Improving protein secondary structure prediction with aligned homologous sequences

V. DI FRANCESCO,¹ J. GARNIER,² AND P.J. MUNSON¹

¹ Analytical Biostatistics Section, Laboratory of Structural Biology, Division of Computer Research and Technology,

National Institutes of Health, Bethesda, Maryland 20892-5626

² Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892-5626

(RECEIVED August 22, 1995; ACCEPTED October 31, 1995)

Abstract

Most recent protein secondary structure prediction methods use sequence alignments to improve the prediction quality. We investigate the relationship between the location of secondary structural elements, gaps, and variable residue positions in multiple sequence alignments. We further investigate how these relationships compare with those found in structurally aligned protein families. We show how such associations may be used to improve the quality of prediction of the secondary structure elements, using the Quadratic-Logistic method with profiles. Furthermore, we analyze the extent to which the number of homologous sequences influences the quality of prediction. The analysis of variable residue positions shows that surprisingly, helical regions exhibit greater variability than do coil regions, which are generally thought to be the most common secondary structure elements in loops. However, the correlation between variability and the presence of helices does not significantly improve prediction quality. Gaps are a distinct signal for coil regions. Increasing the coil propensity for those residues occurring in gap regions enhances the overall prediction quality. Prediction accuracy increases initially with the number of homologues, but changes negligibly as the number of homologues exceeds about 14. The alignment quality affects the prediction more than other factors, hence a careful selection and alignment of even a small number of homologues can lead to significant improvements in prediction accuracy.

Keywords: prediction; protein secondary structure; sequence alignments

For several decades, the protein folding problem has been among the most challenging problems in the biological sciences. In 1994, a bona fide protein structure prediction contest was organized with the aim of assessing the real virtues and defects of “state-of-the-art” methodologies. Analysis of the structures predicted by the contestants (Moult et al., 1995) has generally shown that even the most promising techniques need considerable improvement, and that the protein folding problem should still be considered unresolved. Briefly, *ab initio* calculations, although promising, are feasible only for small-size proteins; there have been no major breakthroughs in the molecular modeling techniques, and threading techniques need further development. Protein secondary structure prediction was reevaluated and recognized as a useful tool for establishing starting points for tertiary structure calculation of protein structures.

Early approaches to protein secondary structure prediction from the primary sequence had a prediction accuracy Q3 (percentage of correctly predicted residues in the three states: α -

helix, β -strand, and coil) of about 57% (Chou & Fasman, 1978; Garnier et al., 1978). Various later attempts to improve the accuracy (Gibrat et al., 1987; Biou et al., 1988; Levin & Garnier, 1988; Holley & Karplus, 1989; Qian & Sejnowski, 1989; King & Sternberg, 1990; Salzberg & Cost, 1992; Stolorz et al., 1992; Zhang et al., 1992; Munson et al., 1994) with innovative artificial intelligence techniques, such as neural networks, machine learning, nearest neighbors, and combined approaches, have not achieved prediction accuracies greater than 66%. The inclusion of evolutionarily related sequences into the prediction scheme has given a significant boost in prediction accuracy, up to values of about 68–72% (Zvelebil et al., 1987; Levin et al., 1993; Rost & Sander, 1993, 1994; Rost et al., 1994a; Di Francesco et al., 1995; Salamov & Solovyev, 1995). In general, the suggested explanation for these improvements in prediction accuracy is that sequence alignments of homologous proteins should emulate as closely as possible the structural alignment. Thus, aligned residues, in particular those in the core of proteins, should belong to the same secondary structure elements. Sequence alignments may be utilized to obtain a consensus from the predictions based on each homologous sequence, or they may be used to build sequence profiles at each aligned position. In addition to the

Reprint requests to: Valentina Di Francesco, NIH/DCRT/LSB, Analytical Biostatistics Section, 12 South MSC 5626, Bldg. 12A – Room 2041, Bethesda, Maryland 20892-5626; e-mail: valedf@helix.nih.gov.

identity of the aligned residues (a feature exploited by all the predictive schemes), other information is available from sequence alignments, such as the location of gaps or the patterns of residue mutation in the aligned protein families. Some authors have used such information to refine their prediction models (Zvelebil et al., 1987; Rost & Sander, 1993; Rost et al., 1994a; Salamov & Solovyev, 1995).

However, the reasons why the inclusion of this additional information improves the quality of the prediction have not been understood. Here we perform an exploratory analysis on a database of aligned sequences to find and understand the relationships between the location of secondary structure elements and the position of insertions or deletions (frequently called indels) and of conserved residue patterns in multiple sequence alignments and to investigate how these relationships compare with those found in structurally aligned protein families. We also investigate how such relationships may be used to improve the quality of the prediction of the secondary structure elements. Furthermore, we analyze the extent to which the number of homologous sequences influences the quality of the prediction. In our investigations, we used the Quadratic-Logistic (QL) method (Munson et al., 1994) as a predictive scheme.

Results and discussion

Table 1 shows the prediction accuracies of the QL method, measured as percentage of correctly predicted residues in the three structural states (Q3), and the improvements obtained by the use of multiple sequence alignments. Sequence alignments were incorporated into the model by calculating the profiles or relative frequency of occurrence of each residue at each position. With database I (see the Methods), the use of profiles increases the accuracy of QL; the improvement in Q3 is only 2.8% for uncrossvalidated results, whereas it is 5.0% for crossvalidated data. Different secondary structure prediction methods attain different improvements in accuracy by the use of profiles alone. Using the same sequences and alignments as in the present study, Rost and Sander (1993) obtain a 6.8% improvement in prediction accuracy with neural networks trained with a balanced database, whereas Salamov and Solovyev (1995) obtain only 3.3% by an explicit use of multiple sequence alignments. The improvement

in Q3 obtained by QL just with the use of profiles falls within the range of improvements obtained by these two groups. QL-crossvalidated results were obtained by sevenfold cross-validation, which consists of dividing the database of protein sequences into seven different training and test sets, so that each protein is included in one of the test sets only once. This type of crossvalidation is preferred because it is less computationally intensive than the jackknife. Moreover, in earlier tests of QL prediction accuracy, jackknife and sevenfold crossvalidations gave results that differed by less than 0.1%. Crossvalidated prediction accuracy ordinarily drops relative to the uncrossvalidated numbers. Using the test sequence only, without profiles, the drop in prediction accuracy is 5.4%. When profiles are used, this decrease is reduced to 3.2%, suggesting that the profiles effectively stabilize the parameter estimates.

By comparing the sensitivity indices (the ratio of the number of correctly predicted residues to those observed) obtained with and without sequence alignments, it is possible to analyze the effects of the use of profiles on the prediction accuracy. Table 2 shows that the prediction quality for all three conformational states takes advantage of the use of sequence alignments, in particular the sensitivity for both helical and extended residues is increased by 7%. For helices, the improvement in sensitivity is derived mainly from a reduction of the number of helical residues being predicted as coil (5%). For strands, the improvement is derived mainly from a reduction of the number of extended residues wrongly predicted as helices (6%).

Use of variability score

Figure 1 shows the distribution of the three observed structural states of the residues by variability score categories (*Varsc*). Note that the proportion of helical residues increases with increasing values of *Varsc*, whereas the proportion of coil residues decreases. Twenty-five percent of highly conserved residue positions (*Varsc* = 0) are helical in conformation and 53% are coil residues. The proportion of helical residues increases to 41% for highly variable residue positions ($0.6 \leq \text{Varsc} < 1$) exceeding the coil proportion (36%). The proportion of extended residues remains quite stable: between 22% and 25% across the *Varsc* cat-

Table 1. Prediction accuracy (and improvements) using various factors with the QL method

Calibration database ^a	Calibration based on	Prediction database ^a	Prediction based on	% Prediction accuracy Q3 uncrossvalidated	% Prediction accuracy Q3 crossvalidated
I	Test sequence	I	Test sequence	67.8 (0.0)	62.4 (0.0)
I	Profiles	I	Profiles	70.6 (+2.8)	67.4 (+5.0)
I	Profiles	I	Profiles + <i>Varsc</i>	70.9 (+3.1)	67.7 (+5.3)
I	Profiles	I	Profiles + <i>Indels</i>	71.2 (+3.4)	68.1 ^b (+5.7)
I	Profiles	I	Profiles + <i>Varsc</i> + <i>Indels</i>	71.5 (+3.7)	68.5 ^b (+6.1)
I	Profiles	II	Test sequence		62.3 (0.0)
I	Profiles	II	Profiles		68.6 (+6.3)
I	Profiles	II	Profiles + <i>Varsc</i>		68.5 (+6.2)
I	Profiles	II	Profiles + <i>Indels</i>		69.0 (+6.7)

^a See the Methods: The databases.

^b These results are crossvalidated with respect to the QL parameters, but not with respect to parameters associated with *Varsc* or *Indels*.

Table 2. Sensitivity indices for prediction of three conformations

Predicted secondary structure state					
Observed secondary structural state	Helix	Strand	Coil	Total	Total
					no. of residues
Test sequence only ^a					
Helix	59	10	31	100	6,436
Strand	20	43	37	100	4,629
Coil	16	10	74	100	9,965
Total number of residues	6,295	3,646	11,089		21,030
Profiles ^b					
Helix	66	8	26	100	6,436
Strand	14	50	36	100	4,629
Coil	14	9	77	100	9,965
Total number of residues	6,246	3,733	11,051		21,030

^a Prediction with QL, calibrated using test sequences only in database I, without *Varsc* and *Indels*; sevenfold crossvalidation results.

^b Prediction with QL, calibrated using profiles in database I, without *Varsc* and *Indels*; sevenfold crossvalidation results.

egories. In positions having indels, the proportions of the three states shows no simple trend across the four categories (Fig. 2). These observations suggest that there is a strong and statistically significant association between structural states and residue variability in aligned positions. To ascertain that our result did not depend upon the particular definition of variability score, we repeated the analysis using a different variability measure. We computed the sum of pairwise similarity scores for each position in the alignment, based on the mutation matrix PAM250

Fig. 1. Distribution of structural states by variability score value (*Varsc*) at residue positions without insertions or deletions in multiple sequence alignments for proteins in database I (18,214 residues of 21,030). The variability score was broken into four roughly equal categories. The proportion of helical residues rises with increasing variability score values, from 25% (*Varsc* = 0) to 41% ($0.6 \leq \text{Varsc} < 1$). The proportion of coil residues decreases with increasing variability score values, from 53% (*Varsc* = 0) to 36% ($0.6 \leq \text{Varsc} < 1$). The proportion of extended residues ranges between 22% and 25%. The association between secondary structural state and residue variability in aligned positions with no indels is strongly significant (Pearson's chi square test = 308, $p < 0.001$).

Fig. 2. Distribution of structural states by variability score value (*Varsc*) at residue positions with insertions or deletions in multiple sequence alignments for proteins in database I (2,816 residues of 21,030). The proportion of helical residues ranges between 9% and 11% across the *Varsc* categories, the proportion of extended residues ranges between 8% and 15%, that of coil residues ranges between 76% and 81%. The association between secondary structural state and residue variability in aligned positions remains moderately significant (Pearson's chi square test = 31, $p < 0.001$).

(Dayhoff, 1978). This second variability measure is highly correlated to *Varsc* (correlation coefficient $\rho = -0.67$). This new analysis confirmed that, in residue positions having no gaps, the proportion of helical residues increases monotonically with increasing variability in aligned residue positions, whereas the proportion of coil residues decreases (data not shown).

From the analysis of structural alignments, it has been observed that sequence variability is most frequent in regions that do not belong to the cores of aligned proteins. These regions usually consist of peripheral elements of secondary structure, such as ends of helices or strands at the edges of β sheets, or loops between major secondary structure elements, which might contain helices and strands (Chothia & Lesk, 1986). Our study shows that, in multiple sequence alignments, highly variable residue positions, indicated by large values of *Varsc*, are more frequent in helical residues. Apart from inaccuracies introduced by sequence alignments (see below), this finding might be explained by noting that helices are often amphipathic, containing many solvent-exposed residues, which are more likely to undergo mutations.

Table 1 shows the prediction accuracy of QL, utilizing the variability score factor. We obtain only a small improvement in Q3 (0.3% in database I) over the use of profiles alone. A similar result was obtained by Rost (Rost & Sander, 1993; Rost et al., 1994a): a Q3 improvement of only about 0.5% using a different definition of the conservation score. This is probably due to the fact that residue variability is already embedded and almost fully exploited in the profiles of the aligned residue positions. Zvelebil et al. (1987) obtained a much greater improvement in prediction accuracy (4.8%) using a conservation score based on physical/chemical properties of the residues. This seeming contradiction is resolved by noting that this method began at a lower accuracy rate (Q3 = 61.3%) and therefore had greater room for improvement than did the QL method in our tests. Their greater improvement may also be ascribed to the fact that they were using a fairly small database of only 11 protein families. Benner's group (Benner et al., 1994; Jenny & Benner, 1994; Jenny et al., 1995)

takes advantage of the patterns of residue conservation within aligned subfamilies, but they do not explicitly define a conservation/variability score. This makes it difficult for us to quantify and compare the effects of such patterns on the predictions.

Use of insertions and deletions

Figure 3 shows the distribution of structural states at residue positions where there are insertions, deletions, or no indels in the multiple sequence alignments. The association between structural states and indels positions is highly statistically significant. From the dramatic increase in the frequency of coil residues, it is clear that, in multiple sequence alignments, insertions and deletions favor the observed coil state. This is in agreement with the observed association between the location of indels and secondary structure from structural alignments (Greer, 1981, 1990; Chothia & Lesk, 1986; Lesk, 1991). One might expect the proportion of each structural state to be the same at residue positions where either insertions or deletions occur. Figure 3 shows that, instead, these proportions change. This asymmetry is probably due to the fact that when an insertion occurs in the aligned sequences, the number of residues in the insert are not included in the proportions of the three states, whereas each deleted residue is counted.

The effect of extending the QL profile with the *Indels* factor (see the Methods) is shown in Table 1. We see an improvement of 0.6–0.7% over the use of profiles alone, using database I, or 0.4% using database II. Although indels are informative about the structural state, 79% of indels are in regions already predicted as coil in database I, thus vitiating the effectiveness of this factor.

Validation database

Because our use of indels and variability scores in the prediction method was developed after the predictions for database I were made, we could not easily crossvalidate the prediction ac-

curacy improvements. We therefore used an additional test set of 20 proteins (database II) that are not homologous to those of the database I, to check whether our findings could be confirmed. As we saw in the previous section, biasing the prediction in favor of the coil state at gap positions improved the overall prediction accuracy for database II (Table 1). Variability scores were included in the method by adjusting the estimated probabilities for the three structural states by factors determined from database I (see the Methods: Use of variability score), but resulted in a slight loss in prediction accuracy. In summary, the greatest improvement in accuracy is due to profiles from the multiple sequence alignment, followed by the inclusion of coil propensities for gaps. The fact that we obtain a higher prediction accuracy with database II (Table 1) is probably due both to the small number of proteins in this set and its higher proportion of coil residues (see the Methods), which are known to be more accurately predicted.

Effects of the number of homologous sequences

Analysis of the prediction accuracy by protein in database II suggests that greater numbers of homologous sequences improve the prediction accuracy. In particular, we previously observed (Di Francesco et al., 1995) that, by considering only those proteins with more than seven homologues (12 proteins; about 60% of the residues of database II), the improvement in Q3 due to profiles was 7.8%, whereas on the average it was 6.7%. Previously we had not used all the available homologues in the *hssp* files (Schneider & Sander, 1991), but only a selection of them (see the Methods: The databases). Here we produce new predictions with all available homologues and the alignments found in the *hssp* files. Table 3 shows the prediction accuracies in database II only for those proteins whose homologous sequences were previously selected. Although the average number of homologues increased almost threefold, from 14 to 36, the prediction accuracy did not change significantly, suggesting that a large number of homologous sequences is not essential. How large is "large" remains to be established: in our case, the choice of seven homologues was purely empirical and looked like a reasonable number. The rationale we used in the selection of the homologues was as follows. The diversification of sequence identity of the homologues to the test sequences seems to have been appropriate and not damaging to the quality of the prediction at all. No specific relationship between the number of homologues and the improvement in prediction quality has been formulated, just a general, positive trend with a diminishing return as the number of homologues increases.

The quality of the prediction and the contribution to the improvement in accuracy of all the factors discussed here is seriously affected by the quality of the alignment, which in our opinion is the most important factor of all, and it is probably one of the reasons why a large variation in the percentage improvement of prediction accuracy is observed among different protein families. To support this thesis, a very enlightening example is given by the secondary structure prediction of a beta-sheet domain of 6-phospho-beta-D-galactosidase (pbdg). Pbdg was one of the target proteins listed in the bona fide structure prediction contest (Moult et al., 1995). The structure of the protein is now known to be a TIM-barrel, with an inserted beta domain consisting of four successive strands between the sixth strand and the helix preceding the seventh strand of the barrel.

Fig. 3. Distribution of structural states at residue positions without and with insertions or deletions in the multiple sequence alignments for proteins in database I. The proportion of helical residues decreases from 33% (no indels position) to 15% if deletions are present, or to 5% if insertions are present. Likewise, strand proportion also decreases from 24% to 14% if deletions occur, or to 6% if insertions occur. Coil proportion increases from 43% to 71% if deletions are present, or to 89% if insertions are present. The association between secondary structural state and presence of indels in the aligned positions is significant (Pearson's chi square test = 1,312, $p < 0.001$).

Table 3. Prediction accuracy versus number of homologues

Protein PDB name	Length	Test sequence only – Q3 (%)	Number of homologues	Profiles ^a Q3 (%)	Number of homologues	Profiles ^b Q3 (%)
lace	481	61.33	10	71.31	24	73.80
lgly	470	64.04	7	69.15	10	69.15
lhdd	57	71.93	15	77.19	50	78.95
lhrh	111	72.07	12	70.27	48	72.07
lnsb	390	55.38	14	67.44	29	66.67
2cpk	336	62.80	17	75.89	50	72.02
2pk4	80	86.25	24	88.75	29	87.50
5p21	166	66.87	16	65.66	50	66.27
Average		63.32	14	71.16	36	71.11

^a Prediction performed using a selected set of homologues to the test sequences. Alignments of the homologues taken from the *hssp* database.

^b Prediction performed without selection of homologues to the test sequences. Alignments of the homologues taken from the *hssp* database.

The PHD server (Rost et al., 1994b) returned a set of 22 homologues to the whole sequence that were used in the prediction. In particular, the PHD secondary structure prediction (Rost & Sander, 1993) of the beta sheet domain sequence fragment consisted mainly of coil residues, except for the first strand that was wrongly predicted as helix for most of its residues (Table 4). Our QL + Profiles and Benner's (Gerloff & Benner, 1995) predictions made the same mistakes: first strand mispredicted as helix and the remaining three strands predicted as coil. By submitting only the beta domain sequence fragment to the PHD server, just three homologues were found (two of those already belonged to the set of the 22 whole sequence homologues). The quality of the PHD prediction improved, correctly predicting the position of three of the four strands in the beta sheet domain. The QL prediction also changed when using the reduced set of homologues, and detected a signal for secondary structure (Table 4). However, many extended residues were predicted as helical residues. The high sensitivity of PHD for predicting extended residues has been documented previously ($Q_{\beta}^{\text{obs}} = 62\%$ (Rost et al., 1994)) and is seen here in the more accurate prediction in this beta sheet domain. The improvement in the prediction based on the reduced set of homologues is supported by the improved quality of the alignment (Table 4, Set B). The alignment of the 22 homologues to pbdg showed instead scattered irregular gaps in the beta domain region (Table 4, Set A). The misalignment of the numerous homologues clearly disrupts and washes out the information from the three close homologues to that sequence domain. This example supports the practice of revising and manually refining automated multiple sequence alignments and demonstrates the need to carefully select the homologues before proceeding with the prediction.

In summary, the use of profiles is the major factor improving prediction accuracy by 5–6%. The use of indels improves prediction accuracy by 0.4–0.7%, possibly due to the significant association between indels positions and coil residues. Residue variability appears to be, at best, a weak signal for secondary structure prediction (0–0.4% improvement), although helical residues remain strongly associated with highly variable regions. We have proposed an explanation for why variability score is an ineffective predictor: variability information can be derived

from profiles and may be redundant in some methods already utilizing profiles. The information in a multiple alignment for structure prediction does not appear to increase after about 14 homologues to the test sequence have been obtained. This suggests the existence of a plateau for prediction accuracy using sequence alignments. Finally, prediction accuracy is limited by the quality of the alignment. Careful attention to both selection of homologues and alignment techniques are required to obtain the highest prediction accuracy.

Methods

The QL method (Munson et al., 1994) uses a maximum-likelihood logistic regression, incorporating both the conventional “linear” effects of residues contained in a 17-residue window with the “quadratic” effect of all residue pair interactions within the same window. The likelihood is based on the trinomial probability model. The QL method produces probability estimates for each of three structural states, given the local sequence of residues:

$$\Pr(\alpha|\text{local seq.}) = \frac{z_{\alpha}}{1 + z_{\alpha} + z_{\beta}};$$

$$\Pr(\beta|\text{local seq.}) = \frac{z_{\beta}}{1 + z_{\alpha} + z_{\beta}};$$

$$\Pr(\text{coil}|\text{local seq.}) = \frac{1}{1 + z_{\alpha} + z_{\beta}},$$

where $z_{\alpha} = \exp(a_0 + a_{11}x_{11} + a_{21}x_{21} + \dots)$, $z_{\beta} = \exp(b_0 + b_{11}x_{11} + b_{21}x_{21} + \dots)$, and x_{ij} is an indicator variable for residue type i in position j , $i = 1 \dots 20$, $j = 1 \dots 17$. We may express z_{α} and z_{β} in matrix form as follows: $z_{\alpha} = \exp(a_0 + A'x)$, $z_{\beta} = \exp(b_0 + B'x)$, where the linear transformation of the data is more evident (A and B are vectors of parameters). In the QL method, the interaction terms of residue pairs in the 17-residue window are included by adding a quadratic term as follows: $z_{\alpha} = \exp(a_0 + A'x + x'Ax)$, $z_{\beta} = \exp(b_0 + B'x + x'Bx)$, where A and B are matrices of the quadratic form. The

Table 4. Beta-sheet domain of pbdg

Set A ^a	300	340
pbdg	YM SDWMQAFDGE TEIIHNGKGE KGSSKYQIKG VGRRVAPDYV	
lacg_staau	YM SDWMRGYDGE SEITHNATGD KGGSKYQLKG VGQREFDQDV	
lacg_lacca	YF SKWLRAYHGK SETIHNGDGT KGSSVARLQG VGEEKLPDGI	
bgla_clotm	YS SEFIK.YDPS SESGFSPANS ILEK.....F	
bgls_calsa	YT RA.VRLYDEN SSWIFPIRWE HPAGEY....	
bgla_bacci	YT SSMNRYNPGE AGGMLSSEAI SMGA.....	
ascb_ecoli	YA SRCASA....E MNANSSAAN VVKSLRNPYL	
bgla_bacpo	YS MSVNR.FNPE AGFLQSEEIN MG.....L	
arbb_erwch	YM TGCVTDEAQ LEKTRGN... ILNMVNPYL	
bgla_thema	YS GHLVKFDPDA PAKVSFVERDL	
bglb_ecoli	YM TGCVS....HD ESINKNAQGN ILNMIPNPHL	
myro_brana	YV TQYAKpyPSE THTAMMDAGV KLTVD...NS RGEFLGPLFV	
myr3_sinal	YV TQYAKpyPSE THTALMDAGV DLTFNNS...RGEYPGPV	
myro_arath	YV TQYAK...NN QTIVPSDVHT ALMDSRTTTL SKNATGHAPG	
bglt_trirp	YS SYAAKiPNA RPAIQTDSL I NATFEHNGKP LGPMAASSWL	
bgls_agrsp	YT P..MRVADDA TPGVEFFPATM PAPAUSDVK.....	
bglb_bacpo	YT RSIIRSTNDA SL.....LQVEQVHMEE	
bgls_trirp	YS SSIISnhGNA KPSYSTNPMT NISFEKHGIP LGPRAASIWI	
myr2_sinal	YV TQYAKPKPNP YPSEKHTAMD DAGVDLTFKN .SRGEYPGPV	
myr1_sinal	YV TQYAKpyPSE THTAMMDAGV DLTFKNS...RGEYPGPV	
lph_human	YT SRLISNAPQN TCI.....PSYDTIGG FSQHVNHVWP	
lph_rabit	YT SRLISKAPED SCI.....PSYDTIGG FSQHTDPAPW	
bglb_micbi	YS HVRLAAEGEP ANRLPGSEGI RFERPTAVTA WPGD.....	
Obs.structure ^b	CC CEEEECCCC CEEC***** **CCEEECC CEEEECCCC	
PHD prediction	HH HHHCCCCC CCCCCCCCC CCCCCCCCC CCCCCCCCC	
Benner's	CH HHHHHHHHH CCCCCCCCC CCCCCCCCC CCCCCCCCC	
QL+Profiles	CC HHHHCCCC CCCCCCCCC CCCCCCCCC CCCCCCCCC	
Set B ^c	300	340
pbdg	YM SDWMQAFDGE TEIIHNGKGE KGSSKYQIKG VGRRVAPDYV	
lacg_staau	YM SDWMRGYDGE SEITHNATGD KGGSKYQLKG VGQREFDQDV	
lacg_lacca	YF SKWLRAYHGK SETIHNGDGT KGSSVARLQG VGEEKLPDGI	
cadc_ecoli	HP LDEKQLAALN TeiVTLPELN NLSIIYQIKA VSALVK....	
Obs.structure ^b	CC CEEEECCCC CEEC***** **CCEEECC CEEEECCCC	
PHD prediction	CC HHHHHHCCCC EEEEECCCC CCCEEEEECC CCCEEECCCC	
QL+Profiles	CC HHHHHHCCCC HHEEECCCC CCCHHHHHC CCCCCCCCC	

^a Set of homologues and their alignment returned by the PHD server (Rost et al., 1994b), when the whole sequence of pbdg was submitted.

^b Observed structure obtained by DSSP (Kabsch & Sander, 1983). The * indicates that the conformation of these residues is unknown.

^c Set of homologues and their alignment returned by the PHD server (Rost et al., 1994b), when only the sequence fragment of the beta domain of pbdg was submitted.

inherent overparameterization of the quadratic model is effectively dealt with using penalized likelihood techniques, parameter selection, and imposition of certain reasonable constraints on the quadratic parameters. In particular, the periodic information inherent in the alpha helix (period 3.6) and beta sheet (period 2.0) is encoded into those constraints. Such quadratic or "pairwise" parameters have been shown (Gibrat et al., 1987; Munson et al., 1994) to significantly augment the predictive power of the linear, "directional" coefficients (Garnier et al., 1978).

Extending QL to incorporate aligned sequences can be done in one of two ways. One may first predict the structure of each homologue, then combine them, or, alternatively, one may first combine the homologous sequences and then make a single prediction. Aligned homologous sequences may be combined by calculating the profiles or relative frequency of occurrence of each amino acid at each position. The original QL method treats each position of the test sequence as a vector of 20 zero-one or

dummy variables (one for each amino acid). With homologous sequences, the inputs to QL are the profiles for each residue position, consisting of a sequence of 20 real numbers assuming values between zero and one, representing the proportions of each residue type (Table 5). The advantages of the QL method are that it uses a likelihood maximization algorithm to obtain statistically optimal parameter estimates and that the use of logistic models for categorical variables is based on a well-developed theory.

Variability score and secondary structure

Multiple sequence alignments provide information about the variability of the residue at each position. To investigate the relationship between patterns of residue conservation and secondary structural states in sequence alignments, we developed a variability score (*Varsc*), based on the variance of a multinomial distribution, as follows. For each residue position:

Table 5. Construction of profiles

Residue position	Homologous sequences ^a			
	1	2	3	4
1	E	E	E	D
2	G	R	R	R
3	M	F	C	A

x_{ij} Test sequence only

Residue position j	Residue type i															
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	

x_{ij} With homologous sequences

1	0	0	1/4	3/4	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	1/4	0	0	0	0	0	0	0	0	3/4	
3	1/4	1/4	0	0	1/4	0	0	0	0	0	1/4	0	0	0	0	

^a Test sequence in column 1.

$$Varsc = 1 - \frac{\sum_{i=1}^{21} n_i^2}{\left(\sum_{i=1}^{21} n_i\right)^2},$$

where n_i is the number of residues of type i . An example of calculation of the variability score is given in Table 6. Note that $0 \leq Varsc < 1$, $Varsc = 0$ for conserved residue positions and $Varsc$ approaches 1 for increasingly variable residue positions. The sum ranges to 21, because missing values (insertions or deletions in the homologous sequences) are dealt with as an additional category, along with the 20 residue types. By definition of $Varsc$, highly conserved residues ($Varsc = 0$) are only in positions where no insertions or deletions are present in the multiple sequence alignments. This definition of $Varsc$ does not take into account any evolutionary or structural similarity among residue types, but the definition can be extended to include the similarity factor among residues by weighting the terms of the sum appropriately. On the other hand, because $Varsc$ is based on the 20 residue types, it allows for all possible patterns of variation to be measured. To simplify the analysis of the relationship between secondary structure states and residue variability, the values of the variability score were divided into four different categories, two of which correspond to the extreme values of the score: zero for no mutations in that residue position, and values ≥ 0.6 for the residue positions that have undergone many mutations. To test the ability of residue variability for secondary structure prediction, we extended the logistic regression of QL with the variability score factor coded as a real number and not as a categorical variable.

Table 6. Calculation of the variability score

Residue position	Protein sequences				Variability score
	1	2	3	4	
1	E	E	E	E	$1 - (4/4)^2 = 0$
2	G	R	G	R	$1 - (2/4)^2 - (2/4)^2 = 1/2$
3	M	F	C	A	$1 - 4 * (1/4)^2 = 3/4$

Insertions/deletions and secondary structure

The crystal structures of protein families show that sequence insertions occur in loop regions between secondary structures (Greer, 1981, 1990; Chothia & Lesk, 1986; Lesk, 1991). To assess whether information about the location of insertions and deletions in multiple sequence alignments affects the prediction accuracy, we analyzed the frequencies of occurrences of insertions and deletions relative to each residue's particular secondary structural state. To take advantage of the observed high propensity for the coil state whenever a gap is present in the alignment, we introduced the *Indels* factor into the model, which consists of setting the estimated probability for the coil state to the value 0.5. This value was determined empirically using database I as the weight giving the highest improvement in the prediction accuracy. The same value was used in database II.

The databases

Database I

The distributions of the variability scores and of indels were analyzed on a set of proteins selected from the database of sequences used in (Rost & Sander, 1993). Test sequences and their homologues were taken from the *hssp* database (Schneider & Sander, 1991). We selected only those sequences with at least two sufficiently divergent homologues. This was necessary because we did not want to have an inflated number of apparently conserved residue positions when in fact there were no or few homologues. We selected a total of 115 polypeptide chains, with 21,030 residues. The DSSP assignments (Kabsch & Sander, 1983) were regarded as the observed secondary structure with the following modification: states H, I, G were assigned as helix (H); state E remained extended (E); and the remaining states were regarded as coil (C). The secondary structure residue composition is 30.6% helix, 22.0% strand, 47.4% coil. The Protein Data Bank designators for the proteins used in this database are: 1acx, 1ak3.A, 1azu, 1bbp.A, 1bds, 1bm.v.1, 1bm.v.2, 1cbh, 1cc5, 1cd4, 1cdt.A, 1crn, 1cse.I, 1eca, 1etu, 1fdl.H, 1fdx, 1fkf, 1fxi.A, 1gd1.O, 1gp1.A, 1hip, 1il8.A, 1l58, 1lap, 1lrd.3, 1mcp.L, 1mrt, 1ovo.A, 1paz, 1ppt, 1pyp, 1r09.2, 1rbp, 1rhd,

1s0l, 1sdh.A, 1sh1, 1tgs.I, 1tnf.A, 1ubq, 1wsy.A, 1wsy.B, 2aat, 2alp, 2cab, 2ccy.A, 2fnr, 2fxb, 2gcr, 2gls.A, 2gn5, 2hmz.A, 2i1b, 2ih4, 2lhb, 2ltm.A, 2ltm.B, 2mev.4, 2mhu, 2or1.L, 2pab.A, 2pcy, 2sns, 2stv, 2tgp.I, 2tsc.A, 2utg.A, 2wrp.R, 3ait, 3b5c, 3blm, 3cla, 3cln, 3ebx, 3gap.A, 3hmg.A, 3hmg.B, 3icb, 3pgm, 3rnt, 3tim.A, 4bp2, 4cms, 4cpv, 4fxn, 4gr1, 4pfb, 4rhv.1, 4rhv.3, 4rhv.4, 4rxn, 4sgb.I, 4ts1.A, 4xia.A, 5cyt.R, 5er2.E, 5hvp.A, 5ldh, 5lyz, 6acn, 6cpa, 6cts, 6dfr, 6hir, 6tmn.E, 7cat.A, 7icd, 7rsa, 8adh, 9api.A, 9api.B, 9ins.B, 9pap, 9wga.A.

Database II

To further validate the predictive power of our algorithm using variability scores and indels, we used a different database of 20 proteins that are less than 25% homologous to the proteins of the previous database. The secondary structure composition was 29.4% helix, 20.9% strands, 49.7% coil. The total number of residues was 3,988. We selected the homologues to these proteins primarily from the *hssp* database, and from Swiss-prot or GenBank. The rationale used in the selection of homologues was to obtain a nearly uniform spread of the percentage of identity to the test sequences, thereby avoiding the inclusion of a large number of nearly identical sequences. We used a homology threshold that depended on the length of the test sequence, as suggested (Schneider & Sander, 1991). We performed the multiple sequence alignment with CLUSTAL (Higgins & Sharp, 1989), with default parameter values for gap penalties. The proteins used in this test set are: 1ace, 1col, 1cox, 1dfn, 1f3g, 1gly, 1gmf, 1hdd, 1hrh, 1msb, 1nsb, 1pi2, 2cpk, 2hip, 2ifb, 2pk4, 2sar, 2scp, 3fgf, 5p21.

References

- Benner SA, Jenny TF, Cohen MA, Gonnet GH. 1994. Predicting the conformation of proteins from sequences. Progress and future progress. *Adv Enzyme Regul* 34:269–353.
- Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. 1988. Secondary structure prediction: Combination of three different methods. *Protein Eng* 2:185–191.
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* 5(4):823–826.
- Chou PY, Fasman GD. 1978. Empirical predictions of protein conformation. *Rev Biochem* 47:251–273.
- Dayhoff MO. 1978. *Atlas of protein sequence and structure*. Washington, DC: National Biomedical Research Foundation.
- Di Francesco V, Munson PJ, Garnier J. 1995. Use of multiple sequence alignments in protein secondary structure prediction. *28th Hawaii International Conference on System Sciences*. IEEE Computer Society Press.
- Garnier J, Osguthorpe DJ, 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.
- Gerloff DL, Benner SA. 1995. A consensus prediction of the secondary structure for the 6-phospho-beta-D-galactosidase superfamily. *Proteins Struct Funct Genet* 21:273–281.
- Gibrat JF, Garnier J, Robson B. 1987. Further development of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J Mol Biol* 198:425–443.
- Greer J. 1981. Comparative model-building of the mammalian serine proteases. *J Mol Biol* 153(4):1027–1042.
- Greer J. 1990. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins Struct Funct Genet* 7:317–334.
- Higgins DG, Sharp PM. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci* 5:151–153.
- Holley LH, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA* 86:152–156.
- Jenny TF, Benner SA. 1994. A prediction of the secondary structure of the pleckstrin homology domain. *Proteins Struct Funct Genet* 20(1):1–3.
- Jenny TF, Gerloff DL, Cohen MA, Benner SA. 1995. Predicted secondary and supersecondary structure for the serine-threonine-specific protein phosphatase family. *Proteins Struct Funct Genet* 21(1):1–10.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- King RD, Sternberg MJ. 1990. Machine learning approach for the prediction of protein secondary structure. *J Mol Biol* 216:441–457.
- Lesk AM. 1991. *Protein architecture—A practical approach*. Oxford: Oxford University Press.
- Levin JM, Garnier J. 1988. Improvements in secondary structure prediction method based on search for local sequence homologies and its use as a model building tool. *Biochim Biophys Acta* 955:283–295.
- Levin JM, Pascarella S, Argos P, Garnier J. 1993. Quantification of secondary structure prediction improvement using distantly related proteins. *Protein Eng* 6(8):849–854.
- Moult J, Judson R, Fidelis K, Pedersen JT. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct Funct Genet* 23:i–iv.
- Munson PJ, Di Francesco V, Porrelli R. 1994. Protein secondary structure prediction using periodic-quadratic-logistic models: Statistical and technical issues. *27th Hawaii International Conference on System Sciences*. IEEE Computer Society Press.
- Qian N, Sejnowski TJ. 1989. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202(4):865–884.
- Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232:584–599.
- Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct Funct Genet* 19:55–72.
- Rost B, Sander C, Schneider R. 1994a. Evolution and neural networks. Protein secondary structure prediction above 71% accuracy. *27th Hawaii International Conference on System Sciences*. IEEE Computer Society Press.
- Rost B, Sander C, Schneider R. 1994b. PHD—An automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* 10(1):53–60.
- Salamov AA, Solov'yev VV. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms, multiple sequence alignments. *J Mol Biol* 247:11–15.
- Salzberg S, Cost S. 1992. Predicting protein secondary structure with a nearest-neighbor algorithm. *J Mol Biol* 227:371–374.
- Schneider R, Sander C. 1991. Database of homology derived structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet* 9:56–68.
- Stolorz P, Lapedes A, Xia Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J Mol Biol* 225:363–377.
- Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. *J Mol Biol* 225:1049–1063.
- Zvelebil MJ, Barton GJ, Taylor WR, Stenberg MJE. 1987. Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J Mol Biol* 195:957–961.