Protein Structure Prediction in 1D, 2D, and 3D

Burkhard Rost

European Molecular Biology Laboratory, Heidelberg, Germany

Abbreviations

1D = one-dimensional; 1D structure = one-dimensional (e.g., sequence or string of secondary structure); 2D = two-dimensional; 2D structure = two-dimensional (e.g., inter-residue distances); 3D = three-dimensional; 3D structure = three-dimensional (coordinates of protein structure); PDB = Protein Data Bank of experimentally determined 3D structures of proteins; SWISS-PROT = database of protein sequences; T = target used for homology modeling (protein of known 3D structure); U = protein sequence of unknown 3D structure (e.g., search sequence).

1 INTRODUCTION

1.1 Proteins are the Machinery of Life

The information for life is stored by a four-letter alphabet in the genes (DNA). Proteins are, among others, the macromolecules that perform all important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, recognition, and transmission of signals. Thus, genes are the blueprints or library, and proteins are the machinery of life. Proteins are formed by joining amino acids by peptide bonds into a stretched chain. This protein sequence comprises a translation of the four-letter DNA alphabet into a 20-letter alphabet of native amino acids. Proteins differ in length (from 30 to over 30 000 amino acids), and in the arrangement of the amino acids (dubbed residues, when joined in proteins). In water, the chain folds up into a unique three-dimensional (3D) structure. The main driving force is the need to pack residues for which a contact with water is energetically unfavorable (hydrophobic residues) into the interior of the molecule. A detailed analysis of the underlying chemistry shows that this is only possible if the protein forms regular patterns of a macroscopic substructure called secondary structure (Figure 1; for an excellent introduction into protein structure, see Ref. 1; for a short review of the basic principles of folding, see Ref. 2).

1.2 Sequence Determines Structure Determines Function

Protein three-dimensional (3D) structure (i.e., the coordinates of all atoms) determines protein function. But what determines 3D structure? The hypothesis that structure (also referred to as 'the fold') is uniquely determined by the specificity of the sequence, has been verified for many proteins.[3] While it is now known that particular proteins (chaperones) often play a role in the folding pathway, and in correcting misfolds,[4] it is still generally assumed that the final structure is at the free-energy minimum. Thus, all information about the native structure of a protein is coded in the amino acid sequence, plus its native solution environment. Can the code be deciphered, i.e., can 3D structure be predicted from sequence? In principle, the code could by deciphered from physico-chemical principles using, for example, molecular dynamics methods.[5] In practice, however, such approaches are frustrated by two principal obstacles. First, energy differences between native and unfolded proteins are extremely small (order of 1 kcal mol^{-1}). Second, the high complexity (i.e., cooperativity) of protein folding requires several orders of magnitudes more computing time than we anticipate to have over the next decades. Thus, the inaccuracy in experimentally determining the basic parameters, and the limited computing resources become fatal for predicting protein structure from first principles.[6] The only successful structure prediction tools are knowledge-based, using a combination of statistical theory and empirical rules.

1.3 The Sequence–Structure Gap is Rapidly Increasing

Currently, databases for protein sequences (e.g., SWISS-PROT[7]) are expanding rapidly, largely because of large-scale genome sequencing projects. The first four entire genome sequences have been published; they represent all three terrestrial kingdoms: (1) prokaryotes: *Haemophilus influenzae*,[8] and *Mycoplasma genitalium*;[9] (2) eucaryotes: yeast;[10] and (3) archeans: *Methanococcus jannaschii*.[11] At least another dozen genomes will be completely sequenced before the end of 1997 (Terry Gaasterland, personal communication); the entire human genome is likely to be known in the year 2003. This implies that the explosion of genome, and hence, protein, sequences is supposedly the only field outgrowing the speed in development of computer hardware. It also implies, that despite significant improvements of structure determination techniques, the gap between the number of proteins for which structure is deposited in public databases (PDB[12]), and the number of proteins for which sequences are known is increasing.

1.4 Can the Egg be Unboiled?

When an egg is boiled, the proteins it contains unfold. Can this procedure be reversed in theory? Can the encrypted code of protein structure be deciphered? Or, can theory help to bridge the sequence–structure gap? Indeed, for over 30 years, there has been an ardent search for methods to predict protein structure from the sequence. Many methods were found which looked initially very promising – but always the hope has been dashed. How well do we do?

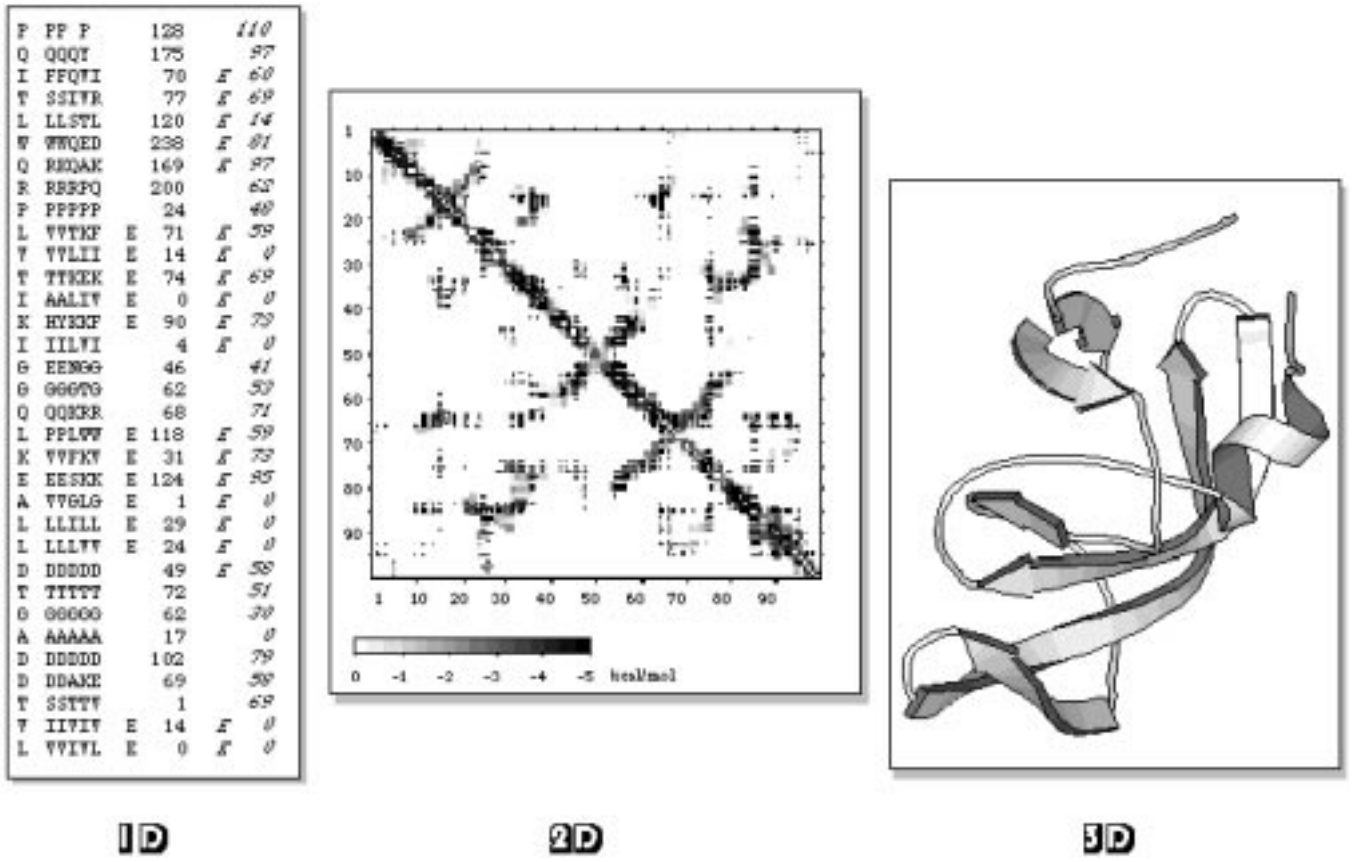

Figure 1 Representation of HIV-1 protease monomer (Protein Data Bank code 1HHP) in one, two, and three dimensions. Each of the representations gives rise to a different type of prediction problem. 1D prediction of secondary structure and solvent accessibility. From left to ion of right: amino acids for the first 33 residues (one letter code, first column); alignment exemplified by 5 sequences (second column); secondary structure[20] (H, helix; E, strand; blank, other: third column), solvent accessibility (measured in Å^2, fourth column,[20]), and a typical prediction by the neural network program PHD[21] for secondary structure and solvent accessibility (in italics, fifth and sixth column). 2D prediction of contact map. The 3D structure can be projected onto a two-dimensional matrix of inter-residue distances or contacts (as shown here). The entry at position *i j* of the matrix gives the contact strength between residue *i* and residue *j*. The stronger a contact, the darker the marker. Horizontal and vertical lines give borders of secondary structure segments. Graph made with CONAN.[22] 3D prediction of three-dimensional coordinates. The ion of trace of the protein chain in 3D is plotted schematically as a ribbon C^{α}-trace. Strands are indicated by arrows, the short helix is on the right towards the end (C-term) of the protein. Graph made with MOLSCRIPT.[23] Prediction not shown

1.5 No General Prediction of Structure from Sequence, Yet

An important experiment has been initiated by John Moult (CARB, Washington): those who determine protein structures submitted the sequences of proteins for which they were about to solve the structure to a 'to-be-predicted' database; for each entry in that database predictors could send in their predictions before a given deadline (the public release of the structure); finally, the results were compared, and discussed during a workshop (in Asilomar, California). Two such experiments have been completed: in December 1994 (*Proteins* special issue, Vol. 23, 1995), and in December 1996 (to be published in *Proteins*, 1998). The results of both experiments demonstrated clearly that the goal to predict structure from sequence has not been reached, yet. So, has there been no improvement despite ardent attempts, and the explosion of knowledge deposited in databases?

Indeed, there is a flood of literature on protein structure prediction attempting to keep track with the expanding databases (reviews;[13,14] books;[15,16] a practical approach to structure prediction and sequence analysis.[17–19] In this review focus will be laid on recent prediction methods that do actually contribute to bridging the sequence structure gap in particular in view of analyzing entire genomes. The first section will provide a brief sketch about where we are today in protein structure prediction. The following sections will sketch the problems, and some of the solutions in database searches, and the prediction of protein structure in 1D, 2D, and 3D (Figure 1).

2 STATE OF THE ART IN PROTEIN STRUCTURE PREDICTION

2.1 Bridging the Sequence–Structure Gap for 10–30% of all Sequences

The gap between the number of known sequences (>170 000[24]) and the number of known structures (about 5000[12]) is widening rapidly. The most successful theoretical approach to bridging this gap is homology modeling. The

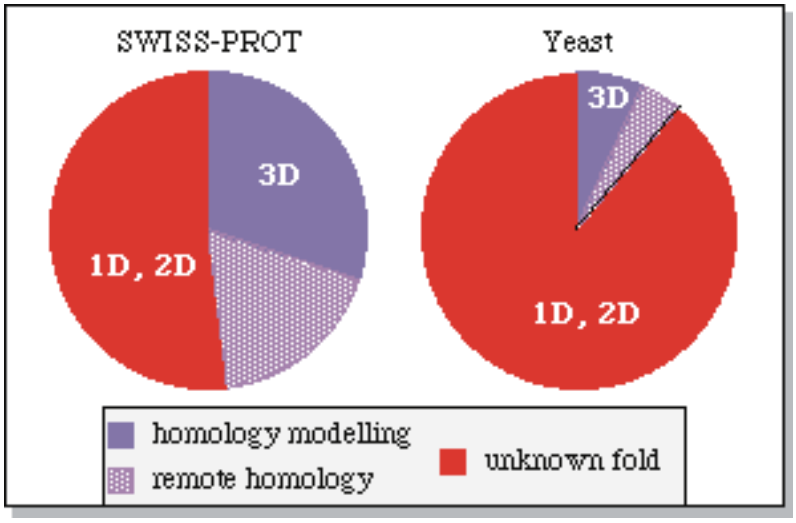

Figure 2 Scope of structure prediction. Given any expressed protein, how likely can theory predict its 3D structure? For example, for 30% of the proteins in the current SWISS-PROT database we can find regions for which homology modelling is applicable,[28] but for the first four entirely sequenced genomes (shown is yeast) this is true for less than 10% of all proteins.[29] Thus, SWISS-PROT contains a bias introduced, e.g., by limitations of previous sequencing techniques. Estimating the contribution of fold recognition or threading techniques is problematic. Margins given are certainly over-estimated in terms of the accuracy of current threading methods, and supposedly under-estimated in terms of the number of remote homologs that could be detected. (Note, however, today threading techniques are not accurate enough for any large-scale prediction of 3D structure!) The remaining region (50–80%) is occupied by unknown folds for which no accurate predictions in 3D can be obtained

principal idea bases on the following observation. Each native protein sequence adopts a unique structure. However, many different sequences can adopt the same basic fold. In other words, proteins with similar sequences tend to fold into similar structures. Indeed, for a pair of naturally evolved proteins, levels of 25–30% pairwise sequence identity (percentage of residues identical between the two proteins) are sufficient to assure that the two proteins fold into similar structures.[25–27] Thus, if a sequence of unknown structure (denoted U) has significant sequence similarity to a protein of known structure (T), it is possible to construct an approximate 3D model for U based on the assumption that U simply has basically the same structure as T. This technique is referred to as homology modeling. It effectively raises the number of 'known' 3D structures from 5000 to over 50 000[28] (Figure 2).

2.2 Widening the Bridge by Threading

Homology modeling allows prediction of 3D structure for 10–30% of all protein sequences. However, there is evidence that most pairs of proteins with similar structure are remote homologs with less than 25% pairwise sequence identity.[30] These remote homologs cannot usually be recognised by conventional sequence alignments, as this level of sequence identity is not significant for structural similarity in the following sense. If one were to collect all pairwise alignments of $<25\%$ sequence identity that result from a search with U against a database of protein sequences, then the vast majority of these pairs would be entirely unrelated proteins. Thus, most similar structures appear to be remote homologs, but most possible pairs at low levels of sequence identity are, in fact, unrelated. Consequently, searching for remote homologs is similar to the task of finding a needle in a haystack.[31] Techniques to manage this difficult task are referred to as 'threading techniques'. Most of these techniques are applicable if, and only if, the remote homolog to U has known structure. Once a remote homology is detected, remote homology modeling may be used to construct a 3D model. This could potentially reduce the sequence–structure gap by an additional 10 000–50 000 proteins (Figure 2). Given a sequence U from one of the complete genome sequences which have recently become available; what is the likelihood that the 3D structure can be predicted for U by homology modeling or remote homology modeling? A conservative answer is: 10%, based on the success of sequence alignment-based homology modeling (Figure 2). A very optimistic estimate is over 50%, assuming all remote homologs could be recognized (Figure 2).

2.3 Accurate Prediction for 1D Aspects of 3D Structure

If no remote homolog 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,[32] and the location of helices for the special class of coiled-coil proteins.[33]

3 SEQUENCE ALIGNMENTS

3.1 Basic Concept

The principal problem of sequence alignments is to find the optimal superposition between two strings of amino (or nucleic) acid sequences, i.e., to optimally align the two strings. The most simple objective is to optimize the percentage

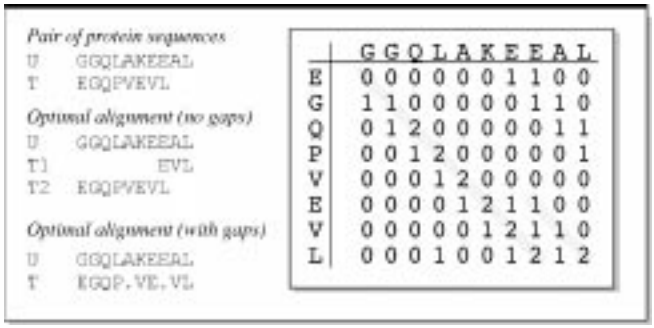

Figure 3 The simple dynamic programming algorithm simply proceeds in the following way. The two sequences to be aligned (U and T) are written into a matrix. Starting from the first element in the matrix, identities are counted and summed along the diagonal. The two best paths are marked by gray lines. The two best alignments match only two identical residues for the example given. However, if insertions (marked by dots) were allowed, the best alignment, actually, matches four residues

of residues that are identical between the two sequences. A dynamic programming *algorithm* is guaranteed to find the optimal solution for the problem in an algorithmic time quadratic in the length of both sequences[34] (Figure 3). For the alignment of protein sequences this simple approach is not sufficient as finding the best alignments, usually, requires to introduce gaps in one sequence, or insertions in the other[35] (Figure 3: rather than placing two dots into sequence T, residues A and E could be deleted in sequence U; note: the gap increases the score from 2 to 4 identical residues). The introduction of such gaps is mathematically treated by adding a constant (gap open penalty) to the final score (here number of identical residues). However, to align protein sequences sensitively, this still is not sufficient. The major addition to the simple approach described so far is to evaluate scores not based on residue identities but based on biochemical properties of amino acid. For example, aligning two hydrophobic residues (I and L) is more beneficial than aligning a hydrophobic and a charged residue (L and K; note: when treating hydrophobic residues as identical, the score for the best gapped alignment in Figure 3 increases from 4 to 6).

3.2 Evolution Distinguishes Signal from Noise

At the level of protein molecules, selective pressure results from the need to maintain function, which in turn requires maintenance of the specific 3D structure. This evolutionary history is the basis for the success in aligning protein (or nucleotide) sequences. Accordingly, conservation and mutation patterns observed in alignments contain very specific information about 3D structure. How much variation is tolerated without loss of structure? Two naturally evolved proteins with more than 25% identical residues (length >80 residues) are very likely to be similar in 3D structure.[27]

3.3 Task Trivial for High Levels of Sequence Identity

Any sequence analysis starts with database searches: all known databases are scanned by sequence alignment procedures for proteins homologous to the search sequence U. When the pairwise sequence identity between U and a putative homolog H is over 25–30% (for more than 80 residues), alignment procedures are usually straightforward.[36–40] For less similar protein pairs, alignments may fail.

3.4 Routine Database Searches by Simplified Procedures

Aligning two sequences by dynamic programming is a matter of seconds on a modern workstation. However, database searches require to repeat this many times, and since the databases grow, CPU time becomes a constraint in everyday sequence analysis. This bottleneck is opened by methods that start to find 'identical words' (sub-strings), and then grow the alignment around such blocks. The most widely used programs of this sort are BLAST and FASTA.[37,39] In practice, advanced alignment algorithms typically proceed by first running a fast scan with BLAST and/or FASTA, and then by applying the full dynamic programming algorithm. To illustrate sequence analysis in practice: aligning the 6000 sequences of yeast against all known proteins was recently accomplished in 72 h on 64 SGI 10 000 processors.[41]

3.5 Multiple Alignments Improve as Data Banks Grow

The most advanced sequence alignment tools base the alignment on profiles derived from databases or particular sequence families.[14,42] One new generation of alignment methods is based on Hidden Markov Models, another on genetic algorithms. These new methods may be more successful in intruding into the twilight zone of sequence alignments (20–30% sequence identity[26]) than advanced profile-based methods. However, this remains to be proven.

3.6 Drawback: Lack of Sufficiently Tested Cut-off Criteria

There are many different alignment methods available for those who need to run database searches for their everyday work. Which method is best? One of the difficulties in comparing different alignment procedures is the lack of well-defined criteria for measuring the alignment quality. Very few papers have attempted to define such measures for the comparison of various methods.[43] The second problem for users is that most methods do not supply a cut-off criterion for distinguishing between homologous and nonhomologous sequences (i.e., false positives). For some large sequence families, remote homologs can be aligned correctly, but for most cases sequences aligned to the search protein U at levels below 25% pairwise sequence identity will be false positives, i.e., will have no structural or functional similarity to U. A simple length-dependent cut-off based on sequence identity is provided by the program MAXHOM.[27] However, this threshold neither quantifies the influence of biochemical similarities between amino acids, nor the occurrence of gaps.

4 PREDICTION IN 1D

4.1 Secondary Structure

4.1.1 Basic Concept

The principal idea underlying most secondary structure prediction methods is the fact that segments of consecutive residues

have preferences for certain secondary structure states.[1,21] Thus, the prediction problem becomes a pattern-classification problem tractable by pattern recognition algorithms. The goal is to predict whether the residue at the centre of a segment of typically 13–21 adjacent residues is in a helix, strand or in neither of the two (no regular secondary structure, often referred to as the 'coil' or 'loop' state). Many different algorithms have been applied to tackle this simplest version of the protein structure prediction problem: physico-chemical principles, rule-based devices, expert systems, graph theory, linear and multi-linear statistics, nearest-neighbor algorithms, molecular dynamics, and neural networks.[21] However, until recently, performance accuracy seemed to have been limited to about 60% (percentage of residues correctly predicted in either helix, strand, or other). The limited accuracy was argued to result from the fact that all methods used only information local in sequence (window of less than 20 adjacent residues). Local information was estimated to account for roughly 65% of the secondary structure formation. Two additional problems were common to all methods developed from 1957 to 1993: (1) strands were predicted at levels of accuracy only slightly superior to random predictions, and (2) predicted secondary structure segments were, on average, only half as long as observed segments. The later two shortcomings could be surmounted by using a particular combination of neural networks.[21]

4.1.2 Evolutionary Information Key to Significantly Improved Predictions

On the one hand, about 75 out of 100 residues can be exchanged in a protein without changing structure. On the other hand, exchanges of 1–5 residues can already destabilize a protein structure. These statements may appear contradictory. However, the explanation is simple: evolution has explored exactly the unlikely exchanges of particular amino acids at particular positions that do not change structure, as a change of structure usually results in a loss of function (and thus would not survive). Thus, the residue exchange patterns extracted from a protein family (i.e., alignments of similar sequences) are highly indicative of the specific structural details for that family. The first method that reached a sustained level of a three-state prediction accuracy above 70% was the profile-based neural network system PHD which uses exactly such evolutionary information derived from multiple sequence alignments as input.[21] By stepwise incorporation of particular evolutionary information, prediction accuracy (Figure 4) has been pushed above 72% accuracy.[21] An interesting, technical detail of this network system is that the use of a global 'descriptor', namely the overall amino acid composition (percentage of occurrence of each of the 20 amino acids) does not affect the local score for accuracy as measured by the percentage of correctly predicted single residues. Using amino acid composition, however, improves the accuracy in terms of a more global score, such as the difference between the percentage of observed and predicted secondary structure.[21] Is the neural network an essential tool for the most accurate secondary structure prediction? A nearest-neighbor algorithm can be used to incorporate evolutionary information in a similar manner as the neural network system; the result is a similar performance.[44] Methods combining statistics, and multiple alignment information have been clearly less successful, so far. In comparison with methods using single sequence information only, methods making use of the growing databases are 6–14 percentage points more accurate. Thus, using evolutionary information secondary structure can now be predicted more accurately and reliably than other features of protein structure.

4.1.3 Secondary Structure Predictions now Extremely Useful, in Practice

How good is a prediction accuracy of 72% in practice? It is certainly reasonably good compared with the prediction of secondary structure by homology modeling.[45] However, prediction accuracy varies between different proteins, i.e., prediction accuracy is $72\% \pm 9\%$ (one standard deviation).[21] For applications this implies that predictions can be as good as $>95\%$, but also as bad as $<54\%$. Can users distinguish one from the other? A few methods successfully use reliability indices allowing one to label residues for which predictions are, on average, likely to be more accurate. Indeed, for the neural network system PHD the correlation between such a reliability index and accuracy is linear.[21] Thus, the reliability index effectively becomes a means to predict prediction accuracy, and hence to assess to which class a protein of unknown structure (U) belongs: to the well predicted, or to the badly predicted ones. Various methods successfully use secondary structure predictions as a first step, e.g., prediction-based threading (indeed one of the problems of the Asilomar 1996 prediction contest was that many developers of threading algorithms used the same PHD secondary structure predictions as a first step), inter-strand, and inter-residue distance predictions. However, the use of secondary structure predictions is not limited to structure prediction. Instead, the results of, for instance, the public prediction service (PredictProtein[21]) have been used to assist the determination of protein structures (chain tracing in X-ray crystallography), as well as to formulate hypotheses about protein structure and function that guided experiments in molecular biology, in general (in particular, prediction of binding sites, homologous proteins, design of residue mutations).

4.1.4 Separate Prediction of Secondary Structure Content not very Useful

Proteins have been partitioned into various structural classes, e.g., based on the percentage of residues assigned to helix, strand, and other.[46] However, such a coarse-grained classification is not well defined.[47] Consequently, given a protein sequence U of unknown structure, attempts to first predict the secondary structure content for U and then to use the result to predict the secondary structural class (i.e., all-α, all-β or intermediates) is of limited practical use. How do alignment-based predictions compare with experimental means of determining the content in secondary structure? For example, PHD is, on average, surprisingly about as accurate as circular dichroism spectroscopy.[47] Of course, this does not imply that predictions can replace experiments. In particular, variation of secondary structure as a result of changes in environmental conditions (e.g., solvent) is generally only accessible experimentally.

4.2 Solvent Accessibility

4.2.1 Basic Concept

It has long been argued that, if the segments of secondary structure could be accurately predicted, the 3D structure

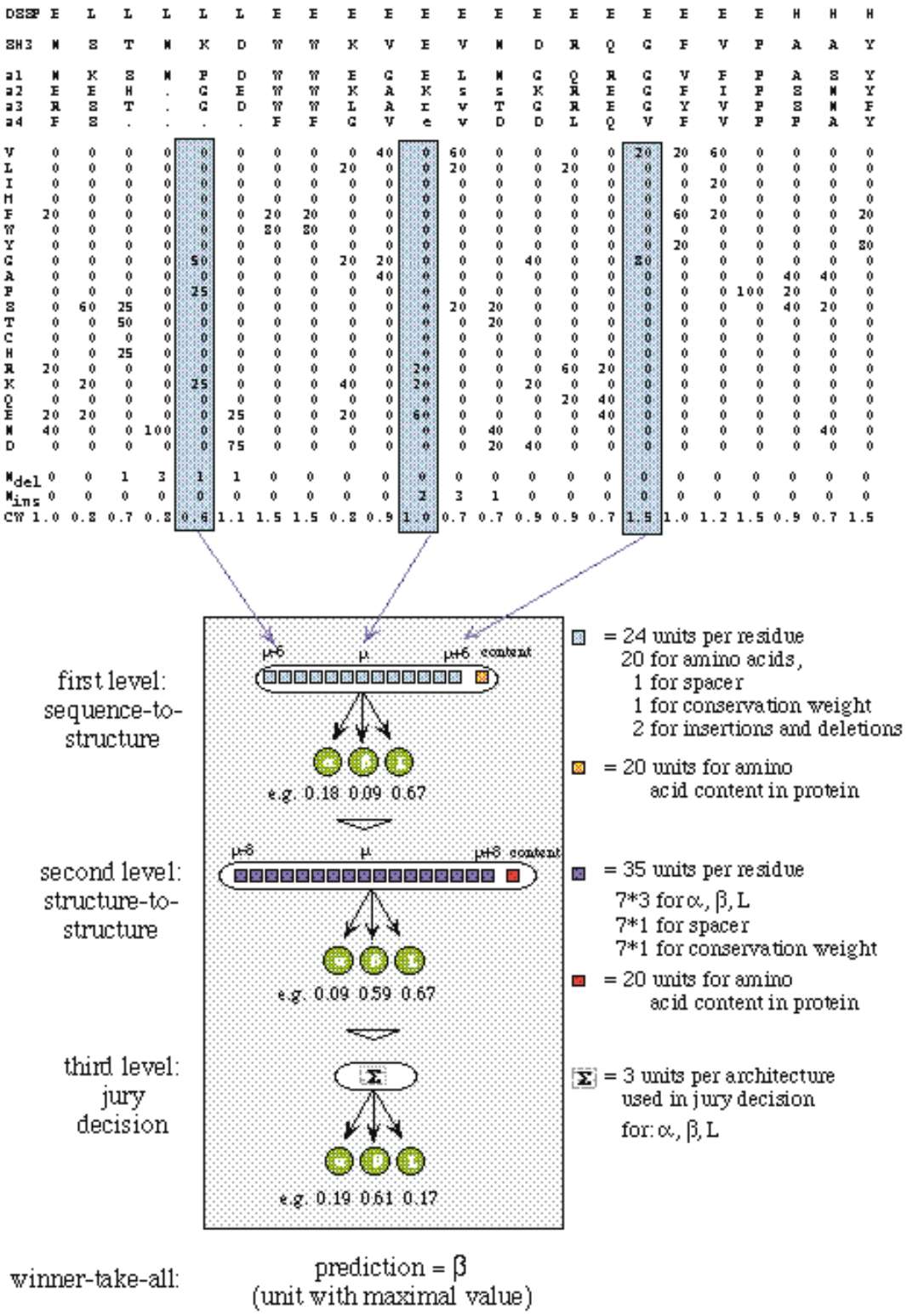

Figure 4 PHDsec: profile-based neural network system for secondary structure prediction.[21] From the multiple alignment (here guide sequence SH3 plus four other proteins a1–a4, note: lower case letters indicate deletions in the aligned sequence) a profile of amino acid occurrences is compiled. To the resulting 20 values at one particular position μ in the protein (one column) three values are added: the number of deletions and insertions, and the conservation weight (CW). 13 adjacent columns are used as input. The whole network system for secondary structure prediction consists of three layers: two network layers and one layer averaging over independently trained networks

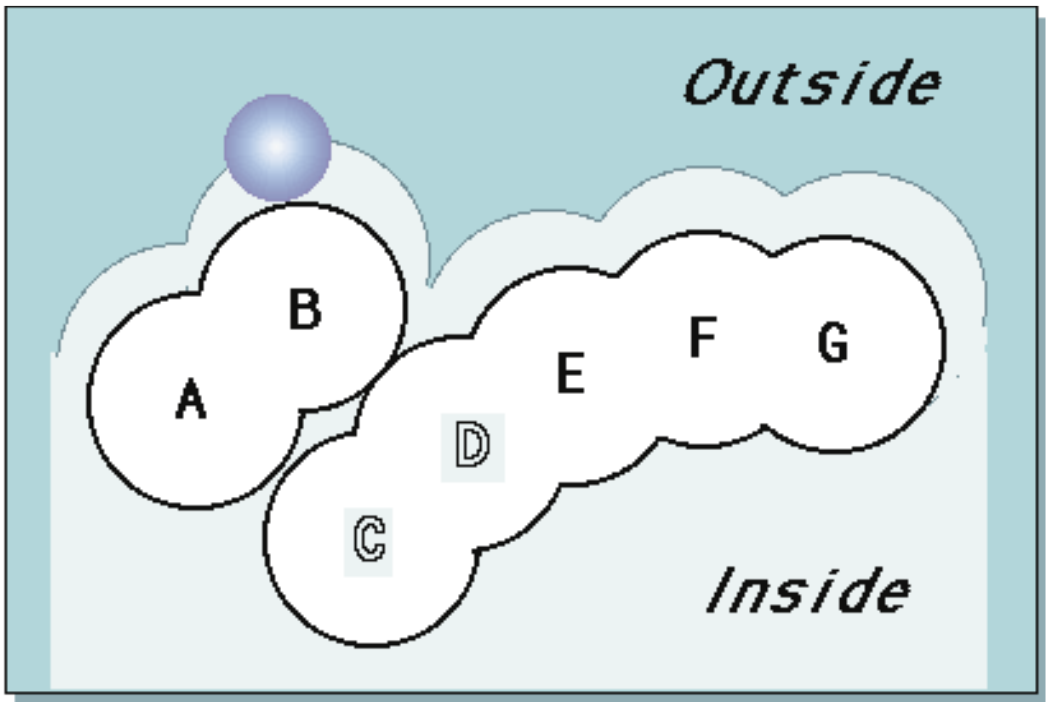

Figure 5 Residue solvent accessibility is usually measured by rolling a spherical water molecule over a protein surface and summing the area that can be accessed by this molecule on each residue (typical values range from 0 to 300 Å^2). To allow comparisons between the accessibility of long extended and spherical amino acids, typically relative values are compiled (actual area as percentage of maximally accessible area). A simplified descriptions distinguishes two states: buried (here residues C and D) and exposed (here residues A, B, E, F, and G) residues. Since the packing density of native proteins resembles that of crystals, values for solvent accessibility provide upper and lower limits to the number of possible inter-residue contacts

could be predicted by simply trying different arrangements of the segments in space.[48] One criterion for assessing each arrangement could be to use predictions of residue solvent accessibility.[49,50] The principal goal is to predict the extent to which a residue embedded in a protein structure is accessible to solvent (Figure 5). Solvent accessibility can be described in several ways.[49,50] The simplest is a two-state description distinguishing between residues that are buried (relative solvent accessibility $<16\%$) and exposed (relative solvent accessibility $\geq 16\%$). The classical method to predict accessibility is to assign either of the two states, buried or exposed, according to residue hydrophobicity. However, a neural network prediction of accessibility has been shown to be superior to simple hydrophobicity analyses.[51]

4.2.2 Evolutionary Information Improves Prediction Accuracy

Solvent accessibility at each position of the protein structure is evolutionarily conserved within sequence families. This fact has been used to develop methods for predicting accessibility using multiple alignment information.[52] Prediction accuracy is about $75 \pm 7\%$, four percentage points higher than for methods not using alignment information. Predictions of solvent accessibility have also been used successfully for prediction-based threading, as a second criterion towards 3D prediction by packing secondary structure segments according to upper and lower bounds provided by accessibility predictions, and as basis for predicting functional sites.[52] More recently, predictions of accessibility were also used successfully to predict sub-cellular location (Andrade and O'Donoghue, personal communication).

4.3 Transmembrane Helices

4.3.1 Basic Concept

Even in the optimistic scenario that in the near future most protein structures will be experimentally determined, one class of proteins will still represent a challenge for experimental determination of 3D structure: transmembrane proteins. The major obstacle with these proteins is that they do not crystallize, and are hardly tractable by NMR spectroscopy. Consequently, for this class of proteins, structure prediction methods are needed even more than for globular water-soluble proteins. Fortunately, the prediction task is simplified by strong environmental constraints on transmembrane proteins: the lipid bilayer of the membrane reduces the degrees of freedom to such an extent that 3D structure formation becomes almost a 2D problem. Two major classes of membrane proteins are known: proteins which insert helices into the lipid bilayer (Figure 6), and proteins that form pores by a barrel of 16 strands (the only known cases of this type are porins[53]). Since there is not much experimental information available on different porin-like membrane proteins, we can hardly estimate prediction accuracy for this class. The situation is quite different for helical membrane proteins. Once the location of transmembrane segments is known for helical transmembrane proteins, 3D structure can be predicted by exploring all possible conformations.[54] Additionally, predicting the locations of these transmembrane helices is a much simpler problem than

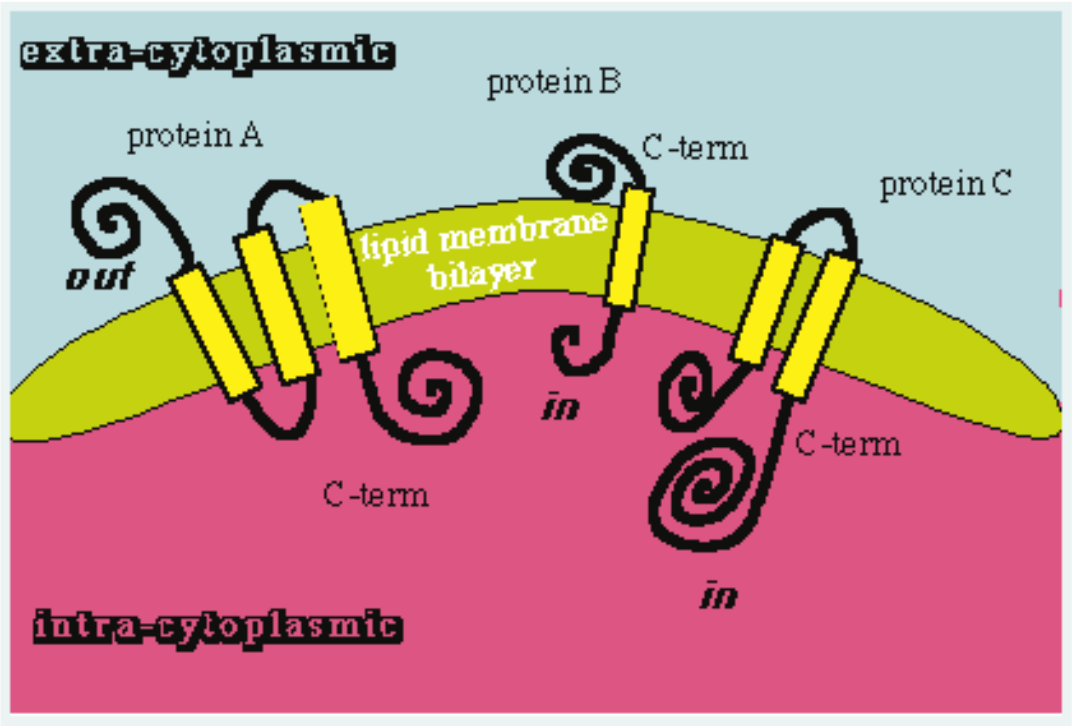

Figure 6 Topology of helical transmembrane proteins. In one class of membrane proteins, typically apolar helical segments are embedded in the lipid bilayer oriented perpendicular to the surface of the membrane. The helices can be regarded as more or less rigid cylinders. The orientation of the helical axes, i.e., the topology of the transmembrane protein, can be defined by the orientation of the first N-terminal residues with respect to the cell. Topology is defined as *out* when the protein N-term (first residue) starts on the extra-cytoplasmic region (protein A), and as *in* if the N-term starts on the intra-cytoplasmic side (proteins B and C)

is the prediction of secondary structure for soluble proteins. Elaborated combinations of expert-rules, hydrophobicity analyses and statistics yields a two-state per-residue accuracy of about 90% (residues predicted correctly as either transmembrane helix, or other).

4.3.2 Evolutionary Information Improves Prediction Accuracy

For two methods the use of multiple alignment information is reported to clearly improve the accuracy of predicting transmembrane helices.[21,55] The best current prediction methods have a similar high accuracy around 95%. One such method uses a system of neural networks similar to the one sketched in Figure 4. In order to predict the orientation of the helices (i.e., the topology Figure 6) a simple rule is applied: positively charged residues occur more often in intra-cytoplasmic than in extra-cytoplasmic regions. The advanced neural network system has been improved significantly by adding a dynamic programming algorithm to the neural network output. The principal idea is to use the neural network output as an energy landscape and to find the optimal path through this landscape.[32] As reliable data for the locations of transmembrane helices exist only for a few proteins, data used for deriving these methods originate predominantly from experiments in cell biology and gene-fusion techniques. Different experimental groups often report different locations for transmembrane regions. Thus, the level of 95% accuracy is not verifiable. Despite this uncertainty in detail, the prediction of transmembrane helices is a valuable tool for quickly scanning entire chromosomes.[32] The classification into membrane/nonmembrane proteins has an expected error rate of less than 2%, i.e., about 2% of the proteins predicted to contain transmembrane regions will probably be false positives. The predictions of transmembrane helices has provided a lower bound to approach the question of how many proteins organisms need for example, communication: the percentage of proteins with transmembrane helices has been estimated to be about 25% for yeast and *Haemophilus influenzae*, and around 10–15% for *Mycoplasma genitalium* and *Methanococcus jannaschii* (Rost, manuscript in preparation; data available at http://www.embl-heidelberg.de/~rost).

5 PREDICTION IN 2D

5.1 Inter-residue Contacts

5.1.1 Prediction Problem is Difficult, but the Stakes are High

Given all inter-residue contacts or distances (Figure 1), 3D structure can be reconstructed by distance geometry or molecular dynamics. This is used for the determination of 3D structures by NMR spectroscopy which produces experimental data of distances between protons.[56] Can inter-residue contacts be predicted? Obviously, some fraction of these contacts can be: helices and strands can be assigned based on hydrogen-bonding pattern between residues. Thus, a successful prediction of secondary structure implies a successful prediction of some fraction of all the contacts. However, contacts predicted from secondary structure assignment are

short-ranged, i.e., between residues nearby in sequence. For a successful application of distance geometry, long-range contacts have to be predicted, i.e., contacts between residues far apart in sequence. A few methods have been proposed for the prediction of long-range inter-residue contacts. Two questions surround such methods: first, how accurate are these prediction methods on average; and second, are all important contacts predicted?

5.1.2 Correlated Mutations can Imply Spatial Proximity

In sequence alignments, some pairs of positions appear to co-vary in a physico-chemically plausible manner, i.e., a 'loss of function' point mutation is often rescued by an additional mutation that compensates for the change.[57] One hypothesis is that compensations would be most effective in maintaining a structural motif if the mutated residues were spatial neighbors. Attempts have been made to quantify such a hypothesis and to use it for contact predictions.[58–60] In general, prediction accuracy is rather poor, with a direct trade-off between predicting enough contacts, and predicting only correct ones, e.g., taking 5% of the best-predicted long-range contacts (sequence separation above 10 residues) the accuracy prediction is about 50% (A. Valencia, personal communication).

5.1.3 Distinction Between Different Models, no Prediction of 3D, Yet

Analyzing correlated mutations is only one way to predict long-range inter-residue contacts. Other methods use statistics, mean-force potentials, or neural networks. So far none of the methods appears to find a path between the Scylla of missing too many true contacts and the Charibdis of predicting too many false contacts. However, some of the methods provide sufficient information to distinguish between alternative models of 3D structure (Valencia, personal communication). The ambitious goal of predicting long-range inter-residue contacts sufficiently accurately will hopefully continue to attract intellectual resources.

5.2 Inter-strand Contacts

5.2.1 Simplifying the Contact Prediction Problem

One simplification of the problem of predicting inter-residue contacts focuses on predicting the contacts between residues in adjacent strands (Figure 1). Such an attempt is motivated by the hope that such interactions are more specific than are sequence-distant (long-range) contacts in general, and hence are easier to predict.

5.2.2 Identifying the Correct β-Strand Alignment

The only method published for predicting inter-strand contacts is based on potentials of mean-force[61] similar to those used in the evaluation of strand–strand threading.[62] Propensities are compiled by database counts for $2 \times 2 \times 2$ classes (parallel/anti-parallel, H-bonded/not H-bonded, N-/C-terminal). Each of the eight classes is divided further into five sub-classes in the following way. Suppose the two strand residues at positions i and j are close in space. Then the following five residue pairs are counted in separate tables: $i/j-2$, $i/j-1$, i/j, $i/j+1$, $i/j+2$. Such pseudo-potentials identify the correct β-strand alignment in 35–45% of the cases.

5.2.3 Using Evolutionary Information to Predict Inter-strand Contacts

Even if the locations of strands in the sequence are known exactly, the pseudo-potentials cannot predict the correct inter-strand contacts in most cases.[61] However, when using multiple alignment information, the signal-to-noise ratio increases such that inter-strand contacts have been predicted correctly for most of the strands inspected in some test cases.[61] For the purpose of reliable contact prediction, this result is inadequate, especially as the locations of the strands are not known precisely. Can the pseudo-potentials handle errors resulting from incorrect prediction of strands? Various test examples using predictions by PHDsec[21] as input to the strand pseudo-potentials indicate that the accuracy in predicting inter-strand contacts drops (T. Hubbard, unpublished), but in some cases is still high enough to be useful for approximate modeling of 3D structure.[63]

6 PREDICTION IN 3D

6.1 Known Folds: Homology Modeling

6.1.1 Basic Concept

An analysis of PDB reveals that all protein pairs with more than 30% pairwise sequence identity (for alignment length >80[27]) have homologous 3D structures, i.e., the essential fold of the two proteins is identical, details such as additional loop regions – regions not in helices or strands – may vary. Structure is more conserved than is sequence. This is the pillar for the success of homology modeling. The principal idea is to model the structure of U (protein of unknown structure) based on the template of a sequence homolog of known structure (T). Consequently, the precondition for homology modeling is that a sequence homolog of known structure is found in PDB. Since homology modeling is currently the only theoretical means successfully to predict 3D structure, this has two implications. First, homology modeling is applicable to 'only' one quarter of the known protein sequences (Figure 2). Second, as the template of a homolog is required, no unique 3D structure can yet be predicted, i.e., no structure that has no similarity to any experimentally determined 3D structure. Suppose, there is a protein with a sequence similar to U in PDB (say T), is homology modeling straightforward?

6.1.2 High Level of Sequence Identity: Atomic Resolution

The basic assumption of homology modeling is that U and T have identical backbones (main chain C). The task is correctly to place the side chains of U into the backbone of T. For very high levels of sequence identity between U and T (ideally differing by one residue only), side chains can be 'grown' during molecular dynamics simulations.[64] For slightly lower levels (still of high sequence similarity), side chains are built based on similar environments in known structures.[65,66] Rotamer libraries (libraries containing all side-chain orientations observed in known structures) are used in the following way. (1) Rotamer distributions are extracted

from a database of nonredundant sequences. (2) Fragments of seven (helix, strand) or five residues (other) are compiled. (3) Fragments of the same length are successively shifted through the backbone of U. (4) For modeling the side chains of U only those fragments from the rotamer library are accepted which have the same amino acid in the center as U, and for which the local backbone is similar to that around the evaluated position). Over the whole range of sequence identity between U and T for which homology modeling is applicable, the accuracy of the model drops with decreasing similarity. For levels of at least 60% sequence identity, the resulting models are quite accurate;[66,67] even for higher values, the models are as accurate as is experimental structure determination. The limiting factor is the computation time required. How accurate is homology modeling for lower levels of sequence identity?

6.1.3 Low Level of Sequence Identity: Loop Regions Sometimes Correct

With decreasing sequence identity between the known structure H and the query protein U, the number of loops that have to be inserted to align the two grows. An accurate modeling of loop regions, however, implies solving the structure prediction problem. The problem is simplified in two ways. First, loop regions are often relatively short and can thus be simulated by molecular dynamics (note the CPU time required for molecular dynamics simulations grows exponentially with the number of residues of the polypeptide to be modeled). Second, the ends of the loop regions are fixed by the backbone of the template structure. Various methods are employed to model loop regions. The best have the orientation of the loop regions correct in some cases.[67] This illustrates the current limitations of molecular dynamics: not even short loop regions can be predicted from sequence. Furthermore, for experimental structure refinement (use of molecular dynamics to improve consistency, and accuracy of experimental data) molecular dynamics is successfully applied to find a better solution when starting from an almost correct structure. However, for homology modeling, molecular dynamics refinement usually reduces prediction accuracy.[67] Below about 40% sequence identity the accuracy of the sequence alignment used as basis for homology modeling becomes an additional problem. Nevertheless, even down to levels of 25–30% sequence identity, homology modeling produces coarse-grained models for the overall fold of proteins of unknown structure.

6.2 Known Folds: Remote Homology Modeling (Threading)

6.2.1 Basic Concept

As noted in the previous section, naturally evolved sequences with more than 30% pairwise sequence identity have homologous 3D structures.[27] Are all others non-homologous? Not at all. In the current PDB database there are thousands of pairs of structurally homologous pairs of proteins with less than 25% pairwise sequence identity (remote homologs). Actually, most similar protein structures are such remote homologs.[30] If a correct alignment between U (sequence of unknown structure) and a remote homolog T (pairwise sequence identity to U <25%) is given, one could build the 3D structure of U by (remote) homology modelling based on the template of T. A successful remote homology modeling must solve three different tasks. (1) The remote homolog (T) has to be detected. (2) U and T have to be correctly aligned. (3) The homology modeling procedure has to be tailored to the harder problem of extremely low sequence identity (with many loop regions to be modeled). Most methods developed so far have been primarily addressed to solve the first, and the second problem. The basic idea is to thread the sequence of U into the known structure of T and to evaluate the fitness of sequence for structure by some kind of environment-based or knowledge-based potential.[68,69] Threading is in some respects a harder problem than is the prediction of 3D structure (NP-complete;[70] no physical connection between remote homologs, as many remotely homologous protein pairs may have originated from different ancestors[30]). However, the stakes are high: solving the threading problem could enable the prediction of thousands of protein structures. Indeed, threading has evolved to become one of the most active fields in the arena of protein structure prediction (with well over 100 annual publications).

6.2.2 Variety of Threading Techniques

The optimism generated by one of the first papers on threading published in the 1990s[72] has boosted attempts to develop threading methods. The principal idea has been to use structural propensities of amino acids (such as preferences for secondary structure formation, hydrophobicity, and polarity), and to then assess whether or not a given sequence with its structural preferences fits into the structural environment of a given structure.[69] A principally different approach has been pushed by Manfred Sippl.[71,73] The idea is to use the rich knowledge deposited in the database of protein structures (PDB) by extracting mean-force potentials. Such potentials monitor the observed distances between residue pairs of particular amino acids, with a particular sequence separation (number of residues between the two). Until 1995, most threading methods used mean-force potentials,[42,68,71] A more recent generation of threading methods is based on 1D predictions:[52] first a 1D structure (secondary structure and solvent accessibility) is predicted for a sequence of unknown structure, then the 1D structure is extracted for a library of known structures, and finally the observed and the predicted 1D structure strings are aligned by typical dynamic programing *algorithms*.[35] Has all this effort enabled the hard nut of threading to be cracked?

6.2.3 Remote Homologs 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 homolog is likely to be found by at least one of them.[74] Now the bad news: so far, no single method has been able to detect the correct remote homolog for more than half of all test cases.[74] For the methods which have been rigorously evaluated using large test sets, the correct remote homolog is detected in less than 40% of all cases.[52] However, this performance is clearly superior to that of traditional sequence alignments at this low level (<25%) of sequence identity. Furthermore, the success of the last Asilomar experiment on structure prediction (*Proteins*, 1998, in press) suggests that the likelihood of detecting the correct remote homolog is reasonably high when the choice is refined by experts.

6.2.4 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 modeling procedures available today cannot correct the mistakes in the alignments. Although the last Asilomar experiment on structure prediction (*Proteins*, 1998, in press) suggested that major improvements have been accomplished over the last two years, there are still very few publications to date which report accurate 3D predictions from threading methods. Currently, the successful use of threading methods requires sceptical, expert user intervention to spot wrong hits and false alignments. It is still possible that threading method will become the most successful structure prediction method, but a lot of detailed work lies ahead.

6.3 Unknown Folds: *Ab Initio* Prediction of Structure?

6.3.1 Recent Breakthrough in Structure Prediction?

In the 1994 Asilomar meeting, none of the 3D *ab initio* methods was able to predict the correct protein structure.[67] 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.[52] Simplified force fields in combination with dynamic *optimization* strategies have yielded promising, but still relatively inaccurate results.[75,76] Srinivasan and Rose have reported very encouraging results with their hierarchical search method.[77] However, the second Asilomar experiment on structure prediction (*Proteins*, 1998, in press) concluded similarly to the first: no prediction of 3D structure from sequence, yet.

6.3.2 Accurate Prediction of 3D Structure for Coiled-coil Proteins

A particular class of proteins are coiled-coils. These are proteins can be defined by a rather simple geometry of long helices, of which two or more wind around one another.[33] Nilges and Brünger[78] 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.[84]

6.3.3 Recognizing Incorrect Structures

The single most important theoretical advance in 3D prediction in recent years may have 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 nonbonded terms which describe the internal energy of the molecule.[6] In contrast, the mean-force potentials, derived from databases of protein structure,[79] attempt to describe the free energy of the molecule. The physical potentials have been used very successfully to refine experimentally determined structures.[56] However, these terms cannot distinguish between a native fold and a grossly misfolded structure.[79] In contrast, mean-force potentials of pairwise residue distances are quite successful in fold recognition, as well as remote homology modeling.[71] 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, best results where obtained by combining both potentials.[84]

6.3.4 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.[80] Simulations and other studies indicate that the free energy balance of hydrogen bond formation is close to zero, or slightly unfavorable,[81,82] and that a specific fold is selected primarily by side-chain interactions.[80] 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.[83] The formalism starts with the following two assumptions: (1) that protein structures can be described by Helmholtz free energies (or mean-force potentials), and (2) 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. (1) Backbone H-bonds (except for the α-helix interaction $O_i \ldots N_{i+4}$) do not contribute to the thermodynamic stability of native folds. (2) H-bond formation (except for $O_i \ldots N_{i+4}$) requires energy input that is regained when H-bonds are formed. Once formed, H-bonds are locked in a deep, narrow minimum. (3) 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 start of a new era of force fields? Further applications of these mean-force potentials will be needed to answer this question.

7 CONCLUSIONS

Native 3D structures of proteins are encoded by a linear sequence of amino acid residues. To predict 3D structure from 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.

7.1 Prediction in 3D: Theory Bridges the Sequence–Structure Gap

The only source for new, unique protein structures (structures for which no homolog exists in the database) is experiments. However, given the amount of time needed to determine a protein structure experimentally, more nonunique

structures can be predicted at atomic resolution by homology modeling in a month than have been determined by experiment over the last three decades. Homology derived models are frequently accurate at the level of atomic resolution. Unfortunately, most models typically have considerable coordinate errors in loop regions. Coarse-grained homology derived models are available for almost one-third of the sequences deposited in the SWISS-PROT database.[28] Threading techniques could increase this ratio considerably by finding more distant homologs. However, for large-scale sequence analyses, threading techniques are not yet reliable.

7.2 Predictions in 1D: Significant Improvement from Larger Databases

The rich information contained in the growing sequence and structure databases has been used to improve the accuracy of predictions of some aspects of protein structure. 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 five years ago. Some methods have indicated that 1D predictions can be useful as an intermediate step on the way to predicting 3D structure (inter-strand contacts; prediction-based threading). Another advantage of predictions in 1D is that they are not very CPU-intensive, i.e., 1D structure can be predicted for the protein sequence of, for example, entire yeast chromosomes overnight.

7.3 Predictions in 2D: so far of Limited Success

The prediction accuracy of chain-distant inter-residue contacts is so far relatively limited. Analysis of correlated mutations can be used to distinguish between alternative models (e.g., for threading techniques). The prediction of inter-strand contacts appears to be useful in some cases. An accurate method for the automatic prediction of contacts between residues not close in sequence remains to be developed.

Most breakthroughs in protein structure prediction were achieved since 1990. 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 2003 when the human genome will be known.

8 RELATED ARTICLES

Circular Dichroism: Electronic; Drug Design; Hydrophobic Effect; Molecular Docking and Structure-based Design; Molecular Dynamics: Techniques and Applications to Proteins; Molecular Surfaces and Solubility; Molecular Surface and Volume; Neural Networks in Chemistry; Protein Data Bank (PDB): A Database of 3D Structural Information of Biological Macromolecules; Protein Design Concepts; Protein Folding and Optimization Algorithms; Protein Modeling; Protein Structure and Stability: Database-derived Potentials and Prediction; Superfamily Analysis: Understanding Protein Function from Structure and Sequence.

9 REFERENCES

1. C. Brändén and J. Tooze, 'Introduction to Protein Structure', Garland, New York, 1991.
2. E. E. Lattman and G. D. Rose, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 439–441.
3. C. B. Anfinsen, *Science*, 1973, **181**, 223–230.
4. F. J. Corrales and A. R. Fersht, *Folding & Design*, 1996, **1**, 265–273.
5. M. Levitt and A. Warshel, *Nature*, 1975, **253**, 694–698.
6. W. F. van Gunsteren, *Curr. Opin. Struct. Biol.*, 1993, **3**, 167–174.
7. A. Bairoch and R. Apweiler, *Nucl. Acids Res.*, 1996, **24**, 21–25.
8. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter, *Science*, 1995, **269**, 496–512.
9. C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, and J. C. Venter, *Science*, 1995, **270**, 397–403.
10. A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, Y. Philippsen, H. Tettelin, and S. G. Oliver, *Science*, 1996, **274**, 546–567.
11. C. J. Bult, O. W. White, G. J. Olsen, L. Z. Zhou, R. D. Fleischmann, G. Granger, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, D. Jeannine, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. T. Tomb, D. Mark, M. D. Adams, C. I. Reich, R. O. Overbeek, E. F. Kirkness, K. G. Weinstock, M. Joseph, J. M. Merrick, A. G. Glodek, J. L. Scott, and N. S. M. Geoghagen, *Science*, 1996, **273**, 1058–1073.
12. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, 1977, **112**, 535–542.
13. B. Rost and C. Sander, *Curr. Opin. Biotechnol.*, 1994, **5**, 372–380.
14. B. Rost and C. Sander, *Annu. Rev. Biophys. Biomol. Struct.*, 1996, **25**, 113–136.
15. R. F. Doolittle, 'Computer Methods for Macromolecular Sequence Analysis', Academic Press, San Diego, CA, 1996.
16. M. J. E. Sternberg, 'Protein Structure Prediction', Oxford University Press, Oxford, 1996.
17. P. Bork and T. J. Gibson, *Meth. Enzymol.*, 1996, **266**, 162–184.
18. B. Rost and R. Schneider, 'Pedestrian Guide to Analysing Sequence Databases', ed. K. Ashman, 'Core Techniques in Biochemistry', Springer, Heidelberg, 1998, in press.
19. B. Rost and A. Valencia, *Curr. Opin. Biotechnol.*, 1996, **7**, 457–461.
20. W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
21. B. Rost, *Meth. Enzymol.*, 1996, **266**, 525–539.
22. M. Scharf, 'CONAN (CONtact ANalysis)', Heidelberg, 1988.
23. P. Kraulis, *J. Appl. Crystallogr.*, 1991, **24**, 946–950.
24. A. Bairoch and R. Apweiler, *Nucl. Acids Res.*, 1997, **25**, 31–36.
25. C. Chothia and A. M. Lesk, *EMBO J.*, 1986, **5**, 823–826.

26. R. F. Doolittle, 'Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences', University Science Books, Mill Valley, CA, 1986.
27. C. Sander and R. Schneider, *Proteins*, 1991, **9**, 56–68.
28. R. Schneider, A. de Daruvar, and C. Sander, *Nucl. Acids Res.*, 1997, **25**, 226–230.
29. G. Casari, A. De Daruvar, C. Sander, and R. Schneider, *Trends Genetics*, 1996, **12**, 244–245.
30. B. Rost, *Folding & Design*, 1997, **2**, S19–S24.
31. B. Rost, R. Schneider, and C. Sander, *J. Mol. Biol.*, 1997, **270**, 471–480.
32. B. Rost, R. Casadio, and P. Fariselli, *Prot. Sci.*, 1996, **5**, 1704–1718.
33. A. Lupas, *Trends Biol. Sci.*, 1996, **21**, 375–382.
34. S. B. Needlman and C. D. Wunsch, *J. Mol. Biol.*, 1970, **48**, 443–453.
35. T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 1981, **147**, 195–197.
36. S. H. Bryant and L. M. Amzel, *J. Int. Pept. Prot. Res.*, 1987, **29**, 46–52.
37. S. F. Altschul and W. Gish, *Meth. Enzymol.*, 1996, **266**, 460–480.
38. D. G. Higgins, J. D. Thompson, and T. J. Gibson, *Meth. Enzymol.*, 1996, **266**, 383–402.
39. W. R. Pearson, *Meth. Enzymol.*, 1996, **266**, 227–258.
40. W. R. Taylor, *Meth. Enzymol.*, 1996, **266**, 343–367.
41. R. Schneider, G. Casari, d. D. Antoine, P. Bremer, M. Schlenkrich, R. Mercille, H. Vollhardt, and C. Sander, 'GeneCrunch: Experiences on the SGI POWER CHALLENGE Array with Bioinformatics Applications. Proceedings of the Supercomputer 96 Seminar, Mannheim', K. G. Saur, pp. 108–119; **http://www.embl-heidelberg.de/~schneide/**
42. S. H. Bryant and S. F. Altschul, *Curr. Opin. Struct. Biol.*, 1995, **5**, 236–244.
43. S. Henikoff and J. G. Henikoff, *Proteins*, 1993, **17**, 49–61.
44. A. A. Salamov and V. V. Solovyev, *J. Mol. Biol.*, 1995, **247**, 11–15.
45. B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.*, 1994, **235**, 13–26.
46. M. Levitt and C. Chothia, *Nature*, 1976, **261**, 552–558.
47. B. Rost and C. Sander, *Proteins*, 1994, **19**, 55–72.
48. F. E. Cohen and S. R. Presnell, 'The Combinatorial Approach', in 'Protein Structure Prediction', ed. M. J. E. Sternberg, Oxford University Press, Oxford, 1996, pp. 207–228.
49. B. K. Lee and F. M. Richards, *J. Mol. Biol.*, 1971, **55**, 379–400.
50. C. Chothia, *J. Mol. Biol.*, 1976, **105**, 1–12.
51. S. R. Holbrook, S. M. Muskal, and S.-H. Kim, *Prot. Eng.*, 1990, **3**, 659–665.
52. B. Rost and S. I. O'Donoghue, *CABIOS*, 1997, **13**, 345–356.
53. G. von Heijne, 'Prediction of Transmembrane Protein Topology', in 'Protein Structure Prediction', ed. M. J. E. Sternberg, Oxford Univ. Press, Oxford, 1996, pp. 101–110.
54. W. R. Taylor, D. T. Jones, and N. M. Green, *Proteins*, 1994, **18**, 281–294.
55. B. Persson and P. Argos, *Prot. Sci.*, 1996, **5**, 363–371.
56. M. Nilges, *Curr. Opin. Str. Biol.*, 1996, **6**, 617–623.
57. D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug, *J. Mol. Biol.*, 1987, **193**, 693–707.
58. U. Goebel, C. Sander, R. Schneider, and A. Valencia, *Proteins*, 1994, **18**, 309–317.
59. E. Neher, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 98–102.
60. W. R. Taylor and K. Hatrick, *Prot. Eng.*, 1994, **7**, 341–348.
61. T. J. P. Hubbard, 'Use of β-strand Interaction Pseudo-potential in Protein Structure Prediction and Modelling', in 'Proceedings of the 27th Hawaii International Conference on System Sciences', ed. L. Hunter, IEEE, New York, 1994, pp. 336–344.
62. S. Lifson and C. Sander, *J. Mol. Biol.*, 1980, **139**, 627–639.
63. T. J. P. Hubbard and J. Park, *Proteins*, 1995, **23**, 398–402.
64. M. Karplus and G. A. Petsko, *Nature*, 1990, **347**, 631–639.
65. N. L. Summers and M. Karplus, *J. Mol. Biol.*, 1990, **216**, 991–1016.
66. A. C. W. May and T. L. Blundell, *Curr. Opin. Biotechnol.*, 1994, **5**, 355–360.
67. J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis, *Proteins*, 1995, **23**, ii–iv.
68. S. J. Wodak and M. J. Rooman, *Curr. Opin. Struct. Biol.*, 1993, **3**, 247–259.
69. D. Fischer, D. W. Rice, J. U. Bowie, and D. Eisenberg, *FASEB J.*, 1996, **10**, 126–136.
70. R. H. Lathrop, *Prot. Eng.*, 1994, **7**, 1059–1068.
71. M. J. Sippl, *Curr. Opin. Struct. Biol.*, 1995, **5**, 229–235.
72. J. U. Bowie, R. Lüthy, and D. Eisenberg, *Science*, 1991, **253**, 164–169.
73. M. J. Sippl and S. Weitckus, *Proteins*, 1992, **13**, 258–271.
74. C. M.-R. Lemer, M. J. Rooman, and S. J. Wodak, *Proteins*, 1995, **23**, 337–355.
75. A. Elofsson, S. M. Le Grand, and D. Eisenberg, *Proteins*, 1995, **23**, 73–82.
76. J. T. Pedersen and J. Moult, *Curr. Opin. Struct. Biol.*, 1996, **6**, 227–31.
77. R. Srinivasan and G. D. Rose, *Proteins*, 1995, **22**, 81–99.
78. M. Nilges and A. T. Brünger, *Proteins*, 1993, **15**, 133–146.
79. M. J. Sippl, *J. Mol. Biol.*, 1990, **213**, 859–883.
80. B. Honig and F. E. Cohen, *Folding & Design*, 1996, **1**, R17–R20.
81. A.-S. Yang and B. Honig, *J. Mol. Biol.*, 1995, **252**, 351–365.
82. A.-S. Yang and B. Honig, *J. Mol. Biol.*, 1995, **252**, 366–376.
83. M. J. Sippl, *J. Mol. Biol.*, 1996, **260**, 644–648.
84. S. I. O'Donoghue and M. Nilges, *Folding & Design*, 1997, **2**, S47–S52.