

Sisyphus and prediction of protein structure

Burkhard Rost and Sean O'Donoghue

Abstract

The problem of predicting protein structure from the sequence remains fundamentally unsolved despite more than three decades of intensive research effort. However, new and promising methods in three-dimensional (3D), 2D and 1D prediction have reopened the field. Mean-force-potentials derived from the protein databases can distinguish between correct and incorrect models (3D). Inter-residue contacts (2D) can be detected by analysis of correlated mutations, albeit with low accuracy. Secondary structure, solvent accessibility and transmembrane helices (1D) can be predicted with significantly improved accuracy using multiple sequence alignments. Some of these new prediction methods have proven accurate and reliable enough to be useful in genome analysis, and in experimental structure determination. Moreover, the new generation of theoretical methods is increasingly influencing experiments in molecular biology.

Introduction

In Greek mythology, Sisyphus is condemned to an eternity of hard labour; his labour is frustrating and fruitless, for just as he is about to achieve his goal, his work is undone and he must start again from the beginning. Those who work in protein structure prediction seem to share the same fate.

For over 30 years, there has been an ardent search for methods to predict the three-dimensional (3D) structure from the sequence. Many methods were found which initially looked very promising, but always the hope has been dashed. The results of the first and the second Asilomar meeting on structure prediction [*Proteins*, 23(Special Issue), 1995] demonstrate clearly that the goal has not been reached, yet.

The search has been driven by the belief that the 3D structure of a protein is determined by its amino acid sequence (Anfinsen, 1973). While it is now known that chaperones often play a rôle in the folding pathway, and in correcting misfolds (Hartl *et al.*, 1994; Corrales and Fersht, 1996), it is believed that the final structure is at the free-energy minimum. Thus, all information needed to predict the native structure of a protein is contained in the amino

acid sequence, plus a knowledge of its native solution environment.

Currently, databases for protein sequence (Bairoch and Apweiler, 1997) and protein structure (Bernstein *et al.*, 1977) are expanding rapidly due to large-scale sequencing projects (Oliver *et al.*, 1992; Fleischmann *et al.*, 1995; Fraser *et al.*, 1995; Goffeau *et al.*, 1996; Johnston *et al.*, 1996) and improvements in experimental determination of 3D structures (Lattman, 1994). Can structure prediction profit from this flood of information?

Indeed, there is a flood of literature on protein structure prediction that is attempting to keep track with the expanding databases. Recent comprehensive reviews include Rost and Sander (1994b) and Rost and Sander (1996); recent books on structure prediction include Doolittle (1996) and Sternberg (1996). For a user-oriented, practical approach to structure prediction and sequence analysis, see Bork and Gibson (1996), Rost and Valencia, (1996) and Rost and Schneider (1997). In this review, we will focus mainly on recent prediction methods designed to exploit the growing databases. We show that, unlike Sisyphus, the predictor of protein structure has actually moved closer to his goal over the years—although he is still far from it; moreover, the labour has not been fruitless. We will discuss some of the prediction methods which have proven to be accurate and reliable enough to be useful in genome analysis and in experimental structure determination.

State of the art in protein structure prediction

Ab initio prediction of protein structure from sequence: not yet

Given only the amino acid sequence, it should be possible in principle to predict protein structure directly from physico-chemical principles using, for example, molecular dynamics methods (Levitt and Warshel, 1975). In practice, however, such approaches are frustrated by the enormous complexity of the calculation (requiring many orders of magnitude more computing time than is currently feasible) and by inaccuracies in the experimental determination of basic parameters (van Gunsteren, 1993; Shortle *et al.*, 1996). Thus, the most successful structure prediction tools are knowledge based, using a combination of statistical theory and empirical rules.

European Molecular Biology Laboratory, Protein Design Group, Postfach, Meyerhofstraße 1, D-69012 Heidelberg, Germany

E-mail: rost@embl-heidelberg.de (<http://www.embl-heidelberg.de/~rost/>);
odonoghue@embl-heidelberg.de

Bridging the sequence-structure gap for more than 30% of all sequences

The gap between the number of known sequences (>100 000; Bairoch and Apweiler, 1997) and the number of known structures (~4000; Bernstein *et al.*, 1977) is widening rapidly. The most successful theoretical approach to bridging this gap is homology modelling. Given a sequence of unknown fold (denoted U), if U has significant sequence similarity to a protein of known structure (i.e. if the pairwise sequence identity is >25%), it is possible to construct an approximate 3D model which has a correct fold but inaccurate loop regions. Homology modelling effectively raises the number of 'known' 3D structures from 4000 to >15 000 (Schneider *et al.*, 1997) [Figure 1; and Figures 2 and 3 in Rost and Schneider (1997)]. However, most pairs of proteins with similar structure are remote sequence homologues with <25% pairwise sequence identity (Rost *et al.*, 1996b). These remote homologues cannot usually be recognized by conventional sequence alignments, but may sometimes be recognized by threading methods. Once a remote homology is detected, remote homology modelling may be used to construct a 3D model. This could potentially reduce the sequence-structure gap by an additional 5000-10 000 proteins (Figure 1). Now suppose we randomly choose a sequence U from one of the complete genome sequences which have recently become available, what is the likelihood that we could predict the 3D structure by homology modelling or remote homology modelling? A conservative answer would be 10%, based on the success of sequence alignment-based homology modelling (Figure 1). A very optimistic estimate would be >50%, assuming all remote homologues could be recognized (Figures 1 and 3).

Accurate prediction for 1D aspects of 3D structure

If no remote homologue can be detected for U, we are forced to simplify the prediction problem. There is a pay-off from making this simplification: using the rich diversity of information in current databases, it is possible to make very accurate 1D predictions from the sequence alone. Automatic prediction services are readily available for secondary structure, solvent accessibility, location and topology for transmembrane helices (Rost, 1996a), and coiled-coils (Lupas, 1996).

Structure prediction for known folds

Sequence alignments

Trivial for high levels of sequence identity. Any sequence analysis starts with database searches for homologous proteins by sequence alignment procedures. When pairwise sequence identity is >25-30% (for >80 residues), alignment

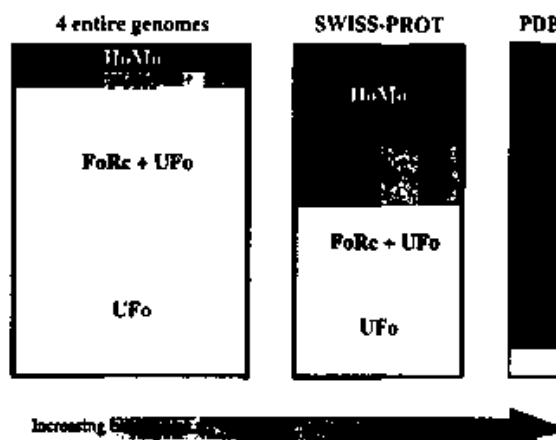


Fig. 1. Scope of structure prediction. Given any expressed protein, how likely can theory predict its 3D structure? For example, for 30% of the proteins in the current SWISS-PROT database, we can find regions for which homology modelling (*HoMo*) is applicable (Schneider *et al.*, 1997), but for the first four entirely sequenced genomes (yeast, *Harmophilus influenzae*, *Mycoplasma genitalium*, *Methanococcus jannaschii*) this is true for <10% of all proteins (Casati *et al.*, 1996). Thus, SWISS-PROT contains a bias introduced, for example, by limitations of previous sequencing techniques (note: PDB is biased to an even higher extent: for 58% of the proteins homology modelling is applicable). Estimating the contribution of fold recognition (*FoRc*) techniques is rather tricky. Here, we used the following procedure: 35% of all proteins in PDB could be subject to fold recognition techniques. This number allows two estimates: (i) the number of fold recognition targets is ~60% (35/58) of the homology modelling targets (and would thus be ~5% for the four genomes); and (ii) fold recognition covers ~80% (35/42) of would be not covered by homology modelling (which would suggest that ~70% of the four genomes could be recognized by threading). The truth, supposedly, lies in between these extremes. (Note: today threading techniques are not accurate enough for any large-scale prediction of 3D structure!) The remaining region (supposedly larger than 50%) is occupied by unknown folds (*UFo*).

procedures are usually straightforward (Bryant and Altschul, 1995; Barton, 1996; Taylor, 1996; Schneider *et al.*, 1997). For less similar protein sequences, alignments may fail (Henikoff and Henikoff, 1993; Bordo *et al.*, 1994; Vingron and Waterman, 1994).

Multiple alignments improve as data banks grow. The goal of sequence alignment procedures is to align related sequence stretches accurately and to avoid aligning unrelated stretches. The most advanced sequence alignment tools base the alignment on profiles derived from databases or particular sequence families (Deperieux and Feytmans, 1992; Vingron and Waterman, 1994; Neuwald *et al.*, 1995; Altschul and Gish, 1996; Barton, 1996; Feng and Doolittle, 1996; Gribskov and Veretnik, 1996; Henikoff and Henikoff, 1996a,b; Higgins *et al.*, 1996; Pearson, 1996; Thompson and Goldstein, 1996a; Tomii and Kanehisa, 1996). A new generation of alignment methods are based on Hidden Markov Models (Eddy, 1995; Hubbard and Park, 1995; Krogh and Mitchison, 1995; Bucher and Hofmann, 1996; Hughley and Krogh, 1996; McClure *et*

Table 1. Abbreviations and notations

1D → 3D	The classification into 1D, 2D and 3D structure is based on physical properties grouped with respect to protein structure prediction (Rost and Sander, 1994b)
3D	Three-dimensional signifies the co-ordinates of atoms, and/or residues that define a protein structure
2D	Two-dimensional describes inter-residue distances, or contacts (note: a correct prediction of all inter-residue distances is equivalent to a 3D prediction)
1D	One-dimensional summarizes properties of single residues that can be written in a 1D string, e.g. sequence, secondary structure, residue solvent accessibility, or hydrophobicity (note: prediction of, for example, helices implies prediction of some local inter-residue contacts; prediction of residue solvent accessibility implies providing an upper limit to the number of possible contact partners; thus, these properties are occasionally referred to as 2D; however, even a perfect prediction of 1D properties is, in general, not equivalent to a prediction in 3D)
NMR	Nuclear magnetic resonance
U	Sequence of unknown structure
T	Target sequence with determined 3D structure predicted to be similar to that of U
Homology modelling	Prediction of 3D structure for a protein U based on a significant pairwise sequence identity (>25%) to a protein of known structure T
Remote homology modelling	Prediction of 3D structure for a protein U based on low levels of pairwise sequence identity (<25%) to a protein of known structure T
Fold recognition	Prediction that two proteins with no significant pairwise sequence identity have similar folds (note: in principle, the fold has to be known explicitly for any of the two proteins)
Significant pairwise sequence similarity	Percentage of residues identical between two naturally evolved proteins A and B guaranteeing that A and B have similar structures; the exact number depends on the alignment length: for more than 80 aligned residues, 25% pairwise sequence identity mostly suffices to guarantee structural similarity (Schneider and Sander, 1991); there are very few exceptions of pairs with >30% sequence identity and dissimilar structures (S. Brenner, personal communication)

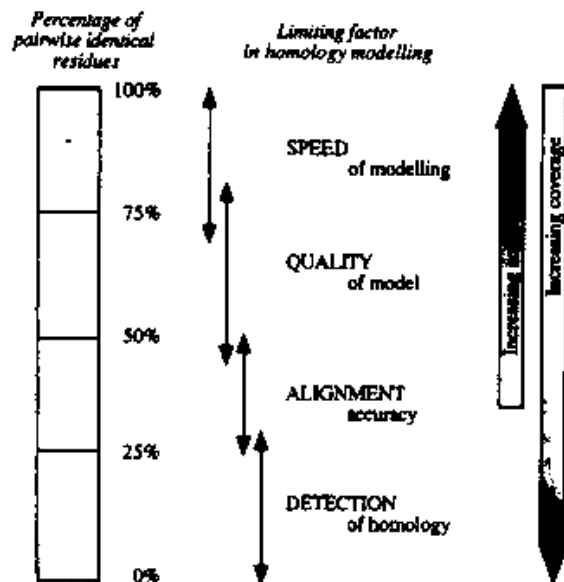


Fig. 2. Limiting steps of homology modelling. The accuracy of homology modelling is proportional to the level of pairwise sequence identity between the protein of unknown structure and its target of known structure. For high levels of identity, CPU time is the major constraint; for lower levels, loop regions become a problem (and thus the quality of the model). Below 40–50% sequence identity, errors in the sequence alignment become fatal. Below 25–30% sequence identity, fold recognition (threading) techniques have to replace (or complement) the sequence alignment procedure. [Note: the figure is partly taken from Holm *et al.* (1994).]

al., 1996) and another on genetic algorithms (Notredame and Higgins, 1996). These new methods may be more successful in the twilight zone of sequence alignments (currently 20–30% sequence identity; Doolittle, 1986) than advanced profile-based methods (Higgins *et al.*, 1996; Taylor, 1996; Thompson and Goldstein, 1996a); however, this remains to be proven.

Homology modelling

Prediction at atomic accuracy for high levels of sequence identity. The basic assumption of homology modelling is that U and the homologous template protein of known structure (T) have nearly identical backbone structure in the aligned regions. The task is to place the side chains of U correctly into the backbone of T. For levels of >70–90% sequence identity, the resulting models are quite accurate (De Filippis *et al.*, 1994; May and Blundell, 1994; Sali and Blundell, 1994; Johnson *et al.*, 1996). The limiting factor is the computation time required (Figure 2).

Prediction at intermediate accuracy for lower levels of sequence identity. For sequence identities down to ~30% sequence identity, U and T will still have the same fold (Sander and Schneider, 1991), but the number of loops inserted grows and the divergence between U and T becomes considerable (De Filippis *et al.*, 1994; May and Blundell, 1994; Chinea *et al.*, 1995; Mosimann *et al.*, 1995; Moulton *et al.*, 1995; Samudrala *et al.*, 1995; Vinals *et al.*, 1995). Modelling of loop regions is still a difficult problem (Cardozo *et al.*, 1995; Mosimann *et al.*, 1995; Sali *et al.*, 1995); even

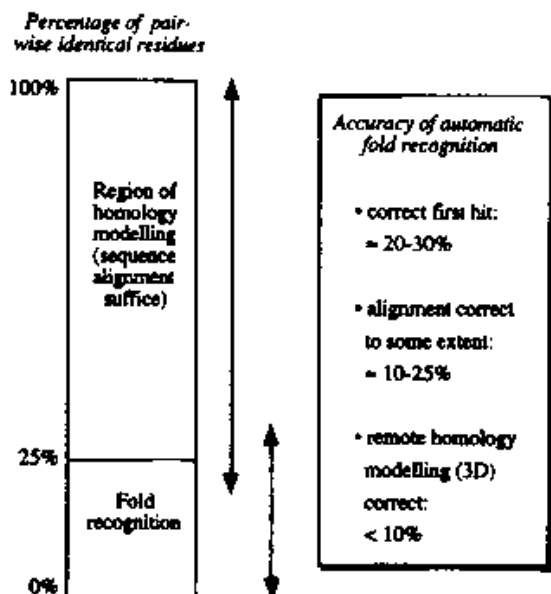


Fig. 3. Expected accuracy of automatic threading. The estimates given are controversial, but tend to be a more conservative estimate of an automatic use of fold recognition methods. The numbers refer to the following experiment. Suppose we run a threading program on all 6000 yeast proteins. For each of them we thread the sequence into a library of (say 800) folds. The program will rank the hits according to the predicted similarity to the search protein (U). How often is the first hit in the resulting list a correctly recognized remote homologue (similar fold with $<25\%$ sequence identity)? And how often is the alignment of the first hit correct? (Note: given a correct alignment, the accuracy of the finally remote homology modelled 3D prediction will be higher for higher levels of sequence identity. However, for most threading programs, the detection error is independent of the level of sequence identity.)

the best methods only rarely achieve atomic accuracy and are often completely different to the correct structure. For lower levels of pairwise sequence identity, the accuracy of the sequence alignment becomes an additional problem. A pessimistic view is that the accuracy of resulting 3D predictions is typically at the level of ribbon plots, i.e. the mutual orientation of elements such as helices and sheets can be identified. The optimistic version is that even down to levels of 30% sequence identity, homology modelling occasionally yields correct predictions at atomic resolution.

Remote homology modelling

Three difficult problems. Remote homology modelling ($<25\%$ pairwise sequence identity between the unknown structure, U, and template, T) has three obstacles to overcome: (i) the remote homology between U and T has to be detected; (ii) U and T have to be aligned correctly; (iii) the homology modelling procedure has to be tailored to the harder problem of extremely low sequence identity. In the early 1990s, there was a great deal of optimism that the first

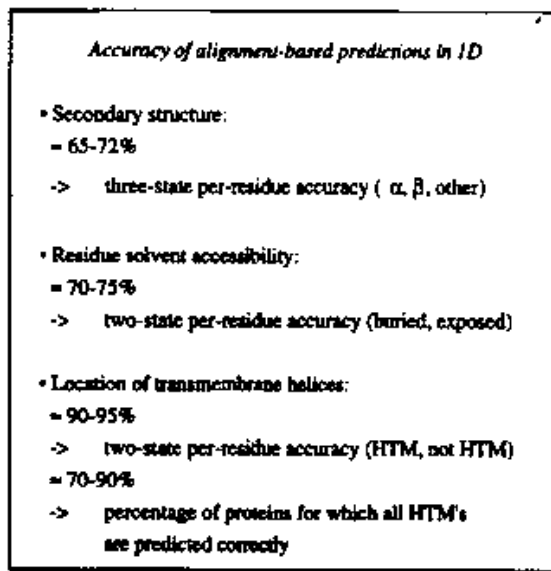


Fig. 4. Expected accuracy of predictions in 1D. Numbers were taken from the major prediction programs that make use of information contained in multiple alignments. All estimates are averages over distributions that are associated with standard deviations of the order of $\pm 10\%$.

obstacle, the detection of similar folds, would be solved by threading methods. The basic idea is to thread the sequence of U into the backbone 3D structure of T, at each step evaluating the 'fitness of sequence for structure' using environment-based (Bowie *et al.*, 1990a,b, 1991, 1996; Eisenberg *et al.*, 1991; Ouzounis *et al.*, 1993; Wilmanns and Eisenberg, 1995; Fischer and Eisenberg, 1996) or knowledge-based mean-force-potentials (Bryant and Altschul, 1995; Sippl, 1995). Most threading methods use mean-force-potentials derived from the PDB (Wodak and Rooman, 1993; Kocher *et al.*, 1994; Lerner *et al.*, 1995; Sippl, 1995). An alternative method, originally proposed a decade ago (Sheridan *et al.*, 1985), is to thread using 1D predictions. The first application of 1D threading was reported 2 years ago (Rost, 1995a); since then, several groups have investigated similar concepts, and refinements to the method (Rost, 1995b; Fischer and Eisenberg, 1996; Fischer *et al.*, 1996b; Rost *et al.*, 1996c; Russell *et al.*, 1996). The general threading problem has been shown to be N-P complete (Lathrop and Smith, 1994); consequently, heuristic algorithms must be employed, and there have been many proposed (Bowie *et al.*, 1990a, 1991; Hendlich *et al.*, 1990; Sippl, 1990; Eisenberg *et al.*, 1991; Casari and Sippl, 1992; Sippl and Weitckus, 1992; Godzik *et al.*, 1993; Lathrop and Smith, 1994; Collura *et al.*, 1995; Flöckner *et al.*, 1995; Hubbard and Park, 1995; Jones *et al.*, 1995; LeGrand *et al.*, 1995; Madej *et al.*, 1995b; Matsuo and Nishikawa, 1995; Rost, 1995b; Wang *et al.*, 1995; Wilmanns

and Eisenberg, 1995; Finkelstein and Reva, 1996; Fischer and Eisenberg, 1996; Johnson *et al.*, 1996; Jones *et al.*, 1996; Miller *et al.*, 1996; Reva and Finkelstein, 1996; Rost *et al.*, 1996c; Russell *et al.*, 1996). Has all this effort achieved any success for remote homology modelling?

Remote homologues can often be detected. First the good news: since the different mean-force-potentials which have been proposed capture different aspects of protein structure, the correct remote homologue is likely to be found by at least one of them (Lerner *et al.*, 1995). Now the bad news: so far, no single method has been able to detect the correct remote homologue for more than half of all test cases (Lerner *et al.*, 1995). For the methods which have been rigorously evaluated using large test sets, the correct remote homologue is detected in <40% of all cases (Rost, 1995b; Fischer *et al.*, 1996a,b; Rost *et al.*, 1996c; Russell *et al.*, 1996). However, this performance is clearly superior to that of traditional sequence alignments at this low level (<25%) of sequence identity (Madej *et al.*, 1995a; Rost, 1995b; Fischer and Eisenberg, 1996; Rost *et al.*, 1996c).

3D prediction by threading is still not reliable. Detecting the remote homology is only the first of the three obstacles. It appears that the second obstacle (correct alignment between U and T) is much more difficult and, unfortunately, there is no general solution so far. Thus, the final step, building a 3D model, usually fails since the modelling procedures available today cannot correct the mistakes in the alignments. As a result, there are very few publications to date which report accurate 3D predictions from threading methods (Flöckner *et al.*, 1995; Lerner *et al.*, 1995; Sippl, 1995; Rost *et al.*, 1996c). Currently, the successful use of threading methods (Hubbard and Park, 1995; Valencia *et al.*, 1995; Hubbard *et al.*, 1996) has required sceptical, expert user intervention to spot wrong hits and false alignments. It is still possible that the threading method will become the most successful structure prediction method; however, three large obstacles remain to be dealt with.

Structure prediction for unknown folds

Exploiting the protein databases

For over three decades, researchers have pursued the goal of predicting the structures of unknown folds. However, the major breakthrough has come about due to the expansion of the protein databases (Bernstein *et al.*, 1977; Bairoch and Apweiler, 1997; Benson *et al.*, 1997). Two main strategies have been developed for exploiting the information in these databases for structure prediction: (i) studying the evolution of protein families from both the sequence and structure databases; (ii) studying the physical principles of protein structure and folding from the structure databases.

Odyssey of evolution teaches us structure prediction

It appears that for most proteins, almost all residues can be changed without affecting the structure (Rost *et al.*, 1996b); however, a single, randomly chosen mutation is more likely to destabilize than to maintain a particular structure. Thus, the precise pattern of amino acid exchanges observed in a multiple sequence alignment of a protein family is highly indicative of the particular structure. These patterns constitute a fossil record of mutations preserving protein structure and function. The importance of such evolutionary information for structure prediction was realized very early (Zuckerlandl and Pauling, 1965), and has long been exploited in exceptional cases by experts (Dickerson *et al.*, 1976; Benner, 1989; Frampton *et al.*, 1989; Benner and Gerloff, 1990; Nardelli *et al.*, 1991; Musacchio *et al.*, 1992; Livingstone and Barton, 1994), as well as in automatic and systematic ways (Maxfield and Scheraga, 1979; Zvelebil *et al.*, 1987). More recently, the use of evolutionary information has grown in importance. This importance was made particularly clear recently when it was shown that the accuracy of secondary structure was improved to >70% due to the use of evolutionary information (Rost and Sander, 1993).

Prediction in 1D

Secondary structure predictions—the most accurate view of an unknown fold. Secondary structure can usually be predicted more accurately and reliably than other features of protein structure (Figure 4). Most of the recent methods for secondary structure prediction rely heavily upon evolutionary information (Rost and Sander, 1993; Livingstone and Barton, 1994; Zimmermann, 1994; Barton, 1995; Geourjon and Deléage, 1995; Mehta *et al.*, 1995; Salamov and Solovyev, 1995; Tuckwell *et al.*, 1995; Di Francesco *et al.*, 1996; Garnier *et al.*, 1996; Riis and Krogh, 1996; Rost, 1996a; Rychlewski and Godzik, 1996). A promising new concept is the use of long-range contact potentials (Frishman and Argos, 1996). From the perspective of a user of a prediction method, it is important to know the accuracy of these different methods. Unfortunately, however, in only relatively few cases has the prediction accuracy been rigorously evaluated with sufficiently large test sets (Rost and Sander, 1993; Salamov and Solovyev, 1995; Di Francesco *et al.*, 1996; Riis and Krogh, 1996; Rost, 1996a); in these cases, the sustained average accuracy of three-state predictions is just over 70%. The typical standard deviation in the prediction accuracy is ~10% [one standard deviation over more than 500–700 unique protein chains (Rost, 1996a,(WWW)c)]. For some of the prediction methods, the strength of the prediction has been shown to correlate with prediction accuracy (Rost and Sander, 1993; Munson *et al.*, 1994; Di Francesco *et al.*, 1996; Garnier *et al.*, 1996; Rost, 1996a). In practice, this enables

users to focus on regions for which predictions are more likely to be correct, e.g. ~45% of all residues are predicted at levels of accuracy comparable to homology modelling (Rost, 1996a). By comparison, earlier methods for secondary structure prediction which did not use evolutionary information, such as GOR (Garnier *et al.*, 1996), had an average accuracy of ~60%, and <10% of residues were predicted at the same accuracy as homology modelling. Thus, for methods that need highly accurate secondary structure predictions as input to predict other properties of protein structure and/or function, the best prediction methods of today are six times more useful than the methods of 5 years ago. A general feature of methods based on information from multiple sequence alignments is that errors in the alignment greatly decrease prediction accuracy (Di Francesco *et al.*, 1996; Rost, 1996a; Rost and Valencia, 1996).

Predicting residue solvent accessibility—the second step towards 3D structure? It has long been argued that if the segments of secondary structure could be accurately predicted, the 3D structure could be predicted by simply trying different arrangements of the segments in space (Cohen *et al.*, 1982; Cohen and Presnell, 1996). One criterion for assessing each arrangement could be to use predictions of residue solvent accessibility (Esposito *et al.*, 1994; Monge *et al.*, 1994; Mumenthaler and Braun, 1995; Nilges, 1995; Galaktionov and Marshall, 1996). Various methods for predicting accessibility have been developed recently (Benner *et al.*, 1994; Rost and Sander, 1994a; Wako and Blundell, 1994; Thompson and Goldstein, 1996b). Although residue solvent accessibility is not as well conserved within structural families as is secondary structure [Rost and Sander, 1994a; Russell and Barton, 1994; Rost, 1996a,(WWW)d], prediction accuracy is much improved by including evolutionary information (Rost, 1996a; Thompson and Goldstein, 1996b). Predictions of solvent accessibility have also been used successfully for prediction-based threading (Rost, 1995b; Rost *et al.*, 1996c; Russell *et al.*, 1996), and as a basis for predicting functional sites (Cornette *et al.*, 1995; Hansen *et al.*, 1995, 1996).

Membrane proteins—successful prediction in the absence of experimental information. Integral membrane proteins are an important class of proteins for which it is very difficult to obtain atomic resolution information about 3D structure. Two main classes of membrane proteins are recognized (von Heijne, 1996): proteins with long (17–27 residues) transmembrane helices spanning the membrane; and porins, 16-fold β -barrel proteins which form a pore through the membrane. Developing prediction methods for the porins is problematic as there is very little experimental information currently available; some attempts have been made using sequence profiles (von Heijne, 1996). Predicting the locations

of the transmembrane helices is a task comparable to secondary structure prediction. Very accurate predictions have been achieved by combining expert rules, hydrophobicity analyses, and statistics (Jones *et al.*, 1994; Persson and Argos, 1994; von Heijne, 1994, 1996; Neuwald *et al.*, 1995; Efremov and Vergoten, 1996; Fariselli and Casadio, 1996; Persson and Argos, 1996; Rost *et al.*, 1996a). A separate task is the prediction of specific peptide signals (Nielsen *et al.*, 1996). The use of multiple alignment information has been shown to improve prediction accuracy (Persson and Argos, 1994, 1996; Rost *et al.*, 1995, 1996a; Rost, 1996a). Currently, the best methods predict all transmembrane helices correctly for ~85–90% of all test proteins (Rost *et al.*, 1996a; von Heijne, 1996). Further methods have been developed for predicting transmembrane-helix topology, i.e. the orientation of the helices with respect to the membrane (Jones *et al.*, 1994; Persson and Argos, 1996; Rost *et al.*, 1996a; von Heijne, 1996).

Prediction in 2D

A hard problem, but the stakes are high. Given only a small fraction of the inter-residue distances, it is possible to calculate the 3D structure using either metric matrix distance geometry or simulated annealing by molecular dynamics (Bohr *et al.*, 1993; Brünger and Nilges, 1993; Aszodi *et al.*, 1995; Galaktionov and Marshall, 1996; Nilges, 1996). Can inter-residue contacts be predicted accurately from the sequence alone? And can evolutionary information help out once again?

Distinction between different models possible. Two attempts have been made to use evolutionary information for the prediction of inter-residue contacts. The first was a method for predicting contacts between β strands from multiple sequence alignments and alignment-based predictions of secondary structure (Hubbard, 1994; Hubbard and Park, 1995). Thus, the method is limited in its applicability, but would be helpful as part of a bouquet of other prediction methods (Hubbard *et al.*, 1996). The second attempt has been in the development of a group of methods for the prediction of inter-residue contacts from correlated mutations (Goebel *et al.*, 1994; Shindyalov *et al.*, 1994; Taylor and Hatrick, 1994; Kreisberg *et al.*, 1995). In general, the prediction accuracy is rather poor, with a direct trade-off between the Scylla of predicting enough contacts, and the Charibdis of predicting only correct ones, e.g. taking 5% of the best-predicted long-range contacts (sequence separation above 10 residues), the accuracy prediction is ~50% (A.Valencia, personal communication). Although this level of accuracy is not high, it is sufficient to distinguish between correct and incorrect alignments in threading experiments (A.Valencia, in preparation).

Prediction in 3D

Sisyphus again? In the 1994 Asilomar meeting, none of the 3D *ab initio* methods were able to predict the correct protein structure. Since that time, new methods have been proposed which indicate possible directions for the future. Several groups have obtained promising results using distance geometry methods (Aszodi *et al.*, 1995; Mumenthaler and Braun, 1995; Nilges, 1995); these methods may be particularly powerful in combination with 2D contact predictions. Nilges and Brünger (1991, 1993) have achieved atomic accuracy in an *ab initio* prediction of the GCN4 leucine zipper using a hybrid molecular dynamics/simulated annealing search strategy. Recently, equally accurate models for three leucine zippers were obtained with faster calculations based on mean-force-potentials (O'Donoghue and Nilges, 1997). Simplified force fields in combination with dynamic optimization strategies have yielded promising, but still relatively inaccurate results (Elofsson *et al.*, 1995; Pedersen and Moulis, 1996a,b). Srinivasan and Rose (1995) have reported very encouraging results with their hierarchical search method; however, they have not repeated the initial claims, so it may be that the initial report was too optimistic. In addition to these methods, many research groups have been working at improving their more established methods since 1994, so we can only wait for the outcome of the next Asilomar meeting.

Recognizing incorrect structures. We consider that the single most important theoretical advance in 3D prediction in recent years has been the development of mean-force-potentials. Before these potentials, structure prediction was normally done with 'physical' potentials, i.e. bonds, angles, torsion angles, and van der Waals as well as electrostatic non-bonded terms which describe the internal energy of the molecule (van Gunsteren, 1993). In contrast, the mean-force-potentials, derived from databases of protein structure (Sippl, 1990), attempt to describe the free energy of the molecule. The physical potentials have been used very successfully to refine experimentally determined structures (Brünger and Nilges, 1993; Nilges, 1996). However, these terms cannot distinguish between a native fold and a grossly misfolded structure (Novotny *et al.*, 1988; Sippl, 1990). In contrast, mean-force-potentials of pairwise residue distances are quite successful in fold recognition, as well as remote homology modelling (Sippl, 1995). It remains to be seen how best to combine these two different potentials. In one pilot study on the use of mean-force-potentials for 3D structure prediction, the best results were obtained by combining both potentials (O'Donoghue and Nilges, 1997).

Extracting principles about structure formation from structures? The mean-force-potential approach has recently been extended to study protein folding. Both the physical basis and

the general characteristics of protein folding remain controversial (Honig and Cohen, 1996; Israelachvili and Wennerström, 1996). Simulations and other studies indicate that the free-energy balance of hydrogen bond formation is close to zero, or slightly unfavourable (Yang and Honig, 1995a,b), and that a specific fold is selected primarily by side-chain interactions (Honig and Cohen, 1996). Recently, Sippl *et al.* have extended the concept of deriving mean-force potentials to a formalism of describing Helmholtz free energies of atom pair interactions (Sippl, 1996; Sippl *et al.*, 1996). The formalism starts with the following two assumptions: (i) that protein structures can be described by Helmholtz free energies (or mean-force-potentials) and (ii) that the distribution of intramolecular distances in experimentally determined protein structures does, on average, not deviate substantially from the corresponding distribution in native proteins. To normalize the absolute free energy contributions, the ideal gas is chosen (no internal interactions). Without any further assumptions or approximations, atom-atom mean-force-potentials are derived from a data set of known protein structures. The resulting Helmholtz mean-force-potentials unravel interesting principles about protein structure formation. (i) Backbone H-bonds (except for the α -helix interaction $O_i \dots N_{i+4}$) do not contribute to the thermodynamic stability of native folds. (ii) H-bond formation (except for $O_i \dots N_{i+4}$) requires energy input that is regained when H-bonds are formed. Once formed, H-bonds are locked in a deep, narrow minimum. (iii) The energy gain of forming one ionic or two hydrophobic contacts can provide roughly the activation energy required for forming a H-bond. Both the eloquence and the conclusions of the approach have prompted strong criticism, even unanimous rejection of these findings. Do we witness an error in a method laid out to spot errors, or the beginning of a new era of force fields? Further applications of these mean-force-potentials will be needed to answer this question.

Uses of prediction methods

Homology modelling (including remote homology modelling, when it works) has proven to be the most useful prediction tool. Homology models are used to suggest mutation experiments to investigate function, or to facilitate crystallization (for X-ray diffraction studies), or to reduce aggregation (for NMR spectroscopy). Furthermore, homology-based models are often used to provide starting structures for molecular replacement in X-ray crystallography.

In the absence of known structures with significant sequence identity to a protein of interest, predictions of residue solvent accessibility can be used to investigate function, and to suggest which residues to mutate to facilitate crystallization or to reduce aggregation. Accurate prediction of secondary structure can help with X-ray diffraction (e.g. the GroEL

crystal structure was derived making use of secondary structure predictions for the molecular replacement search). In principle, the early stages of NMR frequency assignment could also be aided by knowledge of the secondary structure, although this has not been attempted.

1D predictions and predictions of topology (transmembrane, coiled-coil) have proven to be quick and accurate enough for the analysis of entire genomes (Koonin *et al.*, 1996; Odgren *et al.*, 1996; Rost, 1996b; Rost *et al.*, 1996a).

Mean-force-potentials can assist experimental structure determination by spotting stresses or anomalies (often errors) in protein structures (Sippl, 1993). A set of methods based on particular statistics has recently been tailored to manage exactly this task (Laskowski *et al.*, 1993; Gray *et al.*, 1996; Hoof *et al.*, 1996).

Conclusions

Native 3D structures of proteins are encoded by a linear sequence of amino acid residues. To predict 3D structure from the sequence is a task challenging enough to have occupied a generation of researchers. Have they finally succeeded in their goal? The bad news is: no, we still cannot predict structure for any sequence. The good news is: we have come closer, and growing databases facilitate the task.

(i) Evolutionary information is successfully used for predictions of secondary structure, solvent accessibility and transmembrane helices. These predictions of protein structure in 1D are significantly more accurate, and more useful than 5 years ago.

(ii) Databases of protein structure can be used to derive mean-force-potentials. Residue pair mean-force-potentials are extremely valuable for the detection of remote homologues, and for the distinction between alternative models (generated by theory or experiment). Moreover, the database of protein structures contains a record of structure formation that has recently been unravelled by the derivation of atom-atom mean-force-potentials (Sippl, 1996; Sippl *et al.*, 1996).

(iii) Homology modelling allows predictions of 3D structure for about one-tenth of all expressed proteins (Figure 1).

(iv) Recent improvements in fold recognition (threading), and alignment techniques enable remote homology modelling for another considerable fraction of the expressed proteins (Figure 1).

All of the above breakthroughs were achieved in the last 6 years. Thus, although we still cannot solve the general prediction problem, progress has been made. In general, however, we could ask the question: is it worth persevering with structure prediction, given that it is clearly such a difficult task? The answer is: yes. The methods which have spun off from structure prediction have already given us considerable insight into the first four complete genomes.

Perseverance with structure prediction will yield fruit in about 5 years time when the human genome will be known.

Note added in proof

John Moult (CARB, Washington, DC) has initiated a meeting for the assessment of prediction methods: theoreticians made predictions for the structure proteins before the structure was experimentally determined. After experimental determination of some of the target structures, the accuracy of the predictions was assessed in a meeting in Asilomar, California (December 1994). This experiment comprises one of the most important events to separate the chaff from the wheat in the field of structure prediction. We wrote our review a few weeks before the second Asilomar meeting (CASP2). Has CASP2 falsified our optimism or scepticism? We dare the following conclusions: (i) structure prediction remains an unsolved problem; (ii) new methods for predicting unknown folds may still be regarded as promising, but no major breakthrough was witnessed; (iii) threading methods looked more promising in the optimistic light of the few CASP2 threading targets (than they appear in the light of this review); (iv) sufficiently skilled, and motivated, experts (such as Alexei Murzin, LMB, Cambridge) can use today's prediction methods to find and relatively accurately align even remotely homologous proteins.

References

- Altshul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.* **246**, 460-480.
- Ahlinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223-230.
- Azodi,A., Cradwell,M.J. and Taylor,W.R. (1995) In Bohr,H. and Brunak,S. (eds), *Protein Folds: A Distance Based Approach*. CRC Press, Boca Raton, FL, pp. 85-97.
- Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.*, **25**, 31-36.
- Barton,G.J. (1995) Protein secondary structure prediction. *Curr. Opin. Struct. Biol.*, **5**, 372-376.
- Barton,G.J. (1996) In Sternberg,M.J.E. (ed.), *Protein Structure Prediction*. Oxford University Press, Oxford, pp. 31-64.
- Benner,S.A. (1989) Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enzyme Regul.*, **28**, 219-236.
- Benner,S.A. and Gerloff,D. (1990) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.*, **31**, 121-181.
- Benner,S.A., Badoe,I., Cohen,M.A. and Gerloff,D.L. (1994) *Bona fide prediction of aspects of protein conformation. J. Mol. Biol.*, **238**, 926-958.
- Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) GenBank. *Nucleic Acids Res.*, **25**, 1-6.
- Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Bruce,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535-542.
- Bohr,J., Bohr,H., Brunak,S., Cotterill,R.M.J., Fredholm,H., Laurrup,B. and Petersen,S.B. (1993) Protein structures from distance inequalities. *J. Mol. Biol.*, **231**, 861-869.
- Bordo,D., Djinovic,K. and Bolognesi,M. (1994) Conserved patterns in the Cu, Zn superoxide dismutase family. *J. Mol. Biol.*, **238**, 366-386.

- Boek, P. and Gibson, T.J. (1996) Applying motif and profile searches. *Methods Enzymol.* **266**, 162-184.
- Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T. (1990a) Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, **7**, 257-264.
- Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A. and Sauer, R.T. (1990b) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306-1310.
- Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-169.
- Bowie, J.U., Zhang, K., Wilmanns, M. and Eisenberg, D. (1996) Three-dimensional profiles for measuring compatibility of amino acid sequence with three-dimensional structure. *Methods Enzymol.* **266**, 598-616.
- Bringer, A.T. and Nilges, M. (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* **26**, 49-125.
- Bryant, S.H. and Altschul, S.F. (1995) Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236-244.
- Bucher, P. and Hofmann, K. (1996) In States, D., Agarwal, P., Guasterland, T., Hunter, L. and Smith, R.F. (eds), *Fourth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, St Louis, MO, pp. 44-51.
- Bult, C.J. et al. (1996) Complete genome sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058-1073.
- Cardozo, T., Totrov, M. and Abagyan, R. (1995) Homology modeling by the ICM method. *Proteins*, **23**, 403-414.
- Casari, G. and Sippl, M.J. (1992) Structure-derived hydrophobic potential. *J. Mol. Biol.* **224**, 725-732.
- Casari, G., De Daruvar, A., Sander, C. and Schneider, R. (1996) Bioinformatics and the discovery of gene function. *Trends Genet.* **12**, 244-245.
- Chinea, G., Padron, G., Hooft, R.W.W., Sander, C. and Vriend, G. (1995) The use of position-specific rotamers in model building by homology. *Proteins*, **23**, 415-421.
- Cohen, F.E. and Presnell, S.R. (1996) In Sternberg, M.J.E. (ed.), *Protein Structure Prediction*. Oxford University Press, Oxford, pp. 207-228.
- Cohen, F.E., Sternberg, M.J.E. and Taylor, W.R. (1982) Analysis and prediction of the packing of α -helices against a β -sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**, 821-862.
- Collier, W.V.P., Eldridge, M.D., Firth, M.A. and Murray, C.W. (1995) Protein fold recognition by threading: comparison of algorithms and analysis of results. *Protein Eng.* **8**, 1197-1204.
- Comette, J.L., Margalit, H., Bertozsky, J.A. and DeLisi, C. (1995) Periodic variation in side-chain polarities of T-cell antigenic peptides correlates with their structure and activity. *Proc. Natl Acad. Sci. USA*, **92**, 8368-8372.
- Corrales, F.J. and Fersht, A.R. (1996) Kinetic significance of GroEL₁₄ (GroES)₂ complexes in molecular chaperone activity. *Folding Design*, **1**, 265-273.
- De Filippin, V., Sander, C. and Vriend, G. (1994) Predicting local structural changes that result from point mutations. *Protein Eng.* **7**, 1203-1208.
- Deperieu, E. and Feytmans, E. (1992) MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput. Applic. Biosci.* **8**, 501-509.
- Di Francesco, V., Garnier, J. and Munson, P.J. (1996) Improving protein secondary structure prediction with aligned homologous sequences. *Protein Sci.* **5**, 106-113.
- Dickerson, R.E., Timkovich, R. and Almasy, R.J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism. *J. Mol. Biol.* **100**, 473-491.
- Doolittle, R.F. (1986) *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, CA.
- Doolittle, R.F. (1996) *Computer Methods for Macromolecular Sequence Analysis*. Academic Press, San Diego, CA.
- Eddy, S.R. (1995) In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S. (eds), *Third International Conference on Intelligent Systems for Molecular Biology (ISMB)*. Menlo Park, CA, pp. 114-120.
- Efremov, R.G. and Vergoten, G. (1996) Recognition of transmembrane α -helical segments with environmental profiles. *Protein Eng.* **9**, 253-263.
- Eisenberg, D., Luthy, R. and McLachlan, A.D. (1991) Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, **10**, 229-239.
- Einfasson, A., Le Grand, S.M. and Eisenberg, D. (1995) Local moves: an efficient algorithm for simulation of protein folding. *Proteins*, **23**, 73-82.
- Esposito, G., Lesk, A.M., Molinar, H., Mona, A., Nicolai, N. and Pastore, A. (1994) In Bohr, H. and Brunak, S. (eds), *Protein Structure by Distance Analysis*. IOS Press, Amsterdam, pp. 51-63.
- Farielli, P. and Casadio, R. (1996) HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Comput. Applic. Biosci.* **12**, 41-48.
- Feng, D.-F. and Doolittle, R.F. (1996) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* **266**, 368-382.
- Finkelstein, A.V. and Reva, B.A. (1996) Search for the most stable folds of protein chains: I. Application of a self-consistent field theory to a problem of protein three-dimensional structure prediction. *Protein Eng.* **9**, 387-397.
- Fischer, D. and Eisenberg, D. (1996) Fold recognition using sequence-derived properties. *Protein Sci.* **5**, 947-955.
- Fischer, D., Eklöfsson, A., Rice, D. and Eisenberg, D. (1996a) In Hunter, L. and Klein, T. (eds), *Pacific Symposium on Biocomputing*, Hawaii, 1996. World Scientific Publishing Co., Singapore, pp. 300-318.
- Fischer, D., Rice, D.W., Bowie, J.U. and Eisenberg, D. (1996b) Assigning amino acid sequences to 3D protein folds. *FASEB J.* **10**, 126-136.
- Fleischmann, R.D. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496-512.
- Flöckner, H., Braxenthaler, M., Lackner, P., Janz, M., Ortner, M. and Sippl, M.J. (1995) Progress in fold recognition. *Proteins*, **23**, 376-386.
- Frampton, J., Leutz, A., Gibson, T.J. and Graf, T. (1989) DNA-binding domain ancestry. *Nature*, **342**, 134.
- Fraser, C.M. et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397-403.
- Frühman, D. and Argos, P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9**, 133-142.
- Galaktionov, S.G. and Marshall, G.R. (1996) Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D. Preprint, Institute for Biomedical Computing, Washington University, St Louis, MO.
- Garnier, J., Gibrat, J.-F. and Robson, B. (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540-553.
- Geourjon, C. and Deléage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Applic. Biosci.* **11**, 681-684.
- Godzik, A., Kolinski, A. and Skolnick, J. (1993) De novo and inverse folding predictions of protein structure and dynamics. *J. Comput. Aided Mol. Design*, **7**, 397-438.
- Goebel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309-317.
- Goffeau, A. et al. (1996) Life with 6000 genes. *Science*, **274**, 546-567.
- Gray, P.M.D., Kemp, G.J.L., Rawlings, C.J., Brown, N.P., Sander, C., Thornton, J.M., Orreign, C.M., Wodak, S.J. and Richelle, J. (1996) Macromolecular structure information and databases. *Trends Biochem. Sci.* **21**, 251-256.
- Grisham, M. and Veretnik, S. (1996) Identification of sequence patterns with profile analysis. *Methods Enzymol.* **266**, 198-227.
- Hansen, J.E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J.O., Hansen, J.-E.S. and Brunak, S. (1995) Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc: polypeptide N-acetyl-galactosaminyltransferase. *Biochem. J.* **308**, 801-813.
- Hansen, J.E., Lund, O., Nielsen, J.O., Brunak, S. and Hansen, J.-E.S. (1996) Prediction of the secondary structure of HIV-1 gp120. *Proteins*, **25**, 1-11.
- Hartl, F.-U., Hlodan, R. and Langer, T. (1994) Molecular chaperones in protein folding: the art of avoiding sticky situations. *Trends Biochem. Sci.* **19**, 20-25.
- Hendlich, M., Lackner, P., Weickus, S., Flöckner, H., Froschauer, R., Gottschner, K., Casari, G. and Sippl, M.J. (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low

- energy conformations from potentials of mean force. *J. Mol. Biol.*, **216**, 167-180.
- Henikoff, J.G. and Henikoff, S. (1996a) Blocks database and its applications. *Methods Enzymol.*, **266**, 88-104.
- Henikoff, J.G. and Henikoff, S. (1996b) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Applic. Biosci.*, **12**, 135-143.
- Henikoff, S. and Henikoff, J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49-61.
- Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383-402.
- Holm, L., Rost, B., Sander, C., Schneider, R. and Vriend, G. (1994) In Doniach, S. (ed.), *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions*. Plenum Press, New York, pp. 277-296.
- Honig, B. and Cohen, F.E. (1996) Adding backbone to protein folding: why proteins are polypeptides. *Folding Design*, **1**, R17-R20.
- Hooft, R.W.W., Vriend, G., Sander, C. and Abola, E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
- Hubbard, T. et al. (1996) Update on protein structure prediction: results of the 1995 IRBM workshop. *Folding Design*, **1**, R55-R63.
- Hubbard, T.J.P. (1994) In Hunter, L. (ed.), *27th Hawaii International Conference on System Sciences*. IEEE Society Press, Maui, HI, pp. 336-344.
- Hubbard, T.J.P. and Park, J. (1995) Fold recognition and ab initio structure predictions using Hidden Markov models and β -strand pair potentials. *Proteins*, **23**, 398-402.
- Hughley, R. and Krogh, A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Applic. Biosci.*, **12**, 95-107.
- Israelachvili, J. and Wennerström, H. (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature*, **379**, 219-225.
- Johanson, M.S., May, A.C.W., Rodionov, M. and Overington, J.P. (1996) Discrimination of common protein folds: application of protein structure to sequence/structure comparisons. *Methods Enzymol.*, **266**, 575-598.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038-3049.
- Jones, D.T., Miller, R.T. and Thornton, J.M. (1995) Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins*, **23**, 387-397.
- Jones, D.T., Orengo, C.A. and Thornton, J.M. (1996) In Sternberg, M.J.E. (ed.), *Protein Structure Prediction*. Oxford University Press, Oxford, pp. 173-206.
- Kocher, J.-P., Rooman, M.J. and Wodak, S.J. (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.*, **235**, 1598-1613.
- Koonin, E.V., Tatusov, R.L. and Rudd, K.E. (1996) Protein sequence comparison at genome scale. *Methods Enzymol.*, **266**, 295-322.
- Krisberg, R., Buchner, V. and Arad, D. (1995) Paired natural cysteine mutation mapping: aid to constraining models of protein tertiary structure. *Protein Sci.*, **4**, 2405-2410.
- Krogh, A. and Mitchison, G. (1995) In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S. (eds), *Third International Conference on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, Cambridge, UK, pp. 215-221.
- Laskowski, R.A., Moss, D.S. and Thornton, J.M. (1993) Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, **231**, 1049-1067.
- Lathrop, R.H. and Smith, T.F. (1994) In Hunter, L. (ed.), *27th Hawaii International Conference on System Sciences*, Los Alamos, CA. IEEE Computer Society Press, Wailea, HI, pp. 365-374.
- Lattman, E.E. (1994) Protein crystallography for all. *Proteins*, **18**, 103-106.
- LeGrand, S., Elofsson, A. and Eisenberg, D. (1995) In Bohr, H. and Brunak, S. (eds), *Protein Folds: A Distance Based Approach*. CRC Press, Boca Raton, FL, pp. 105-113.
- Lerner, C.M.-R., Rooman, M.J. and Wodak, S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins*, **23**, 337-355.
- Levin, M. and Warshel, A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694-698.
- Livingstone, C.D. and Barton, G.J. (1994) Secondary structure prediction from multiple sequence data: blood clotting factor XIII and Yersinia protein-tyrosine phosphatase. *Int. J. Peptide Protein Res.*, **44**, 239-244.
- Lupas, A. (1996) Coiled coils: new structures and new functions. *Trends Biochem. Sci.*, **21**, 373-382.
- Madej, T., Boguski, M.S. and Bryant, S.H. (1995a) Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Lett.*, **373**, 12-18.
- Madej, T., Gibral, J.-F. and Bryant, S.H. (1995b) Threading a database of protein cores. *Proteins*, **23**, 356-369.
- Matsuo, Y. and Nishikawa, K. (1995) Assessment of a protein fold recognition method that takes into account four physicochemical properties: side-chain packing, solvation, hydrogen-binding, and local conformation. *Proteins*, **23**, 370-375.
- Maxfield, F.R. and Scheraga, H.A. (1979) Improvements in the prediction of protein topography by reduction of statistical errors. *Biochemistry*, **18**, 697-704.
- May, A.C.W. and Blundell, T.L. (1994) Automated comparative modelling of protein structures. *Curr. Opin. Biotech.*, **5**, 353-360.
- McClure, M.A., Smith, C. and Elton, P. (1996) In States, D., Agarwal, P., Gausterland, T., Hunter, L. and Smith, R.F. (eds), *Fourth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, St Louis, MO, pp. 135-154.
- Mehta, P.K., Heringa, J. and Argos, P. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.*, **4**, 2517-2525.
- Miller, R.T., Jones, D.T. and Thornton, J.M. (1996) Protein fold recognition by sequence threading: tools and assessment techniques. *FASEB J.*, **10**, 171-178.
- Monge, A., Friesner, R.A. and Honig, B. (1994) An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA*, **91**, 5027-5029.
- Moumouni, S., Melisshko, R. and James, M.N.G. (1995) A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins*, **23**, 301-317.
- Moult, J., Pedersen, J.T., Judson, R. and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii-iv.
- Mumenthaler, Ch. and Braun, W. (1995) Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.*, **4**, 863-871.
- Munson, P.J., Di Francesco, V. and Porrelli, R. (1994) In Hunter, L. (ed.), *27th Hawaii International Conference on System Sciences*. Los Alamos, CA. IEEE Computer Society Press, Wailea, HI, pp. 375-384.
- Musacchio, A., Gibson, T., Lehto, V.-P. and Saraste, M. (1992) SH3—an abundant protein domain in search of a function. *FEBS Lett.*, **307**, 55-61.
- Nardelli, J., Gibson, T.J., Vesque, C. and Charnay, P. (1991) Base sequence discrimination by zinc-finger DNA-binding domains. *Nature*, **349**, 175-178.
- Neuwald, A.F., Liu, J.S. and Lawrence, C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618-1631.
- Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **24**, 165-177.
- Nilges, M. (1995) Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.*, **248**, 645-660.
- Nilges, M. (1996) Structure calculation from NMR data. *Curr. Opin. Struct. Biol.*, **6**, 617-623.
- Nilges, M. and Brünger, A.T. (1991) Automated modelling of coiled coils. Application to the GCN4 dimerization region. *Protein Eng.*, **4**, 649-659.
- Nilges, M. and Brünger, A.T. (1993) Successful prediction of the coiled coil geometry of the GCN4 leucine zipper domain by simulated annealing: comparison to the X-ray structure. *Proteins*, **6**, 133-146.
- Notredame, C. and Higgins, D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515-1524.
- Novotny, J., Rashin, A.A. and Brucoleri, R.E. (1988) Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, **4**, 19-30.

- Odgren, P.R., Harvie, L.W.J. and Fey, E.G. (1996) Phylogenetic occurrence of coiled coil proteins: implications for tissue structure in metazoa via a coiled coil tissue matrix. *Proteins*, **24**, 467-484.
- O'Donoghue, S.I. and Nilges, M. (1997) Use of mean force potentials for tertiary structure prediction: successful prediction of leucine zippers. *Folding Design*, submitted.
- Oliver, S. et al. (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38-46.
- Ouzounis, C., Sander, C., Scharf, M. and Schneider, R. (1993) Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.*, **232**, 805-825.
- Pearson, W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.*, **266**, 227-258.
- Pedersen, J.T. and Moult, J. (1996a) Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.*, **6**, 227-231.
- Pedersen, J.T. and Moult, J. (1996b) Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.*, submitted.
- Persson, B. and Argos, P. (1994) Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. *J. Mol. Biol.*, **237**, 182-192.
- Persson, B. and Argos, P. (1996) Topology prediction of membrane proteins. *Protein Sci.*, **5**, 363-371.
- Reva, B.A. and Finkelstein, A.V. (1996) Search for the most stable folds of protein chains: II. Computation of stable architectures of β -proteins using a self-consistent molecular field theory. *Protein Eng.*, **9**, 399-411.
- Rits, S.K. and Krogh, A. (1996) Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol.*, **3**, 163-183.
- Rost, B. (1995a) In Bohr, H. and Brunak, S. (eds), *Protein Folds: A Distance Based Approach*. CRC Press, Boca Raton, FL, pp. 132-151.
- Rost, B. (1995b) In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S. (eds), *Third International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, CA, AAAI Press, Cambridge, UK, pp. 314-321.
- Rost, B. (1996a) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525-539.
- Rost, B. (1996b) Sneaking in genomes for helical transmembrane proteins. Talk presented at 'Distance based approaches to protein structure prediction III', Copenhagen, Denmark. EMBL, Heidelberg, Germany, manuscript in preparation.
- Rost, B. (1996c) Expected prediction accuracy of PHD. WWW document (<http://www.embl-heidelberg.de/~rost/Res/96D-ExpAccuracyPHD.html>). EMBL Heidelberg, Germany.
- Rost, B. (1996d) Protein fold recognition by merging 1D structure prediction and sequence alignments. WWW document (<http://www.embl-heidelberg.de/~rost/Papers/96PrTopis.html>). EMBL Heidelberg, Germany.
- Rost, B. and Sander, C. (1993) Secondary structure prediction of all-helical proteins in two states. *Protein Eng.*, **6**, 831-836.
- Rost, B. and Sander, C. (1994a) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216-226.
- Rost, B. and Sander, C. (1994b) Structure prediction of proteins—where are we now? *Curr. Opin. Biotech.*, **5**, 372-380.
- Rost, B. and Sander, C. (1996) Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 113-136.
- Rost, B. and Schneider, R. (1996) In Ashman, K. (ed.), *Core Techniques in Biochemistry*. Springer, Heidelberg, in press.
- Rost, B. and Valencia, A. (1996) Pitfalls of protein sequence analysis. *Curr. Opin. Biotech.*, **7**, 457-461.
- Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.*, **4**, 521-533.
- Rost, B., Casadio, R. and Fariselli, P. (1996a) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704-1718.
- Rost, B., O'Donoghue, S. and Sander, C. (1996b) Protein structures evolve at random—almost. Submitted.
- Rost, B., Schneider, R. and Sander, C. (1996c) Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, in press.
- Russell, R.B. and Barton, G.J. (1994) Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.*, **244**, 332-350.
- Russell, R.B., Copley, R.R. and Barton, G.J. (1996) Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.*, **259**, 349-365.
- Rychlewski, L. and Godzik, A. (1996) Secondary structure predictions: in quest of forces that shape the local protein structure. Preprint, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, USA.
- Salamov, A.A. and Solovyev, V.V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *J. Mol. Biol.*, **247**, 11-15.
- Sali, A. and Blundell, T. (1994) In Bohr, H. and Brunak, S. (eds), *Protein Structure by Distance Analysis*. IOS Press, Amsterdam, pp. 64-87.
- Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318-326.
- Samudran, R., Pedersen, J.T., Zhou, H.-B., Luo, R., Fidelis, K. and Moult, J. (1995) Confronting the problem of interconnected structural changes in the comparative modeling of proteins. *Proteins*, **23**, 327-336.
- Sander, C. and Schneider, R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56-68.
- Schneider, R., de Ruyvar, A. and Sander, C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226-230.
- Sheridan, R.P., Dixon, J.S. and Venkataraghavan, R. (1985) Generating plausible protein folds by secondary structure similarity. *Int. J. Peptide Protein Res.*, **25**, 132-143.
- Shindyalov, I.N., Kolchanov, N.A. and Sander, C. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, **7**, 349-358.
- Shortle, D., Wang, Y., Gillespie, J. and Wrabl, J.O. (1996) Protein folding for realists: a timeless phenomenon. *Protein Sci.*, **5**, 991-1000.
- Sippl, M.J. (1990) The calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures of globular proteins. *J. Mol. Biol.*, **213**, 859-883.
- Sippl, M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355-362.
- Sippl, M.J. (1995) Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229-235.
- Sippl, M.J. (1996) Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biol.*, **260**, 644-648.
- Sippl, M.J. and Weickus, S. (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, **13**, 258-271.
- Sippl, M.J., Ormer, M., Jaritz, M., Lackner, P. and Flöckner, H. (1996) Helmholtz free energies of atom pair interactions in proteins. *Folding Design*, **1**, 289-298.
- Srinivasan, R. and Rose, G.D. (1995) LINUS: a hierarchical procedure to predict the fold of a protein. *Proteins*, **22**, 81-99.
- Sternberg, M.J.E. (1996) *Protein Structure Prediction*. Oxford University Press, Oxford.
- Taylor, W.R. (1996) Multiple protein sequence alignment: algorithms and gap insertion. *Methods Enzymol.*, **266**, 343-367.
- Taylor, W.R. and Harbeck, K. (1994) Compensating changes in protein multiple sequence alignment. *Protein Eng.*, **7**, 341-348.
- Thompson, M.J. and Goldstein, R.A. (1996a) Constructing amino acid residue substitution classes maximally indicative of local protein structure. *Proteins*, **25**, 28-37.
- Thompson, M.J. and Goldstein, R.A. (1996b) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins*, **25**, 38-47.
- Tomii, K. and Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27-36.
- Tuckwell, D.S., Humphries, M.J. and Brass, A. (1995) Protein secondary structure prediction by the analysis of variation and conservation in multiple alignments. *Comput. Applic. Biosci.*, **11**, 627-632.
- Valencia, A., Hubbard, T.J., Muga, A., Ba-ueks, S., Llorca, O., Carrascosa, J. and Valpuesta, J.M. (1995) Prediction of the structure of GroES and its interaction with GroEL. *Proteins*, **22**, 199-209.
- van Gunsteren, W.F. (1993) Molecular dynamics studies of proteins. *Curr. Opin. Struct. Biol.*, **3**, 167-174.
- Vinals, C., De Boffe, X., Depiereux, E. and Feytmans, E. (1995) Knowledge-

- based modeling of the D-lactate dehydrogenase three-dimensional structure. *Proteins*, **21**, 307-318.
- Vingron, M. and Waterman, M.S. (1994) Sequence alignment and penalty choice. *J. Mol. Biol.*, **238**, 1-12.
- von Heijne, G. (1994) Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167-192.
- von Heijne, G. (1996) In Sternberg, M.J.E. (ed.), *Protein Structure Prediction*. Oxford University Press, Oxford, pp. 101-110.
- Wako, H. and Bhandell, T.L. (1994) Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins I. Solvent accessibility classes. *J. Mol. Biol.*, **238**, 682-692.
- Wang, Y., Lai, L., Han, Y., Xu, X. and Tang, Y. (1995) A new protein folding recognition potential function. *Proteins*, **21**, 127-129.
- Wilamowska, M. and Eisenberg, D. (1995) Inverse protein folding by the residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences. *Protein Eng.*, **8**, 627-639.
- Wodak, S.J. and Rooman, M.J. (1993) Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, **3**, 247-259.
- Yang, A.-S. and Honig, B. (1995a) Free energy determinants of secondary structure formation. 1. Alpha-helices. *J. Mol. Biol.*, **252**, 351-365.
- Yang, A.-S. and Honig, B. (1995b) Free energy determinants of secondary structure formation. 2. Antiparallel beta-sheets. *J. Mol. Biol.*, **252**, 366-376.
- Zimmermann, K. (1994) When awaiting 'bio' champion: dynamic programming regularization of the protein secondary structure predictions. *Protein Eng.*, **7**, 1197-1202.
- Zuckerklund, E. and Pauling, L. (1965) In Bryson, V. and Vogel, H.J. (eds), *Evolving Genes and Proteins*. Academic Press, New York, pp. 97-166.
- Zvelebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E. (1987) Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957-961.

Received on December 12, 1996; revised on February 10, 1997; accepted on February 10, 1997