Protein Structure Modelling for Structural Genomics

a report by
Dr Marc A Marti-Renom

Adjunct Assistant Professor, Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and The California Institute for Quantitative Biomedical Research, University of California at San Francisco

Introduction

The functional characterisation of protein sequences is central to many problems in biology. This task is usually facilitated by an accurate three-dimensional (3-D) structure of the protein of interest. In the absence of an experimentally determined structure, comparative or homology modelling can provide a useful 3-D model for a protein that is related to at least one known structure. Comparative modelling predicts the 3-D structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target–template alignment, model-building and evaluation of models (see *Figure 1*). The number of protein sequences that can be modelled accurately is increasing steadily because of the growth in the number and variety of experimentally determined structures and because of improvements in the modelling software. It is currently possible to model with useful accuracy significant parts of approximately one-half of all known protein sequences.[1]

Despite progress in *ab initio* protein structure prediction,[2] comparative modelling remains the most reliable method to predict the 3-D structure of a protein with an accuracy comparable with a low-resolution, experimentally determined structure.[3] Even models with errors can be useful because some aspects of function can be predicted from coarse structural features.

Fold Assignment

Fold assignment is the first step in comparative protein structure modelling. All known protein structures related to the target sequence are identified and those that can be used as templates are selected. Templates can be found using the target sequence as a query for searching structure databases such as Class, Architecture, Topography and Homologous Superfamily (CATH) (http://www.biochem.ucl.ac.uk/bsm/cath_new), Dali (http://www2.ebi.ac.uk/dali), Protein Data Bank (PDB) (http://www.rcsb.org/pdb) and Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk/scop).

Once a list of all related protein structures has been obtained, templates that are appropriate for the given modelling problem have to be selected. Usually, a higher overall sequence identity between the target and the template sequence yields a better template.

Sequence–Structure Alignment

Having recognised the template structure(s), the next step in the modelling of the target sequence is to align it with the template structures. Since the model-building step depends crucially on the alignment, alignment inaccuracies usually lead to unrecoverable errors in the model. Sometimes, the alignment constructed during template identification is appropriate, but template identification is usually not optimised for alignment accuracy. The sequence–structure alignment therefore needs a more specialised approach, especially in cases of low sequence identity.

Another alignment strategy is to build models based on many alignments and then rank the alignments by the corresponding model assessment scores. This step can be iterated to improve the initial alignments. Although such a procedure can be time-consuming, it can improve the resulting comparative models significantly.

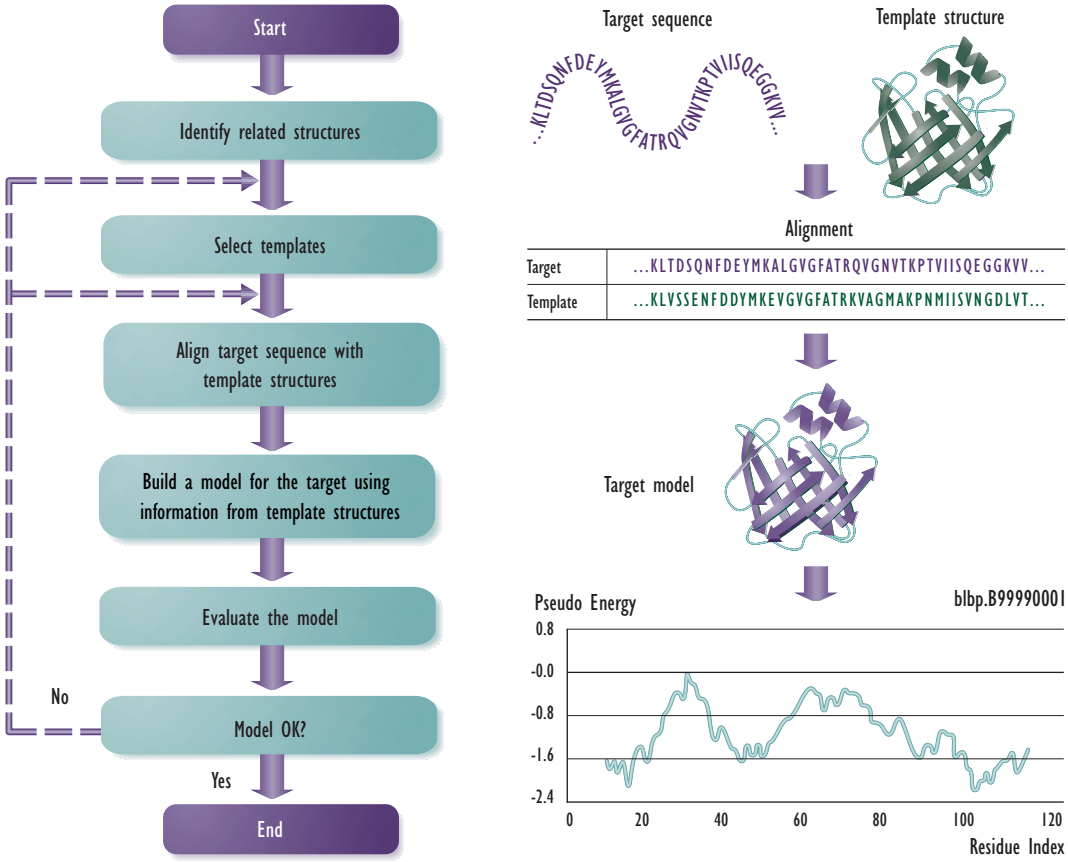
Model-building

Modelling by satisfaction of protein restraints begins by generating many constraints or restraints on the structure of the target sequence using its alignment to

Dr Marc A Marti-Renom is Adjunct Assistant Professor at the Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and the California Institute for Quantitative Biomedical Research at the University of California at San Francisco. Most of his current work involves improving the accuracy of protein 3-D models, focusing on sequence–structure alignment methods. Recently, his interest has been focused on understanding protein structure evolution and its application to comparative protein structure prediction. Dr Marti-Renom was appointed Research Associate at the Laboratory of Molecular Biophysics (Sali Lab) at The Rockefeller University in 2002. He received his BSc in Genetics from the Autonomous University of Barcelona, Catalonia, in 1994 and a PhD in Biophysics in 1999, focusing on the development of methods for folding studies using molecular dynamics.

1. *U Pieper, N Eswar, V A Ilyin, A Stuart and A Sali, "ModBase, a database of annotated comparative protein structure models",* Nucleic Acids Res., *30 (2002), pp. 255–259.*
2. *D Baker, "A surprising simplicity to protein folding",* Nature, *405 (2000), pp. 39–42.*
3. *M A Marti-Renom, A Stuart, A Fiser, R Sanchez, F Melo and A Sali, "Comparative protein structure modeling of genes and genomes",* Annu. Rev. Biophys. Biomol. Struct., *29 (2000), pp. 291–325.*

Figure 1: Steps in Comparative Protein Structure Modelling



related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from nuclear magnetic resonance (NMR)-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles and non-bonded atom–atom contacts that are obtained from a molecular mechanics force field. The model is then derived by minimising the violations of all the restraints. This optimisation can be achieved either by distance geometry or real-space optimisation.

Predicting Model Accuracy

The quality of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of 3-D protein models in the absence of the known structures is essential for interpreting them. The model can be evaluated as a whole as well as by the individual regions.

The first step in evaluating models is to determine if the model has the correct fold. A model will have the correct fold if the correct template is picked and if that template is aligned at least approximately correctly with the target sequence. The confidence in the fold of a model is generally increased by a high-sequence similarity with the closest template, an energy-based Z-score[4] or by conservation of the key functional or structural residues in the target sequence.

Once the fold of a model is accepted, a more detailed evaluation of the overall model accuracy can be obtained based on the similarity between the target and template sequences.[5] Sequence identity above 30% is a relatively good predictor of the expected accuracy because the deviation from the least-squares curve relating sequence identity to the accuracy is relatively small. The reasons for this are the well-known relationship between structure and sequence similarities of two proteins,[6] the 'geometrical' nature of modelling that forces the model to be as close to the template as possible[7] and the inability of any

4. *M J Sippl, "Recognition of errors in three-dimensional structures of proteins",* Proteins, *17 (1993), pp. 355–362.*
5. *R Sanchez and A Sali, "Large-scale protein structure modeling of the Saccharomyces cerevisiae genome",* Proc. Natl. Acad. Sci. USA, *95 (1998), pp. 13,597–13,602.*
6. *C Chothia and A M Lesk, "The relation between the divergence of sequence and structure in proteins",* EMBO Journal, *5 (1986), pp. 823–826.*
7. *A Sali and T L Blundell, "Comparative protein modelling by satisfaction of spatial restraints",* J. Mol. Biol., *234 (1993), pp. 779–815.*

current modelling procedure to recover from an incorrect alignment.[8]

Applications

Fortunately, a 3-D model does not have to be absolutely perfect to be helpful in biology. The type of question that can be addressed with a particular model does depend on its accuracy.

At the low end of the accuracy spectrum, there are models that are based on less than 25% sequence identity and have sometimes less than 50% of their C_α atoms within 3.5Å of their correct positions. However, such models still have the correct fold. Even knowing only the fold of a protein may sometimes be sufficient to predict its approximate biochemical function. Models in this low range of accuracy, combined with model evaluation, can be used for confirming or rejecting a match between remotely related proteins.

In the middle of the accuracy spectrum are the models based on approximately 35% sequence identity, corresponding to 85% of the C_α atoms modelled within 3.5Å of their correct positions. Fortunately, the active and binding sites are frequently more conserved than the rest of the fold and are therefore modelled more accurately. In general, medium-resolution models frequently allow a refinement of the functional prediction based on sequence alone because ligand binding is determined most directly by the structure of the binding site rather than its sequence. It is frequently possible to predict correctly important features of the target protein that do not occur in the template structure.

Medium-resolution models can also be used to construct site-directed mutants with altered or destroyed binding capacity, which, in turn, could test hypotheses about the sequence–structure–function relationships. Other problems that can be addressed with medium-resolution comparative models include designing proteins that have compact structures without long tails, loops and exposed hydrophobic residues for better crystallisation, or designing proteins with added disulphide bonds for extra stability.

The high end of the accuracy spectrum corresponds to models based on 50% sequence identity or more. The average accuracy of these models approaches that of low-resolution X-ray structures (3Å resolution) or medium-resolution NMR structures (10 distance restraints per residue). The alignments on which these models are based generally contain almost no errors. In addition to the applications already listed, high-quality models can be used for docking small ligands or whole proteins onto the given protein.

Structural Genomics and Comparative Modelling

The complete genomes of a number of organisms have been sequenced and many more are under way. Structural biology now faces the arduous task of characterising the shapes and dynamics of the encoded proteins to facilitate the understanding of their functions and mechanisms of action. Recent developments in the techniques of structure determination at atomic resolution, X-ray diffraction and NMR spectroscopy have enhanced the quality and speed of structural studies.[9] Nevertheless, current statistics still show that the known protein sequences (~1,000,000)[10] vastly outnumber the available protein structures (~20,000).[11] Fortunately, domains in protein sequences are gradually evolving entities that can be clustered into a relatively small number of families of domains with similar sequences and structures (i.e. folds). These evolutionary relationships enable the use of computational methods such as threading[12] and comparative protein structure modelling[13] to predict the structures of protein sequences based on their similarity to known protein structures.

Many structural genomics efforts combine the experimental structure determination methods and the computational modelling techniques to determine a sufficient number of appropriately selected

8. *R Sanchez and A Sali, "Advances in comparative protein-structure modelling",* Curr. Opin. Struct. Biol., *7 (1997), pp. 206–214.*
9. *C Zhang and S H Kim, "Overview of structural genomics: from structure to function",* Curr. Opin. Chem. Biol., *7 (2003), pp. 28–32.*
10. *B Boeckmann, A Bairoch, R Apweiler, M C Blatter, A Estreicher, E Gasteiger, M J Martin, K Michoud, C O'Donovan, I Phan, S Pilbout and M Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003",* Nucleic Acids Res., *31 (2003), pp. 365–370.*
11. *H M Berman, T Battistuz, T N Bhat, W F Bluhm, P E Bourne, K Burkhardt, Z Feng, G L Gilliland, I Iype, S Jain, P Fagan, J Marvin, D Padilla, V Ravichandran, B Schneider, N Thanki, H Weissig, J D Westbrook and C Zardecki, "The Protein Data Bank",* Acta Crystallogr. D. Biol. Crystallogr., *58 (2002), pp. 899–907.*
12. *F S Domingues, W A Koppensteiner and M J Sippl, "The role of protein structure in genomics",* FEBS Lett, *476 (2000), pp. 98–102.*
13. *T L Blundell, B L Sibanda, M J Sternberg and J M Thornton, "Knowledge-based prediction of protein structures and the design of novel molecules",* Nature, *326 (1987), pp. 347–352.*

structures so that most other sequences can be placed within a modelling distance of at least one known structure. To maximise the number of proteins that can be modelled reliably, a concerted effort towards structure determination of new folds by X-ray crystallography and NMR spectroscopy is in order, as envisioned by structural genomics.[14] It has been estimated that 90% of all globular and membrane proteins can be organised into approximately 16,000 families containing protein domains with more than 30% sequence identity to each other.[15] Of these families, 4,000 are already defined structurally; the others present suitable targets for structural genomics.

The full potential of the genome sequencing projects will only be realised once all protein functions are assigned and understood. This aim will be facilitated by integrating genomic sequence information with databases arising from functional and structural genomics. Comparative modelling will play an important bridging role in these efforts.

Future Perspectives

There has been a gradual increase in both the accuracy of comparative models and the fraction of protein sequences that can be modelled.[1,16,17] The magnitude of errors in fold assignment, alignment and the modelling of side chains and loops has decreased measurably. These improvements are a consequence of better techniques and a larger number of known protein sequences and structures. Nevertheless, the errors remain significant and demand methodological improvements. In addition, a great need exists for more accurate detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models. ■

Acknowledgments

The author is grateful to the Sali Group members for many discussions about comparative protein structure prediction. This report is partially based on reviews.[3,16]

14. *A Sali, "100,000 protein structures for the biologist",* Nat. Struct. Biol., *5 (1998), pp. 1,029–1,032.*
15. *D Vitkup, E Melamud, J Moult and C Sander, "Completeness in structural genomics",* Nat. Struct. Biol., *8 (2001), pp. 559–566.*
16. *M A Marti-Renom, B Yercovich and A Sali, "Modeling protein structure from its sequence",* Current Protocols in Bioinformatics, *in press (2003).*
17. *D Baker and A Sali, "Protein structure prediction and structural genomics",* Science, *294 (2001), pp. 93–96.*