Evaluation and improvement of multiple sequence methods for protein secondary structure prediction

James A. Cuff and Geoffrey J. Barton†

Laboratory of Molecular Biophysics,
Rex Richards Building,
South Parks Road,
Oxford, OX1 3QU, UK

and

European Molecular Biology Laboratory Outstation
The European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton,
Cambridge, CB10 1SD, UK

Keywords: protein; secondary structure prediction; combination of methods; benchmarks

†Corresponding Author: G. J. Barton, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Abstract

A new dataset of 396 protein domains is developed and used to evaluate the performance of the protein secondary structure prediction algorithms DSC, PHD, NNSSP and PREDATOR. The maximum theoretical Q_3 accuracy for combination of these methods is shown to be 78%. A simple consensus prediction on the 396 domains, with automatically generated multiple sequence alignments gives an average Q_3 prediction accuracy of 72.9%. This is a 1% improvement over PHD, which was the best single method evaluated. Segment Overlap Accuracy (SOV) is 75.4% for the consensus method on the 396 protein set. The secondary structure definition method DSSP defines 8 states, but these are reduced by most authors to 3 for prediction. Application of the different published 8 to 3 state reduction methods shows variation of over 3% on apparent prediction accuracy. This suggests that care should be taken to compare methods by the same reduction method. Two new sequence datasets (CB513 and CB251) are derived which are suitable for cross-validation of secondary structure prediction methods without artifacts due to internal homology. A fully automatic World Wide Web service that predicts protein secondary structure by a combination of methods is available *via* http://barton.ebi.ac.uk/

Introduction

The most successful techniques for prediction of the protein three dimensional structure rely on aligning the sequence of a protein of unknown structure to a homologue of known structure (e.g. see Sali for review[1]). Such methods fail if there is no homologue in the structural database, or if the technique for searching the structural database is unable to identify homologues that are present. While absence of a homologue must await further X-ray or NMR structures, up to 4/5 of known homologues may be missed even by the best conventional pairwise sequence comparison methods[2].

Techniques that exploit evolutionary information from protein families[3, 4, 5, 6, 7, 8, 9] or use empirical pair-potentials[10, 11] can normally detect more homologues than pairwise sequence comparison methods. An even greater challenge is to detect proteins that share similar folds, but are not clearly derived from a common ancestor (e.g. Rossman fold domains of lactate dehydrogenase and glycogen phosphorylase, and SH2-BirA[12])

Techniques for the prediction of protein secondary structure provide information that is useful both in *ab initio* structure prediction and as an additional constraint for fold-recognition algorithms[13, 14, 15]. Knowledge of secondary structure alone can help in the design of site-directed or deletion mutants that will not destroy the native protein structure. However, for all these applications it is essential that the secondary structure prediction be accurate, or at least that, the reliability for each residue can be assessed.

The majority of secondary structure prediction algorithms derive parameters or rules from an analysis of proteins of known three dimensional structure. The parameters are then applied by the algorithm to the sequence of unknown structure. Such approaches rely on having sufficient data to obtain reliable parameters and to avoid over-training for a specific data set.

Early algorithms to predict protein secondary structure[16, 17, 18] claimed high accuracy for prediction, but on small datasets that were also used in training the methods. For example, Lim (1974)[16] quoted 70% Q_3[19] accuracy on a dataset of 25 proteins, Garnier *et al.*(1978)[18] achieved 63% accuracy for a different set of 26 proteins, and Chou & Fasman (1974)[17] quoted 77% for yet another different set of 19 proteins.

The use of different datasets in training and testing each algorithm makes it difficult to make an objective comparison of methods. For this reason, Kabsch & Sander (1983)[20] carried out a test of prediction methods by applying the algorithms to proteins that were not used in their development. In this independent test, the GOR[18] accuracy reduced by 7% to 56%. The Lim[16] accuracy reduced by 14% to 56%, and Chou-Fasman[17] dropped by 27% to 50%. Cross-validation techniques, where test proteins are removed from the training set, have allowed more realistic evaluation of prediction accuracy to be obtained.

Prediction from a multiple alignment of protein sequences rather than a single sequence has long been recognised as a way to improve prediction accuracy[18]. During evolution, residues with similar physico-chemical properties are conserved if they are important to the fold or function of the protein. This makes patterns of hydrophobic residues characteristic of particular secondary structures easier to identify[21]. Analysis of conservation in protein families has been effective in many secondary structure predictions performed before knowledge of the protein structure[22, 23, 24, 25]. Zvelebil *et al.*(1987)[26] developed an automatic procedure that showed a 9% improvement in prediction accuracy on a small set of protein families when multiple sequence data was included. Most current secondary structure prediction algorithms exploit similar principles to gain higher accuracy than is possible from a single sequence[27, 28, 29, 30]. The recent CASP[31] series of experiments in which predictions are made blind have shown that recent claims for secondary structure prediction algorithms[32] are within reasonable limits.

Prediction accuracy has also been improved by combining more than one algorithm on a single sequence[33, 34, 35, 36, 37]. For example, Zhang *et al.*(1992)[34] obtained 66.4% accuracy on a set of 107 proteins, an improvement of 2% over the best method they considered.

In this paper we describe datasets and procedures for the evaluation of current techniques for secondary structure prediction. We discuss the effects of homology within the training and test datasets and describe new non-redundant datasets appropriate for developing secondary structure prediction algorithms. We evaluate the accuracy of four recently published algorithms that exploit multiple sequence data NNSSP[30], PHD[27], DSC[29] and PREDATOR[28] and two older methods, ZPRED[26] and MULPRED (Barton, unpublished). We develop an algorithm that combines the predictions of PHD, DSC, PREDATOR and NNSSP and show that it gives a 1% improvement in average accuracy over the best single method. Finally, we investigate the effect of the quality of multiple sequence alignment used in prediction, the effect of secondary structure assignment algorithm (DSSP[38], DEFINE[39] and STRIDE[40]) and influence of redundancy in the multiple alignments.

Methods

The problem of objectively testing secondary structure prediction methods

If a protein sequence shows clear similarity to a protein of known three dimensional structure, then the most accurate method of predicting the secondary structure is to align the sequences by standard dynamic programming algorithms[41], as homology modeling is much more accurate than secondary structure prediction for high levels of sequence identity. Secondary structure prediction methods are of most use when sequence similarity to a protein of known structure is undetectable. Accordingly, it is important that there is no detectable sequence similarity between sequences used to train and test secondary structure prediction methods.

Most secondary structure prediction methods include a set of parameters that must be estimated. Values for the parameters are obtained by statistical analysis or learning from a set of proteins for which the tertiary structure is known. This is the *training set* of proteins. Testing predictive accuracy on the training set leads to unrealistically high accuracies. An objective test of a secondary structure prediction method will predict the structures of a *test set* of proteins that are not in the training set and show no detectable sequence similarity with the training set. If the test is to be balanced, then both training an test sets should have a similar distribution of secondary structure classes and types.

Since the number of proteins of known structure is limited, it is normal to develop secondary structure prediction methods by cross-validation techniques, or jack-knife. In a full jack-knife test of N proteins, one protein is removed from the set, the parameters are developed on the remaining $N-1$ proteins, then the structure of the removed protein is predicted and its accuracy measured. This process is repeated N times by removing each protein in turn. Since some training techniques are very time consuming, a more limited cross-validation is often performed. The set of proteins might be split into M equally balanced subsets rather than N. Parameters are developed on $(M-1)N/M$ proteins, then tested on the remaining N/M proteins. This process is repeated M times, once for each subset. As described the jack-knife process may also be referred to as a leave-one-out technique, although the two terminologies have become some what synonymous.

Cross-validation appears to remove the problem of a limited data set for training and test. However, artificially high accuracies can be obtained for some methods if the set of proteins used in the cross-validation show sequence similarity to each other. Accordingly, cross-validation sets must be pruned stringently to remove internal sequence similarities, or if this is not possible, then a completely independent test set must be used.

Selection of suitable test and training sets rests with the definition of 'undetectable' sequence similarity. Appropriate measures of sequence similarity are discussed in the following section.

There are now available ≈ 500 sequence dissimilar proteins of known three dimensional structure, suitable for developing and testing secondary structure prediction techniques. However, many of the current generation of secondary structure prediction methods were developed on a set of 126 protein chains proposed by Rost & Sander[27] (referred to here as RS126). In this paper we develop a new, non-redundant set of 396 protein domains (the CB396 set) that does not include proteins from the RS126 set.

Training and test sets of protein structures

Rost & Sander (1993) selected 126 proteins with which to train and test secondary structure prediction algorithms[27]. They defined non-redundancy to mean that no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues. Unfortunately, as shown below, the RS126 set contains pairs of proteins that are clearly sequence similar when compared by more sophisticated methods than percentage identity.

Percentage identity has long been known to be a poor measure of sequence similarity, particularly for values below 30%. Percentage identity is dependent upon both the length of the alignment[42] and the composition of the sequences. Thus, two sequences of similar unusual amino acid composition may give high values of percentage identity, even when unrelated.

Recently, deficiencies in percentage identity have been quantified by Brenner *et al.*(1998) when scoring protein sequence database searches[2]. Even with the length correction suggested by Sander & Schneider (1991)[42], percentage identity was significantly worse than measures that consider conservative substitutions as well as identities, and attempt corrections for length and composition.

Fortunately, techniques exist that overcome the deficiencies of percentage identity or other simple measures of sequence similarity. A long established method[43, 6] to measure the similarity between two protein sequences A and B is first to align the proteins by a standard dynamic programming algorithm (e.g. Needleman & Wunsch (1970)[44]) and obtain the score for the alignment V. The order of amino acids in each protein sequence is then randomised and a dynamic programming alignment of the randomised sequences performed. This process is repeated typically 100 or more times and the mean $\bar{x}$ and standard deviation σ of the scores for comparison of the randomised sequences is calculated. The SD score, or Z score for comparison of the native sequences is given by: $\frac{V-\bar{x}}{\sigma}$. Unlike the percentage identity, SD score corrects for bias due to the length and composition of the sequences. Accordingly, we use SD scores to derive our non-redundant test set of protein sequences.

PHD[27], NNSSP[30], DSC[29], and PREDATOR[28] have been trained on the Rost & Sander set of 126 proteins. The release versions of PREDATOR and NNSSP available for this analysis were trained on larger sets, that included the 126 proteins. In principle, this should give PREDATOR and NNSSP an advantage over PHD.

The sequences in the test set developed here came from the 3Dee[45] database of structural domain definitions. In 3Dee, a non-redundant sequence set was created by the use of a sensitive sequence comparison algorithm and cluster analysis, rather than a simple percentage identity cutoff. This provided a set of 1233 domains where no pair shared obvious sequence similarity. The new test set was derived from these domains by first removing multi-segment domains, to reduce the set size from 1233 to 988 sequences. The sequences were then filtered only to permit X-ray crystal structures with resolutions of ≤ 2.5 Angstroms. This left a representative set of 554 domain sequences, referred to as

CB554.

To ensure that the CB554 domain set had no sequence similarity to the RS126 set, the two sets were combined and all pairs of sequences compared by AMPS[6] with a blosum62 matrix, and gap penalty of 10. Alignments with an SD score of ≥ 5 were regarded as sequence similar[6, 46]. According to this stringent definition of similarity, there were 11 sequence-similar pairs within the RS126 protein set, 119 pairs between CB554 domain set and RS126, and 21 pairs within CB554. Thus, there were 140 sequences in CB554 that matched either a sequence in CB554, or in the RS126 protein set. Of the 140, 3 sequences matched more than once, leaving 137 unique sequences. The 137 sequences were removed from CB554, leaving 417 sequences that were not sequence similar either to any sequence within the set of 417 sequences, or the RS126 sequence set. Of the 417 domain sequences remaining, 21 that did not have 'full DSSP definitions', (*i.e.* those with more than 9 consecutive residues with incomplete backbones for which DSSP[38] does not define a state), were also removed, leaving a test set of 396 proteins (CB396).

The process of deriving CB396 showed up homologies in the RS126 set, with 11 proteins showing sequence similarity to at least one other within the RS126 set. These pairs are summarised in Table 1. Table 1 shows each pair to have the same fold according to SCOP[47]. For example, 4cms[48] and 5er2e[49] are present in the RS126 set, yet have an SD score of 15.9. Both proteins are acid proteases with an all β, closed barrel structure.

Although not applied in this paper, three further non-redundant datasets suitable for cross-validated training and testing of secondary structure prediction methods were generated. The CB396 and RS126 sequence sets were combined. One of each of the 11 pairs that had an SD score of ≥ 5 were removed from the RS126 set. Since 2pcy and 1lhb matched more than one protein in this subset, this left 9 unique homologues (1mcpl[50], 1tgsi[51], 2lhb[52], 2pcy[53], 3ebx[54], 4cms[48], 4cpv[55], 5hvpa[56], 8abp[57]) that were removed from

RS126. This set added to CB396 gave CB513. Protein chains of $\leq$ 30 residues often do not have well defined secondary structure. The CB497 set was constructed by removing the 16 domains from CB513 of $\leq$ 30 residues.

The 5SD cutoff used to derive the sets CB396, CB497 and CB513 is more stringent than scores used in previous studies of secondary structure prediction. However, although the SD score is a good measure of pairwise sequence similarity, it still will not identify all known homologues within the data set. In the SCOP[47] classification of protein structure superfamilies are defined from careful analysis of structure, evolution and function. The SCOP superfamilies contain protein domains that have the same fold and are likely to have evolved from a common ancestor. Accordingly, we derived a further dataset from an analysis of all domains in SCOP_1.37. We took a representative domain from each superfamily, screened out multi-segment domains, NMR structures and those with a resolution $\geq$ 2.5 Angstroms to give the CB251 dataset.

All datasets, including secondary structure definitions and automatically generated multiple sequence alignments will be distributed *via* http://barton.ebi.ac.uk/.

Generating the multiple sequence alignments

With the exception of PREDATOR[58, 28] all methods considered here, required a multiple sequence alignment as input, where as PREDATOR only required the multiple sequences in an unaligned format. In order to simplify the generation of multiple sequence alignments for large numbers of proteins, in this study we developed an automatic procedure.

We first perform a BLAST[59] database search of the OWL v29.4 database, which contains 198,742 entries[60]. The BLAST output is then screened by SCANPS, an implementation of the Smith Waterman dynamic programming algorithm[61, 62], with length dependent statistics. Sequences are rejected if their SCANPS probability score is higher than $1\text{x}10^{-4}$.

Sequences are also rejected if they do not fit a length cutoff of 1.5. For example, if the query sequence is 90 residues long, the sequence length would have to range between 60 and 135 residues to be included. If sequences exceed the length criterion, they are truncated by removing end residues until the length of the sequence satisfies the cut off value. Sequences falling short of the lower length limit are discarded. The value of 1.5 for the length cutoff was reached by visual inspection of a number of multiple sequence alignments, produced with different cut-off values. The method removes both ridiculously long, short and unrelated sequences. However it does allow sequences that are longer than the query, and are related, to be included after truncation. The sequence similar proteins selected by this method, are then aligned by CLUSTALW (version 1.7)[63], with default parameters.

The multiple sequence alignments are modified so that they do not contain gaps in the first or 'query' sequence, since with the current algorithms, gaps in the first sequence tend to reduce the accuracy of the prediction, or cause the program to fail to execute (NNSSP[30]). A slightly different method is used for PHD[64], whereby only gaps at the end of the target sequence are removed. Without this modification, the conversion of MSF to HSSP file format fails, as a correct insertion table is not constructed.

The reference secondary structure for each domain was defined by DSSP[38], STRIDE[40] and DEFINE[39]. All definitions were reduced to 3 state models, as follows:

1. DSSP: H and G to H, E and B to E, all other states to C

2. STRIDE: H and G to H, E and b to E, all other states to C

3. DEFINE: H and G to H, E to E, all other states to C

Where H is $\alpha-$ helix, G is $3_{10}-$ helix, B and b are isolated $\beta-$ bridge and E is $\beta-$ strand.

The effect of alternative reduction methods for the DSSP algorithm is discussed in the results section.

Prediction methods analysed

Six different secondary structure prediction methods were run on the alignments, each is briefly described here.

PHD[64] is a 3 level artificial neural network. The different levels consist of a sequence to secondary structure network, with a window of 13 amino acids, a structure to structure network, with a window of 17 amino acids, and finally an arithmetic average over a number of independently trained networks. The structure to structure network, improves prediction of the final length distributions of secondary structures. The arithmetic average has the effect of smoothing random noise that is seen in all artificial neural networks. The method also applies balanced training, percentage amino acid composition and conservation, sequence length, and insertions and deletions (indels) to enhance prediction accuracy.

DSC[29] applies GOR[18] residue attributes, with the addition of hydrophobicity and amino acid position, which are combined with information from the multiple sequence alignment (conservation and indels). Optimal weights are deduced by linear discrimination[65], with filtering applied to remove erroneous predictions. This method has an advantage in that the prediction method is both implicit and effective.

NNSSP[30] is a scored nearest neighbour method. It is based upon the environmental scoring scheme proposed by Bowie[66]. The NNSSP method extends the Bowie method by considering N and C terminal positions of $\alpha-$ helices and $\beta-$ strands. The size of the database used for scanning is also altered to reflect similarity to the query sequence, reducing computation time, and improving the final accuracy.

PREDATOR[28] is slightly different to other methods discussed here, in that it uses an internal pairwise alignment method, rather than reading a global multiple sequence alignment. The SIM software[67] is applied to produce local alignments between sequence pairs. The original PREDATOR algorithm[58] is then used to predict the secondary structure segments. This algorithm also includes propensities for hydrogen bonding characteristics of $\beta-$ sheets. Seven different secondary structure propensities are generated for the query sequence, with a nearest neighbour implementation applied to calculate propensities for $\alpha-$ helix, $\beta-$ strand and coil.

ZPRED[26] is also based on the GOR[18] method, but with the addition of weights from calculated conservation values. The conservation value is calculated from amino acid properties as proposed by Taylor[68]. The ZPRED method improved the accuracy of the GOR method by noting that insertions and high sequence variability tend to occur in loop regions.

MULPRED (Barton, unpublished) is a combination of single sequence methods that are combined to give a prediction profile, from which a consensus is taken. The methods within MULPRED are Lim[16], GOR[18], Chou-Fasman[17], Rose[69] and Wilmot & Thornton[70] turn prediction methods.

Consensus Prediction Method

The observed Q_3 accuracy of ZPRED and MULPRED was between 3 and 8% lower than the other methods, so a consensus was calculated only from DSC, PHD, PREDATOR and NNSSP. The standard consensus was calculated by examining the prediction for each method, at each position and taking the most popular state. For example if a residue had the following predictions:

NNSSP = Helix, PREDATOR = Helix, DSC = Strand, PHD = Helix

the consensus prediction would be Helix. If there was no consensus for a particular residue, the result from the PHD method was used. More complex methods for combining the different predictions were investigated and tested, as discussed in the results section.

Assessment of accuracy

Two methods were applied to assess the accuracy of the predictions. Average Q_3[19], and Segment Overlap[64].

Q_3 is a measure of the overall percentage of predicted residues, to observed:

$$Q_3 = \sum_{(i=H,E,C)} \frac{predicted_i}{observed_i} \times 100 \tag{1}$$

Segment overlap calculation[64] was performed for each data set. Segment overlap values attempt to capture segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average 90% level for homologous protein pairs. Segment overlap is calculated by:

$$Sov = \frac{1}{N} \sum_{s} \frac{minov(s_{obs}; s_{pred}) + \delta}{maxov(s_{obs}; s_{pred})} \times len(s_1) \tag{2}$$

Where N is the total number of residues, minov is the actual overlap, with maxov is the extent of the segment. δ is the accepted variation which assures a ratio of 1.0 where there are only minor deviations at the ends of segments[64].

Results and Discussion

Comparison of secondary structure definition methods

All secondary structure prediction methods are trained and tested on secondary structure definitions from known structures. Defining secondary structure from co-ordinates is an inexact process due to differences in the concept of what is a secondary structure, as well as errors and inconsistencies in the experimental structure. This was illustrated in the comparison by Colloc'h *et al* of DSSP[38], DEFINE[39] and P-curve[71] on a non-redundant set of 154 proteins where all three methods agreed at only 63% of positions.

Here we show the differences between DSSP[38], DEFINE[39] and STRIDE[40] definitions on the RS126 protein set. When compared pairwise, DSSP and STRIDE agree to 95%, whereas DSSP and DEFINE agree at 73%, with STRIDE and DEFINE agreeing at 74%. All three methods agree at only 71% of positions.

Table 2 shows that DEFINE defines more sheet. 7.1% relative to DSSP and 5.6% relative to STRIDE. Helix is also defined more often by DEFINE. As a consequence of these two factors, DSSP and STRIDE define more coil than DEFINE, at 8.4% and 6.1% respectively.

The length of secondary structure elements as defined by DSSP, STRIDE and DEFINE is summarised in Table 3 and Figure 1. DEFINE does not define sheet regions of less than 4 residues. The mean segment length values for DEFINE are higher than those of STRIDE and DSSP for all secondary structure states. Figure 1 shows DSSP to have a peak in the helix distribution at 4 residues. However, this is not found with the STRIDE or DEFINE definitions. With the exception of the peak at 4, the overall shape of DSSP and STRIDE length distributions are similar.

When assessing prediction methods, the average Q_3 was calculated for all the definition methods, for all runs, but because DEFINE is so dissimilar to DSSP and STRIDE, all results from DEFINE have been omitted from discussion.

Analysis of the test and training alignments

Table 4 summarises an analysis of the automatic multiple sequence alignments that were generated for the RS126 and CB396 sets. Both sets have a similar average length of sequence, and average percentage identity within the set. However, there is a significant difference between the average number of sequences per alignment between the two sets, even though both sets of alignments were generated using the same method. The older

RS126 protein set has significantly (1.6 times) more sequence similar proteins in each alignment. The distribution of the number of sequences in the RS126 protein set was not biased by one or two large families.

A comparison between the CB396 set and the RS126 set showed the same distribution. The difference is therefore that each sequence family in the RS126 set is on average larger than any found in the CB396 set. This observation may simply reflect the fact that RS126 was derived from protein families whose first known members were characterised longer ago.

To verify that there was no bias to a particular structural class, the SCOP[47] classifications were examined for the proteins within the two sets as shown in Table 5. There is a higher proportion of small proteins in the RS126 protein set (14% against 7%), while the CB396 protein set has a higher proportion of $\alpha+\beta$ proteins (26% against 13%). However the overall composition within each of the two sets is balanced.

Alignment quality

The RS126 set of proteins was used to check that the alignments generated by our method could reproduce previous results for PHD[27], and DSC[29]. PHD[27] (run in cross-validation mode) increased in average Q_3 accuracy by 1.9% from 71.6 to 73.5%, over the published accuracy. The accuracy obtained here for DSC improved by 1.0%. However, the published value of 70.1% for DSC was for a full jack-knife test, whereas our test utilised all the data. These results confirm that the automatic method used to build multiple alignments in this study is appropriate for secondary structure prediction.

Comparison of prediction accuracy for the CB396 test and RS126 training sets

Table 6 shows the differences between the RS126 and CB396 set of proteins. For all methods, the average accuracy drops by between 1.3% (NNSSP) and 2.7% (DSC) for the CB396 protein set. The NNSSP and PREDATOR programs used in this analysis were trained on larger numbers of proteins than RS126, and so should be less degraded by evaluation on a different test set. However, these methods still show a decrease in accuracy with the CB396 set.

Table 6 illustrates that the percentages for the segment overlap correlate well with the Q_3 values. The SOV score for the consensus method is somewhat higher (74.5%) than the previous published value for PHD of 72%[64]. Table 6 also shows that the PHD method does exceedingly well at predicting segments. As measured by the SOV method, it is on average 2% better than any of the other methods tested. The difference between the consensus and PHD segment overlap scores is smaller than the corresponding Q_3 value. This may be due to segment overlap being a more sensitive method to assess secondary structure predictions, or that the PHD method is the only one that has been optimised to predict segments scored by Equation 2.

Table 6 summarises the differences between SOV and Q_3 accuracies for each method on the RS126 and CB396 sets. Although Q_3 shows a consistent reduction on moving from RS126 to CB396, SOV shows a general improvement. Of the individual methods, NNSSP increases in accuracy the most (0.7%) while the consensus method increases by 0.9% to 75.4% SOV accuracy.

Effect on Q_3 of changing the number of related sequences

Prediction methods that use multiple sequence alignments gain accuracy over single-sequence methods by exploiting the patterns of residue conservation that are seen in protein families. Inclusion of more distantly related sequences in the alignment should improve the clarity of such patterns, but in an automated alignment building procedure, the risk is that unrelated protein sequences will pollute the alignment. Here, we investigated the effect of using a more permissive BLAST p-value cutoff[59] in the first phase of our alignment building procedure. The cutoff was lowered from $1\text{x}10^{-10}$ to $1\text{x}10^{-2}$ while leaving thresholds for SCANPS alone.

Table 7 shows that the change in p-value cutoff increased the total number of residues after filtering with SCANPS by 297,276, and the total number of sequences by 1,961. This gives an increase in the average number of sequences per alignment of 15. Table 8 shows that increasing the number of sequences improves Q_3 by approximately 1% for all methods.

Table 8 also shows the marked difference between the prediction methods. The older methods, ZPRED and MULPRED were between 3 and 8 percent worse than the newer methods.

Effect on Q_3 of reducing redundancy in multiple alignments

While all sequences that are not 100% identical in a multiple sequence alignment will contribute to the prediction, the most informative sequences are those with the greatest variation from the query. Here we test the effect of systematically removing sequences from the alignment that were similar to the query sequence at better than 95, 80, 75 and 60% identity. Table 9 summarises the effect of these thresholds. The average Q_3 accuracy improves slightly as the percentage identity threshold is reduced. The consensus method improves by 0.3% at the 75% level. Since the predictions do not get any *worse* by removing redundant sequences and prediction methods run faster with fewer sequences, the 75% cutoff was used for all predictions, other than those shown in Table 8.

The average Q_3 for each prediction when compared to DEFINE secondary structure definitions was between 3 and 8% worse than for DSSP and STRIDE definitions. As none of the prediction methods examined here were trained on DEFINE definitions, we do not consider comparison of predictions to DEFINE definitions any further.

The effect on accuracy of alternative 8 to 3 state reductions

DSSP[38] provides an 8 state assignment of secondary structure denoted by single letter codes: H ($\alpha-$ helix), T ($\beta-$ turn), S (bend), I (π helix) , G ($3_{10}-$ helix), E ($\beta-$ strand), B ($\beta-$ bridge) and C (not HTSIGE or B). However, prediction methods are normally trained and assessed for only 3 states (H,C,E), so the 8 states must be reduced to 3. Here we consider the effect on accuracy of applying three different published 8 to 3 state reduction methods when testing.

- **Method A**: E and B to E, G and H to H. Rest to coil[27]
- **Method B**: E as E, H as H. Rest to coil including EE and HHHH[28]
- **Method C**: GGGHHHH redefined as HHHHHHH, then B, and GGG to coil, with H to H and E to E[30]

Method **A** treats both isolated $\beta-$ bridges and residues that are part of a $\beta-$ sheet as 'extended'. $3_{10}-$ helix and $\alpha-$ helix are treated as 'helix', and all other states treated as 'coil'. Method **B** translates more secondary structures into coil. This includes all $3_{10}-$ helix, $\alpha-$ helix that is a single turn, isolated $\beta-$ bridges and $\beta-$ strands that are only two residues long. The rationale for this reduction is that short secondary structures are normally of marginal stability and also variable within protein families. Table 10 shows reduction Methods **B** and **C** on average to contain 4% more coil, 3% less helix and 2% less strand than Method **A**.

Table 11 summarises the difference in Q_3 accuracy obtained by varying the reduction of 8 state DSSP definition to 3 states. The original value for the consensus method was 74.8% (Method A). This increases to 77.9% for Method B. Table 11 shows that reducing both single strand 'B' to coil and $3_{10}-$ helix 'G' to coil, gives the greatest stepwise increase (2.7%).

Table 12 shows that *all* prediction methods appear to improve in accuracy with comparison to Method A, when one uses Method B, as the method for the 8-state reduction. The apparent improvement is between 2.2 and 4.9%. The PREDATOR[28] method improves by the largest extent (4.9%), as a reflection of PREDATOR being trained with Method B as the reduction scheme. STRIDE [40] values have also been included in Table 12 for comparison. Prior to this study, PHD[64] and DSC[29] used the same reduction method (Method **A**), but NNSSP[30] and PREDATOR[58, 28] used methods **B** and **C** respectively. Unless the same reduction method is used, an objective comparison can not be made.

Single sequence prediction methods

The prediction methods we have carefully selected for this work represent current state-of-the-art prediction methods, that use multiple sequences. However for completeness, the SIMPA[72], SOPM[73], and GORIV[74], single sequence methods were also examined. These methods do not have pre-calculated propensity tables, and as such we could perform a full jack-knife test with the new datasets. We only compare the results for the single sequence methods to those obtained for the PHD algorithm, as PHD was the only other method for which we were able to carry out cross validation. The SIMPA[72], SOPM[73] and GORIV[74] methods have quoted accuracies based on removing helices shorter than 4 residues and strands less than 2 residues. For testing these methods, we used method A as the reduction method and also converted G and B states to coil. If reduction method A alone is used, SIMPA[72], SOPM[73] and GORIV[74] reduce in accuracy from those shown in table 13 by 2-4%.

The difference between the single sequence methods we examined and PHD ranges between 23.3%, and 6.6% depending upon the method and database used. Table 13 shows the GORIV method to improve remarkably (11.3%) with an increased database size 126 proteins to 396 proteins). This is to be expected as GORIV no longer uses 'dummy frequencies'[74] instead relying on a large database to calculate its propensity tables. To examine if this feature scaled, we also applied the GORIV method to the CB513 dataset. The accuracy improved by 1.1% from 64.6% to 65.7%. SOPM only achieved 66.8% on the RS126 protein set, and 64.6% for the CB396 set. The authors of the SOPM method quoted 69%[73]. However, the database used in their study, was non-redundant at 50% sequence identity, and so included a number of clear homologues.

Improving the consensus prediction

In order to establish the upper limit of accuracy possible by combining the prediction methods, we took the most accurate prediction for each residue in the RS126 data set by PHD, DSC, NNSSP or PREDATOR. This gave the theoretical best accuracy for a combination of these methods of $Q_3 = 78\%$.

We investigated a variety of techniques for combining the prediction methods, in an attempt to raise the average Q_3 on RS126 from 74.8% towards 78%. All possible combinations of methods were tried to calculate the consensus, but no combination of methods improved upon the average Q_3 of the consensus of DSC, PREDATOR, NNSSP and PHD, with PHD taken if there was a tie. However, the next highest combination was only 0.3%

worse at 74.2% and used NNSSP, PREDATOR and DSC, predictions relying on PREDATOR's definition if there was no consensus. Experiments with filtering single residue helix predictions and other unlikely secondary structures did not improve the overall Q_3.

The reliability information from the PHD and PREDATOR predictions was also investigated. When a method predicted with a reliability of greater than 7, that prediction was taken. No further increase in average Q_3 accuracy could be achieved using this approach.

The predictions for each method were weighted by adding constants. All combinations of all values from 1 to 10 were applied to all predictions for each method. The consensus was then calculated in the same manner as before, but now using the weighted predictions. The optimal weighting scheme was 2,1,2,2 where PREDATOR was down weighted by one point. The Q_3 accuracy for this approach was no higher than that of the non weighted majority wins method.

An artificial neural network, with 9 hidden nodes was trained with the output from the NNSSP, PHD, DSC and PREDATOR methods. A 17 residue window was used. The inputs were coded as binary, with 001, 010 and 100 representing the helix, strand and coil states respectively. Seven fold cross validation was performed. This yielded 73.2% for the 126 protein set. This result was still lower than the simple consensus approach. No further improvement in accuracy was seen by changing the free parameters of the network, for example, hidden nodes or number of training epochs. The target sequence was also included in the input layer, but this also proved unsuccessful. We suggest that better accuracies may be achieved if propensities for the different states are used, rather than the binary input, and this idea forms the basis of future work.

When the lower accuracy predictions from ZPRED, MULPRED were included, the overall accuracy of the consensus method was reduced. SIMPA, SOPM and GORIV were not included at any stage in the consensus method. Further work aims to discover if the single sequence prediction methods can be incorporated into a more accurate consensus method.

Summary and Conclusions

In this study we have developed a new, non redundant test set of 396 protein domains (CB396). The set does not include any of the 126 proteins with which many current methods have been trained, nor does it contain homologues of those 126 proteins as measured by a stringent test of sequence similarity. We have shown that by combining four secondary structure prediction methods DSC[29], PHD[27], PREDATOR[28] and NNSSP[30] by a simple majority wins method, the average three-state Q_3 prediction accuracy can be improved by 1% from 71.9% (PHD) to 72.9% on the CB396 set. A fair comparison of the accuracy of the constituent methods is only possible for PHD[27] and DSC[29] as all other algorithms included some of our test proteins in their training set. Despite this, PHD[27] still gave the highest accuracy on the new test set (71.9%) of any of the methods considered.

An automatic procedure for database searching to build a multiple sequence alignment has been developed. Alignments from this procedure give a 1.9% increase in the average accuracy of prediction compared to previous published results for the PHD algorithm on the 126 protein set[64]. The increase may be attributed to better alignments and the increased size of the current sequence databases.

In the literature there are different standards for reducing DSSP[38] 8-state (H,C,B,E,T,S,G,I) assignments to 3 states (H,C,E). It was found that changing the reduction method can alter the apparent prediction accuracy by over 3% on average. Although we were unable to train the methods using different 8 to 3 state reductions, testing all methods with different reduction methods showed that Method B[58] consistently gave higher accuracy. This may be attributed to Method B assigning more of the protein to Coil (C).

Secondary structure definition methods DSSP[38], DEFINE[39] and STRIDE[40] were compared. All three agree at only 75% of positions. This is mainly due to differences between DEFINE and DSSP/STRIDE. DSSP and STRIDE agree at 95% of positions, though DSSP defines many more 4 residue helices than STRIDE.

In summary, with the alignment method presented here, the method with the highest average accuracy on the new non-redundant test set of 396 proteins was PHD[64] with 71.9%. While the new combination of NNSSP[30], PHD [27], DSC[29] and PREDATOR[28] presented here improves upon this figure by 1% to 72.9%.

The non-redundant datasets constructed during this analysis will facilitate the future development and testing of secondary structure prediction methods. The datasets, alignments and definitions are available *via* http://barton.ebi.ac.uk.

Acknowledgements

We would like to thank the developers of the algorithms we have considered for their helpful discussions and cooperation in this study. In particular Drs B. Rost, D. Frishman V. Solovyev, R. King and M. Zvelebil, for providing their software. We also thank Matt Finlay for coding many of the routines to parse prediction output, as well as Dr. A.S. Siddiqui for his automatic alignment building algorithm. We thank Prof. L. N. Johnson. for her encouragement and support. This work was supported in part by grants from the Medical Research Council and the Royal Society. JC is an Oxford Centre for Molecular Sciences/MRC student.

References

1. A. Sali. Modelling mutations and homologous proteins. *Current Opinion in Biotechnology*, 6:437–451, 1995.

2. S. E. Brenner, C. Chothia, and T. J. P. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Nat. Acad. Sci.*, 95:6073–6078, 1998.

3. W. R. Taylor. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, 188:233–258, 1986.

4. M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *Proc. Nat. Acad. Sci.*, 84:4355–4358, 1987.

5. G. J. Barton. Protein multiple sequence alignment and flexible pattern matching. *Meth. Enz.*, 183:403–428, 1990.

6. G.J. Barton and M.J.E. Sternberg. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, 198:327–337, 1987.

7. S. R. Eddy. Hidden markov models. *Current Opinion Structural Biol.*, 6:361–365, 1996.

8. K. Karplus, K. Sjolander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm, and C. Sander. Predicting protein structure using hidden markov models. *Proteins*, Suppl. 1:134–139, 1997.

9. J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of distant sequence homologues. *J. Mol. Biol.*, 273:349–354, 1997.

10. D. T. Jones, W. R Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

11. C. Lemer, M. J. Rooman, and S. J. Wodak. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins*, 23:337–355, 1996.

12. R. B. Russell and G. J. Barton. An SH2-SH3 domain hybrid. *Nature*, 364:765, 1993.

13. R. B. Russell, R. R. Copley, and G. J. Barton. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.*, 259:349–365, 1996.

14. B. Rost. TOPITS: Threading one-dimensional preditions into three-dimensional structures. *Proc. 3rd. Int. Conf. Intel. Sys. Mol. Biol.*, pages 314–321, 1995.

15. B. Rost. Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, 270:1–10, 1997.

16. V. I. Lim. Algorithms for prediction of α helices and β structural regions in globular proteins. *J. Mol. Biol.*, 88:873–894, 1974.

17. P. Y. Chou and G. D. Fasman. Conformational parameters for amino acids in helical, β–sheet, and random coil regions calculated from proteins. *Biochem.*, 13:211–222, 1974.

18. J. Garnier, D. J. Osguthorpe, and B. Robson. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120:97–120, 1978.

19. G. E. Schulz and R. H. Schirmer. *Principles of Proteins Strcuture.* Springer–Verlag, New York, 1979.

20. W. Kabsch and C. Sander. How good are predictions of protein secondary structure? *FEBS Letters*, 155:179–182, 1983.

21. C. D. Livingstone and G. J. Barton. Identification of functional residues and secondary structure from protein multiple sequence alignment. *Meth. Enz.*, 266:497–512, 1996.

22. I. P. Crawford, T. Niermann, and K Kirchner. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase. *Proteins*, 2:118–129, 1987.

23. G. J. Barton, R. H. Newman, P. F. Freemont, and M. J. Crumpton. Amino acid sequence analysis of the annexin super-gene family of proteins. *European J. Biochem.*, 198:749–760, 1991.

24. R. B. Russell, J. Breed, and G. J. Barton. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters*, 304:15–20, 1992.

25. S.A. Benner and D. Gerloff. 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. Enz. Reg.*, 31:121–181, 1990.

26. M. J. J. M. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. E. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, 195:957–961, 1987.

27. B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

28. D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, 27:329–335, 1997.

29. R. D. King and M. J. E. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.*, 5:2298–2310, 1996.

30. A. A. Salamov and V. V. Solovyev. Prediction of protein secondary structure by combining nearest- neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, 247:11–15, 1995.

31. *Proteins*, Suppl. 1:1–230, 1997.

32. B. Rost. Better 1D predictions by experts with machines. *Proteins*, Suppl. 1:192–197, 1997.

33. V. Biou, J. F. Gilbrat, B. Robson, and J. Garnier. Secondary structure prediction: combination of three different methods. *Prot. Eng.*, 2:185–191, 1995.

34. X. Zhang and D. Mesriov, J.and Waltz. A hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, 225:1049–1063, 1992.

35. K. Nishikawa and T. Ooi. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochem.. Biophys. Acta*, 871:45–54, 1986.

36. K. Nishikawa and T. Nogughi. Predicting protein secondary structure based on amino acid sequence. *Meth. Enz.*, 202:31–44, 1995.

37. C. Geourjon and G. Deleage. Sopma : Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comp. App. Biosci.*, 11:681–684, 1995.

38. W. Kabsch and C. Sander. A dictionary of protein secondary structure. *Biopolymers*, 22:2577–2637, 1983.

39. F. M. Richards and C. E. Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first–level supersecondary structure. *Proteins*, 3:71–84, 1988.

40. D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23:566–579, 1995.

41. P. E. Boscott, G. J. Barton, and W. G Richards. Secondary structure prediction for modelling by homology. *Prot. Eng.*, 6:261–266, 1993.

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

43. D. F. Feng, M. S. Johnson, and R. F Doolittle. Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, 21:112–125, 1985.

44. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

45. A. Siddiqui and G. J. Barton. 3Dee — database of protein domain definitions. submitted., 1998.

46. G. J. Barton and M. J. Sternberg. Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.*, 1:89–94, 1987.

47. A. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database and the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

48. M. Newman, C. Frazao, G. Khan, I. J. Tickle, T. L. Blundell, M. Safro, N. Andreeva, and A. Zdanov. X-ray analyses of Aspartic Proteinases. structure and refinement at 2.2 Angstroms resolution of Bovine Chymosin. *J. Mol. Biol.*, 221:1295, 1991.

49. A. Sali, B. Veerapandian, J. B. Cooper, S. I. Foundling, D. J. Hoover, and T. L. Blundell. High resolution x-ray diffraction study of the complex between endothiapepsin and an oligopeptide inhibitor. the analysis of the inhibitor binding and description of the ridgid body shift in the enzyme. *EMBO J.*, 8:2179, 1989.

50. Y.Satow, G.H.Cohen, E.A.Padlan, and D.R.Davies. Phosphocholine binding Immunoglobulin study at 2.7 angstroms. *J. Mol. Biol.*, 190:593, 1987.

51. M.Bolognesi, G.Gatti, E.Menegatti, M.Guarneri, M.Marquart, E.Papamokos, and R.Huber. Three dimensional structure of the complex between pancreatic secretory inhibitor (kazal type) and trypsinogen at 1.8 angstroms resolution. *J. Mol. Biol.*, 162:839, 1982.

52. R.B.Honzatko, W.A.Hendrickson, and W.E.Love. Refinement of a molecular model for Lamprey Hemoglobin from Perromyzon Marinus. *J. Mol. Biol.*, 184:147, 1985.

53. T.P.J.Garrett, J.M.Guss, and H.C.Freeman. The crystal structure of Poplar Apoplastocyanin at 1.8 Angstroms resolution. *J. Biol. Chem.*, 259:2822, 1984.

54. J.L.Smith, P.W.R.Corfields, W.A.Hendrickson, and B.W.Low. Refinement at 1.4 Angstroms resolution of a model of Erabutoxin B. treatment of ordered olvent and discrete order. *Acta Cryst.*, 44:357, 1988.

55. V.D.Kumar, L.Lee, and B.F.P.Edwards. Refined crystal structure of Calcium liganded Carp Paravalbumin 4.25 at 1.5 Angstroms resolution. *Biochem.*, 29:1404, 1990.

56. P.M.D.Fitzgerald, B.M.Mc Keever, J.F.Van Middlesworth, and J.P.Springer. Crystallographic analysis of a complex between Human Immunodeficiency Virus Type 1 Protease and Acetyl Pepstatin at 2.0 Angstroms resolution. *J. Biol. Chem.*, 265:14209, 1990.

57. F.A.Quiocho, D.K.Wilson, and N.K.Vyas. Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature*, 340:404, 1989.

58. D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Prot. Eng.*, 9:133–142, 1996.

59. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

60. A. J. Bleasby, D. Akrigg, and T. K. Attwood. OWL — A non-redundant, composite protein sequence database. *Nuc. Ac. Res.*, 22:3574–3577, 1994.

61. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

62. G. J. Barton. Alscript: A tool to format multiple sequence alignments. *Prot. Eng.*, 6:37–40, 1993.

63. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sesitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weigh matrix choice. *Nuc. Ac. Res.*, 22:4673–4680, 1994.

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

65. S.M. Weiss and C.A. Kulikowski. San Mateo, 1991.

66. J. U. Bowie, R. Luthy, and D. Eisnenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

67. X. Huang and W. A. Miller. *Adv. Appl. Math.*, 12:337–357, 1991.

68. W. R. Taylor. Classification of amino acid conservation. *J. Theor. Biol.*, 119:205–218, 1986.

69. G. D. Rose. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature*, 272:586–591, 1978.

70. A. C. M. Wilmot and J. M. Thornton. Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.*, 203:221–232, 1988.

71. H. Sklenar, C. Etchebest, and R. Lavery. *Proteins*, 6:46–60, 1989.

72. J. M. Levin. Exploring the limits of nearest neighbour secondary structure prediction. *Prot. Eng.*, 10:771–776, 1997.

73. C. Geourjon and G. Deleage. SOPM : a self optimised method for protein secondary structure prediction. *Prot. Eng.*, 7:157–164, 1994.

74. B. Robson J. Garnier, J. Gibrat. GOR method for predicting protein secondary structure from amino acid sequence. *Meth. Enz.*, 266:540–553, 1996.

75. W. Steigemann and E. Webber. Structure of Erythrocruorin in different ligand states refined at 1.4 Angstroms resolution. *J. Mol. Biol.*, 127:309, 1979.

76. E.T.Adman, L.C.Sieker, and L.H.Jensen. Structural features of Azurin at 2.7 Angstroms. *Isr. J. Chem.*, 21:8, 1981.

77. A. Wlodawer, M.Miller, and M.Jaskolski. Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature*, 337:576, 1989.

78. K.Petratos, Z.Dauter, and K.S.Wilson. Refinement of the structure of Pseudoazurin from Alcaligenes Faecalis S-6 at 1.55 Angstroms. *Acta Cryst.*, 44:628, 1988.

79. W.E.Royer (Jr.). High resolution crysallographic analysis of a cooperative dimeric Haemoglbin. *J. Mol. Biol.*, 657:657, 1994.

80. B.Rees, A.Bilwes, J.P.Samama, and D.Moras. Cardiotoxin from Naja Mossanmica: The refined crystal structure. *J. Mol. Biol.*, 214:281, 1990.

81. Y.S.Babu, C.E.Bugg, and W.J.Cook. Structure of Calmodulin refined at 2.2 Angstroms resolution. *J. Mol. Biol.*, 204:191, 1988.

82. N.K.Vyas, M.N.Vyas, and F.A.Quiocho. Sugar and signal transducer binding sites of the escherichia coli galactose chemoreceptor protein. *Science*, 242:1290, 1988.

83. E.Weber, E.Papamokos, W.Bode, R.Huber, I.Kato, and M. Laskowski. Ovomucoid, a Kazal-type inhibitor, and model building studies of complexes with serine proteases. *J. Mol. Biol.*, 158:515, 1982.

84. T.O.Fischmann and R.J.Poljak. Crystallographic refinement of the three-dimensional structure of FAB D1.2 Lysozyme complex at 2.5 Angstroms. *J. Biol. Chem.*, 266:12915, 1991.

(1)	(2)	SD score	Fold (1)	Fold(2)
1eca[75]	2lhb[52]	5.12	Globin-like	Globin-like
1azu[76]	2pcy[53]	5.40	Cupredoxins	Cupredoxins
2rspa[77]	5hvpa[56]	5.81	Acid proteases	Acid proteases
1paz[78]	2pcy[53]	7.22	Cupredoxins	Cupredoxins
2lhb[52]	4sdha[79]	7.70	Globin-like	Globin-like
1cdta[80]	3ebx[54]	8.26	Snake toxin-like	Snake toxin-like
3cln[81]	4cpv[55]	8.27	EF Hand-like	EF Hand-like
2gbp[82]	8abp[57]	8.86	Periplasmic binding	Periplasmic binding
1ovoa[83]	1tgsi[51]	9.45	Ovomucoid/PCI-1 like	Ovomucoid/PCI-1 like
1fdlh[84]	1mcpl[50]	12.66	Immunoglobulin	Immunoglobulin
4cms[48]	5er2e [49]	15.98	Acid proteases	Acid proteases

Table 1: Pairs in the RS126 set that have an SD score of greater than 5. Alignments were generated by the AMPS package[6] a blosum62 matrix, and gap penalty of 10, with 100 randomisations. Fold definitions come from the current release (1.37) of the SCOP database [47]

	DSSP[38]	STRIDE[40]	DEFINE[39]
Helix	28.9	29.8	30.2
Sheet	22.9	24.4	30.0
Coil	48.1	45.8	39.7

Table 2: Percentages of secondary structural state per secondary structure definition method

State	Method	Min	Mean	Max	Total number of secondary structures
Helix	DSSP[38]	3	9	54	817
	STRIDE	2	10	51	753
	DEFINE[39]	5	14	65	553
Strand	DSSP	1	4	19	1302
	STRIDE[40]	1	4	19	1303
	DEFINE	4	6	26	1030

Table 3: Ranges of length for secondary structural elements as defined by DSSP[38] STRIDE[40] and DEFINE[39] for the RS126 set

	Ave. % ID. between sequences	*Ave. sequence length*	*Ave. No of sequences per alignment*
CB396 set	34	157 residues	18
RS126 set	31	185 residues	30

Table 4: Summary statistics of the alignments used in the predictions

Class definition	*RS126 set No. (%)*	*CB396 set No. (%)*
Alpha and beta (a/b)	25 (20)	107 (27)
Alpha and beta (a+b)	17 (13)	101 (26)
All alpha	27 (21)	68 (17)
All beta	38 (20)	78 (20)
Multi-domain	3 (2)	0 (0)
Small proteins	18 (14)	27 (7)
Membrane	1 ($\leq$1)	3 ($\leq$1)
Peptides	1 ($\leq$1)	12 (3)

Table 5: Data for class types used for the predictions

Method	*RS126 Protein set*		*CB396 protein set*	
	Q3	SOV	Q3	SOV
PHD[64]	73.5	73.5	71.9	75.3
DSC[29]	71.1	71.6	68.4	72.0
PREDATOR[28]	70.3	69.9	68.6	69.8
NNSSP[30]	72.7	70.6	71.4	71.3
CONSENSUS	74.8	74.5	72.9	75.4

Table 6: Q_3 and segment overlap results for the set of RS126, and CB396 proteins

	p-Value cutoff 10^{-10}	p-Value cutoff 10^{-2}
Total Number of Residues	1716356	2013632
Total Number of Sequences	7013	8974
Average Number of Sequences per Family	55.6	71.2

Table 7: Family size for the automatically generated alignments for the RS126 protein set, considering 2 levels of BLAST[59] p-value cutoff

	p-value cutoff	10^{-10}	p-value cutoff	10^{-2}
Method	*DSSP*	*STRIDE*	*DSSP*	*STRIDE*
PHD[64]	72.4	72.4	73.2	73.2
DSC[29]	70.2	70.0	71.0	70.7
PREDATOR[58]	69.7	69.3	70.7	70.3
NNSSP [30]	71.8	71.2	72.4	71.7
MULPRED	66.7	65.4	67.2	66.8
ZPRED [26]	65.5	64.7	66.7	65.9
CONSENSUS	**73.9**	73.7	**74.5**	74.3

Table 8: Comparison of the of Q_3 accuracy for a decrease in the BLAST[59] P-value cut-off from 10^{-10} to 10^{-2} with the RS126 set. (The alignments used for these predictions did not use a percentage identity filter)

	100%	95%	80%	75%	60%
PHD[64]	73.2	73.3	73.4	73.5	73.3
DSC[29]	71.0	71.0	71.0	71.1	70.9
PREDATOR[28]	70.7	70.4	70.1	70.3	70.4
NNSSP[30]	72.4	72.5	72.7	72.7	72.8
CONSENSUS	74.5	74.5	74.6	74.8	74.7
Total No. of Sequences	8974	5907	4320	3833	2681
Ave No. of Sequences/Alignment	71	47	34	30	20

Table 9: Effect on Q_3 accuracy by removing all sequences similar to the query at different % identity thresholds, using the RS126 protein set

	Mean % of Helix	Mean % of Sheet	Mean % of Coil
Method A	28.9	22.9	48.1
Method B	25.3	21.2	52.6
Method C	25.6	21.2	52.3

Table 10: Mean percentages of secondary structure state defined by DSSP[38] when different 8 to 3 state reduction methods are used

Change	Q_3
Reduction method A[64]	74.8
B $\rightarrow$ Coil only	75.7
G $\rightarrow$ Coil only	76.6
B and G $\rightarrow$ Coil	77.5
GGGHHHH $\rightarrow$ HHHHHHH, B and G $\rightarrow$ Coil (Method C)[30]	77.5
B, G and HHHH EE $\rightarrow$ Coil (Method B)[58]	77.9

Table 11: Changing 8 to 3 state reduction, for DSSP and resultant Q_3 accuracy for the consensus method, based on the RS126 set of proteins

Method	*DSSP (A)*	*STRIDE (A)*	*DSSP (B)*	*STRIDE (B)*
PHD[64]	73.5	73.5	76.3	76.3
DSC[29]	71.1	70.9	73.3	73.4
PREDATOR[28]	70.3	69.6	75.2	74.0
NNSSP[30]	72.7	72.2	77.3	76.5
CONSENSUS	**74.8**	74.7	**77.9**	77.9

Table 12: Results for the RS126 protein set, by reducing the definition to 3 state by methods A and B

Method	*RS126*	*CB396*	*Author*
PHD[64]	76.3	74.2	-
SIMPA[72]	67.3	67.6	67.7
GOR IV[74]	53.3	64.6	64.4
SOPM[73]	66.8	64.7	69.0

Table 13: Results for single sequence prediction methods via a full jack-knife test. The column 'Author' is the authors jack-knife value for the method with their dataset, and definition reduction method. All results are calculated using reduction method A, and also converting G and B states to coil. For PHD[64] the authors quote 71.6% as their cross-validated accuracy. However, G and B states were considered in the accuracy calculation for PHD[64]

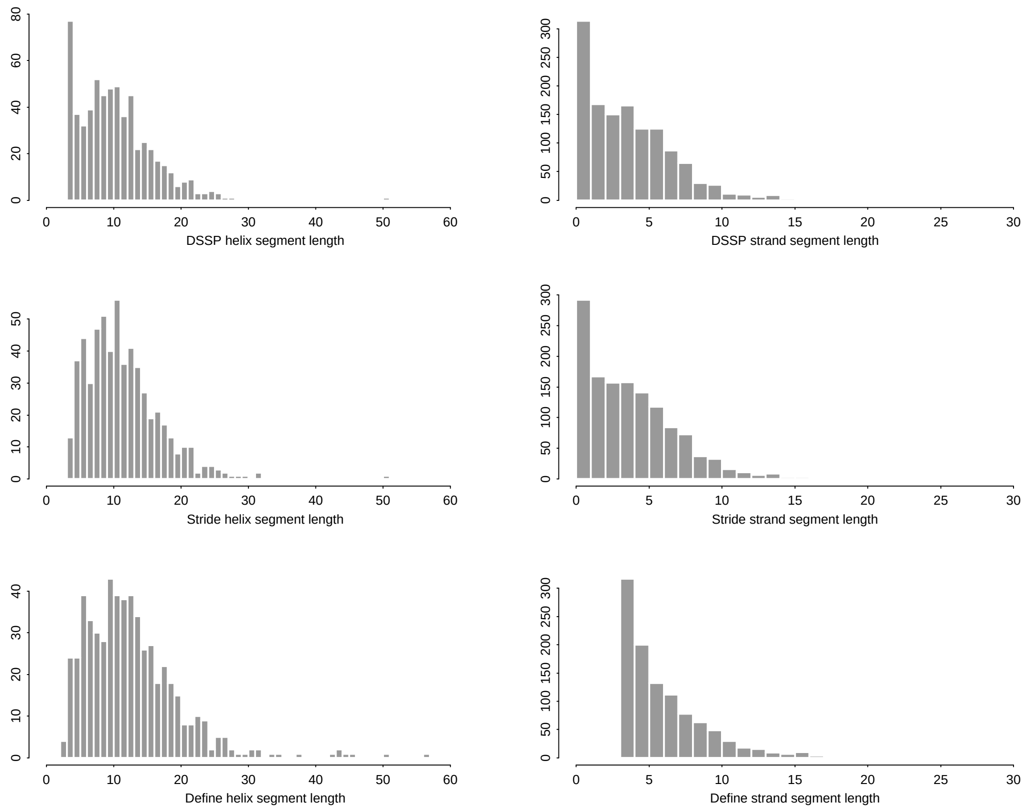

Figure 1: Comparison of segment length distributions for each definition method