ELSEVIER

A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction
John Moult

For the past ten years, CASP (Critical Assessment of Structure Prediction) has monitored the state of the art in modeling protein structure from sequence. During this period, there has been substantial progress in both comparative modeling of structure (using information from an evolutionarily related structural template) and template-free modeling. The quality of comparative models depends on the closeness of the evolutionary relationship on which they are based. Template-free modeling, although still very approximate, now produces topologically near correct models for some small proteins. Current major challenges are refining comparative models so that they match experimental accuracy, obtaining accurate sequence alignments for models based on remote evolutionary relationships, and extending template-free modeling methods so that they produce more accurate models, handle parts of comparative models not available from a template and deal with larger structures.

Addresses
Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850, USA

Introduction
This article reviews progress in modeling protein structure from amino acid sequence over the past ten years, as monitored by CASP (Critical Assessment of Structure Prediction); another independent useful review is [1•]. CASP is a community-wide experiment with the primary aim of assessing the effectiveness of modeling methods. This review covers CASP experiments 1 (held in 1994) through 6 (2004). The emphasis is on what has been learnt about the strengths and weaknesses of prediction methods, what progress has been made, where there are serious bottlenecks to further progress and how these may eventually be removed. Reviews and reports on the CASP6 experiment were not yet available at the time of writing. Readers are advised to check the literature for newer material.

The CASP experiment
CASP is a large-scale community experiment, conducted every two years. The key feature is that participants make *bona fide* blind predictions of structures. Over 200 prediction teams from 24 countries participated in CASP6. Information about soon-to-be experimentally determined protein structures is collected and passed on to registered predictors. Over 30 000 predictions for 64 protein targets divided into 90 domains were collected and evaluated. Predictors fall into two categories: teams of participants who devote considerable time and effort to modeling each target, usually having a period of several weeks to complete their work; and automatic servers, which must return a model within 48 hours, in principle without human intervention. Servers provide information on what can be achieved by robust and rapid computer methods alone. Predictions are evaluated using a battery of numerical criteria [2] and, more importantly, are carefully examined by independent assessors. A conference is then held to discuss the results and a special issue of the journal *Proteins* is published, with articles by the assessors and by some of the more successful prediction teams. Further details can be found in the most recent special issue, for the fifth experiment. In particular, articles by the three assessment groups [3••–5••] provide a detailed overview of the state of the art at that time and another article puts the results into the context of previous CASP experiments [6••]. The issue for the sixth experiment will appear in the autumn of 2005. Participant registration, target management, prediction collection and numerical analysis are all handled by the Protein Structure Prediction Center [2]. That web site (predictioncenter.llnl.gov) provides access to details of the experiment and all results. A second web site (www.forcasp.org) provides a discussion forum for the CASP community.

Classes of structure prediction difficulty
Early work in the structure modeling field primarily focused on understanding the nature of the natural folding process and on the development of physics-based force fields to determine the relative free energy of any conformation of a polypeptide chain. These methods were much in evidence at the first CASP, but have largely been supplanted by more successful 'knowledge-based' approaches, which utilize the large and rapidly growing number of experimentally determined structures and sequences in a variety of ways. As a consequence, the accuracy of models depends on similarity to already known structures. Using this principle, CASP divides modeling difficulty into four classes.

Comparative modeling based on a clear sequence relationship

When there is an easily detectable sequence relationship between a target protein and one or more of known structure (a high significance score from a BLAST search), an accurate core model (typically 2–3 Å rms error for Cα atoms) can be obtained by copying from the structural template or templates [3[••]]. Copying is often non-trivial, requiring a correct alignment of the target and template sequences. Improvements over the course of the CASP experiments have resulted in largely correct alignments for this modeling class. Alignment methods are discussed below. A single template rarely provides a complete model. Alternative template structures may provide some additional structural features. Short regions of chain ('loops') are sometimes modeled in an approximately correct manner. Generally, reliably building regions of the structure not present in a template remains a challenge. Sidechain conformations are very tightly correlated with backbone conformation [7], so, not surprisingly, sidechain accuracy of these approximate models is poor.
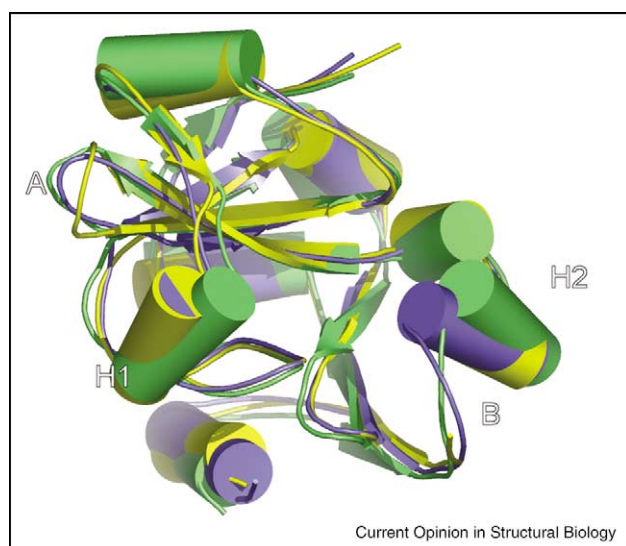
A typical model is shown in Figure 1, CASP6 target 266, an *Aeropyrum pernix* homolog of a *Haemophilus influenzae* proline tRNA editing enzyme [8]. For large regions of the structure, the template provides an accurate guide, resulting in good overall quality. Two non-template loop regions (A and B) are successfully modeled. The largest differences between the template and the target are in two helices (H1 and H2) flanking the active site, suggesting different substrate specificities (PDB entry 1wdv). The best models leave the helices in the template orientation, so it is not possible to analyze any specificity difference. In general, although the structure around active sites is usually well conserved among proteins with the same specificity, it is often the least conserved when the specificities differ.

Although large parts of the models are approximately correct, they require refinement to be competitive with experiment and to reproduce key functional features. Early hopes that molecular dynamics methods would allow refinement have not been fulfilled. Reasons for this are a matter of hot debate within the field, with three suggested inter-related explanations: inadequate sampling of alternative conformations, insufficiently accurate description of the interatomic forces and too short trajectories. Refinement remains the principal bottleneck to progress.

In spite of limitations, this class of model is very useful for a variety of purposes, such as identifying which members of a protein family have the same detailed function and which are different [9[••]].

Modeling based on more distant evolutionary relationships

A second class of model quality is provided by cases that often require the use of more sophisticated methods than BLAST to detect an evolutionary relationship. The core of these methods is the alignment of a set of sequences. A major step forward in this area was the introduction of PSI-BLAST [10] and hidden Markov models [11–13]. Both of these methods compile a profile based on a single input sequence. A profile may be built from a target sequence and the sequences of all known structures tested for compatibility with it, or the target sequence may be compared with profiles compiled for all known structures. More recently, profile-profile methods, whereby a profile built from a target sequence is compared with profiles built for all known structures, have been introduced [14[•],15,16], further increasing the sensitivity and allowing more remote homologs to be detected and aligned [17]. Structural information is used in several ways to enhance the detection of homologs. If multiple structures are known for a protein family, sequence alignment based on structure superposition leads to more reliable profiles [18]. Local structural features are used to influence the introduction of gaps in alignments (e.g. [19,20]). Also, predicted local structural features may be compared with those of a candidate fold, sometimes integrated with the sequence-profile methods

Figure 1

Current Opinion in Structural Biology

CASP6 target 266, an example of a model based on a relatively close evolutionary relationship. The best model is blue, the experimental structure (PDB code 1wdv) is green and the available template structure (28% sequence identity to the target, PDB code 1dbu_A) is yellow. Where template and target are similar (yellow and green superpose), the model is accurate. Two loop regions not available in the template (A and B) are also reasonably correct. Helices H1 and H2 have different orientations in template and target, not corrected in the model. These structural features may be related to ligand specificity differences. Refinement of these models to rival experiment remains a central challenge, with signs of recent progress.

[21,22]. Several scoring functions have been applied to evaluating and choosing alternative alignments and structures (e.g. [23,24]). Varying the alignment parameters provides one means of generating alternatives and identifying regions where the alignment is least reliable [25].

Improvements in the quality of this class of model have been incremental but steady over the course of the CASP experiments. Noteworthy progress between CASP4 and CASP5 was partly due to the effective use of metaservers. These automatic servers collect models from other servers and use that input to produce consensus structures [26[••]]. By CASP6, many successful human prediction teams used starting models produced by metaservers.

Models based on the detection of these weaker sequence relationships are limited in accuracy by four main factors. First, a correct template may not be identified. The increasing size of the pool of known sequences and the improved methods outlined above have led to a steady improvement in recognition. In the past few CASP experiments, there have been very few targets for which one or more modeling groups did not identify the evolutionary relationship (there was one such target in CASP6). So far, no single group is able to recognize all relationships. Second, even though an evolutionary relationship has been detected, aligning the target sequence onto the template structure or structures is challenging, and typically results in very significant errors. There has been considerable progress in alignment accuracy over the course of CASP [6[••]], continuing in CASP6. Third, although evolutionarily related structures share many features, a significant fraction of residues in a target will have no structural equivalent in an available template. For CASP targets, this varies from about 10% at the close homolog end of the range to about 50% at the remote homolog limit [6[••]]. So far, there are occasional successes with modeling these additional regions, for example, the inclusion of a non-template helix in CASP6 target 205. Fourth, although the equivalent regions of evolutionarily related structures are similar, they are not identical and the more remote the relationship, the larger the differences between related regions of the backbone.

Further progress in modeling features not present in a template requires the more systematic application of template-free modeling methods. As discussed below, these are now powerful enough in many cases. The energy difference between alternative alignments is determined by the detailed atomic interactions of adjacent regions of polypeptide chain. These interactions are not correctly represented in this class of model and so further substantial improvements in alignment accuracy may not be possible until effective all-atom refinement procedures are developed. It will then be necessary to refine models based on each possible alignment variant and compare the energetics to determine which is correct.

For the reasons listed above, details of these models are not accurate. Nevertheless, they are useful for providing an overall idea of what a structure is like, helping to choose residues for mutagenesis experiments, for example. They may also identify the superfamily that a protein belongs to and so provide valuable, although approximate, information about molecular function. Figure 2 shows an example from CASP6.

Modeling based on non-homologous fold relationships

Even when no relationship can be detected, the fold of a target protein may be similar to that of a known structure, either because of an evolutionary relationship too remote for detection or because the two folds have converged to be similar. During the 1990s, methods were developed that focused on identifying and utilizing such structural relationships (usually referred to as 'threading' [27]). In recent CASP experiments, these methods have not been competitive with template-free approaches, although some packages retain them as one component of an integrated strategy [20,28].

Template-free modeling

For targets with folds that have not previously been identified or for which no relationship with a protein of known structure can be detected, a different set of

Figure 2


Current Opinion in Structural Biology
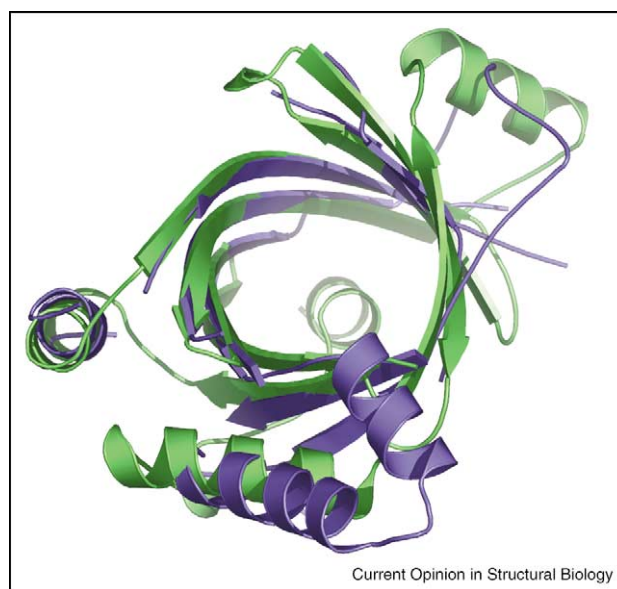
CASP6 target 197, an example of comparative modeling based on a distant evolutionary relationship. The best model is blue and the experimental structure (PDB code 1xkc) is green. Accurately modeled regions of the β barrel reflect available template information. Other regions, outside the β barrel, have different conformations from the template and are not accurately modeled. These structural features are probably related to detailed functional differences. In spite of limited accuracy, structure-assisted recognition of these evolutionary relationships does provide valuable information about function, in this case probable involvement in RNA editing.

methods are needed. Traditionally, this was the area in which physics-based approaches were used. These methods are still used by a few CASP participants, but have been largely displaced. Newer methods primarily utilize the fact that, although we are far from observing all folds used in biology [29], we probably have seen nearly all substructures [30]. Methods make use of these substructure relationships, on a range of scales [31] — from a few residues [32••], through secondary structure units, to supersecondary units [33]. Structure fragments are chosen on the basis of compatibility of the substructure with the local target sequence and compatibility of secondary structure propensity. As the sequence/structure relationship is rarely strong enough to completely determine the structure of fragments [34], a range of possible conformations of each fragment are usually selected and many possible combinations of substructures are considered. Relationships between substructures may be determined by approximate potentials that guide a conformational search process and the prediction of residue contacts [5••]. Large numbers of possible complete structures (1000–100 000) are usually generated. The most successful package using this strategy (ROSETTA) has recently been reviewed [32••]. For proteins of less than about 100 residues, these procedures may produce one or a few approximately correct structures (4–6 Å rmsd for Cα atoms). Selecting the most accurate structures from the large set of candidates remains to be solved and most methods rely on clustering procedures, selecting representative structures at the center of the largest clusters of generated candidates [35]. Structure-based scoring functions, whether based on physics or 'knowledge based' (utilizing features of experimental structures), are not adequately effective. The probable reason for this is that the information determining the preferred structure primarily resides in the detailed atomic interactions. As a consequence, reliable identification of accurate models will require the use of refined all-atom models. Thus, for this class of modeling too, the development of atomic-level refinement methods is probably crucial to major progress.

In CASP1, all new fold models were close to random. There has been steady improvement over the course of CASP and, by CASP6, most non-homology targets of less than 100 residues have models that visual inspection shows to be closely related to experiment. Examples are targets 198, 201, 215, 248 domain 2 and 281 (Figure 3). Between CASP5 and CASP6, most progress has been made in producing reasonable models for proteins with a significant amount of β structure. Models for larger proteins or domains are still rarely usefully accurate. One contributing difficulty with larger structures is parsing into domains. Separate evaluation of this in CASP6 showed that it is problematic if no homologies can be detected. Thus, although there is very impressive progress for small proteins, there is still a long way to go before all proteins can be modeled at that level. Also,

Figure 3

Current Opinion in Structural Biology
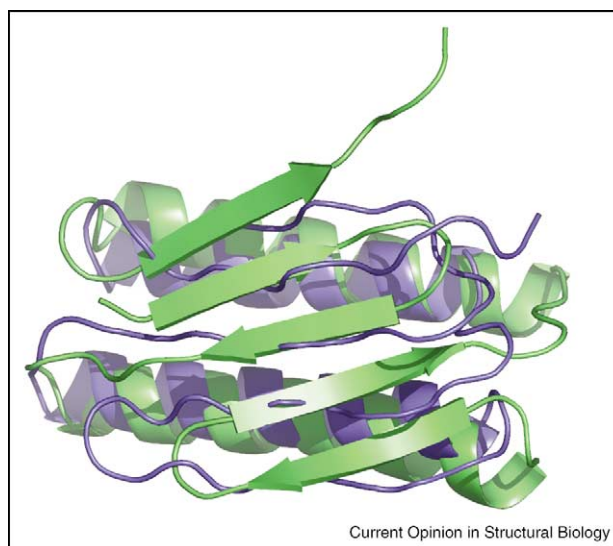
CASP6 target 201, an example of modeling a previously unknown fold. The best model is blue and the experimental structure (PDB code 1s12) is green. The helical regions are accurately modeled and the general features of the β sheet are correct, although there is a topology error and the sheet is slightly misoriented. This quality of model is now often obtained for small structures.

although topologically pleasing, these models often have significant alignment and other errors.

Major current challenges

As detailed above, overcoming four of the current major bottlenecks — close evolutionary relationship models approaching experimental accuracy, improved alignments, refinement of remote evolutionary relationship models and reliable discrimination among possible template-free models — is dependent on the development of effective all-atom structure refinement procedures. The 'refinement' problem has received increasing attention in recent years (http://www.nigms.nih.gov/psi/reports/comparative_modeling.html). At CASP6, for the first time, there was a report of an initial model refined from a backbone rmsd of about 2.2 Å to about 1.6 Å, with many of the core sidechains correctly oriented. This result is for a small protein (target 281) and, at present, the method does not scale to larger structures. Nevertheless, it is an encouraging signal of progress in this key area.

Acknowledgements

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Bourne PE: **CASP and CAFASP experiments and their findings**.
• *Methods Biochem Anal* 2003, **44**:501-507.
An independent review of the state of the art in protein structure modeling.

2. Zemla A, Venclovas, Moult J, Fidelis K: **Processing and evaluation of predictions in CASP4**. *Proteins* 2001, **suppl 5**:13-21.

3. Tramontano A, Morea V: **Assessment of homology-based**
•• **predictions in CASP5**. *Proteins* 2003, **53(suppl 6)**:352-368.
An assessment of the state of the art in comparative modeling of protein structure.

4. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI,
•• Qi Y, Pei J, Cheng H, Grishin NV: **CASP5 assessment of fold recognition target predictions**. *Proteins* 2003, **53(suppl 6)**:395-409.
An assessment of the state of the art in fold-recognition-based modeling.

5. Aloy P, Stark A, Hadley C, Russell RB: **Predictions without**
•• **templates: new folds, secondary structure, and contacts in CASP5**. *Proteins* 2003, **53(suppl 6)**:436-456.
An assessment of the state of the art in template-free modeling.

6. Venclovas C, Zemla A, Fidelis K, Moult J: **Assessment of**
•• **progress over the CASP experiments**. *Proteins* 2003, **53(suppl 6)**:585-595.
An analysis of progress over the course of the CASP experiments.

7. Chung SY, Subbiah S: **The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins**. *Protein Sci* 1995, **4**:2300-2309.

8. An S, Musier-Forsyth K: **Trans-editing of Cys-tRNAPro by *Haemophilus influenzae* YbaK protein**. *J Biol Chem* 2004, **279**:42359-42362.

9. DeWeese-Scott C, Moult J: **Molecular modeling of protein**
•• **function regions**. *Proteins* 2004, **55**:942-961.
An analysis of the usefulness of comparative modeling in interpreting details of molecular function.

10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

11. Karplus K, Hu B: **Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set**. *Bioinformatics* 2001, **17**:713-720.

12. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies**. *Bioinformatics* 1998, **14**:846-856.

13. Kahsay RY, Wang G, Gao G, Liao L, Dunbrack R: **Quasi-consensus based comparison of profile hidden Markov models for protein sequences**. *Bioinformatics* 2005, doi:10.1093/bioinformatics/bti374.

14. Ohlson T, Wallner B, Elofsson A: **Profile-profile methods provide**
• **improved fold-recognition: a study of different profile-profile alignment methods**. *Proteins* 2004, **57**:188-197.
An evaluation of profile-profile alignment methods.

15. Wang G, Dunbrack RL. Jr: **Scoring profile-to-profile sequence alignments**. *Protein Sci* 2004, **13**:1612-1626.

16. Wallner B, Fang H, Ohlson T, Frey-Skott J, Elofsson A: **Using evolutionary information for the query and target improves fold recognition**. *Proteins* 2004, **54**:342-350.

17. Marti-Renom MA, Madhusudhan MS, Sali A: **Alignment of protein sequences by their profiles**. *Protein Sci* 2004, **13**:1071-1087.

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

19. Wrabl JO, Grishin NV: **Gaps in structurally similar proteins: towards improvement of multiple sequence alignment**. *Proteins* 2004, **54**:71-87.

20. McGuffin LJ, Jones DT: **Improvement of the GenTHREADER method for genomic fold recognition**. *Bioinformatics* 2003, **19**:874-881.

21. Przybylski D, Rost B: **Improving fold recognition without folds**. *J Mol Biol* 2004, **341**:255-269.

22. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R: **Combining local-structure, fold-recognition, and new fold methods for protein structure prediction**. *Proteins* 2003, **53(suppl 6)**:491-496.

23. von Grotthuss M, Pas J, Wyrwicz L, Ginalski K, Rychlewski L: **Application of 3D-Jury, GRDB, and Verify3D in fold recognition**. *Proteins* 2003, **53(suppl 6)**:418-423.

24. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins**. *Proteins* 1993, **17**:355-362.

25. Venclovas C: **Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance**. *Proteins* 2003, **53(suppl 6)**:380-388.

26. Rychlewski L, Fischer D: **LiveBench-8: the large-scale,**
•• **continuous assessment of automated protein structure prediction**. *Protein Sci* 2005, **14**:240-245.
A report on the performance of servers for protein structure modeling, including metaservers.

27. Godzik A: **Fold recognition methods**. *Methods Biochem Anal* 2003, **44**:525-546.

28. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D: **TOUCHSTONE: a unified approach to protein structure prediction**. *Proteins* 2003, **53(suppl 6)**: 469-479.

29. Coulson AF, Moult J: **A unifold, mesofold, and superfold model of protein fold use**. *Proteins* 2002, **46**:61-71.

30. Du P, Andrec M, Levy RM: **Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update**. *Protein Eng* 2003, **16**:407-414.

31. Bystroff C, Shao Y, Yuan X: **Five hierarchical levels of sequence-structure correlation in proteins**. *Appl Bioinformatics* 2004, **3**:97-104.

32. Rohl CA, Strauss CE, Misura KM, Baker D: **Protein structure**
•• **prediction using Rosetta**. *Methods Enzymol* 2004, **383**:66-93.
A detailed description of one of the more successful packages for template-free modeling of protein structure.

33. Jones DT, McGuffin LJ: **Assembling novel protein folds from super-secondary structural fragments**. *Proteins* 2003, **53(suppl 6)**:480-485.

34. Bystroff C, Simons KT, Han KF, Baker D: **Local sequence-structure correlations in proteins**. *Curr Opin Biotechnol* 1996, **7**:417-421.

35. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M: ***Ab initio* protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement**. *Proteins* 2001, **suppl 5**:149-156.