Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round 6

John Moult,[1]* Krzysztof Fidelis,[2] Burkhard Rost,[4] Tim Hubbard,[5] and Anna Tramontano[3]
[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California*
[3]*Istituto Pasteur–Fondazione Cenci Bolognetti, University of Rome "La Sapienza," Rome, Italy*
[4]*CUBIC, Columbia University, Department of Biochemistry and Molecular Biophysics, New York, New York*
[5]*Sanger Institute, Wellcome Trust Genome Campus, Cambridgeshire, United Kingdom*

ABSTRACT **This article is an introduction to the special issue of the journal *Proteins*, dedicated to the sixth CASP experiment to assess the state of the art in protein structure prediction. The article describes the conduct of the experiment and the categories of prediction included, and outlines the evaluation and assessment procedures. A brief summary of progress over the decade of CASP experiments is also provided. Proteins 2005;Suppl 7:3–7.**

Key words: protein structure prediction; community-wide experiment; CASP

INTRODUCTION

This issue of *Proteins* is devoted to articles reporting the outcome of the sixth community-wide experiment to assess methods of protein structure prediction (CASP6), and related activities. There have been five previous CASP experiments, in 1994, 1996, 1998, 2000, and 2002, and these were reported in previous special issues of *Proteins*.[1–5] Other discussions of CASP6 have also appeared.[6–8]

The primary goals of CASP are to establish the capabilities and limitations of current methods of modeling protein structure from sequence, to determine where progress is being made, and to determine where the field is held back by specific bottlenecks. With a decade of CASP experiments now complete, bottlenecks and progress have become more important. Methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure.

This article outlines the structure and conduct of the experiment, and is followed by a description of the CASP6 target proteins. There are articles by the assessment teams in each of the primary prediction categories—Comparative Modeling, Fold Recognition, and New Fold Methods—followed by articles from some of the more successful prediction teams. Two articles describing results using automated structure prediction servers are included, and this continues to be an important area. Whereas CASP "human" predictions may be made with any combination of computational and human methods, the server section captures predictions directly from fully automatic servers. There are then a series of articles describing more specialized prediction areas. For the second time, prediction of disordered regions was included in CASP. A number of experimental studies have established that not all proteins have a single, ordered, three dimensional (3D) structure.[9] Thus, the ability to predict disorder is of considerable importance. One article describes evaluation in that area, and another reports results by one of the prediction teams. There were two new components in this CASP. One is prediction of domain boundaries, crucial to the modeling of large structures, and there is an article describing the evaluation in that area. The second new area is the prediction of the function of proteins, and there is an article describing that experiment. Mostly because of structural genomics, there are now a substantial number of experimentally determined protein structures with no or incomplete characterization of molecular function. There is also an article describing results of predicting 3D contacts. The final article is the latest in a series assessing progress over the course of the CASP experiments. With a decade of results now accumulated, discovering where progress has been made and also where there are bottlenecks to progress has become increasingly worthwhile. The assessors' articles are probably the most important in the whole issue, and describe the state of the art as they found it in CASP6.

THE CASP6 EXPERIMENT

The structure of the experiment was very similar to that of the earlier ones and consisted of three steps:

Grant sponsor: National Library of Medicine; Grant number: LM07085 (to the Livermore Prediction Center). Grant sponsor: U.S. Department of Energy; Grant number: W-7405-Eng-48 (for work performed by the University of California, Lawrence Livermore National Laboratory). Grant sponsor: National Institutes of Health, Institute of General Medical Sciences; Grant number: GM072354 (providing meeting support). Grant sponsor: BioSapiens Network of Excellence (funded by the European Commission FP6 Programme); Grant number: LHSG-CT-203-503265. Grant sponsor: European Molecular Biology Organization (EMBO). Grant sponsor: Istituto Pasteur–Fondazione Cenci Bolognetti.

*Correspondence to: John Moult, University of Maryland Biotechnology Institute, Center for Advanced Research in Biotechnology, 9700 Gudelsky Drive, Rockville, MD 20850. E-mail: moult@umbi.umd.edu

Received 29 June 2005; Accepted 29 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20716

The article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

1. Information about "soon to be solved" structures was collected from the experimental community and passed on to the prediction community. Target information was made available through the CASP website and sent directly to registered servers.
2. Prediction teams deposited models of the structures before the experimental results were public. For human prediction teams, deposition was required by a specified deadline. Servers were required to respond within 48 hours.
3. The models were compared with the experiment, using numerical evaluation techniques and human assessment, and a meeting was held to discuss the significance of the results.

MANAGEMENT AND ORGANIZATION

CASP has a multilevel structure, intended to ensure substantial input from the prediction community:

A. Organizers. The authors of this article, responsible for all aspects of the organization of the experiment and meeting.
B. Consultancy groups. Three groups of approximately 10 veteran CASP predictors each, one for each of the three primary prediction categories. These groups, first introduced in CASP3, are involved in the selection of the independent assessors, are influential in the choice of numerical evaluation methods, and provide advice on other aspects of the experiment.
C. Predictors' meeting at Asilomar. During each CASP conference, there is a predictors' meeting with votes on issues of CASP policy, particularly the timing of the next experiment, the organization team for the next experiment, and major changes and extensions of the CASP process.
D. Independent assessors. The independent assessors have primary responsibility for judging the quality of the predictions received, and commenting on the current state of the art. Assessors are provided with numerical analysis data generated using approved methods, and may also add their own numerical methods.
E. Protein Structure Prediction Center at Lawrence Livermore Laboratory. The prediction center is responsible for all data management aspects of the experiment, including the distribution of target information, collection of predictions, generation of numerical evaluation of predictions, collection of numerical evaluations from other workers, and maintenance of a website where all data are available. Details of these aspects of the experiment are described by Krystafovych et al.[10]
F. The FORCASP website (www.forcasp.org). FORCASP provides a forum where members of the prediction community may discuss aspects of the CASP experiment.

COLLECTION OF TARGETS

X-ray crystallographers and NMR spectroscopists were solicited to provide information about structures that were either expected to be solved shortly or that were already solved but had not yet been discussed in public. Structural genomics projects were also asked to contribute prediction targets, and substantially more than half of the targets came from that source. Target information was made available to predictors through a web interface. Details of 87 structures were obtained. Information on 11 of these targets was released prematurely, causing them to be cancelled, and information on a further 12 targets was not available in time for assessment, so that a total of 64 targets, divided into 90 domains, formed the set included in the experiment. Significant attrition through premature release is a new problem in CASP, and is largely due to the high throughput of structural genomics centers. Because of this, and the structures not available in time, about a quarter of the predictors' work could not be included in the results. Nevertheless, a total of 90 domain-level targets is close to our longtime goal of 100, and more data were assessed than in previous CASPs.

CATEGORIES OF STRUCTURE PREDICTION

The quality of a structure model depends on how much information from already known structures can be used: At one extreme, models competitive with the experiment can be produced for proteins with sequences very similar to that of a known structure. At the other, models for proteins with no detectable sequence or structure relationship to one of known structure are still at best approximate. In all the CASPs so far, targets have been divided into three broad categories, reflecting how extensively models could be based on knowledge of other structures. This system was partly revised in CASP6, with a changed meeting format (see below). However, the traditional divisions were maintained in that there were again three assessment teams, one for each of the usual categories. The three categories are as follows:

1. Comparative or Homology Modeling

When the sequence of the target structure is clearly related to that of one or more structures, the structures will also be similar. Thus, an approximate model can be created simply by copying related regions of polypeptide from the parent structure or structures and changing the side-chains where necessary. There were a total of 46 target domains considered by the assessors to be in the comparative modeling category. These domains were divided into two finer categories: the 27 that could be related to known structures using a simple BLAST search [high sequence identity: an *E*-score of 0.01 or better against a Protein Data Bank (PDB) library], and the 19 where a relationship to a known structure could be identified using moderately sophisticated PSI-BLAST searches (low sequence identity: *E*-score 0.01 or better, using a Swiss-Prot/TREMBL profile against the set of PDB sequences). Some of the models for the high sequence identity set were analyzed in more detail than the rest, considering the accuracy of side-chains, the construction of regions not present in available template structures, and whether the overall backbone accuracy is higher than that obtained by simply copying the best template.

2. Fold Recognition (FR)

Increasingly, new structures deposited in the PDB turn out to have folds that have been seen before, even though conventional sequence searches with BLAST and PSI-BLAST fail to find the relationship. With increasing diversity and power of sequence search methods, and the emergence of effective hybrid sequence–structure approaches, the division between this category and comparative modeling has become arbitrary. Nevertheless, we maintained this distinction in CASP6 assessment. Targets were assigned to this category if the target structure was found to be similar to one or more already in the PDB and did not meet the criteria for comparative modeling.

Targets in this category are subdivided into those that are considered to have diverged from a common ancestor of known structure–homologous folds (FR/H), and those that are considered more likely to resemble known structures as a result of convergent evolution–analogous folds (FR/A). FR/H domains are those where a search of the sequence profile of the target against profiles of all PDB entries delivers a significant hit,[11] or there is evidence of a functional relationship between the target and related structures. For FR/H targets, evaluation of the quality of the models has common components with comparative modeling, specifically, alignment accuracy. In recent CASPs, template-free modeling methods have been very competitive with fold recognition for FR/A targets, and so these were also considered by the New Fold assessor. There are 23 FR/H targets and 14 FR/A targets.

3. New Fold Methods

In early CASPs, targets where there was no relationship to an already known complete structure were described as "ab initio." The name implies that there is no reliance on known structures in building models. In practice, most of the methods used do make extensive use of available structural information, both in devising scoring functions to distinguish between correct and incorrect predictions, and in choosing fragments to incorporate in the model. For this reason, the category was renamed, starting in CASP4. A wide range of knowledge-based techniques are used: well-established secondary structure prediction tools; sequence-based identification of sets of possible conformations for short fragments of chain; methods that assemble 3D folds from candidate fragments and predicted secondary structure; prediction of which residues are in contact in the structure; "minithreading" methods that identify supersecondary structure motifs; and full-domain fold recognition methods that may establish an approximate or partial topology. These approaches are sometimes combined with numerical search methods such as molecular dynamics, Monte Carlo, and genetic algorithms. There are a few "pure" ab initio methods, usually based on some form of numerical simulation techniques together with more traditional empirical potentials.

Important evaluation criteria in the new fold category are the fraction of the structure that is predicted below a specified error level, and recognition of success in identifying general architecture. As noted above, the same methods generally work best for FR/A targets as well. Ten "New Fold" targets were evaluated together with the 14 FR/A ones.

LEVEL OF PARTICIPATION

A high level of participation from the prediction community is critical to the success of the experiment. As usual, participation was solicited through announcements in published articles and news groups, a website, and direct approaches to known prediction groups. Overall participation has steadily increased over the CASPs from 34 groups in CASP1, then 70, 163, 98, 216, and in CASP6, 266. Figures for the last four CASPs are a sum of human and server groups, and include some overlap.

COLLECTING AND VALIDATING PREDICTIONS

There were a total of 41,283 models deposited in CASP6, of which 32,703 could be assessed. Of the assessed ones, 23,119 are 3D coordinate sets. A further 4484 are alignments that are converted into coordinates for assessment. The remainder are residue–residue contacts (1397), domain assignments (1293), disorder predictions (1769), and function predictions (990). As before, all predictions were required to be in a machine readable format. All submissions were processed by the Prediction Center at the Lawrence Livermore Laboratory.[10] Accepted submissions were issued an accession number that served as the record that a prediction had been made by a particular group on a particular target. Human predictions were submitted through the Web interface, or by e-mail. A final acceptance time was established for predictions on each target, determined by the expected release date of the experimental structure, or other factors. Target queries were sent to servers directly from the CASP distribution server, and the returned models were immediately processed by the CASP verification software. Servers had 48 hours in which to respond. In previous CASPs, server predictions were collected as part of the parallel CAFASP experiment.[12] Because of procedural differences, that was not possible at CASP6. Nevertheless, we are grateful to the organizers of CAFASP for establishing the principles by which the system works. The prediction season ran from June through early September. As in previous experiments, predictors were limited to a maximum of five models per target, and were instructed that most emphasis would be placed on the model they designated as the best (referred to as "Model 1").

NUMERICAL EVALUATION OF PREDICTIONS

CASP evaluation is based on comparison of each model with the corresponding experimental structure. Numerical evaluation criteria have been moderately stable for the last few CASPs. In CASP6, the GDT_TS[13] (Global Distance Test Total Score) measure has again been used by all three assessors as the principal metric of main-chain accuracy. In comparative modeling, alignment accuracy is also of primary importance, and for the high sequence identity targets, side-chain accuracy, and the accuracy of "loop" region main-chain were also considered. In fold

recognition, the assessors once more experimented with a number of additional measures but found that GDT_TS was a reasonable consensus. There, sequence alignment accuracy was also an important measure. The "New Fold" assessors found that GDT_TS was the best single measure, but GDT_TS rankings of accuracy were sometimes modified by visual inspection. The new fold assessors in the previous two CASPs reached similar conclusions, but so far, no better measure has emerged. In general, numerical evaluation of model structures remains an imperfect science.

ASSESSMENT

All CASP experiments have placed the primary responsibility for assessing the significance of the results in the hands of independent assessors. The CASP6 assessors were Alfonso Valencia, assisted by Michael Tress, for comparative modeling; Roland Dunbrack, assisted by Guoli Wang and Yumi Jin, for fold recognition; and B. K. Lee, assisted by Bangalore Sathyanarayana and Chin-Hsien Tai, for the new fold category. The articles by the assessment teams in the special issue constitute the most thorough and authoritative analysis available. As usual, the identities of the prediction teams were not known to assessors until they had completed an analysis and ranking of the results. Alfonso Valencia and Michael Tress also assessed residue–residue contact predictions; B. K. Lee considered domain predictions; Roland Dunbrack evaluated disorder; and Anna Tramontano, function predictions.

STATISTICAL SIGNIFICANCE OF RESULTS

The primary goal of the CASP experiments is to assess the state of the art in protein structure prediction. In general, with a large number of prediction teams taking part, and an increased number of prediction targets, the results do provide a sound basis for drawing conclusions concerning the accuracy of models in particular prediction categories, and for determining where significant progress has or has not been made. And also, in general, there are enough data to indicate which prediction teams are producing the most accurate models in each category. However, there are not enough data to reliably rank closely performing predictors. Although this is not an objective of CASP, understandably, predictors are very sensitive to any perceived misranking. Over the CASP experiments, assessors have become more aware of this issue, and once more, in CASP6, all three teams have taken considerable care to evaluate the reliability of rankings, and to address this issue in their articles, and in the choice of prediction groups invited to submit articles to the special issue.

MEETINGS, WEBSITE, AND PUBLICATIONS

Following the closing of the prediction season, two planning meetings involving the assessment teams and the organizers were held, one before any assessment of the predictions, and the other when a full assessment was complete. The first of these meetings was also attended by several assessors from earlier CASPs, and the primary aim was to provide guidance to the CASP6 assessors. At the second meeting, the assessors presented the results of their work, including a full ranking of prediction teams, and these were extensively discussed. Only then were the names of the prediction teams made known to the assessors.

The meeting to discuss the outcome of the experiment was held at a hotel in Gaeta, Italy, a change from the customary venue of Asilomar, California. The format of the meeting was changed from previous that of CASPs. At previous meetings, one day was primarily addressed to each of the modeling categories of comparative modeling, fold recognition, and new fold methods. At the CASP6 meeting, one day was devoted to prediction results in the comparative modeling and homologous fold recognition categories, that is, all classes of homology-based prediction, and one day was devoted to analogous fold recognition and new fold predictions, primarily template-free modeling. This regrouping better reflects the divisions in methodology. The organizers and a number of participants in CASP have felt that we should do more to encourage methods development, so the third day emphasized modeling methods. The final half-day of the meeting dealt with the three auxiliary prediction areas—domains, disorder, and function. The assessors selected prediction teams to talk at the meeting, and to write prediction reports, based on their judgment of who had done the most significant work. Both at the meeting and in the articles, participants have been urged to concentrate on what went right, what went wrong, and where possible, to explain why, and what they learned as a result. Because of space limitations, details of the methods are often absent, and readers are requested to turn to the references for more information. All the prediction and assessment papers in this issue have been peer reviewed. The CASP website (http://predictioncenter.org) provides extensive details of the targets, the predictions, and the numerical analyses. Discussions of a number of issues can also be found on the FORCASP site (www.forcasp.org). Many possible views may be taken of the results, and the interested reader is encouraged to consult other sources for alternative points of view.

PROGRESS OVER THE CASPS

How much progress has been made over the decade of CASP experiments? The final article in this issue[14] addresses that in some detail. Overall, the picture is different in the different categories of prediction. Least progress has been made in comparative modeling from relatively high sequence identity templates. Here, there has been little detectable improvement since CASP2. There is general agreement that progress requires the development of appropriate refinement techniques and potentials, capable of making adjustments on an atomic scale. This problem is now receiving considerable attention in the CASP community, and there were some encouraging signs of progress in CASP6. Hopefully, CASP7 may see a breakthrough in this area. In contrast to that, there is steady but modest progress in difficult comparative modeling and homolo-

gous fold recognition, in terms of the extent of sequence dependent superposition between model and target (as measured by GTD_TS), and in alignment accuracy. Scores for both measures have approximately doubled from CASP1 to CASP6, and another decade of this level of progress would result in excellent models. However, as discussed in the progress article, different sorts of problems may have to be confronted for this further progress to be made. The most dramatic advances have been made in the "new fold" or template-free modeling regimen. Here, at CASP1, there was little sign of any relationship to experiment in any of the models. The situation has steadily improved, and since CASP4, a few small targets have had at least one topologically pleasing and occasionally quite accurate models. Progress has perhaps slowed in the last two CASPs, though visual inspection suggests there was progress this time for small targets. There is also evidence of more sustained performance by more prediction groups. The quality of models for large targets remains generally very poor. One of the difficulties in this area is identifying domains in a structure. Assessment of domain identification in CASP6[15] shows that this is far from a solved problem.

The quality of predictions from automatic servers has also improved steadily. Humans still do better, but this may be an unfair comparison, since it is usual to start human modeling from the best available server output.

CASP CHALLENGES

As discussed earlier, CASP is adjusting its format to put more emphasis on methods development, and particularly methods that will remove some of the bottlenecks to further progress. At the CASP6 community meeting in Gaeta, it was agreed to work on four challenges in the next 2 years, with a review of results at the CASP7 meeting. These challenges are as follows: First, modeling the structure of single-residue mutants. The challenge here is correct modeling of individual or groups of side-chains, and focus on this problem should help develop refinement methods. Second, modeling structure changes associated with specificity changes within protein families. This is a very difficult but important task. A central goal of experimental structural genomics is to provide representative structures for as many families as possible. Making optimum use of these structures will require modeling the structural differences that determine the different specificities. This is a second challenge requiring the development of refinement methods. The third challenge is directly focused on refinement, and is to produce a 0.5 Å root-mean-square deviation (RMSD) improvement in the Cα accuracy of models based on 30% or higher sequence identity. The final challenge addresses a bottleneck in the new fold area, that is, to devise scoring functions that will reliably pick the most accurate models from a set of candidate structures produced by current new fold methods.

FUTURE DEVELOPMENTS

There will be a CASP7 experiment, running from Spring 2006, and culminating in a meeting in December of that year. The meeting is planned to take place in Asilomar this time. There will be some changes of format, helping to strengthen emphasis on methods, and focusing on those predictions that represent real progress rather than just being among the best. Those interested should check the CASP website for further announcements.

ACKNOWLEDGMENTS

As always, this CASP experiment depended on the cooperation, hard work, and support of a large number of people. We are grateful to the members of the experimental community, particularly the structural genomics centers, who agreed to provide targets. Taking part required courage and hard work on the part of all the predictors. Without the careful and critical assessment of the results by the assessment teams, the experiment would have failed. We thank the editor of this journal, Ed Lattman, for once again providing a mechanism for peer reviewed publication of the results. Many thanks to Andriy Kryshtafovych, who was in charge of data management at Livermore in CASP6, and to Volker Eyrich, who handled the collection and first stage processing of server predictions. We are also grateful to Italtech Solutions and to IBM for technical support.

REFERENCES

1. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii–v.
2. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins 1997;29(Suppl 1):2–6.
3. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;37(Suppl 3):2–6.
4. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;45(Suppl 5):2–7.
5. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003;53(Suppl 6):334–339.
6. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
7. Rychlewski L, Fischer D. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 2005;14:240–245.
8. Cozzetto D, Di Matteo A, Tramontano A. Ten years of predictions… and counting. FEBS Lett 2005;272:881–882.
9. Dyson HJ, Wright PE. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. FEBS Journal 2002;62:311–340.
10. Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P, Fidelis K. CASP6 data processing and automatic evaluation at the Protein Structure Prediction Center. Proteins 2005;61(Suppl 7):19–23.
11. Wang G, Dunbrack RL Jr. Scoring profile-to-profile sequence alignments. Protein Sci 2004;13:1612–1626.
12. Bourne PE. CASP and CAFASP experiments and their findings. Methods Biochem Anal 2003;44:501–507.
13. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;45(Suppl 5):13–21.
14. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61(Suppl 7):225–236.
15. Tai CH, Lee WJ, Vincent JJ, Lee BK. Evaluation of domain prediction in CASP6. Proteins 2005;61(Suppl 7):183–192.