Protein Science (1996), *5*:2298–2310. Cambridge University Press. Printed in the USA.

Identification and application of the concepts important for accurate and reliable protein secondary structure prediction

ROSS D. KING AND MICHAEL J.E. STERNBERG
Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, London, WC2A 3PX, United Kingdom

(RECEIVED June 4, 1996; ACCEPTED August 13, 1996)

Abstract

A protein secondary structure prediction method from multiply aligned homologous sequences is presented with an overall per residue three-state accuracy of 70.1%. There are two aims: to obtain high accuracy by identification of a set of concepts important for prediction followed by use of linear statistics; and to provide insight into the folding process. The important concepts in secondary structure prediction are identified as: residue conformational propensities, sequence edge effects, moments of hydrophobicity, position of insertions and deletions in aligned homologous sequence, moments of conservation, auto-correlation, residue ratios, secondary structure feedback effects, and filtering. Explicit use of edge effects, moments of conservation, and auto-correlation are new to this paper. The relative importance of the concepts used in prediction was analyzed by stepwise addition of information and examination of weights in the discrimination function. The simple and explicit structure of the prediction allows the method to be reimplemented easily. The accuracy of a prediction is predictable a priori. This permits evaluation of the utility of the prediction: 10% of the chains predicted were identified correctly as having a mean accuracy of >80%. Existing high-accuracy prediction methods are "black-box" predictors based on complex nonlinear statistics (e.g., neural networks in PHD: Rost & Sander, 1993a). For medium- to short-length chains (≥90 residues and <170 residues), the prediction method is significantly more accurate ($P < 0.01$) than the PHD algorithm (probably the most commonly used algorithm). In combination with the PHD, an algorithm is formed that is significantly more accurate than either method, with an estimated overall three-state accuracy of 72.4%, the highest accuracy reported for any prediction method.

Keywords: prediction; secondary structure; statistics

The relationship between protein amino acid sequence and secondary structure is complex, reflecting the intricate thermodynamic and kinetic process of protein folding. Yet most, if not all, of the information necessary to predict secondary structure lies in the primary structure. The problem in secondary structure prediction is to extract the maximum information from the primary sequence in the absence of a tertiary structure model.

There are two broad approaches to tackling complex prediction problems in science. The traditional approach is to transform the representation of the problem so as to decompose it into discrete understandable features that can be combined simply for prediction. This approach has been taken by a number of secondary structure prediction methods (Chou & Fasman, 1974; Lim, 1974b; Robson, 1976; Cohen et al., 1983; King & Sternberg, 1990; Muggleton et al., 1992). The alternative approach is to use sophisticated nonlinear statistical methods for prediction, and to omit an explicit understanding of the problem. Most of the currently successful secondary structure prediction programs take the nonlinear statistical approach. They are often based on neural network (e.g., Qian & Sejnowski, 1988; Kneller et al., 1990; Rost & Sander, 1993a) or k-nearest-neighbor (e.g., Biou et al., 1988; Zhang et al., 1992; Yi & Lander, 1993; Geourjon & Deleage, 1994; Salamov & Solovyev, 1995). The best such prediction methods use very complicated statistical procedures (elaborate architectures and distance measures) and have Q_3 accuracies of ≈70% (see Materials and methods) on a standard database of aligned sequences (Rost & Sander, 1993a; Salamov & Solovyev, 1995). These nonlinear prediction methods are "black-box" predictors (Michie et al., 1994; King et al., 1995). They do not make the basis of their prediction explicit, nor do they provide insight into the principles governing the formation of secondary structure. A separation has occurred between the understanding and the prediction of protein structure.

This criticism of "black-box" predictors is echoed in the work of Benner et al., who complain that there is no explanation "why" neural network predictions work (Benner & Gerloff, 1993). Benner et al. take a different approach to prediction based on hand-

Reprint requests to: Michael J.E. Sternberg, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, P.O. Box 123, London, WC2A 3PX, United Kingdom; e-mail: m.sternberg@icrf.icnet.uk.

crafted predictions on individual proteins by experts on protein structure, e.g., Benner and Gerloff (1990) and Benner et al. (1992). The evaluation of the success of this work is complicated by the subjective nature of the prediction method, but comparable accuracies to the best automatic algorithms can be obtained. One problem of extending this approach for general use is that it is notoriously difficult for experts in a field to articulate the reasons for many of their judgments (Michie, 1986).

To produce understandable predictions, it is therefore essential to avoid use of both complicated nonlinear statistical techniques and human intervention in prediction. However, statistical techniques that produce understandable predictions have not been powerful enough to achieve high accuracy (Chou & Fasman, 1974; Gibrat et al., 1987; Dowe et al., 1993; Solovyev & Salamov, 1994). Consequently, to produce understandable and accurate predictions, it is necessary to transform the representation of the problem from one based solely on simple sequences of residues, to one based on the important underlying concepts. These are the concepts that are implicitly used by human experts and nonlinear statistical methods. This transformation would allow simple statistical methods to be accurate and provide insight.

What then are the important concepts in secondary structure prediction? The most basic concept is the propensity of particular residues for particular secondary structures, e.g., it has long been recognized that alanine residues favor formation of α-helices. Residues are also associated with certain positions within secondary structures (Richardson & Richardson, 1988; Wako & Blundell, 1994). At a higher level, it has been recognized that patterns of hydrophobicity are important. Lim (1974a, 1974b) identified patterns of hydrophobicity associated with different types of secondary structure. Eisenberg (1984) identified hydrophobic moment as an important component in structure prediction. The introduction of the idea of using aligned homologous sequences in secondary structure prediction (Zvelebil et al., 1987) allows the incorporation of other types of information. The use of aligned sequence allows better application of residue propensities and identification of patterns of conservation. The position of insertions and deletions also gives valuable information, because the tolerance of such mutations varies with secondary structures class. Secondary structure is also auto-correlated, that is, discrete secondary structure elements occur (knowing the state of position $i - 1$, helps to predict position i).

In this paper, a transparent, accurate, and reliable protein secondary structure prediction method, Discrimination of Secondary structure Class (DSC), is described and evaluated. DSC is based on decomposing secondary structure prediction into the basic concepts and then using simple and linear statistical methods to combine the concepts for prediction. This makes the prediction method comprehensible and allows the relative importance of the different sources of information used to be measured.

The DSC prediction method is summarized in Figure 1. For every residue position, the following are calculated: mean GOR potential for each secondary structure class, distance to end of chain, mean moment of hydrophobicity assuming α-helix and β-strand, existence of insertions and deletions, and the mean moment of conservation assuming α-helix and β-strand. These attributes are then smoothed and a linear discrimination function is applied to make a level-one prediction for each residue position. The fraction of residues predicted to be α-helix and β-strand per protein were then calculated, as well as the fractional content of certain residues. This information is then used, with the level-one information, to make a refined prediction using a second linear discrimination function. The prediction is then filtered to give a final prediction (Tables 1, 2).

Results

DSC results

The estimated values of residue propensities for secondary structure class calculated from all 126 proteins are given in Tables 3,

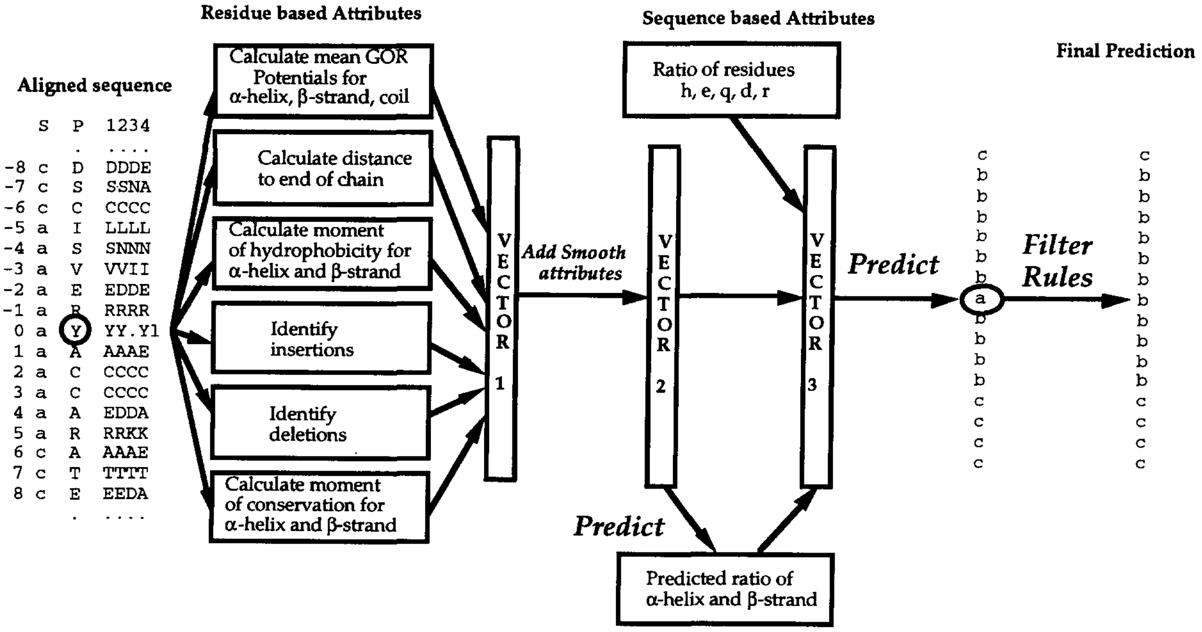

Fig. 1. DSC prediction method. For the aligned sequence: S is the observed secondary structure of the primary sequence, P. The residue at position 0 is predicted (circled).

Table 1. *Accuracy obtained using different sources of information*[a]

Method	Q_3 Acc.%	Qa%	Qb%	Qc%	Ma	Mb	Mc
Run1: Standard GOR on aligned sequences	63.5	68.4	61.3	61.1	0.48	0.43	0.43
Run2: GOR + attributes	67.8	67.2	60.9	70.1	0.53	0.46	0.47
Run3: GOR + attributes + smoothing	68.3	67.9	62.0	70.1	0.54	0.47	0.48
Run4: GOR + attributes + smoothing + feedback	69.4	69.7	64.2	71.2	0.56	0.50	0.48
Run5: DSC	70.1	73.5	64.9	70.3	0.58	0.51	0.48

[a] Q_3 is the predicted (3 state) accuracy. Qa, Qb, and Qc are the accuracy for α-helix, β-strand, and coil, respectively. Ma, Mb, and Mc are the correlation coefficients for α-helix, β-strand, and coil, respectively.

4, 5 (predictions were made out using leave-one-out cross-validated versions of the tables). The results obtained are broadly similar to those found by Gibrat et al. (1987), but they vary on many specifics.

The relative importance of the different attributes for prediction can be assessed using the learned discrimination functions (Tables 6, 7, 8, and 9). The functions shown were formed using the attribute vectors from all 126 proteins. Actual prediction was done using functions learned using leave-one-out cross-validation.

The per residue Q_3 accuracy of DSC is 70.1% (see Tables 1 and 2), and the distribution of accuracies is given in Figure 2. Only two other prediction methods have comparably high prediction accuracies, Profile network from Heidelberg (PHD) (Rost & Sander, 1993a) and Nearest Neighbor Secondary Structure Prediction (NNSSP) (Salamov & Solovyev, 1995). The recent paper by Mehta et al. (1995), which reports an accuracy of 70.9%, is for a much smaller data set of sequences with large amounts of aligned sequences. On exactly the same set of proteins, the PHD method had

Table 2. *Q_3 accuracy and predicted accuracy for the 126 protein chains used in this study (Run5)*[a]

ID	Acc.%	Pred. Acc.%	ID	Acc.%	Pred. Acc. %	ID	Acc.%	Pred. Acc.%	ID	Acc.%	Pred. Acc.%
1acx	72.9	63.3	1pyp	65.0	67.1	2or1_l	82.5	77.0	4gr1	69.4	68.1
1ak3_a	78.5	73.1	1r09_2	63.1	64.8	2pab_a	59.6	69.6	4pfk	75.2	70.3
1azu	66.1	68.1	1rbp	59.8	63.6	2pcy	81.8	83.7	4rhv_1	69.2	70.4
1bbp_a	64.7	65.2	1rhd	68.3	68.1	2phh	56.5	67.8	4rhv_3	75.8	68.5
1bds	74.4	74.8	1s01	73.1	72.8	2rsp_a	67.0	70.2	4rhv_4	30.0	64.3
1bmv_1	76.2	73.2	1sdh_a	81.5	76.6	2sns	62.4	67.8	4rxn	63.0	77.1
1bmv_2	67.6	63.7	1sh1	47.9	65.5	2sod_o	85.4	78.1	4sgb_i	82.4	75.3
1cbh	63.9	59.4	1tgs_i	62.5	74.6	2stv	69.6	61.2	4ts1_a	71.6	69.1
1cc5	67.5	77.4	1tnf_a	68.4	68.7	2tgp_i	77.6	63.7	4xia_a	75.6	71.6
1cd4	68.8	69.8	1ubq	69.7	66.3	2tmv_p	63.6	65.4	5cyt_r	73.8	66.4
1cdt_a	68.3	75.0	1wsy_a	83.5	74.6	2tsc_a	65.2	62.6	5er2_e	66.4	67.3
1crn	58.7	67.8	1wsy_b	71.7	71.8	2utg_a	82.9	79.6	5hvp_a	78.8	69.7
1cse_i	76.2	78.1	256b_a	84.0	81.2	2wrp_r	89.4	74.6	5ldh	60.7	69.9
1eca	78.7	84.9	2aat	71.2	68.3	3ait	81.1	77.2	5lyz	66.7	62.5
1etu	75.3	71.0	2alp	69.7	62.4	3b5c	62.4	70.4	6acn	62.9	65.5
1fc2_c	60.5	59.9	2cab	73.4	70.6	3blm	64.2	69.6	6cpa	73.6	64.4
1fdl_h	80.3	80.8	2ccy_a	93.7	79.2	3cla	67.6	58.9	6cpp	57.0	68.0
1fdx	75.9	85.3	2cyp	62.5	65.8	3cln	85.3	87.5	6cts	79.1	67.5
1fkf	78.5	73.7	2fnr	73.3	71.8	3ebx	71.0	75.2	6dfr	67.7	72.9
1fxi_a	85.4	73.6	2fxb	76.5	70.9	3gap_a	60.1	67.8	6hir	75.5	80.7
1gd1_o	63.8	70.4	2gbp	66.7	67.6	3hmg_a	65.2	72.1	6tmn_e	63.0	61.2
1gp1_a	71.0	70.2	2gcr	71.1	67.8	3hmg_b	60.6	67.7	7cat_a	67.3	67.9
1hip	52.9	66.4	2gls_a	70.5	68.9	3icb	85.3	83.5	7icd	71.0	68.4
1il8_a	73.2	70.4	2gn5	39.1	76.3	3pgm	68.7	68.0	7rsa	68.5	75.6
1l58	76.2	72.2	2hmz_a	83.3	74.6	3rnt	75.0	64.0	8abp	63.9	69.2
1lap	71.0	67.5	2i1b	80.4	75.5	3tim_a	82.3	71.2	8adh	66.0	71.0
1lrd_3	75.9	79.9	2lh4	82.4	83.3	4bp2	59.3	60.1	9api_a	54.9	68.5
1mcp_l	80.0	76.0	2lhb	79.9	73.4	4cms	65.1	67.3	9api_b	83.3	76.8
1mrt	83.9	71.3	2ltn_a	81.8	76.7	4cpa_i	75.7	73.8	9ins_b	83.3	68.9
1ovo_a	55.4	73.6	2ltn_b	74.5	73.0	4cpv	88.9	80.0	9pap	71.7	74.3
1paz	83.3	81.1	2mev_4	46.6	65.2	4fxn	82.6	74.7	9wga_a	66.7	78.1
1ppt	97.2	86.8	2mhu	90.0	80.1						

[a] ID, Brookhaven identity and chain; Acc., actual accuracy obtained; Pred. Acc., accuracy predicted using regression.

Table 3. *Directional informational parameters: I(Sj = x:x'; Rj + m) for residue position versus residue type for α-helices*[a]

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	47	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-55	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	17	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	0	0	1	1	2	-5	1	2	1	7	-6	-3	10	8	6
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	0	6	9	16	30	33	45	47	51	53	37	32	30	25	18
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-17	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-30	-48	-82	-195	-145	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

[a] Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix at position j.

an estimated per residue Q_3 accuracy of 70.8% (which has been subsequently upgraded to 71.6%), and NNSSP has an estimated per residue Q_3 accuracy of 72.2%. It is difficult to test if there is a significant statistical difference in accuracy between DSC, PHD, and NNSSP because the individual residue predictions are not available. If they were available, each residue prediction would not be independent, making the correct statistical test difficult to determine.

Table 4. *Directional informational parameters for residue position versus residue type for β-strands*

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-8	-7	-13	-17	-23	-33	-26	-32	-43	-37	-30	-30	-26	-27	-26	-25	-25
c	3	13	-9	-20	-15	-3	9	33	47	51	21	19	9	-5	7	-5	-14
d	-7	-5	0	-9	-4	-14	-42	-73	-83	-59	-21	10	22	24	16	11	13
e	-14	-5	-5	-11	-21	-27	-45	-44	-57	-54	-46	-29	-25	-12	-12	-2	0
f	-9	-20	-32	-34	-30	-12	24	44	49	39	24	2	-9	-23	-24	-29	-23
g	-3	9	24	29	34	30	18	-23	-48	-27	6	27	39	38	33	23	23
h	6	11	17	22	12	16	0	-2	3	-2	5	3	8	4	-1	1	-3
i	-21	-30	-31	-21	-12	-3	26	58	76	64	33	11	-14	-24	-20	-14	-11
k	20	12	15	14	8	4	-8	-14	-25	-40	-39	-27	-20	-24	-20	-15	-15
l	-2	-10	-18	-27	-30	-27	-6	15	27	21	2	-19	-31	-29	-28	-26	-25
m	-22	-26	-29	-40	-31	-17	-7	23	24	28	17	2	-15	-31	-53	-36	-16
n	1	8	14	5	0	-6	-30	-65	-62	-28	-6	11	18	21	16	10	3
p	9	7	12	24	20	8	-22	-65	-108	-64	-8	17	25	30	32	31	21
q	6	12	8	16	8	-5	-22	-27	-30	-52	-49	-34	-22	-17	-9	2	20
r	0	8	3	-3	5	2	1	-14	-26	-32	-30	-35	-27	-26	-25	-25	-21
s	16	14	17	19	14	5	-3	-13	-15	-4	15	27	32	32	31	28	21
t	6	8	14	15	16	21	19	25	31	22	13	9	12	25	34	34	34
v	1	-11	-15	-11	4	25	51	75	91	81	49	19	-6	-12	-16	-11	-11
w	-8	-8	-28	-19	-9	5	23	44	45	30	13	-18	-22	-40	-15	-7	-9
y	13	13	4	14	12	20	24	37	48	31	20	-1	2	11	7	0	-4

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-12	-15	-12	-12	-17	-13	-25	-24	-32	-35	-32	-29	-24	-20	-12	-5	-6
c	36	26	41	50	45	31	29	19	7	5	27	29	38	48	41	45	59
d	-8	-10	-13	-8	-13	-10	12	25	50	43	39	27	7	-7	-4	-9	-5
e	-3	-11	-10	-11	-10	-7	-5	-23	-26	-23	-2	5	-1	-3	3	-5	-9
f	22	25	28	25	21	9	-23	-34	-49	-40	-29	-12	9	20	13	18	13
g	-3	-8	-18	-17	-7	2	26	68	97	58	19	-2	-18	-14	-18	-11	-11
h	15	9	-4	-7	8	-2	12	8	8	5	-4	1	-3	-5	-5	-10	-9
i	7	12	19	14	7	1	-21	-42	-66	-55	-26	-14	14	18	4	2	1
k	-12	-7	-10	-9	-1	5	11	5	0	9	5	-8	-20	-15	-7	-10	-12
l	2	8	11	11	11	2	-23	-42	-65	-63	-52	-39	-15	-11	-10	-6	0
m	11	14	4	3	-9	-16	-33	-52	-62	-77	-71	-54	-32	-7	3	9	9
n	-2	-8	-11	1	8	12	32	51	61	31	18	6	-6	-8	-4	2	2
p	4	8	4	-1	5	15	39	76	120	159	98	59	32	17	11	3	0
q	-1	-11	-12	-15	-17	-4	5	-5	-13	1	1	2	-2	-5	-1	-9	-20
r	-4	-9	-8	-10	-10	-13	-18	-16	-14	-9	-14	-16	-14	-11	-5	-3	-2
s	-3	-4	-4	-4	4	11	22	26	41	31	20	13	3	5	4	8	11
t	-5	-5	-5	-4	-7	-5	0	2	15	21	29	30	19	7	3	-4	-5
v	3	17	20	20	8	-2	-26	-46	-68	-51	-20	3	25	24	23	15	11
w	5	9	28	28	12	-16	-32	-46	-53	-38	-20	5	13	30	9	2	16
y	10	7	12	7	6	3	7	-1	-31	-14	-11	11	13	1	3	12	15

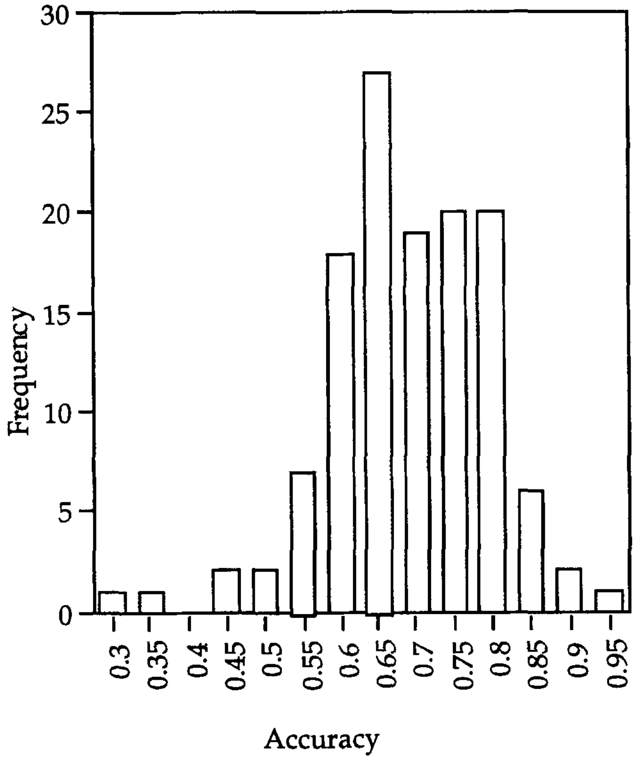

Fig. 2. Histogram of Q_3 accuracies obtained using DSC.

A few proteins sequences have low Q_3 accuracy: 2gn5 (gene V protein), 4rhv_4 (fourth chain in rhinovirus coat protein), 2mev_4 (fourth chain in cardio picornavirus coat protein), and 1shl (sea anemone neurotoxin 1) all have a $Q_3 < 50\%$. Protein 2gn5 is an incorrect structure, and the structure 1bgdh has replaced it. The true Q_3 accuracy of DSC on gene V protein is 73.6% (the structure of 2gn5 had only two β-stands of length two, 1bgdh has 7 strands). The two viral coat protein chains are not globular and probably should not have been included in the data set. Removal of the two viral coat protein and replacement of 2gn5 with 1bgdh increases the overall Q_3 accuracy of DSC to 70.3%. The protein 1shl is an averaged NMR structure; it has a small disulfide-rich structure with an unusually large amount of β-strand.

Comparison with the PHD prediction method

The prediction accuracies for the PHD prediction method (Rost & Sander, 1993a) based only on the nonredundant database of 126 protein chains were obtained from the authors. This allowed a direct comparison of DSC with the popular PHD algorithm. A comparison with predictions obtained from the PHD e-mail server would have been biased against DSC, because the server uses information from more proteins (including the test proteins). Unfortunately, it was not possible to obtain the actual residue predictions made by PHD. This would have allowed a more sensitive statistical comparison to have been made, although care would have had to be taken in interpreting such results because residues within a particular protein are far from being statistically independent of each other. This is not a problem when comparisons are done at the chain level.

On a protein by protein basis, DSC had a mean chain accuracy of 71.3%, and PHD has an accuracy of 71.7%. The standard deviation of chain accuracy for DSC is 10.5%, that of PHD, 8.9%; the two prediction measures have a correlation of 0.72. There was no statistical difference at $P < 0.05$ in the accuracies of the two methods (using a two-tail binomial test based on greater prediction accuracy and a normal approximation). Removal of the incorrect structure 2gn5 and the two nonglobular viral chains increases the chain accuracy of DSC to 72.1% and that of PHD to 72.0%.

For medium-length chains (≥90 residues and <170 residues), DSC was more accurate than PHD ($P < 0.01$). For short chains (<90 residues), there is no significant difference between the methods, and for long chains (≥170 residues), PHD is significantly more accurate than DSC ($P < 0.01$). It is not understood why there should be differences in prediction accuracy based on chain length. It may have to do with differences in residue frequency (White, 1992), or it may be that the decomposition of concepts used by DSC does not take into account domain boundaries.

The combined DSC-PHD prediction algorithm: if medium length, then run DSC, else run PHD, has an overall three-state accuracy of 72.4%—the *highest* accuracy reported for any prediction method. This prediction method has a mean accuracy of 72.9% for protein chains and standard deviation of 9.2%.

Predictive accuracy with increasing information

Secondary structure prediction was conducted using successively more information (Table 1). This step-by-step addition of new information allowed estimation of the relative importance of the different concepts and sources of information used (all accuracy estimates were made using leave-one-out cross-validation).

1. Prediction was performed using residue propensities from aligned homologous sequence information (Run1). This is the standard GOR method extended for homologous sequences. This has an estimated Q_3 of 63.5%.
2. Linear discrimination was then applied using the following attributes: residue propensities, distance from chain, hydrophobic moments, insertions, deletions, and conservation moments (Run2). This increased the estimated Q_3 accuracy to 67.8%.
3. Auto-correlation of protein structure was taken into account by using smoothed residues (Run3). Auto-correlation is when neighboring states are correlated, e.g., in an α-helix. This increased the estimated Q_3 accuracy by 0.5 percentage points to 68.3%.
4. The fractions of residues predicted to be α-helix and β-strand secondary structure are then added, along with the fraction of residues of type histidine, glutamate, glutamine, aspartate, and arginine (Run4). This increased estimated Q_3 accuracy to 69.4%. The residues were selected using stepwise linear regression; note that they are all highly hydrophilic residues. The attributes in step four are full-sequence based, not individual-residue based.
5. The predictions were finally filtered to produce the final predictions with an estimated Q_3 accuracy of 70.1% (Run5).

Prediction of expected accuracy

It was investigated whether it was possible to predict the likely accuracy of a secondary structure prediction. This would allow the accuracy and hence usefulness of a prediction to be assessed before use.

Standard stepwise linear regression was used with the dependent variable being accuracy, and the independent variables selected were: mean difference in probability (Dp), fraction of residues that are valine (Rv), and fraction of residues that are glutamic acid (Re). The mean difference in probability for a protein is defined to be the mean difference between the probability of the most likely state and the second most likely state at each residue. This produced the simple regression equation:

$$\text{Predicted accuracy} = 0.713 + (0.0612 * \text{Dp}) + (0.0236 * \text{Rv}) + (0.0180 * \text{Re}).$$

This equation predicts accuracy quite well, with a correlation of 0.58 (Fig. 3). The predicted accuracies, along with the actual accuracies, are given in Table 2. The residue frequencies used in the equation are "centered" to show their relative importance (see below). The constant in the equation is the mean accuracy, and the most important variable is Dp. Valine has the highest propensity of any residue for β-strands and glutamic acid has the highest propensity of any residue for α-helices (see Tables 3 and 4).

Proteins that are predicted to have an accuracy ≥80% have an average accuracy of 82.4% (6% of the database); proteins that are predicted to have an accuracy ≥75% have an average accuracy of 77.6% (20% of the database); and proteins that are predicted to have an accuracy ≥70% have an average accuracy of 72.7% (52% of the database).

It is also interesting to note that proteins that are predicted to have low ratios of β-strand (<3%) have an average accuracy of 80% (10% of database). Such proteins are likely to have alpha-type domains. A number of authors have noted that it is possible to predict alpha-type domain proteins with an accuracy ≈80% (Muggleton et al., 1992; Rost & Sander, 1993b). However, such work was based on a priori knowledge of the domain type before prediction. The present accuracy of 80% was obtained without knowledge of domain type.

The proteins with Q_3 accuracy < 50% (2gn5, 4rhv_4, 2mev_4, and 1shl) are identified as outliers in the regression equation (see above). It might be possible to use gross discrepancy between predicted secondary structure accuracy and actual accuracy to identify errors in structure. The true Q_3 accuracy of DSC on 2gn5 (gene V protein) is 73.6%, close to the predicted accuracy of 76.4%.

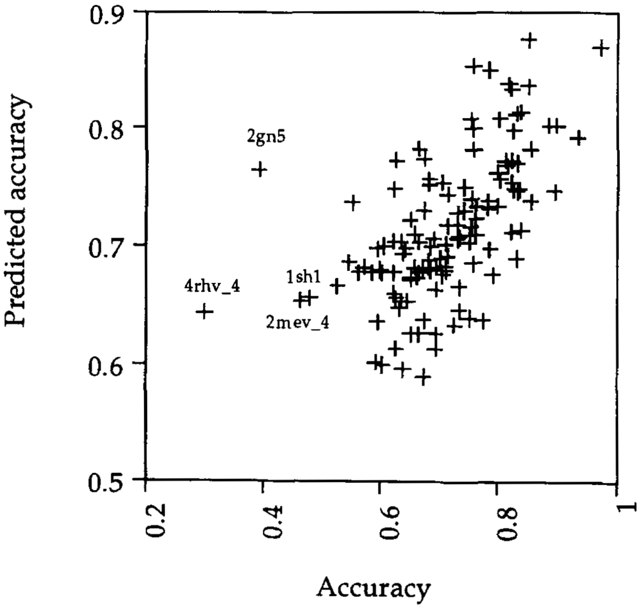

Fig. 3. Scatter diagram of actual Q_3 accuracy versus predicted accuracy.

Discussion

Relative importance of the amino acid residues for prediction

The Garnier Osguthorpe Robson (GOR) method (Gibrat et al., 1987) generates information tables that can be interpreted to provide information about the relative importance of the different residues in formation of secondary structure (Tables 3, 4, 5). This method of representing specificity for initiating and terminating particular secondary structures differs from the technique used most commonly of calculating the propensity of a residue for being at the start or end of a particular secondary structure (e.g., Richardson & Richardson, 1988; Wako & Blundell, 1994). However, such studies have problems in defining where the beginnings and ends occur and in gathering enough statistical data. They also do not take into account neighborhood effect. The GOR approach takes into account such effects and works well for structures ≤17 residues long ($i - 8$ to $i + 8$). For longer structures, there is overlap between central regions of structure and initiation and termination regions.

The residues that most strongly favor and disfavor particular states are given in Table 6. Histidine is the least informative residue, i.e., the residue that has the lowest mean preference for any one particular secondary structure. This means that, for a protein of unknown structure, mutating a residue into a histidine is the least likely to disrupt the secondary structure of the protein. It is interesting that a number of the residues that disrupt β-strand conformation (p, d, n, g) only do so locally ($i - 2$ to $i + 2$), and that these residues favor β-strand conformation in more distant residues (probably by formation of turns). Note also the large role that charged residues play in initiating and terminating secondary structure. This role seems to be reversed between α-helices and β-strands, with histidine and lysine terminating α-helix conformation and initiating β-strand conformation, and aspartic acid initiating α-helix conformation and terminating β-strand conformation. The results for α-helices are explained by existence of a dipole with the N-terminus being negative. The reason for the β-strand results are unclear, but similar results are found in Colloc'h and Cohen (1991).

It is interesting to compare the information measures with results obtained investigating residue α-helix propensities (Bryson et al., 1995), both previous statistical work (e.g., Gibrat et al., 1987; Williams et al., 1987), and the larger literature based on direct experimentation, e.g., using directed mutagenesis (Horovitz et al., 1992; Blaber et al., 1993) and model peptides (Padmanabhan et al., 1990). The correlation of the DSC intraresidue information for the 20 residues with α-helix propensities of the other methods is: Gibrat et al. (1987), 0.94; Williams et al. (1987), 0.93; Blaber et al. (1993), 0.81; Horovitz et al. (1992), 0.67 (19 values); and Padmanabhan et al. (1990), 0.87 (5 values). The use of statistical surveys has the advantage over physical-based methods in being based on many more proteins, and thus averaged over all types of protein environment. This allows greater confidence that the results obtained will hold for new proteins, and that they are not the indirect effect of some peculiarity of the protein under study. The larger amount of data used also has the advantage that higher level correlations (such as between pairs of residues) can be investigated, not just the simple relationship between residue type and secondary structure. Statistical analysis has the disadvantage that it is harder to discern direct physical mechanisms for propensities and dissect out tertiary from secondary structure effects.

Table 6. *Residues that favor and disfavor particular secondary structure classes*

	α-Helix	β-Strand
Favor	e, a, l, m	v, i, f, y, c
Disfavor	p, g, c, t	p, d, n, e, g, a
N-Cap	d	h, k, f
C-Cap	k, r, h	p, d

Relative importance of the different predictive attributes

In a discrimination function, the larger the modulus of a weight, the more important (informative) the attribute is for predicting a particular secondary structure. Positive values favor a particular prediction state, and negative values disfavor a state. To allow direct comparison between the different attribute types, all theattributes values were centered on a standard scale. This is done for an attribute type by subtracting from each value the mean of the attribute values and dividing by the standard deviation. Centering does not affect the predictions of linear discrimination. The constants in the function correspond to the logs of the prior probabilities of the different states. If no observations have been made about a residue, then the residue is a priori most likely to be in coil conformation and least likely to be in β-strand conformation. The prior probabilities are given to the system.

1. The most important attributes are the GOR information parameters with values of ≈0.7 for α-helix and β-strands. This shows that the first two terms of the GOR decomposition captures most of the information used in prediction.

2. Residues close to the end of a chain are most likely to be in chain conformation (note that distance is calculated *away* from the edge). The termini residues are charged, and the terminal regions are generally more mobile and more variable in sequence than other parts of the sequence (Thornton & Sibanda, 1983). Terminal residues probably have difficulty in tightly packing to the core and in forming hydrogen bonds.

3. The hydrophobic moment pattern of α-helices is more pronounced than that of β-strand. This may be because there are two distinct types of β-strand, those fully buried and those that show amphipathic behavior (see Lim, 1974b). There are relatively few fully buried helices. As expected, the hydrophobic moment discriminates mainly between α-helices and β-strands (the weights are near zero for coil).

4. Insertions and deletions, as expected, favor coil conformation. Insertions favor coil more than deletions do (i.e., they are more informative)—this is not an artifact of their slightly different coding. It is also interesting that deletions disrupt α-helix structure much less than insertions, whereas they are equally disruptive of β-strand. The more disruptive nature of insertions suggests that mutations causing insertions may be less likely than deletions, and insertions and deletions should not have the same weighting in homologous alignment (see below).

5. The conservation moment pattern of α-helices and β-strands is about equal; it discriminates mainly between α-helices and β-strands.

6. Smoothing has the most effect on β-strand formation, and this can be explained by cooperative effects in β-strand formation. The hydrophobic moment and conservation moment are the attributes most effected by smoothing.
7. The residue fractions that were identified as being most important are glutamate and arginine.
8. The strongest feedback effect from predicted fraction of secondary structure is for β-strands to negatively feedback the formation of α-helices and to positively feedback formation of β-strands. This, like the smoothing effects, can be explained by cooperative effects in β-strand formation.

Consensus predictions

Many workers use aligned homologous sequences to make consensus secondary structure prediction (Russell & Barton, 1993; Jenny & Benner, 1994; Rost et al., 1994). That is, they make some form of average structural prediction for the aligned sequences. This is the correct procedure if the prediction is of some structural property of all the sequences, for example, the fold type of the sequences. It is not appropriate if interest is focused on one particular sequence, for example, in modeling a protein.

In the present work, we are interested in predicting the secondary structures of particular protein chains, not in making consensus predictions. Two homologous sequences may have variable secondary structure and, consequently, it has been suggested that there is a lower limit on Q_3 accuracy than 100%. This is correct for consensus predictions, but does not apply for the prediction of single sequences using aligned sequences. Because homologous sequences may have different secondary structures, it is to be expected and hoped that homologous sequences can have different predicted secondary structures. This is the case with the DSC program. To see why this is so, consider the simple case of two aligned sequences:

```
  12345678901234567
A ETARACCAYREVSICSD
B ETARACCA.REVSICSD
```

The attribute vectors formed for these residues will be different. The calculated GOR potentials will be different because the sequences are different, e.g., in sequence A for position 8, the $i + 1$ residue is a tyrosine, and for sequence B, it is an arginine. With respect to sequence A, sequence B has a deletion at position 9; and with respect to sequence B, sequence A has an insertion at position 8. All the other attributes will differ as well.

The same argument extends to the alignment of sequences. If the role of the alignment is for predicting a single sequence, it might be better to form an alignment optimized for that particular sequence, not an overall optimization, i.e., different alignments would be formed depending on which protein was chosen as seed.

Measures of prediction success

Many alternative methods of measuring the success of prediction have been proposed (Schulz & Schirmer, 1979; Rost & Sander, 1993a; Rost et al., 1994; King, 1996). These methods, like prediction accuracy, are based on the assumptions that there exists a single true secondary structure assignment for all residues and we know what the assignment is. These assumptions are incorrect. In protein structures, there are intermediate and difficult-to-classify regions and these residues cannot be assigned unambiguously to any one class of secondary structure (Colloc'h et al., 1993). Secondary structure assignment should be probabilistic—with probabilities near 1.0 for well-defined secondary structures and high-resolution coordinates, and correspondingly lower probabilities for the edges of secondary structures in lower resolution structures. Prediction of secondary structure should also be probabilistic, because this gives the user of the prediction more information. The PHD prediction algorithm produces a "reliability" measure from 0 to 9, which can be interpreted as a probability. The DSC prediction method makes full probabilistic predictions.

If both assignment and prediction of secondary structure are probabilistic, there is natural measure of prediction success.

$$D = \sqrt{\frac{\left(\sum_{1}^{N} a^2\right)}{N}} + \sqrt{\frac{\left(\sum_{1}^{N} b^2\right)}{N}} + \sqrt{\frac{\left(\sum_{1}^{N} c^2\right)}{N}}.$$

N is the number of residues, a the difference between actual and predicted probability for an α-helix, b the difference for β-strands, and c the difference for coil (1 of a, b, or c is redundant). The minimum this function can take is 0 (optimal prediction) and the maximum 3.

This measure gives, for the DSC predictions, a value of 1.07 (0.495 from α-helix, 0.234 from β-strand, and 0.347 from coil). This is a large overestimate because it is based on all assigned secondary structures having a probability of 1.0.

The prediction measure could also be simply extended to take into account differing costs of predictive error. This could be done by multiplying the costs by the squared difference in probabilities. Differing costs would be important if different types of error were considered to make successful conformational prediction less likely (Jenny & Benner, 1994; King, 1996).

GOR information

The possibility of using the next term in GOR decomposition was investigated. This term refers to information a residue caries about another residue's secondary structure, which depends on the other residues type, (side-chain–side-chain interactions) (Robson, 1976). This type of information was used by Gibrat et al. (1987). Following their procedure, 915 pairs of side-chain–side-chains were found to be significant of 20,400 (20 * 20 * 3 * 17) possible pairs at ($P < 0.01$) using a χ^2 test. It was not found possible to improve overall predictive accuracy by inclusion of side-chain–side-chain information—even when resampling and Bayesian statistical methods were used to try to better estimate the information values. The reasons for this are unclear.

Conclusions

The DSC prediction method is simple, accurate, and can be programmed easily. This combination is achieved by identification of a set of concepts important for prediction followed by the use of linear statistics, allowing the relative importance of the different features used to be compared and measured.

The general DSC approach of identifying important concepts and using linear discrimination for prediction could be extended easily to include new features that are considered important. For example, if CD information was available, this could be added as two new attributes in the attribute vector, and the relative usefulness of this feature measured against that of the other features. Any other relevant feature could be treated similarly.

The DSC prediction method compares favorably with existing methods based on expert manual intervention or complex nonlinear statistics. Compared with manual intervention methods, DSC can be used by non-experts and its results are fully reproducible. Many of the concepts used by experts in protein structure prediction are quantified and made explicit in DSC. The DSC prediction method is much simpler than those based on nonlinear statistical prediction methods. The DSC method has ≈1,000 variables, whereas the neural-network PHD method (Rost & Sander, 1993a) has ≈25,000 variables, and the nearest-neighbor NNSSP method (Salamov & Solovyev, 1995) requires use of the whole data set, ≈500,000 aligned residues.

The simplicity of DSC makes it possible to reimplement the method directly from the information in this paper. It is not much more complicated to implement than the GOR method, which is used widely in molecular biology software packages. The complete reimplementation of a secondary structure prediction method is desirable if the prediction method is to be employed in a larger program, for example, as a module in a threading algorithm (Wodak & Rooman, 1993), hence progressing from secondary to tertiary structure prediction.

Materials and methods

Data

The protein data set used in this study was the 126 representative globular protein chains used in the study of Rost and Sander (1993a); we did not include the four membrane protein chains considered by Rost and Sander. This data set was also used by Salamov and Solovyev (1995). The protein chains have less than 25% pairwise similarity for lengths >80. There are 23,336 residues, 7,409 (31.7%) in α-helix conformation, 5,044 residues (21.6%) in β-strand conformation, and 10,883 (46.7%) in coil conformation (see Rost & Sander, 1993a for details of assignment of secondary structure). Each protein chain has associated with it zero or more aligned sequences. These derive from the database of Homology-derived Structures and Sequence alignments of Proteins (HSSP) (Sander & Schneider, 1991) and were used in the predictions of Rost and Sander (1993a) and Salamov and Solovyev (1995).

Prediction measurements

The main measure of prediction success was standard per residue Q_3 prediction accuracy. This is defined as {(number of residues correctly predicted)/number of residues} * 100. This measures the expected accuracy of predicting an unknown residue. A more relevant variant of this measure is mean per protein Q_3 accuracy. In this measure, the mean of the Q_3 accuracies for each individual protein is taken. This measures the expected accuracy of predicting an unknown protein. The accuracy of prediction for the three types of secondary structure was also measured, and, in addition, the Matthews' correlation coefficient (Matthews, 1975) was calculated for each secondary structure type:

$$C = \frac{pn - uo}{\sqrt{(p + u)(p + o)(n + u)(n + o)}}$$

with p being the number of residues correctly positively predicted, n being the number of residues correctly negatively predicted, u the number of false negatives, and o the number of false positives.

The measures of success were estimated using leave-one-out cross-validation (jack-knifing). This method was also used by Salamov and Solovyev (1995). It is less biased than sevenfold cross-validation, the method adopted by Rost and Sander (1993a) to estimate the accuracy of the PHD algorithm (chosen because of the slow training speeds of neural networks).

Residue propensities and GOR

The simplest concepts used in prediction were the propensities of residues for particular secondary structure states. These residue propensities were calculated using the method developed in the GOR secondary structure prediction method (Robson & Suzuki, 1976; Garnier et al., 1978; Gibrat et al., 1987). The GOR method provides an elegant technique of decomposing the various ways residues can interact to form secondary structure by order of simplicity—single residues, pairs, etc.

Ideally, the secondary structure of a residue would be calculated using the propensities (information terms) from all possible terms in the decomposition: this would be equivalent to calculating the Bayesian optimal prediction rule (Weiss & Kulikowski, 1991). However, this is unfeasible because it would require a vast amount of structural information to estimate accurately all the terms in the decomposition. Currently, there is only enough data to use the first two terms in the decomposition. These are: (1) information a residue carries about its own secondary structure—intraresidue information, (side-chain–own backbone interaction); and (2) information a residue carries about another residue's secondary structure that does not depend on the other residue's type—directional information (Robson & Suzuki, 1976). Ignoring the other terms can be thought of as assuming that residues do not interact in any other way in forming secondary structure. The GOR method can probably be best understood as a variety of the "naive" Bayesian statistical method (Weiss & Kulikowski, 1991).

The directional information measures were calculated using the data set of 126 chains (information from the aligned sequences was not used for this because it is not statistically independent). As in the GOR method (Garnier et al., 1978), information parameters were calculated for the 20 residues for the three conformation states at positions $i - 8$ to $i + 8$, giving $20 * 3 * 17 = 1{,}020$ parameters. The information measures were estimated directly from frequencies, because the sample size is large enough to preclude the need for a Bayesian estimation method (as recommended originally [Robson & Suzuki, 1976]). These information measures are closely related to probabilities, but they have the advantage of being simply additive (because the decomposition ensures that the same information is not counted twice and they are based on logs). To predict the secondary structure of a residue, the relevant information terms are gathered together and summed, and the secondary structure with the highest information is then predicted.

Other residue-based concepts in prediction

Apart from the first two terms in the GOR decomposition of residue interaction, it was possible to identify two other concepts based on primary structure that are important for prediction of secondary structure. These are: distance from the end of the chain, and the moments of hydrophobicity. The distance to the end of chain is important because residues near the end of a chain have fewer structural constraints, allowing greater flexibility. This concept has, to our knowledge, not been explicitly used in secondary structure before.

1. Distance from end of chain is calculated as the number of resides (to a maximum of 5) to the nearest end of chain.
2. The moment of hydrophobicity (Eisenberg, 1984) is calculated for each residue under the assumption that it, and the three neighboring residues in each direction, are in α-helix conformation (100°); the Eisenberg hydrophobicity scale is used. The moment of hydrophobicity is also calculated assuming β-strand conformation (180°). This is informative because, if the hydrophobicity profile suits a particular secondary structure conformation, a large value will be produced. Similar information is calculated in Wako and Blundell (1994).

Information from aligned sequences

Aligning homologous sequences provides additional information for predicting secondary structure. The simplest way this information was used was to calculate the mean of the summed GOR information terms for aligned residues. This is equivalent to extending the GOR prediction method to include homologous information (Zvelebil et al., 1987). It may have been possible to produce more accurate results by a more sophisticated method of combining the information in the sequence (Russell & Barton, 1993). The moment of hydrophobicity was also simply extended for aligned sequences by taking the mean value for the sequences.

Three other ways of using aligned sequence information were identified. These are: aligned deletions, aligned insertions, and the moments of conservation for α-helix and β-sheet.

1. Deletions are relative to the predicted primary structure, i.e., the homologous sequence has a missing residue. Deletions are treated as "indicator" variables, represented by "1" if an insertion is observed at that position in any homologous sequence, and by "0" if no insertions are observed.
2. Insertions are also relative to the predicted primary structure, i.e., the homologous sequence has one or more extra residues. Insertions are treated in a similar way to deletions, with an indicator of "1" for the residue where the insertion starts and "0.5" for its direct neighbors.
3. The moment of conservation is calculated in an analogous manner with moment of hydropathy, with the conservation measure of entropy used in place of hydrophobicity. Entropy is a robust measure of the degree of variability of residue type at a position. Entropy is defined as $-\Sigma p_r * \log2(p_r)$ (where p_r is the probability of a residue type at the position). The moment of conservation is a quantification of the important concept used in visual inspection of multiple sequences. This concept has, to our knowledge, not been used explicitly in secondary structure before.

Attribute vectors

For each residue position, an "attribute vector" was formed using the information from the different calculated quantities. For example, in the first residue in protein 1acx (actinoxathnin), the attribute vector for Run2 is, before centering:

$$[-2.170409, -0.30941, 1.31876, 1, 1.21334, 0.5480, 0, 0, 1.88054, 0.72193].$$

The first three values are the summed GOR predicted information measures (averaged over all homologous sequences), in order α-helix, β-strand, coil. The high value for coil indicates that, using the GOR prediction measure, the residue is predicted to be in coil conformation. The residue is at the edge (position 1). The hydrophobic moment assuming α-helix is 1.21334, assuming β-strand is 0.5480. There are no insertions or deletions at this position, and the conservation moment assuming α-helix is 1.88054; assuming β-strand, is 0.72193. Centering the attributes produces the vector:

$$[-0.83046, -0.72167, 1.04627, -5.2387, -0.59548, -1.06786, -0.0967, -0.41774, 0.65233, -0.68956].$$

Linear discrimination

Prediction of secondary structure was made from the attribute vectors using linear discrimination (Weiss & Kulikowski, 1991; Michie et al., 1994) (Fig. 1). The secondary structure of each position was predicted using a leave-one-out cross-validated linear discrimination function. The equivalent functions formed using all 126 chains are given in Tables 7, 8, and 9. The Minitab statistical package was used to apply linear discrimination (Minitab Inc., Pennsylvania State University, Pa).

Linear discrimination is probably the most commonly used statistical prediction method. It is robust and it produces simple-to-understand output (King et al., 1995). In linear discrimination, as the name suggests, a linear combination of evidence (the attributes) is used to separate or discriminate between classes and to assign a new example. For a problem involving n features, this means that the separating surface between the classes will be a $(n - 1)$ dimensional hyperplane. The general form of classifier is: $w_1e_1 + w_2e_2 + \ldots + w_ne_n + w_0$. This discrimination function is optimal assuming a multivariate normal distribution and pooled covariance matrix (Weiss & Kulikowski, 1991). For each class to be discriminated, a number is calculated (related to a probability using the linear function and the attribute vector. For example, the number A for α-helix using the above centered attribute vector and the function in Table 7 would be calculated:

$$\begin{aligned} A = {} & -1.5862 + (0.6589 * -0.83046) + (-0.1734 * -0.72167) \\ & + (-0.2274 * 1.04627) + (0.1903 * -5.23387) \\ & + (0.2848 * -0.59548) + (-0.1881 * -1.06786) \\ & + (-0.0815 * -0.0967) + (-0.2319 * -0.41774) \\ & + (0.1600 * 0.65233) + (-0.0477 * -0.68956). \end{aligned}$$

Table 7. *Discrimination function for the three secondary structure classes formed in the Run2 predictions*

Parameters	α-Helix	β-Strand	Coil
Constant	−1.5862	−2.0802	−0.9931
Mean potential for α-helix	0.6589	−0.4006	−0.2629
Mean potential for β-strand	−0.1734	0.7275	−0.2629
Mean potential for coil	−0.2274	−0.5236	0.3975
Distance to edge	0.1903	0.1528	−0.2004
Hydrophobic moment α-helix	0.2848	−0.3173	−0.0468
Hydrophobic moment β-strand	−0.1881	0.2107	0.0304
Deletions	−0.0815	−0.1483	0.1242
Insertions	−0.2319	−0.1655	0.2346
Conservation moment α-helix	0.1600	−0.1520	−0.0385
Conservation moment β-strand	−0.0477	0.1301	−0.0279

Table 8. *Discrimination function for the three secondary structure classes formed in the Run3 predictions*

Parameters	α-Helix	β-Strand	Coil
Constant	−1.6109	−2.1138	−0.9969
Mean potential for α-helix	0.5093	−0.5412	−0.0959
Mean potential for β-strand	−0.0612	−0.0787	0.0781
Mean potential for coil	−0.2205	−0.6998	0.4744
Distance to edge	0.1526	0.3724	−0.0687
Hydrophobic moment α-helix	−0.0936	−0.0359	0.0803
Hydrophobic moment β-strand	0.0339	−0.0511	0.0006
Deletions	−0.1763	−0.1259	0.1784
Insertions	−0.1295	−0.2019	0.1818
Conservation moment α-helix	−0.0211	−0.0288	0.0277
Conservation moment β-strand	−0.0078	0.0190	−0.0035
Smoothed mean potential for α-helix	0.2637	0.0181	−0.1879
Smoothed mean potential for β-strand	−0.0308	0.7174	−0.3115
Smoothed mean potential for coil	0.1162	0.0039	−0.0809
Smoothed distance to edge	0.3445	−0.2241	−0.1307
Smoothed hydrophobic moment α-helix	0.4178	−0.3148	−0.1385
Smoothed hydrophobic moment β-strand	−0.2461	0.2892	0.0335
Smoothed deletions	0.1136	−0.0296	−0.0636
Smoothed insertions	−0.1081	0.0287	0.0603
Smoothed conservation moment α-helix	0.2222	−0.1794	−0.0681
Smoothed conservation moment β-strand	−0.0841	0.1930	−0.0322

Table 9. *Discrimination function for the three secondary structure classes formed in the Run4 predictions*

Parameters	α-Helix	β-Strand	Coil
Constant	−1.6910	−2.2278	−1.0004
Mean potential for α-helix	0.6587	−0.7654	−0.0937
Mean potential for β-strand	−0.0253	−0.1277	0.0764
Mean potential for coil	−0.0801	−0.8902	0.4671
Distance to edge	−0.2635	0.4183	−0.0144
Hydrophobic moment α-helix	−0.0904	−0.0418	0.0809
Hydrophobic moment β-strand	0.0236	−0.0394	0.0022
Deletions	−0.1690	−0.1341	0.1772
Insertions	−0.1359	−0.1892	0.1802
Conservation moment α-helix	−0.0127	−0.0381	0.0263
Conservation moment β-strand	−0.0097	0.0215	−0.0034
Smoothed mean potential for α-helix	0.2982	−0.0068	−0.1998
Smoothed mean potential for β-strand	0.0707	0.5340	−0.2956
Smoothed mean potential for coil	0.1674	−0.1170	−0.0597
Smoothed distance to edge	0.4765	−0.2852	−0.1922
Smoothed hydrophobic moment α-helix	0.4186	−0.3070	−0.1427
Smoothed hydrophobic moment β-strand	−0.2271	0.2737	0.0277
Smoothed deletions	0.0787	−0.0046	−0.0515
Smoothed insertions	−0.0954	−0.0010	0.0654
Smoothed conservation moment α-helix	0.1820	−0.1425	−0.0579
Smoothed conservation moment β-strand	−0.0588	0.1532	−0.031
Fraction predicted α-helix	0.1175	−0.2487	0.0353
Fraction predicted β-strand	−0.4374	0.4402	0.0938
Fraction of residues histidine	0.0040	−0.0488	0.0199
Fraction of residues glutamate	−0.2106	0.1835	0.0584
Fraction of residues glutamine	−0.0701	0.0616	0.0192
Fraction of residues aspartate	−0.0953	0.0902	0.0231
Fraction of residues arginine	−0.1198	0.1094	0.0308

The class with the largest number from its function is predicted to be present.

Post-processing

A linear discrimination function cannot capture all information necessary for prediction. In particular, it cannot directly include auto-correlation, secondary structure feedback effects, and neighborhood constraints on secondary structures. For this reason, the predictions from the second-level linear discrimination function were filtered to produce the final predictions.

During the folding process, stretches of secondary structure interact and affect the formation of other secondary structures. These interactions may be positive or negative. Such feedback interactions cannot be captured in a linear model based on the attributes described above. Therefore, feedback was modeled in two stages, by use of smoothed attributes and by use of the fraction of residues predicted to be in α-helix and β-strand conformation (the ratio of coil is redundant because it is linearly dependent on the ratios of α-helix and β-strand). The smoothing method used was the standard one in the Minitab statistical package. It consists of a running median of 4, then 2, then 5, then 3, followed by a Hanning smooth $\{(0.25 * i - 1) + (0.5 * i) + (0.25 * i + 1)\}$.

The fraction of residues of particular types has been recognized previously to have a role in secondary structure prediction (Rost & Sander, 1994). The role of these types seems to be in determining the structural class of the chain. The fractional content of not all residues are important. The important ones were determined by stepwise linear regression and are: histidine, glutamate, glutamine, aspartate, and arginine. All these residues are highly hydrophilic. All these residues, with the exception of histidine, favor β-strand formation.

The final predictions were filtered/smoothed to make them more realistic by removing physically unlikely sequences of conformation. Filtering is now standard in secondary structure prediction, and is used in the most successful methods (Rost & Sander, 1993a; Salamov & Solovyev, 1995). The following if-then rewrite rules were used for filtering:

[¬a,¬a, c, **b**, *,¬b] → **c**

[¬a, *, *, **a**, b] → **b**

[¬a, *, *, **a**, c] → **c**

[a, *, *, **a**, c, *,¬c] → **c**

[¬a,¬a, **a**, a, c,¬a] → **c**

$$[\neg a, c, \neg c, \mathbf{a}, a, c, \neg a] \rightarrow \mathbf{c}$$

$$[\neg a, c, c, \mathbf{a}, a, \neg b, \neg a] \rightarrow \mathbf{c}$$

$$[a, c, *, \mathbf{a}, a, a, \neg a] \rightarrow \mathbf{c}$$

$$[*, c, *, \mathbf{a}, a, b, \neg a] \rightarrow \mathbf{c}$$

$$[c, b, b, \mathbf{a}, a, *, a] \rightarrow \mathbf{b}$$

$$[c, *, \mathbf{a}, a, \neg a, a] \rightarrow \mathbf{c}$$

a = α-helix, b = β-strand, c = coil, * = wildcard (α-helix or β-strand or coil), ¬ = not.

If the pattern on the left is met in a prediction, then the secondary structure in bold on the left is rewritten as the secondary structure on the right of the rule. For example:

$$[b, b, b, \mathbf{a}, c] \rightarrow [b, b, b, \mathbf{c}, c]$$

$$[b, b, c, \mathbf{a}, c] \rightarrow [b, b, c, \mathbf{c}, c]$$

$$[b, b, b, \mathbf{a}, b, b, b] \rightarrow [b, b, b, \mathbf{b}, b, b, b].$$

The filtering rules were found using the machine learning algorithm CART with 10-fold cross-validation (Breiman et al., 1984); as in other prediction methods, the rules were taken as given a priori (Rost & Sander, 1993a; Salamov & Solovyev, 1995). It is interesting that α-helix structure is the type of structure most in need of filtering.

Availability

The DSC program for protein secondary structure prediction is available via a server on the Internet: http://www.icnet.uk/bmm/dsc_read_align.html. The DSC program (with small manual and source code in C) is also freely available from ftp://ftp.icnet.uk/icrf-public/bmm/king/dsc; this allows use of DSC without recourse to the Internet. The databases used to develop DSC are available on request from rd_king@icrf.icnet.uk, allowing re-implementations of the algorithm.

Acknowledgments

We thank B. Rost and C. Sander for providing the aligned data set and the results of PHD trained on this data set.

References

Benner SA, Cohen MA, Gerloff D. 1992. Correct structure prediction. *Nature 359*:781.

Benner SA, Gerloff D. 1990. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. *Adv Enz Reg 31*:121–181.

Benner SA, Gerloff DL. 1993. Predicting the conformation of proteins. *FEBS Lett 325*:29–33.

Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. 1988. Secondary structure prediction: Combination of three different methods. *Protein Eng 2*:185–191.

Blaber M, Xue-jun Z, Matthews BW. 1993. Structural basis of amino acid α-helix propensity. *Science 260*:1637–1640.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and regression trees.* Wadsworth: Belmont.

Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF. 1995. Protein design: A hierarchic approach. *Science 270*:935–941.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Cohen FE, Abarbanel RM, Kuntz ID, Fletterick RJ. 1983. Secondary structure-assignment for alpha/beta proteins by a combinatorial approach. *Biochemistry 22*:4894–4904.

Colloc'h N, Cohen FE. 1991. β-Breakers: An aperiodic secondary structure. *J Mol Biol 221*:603–613.

Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins the advantages of consensus assignment. *Protein Eng 6*:377–382.

Dowe LD, Oliver J, Dix T, Allison L, Wallace CS. 1993. A decision graph explanation of protein secondary structure prediction. In: Mudge TN, Milutinovic V, Hunter L, eds. *Proceedings of the 26th Annual Hawaii International Conference on System Sciences.* IEEE Computer Society Press. pp 669–678.

Eisenberg D. 1984. Three-dimensional structure of membrane and surface proteins. *Annu Rev Biochem 53*:595–623.

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

Geourjon C, Deleage G. 1994. SOPM: A self optimised prediction method for protein secondary structure prediction. *Protein Eng 7*:157–164.

Gibrat JF, 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.

Horovitz A, Matthews JM, Fersht AR. 1992. α-Helix stability in proteins. II. Factors that influence stability at an internal position. *J Mol Biol 227*:560–568.

Jenny TF, Benner SA. 1994. Evaluating predictions of secondary structure in proteins. *Biochem Biophys Res Commun 200*:149–155.

King RD. 1996. Secondary structure prediction. In: Sternberg MJE, ed. *Protein structure prediction: A practical approach.* Oxford: Oxford University Press. Forthcoming.

King RD, Feng C, Sutherland A. 1995. StatLog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence 9*:289–335.

King RD, Sternberg MJE. 1990. Machine learning approach for the prediction of protein secondary structure. *J Mol Biol 216*:441–457.

Kneller DG, Cohen FE, Langridge R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol 214*:171–182.

Lim VI. 1974a. Structural principles of the globular organisation of protein chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol 80*:857–872.

Lim VI. 1974b. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol 88*:873–894.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta 405*:442–451.

Mehta PK, Heringa J, Argos P. 1995. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci 4*:2517–2525.

Michie D. 1986. The superarticulacy phenomenon in the context of software manufacture. *Proc R Soc Lond (A) 405*:185–212.

Michie D, Spiegelhalter DJ. Taylor CC. 1994. *Machine learning, neural and statistical classification.* London: Ellis Horwood.

Muggleton S, King RD, Sternberg MJE. 1992. Protein secondary structure prediction using logic. *Protein Eng 5*:647–657.

Padmanabhan S, Marqusee S, Ridgeway T, Laue TM, Baldwin RL. 1990. Relative helix-forming tendencies of nonpolar amino acids. *Nature 344*:268–270.

Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 202*:865–884.

Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of alpha helices. *Science 240*:1648–1652.

Robson B. 1976. Conformational properties of amino acid residues in globular proteins. *J Mol Biol 107*:327–356.

Robson B, Suzuki E. 1976. Conformational properties of amino acid residues in globular proteins. *J Mol Biol 107*:327–356.

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

Rost R, Sander C. 1993b. Secondary structure prediction of all-helical proteins in two states. *Protein Eng 8*:831–836.

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

Russell BR, Barton GJ. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J Mol Biol 234*:951–957.

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

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

Schulz GE, Schirmer RH. 1979. *Principles of protein structure*. New York: Springer-Verlag.

Solovyev VV, Salamov AA. 1994. Predicting α-helix and β-strand segments of globular proteins. *CABIOS 10*:661–669.

Thornton JM, Sibanda BL. 1983. Amino and carboxy-terminus regions in globular proteins. *J Mol Biol 167*:433–460.

Wako H, Blundell TL. 1994. Use of amino acid environment-dependent substitution tables and conformation propensities in structure prediction from aligned sequence of homologous proteins. II. Secondary structures. *J Mol Biol 238*:693–708.

Weiss SM, Kulikowski CA. 1991. *Computer systems that learn*. San Mateo: Morgan Kaufmann.

White SH. 1992. Amino acid preferences in small proteins. *J Mol Biol 227*:991–995.

Williams RA, Chang A, Juretic D, Loughran S. 1987. Secondary structure predictions and medium range interactions. *Biochim Biophys Acta 916*:200–204.

Wodak SJ, Rooman MJ. 1993. Generating and testing protein folds. *Curr Opin Struct Biol 3*:247–259.

Yi T, Lander ES. 1993. Protein secondary structure prediction using nearest-neighbour methods. *J Mol Biol 232*:1117–1129.

Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for predicting secondary structure prediction. *J Mol Biol 225*:1049–1063.

Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol 195*:957–961.