PROTEINS: Structure, Function, and Genetics 23:ii–iv (1995)

INTRODUCTION

A Large-Scale Experiment to Assess Protein Structure Prediction Methods

Methods for obtaining information about structure from amino acid sequence have apparently been advancing rapidly. But just what can these methods currently deliver? The following papers present the results of a large scale experiment that we have orchestrated to determine the current state of the art in protein structure prediction. We consider that the only way to objectively assess the usefulness of prediction methods is to ensure that predictions are made without any knowledge of the answers. We therefore set out to provide a framework in which a large number of such blind predictions could be made and evaluated. The procedure consisted of three parts: the collection of targets for prediction from the experimental community, the collection of predictions from the modeling community, and the assessment and discussion of the results.

COLLECTING PREDICTION TARGETS

Information was solicited from X-ray crystallographers and NMR spectroscopists about structures that were either expected to be solved shortly or that had been solved already but not discussed in public. Targets were identified through personal contacts, blanket emailing, and appeals at scientific meetings. The collecting and management of prediction targets proved to be a difficult undertaking. In all, information on 33 different proteins was obtained. Some of these were not solved in time for the prediction experiment and some were made public without sufficient notice to the predictors. Finally, one or more predictions were received on 24 of these targets.

CATEGORIES OF PREDICTION

The difficulty of prediction depends on the extent of the relationship of the target protein to already known structures. For this reason, predictions were divided into three types:

1. Comparative modeling: Cases where there is a clear relationship between the sequence of the target protein and one or more known structures. In these circumstances, it is assumed that the tertiary structures are similar, and an initial model may be based on the structure with the most similar sequence. Thus, an approximately correct fold is assured. The prediction challenge is then in devising techniques that can determine the detailed structural differences between the target and the known related structures. These techniques deal with the alignment of the target sequence on the templates, the best choice of template structure for each part of the chain, small (of the order of 1 or 2 Å) adjustments of main chain position, the orientation of side chains, and the conformation of stretches of chain not related to any of the template structures (the "loops").

2. Threading, or fold identification: Even when there is no detectable sequence relationship between two proteins, they may have closely related folds. Threading techniques attempt to identify the fold a sequence will adopt by considering its fit to each member of a library of known folds. That is, the sequence is "threaded" onto each fold, and the suitability of the interactions thus created is evaluated using a fitness function. This is a relatively new technique, made possible by the rapidly increasing size of the set of known folds. A wide variety of scoring functions and sequence structure alignment methods are currently being developed. The primary challenge is to unambiguously identify an equivalent fold to the target protein in the database, if one exists. In cases where this can be done, subsidiary questions concern how reliable a model based on the fold similarity would be. For example, is the alignment of the target protein sequence on the related structure correct? At this stage, it is probably not possible to produce as detailed a model as in comparative modeling.

3. Ab initio predictions: All methods that do not rely in a direct way on database approaches. The classical view of the structure prediction problem: presented with nothing but a sequence and some knowledge of the interactions between amino acids, predict the three-dimensional fold. Methods include secondary structure prediction, the use of rules about protein topology, lattice based simulations, and molecular dynamics and Monte Carlo methods.

Received May 30, 1995; accepted June 2, 1995.
Address reprint requests to John Moult, Center for Advance Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850.

Many different types of empirical force field and levels of structure description are in use. This is certainly the most difficult category of prediction, and as the following papers show, any significant relationship between the prediction and the experimental structure should be considered a success. This includes a very rough topological description rather than a numerical assessment based on root mean square deviations between structures. Although some methods aspire to predicting small pieces of proteins in a detailed way, complete detailed models are not to be expected at present. Predictions of just secondary structure were not encouraged since the experiment was focused on tertiary structure. In the end, however, quite a few such predictions were accepted.

SOLICITING THE PARTICIPATION OF PREDICTORS

An experiment of this type is only meaningful if a representative sample of the prediction community can be persuaded to take part. We therefore made a concerted effort to make sure the experiment was widely publicized by email, usenet announcements, and journal advertisements. We also directly approached the predictors we were aware of. The more prominent people in the field were particularly courted. Details of participating groups can be found in the papers by the assessment teams. The success of this strategy varied by category of prediction. In the comparative modeling segment, all but one of those groups we considered key took part, and the more popular commercial packages (from Biosym, MSI, Tripos, and Oxford Molecular) are represented. In the threading category, all known predictors agreed to participate. We were less successful in the ab initio area. Although a number of very interesting and significant predictions were received, four people we regarded as key did not take part. Two did not agree to participate, the other two agreed, but in the end did not submit any predictions. Readers must draw their own conclusions as to the significance of this. On the one hand, taking part in the experiment required a large commitment of time, and it was difficult for a predictor to judge how significant the outcome would be. On the other hand, nearly all of the key groups in the other two categories did participate. A point that needs stressing is that taking part required courage from all predictors. They had to agree to put their reputations on the line, through discussion of their work at a meeting, publication in this journal, and public availability of the results. In all, 35 prediction teams took part in the experiment.

COLLECTING PREDICTIONS

Information about the target structures (the sequences, and any relevant references) was made available to the prediction community via an anonymous ftp account. A list of interested predictors was maintained, and they were informed of new targets and the expiration date of targets. Timing is critical in a procedure of this kind. To be considered part of the experiment, predictions had to be received by the specified expiration dates. Each target was assigned an initial date based on information provided by the experimentalists. Expiration dates were then updated as the structure solution proceeded. The prediction period began in March 1994, and finished at the end of October, with different targets expiring throughout that period. Most predictions were received toward the end of the process. Some 34 predictions were obtained on seven different targets for the comparative modeling category. There are 66 predictions using threading methods and 29 tertiary structure ab initio predictions, both on about 20 targets. The predictions can be obtained via the internet at ftp:http://iris4.carb.nist.gov/pub/model_database.

Assessment of the predictions was done by three independent teams, one for each category. These teams were led by Michael James, University of Alberta, for comparative modeling; Shoshana Wodak, Free University of Brussels, for threading; and Fred Cohen, University of California at San Francisco, for ab initio predictions. Initial guidelines to predictors outlining the basis on which assessment would be made were provided by the organizers. These were extended and modified as required by the assessors. The assessors had a period of 1 month between the end of the prediction period and the meeting to complete their evaluations. The outcome of their work constitutes the primary results of the experiment. There is a paper by each assessment team in the following pages.

ASSESSMENT MEETING

In December 1994, a meeting was held at the Asilomar conference center in California to examine what went right with the predictions, what went wrong, and, where possible, to understand why. Approximately 1 day was devoted to each of the prediction categories. Each day began with a review lecture by the leader of an assessment team, followed by lectures by some of the predictors in that category. Speakers were selected by the assessment teams on the basis of the interest and accuracy of their predictions. In the afternoons, participants were able to investigate many of the methods interactively on computer workstations. In the evenings, there was an extensive discussion of the day's results.

SELECTION OF PREDICTIONS FOR PUBLICATION

In addition to review papers by the three assessment teams, this issue contains short papers by those predictors who were invited to make oral pre-

sentations at the assessment meeting last December. All the papers have been peer reviewed. Authors were instructed to focus on what went right with their predictions, what went wrong, why, and what they learned about the prediction methods. Reviewers were asked to consider how well these issues are addressed.

LIMITATIONS

This is the first experiment of its kind on such a large scale. We consider that much was learned, but it should be realized there are limitations to the significance of the results. It was hard for predictors to gauge how seriously the community would view the outcome, and therefore how much effort to devote to the task. Some unevenness in results may arise from that factor, rather than real differences between the effectiveness of the methods. It is impossible to assess the quality of a method on the basis of one example. Although we tried to insist that all predictors made two and preferably more predictions, there are some exceptions. Methods of assessment evolved during the experiment, and initially it was not clear what information should be required from predictors about their methods and results. Some of these gaps were filled in along the way, but not all. As noted above, in the ab initio category, we were not completely successful in soliciting the participation of all the predictors we would have liked. Finally, the results represent a snapshot in time in the development of the methods. A year earlier or a year later would produce quite a different picture. For all of these reasons, the results should not be used to condemn or exult any particular group and their methods.

SIGNIFICANCE OF THE RESULTS

There is an adage in the molecular modeling field attributed to one of its wisest founding fathers, Shneior Lifson: "You don't learn any thing until something goes wrong." Plenty went wrong with these predictions, and therein lies the principal value of the experiment. At the assessment meeting and in the papers in this issue, a rather precise picture of the capabilities and deficiencies of the methods has been obtained. We hope there will be three main outcomes. For members of the structural biology community not directly involved in structure prediction, the results should provide a reasonable guide to the current state of the art. For most of us predictors this was a cathartic experience, and we have emerged from it with a new and sharper sense of direction. In many labs, new work focused on overcoming the current bottlenecks is now in progress. We hope that this stimulus to and directing of research will in time prove to be a valuable outcome. Finally, we believe the experiment has shown that objective testing of structure prediction methods is both practical and necessary. Thus, in the future, any algorithm which claims to have predictive abilities should be required to demonstrate that in this manner.

FUTURE DEVELOPMENTS

A second experiment along the same lines is just beginning. A call for targets will be issued in the fall of 1995, and predictions will be collected through September 1996. An assessment meeting is planned for in December 1996. We are also attempting to establish a prediction database that will provide a continuous supply of targets, register predictions, and provide prediction assessment services and software. Details on these developments will be posted on the web page(http://iris4.carb.nist.gov/pub/model_database).

ACKNOWLEDGMENTS

This experiment was possible only because of the cooperation and work of a large number of people. We are very grateful to the members of the experimental community who agreed to provide targets. Without their cooperation the experiment would not have been possible. As noted before, taking part required courage and commitment on the part of all the predictors. Careful and authoritative assessment is a critical component of the experiment, and without a great deal of work by each of the assessment teams, it would have failed. We are grateful to the editor of this journal, Ed Lattman, for providing a mechanism for peer reviewed publication of the results. We gratefully acknowledge financial support from Lawrence Livermore National Laboratory (contract W-7405-ENG-48 and LDRD award 93-DI-003), the National Institute of Standards and Technology, Sandia National Laboratory, and the Department of Energy Office of Health and Environmental Research. We thank Rod Balhorn, Walt Stevens, and Nick Winter for their participation in the organization of the experiment. Thanks to MSI and Biosym for additional meeting support.

John Moult
Jan T. Pedersen
University of Maryland Biotechnology Institute
Rockville, Maryland

Richard Judson
Sandia National Laboratories
Livermore, California

Krzysztof Fidelis
Lawrence Livermore National Laboratory
Livermore, California