Redefining the Goals of Protein Secondary Structure Prediction

Burkhard Rost, Chris Sander and Reinhard Schneider

EMBL Heidelberg
Meyerhofstraße 1, 69117 Heidelberg, Germany

Secondary structure prediction recently has surpassed the 70% level of average accuracy, evaluated on the single residue states helix, strand and loop (Q3). But the ultimate goal is reliable prediction of tertiary (three-dimensional, 3D) structure, not 100% single residue accuracy for secondary structure. A comparison of pairs of structurally homologous proteins with divergent sequences reveals that considerable variation in the position and length of secondary structure segments can be accommodated within the same 3D fold. It is therefore sufficient to predict the approximate location of helix, strand, turn and loop segments, provided they are compatible with the formation of 3D structure. Accordingly, we define here a measure of segment overlap (Sov) that is somewhat insensitive to small variations in secondary structure assignments. The new segment overlap measure ranges from an ignorance level of 37% (random protein pairs) via a current level of 72% for a prediction method based on sequence profile input to neural networks (PHD) to an average 90% level for homologous protein pairs. We conclude that the highest scores one can reasonably expect for secondary structure prediction are a single residue accuracy of Q3 > 85% and a fractional segment overlap of Sov > 90%.

Keywords: secondary structure prediction; prediction accuracy; secondary structure segments; evaluation; homologous proteins

1. Introduction

(a) Simplify the prediction problem

Protein three-dimensional (3D†) structure is determined by the sequence (Epstein et al., 1963; Anfinsen, 1973). The 3D structure of a new sequence can be predicted from the sequence fairly accurately if a homologue with significant sequence similarity exists in the data bank of experimentally solved 3D structures (Chothia & Lesk, 1986; Taylor & Orengo, 1990a; Overington et al., 1990; Summers & Krüger, 1990; Schneider & Sander, 1991; Friend & Sander, 1991; Holm & Sander, 1992; Levitt, 1992; Taylor, 1992). However, for probably more than 80% of the proteins with known sequence (Bairoch & Boeckmann, 1992), there is no homologue of known 3D structure (Bernstein et al., 1977; Schneider & Sander, 1991). For these proteins the prediction of the 3D structure poses insurmountable difficulties. The way out is to reduce the problem to a simpler one that is amenable to a partial solution.

(b) Secondary structure at 70% single residue accuracy

One way of simplifying the prediction problem is to project the very complicated 3D structure onto one dimension, i.e. onto a string of secondary structure assignments for each residue. Such a reduction is possible, because proteins form local conformational patterns, such as helices and strands (Pauling & Corey, 1951; Schirmer, 1979; Brändén & Tooze, 1991). Various prediction methods have been developed over the last two decades (Pain & Robson, 1970; Finkelstein & Ptitsyn, 1971; Robson & Pain, 1971; Kabat & Wu, 1973a,b; Nagano, 1973, 1977; Burgess et al., 1974; Chou & Fasman, 1974, 1978; Lim, 1974; Nagano & Hasegawa, 1975; Maxfield & Scheraga, 1976, 1979; Chou & Fasman, 1978; Robson, 1976; Nagano, 1977; Garnier et al., 1978; Cohen et al., 1983, 1986; Ptitsyn & Finkelstein, 1983; Taylor & Thornton, 1983; Gibrat et al., 1987; Zvelebil et al., 1987; Biou et al., 1988; Gasco & Golmard, 1988; Qian & Sejnowski, 1988; Holley & Karplus, 1989; Taylor & Orengo, 1990b; King & Sternberg, 1990; Kneller et al., 1990;
Nishikawa & Noguchi, 1991; Rooman et al., 1991; Muggleton et al., 1992; Salzberg & Cost, 1992; Zhang et al., 1992; Maclin & Shavlik, 1993; Rost & Sander, 1993a). The implied goal for such methods is to reach 100% accuracy, evaluated in terms of single residue states. For three-state predictions an average accuracy of $Q_3 > 70\%$ has been reached, an improvement of about five percentage points (Rost & Sander, 1993b). What is next?

(c) Function and 3D structure are more conserved than secondary structure

The ultimate goal of secondary structure prediction is to predict those aspects of the 3D structure that are important for function. Studies of protein evolution show that the overall 3D fold and the precise position of functional residues are conserved, and thus important. The record of evolution also reveals that there is considerable variation of secondary structure within one 3D family. So the precise extent of secondary structure is apparently not essential for the formation of 3D structure. By analogy, for a correct prediction of the 3D fold it may be sufficient to predict secondary structure at less than 100% accuracy.

(d) Comparison of secondary structure in proteins of known structure

A detailed comparison of the secondary structure of proteins belonging to the same structural family shows (1) that a reasonable goal for secondary structure prediction is not to arrive at 100% in the overall three-state per-residues accuracy, but to reach some 80 to 85%, and (2) that the per-residue comparison is not sufficient to assess the presence of segments in 3D (see section 2). In terms of a segment measure, defined here, 3D homologue pairs are closer to 100% identity in their secondary structure strings, but there is still some variation (see section 3). This variation remains even if the comparison is restricted to the cores of the proteins (see section 4).

2. Single Residue Measure: Proteins with the Same 3D Fold Differ by 12% in Secondary Structure

(a) A test set of pairs of proteins of similar 3D structure

Do proteins with similar 3D structure have identical secondary structure? A simple way to check this is to compare proteins within one structural family. A structural family can be defined as consisting of proteins that have the same 3D fold, as judged either visually or by structural alignment (Holm et al., 1993). Here, we judge structural similarity on the basis of sequence criteria, as calibrated on structural alignment in earlier work (Chothia & Lesk, 1986; Schneider & Sander, 1991). To assemble a test set of pairs of proteins of similar structure we took all protein chains from Protein Data Bank (PDB) that have a sequence identity exceeding 30% relative to a representative set of protein chains (Hobohm et al., 1992). Two examples are shown in Figure 3. The representative set is the same as that used in an earlier study for the evaluation of the prediction method PHD (Rost & Sander, 1993b). The test set of proteins thus assembled (dubbed PDB95, Table 1) comprises 140 aligned pairs. The secondary structure state was assigned based on the 3D co-ordinates according to DSSP (Kabsch & Sander, 1983a). To allow residue by residue comparison, the secondary structures were brought into reliable alignment using the amino acid sequences. The alignment was performed using a standard dynamic programming alignment algorithm, with insertions and deletions confined to loops and the ends of helices and strands (Schneider & Sander, 1991). Although sequence alignment has some inherent inaccuracies compared to 3D structure alignment, the accuracy of alignment was deemed sufficient for the purposes of the statistical investigation of this paper.

(b) Similar 3D structure, yet different secondary structure, 12% on average

For the 140 protein pairs, the percentage of identical secondary structure symbols between two strings is $Q_3 = 88.4\%$ (Table 1: as a control, the result of alignments between proteins of dissimilar 3D structure, dubbed RAN, is also given). A surprising result is that the secondary structure identity varies considerably around this average, with a standard deviation of nine percentage points (distribution in Fig. 2(a)). This deviation is an intrinsic feature of protein families (Chothia & Lesk, 1986). The value is comparable to the standard deviations of single residue accuracy for various prediction methods (Robson & Garnier, 1993; Rost & Sander, 1993b). Methods that predict an average secondary structure for a family of homologous sequences therefore cannot be expected to be more precise than the natural variation observed in structural families.

(c) Unavoidable variation in secondary structure

What is the reason for 12% average dissimilarity in secondary structure symbols for homologous 3D structures? The principal source of variation is the plastic response of protein structures to variations in amino acid sequence within one structural family (Chothia & Lesk, 1986). The variation in secondary structure is attributed to a sensitivity of the backbone to interactions between residues far apart in sequence and close in 3D (tertiary interactions: Kotelnchuck & Scheraga, 1968, 1969; Anfinsen & Scheraga, 1975; Robson, 1976; Maxfield & Scheraga, 1979). The formation of secondary structure can be thought of to be influenced by two terms:

$$C = C_{local} + C_{global}.$$
Table 1

<table>
<thead>
<tr>
<th></th>
<th>N(^{t})</th>
<th>N(_{prot}\ber^{g})</th>
<th>Q(_{3}\ber)</th>
<th>Q(_{5}\ber)</th>
<th>Q(_{p}\ber)</th>
<th>Q(_{L}\ber)</th>
<th>Corr(_{h}\ber)</th>
<th>Corr(_{p}\ber)</th>
<th>Corr(_{L}\ber)</th>
<th>Info\ber)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB98\ber*</td>
<td>24291</td>
<td>140</td>
<td>88.4</td>
<td>88</td>
<td>99</td>
<td>0.84</td>
<td>0.85</td>
<td>0.77</td>
<td>0.616</td>
<td></td>
</tr>
<tr>
<td>PDB98 core\ber*</td>
<td>6948</td>
<td>140</td>
<td>91.7</td>
<td>92</td>
<td>91</td>
<td>0.88</td>
<td>0.87</td>
<td>0.83</td>
<td>0.882</td>
<td></td>
</tr>
<tr>
<td>RAN\ber*</td>
<td>12162</td>
<td>94</td>
<td>35.2</td>
<td>23</td>
<td>26</td>
<td>47</td>
<td>-0.09</td>
<td>0.01</td>
<td>0.04</td>
<td>0.006</td>
</tr>
<tr>
<td>PHD\ber#</td>
<td>23209</td>
<td>126</td>
<td>70.8</td>
<td>72</td>
<td>66</td>
<td>72</td>
<td>0.40</td>
<td>0.52</td>
<td>0.51</td>
<td>0.261</td>
</tr>
<tr>
<td>PHD core\ber#</td>
<td>6175</td>
<td>126</td>
<td>72.5</td>
<td>70</td>
<td>72</td>
<td>74</td>
<td>0.62</td>
<td>0.57</td>
<td>0.55</td>
<td>0.287</td>
</tr>
</tbody>
</table>

\*PDB98, a set of 126 pairwise non-homologous protein chains is taken from protein data bank of 3D structures. PDB is scanned for sequences homologous to the 126 chains. PDB98 contains all pairs with an alignment length of >30 residues, and the sequence identity <98% and above the length dependent cut-off given by HSSP for similarity in 3D structure (Schneider & Sander, 1991).

\*PDB98 core. values for the core segments of PDB98, i.e. all segments for which the relative accessibility of at least half of all residues is <=0.10 (=accessibility/maximal accessibility, where the values have been taken from DSSP (Kabsch & Sander, 1983a).

\#RAN, from the set of 126 proteins alignment pairs were taken that had between 5 and 10% pairwise sequence similarity over an alignment length of more than 80 residues. Thus, these pairs are largely dissimilar in their 3D structure.

\#PHD, neural network system prediction of secondary structure in 3-states. PHD was tested with multiple cross-validation on the same set of unique 126 chains used to generate the set of PDB pairs (Rost & Sander, 1993b).

\#PHD core, values for the core segments of PHD (see above).

\(N\), number of residues in the data set; \(N_{prot}\ber\), number of proteins in the data set.

\(Q_{3}\ber\), \(Q_{5}\ber\), \(Q_{p}\ber\), \(Q_{L}\ber\), percentages of per-residue identity between 2 secondary structure strings for all 3 states, and for helix, strand and loop.

\(h(Corr_{h}\ber\), Corr\(_{p}\ber\), Corr\(_{L}\ber\), give the Matthews correlation coefficients (Matthews, 1975).

\(q_{nfo}\ber\), measures the information defined by:

\[
\text{Info} = 1 - \frac{\sum_{i=1}^{N} a_{i} \ln a_{i} - \sum_{i=1}^{N} A_{i} \ln A_{i}}{N \ln N - \sum_{i=1}^{N} b_{i} \ln b_{i}}
\]

where \(N\) is the number of residues in the data bank, \(a_{i}\), the number of residues predicted to be in secondary structure \(i\), \(b_{i}\) the number of residues observed to be in \(i\), and \(A_{i}\) the number of residues predicted to be in \(i\) and observed to be in \(j\) (Rost & Sander, 1993b).

where \(C\ber\) is the secondary structure conformation of a protein, \(C_{local}\ber\) describes the part that is determined by local interactions, and \(C_{global}\ber\) the one influenced by global interactions. Of course, the assumption of additivity is nothing but a rough first order approximation. Given this concept, the variation in secondary structure between homologues can be split into two terms:

identity in secondary structure for a pair of homologous proteins = identity in locally formed secondary structure + identity in globally formed secondary structure.

The hypothesis is that the differences in secondary structure for pairs from the same 3D family attribute mainly to the second term (Pohl, 1971, 1980; Robson, 1974; Robson & Pain, 1974a,b,c).

In addition, the concept of secondary structure is somewhat imprecise. In particular, different authors arrive at different secondary structure assignments for the same protein, either by visual inspection or by different feature extraction algorithms (Sklenar et al., 1989; Woodcock et al., 1992; Colloc’h et al., 1993). To reduce the influence of this imprecision, we use here an automatic method of secondary structure assignment from the 3D co-ordinates that is based on detection of periodic repeats of hydrogen bonded structure (Kabsch & Sander, 1983a). However, as a result of a sharp cutoff in the hydrogen bond energy, very small variations in 3D co-ordinates can lead to the addition or deletion of one or more residues at the ends of a helix or strand. Even for sequence-identical structures, such variations can occur as the result of crystal packing in different crystal forms or as a result of experimental error in structure determination (Brändén & Jones, 1990). All of these factors also contribute to the variation of secondary structure assignments for sequence-dissimilar proteins within one structural family and thus introduce an annoying technical complication into the field of secondary structure prediction. The main variation, however, occurs as the result of sequence changes between homologous proteins. Any method that predicts an average structure for an entire family has to take this inherent imprecision into account, on average nine percentage points in the single residue measure (\(Q_{3}\ber\)). We now discuss if the comparison of secondary structure in terms of segments can reduce this variation of 12(\(\pm\)9)% (Fig. 1).


(a) Simple assessment of segment accuracy

There are three simple measures for assessing the quality of predicted secondary structure segments:
Figure 1. Goal of segment based measures: reduction of variation in secondary structure identity for protein pairs of the same structure family. The secondary structure strings differ for protein pairs from the same structure family. One goal for defining segment based measures is that this variation is smaller for a segment than for a per-residue comparison. \( Q \), Variation in secondary structure within one protein family of the same structure (related to the family representative); \( O \), representative of structure family (used as reference for computation of the variation in secondary structure); \( Q \), sequences with homologous 3D structure to the representative; outer circle, variation in secondary structure for a per-residue comparison \( (Q_a) \); inner circle, variation in secondary structure for a segment based comparison; ~, the arrows indicate that the segment-based measure should be chosen such that the variation (radius of circle) decreases.

(1) the number of segments in the protein, (2) the average segment length \( <L> \), for \( i = \alpha \) (helix), \( \beta \) (strand) and \( L \) (loop), i.e. the average number of residues in a helix (= 8.5 for PDB98), strand (5-1), or loop (6-0) segment, and (3) the distribution of the number of segments with length (histogram). These measures are related. They are useful in characterizing prediction methods, in particular, methods with fairly high single residue accuracy \( (Q_2) \), yet an unrealistic distribution of segments \( (Rost \& Sander, 1993b) \).

(b) When are two segments identical?

A more complicated alternative is the count of identical segments. However, the attempt immediately leads to difficulties. In the example of Figure A1(a), should the helix segments in prediction 1 and 2 both be counted as identical to the observed helix? Or neither of the two? How should the two helical fragments in prediction 3 be evaluated? There is obviously more than one way of evaluating agreement of different segment assignments. As a result, no generally accepted segment measure has emerged in the literature, although the need for evaluating secondary structure on a segment base has been pointed out by quite a number of authors \( (Cohen \ et \ al., \ 1983, \ 1986; \ Taylor, \ 1984; \ Taylor \ & \ Thornton, \ 1984; \ Cohen \ & \ Kuntz, \ 1989; \ Benner, \ 1992; \ Pressnell \ et \ al., \ 1992; \ Sternberg, \ 1992; \ Thornton \ et \ al., \ 1992; \ Benner \ et \ al., \ 1993) \). We have attempted to systematically investigate a number of different definitions. These are described in detail in the Appendices.

(c) A new measure: fractional segment overlap \( (S_{ov}) \)

The following segment measure strikes a good compromise between permissiveness and precision. It is based on several ideas: (1) allow some variation at the ends of segments; (2) provide a sliding scale of segment overlap that gives intuitively expected values in extreme cases; and (3) give low values for random predictions. The measure simply counts the fractional extent to which two segments overlap, with some allowance for non-matching residues at the ends. There is only one adjustable parameter, the amount of allowance.

In detail, given two strings of secondary structure symbols, we define (for illustration, Fig. A1(c); for a detailed definition, eqn (A3) in Appendix I):

\[
S_{ov} = \frac{1}{N} \sum_{s} \min_{\text{ov}}(s_1; s_2) + \delta \times \max_{\text{ov}}(s_1; s_2) 
\]

where \( N \) is the total number of residues. The subscript 1 or 2 labels either the proteins being compared or the observed (usually subscript 1) and predicted (usually subscript 2) structure. The sum is taken over all segment pairs \( s = \{s_1, s_2\} \), where \( s_1 \) and \( s_2 \) are two segments that have in common at least one residue position in the same secondary structure. Asymmetry between the two segments is introduced in that the weight \( \text{len}(s_1) \) is the length of \( s_1 \), usually chosen to be the segment in the experimental structure \( (S_{oy} \ of \ observed) \). The actual overlap between the two segments is \( \min_{\text{ov}} \), i.e. the number of residues for which both segments have, e.g. an \( \text{H} \) (helix) in common, while \( \max_{\text{ov}} \) is the total extent of both segments, i.e. the number of residues for which either of the two has, say, the assigned state \( \text{H} \) (Fig. A1(c)). The accepted variation \( \delta \) assures a ratio of 1:0 when there are only minor deviations at the ends of segments, as often observed in structural homologues \( (Taylor \ & \ Orengo, \ 1989a; \ Wilmot \ & \ Thornton, \ 1990; \ Perczel \ et \ al., \ 1992; \ Hutchinson \ & \ Thornton, \ 1993) \); it is chosen such to be smaller than \( \min_{\text{ov}} \) and smaller than half the length of segment \( s_1 \) \( (\delta = 1, 2, 3 \ for \ short, \ intermediate, \ long \ segments) \). The ratio \( \min_{\text{ov}}/\max_{\text{ov}} \) is constrained to a maximum value of 1:0, i.e. the allowance cannot lead to a "more than perfect" value of fractional overlap.

(d) Fractional segment overlaps are more informative than single residue accuracy

Helices and sheets for the most part constitute the core of a globular protein structure and are more
Figure 2. Distribution of secondary structure identity per alignment pair for PDB98. (a) The distribution of
the per-residue identity per PDB pair and (b) the identity in segments (Sov) are shown. The average identity in $Q_3$ per
pair is 88.8% with a standard deviation of 9.1% (c) and (d). The dependency of the scores on the sequence similarity for $Q_3$ and $Sov$
are illustrated (e).

conserved than loop regions in evolution. This is borne out when comparing the experimentally
known structure of homologues: the segment overlap $Sov$ of 3D pairs is clearly higher for helices
(91%) and for strands (93%) than for loop regions (87%). The average $Sov$ over all 140 pairs is about
90%. More importantly, the fact that $Sov$ scores are shifted toward the 100% mark relative to the single
residue scores (Fig. 1) indicates that the segment measure is more informative about the 3D structure
than per-residue ratios, as an ideal scoring system would give 100% for any pair with the same 3D
fold. Indeed, for some 10% of the protein pairs the segment overlap $Sov$ becomes 100%, a value never
reached for the single residue measure $Q_3$, and for the majority of the pairs $Sov$ is larger than 95% 
(Fig. 2(a) to (e)). The two examples in Figure 3 clearly illustrate the advantage of comparing the
secondary structure on the basis of segments rather than single residues.

4. The Cores of Proteins with the Same 3D Fold
Differ Less in Secondary Structure

Intuition suggests that the variation in secondary
structure within a structure family is less in the core of the protein and that restricting secondary struc-
ture comparison to protein cores may be a good way
to capture the essence of 3D similarity. We have
quantified this intuitive idea and show that the
effect is real, but not large enough to warrant defini-
tion and standard usage of a core-based measure.
Protein Secondary Structure: Redefining Prediction

(a) A simple definition of protein cores

The 30% least exposed segments were taken to represent protein cores. We selected all segments for which the relative solvent accessibility was <0.10 for more than half of the residues of that segment (values defined as in DSSP: Kabach & Sander, 1983a; Baumann et al., 1989). The secondary structures of the 140 protein pairs of similar 3D fold (PDB98) were then evaluated in terms of single residue identity and segment overlap.

(b) Less structural variation in protein cores

The segment overlap scores increased by about three percentage points (Table 1, Fig. 4) which leads to \( \text{Sov} = 93\% \) averaged over the three secondary structure states, and to 98\% for helices and 93\% for sheets (data not shown). By restricting the evaluation to protein core, the scores of helices and sheets increased more than is the case for single-residue comparisons. However, a measure yielding 100\% for pairs from the same structure family is not achieved this way.

(c) Restrict evaluation of prediction methods to the protein core?

When the evaluation of secondary structure is applied to comparison of observed and predicted strings, the scores for the core segments are also higher than for the protein as a whole (a similar increase in accuracy for predicting core residues is also described for a different method: Rooman et al., 1992; Rooman & Wodak, 1992). However, the increase is smaller for the prediction method than for the 3D similar pairs (Table 1, Fig. 4). This result
confirms the fact that any secondary structure prediction method is on average worse in capturing 3D information than 3D structure predicted on the basis of alignment of clearly homologous sequences. In summary, we do not see a strong reason to introduce the comparison of cores as a standard for evaluation of secondary structures.

5. Conclusions

(a) From 100% to 88%

First, careful comparison of the secondary structure assignments of structural homologues, i.e. of proteins with essentially identical 3D structure, reveals a striking deviation from 100% identity: the per-residue identity in secondary structure for 140 pairs of homologous structure is about \( Q_3 = 88(\pm 9)\% \). This value sets an upper limit in per-residue accuracy for what secondary structure prediction can reasonably be expected to achieve. A lower limit is given by a comparison of the secondary structure of 3D dissimilar sequences: 35%. So the target range for secondary structure accuracy for single residue counts is redefined from the previous range 35% < \( Q_3 < 100\% \) to the new range 35% < \( Q_3 \leq 88\% \) (Fig. 5).

(b) From single residue measure to segment measure

Second, segment based comparisons capture better the reality of secondary structure segments as flexible elements than do per residue accuracies. To select a “best” measure for segment comparison, we have compared different ways of defining segment measures. The best trade-off between the goal of shrinking the variation in secondary structure segments as close as possible to the 90% value for the segment overlap.

**Figure 5.** Redefining the goal of secondary structure prediction. Secondary structure prediction methods (like PHD) range between the lower limit given by a random prediction of 35 to 37% and the upper limit given by the comparison of pairs from the same 3D family, 88 to 90%. The goal of prediction is not to reach 100%, but to come as close as possible to the 90% value for the segment overlap.

**Figure 6.** Per-residue and segment comparison of prediction accuracy. Each point in the scatter plot represents the 2 accuracy values for a particular prediction method: per-residue accuracy \( Q_3 \) and segment overlap \( S_{ov} \). The 2 series of points (circles and triangles) are for 2 ways of calculating the segment accuracy: for \( S_{ov}^{obs} \), the average over segments is weighted with the length of observed segments (eqn (1)), for \( S_{ov}^{pred} \) of the predicted segments. The \( S_{ov}^{pred} \) points refer to the probability that predicted segments are correct. The data set and the methods are given in the legend to Table 2.

The following questions remain open for further analysis. Does the deviation of \( S_{ov} \) from the ideal
value of 100% decrease if instead of a particular method for the assignment of secondary structure (here DSSP) a consensus of various methods, as recently analysed by Colloc'h et al. (1993), is used? Can the residual variation be regarded as a "non-stationary", "uniform noise" (Robson, 1974; Robson & Pain, 1974a,b,c)? Does a measure exist that is better in capturing the aspects of 3D structure contained in one-dimensional strings of secondary structure than the combination of per-residue and segment measures proposed here? Our analysis shows that it is not straightforward to develop such measures, and that some simple measures (\(Q_{\text{Taylor}}\), \(Q_{\text{Taylor}}\), see Appendix I and II) tend to overestimate the performance of prediction methods.

(c) The ultimate goal of pure secondary structure predictions?

For more than two decades researchers have attempted to reach secondary structure prediction of 100% accuracy. Now that a single residue accuracy of \(Q_s = 70\%\) has been exceeded, the goal is shifting. Reaching 100% in per-residue score is not a reasonable goal of secondary structure. Instead, we propose to redefine the goal in terms of a segment and a per-residue score. On the way to tertiary structure prediction, a reasonable next goal of secondary structure methods is to reach three-state per-residue and segment scores closer to 80% (on a representative data set of at least 100 unique proteins and with multiple-cross validation).

### Table 2

<table>
<thead>
<tr>
<th>Measure</th>
<th>Segment*</th>
<th>GORIII*</th>
<th>SIMPA*</th>
<th>PHD*</th>
<th>PDB98</th>
<th>RAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_s)</td>
<td>59.9</td>
<td>58.3</td>
<td>60.7</td>
<td>72.5</td>
<td>88.4</td>
<td>35.2</td>
</tr>
<tr>
<td>(Q_s)</td>
<td>62.1</td>
<td>59.6</td>
<td>62.7</td>
<td>72.8</td>
<td>88.0</td>
<td>23.6</td>
</tr>
<tr>
<td>(Q_s)</td>
<td>41.3</td>
<td>36.6</td>
<td>35.2</td>
<td>38.4</td>
<td>88.0</td>
<td>26.6</td>
</tr>
<tr>
<td>Info</td>
<td>0.18</td>
<td>0.10</td>
<td>0.12</td>
<td>0.20</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>(S_{\text{helix}})</td>
<td>61.2</td>
<td>57.2</td>
<td>61.7</td>
<td>75.6</td>
<td>80.7</td>
<td>30.6</td>
</tr>
<tr>
<td>(S_{\text{helix}})</td>
<td>57.9</td>
<td>54.4</td>
<td>58.3</td>
<td>75.3</td>
<td>92.2</td>
<td>34.2</td>
</tr>
<tr>
<td>(S_{\text{str}})</td>
<td>72.8</td>
<td>67.6</td>
<td>69.3</td>
<td>82.2</td>
<td>94.3</td>
<td>47.1</td>
</tr>
<tr>
<td>(S_{\text{str}})</td>
<td>30.6</td>
<td>20.1</td>
<td>27.4</td>
<td>44.0</td>
<td>73.2</td>
<td>13.7</td>
</tr>
<tr>
<td>(S_{\text{Taylor}})</td>
<td>62.9</td>
<td>57.9</td>
<td>62.2</td>
<td>83.2</td>
<td>92.8</td>
<td>38.4</td>
</tr>
<tr>
<td>(&lt;L&gt;_p)</td>
<td>13.3</td>
<td>4.3</td>
<td>6.0</td>
<td>11.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;L&gt;_p)</td>
<td>4.9</td>
<td>2.5</td>
<td>3.0</td>
<td>5.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX I

Definition of Three Measures Based on Secondary Structure Segments

Here, we define three measures for the overlap of segments that are easy to implement. First, a loose criterion \( (\text{ov}^{\text{loose}}) \) describing what an expert upon sight would term as about identical (a comparable measure was proposed by others: Zhang et al., 1992; Gerloff et al., 1993; Stultz et al., 1993). Second, a strict criterion \( (\text{ov}^{\text{strict}}) \) for matching segments which treats the segments as fixed objects accepting marginal deviation at the edges (Fig. A1(b), an even more stringent measure has been introduced for helix/non helix predictions by Presnell et al. (1992)). Third, the fractional overlap \( (\text{Sov}) \), e.g. for a helical segment, this is the number of residues for which there is a helix in both strings divided by the number of residues for which either of the two strings has a helix (Fig. A1(c)). Additionally, a measure introduced earlier in the literature \( (\text{ov}^{\text{Taylor}}) \) is described.

### (a) Loose criterion for overlapping segments \( \text{ov}^{\text{loose}} \)

The loose criterion for matching segments is the length weighted percentage of all roughly overlapping segments:

\[
\text{ov}^{\text{loose}} = \frac{1}{|N|} \sum_s \theta^{\text{loose}}(s_1, s_2) \cdot |\text{len}(s_1)|, \tag{A1}
\]

where \( N \) is the number of all residues of the protein, or the aligned fragment, \( \text{len}(s_i) \) the length of segment \( s \) for sequence \( i \). \( \theta^{\text{loose}} \) is a step function, which for helix and strand is given by:

\[
\theta^{\text{loose}}(s_1, s_2) = \begin{cases} 
1, & \text{if segment } s \text{ of sequence } 2 \text{ overlaps with} \\
& \text{at least half of the segment } s \text{ of sequence } 1 \\
0, & \text{else,}
\end{cases}
\]

and for loop regions:

\[
\theta^{\text{loose}}(s_1, s_2) = \begin{cases} 
1, & \text{if the loop } s \text{ of sequence } 2 \text{ overlaps with at least} \\
& \text{two residues in the loop region } s \text{ of sequence } 1 \\
0, & \text{else.}
\end{cases}
\]

The distinction between helix and strand on the one hand and loop on the other, stems from the evidence that loop regions are more variable inside one structure family than helices and strands (Brändén & Tooze, 1991; Lesk, 1991). A consequence of the definition is that two pieces of helix are not regarded as a consecutive segment even if the gap in between is small. For example, the example for prediction 3 in Figure A1(a) yields \( \text{ov}^{\text{loose}} = 0\% \) because none of the two helix segments overlap with three or more residues of the observed helix. The other two predictions in Figure A1(a) result in 100\% correct helix segments.

### (b) Strict criterion for overlapping segments \( \text{ov}^{\text{strict}} \)

A far more strict criterion for correctly matching segments is:

\[
\text{ov}^{\text{strict}} = \frac{1}{|N|} \sum_s \theta^{\text{strict}}(s_1, s_2) \cdot |\text{len}(s_1)|, \tag{A2}
\]

with the step function here chosen as:

\[
\theta^{\text{strict}}(s_1, s_2) = \begin{cases} 
1, & \text{if: } |\text{len}(s_1) - \text{len}(s_2)| \leq \delta^{\text{strict}} \text{ and} \\
& |\text{beg}(s_1) - \text{beg}(s_2)| \leq \delta^{\text{strict}} \text{ and} \\
& |\text{end}(s_1) - \text{end}(s_2)| \leq \delta^{\text{strict}}, \\
0, & \text{else,}
\end{cases}
\]

where \( \text{beg}(s_i) \) marks the first residue of segment \( s \) in sequence \( i \), \( \text{end}(s_i) \) the last residue of this segment. \( \delta^{\text{strict}} \) is the accepted deviation. The step function \( \theta^{\text{strict}} \) assures that only segments with equivalent length (\( \pm \delta^{\text{strict}} \)) and with exactly equal placement.
Protein Secondary Structure: Redefining Prediction

(±δstrict) are counted. We set δstrict to:

<table>
<thead>
<tr>
<th>len(s₁)</th>
<th>1 to 5</th>
<th>6 to 10</th>
<th>11 to 15</th>
<th>15 to 20</th>
<th>20 to 25</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>δstrict</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

The acceptance of a deviation is justified by the potential ambiguity of assigning the secondary structure at the edges of segments. Examples 1 and 3 in Figure A1(a) yield ovstrict = 0%, example 2 ovstrict = 100%.

Both, the loose (ovloose) and the strict (ovstrict) measure for matching segments require arbitrary choices for free parameters (θloose, θstrict, δstrict). Furthermore, two predicted segments of, e.g. helix in one string with one loop residue in between are not evaluated as one. Thus, one of the helical pieces is effectively ignored by both measures.

(c) Fractional overlap of segments Sov

A less arbitrary measure is the fraction of the number of residues that overlap in the two segments and the number of residues spanned by both segments (Fig. A1I(c)):

\[
Sov = \frac{1}{N} \sum_i \min\{\text{end}(s_1); \text{end}(s_2)\} - \max\{\text{beg}(s_1); \text{beg}(s_2)\} + 1 + \delta \times \text{len}(s_1). \tag{A3}
\]

where min \{a; b\} is the minimum of the two integers a and b, and max \{a; b\} the maximum. δ is a parameter for the accepted (maximal) deviation capturing the observation that definition of secondary structure at the edges of segments might not sufficiently capture the 3D reality (Taylor & Orengo, 1989a; Wilmot & Thornton, 1990; Perczel et al., 1992; Hutchinson & Thornton, 1993). δ is either = 0 (Sov0), or it is restricted by (Sov50%):

\[
\delta \leq \min\{\frac{\text{maxov}(s_1, s_2) - \text{minov}(s_1, s_2)}{2}; \text{minov}(s_1, s_2); \frac{\text{len}(s_1)}{2}\}.
\]

The first describes the correctness in predicting the observed segments, the second the probability that a predicted segment is correct (Kabsch & Sander, 1983b).

(d) Overlapping segments defined by Taylor ovTaylor

An algorithm for an alternative segment accuracy measure has been developed by Taylor (1984). In his procedure, the first step is to exclude too short segments, i.e. helices shorter than five residues and strands shorter than three residues. The second step is the following conversion for the predicted structure: ‘XXYXX’ → ‘XXXXX’, where either X = H and Y = E, or X = E and Y = H. The third step is to count (starting from the C terminus) identical segments of regular secondary structure (helix or strand) in the following way. Supposing, segment i of sequence 1 is an X (= H, or E). Then, if sequence 2 has at any position of that segment as well an X, the segment is counted as identical and the algorithm proceeds with the next segment i-1 in sequence 1. We shall refer to this measure as ovTaylor.

APPENDIX II

The Merits and Pitfalls of the Four Measures

The measures introduced in Appendix I interpret secondary structure segments as flexible objects that are counted as identical even if slightly displaced, stretched or compressed. Do these measures show that the difference in secondary structure for 3D homologous proteins is caused only by the edges of segments?

The first result is that for the PDB pairs, only ovloose (eqn (A1)) and Sov50% (eqn (A3)) reduce the variation in the secondary structure of 3D homologues. However, even for the strict measure ovstrict (eqn (A2)) 10% of all protein pairs score at 100% (data not shown), whereas no pair reaches a Q3 of 100% (Fig. A2).

The second result is that none of the measures reaches the ideal score of 100% (Fig. A2). Closest to this line come the loose criterion, ovloose = 94%. Does this mean that this measure is the best to reveal the degree of correlation between similarity in 3D and secondary structure? Or are the 90% values gener-
candidates introduced in Appendix I. For the fractional overlap 2 different choices for the accepted deviation (Sov°, and Sov5°%) are given. The measures are given as percentages.

Figure A2. Per-residue identity and segment similarity for PDB pairs. For all PDB pairs the per residue 3-state accuracy is compared to the 3 measures for matching segments introduced in Appendix I. For the fractional overlap 2 different choices for the accepted deviation (Sov°, and Sov5°%) are given. The measures are given as percentages.

Figure A3. Ratio of segment similarity and per residue identity. Plotted are the quotients: (segment measure/per-residue measure)-1. Thus positive values indicate that the segment measure scores higher than the single-residue quantity. The quotients are chosen to enable a comparison between results with different absolute values, e.g. the prediction method PHD reaches scores (averaged over the data set used) more than 10 percentage points below the values for PDB98 (abbreviations are as in Table 1). To make the Sov measure comparable, it also can be evaluated on helices and strands only, given as Sovh, s.

The conclusion is that segment criteria do not score at 100% for similar 3D proteins, but they capture information about the correlation between secondary and 3D structure that cannot be identified on the level of per-residue comparisons. An excellent prediction of three-state secondary structure would yield about the following scores: Qa > 85%, Info > 65%. For the relaxed fractional overlap, Sov5°% for helical pairs (PDB98) and 3D dissimilar pairs RAN score relatively higher (Fig. A3). For ov loose, this mainly stems from the increase for loop regions that are counted as identical if only two residues overlap (Fig. A2). Surprising is, that the acceptance of a deviation in Sov, although best reflecting the flexibility of the segments, seems to reduce the strength of this measure in extracting the information about the 3D relevant information contained in the secondary structure strings. All three, the loose criterion ov loose, ov Taylor, and the fractional overlap Sov5°% seem to have a tendency to introduce artifacts. However, for Sov5°% this tendency is still very weak. This is confirmed by the results for helix and strand only (Fig. A3).

The conclusion is that segment criteria do not score at 100% for similar 3D proteins, but they capture information about the correlation between secondary and 3D structure that cannot be identified on the level of per-residue comparisons. An excellent prediction of three-state secondary structure would yield about the following scores: Qa > 85%, Info > 65%. For the relaxed fractional overlap, Sov5°%, the prediction method PHD and the 3D dissimilar pairs RAN score relatively higher (Fig. A3). For ov loose, this mainly stems from the increase for loop regions that are counted as identical if only two residues overlap (Fig. A2). Surprising is, that the acceptance of a deviation in Sov, although best reflecting the flexibility of the segments, seems to reduce the strength of this measure in extracting the information about the 3D relevant information contained in the secondary structure strings. All three, the loose criterion ov loose, ov Taylor, and the fractional overlap Sov5°% seem to have a tendency to introduce artifacts. However, for Sov5°% this tendency is still very weak. This is confirmed by the results for helix and strand only (Fig. A3).

The question arises whether the defined measures are sharp enough to reveal essential similarities in 3D structure on the level of secondary structure strings. To find out we take a prediction method (PHD, Table 1), and a set of protein pairs with pairwise dissimilar 3D structure (dubbed RAN, Table 1). If the segment based measure does reveal features of the 3D structure better than single residue based measures, the quotients 'segment/per-residue' measure should be higher for 3D homologous pairs (PDB98) than for 3D dissimilar ones (RAN). Since methods for the prediction of secondary structure from the sequence are still less suitable to predict essential features of 3D structure than alignment procedures using the sequence similarity with known 3D structures, the quotients (segment/per-residue measure) should also tend to be smaller for the prediction methods.

The quotients between the segment overlaps and per residue accuracy are highest for 3D homologous PDB pairs only for ov Strict and Sov°. For the loose criterion ov loose, for ov Taylor and for the relaxed fractional overlap, Sov5°% the prediction method PHD and the 3D dissimilar pairs RAN score relatively higher (Fig. A3). For ov loose, this mainly stems from the increase for loop regions that are counted as identical if only two residues overlap (Fig. A2). Surprising is, that the acceptance of a deviation in Sov, although best reflecting the flexibility of the segments, seems to reduce the strength of this measure in extracting the information about the 3D relevant information contained in the secondary structure strings. All three, the loose criterion ov loose, ov Taylor, and the fractional overlap Sov5°% seem to have a tendency to introduce artifacts. However, for Sov5°% this tendency is still very weak. This is confirmed by the results for helix and strand only (Fig. A3).

The conclusion is that segment criteria do not score at 100% for similar 3D proteins, but they capture information about the correlation between secondary and 3D structure that cannot be identified on the level of per-residue comparisons. An excellent prediction of three-state secondary structure would yield about the following scores: Qa > 85%, Info > 65%. For the relaxed fractional overlap, Sov5°% the prediction method PHD and the 3D dissimilar pairs RAN score relatively higher (Fig. A3). For ov loose, this mainly stems from the increase for loop regions that are counted as identical if only two residues overlap (Fig. A2). Surprising is, that the acceptance of a deviation in Sov, although best reflecting the flexibility of the segments, seems to reduce the strength of this measure in extracting the information about the 3D relevant information contained in the secondary structure strings. All three, the loose criterion ov loose, ov Taylor, and the fractional overlap Sov5°% seem to have a tendency to introduce artifacts. However, for Sov5°% this tendency is still very weak. This is confirmed by the results for helix and strand only (Fig. A3).

The conclusion is that segment criteria do not score at 100% for similar 3D proteins, but they capture information about the correlation between secondary and 3D structure that cannot be identified on the level of per-residue comparisons. An excellent prediction of three-state secondary structure would yield about the following scores: Qa > 85%, Info > 65%. For the relaxed fractional overlap, Sov5°% the prediction method PHD and the 3D dissimilar pairs RAN score relatively higher (Fig. A3). For ov loose, this mainly stems from the increase for loop regions that are counted as identical if only two residues overlap (Fig. A2). Surprising is, that the acceptance of a deviation in Sov, although best reflecting the flexibility of the segments, seems to reduce the strength of this measure in extracting the information about the 3D relevant information contained in the secondary structure strings. All three, the loose criterion ov loose, ov Taylor, and the fractional overlap Sov5°% seem to have a tendency to introduce artifacts. However, for Sov5°% this tendency is still very weak. This is confirmed by the results for helix and strand only (Fig. A3).
choice of the segment overlap measure (above) was mainly motivated by the desire to reduce the variation in the secondary structure of protein pairs from the same 3D family (Figs 1 and 4).

We thank three colleagues at EMBL: Gerrit Vriend and Ulrike Göbel for helpful discussions, and Michael Scharff for contributing software tools. We also thank the referees for their constructive criticism. Last, not least, we wish to express our gratitude to all those who make co-ordinates available.


Protein Secondary Structure: Redefining Prediction


*Edited by F. Cohen*

*(Received 4 June 1993; accepted 24 August 1993)*