A brighter future for protein structure prediction

Patrice Koehl and Michael Levitt

The most recent critical assessment of structure prediction meeting (CASP3) revealed significant progress in predicting the three-dimensional folds of proteins with unknown structures.

In recent alternating years, protein structure predictors have met in Asilomar, California, where their predictions and methods have been assessed. The most recent critical assessment of structure prediction (CASP) meeting* was held this past December, and the results presented at this conference reveal clear progress in the areas of accurate alignments, sensitive fold recognition, reasonable *ab initio* structure modeling and improved secondary structure prediction. These advances give hope that protein structure prediction is indeed a tractable problem. They also have implications for ongoing efforts to assign structure and function to the thousands of novel genes that are sequenced each year.

Predicting a protein's structure from its amino acid sequence is a 'holy grail' for the structural biology community. In spite of decades of effort, it remains an extremely difficult problem both because the folded three-dimensional structure of a protein is complicated and because the structure is defined by many degrees of freedom. All theoretical and computational approaches must be validated by testing. This is usually accomplished by attempts to reproduce the known folds of a number of small proteins. Such tests can be criticized in that they 'remember' the known structures that are being 'predicted'. For truly objective assessment, one needs a system of blind prediction. Five years ago, such a scheme was devised by John Moult (CARB, University of Maryland Biotechnology Institute, USA), who initiated the CASP meetings.

CASP works as follows. Protein crystallographers and NMR spectroscopists are solicited for proteins whose structures are likely to be completed before the next CASP meeting. The sequences of these target proteins are made available on a web server. Researchers interested in taking part in the CASP experiment then submit up to five predictions for each target before a given deadline. Assessors chosen by the organizers critically analyze these predictions. The results are presented at Asilomar and papers by the three assessors and the 18 groups chosen to speak are published in a special issue of *Proteins: Structure, Function & Genetics.* The first meeting (CASPl, see ref. 1) took place in 1994, assessing 135 predictions made by 35 groups for 33 different protein targets. The second meeting (CASP2, see ref. 2), held in 1996, was marked by a dramatic increase in the interest of the community (947 predictions made by 76 groups for 42 targets). Recently, some 250 anxious predictors gathered again in Asilomar for the CASP3 meeting, organized by John Moult, Tim Hubbard, (Sanger Center, UK), Krzysztof Fidelis, (Lawrence Livermore, USA) and Jan Pedersen, (Acadia Pharmaceuticals, Denmark). This time, 3,807 predictions, including 1,256 three-dimensional models, were made by 98 groups for 43 targets . Experimental structures for 36 of the 43 targets were solved in time for the meeting. Here we summarize what was revealed during the four days of CASP3. Complete information can be found at the Prediction Center web site (http://predictioncentetllnLgov/casp3).

CASP considers three categories of structure prediction: (i) comparative modeling using a known template structure; (ii) fold recognition using a library of known protein folds; and (iii) *ab initio* prediction using principles of atomic interactions and protein architecture. Previously, each target was initially assigned to a prediction category but at CASP3, categories were defined at the meeting. Such categorization is somewhat arbitrary and it is fortunate that the 36 solved structures spanned the full range of difficulty and challenged all categories of predictors (Table 1).

Comparative modeling

Alwyn Jones (Uppsala, Sweden), a pioneer of interactive modeling software and an experienced X-ray crystallographer, was assessor together with Gerard Kleywegt (Uppsala) for the comparative modeling category. The goal of comparative modeling is to build a structural model of a protein on the basis of close sequence similarity to a template protein of known structure. At least one group built a high quality model (Cα r.m.s. deviation < 2 A for a substantial fraction of the structure) for every target for which a homolog with a known structure could be detected by sequence comparison methods like FASTA or PSI-BLAST. This was even true for cases with sequence identity below 25%, where accurate modeling was possible for 60-75% of the structure (Table 1). Getting the alignment correct at such low levels of sequence identity is a difficult problem. It was solved in different ways by the speakers for the selected groups including: Sternberg (Imperial Cancer Research Fund, UK); Blundell (Biochemistry, Cambridge, UK); Fidelis (Lawrence Livermore); Yang (Honig Lab, Columbia University), Fischer (Ben Gurion University, Israel); and Dunbrack (Fox Chase Cancer Center) . These groups commonly made use of multiple sequence alignments, alignment of actual and predicted secondary structures and hand-adjustments.

Once a known structure has been identified as a homolog of the target, a model is built by copying backbone elements from this template. A wide variety of methods are used for loop building and side chain reconstruction. While it is hard to evaluate the quality of the loop building due to insufficient data, side chain modeling reached a high level of accuracy for buried side chains: from >80% correct (to within 30°) for T47, T58 and T60 ('T' stands for 'target'; the targets are numbered consecutively from CASP2) to -45% correct for T57, T68 and T70. Some of the consistently successful methods for placing side chains include backbone-dependent rotamer libraries (Dunbrack), segment matching followed by energy minimization (Levitt) and self-consistent mean field optimization (Sternberg). Methods for comparative modeling were criticized at earlier CASP meetings for their inability to provide a final model of the target closer to the experimental structure than the template structure from

which it is built. At CASP3, most models were not refined. While this led to better models, it skirts a central embarrassment of molecular mechanics, namely that energy minimization or molecular dynamics generally leads to a model that is less like the experimental structure. Keeping the template fixed works badly for proteins in the calmodulin family, which show a great deal of structural variability for very similar sequences (T74 and T76; Table 1).

Knowing the structure of a protein is useful for predicting, analyzing and designing its function, but it is expensive to determine experimentally the structure of every protein, Structural genomics focuses structure determination of a well-chosen subset of proteins that should put all other protein sequences within the range of comparative modeling. This was discussed at a recent meeting held in Avalon, New Jersey and reviewed in *Nature Structural* Biology by Andrej Sali[3]. The results of CASP3 are directly relevant to structural genomics: they confirm it is possible to build a reasonable model when a proper template can be identified by sequence methods. How many experimental structures should be determined in order to fulfill the objective of structural genomics? At Avalon, estimates of between 10,000 and 100,000 structures were given. The CASP3 results show that the best current methods can build a reasonable model (70% of residues to 2 Å Cα r.m.s. deviation) when the sequence similarity is significant (chances of a false match below 1%, FASTA or PSI-BLAST E-value < 0.01). This could lower the previous estimates for the number of template structures needed to cover sequence space.

Fold recognition

Alexey Murzin (MRC, Cambridge, UK), who is recognized for his unparalleled understanding of protein structure, was the assessor for the fold recognition category. Because both comparative modeling and fold recognition aim to find a template structure on which to model the target protein, the categories overlap. At CASP3, the distinction was clear as sequence comparison methods could not confidently identify the correct template for any of the fold-recognition targets. Fold recognition methods, also referred to as threading or three-dimensional profile matching, use structure to assess the compatibility of the target sequence with each member of a library of known protein folds. These methods had limited success at both CASPl and CASP2. Among all predictors in this category at CASP2, the team formed by Alexey Murzin and Alex Bateman (MRC, Cambridge) did best. This was a surprise as they relied much more on careful analysis of sequences, a unique knowledge of protein structure and study of the functional literature, than on massive computation.

At CASP3, 17 targets could be assigned to the fold-recognition category in that they were structurally similar to other proteins defined as superfamilies or folds in the scop database[4]. The correct folds for 13 of these targets were recognized by at least one predictor. This is impressive as none of the 17 folds could have been recognized by the best standard sequence comparison methods, such as FASTA or PSI-BLAST (Table 1). Three of the speakers used their programming systems: Procyon (Sippl, Salzburg, Austria), Threader (Jones, Warwick, UK), and the NCBI threading program (Bryant, NCBI, USA). Human expertise plays a major role in some of these highly computerized approaches. This was clearly demonstrated by the success of the UNAGI team, which consists of four independent predictors (Ota, Kawabata, Kinjo and Nishikawa, National Institute of Genetics, Japan) who exchanged information about the target proteins, compared results of their own threading programs, and agreed on submissions. Other groups who did well included Koretke, Russel, Lupas and Copley (Smith-Kline Beecham) and Karplus *et al.* (UC Santa Cruz, USA), who used sequence-based hidden Markov models.

Fold recognition is the first step in defining a model for the target protein. For all proteins in the fold recognition category, main chain models were generated using homology modeling based on the selected template. The easiest model in this category (T81, Table 1) is as good as the most difficult model in the comparative modeling category (T74). The four most difficult targets gave much less accurate models (~60 residues to 6 Å Cα r.m.s. deviation). For the other seven fold-recognition targets judged by Murzin to be homologous to a protein of known structure, the models were more accurate (between 64% modeled to 2 Å and 44% modeled to 4.8 A).

Ab initio prediction

Christine Orengo (University College, London), a pioneer in automatic structure comparison and classification, was assessor for the *ab initio* prediction category. The goal here is to build a model for the target protein without using a specific template structure. There was considerable overlap between the *ab initio* and fold recognition categories and Orengo assessed 15 targets. In fact, the best *ab initio* models were clearly better than any fold-recognition models for three targets with folds already represented in the database (T61, T67 and T77).

Four groups selected to speak used the predicted secondary structure in their *ab initio* predictions. (i) Baker *et al.* (University of Washington, Seattle) assembled fragments of known structures chosen to match both sequence and secondary structure and then minimized a knowledge-based energy function by simulated annealing with moves that change fragments. (ii) Samudrala *et al.* (Levitt Lab, Stanford) generated all possible folds of the chain using a simplified three-state lattice model. The set of structures was pruned by repeated applications of knowledge-based scoring functions, proceeding from a simple point residue representation to an all-atom model with the predicted secondary structure. (iii) Skolnick *et* al. (Scripps Institute, San Diego) derived a list of constraints from secondary structure prediction and multiple sequence alignment, and then optimized the conformation by Monte Carlo moves on a simplified lattice. (iv) Lomize *et* al. (University of Michigan) used a thermodynamic model for secondary structure prediction followed by manual docking of pre-formed helices and strands and minimization using distance restraints.

Two groups selected to speak made less or no use of information derived from protein structures. (i) Scheraga *et al.* (Cornell University, Ithaca) followed the route he pioneered over 35 years ago: minimizing a physical potential energy function with a simplified representation of the protein to keep the computation tractable. (ii) Osguthorpe (University of Bath, UK) also used a simplified protein model and minimized a physical energy function, but included some secondary structure propensities for turns.

Overall, the quality of the models for the five easiest *ab initio* targets was as good as that of the four most difficult fold-recognition targets, with 60 residues predicted to 6 Å (Table 1). The best model for T80 is so poor that Murzin concluded that no group had correctly recognized the fold. A similar conclusion could be reached about the three *ab initio* targets modeled least well (like T80, the targets T52, T59 and T75 all modeled less than 55 residues to 6 A). Nevertheless, several

Table 1 Best results with CASP3 targets in order of difficulty

Target code	Number of residues	Percent identity	Number residues modeled	% residues modeled[1]	Cα r.m.s. deviation (Å)	scop classification[2]
Comparative modeling						
T473	158	**66**	158	**100**	1.3	Known family
T73	47	**63**	47	**100**	2.0	Known family
T58	225	**58**	225	**100**	1.5	Known family
T60	115	**33**	115	**100**	1.5	Known family
T76	140	**35**	69	49	1.5	Known family
T82	190	**34**	163	**86**	2.0	Known family
T48	116	**30**	99	**84**	**1.4**	Known family
T64	100	**30**	55	**55**	1.9	Known family
T49	375	**30**	240	**64**	2.0	Known family
T55	125	**29**	125	**100**	2.4	Known family
T57	340	**26**	207	**61**	2.0	Known family
T68	375	**25**	278	**74**	2.0	Known family
T70	330	**17**	241	**73**	2.0	Known family
T74	95	**15**	57	**60**	2.0	Known family
Fold recognition						
T81	150	-[4]	96	**64**	**2.0**	Known superfamily
T54	200	–	**76**	**38**	**2.0**	Known superfamily
T83	155		**57**	**37**	**2.0**	Known superfamily
T53	255	–	**174**	**66**	**4.0**	Known superfamily
T44	335	–	**102**	**31**	**4.0**	Known superfamily
T85	210		**55**	**26**	**4.0**	Known superfamily
T63	135	-	**60**	**44**	**4.8**	Known superfamily
T46	115	–	**97**	**84**	**6.0**	Known fold
T43	155	–	**65**	**42**	**6.0**	Known fold
T79	115	–	**62**	**54**	**6.0**	Known superfamily
T80	200	–	**54**	**27**	**6.0**	Known superfamily
Ab initio						
T77	**100**	–	**63**	**63**	**4.0**	Known fold
T61	**75**		**57**	**76**	**4.0**	Known fold
T56	**110**	-	**67**	**61**	**6.0**	New fold
T71	**235**	-	**63**	**27**	**6.0**	Known fold
T67	**185**		**61**	**33**	**6.0**	Known fold
T75	**85**	–	**48**	**57**	**6.0**	Known fold
T59	**70**	–	**45**	**64**	**6.0**	Known fold
T52	**100**		**40**	**40**	**6.0**	New fold

[1]The number of residues modeled and Cα root mean square (r.m.s.) deviation of the best model of for the particular target are taken from the r.m.s. deviation versus coverage plots provided online by Hubbard. These r.m.s. deviation values depend almost linearly on the number of residues modeled and we have chosen to set the r.m.s. value to 2 A, 4 Å or 6 Å and to tabulate the corresponding number of modeled residues.
[2]For definitions of scop nomenclature see ref. 4.
[3]More details of the target protein including the name of the experimental group who kindly made the coordinates available is available at the CASP3 web site given above. 'T' stands for 'target'; the targets are numbered consecutively form CASP2.
[4]If no significant sequence match (E-value <0.01) can be found by either FASTA or PSI-BLAST, the percent identity is marked "-". For T79 and T85, the correct template was found by PSI-BLAST at an E-value of 0.1 and for T63 at an E-value of 8.3..

groups did show sustained performance on several targets using the different methods described above. While a 6 Å Cα r.m.s. deviation model is unlikely to be useful for experiments or functional studies, this advance in *ab initio* prediction was one of the highlights of CASP3.

Secondary structure prediction

Knowing what parts of the sequence adopt a-helix or P-strand secondary structure is an important part of almost all protein structure prediction methods. Because of its success at CASP2, most predictors used PHD, Burkhard Rost's secondary structure prediction program based on neural networks and multiple sequence alignment. Several groups have implemented variants of PHD. At CASP3, David Jones (Warwick) did best with 77% of residues correctly predicted versus 70% for PHD; the corresponding numbers for the subset of difficult targets are 73% and 67%, respectively. Jones' method trains the neural net on PSI-BLAST profiles of over 1,800 known structures.

The lessons of CASP

Assessing predictions is a challenging task for which no automated scheme exists. All three assessors at CASP3 exercised exceptional judgment and also used evaluation methods developed at earlier CASP meetings. Two very useful evaluation tools were applied to all CASP3 results: the r.m.s. deviation versus coverage plots made by Hubbard and the ProSub tables made by Sippl et al. These help to rank the predictions for a particular target. Assessing sustained performance is much

more difficult. Each assessor struggled with the question of whether it is better to predict more targets moderately well or a few targets exceptionally well. In the context of the Olympic Games, what is better: six silver medals or three gold medals? Also, should 'difficult' targets score more? The format of the meeting gives 18 groups the opportunity to speak and write about their work. However, a number of other groups also did well, including: Godzik (La Jolla Institute, San Diego); Uberbacher et al. (Oak Ridge national Laboratory, USA); and Sjolander and Thomas, (Molecular Applications Group, USA).

The best predicted models have been characterized in terms of Cα r.m.s. deviation values of 2.4 and 6 A. A 2 Å model is likely to be useful for functional studies, but the key residues could be less well determined if they are in loops. A 4 Å model will indicate what residues are on the same side of the molecule and could help in planning experiments. A 6 Å model is a very rough hint at the general chain path but is not likely to be useful for functional inference or experimental work.

Although CASP provides an unbiased assessment of the ability of the best practitioners in the field to predict protein structure, it does not test how well other scientists can expect to do, nor how well totally automated methods would do. Fischer (Ben Gurion University) has proposed CAFASP, critical assessment of fully automatic structure prediction, to test automated prediction methods on Web servers provide by the participants. Such automation is particularly important as we enter the post-genomic era and need to assign structure and function to the huge number of novel genes.

Some of the progress we have reported here has resulted from the increased number of protein sequences and structures as well as the greater availability of computer resources. Nevertheless, it is clear that new and improved techniques did contribute to the significant progress at CASP3. Protein structure prediction remains a very difficult problem and much remains to be done. The CASP meetings provide more than a framework for unbiased assessment. As most research groups in the field are active participants in CASP, there is an unprecedented level of enthusiasm and information exchange that will surely catalyze future advances.

Patrice Koehl and Michael Levitt are in the Department of Structural Biology, Stanford Medical School, Stanford, California 94305, USA, email: koehl@zen.stanford.edu or levitt@zen.stanford.edu.

* **Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP3). Asilomar Conference Center. December 13-17.1998.**

1. **Shortle, D. Nature Struct. Biol. 2. 91-93 (1995)**
2. **Eisenbera. D. Nature Struct. *Biol.* 4.95-97 (1997)**
3. **Sali A. Nature Struct. Biol. 5, 1029-1032 (1998)**
4. **Murzin, A.G., Brenner, S.F., Hubbard, T. & Chothia, C. J. Mol. *Biol. 297,* 536-540 (1995). *http://scop.mrc-lmb.cam.ac.uk/scop***