Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling

Donald Petrey, Zhexin Xiang, Christopher L. Tang, Lei Xie, Marina Gimpelev, Therese Mitros, Cinque S. Soto, Sharon Goldsmith-Fischman, Andrew Kernytsky, Avner Schlessinger, Ingrid Y.Y. Koh, Emil Alexov, and Barry Honig*

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University New York, New York

ABSTRACT **We participated in the fold recognition and homology sections of CASP5 using primarily in-house software. The central feature of our structure prediction strategy involved the ability to generate good sequence-to-structure alignments and to quickly transform them into models that could be evaluated both with energy-based methods and manually. The in-house tools we used include: a) HMAP (Hybrid Multidimensional Alignment Profile)—a profile-to-profile alignment method that is derived from sequence-enhanced multiple structure alignments in core regions, and sequence motifs in non-structurally conserved regions. b) NEST–a fast model building program that applies an "artificial evolution" algorithm to construct a model from a given template and alignment. c) GRASP2–a new structure and alignment visualization program incorporating multiple structure superposition and domain database scanning modules. These methods were combined with model evaluation based on all atom and simplified physical-chemical energy functions. All of these methods were under development during CASP5 and consequently a great deal of manual analysis was carried out at each stage of the prediction process. This interactive model building procedure has several advantages and suggests important ways in which our and other methods can be improved, examples of which are provided. Proteins 2003;53:430–435.** **© 2003 Wiley-Liss, Inc.**

Key words: homology modeling; fold recognition; protein structure alignment; profile-profile alignment

INTRODUCTION

In the past few years our group has been involved in the development of tools devoted to the analysis and prediction of protein structure and function. CASP5 provided us with an opportunity both to test our methods and to refine them so as to meet realistic challenges. The philosophy of our group with regard to software development has assumed that computational results are most valuable when used as input to a human brain where (hopefully) more sophisticated analysis can be carried out. As an example, we are more interested in building high quality homology models for proteins that are associated with problems related to our biological interests than in developing high throughput modeling procedures. The tools described in this paper were primarily designed with the expert user in mind although most of the authors of the paper were novices in structure prediction prior to the beginning of the CASP5 experiment. However, the combined availability of powerful computational algorithms with user-friendly graphical tools raised the level of expertise of each one of us as the rather intense summer proceeded.

Inevitably, improved understanding leads to better automation and our hope is that as we improve our ability to build accurate models with manual intervention, this will lead to new ideas for automation. Some of these ideas will be reflected in a discussion of selected targets that appears below. The programs described in this paper are, or will soon be, available on our web site (trantor.bioc.columbia.edu) either as server applications or as downloadable programs.

The central feature of our strategy involved the ability to generate alignments and to quickly transform them into models that could be evaluated both with energy-based methods and manually. We participated in the fold recognition and homology sections of the experiment using primarily in-house software. Our strategies for fold recognition and homology modeling were very similar. For fold recognition we tried only targets where HMAP (see below) detected templates with a reasonable statistical score, or where HMAP appeared to improve the alignments that came from the CAFASP servers. If we felt we had nothing to add beyond what the servers listed, we decided not to submit that target. Overall we submitted 53 targets and placed most of our focus on the first prediction. Multiple

Grant sponsor: National Science Foundation; Grant number: DBI-9904841; Grant sponsor: National Institutes of Health; Grant number: GM-30518.

Donald Petrey and Zhexin Xiang contributed equally to the work described in this manuscript.

*Correspondence to: Barry Honig, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032. E-mail: bh6@columbia.edu

Received 14 February 2003; Accepted 17 March 2003

predictions were primarily submitted for retrospective analysis. The large majority of our best predictions were more accurate than the others we submitted providing some validation for the methods used in making our choices.

For each fold recognition and homology target we built several 3D models and evaluated them using physical-chemical energetic criteria and with Verify-3D. Different models were generated by identifying regions of the target sequence where there was variability in the alignments to the templates, or by using a multiple structure alignment of template family members (where there was structural variability in the templates themselves). Since NEST works so rapidly we were able to try many alternate models, evaluate them with sequence and energetic criteria, borrow regions from other proteins where we believed they provided better local templates, and then link them together with loop closure procedures. In general, we did not try to keep the target as close as possible to the template. We realized that this was a risky procedure but we felt it important to test our ability to relax the structure using the refinement module of NEST. Modifications to the template were sometimes done with manual input. For example, we always checked for buried charges and unless we could visually identify a potential ion-pairing partner we would either change the alignment or try to change the structure.

MATERIALS AND METHODS

Template Selection and Alignment

Template identification and alignment was primarily based on HMAP (Hybrid Multidimensional Alignment Profile). HMAP is a profile-based sequence alignment method that we recently developed.[1] Attention was also given to possible templates found by the CAFASP servers. Any differences between HMAP and CAFASP, either in the identity of the templates or in the alignments to particular templates, were manually examined and assessed using the structure and alignment visualization tools described below.

Figure 1 shows a flow-chart describing the steps in the creation of an HMAP profile. A central feature of the procedure is a multiple structure alignment of proteins that are structural neighbors of a given template seed. These neighbors are defined as those structures with a protein structure distance (PSD) from the seed of less than one.[2] The sequence of each structural neighbor is used in turn as a seed sequence for a PSI-BLAST search. The PSI-Blast profiles are then combined using the multiple structure alignment as a guide to yield a 3D profile.[3] PSI-Blast is also used to generate a 1D profile for the template. The template sequence is partitioned into structural core and non-core regions. Hybrid 3D-1D profiles are generated using 3D profiles in core regions where it is expected that structure based alignments will be more accurate and 1D profiles in non-core regions, where it is expected that sequence alignments or sequence motif identification will be more accurate. The weight of each position and its classification as core or non-core is determined by the secondary structure composition at that position.

To search a database of HMAPs for possible templates for a given target, a 1D profile is generated for that target. Secondary structure assignments are taken from a consensus of three different secondary structure prediction algorithms: PHD,[4] JPred,[5] and PSIPRED.[6] Profile-profile alignments of the two HMAPs are carried out using methods similar to those reported previously.[7] An HMAP profile for each template in SCOP40[8] was constructed and stored in a database. Since not all proteins in the PDB were in the database, some templates were inevitably missed. When this occurred, HMAP profiles for CAFASP-identified templates were created "on the fly" and the significance of a hit between target and template, as well as the alignment, was decided on this basis.

Model Building

All models were built using the program NEST (Xiang & Honig, in preparation). NEST is an integrated model-building application that takes as input a sequence alignment of a target to one or multiple template PDB files and produces a model. NEST incorporates modified versions of the software provided in the program LOOPY[9] for loop prediction and SCAP[10] for side-chain prediction. These rapid loop and side-chain modeling methods are included in an algorithm called "artificial evolution" that is used to build the final model. In this algorithm, changes to the template structure, such as residue mutation, insertions or deletions, are made one at a time. After each change is made, a torsional energy minimizer is applied and an energy is calculated based on a simplified potential function that includes van der Waals, hydrophobic, electrostatic, torsion angle, and hydrogen-bond terms. The change to the template structure that produces the least unfavorable change in energy is accepted and the resulting structure is again subjected to a slight minimization. This process is repeated until the target sequence is completely modeled. Evaluations of this method of model building are the subject of ongoing research in our group.

The user can set a wide variety of parameters in NEST that specify the way in which models are built. Different levels of side-chain refinement can be applied during this procedure. At the minimum setting only steric clashes are removed and unrefined models can be built in times on the order of a minute on an SGI R10000 processor. This feature of NEST was frequently exploited during CASP to allow for visual examination of models built using different parameters. Subsequent levels of refinement can be applied to loops only, or to both secondary structure elements and loops.

Model Visualization, Alignment Visualization, and Multiple Structure Alignment

A new version of the program GRASP,[11] has been developed by our group. This program, GRASP2,[12] was extensively used to display and analyze alignments and models during CASP5. The graphical user interface to GRASP2 is shown in Figure 2. Many new features have been added to the original GRASP program and GRASP2

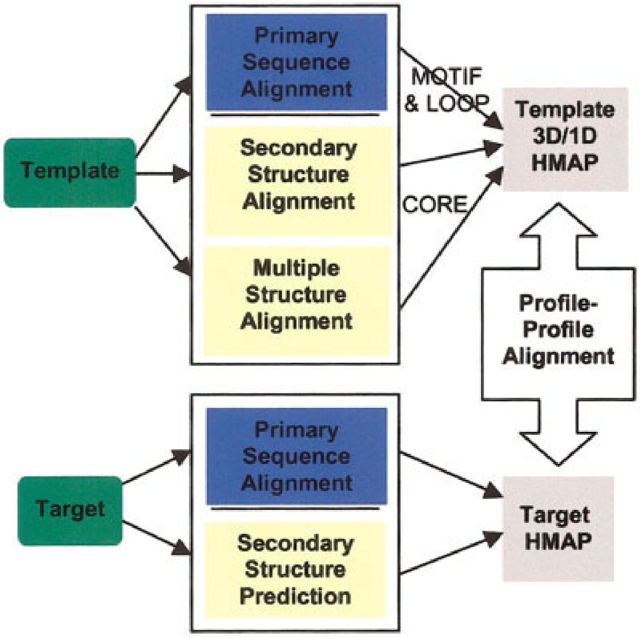
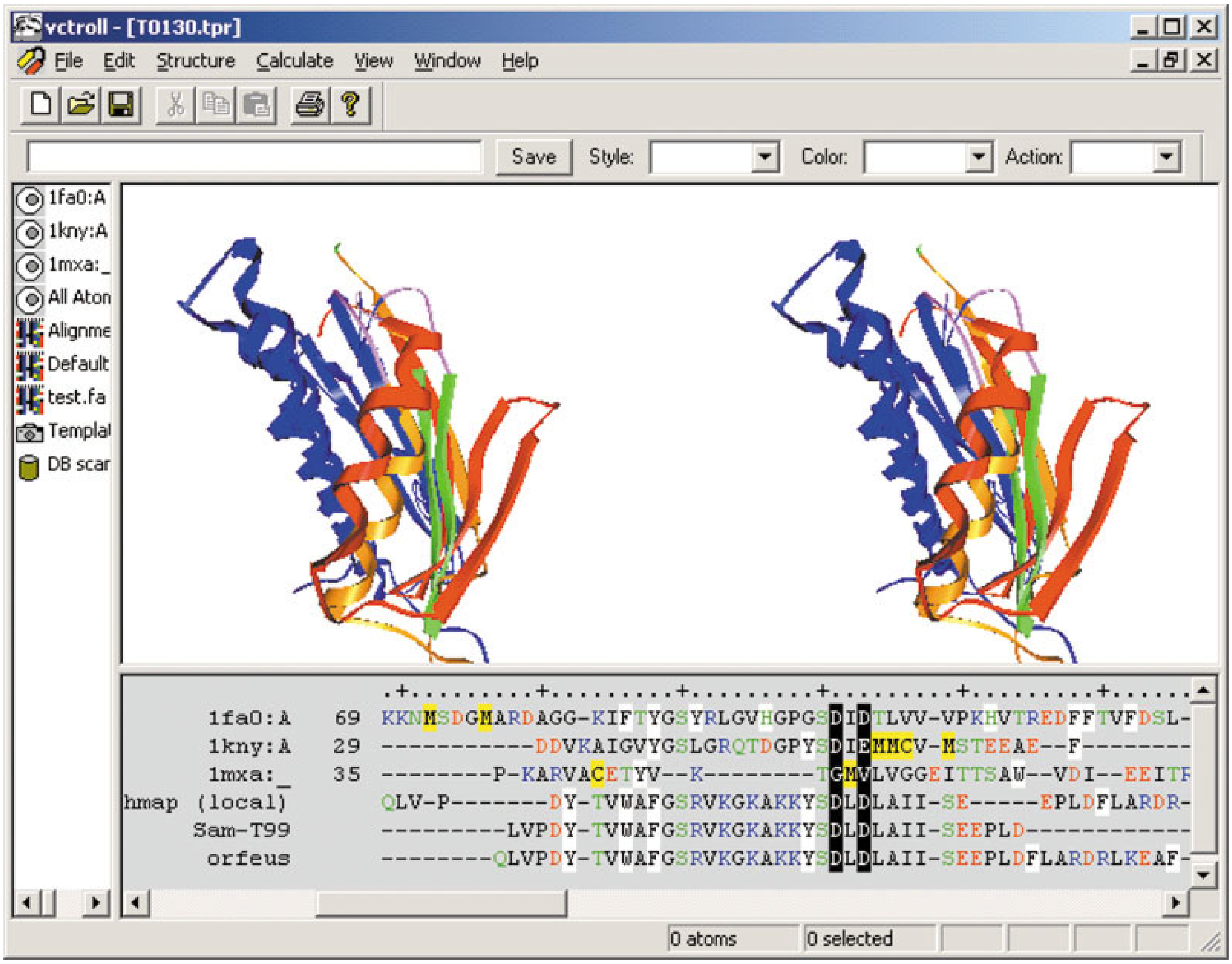

Template
Primary Sequence Alignment
Secondary Structure Alignment
Multiple Structure Alignment
MOTIF & LOOP
CORE
Template 3D/1D HMAP
Profile-Profile Alignment
Target
Primary Sequence Alignment
Secondary Structure Prediction
Target HMAP

Fig. 1.

Fig. 1. The creation of a 3D/1D HMAP for a template structure and target sequence. Pure sequence information (blue) is used in non-core regions of the profile. Multiple structure alignments and secondary structure information (yellow) are used in core regions of the profile. Core/non-core positions in the profile are determined from the percentage of residues contained in secondary structure elements in the multiple structure alignment. For a target sequence, secondary structure information comes from a consensus of three secondary structure prediction algorithms. Profile-profile alignment methods are used to generate the final sequence alignment of a target to a template.

Fig. 2. GUI for GRASP2 showing the project file for T0130. The leftmost window displays icons for each object stored in the project file. For T0130, this includes predefined subsets of the molecules, alignments to the templates and results from a scan of a domain database. The large window is the molecular graphics window, which shows a stereo view of a multiple structure alignment of the three templates that we used for this target. Red, brown and green represent the regions of structural diversity including the deleted beta-hairpin (the red strands). Blue corresponds to regions that superimpose well for all three proteins. The sequence window on the bottom left shows the structure-based sequence alignment of the three templates. Sequence alignments from three different servers to one of the templates (1fa0) have been merged into it.



```
                .+.........+.........+.........+.........+.........+.....
     1fa0:A   69 KKNMSDGMARDAGG-KIFTYGSYRLGVHGPGSDIDTLVV-VPKHVTREDFFTVFDSL-
     1kny:A   29 -----------DDVKAIGVYGSLGRQTDGPYSDIEMMCV-MSTEEAE--F--------
     1mxa:_   35 --------P-KARVACETYV--K--------TGMVLVGGEITTSAW--VDI--EEITR
hmap (local)     QLV-P-------DY-TVWAFGSRVKGKAKKYSDLDLAII-SE-----EPLDFLARDR-
     Sam-T99     ---------LVPDY-TVWAFGSRVKGKAKKYSDLDLAII-SEEPLD------------
      orfeus     --------QLVPDY-TVWAFGSRVKGKAKKYSDLDLAII-SEEPLDFLARDRLKEAF-
```

Fig. 2.

operates under Windows. New modules for the creation and visualization of multiple structure and sequence alignments and for domain database scanning are most relevant to the work we discuss here. The simultaneous display of the template structures, the template sequences and alignments of the target to the templates allowed for very convenient analysis, comparison and editing of the alignments. These features helped to define our overall strategy during CASP5: templates identified by HMAP are loaded into GRASP2; a domain database is searched and structural neighbors are also loaded into the program; a multiple structure superposition of the possible templates is then carried out; possible alignments of the target to these templates are then merged into the multiple structure alignment and this alignment with all the templates is stored in a project file; as models are generated, they are also loaded into the program and merged with the alignment. Figure 2 is a screen shot of the graphical user interface (GUI) to GRASP2 showing the project file for T0130. The GUI contains three windows: an object window, a molecular graphics window and a sequence window. A description of the contents of each window for the T0130 project file is given in the figure caption.

The search for structural homologues of possible templates was an important step in our structure prediction pipeline. Primarily, it served to identify regions of structural variability, knowledge that helped in selecting the appropriate alignment. This step also helped us ensure that all possible templates had been identified. For many of the targets we were able to use this additional template information to experiment with the construction of HMAPs using different multiple structure alignments and altering various parameters. In some cases, this resulted in a significant improvement in the alignment. SCOP was used to define domains in our database; however, rather than using a pre-defined set of families to identify structural neighbors, we carried out real time searches of our domain database for structural homologues of possible templates using the structure database scanning feature of GRASP2. A search of approximately 4,000 domains takes on the order of minutes on a 1.2 GHz Pentium processor. The algorithm used to search the database and the database itself are variations and enhancements of previously developed algorithms.[2,13]

Model Evaluation

Models were evaluated using a variety of procedures. Alignments and template structures were displayed simultaneously in GRASP2 and an attempt was made to assess the alignment quality. For example, gaps and insertions were mapped to the structures to verify that they made sense geometrically. Where residue substitutions occurred, we verified that structural features such as hydrophobic packing were maintained and verified that active site residues and any other features of the target identified from the literature were conserved. Empirical and physical-chemical energy functions were also used to evaluate the models. Verify-3D[14] was regularly used to help identify possible regions where a model might be improved. A detailed calculation of the conformational free energy of the models using a protocol developed in our group and simplified effective energy functions were also employed.[15]

RESULTS AND DISCUSSION

Analysis of Selected Targets

The modeling of T0130 is a good illustration of the successful use of our methods. T0130 was unambiguously identified as an NTP-transferase and a template was available (1fa0) with reasonable sequence identity and similar function. However, the alignment to the template was ambiguous in certain regions in that only two of three acidic residues in T0130, which were known from the literature to be important for function, were conserved in these alignments. Examination of a multiple sequence alignment of this family showed that the third conserved acid was usually found in a cluster of hydrophobic residues. Such a cluster was present in the sequence of T0130. Moreover, HMAP aligned this cluster to a similar cluster in the template, although the gaps and insertions in the alignment did not make geometric sense. Visual examination of a multiple structure alignment using GRASP2 showed that an alignment that deleted a beta-hairpin in 1fa0 would make geometric sense. Finally, a search of our domain database using GRASP2 revealed that there was a protein with a similar fold and function (1mxa, domain 1) with just such a deletion as well as an additional strand on the other side of the sheet.

NEST was then used to build multiple models using the original alignments generated by HMAP as well as new alignments, which deleted strands 3 and 4 of the template. These models were evaluated using both our all-atom conformational free energy function and Verify3D. Both methods predicted that the model with the beta-hairpin deletion was preferred. Since the template with the predicted deletion also had an additional helix and strand at the C-terminus, and since this configuration of SSEs agreed with the secondary structure predictions reported by CAFASP, they were added to the final model. All these adjustments to the model proved to be correct.

T0132 and T0142 illustrate the ambiguities associated with the decision about what constitutes a “good” model and the importance of incorporating structural information into the evaluation of the alignments. In particular, there is an ambiguity in the register of a particular strand in both targets. For T0132, the ambiguity was suggested because different alignment methods predicted a different register. The misalignment for T0142 was suggested by a low Verify3D score in that region, but neither HMAP nor any server in the CAFASP jury predicted the correct alignment. A clear reason for the requirement that the alignments be adjusted could be found based on a visual inspection of the models generated for both targets: ionizable residues would be buried in the hydrophobic core of the protein using the incorrect alignments.

For T0132, we were able to identify the correct alignment using simplified potential functions that measure the quality of hydrophobic packing and the charge distribution.[15] The model with the correct strand register had a significantly better total simplified energy than the other models. Both the electrostatic term and the hydrophobic term individually favored the correct alignment, but the difference was greater in the hydrophobic packing term, suggesting that we had properly formed the hydrophobic core.

In contrast, we were not thorough enough in our analysis of T0142. Our adjustment to the alignment in this case was to shift it by two residues in the strand based on what appeared to be a better sequence identity after the adjustment although in hindsight several possible shifts should have been tried. Following CASP5 we generated a model that contained the correct single residue shift in the alignment and, as with T0132, this model had a more favorable simplified energy. It is interesting to note that, of the top five models for targets T0132 and T0142 based on GDT_TS, only one method predicted the correct register of the strand in the case of T0132 and none predicted the correct register for T0142. These results suggest that a more detailed physical-chemical evaluation of models, such as that described above, can offer an effective means of improving alignments.

Targets 138 and 142 provide interesting examples of correct identification of problematic regions in the alignments but incorrect solutions to the problems identified. Several approximately equivalent templates, in terms of sequence identity to the target, were available for T0138. HMAP's alignments to these templates differed in the region of helix 4 where essentially two choices were possible: an alignment that placed two prolines in the first turn, or an alignment that placed them in a loop preceding the helix. A similar situation existed for T0142 where HMAP's alignment would again have placed a proline in the first turn of helix 1 of its template. In both cases, the alignment that avoided building a physically unreasonable model was chosen based on a calculation of the all-atom conformational free energy of the models. In fact, the helices in both these targets that were predicted to contain the prolines simply unraveled in the target structure and the correct solution, perhaps obvious in hindsight, was not to make any adjustment in the alignment but simply to relax the template structure. In the case of T0138, the equivalent of helix 4 in the template structure should have been modeled as an extended loop. Had we done this correctly, a high quality model would have been produced since our second choice for this target, generated using an unmodified alignment, was indeed one of the best models submitted. In T0142, the equivalent of helix 1 in the template structure should have been shortened by 1 turn.

What Went Right

Analysis of the data provided by the assessors and of the Hubbard plots provided by the organizers (predictioncenter.llnl.gov/casp5/ResultS/CASP_BROWSER) indicates that our group was uniformly successful in homology modeling and the easier fold recognition targets. Our analysis suggests that this was due to the following factors; 1) the high quality alignments from HMAP, 2) the ability offered by GRASP2 to identify problematic regions in multiple structure alignments and to focus sequence alignment efforts in these regions, 3) the ability to carry out multiple structure alignments interactively and to select regions of different proteins to be spliced so as to produce a composite model, 4) the ability to rapidly generate alternate homology models with NEST and in this way to generate models for alternate alignments, 5) the visual assessment of the energetic feasibility of alternate models with GRASP2 and 6) the quantitative evaluation of conformational free energies with the procedures described in the paper by Petrey and Honig (2000).

What Went Wrong

Our biggest failures seem to have resulted from not always knowing how to fix problems when we found them. Examples include changing alignments rather than perturbing secondary structure and making incorrect predictions for long loops, especially in cases where entire secondary structure elements were contained in an insertion. Refining structures so that they correctly move away from the template remains a major problem. For example, in addition to the targets described above, a tryptophan in our model of T0190 that was not in a loop was clearly too close to proximal loop residues. Our solution was to build a model where the side chain of the tryptophan was placed so as to avoid clashes with the loop. Examining the experimentally determined structure following CASP5, it is clear that we should have allowed the loop to relax during side-chain addition rather than assuming that the loop should be kept fixed. We also encountered difficulties in positioning terminal helices and strands correctly in the structure. One common problem is that the information that is needed to properly adjust alignments is associated with structural features that are unique to a particular target and is thus unlikely to be accounted for in general purpose algorithms. These results suggest that a significant improvement in the accuracy of prediction algorithms may result from the development of methods that incorporate family-specific information which can in principle be automated, but may at times require manual user input.

CONCLUSION

Our general conclusion based on our own performance at CASP5, and that of others, is that the main requirements for further improvements in structure prediction are refinement and sampling. More alignments need to be sampled and each alignment must be used to generate a family of models that can in turn be evaluated with some scoring function. From our perspective the good news is that in a post CASP5 evaluation, the experimentally determined structure was lower in conformational free energy than our best model in all cases but one. The bad news is that we did not sample sequence and structure space extensively enough. One approach to the problem suggested by our results is to develop methods to find local templates in the structural database. As an example, for T0130, the template that led to the correct placement of the last strand and helix in the target structure would not have been considered because of its low sequence identity. Yet it was the best choice for part of the structure.

A problem that is related to sampling is the need to improve structural refinement, especially of the backbone.

Some of the examples discussed above involved apparently reasonable alignments that produced problematic structures. Structures need to be allowed to relax with a refinement procedure before a model is accepted or rejected. Exploiting both information available in databases and our increasing understanding of the physical-chemical factors that determine protein structure and function appears to provide a path forward.

The major bottleneck appears to involve overcoming the computational complexity of sampling a sufficient number of conformations to arrive at a structure with an RMSD close enough to the native that it has a favorable enough energy to be distinguishable from less accurate models. This procedure might be simplified using great deal of information that can be obtained from an analysis of the templates, their alignments to the target, and a multiple structure alignment of their homologues. The examples discussed above describes a manual procedure for carrying out such an analysis but clearly more automated methods are necessary. Work in this area is currently underway in our group.

ACKNOWLEDGMENTS

Helpful conversations with other members of our group and with Richard Friesner and Matt Jacobsen are also acknowledged.

REFERENCES

1. Tang CL, Xie L, Koh I, Posy S, Alexov E, Honig B. On the role of structure and sequence information for remote homology detection and sequence alignment: new methods using hybrid sequence profiles. Submitted.
2. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. J Mol Bio 2000;301:665–678.
3. Al-Lazikani B, Sheinerman FB, Honig B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc Nat Acad Sci 2001;98:14796–14801.
4. Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comp. Appl. Biosci. 1994;10:53–60.
5. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinform. 1998; 14:892–3.
6. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinform. 2000;16:404–5.
7. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 1998;7:1431–1440.
8. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Bio 1995;247:536–540.
9. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Nat Acad Sci 2002;99:7432–7437.
10. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Bio 2001;311:421–430.
11. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.
12. Petrey D, Honig B. GRASP2: visualization, surface properties and electrostatics of macromolecular structures and sequences. Meth Enzymol In press.
13. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. 1996;266:617–635.
14. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
15. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 2000;9:2181–2191.