Protein Homology Modeling

Manuel C Peitsch, *Novartis Pharma AG and Swiss Institute of Bioinformatics, Basel, Switzerland*

Torsten Schwede, *Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Basel, Switzerland*

Alexander Diemand, *Glaxo Wellcome Experimental Research and Swiss Institute of Bioinformatics, Geneva, Switzerland*

Nicolas Guex, *Glaxo Wellcome Experimental Research and Swiss Institute of Bioinformatics, Geneva, Switzerland*

Advanced article

Article contents

doi: 10.1038/npg.els.0005273

Protein homology modeling is the prediction of the three-dimensional structure of proteins by comparative methods that use the known three-dimensional structure of related proteins. This method will play a major role in the functional analysis of the genes (and their protein transcripts) discovered in fully sequenced genomes.

Introduction

Understanding the function and physiological role of proteins is a basic requirement for the discovery of novel medicines (small molecules) and 'biologicals' (protein-based products) with medical, industrial or commodity applications. Although the sequence of the human genome has been deciphered, we are very far from understanding the function and the physiological role of the gene products it encodes. Indeed, being able to read the letters and the words is not connected with understanding their meaning. Therefore, the attention of many biologists is now shifting to functional genomics which aims to discover the function and physiological role of the gene products encoded in the genome. Functional genomics is a very complex field and requires a combination of technologies. Consequently, new experimental approaches, and their automation for large-scale applications, will need development. Concurrently, and in order to maximize the value of large data sets, one will witness the development of new data mining methods and mathematical models for the simulation of biological processes. One area where this development has started is structural genomics, or the elucidation of the three-dimensional (3D) structure of proteins discovered in the genome sequences. All the major steps in experimental protein structure determination will be optimized and automated as much as possible. Protein homology modeling will allow for the rapid generation of new protein models based on this large body of newly elucidated structures, effectively increasing the impact of these research programs.

A protein's function is tightly linked to its 3D structure. As residues located far apart in the primary sequence can be very close in space, and only a few residues are generally responsible for a protein's function, insights into the 3D structure of a protein can represent a key component of the process of functional analysis. Consequently, an atomic-level 3D representation to assign roles to specific residues is a major asset, both for planning experiments and for explaining observations.

The 'folding' process of a protein is very complex, and no objective and reliable way to determine it from the sequence alone has as yet been developed. Scientists are thus dependent on experimental elucidation of protein structure. The usual approaches, both X-ray diffraction and nuclear magnetic resonance (NMR), are, however, hampered by many technical hurdles and limitations. Consequently, several concerted 'structural genomics' efforts are being launched in both the private and public sector to address these difficulties and increase the throughput of experimental elucidation of structure. However, these efforts will not be sufficient to elucidate the structure of all proteins of interest. While in early 2002 the protein sequence databases SWISS-PROT and trEMBL contain details of over 250 000 proteins, only about 10 000 of them have known 3D structures. Furthermore, SWISS-PROT and trEMBL hold fewer than 20 000 human proteins, meaning that many others will be added in the near future. The direct consequence of this is that only the 'highly interesting' proteins will be elucidated experimentally in the foreseeable future.

In this context, comparative modeling methods (homology based) have been developed and have matured to a point where many of the resulting models yield enough insights into a protein's 3D structure to be useful in functional analysis (Westhead and Thornton, 1998).

What is Comparative Protein Modeling?

Proteins from different sources and with sometimes diverse biological functions can have similar sequences, and it is generally accepted that high sequence similarity is reflected by distinct structural similarity. Indeed, the relative mean square deviation (rmsd) of the α-carbon coordinates for protein cores sharing 50% residue identity is expected to be around 1 Å. Thus the most reliable prediction methods, called comparative protein modeling (also often called modeling by homology), consist of the extrapolation of the structure for a new (or target) sequence from the known 3D structure of related family members (or templates). This process is guided by a sequence alignment between target and template sequence (for an overview of the modeling method, see Bajorath *et al.* (1993)). Membrane proteins are almost completely excluded from this approach, as there are only a very few experimentally elucidated template structures available (see the previous section). In this article we will mainly focus on the accuracy and applicability of protein models derived from these methods.

How Does One Build a Model?

Protein modeling is based on a combination of methods used in bioinformatics and computational chemistry, including searching sequence databases, sequence threading, sequence alignment and force field computations (energy minimization and molecular dynamics). Four major steps can be distinguished and these are broadly outlined below.

Identification of modeling templates

Comparative protein modeling requires at least one sequence of known 3D structure with significant similarity to the target sequence. In order to determine if a modeling request can be carried out, one generally compares the target sequence with a database of sequences derived from the Brookhaven Protein Data Bank (PDB). This can be performed using sequence database search tools (FastA, BLAST and PSI-BLAST, HMMER, PROFILE searches). Generally speaking, the choice of template structures should be restricted to those that share at least 25% residue identity with 40% of the target sequence. These limits are imposed by the current accuracy of the modeling methods in general.

The above procedure might allow the selection of several suitable templates, which can all be used but must be optimally superposed in 3D. From this set of superposed structures, a structurally corrected multiple sequence alignment can then be derived.

Aligning the target sequence with the template sequence(s)

The target sequence can then be aligned with the template sequence or, if several templates are selected, with the structurally corrected multiple sequence alignment using the best-scoring diagonals obtained by sequence alignment algorithms. Residues, which should not be used for model building, for example those located in nonconserved loops, should be ignored during the modeling process. Thus, the common core of the target protein and the loops completely defined by at least one supplied template structure will be built.

Building the model

Two very distinct classes of methods have been developed to build models. One is based on the satisfaction of spatial restraints derived from the alignment between the target sequence and its 3D templates. The other is based on an averaged framework derived from the coordinates of the templates.

Rebuilding nonconserved loops can be performed using a 'spare-parts' algorithm. Although most of the known 3D structures available share no sequence or structural similarity with the target and templates, there might be similarities in the loop regions that can be inserted into the protein model. Each loop is defined by its length and the geometry of its 'stems', namely the coordinates of the α-carbon (Cα) atom of the four residues preceding and following the loop. The fragments that correspond to the above loop definition are extracted from the PDB entries if the rmsd computed for their 'stems' is lower than a specified cutoff value. Furthermore, only fragments that do not overlap with neighboring parts of the structure are considered possible candidates. The accepted 'spare parts' are sorted according to their rmsd and their degree of sequence similarity with the target. The best-fitting fragment is then added to the model.

Because the 'spare-parts' algorithm does not always lead to convincing solutions, one can also use an approach based on a conformational space search driven by the satisfaction of stereochemical, distance and steric constraints. Loops modeled with these methods are filtered according to criteria such as the surface exposure of hydrophobic moieties and relative conformational energies.

The final step of coordinate generation is the correction and completion of the side chains, using

stereochemical criteria and libraries of allowed side-chain conformations.

Model refinement

The final step of the model building process is the idealization of the stereochemistry of the model, and consists mainly of the optimization of bond geometry and the removal of unfavorable nonbonded contacts. This step is performed using energy minimization methods as implemented in several force-field computation packages. Excessive energy minimization will cause the model to deviate markedly from the original model, which is not suitable and should be avoided.

What Defines the Accuracy of a Model?

The quality of a model is determined by two distinct criteria, which will determine its applicability. First, the correctness of a model is dictated by the quality of the sequence alignment used to guide the modeling process. If the sequence alignment is wrong in some regions, then the spatial arrangement of the residues in this portion of the model will be incorrect. The first edition of the community-wide experiment known as Critical Assessment of Protein Structure Prediction (CASP) has already underscored that most severe modeling errors can be traced back to mistakes in sequence alignment (Mosimann *et al.*, 1995). Despite many efforts to address this issue (Jones and Kleywegt, 1999), it remains the main weakness of comparative protein modeling. Second, the accuracy of a model is essentially limited by the deviation of the template structure(s) used relative to the experimental control structure. This limitation is inherent to the methods used, since models result from an extrapolation. As a consequence, the core Cα atoms of protein models which share 35–50% sequence identity with their templates will generally deviate by 1.5–1.0 Å from their experimental counterparts, as do experimentally elucidated structures. One should, however, not overlook the contributions of the templates to the accuracy of the model. The templates, which are obtained through experimental approaches, are subject to structural variations caused not only by experimental errors and differences in data collection conditions – such as the temperature – but also by different crystal lattice contacts and the presence or absence of ligands. Furthermore, X-ray crystallography and NMR generally yield 3D structures with an even broader rmsd spread. This is well illustrated by a typical example: the structure of interleukin 4 (IL-4) (Harrison *et al.*, 1995), a cytokine consisting of a 130-residue four-helix bundle, was elucidated by X-ray crystallography as well as by NMR. The backbones of three IL-4 crystal structures (PDB entries 1RCB, 2INT and 1HIK) show rms deviations of 0.4–0.9 Å, while those of three IL-4 NMR forms (PDB entries 1ITM, 1CYL and 2CYK) deviated by 1.2–2.6 Å. These values illustrate the structural differences due to experimental procedures and the molecular environment at the time of data collection. It is thus crucial to know the experimental conditions under which the modeling templates were collected, as this has a direct impact on the accuracy of the derived models and thereby on their potential use.

Almost every protein model contains nonconserved loops, which are expected to be the least reliable portions of a protein model. Indeed, nonconserved loops often deviate markedly from experimentally determined control structures. In many cases, however, these loops also correspond to the most flexible parts of the structure, as evidenced by their high crystallographic temperature factors (or multiple solutions in NMR experiments). On the other hand, the core residues – the least variable in any given protein family – are usually found in essentially the same orientation as in experimental control structures, while far larger deviations are observed for surface amino acids. This is expected since the core residues are generally well conserved and the rotamers of their side chains are constrained by neighboring residues. In contrast, the more variable surface amino acids will tend to show more deviations since there are few steric constraints imposed upon them.

Some structural aspects of a protein model can be verified using methods based on the inverse folding approach. Two of them, namely the 3D profile-based verification method and the Prosa suite of programs developed by Manfred Sippl, are widely used. The 3D profile of a protein structure is calculated by adding the probability of occurrence for each residue in its 3D context. Each of the 20 amino acids has a certain probability of being located in one of the 18 environmental classes (defined by criteria such as the amount of surface accessible to solvent, buried polar and exposed nonpolar area, and secondary structure) defined by Luthy *et al.* (1992). In contrast, ProsaII relies on empirical pseudoconformational energy potentials derived from the pairwise interactions observed in well-defined protein structures. These terms are summed over all residues in a model and result in a more (more negative) or less (more positive) favorable energy. Both methods can detect a global sequence to structure incompatibility and errors corresponding to topological differences between template and target. They also allow the detection of more localized errors such as β strands that are 'out of register' or buried charged residues. These methods

are, however, unable to detect the more subtle structural inconsistencies often localized in nonconserved loops, and cannot provide an assessment of the correctness of their geometry.

About the Use of Protein Models

Protein models obtained using comparative modeling methods can be classified into three broad categories:

- Models that are based on incorrect alignments between target and template sequences. Such alignment errors, which generally reside in the inaccurate positioning of insertions and deletions, are caused by the weaknesses of the alignment algorithms and often cannot be resolved in the absence of a control experimental structure. It is, however, often possible to correct such errors by producing several models based on alignment variants and by selecting the most 'sensible' solution. Nevertheless, it turns out that such models are often useful as the errors are not located in the area of interest, such as within a well-conserved active site.
- Models based on correct alignments are of course much better, but their accuracy can still be medium to low as the templates used during the modeling process have a medium to low sequence similarity with the target sequence. However, such models as the ones described above are very useful tools for the design of rational mutagenesis experiments. They are not of great assistance during detailed ligand-binding studies.
- The last category of models comprises all those based on templates that share a high degree of sequence identity ($> 70\%$) with the target. Such models have proved to be useful during drug design projects and have allowed key decisions to be taken in compound optimization and chemical synthesis. For instance, models of several species variants of a given enzyme can guide the design of more specific nonnatural inhibitors.

However, nothing is absolute and there are numerous occasions in which models falling into any of the above categories could either not be used at all or, in contrast, have proved to be more useful and correct than initially thought. In our experience, several applications of medium-accuracy models have proved to be successful. These can be classified into three categories detailed in the following subsections.

Interpreting the impact of mutations on protein function: a potential link to diseases

One of the first uses that one can make of a model structure is to interpret the impact a mutation can have on the overall function of a protein. Although the development of objective scoring functions has begun only recently, 'visual inspection' associated with a good knowledge of the rules underlying protein structure has proved useful in defining the broad reasons for mutant malfunction (Notarangelo *et al.*, 1996). When high-throughput production of single-nucleotide polymorphisms (SNPs) is achieved, objective scoring functions will be crucial in making maximum use of the information. Indeed, a sizeable proportion of the SNPs will alter the translated protein sequences, and thus interpreting the potential functional effects of these mutants will be crucial in elucidating the molecular basis of human diseases.

Prioritization of residues to mutate to determine protein function

As mentioned previously, the discovery of gene function in the genomic era will require a sustained experimental effort, which includes the creation of molecular mutants. The prioritization of residues to mutate will be greatly optimized by considering the 3D structure of the target protein (Schneider *et al.*, 1997).

Providing hints about protein function

This is probably the broadest and least well-defined spectrum of potential applications for 3D models. The common feature of these applications is that models can be used to formulate a hypothesis around a protein, which can then be tested in experimental settings. It is well known that low, yet significant, degrees of sequence similarity are often not sufficient to attribute a function to a protein. In such cases, protein modeling can provide useful insights and help determine or confirm a potential functional assignment (Duret *et al.*, 1998). Furthermore, one can use models to create a hypothesis about potential enzymatic activities (Peitsch and Boguski, 1991) and possible ligand-binding functions (Peitsch and Boguski, 1990).

Membrane Protein Models

Membrane proteins remain a class of proteins that represent an even greater challenge to modelers. G-protein-coupled receptors, in particular, represent a group of molecules of special interest to the pharmaceutical industry, as a very large proportion of today's medicines are modulators of their activities. Modeling such proteins has thus been attempted on many occasions, and both *de novo* (Thomas, 1996) and

comparative approaches (Thomas, 1996) have been used. The two main steps along the path to a model have been automated: algorithms have been developed to identify the transmembrane domains (Persson *et al.*, 1996) and to generate 3D models using *de novo* approaches (Herzyk and Hubbard, 1995) and comparative methods. In all cases, however, the steps of sequence analysis and coordinate generation were separated and could not be linked automatically because of the relatively low reliability of the first step. Consequently this group of proteins is not yet amenable to high-throughput model building. This will of course dramatically change with the future availability of an experimentally determined structure for one of their family members.

References

Bajorath J, Stenkamp R and Aruffo A (1993) Knowledge-based model building of proteins: concepts and examples. *Protein Science* **2**: 1798–1810.

Duret L, Guex N, Peitsch MC and Bairoch A (1998) New insulin-like protein with atypical disulphide bond pattern characterised *Caenorhaditis elegans* by comparative analysis and homology modeling. *Genome Research* **8**: 348–353.

Harrison RW, Chatterjee D and Weber IT (1995) Analysis of six protein structures predicted by comparative modeling techniques. *Proteins: Structure, Function and Genetics* **23**: 463–471.

Herzyk P and Hubbard RE (1995) Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophysical Journal* **69**: 2419–2442.

Jones TA and Kleywegt GJ (1999) CASP3 comparative modeling evaluation. *Proteins* **S3**: 30–46.

Lüthy R, Bowie JU and Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**: 83–85.

Mosimann S, Melshko R and James MNG (1995) A critical assessment of comparative modeling of tertiary structure of proteins. *Proteins* **23**: 301–317.

Notarangelo LD, Peitsch MC and Tore G Abrahamsen (1996) CD40Lbase: a database of CD40L gene mutations causing X-linked hyper-IgM syndrome. *Immunology Today* **17**: 511–516.

Peitsch MC and Boguski MS (1990) Is apolipoprotein D a mammalian bilin-binding protein? *New Biology* **2**: 197–206.

Peitsch MC and Boguski MS (1991) The first enzyme among the lipocalin family. *Trends in Biological Science* **16**: 363.

Persson B, Milpetz F and Argos P (1996) Prediction of transmembrane segments in proteins using multiple sequence alignments. In: Findlay JBC (ed.) *Membrane Protein Models*, pp. 1–25. Oxford: BIOS Scientific.

Schneider P, Bodmer JL, Holler H, *et al.* (1997) Characterization of the Fas (Apo-1, CD-95)–Fas ligand (Apo-1L, CD95L) interaction. *Journal of Biological Chemistry* **272**: 18 827–18 833.

Thomas P (1996) Making and breaking models of G protein-coupled receptors. In: Findlay JBC (ed.) *Membrane Protein Models*, pp. 73–89. Oxford: Bios Scientific.

Westhead DR and Thornton JM (1998) Protein structure prediction. *Current Opinion in Biotechnology* **9**: 383–389.

Further Reading

Chothia C and Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**: 823–826.

Herzyk P and Hubbard RE (1998) Using experimental information to produce a model of the transmembrane domain of the ion channel phospholamban. *Biophysical Journal* **74**: 1203–1214.

Martin ACR, MacArthur MW and Thornton JM (1997) Assessment of comparative modeling in CASP2. *Proteins* **S1**: 14–18.

Peitsch, MC (1997) Large scale protein modeling and model repository. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, vol. 5, pp. 234–236. Menlo Park, CA: AAAI Press.

Peitsch MC, Herzyk P, Wells TNC and Hubbard RE (1996) Automated modeling of the transmembrane region of G-protein coupled receptor by Swiss-Model. *Receptors and Channels* **4**: 161–164.

Peitsch MC and Tschopp J (1995) Comparative molecular modeling of the Fas-ligand and other members of the TNF family. *Molecular Immunology* **32**: 761–772.

Sankararamakrishnan R and Sansom MSP (1996) α-helix bundles and ion channels. In: Findlay JBC (ed.) *Membrane Protein Models*, pp. 55–72. Oxford: BIOS Scientific.

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function and Genetics* **17**: 355–362.

Tilton RF, Dewan JC and Petsko GA (1992) Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. *Biochemistry* **31**: 2469–2481.

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology* **225**: 487–494.