REVIEW

Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era

Bruno Contreras-Moreira, Paul W Fitzjohn, Paul A Bates

Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, London, United Kingdom

Abstract: The gap between the number of protein sequences and protein structures is increasing rapidly, exacerbated by the completion of numerous genome projects now flooding into public databases. To fill this gap, comparative protein modelling is widely considered the most accurate technique for predicting the three-dimensional shape of proteins. High-throughput, automatic protein modelling should considerably increase our access to protein structures other than those determined by experimental techniques such as X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy. The uses for these complete three-dimensional models are growing rapidly, ranging from guiding site-directed mutagenesis experiments to protein–protein interaction predictions. In recognition of this, a number of very useful comparative modelling servers have begun to emerge on the Web. Molecular biologists now have a powerful web-based toolkit to construct models, assess their accuracy, and use them to explain and predict experiments. There is, however, still much to do by those engaged in algorithmic development if comparative modelling is to compete on an equal footing with experimental protein structure determination techniques.

Keywords: protein structure prediction, comparative modelling, homology modelling, methods for protein modelling, web-based protein modelling

Abbreviations:

CASP	Critical Assessment of techniques for protein Structure Prediction[1] (http://predictioncenter.llnl.gov/)
EVA	continuous automatic evaluation of protein structure prediction servers (http://cubic.bioc.columbia.edu/eva/)
NMR	nuclear magnetic resonance
PDB	Protein Data Bank, currently hosted at the Research Collaboratory for Structural Bioinformatics (http://www.rcsb.org/pdb)
RMSD	root mean square deviation
SCOP	structural classification of proteins (http://scop.mrc-lmb.cam.ac.uk/scop/)

Introduction

Ever since Anfinsen's classic demonstration of the reversible denaturation of ribonuclease established that the tertiary structure of proteins in solution may be determined primarily by their amino acid sequence (Anfinsen 1973), there has been intense interest in predicting protein structure from sequence – sometimes referred to as 'cracking the protein code'. However, almost three decades on from this experiment we are unable to routinely decode protein sequences to reveal their underlying structure. Nevertheless, the field has steadily made progress and can currently be divided into three main areas of active research: *ab initio*, classically defined as the folding of the protein sequence according to physical principles; *fold recognition* (or threading), recognising that a protein sequence may represent a protein fold already classified by experimental techniques; and *comparative protein modelling* (herein referred to as comparative modelling), a method of protein modelling encompassing the fact that the structural templates found to model the protein sequence of interest (query sequence) could either be related by homology (common ancestor) or by analogy (common protein fold but not obviously evolutionary related) (Russell et al 1997).

Structural genome projects are leading biologists to a complete understanding of the cell by describing all proteins at the atomic level (Rost et al 2002). Predicted protein

Correspondence: Paul A Bates, Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London, WC2A 3PX, United Kingdom; tel +44 20 7269 3223; fax +44 20 7269 3534; email Paul.Bates@cancer.org.uk; http://www.bmm.icnet.uk/

interactions can then be tested, even simulated, and their associated cellular mechanisms understood. However, currently there are approximately 60 times more protein sequences than protein structures, hence structural coverage of any one particular genome is rather sparse (this figure was calculated from the number of nonredundant protein sequences and structures). Current comparative modelling methods can potentially alleviate this problem since they have been estimated to provide up to a twentyfold increase in structural coverage (Baker and Sali 2001; Vitkup et al 2001) over the experimental data within the PDB database (Berman et al 2000). The main reason for this is the large number of fully sequenced genomes, including the human (Venter et al 2001), incorporated into public sequence databases. This raises the accuracy of essential sequence-based tools used by comparative modelling, for example secondary structure prediction (Przybylski and Rost 2002). On the other hand, the contribution that predicted structure itself makes to the understanding of protein function is being debated, with many experts suggesting caution when transferring functional features even between homologous proteins (Devos and Valencia 2000; Thornton et al 2000; Irving et al 2001; Rost 2002).

Early genome projects, apart from sequence-based protein function annotation, have permitted large-scale structural modelling projects (Sanchez and Sali 1998, 1999). Such efforts provide molecular biologists with instantly accessible models for a proportion of proteins within each sequenced genome. In addition, recent novel methodologies (Aloy and Russell 2002) will permit the discovery of genomic protein–protein interaction networks, although molecular models for such interactions are not currently available.

Other problems which are tentatively being tackled include docking protein models (Tovchigrechko et al 2002), mapping protein motions (Hayward 1999; Karplus and McCammon 2002), using models to help understand the interplay of complex metabolic networks (Alves et al 2002) and probing the specificities of the immune system (Oliva et al 1998). In addition, the screening of large protein model databases with even larger small molecule databases should one day prove useful, not just in terms of designing drugs to modulate protein function (Peitsch 2002), but also in calculating the potential side effects of those drugs, ie unintended modulation of protein function (Rockey and Elcock 2002). There is, therefore, a pressing need for highly accurate, high-throughput and automatic comparative modelling software.

However, as the results from four CASP experiments[1] have shown, little progress seems to have been made in algorithmic developments that have directly improved the overall accuracy of the comparative modelling approach (Tramontano et al 2001). Nevertheless, essentially due to increases in various protein database sizes, particularly protein sequences, many useful models can now be predicted even at very low sequence similarity between the query and best template sequences. The possible reasons that comparative modelling is not able to obtain a consistently high level of accuracy will be outlined, but first the current comparative modelling protocols and underlying algorithms must be described.

Comparative modelling protocols

Figure 1 outlines the key generic model building steps used by most developers in the field. These steps shown are common to the two main modelling protocols; satisfaction of spatial restraints (Sali and Blundell 1993) and building up a protein by inheriting segments of other proteins (Greer 1981; Jones and Thirup 1986). However, some of the steps may be executed concurrently or in a different order.

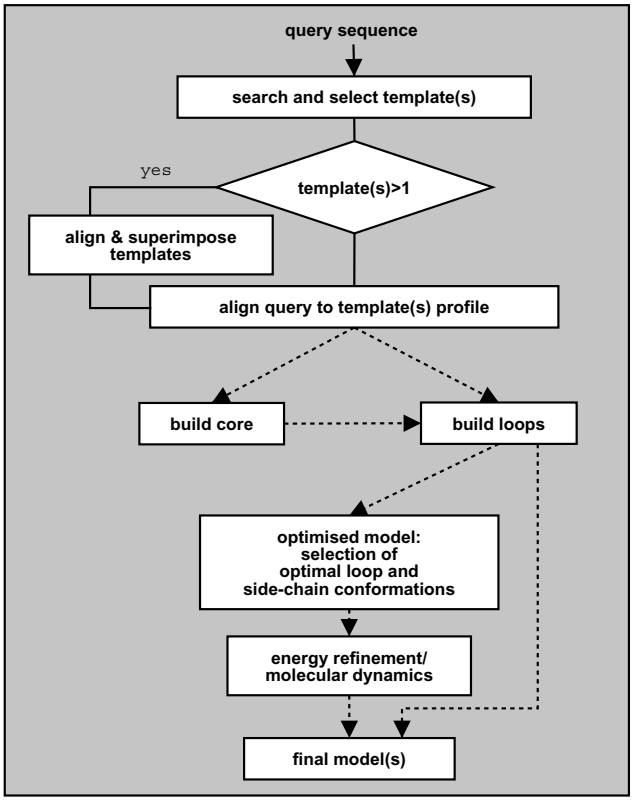

Figure 1 Generic steps in comparative modelling protocols. Dotted lines indicate optional or parallel steps.

Finding the best templates

Templates can be found by sequence similarity alone or by using additional sources of structural information, such as secondary structure. The former approach is used by the BLAST (Altschul et al 1997) and FASTA (Pearson and Lipman 1988) families of programs, where a query sequence is scanned against a database of template sequences using broad-spectrum matrices, such as BLOSUM (Henikoff and Henikoff 1993) or PAM (Schwartz and Dayhoff 1978), to score the alignments. Increased sensitivity can be gained by using the information of protein families (represented as position-specific scoring matrices or hidden Markov models) as family-specific matrices and by using intermediate sequence searching procedures (Baldi et al 1994; Krogh et al 1994; Eddy 1996; Park et al 1998; Schaffer et al 2001). Still further sensitivity can sometimes be gained by including structural information such as residue solvent accessibility and secondary structure (Rost 1995; Kelley et al 2000; Shi et al 2001), or by combining different alignment strategies (Elofsson 2002). However, as low sequence similarity templates generally yield low accuracy models (Vitkup et al 2001), some comparative modelling programs, for example SWISS-MODEL (Guex et al 1999), use less ambitious and simpler methods to assure the quality of their results at the risk of missing some modelling targets (see Table 1).

Table 1 Freely available comparative modelling Web servers and programs[a]

Server/program name and URL	*Modelling method*	*References*
3D-JIGSAW http://www.bmm.icnet.uk/servers/3djigsaw	Looks for homologous templates and splits the query sequence into domains. If good templates are found the best-covered domains are modelled, currently using a maximum of two templates. Different loops are tried to connect secondary structure elements taken from the templates. The best model within the ensemble is then selected and refined.	(Bates et al 2001; Contreras-Moreira and Bates 2002)
CPHmodels http://www.cbs.dtu.dk/services/CPHmodels	A neural network based method to predict C-α contacts to drive the sequence alignment. No side chains are constructed.	(Lund et al 1997)
ESyPred3D http://www.fundp.ac.be/urbm/bioinfo/esypred	Exploits a new alignment strategy using neural networks. Complete models built with MODELLER.	(Lambert et al 2002)
Nest[b] http://trantor.bioc.columbia.edu/~xiang/jackal/#nest	Allows building of models with one or several templates tuning their alignments and permitting artificial evolution.	
MODELLER[b] http://guitar.rockefeller.edu/modeller	Builds a complete model based on alignments prepared by the user. The procedure is based on satisfying spatial restraints (automatically computed from the templates used). Models are refined using a variety of algorithms.	(Sali and Blundell 1993; Fiser et al 2000)
Modzinger Z http://peyo.ulb.ac.be/mz/	Templates are aligned to the query sequence to build a library of backbone fragments. Fragments are then combined to build alternate models and scored. Finally side chains are added.	
Pcomb http://www.sbc.su.se/~arne/pcomb	Pcomb uses a combination of several sequence-profile and profile-sequence searches. Final models are produced using MODELLER.	
Protinfo http://protinfo.compbio.washington.edu	A core model is built for each template found by sequence similarity to the query. Loops and side chains are then added to the best scoring models.	
SDSC1 http://cl.sdsc.edu/hm.html	Templates are found using intermediate sequences primarily found by BLAST. Phylogenetic trees are used to weight pairwise alignments. Only backbone coordinates are returned.	
SWISS-MODEL http://www.expasy.org/swissmod	Templates found by BLAST are superimposed and then aligned to the query sequence excluding loop regions. The core is then calculated as a weighted average of the templates. Loops are then added and the final model is refined.	(Guex et al 1999)
TSUNAMI http://www.pirx.com/tsunami	Fragments of templates found by a BLAST-like algorithm are assembled and the final model is evaluated using statistical potentials.	

[a] These programs return atomic coordinates to the user. Most fold-recognition servers return only alignments and therefore are not listed here.
[b] Downloadable software.
NOTE: All websites accessed 29 January 2003.

Most of the above methods for identifying suitable templates perform local alignments by finding maximum scoring sequence patches, which do not necessarily correspond to complete protein domains. For this reason, databases of protein structural domains, for example SCOP (Murzin et al 1995) or CATH (Orengo et al 1997), have been used to define templates (Kelley et al 2000; Contreras-Moreira and Bates 2002). For the same reason, multi-domain proteins remain a problem for comparative modelling programs, and despite preliminary efforts (Contreras-Moreira and Bates 2002) most servers rely on the user's knowledge of how to split their query sequence into domains before submission.

Aligning the templates and query

Once the complete set of possible template(s) has been found, it is necessary to select a subset from which to build the actual model. Modellers have long preferred to use several templates where available (Sali and Blundell 1993; Guex et al 1999; Bates et al 2001; Venclovas 2001), but the practical advantage of this approach has not yet been proven (Tramontano et al 2001). Indeed, most methods would perform better if the single ideal template could be recognised, but unfortunately pairwise sequence identity is not a consistent criterion by which to address this question (Wood and Pearson 1999; Koehl and Levitt 2002). If several templates are to be used they have to be optimally aligned to drive the process of model building. ClustalX (Thompson et al 1994), T-Coffee (Notredame et al 2000) and similar programs can be used for this, despite the fact they can only produce approximations to optimal solutions for more than two sequences. But because sequence similarity between templates can be very low, it may be necessary to use their structural similarity to align them. In this case, programs such as SSAP (Taylor and Orengo 1989), STAMP (Russell and Barton 1992) or CE (Shindyalov and Bourne 1998) may be used.

Finally, the query sequence needs to be accurately aligned to the template(s); again sequence and structural information is often used. Typically the alignment procedure must exclude gaps in secondary structure elements and anchor the alignment in non-loop regions. In addition, key functional motifs should also be correctly aligned, for example P-loops (Walker et al 1982), EF-calcium-binding loops (Kawasaki and Kretsinger 1995) and catalytic triads. Databases of such motifs have been constructed, including PRINTS (Attwood et al 1998) and BLOCKS (Henikoff et al 1999); however, we are unaware of any automatic modelling procedure that takes advantage of these extremely useful sources of information.

Modelling by satisfaction of spatial restraints

This family of approaches was first proposed in the mid-eighties (Braun and Go 1985; Havel and Snow 1991; Sali and Blundell 1993) and consists of computing geometrical and biochemical restraints from the set of superimposed templates that the aligned query sequence will have to optimally satisfy. This method considers the possible templates as a sample of the folding space for a group of homologous proteins. Since the query sequence is believed to be another homologous member of the group, it will have to fulfil the restraints dictated by its relatives. As a consequence, models built using this method are derived from every template used and do not directly inherit backbone segments from any one template. Optimisation of possible conformations according to the restraints can be done in a variety of ways, including conjugate gradient minimisation (Sali and Blundell 1993), simulated annealing (Ogata and Umeyama 2000) and graph theory (Samudrala and Moult 1998). The weakness of the method is that templates need to be reasonably superimposable to define the restraints and that some regions are poorly restrained. Its strength however, is that it can directly model an entire protein structure as a continuous chain. Methods which essentially apply distance constraints to reconstruct the protein backbone, such as neural networks (Lund et al 1997), also fall into this category.

Modelling by fragment building approaches

This has historically been the most popular approach for comparative modelling and is based on grafting protein fragments from the template(s) to build up the query structure (Greer 1981; Jones and Thirup 1986; Blundell et al 1987; Sutcliffe et al 1987; Bates et al 2001). This method has clear limitations in modelling sections which differ widely between templates, such as loops, because matching of the selected fragments is non-trivial and often requires additional modelling steps (see below). However, the benefit of the approach is that sections confidently inherited from the templates (good agreement between templates) have intrinsically good geometry and require minimum subsequent optimisation. A related but novel methodology has recently been applied to ab initio protein structure prediction. This uses small protein fragments extracted from

templates that are not necessarily homologous (Unger et al 1989; Simons et al 1997; Kolodny et al 2002), allowing models to be built where no significant sequence similarity is found to any template.

Optimisation

Once the basic model has been constructed, most protocols then investigate loop and side-chain optimisation. In the context of a protein, a loop can be defined as a region of variable length and irregular shape connecting secondary structure elements (Branden and Tooze 1999).

If there is a high sequence similarity with the template then these homologous loops may be modelled in a similar way to the rest of the protein (Greer 1981). The methods for constructing loops for less conserved regions fall into two main categories: database searches and ab initio methods.

Database searches are based on grouping observed loops in the PDB and building a library. This method relies on the assumption that the set of structures used is large enough to produce a database that covers all possible geometrical configurations that protein loops can adopt. However, as segments of up to nine residues with the same sequence can have completely unrelated conformations in different proteins (Sander and Schneider 1991; Mezei 1998), sequence alone cannot be used as a method of defining useful groups. Early classification systems relied on manual investigation of loops within specific environments, such as β-turns (Ventkatachalam 1968), γ-turns (Rose et al 1985; Milner-White 1987) and α-α, α-β, β-α and α-α arches (Edwards et al 1987; Rice et al 1990; Colloch 1991; Efimov 1991). More recently, automatic classification systems have been used, which classify the loops according to the local environment and intra group RMSD (Kwasigroch et al 1996; Wintjens et al 1996). More specific and tighter clusters have also been generated by specifically taking into account bracing geometry, Ramachandran patterns and sequence (Oliva et al 1997).

The ab initio loop prediction methods are based on a conformational search of the space to be filled. There are many methods that use different search algorithms and different energy functions. Some of the search algorithms used include the minimum perturbation random tweak method (Fine et al 1986; Shenkin et al 1987; Smith and Honig 1994), systematic conformational searches (Bruccoleri and Karplus 1987; Bruccoleri et al 1988), molecular dynamics simulations (Bruccoleri and Karplus 1990; Rao and Teeter 1993; Nakajima et al 2000), energy minimisations (Lambert and Scheraga 1989; Dudek and Scerage 1990; Dudek et al 1998; Fiser et al 2000), genetic algorithms (McGarrah and Judson 1993), Monte Carlo techniques (Collura et al 1993; Evans et al 1995; Carlacci and Englander 1996; Thanki et al 1997), scaling relaxation (Zheng et al 1993; Rosenbach and Rosenfeld 1995; Zheng and Kyle 1996) and dynamic programming (Vajda and DeLisi 1990).

The jury remains out as to whether database or ab initio methods are the more accurate for small to medium size loop construction. For example, in 1994 a study assessing the effectiveness of database methods concluded that they were only sufficient for loops of up to 4 residues (Fidelis et al 1994). However, later work showed that with some optimisation of the loops, the limit for database searches could be raised to 9 residues (van Vlijmen and Karplus 1997). For a loop of this size, ab initio methods need to generate substantial numbers of loop configurations to fully sample conformational space. What is clear is that in both the database and ab initio methods a scoring function is required to select the correct loop from the ensemble searched. Many scoring functions have been tried and the effectiveness of these dictates the final accuracy that can be attained. Scoring functions remain a problem and may require a deeper consideration of complete free energy summations that include appropriately weighted terms, for example loop entropy (Xiang et al 2002) and desolvation (Janardhan and Vajda 1998).

Usually the second phase in optimising a model is the addition and refinement of the side chains. Side-chain prediction algorithms almost exclusively use a database of rotamers, as this significantly reduces the complexity of refining all the side chains in a protein at the same time. Some early work (Lee and Subbiah 1991) was reasonably successful at predicting the core side chains using simulated annealing. A significant reduction in the number of combinations of rotamers to search was made possible by the dead-end elimination method (Desmet et al 1992; Lasters and Desmet 1993; De Maeyer et al 2000), which allows the early elimination of impossible combinations. Early work noted that there was a significant tendency for side chains to prefer certain rotameric states depending on secondary structure (McGregor et al 1987). Similar investigations led to the production of backbone dependent rotamer libraries (Dunbrack and Karplus 1993; Bower et al 1997). Methods for searching the possible combinations were also being developed, one of the most widely used being the self-consistent mean-field approach (Koehl and Delarue 1994).

Many of these approaches are often tested on known crystal structures with the side chains removed. Whilst this is fine for checking the accuracy of the methods, it does not check the accuracy when used for predicting side chain conformations for a comparative model that has backbone errors inherited from the modelling process. Desjarlais and Handel (1999) developed a method that allowed flexibility in the backbone at the same time as the selection of the side chains. This showed that even in core regions, significant changes to the backbone inherited from homologous proteins can occur to accommodate the new side chains, and current methods that do not include backbone flexibility would be severely impeded in choosing the correct rotamers. It was also assumed that core regions were exclusively dictated by van der Waals packing. However, this has been shown to be insufficient on its own to define these regions (Kussell et al 2001).

Recent work (Xiang and Honig 2001) has concluded that there is no combinatorial problem in the choice of the correct side chain on a correct backbone, but that as long as a highly detailed rotamer library is used the limiting factor becomes the scoring function. A detailed study (Jacobson et al 2002) into surface side chains has shown that the crystal environment has significant effect on the final conformation adopted. In addition, limits for the maximum accuracy were also calculated which showed that while it should be possible to predict core regions to 90% accuracy compared with the X-ray structure, many surface side chains adopted many different conformations dependent on their environment. Therefore, predicting single rotamer states for exposed side chains is not justified. Given these constraints, many modern methods do manage to achieve a reasonable level of accuracy and even reach the limit in the core regions (Mendes et al 1999; Petrella and Karplus 2001; Liang and Grishin 2002).

Energy refinement and molecular dynamics

As a final step, some form of energy refinement is usually performed on the models. This can be achieved by using one of the energy minimisation software packages such as CHARMM (Brooks et al 1983). Such refinements usually have a small radius of convergence and are used simply to remove steric clashes, particularly between side chains, and ensure sensible covalent geometry is maintained around each atom. Often this achieves little more than improving the appearance of the model (Schonbrun et al 2002). Indeed, there has been little work done to show if energy refinement does in general slightly refine models in the correct direction. A technique that enables a larger radius of convergence, compared to standard energy minimisation, is molecular dynamics. However, in a recent study on a small number of protein models using state-of-the-art explicit solvent molecular dynamics and implicit solvent for energy calculations, only limited success was achieved in refining some of the models closer to the native state (Lee et al 2001).

Error analysis

What are the most common errors in comparative models? Following previous papers (Marti-Renom et al 2000; Bates et al 2001; Tramontano et al 2001) and according to our experience, three major sources of errors in comparative models can be identified: template selection, sequence alignment and loop/side-chain building.

Selecting templates becomes especially difficult when their sequence similarity to the query is low (less than 25%–30% of sequence identity). In these circumstances even statistically significant sequence matches, for example found by BLAST, can identify totally different folds.

As explained in detail above, there are many different sequence alignment methods but so far none can be considered optimal. However, whilst sequence identity is not a consistent measure of expected alignment accuracy (Tramontano et al 2001), alignments with over 40% of sequence identity between query and template can be considered confident (Marti-Renom et al 2000). Below this threshold, alignments tend to accumulate errors. Unfortunately these errors are inherited by the rest of the modelling process and current protocols are not able to detect them. A possible solution to this has been investigated by building models from several alternative alignments and then choosing the best, based upon energetic or statistical potentials (Melo et al 2002). Finally, whilst no method is perfect, it has been shown that by using several protocols the optimal alignment may be obtained. The problem is then reduced to being able to routinely identify this alignment (Elofsson 2002).

Even in confident regions of sequence similarity, quite different backbone conformations can be present in a comparative model compared to the native or target structure. These can confuse rational experimental design and occur essentially because proteins are flexible (see Figure 2a); proteins can exhibit different conformations depending on their environment (Branden and Tooze 1999; Liu et al 2002). A clear example of this problem is seen in globular proteins that build the 30S ribosome. Many of them have been solved independently and as part of the ribosome,

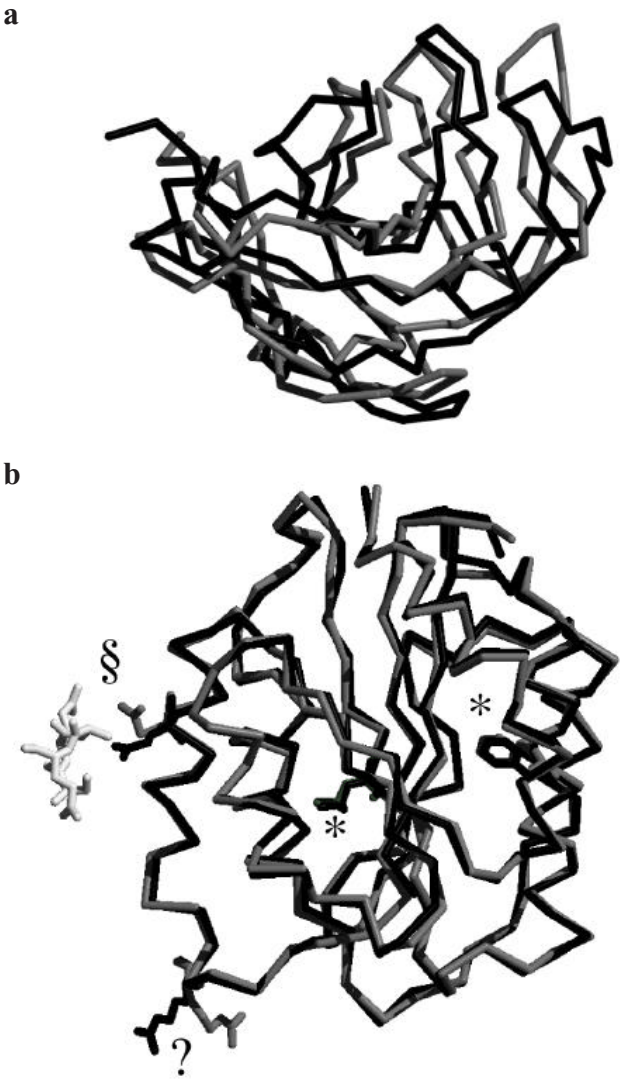

Figure 2 (a) An example from the authors' automatic server (3D-JIGSAW) showing a model (black), based on a NMR template, optimally superimposed onto the high resolution structure of the same protein eventually solved by X-ray crystallography (grey). The NMR (template) and X-ray structures have identical sequences. Interestingly, there are many conformational differences throughout the fold (not just loop regions) giving a final RMSD of 2.5 Å. **(b)** The backbone of a model (black) showing minor deviations from the experimental X-ray structure (grey) modelled (3D-JIGSAW) from a 95% identical template. Predicted core side chains (*) agree well with the observed. However, exposed side chains can show significant differences in their rotameric states due to crystal contacts (§), indicated here by the white side chains, or simply because they are exposed to solvent (?), indicating that they probably have multiple rotameric states.
NOTE: These proteins can be downloaded or interactively viewed at http://www.bmm.icnet.uk/supplementary/review2003

and they show important differences in exposed loops and N- and C-termini that seem to be important for function (Brodersen et al 2002). If these structures are used as templates they will yield different models for the same protein.

If we are sure that the above alignment problems do not affect the model under construction, we can then consider loop building errors as the next major problem. Loops can be confidently modelled if they are only up to 5 or 6 residues long (Martin et al 1997). In fact, as mentioned previously, loops of this size tend to form conformational clusters (Oliva et al 1997; Branden and Tooze 1999). Longer flexible fragments are usually not accurately modelled and indeed some modelling protocols simply do not attempt to model these regions (Venclovas 2001). However, since loops are frequently important for protein function (Oliva et al 1997) and are sometimes difficult to 'see', even for X-ray or NMR structure determination experiments, we must look further for solutions to this essentially mini protein folding problem. One possible solution to this could be to consider an ensemble of low energy loop conformations within a broad free energy minimum (Xiang et al 2002).

The next level of uncertainty in models is at the side chain level. As discussed earlier, provided the modelled backbone quality is high, side chains are usually well placed in the protein core but are subject to variations at the surface, as shown in Figure 2b. The uncertainty in surface side-chain rotamers can sometimes be resolved when considering protein–protein interactions, as these reduce their degree of flexibility.

Finally, a common problem in comparative modelling is calculating exact relative domain orientations in multi-domain proteins. Surprisingly, given the large RMSD errors involved, this appears to be a subject for which a comprehensive study has not yet been performed. Molecular dynamics and protein docking techniques may aid the solution to this domain-packing problem.

Quality control

What kind of RMSDs are we likely to expect between the model and the experimentally determined structure? Chothia and Lesk (1986) studied the sequence and structural variability within protein families and observed that as the sequence similarity between proteins decreased, the RMSDs between their superimposed structures increased exponentially. Based on the results from CASP experiments, similar studies have been conducted on protein model quality relative to closest template (Vitkup et al 2001). Figure 3 shows the latest results from the EVA experiment (discussed below) (Eyrich et al 2001) plus the authors' own in-house benchmark of model accuracy. In general, regardless of the servers used, for protein sequences around 95% identical the backbone RMSD is expected to be under 1 Å; when the sequence identity drops to 30%, the expected RMSD is around 4 Å. As can be seen in Figure 3, there is

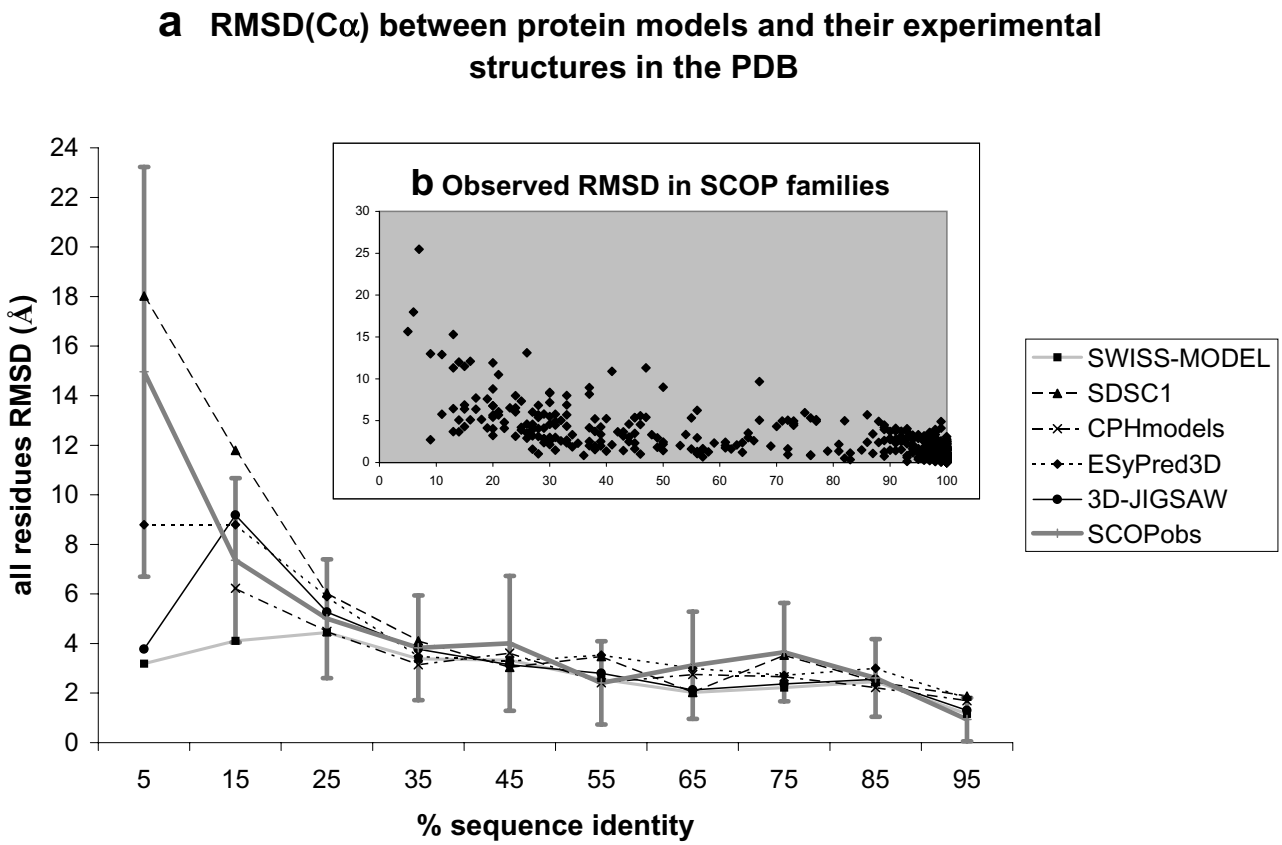

Figure 3 **(a)** Comparison of observed accuracy for models returned to the assessors for the EVA experiment. Backbone RMSDs are reported versus % sequence identity to the closest template. Results from five servers are plotted, indicated by the first five labels in the figure key (see Table 1 for more details), plus a benchmark plot from pairs of SCOP family members (SCOPobs). The error bars show the extent of variation expected for each sequence identity subgroup (binned every 10%). **(b)** Individual observations in the plot of pairwise SCOP families used in the calculation of error bars for (a).

an increasing range of variability around these error estimates towards lower sequence identities.

There is a formal quality control procedure to test and evaluate new prediction techniques every two years – the CASP experiments. Because the number of protein structures predicted in each CASP experiment has been small, the statistical significance of ranking the prediction methods has been brought into question (Marti-Renom et al 2002). However, the value of human expert analysis should not be underestimated, as developers gain additional insights into further developing their algorithms beyond that given by pure numerical analysis. For example, advantageous ways to mix current algorithms may be suggested.

To address the statistical weakness of CASP and to help modellers test their algorithms on a more frequent basis, two continuous assessment projects have recently started: EVA (Eyrich et al 2001) and LiveBench (focused more on fold recognition programs, http://bioinfo.pl/LiveBench/) (Bujnicki et al 2001). In these experiments, sequences of proteins about to be released in the PDB database (determined experimentally) are automatically sent to participant servers around the world, which in turn send back automatically built protein models. The benefit of such on-line experiments is that the evaluation of model quality is also fully automatic, enabling the results for each server in the experiment to be posted on the Web very quickly and at regular intervals; EVA results for example are tabulated weekly. This enables molecular biologists to determine which server(s) are currently likely to produce the more accurate models and helps developers rapidly benchmark and rank their new modelling algorithms against others in the field. The handicap of these methods is that although an extensive numerical analysis is performed, there is no human overview of the interplay between these results and the variety of complex methods used to obtain them.

Apart from the grosser limitations to the use of protein models dictated by sequence similarity to the templates, the user can check the stereochemical and thermodynamical quality of models by using programs such as PROCHECK (Laskowski et al 1993) and WHATCHECK (Hooft et al 1996). However, until a rigorous ranking scheme for model

accuracy can be found, the final indication of the correctness of a model protein will always lie in the hands of the experimentalist.

Applications

As a consequence of the above quality controls, it is possible to enumerate the applications for which protein models are likely to be useful according to the sequence identity between query and template (Marti-Renom et al 2000; Baker and Sali 2001). Traditionally, molecular biologists have used protein models to design site-directed mutagenesis experiments and to understand mutant phenotypes in the light of protein structure. Even very low sequence identity templates yield useful models, some of which have given insights into potential protein functions (see for example Garmendia et al (2001) and Devos et al (2002)). Apart from functional study applications, low resolution models are also being used to build supramolecular structures (Zhang et al 2000; Wriggers and Chacon 2001; Aloy et al 2002; Elcock 2002). Mid-resolution models, derived from templates around 50%–60% identity level, can be valuable as models for use in molecular replacement (X-ray crystallography) and the rational design of more stable proteins, for example the addition of a disulphide bond (Mansfeld et al 1997). Finally, high resolution models, those typically obtained from templates over 90% identical in sequence, are being routinely used as receptors to dock and rank small molecules for potential pharmaceutical use (Mangoni et al 1999; Schafferhans and Klebe 2001; Peitsch 2002). In addition, it is accepted that the growing interest in unveiling protein–protein interactions can benefit from the contributions of comparative modelling and docking programs (Tovchigrechko et al 2002).

In terms of finding disease-related proteins, and for preliminary investigations of potential drugs to modulate the functions of these proteins, the most important genome to generate complete three-dimensional models for is obviously our own human genome. Figure 4 shows the number of human proteins with at least one domain that can be modelled using comparative modelling techniques. We estimate that up to 38% of the translated genome contains domains which can be modelled using templates

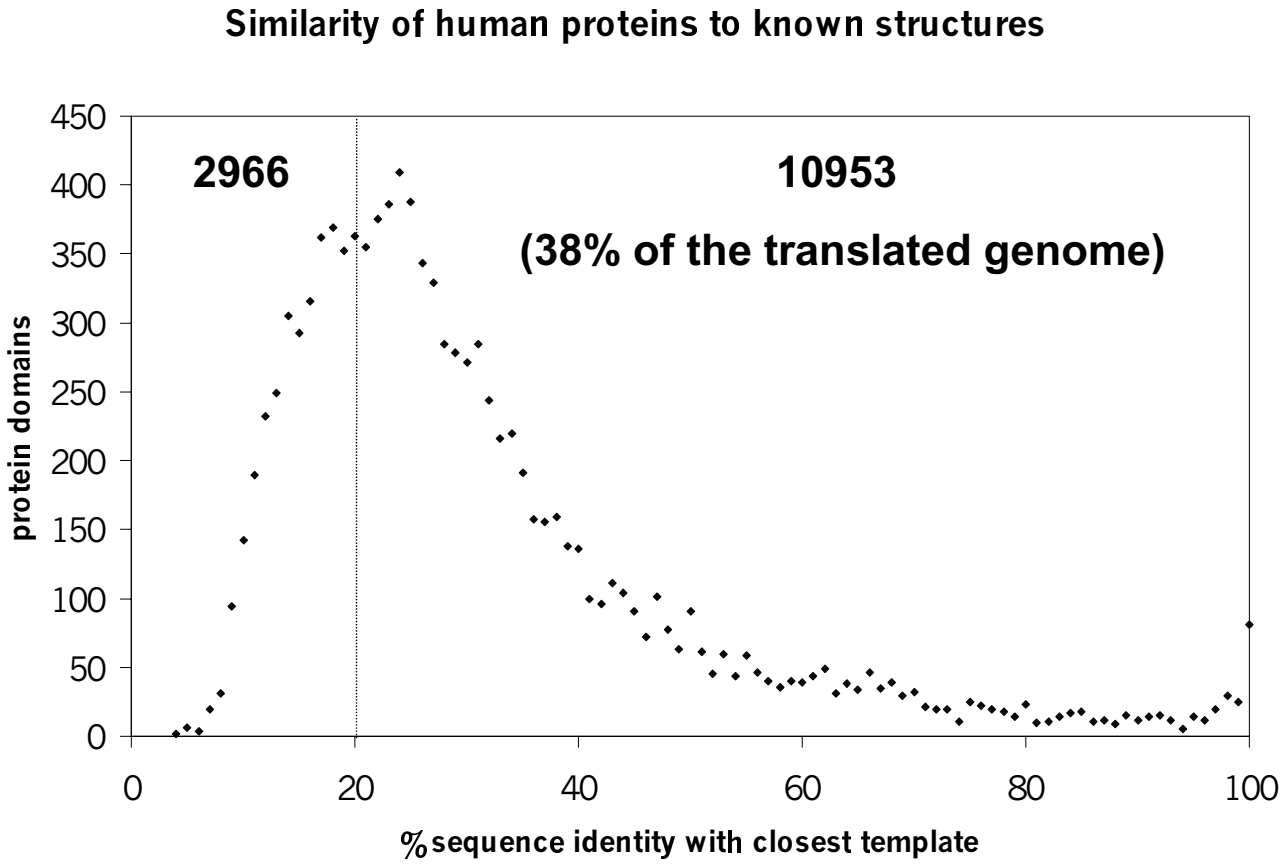

Figure 4 Distribution of human proteins containing at least one domain with significant sequence similarity to SCOP domains. The vertical line separates the fraction that can be modelled to at least a level of resolution that may be useful for experimental design such as site-directed mutagenesis. Over half of the human genome (proteins not represented in the plot) cannot confidently be assigned to known protein folds. These assignments were made using mainly PSI-BLAST.

of at least 20% sequence identity. This would mean a level of expected accuracy for each model of between 0.9 Å and 4.0 Å RMSD. These models could be used for any of the tasks mentioned above, or to understand the structural effects on proteins due to single nucleotide polymorphisms (Wang and Moult 2001) or genetically characterised diseases at the molecular level (Hogg and Bates 2000; Huyton et al 2000).

Problems and potential solutions

As the CASP experiments have shown, comparative modelling involving some form of human intervention still produces models of higher quality than models produced from completely automatic procedures. Intervention seems to be particularly critical in selecting adequate templates and tweaking the alignments (Bates et al 2001; Venclovas 2001). Therefore, more algorithmic development is required if we are to automatically select optimal templates and alignments. Some progress has recently been made with the former problem by selecting templates from large ensembles of sequences, theoretically generated according to their structural compatibility with a template (Koehl and Levitt 2002). Recently the latter problem has also been addressed by consideration of a weighted contribution of a number of current sequence alignment protocols (Elofsson 2002). However, a full appreciation of the power of these new approaches will probably have to wait until the results of CASP5.

Irrespective of the above problems, increasingly more is being asked of comparative modellers. For example, at CASP4 they were expected to model as low as 13% sequence identity with the closest template, and for CASP5 (results not known at the time of writing), of the 38 targets considered to be within reach of comparative modelling, 10 have only between 10%–20% similarity to the closest template. Many of the algorithms designed for comparative modelling were not specifically designed to model at these very remote levels, as this was then considered more the domain of fold recognition experts. Interestingly, this is leading to a progressive merging of the fold recognition and comparative modelling fields. Comparative modellers are learning from the fold recognition community how best to detect very remote sequence relationships and how best to align the query structure to those templates once identified. Equally, those in the fold recognition community are keen to learn how to generate full three-dimensional models from their fold recognition and alignment algorithms. Hopefully this will create a second generation of algorithms, or a blend of algorithms, that are more likely to be successful across a wide range of sequence similarity between query and template sequences. Together with this convergence of algorithms, and on the assumption that only a limited number of protein folds exist, rational structural genomics efforts may be the key to allow three-dimensional modelling of any sequence in a matter of years (Baker and Sali 2001; Vitkup et al 2001). However, the endgame of protein modelling, refining medium resolution models to high levels of atomic accuracy (levels of accuracy routinely obtained in X-ray structures), may take considerably longer as more sophisticated force fields (Halgren and Damm 2001) and substantially more computer power at the fingertips of developers may be required.

Web-based modelling

Although there are a number of well-maintained, downloadable comparative modelling software packages available, the future of comparative modelling as an essential tool for biologists is the growing number of web-based servers. Table 1 summarises the tools that are currently freely available for academic use. The advantage of Web tools is that they are very easy to run, even across different computer platforms, often only requiring the query sequence and user's email address. In addition, the sequence and structural databases that the algorithms require are usually maintained by the developer, thus, linking software to the appropriate up-to-date databases is not a problem. Several of these servers are now allowing some user intervention in the model building process, for example, SWISS-MODEL allows choice of templates and the authors' own server, 3D-JIGSAW, allows both template selection and manual adjustments of the query to template alignments.

Conclusion

There is little doubt that comparative modelling, if it is not already considered to be so, will become an essential tool for molecular biologists and those involved in rational drug design. It is therefore essential that comparative modelling tools are readily accessible, both in terms of downloadable, easy to use software packages and versatile, quick response web-based tools. Due to the high importance of this field, algorithmic developments on all aspects of comparative modelling must be encouraged. These necessary developments range from template selection and sequence alignments, to energy optimisation and movement analysis

of the constructed three-dimensional models. This will require dedicated efforts from scientists within a wide range of disciplines, particularly mathematicians, physicists and computer scientists. These developments are essential if we are to routinely refine useful, but often low resolution models, to the atomic resolution found within most X-ray structures.

Acknowledgements

Thanks to Arne Müller for his help to calculate the coverage of the human genome and to the BMM group for helpful comments and discussions.

Notes

[1] This experiment, held every two years, is where the CASP organisers (see for example Moult et al (2001)) send protein modellers the sequences of recently determined structures before those structures are actually published. Modellers then make predictions for those structures and a committee of external assessors evaluates the quality of each model. Finally, in December of that year, participants attend an evaluation conference where the failures and successes of the modelling protocols used, and possible improvements to them, are discussed. At the time of writing, four CASP experiments have been completed and the fifth, for which all predictions have been submitted, is currently being assessed.

References

Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G, Bork P, Bottcher B, Russell RB. 2002. A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep*, 3:628–35.

Aloy P, Russell RB. 2002. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA*, 99:5896–901.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25: 3389–402.

Alves R, Chaleil RA, Sternberg MJ. 2002. Evolution of enzymes in metabolism: a network perspective. *J Mol Biol*, 320:751–70.

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science*, 181:223–30.

Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN. 1998. The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res*, 26:304–8.

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. *Science*, 294:93–6.

Baldi P, Chauvin Y, Hunkapiller T, McClure MA. 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA*, 91:1059–63.

Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, 45 Suppl 5:39–46.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res*, 28:235–42.

Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347–52.

Bower MJ, Cohen FE, Dunbrack RL. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol*, 267:1268–82.

Branden C, Tooze J. 1999. Introduction to protein structure. New York: Garland.

Braun W, Go N. 1985. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J Mol Biol*, 186:611–26.

Brodersen DE, Clemons WM, Carter AP, Wimberly BT, Ramakrishnan V. 2002. Crystal structure of the 30 S ribosomal subunit from *Thermus thermophilus*: structure of the proteins and their interactions with 16 S RNA. *J Mol Biol*, 316:725–68.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: a program for macromolecular energy, minimization and dynamics calculation. *J Comp Chem*, 4:187–217.

Bruccoleri RE, Haber E, Novotny J. 1988. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature*, 335:564–8.

Bruccoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 26:137–68.

Bruccoleri RE, Karplus M. 1990. Conformational sampling using high-temperature molecular dynamics. *Biopolymers*, 29:1847–62.

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. 2001. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, 45 Suppl 5:184–91.

Carlacci L, Englander SW. 1996. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. *J Comp Chem*, 17:1002–12.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5:823–6.

Colloch N, Cohen FE. 1991. Beta-Breakers: an aperiodic secondary structure. *J Mol Biol*, 221:603–13.

Collura V, Higo J, Garnier J. 1993. Modeling of protein loops by simulated annealing. *Protein Sci*, 2:1502–10.

Contreras-Moreira B, Bates PA. 2002. Domain fishing: a first step in protein comparative modelling. *Bioinformatics*, 18:1141–2.

De Maeyer M, Desmet J, Lasters I. 2000. The dead-end elimination theorem: mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol Biol*, 143:265–304.

Desjarlais JR, Handel TM. 1999. Side-chain and backbone flexibility in protein core design. *J Mol Biol*, 290:305–18.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–42.

Devos D, Garmendia J, de Lorenzo V, Valencia A. 2002. Deciphering the action of aromatic effectors on the prokaryotic enhancer-binding protein XylR: a structural model of its N-terminal domain. *Environ Microbiol*, 4:29–41.

Devos D, Valencia A. 2000. Practical limits of function prediction. *Proteins*, 41:98–107.

Dudek MJ, Ramnarayan K, Ponder JW. 1998. Protein structure prediction using a combination of sequence homology and global minimization II. Energy functions. *J Comp Chem*, 19:548–73.

Dudek MJ, Scerage HA. 1990. Protein structure prediction using a combination of sequence homology and global energy minimization I. Global energy minimization of surface loops. *J Comp Chem*, 11: 121–51.

Dunbrack RL, Karplus M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*, 230: 543–74.

Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol*, 6:361–5.

Edwards MS, Sternberg JE, Thornton JM. 1987. Structural and sequence patterns in the loops of beta alpha beta units. *Protein Eng*, 1:173–81.

Efimov AV. 1991. Structure of alpha-alpha-hairpins with short connections. *Protein Eng*, 4:245–50.

Elcock AH. 2002. Modeling supramolecular assemblages. *Curr Opin Struct Biol*, 12:154–60.

Elofsson A. 2002. A study on protein sequence alignment quality. *Proteins*, 46:330–9.

Evans JS, Mathiowetz AM, Chan SI, Goddard WA. 1995. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. *Protein Sci*, 4:1203–16.

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. 2001. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, 17:1242–3.

Fidelis K, Stern PS, Bacon D, Moult J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng*, 7:953–60.

Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins*, 1:342–62.

Fiser A, Do RK, Sali A. 2000. Modeling of loops in protein structures. *Protein Sci*, 9:1753–73.

Garmendia J, Devos D, Valencia A, de Lorenzo V. 2001. A la carte transcriptional regulators: unlocking responses of the prokaryotic enhancer-binding protein XylR to non-natural effectors. *Mol Microbiol*, 42:47–59.

Greer J. 1981. Comparative model-building of the mammalian serine proteases. *J Mol Biol*, 153:1027–42.

Guex N, Diemand A, Peitsch MC. 1999. Protein modelling for all. *Trends Biochem Sci*, 24:364–7.

Halgren TA, Damm W. 2001. Polarizable force fields. *Curr Opin Struct Biol*, 11:236–42.

Havel TF, Snow ME. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol*, 217:1–7.

Hayward S. 1999. Structural principles governing domain motions in proteins. *Proteins*, 36:425–35.

Henikoff S, Henikoff JG. 1993. Performance evaluation of amino acid substitution matrices. *Proteins*, 17:49–61.

Henikoff S, Henikoff JG, Pietrokovski S. 1999. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15:471–9.

Hogg N, Bates PA. 2000. Genetic analysis of integrin function in man: LAD-1 and other syndromes. *Matrix Biol*, 19:211–22.

Hooft RWW, Sander C, Vriend G. 1996. Verification of protein structures: side-chain planarity. *J Appl Cryst*, 29:714–16.

Huyton T, Bates PA, Zhang X, Sternberg MJ, Freemont PS. 2000. The BRCA1 C-terminal domain: structure and function. *Mutat Res*, 460:319–32.

Irving JA, Whisstock JC, Lesk AM. 2001. Protein structural alignments and functional genomics. *Proteins*, 42:378–82.

Jacobson MP, Friesner RA, Xiang Z, Honig B. 2002. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol*, 320:597–608.

Janardhan A, Vajda S. 1998. Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. *Protein Sci*, 7:1772–80.

Jones TA, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J*, 5:819–22.

Karplus M, McCammon JA. 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, 9:646–52.

Kawasaki H, Kretsinger RH. 1995. Calcium-binding proteins 1: EF-hands. *Protein Profile*, 2:297–490.

Kelley LA, MacCallum RM, Sternberg MJ. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, 299:499–520.

Koehl P, Delarue M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239:249–75.

Koehl P, Levitt M. 2002. Sequence variations within protein families are linearly related to structural variations. *J Mol Biol*, 323:551–62.

Kolodny R, Koehl P, Guibas L, Levitt M. 2002. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, 323:297–307.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, 235:1501–31.

Kussell E, Shimada J, Shakhnovich EI. 2001. Excluded volume in protein side-chain packing. *J Mol Biol*, 311:183–93.

Kwasigroch JM, Chomilier J, Mornon JP. 1996. A global taxonomy of loops in globular proteins. *J Mol Biol*, 259:855–72.

Lambert C, Leonard N, De Bolle X, Depiereux E. 2002. ESyPred3D: prediction of proteins 3D structures. *Bioinformatics*, 18:1250–6.

Lambert MH, Scheraga HA. 1989. Pattern recognition in the predicition of protein structure. *J Comp Chem*, 10:770–831.

Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst*, 26:283–91.

Lasters I, Desmet J. 1993. The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng*, 6:717–22.

Lee C, Subbiah S. 1991. Prediction of protein side-chain conformation by packing optimization. *J Mol Biol*, 217:373–88.

Lee MR, Tsai J, Baker D, Kollman PA. 2001. Molecular dynamics in the endgame of protein structure prediction. *J Mol Biol*, 313:417–30.

Liang S, Grishin NV. 2002. Side-chain modeling with an optimized scoring function. *Protein Sci*, 11:322–31.

Liu J, Tan H, Rost B. 2002. Loopy proteins appear conserved in evolution. *J Mol Biol*, 322:53.

Lund O, Frimand K, Gorodkin J, Bohr H, Bohr J, Hansen J, Brunak S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng*, 10:1241–8.

Mangoni M, Roccatano D, Di Nola A. 1999. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, 35:153–62.

Mansfeld J, Vriend G, Dijkstra BW, Veltman OR, Van den Burg B, Venema G, Ulbrich-Hofmann R, Eijsink VG. 1997. Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J Biol Chem*, 272:11152–6.

Martin AC, MacArthur MW, Thornton JM. 1997. Assessment of comparative modeling in CASP2. *Proteins*, 29 Suppl 1:14–28.

Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. 2002. Reliability of assessment of protein structure prediction methods. *Structure (Camb)*, 10:435–40.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325.

McGarrah DB, Judson RS. 1993. Analysis of the genetic algorithm method of molecular conformation determination. *J Comp Chem*, 14: 1385–95.

McGregor MJ, Islam SA, Sternberg MJ. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol*, 198:295–310.

Melo F, Sanchez R, Sali A. 2002. Statistical potentials for fold assessment. *Protein Sci*, 11:430–48.

Mendes J, Baptista AM, Carrondo MA, Soares CM. 1999. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins*, 37:530–43.

Mezei M. 1998. Chameleon sequences in the PDB. *Protein Eng*, 11: 411–4.

Milner-White EJ, Poet R. 1987. Loops, bulges, turns and hairpins in proteins. *Trends Biochem Sci*, 12:189–92.

Moult J, Fidelis K, Zemla A, Hubbard T. 2001. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, 45 Suppl 5:2–7.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–40.

Nakajima N, Higo J, Kidera A, Nakamura H. 2000. Free energy landscapes of peptides by enhanced conformational sampling. *J Mol Biol*, 296:197–216.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302:205–17.

Ogata K, Umeyama H. 2000. An automatic homology modeling method consisting of database searches and simulated annealing. *J Mol Graph Model*, 18:258–72, 305–6.

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. 1997. An automated classification of the structure of protein loops. *J Mol Biol*, 266: 814–30.

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. 1998. Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J Mol Biol*, 279:1193–210.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH- a hierarchic classification of protein domain structures. *Structure*, 5:1093–108.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 284:1201–10.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85:2444–8.

Peitsch MC. 2002. About the use of protein models. *Bioinformatics*, 18:934–8.

Petrella RJ, Karplus M. 2001. The energetics of off-rotamer protein side-chain conformations. *J Mol Biol*, 312:1161–75.

Przybylski D, Rost B. 2002. Alignments grow, secondary structure prediction improves. *Proteins*, 46:197–205.

Rao U, Teeter MM. 1993. Improvement of turn structure prediction by molecular dynamics: a case study of alpha 1-purothionin. *Protein Eng*, 6:837–47.

Rice PA, Goldman A, Steitz TA. 1990. A helix-turn-strand structural motif common in alpha-beta proteins. *Proteins*, 8:334–40.

Rockey WM, Elcock AH. 2002. Progress toward virtual screening for drug side effects. *Proteins*, 48:664–71.

Rose GD, Gierasch LM, Smith JA. 1985. Turns in peptides and proteins. *Adv Protein Chem*, 37:1–109.

Rosenbach D, Rosenfeld R. 1995. Simultaneous modeling of multiple loops in proteins. *Protein Sci*, 4:496–505.

Rost B. 1995. TOPITS: threading one-dimensional predictions into three-dimensional structures. *Proc Int Conf Intell Syst Mol Biol*, 3:314–21.

Rost B. 2002. Enzyme function less conserved than anticipated. *J Mol Biol*, 318:595–608.

Rost B, Honig B, Valencia A. 2002. Bioinformatics in structural genomics. *Bioinformatics*, 18:897–8.

Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, 14:309–23.

Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol*, 269:423–39.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234:779–815.

Samudrala R, Moult J. 1998. A graph-theoretic algorithm for comparative modeling of protein structure. *J Mol Biol*, 279:287–302.

Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci USA*, 95:13597–602.

Sanchez R, Sali A. 1999. ModBase: a database of comparative protein structure models. *Bioinformatics*, 15:1060–1.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68.

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*, 29:2994–3005.

Schafferhans A, Klebe G. 2001. Docking ligands onto binding site representations derived from proteins built by homology modelling. *J Mol Biol*, 307:407–27.

Schonbrun J, Wedemeyer WJ, Baker D. 2002. Protein structure prediction in 2002. *Curr Opin Struct Biol*, 12:348–54.

Schwartz RM, Dayhoff MO. 1978. Matrices for detecting distant relationships. In Dayhoff MO, ed. Volume 5 Supplement 3, Atlas of protein sequence and structure. Washington: Natl Biomed Res Found. p 353–8.

Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. 1987. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers*, 26:2053–85.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310:243–57.

Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11:739–47.

Simons KT, Kooperberg C, Huang E, Baker D. 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.

Smith KC, Honig B. 1994. Evaluation of the conformational free energies of loops in proteins. *Proteins*, 18:119–32.

Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins, Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, 1:377–84.

Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol*, 208:1–22.

Thanki N, Zeelen JP, Mathieu M, Jaenicke R, Abagyan RA, Wierenga RK, Schliebs W. 1997. Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop. *Protein Eng*, 10:159–67.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–80.

Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. 2000. From structure to function: approaches and limitations. *Nat Struct Biol*, 7 Suppl 11:991–4.

Tovchigrechko A, Wells CA, Vakser IA. 2002. Docking of protein models. *Protein Sci*, 11:1888–96.

Tramontano A, Leplae R, Morea V. 2001. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, 45 Suppl 5: 22–38.

Unger R, Harel D, Wherland S, Sussman JL. 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5:355–73.

Vajda S, DeLisi C. 1990. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers*, 29:1755–72.

van Vlijmen HW, Karplus M. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol*, 267:975–1001.

Venclovas C. 2001. Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins*, 45 Suppl 5:47–54.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HA, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. *Science*, 291:1304–51.

Ventkatachalam CM. 1968. Stereochemical criteria for polypeptides and proteins. Conformation of a system of three linked peptide units. *Biopolymers*, 6:1425–36.

Vitkup D, Melamud E, Moult J, Sander C. 2001. Completeness in structural genomics. *Nat Struct Biol*, 8:559–66.

Walker JE, Saraste M, Runswick MJ, Gay NJ. 1982. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J*, 1:945–51.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum Mutat*, 17:263–70.

Wintjens RT, Rooman MJ, Wodak SJ. 1996. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol*, 255: 235–53.

Wood TC, Pearson WR. 1999. Evolution of protein sequences and structures. *J Mol Biol*, 291:977–95.

Wriggers W, Chacon P. 2001. Modeling tricks and fitting techniques for multiresolution structures. *Structure (Camb)*, 9:779–88.

Xiang Z, Honig B. 2001. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*, 311:421–30.

Xiang Z, Soto CS, Honig B. 2002. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci USA*, 99:7432–7.

Zhang X, Shaw A, Bates PA, Newman RH, Gowen B, Orlova E, Gorman MA, Kondo H, Dokurno P, Lally J et al. 2000. Structure of the AAA ATPase p97. *Mol Cell*, 6:1473–84.

Zheng Q, Kyle DJ. 1996. Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings. *Proteins*, 24:209–17.

Zheng Q, Rosenfeld R, Vajda S, DeLisi C. 1993. Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci*, 2:1242–8.