5

Third Generation Prediction of Secondary Structures

Burkhard Rost and Chris Sander

1. Introduction

The sequence–structure gap is rapidly increasing. Currently, databases for protein sequences (e.g., SWISS-PROT ***[1]***) are expanding rapidly, largely due to large-scale genome sequencing projects: at the beginning of 1998, we know already all sequences for a dozen of entire genomes ***(2)***. This implies that despite significant improvements of structure determination techniques, the gap between the number of protein structures in public databases (PDB ***[3]***), and the number of known protein sequences is increasing. The most successful theoretical approach to bridging this gap is homology modeling. It effectively raises the number of "known" 3D structures from 7000 to over 50,000 ***(4,5)***.

No general prediction of structure from sequence, yet. John Moult (Center for Advance Research in Biotechnology [CARB], Washington) has initiated an important experiment: 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, CA). The results of the first two critical assessment of protein structure prediction (CASP) experiments ***(6,7)*** demonstrated clearly that we still cannot predict structure from sequence.

Simplifying the structure prediction problem. The rapidly growing sequence–structure gap has enticed theoreticians to solve simplified prediction problems ***(4)***. An extreme simplification is the prediction of protein structure in one dimension (1D), as represented by strings of, e.g., secondary-structure or residue solvent accessibility. Theoreticians are lucky in that a simplified

From: *Methods in Molecular Biology*, vol. 143: *Protein Structure Prediction: Methods and Protocols*
Edited by: D. Webster

predictions in 1D (e.g., secondary-structure or solvent accessibility ***[4,8,9]***) even when only partially correct — are often useful, e.g., for predicting protein function, or functional sites.

In this review we focus on recent secondary-structure prediction methods (for reviews on older methods ***(10–17)***, for reviews on other prediction methods in 1D ***[4,5,18]***). We present some of the new, successful concepts and a few "hints for the user" based on the currently most widely used secondary-structure prediction method: PHD.

2. Materials

Assignment of secondary-structure. Secondary-structure is most often assigned automatically based on the hydrogen bonding pattern between the backbone carbonyl and NH groups (e.g., by Dictionary of Secondary Structure assignment of Proteins [DSSP] ***[19]***). DSSP distinguishes eight secondary-structure classes which are often grouped into three classes: H = helix, E = strand, and L = non-regular structure. Typically the grouping is as follows: H (α-helix) -> H, G (3_{10}-helix) -> H, I (π-helix) -> H, E (extended strand) -> E, and B (residue in isolated b-bridge) -> E, T (turn) -> L, S (bend) -> L, (blank = other) -> L, with the "corrections": B -> EE, but B_B -> LLL. Note that developers often use different projections of the eight DSSP classes onto three predicted classes; most of these yield seemingly higher levels of prediction accuracy. For example, short helices are more difficult to predict ***(20)*** (*see also* **Fig. 5**); thus, converting GGG -> LLL results, on average, in higher levels of prediction accuracy.

Per-residue prediction accuracy. The simplest and most widely used score is the three-state-per-residue accuracy, giving the percentage of correctly predicted residues predicted correctly in either of the three states: helix, strand, other:

$$Q_3 = 100 \cdot \sum_{i=1}^{3} c_i / N \quad (1)$$

where c_i is the number of residues predicted correctly in state i (H, E, L), and N is the number of residues in the protein (or in a given data set). Because typical data sets contain about 32% H, 21% E, and 47% L, correct prediction of the nonregular class tends to dominate the three-state accuracy. More fine-grained methods that avoid this shortcoming are defined in detail elsewhere ***(21,22)***.

Per-segment prediction accuracy. Measures for single-residue accuracy do not completely reflect the quality of a prediction ***(14,22–26)***. There are three simple measures for assessing the quality of predicted secondary-structure segments: (1) the number of segments in the protein, (2) the average segment length, and (3) the distribution of the number of segments with length ***(27)***. These measures are related. They are useful in characterizing prediction meth-

ods, in particular, methods with fairly high per-residue accuracy, yet an unrealistic distribution of segments. However, there is a more elaborated score base on the overlap between predicted and observed segments ***(22)***.

Conditions for evaluating sustained performance. A systematic testing of performance is a precondition for any prediction to become reliably useful. For example, the history of secondary-structure prediction has partly been a hunt for highest accuracy scores, with over-optimistic claims by predictors seeding the skepticism of potential users. Given a separation of a data set into a training set (used to derive the method) and a test set (or crossvalidation set, used to evaluate performance), a proper evaluation (or crossvalidation) of prediction methods needs to meet four requirements: (1) no significant pairwise sequence identity between proteins used for training and test set, i.e., < 25% (length-dependent cutoff ***[28]***); (2) all available unique proteins should be used for testing, as proteins vary considerably in structural complexity; certain features are easy to predict, others harder; (3) no matter which data sets are used for a particular evaluation, a standard set should be used for which results are also always reported; (4) 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.

Number of crossvalidation experiments of NO meaning. Most methods are evaluated in *n*-fold crossvalidation experiments (splitting the data set into *n* different training and test sets). How many separations should be used, i.e., which number of *n* yields the best evaluation? A misunderstanding is often spread in the literature: the more separations (the larger *n*) the better. However, the exact number of *n* is not important provided the test set is representative, and comprehensive and the crossvalidation results are not misused to again change parameters. In other words, the choice of *n* has no meaning for the user.

3. Methods

3.1. The Dinosaurs of Secondary Structure Prediction Are Still Alive

First generation: single-residue statistics. The first experimentally determined 3D structures of hemoglobin and myoglobin were published in 1960 ***(29,30)***. Almost a decade before, Pauling and Corey suggested an explanation for the formation of certain local conformational patterns such as α-helices and β-strands ***(31,32)***. Shortly later (and still prior to the first published structure), the first attempt was made to (positively) correlate the content of certain amino acids (e.g., proline) with the content of an α-helix ***(33)***. The idea was expanded to correlate the content for all amino acids with that of the α-helix and the β-strand structure ***(34,35)***. The field of predicting secondary-structures had been opened. Most methods of the first generation based on single-residue

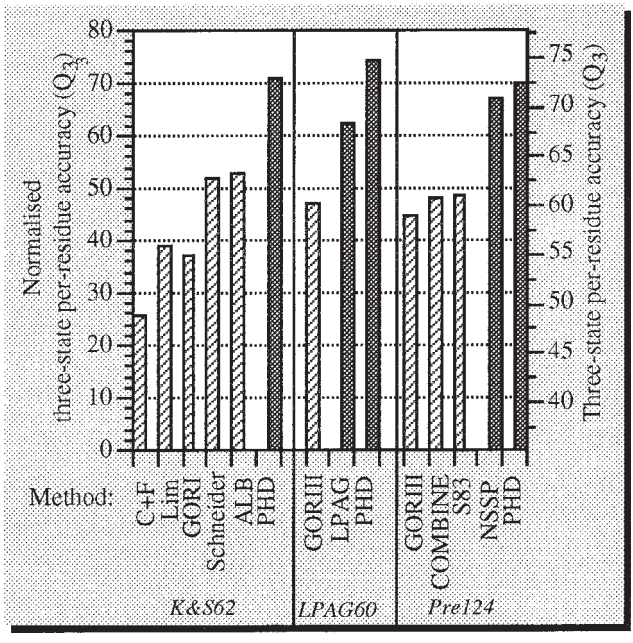

Fig. 1. Three-state-per-residue accuracy of various prediction methods. Shaded bars: methods of first and second generation; filled bars: methods of third generation. The left axis showed the normalized three-state-per-residue accuracy, for which a random prediction would rate 0%, and an optimal prediction by homology modeling would rate as 100% (unnormalized values according to **Eq. 1**, shown on the right axis):

$$\text{norm } Q3 = 100 \cdot Q_3^{\text{method}} - Q_3^{\text{RAN}} / Q_3^{\text{HM}} - Q_3^{\text{RAN}} \quad (1)$$

$$\text{with } Q_3^{\text{HM}} = 88.4\%, \text{ and } Q_3^{\text{RAN}} = 35.2\%$$

Only methods were included for which the accuracy had been compiled based on comparable data sets, the sets in particular are *K&S62*, 62 proteins taken from **ref. *45***; *LPAG60*, 60 proteins taken from **ref. *128***; *Pre124*, 124 unique proteins taken from **ref. *48***. The methods were: *C + F* Chou & Fasman (first generation) ***(42,148)***; *Lim* (first) ***(43)***; *GORI* (first) ***(53)***; *Schneider* (second) ***(87)***; *ALB* (second) ***(62)***; *GORIII* (second) ***(54)***; *LPAG* (third) ***(128)***; *COMBINE* (second) ***(17)***; *S83* (second) ***(86)***; *NSSP* (third) ***(84)***; *PHD* (third) ***(48)***. Most values were recompiled — only those for NSSP and LPAG were taken from the original publications. The scores for PHD on the three different data sets illustrated that data sets can be tuned to give more optimistic (LPAG62), or more realistic estimates for prediction accuracy. The first two structure prediction contests have indicated that the most conservative estimates of this graph (Pre124) tend to be slightly too optimistic, still PHD rates at an average accuracy of about 72% (as originally estimated ***[18,48]***).

statistics, i.e., from the limited databases evidence was extracted for the preference of particular residues for particular secondary-structure states ***(36–44)***. By 1983, it became clear that the performance accuracy had been overstated ***(45)*** (*see* **Fig. 1**).

Second generation: segment statistics. The principal improvement of the second generation of prediction tools was a combination of a larger database of protein structure and the usage of statistics based on segments: typically 11–21 adjacent residues were taken from a protein and statistics were compiled to evaluate how likely the residue central in that segment was in a particular secondary-structure state. Similar segments of adjacent residues were also used to base predictions on more elaborated algorithms, some of which were spun off from artificial intelligence ***(46)***. Almost any algorithm has meanwhile been applied to the problem of predicting secondary-structures; all were limited to accuracy levels slightly higher than 60% (*see* **Fig. 1**; reports of higher levels of accuracy were usually based on too small, or non-representative data sets ***[21,25,47,48]***). The most-used algorithms were based on (1) statistical information ***(49–61)***, (2) physicochemical properties ***(62)***, (3) sequence patterns ***(63–65)***, (4) multilayered (or neural) networks ***(66–73)***, (5) graph theory ***(74,75)***, (6) multivariate statistics ***(76,77)***, (7) expert rules ***(75,78–82)***, and (8) nearest-neighbor algorithms ***(83–85)***.

Problems with first- and second-generation methods. All methods from the first and second generation shared, at least, two of the following problems (most all three): (1) three-state per-residue accuracy was below 70%, (2) β-strands were predicted at levels of 28–48%, i.e., only slightly better than random, and (3) predicted helices and strands were too short.

The first problem (<100% accuracy) has two sources: (1) secondary-structure assignments differ even between different crystals of the same protein, and (2) secondary-structure formation is partially determined by long-range interactions, i.e., by contacts between residues that are not visible by any method based on segments of 11–21 adjacent residues. The second problem (β-strands <50% accuracy) has been explained by the fact that b-sheet formation is determined by more nonlocal contacts than is α-helix formation. The third problem was basically overlooked by most developers (for exceptions, *see* **refs. *86*** and ***87***). This problem makes predictions very difficult to use in practice (*see* **Fig. 2**). As we show in the next paragraph, some of the prediction methods of the third generation address all three problems simultaneously, and are clearly superior to the old methods (*see* **Fig. 1**). Nevertheless, many of the secondary-structure prediction methods available today (e.g., in University of Wisconsin Genetics Computer Group (GCG) ***[88]***, or from Internet services ***[89]***) are unfortunately still using the dinosaurs of secondary-structure prediction.

```
SEQ   KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
OBS     EEEE          E--E    EEEEEE      EEEEEE  EEEEEEEHHHEEEE
TYP     EEHHH         EE    EEEE   EE     HHHEE       EEEHH
```

Fig. 2. Example for typical secondary structure prediction of the second generation. The protein sequence (*SEQ*) given was the SH3 structure ***(131)***. The observed secondary structure (*OBS*) was assigned by DSSP ***(19)*** (H = helix; E = strand; blank = nonregular structure; the dashes indicate the continuation of the second strand that was missed by DSSP). The typical prediction of too short segments (*TYP*) poses the following problems in practice: (1) Are the residues predicted to be strand in segments 1, 5, and 6 errors, or should the helices be elongated? (2) Should the second and third strand be joined, or should one of them be ignored, or does the prediction indicate two strands here? Note: the three-state-per-residue accuracy is 60% for the prediction given.

3.2. Breakthrough By Using Evolutionary Information

3.2.1. Is Evolutionary Odyssey Informative?

Variation in sequence space. The exchange of a few residues can already destabilize a protein ***(90)***. This implies that the majority of the 20^N possible sequences of length N form different structures. Has evolution really created such an immense variety? Random errors in the DNA sequence lead to a different translation of protein sequences. These "errors" are the basis for evolution. Mutations resulting in a structural change are not likely to survive, as the protein can no longer function appropriately. Furthermore, the universe of stable structures is not continuous: minor changes on the level of the 3D structure may destabilize the structure (due to high complexity). Thus, residue exchanges conserving structure are statistically unlikely. However, the evolutionary pressure to conserve function has led to a record of this unlikely event: structure is more conserved than sequence ***(91–93)***. Indeed, all naturally evolved protein pairs that have 35 of 100 pairwise identical residues have similar structures ***(28,94)***. However, the attractors of protein structures are even larger: the majority of protein pairs of similar structures has levels of below 15% pairwise sequence identity ***(95,96)***.

Long-range information in multiple sequence alignments. The residue substitution patterns observed between proteins of a particular family, i.e., changes that conserved structure, are highly specific for the structure of that family. Furthermore, multiple alignments of sequence families implicitly also contain information about long-range interactions: suppose residues i and i + 100 are close in 3D, then the types of amino acids that can be exchanged (without changing structure) at position i are constrained by that their physicochemical characteriztics have to fit the amino acid types at position i + 100 ***(97,98)***.

3.2.2. Can Evolutionary Information Be Used?

Expert predictions: visual use of alignment information. The first method that used information from family alignments was proposed in the 1970s already ***(99)***. Furthermore, experts have based single-case predictions successfully on multiple alignments ***(99–116)***.

Automatic use of alignment information. The simplest way to use alignment information automatically was first proposed by Maxfield and Scheraga and by Zvelebil et al. ***(117,118)***: predictions were compiled for each protein in an alignment, and then averaged over all proteins. A slightly more elaborated way of automatically using evolutionary information is to directly base prediction on a profile compiled from the multiple sequence alignment ***(18,21,48)***. The following steps are applied in particular for the PHD method ***(18,119)*** (*see* **Fig. 3**): (1) A sequence of unknown structure (*U*) is quickly aligned against the database of known sequences (typically by BLAST ***[120]***) (i.e., no information of structure required); (2) proteins with sufficient sequence identity to *U* to assure structural similarity are extracted and realigned by a multiple alignment algorithm *MaxHom* ***(121)***; (3) for each position, the profile of residue exchanges in the final multiple alignment is compiled, and used as input to a neural network.

3.2.3. Third Generation: Evolution to Better Predictions

Example chosen: PHD. We illustrated the principal concepts of third generation methods based on the particular neural network-based method PHD because it is currently the most accurate method ***(7)***, and because most of these concepts were introduced by this method ***(21,48)***. Meanwhile, several other methods have reported and/or achieved similar levels of performance ***(16,18,21,48,84,114,122–129)***.

Multiple levels of computations. PHD processes the input information on multiple levels (*see* the neural network in **Fig. 3**). The first level is a feed-forward neural network with three layers of units (input, hidden, and output). Input to this first-level sequence-to-structure network consists of two contributions: one from the local sequence, i.e., taken from a window of 13 adjacent residues, and another from the global sequence. Output of the first-level network is the 1D structural state of the residue at the center of the input window. The second level is a structure-to-structure network. The next level consists of an arithmetic average over independently trained networks (jury decision). The final level is a simple filter.

Balanced predictions by balanced training. The distribution of the training examples (known structures) is rather uneven: about 32% of the residues are observed in helix, 21% in strand, and 47% in loop. Choosing the training

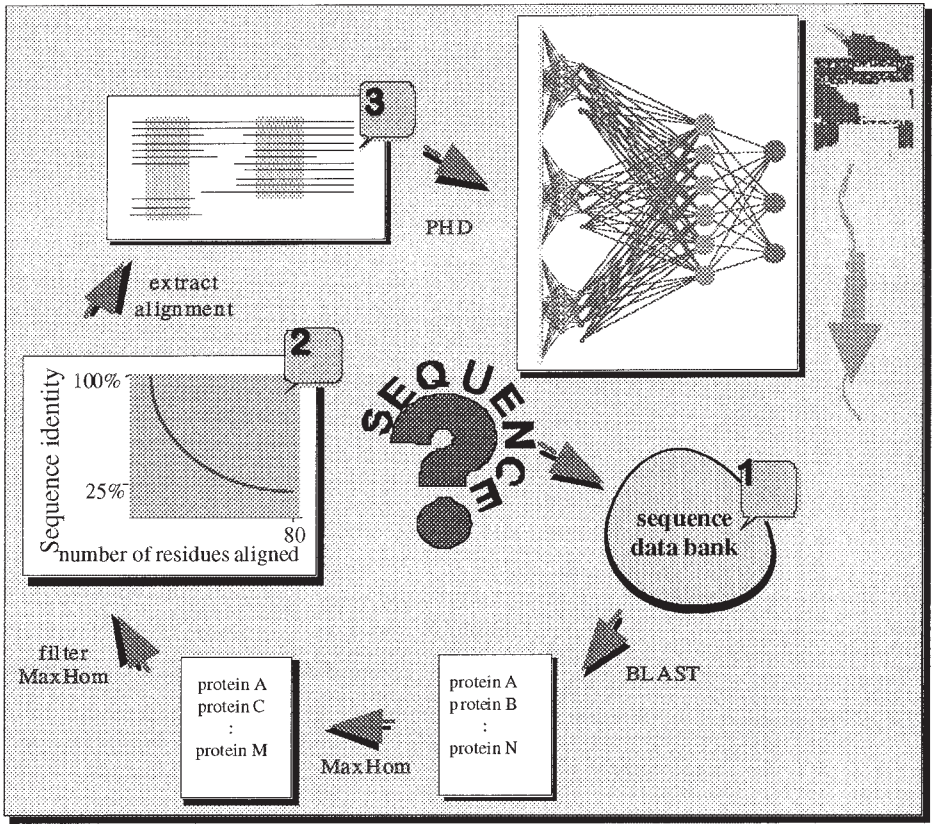

Fig. 3. Using evolutionary information to predict secondary structure. Starting from a sequence of unknown structure (SEQUENCE) the following steps are required to finally feed evolutionary information into the PHD neural networks (upper right): (1) a database search for homologues (method BLAST ***[120]***), (2) a refined profile-based dynamic-programming alignment of the most likely homologues (method *MaxHom* ***[121]***), (3) a decision for which proteins will be considered as homologues (length-dependent cutoff for pairwise sequence identity ***[28,92]***), and (4) a final refinement, and extraction of the resulting multiple alignment. Numbers 1–3 indicate the points where users of the *PredictProtein* service ***(18)*** can interfere to improve prediction accuracy without changes made to the final prediction method PHD.

examples proportional to the occurrence in the data set (unbalanced training) results in a prediction accuracy that mirrors this distribution, e.g., strands are predicted inferior to helix or loop ***(20,21,48)***. A simple way around the database bias is a balanced training: at each time step one example is chosen from each class, i.e., one window with the central residue in a helix, one with the

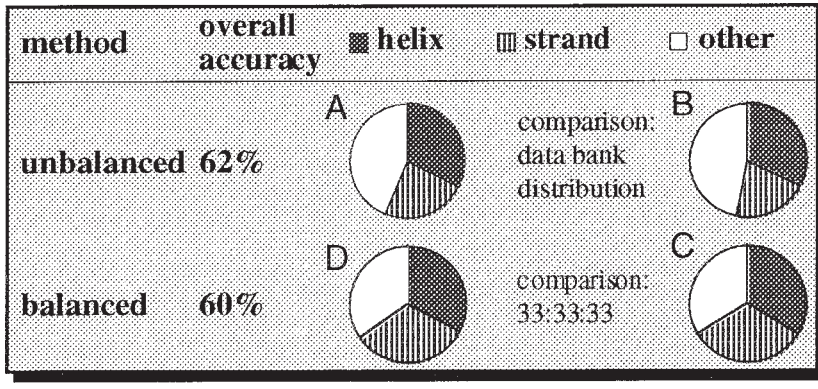

Fig. 4. Prediction balanced between three secondary structure states. The pies were valid for a simple neural network prediction not using evolutionary information (second generation). The entire pies represented 100% of **(A + D)** all correctly predicted residues, **(B)** all residues in a representative subset of PDB, and **(C)** all residues presented during balanced training. The basic message is that the prediction of strand is not inferior to the one for helix for second-generation methods **(A)** because strand formation is more dominated by long-range interactions (as previously argued) but because the database distributions differ between the three states **(B)**. Simply skewing the distribution **(C)** resulted in an equally accurate prediction for all three states **(D)**.

central residue in a strand and one representing the loop class. This training results in a prediction accuracy well balanced between the output states (*see* **Fig. 4**).

Better segment prediction by structure-to-structure networks. The first level sequence-to-structure network uses as input the following information from 13 adjacent residues: (1) the profile of amino acid substitutions for all 13 residues, (2) the conservation weights compiled for each column of the multiple alignment, (3) the number of insertions, and the number of deletions in each column, (4) the position of the current segment of 13 residues with respect to the N- and C-term, (5) the amino acid composition, and (6) the length of the protein. Output consists of three units coding for helix, strand, and nonregular structure. The output coding for the second level network is identical to the one for the first. The dominant input contribution to the second level structure-to-structure network is the output of the first-level sequence-to-structure network. The reason for introducing a second level is the following. Networks are trained by changing the connections between the units such that the error is reduced for each of the examples successively presented to the network during training. The examples are chosen at random. Therefore, the examples taken at time step t and at time step $t + 1$ are usually not adjacent in sequence. This implies

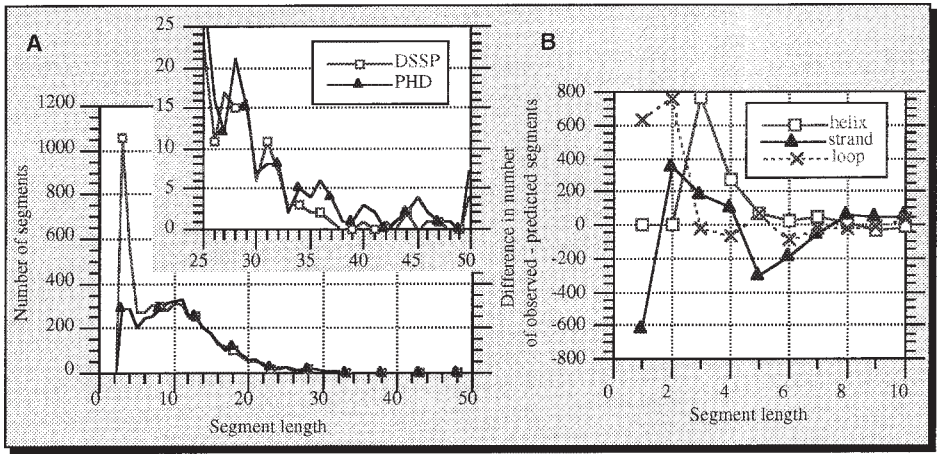

Fig. 5. Distribution of segment length **(A)** The number of helical segments observed (open squares; according to DSSP ***[19]***) and predicted (filled triangles; by PHD ***[18]***) is plotted against their length. Obviously, most short helices are missed by the prediction. The inlet zooms on longer helices, revealing that PHD predicts slightly too long helices. Figures for strands and nonregular structures are not given, as the observed and predicted distributions agree relatively well, for longer segments at least. However, there are important differences for shorter segments: **(B)** plots the differences between the numbers of observed–predicted segments at given lengths (helices: open squares, strands: filled triangles, nonregular structure: dashed line with crosses). In particular, strands of a single residue are overpredicted; short loop regions and three helices ***(10)*** (three residues) are underpredicted.

that the network cannot learn, e.g., that helices contain at least three residues. The second-level structure-to-structure network introduces a correlation between adjacent residues with the effect that predicted secondary-structure segments have length distributions similar to the ones observed ***(27)***. Problems arise, in particular, for short segments (*see* **Fig. 5**).

3.3. State-of-the-Art Secondary Structure Prediction

3.3.1. Estimates of Prediction Accuracy

Difference between 60% and 70% accuracy may matter a lot. Some of the third-generation methods for secondary-structure prediction are clearly superior to previous methods: β-strands are predicted more accurately; predicted segments look like those observed; and the overall accuracy is about 10 percentage points higher. The advantage in practice is illustrated in **Fig. 6**. Not only does the third-generation method (here PHD) gets most segments right,

Fig. 6. Example for secondary structure prediction of first–third generation. The protein sequence (*SEQ*) given was the SH3 structure ***(131)***. The observed secondary structure (*OBS*) was assigned by DSSP ***(19)*** (H = helix; E = strand; blank = nonregular structure; the dashes indicated the continuation of the second strand that was missed by DSSP). The methods are first generation: *C + F* ***(42)***; second generation: *GOR* ***(17)*** (= GORIII), and third generation: PHD ***(18)***. The levels of three-state accuracy were: C + F = 59%; GOR = 65%; and PHD = 72%. Whereas the first- and second-generation methods performed above their average accuracy (**Fig. 1**) for this protein, the PHD prediction was average (*see* **Figs. 1** and **7**). The strength of the PHD prediction was reflected in the one-digit reliability index (*Rel*, 0 = low, 9 = high), correlated with prediction accuracy. All residues predicted at values of Rel > 4 (marked by *) were predicted correctly.

but it also enables one to focus on more reliably predicted residues. The reliability index (*Rel* in **Fig. 6**) is compiled as the difference between the output unit with highest value (winner unit) and the output unit with the next highest value (normalized to a scale from 0 [low] to 9 [high]). All strongly predicted residues (* in **Fig. 6**) are predicted correctly.

Values for expected prediction accuracy are distributions. Statements such as "secondary-structure is about 90% conserved within sequence families" ***(22)*** refer to averages over distributions. The same holds for the expected prediction accuracy (*see* **Fig. 7**). Such distributions explain why some developers have overestimated the performance of their tools using data sets of only tens of proteins (or even fewer). In general, single sequences yield accuracy values about 10 percentage points lower than multiple alignments ***(21,25,48)***. Note that for most proteins some helix and strand residues are confused (refer to **Fig. 7**).

Reliability of prediction correlates with accuracy. For the user interested in a particular protein *U*, the fact that prediction accuracy varies with the protein (*see* **Fig. 7**) implies a rather unfortunate message: the accuracy for *U* could be lower than 40%, or it could be higher than 90% (*see* **Fig. 7**). Is there any way to provide an estimate at which end of the distribution the accuracy for *U* is likely to be? Indeed, the reliability index correlates with accuracy. In other words, residues with a higher reliability index are predicted with higher accuracy

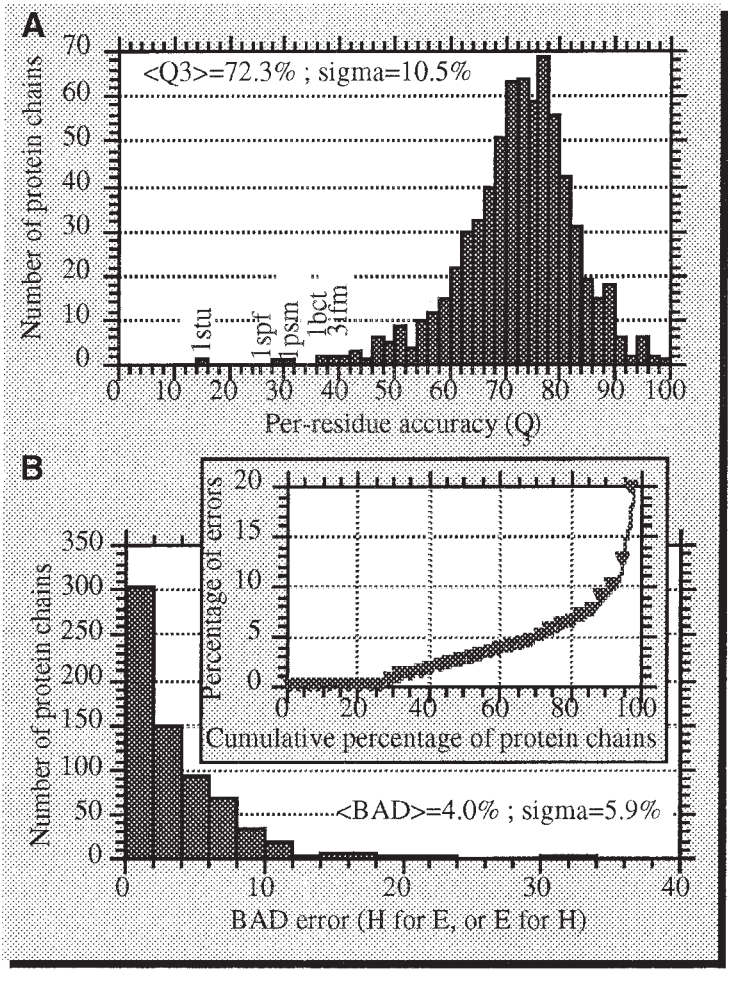

Fig. 7. Expected variation of prediction accuracy with protein chain. **(A)** Three-state per-residue accuracy (*see* **Eq. 1**; PDB identifier given for the proteins predicted worst); **(B)** percentage of BAD predictions, i.e., residues either predicted in helix and observed in strand, or predicted in strand and observed in helix (introduced by **ref. *14***); (B inset) cumulative percentage of proteins with BADly predicted residues (e.g., for 80% of the proteins the percentage of confusing helix and strand residues is <7%; however, for only for 30% of all proteins such a confusion never happened). Given: distributions (over 721 unique protein chains), averages, and one standard deviation.

(18,21,48). Thus, the reliability index offers an excellent tool to focus on some key regions predicted at high levels of expected accuracy. Furthermore, the reliability index averaged over an entire protein correlates with the overall pre-

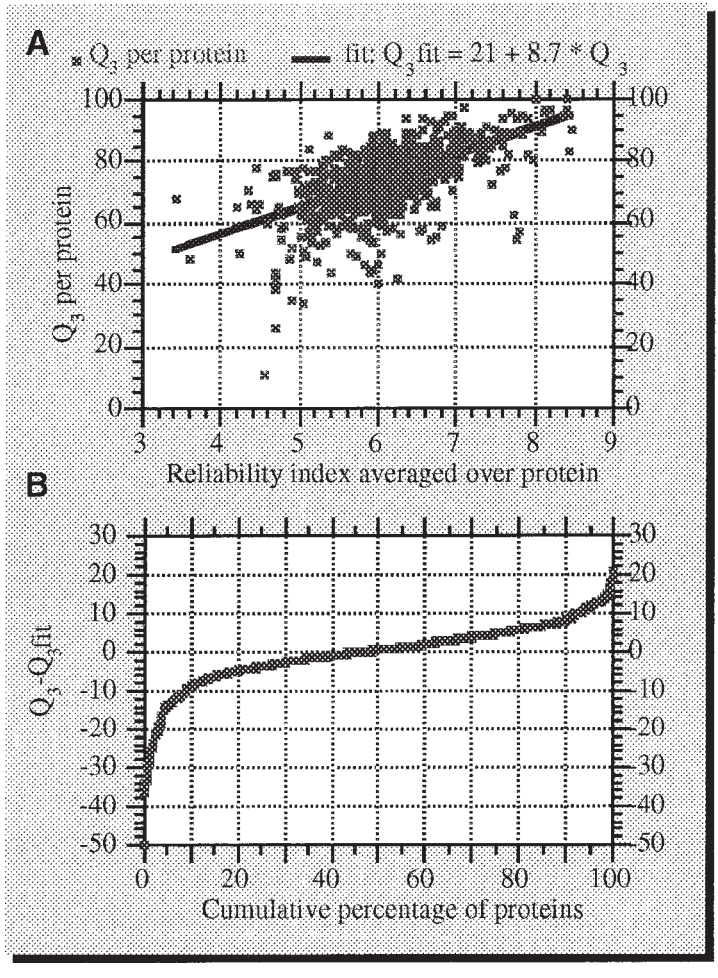

Fig. 8. Correlation between reliability and accuracy. Residues predicted at higher reliability are predicted more accurately ***(18,21,48)***. Here, we plotted the reliability index averaged over a protein with the overall accuracy for that protein **(A)**. Even a simple linear fit **(A)** provided a reasonably accurate estimate of the performance: for more than 80% of all proteins the linear fit yielded estimates in the range of less than ±10% accuracy **(B)**.

diction accuracy for this protein (*see* **Fig. 8**) (Note however, that reliability indices tend to be unusually high for alignments of sequence families without very divergent sequences.)

Do we understand why certain proteins are predicted poorly? For some of the worst predicted proteins, the low level of accuracy could be anticipated from

their unusual features, e.g., for crambin, or the antifreeze glycoprotein type III. However, this procedure turned out to be rather arbitrary. First, some proteins with the same "unusual features" are predicted at high levels of accuracy. Second, occasionally similar proteins are predicted at very different levels of accuracy, e.g., both the phosphotidylinonitol 3-kinase ***(130)*** and the Src-homology domain of cytoskeletal spectrin have homologous structures ***(131)***, but prediction accuracy varies between less than 40% (pik) and more than 70% (spectrin). None of the conclusions from studying poor predictions has yet yielded a way to better predictions. Nevertheless, two observations may be added. First, bad alignments (i.e., noninformative and/or falsely aligned residues) result in bad predictions. Second, the BAD predictions (*see* **Fig. 7B**), i.e., the confusion of helix and strand, are frequently observed in regions that are stabilized by long-range interactions. For example, the peptide around the fourth strand of SH3 (*see* **Fig. 6**) forms a helix in solution (L. Serrano, personal communication). Furthermore, helices and strands that are confused despite a high reliability index often have functional properties, or are correlated to disease states (B. Rost, unpublished data).

3.3.2. Availability of Methods

Internet prediction services for secondary-structure, in general. Programs for the prediction of secondary-structure available as Internet services have mushroomed since the first prediction service PredictProtein went online in 1992 ***(119,132)*** (a list of links is found in **ref. *133***). Unfortunately, not all services are sufficiently tested. In general, prediction accuracy is significantly superior if predictions are based on multiple alignments ***(4,13,16)***.

Completely vs. almost automatic. The PHD prediction method is automatically available via the Internet service PredictProtein ***(18)*** (send the word *help* to `PredictProtein@Columbia.edu`, or use the World Wide Web interface ***[132]***). Users have the choice between the fully automatic procedure taking the query sequence through the entire cycle, or expert intervention into the generation of the alignment. Indeed, without spending much time the result was that predictions could be easily improved ***(134)***.

4. Notes

The following notes result from the experiences one of us (BR) has gathered by offering, and running the PredictProtein ***(132)*** service and during various structure prediction workshops ***(135)***. Some comments apply in particular to the PHD methods ***(18,136)***; however, most hold also for using other secondary-structure prediction methods (we strongly recommend reading the detailed "hints" on the PredictProtein WWW page: ***[132]***).

4.1. What Can You Expect From Secondary Structure Prediction?

How accurate are the predictions? The expected levels of accuracy ($Q_3 = 72 \pm 11\%$) are valid for typical globular, water-soluble proteins when the multiple alignment contains many and diverse sequences. High values for the reliability indices indicate more accurate predictions (Note: for alignments with little variation in the sequences, the reliability indices adopt misleadingly high values.) PHD predictions tend to be relatively accurate for porins ***(18)***; however, for helical membrane proteins, other programs ought to be used ***(5,18,136)***.

How useful are the predictions? The prediction of secondary-structures can be accurate enough to assist chain tracing. Furthermore, PHD predictions are being used as a starting point for modeling 3D structure and predicting function ***(115,116,122,137–143)***.

Is there confusion between strand and helix? PHD (as well as other methods) focuses on predicting hydrogen bonds. Consequently, occasionally strongly predicted (high reliability index) helices are observed as strands and vice versa (*see* **Fig. 7B**).

Is there a strong signal from secondary-structure caps? The ends of helices and strands contain a strong signal. However, on average PHD predicts the core of helices and strands more accurately than do the caps ***(20)***. This seems to also hold for other methods.

Are internal helices poorly predicted? Steven Benner has indicated that internal helices are difficult to predict ***(24,107)***. On average, this is not the case for PHD predictions ***(144)***.

What about protein design and synthesized peptides? The PHD networks are trained on naturally evolved proteins. However, the predictions have been useful in some cases to investigate the influence of single mutations (e.g., for Chameleon ***[145,146]***, or for Janus ***[147]***; B. Rost, unpublished). For short polypeptides, users should bear in mind that the network input consists of 17 adjacent residues. Thus, shorter sequences may be dominated by the ends (which are treated as solvents by the current version of PHD).

4.2. How Can You Avoid Pitfalls?

70% correct implies 30% incorrect. The most accurate methods for predicting secondary-structure reach sustained levels of about 70% accuracy. When interpreting predictions for a particular protein, it is often instructive to mark the 30% of the residues you suspect to be falsely predicted.

Spread of prediction accuracy. An expected accuracy of 70% does *not* imply that for your protein *U* 70% of all residues are correctly predicted. Instead, values published for prediction accuracy are averaged over hundreds of unique proteins. An expected accuracy of $70 \pm 10\%$ (one standard deviation) implies

that, on average, for two-thirds of all proteins between 60 and 80% of the residues will be predicted correctly (*see* **Fig. 7**). Thus, prediction accuracy can be higher than 80% or lower than 60% for your protein. Few methods supply well-tested indices for the reliability of predictions (*see* **Fig. 8**; ***[18,134]***). Such indices can help to reduce or increase your trust in a particular prediction.

Special classes of proteins. Prediction methods are usually derived from knowledge contained in proteins from subsets of current databases. Consequently, they should not be applied to classes of proteins not included in these subsets, e.g., methods for predicting helices in globular proteins are likely to fail when applied to predict transmembrane helices. In general, results should be taken with caution for proteins with unusual features, such as proline-rich regions, unusually many cysteine bonds, or for domain interfaces.

Better alignments yield better predictions. Multiple-alignment-based predictions are substantially more accurate than single-sequence-based predictions. How many sequences do you need in your alignment for an improvement? How sensitive are prediction methods to errors in the alignment? The more divergent sequences contained in the alignment, the better (two distantly related sequences often improve secondary-structure predictions by several percentage points). Regions with few aligned sequences yield less reliable predictions. The sensitivity to alignment errors depends on the methods, e.g., secondary-structure prediction is less sensitive to alignment errors than accessibility prediction.

Better + worse = even better? Today, several automatic services accomplish secondary-structure predictions. Some users fall into the what-is-common-is-correct trap, i.e., they average over all prediction methods and consider identical regions as more reliable. Such a majority vote may be beneficial. However, the result will frequently be the worst-of-all prediction. Often, it is preferable to use reliability indices provided by some methods. Such indices answer the question: how reliably is the tryptophan at position 307 predicted in a surface loop? (Note: the correlation between such indices and prediction accuracy is sufficiently tested for only a few methods.)

1D structure may or may not be sufficient to infer 3D structure. Say you the following as a prediction for a regular secondary-structure: helix-strand-strand-helix-strand-strand (H-E-E-H-E-E). Assume that you find a protein of known structure with the same motif (H-E-E-H-E-E). Can you conclude that the two proteins have the same fold? Yes and no; your guess may be correct, but there are various ways to realize the given motif by completely different structures. For example, at least 16 structurally unrelated proteins contain the secondary-structure motif H-E-E-H-E-E.

Addendum

At the third meeting for the Critical Assessment of Structure Prediction (CASP) in December 1998, David Jones presented a method that extended the basic idea of 3rd generation prediction methods, i.e., using evolutionary information, by replacing previously used sequence alignment procedures with an iterated PSI-BLAST profile ***(149)***. The resulting method PSI-PRED appears to be more than 2–3 percentage points more accurate than any other method published so far ***(150)***. About one percentage point of this improvement can be achieved by simply replacing the alignment profiles (Rost, unpublished).

However, the major step appears to be attributed to the fact that the databases have grown, and developing prediction methods can now be based on data sets more than 10 times larger than those used to develop the first 3rd generation tools (Rost, unpublished). The work of David Jones has reactivated the field, at least one other novel method (JNET: Cuff & Barton, unpublished) appears clearly moved accurate than the original PHD1 referred to in our review.

References

1. Bairoch, A. and Apweiler, R. (1997) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.* **25,** 31–36.
2. Gaasterland, T. (1997) Genome sequencing projects. WWW document (http://genomes.rockefeller.edu):Rockefeller University.
3. Bernstein, F. C., et al. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542.
4. Rost, B. and Sander, C. (1996) Bridging the protein sequence–structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25,** 113–136.
5. Rost, B. and O'Donoghue, S. I. (1997) Sisyphus and prediction of protein structure. *CABIOS* **13,** 345–356.
6. CASP1. (1995) Special issue of *Proteins* **23,** 295–462.
7. CASP2. (1997) Special issue of *Proteins* **Suppl 1,** 1–230.
8. Rost, B. and Sander, C. (1994) Structure prediction of proteins — where are we now? *Curr. Opin. Biotech.* **5,** 372–380.
9. Rost, B. (1998) Protein structure prediction in 1D, 2D, and 3D, in *Encyclopedia of Computational Chemistry* (von Ragué Schleyer, P., et al., eds.), Wiley, Chichester, UK, pp. 2242–2255.
10. Fasman, G. D. (1989) *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum, New York, London.
11. Sternberg, M. J. E. (1992) Secondary structure prediction. *Curr. Opin. Struct. Biol.* **2,** 237–241.
12. Presnell, S. R. and Cohen, F. E. (1993) Artificial neural networks for pattern recognition in biochemical sequences. *Annu. Rev. Biophys. Biomol. Struct.* **22,** 283–298.

13. Barton, G. J. (1995) Protein secondary structure prediction. *Curr. Opin. Struct. Biol.* **5,** 372–376.
14. Defay, T. and Cohen, F. E. (1995) Evaluation of current techniques for *ab initio* protein structure prediction. *Proteins* **23,** 431–445.
15. Russell, R. B. and Sternberg, M. J. E. (1995) How good are we? *Curr. Biol.* **5,** 488–490.
16. Di Francesco, V., Garnier, J., and Munson, P. J. (1996) Improving protein secondary structure prediction with aligned homologous sequences. *Protein Sci.* **5,** 106–113.
17. Garnier, J., Gibrat, J.-F., and Robson, B. (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266,** 540–553.
18. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266,** 525–539.
19. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22,** 2577–2637.
20. Rost, B. and Sander, C. (1994) 1D secondary structure prediction through evolutionary profiles, in *Protein Structure by Distance Analysis* (Bohr, H., and Brunak, S., eds.), IOS Press, Amsterdam, Oxford, Washington, pp. 257–276.
21. Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–599.
22. Rost, B., Sander, C., and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235,** 13–26.
23. Thornton, J. M., et al. (1992) Prediction of progress at last. *Nature* **354,** 105–106.
24. Benner, S. A. and Gerloff, D. L. (1993) Predicting the conformation of proteins: man versus machine. *FEBS Lett.* **325,** 29–33.
25. Rost, B., Sander, C., and Schneider, R. (1993) Progress in protein structure prediction? *TIBS* **18,** 120–123.
26. Russell, R. B. and Barton, G. J. (1993) The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* **234,** 951–957.
27. Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA* **90,** 7558–7562.
28. Sander, C. and Schneider, R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9,** 56–68.
29. Kendrew, J. C., et al. (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2Å resolution. *Nature* **185,** 422–427.
30. Perutz, M. F., et al. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5. 5Å resolution, obtained by X-ray analysis. *Nature* **185,** 416–422.
31. Pauling, L. and Corey, R. B. (1951) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl. Acad. Sci. USA* **37,** 729–740.

32. Pauling, L., Corey, R. B., and Branson, H. R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **37,** 205–234.
33. Szent-Györgyi, A. G. and Cohen, C. (1957) Role of proline in polypeptide chain configuration of proteins. *Science* **126,** 697.
34. Blout, E. R., et al. (1960) Dependence of the conformation of synthetic polypeptides on amino acid composition. *J. Am. Chem. Soc.* **82,** 3787–3789.
35. Blout, E. R. (1962) The dependence of the conformation of polypetides and proteins upon amino acid composition, in *Polyamino Acids, Polypeptides, and Proteins* (Stahman, M., ed.), University of Wisconsin Press, Madison, WI, pp. 275–279.
36. Scheraga, H. A. (1960) Structural studies of ribonuclease III. A model for the secondary and tertiary structure. *J. Am. Chem. Soc.* **82,** 3847–3852.
37. Davies, D. R. (1964) A correlation between amino acid composition and protein structure. *J. Mol. Biol.* **9,** 605–609.
38. Schiffer, M. and Edmundson, A. B. (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* **7,** 121.
39. Pain, R. H. and Robson, B. (1970) Analysis of the code relating sequence to secondary structure in proteins. *Nature* **227,** 62–63.
40. Finkelstein, A. V. and Ptitsyn, O. B. (1971) Statistical analysis of the correlation among amino acid residues in helical, β-structural and non-regular regions of globular proteins. *J. Mol. Biol.* **62,** 613–624.
41. Robson, B. and Pain, R. H. (1971) Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.* **58,** 237–259.
42. Chou, P. Y. and Fasman, U. D. (1974) Prediction of protein conformation. *Biochemistry* **13,** 211–215.
43. Lim, V. I. (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88,** 857–872.
44. Rose, G. D. (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* **272,** 586–590.
45. Kabsch, W. and Sander, C. (1983) How good are predictions of protein secondary structure? *FEBS Lett.* **155,** 179–182.
46. Rost, B. (2000) Neural networks for protein structure prediction: hype or hit? Preprint, (http://cubic.bioc.columbia.edu/papers/Pre 1999_tics/) Columbia, University, New York.
47. Rost, B. and Sander, C. (1993) Secondary structure prediction of all-helical proteins in two states. *Protein Eng.* **6,** 831–836.
48. Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19,** 55–72.
49. Kabat, E. A. and Wu, T. T. (1973) The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of β-sheets in concanavalin A. *Proc. Natl. Acad. Sci. USA* **70,** 1473–1477.

50. Maxfield, F. R. and Scheraga, H. A. (1976) Status of empirical methods for the prediction of protein backbone topography. *Biochemistry* **15,** 5138–5153.
51. Robson, B. (1976) Conformational properties of amino acid residues in globular proteins. *J. Mol. Biol.* **107,** 327–356.
52. Nagano, K. (1977) Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol.* **109,** 251–274.
53. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and Implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97–120.
54. Gibrat, J.-F., Garnier, J., and Robson, B. (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198,** 425–443.
55. Biou, V., et al. (1988) Secondary structure prediction: combination of three different methods. *Protein Eng.* **2,** 185–91.
56. Gascuel, O. and Golmard, J. L. (1988) A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *CABIOS* **4,** 357–365.
57. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164.
58. Viswanadhan, V. N., Denckla, B., and Weinstein, J. N. (1991) New joint prediction algorithm (Q7-JASEP) improves the prediction of protein secondary structure. *Biochemistry* **30,** 11,164–11,172.
59. Juretic, D., et al. (1993) Conformational preference functions for predicting helices in membrane proteins. *Biopolymers* **33,** 255–273.
60. Mamitsuka, H. and Yamanishi, K. (1993) Protein α-helix region prediction based on stochastic-rule learning, in *26th Annual Hawaii International Conference on System Sciences* (eds.), IEEE Computer Society, Maui, HI, pp. 659–668.
61. Donnelly, D., Overington, J. P., and Blundell, T. L. (1994) The prediction and orientation of α-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules. *Protein Eng.* **7,** 645–653.
62. Ptitsyn, O. B. and Finkelstein, A. V. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* **22,** 15–25.
63. Taylor, W. R. and Thornton, J. M. (1983) Prediction of super-secondary structure in proteins. *Nature* **301,** 540–542.
64. Cohen, F. E. and Kuntz, I. D. (1989) Tertiary structure prediction, in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., eds.), Plenum, New York, London, pp. 647–706.
65. Rooman, M. J., Kocher, J. P., and Wodak, S. J. (1991) Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *J. Mol. Biol.* **221,** 961–979.
66. Qian, N. and Sejnowski, T. J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202,** 865–884.
67. Bohr, H., et al. (1988) Protein secondary structure and homology by neural networks. *FEBS Lett.* **241,** 223–228.

68. Holley, H. L. and Karplus, M. (1989) Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
69. Kneller, D. G., Cohen, F. E., and Langridge, R. (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214,** 171–182.
70. Stolorz, P., Lapedes, A., and Xia, Y. (1992) Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* **225,** 363–377.
71. Zhang, X., Mesirov, J. P., and Waltz, D. L. (1992) Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225,** 1049–1063.
72. Maclin, R. and Shavlik, J. W. (1993) Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning* **11,** 195–215.
73. Chandonia, J.-M. and Karplus, M. (1995) Neural networks for secondary structure and structural class predictions. *Protein Sci.* **4,** 275–285.
74. Mitchell, E. M., et al. (1992) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212,** 151–166.
75. Geourjon, C. and Deléage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* **11,** 681–684.
76. Kanehisa, M. (1988) A multivariate analysis method for discriminating protein secondary structural segments. *Protein Eng.* **2,** 87–92.
77. Munson, P. J. and Singh, R. K. (1997) Multi-body interactions within the graph of protein structure, in *Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., et al., eds.), AAAI Press, Halkidiki, Greece, pp. 198–201.
78. King, R. D., et al. (1992) Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA* **89,** 11,322–11,326.
79. Muggleton, S., King, R. D., and Sternberg, M. J. E. (1992) Protein secondary structure prediction using logic-based machine learning. *Protein Eng.* **5,** 647–657.
80. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* **23,** 566–579.
81. Zhu, Z.-Y. and Blundell, T. L. (1996) The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J. Mol. Biol.* **260,** 261–276.
82. Asogawa, M. (1997) Beta-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network, in *Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., et al., eds.), AAAI Press, Halkidiki, Greece, pp. 48–51.
83. Yi, T.-M. and Lander, E. S. (1993) Protein secondary structure prediction using nearest-neigbor methods. *J. Mol. Biol.* **232,** 1117–1129.
84. Solovyev, V. V. and Salamov, A. A. (1994) Predicting α-helix and β-strand segments of globular proteins. *CABIOS* **10,** 661–669.

85. Salamov, A. A. and Solovyev, V. V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *J. Mol. Biol.* **247,** 11–15.
86. Kabsch, W. and Sander, C. (1983) Segment83. unpublished.
87. Schneider, R. (1989) Sekundärstrukturvorhersage von Proteinen unter Berücksichtigung von Tertiärstrukturaspekten. Diploma thesis: Department of Biology, University of Heidelberg, Heidelberg, Germany.
88. Devereux, J., Haeberli, P., and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12,** 387–395.
89. Rost, B. and Schneider, R. (1998) Pedestrian guide to analysing sequence databases, in *Core Techniques in Biochemistry* (Ashman, K., ed.) http://cubic.bioc.columbia.edu/papers/1999_pedestrian/, Springer, Heidelberg, pp. in press.
90. Dao-pin, S., et al. (1991) Contributions of surface salt bridges to the stability of bacteriophage T4 lysozyme determined by directed mutagenesis. *Biochemistry* **30,** 7142–7153.
91. Doolittle, R. F. (1986) *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley CA.
92. Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826.
93. Lesk, A. M. (1991) *Protein Architecture — A Practical Approach*. Oxford University Press, Oxford, New York, Tokyo.
94. Rost, B. (1998) Twilight zone of protein sequence alignments. *Prof. Enginering* **12,** S85–S94.
95. Rost, B. (1997) Protein structures sustain evolutionary drift. *Fold. Des.* **2,** S19–S24.
96. Rost, B., O'Donoghue, S., and Sander, C. (1998) Midnight zone of protein structure evolution. Preprint (http://cubic.bioc.columbia.edu/papers/Pre 1998_midnight/) Columbia University, New York.
97. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257,** 342–358.
98. Pazos, F., et al. (1997) Comparative analysis of different methods for the detection of specificity regions in protein families, in *BCEC97: Bio-Computing and Emergent Computation* (Olsson, B., Lundh, D., and Narayanan, A., eds.), World Scientific, Skövde, Sweden, pp. 132–145.
99. Dickerson, R. E., Timkovich, R., and Almassy, R. J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism. *J. Mol. Biol.* **100,** 473–491.
100. Dickerson, R. E. (1971) The structure of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* **1,** 26–45.
101. Frampton, J., et al. (1989) DNA-binding domain ancestry. *Nature* **342,** 134.
102. Benner, S. A. (1989) Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enzyme Regul.* **28,** 219–236.
103. Bazan, J. F. (1990) Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci. USA* **87,** 6934–6938.
104. Benner, S. A. and Gerloff, D. (1990) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.* **31,** 121–181.

105. Niermann, T. and Kirschner, K. (1990) Improving the prediction of secondary structure of "TIM-barrel" enzymes. *Protein Eng.* **4,** 137–147.
106. Barton, G. J., et al. (1991) Amino acid sequence analysis of the annexin supergene family of proteins. *Eur. J. Biochem.* **198,** 749–760.
107. Benner, S. A. (1992) Predicting de novo the folded structure of proteins. *Curr. Opin. Struct. Biol.* **2,** 402–412.
108. Gibson, T. J. (1992) Assignment of α-helices in multiply aligned protein sequences — applications to DNA binding motifs, in *Patterns in Protein Sequence and Structure* (Taylor, W. R., ed.), Berlin-Heidelberg, Springer-Verlag, pp. 99–110.
109. Musacchio, A., et al. (1992) SH3 — an abundant protein domain in search of a function. *FEBS Lett.* **307,** 55–61.
110. Barton, G. J. and Russell, R. B. (1993) Protein structure prediction. *Nature* **361,** 505–506.
111. Boscott, P. E., Barton, G. J., and Richards, W. G. (1993) Secondary structure prediction for homology modelling. *Protein Eng.* **6,** 261–266.
112. Gerloff, D. L., et al. (1993) The nitrogenase MoFe protein. *FEBS Lett.* **318,** 118–124.
113. Gibson, T. J., Thompson, J. D., and Abagyan, R. A. (1993) Proposed structure for the DNA-binding domain of the Helix-Loop-Helix family of eukaryotic gene regulatory proteins. *Protein Eng.* **6,** 41–50.
114. Livingstone, C. D. and Barton, G. J. (1994) Secondary structure prediction from multiple sequence data: blood clotting factor XIII and versinia protein-tyrosine phosphatase. *Int. J. Peptide Protein Res.* **44,** 239–244.
115. Hansen, J. E., et al. (1996) Prediction of the secondary structure of HIV-1 gp120. *Proteins* **25,** 1–11.
116. Valencia, A., et al. (1995) Prediction of the structure of GroES and its interaction with GroEL. *Proteins* **22,** 199–209.
117. Maxfield, F. R. and Scheraga, H. A. (1979) Improvements in the prediction of protein topography by reduction of statistical errors. *Biochemistry* **18,** 697–704.
118. Zvelebil, M. J., et al. (1987) Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* **195,** 957–961.
119. Rost, B., Sander, C., and Schneider, R. (1994) PHD — an automatic server for protein secondary structure prediction. *CABIOS* **10,** 53–60.
120. Altschul, S. F. and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.* **266,** 460–480.
121. Schneider, R. (1994) Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen. PhD thesis: University of Heidelberg, Heidelberg, Germany.
122. Hubbard, T. J. P. and Park, J. (1995) Fold recognition and *ab initio* structure predictions using Hidden Markov models and β-strand pair potentials. *Proteins* **23,** 398–402.
123. Mehta, P. K., Heringa, J., and Argos, P. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.* **4,** 2517–2525.

124. Riis, S. K. and Krogh, A. (1996) Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol.* **3,** 163–183.
125. Gerloff, D. L. and Cohen, F. E. (1996) Secondary structure prediction and unrefined tertiary structure prediction for cyclin A, B, and D. *Proteins* **24,** 18–34.
126. Frishman, D. and Argos, P. (1997) 75% accuracy in protein secondary structure prediction. *Proteins* **27,** 329–335.
127. Salamov, A. A. and Solovyev, V. V. (1997) Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **268,** 31–36.
128. Levin, J. M., et al. (1993) Quantification of secondary structure prediction improvement using multiple alignment. *Protein Eng.* **6,** 849–854.
129. King, R. D. and Sternberg, M. J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5,** 2298–2310.
130. Koyama, S., et al. (1993) Structure of the PI3K SH3 domain and analysis of the SH3 family. *Cell* **72,** 945–952.
131. Musacchio, A., et al. (1992) Crystal structure of a Src-homology 3 (SH3) domain. *Nature* **359,** 851–855.
132. Rost, B. (1997) PredictProtein — internet prediction service. WWW document (http://cubic.bioc.columbia.edu/predictprotein):Columbia University, New York.
133. Rost, B. and Schneider, R. (1996) WWW services for sequence analysis. WWW document (http://cubic.bioc.columbia.edu/doc/links_index.html):Columbia University, New York.
134. Rost, B. (1998) Better 1D predictions by experts with machines. *Proteins*, in press.
135. Rost, B. and Valencia, A. (1996) Pitfalls of protein sequence analysis. *Curr. Opin. Biotechnol.* **7,** 457–461.
136. Rost, B., Casadio, R., and Fariselli, P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5,** 1704–1718.
137. Meitinger, T., et al. (1993) Molecular modelling of the Norrie disease protein predicts a cysteine knot growth factor tertiary structure. *Nature* Gen., **5,** 376–380.
138. Rawlings, D. J., et al. (1993) Mutation of unique region of Bruton's tyrosine kinase in immunodeficient XID M mice. *Science* **261,** 358–361.
139. Lupas, A., et al. (1994) Predicted secondary structure of the 20S proteasome and model structure of the putative peptide channel. *FEBS Lett.* **354,** 45–49.
140. Viguera, E., et al. (1994) Mammalian L-amino acid decarboxylases producing 1,4-diamines: analogies among differences. *TIBS* **19,** 318–319.
141. Fischer, D. and Eisenberg, D. (1996) Fold recognition using sequence-derived properties. *Protein Sci.* **5,** 947–955.
142. Rost, B., Schneider, R., and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270,** 471–480.
143. Springer, T. A. (1997) Folding of the N-terminal, ligand-binding region of integrin α-subunits into a β-propeller domain. *Proc. Natl. Acad. Sci. USA* **94,** 65–72.
144. Rost, B. (1996) Accuracy of predicting buried helices by PHDsec. WWW document (http://cubic.bioc.columbia.edu/results/1996/PredBuriedHelices.html). Columbia University, New York.

145. Minor, D. L. J. and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380,** 730–734.
146. Rost, B. (1996) 1D structure prediction for Chameleon (IgG binding domain of protein G) WWW document (http://cubic.bioc.columbia.edu/results/1996/PredCameleon.html):Columbia University, New York.
147. Dalal, S., Balasubramanian, S., and Regan, L. (1997) Protein alchemy: changing β-sheet into α-helix. *Nat. Struct. Biol.* **4,** 548–552.
148. Chou, P. Y. and Fasman, G. D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* **47,** 45–148.
149. PSI-BLAST: Altschul, S., et al. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389–3402.
150. PSI-PRED: Jones, D. T. (1999) Protein secondary structure prediction based on position specific scoring matrices. *J. Mol. Biol.* **292,** 195–202.