. 3. The

largest difference between true and false positives occurs at a PPRODO score of 0.83. Thus, we assert this value to be the cutoff. At this cutoff score, there are 66% correct and 34% incorrect predictions.

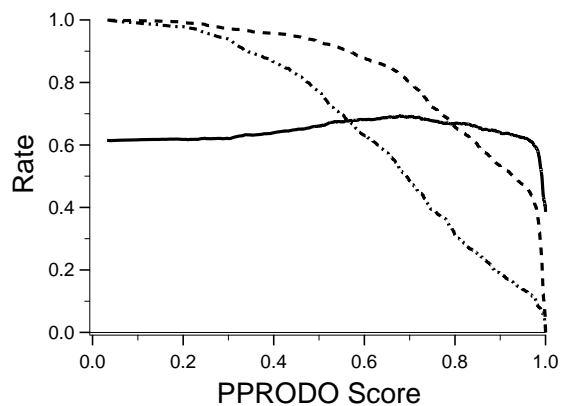


Figure 2. Statistics of the predictive capabilities of the PPRODO program. Legend: dark line – accuracy, dashed line – true positive rate, dash-dotted line – false positive rate. Accuracy is measured as the sum of the true positives and true negatives divided by the total number of predictions.

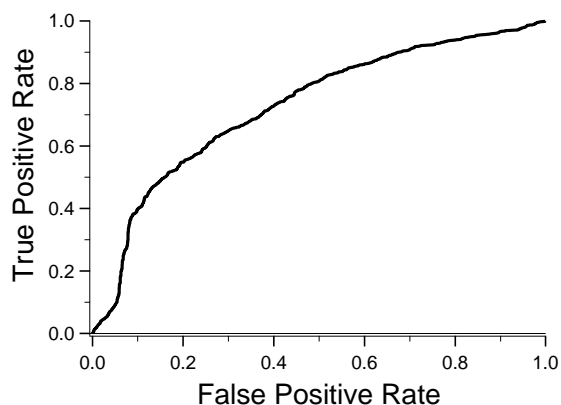


Figure 3. The receiver operating characteristic of the PPRODO results: true positive rate vs. false positive rate. A graph with an upside-down L shape would imply a perfect predictor. On the other hand, a $y=x$ line would indicate no predictive value.

3.2 Template-based models

PDB-BLAST/Nest and PROSPECT/Nest structural model results are presented in Figs. 4-8 and Table 1 for a test of 51 sequences. Sequences in this set were chosen from the SCOP 1.69 database (Andreeva, Howorth et al. 2004) and a list of PDB sequences that have low sequence similarity to every other sequence in the PDB. Figure 4

shows the best models as ranked by GDT for each query sequence. Above 50% sequence homology, the best models generated are very accurate (GDT > 80) and probably suitable for drug/vaccine design. Nonetheless, side-chain placement, which is critical for design, has not been assessed in this work. Between 20% and 50% sequence similarity, the average quality decays monotonically with a wide range of possible accuracies.

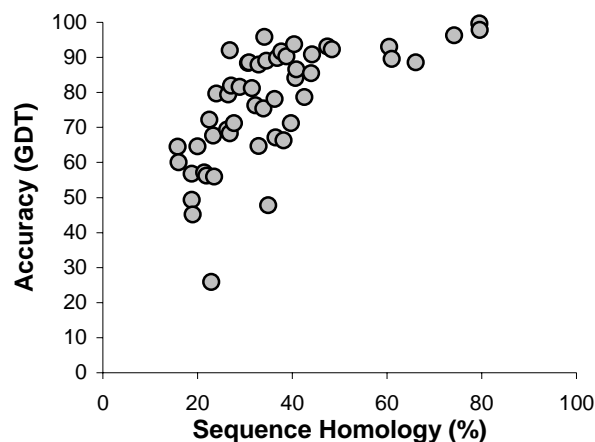


Figure 4. GDT vs. sequence homology for the most similar template-based models to the native based on GDT scores.

Because the most similar template cannot be known in advance of the actual native structure, different criteria need to be assessed. For example, the homology of the query sequence to each proposed template is a good descriptor. In Fig. 5, it can be seen that choosing the most homologous template from PDB-BLAST to form a comparative model leads to overall worse accuracy than Fig. 4. However, a clear quality vs. homology trend can still be ascertained. Many of the outliers to the average trend in Figs. 4 and 5 can be attributed to large gaps in the sequence alignment between the query and template.

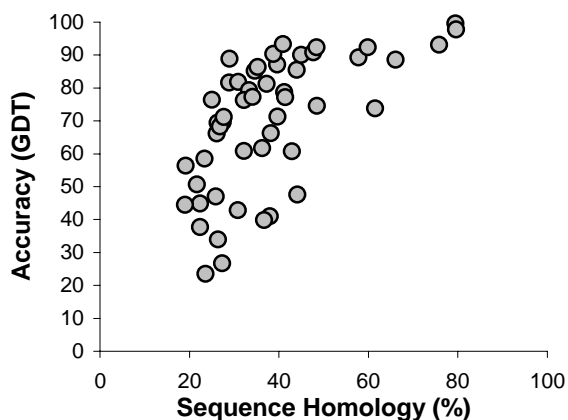


Figure 5. GDT vs. sequence homology for the most homologous PDB-BLAST/Nest structure.

Model quality is much more questionable in Fig. 6, which shows the GDT scores of the top sequence-homologous PROSPECT/Nest model. The main problem outlined in this graph is that there is a small fraction of models that are much poorer than Figs. 4 and 5 for higher sequence homologies (40 to 60%). In addition, several structures have GDT scores of 30 or lower. One of the reasons for this result is that large query sequences cannot be adequately matched with templates whose sizes are roughly 200 residues or less. If a choice must be made with no other criterion available, the PDB-BLAST/Nest models should always be considered more reliable. Despite these results, the main goal of fold recognition programs, such as PROSPECT, is to detect fold similarities with known structures even if very weak homologies exist. This feature was not adequately explored in this work.

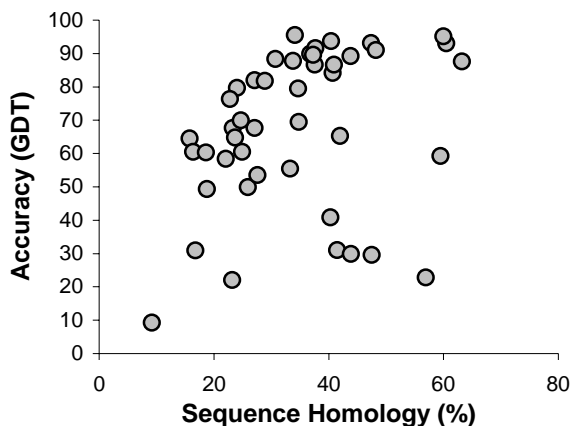


Figure 6. GDT vs. sequence homology for the top homology PROSPECT/Nest model.

Figs. 7 and 8 illustrate the detection capabilities of the GB22 and DFIRE scoring functions, respectively. There are similar trends as in Figs. 4 and 5, except for a few outliers. It is interesting to note that GB22 determined 7 PROSPECT models to be the best scoring, while DFIRE selected 25 PROSPECT models (results not shown.) In Table 1, we see that 60% of the time, the best homology PDB-BLAST model was within 5% GDT of the best overall model for each protein. Generally speaking, the detection abilities of these criteria are similar. One curious exception is that GB22 tended to perform better than the other criteria in the regime of higher sequence similarities: 18 of its successes occurred in the 26 sequences with the highest homologies to known PDB structure (results not shown). Also, it was found that for every sequence, at least one of these four criteria picked out a model within 5% of the best available model. This suggests it is worthwhile to present all of these

detection criteria to the user. In addition, development of a consensus function of criteria might be worthwhile.

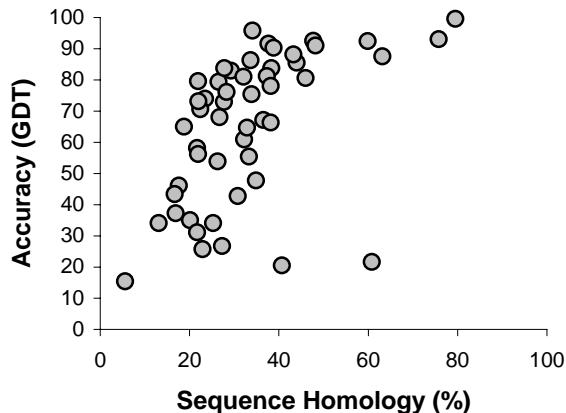


Figure 7. GDT vs. sequence homology for the model with the best GB22 score.

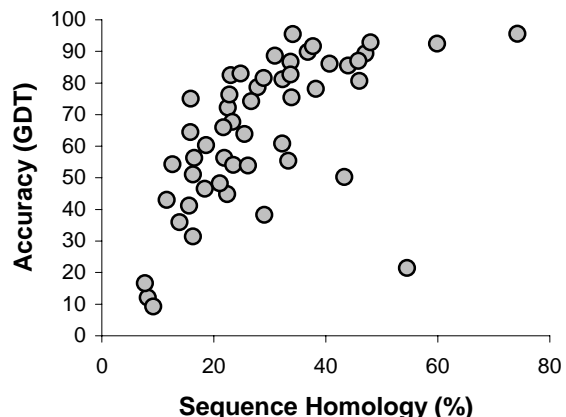


Figure 8. GDT vs. sequence homology for the model with the best DFIRE score.

Method	# within 5% GDT of the best model
PDB-BLAST ^a	31
PROSPECT ^a	24
GB22	27
DFIRE	25

Table 1. Number of models detected by different criteria within 5% GDT of the highest GDT model for each query sequence (51 total sequences). ^aThe highest sequence homology from this method is used for comparison.

3.3 *De novo* models

Lastly, we investigated the accuracy of the *de novo* structure prediction program, Rosetta. First, we found that our Rosetta protocol was most accurate in the regime of

less than ~100 amino acids (results not shown). Beyond this sequence size, the structures produced were much less reliable (Bradley, Malmstrom et al. 2005). Second, we determined that GB22 was more accurate vs. DFIRE in detecting structures closer to the native structure as seen in Table 2. Furthermore, Rosetta can often produce a few structures with good backbone RMSDs to the native (<3Å), but these may go undetected by our current scoring functions. A recent work (Bradley, Misura et al. 2005) seems to show much better performance than our results. However, there are two differences between their protocol and ours. First, the computational expense of their approach is much greater (see Table 2) due to a post-processing refinement step. Second, they used a recently developed scheme whereby query sequences are substituted by homologous sequences to generate a greater diversity of models.

We also evaluated the extent to which medium to low-resolution Rosetta structure predictions can be used to deduce the correct topological fold (Table 3). In our small assessment, the reliability is around 40% for sequences of ~60 amino-acids.

Protocol	CPU time	Top scorer (RMSD)	Top 5 scorers: best RMSD
Bradley, et al.	100-150 days	4.9 Å	4.0 Å
DFIRE	1 day	7.6 Å	6.0 Å
GB22	2 days	6.5 Å	5.0 Å

Table 2. A comparison of different post-processing protocols using Rosetta on 14 small proteins.

PDB ID ^a	RMSD to native (Å)	Rank of correct fold family
1dtja	1.3	1
1r69	2.9	1
1tig_	4.4	16
1di2A	4.9	1
1shfA	7.8	1
1af7_	7.3	12
1mlA2 ^b	8.3	4
1ogwA	8.7	12
1csp	5.7	14
1tif_	5.7	25

Table 3. An assessment of whether the top scoring Rosetta structural model for a given protein can be used to identify the correct fold family. ^a lowercase letters reflect

PDB identifier, uppercase letter or underscore indicates chain. ^b “2” refers to the second domain of this protein as defined by SCOP.

4. DISCUSSION

Our assessment of the capabilities of the protein structure-prediction suite is consistent with other reviews in the literature (Kryshtafovych, Venclovas et al. 2005). Our domain prediction component, PPRODO, is exemplary of what is currently available in the field; however, there is still room for improved algorithms. For the template-based approaches, PDB-BLAST/Nest and PROSPECT/Nest, there is correlation between the best sequence homology and the closeness of the resultant model to the native structure.

The PROSPECT/Nest protocol appears to be worse than PDB-BLAST/Nest even in the lower homology ranges. However, PROSPECT was not tested thoroughly in the regime where PDB-BLAST fails to uncover any matching templates. Other studies by the PROSPECT authors highlight this important capability (Kim, Xu et al. 2003; Guo, Ellrott et al. 2004). In addition, the sequence alignment quality in PROSPECT is at times questionable. This can be expected, as PROSPECT performs sequence-structure alignment, which is global with respect to the sequences and local with respect to the template and thus may introduce a significant fraction of gaps for large sequences. Because this is usually the case for multiple-domain proteins, better domain prediction should lead to a performance improvement.

In the newest version of PROSPECT, called OpenProspect (Ellrott, Guo et al. 2006), which is not yet publicly available, alignment algorithms have been improved and alternative scoring functions have been implemented. In addition, for low-prediction reliability, structural models can be refined with replica exchange molecular dynamics (Zhang, Kolinski et al. 2003). Because the refinement of structural models is considered more promising than improving alignments (Dunbrack 2006), future work may consider the possibility of refining some of the structural models generated in the pipeline (Misura and Baker 2005).

The use of the atomistic scoring functions, DFIRE and GB22, to detect good templates is a unique addition to our pipeline in contrast to other approaches (Ginalski, Elofsson et al. 2003). Template detection using alternative scoring functions is currently an active field of study (Wallner and Elofsson 2006).

For the *de novo* approaches, the size of the protein is the most important variable in whether accurate structures may be found. It is evident from other studies that increasing the number of Rosetta-generated models increases the likelihood of finding good structures (Tsai, Bonneau et al. 2003). Discounting the fact that we did not

- definition and its application to generalized Born calculations, *J Comput Chem*, **24**, 1348-56.
- Lee, M. S. and M. A. Olson, 2006: Calculation of absolute protein-ligand binding affinity using path and endpoint approaches, *Biophys J*, **90**, 864-77.
- Mackerell, A. D., Jr., D. Bashford, et al., 1998: All-atom empirical potential for molecular modeling and dynamics studies of proteins, *Journal of Physical Chemistry B*, **102**, 3586-3616.
- Madhusudhan, M. S., et al. (2005). Comparative Protein Structure Modeling. *The Proteomic Protocols Handbook*. J. M. Walker. Totowa, NJ., Humana Press.
- Misura, K. M. and D. Baker, 2005: Progress and challenges in high-resolution refinement of protein structure models, *Proteins*, **59**, 15-29.
- Petrey, D., Z. Xiang, et al., 2003: Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling, *Proteins*, **53 Suppl 6**, 430-5.
- Shi, J., T. L. Blundell, et al., 2001: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J Mol Biol*, **310**, 243-57.
- Sim, J., S. Y. Kim, et al., 2005: PPRODO: prediction of protein domain boundaries using neural networks, *Proteins*, **59**, 627-32.
- Simons, K. T., C. Kooperberg, et al., 1997: Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J Mol Biol*, **268**, 209-25.
- Summa, C. M., M. Levitt, et al., 2005: An atomic environment potential for use in protein structure prediction, *J Mol Biol*, **352**, 986-1001.
- Tsai, J., R. Bonneau, et al., 2003: An improved protein decoy set for testing energy functions for protein structure prediction, *Proteins*, **53**, 76-87.
- Wallner, B. and A. Elofsson, 2006: Identification of correct regions in protein models using structural, alignment, and consensus information, *Protein Sci*, **15**, 900-13.
- Zhang, Y., A. K. Arakaki, et al., 2005: TASSER: an automated method for the prediction of protein tertiary structures in CASP6, *Proteins*, **61 Suppl 7**, 91-8.
- Zhang, Y., A. Kolinski, et al., 2003: TOUCHSTONE II: a new approach to ab initio protein structure prediction, *Biophys J*, **85**, 1145-64.
- Zhou, H. and Y. Zhou, 2002: Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci*, **11**, 2714-26.