Alignments Grow, Secondary Structure Prediction Improves

Dariusz Przybylski and Burkhard Rost*
Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

ABSTRACT **Using information from sequence alignments significantly improves protein secondary structure prediction. Typically, more divergent profiles yield better predictions. Recently, various groups have shown that accuracy can be improved significantly by using PSI-BLAST profiles to develop new prediction methods. Here, we focused on the influences of various alignment strategies on two 8-year-old PHD methods. The following results stood out. (i) PHD using pairwise alignments predicts about 72% of all residues correctly in one of the three states: helix, strand, and other. Using larger databases and PSI-BLAST raised accuracy to 75%. (ii) More than 60% of the improvement originated from the growth of current sequence databases; about 20% resulted from detailed changes in the alignment procedure (substitution matrix, thresholds, and gap penalties). Another 20% of the improvement resulted from carefully using iterated PSI-BLAST searches. (iii) It is of interest that we failed to improve prediction accuracy further when attempting to refine the alignment by dynamic programming (MaxHom and ClustalW). (iv) Improvement through family growth appears to saturate at some point. However, most families have not reached this saturation. Hence, we anticipate that prediction accuracy will continue to rise with database growth. Proteins 2002;46:197–205.**

Key words: protein structure prediction; solvent accessibility; evolutionary information; profiles-based multiple alignments; dynamic programming; neural networks; PSI-BLAST

INTRODUCTION

Evolutionary information improves structure prediction. Proteins with similar sequences adopt similar structures.[3,8] In fact, proteins can exchange >70% of all their residues without altering the basic fold.[9–12] However, the vast majority of possible sequences supposedly do not adopt globular structures at all. Rather, the exact substitution pattern of which residues can be exchanged against which other is indicative of particular structural details. Consequently, the evolutionary information contained in sequence alignments can aid structure prediction. This finding has been realized for a long time.[9,13–17] The breakthrough in automatically using this information was achieved by applying neural networks to the problem of secondary structure prediction.[18,19] Replacing single sequences by family profiles improved prediction accuracy by about 5%.[19,20] The success in using evolutionary information for secondary structure prediction was not restricted to neural networks.[21–27] Furthermore, evolutionary information proved also beneficial for predicting other aspects of protein structure.[5,28–42]

More divergence yields better predictions. How much divergence in a family is needed to improve prediction accuracy? The more, the better! In the extreme: if we could use structural alignments to identify remote homologues and to build profiles, we would get better improvements.[43] The trouble with this promising concept is, of course, that we cannot structurally align proteins of unknown structure. However, the iterated, profile-based PSI-BLAST program[6] achieved the breakthrough, in practice, of another old idea: use profiles to refine database searches. PSI-BLAST identifies more distant relations than pairwise alignment methods do.[11] This increased detection of very diverged family members has been used successfully to improve prediction accuracy by training neural networks on the PSI-BLAST profiles.[42,44] The impressive improvement pioneered by David Jones[44] is based on developing a new prediction method. Here, we tried to isolate the causes for the recent improvement. Although Cuff and Barton[42,45] investigated how a new method could benefit from particular alignment strategies, we wanted to estimate how grown databases and better search tech-

Abbreviations: BIG, database merging SWISS-PROT + TrEMBL + PDB; BLAST, fast alignment method[1]; ClustalW, profile-based dynamic programming alignment method[2]; MaxHom, dynamic programming algorithm for conservation weight-based multiple sequence alignment[3]; PDB, Protein Data Bank of experimentally determined protein structures[4]; PHD, profilbased neural network prediction of secondary structure (PHDsec[5]) and solvent accessibility (PHDacc[5]); PSI-BLAST, position-specific iterated BLAST database search[6]; SWISS-PROT, annotated database of protein sequences[7]; TrEMBLE, translation of the EMBLE-nucleotide database coding DNA to protein sequences.[7]

Grant sponsor: National Institutes of Health; Grant number: RO1-GM63029-01.

Availability: PHD and our protocol for automatic iterated PSI-BLAST searches are available through PredictProtein http://cubic.bioc.columbia.edu/predictprotein; the filtered PSI-BLAST database and additional scripts are available upon request from the authors.

*Correspondence to: Burkhard Rost, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY. E-mail: rost@columbia.edu

Received 26 February 2001; Accepted 21 August 2001

DOI 10.1002/prot.10029

niques would improve existing methods. In particular, we chose the 8-year-old method PHDsec as an example for "any method." Incidentally, Barton and Cuff used PHDsec to underline their conclusion that using more informative alignments alone did not significantly improve prediction accuracy.[42]

MATERIALS AND METHODS

Alignment and Prediction Methods

We obtained the following alignment methods from the respective public web sites: PSI-BLAST[1,6]: ftp://ncbi.nlm.nih.gov/blast/; ClustalW[2,46] ftp://ftp.ebi.ac.uk/pub/software/. The source code for MaxHom[3,47] was obtained from Reinhard Schneider (LION Biosciences, Boston, MA). All predictions were obtained from the publicly available PHD programs: PHDsec for secondary structure[5,19,20] and PHDacc for solvent accessibility.[5,32]

Prefiltering the Database for PSI-BLAST

David Jones reported the importance of prefiltering the database used for running PSI-BLAST.[44] We followed that concept. First, we combined SWISS-PROT,[7] TrEMBL,[7] and PDB[4] into one big database (referred to as BIG). Second, we marked low-complexity regions by using SEG.[48,49] Third, we marked coiled coil regions with COILS.[50,51]

Building Families With PSI-BLAST and Filtering the Output

First, we searched with PSI-BLAST against our prefiltered database restricting the number of iterations to three (if not stated otherwise). Second, we included all proteins into the family that were below a certain BLAST E-value. Before using the final alignment for prediction, we reduced the redundancy by omitting all proteins >80% identical to any protein previously taken. Finally, we realigned all proteins found by using ClustalW and MaxHom.

Monitoring Changes by Simple Per-Residue Scores

The most established measure for secondary structure prediction accuracy is the three-state per-residue accuracy, often referred to as Q_3.[19,52] Q_3 reports the percentage of residues correctly predicted in one of the three states helix (H), strand (E), or other (L):

$$Q_3 = \frac{1}{\text{Nprot}} \sum_{k}^{\text{Nprot}} 100 \cdot \frac{\text{number of residues predicted in [HEL]}}{\text{number of residues in protein } k} \qquad (1)$$

where Nprot was the number of proteins in the data set. We assigned secondary structure from PDB files by the default DSSP 53, with the following conversions of the 8 DSSP states: [HGI] → H (helix), [EB] → E (strand), [ST] → L (other, nonregular). More elaborate scores[19,21,54] yielded qualitatively similar results (data not shown). Solvent accessibility was also taken from DSSP, with the standard conversion of accessibility to relative values.[32] To monitor prediction accuracy, we used a two-state per-residue score, giving the percentage of residues correctly predicted as either buried (<16% accessible) or exposed (≥16% accessible). Many other measures for prediction accuracy are beneficial to evaluate methods.[21,32,52,5,55] However, here we were only concerned about separating various sources of improvement.

Data Sets for Evaluation

We provided mostly relative values, that is, "improvement relative to using standard PHD alignments." One reason for this was that relative values did not vary between different data sets chosen to estimate prediction accuracy. Nevertheless, we attempted to base all numbers on data sets "as clean as possible." In particular, none of the proteins used for evaluation had significant levels of sequence identity to any protein used for developing PHD. As a threshold for "significant identity," we used a mark of 5 percentage points below our previously established line implying structural similarity[56] (this roughly translates to <25% sequence identity over >100 residues). This particular cutoff implies that pairwise sequence searches have >90% wrong hits. Thus, our data sets were fairly conservative in terms of "distance to known structures." We tested three different data sets. First, 199 proteins added to PDB between June and December 1999 (dubbed "set_199"). Second, 264 proteins added from April 2000 to January 2001 and used by the EVA server (dubbed "set_EVA264"[57]). Note for this particular set we also had blind results from other methods available by fulfilling the criterion that "no protein was similar to any protein used to develop that method." Third, 1136 proteins added to PDB between 1994—when PHD was developed—and 1999 (dubbed "set_1136"). All three sets were nonredundant in the sense that no pair in the respective set had significant sequence identity.

Significant Differences

Plotting Q_3 for many proteins yields a Gaussian-like distribution. Thus, whether a difference in prediction accuracy is significant depends on the standard deviation of that distribution and on the size of the data set. Here, we used the following rule-of-thumb to refer to a difference in accuracy (ΔQ) as "significant":

$$|\Delta Q_{\text{significant}}| \geq \frac{\sigma}{\sqrt{N}} \qquad (2)$$

where N was the number of proteins in the data set and σ is the standard deviation of Q over that set. Typically, standard deviations for Q_3 are in the order of 10. Note that Eq. 2 typically is known as the "standard error." Hence, differences of 2 percentage points for 9 proteins are not significant ($2 < 10/3 = 3.33$), whereas differences of 1 percentage point are when based on 225 proteins ($1 > 10/15 = 0.66$). Thus, for our largest data set "set_1136," differences above 0.3 were significant.

TABLE I. Improvement by Using Different Methods to Realign Pairwise BLAST Hits[†]

	SWISS-PROT[a]		BIG[b]	
Method	$E <$ 1	$E <$ 10^{-3}	$E <$ 1	$E <$ 10^{-3}
BLAST	8.2	7.6	9.7	9.2
Simple ClustalW[c]	4.4	5.4		
Profile ClustalW[d]	5.4	7.1		
MaxHom with McLachlan[e]	7.2	7.5	9.0	8.9
MaxHom with BLOSUM62[f]	8.3	7.9	9.5	9.1
BLAST-filter[g]	7.9	7.6	9.5	9.2
Profile-based BLAST[h]	8.2	7.8	9.6	9.1
Significant difference	>0.44	>0.44	>0.44	>0.44

[†] Given are percentage points by which PHD improved over single sequence-based predictions by using the respective alignment methods three-state per-residue accuracy Q_3. The baseline reflected the performance of PHD on single sequences (PHDsec = 66.3%, PHDacc = 71.0%). In all cases, we used pairwise BLAST[1] searches to select the proteins to be aligned with a BLAST E-value threshold of 1 (left columns: more homologues, more false positives) and of 10^{-3} (right columns: fewer homologues, fewer false positives). All results are based on a set of 199 proteins.

[a] Using only homologues found in SWISS-PROT.[7]

[b] Using homologues found in a "non-redundant" database merging SWISS-PROT + TrEMBL + PDB.[4,7]

[c] Dynamic programming with default parameters[2] (note: we could not test ClustalW on BIG since our CPU time was too limited).

[d] Dynamic programming with gradual profile alignment by default parameters, sequences are brought into alignment one by one[2] (note: we could not test ClustalW on BIG since our CPU time was too limited).

[e] Dynamic programming with McLachlan[63] matrix and gap penalties same as those used in generation of HSSP database.[3]

[f] Dynamic programming with BLOSUM62[60] and gap penalties of 10 + k, where k is a length of a gap.

[g] Homology reduction of BLAST alignments using a modified HSSP curve.[56]

[i] Sequences found by BLAST are realigned by using a position-specific scoring matrix produced by PSI-BLAST[6] on BIG (note: no sequence found through iterations was used here).

TABLE II. Improvement by Using Different Pairwise BLAST Thresholds[†]

E-value[a]	PHDsec[b]	PHDacc[c]
100	8.7	4.4
20	9.1	4.9
10	9.5	5.0
1	9.7	5.2
10^{-1}	9.5	5.3
10^{-2}	9.2	5.3
10^{-3}	9.1	5.2
10^{-4}	8.9	5.2
10^{-7}	8.5	5.0
10^{-20}	6.9	4.5
Significant difference	>0.44	>0.39

[†] Given are percentage points by which PHD improved over single sequence-based predictions by using the pairwise BLAST[1] searches on BIG. Data set as in Table I.

[a] Maximal E-value of sequences included in final alignment.

[b] Secondary structure prediction accuracy (PHDsec) improvements as given by an increase in the three-state per-residue accuracy Q_3.

[c] Solvent accessibility prediction accuracy (PHDacc) improvements as given by an increase in the two-state per-residue accuracy Q_2.

RESULTS

Better Secondary Structure Prediction by Using BLAST

First, we searched with pairwise BLAST and the Blosum62 matrix against SWISS-PROT.[7] This improved accuracy over our previous strategy (MaxHom with McLachlan) significantly by about 1 percentage point (Q_3, Eq. 1, Table I). Encouraged by the immediate success, we tried to further improve by using a variety of other alignment methods. We failed (Table I). Next, we applied the full Smith–Waterman alignment algorithm[58] implemented in MaxHom[3] to sequences identified by BLAST. The gain was insignificant (Table I). It was surprising that when we generated multiple alignments with ClustalW,[2] prediction accuracy decreased significantly (simple ClustalW in Table I). A similar tendency was reported by Cuff and Barton.[42] This may be due to the sensitivity of ClustalW to including proteins of unrelated structure in the alignment (false positives). In fact, many of the proteins found at high E-values were likely false positives.[11,12] Such errors may affect the quality of the multiple alignment, especially when close to a root of the family dendrogram used by ClustalW to align a family. The effect decreased when we built the multiple alignment gradually, starting from a query sequence and proceeding toward more distant homologues (profile ClustalW in Table I). Finally, we tried to filter out possible false positives by using our extension of the HSSP-curve.[3,56] This step decreased prediction accuracy albeit insignificantly (BLAST-filter in Table I). Hence, it was beneficial to include more distant homologues in the alignment even if some of them were false positives. More precisely, the increase in divergence was more beneficial than the inclusion of false positives was detrimental.

Significant Improvement Through Larger Database

The PHD methods were developed, analyzed, and distributed on the basis of alignments generated from the SWISS-PROT database (currently containing about 90,000 sequences). When we switched from SWISS-PROT to a large "non-identical" database (BIG = SWISS-PROT + TrEMBL + PDB) of about 500,000 sequences, predictions increased significantly by an additional 1.5 percentage points (Table I). On average, we aligned about 2.7 times more proteins to each query sequence when using BIG. We observed a strong dependence of the performance on the threshold chosen to include sequences into the family: alignments containing sequences with E-values ≤ 1 (corresponding to P-value of about 0.63) yielded the highest prediction accuracy (Table II). Although adding sequences to the families improved accuracy for most proteins, occasionally accuracy dropped [Fig. 1(A)]. In fact, for some proteins, predictions based on single sequences were more accurate than those based on alignments (negative values in Fig. 1). This effect persisted even when including only proteins with E-values $< 10^{-20}$ [Fig. 1(B)], suggesting that the drop in accuracy was not only caused by false positives.

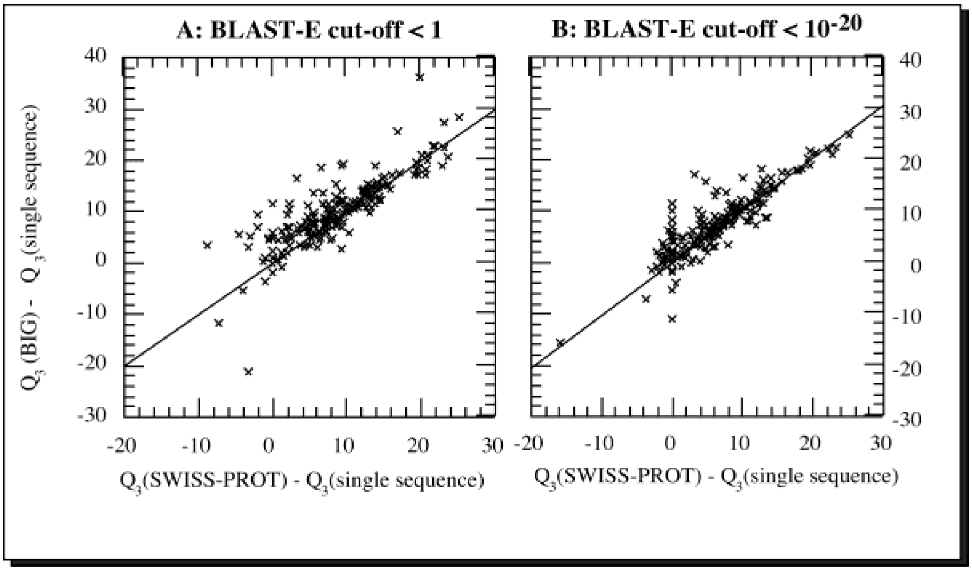

Fig. 1. Influence of database size on prediction accuracy. The improvement in the three-state per-residue accuracy is given as differences between predictions based on alignments vs. predictions based on single sequences. Thus, negative numbers imply that evolutionary information was detrimental. All alignments were generated by simple pairwise BLAST searches using E-value thresholds for including sequences of one (**A**) and 10^{-20} (**B**). The diagonal lines mark proteins predicted equally well when searching through SWISS-PROT and BIG. Points well above the diagonal mark proteins for which the larger database was highly beneficial. The observation of points below the diagonal for the conservative cutoff threshold (B) suggested that the decrease from using the larger database was not caused by accumulating false positives.

PSI-BLAST Improved Secondary Structure Prediction Accuracy Slightly

PSI-BLAST finds more distantly related homologues than pairwise search methods.[11] These extended profiles have been reported to improve prediction accuracy significantly.[44,45] In contrast, we noticed only a marginal improvement of about 0.4 percentage points through using PSI-BLAST (Table III). In fact, this was below the mark of 0.44 for significant differences. However, we observed consistently positive effects by using PSI-BLAST on other data sets. In particular, for the set of new proteins used by EVA,[57,59] PSI-BLAST improved by about 0.6 percentage points. Given a standard error on that set of around 0.6 (Eq. 2), this was at the edge of being statistically significant, albeit rather small in comparison to the difference between using SWISS-PROT and BIG with a simple BLAST (1.8 percentage points). Similarly, we observed an increase of 0.4 percentage points for "set_1136" for which differences above 0.3 were statistically significant.

TABLE III. Improvement by Using PSI-BLAST[†]

Iteration E-value[a]	PHDsec[b]	PHDacc[c]
10	7.3	3.0
1	9.3	4.2
10^{-1}	10.1	4.8
10^{-2}	10.1	5.0
10^{-3}	10.0	5.0
10^{-4}	10.1	5.1
10^{-7}	9.9	5.1
10^{-20}	9.6	5.2
10^{-60}	9.4	5.0
Significant difference	>0.44	>0.3

[†] Given are percentage points by which PHD improved when using iterated PSI-BLAST over single sequence-based predictions (three iterations). The iteration parameter determines which proteins to include when building the profile used for the next iteration step. Data set as in Table I.

[a] Maximum E-value of sequences used in refinement of position specific scoring matrix for PSI-BLAST (final alignment maximum E-value is set to 1).

[b] Secondary structure prediction accuracy (PHDsec) improvements as given by an increase in the three-state per-residue accuracy Q_3.

[c] Solvent accessibility prediction accuracy (PHDacc) improvements as given by an increase in the two-state per-residue accuracy Q_2.

Accuracy Was Stable Over a Wide Range of Iteration Parameters

It is "surprising that prediction accuracy on PSI-BLAST alignments was not very sensitive to the choice of the E-value limiting inclusion of sequences into the position-specific scoring matrix during iteration. Only rather extreme values were clearly worse (Table III). In contrast, the number of iterations appeared more crucial: When iterating more than three times, accuracy dropped significantly when using a permissive h-parameter (proteins included when E-value $< 10^{-4}$) and decreased slightly when using a restrictive h-parameter (10^{-10}; Table IV). Finally, we investigated the influence of gap parameters and substitution matrices. In particular, we found that gap open values of 10–12 did not change accuracy and that predictions were similar when replacing the default BLOSUM62[60] matrix by BLOSUM80 or BLOSUM45 matrices (data not shown).

TABLE IV. Improvement by Using PSI-BLAST With Different Numbers of Iterations[†]

Number of iterations[a]	h 10^{-4}[b]		h 10^{-10}[b]	
	Filtered[c]	Nonfiltered[c]	Filtered[c]	Nonfiltered[c]
1	9.5	9.7	9.5	9.7
2	9.9	10.0	9.8	10.0
3	10.1	9.8	10.1	10.0
4	9.6	9.3	10.1	9.8
6	9.3	8.8	9.9	9.7
10	8.1	7.4	9.7	9.5
Significant difference	>0.44	>0.44	>0.44	>0.44

[†] Given are percentage points by which PHDsec improved when using iterated PSI-BLAST over single sequence-based predictions (three-state per-residue accuracy, Q_3). For all runs, we included all proteins found below E-values of 1 in the final iteration. Data set used as in Table I.

[a] Number of PSI-BLAST iterations.

[b] PSI-BLAST iteration parameter (h-value) set to 10^{-4} (10^{-10}), that is, only sequence with E values $<10^{-4}$ ($<10^{-10}$) were considered when compiling the profile.

[c] Filtered refers to filtering the database used for the search (Materials and Methods); nonfiltered refers to not filtering the database.

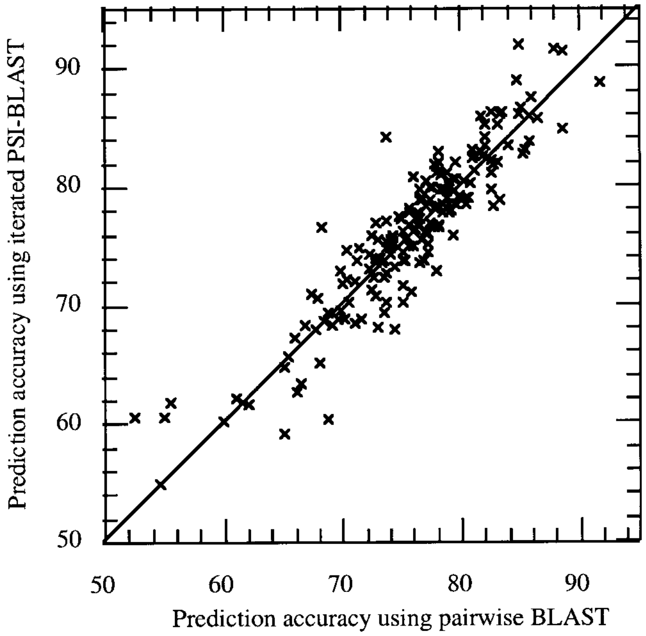

Fig. 2. Iterated PSI-BLAST versus pairwise BLAST. All searches against BIG with an E-value threshold of 10^{-4} for the iteration and a threshold of 1 for including the proteins into the final family. Proteins for which PSI-BLAST yielded better predictions than BLAST fall above the diagonal. Obviously, predictions based on iterated PSI-BLAST searches were not consistently more accurate than those based on iterated BLAST searches.

Prefiltering Database Was Not Vital for PSI-BLAST

It is surprising that we could not establish that filtering the database for PSI-BLAST, as proposed by David Jones,[44] was crucial for secondary structure prediction. Although the tendency was that filtering the database improved, the improvement was not significant (Table IV). Nevertheless, the numbers for the unfiltered database were consistently lower for all the E and h parameters (Table IV).

PSI-BLAST Versus BLAST: Combination Could Be Best

We found, on average, 2.4 times more proteins by PSI-BLAST than by BLAST. This family growth was similar to that obtained by switching from SWISS-PROT to BIG; the improvement was not (Table I; Table III). Furthermore, for many proteins, BLAST yielded better predictions than PSI-BLAST (Fig. 2). If we could decide from looking at the BLAST and PSI-BLAST alignments, which one is "better," accuracy would increase by an additional percentage point (data not shown). Possibly, we could approach this improvement by a more elaborated protocol for running PSI-BLAST.

Prediction Accuracy Related to Number of Sequences in Alignment

The 25% of proteins with highest prediction accuracy had, on average, twice as many proteins in their alignments as the lowest scoring 25%. For small families (<10 proteins), prediction accuracy improved about 5 percentage points over single sequences; for large families (>100 proteins), >11 percentage points. Qualitatively, this dependence was apparent when plotting the improvement versus the number of proteins aligned ["set_1136"; Fig. 3(A)]. Noticeably, the most significant gain resulted from adding a few proteins to the alignment [steep slope for small numbers in Fig. 3(A)]. For large families, the improvement appeared to saturate. However, the number of proteins was not a perfect indicator of prediction accuracy because the slopes differed between different methods and databases [Fig. 3(A)]. In particular, the data plotted in Figure 3A appeared to suggest that BLAST searches against SWISS-PROT yielded more improvement than BLAST or PSI-BLAST searches against BIG. However, the average was not compiled over the same families because BLAST found fewer homologues in SWISS-PROT than in BIG. We corrected for this by labeling the families according to the number of SWISS-PROT proteins in each alignment [Fig. 3(B)]. This revealed that for any family size (i) searches against BIG were better than against SWISS-PROT and (ii) PSI-BLAST performed better than BLAST. Despite the saturation for large families, we observed improvements from adding sequences even for family sizes between 200 and 500 (largest family found in SWISS-PROT).

Larger Alignments Improve Accessibility Prediction Marginally

Solvent accessibility predictions did not improve as much as secondary structure predictions by the various protocols we investigated. The two-state accuracy (see Materials and Methods) increased by <1 percentage point when using a larger database. It is interesting that this improvement was more sensitive to the particular threshold used to include sequences (Table II). Hence, predictions of accessibility appeared more sensitive to false positives than did predictions of secondary structure. It

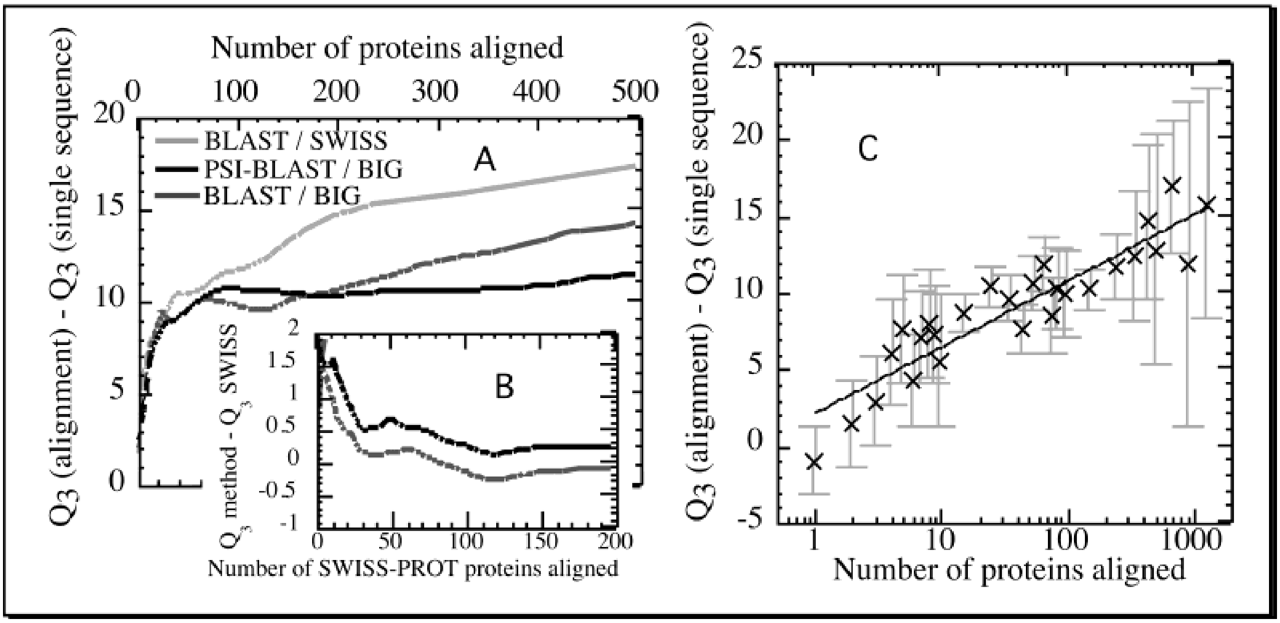

Fig. 3. Prediction accuracy versus family size. The improvement given as differences between predictions based on alignments and predictions based on single sequences. Curves represented an average weighted fit. Different methods produced somewhat different shapes for the fit (**A**). SWISS-PROT based alignments were on average most efficient. On the other hand, alignments produced on BIG by both BLAST and PSI-BLAST improved over SWISS-PROT for all sizes. **B:** Families labeled by number of SWISS-PROT homologues found. Prediction accuracy improved most significantly for small families and saturated for very large families. The average variation of the prediction accuracy was considerable (standard deviation of about 8). Thus, the number of aligned sequences was not a very good indicator of prediction accuracy for any given sequence. Nevertheless, the average trend was obvious. **C:** Predictions based on BLAST searches against BIG; note the straight line indicates a logarithmic fit; error bars correspond to 95% confidence intervals.

was not surprising then that using PSI-BLAST did not improve accessibility prediction either (Table III).

DISCUSSION

PSI-BLAST Better Than BLAST?

The average growth of the family size identified by PSI-BLAST was comparable to the growth achieved by searching through a larger database with a simple pairwise BLAST (factor 2.4 vs 2.7). However, the resulting gain in performance was significantly smaller for PSI-BLAST. Proteins added to the families by PSI-BLAST were more distantly related to the query sequence than those found by BLAST. Thus, we anticipated a larger number of false positives for the PSI-BLAST families. When we based the prediction only on the remote homologues exclusively identified by PSI-BLAST, accuracy improved over single sequence-based predictions overall by about 6 percentage points (data not shown). This value was about 3 percentage points lower than the gain through using pairwise BLAST on BIG (Table I). Partially, this might be explained by the different distributions of family sizes between the different methods and databases (Fig. 4). For example, about 40% of the PSI-BLAST searches on BIG added <10 proteins to the family identified by pairwise BLAST, whereas only 20% of the families had <10 proteins using pairwise BLAST searches. The most important increase in prediction accuracy resulted from the first 10 proteins included in an alignment [Fig. 3(A)]. However, PSI-BLAST found only 4% more families with >10 members than BLAST (Fig. 4). This suggested the following explanation for the relatively small improvement through PSI-BLAST. PSI-BLAST needs a reasonable number of first iteration hits (hits from pairwise BLAST) to find additional family members. However, when this number is sufficiently large to unravel the full “power” of the PSI-BLAST search, we reached the saturation of family sizes that improved prediction accuracy. In contrast, when we relaxed the stringent criteria for iterating PSI-BLAST (more iterations, lower h-parameter), prediction accuracy dropped. This could be explained by the effects of “drift” and “pollution” associated with PSI-BLAST[61]: the final profile is not centered around the original search sequence (drift) for which secondary structure is predicted, and many profiles contain proteins of different structure (pollution) than the one predicted.

Is Waiting for Databases to Grow More Successful Than Developing New Methods?

Overall, our combination of PSI-BLAST and larger databases improved our decade-old prediction method PHD by about 3 percentage points over the previous protocol of using dynamic programming (MaxHom on SWISS-PROT). Because we were not aware of making use of any particular feature of PHD to reach this value, we expect that similar improvements could be obtained for any old prediction method. It is surprising that most of this improvement resulted from the growth of the databases (BIG vs SWISS-PROT) and not from better search meth-

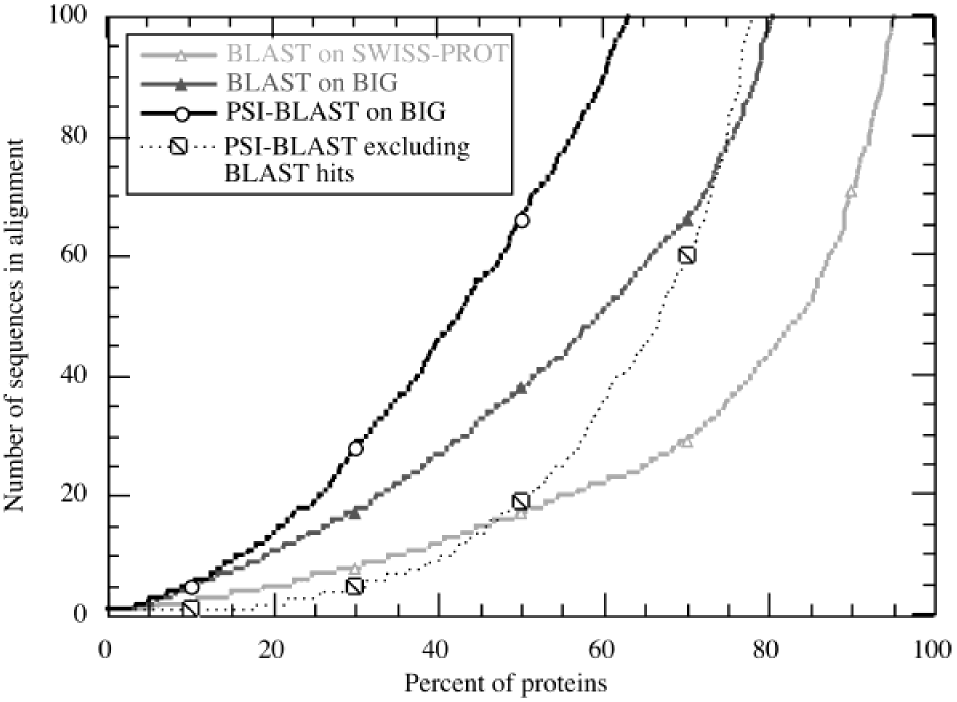

Fig. 4. Distributions of family size. Here, we focused on families with fewer than 100 homologues; for these prediction accuracy increased most markedly (Fig. 3). Understandably, PSI-BLAST did not identify many more homologues than pairwise BLAST for small families. Supposedly, this caused the observation that prediction accuracy differed only marginally between PSI-BLAST and BLAST searches. However, around family sizes of 50, PSI-BLAST added as many family members as BLAST identified. Furthermore, although about 40% of the PSI-BLAST families had >100 homologues, only 5% of the pairwise BLAST families found as many homologues when searching against SWISS-PROT.

ods (PSI-BLAST vs BLAST). Another surprise to us was the success of the popular BLAST algorithms in comparison with more CPU intensive dynamic programming searches (Table I). Could we gain further by retraining on larger databases and on PSI-BLAST profiles? Contrary to earlier claims,[42] PHD did profit substantially from extended profiles and did even perform on par with JNet/JPred2 developed on large data sets of extended PSI-BLAST profiles.[42,57,59] In contrast, PSIPRED[44] and PROFsec (B. Rost, unpublished) reached a sustained level about 1.5 percentage points above this value on a data set used by EVA (data not shown).[57,59] The better performance of PROFsec resulted entirely from improving the method. However, for the case of PSIPRED, we suspect that part of the better performance of the public PSIPRED server[62] resulted from a better protocol in running PSI-BLAST. In fact, when we used our PSI-BLAST alignments for our local version of PSIPRED, accuracy dropped significantly with respect to the server results.

How Fatal Are Errors in the Database?

We assume that SWISS-PROT contains fewer errors than TrEMBL. Could we see this effect in the accuracy of secondary structure prediction? Too many overlapping effects prevented a conclusive answer to this question. Figure 3(A) appeared to suggest that SWISS-PROT improved prediction accuracy more than TrEMBL. However, this observation might have been caused mainly by the overlying saturation effect. Hence, we shall have to wait for SWISS-PROT to double before we can answer more precisely.

Will Accuracy Rise With Future Database Growth?

A roughly sixfold growth of the database (SWISS-PROT vs BIG) improved prediction accuracy by about 1.5 percentage points. However, the increase saturated for families with more than 100–200 proteins (Fig. 3). Hence, will future growth improve performance considerably? We anticipate an affirmative answer, because <30% of all families contain >100 proteins through pairwise alignments (Fig. 4). Our assumption here is that newly sequenced proteins will not differ considerably from the proteins we already know from projects sequencing entire organisms.

CONCLUSIONS

Recent improvements in secondary structure prediction seem to be due to various sources. Our results indicated the following oversimplified formula. More than half of the recent improvement of secondary structure prediction resulted from the growth of sequence databases (from 90 K in SWISS-PROT to 500 K in BIG). Less than one fifth of the improvement was achieved through better database search methods (PSI-BLAST over BLAST). The remaining one third of the improvement was due to better methods. Hence, the most crucial tool to improve secondary structure prediction proved to be the new BLAST/PSI-BLAST searching tools. The contribution of PSI-BLAST was probably smaller than usually assumed due to saturating nature of the dependence of prediction accuracy on alignment size (Figs. 3 and 4). Most of the proteins did not reach this saturation point yet. Hence, we anticipate that prediction accuracy will rise continuously with every protein added to the databases.

ACKNOWLEDGMENTS

We thank Jinfeng Liu (Columbia) for computer assistance and the BLAST team at NCBI for assistance. DP and BR acknowledge support from the National Institutes of Health grant RO1-GM63029-01. We also thank Amos Bairoch (Geneva), Rolf Appweiler (EBI, Hinxton), Phil Bourne (San Diego) and their teams for maintaining vital public databases, and to all structural biologists who make the fruits of their efforts publicly available.

REFERENCES

1. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;266:460–480.
2. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383–402.
3. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
4. Berman HM, Westbrook J, Feng Z, Gillliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
5. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–539.
6. 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.
7. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
8. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
9. Benner SA, Gerloff D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv Enzyme Regul 1991;31:121–181.
10. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
11. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
12. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
13. Dickerson RE, Timkovich R, Almassy RJ. The cytochrome fold and the evolution of bacterial energy metabolism. J Mol Biol 1976;100: 473–491.
14. Maxfield FR, Scheraga HA. Improvements in the prediction of protein topography by reduction of statistical errors. Biochemistry 1979;18:697–704.
15. Sweet RM. Evolutionary similarity among peptide segments is a basis for prediction of protein folding. Biopolymers 1986;25:1565–1577.
16. Crawford IP, Niermann T, Kirchner K. Prediction of secondary structure by evolutionary comparison: application to the a subunit of tryptophan synthase. Proteins 1987;2:118–129.
17. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using alignment of homologous sequences. J Mol Biol 1987;195:957–961.
18. Rost B, Sander C. Jury returns on structure prediction. Nature 1992;360:540.
19. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
20. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
21. Defay T, Cohen FE. Evaluation of current techniques for ab initio protein structure prediction. Proteins 1995;23:431–445.
22. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
23. Mehta PK, Heringa J, Argos P. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci 1995;4:2517–2525.
24. Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. J Mol Biol 1995;247:11–15.
25. Di Francesco V, Garnier J, Munson PJ. Improving protein secondary structure prediction with aligned homologous sequences. Protein Sci 1996;5:106–113.
26. Riis SK, Krogh A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J Comp Biol 1996;3:163–183.
27. Levin JM. Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng 1997;10:771–776.
28. Benner SA, Badcoe I, Cohen MA, Gerloff DL. Bona fide prediction of aspects of protein conformation. J Mol Biol 1994;235:926–958.
29. Hubbard TJP. Use of β-strand interaction pseudo-potential in protein structure prediction and modelling. In: Hunter L, editors. 27th Hawaii International Conference on System Sciences. Maui, Hawaii: IEEE Society Press; 1994. p 336–344.
30. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 1994;33:3038–3049.
31. Persson B, Argos P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 1994; 237:182–192.
32. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
33. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J Mol Biol 1994;238:682–692.
34. Gerloff DL, Chelvanayagam G, Benner SA. A predicted consensus structure for the protein-kinase c2 homology (c2h) domain, the repeating unit of synaptotagmin. Proteins 1995;22:299–310.
35. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 1995;4:1618–1631.
36. Rost B, Casadio R, Fariselli P, Sander C. Prediction of helical transmembrane segments at 95% accuracy. Protein Sci 1995;4:521–533.
37. Persson B, Argos P. Topology prediction of membrane proteins. Protein Sci 5:363–371.
38. Rost B, Casadio R, Fariselli P. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5: 1704–1718.
39. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
40. Cserzö M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane a-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 1997;10:673–676.
41. Sonnhammer ELL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In: editors. Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB98); 1998. p 175–182.
42. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
43. Levin JM, Pascarella S, Argos P, Garnier J. Quantification of secondary structure prediction improvement using multiple alignment. Protein Eng 1993;6:849–854.
44. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.

45. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
46. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4690.
47. Schneider R. Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen. PhD: Univ. of Heidelberg, 1994.
48. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Submitted for publication.
49. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
50. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
51. Lupas A. Coiled coils: new structures and new functions. Trends Biochem Sci 1996;21:375–382.
52. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 235:13–26.
53. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
54. 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.
55. 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.
56. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
57. Eyrich WWWV, Martí-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. http://cubic.bioc.columbia.edu/eva/.
58. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
59. 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. Forthcoming.
60. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
61. Rost B. Protein secondary structure prediction continues to rise. Structural Bioinformatics. Forthcoming.
62. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
63. McLachlan AD, Staden R, Boswell DR. A method for measuring the non-random bias of a codon usage table. Nucleic Acids Res 12:9567–9575.