

EVA: Large-Scale Analysis of Secondary Structure Prediction

Burkhard Rost^{1*} and Volker A. Eyrich^{1,2}

¹CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

²Department of Chemistry, Columbia University, New York, New York

ABSTRACT EVA is a web-based server that evaluates automatic structure prediction servers continuously and objectively. Since June 2000, EVA collected more than 20,000 secondary structure predictions. The EVA sets sufficed to conclude that the field of secondary structure prediction has advanced again. Accuracy increased substantially in the 1990s through using evolutionary information taken from the divergence of proteins in the same structural family. Recently, the evolutionary information resulting from improved searches and larger databases has again boosted prediction accuracy by more than 4% to its current height around 76% of all residues predicted correctly in one of the three states: helix, strand, or other. The best current methods solved most of the problems raised at earlier CASP meetings: All good methods now get segments right and perform well on strands. Is the recent increase in accuracy significant enough to make predictions even more useful? We believe the answer is affirmative. What is the limit of prediction accuracy? We shall see. All data are available through the EVA web site at {cubic.bioc.columbia.edu/eva/}. The raw data for the results presented are available at {eva/seq/bup_common/2001_02_22/}. Proteins 2001;Suppl 5:192–199. © 2002 Wiley-Liss, Inc.

Key words: automatic evaluation; large-scale assessment; protein structure prediction

INTRODUCTION

Secondary structure is at the heart of structure prediction. The rapidly growing sequence-structure gap (number of known protein structures vs. number of known protein sequences) has enticed theoreticians to solve simplified prediction problems.^{10,11} An extreme simplification is the prediction of protein structure in one dimension (1D), as represented by strings of secondary structure. Theoreticians are lucky in that this relatively simple task comprises a goal relevant for prediction of protein structure and function, in general. Almost any imaginable algorithm has been applied to this task. The result is that we have come a long way since the first method published 44 years ago.¹² The most important aspect of third-generation methods demonstrating their breakthrough at the first CASP meetings was the automatic use of evolutionary information.^{13,14} In fact, secondary structure prediction may have been the most successful discipline of protein

structure prediction over the last 40 years. Is the field still alive?

Here, we focus on presenting various aspects of the performance of recent secondary structure prediction methods. We analyze automatic methods based on large data sets. The machinery allowing such a large-scale assessment is the automatic, continuous, and objective web server EVA.^{15,16}

MATERIALS AND METHODS

Current Implementation of EVA

EVA assessment in four prediction categories. Currently, EVA¹⁵ evaluates four different categories of structure prediction servers (URLs at¹⁶): (a) comparative modeling, (b) fold recognition and threading, (c) secondary structure prediction, and (d) inter-residue contact predictions. The following groups agreed to let their public secondary structure prediction servers be evaluated by EVA: James Cuff & Geoff Barton (JPred2)^{2,17}; Mohammed Ouali & Ross King (PROFking)⁶; David Jones (PSIPRED)^{7,18}; Gajendra Raghava (PSSP, unpublished), Kevin Karplus (SAM-T99sec)⁸, Pierre Baldi & Gianluca Pollastri (SSpro).⁹ Our group contributed PHDsec,^{3,19,20} PHDpsi,⁴ and PROFphd (Rost, unpublished).

Results are updated every week. Everyday, EVA obtains the latest experimentally determined structures from the PDB²¹ web site. These structures are parsed into chains by using the DSSP program.¹ The sequence of each chain is submitted immediately to the prediction servers using

Abbreviations: 3D, three-dimensional; 1D, one-dimensional (e.g., string of secondary structure); DSSP, programs and database assigning secondary structure from 3D coordinates¹; JPred2, divergent profile (PSI-BLAST) based neural network prediction²; PHD, profile-based neural network prediction of secondary structure (PHDsec), solvent accessibility (PHDacc), and transmembrane helices (PHDhtm)³; PHDpsi, divergent profile (PSI-BLAST)-based neural network prediction⁴; PSI-BLAST, position-specific iterated database search⁵; PDB, Protein Data Bank of experimentally determined 3D structures of proteins; PROFphd, advanced profile-based neural network prediction of secondary structure (Rost, unpublished); PROFking, cascaded statistic-based secondary structure prediction method⁶; PSIPRED, divergent profile (PSI-Blast)-based neural network prediction⁷; SAM-T99sec, neural network prediction, using Hidden Markov models as input⁸; SSpro, profile-based advanced neural network prediction method.⁹

*Correspondence to: Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, New York, NY 10032. E-mail: rost@columbia.edu, http://cubic.bioc.columbia.edu/

Received 12 March 2001; Accepted 19 October 2001

META-PP.²² Predictions are collected and sent for evaluation to the EVA satellites: to Rockefeller University for comparative modeling, to CNB Madrid for contact predictions, and to CUBIC at Columbia University for all other predictions. Depending on the category, the assessments are made available within hours to days. The central EVA site at Columbia downloads all HTML pages produced by the satellites and builds up the “latest week” results that are then mirrored at the Rockefeller University and at the CNB Madrid.

Measuring Secondary Structure Prediction Accuracy¹

Selection of data sets. Currently, EVAsec uses only proteins with new structures to evaluate secondary structure prediction. We use the following operational definition: if a pairwise alignment search detects the similarity at a level at which it detects 50% false positives, the sequence similarity is deemed “not significant.” This concept translates to a threshold above 28 identical in 100 aligned residues.^{23,24} This implied that we presented only results from the “fold recognition” and “new fold” categories in CASP. We reported results for two sets; the first set_218 contained all 218 protein chains with new structures added to PDB between June 2000 and February 2001 for which we had results for six methods. The second set_99 was a subset of set_218 with 99 chains for which we had predictions for all nine methods evaluated.

Assigning secondary structure from 3D coordinates. EVAsec uses secondary structure assignments from DSSP.¹ The eight DSSP states are converted to three states by using the following transformation: DSSP [HGI] → helix (H), DSSP [EB] → strand (E), all other DSSP states [TS] → other (L). Note that occasionally developers convert 3₁₀ helices (DSSP G), pi-helices (DSSP I), or beta-bridges (DSSP B) to the “other” state. Such a conversion seemingly increases accuracy, because these states are more difficult to predict.¹⁷

Scoring per-residue accuracy. The three-state per-residue accuracy (Q_3) is the most widely used score for evaluating secondary structure predictions. Q_3 gives the percentage of residues correctly predicted in one of the three states: helix, strand, or other. Most residues are observed in the “other” state. Hence, Q_3 can be high even for methods predicting helices and strands inaccurately. One way around this problem is to measure the percentages of residues observed in state i (HEL) predicted correctly in state i ($Q_i^{\%obs}$) and the percentage of residues predicted in i and predicted correctly in i ($Q_i^{\%prd}$). Another way around are the Matthews correlation coefficients²⁵ and the information index.^{19,26} Some methods predict 3D structure starting from rigid body secondary structure segments. These methods need predictions with a low percentage of residues confused between strand and helix as measured by the BAD score.²⁷

Scoring per-segment accuracy. In practice, methods that get most of the segment cores right are more useful than those that get some of the entire segments right. Per-residue scores cannot distinguish between these two. Many segment-based measures have been proposed²⁶; the one that appears to distinguish best between good and bad predictions is the average overlap between segments (SOV).^{26,28}

Scoring accuracy in predicting secondary structure class. A coarse-grained classification of protein structures bases on secondary structure composition.^{29,30} Hence, secondary structure predictions also imply predictions of secondary structural class. EVAsec reports the percentage of proteins correctly predicted in one of the following four classes: all-alpha (length > 60, helix > 45%, strand < 5%), all-beta (length > 60, helix < 5%, strand > 45%), alpha/beta (length > 60, helix > 30%, strand > 20%), other. The thresholds were chosen by intuition^{20,31,32} because these simplified structural classes are not separated well.³³ EVAsec also reports differences between observed and predicted overall content to measure the accuracy in predicting secondary structure composition independently of thresholds.

Ranking Methods

Methods are not ranked based on too few test proteins! EVAsec does not rank prediction methods based on too few test proteins. For example, because the accuracy of secondary structure prediction varies between proteins, accuracy estimates typically constitute averages over many test sequences, with standard deviations usually >10%. We use this standard deviation to estimate the error of the average accuracy as a function of the test set size.

A significant difference (ΔQ) between two methods, or the error of the accuracy estimate for one method, is:

$$\Delta Q = \frac{\sigma(Q, N_{\text{protLarge}})}{\sqrt{N_{\text{prot}}}}$$

where Q is the measure for accuracy, N_{prot} is the number of proteins used in the test set, $N_{\text{protLarge}}$ is the number of proteins used in a larger, representative set (>100 proteins), and $\sigma(Q, N_{\text{protLarge}})$ is the standard deviation of variable Q in a large, representative test set (assuming a Gaussian distribution of variable Q). The observation that different prediction methods typically have similar standard deviations provides a necessary justification for this approach. For example, when a method correctly predicts 75% of the residues in a test set of 16 proteins with a standard deviation of 10%, a difference relative to another method that is smaller than 2.5% (i.e., $\Delta Q = 10/\sqrt{16}$) is not significant. Thus, we cannot distinguish between two methods that predict correctly 75% and 73% of all residues, respectively.

EVAsec uses this estimate to rank methods in the following way. Assume four methods have accuracy levels of A = 75, B = 73, C = 71, and D = 68. D can be distinguished from all other methods ($\Delta Q > 2.5$ to all). Hence, it ranks last. C can be distinguished from A ($\Delta Q = 4 > 2.5$). However, A cannot be distinguished from B

¹Explicit definitions of the scores are given at: http://cubic.bioc.columbia.edu/eva/doc/measure_sec.html

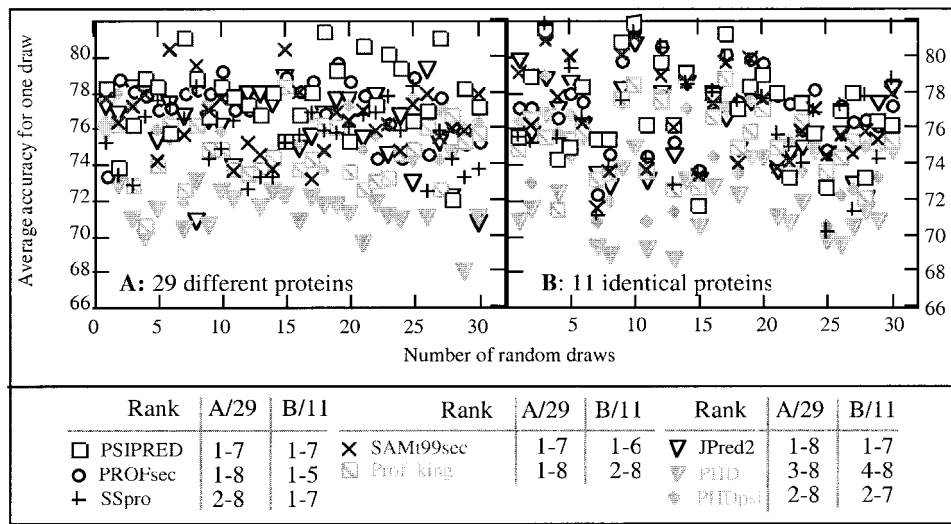


Fig. 1. Significance of averages and ranks from small numbers. CAFASP2 had 29 sequence-unique proteins to evaluate secondary structure prediction; for 11 of these we had CAFASP2 results for all methods. Each point in the graphs is an average over 29 (A) and over 11 (B) proteins for one method. Proteins are selected at random from a set of 99 protein chains. The x axes give the number of different random draws from the set of 99 proteins. The subsets of 11 proteins are constrained to be identical between all methods for every subset (B), whereas the 29 (A) are not subjected to this constraint, that is, they differ on average between the methods. Below, the spread of ranks are given that would result from the respective averages. For example, JPred2 would be the winner at one random draw and come in the last at another when using 29 incomparable proteins (A: 1–8); it would never rank last when using 11 identical proteins (B: 1–7). We conclude first that 11 proteins are clearly not enough to rank methods. Second, if we have results for all methods for 29 proteins but these are not identical 29, results are even less significant. Hence, using identical subsets is the better strategy even if it implies discarding most possible targets.

($\Delta Q = 2 < 2.5$), and B cannot be distinguished from C ($\Delta Q = 2 < 2.5$). This situation results in a dilemma that has four different possible solutions: (a) A, B, and C get the same rank ascertaining that no two methods are ranked differently that cannot be distinguished. (b) A and B get rank 1, and C rank 2 ensuring that no two methods are ranked equally that can be distinguished. (c) A gets rank 1, B rank 2 and C rank 3, ignoring that we cannot distinguish between A and B, nor between B and C. (d) Do not rank. None of these solutions is correct. During the first three CASP experiments, solution 3 was practiced (higher average results in higher rank). The evaluation of secondary structure prediction performance for CASP4 effectively implemented a concept more similar to solution a.³⁴ The first solution is also realized by EVAsec. For the example given, this implies that A, B, and C are ranked 1, and D is ranked 2.

Comparing Prediction Methods

Bootstrap experiment to test effect of small data sets. CASP4 had 53 targets for 43 structures available at the meeting in December 2000. Twenty-nine of these 43 proteins constituted fold recognition/novel fold targets at CASP (see Materials and Methods). For 11 of these 29 all methods participating at CAFASP predicted secondary structure.³⁵ Are these sets sufficient to rank methods? Although such ranking based on too small sets was carefully avoided at CASP4,³⁴ we still wanted to address this question by the following bootstrap experiment. (a)

Take a data set of proteins predicted by all M methods (here 99 protein chains from 10 methods). (b) Select at random (i) a set of 11 proteins predicted by all methods (common sets) and (ii) another M sets of 29 proteins that may differ between the methods (incomparable sets). (c) Measure the average performance on each set.

Methods best evaluated on identical subsets. The first observation from such an experiment was that methods differed considerably between different random draws, that is, between different hypothetical CASP experiments (Fig. 1). Remarkably, using 29 proteins from incomparable subsets [Fig. 1(A)] resulted in slightly higher variation than using 11 proteins from common subsets [Fig. 1(B)]. This suggested that, on average, it is a better strategy to base comparisons of methods on identical common subsets rather than on all available predictions even if the constraint to have a common subset reduces the available data from 29 to 11.

Ranking methods based on small subsets can be misleading. The variation between different random draws of data sets became even more dramatic when we ranked the methods based on the average performance: most methods did rank best and worst in one of the random draws. Ranks varied more for the 29 proteins from incomparable sets than for the 11 proteins from the common sets. Assuming that the 99 proteins constitute a representative set (which is wrong, as indicated below and in Table I), we can compile a cumulative average of one method over many random draws. How many draws do we need before the

TABLE I. Accuracy on a Common Set of 99 New Protein Chains[†]

Method	CASP	Q_3	SOV	$Q_H^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%obs}$	$Q_E^{\%prd}$	BAD	Info	C_H	C_E	Class	ΔH	ΔE
JPred2	102	75.5	67.6	72.2	84.7	58.0	75.0	1.9	0.34	0.67	0.59	80.8	6.7	5.6
PHDpsi	385	74.5	67.9	78.7	81.9	63.8	69.0	2.9	0.23	0.68	0.58	80.8	6.6	4.8
PROFphd	402	77.0	71.6	80.5	84.4	68.8	71.6	2.3	0.36	0.72	0.63	81.8	5.9	3.9
PROFking	214	74.4	67.4	74.0	86.4	70.9	66.6	2.7	0.32	0.69	0.61	84.8	7.6	6.7
PSIPRED	258	76.8	72.2	81.9	82.9	68.5	72.3	2.5	0.37	0.71	0.63	84.8	5.5	4.6
SAM-T99sec	111	76.1	70.8	84.7	80.0	62.0	76.9	1.9	0.35	0.71	0.62	80.8	6.7	4.8
SSpro	115	76	69.1	81.2	82.1	63.2	72.6	2.4	0.35	0.70	0.60	77.8	6.5	5.8
PHD	142	71.7	67.3	75.9	77.7	61.1	62.9	3.8	0.25	0.62	0.53	77.8	7.9	5.7
PSSP	510	64.3	58.7	64.4	71.5	55.6	50.0	5.9	0.20	0.50	0.40	74.7	9.8	7.8

[†]Data set and sorting: All methods have been tested on the same set of 99 new protein chains (EVA version February 2001). None of these structures was similar to any protein used to develop the respective method. This set comprised the largest such set by February 23, 2001, for which we had results. Sorting and grouping reflects the following concept: if the data set is too small to distinguish between two methods, these two are grouped. For the given set of 99 proteins, this yielded three groups. Inside of each group, results are sorted alphabetically. Note that groups are separated by an empty line; 99 proteins did not suffice to separate between the first seven methods (see Table II for a larger set). Method: See abbreviations on top of article. Scores^{19,58}: Q_3 : three-state per-residue accuracy, that is, number of residues predicted correctly in either of the three states: helix, strand, other; SOV: three-state per-segment score measuring the overlap between predicted and observed segments^{26,28}; $Q_H^{\%obs}$: residues predicted correctly in helix (or strand) as percentage of residues observed in helix (or strand); $Q_H^{\%prd}$: residues predicted correctly in helix (or strand) as percentage of residues predicted in helix (or strand); BAD: percentage of helical residues predicted as strand, and of strand residues predicted as helix²⁷; Info: per-residue information content¹⁹; C_H : Matthew's correlation coefficient for state helix⁵⁹; C_E : Matthew's correlation for state strand⁵⁹; Class: percentage of proteins correctly sorted into one of the four classes: all-alpha, all-beta, alpha/beta, other; ΔH : difference between predicted and observed secondary structure content in helix; ΔE : difference between predicted and observed secondary structure content in strand.

TABLE II. Accuracy on a Set of 218 Identical Proteins[‡]

Method ^a	Q_3	SOV	$Q_H^{\%obs}$	$Q_H^{\%prd}$	$Q_E^{\%obs}$	$Q_E^{\%prd}$	BAD	info	C_H	C_E	Class	ΔH	ΔE
PROFsec	76.8	72.8	80.5	84.4	68.8	71.6	2.2	0.36	0.72	0.63	82.1	5.6	4.0
PSIPRED	76.4	72.0	81.9	82.9	68.5	72.3	2.5	0.37	0.71	0.63	79.8	5.4	4.4
SSpro	76.1	71.2	81.2	82.1	63.2	72.6	2.5	0.35	0.70	0.60	81.2	6.0	5.2
JPred2	74.8	69.3	72.2	84.7	58.0	75.0	2.4	0.34	0.67	0.59	76.1	7.6	5.7
PHDpsi	74.7	69.6	78.7	81.9	63.8	69.0	3	0.29	0.68	0.58	79.8	6.2	4.9
PHD	71.4	67.4	75.9	77.7	61.1	62.9	4.2	0.25	0.62	0.53	76.1	7.7	5.9

[‡]Symbols as in Table I; results are based on 218 proteins not used for developing the methods.

methods will reach the averages of the original set? About 30 draws for the common subsets of 11 proteins and more than 60 for the incomparable subsets of 29 proteins (data not shown). In our experiment, we forced all methods to predict equal numbers of proteins. Most methods provide estimates for the reliability of the prediction for each residue (see Fig. 3). What if methods submitted only their seemingly best predictions? We used such an index to select only the most reliable predictions from one method while forcing all other methods to predict for all proteins. Note that this selection was realized without knowing the accuracy of a particular prediction. It is surprising that we could make *every* method for which we had such an index become the winner at *any* of the random draws by submitting only the most reliable predictions! Our bootstrap experiments underlined what most CASP evaluators practiced: ranking methods based on small subsets may not be appropriate.

RESULTS

Better alignments improved secondary structure predictions significantly. The set of 99 new protein chains for which EVA collected results for all methods did not suffice to distinguish between all methods. However, some trends became apparent. A number of methods in 2001 predict

secondary structure more accurately than did the best method of 1996. In fact, all methods using alignments not restricted to pairwise comparisons performed significantly better than PHD using only pairwise alignment information (Table I). Simply replacing the pairwise alignments input to PHD by PSI-BLAST profiles⁵ made the resulting PHDpsi rank in the “winner” group. (Upon closer look: most of the improvement of PHDpsi over PHD resulted from using larger databases, rather than from using PSI-BLAST.⁴) Did this imply that nothing has changed but the databases and the search methods? Ninety-nine protein chains did not suffice to tell.

And the winners are . . . For a few of the methods EVA had 2–3 times larger data sets. In particular, 218 protein chains sufficed to distinguish between some of the methods indistinguishable when evaluated on 99 chains: PROFphd, PSIPRED, and SSpro were significantly more accurate than JPred2 and PHDpsi (Table II). All three “winners” were equally balanced in predicting strand and helix and had a similar level of performance in predicting secondary structure content. Unlike the bulk of methods presented during the early CASP meetings, all seven methods shown in Table I predicted β -strands on average more accurately than residues in nonregular structure. Although SSpro was significantly less accurate in predicting segments

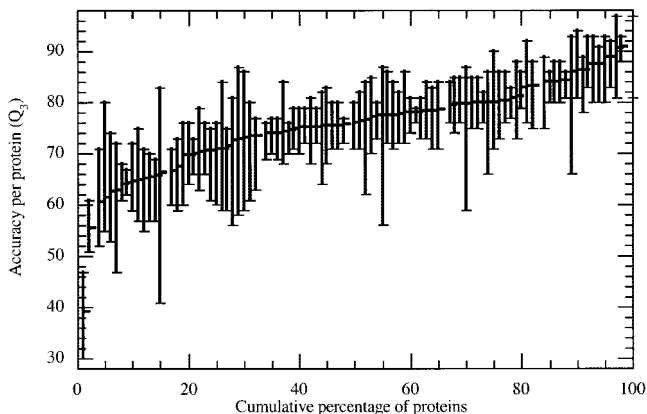


Fig. 2. Best, worst, and average predictions for each protein. Each cross is the average over all seven prediction methods that ranked best on 99 protein chains (Table I); triangles give the lowest and highest accuracy for a particular protein. The x axis describes for which fraction of the 99 proteins the average was above a certain value. For example, the average accuracy was <70% for 20% of all proteins, whereas only for 6 of 99 the best method did not reach 70% accuracy. Furthermore, counting the filled triangles below the 70% line revealed that for half the proteins (48 of the 99), even the worst method surpassed 70% accuracy.

than PROFphd, all three best methods predicted segments much better than the two second best methods JPred2 and PHDpsi.

Some proteins predicted well by all methods. The per-protein average over the seven methods that performed best on the set of 99 chains (Table I) varied as strongly between proteins as did each of the methods. Did this finding confirm the notion that some proteins were easier to predict than others? On average, the worst prediction method for a particular protein chain had a higher accuracy when the average over all methods was higher (Fig. 2). In other words, some chains were predicted by all methods more accurately than others, in particular, the average over all methods was <70% (Q_3) for 20% of the chains, >80% for another 20% of the chains, and between 70 and 80 for all other chains (Fig. 2). However, for many chains, the average over all methods was high, although some method had a very low accuracy. The best methods reached about 77% accuracy. For >80% of all proteins, one method performed better than this. Hence, most proteins were predicted above average by at least one method. For only two of the 99 chains, all methods reached <68% accuracy (Appendix). The worst predictions were obtained for the short peptide of the human apolipoprotein II (1by6:A), the structure of which was determined by NMR. By default, we used the first NMR model to determine prediction errors, although other models correlated better with the predictions (data not shown). The other bad predictions were for the endonuclease I-PPOI complexed to DNA (1evw:A). The major problems were that most methods missed the two helices reaching into the DNA on opposite sites of the molecule and overpredicted a long strand for two parallel nonhydrogen-bonded stretches on the opposite site of the DNA binding. For 66 of the 99 protein chains, one method had >80% of the residues predicted correctly, whereas the average over all methods reached

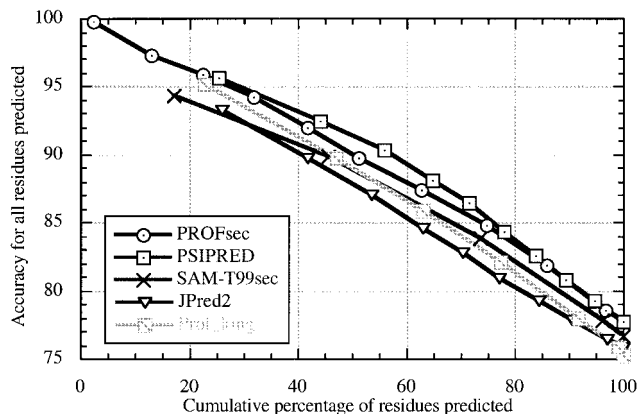


Fig. 3. Prediction strength correlated well with accuracy. Residues predicted at higher reliability are predicted more accurately.^{3,19} Reliability indices are now used by most methods. Shown are cumulative values, that is, the accuracy (Q_3) for all residues predicted above a given reliability. For example, for all methods, 90% of the 40% most strongly predicted residues are predicted correctly. Results were based on the set_99 also used for Table I.

this level for only 25 of 99. Could we anticipate which method did best on which protein without knowing the structure?

Prediction accuracy did not correlate well with experimental resolution. We used our largest set of 218 chains to analyze whether prediction accuracy correlated with the resolution of the respective structure. It is surprising that we could not find a strong correlation between accuracy and resolution. Nevertheless, prediction accuracy was about 3% higher for the X-ray structures than for the first NMR models. Furthermore, when averaging all predictions for the quarter of all chains with highest resolution, we found levels of accuracy about 4% above the average over the quarter with lowest resolution.

Reliability indices estimated prediction accuracy accurately. Prediction methods typically use three output states for helix, strand, and other and predict the state with the highest value as the secondary structure of the respective residue. Assume the output for one residue is (0.3, 0.4, 0.3); that for another (0.1, 0.8, 0.1); both residues are predicted as strand. However, the second prediction is much stronger. This difference can be carved into an index describing the reliability of a secondary structure prediction for each residue.^{19,36} After the successes of such indices at the first CASP meeting, almost all methods now implement estimates for the reliability of the prediction for each residue. For all methods tested, these indices correlated surprisingly well with accuracy (Fig. 3). For example, PSIPRED and PROFphd reached levels >90% for the half of the residues predicted most strongly.

Secondary structural class predicted almost as accurately as by experiment. Grouping proteins into secondary structure classes (all-alpha, all-beta, alpha/beta, other) appears a useful initial approach toward classifying proteins.^{37,38} Such classes can be predicted successfully merely on the basis of the overall amino acid composition of a protein.³⁹⁻⁴¹ More and more increasingly complex and genial methods address this reduced goal; reported levels

of prediction accuracy approach 100%. Recently, Wang and Yuan explained these high values by insufficient testing schemes and challenged that a four-state accuracy of around 60% comprises the maximum for methods based solely on composition.⁴¹ Obviously, it is much easier to predict class starting from the detailed information about evolutionary profiles for the entire sequence than by restricting the input to composition. In fact, today's best general prediction methods also predict secondary structure class better (Tables I and II). The differences between observed and predicted secondary structure composition are now <6% for helix and strand. This performance is similar to what experimental low-resolution (circular dichroism, Fourier transform induced spectroscopy) methods achieve at their best.^{20,42}

Homologues of known structures predicted marginally better. All current top-of-the-line methods somehow learn the secondary structure for proteins of known structure. In particular, no method relies entirely on "first principles" (e.g., one of the best methods of the second generation ALB did⁴³). Consequently, today's methods somehow depend on residual sequence similarity between target and known structures. Did this imply that prediction accuracy was significantly higher for proteins with homologues of known structure? For example, PSIPRED reached a level close to 80% accuracy for a set of 223 protein chains with significant sequence similarity to known structures (data not shown) compared to about 76–77% for new structures (Table II). A similar trend persisted for all prediction methods analyzed (data not shown). How would these values compare to inferring secondary structure through comparative modeling? We did not yet compile data to address this question explicitly. However, for structural alignments, the respective secondary structure assignments agree for >88% of all residues.^{26,44} Hence, when we know a protein of known structure that is similar to a target, we supposedly still best use comparative modeling to predict secondary structure. A similar conclusion was suggested by analyzing the CASP4 results for comparative modeling.³⁴

CONCLUSIONS

CASP and EVA: both are needed. For CAFASP, all methods predicted secondary structure for the same 11 proteins, and some methods for another 18. Our bootstrap experiment provided some numbers illustrating problems with ranking methods based on too small data sets. Can we conclude anything from small sets? Certainly, but the level of detail depends on the data set. For example, 99 proteins sufficed to conclude that seven methods were more accurate than was pairwise PHD (Table I). However, we needed >200 proteins to distinguish between some of the seven best (Table II). EVA continues to assess prediction methods automatically on as many proteins every month as does CASP every two years.¹⁵ Should we then have CASP without, for example, secondary structure prediction? We perceive that the advance in prediction methods over the last 8 years has been influenced strongly by CASP. Further advances might stall with no secondary

structure prediction present at CASP. We suggest that expert predictions and automatic methods based on the limited number of CASP targets be compared and relate the performance to the larger EVA sets. What do secondary structure predictions teach us about protein function? This is a kind of problem that EVA cannot address. We need expert evaluations to learn what to measure. Finally, all tables given here are available through the EVA web site; the interpretation of the data is not.

The field advanced significantly. Growing databases and improved search techniques yielded a substantial improvement in secondary structure prediction over the last 4 years. The best methods now reach sustained levels of 76% (Tables I and II). For almost every second protein, even the worst of the seven best methods (Table I) surpassed 70% accuracy, and for <10% of the proteins, the 70% level was not reached by the best method (Fig. 2). Even more impressively, about 60% of all residues are predicted at levels similar to structural alignments of homologues (Fig. 3).

Eighty-eight percent is a limit, but shall we ever reach close to there? Protein secondary structure formation is influenced by long-range interactions^{45–47} and by the environment.^{48,49} Consequently, stretches of up to 11 adjacent residues (dubbed chameleon after⁴⁵) can be found in different secondary structure states.^{50–52} Implicitly, such nonlocal effects are contained in the exchange patterns of protein families. This is reflected by the fact that strand is predicted almost as accurately as helix (Table I), although sheets are stabilized by more nonlocal interactions than helices. Local evolutionary profiles can even suffice to identify structural switches.^{48,53} It is surprising that we can find some traces of folding events in secondary structure predictions.⁵⁴ Even more amazing is a study suggesting that alignment-based methods achieve similar levels of accuracy for chameleon regions as for all other regions.⁵¹ Secondary structure assignments may vary for two versions of the same structure. One reason is that protein structures are no rocks but dynamic objects with some regions more mobile than others. Another reason is that any assignment method has to choose particular thresholds. Consequently, assignments differ by about 5–10% between different NMR models for the same protein⁴⁴ and by about 12% between structural homologues.²⁶ The latter number provides an upper limit for secondary structure prediction of error-free comparative modeling. After the recent advances, we have reached >76%. Thus, we need to mount another 12% (or even less).

Is the major obstacle the size of the experimental database as suggested by Pan et al.⁵⁰ PHDpsi was trained on 200 proteins; when using PSI-BLAST input, it was almost as accurate as PSIPRED trained on 2000 proteins (Table II). Hence, the database growth may not suffice. Will the current explosion of sequences boost accuracy? In fact, current databases have <10 homologues for more than one third of the proteins and >100 for only 20% of the proteins. Although based on a too small set for conclusions, for these 20% highly populated families the accuracy of PROFphd was 4% above average (data not shown). Thus,

larger databases may get us 6% higher, and it may not. The answer remains nebulous.

What are the major problems of the field? Most major problems prominent in many of the predictions submitted to the first two CASP meetings have been solved. The most important task may now be to correlate predicted secondary structure to aspects of protein function. One method has successfully related secondary structure predictions automatically to functional aspects.^{48,53} However, secondary structure-based identifications of binding sites or other functional aspects is still restricted to single-case expert analyses. Other than this, a number of loose ends remained. (a) All methods still have problems predicting the precise termini of regular secondary structure segments. (b) Frequently, the number of helices and strands is not predicted correctly. (c) We know that evolutionary information improves prediction accuracy. However, we still have not succeeded to correlate the information contained in a particular alignment with the resulting improvement in prediction accuracy. (d) Rather than improving existing methods even further, the field should possibly attempt to expand the concept of secondary structure by predicting other states (e.g., turns⁵⁵) or different descriptions of supersecondary structure (e.g., as used in Isites⁵⁶). Write about bad predictions!

And now we run human? The field has advanced considerably; more improvement appears to lie ahead. Prediction methods are fast enough to analyze entire genomes, and for particular examples, the resulting classifications are relevant to structural and functional genomics.^{38,57} Nevertheless, to play the devil's advocate: We are missing a variety of approaches relating secondary structure predictions explicitly to function. Obviously, this remark may apply to bioinformatics, in general: The new millennium began with the publication of the entire human genome; we must rush to get ready for the data flood.

ACKNOWLEDGMENTS

We thank the EVA teams at the Rockefeller University (Marc A. Martí-Renom, András Fiser, and Andrej Sali) and at the CNB in Madrid (Florencio Pazos and Alfonso Valencia) for joining us in pursuing a laborious idea. We thank also Jinfeng Liu and Dariusz Przybylski (CUBIC, Columbia) for helping with software and hardware. Furthermore, we are grateful to Phil Bourne (UCSD) for his support and to Kevin Karplus (UCSC) for numerous suggestions. Last not least, we thank all the developers who accepted EVA submissions: Pierre Baldi (Irvine), Phil Bourne (UCSD), Søren Brunak (Copenhagen), James Cuff (London), Piero Fariselli (Bologna), Mitsuo Iwadate (Kitasoto), Ross King (Aberystwyth), Ole Lund (Copenhagen), Jarek Meller (Ithaca), David Jones (Uxbridge), Kevin Karplus (UCSC), Osvaldo Olmea (Madrid), Gianluca Pollastri (Irvine), Gajendra Raghava (Chandigarh), Kristoffer Rapacki (Copenhagen), Torsten Schwede (Geneva).

REFERENCES

- Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
- Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 1999;34:508–519.
- Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol* 1996;266:525–539.
- Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. *Proteins* 2002;46:197–205.
- Altschul S, Madden T, Shaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
- Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci* 2000;9:1162–1176.
- Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
- Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. *Proteins* 1999;33:121–125.
- Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 1999;15:937–946.
- Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 1996;25:113–136.
- Rost B, O'Donoghue SI. Sisyphus and prediction of protein structure. *CABIOS* 1997;13:345–356.
- Szent-Györgyi AG, Cohen C. Role of proline in polypeptide chain configuration of proteins. *Science* 1957;126:697.
- Rost B, Sander C. Progress of 1D protein structure prediction at last. *Proteins* 1995;23:295–300.
- Rost B. Better 1D predictions by experts with machines. *Proteins Suppl* 1997;1:192–197.
- Eyrich V, Martí-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* 2001. Forthcoming.
- Eyrich V, Martí-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. WWW document (<http://cubic.bioc.columbia.edu/eva/>): Columbia University, 2001.
- Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. *Bioinformatics* 1998;14:892–893.
- McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;16:404–405.
- Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584–599.
- Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 1994;19:55–72.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
- Eyrich V, Rost B. The META-PredictProtein server. WWW document (http://cubic.bioc.columbia.edu/predictprotein/submit_meta.html): CUBIC, Columbia University, Department of Biochemistry and Molecular Biophysics, 2000.
- Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 1991;9:56–68.
- Rost B. Twilight zone of protein sequence alignments. *Prot Eng* 1999;12:85–94.
- Mathews FS. The structure, function and evolution of cytochromes. *Prog Biophys Mol Biol* 1985;45:1–56.
- Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol* 1994;235:13–26.
- Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. *Proteins* 1995;23:431–445.
- Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 1999;34:220–223.
- Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* 1976;104:59–107.
- Levitt M, Chothia C. Structural patterns in globular proteins. *Nature* 1976;261:552–558.
- Kneller DG, Cohen FE, Langridge R. Improvements in protein

- secondary structure prediction by an enhanced neural network. *J Mol Biol* 1990;214:171–182.
32. Zhang C-T, Chou K-C. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci* 1992;1:401–408.
 33. Rost B. Observed secondary structure content for 721 proteins. WWW document (<http://cubic.bioc.columbia.edu/results/1996/SecStrContent.html>): EMBL Heidelberg, Germany, 1996.
 34. Lesk AM, Lo Conte L, Hubbard TJP. Assessment of novel folds targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins* 2001; Suppl 5:98–118.
 35. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL. CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* 2001; Suppl 5:171–183.
 36. Rost B, Sander C. Jury returns on structure prediction. *Nature* 1992;360:540.
 37. Gerstein M, Levitt M. A structural census of the current population of protein sequences. *Proc Natl Acad Sci USA* 1997;94:11911–11916.
 38. Przytycka T, Aurora R, Rose GD. A protein taxonomy based on secondary structure. *Nat Struct Biol* 1999;6:672–682.
 39. Liu W, Chou KC. Prediction of protein secondary structure content. *Prot Eng* 1999;12:1041–1050.
 40. Zhang CT, Zhang R. Skewed distribution of protein secondary structure contents over the conformational triangle. *Prot Eng* 1999;12:807–10.
 41. Wang Z-X, Yuan Z. How good is prediction of protein structural class by the component-coupled method? *Proteins* 2000;38:165–175.
 42. Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. *Proteins* 1999;35:293–306.
 43. Ptitsyn OB, Finkelstein AV. Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* 1983;22:15–25.
 44. Andersen CAF, Palmer AG, Brunak S, Rost B. Continuous secondary structure assignment correlates with protein flexibility. *Structure*. 2001. Submitted for publication.
 45. Minor DLJ, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. *Nature* 1996;380:730–734.
 46. Muñoz V, Cronet P, López-Hernández E, Serrano L. Analysis of the effect of local interactions on protein stability. *Folding Design* 1996;1:167–178.
 47. Villegas V, Zurdo J, Filimonov VV, Aviles FX, Dobson CM, Serrano L. Protein engineering as a strategy to avoid formation of amyloid fibrils. *Protein Sci* 2000;9:1700–1708.
 48. Young M, Kirshenbaum K, Dill KA, Highsmith S. Predicting conformational switches in proteins. *Protein Sci* 1999;8:1752–1764.
 49. Krittanai C, Johnson WCJ. The relative order of helical propensity of amino acids changes with solvent environment. *Proteins* 2000;39:132–141.
 50. Pan XM, Niu WD, Wang ZX. What is the minimum number of residues to determine the secondary structural state? *J Protein Chem* 1999;18:579–584.
 51. Jacoboni I, Martelli PL, Fariselli P, Compiani M, Casadio R. Predictions of protein segments with the same amino acid sequence and different secondary structure: a benchmark for predictive methods. *Proteins* 2000;41:535–544.
 52. Zhou X, Alber F, Folkers G, Gonnet GH, Chelvanayagam G. An analysis of the helix-to-strand transition between peptides with identical sequence. *Proteins* 2000;41:248–256.
 53. Kirshenbaum K, Young M, Highsmith S. Predicting allosteric switches in myosins. *Protein Sci* 1999;8:1806–1815.
 54. Compiani M, Fariselli P, Martelli PL, Casadio R. Neural networks to study invariant features of protein folding. *Theoret Chem Acc* 1999;101:21–26.
 55. Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Sci* 1999;8:1045–55.
 56. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* 2000;301:173–190.
 57. Teichmann SA, Chothia C, Gerstein M. Advances in structural genomics. *Curr Opin Struct Biol* 1999;9:390–399.
 58. Rost B. EVA measures of secondary structure prediction accuracy. WWW document (http://cubic.bioc.columbia.edu/eva/doc/measurement_sec.html): EMBL, 2001.
 59. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;405:442–451.