
CASP AND CAFASP EXPERIMENTS AND THEIR FINDINGS●

ED1

Philip E. Bourne

The prediction of the three-dimensional (3D) structure of a protein from its one-dimensional (1D) protein sequence is a much published and debated area of structural bioinformatics. This prediction involves the kind of fold that the given amino acid sequence may adopt; in other words, whether it takes a *new fold* or one of the existing folds? If the sequence takes one of the existing folds, which is the most suitable fold among the known folds (*fold recognition*). When fold recognition is apparent because of good sequence similarity to one of the known structures, then the question is how best can one model the structure of the given sequence, taking the relevant information from existing homologous structures in the protein data bank (*comparative modeling*). Alternatively, what if none of this information is available and the structure is modeled from first principles (*ab initio*)?

What makes structure prediction (at least to this author's knowledge) unique among scientific endeavors is the manner in which progress in the field is measured. The Critical Assessment of Structure Prediction (CASP) and the Critical Assessment of Fully Automated Structure Prediction (CAFASP) experiments provide a measure of this progress by attempting to measure in a quantitative way the success of many groups on a predefined set of structures. Beyond a measure of progress, as is true of all good experiments, they● suggest new ways of addressing the problem, and influence the results presented as subsequent CASPs and CAFASPs.

Q1

The approach is for an independent group to solicit protein targets for use in CASP and CAFASP months in advance on their availability. The targets are NMR and X-ray protein structures comprising one or more domains either determined and not published or anticipated to be determined in time for review and to provide the sequence of those targets to the groups competing in CASP and CAFASP. These groups then make a

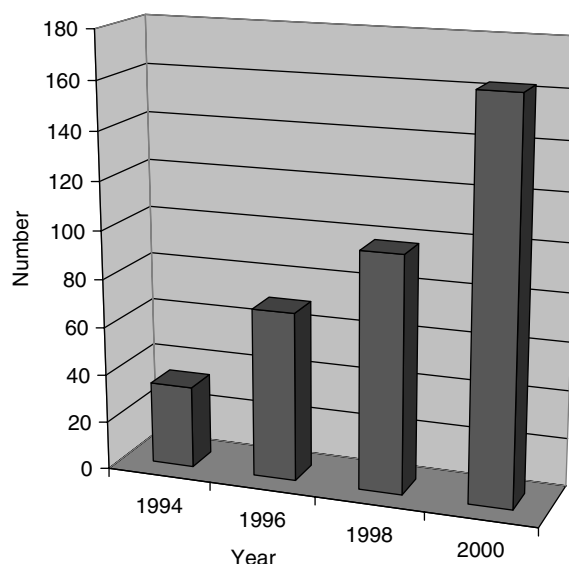
Structural Bioinformatics

Edited by Philip E. Bourne and Helge Weissig

ISBN 0-471-20199-5 Copyright © 2003 by Wiley-Liss, Inc.

series of blind predictions of the 3D structure based on the protein sequence and submit those results to a server for independent and comparative review. CAFASP predictions are collected from prediction servers registered with CAFASP in an automated manner within a very short (48 hours) period following the release of the target and may also be used for subsequent CASP experiments by some groups (all CAFASP predictions are available from the CAFASP site). CASP predictions take longer and the results of both are ranked by a number of criteria that depend on the rules decided by the official CASP organizers and then members of the groups gather at the biannual Asilomar meeting to review their results and hear from members of the groups who did best or came up with a significantly new methodology and those who assessed the different classes of predictions.

I do not know another scientific endeavor that is so open and, from a predictor's perspective, so blatantly competitive. I see it as a testament to the character of those scientists in the field of structural bioinformatics who are willing to share their formative ideas and, data and have them all openly reviewed. Having a laboratory that has competed in a minor way in two past CASPs I can tell you that it becomes a compelling exercise and, when arriving at Asilomar and getting the proceedings containing the results, many different emotions—surprise, joy, disappointment—all bubble to the surface. My laboratory is not alone in participating in this endeavor. Figure 24.1 indicates the number of groups participating in the four CASPs held since 1994. Clearly structure prediction is considered a compelling area of study by many, who are affectionally known as “CASPerS.” It is human nature to at least in some measure regard this as an individual or at least a team effort against stiff competition. However, to the organizers, assessors, and the U.S. federal agencies (National Institutes of Health, National Library of Medicine and Department of Energy) that fund these efforts, it provides a measure of global improvement and an indication of what specific areas of protein structure prediction need to be improved. Consider an



Q2

Figure 24.1. Number of groups participating in CASP in 1994, 1996, 1998, and 2000.

example. CASP4 reports that the protein's structure prediction community has been unable to significantly improve the ability to derive a model structure from a target recognized by comparative modeling that is significantly closer to the real structure than the target from which it was derived. This result● is an example of a bottleneck in our progress and will likely receive significant attention in CASP5. Other bottlenecks are outlined below.

Q3

Much has been written about CASP and CAFASP results. There is no intent here to repeat or even review what has already been written. Here I provide a synopsis of what CASP and CAFASP mean to the field of structural bioinformatics and as such serves as a forward to the three chapters that follow; chapters that review the three key areas of protein structure prediction: comparative (also known as homology) modeling, fold recognition (also known as threading) and ab initio, the latter having been reclassified as "new fold methods" in CASP4 in recognition of the fact that existing structure information is used in some way even by these methods.

The CASP experiments were run in 1994, 1996, 1998, and 2000, and CASP5 will be run in the summer of 2002●. CASPs 2–4 have been extensively documented in supplements to the journal *Proteins: Structure Function and Genetics* (CASP2 29(S1); CASP3 37(S3); CASP4 45(S5)). CAFASP arose in part through recognition of the emergence of high-throughput structure prediction techniques. Although a better prediction of a single structure might be made by an expert with reference to the literature, this method● takes time and does not scale to, for example, predicting structure of all genes or open reading frames in the genome of a higher organism. Fully automated procedures are practical for large-scale predictions and they too should be assessed, hence the emergence of CAFASP run independently for the same set of targets as available as CASP4. In 2002● CAFASP will be more fully integrated with CASP.

Q4

Q5

Q4

WHAT PROGRESS HAS BEEN MADE?

Progress over all CASPs is summarized by Venclovas et al. (2001). In summary, some areas have advanced and some have remained almost static. We explore this further below for each major methodology. At CASP3 the question was posed, "When will we be able to reliably predict a protein structure?" The answer returned at that time was when we have determined all structures experimentally or at least have a representative of each fold. A statement that in part led to the emergence of the structure genomics initiative (Chapter 29), which has as one objective the filling in of protein fold space so that comparative modeling can be more useful. As we see below, the situation is not quite that simple. Then, as if in some form of retort, the quality of new fold predictions improved in CASP4 with some of the contact predictions approaching a useful level of accuracy. Detection of homologous folds at lower levels of sequence identity has also improved. Nevertheless, the best models are still not good enough for defining function. In part this relates to the follow-on steps of homology modeling, which are discussed in detail in Chapter 25. For example, both comparative modeling and fold recognition alignment of the proposed structure with the template remain a problem, as well as subsequent improvement of the model with the template as a starting point.

Asking what progress has been made is a problem in itself, since some would say it depends on how you measure success or failure. The CASP organizers have been innovative in their evolving approach to this problem. Nevertheless, a quantitative answer is hard to come by. Consider a couple of crude examples. First, for a multidomain

Q6

protein, a group might do very well on predicting a single domain, but when measured over the whole protein the rmsd is poor. The assessors have attempted to address this result with a variety of measures relating to, for example, the percentage of correctly aligned residues, measurement by domain, correct positioning of biologically important residues, positions of side chains in the core and overall, and so forth. Second, a complete genome analysis done in batch mode identifies the structure of a protein by matching it to a template using comparative modeling. A group spends all summer on a single prediction of the same protein using manual techniques and all available biological knowledge. When superimposed the two models differ by approximately 1 Å rmsd overall. However, the “summer” group were able to detect features of the active site that led to some functional predictions of value to biologists. Meanwhile, the “high-throughput” group missed the functional prediction, but they made useful discoveries by modeling easier proteins located elsewhere in the genome. How do you measure which approach is better in such a context? The simple answer is that both approaches have merit and eventually each will contribute to progress in the other.

COMPARATIVE MODELING

As described in Chapter 25 comparative modeling can be used when there is a clear relationship between the sequence of a protein of unknown structure to that of a sequence of a known structure, most likely found in the Protein Data Bank (PDB; Chapter 9). The most recent discussion on the results of comparative modeling comes from the CASP4 experiment by Tramontano, Leplae, and Morea (2001) who undertook a detailed analysis of those predictions in this category. While they rightly took great pains to emphasize the difficulties in making assessments, they concluded:

Q7

Q8

- Overall little progress was made since CASP3.
- Alignment of target to template remains a problem and more importantly the quality of the alignment does not correlate well with the level of sequence identity between template and target even at levels of sequence identity approaching 50%. The best methods rarely achieve over an 80% correct alignment with sequence identities below 50%.
- On average biologically important regions are predicted better than the protein as a whole. However, this finding has more to do with the spatial conservation of key residues important to function than a testament of the methods applied. This assumes of course that the best template is chosen on which to model.
- Loop modeling remains a significant problem.
- Improvements could occur as the database of available targets continues to grow. The plus side is that a better template may be devised from multiple experimental structures; the negative side is that there is more opportunity to select a completely incorrect template. Thus, correct template(s) selection becomes a greater challenge as the databases of experimental structures increase.
- Some automated servers perform as well as individual efforts.
- Prediction of the relative orientation of domains relative to those seen in the templates remains elusive.

In summary, we have a way to go before comparative models prove consistently useful surrogates for experimental structures. At this time it would appear impossible

to consistently use such models in rational drug design experiments or mutagenesis experiments other than in active site regions, especially when the sequence identity to the template is low.

FOLD RECOGNITION

As described in Chapter 26, fold recognition techniques deal with finding relationships between sequence and structure that do exist, but are not immediately obvious, that is, a successful model will be proven to have structural similarity to a known fold, but no immediately obvious sequence similarity. Targets in this category generally fall within the twilight and midnight zones of sequence–structure relationships (Rost, 1999). Thus, fold recognition depends on advanced sequence comparison methods, comparisons of secondary structure, and the threading of sequences onto a variety of templates looking for a favorable hit. One measure of the popularity of the approach is that from CASP3 to CASP4 the number of predictions in this category rose from 3807 to 11,136. Conclusions from CASP4 in the fold recognition category are:

- Several groups submitted models that were much closer to the true structure than any of the existing templates within the PDB. But at the same time some of the same groups made completely incorrect predictions. Nevertheless, there was a qualitative assessment that the top scoring groups had made significant progress since CASP3.
- As is true for comparative modeling, prediction of multidomain proteins is more difficult than that for single-domain proteins.
- Predictions varied widely, with a large number of poor predictions. Several of the public servers performed better than more than one-half of the predicting groups. As the assessors pointed out, this result would seem to indicate some groups are less interested in their relative performance than the low probability that they will achieve a valuable prediction. Taking this inference further, there would seem to be a relatively small number of predictors with significant experience in both the process and the techniques for good prediction. Or even further, the best predictors use or have been able to provide their own methodologies, at least to some extent, in automated servers for the benefit of the whole community.

Q6

NOVEL FOLD RECOGNITION

This class of prediction was known in earlier CASPs as *ab initio* fold prediction, but was renamed in CASP4 to better define the current methodologies that are being applied, particularly, to separate methodologies that are using sequence homology from those that are not, by simply testing the latter ones on targets where no sequence homology is actually known. Now *ab initio* is reserved for those methods that rely only on physical principles and not on any existing structure or sequence data. Clearly this is a fine line since those physical principles are themselves derived from known structure and sequence, but it is meant to imply that they are used to define general principles, rather than used directly. Chapter 28 discusses this further. The results from CASP4 for novel fold recognition can be found in Lesk, Lo Conte, and Hubbard (2001). Success in this category was measured in terms of tertiary structure prediction,

Q8

secondary structure prediction, and residue-residue contacts. This● is reflected in the conclusions here:

- Progress has been made in the areas of tertiary structure prediction and in contact prediction, both using *ab initio* methods and knowledge-based methods.
- Secondary structure prediction still breaks down with the appearance of unusual secondary structures, for example, very long helices that were broken into fragments by all participating groups.
- Assessment should be performed with some consideration for the difficulty of the target even though this difficulty is hard to measure. Specifically, with the continuity in fold space there is ambiguity in what can be considered a new fold, but if a fold is believed to be truly new that should be weighted higher than a fold that at least has partial similarity to a known fold.

CAFASP

In recognition of the value of automated prediction servers—which in part reflect progress in structure prediction influenced by previous CASPs—the results of CAFASP2 were published along with the CASP4 results (Fischer et al., 2001). Overall, according to the CAFASP assessors, only 11 groups in CASP performed better than the automated servers and a number of those groups clearly used the automated servers as part of their prediction strategy. However, the best human predictions do much better than the best automated predictions. Moreover, perhaps stating the obvious, difficult targets for humans are also difficult targets for automated methods, and there is much room for improvement in both categories. This comparison between CASP and CAFASP results is useful in a number of ways; most notably, it indicates to structure predictors what elements of the expert contribution need still to be added to automated approaches—clearly a nontrivial exercise—and for a biologist how valuable are the predictions compared to the best expert opinion and which Internet-accessible servers perform best.

CAFASP2 characterized five classes of server: fold recognition (19), secondary structure prediction (8), contacts prediction (2), *ab initio* (2), and homology modeling (3). The numbers in parentheses indicate how many servers were in each category. Some general observations are:

- Targets were divided into two classes: homology modeling (15) and fold recognition (26). The top ranking servers produced correct models for all homology modeling targets, but for only 5 of the 26 fold recognition targets (of the 21 not well predicted, 4 with new folds).
- In the fold recognition server group, the servers combined found approximately twice the number of targets than did any server alone, speaking to the value of a well-evaluated consensus approach.
- Secondary structure prediction accuracy was measured at 76% overall, but there was insufficient data to provide a detailed comparative analysis.

SUMMARY

This short introductory chapter is intended simply to introduce a sense of the progress, limitations, challenges, and likely future developments in the field of protein structure

prediction through what seems to be a unique scientific process. CASP and CAFASP represent a direct challenge and careful assessment of a field of study that has captured the interest of many scientists. Three of the best scientists in the field and their colleagues provide a more detailed description of the field and how it is developing in Chapters 25, 26, and 27.

As prediction methods have advanced the distinction between comparative modeling, fold recognition, and novel fold recognition have blurred somewhat. It is a testament to the community as a whole that the knowledge of the algorithms developed, World Wide Web servers providing access to them fold libraries, and so on are shared by the community, thus, making it relatively straightforward for any investigator to apply a melting pot of methods to the prediction process. What all approaches need are more targets and a continued refinement to the evaluation process. The first need is being met in part by the PDB, which is, with depositors' approval, releasing sequences ahead of structure release (see <http://www.rcsb.org/pdb/status.html>). Further, the structural genomics projects are reporting their progress for all targets on a weekly basis (see <http://targetdb.pdb.org/>). While there is no indication that the sequences of the latter will lead to a structure, it is a rich source of targets (14,000 in March 2002).

Q9

Not only do CASP and CAFASP measure progress, they help define where efforts should be directed to move the field forward. It is a testament to how far the field has come that investigators are now turning to the unknown. Although attempting to predict a structure that will appear experimentally helps improve the methods applied to structure prediction, it does not further our understanding of living systems directly. Attempts at defining the "The Most Wanted" (Abbott, 2001)—the structures most in need of prediction to help further our understanding of the biology, and the efforts to make those predictions, speak to a healthy future for the field of protein structure prediction. To the many individuals who help define the CASP and CAFASP processes and compete in the experiments this is a tribute.

REFERENCES

- Abbott A (2001): Computer modelers seek out 'Ten Most Wanted' proteins. *Nature* 409(6816):4.
- Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001): CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* 45 Suppl 5:171–83.
- Lesk AM, Lo Conte L, Hubbard TJ (2001): Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins* 45 Suppl 5:98–118.
- Rost B (1999): Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85–94.
- Tramontano A, Leplae R, Morea V (2001): Analysis and assessment of comparative modeling predictions in CASP4. *Proteins* 45 Suppl 5:22–38.
- Venclovas ZA, Fidelis K, Moult J (2001): Comparison of performance in successive CASP experiments. *Proteins* 45 Suppl 5:163–70.

QUERIES TO BE ANSWERED BY AUTHOR (SEE MARGINAL MARKS)

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Query No.	Query
Q1	By “they” do you mean the groups?
Q2	Figure legend ok? all figures must have legend.
Q3	“Result” ok? or “finding”? Please be specific
Q4	Book will be published in 2003. Can these sections be provided?
Q5	“method” ok?
Q6	“Result” ok?
Q7	“Finding” ok?
Q8	“this” what?
Q9	Can you use a later count?

QUERIES TO BE ANSWERED BY EDITOR (SEE MARGINAL MARKS)

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Query No.	Query
-----------	-------

ED1	Ok to have acronyms in Chapter title?
-----	---------------------------------------
