BRIDGING THE PROTEIN SEQUENCE–STRUCTURE GAP BY STRUCTURE PREDICTIONS

Burkhard Rost and Chris Sander

European Molecular Biology Laboratory, 69012 Heidelberg, Germany

KEY WORDS: multiple alignments, secondary structure, solvent accessibility, transmembrane helices, interresidue contacts, homology modeling, threading, knowledge-based mean-force potentials

---

ABSTRACT

The problem of accurately predicting protein three-dimensional structure from sequence has yet to be solved. Recently, several new and promising methods that work in one, two, or three dimensions have invigorated the field. Modeling by homology can yield fairly accurate three-dimensional structures for approximately 25% of the currently known protein sequences. Techniques for cooperatively fitting sequences into known three-dimensional folds, called threading methods, can increase this rate by detecting very remote homologies in favorable cases. Prediction of protein structure in two dimensions, i.e. prediction of interresidue contacts, is in its infancy. Prediction tools that work in one dimension are both mature and generally applicable; they predict secondary structure, residue solvent accessibility, and the location of transmembrane helices with reasonable accuracy. These and other prediction methods have gained immensely from the rapid increase of information in publicly accessible databases. Growing databases will lead to further improvements of prediction methods and, thus, to narrowing the gap between the number of known protein sequences and known protein structures.


CONTENTS

113

INTRODUCTION

Large-scale sequencing projects produce data of gene and, hence, pro-
tein sequences at a breathtaking pace. Although determination of pro-
tein three-dimensional structure by crystallography has become more
efficient (51), the gap between the number of known sequences (45,000;
5, 7) and the number of known structures (3000; 8) is increasing rapidly.
For many proteins, sequence determines structure uniquely, i.e. the
entire information for the details of three-dimensional structure is con-
tained in the sequence (4). In principle, therefore, protein structure
could be predicted from physicochemical principles given only the se-
quence of amino acids. In practice, however, prediction from first prin-
ciples, e.g. by molecular dynamics, is prevented by the high complexity
of protein folding (with required computing time orders of magnitude
too high) and by the inaccuracy of the experimental determination of
basic parameters (93). Most protein structure prediction tools, therefore,
are knowledge based, using a combination of statistical theory and
empirical rules. Given a protein sequence of unknown structure (dubbed
U), what can we uncover regarding the structure of U by using theoreti-
cal tools, or what can theory contribute to bridging the sequence-struc-
ture gap?

The most successful tool for predicting three-dimensional structure
is homology modeling. An approximate three-dimensional model
(which has a correct fold but inaccurate loop regions) can be constructed
if U has significant similarity to a protein of known structure, evaluated
in terms of pairwise sequence identity (i.e. by alignment) or sequence-
structure fitness (i.e. threading). Homology modeling effectively raises
the number of ''known'' three-dimensional structures from 3000 to
approximately 10,000 (80). Threading methods may be used to make
tentative predictions of three-dimensional structure for approximately
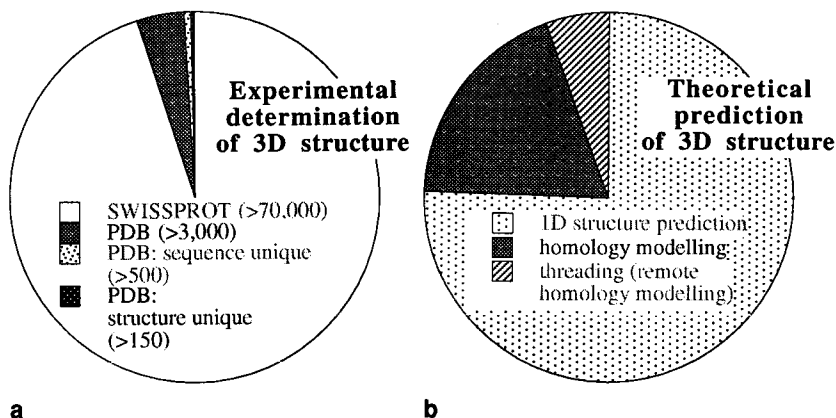
Figure 1   Bridging the sequence-structure gap by experiment and theory. The full-clock cycle corresponds to all protein sequences stored in the database SWISSPROT (release 31 with 44,000 sequences). (*a*) Fraction of proteins for which three-dimensional structure has been experimental determined (sequence unique: < 25% pairwise sequence identity; structure unique: unique overall fold-type, as defined by Reference 36). (*b*) Fraction of proteins for which three-dimensional structure can be predicted by homology modeling (estimated for threading). Note: unique three-dimensional structures cannot be predicted, yet.

an additional 3000 proteins. Consequently, theory-based tools already contribute significantly to bridging the sequence-structure gap (Figure 1). If U has no homologue of known three-dimensional structure, however, we are forced to resort to simplifications of the prediction problem. In the process, we can use the rich diversity of information in current databases. In this review, we focus on generic methods for prediction at three different levels of simplification (Figure 2), namely one, two, and three dimensions (Figure 3). We have included only methods that are available by automatic prediction services or programs and, thus, could be used to analyze large numbers of sequences, e.g. entire chromosomes (25, 42). The underlying question for every method is, What is the practical contribution of the method to the problems of protein structure prediction and analysis?

SEQUENCE ALIGNMENTS

At the level of protein molecules, selective pressure results from the need to maintain function, which in turn requires maintenance of the specific three-dimensional structure (21). This process is the basis for attempts to align protein sequences, i.e. to detect equivalent positions

```
....,....1....,....2....,....3
SeqKEGYLVKKSDGCKYGCLKLGENEGCDTECK
Sec    EE                HHHHHHHH
Acc67240259850365839876949 7048506

....,....4....,....5....,....6....,
SeqAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSC
Sec        EEEEE    EEEEE
Acc59825386051877124056168936468398688
```
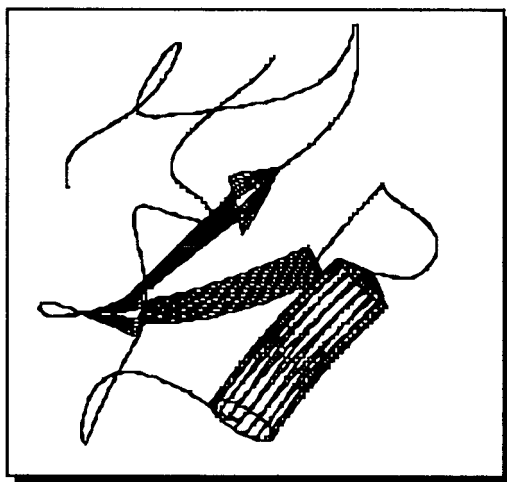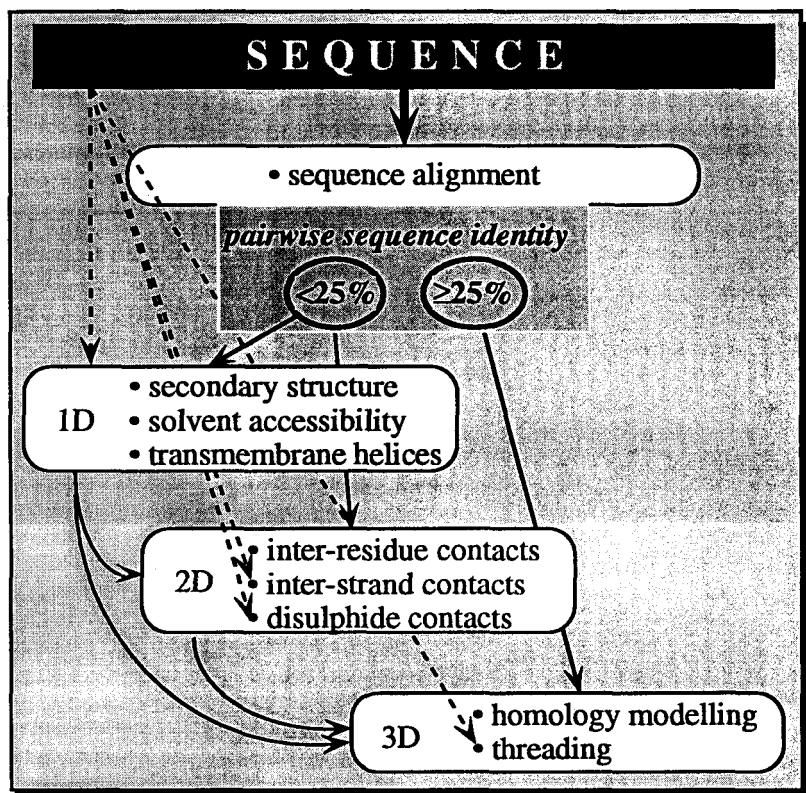
1 D

2 D

3 D

116
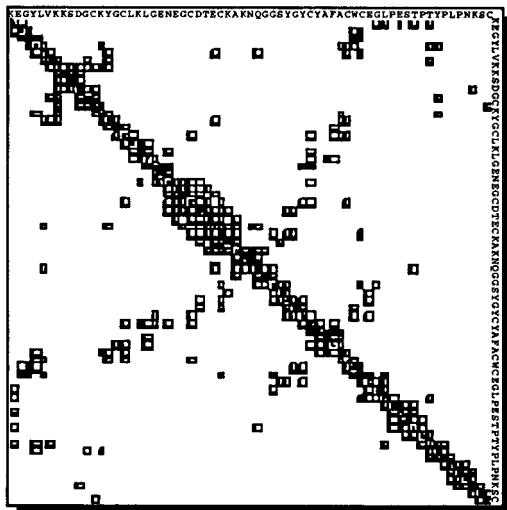
Figure 3   Summary of the tools available for sequence analysis reviewed here. The arrows indicate the input information used for a given method, e.g. secondary structure can be predicted from single sequences and alignments; the two-dimensional prediction can be used, in turn, for prediction of interstrand contacts and threading.

Figure 2   Representation of scorpion neurotoxin (PDB code 2sn3) in one, two, and three dimensions. Each of the representations gives rise to a different type of prediction. (*1D*) Seq, sequence in one-letter alphabet; Sec, secondary structure, with H for helix, E for strand, and blank for other; Acc, relative solvent accessibility (note: integer n codes for a relative accessibility of $n \times n\%$). (*2D*) Interresidue contact-map (sequence positions 1–65 plotted from left to right and from up to down); squares indicate that the respective residue pair is in contact. (*3D*) The trace of the protein chain in three dimensions is plotted schematically as a ribbon α-carbon trace. The two strands are indicated by arrows, the helix is marked by a cylinder. Graphs were generated with the use of WHAT IF, a molecular graphics package with modules for homology modeling, drug design, and protein structure analysis (96).

in strings of amino acid letters optimally. Accordingly, conservation and mutation patterns observed in alignments contain very specific information regarding three-dimensional structure. Surprisingly, 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 three-dimensional structure (79). Even so, structure may be conserved in spite of much higher divergence (36). One naturally wonders how much data are required to detect structure-specific sequence motifs (67) and to align correctly even remote homologues (i.e. sequences with fewer than 25% pairwise identical residues)?

When the level of pairwise sequence identity is sufficient (say, > 40%), alignment procedures are (more or less) straightforward (24, 44, 79). With the use of fast alignment tools, one can scan entire databases that contain 100,000 sequences in minutes. Two fast sequence alignment programs are FASTA (65) and BLAST (3). For less similar protein sequences, however, alignments may fail (30, 94). The art of sequence alignment is to align related sequence segments accurately and to avoid aligning unrelated sequence stretches (20, 22, 31, 48, 52, 57, 75, 79, 92). Alignment techniques can be improved by incorporating information derived from three-dimensional structures (30). Profile-based multiple alignments appear to be sensitive and fast enough to scan entire databases if implemented on parallel machines (80).

One of the difficulties in comparing different alignment procedures is the lack of well-defined criteria for measuring the quality of an alignment. Very few papers have attempted to define such measures for the comparison of various methods (22, 30). The second problem for users is that most methods do not supply a cutoff criterion for distinguishing between homologous and nonhomologous sequences (i.e. false positive sequences). For some large sequence families, remote homologues can be aligned correctly (57, 92); for most cases, however, sequences with less than 25% sequence identity will be false positive, i.e. will have no structural or functional similarity to the guide sequence. A simple, length-dependent cutoff based on sequence identity is provided by MAXHOM, which is a profile-based, multiple-sequence alignment program that also runs in parallel complexes (79). This program, however, does not quantify the influence of (more subtle) similarities and of the occurrence of gaps.

EVALUATION OF PREDICTION METHODS

A systematic testing of performance is a precondition for any prediction to become reliably useful. For example, the history of secondary struc-

ture prediction has partly been a hunt for highest accuracy scores, with overly optimistic claims by predictors seeding the skepticism of potential users. In 1994, one major point about prediction methods became clear at the first international meeting for the evaluation of these methods in Asilomar, California (18): Exaggerated claims are more damaging than genuine errors. Even a prediction method of limited accuracy can be useful if the user knows what to expect. For the editors of scientific journals, this statement implies that a protein structure prediction method should be published only if it has been sufficiently cross-validated. This raises the difficult question of how to evaluate prediction methods.

When a data set is separated into a training set (used to derive the method) and a test set (or cross-validation set, used to evaluate performance), a proper evaluation (or cross-validation) of prediction methods needs to meet four requirements:

1. No significant pairwise sequence identity between training and test set. The proteins used for setting up a method (training set) and those used for evaluating it (test set) should have a pairwise sequence identity of less than 25% [length-dependent cutoff (79)], otherwise homology modeling could be applied that would be much more accurate than ab initio predictions (74, 76).
2. Comprehensive tests through using a large data set. All available unique proteins should be used for testing [currently > 400 (32)]. The reason for taking as many proteins as possible is simply that proteins vary considerably in structural complexity; certain features are easy to predict, others are harder (see Figure 5).
3. Avoid comparing apples with oranges. No matter which data sets are used for a particular evaluation, a standard set should be used for which results are also always reported (see Figure 4).
4. No optimization with respect to the test set. A seemingly trivial—and often violated—rule is that methods should never be optimized with respect to the data set chosen for final evaluation. In other words, the test set should never be used before the method is set up. (For example, using a cross-validation set to indicate when overtraining on the training data has occurred or to find out how many parameters should be used to describe the model is an implicit use of the cross-validation set in parameter optimization. The data reserved to test the method, therefore, should never be used in two ways.)

Most methods are evaluated in n-fold cross-validation experiments (splitting the data set into n different training and test sets). How many

separations should be used, i.e. which value of *n* yields the best evalua-
tion? A misunderstanding is often spread in the literature: the more
separations (the larger *n*) the better. The exact value of *n*, however,
is not important, provided that the test set is representative and compre-
hensive and that the cross-validation results are not misused to change
parameters again. In other words, the choice of *n* is meaningless for
the user.

PREDICTION IN ONE DIMENSION

Secondary Structure

The principal idea underlying most secondary structure prediction
methods is the fact that segments of consecutive residues have prefer-
ences for certain secondary structure states (46). The prediction prob-
lem, therefore, becomes a pattern-classification problem tractable by
computer algorithms. The goal is to predict whether the residue at the
center of a segment of typically 13–21 adjacent residues is in a helix,
a strand, or in no regular secondary structure. Many different algorithms
have been applied to tackle this simplest version of the protein-structure
prediction problem (70, 72). Until recently, however, performance ac-
curacy seemed to have been limited to approximately 60% (percentage
of residues correctly predicted in either α-helix, β-strand, or another
conformation).

   The use of evolutionary information in sequence has improved pre-
diction accuracy significantly. The first method that reached a sustained
level of a three-state prediction accuracy greater than 70% was the
profile-based neural network program PHD, which uses multiple se-
quence alignments as input (70). By stepwise incorporation of more
evolutionary information, prediction accuracy can be pushed to greater
than 72% (72). A nearest-neighbor algorithm can be used to incorporate
the same information with a similar performance (77) (Figure 4). A
method that combines statistics and multiple alignment information
(53) is clearly less accurate (Figure 4). Compared with methods that
use single-sequence information only, methods that use the growing
databases are 6–14 percentage points more accurate (Figure 4).

   How good is a prediction accuracy of 72%? It is certainly reasonably
good compared with the prediction of secondary structure by homology
modeling (16, 74, 76). In addition, some residues within a structure
are predicted at higher levels of accuracy than the mean value, i.e.
prediction accuracy is 72% ± 9% (one standard deviation; Figure 5).
Various applications of improved secondary-structure predictions prove
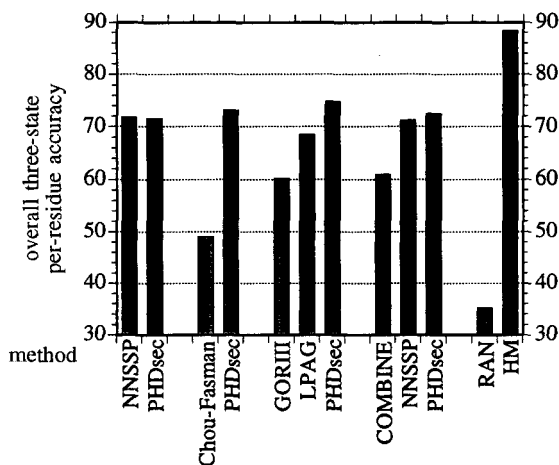
Figure 4   Accuracy of secondary structure prediction for various prediction methods. Abbreviations: RAN and HM, for comparison the results of the worst (random) and the best (homology modeling) possible predictions are given (74); Chou-Fasman, GORIII, and COMBINE, early prediction method based on single-sequence information (9, 15, 28) (these methods are still widely used by standard sequence analysis packages); LPAG, multiple alignment-based method using statistics (53); NNSSP, multiple alignment-based method using nearest neighbor algorithms (77); PHDsec, multiple alignment-based neural network prediction (72). The groups indicate identical test sets, e.g. GORIII is approximately eight percentage points less accurate than LPAG using the same algorithm but additional multiple alignments, and PHDsec is another six percentage points more accurate than LPAG by using neural networks instead of statistics.

that predictions are accurate enough to be of practical use [prediction-based threading, (40, 68); interstrand contact prediction, (39); chain tracing in X-ray crystallography; design of residue mutations]. One way to increase the 72% $\pm$ 9% accuracy level might be to predict secondary-structure content (proportion of residues in α-helix, β-strand, and other) and then use this initial classification to refine secondary structure prediction.

Proteins have been partitioned into various structural classes, e.g. on the basis of percentage of residues assigned to α-helix, β-strand, and other conformations (55). Such a coarse-grained classification, however, is not well defined (36). Consequently, given a protein sequence U, attempts to 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. Alignment-based predictions compare favorably with experimental means of determining the content in secondary structure. Surprisingly, PHD is, on average,

(a) Distribution of prediction accuracy
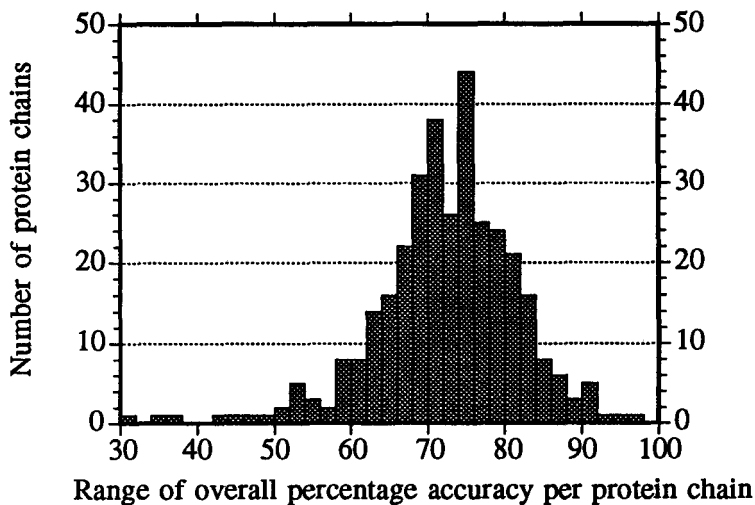
(b) Reliability of prediction
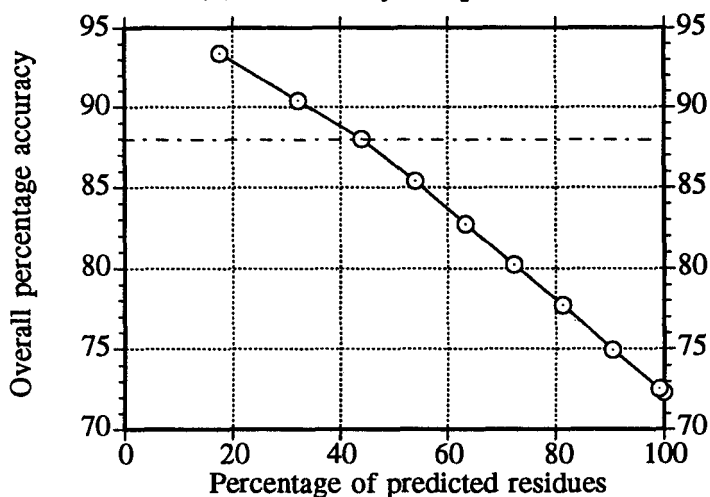
about as accurate as circular dichroism spectroscopy (70, 72). Of course, this finding does not imply that predictions can replace experiments. In particular, variation of secondary structure as a result of changes in environmental conditions (e.g. solvent) is generally accessible only experimentally.

One attempt to improve secondary structure predictions was to develop methods specifically for all-α helix proteins. Two points often have been confused in the literature. First, a two-state accuracy (helix, nonhelix) is not comparable to a three-state accuracy (helix, strand, other). For example, PHD of secondary structure (PHDsec) has an expected three-state accuracy of approximately 72% and an expected two-state accuracy of approximately 82% (71). Second, before a method specialized on all-α proteins can be applied to U, the structure type of U has to be predicted. Such a prediction has an expected accuracy of 70–80% (72). Even if the accuracy for determining whether U belongs to the all-α class reaches almost 100% (99), as recently claimed, specialized methods are still not very useful, as the improvement in accuracy by specializing on one class has been only marginal (71).

Solvent Accessibility

The principal goal is to predict the extent to which a residue embedded in a protein structure is accessible to solvent. Solvent accessibility can be described in several ways (73). The simplest is a two-state description distinguishing between residues that are buried (relative solvent accessibility $<16\%$) and exposed (relative solvent accessibility $\leq$ 16%). The classic method is to assign either of the two states, buried or exposed, according to residue hydrophobicity (for overview, see 73). A neural network prediction of accessibility, however, has been shown to be superior to simple hydrophobicity analyses (33).

Solvent accessibility at each position of the protein structure is conserved evolutionarily within sequence families (73). This fact has been used to develop methods for predicting accessibility using multiple

Figure 5 Secondary structure prediction accuracy for PHDsec evaluated on 337 protein families. (*a*) Prediction accuracy varies considerably between protein families. One standard deviation is nine percentage points, so prediction accuracy for most sequences is 63–81%, and the average accuracy is 72%. Because of this significant variation, prediction methods have to be evaluated on a sufficiently large set of unique proteins. (*b*) Residues with a higher reliability index are predicted with higher accuracy. For example, for 44% of all residues prediction accuracy is, on average, 88% (*dashed line*), i.e. comparable to homology modeling if it were applicable. In practice, attention should be focused on the most reliably predicted residues.
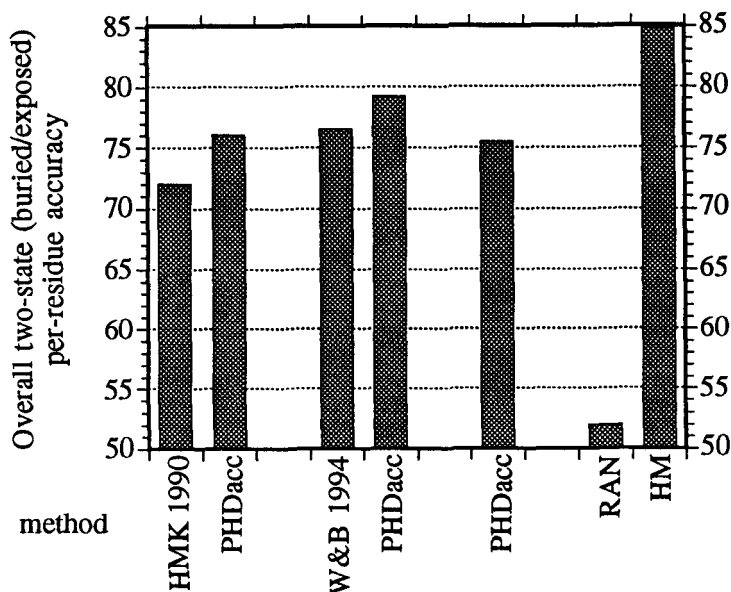
Figure 6 Two-state accuracy of predicting relative accessibility. Abbreviations: RAN and HM, for comparison the results of the worst (random) and the best (homology modeling) possible predictions are given (73); HMK 1990, neural network using single-sequence input (33); W&B 1994, multiple alignment–based prediction method using rather sophisticated expert rules and statistics (97); PHDacc, multiple alignment-based neural network prediction (73). The groups indicate identical test sets.

alignment information (6, 73, 97). Prediction accuracy is approximately 75% ± 7%, four percentage points higher than for methods not using alignment information (Figure 6). Predictions are accurate enough to be used as a seed for predicting secondary structure (6, 97) but not accurate enough to become useful as secondary structure predictions (68).

Transmembrane Helices

Even in the optimistic scenario that, in the near future, most protein structures will be either determined experimentally or predicted theoretically, one class of proteins will still represent a challenge for experimental determination of three-dimensional structure: transmembrane proteins. The major obstacle with these proteins is that they do not crystallize and are hardly tractable by NMR spectroscopy. For this class of proteins, therefore, structure prediction methods are needed even more than for globular water-soluble proteins. Fortunately, the predic-

tion 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 three-dimensional structure formation becomes almost a two-dimensional problem. Once the location of transmembrane segments is known for helical transmembrane proteins, three-dimensional structure can be predicted by exploring all possible conformations (91). Additionally, the prediction of the locations of these transmembrane helices is a much simpler problem than is the prediction of secondary structure for soluble proteins. Elaborated combinations of expert rules, hydrophobicity analyses, and statistics yield a two-state per-residue level of accuracy greater than 90% (43, 69, 83, 95).

Evolutionary information further improves prediction accuracy. For two methods, the use of multiple alignment information is reported to improve the level of accuracy of predicting transmembrane helices (66, 69). The best current prediction methods have a similar high level of accuracy of approximately 95%. 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 authors often report different locations for transmembrane regions. Thus, the 95% level of accuracy is not verifiable. Despite this uncertainty in detail, the prediction of transmembrane helices is a valuable tool to scan entire chromosomes quickly (69). The classification into membrane/nonmembrane proteins has an expected error rate of less than 5%, i.e. approximately 5% of the proteins predicted to contain transmembrane regions will probably be false positive.

Cytoplasmic and extracellular regions have different amino acid compositions (61, 95). This difference allows for a successful prediction of the orientation of transmembrane helices with respect to the cell (pointing inside or outside the cell; 43, 83). Such predictions are estimated to be correct in more than 75% of all proteins (43). Going one step further, Taylor and colleagues (91) have correctly predicted the three-dimensional structure for the membrane-spanning regions of G-coupled receptors (seven helices) when starting from the known locations of the helices. For a successful automatic prediction of three-dimensional structure from sequence, the N- and C-terminal ends of transmembrane helices have to be predicted very accurately. It remains to be tested whether current prediction methods for the location of transmembrane helices are sufficiently accurate to predict three-dimensional structure of integral membrane proteins automatically.

PREDICTION IN TWO DIMENSIONS

Interresidue Contacts

Given all interresidue contacts or distances (see Figure 2), three-dimensional structure can be reconstructed by distance geometry (13, 63). Distance geometry is used for the determination of three-dimensional structures by NMR spectroscopy, which produces experimental data of distances between protons (13). Some fraction of interresidue contacts can be predicted. Helices and strands can be assigned on the basis of hydrogen-bonding patterns between residues (45). Thus, a successful prediction of secondary structure implies a successful prediction of some fraction of all the contacts. Contacts predicted from secondary structure assignment, however, 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 the sequence. A few methods have been proposed for the prediction of long-range interresidue contacts. Two questions surround such methods: First, how accurate are these prediction methods on average. Second, are all important contacts predicted?

In sequence alignments, some pairs of positions appear to co-vary in a physicochemically plausible manner, i.e. a ''loss of function'' point mutation often is rescued by an additional mutation that compensates for the change (2). One hypothesis is that compensation 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 (62, 90) and to use it for contact predictions (29, 81). By applying a stringent significance cutoff in the prediction of contacts by correlated mutations, a small number of residue contacts can be predicted between 1.4 and 5.1 times better than random (29); further slight improvements are possible (D Thomas, unpublished data). These predictions are still not accurate enough to apply distance geometry to the results.

Analyzing correlated mutations is only one way to predict long-range interresidue contacts. Other methods use statistics (26), mean-force potentials (X Tamames and A Valencia, unpublished data), or neural networks (10). So far, none of the methods appears to find a path between the Scylla of missing too many true contacts and the Charybdis of predicting too many false contacts. Some of the methods, however, may provide sufficient information to distinguish between alternative models of three-dimensional structure (A Valencia, unpublished data). The ambitious goal to predict long-range interresidue contacts accurately enough will hopefully continue to attract intellectual resources.

Interstrand Contacts

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

The only method published for predicting interstrand contacts is based on potentials of mean force (39) similar to those used in the evaluation of strand-strand threading (56). Propensities are compiled by database counts for $2 \times 2 \times 2$ classes (parallel/antiparallel, H-bonded/non–H-bonded, N-/C-terminal). Each of the eight classes is divided further into five subclasses in the following way: Suppose the two strand residues at positions i and j are in 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.

Even if the locations of β-strands in the sequence are known exactly, the pseudo-potentials cannot predict the correct interstrand contacts in most cases (39). When using multiple alignment information, however, the signal-to-noise ratio increases such that interstrand contacts have been predicted correctly for most of the strands inspected in some test cases (39). For the purpose of reliable contact prediction, this result is inadequate, especially as the locations of the strands are not known precisely. The pseudo-potentials apparently can handle errors resulting from incorrect prediction of strands. Various test examples using predictions by PHDsec (72) as input to the β-strand pseudo-potentials indicate that the accuracy in predicting interstrand contacts drops (T Hubbard, unpublished data) but, in some cases, is still high enough to be useful for approximate modeling of three-dimensional structure (40).

Intercysteine Contacts

An extreme simplification of the contact prediction problem focuses on predicting contacts between cysteine residues (disulfide bridges). Previously, such contacts were obtained by experimental protein sequencing techniques. In the age of gene-sequencing projects, however, disulfide bridges are no longer part of the sequence information. Disulfide bond predictions are interesting for two reasons: First, disulfide bridges are crucial for structure formation of many proteins. Second, contacts between cysteines account for the most dominant signal in predicting interresidue contacts by mean-force potentials. The prediction of cysteine-bridges, therefore, is a subject of current interest.

One method for the prediction of disulfide-bonds uses a neural net-

work to predict the bonding state of single cysteines (60), i.e. the goal is not to predict which cysteine pair is in contact but whether a cysteine residue is in contact to any other one. Strictly speaking, therefore, the method operates in one dimension. One result is that the cysteine bonding state appears to be influenced by the local sequence environment of up to 15 adjacent residues. Prediction accuracy in two states is claimed to be approximately 80%. The result, however, may be overly optimistic for two reasons. First, the test set was rather small (140 examples). Second, in the cross-validation experiments, training and testing examples were not separated on the basis of the level of pairwise sequence identity. Therefore, the question of how accurately intercysteine contacts can be predicted remains to be answered.

PREDICTION IN THREE DIMENSIONS

Homology Modeling

An analysis of the Protein Data Bank (PDB) of experimentally determined structures of protein reveals that all protein pairs with more than 30% pairwise sequence identity (for alignment length > 80; 79) have homologous three-dimensional structures, i.e. the essential fold of the two proteins is identical, but such details as additional loop regions may vary. Structure is more conserved than is sequence. This finding is the pillar for the success of homology modeling. The principal idea is to model the structure of U on the basis of the template of a sequence homologue of known structure. Consequently, the precondition for homology modeling is that a sequence homologue of known structure is found in PDB. Because homology modeling is currently the only theoretical means to predict three-dimensional structure successfully, this finding has two implications. First, homology modeling is applicable to "only" one quarter of the known protein sequences (see Figure 1). Second, as the template of a homologue is required, no unique three-dimensional structure can yet be predicted, i.e. no structure that has no similarity to any experimentally determined three-dimensional structure. If there is a protein with a sequence similar to U in PDB (say HU), is homology modeling straightforward?

The basic assumption of homology modeling is that U and HU have identical backbones. The task is to place the side chains of U into the backbone of HU correctly. For very high levels of sequence identity between U and HU (ideally differing by one residue only), side chains can be "grown" during molecular dynamics simulations (17, 47). For slightly lower levels (still of high-sequence similarity), side chains are

built on the basis of similar environments in known structures (19, 23, 54, 58, 78, 89, 96). Rotamer libraries are used in the following way (19): 1. Rotamer distributions are extracted from a database of sequences that are not redundant. 2. Fragments of seven (helix, strand) or five residues (other) are compiled. 3. Fragments of the same length are shifted successively through the backbone of U. 4. For modeling the side chains of U, only those fragments from the rotamer library that have the same amino acid in the center as U, and for which the local backbone is similar to that around the evaluated position, are accepted. Over the whole range of sequence identity between U and HU 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 (19). (For even higher values, the models are as accurate as is experimental structure determination.) The limiting factor is the computation time required (34). How accurate is homology modeling for lower levels of sequence identity?

With decreasing sequence identity, the number of loops inserted 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 central processing unit (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 used to model loop regions. The best have the orientation of the loop regions correct in some cases (e.g. 1). With less than approximately 40% sequence identity, the accuracy of the sequence alignment used as the basis for homology modeling becomes an additional problem. Even down to levels of 25–30% sequence identity, however, homology modeling produces coarse-grained models for the overall fold of proteins of unknown structure.

Remote Homology Modeling (Threading)

As noted in the previous section, naturally evolved sequences with more than 30% pairwise sequence identity have homologous three-dimensional structures (79). Are all others nonhomologous? 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 homologues) (36). If a correct alignment between U and a remote homologue RU (pairwise sequence identity to U < 25%) is given, one could build the three-dimensional structure of U by homology modeling on the basis of the template of RU (remote homology

modeling). A successful remote homology modeling must solve three different tasks: 1. RU has to be detected. 2. U and RU have to be aligned correctly. 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 addressed primarily to detect similar folds. The basic idea is to thread the sequence of U into the known structure of RU and to evaluate the fitness of sequence for structure by some kind of environment-based or knowledge-based potential (14, 86). Threading is, in some respects, a harder problem than is the prediction of three-dimensional structure (50, 86). Solving it, however, would enable the prediction of thousands of protein structures (see Figure 1). Can this hard nut be cracked?

The optimism generated by one of the first papers on threading published in the 1990s (11) has boosted attempts to develop threading methods (86). Most methods are based on pseudo-potentials and differ in the way such potentials are derived from PDB (98). One alternative is to use one-dimensional predictions for the threading procedure (68; G Barton, unpublished data; F Drabløs, unpublished data). The good news, after half a decade of intensive research by dozens of groups, is that all potentials capture different aspects, and it is likely that the correct remote homologue is found by at least one of these groups (82). The bad news is that no single method is accurate enough to identify the remote homologue correctly in most cases (82). Instead, evaluated on a larger test set, the correct remote homologue appears to be detected in approximately 30% of all cases (68). Unfortunately, this is only the first of the three tasks for successful remote homology modeling, the second (correct alignment of U and RU) is even harder. In many of the cases for which RU is identified correctly as a remote homologue of U, the alignment of U and RU is flawed in significant ways (unpublished data). This is fatal for the third step, the model-building procedure. Thus, is threading useful, at all?

Like all prediction methods, threading techniques are not error proof. One of the practical disadvantages of current tools is the lack of a successful measure for prediction reliability, such as that established for secondary structure prediction (see Figure 5). The conclusion seems to be that threading methods can be useful in the hands of rather skeptical expert users who can spot wrong hits and false alignments, even when the prediction method suggests a high confidence value for the error it generates. Three points may be added. First, threading techniques can clearly widen the range of successful sequence alignments (68). Second, some methods are accurate enough to be used in scanning entire chromosomes for remote homologues (12). Third, threading tech-

niques may still become one of the most successful tools in structure prediction, but a lot of detailed work lies ahead.

ANALYSIS OF THREE-DIMENSIONAL STRUCTURES

A successful idea was to replace inductive force fields that capture the heuristics of physical principles by deductive, knowledge-based, mean-force potentials (e.g. 84). Such potentials, as well as more expert-knowledge–oriented approaches (49, 96), enable the detection of subtle stresses or possible errors in both experimentally determined three-dimensional structures and predicted models (85). Knowledge-based potentials of mean force appear to be valid even for proteins with properties not used for deriving the potentials [membrane proteins (85); coiled-coils, S O'Donoghue, unpublished data]. Because of this success, quality control tools that use these potentials are becoming a routine check applied to any experimentally determined structure or any structure predicted by homology modeling.

More and more frequently, a newly determined structure is identified to be remotely homologous to a known structure (38). Recently developed algorithms enable routine scans for possible remote homologues in PDB for any new structure (35, 41, 59, 64, 75, 88). Such searches are beginning to rival sequence database searches as a tool for discovering biologically interesting relationships (38). Similar techniques can often be exploited to determine domains in known structures (27, 37, 87).

CONCLUSION

Three-dimensional structure cannot yet be predicted reliably from sequence information alone. In other words, the only source for new, unique structures (structures for which no homologue exists in the database) are experiments. Given the amount of time needed to determine a protein structure experimentally, however, more non-unique structures can be predicted at atomic resolution by homology modeling in 1 month than have been determined by experiment during the past 3 decades. Unfortunately, such models typically have considerable coordinate errors in loop regions, and remote homology modeling (i.e. homology modeling for $< 25\%$ pairwise sequence identity) is not yet reliable. For a few cases, however, threading techniques already have resulted in accurate modeling of the overall fold (86).

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. Predictions of secondary structure, solvent accessibility, and transmembrane helices are becoming increasingly useful. This success is the result of both a better performance of multiple alignment–based methods and the ability to focus on more reliably predicted regions. Some methods have indicated that one-dimensional predictions can be useful as an intermediate step on the way to predicting three-dimensional structure (interstrand contacts, prediction-based threading). Another advantage of predictions in one dimension is that they are not very CPU-intensive, i.e. one-dimensional structure can be predicted for the protein sequence of, for example, entire yeast chromosomes overnight.

The prediction accuracy of chain-distant interresidue contacts is relatively limited so far. Analysis of correlated mutations can be used to distinguish between alternative models (e.g. for threading techniques). The prediction of interstrand 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.

Another encouraging development is the improvement of tools for the analysis of protein structures. Experimental inconsistencies can be spotted, and predicted models can be tested. The ease of scanning structure databases for remote homologues yields a rich amount of information with an effect on our understanding of protein structure and function.

Literature Cited

1. Abagyan R, Totrov M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235:983–1002
2. Altschuh D, Vernet T, Moras D, Nagai K. 1988. Coordinated amino acid changes in homologous protein families. *Protein Eng.* 2:193–99
3. Altschul SF. 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* 36:290–300
4. Anfinsen CB, Scheraga HA. 1975. Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* 29:205–300
5. Bairoch A, Boeckmann B. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.* 22:3578–80
6. Benner SA, Badcoe I, Cohen MA, Gerloff DL. 1994. Bona fide prediction of aspects of protein conformation. *J. Mol. Biol.* 235:926–58
7. Benson D, Lipman DJ, Ostell J. 1993. GenBank. *Nucleic Acids Res.* 21: 963–75

8. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, et al. 1977. The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–42

9. Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. 1988. Secondary structure prediction: combination of three different methods. *Protein Eng.* 2: 185–91

10. Bohr H, Bohr J, Brunak S, Fredholm H, Lautrup B, et al. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.* 261: 43–46

11. Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164–69

12. Braxenthaler M, Sippl M. 1995. Screening genome sequences for known folds. In *Protein Structure by Distance Analysis*, ed. H Bohr, S Brunak. Boca Raton, FL: CRC

13. Brünger AT, Nilges M. 1993. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* 26: 49–125

14. Bryant SH, Altschul SF. 1995. Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5:236–44

15. Chou PY, Fasman GD. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47:45–148

16. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J-P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* 6: 377–82

17. Cornell WD, Howard AE, Kollman P. 1991. Molecular mechanical potential functions and their application to study molecular systems. *Curr. Opin. Struct. Biol.* 1:201–12

18. Defay T, Cohen FE. 1995. Evaluation of current techniques for ab initio protein structure prediction. *Proteins.* 23: 431–45

19. De Filippis V, Sander C, Vriend G. 1994. Predicting local structural changes that result from point mutations. *Protein Eng.* 7:1203–8

20. Depereux E, Feytmans E. 1992. MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput. Appl. Biosci.* 8:501–9

21. Doolittle RF. 1994. Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* 19:15–8

22. Eddy SR. 1995. Multiple alignment using hidden Markov models. See Ref. 66a, pp. 114–20

23. Eisenmenger F, Argos P, Abagyan R. 1993. A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* 231:849– 60

24. Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* 2:1811–26

25. Gaasterland T, Selkov E. 1995. Reconstruction of metabolic networks using incomplete information. See Ref. 66a, pp. 127–35

26. Galaktionov SG, Marshall GR. 1994. Properties of intraglobular contacts in proteins: an approach to prediction of tertiary structure. In *27th Hawaii Int. Conf. System Sciences, Wailea HI*, ed. L Hunter, pp. 326–35. Los Alamitos, CA: IEEE Comput. Soc.

27. Gerstein M, Sonnhammer ELL, Chothia C. 1994. Volume changes in protein evolution. *J. Mol. Biol.* 236: 1067–78

28. Gibrat J-F, Garnier J, Robson B. 1987. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* 198:425–43

29. Goebel U, Sander C, Schneider R, Valencia A. 1994. Correlated mutations and residue contacts in proteins. *Proteins* 18:309–17

30. Henikoff S, Henikoff JG. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* 17: 49–61

31. Henikoff S, Henikoff JG. 1994. Position-based sequence weights. *J. Mol. Biol.* 243:574–78

32. Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3:522–24

33. Holbrook SR, Muskal SM, Kim S-H. 1990. Predicting surface exposure of amino acids from protein sequence. *Protein Eng.* 3:659–65

34. Holm L, Rost B, Sander C, Schneider R, Vriend G. 1994. Data based modeling of proteins. In *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions*, ed. S Doniach, pp. 277–96. New York: Plenum

35. Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–38

36. Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* 22:3600–9

37. Holm L, Sander C. 1994. Parser for protein folding units. *Proteins* 19:256–68

38. Holm L, Sander C. 1994. Searching protein structure databases has come of age. *Proteins* 19:165–73

39. Hubbard TJP. 1994. Use of β-strand interaction pseudo-potential in protein structure prediction and modelling. In *27th Hawaii Int. Conf. System Sciences, , Maui, HI,* ed. L Hunter, pp. 336–44. Los Alamitos, CA: IEEE Comput. Soc.

40. Hubbard TJP, Park J. 1995. Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potentials. *Proteins.* In press

41. Johnson MS, Overington JP, Blundell TL. 1993. Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* 231:735–52

42. Johnston M, Andrews S, Brinkman R, Cooper J, Ding H, et al. 1994. Complete nucleotide sequence of *Saccaromyces cerevisiae* chromosome VIII. *Science* 265:2077–82

43. Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature* 358:86–89

44. Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8:275–82

45. Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 22:2577–637

46. Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA* 81:1075–78

47. Karplus M, Petsko GA. 1990. Molecular dynamics simulations in biology. *Nature* 347:631–39

48. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. 1994. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* 235:1501–31

49. Laskowski RA, Moss DS, Thornton JM. 1993. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* 231:1049–67

50. Lathrop RH. 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* 7:1059–68

51. Lattman EE. 1994. Protein crystallography for all. *Proteins* 18:103–6

52. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–14

53. Levin JM, Pascarella S, Argos P, Garnier J. 1993. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* 6:849–54

54. Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226:507–33

55. Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature* 261:552–58

56. Lifson S, Sander C. 1980. Specific recognition in the tertiary structure of β-sheets in proteins. *J. Mol. Biol.* 139:627–39

57. Livingstone CD, Barton GJ. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* 9:745–56

58. May ACW, Blundell TL. 1994. Automated comparative modelling of protein structures. *Curr. Opin. Biotechnol.* 5:355–60

59. Mitchell EM, Artymiuk PJ, Rice DW, Willett P. 1992. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* 212:151–66

60. Muskal SM, Holbrook SR, Kim S-H. 1990. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng.* 3:667–72

61. Nakashima H, Nishikawa K. 1992. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett.* 303:141–46

62. Neher E. 1994. How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. USA* 91:98–102

63. Nilges M. 1993. A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins* 17:297–309

64. Orengo CA, Brown NP, Taylor WT.

1992. Fast structure alignment for protein databank searching. *Proteins* 14: 139–67

65. Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444–48

66. Persson B, Argos P. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* 237:182–92

66a. Rawlings C, Clark D, Altman R, Hunter L, Lengauer T, et al, eds. 1995. *3rd Int. Conf. on Intelligent Systems for Molecular Biology (ISMB), Cambridge, England.* Menlo Park, CA: AAAI

67. Rooman M, Wodak SJ. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature* 335:45–49

68. Rost B. 1995. TOPITS: Threading one-dimensional predictions into three-dimensional structures. See Ref. 66a, pp. 314–21

69. Rost B, Casadio R, Fariselli P, Sander C. 1995. Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* 4:521–33

70. Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584–99

71. Rost B, Sander C. 1993. Secondary structure prediction of all-helical proteins in two states. *Protein Eng.* 6: 831–36

72. Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55–72

73. Rost B, Sander C. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20: 216–26

74. Rost B, Sander C, Schneider R. 1994. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13–26

75. Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14:309–23

76. Russell RB, Barton GJ. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* 234: 951–57

77. Salamov AA, Solovyev VV. 1995. Prediction of protein secondary structure by combining nearest-neighbor

algorithms and multiple sequence alignments. *J. Mol. Biol.* 247:11–15

78. Sali A, Blundell T. 1994. Comparative protein modelling by satisfaction of spatial restraints. In *Protein Structure by Distance Analysis,* ed. H Bohr, S Brunak, pp. 64–87. Amsterdam/Oxford/Washington: IOS Press

79. Sander C, Schneider R. 1991. Database of homology-derived structures and the structurally meaning of sequence alignment. *Proteins* 9:56–68

80. Sander C, Schneider R. 1994. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* 22:3597–99

81. Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* 7:349–58

82. Shortle D. 1995. Protein fold recognition. *Nature Struct. Biol.* 2:91–92

83. Sipos L, von Heijne G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* 213: 1333–40

84. Sippl MJ. 1993. Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Design* 7:473–501

85. Sippl MJ. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355–62

86. Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–35

87. Sternberg MJE, Hegyi H, Islam SA, Luo J, Russell RB. 1995. Towards an intelligent system for the automatic assignment of domains in globular proteins. See Ref. 66a, pp. 376–83

88. Subbiah S, Laurents DV, Levitt M. 1993. Structural similarity of DNA-binding domains of bacterophage repressors and the globin core. *Curr. Biol.* 3:141–48

89. Summers NL, Karplus M. 1990. Modeling of globular proteins. *J. Mol. Biol.* 216:991–1016

90. Taylor WR, Hatrick K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng.* 7: 341–48

91. Taylor WR, Jones DT, Green NM. 1994. A method for α-helical integral membrane protein fold prediction. *Proteins* 18:281–94

92. Thompson JD, Higgins DG, Gibson TJ. 1994. Improved sensitivity of profile searches through the use of se-

quence weights and gab excision. *Comput. Appl. Biosci.* 10:19–29

93. van Gunsteren WF. 1993. Molecular dynamics studies of proteins. *Curr. Opin. Struct. Biol.* 3:167–74

94. Vingron M, Waterman MS. 1994. Sequence alignment and penalty choice. *J. Mol. Biol.* 235:1–12

95. von Heijne G. 1992. Membrane protein structure prediction. *J. Mol. Biol.* 225:487–94

96. Vriend G, Sander C. 1993. Quality of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.* 26:47–60

97. Wako H, Blundell TL. 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.* 238:682–92

98. Wodak SJ, Rooman MJ. 1993. Generating and testing protein folds. *Curr. Opin. Struct. Biol.* 3:247–59

99. Zhu Z-Y. 1995. A new approach to the evaluation of protein secondary structure predictions at the level of the elements of secondary structure. *Protein Eng.* 8:103–8

Annual Review of Biophysics and Biomolecular Structure
Volume 25, 1996

CONTENTS