

Redefining the Goals of Protein Secondary Structure Prediction

Burkhard Rost, Chris Sander and Reinhard Schneider

**Appendix I: Definition of Three Measures Based on
Secondary Structure Segments**

Appendix II: The Merits and Pitfalls of the Four Measures

REVIEW ARTICLE

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 (Q_3). 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 $Q_3 > 85\%$ 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, 1989a; Overington *et al.*, 1990; Summers & Karplus, 1990; Schneider & Sander, 1991; Vriend & 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; Schulz & 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; Gascuel & Golmard, 1988; Qian & Sejnowski, 1988; Holley & Karplus, 1989; Taylor & Orengo, 1989b; King & Sternberg, 1990; Kneller *et al.*, 1990;

† Abbreviations used: 3D, three-dimensional; Q_3 , overall per-residue identity in 3 states (helix, strand, loop); PDB, Protein Data Bank of known three dimensional structures; PHD, Profile network from Heidelberg (3 levels of networks for the prediction of secondary structure); DSSP, Dictionary of Secondary Structures of Proteins; *Sov*, segment overlap; Swissprot, data bank of known sequences; HSSP, database of Homology-derived Structures and Sequence alignments of Proteins

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 (Hohohm *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 pairs thus assembled (dubbed PDB98, 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: Kotelchuck & Scheraga, 1968, 1969; Anfinson & Scheraga, 1975; Robson, 1976; Maxfield & Scheraga, 1979). The formation of secondary structure can be thought of to be influenced by two terms:

$$C = C_{\text{local}} + C_{\text{global}}$$

Table 1
Per residue comparisons for PDB pairs and prediction methods

	N^a	N_{prot}^f	Q_2^g	Q_3^g	Q_β^g	Q_α^g	Corr_c^h	Corr_β^h	Corr_α^h	Info^i
PDB98 ^a	24291	140	88.4	88	86	99	0.84	0.85	0.77	0.616
PDB98 core ^b	8048	140	91.7	92	91	92	0.88	0.87	0.83	0.682
RAN ^c	12182	94	35.2	23	26	47	0.09	0.01	0.04	0.006
PHD ^d	23200	129	70.8	72	66	72	0.60	0.52	0.51	0.201
PHD core ^e	6175	128	72.5	70	72	74	0.62	0.57	0.55	0.287

^aPDB98, a set of 126 pairwise non-homologous protein chains is taken from protein data bank of 3D structures (PDB). 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 DSSP for similarity in 3D structure (Schmidler & Sander, 1991).

^bPDB98 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)).

^cRAN, 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.

^dPHD, 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).

^ePHD core, values for the core segments of PHD (see above).

^f N , number of residues in the data set; N_{prot} , number of proteins in the data set.

^g Q_2 , Q_3 , Q_β , Q_α , percentages of per residue identity between 2 secondary structure strings for all 3 states, and for helix, strand and loop.

^h Corr_c , Corr_β , Corr_α , give the Matthews correlation coefficients (Matthews, 1975).

ⁱ Info , measures the information defined by:

$$\text{Info} = 1 - \frac{\sum_{i=1}^3 a_i \cdot \ln a_i - \sum_{j=1}^3 A_{ij} \cdot \ln A_{ij}}{N \cdot \ln N - \sum_{i=1}^3 b_i \cdot \ln b_i}$$

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_{ij} the number of residues predicted to be in i and observed to be in j (Rost & Sander, 1993b).

where C is the secondary structure conformation of a protein, C_{local} describes the part that is determined by local interactions, and C_{global} the one influenced by global interactions. Of course, the assumption of additiveness 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 & Poin, 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 hydro-

gen 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 those 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). We now discuss if the comparison of secondary structure in terms of segments can reduce this variation of 12 (± 0)% (Fig. 1).

3. Segment Overlap Measure: An Attempt to Evaluate Similarity of 3D Structure at the Secondary Structure Level

(a) Simple assessment of segment accuracy

There are three simple measures for assessing the quality of predicted secondary structure segments:

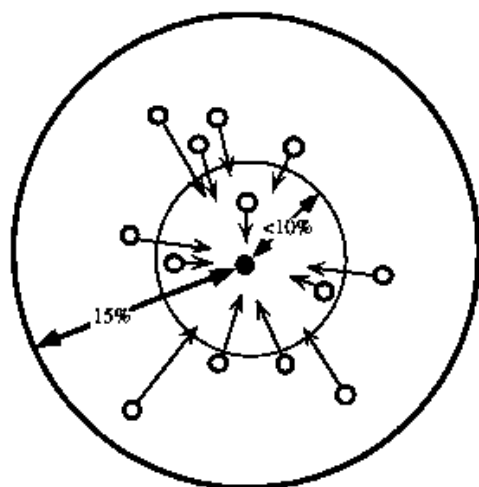


Figure 1. Goal of segment based measures, reduction of variation in secondary structure identity for protein pairs of the same structure family. The secondary structure strings differ for protein pairs from the same structure family. One goal for defining segment based measures is that this variation is smaller for a segment than for a per-residue comparison. ●, Variation in secondary structure within one protein family of the same structure (related to the family representative); ○, representative of structure-family (used as reference for computation of the variation in secondary structure); ○, sequences with homologous 3D structure to the representative; outer circle, variation in secondary structure for a per-residue comparison (Q_2), 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 $\langle L \rangle_i$, for $i = \alpha$ (helix), β (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_3), 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, 1980; Benner, 1992; Presnell *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 (Sov)*

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(e); for a detailed definition, eqn (A3) in Appendix 1)

$$Sov = \frac{1}{N} \sum_s \frac{\min ov(s_1; s_2) + \delta \times len(s_1)}{\max ov(s_1; s_2)} \quad (1)$$

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 $len(s_1)$ is the length of s_1 , usually chosen to be the segment in the experimental structure (Sov of observed). The actual overlap between the two segments is $\min ov$, i.e. the number of residues for which both segments have, e.g. an H (helix) in common, while $\max ov$ is the total extent of both segments, i.e. the number of residues for which either of the two has, say, the assigned state H (Fig. A1(e)). The accepted variation δ 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; Percec *et al.*, 1992; Hutchinson & Thornton, 1993); it is chosen such to be smaller than $\min ov$ and smaller than half the length of segment s_1 ($\delta = 1, 2, 3$ for short, intermediate, long segments). The ratio $\min ov / \max ov$ 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

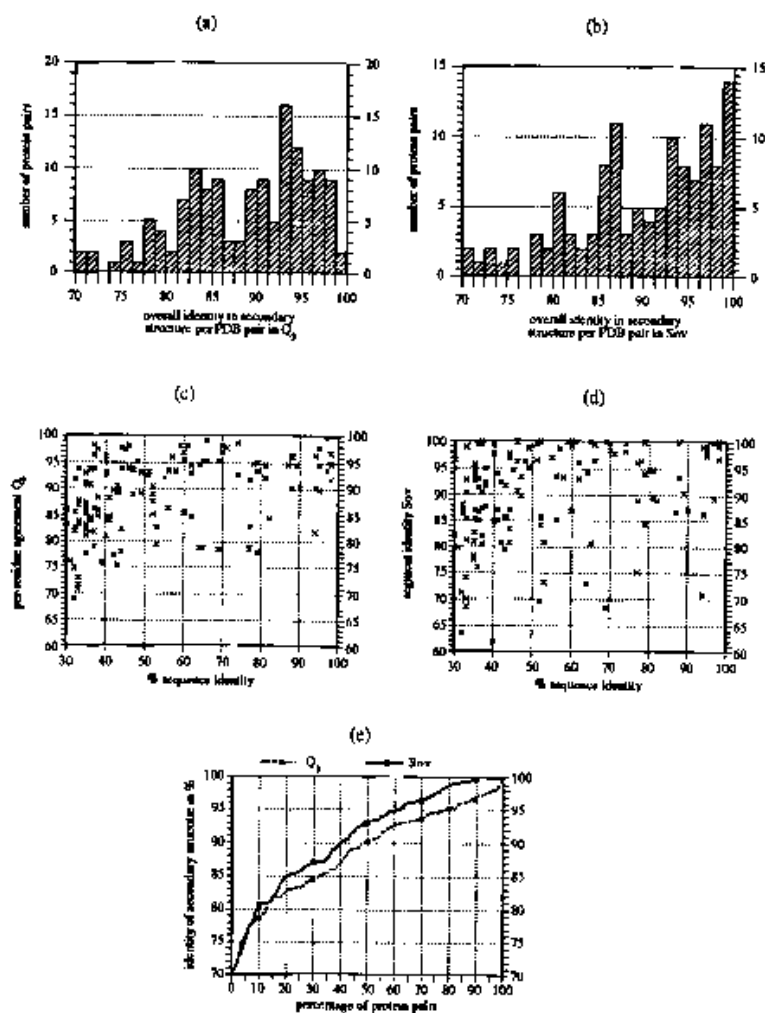


Figure 2. Distribution of secondary structure identity per alignment pair for PDB88. (a) The distribution of the per-residue identity per PDB pair and (b) the identity in segments (Sov) are shown. The average identity in Q_3 per pair is 88.8%, with a standard deviation of 9.1%. For the segment measure, Sov , the average is 89.5%, with a standard deviation of 9.4% (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 93%

(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 structure 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 definition and standard usage of a core-based measure.

6dfr	...	HHHHHHHHHH	...	EEEEHHHHHH	EEEE	EEE	HHHHHH
3dfr	...	HHHHHHHHH	...	EEEEHHHHHH	EEEE	EEE	HHHHHHHHH
6dfr	...	EEE	...	HHHHHHHHH	...	EEEEEE	HHHEEEEEEE	EEEEEE
3dfr	...	EEE	...	HHHHHH	...	EEEEEE	EE	HHHEEEEEEE
scores:		per-residue		86% (87/97/79)							
		segment overlap		97% (96/98/97)							
(a)											
1fd1	...	HHHHHHHHH	HHHHHHHHHH	EEEE	...	EEEE	...	EE
2lz2	...	HHHHHHHHH	HHHHHHHHH	EEE	...	EEE	...	EE
1fd1	HHHH	...	HHHHHHHHHH	...	HHH	...	HHH	...	HHHH
2lz2	HHH	...	HHHHHHH	...	HHH	...	HHHHH	...	HHH
scores:		per-residue		90% (82/80/97)							
		segment overlap		98% (100/100/95)							
(b)											

Figure 3. Pairs of secondary structure strings from homologous proteins. The sequence identity between the aligned fragment of (a) 6dfr (dihydrofolate reductase, *Escherichia coli*) and 3dfr (dihydrofolate reductase, *Lactobacillus c*) is 30%, that between the fragments of (b) 1fd1 (IG*G1 FAB fragment antilysozyme antibody) and 2lz2 (lysozyme) 95%. Given are the per-residue scores for 3-states and for each secondary structure type (in brackets, helix:strand/loop) and the segment overlap for 3-states and for each secondary structure type (in brackets, helix:strand/loop). The abbreviations used are: H, helix; E, strand; L, loop (rest).

(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 $Sw = 93\%$ 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

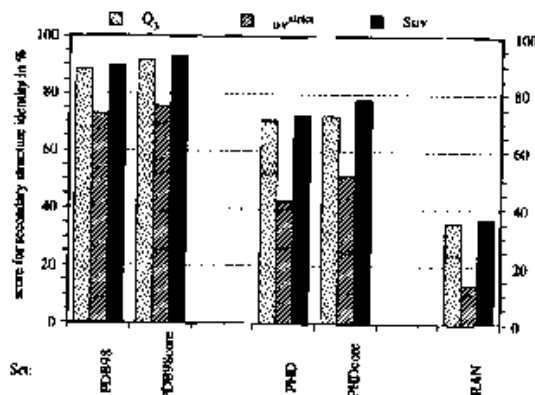


Figure 4. Absolute values for measures for PDB pairs and prediction methods. Percentages of per-residue accuracy and the segment based measures defined in Appendix 1. For the abbreviations see Table 1

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

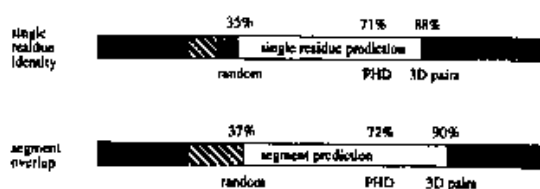


Figure 5. Redefining the goal of secondary structure prediction. Secondary structure prediction methods (like PHD) range between the lower limit given by a random prediction of 35 to 37%, and the upper limit given by the comparison of pairs from the same 3D family, 88 to 90%. The goal of prediction is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

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 within a 3D structure family and the temptation to introduce an excessively permissive criterion is given by the fractional segment overlap, Sov . This score measures the extent of segment overlap with a maximal deviation of about two residues at both ends (more for longer segments, eqn (A3)). Sov scores are at about 90% within a 3D

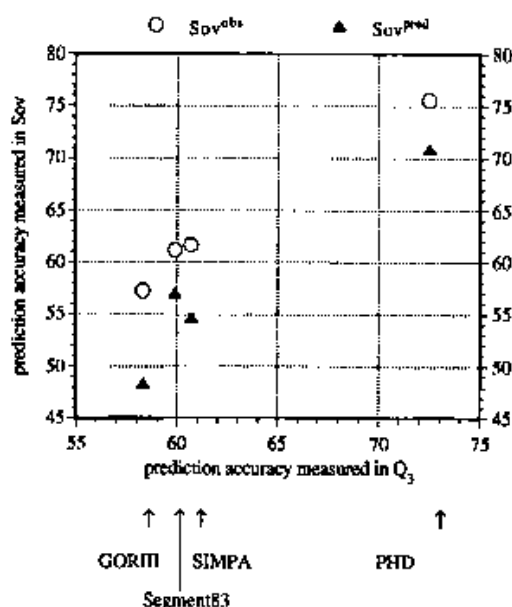


Figure 6. Per-residue and segment comparison of prediction accuracy. Each point in the scatter plot represents the 2 accuracy values for a particular prediction method: per-residue accuracy Q_3 and segment overlap Sov . The 2 series of points (circles and triangles) are for 2 ways of calculating the segment accuracy: for Sov^{obs} , the average over segments is weighted with the length of observed segments (eqn (1)), for Sov^{pred} of the predicted segments. The Sov^{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.

family. The measure is invariant against limited displacement, lengthening and shortening of the segments. Given the end tolerance, the residual variance of some 10% indicates that the loss of information by using secondary structure rather than full 3D structure is not caused by ambiguities in assigning the edges of segments. In practical applications, be it comparison of two experimental structures or comparison of a predicted and an observed structure, the target range for fractional segment overlap is $37\% < Sov \leq 90\%$ (Fig. 5).

The fractional segment overlap measures, Sov^{obs} , i.e. the fractional overlap weighted by the length of the observed segment and Sov^{pred} , i.e. the fractional overlap weighted by the length of the predicted segment, provide a reasonable step in this direction by attempting to capture the intuitive notion of segment similarity. We therefore propose to apply Sov for the evaluation of predictions along with the single residue measure Q_3 . Examples of the analysis of prediction methods using per-residue and segment measures are shown in Table 2 and Figure 6.

The following questions remain open for further analysis. Does the deviation of Sov from the ideal

Table 2
Per residue and segment comparisons for prediction methods

Measure	Segment83*	GORIII*	SIMPA*	PHD*	PDB98	RAN
Q_1	59.8	56.3	60.7	72.5	88.4	35.2
Q_2	62.1	59.6	62.7	72.8	88.0	23.6
Q_p	41.3	30.6	35.2	58.4	88.0	26.6
$1nf_0$	0.18	0.10	0.12	0.28	0.62	0.01
Nov_2	61.2	57.2	61.7	75.6	89.7	30.8
$Nov_{2,p}$	57.9	54.4	58.5	75.3	92.2	34.2
ov_{homo}	72.8	67.6	69.3	82.2	94.3	47.1
ov_{hetero}	30.6	26.1	27.4	44.0	79.2	13.7
ov_{Taylor}	66.9	67.9	63.2	83.2	92.8	38.4
$<L>_p$	13.3	4.3	6.0	11.1	—	—
$<L>_p$	4.9	2.5	3.0	5.2	—	—

The prediction methods are: Segment 83 (Kabsch & Sander, unpublished results), SIMPA (Lavin *et al.*, 1996), GORIII (Gibrat *et al.*, 1987), and PHD (Rost & Sander, 1993b). These are marked with a * to emphasize that they were evaluated on the set, given below (note: for Table 1 and the text the numbers given for PHD apply to another set, described in the legend to Table 1). The measures are explained in the legend to Table 1 and in Appendix I. $Nov_{2,p}$ gives the scores for the fractional overlap of only the helix and strand segments. Thus, Nov becomes directly compatible to ov^{Taylor} as the latter uses helix and strand only. $<L>_p$ is the average length of helical segments, $<L>_s$ the average length of strands. For comparison, the observed experimental averages (DSSP) in the same set are: $<L>_p = 9.9$ and $<L>_s = 5.1$. The accuracy values for homologous pairs (PDB98, Table 1) and random pairs (RAN, Table 1) provide a point of reference (the range between a perfect prediction (PDB98 same 3D structure) and a random one (RAN)).

Evaluation set *, a set of 48 newly determined protein structures. These 48 protein chains were chosen from a much larger Protein Data Bank "pre-release" set such that they all have less than 25% (for length > 80) similarity to any of the proteins used for the development of the methods (18,799 residues with 34% α , 23% β and 44% L). The proteins are listed by name and PDB identifier (where available): face, acetyl cholinesterase; luff, antifreeze polypeptide type A; lcol, antibacterial protein; colicin A (c-terminal domain); lcox, cholesterol oxidase; lcpk E, cAMP dependent protein kinase; ldfn B, defensin HNP-3; lglu, glutathione synthase; lffg, phosphocarrier; lglu, glucoamylase; lgnf A, granulocyte macrophage colony stimulating factor; lher, glycoprotein; ltkc, complement control protein of factor b; ltkd, C, engrailed homeodomain complex with DNA; ltrh, ribonuclease H domain of HIV-1 reverse transcriptase; ltrc, heat shock protein hsc70; lth, intestinal fatty acid binding protein; lmsb_A, mannose binding protein A (lectin domain); lmsb_B, neuraminidase sialidase; lpu2, serine proteinase inhibitor; lrpq, ColE1 repressor of primer; lsar A, endoribonuclease SA; lsny, stndbis virus capsid protein; lfgf, basic fibroblast growth factor; lgb1, protein G (b1 domain); lpk4, human plasminogen kringle 4; lhpj, B, high potential iron sulfur protein; lsep A, sarcoplasmic calcium binding protein; lztb_A, GCN4 leucine zipper; ltrx, thioredoxin; lznf, zinc-finger DNA binding domain; lndl, enolase; lnp21, CH-ras p21 protein (amino acids 1 to 169); lnpf, anpe maum, antifreeze glycoprotein type III; laci, actin (complex with DNase I); larc, Arc repressor DNA-binding protein; lck2, cyclin dependent kinase; lsrc, sh3 domain of tyrosine kinase src; lsh3, cell, β subunit of *Escherichia coli* DNA polymerase III holoenzyme; lsk, yeast hexokinase b; luff phd, flavoprotein related to bacterial luciferase; lmlk, nitrogenase molybdenum-iron; lpk3, phosphatidylinositol 3 kinase; lpr, phtalate thoxylase reductase; lpot, POU-specific domain; lrrx-su, retinoid X receptor α DNA binding domain; lsh2, v src tyrosine kinase transforming protein; lsh3, spectrin SH 3 homologue domain; lta, RNA-binding domain of U1 small nuclear ribonucleoprotein A.

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 (ov^{Jones} , S\TCP5 and ov^{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 $Q_3 = 70\%$ 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).

APPENDIX 1

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 (ov^{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 (ov^{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 (Fov), 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 (ov^{Taylor}) is described.

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

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

$$ov^{loose} = \frac{1}{N} \sum_s \theta^{loose}(s_1, s_2) * len(s_1), \quad (A1)$$

where N is the number of all residues of the protein, or the aligned fragment, $len(s_1)$ the length of segment s for sequence 1, θ^{loose} is a step function, which for helix and strand is given by:

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

and for loop regions:

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

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 $ov^{loose} = 0\%$, 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.

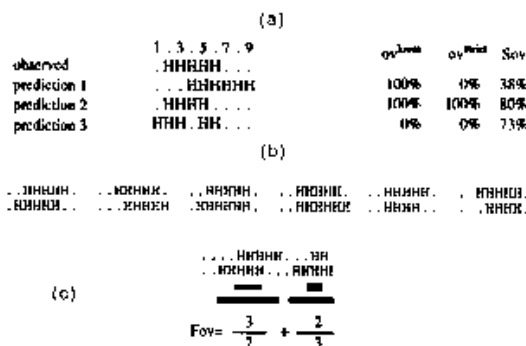


Figure A1. Three examples for predictions. The Table gives the segment measures as defined in eqns (A1) to (A3) for the helices shown. (b) Illustration of strict overlap ov^{strict} . Six different pairs of strings are shown. Each of these scores at 100% for ov^{strict} . (c) Illustration of fraction overlap Fov . The fractional overlap of secondary structure segments is the sum over all quotients between the number of consecutive residues for which both strings have, e.g. an 'H' at the same position and the number of residues for which either of the 2 strings has, e.g. an H.

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

A far more strict criterion for correctly matching segments is:

$$ov^{strict} = \frac{1}{N} \sum_s \theta^{strict}(s_1, s_2) * len(s_1), \quad (A2)$$

with the step function here chosen as:

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

where $beg(s_1)$ marks the first residue of segment s in sequence 1, $end(s_1)$ the last residue of this segment. δ^{strip} is the accepted deviation. The step function θ^{strict} assures that only segments with equivalent length ($\pm \delta^{strip}$) and with exactly equal placement

($\pm \delta^{\text{strict}}$) are counted. We set δ^{strict} to:

$len(s_1)$	1 to 5	6 to 10	11 to 15	15 to 20	...
δ^{strict}	1	2	3	3	3

The acceptance of a deviation is justified by the potential ambiguity of assigning the secondary

$$S_{\text{ov}}^{\text{obs}} = S_{\text{ov}} \text{ as in eqn (A3), with } s_1 \text{ being the observed segments,}$$

$$S_{\text{ov}}^{\text{pred}} = S_{\text{ov}} \text{ as in eqn (A3), with } s_1 \text{ being the predicted segments.}$$

structure at the edges of segments. Examples 1 and 3 in Figure A1(s) yield $\sigma^{\text{strict}} = 0\%$, example 2 $\sigma^{\text{strict}} = 100\%$.

Both, the loose (σ^{loose}) and the strict (σ^{strict}) measure for matching segments require arbitrary choices for free parameters (Θ^{loose} , Θ^{strict} , δ^{strict}). 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 S_{ov}

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. A1(c)):

$$S_{\text{ov}}^{\delta} = \frac{1}{N} \sum_{s_1} \frac{\min\{\text{end}(s_1); \text{end}(s_2)\} - \max\{\text{beg}(s_1); \text{beg}(s_2)\} + 1 + \delta \cdot \text{len}(s_1)}{\max\{\text{end}(s_1); \text{end}(s_2)\} - \min\{\text{beg}(s_1); \text{beg}(s_2)\} + 1} \quad (\text{A3})$$

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 & Orango, 1988a; Wilmut & Thornton, 1990; Perzel et al., 1992; Hutchinson & Thornton, 1993). δ is either = 0 (S_{ov}^0), or it is restricted by ($S_{\text{ov}}^{50\%}$):

$$\delta \leq \min\left\{(\text{maxov}(s_1, s_2) - \text{minov}(s_1, s_2)); \frac{\text{len}(s_1)}{2}\right\},$$

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 σ^{loose} (eqn (A1)) and $S_{\text{ov}}^{50\%}$ (eqn (A3)) reduce the variation in the secondary structure of 3D homo-

Where minov describes the nominator of equation (A3), i.e. the region for which both strings have the same symbol (e.g. H), maxov the denominator, i.e. the region for which either of the two have e.g. an H. For the examples in Figure A1(a) S_{ov}^0 yields 38 to 80%.

For the comparison of a predicted with the observed secondary structure, two versions are meaningful:

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 or Taylor

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' \rightarrow '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 σ^{Taylor} .

gues. However, even for the strict measure σ^{strict} (eqn (A2)) 10% of all protein pairs score at 100% (data not shown), whereas no pair reaches a Q_3 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, $\sigma^{\text{loose}} = 94\%$. 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-

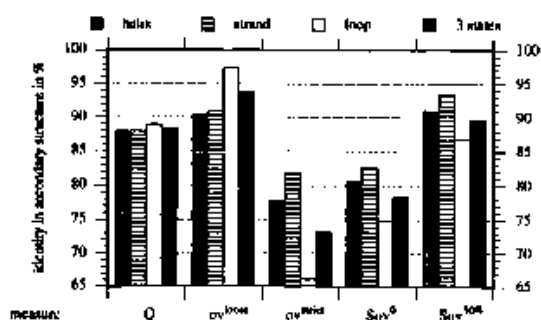


Figure A2. Per-residue identity and segment similarity for PDB pairs. For all PDB pairs the per residue 3-state accuracy is compared to the 3 measures for matching segments introduced in Appendix 1. For the fractional overlap 2 different choices for the accepted deviation (Sov^0 and Sov^{100}) are given. The measures are given as percentages.

ated by an artifact resulting from unrealistic displacements, or extensions of the segments? The goal is that the secondary structure strings of a protein pair are evaluated as identical if, and only if, the pair has similar 3D structure.

Helices and sheets are evolutionary more conserved than loop regions. This cannot be made out on the level of the commonly used single-residue quotients, the loop regions yield a slightly higher percentage of identical residues than the helices and strands (Fig. A2). The loose segment measure amplifies this effect. In contrast, the strict criterion and the fractional overlap result in higher values for helices ($Sov^{100} = 91\%$) and for strands ($Sov^{100} = 93\%$) than for loop regions, thus clearly showing the higher degree of conservation inside a structure family for helix and strand. This result suggests that both the loose segment criterion and the per-residue identity fail to capture an important aspect of 3D structures.

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^0 . For the loose

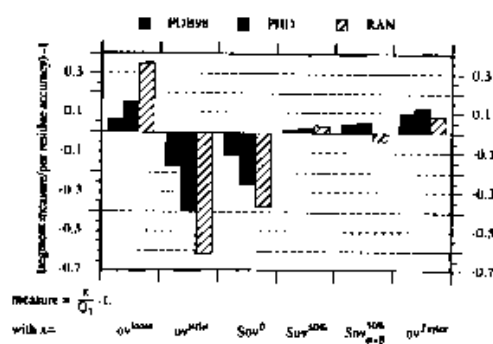


Figure A3. Ratio of segment similarity and per residue identity. Plotted are the quotients: (segment measure/per-residue measure)-1. Thus positive values indicate that the segment measure scores higher than the single-residue quantity. The quotients are chosen to enable a comparison between results with 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 $Sov_{h,s}^{50\%}$.

criterion ov^{loose} , for ov^{Taylor} and for the relaxed fractional overlap, $For^{50\%}$, 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 $Sov^{50\%}$, seem to have a tendency to introduce artifacts. However, for $Sov^{50\%}$ 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: $Q_3 > 85\%$; $Info > 0.5$ (see captions of Table 1); $ov^{strict} > 70\%$; $Sov^{50\%} \geq 85\%$. $Sov^{50\%}$ offers a reasonable trade-off between the goal to reduce the variation in secondary structure inside a structure family and the attempt to avoid not sufficiently discriminatory measures stemming from too relaxed constraints.

The analysis clearly shows that the relaxed measures (ov^{loose} , ov^{Taylor}) tend to overestimate the identity of segments of two secondary structure strings. However, it is not clear cut, whether or not the strict overlap, ov^{strict} , is more sensitive to false predictions than the segment overlap, Sov . The

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 Gübel for helpful discussions, and Michael Scharf 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 coordinates of experimentally determined protein 3D structures available.

References

- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.
- Anfinsen, C. B. & Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Advan. Protein Chem.* **29**, 205-300.
- Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **20**, 2010-2022.
- Baumam, G., Krümmel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329-334.
- Benner, S. A. (1992). Predicting de novo the folded structure of proteins. *Curr. Opin. Struct. Biol.* **2**, 402-412.
- Benner, S. A., Cohen, M. A. & Gerloff, D. (1991). Predicted secondary structure for the Src homology 3 domain. *J. Mol. Biol.* **229**, 295-305.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Bric, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
- Biau, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. (1988). Secondary structure prediction: combination of three different methods. *Protein Eng.* **2**, 185-191.
- Brändén, C.-I. & Jones, T. A. (1990). Between objectivity and subjectivity. *Nature (London)*, **343**, 687-689.
- Brändén, C.-I. & Tooze, J. (1991). *Introduction to Protein Structure*. Garland Publishers, New York, London.
- Burgess, A. W., Ponnuswamy, P. K. & Scheraga, H. A. (1974). Analysis of conformations of amino acid residues and prediction of backbone topography in proteins. *Israel J. Chem.* **12**, 239-286.
- Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.
- Chou, P. Y. & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, **13**, 211-215.
- Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol.* **47**, 45-148.
- Cohen, F. E. & Kuntz, I. D. (1989). Tertiary structure prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), pp. 647-706. Plenum, New York, London.
- Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1983). Secondary structure assignment for α/β proteins by a combinatorial approach. *Biochemistry*, **22**, 4894-4904.
- Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986). Turn prediction in proteins using a pattern-matching approach. *Biochemistry*, **25**, 266-275.
- Collin'h, N., Bichebest, C., Thoreau, E., Hourissat, B. & Moron, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* **6**, 377-382.
- Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. (1963). The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harbour Symp. Quant. Biol.* **28**, 439-449.
- Finkelstein, A. V. & Pitsyn, O. B. (1971). Statistical analysis of the correlation among amino acid residues in helical, β -structural and non-regular regions of globular proteins. *J. Mol. Biol.* **62**, 613-624.
- Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.
- Gascuel, O. & Golmard, J. L. (1988). A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *CABIOS*, **4**, 357-365.
- Gerloff, D. L., Jonny, T. E., Knecht, L. J., Gonnet, G. H. & Benner, S. A. (1993). The nitrogenase MoFe protein. *FEBS Letters*, **318**, 118-124.
- Gibrat, J.-F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425-443.
- Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.
- Holley, H. L. & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 152-156.
- Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins*, **14**, 213-223.
- Holm, L., Ozounik, C., Sander, C., Tuparev, G. & Vriend, G. (1993). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691-1698.
- Hutchinson, E. G. & Thornton, J. M. (1993). The Greek key motif: extraction, classification and analysis. *Protein Eng.* **6**, 233-245.
- Kabat, E. A. & Wu, T. T. (1973a). The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of β -sheets in conalbumin A. *Proc. Nat. Acad. Sci., U.S.A.* **70**, 1473-1477.
- Kabat, E. A. & Wu, T. T. (1973b). The influence of nearest-neighboring amino acid residues on aspects of secondary structure of proteins. Attempt to locate α -helices and β -sheets. *Biopolymers*, **12**, 751-774.
- Kabsch, W. & Sander, C. (1983a). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.
- Kabsch, W. & Sander, C. (1983b). How good are predictions of protein secondary structure? *FEBS Letters*, **155**, 179-182.
- King, R. D. & Sternberg, M. J. (1990). Machine learning approach for the prediction of protein secondary structure. *J. Mol. Biol.* **216**, 441-457.
- Kueller, D. G., Cohen, F. E. & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171-182.

- Kotelchuck, D. & Scheraga, H. A. (1968). The influence of short-range interactions on protein conformation. I. Side chain-backbone interactions within a single peptide unit. *Proc. Nat. Acad. Sci., U.S.A.* **61**, 1163-1170.
- Kotelchuck, D. & Scheraga, H. A. (1969). The influence of short range interactions on protein conformation. II. A model for predicting the α -helical regions of proteins. *Proc. Nat. Acad. Sci., U.S.A.* **62**, 14-21.
- Leak, A. M. (1991). *Protein Architecture: A Practical Approach*. Oxford University Press, Oxford, New York, Tokyo.
- Levin, J. M., Robson, B. & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, **205**, 303-308.
- Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507-533.
- Liu, V. I. (1974). Structural principles of the globular organization of protein chains: A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88**, 857-872.
- MacIn, R. & Shavlik, J. W. (1993). Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, **11**, 195-215.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442-451.
- Maxfield, F. R. & Scheraga, H. A. (1976). Status of empirical methods for the prediction of protein backbone topography. *Biochemistry*, **15**, 5138-5153.
- Maxfield, F. R. & Scheraga, H. A. (1979). Improvements in the prediction of protein topography by reduction of statistical errors. *Biochemistry*, **18**, 697-704.
- Muggleton, S., King, R. D. & Sternberg, M. J. E. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Eng.* **5**, 647-657.
- Nagano, K. (1973). Logical analysis of the mechanism of protein folding. *J. Mol. Biol.* **75**, 401-420.
- Nagano, K. (1977). Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol.* **109**, 251-274.
- Nagano, K. & Hasegawa, K. (1975). Logical analysis of the mechanism of protein folding. *J. Mol. Biol.* **94**, 257-281.
- Nishikawa, K. & Noguchi, T. (1991). Predicting protein secondary structure based on amino acid sequence. *Methods Enzymol.* **202**, 31-44.
- Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Roy. Soc. London*, **241**, 132-145.
- Pain, R. H. & Robson, B. (1970). Analysis of the code relating sequence to secondary structure in proteins. *Nature (London)*, **227**, 62-63.
- Pauling, L. & Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc. Nat. Acad. Sci., U.S.A.* **37**, 729-740.
- Perrzel, A., Park, K. & Fasman, G. D. (1992). Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the anti-parallel β -sheet in proteins. *Proteins*, **13**, 57-69.
- Pohl, F. M. (1971). Empirical protein energy maps. *Nature New Biol.* **234**, 277-278.
- Pohl, F. M. (1980). Statistical analysis of protein structure. In *Protein Folding* (Jaenicke, R., ed.), pp. 183-197. Elsevier/North-Holland Biomedical Press, Amsterdam, New York.
- Presnell, S. R., Cohen, B. I. & Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. *Biochemistry*, **31**, 983-993.
- Ptitayn, O. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15-25.
- Qian, X. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865-884.
- Robson, B. (1974). Analysis of the code relating sequence to conformation in globular proteins: theory and application of expected information. *Biochem. J.* **141**, 853-867.
- Robson, B. (1976). Conformational properties of amino acid residues in globular proteins. *J. Mol. Biol.* **107**, 327-366.
- Robson, B. & Garnier, J. (1993). *Nature (London)*, **361**, 506.
- Robson, B. & Pain, R. H. (1971). Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.* **58**, 237-259.
- Robson, B. & Pain, R. H. (1974a). Analysis of the code relating sequence to conformation in globular proteins: an informational analysis of the residue in determining the conformation of its neighbours in the primary sequence. *Biochem. J.* **141**, 883-907.
- Robson, B. & Pain, R. H. (1974b). Analysis of the code relating sequence to conformation in globular proteins—development of a stereochemical alphabet on the basis of intra-residue information. *Biochem. J.* **141**, 869-882.
- Robson, B. & Pain, R. H. (1974c). Analysis of the code relating sequence to conformation in globular proteins: the distribution of residue pairs in turns and kinks in the backbone chain. *Biochem. J.* **141**, 899-904.
- Rooman, M. J. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. *Biochemistry*, **31**, 10239-10249.
- Romman, M. J., Koehler, J. P. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *J. Mol. Biol.* **221**, 901-970.
- Romman, M. J., Koehler, J.-P. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, **31**, 10226-10238.
- Rost, B. & Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 7558-7562.
- Rost, B. & Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.
- Salzberg, S. & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.* **227**, 371-374.
- Schneider, R. & Sander, C. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56-68.
- Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*. Springer, New York, Berlin.
- Sklenar, H., Etchebest, C. & Lavery, R. (1989).

- Describing protein structure: a general algorithm yielding complete helical parameters and a unique overall axis. *Proteins*, **6**, 46-60.
- Sternberg, M. J. E. (1992). Secondary structure prediction. *Curr. Opin. Struct. Biol.*, **2**, 237-241.
- Stulz, C. M., White, J. V. & Smith, T. F. (1993). Structural analysis based on state-space modeling. *Protein Sci.*, **2**, 305-314.
- Summers, N. L. & Karplus, M. (1990). Modeling of globular proteins. *J. Mol. Biol.*, **216**, 991-1018.
- Taylor, W. R. (1992). New paths from dead ends. *Nature (London)*, **356**, 478-480.
- Taylor, W. R. (1984). An algorithm to compare secondary structure predictions. *J. Mol. Biol.*, **173**, 512-521.
- Taylor, W. R. & Orengo, C. A. (1989a). A holistic approach to protein structure alignment. *Protein Eng.*, **2**, 505-519.
- Taylor, W. R. & Orengo, C. A. (1989b). Protein structure alignment. *J. Mol. Biol.*, **208**, 1-22.
- Taylor, W. R. & Thornton, J. M. (1983). Prediction of super-secondary structure in proteins. *Nature (London)*, **301**, 540-542.
- Taylor, W. R. & Thornton, J. M. (1984). Recognition of super-secondary structure in proteins. *J. Mol. Biol.*, **173**, 487-514.
- Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1992). Prediction of progress at last. *Nature (London)*, **354**, 105-108.
- Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins*, **11**, 52-58.
- Wilmot, C. M. & Thornton, J. M. (1990). β -Turns and their distortions: a proposed new nomenclature. *Protein Eng.*, **3**, 479-493.
- Woodcock, S., Murnon, J. P. & Henriessat, B. (1992). Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.*, **5**, 629-635.
- Zhang, X., Medvedev, J. P. & Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, **225**, 1040-63.
- Zeelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.*, **195**, 957-981.

Edited by F. Cohen

(Received 4 June 1993; accepted 24 August 1993)