Volume 325, number 1,2, 29–33 FEBS 12572 June 1993

Minireview

Predicting the conformation of proteins

Man versus machine

Steven A. Benner and Dietlind L. Gerloff

Laboratory for Organic Chemistry, E.T.H. Zurich, CH-8092 Zurich, Switzerland

Received 14 April 1993; revised version received 26 April 1993

Two types of approaches for predicting the conformation of proteins from sequence data have lately received attention: 'black box' tools that generate fully automated predictions of secondary structure from a set of homologous protein sequences, and methods involving the expertise of a human biochemist who is assisted, but not replaced, by computer tools. A friendly controversy has emerged as to which approach offers a brighter future. In fact, both are necessary. Nevertheless, a snapshot of the controversy at this instant offers much insight into the structure prediction problem itself.

Protein; Structure prediction; Evolution

1. INTRODUCTION

Almost every biochemist knows of the protein structure prediction problem: How does one take a protein sequence as input and produce a model of the protein's conformation (secondary or tertiary structure) as output? Many also know of the Chou–Fasman [1] and GOR [2] methods for obtaining secondary structure predictions. And some know that these (and other) classical methods rarely yield secondary structure predictions that could support an effort to model tertiary structure, at least from a single sequence. As a result, the general view is that the protein structure prediction problem, if not insoluble, is likely to remain unsolved for a very long time [3].

This view has been changed by a number of recent events. First, several bona fide predictions [4–14], those made and announced before experimental structures are known, have proven to be intriguingly accurate when compared with subsequently determined crystal structures [15–25]. Most of these have been made using a blend of human expertise and computer assistance; several included guesses of tertiary structure as well as secondary structure.

Second, some fully automated methods for predicting secondary structures have been reported to have approached or broken the '70%' barrier [26,27]. This implies that when made 'blind' (without the computer having knowledge of the correct structure), over 70% of the residue assignments correspond to experimental assignments using a standard three state scoring scheme (helix, strand, or neither).

In both cases, structural information is generally derived from a *set* of aligned homologous sequences, rather than from a single sequence. Such predictions therefore necessarily assume that homologous proteins have similar conformations [28]. Because this assumption is only an approximation, the structures produced are consensus models. They do not apply exactly to any individual protein in a family, but may be used as starting points for homology modeling of individual family members [29].

The contrast between human-based and machine-based prediction strategies reflects an underlying rift within the structure prediction community. On one side, scientists seek to understand *why* predictions work when they work (and why they fail when they fail) in terms of underlying structural and evolutionary models [30,31]. The tradition finds its roots in organic chemistry, where conformational problems are approached one at a time, and where a high value is placed on understanding one system in detail before proceeding to the next. The other, aware of the ease with which humans are prone to self-deception, focuses on automation and reproducibility [32]. The tradition finds its roots in computer science, and places a high value on testing methods with statistically large numbers of proteins, sampled properly, examined blindly, and scored automatically.

The last year has seen the emergence of a friendly

Correspondence address: S.A. Benner, Laboratory for Organic Chemistry, E.T.H. Zurich, CH-8092 Zurich, Switzerland. Fax: (41) (1) 262 2437.

disagreement [33,34] between two laboratories representing these two cultures, one in Zurich, the other in Heidelberg. As expected from its culture, the first has focused on developing and publishing individual bona fide predictions: for the quaternary structure of alcohol dehydrogenase, for protein kinase, for the SH3 domain, for the venom allergens, for the MoFe nitrogenase protein, for the hemorrhagic metalloproteases, and for protein phosphatase. The second, consistent with *its* culture, has focused instead on developing a neural network accessible to the public by server and testing it on large numbers of proteins in the structure database. Neither group is alone in its approach, of course. Further, it is impossible at this time to evaluate the relative merits of the two approaches. Indeed, it is likely that the two will coexist for a long time, complementing each other in prediction work. Nevertheless, a snap shot for *this* instant in time offers much insight into the structure prediction problem itself, and serve as a guide for how the field will develop in the future.

At the outset, we should point out that we are advocates of the human-based methods [7]; the reader may wish to discount aspects of this overview accordingly. Further, we must acknowledge the crystallographers and NMR spectroscopists whose hard work makes the prediction game possible.

2. PREDICTIONS: THE FIRST ROUND

This story begins when the Zurich group was challenged by A. Musacchio to predict the secondary structure of the SH3 domain before its experimental structure appeared in print [23]. The time was short, as the experimental paper was already in press. Nevertheless, through efforts of several editors, a manuscript containing an unrefined secondary structure prediction was refereed and accepted in the *Journal of Molecular Biology* [13] before the crystal structure of the spectrin SH3 domain appeared in *Nature* [23]; the editors of *Nature* also arranged to publish concurrently a Scientific Correspondence summarizing the prediction [35]. The prediction is shown in Fig. 1, together with the secondary structure assigned from the experimental data.

Shortly thereafter, *Nature* published a Scientific Correspondence from Rost and Sander [32], who served as a jury to evaluate the prediction made in Zurich in the light of the crystal structure. Rost and Sander provided their own secondary structure assignment for the SH3 domain family made using a neural network server developed in Heidelberg [27]; the input for this prediction was the spectrin domain sequence together with homologous sequences that met specified levels of sequence similarity. The jury noted that in both the Zurich and Heidelberg predictions, a single helix had been assigned to a region where no helix was reported in the spectrin domain. Thus, they concluded that both predictions contained one error and a 'per segment' score of 80%. However, because the Heidelberg method was fully automated (and the Zurich method not), the jury considered it superior [33], consistent with its culture.

The jury's verdict contained three complicating details. First, early in the sequence was a region where no secondary structure was assigned by the crystallographers [23], but where both groups predicted a β strand [13,32]. The crystallographers had noted, however, that the segment was extended, and might have been assigned as a β strand had certain key hydrogen bonds been observed. Thus, this region could not be counted as an error for either prediction.

Second, the Heidelberg group remarked that the tertiary structure prediction made in Zurich was 'wrong', while the Zurich group insists that it did not predict a tertiary structure. There has ensued an exchange of correspondence where the Heidelberg group has suggested that the Zurich group implied a tertiary structure in its Scientific Correspondence in *Nature*, even though the JMB prediction paper denied any attempt to predict a tertiary structure from the unrefined secondary structure. The jury had neglected to ask the Zurich group for a copy of the prediction manuscript before delivering a verdict.

Finally and more generally, the jury based their verdict in part on the per residue scores achieved by the various prediction methods. These are rather uninformative in assessing how useful a prediction is as the starting point for assembling a tertiary structure. A discussion has ensued regarding the appropriateness of such scoring schemes for evaluating consensus predictions. For example, Thornton noted that although the residue-by-residue score achieved in one prediction made in Zurich [7] was not much better than that obtained using classical methods, 'the possibility of extending [the] prediction to a tertiary fold is much better with the Benner–Gerloff method' [36]. Largely through a disagreement on scoring methods, the same jury (in another venue) concluded that such evaluations were 'misleading' and 'exaggerated' [34].

3. WHY WAS A HELIX MISASSIGNED IN THE SH3 DOMAIN IN ZURICH?

This was not the first time that a helix had been misassigned in a bona fide prediction made in Zurich. In particular, an interior helix was misassigned in protein kinase [7]. Consistent with its culture, the Zurich group looked to explain this misassignment in terms of the formalism that it uses to assign secondary structures. Central to this formalism are assignments of positions in an alignment to the surface or the interior of the folded protein. A helix is predicted when these assignments display 3.6 residue periodicity. Fig. 2a shows a segment of the protein kinase prediction where this 3.6 residue periodicity is obvious; the predicted helix was found in the subsequently determined crystal structure.

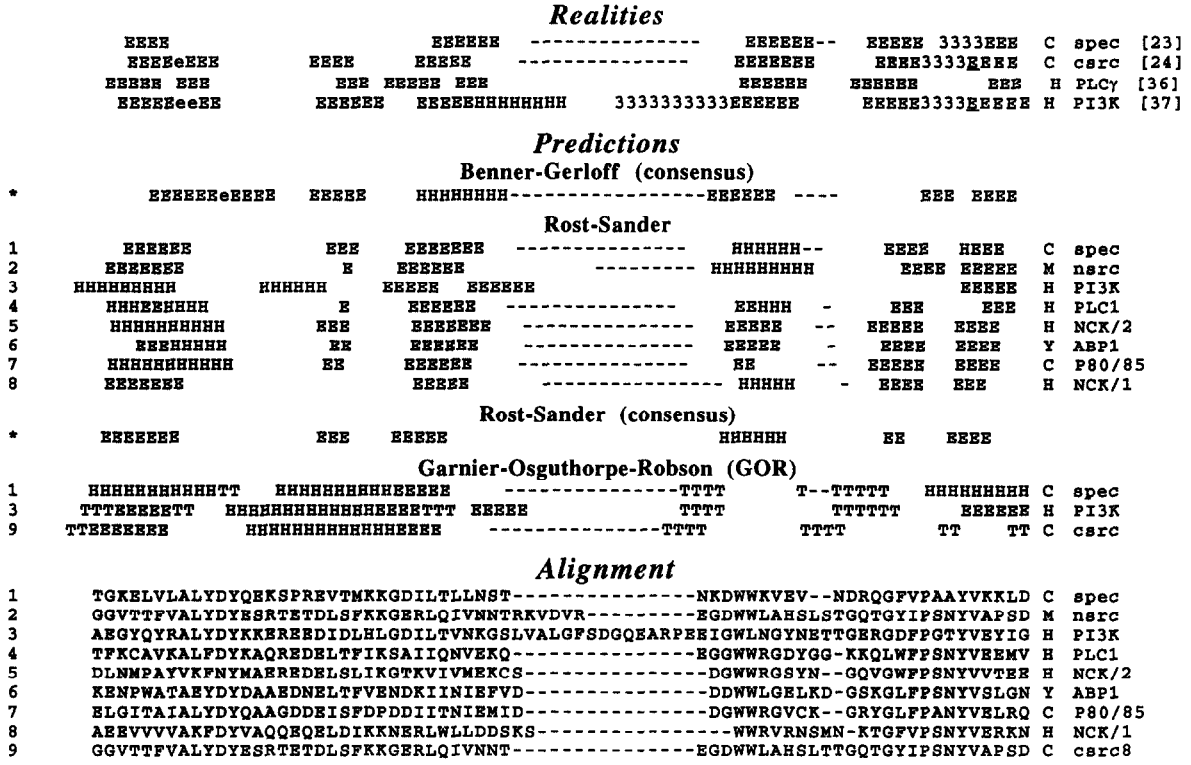

Fig. 1. The SH3 domain. Comparison of the experimentally determined structures of three homologous SH3 domains (using secondary structural assignments provided by experimentalists) and predictions made by several methods. Alignment from [38]. Underlined positions indicate residues assigned both as part of a 3_{10} helix and a β strand. The Zurich method provides a single unrefined consensus prediction [13] for the entire protein family. Predictions obtained by the GOR method [2] and the Heidelberg server [27] for various homologs as indicated. The Rost–Sander consensus prediction, communicated for this article, was based on 65 SH3 domains (not listed). Dashes indicate indels. H indicates α helix; E indicates β strand; e indicates β bulge, T indicates turn, 3 indicates 3_{10} helix. Gaps indicate regions where no secondary structure is assigned.

Fig. 2b shows the helical segment missed in protein kinase. It is clear why the helix was missed; it lies entirely within the folded structure, there are no surface residues, and therefore no pattern of 3.6 residue periodicity can be observed. Prompted by this misassignment, new formalisms were developed to identify internal helices in Zurich.

For the SH3 domain, the issue was not so clear. The predicted helix is short, meaning that the 3.6-residue periodicity could not be well established. Further, the domain is small and the strand lying in this region participates in both β sheets in the protein [23]. These together yield a pattern of surface and internal positions in a beta strand that could be mistaken for a short helix (Fig. 1). Regardless of whether this explains the misassignment (see below), it underscores a difference in the two cultures. To the human-based predictors, such observations are interesting hypotheses that lead to further work. To machine-based predictors, they are ad hoc excuses designed to cover over failure. There is no need to resolve this conflict, of course; formalisms modified in light of past misassignments can be tested by making more bona fide predictions.

4. SOME MORE STRUCTURES EMERGE

Fortunately, the discussion moved forward due to the introduction of experimental structures for three additional SH3 domains, the Src tyrosine kinase [24], the phospholipase C-γ (PLC-γ) [37], and the phosphoinositol-3′-kinase (PI3K) domains [38]. These are shown in Fig. 1. Two points are evident. First, the conformation of the SH3 domain family has undergone considerable divergence. Second, the PI3K SH3 domain contains a helix in the general region where the Zurich group had predicted one. Indeed, Fig. 1 suggests that once divergence in conformation between SH3 domains is considered, the Zurich prediction was not so bad [33,39].

This observation underscores a problem, however, in scoring a consensus prediction: Which experimental structure should be used? A consensus model might be viewed as representing the structure of the most recent common ancestor of the protein family being examined [40]. In this view, the ancestor of the sequences in the multiple alignment used to make the SH3 prediction contained a helix that was retained in the PI3K SH3 domains (not included in this alignment) and lost in the

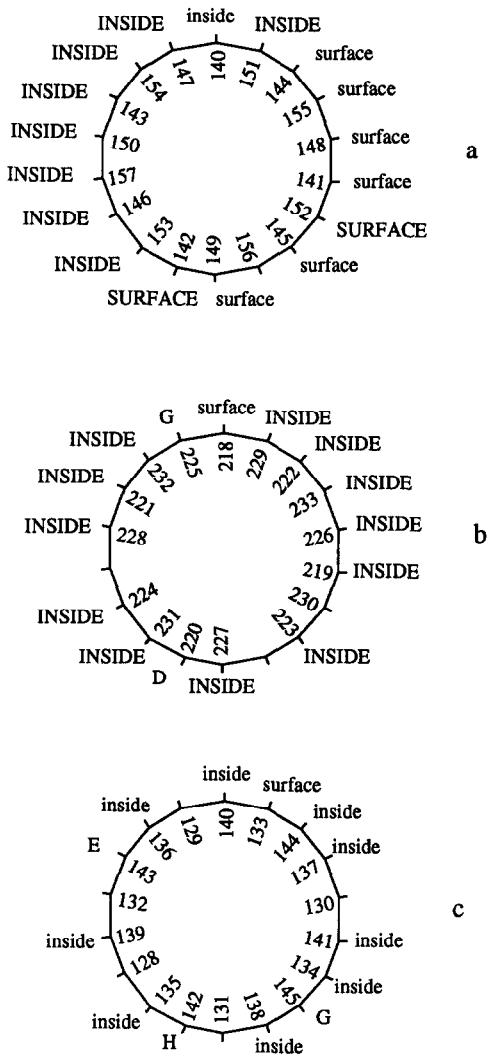

Fig. 2. The value of error. Helical wheels for (a) a surface helix of protein kinase assigned correctly in the bona fide prediction [7], (b) an interior helix misassigned in the prediction [7], and (c) an interior helix assigned in the hemorrhagic metalloproteases using heuristics developed as a result of the error in (b).

other domains. This implies that one must retrieve the ancient protein and determine its structure to evaluate a consensus structural model. While not impossible [40], this is certainly inconvenient. Nevertheless, this experience shows why a jury evaluating a consensus prediction must find and consider several experimental structures [39] before rendering a strong verdict. Weak verdicts can, of course, be rendered at any time.

5. A HUMAN EXAMINES THE MACHINE

The Heidelberg server presumably yields a consensus prediction for a set of homologous protein sequences, just as in Zurich. This implies that the Heidelberg server should give similar secondary structure predictions no matter which member of a family of homologous proteins is used as a 'guide sequence'. This proposition was simple to test; sequences of some homologous SH3 domains were sent to the server and the secondary predictions retrieved. These are shown (together with the guide sequences) in Fig. 1.

The results were initially surprising. The server produced different (and often quite different) secondary structure predictions when challenged with different sequences within the same protein family (Fig. 1). These results were communicated to the Heidelberg group, which suggested an explanation. The server constructs multiple alignments by a pairwise comparison procedure. Thus, it is possible that different guide sequences (which begin the process of alignment construction) yield different multiple alignments, certainly if they retrieve different sets of homologs from the database, and possibly even if they ultimately retrieve the same set of sequences from the database. The discrepancies in the predictions made with different homologs were therefore attributed to difficulties in alignment procedure, rather than the difficulties in the prediction procedure.

For this minireview, Sander and his coworkers kindly provided a consensus prediction for the SH3 domain family starting with the 9 sequence strips from Fig. 1. The database was searched, sequences meeting a threshold collected, sequences probably not SH3-like eliminated by hand, and the remaining sequences aligned. This consensus was considered more appropriate than the consensus prediction that would be obtained simply by averaging the individual predictions shown in Fig. 1.

It was clear, however, that in not all cases were predictions for guide sequences different due to different multiple alignments. For example, the same 7 hemorrhagic metalloproteases [41] yielded essentially identical multiple alignments when used individually as guide sequences. Yet the predictions yielded by the server were different. A flurry of correspondence just before the appearance of this review suggested that this arose if the guide sequence also appeared in the database; in this case, the guide sequence was apparently counted twice in the prediction. This presumably seriously effects the prediction only with small alignments and in regions where the secondary structure predictions are not strong.

6. MAN OR MACHINE?

This snapshot is complete. Remarkably, there has been an agreement. A fully automated black box has remarkable advantages. When the group who developed it makes it conveniently available to other groups by server, it can be probed by almost anyone with a turn around time of just a few hours. This allows those of us from the other culture to do freely what we do best; ask questions about why the predictions are the way they are. In contrast, predictions made through human inter-

vention require a human with expertise and time, first to read the long (and somewhat boring) papers [7,13] that describe the prediction as it passes through all intermediate stages, and then to apply what is learned to a new protein family. Both methods are useful, however, especially when used in parallel. The next step will be a joint *bona fide* prediction made by the Zurich and Heidelberg groups.

In Zurich, more crystal structures are awaited. In Heidelberg, better neural networks are expected. Still more predictions must be made. Above all, better methods are needed for scoring consensus predictions in a way that allows an assessment of their utility in building tertiary structure models; the classical three state scoring scheme is clearly inadequate. We will keep you posted.

REFERENCES

[1] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.
[2] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[3] Hunt, T. and Purton, M. (1992) Trends Biochem. Sci. 17, 273.
[4] Cohen, F.E., Kosen, P.A., Kuntz, I.D., Epstein, L.B., Ciardelli, T.L. and Smith, K.A. (1986) Science 234, 349–352.
[5] Crawford, I.P., Niermann, T. and Kirschner, K. (1987) Proteins, 2, 118–129.
[6] Bazan, J.F. (1990) Proc. Nat. Acad. Sci. USA 87, 6934–6938.
[7] Benner, S.A. and Gerloff, D. (1991) Adv. Enzyme Regulat. 31, 121–181.
[8] Moe, G.R. and Koshland, Jr., D.E. (1986) in: Current Communications in Molecular Biology, Microbial Energy Transduction, Genetics, Structure and Function of Membrane Proteins, 163–168, Cold Spring Harbor Laboratory Press.
[9] Russell, R.B., Breed, J. and Barton, G.J. (1992) FEBS Lett. 304, 15–20.
[10] Barton, G.J., Newman, R.H., Freemont, P.S. and Crumpton, M.J. (1991) Eur. J. Biochem. 19, 749–760.
[11] Musacchio, A., Gibson, T., Lehto, V.-P. and Saraste, M. (1992) FEBS Lett. 307, 55–61.
[12] Benner, S.A., Badcoe, I., Cohen, M.A. and Gerloff, D. (1992) Nature 359, 781.
[13] Benner, S.A., Badcoe, I., Cohen, M.A. and Gerloff, D. (1993) J. Mol. Biol. 229, 295–305.
[14] Gerloff, D.L., Jenny, T.F., Knecht, L., Gonnet, G.H. and Benner, S.A. (1993) FEBS Lett. 318, 118–124.
[15] Bazan, J.F. (1992) Science 257, 410–412.
[16] McKay, D.B. (1992) Science 257, 412–413.
[17] Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W. and Davies, D.R. (1988) J. Biol. Chem. 263, 17857–17871.
[18] de Vos, A.M., Ultsch, M. and Kossiakoff, A.A. (1992) Science 255, 306–312.
[19] Knighton, D.R., Zheng, J., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S. and Sowadski, J.M. (1991) Science 253, 407–414.
[20] Milburn, M.V. Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, Jr., D.E. and Kim, S.-H. (1991) Science 254, 1342–1346.
[21] Waksman, G., Kominos, D., Robertson, S.C., Pant, N., Baltimore, D., Birge, R.B., Cowburn, D., Hanafusa, H., Mayer, B.J., Overduin, M., Resh, M.D., Rios, C.B., Silverman, L. and Kuriyan, J. (1992) Nature 358, 646–653.
[22] Huber, R., Romisich, J. and Paques, E. -P. (1990) EMBO J. 9, 3867–3874; Huber, R., Schneider, M. Mayr, I., Romisich, J. and Paques, E.-P. (1991) FEBS Lett. 275, 15–21.
[23] Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. and Saraste, M. (1992) Nature 359, 851–855.
[24] Yu, H., Rosen, M.K., Shin, T.B., Seidel-Dugan, C., Brugge, J.S. and Schreiber, S.L. (1992) Science, 258, 1665.
[25] Kim, J. and Rees, D.C. (1992) Nature 360, 553–560.
[26] Salzberg, S., Cost, S. (1992) J. Mol. Biol. 227, 371–374.
[27] Rost, B. and Sander, C. (1993) J. Mol. Biol., in press
[28] Chothia C. and Lesk, A.M. (1986) EMBO J. 5, 823–826.
[29] Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) Nature 326, 347–352.
[30] Benner, S.A. (1989) Adv. Enzyme Reg. 28, 219–236.
[31] Benner, S.A. and Ellington, A.D. (1990) Bioorganic Chemistry Frontiers 1, 1–70 (1990)
[32] Robson, B. and Garnier, J. (1993) Nature 361, 506.
[33] Rost, B. and Sander, C. (1992) Nature 360, 540.
[34] Rost, B., Schneider, R. and Sander, S. (1993) Trends Biochem. Sci., 18, 120–123.
[35] Benner, S.A., Cohen, M.A., Gerloff, D. (1992) Nature 359, 781.
[36] Thornton, J.M., Flores, T.P., Jones, D.T. and Swindells, M.B. (1991) Nature 354, 105–106 (1991).
[37] Kohda, D., Hatanaka, H., Odaka, M., Mandiyan, V., Ullrich, A., Schlessinger, J. and Inagaki, F. (1993) Cell 72, 953–960.
[38] Koyama S., Yu, H., Dalgarno, D.C., Shin, T.B., Zydowsky, L.D. and Schreiber, S.L. (1993) Cell 72, 945–952.
[39] Barton, G.J., Russell, R.B. (1993) Nature 361, 505–506.
[40] Stackhouse, J., Presnell, S.R., McGeehan, G., Nambiar, K.P. and Benner, S.A. (1990) FEBS Lett. 262, 104–106.
[41] Takeya, H., Arakawa, M., Miyata, T., Iwanaga, S. and Omori-Satoh, T. (1989) J. Biochem. (Tokyo) 106, 151–157.