Protein Secondary Structure Prediction Using Two-level Case-based Reasoning

Bing Leng[1], Bruce G. Buchanan[1] and Hugh B. Nicholas[2]

[1]Intelligent Systems Laboratory
Department of Computer Science
[2]Pittsburgh Supper Computing Center
University of Pittsburgh
Pittsburgh, PA 15260

leng@cs.pitt.edu buchanan@cs.pitt.edu nicholas@psc.edu

Abstract

We have developed a two-level case-based reasoning architecture for predicting protein secondary structure. The central idea is to break the problem into two levels: first, reasoning at the object (protein) level, and using the global information from this level to focus on a more restricted problem space; second, decomposing objects into pieces (segments), and reasoning at the internal structures level; finally, synthesizing the pieces back to the objects. The architecture has been implemented and tested on a commonly used data set with 69.3% predictive accuracy. It was then tested on a new data set with 67.3% accuracy. Additional experiments were conducted to determine the effects of using different similarity matrices.

The Problem

Predicting protein secondary structure from its primary sequence, given a library of proteins of known structure, is a well known problem in computational biology. However, the problem is difficult in part because biological knowledge of protein structure has significant exceptions (*cf.* Hunter, AAAI-92). That is, the theory of protein secondary structure prediction has not been well formed yet. As a consequence, the current concept description languages used by machine learning researchers are far from adequate, which results in large numbers of disjuncts in concept definitions learned from examples (Rendell & Cho, 1990). Most machine learning work assumes that examples (segments of proteins) are independent, causing further difficulty, because the relationships between a protein and its segments, and relationships among segments are lost. To address these problems, we have designed a general architecture, which we call two-level case-based reasoning.

The Method

The method is similar to case-based reasoning in using reference cases from a library to make a prediction for a new case. This is often contrasted with inferring general rules by induction and using those rules for prediction. The two-level method is different from most case-based reasoning in two important respects, however: (1) more than one reference case is selected from the library, (2) mappings to the unknown case are made with respect to an analysis of the segments of the reference cases that are most similar to segments of the unknown.

The architecture has been implemented for the protein secondary structure prediction task. First, k similar proteins are selected from the protein structure library. The number of cases to be selected is dependent on the precision of the similarity metric used to select them, as discussed in the next section. With a measure based only on fractions of amino acids of each of 20 types, a large value of k is needed. When more domain knowledge is encoded in the similarity metric, as in a PAM matrix, smaller values of k will suffice. However, except in the trivial case of using homologous proteins to predict secondary structure, a single case from the library ($k = 1$) will not suffice because there is not enough knowledge to guide a similarity metric to exactly the case whose structure is closest to that of the unknown. Since many similarity matrices have been published in the biology literature and they all encode some biological knowledge, we decided to represent domain knowledge as a similarity matrix. The similarity value of each protein selected is used as its weight, and this weight is carried over to the second level to reflect the global information.

The second step of the method is to decompose the whole sequence into pieces (all segments of length w). Then we examine reference case to determine where the secondary structure associated with a piece can be used to predict the secondary structure of a corresponding piece of the unknown. Correspondence is determined by a similarity metric - either the same one used in step one or a different one. Then, evidence weights are combined. Each segment of a reference protein will assign its structure to a segment of the unknown with a weight equal to the product of its similarity value and the similarity weight of the protein it comes from as determined in step one. For each amino acid in the unknown weights are accumulated for three classes of structures: α-helix, β-strand, and coil. At any position along the unknown protein, the structure is

predicted as the one with the highest accumulated weight. The final step of the method is to use additional domain knowledge to "smooth" the collected predictions for each amino acid in the unknown into a coherent whole. This step uses rules to fill in gaps in α-helices and β-strands and to change isolated predictions of helices or strands to coil.

Experiments & Results

Four sets of experiments were conducted: one for the development and a preliminary validation of the method on a commonly used data set; one for further validation; the last two for the test of robustness of the method, as well as the effects of different similarity matrices.

In these experiments, we used the commonly accepted performance measures: Q_3, predictive accuracy:

$$Q_3 = \frac{q_\alpha + q_\beta + q_{coil}}{N}$$

where N is the total number of amino acids in the test data set. q_s is the number of amino acids of secondary structure type s that are predicted correctly. (where s is one of α-helix, β-strand, or coil.) Correlation coefficients (Matthews, 1975) C_α, C_β and C_{coil}, were used to measure the quality of the predictions. For secondary structure type s,

$$C_s = \frac{(p_s n_s) - (u_s o_s)}{\sqrt{(n_s + u_s)(n_s + o_s)(p_s + u_s)(p_s + o_s)}}$$

where p_s is the number of positive cases that were correctly predicted; n_s is the number of negative cases that were correctly rejected; o_s is the number of over-predicted cases (false positives), and u_s is the number of underpredicted cases (misses). A leaving-one-out statistical test (Lachenbruch & Mickey, 1968) was performed on predictive accuracy, Q_3.

Qian & Sejnowski's Data Set

One commonly used set of proteins was originally selected by Qian & Sejnowski (1988) from the Brookhaven National Laboratory using the method of (Kabsch & Sander, 1983) and has been used as a reference set for various learning experiments. This data set contains 106 proteins with a total of 21,625 amino acids, of which 25.2% are α-helix, 20.3% β-sheet and 54.5% coil. Throughout this experiment, the number, k, of similar proteins selected to be in the reference set has been set to 55, the window size, w, set to 22. A AA composition similarity metric† was used at the top level to select the reference set. The Dayhoff 250 PAM matrix was used at the bottom level to measure the similarity of segments (of length 22), with a cutoff point of 264 (where identical segments of length 22 would score 350). We increased all of the values in this matrix by eight so that there are no negative values. Table 1 shows our results along with the results from other methods that used the same data set. (see section Related Work for brief descriptions of the other methods.)

Table 1
Performance of different methods on the data set in (Qian & Sejnowski, 1988)

Method	Q_3	C_α	C_β	C_{coil}	No.runs
Leng Buchanan Nicholas	69.3%	0.46	0.50	0.39	106
Salzberg* Cost	65.1%	na	na	na	10
Qian Sejnowski	64.3%	0.41	0.31	0.41	1
Maclin Shavlik	63.4%	0.37	0.33	0.35	10

The column "No. runs" indicates the number of trials averaged into the Q_3 value. *Salzberg and Cost got Q_3 = 71.0% in a single run, and they considered that result misleading as their average over 10 trials was 65.1% (Salzberg & Cost, 1992).

To quantify how greatly the predictive accuracy varies for different unknown proteins we performed a t-test on Q_3 based on 106 runs. The 90% confidence limits for our predictive accuracy are 69.3% ± 3.1%. We determined that at a 99% confidence level (Daniel, 1987) our predictive accuracy of 69.3% is statistically better than the 65.1% predictive accuracy that was the previous best result on this data set.†† Two examples are provided in Appendix.

† That is the fractions of amino acids of each of 20 types. We also used the Dayhoff 250 PAM matrix to select reference proteins, the results were comparable with the results reported here.

†† We computed the confidence level by solving for α in the following equation: $t = z(1-\alpha/2)\sqrt{\frac{p_1(1-p_1)}{r_1} + \frac{p_2(1-p_2)}{r_2}}$ where $t = |p_1 - p_2|$, p_1, p_2 are the predictive accuracies of two methods tested on random data sets of r_1 and r_2 amino acids, respectively. α is the confidence level, z is the inverse cumulative normal distribution. For example, when $r_1 = r_2 = 20{,}000$, we can, with 99% (α = 0.99) confidence, state that one method being 1.2% better than another (t = 1.2%) is statistically significant (*cf.* Zhang *et al.*, 1992). Thus, our improvement is significant at the 99% level, given that we all used the same data set and its size is 21,625.

Table 2
Performance on other data sets

Method	Q_3	C_α	C_β	C_{coil}	No.runs	Note
Levin Garnier	87.2%	na	na	na	na	The pairwise similarity among 29 proteins ranges from 30% - 64%.
Kneller Cohen Langridge	79.0%	0.55	0.00	0.54	22	About half of 22 proteins have pairwise similarity > 40%. All are all-α.
Leng* Buchanan Nicholas	67.3%	0.42	0.51	0.39	74	The pairwise similarity among the 74 proteins is < 38%. However, only 6 pairs are above 30%, the average similarity is 12.3% ± 3%.
Zhang Mesirov Waltz	66.4%	0.47	0.387	0.429	8	The pairwise similarity among the 107 proteins is < 50%.
Nishikawa Ooi	60.0%	0.40	0.30	0.37	na	
Sweet	59.0%	0.27	0.27	na	na	

Because the results were from different data sets and under different circumstances, a fair comparison is hard to conduct, therefore, we leave them in the table with available information.

*After this initial run, with the same parameter settings as developed for the previous experiments, we tuned the system to achieve 68.4% average accuracy over 74 trials, leaving one out each time. The parameters for this run were window size 20, cutoff point 242, reference proteins 24.

Other Data Sets

Since we developed our method based on Qian & Sejnowski's data set selected in 1988 and this data set contains 12 pairs of homologous proteins (among 106 proteins), we have selected an additional set of proteins from Brookhaven data bank to further validate our method. There are three criteria we used to select this set: first, it should be different from Qian & Sejnowski's data set; second, it should preserve similar α-helix, β-sheet and coil distributions as Qian and Sejnowski's; and finally, it should keep pairwise sequence similarity (identity) low so there are no homologous pairs. As a result, 74 proteins were selected † from among recent entries to the PDB with a total of 15,012 amino acids, of which 21.3% were α-helix, 24.2% β-sheet, and 54.5% coil. We also tried to put the identical proteins in the training set, the predictive accuracy was, as expected, 99.8%. Since the degree of a match is proportional to the similarity level, and so is the performance to the homology.

The same leaving-one-out test was performed on this data set using the same Dayhoff matrix under the same parameter settings. The performance compares favorably to the results on Qian & Sejnowski's data set. The results are given in Table 2. Also shown in the table are the results of some other methods studied on different data sets. (see Related Work section for brief descriptions of the other methods.)

† The Brookhaven codes of these proteins are: *155c, 156b, 1aap, 1apd, 1apk, 1bus, 1cms, 1csei, 1dhf, 1fb4h, 1fb4l, 1fc1a, 1fc2c, 1fdx, 1fkf, 1fnr, 1fx1, 1fxb, 1fxi, 1gcn, 1gcr, 1gd1o, 1gox, 1gpdg, 1lz1, 1mcph, 1mcpl, 1omd, 1paz, 1pcy, 1pp2l, 1r08, 1r1a, 1rnt, 1sgt, 1snc, 1ton, 1ubq, 2abxa, 2apr, 2azab, 2cgaa, 2ci2i, 2cln, 2cna, 2cpp, 2hhba, 2hhbb, 2hmg, 2ldx, 2mev, 2prk, 2sni, 2ts1, 351c, 3app, 3ebx, 3est, 3gapa, 3grs, 3pep, 3rxn, 3wgab, 3xia, 4adh, 4atcb, 4cpai, 4fd1, 4ldh, 4ptp, 4tnc, 7cata, 7lyz, 9pap.*

Different Parameters

The third set of experiments was designed to determine the dependence of the performance on the window sizes and cutoff points, the results are shown in Figure 1.

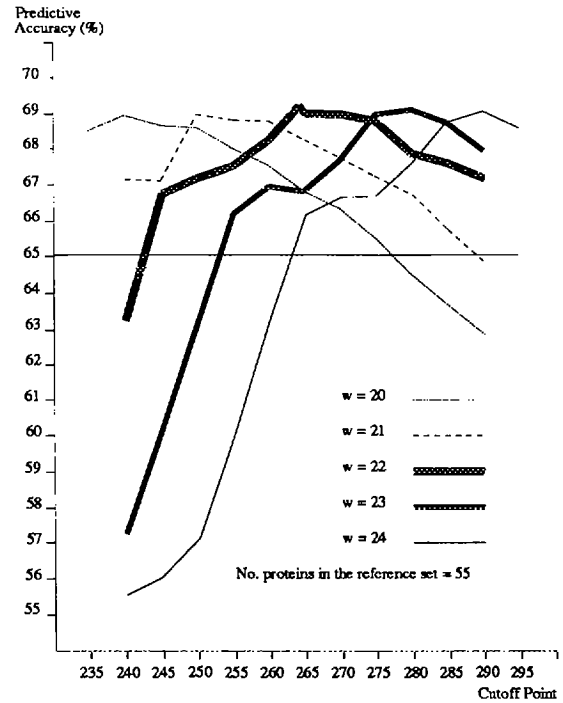

Figure 1. Dependence of predictive accuracy on window size and cutoff point

From Figure 1 we see that for a fixed window size there

is an optimum cutoff point. On either side of this optimum the predictive accuracy goes down (the analysis of the effect of window size will be given below). This is due to the fact that when the cutoff point is too large, then no segments in the reference structures will be similar, which leads to underprediction. If the cutoff point is too small, then too many spuriously segments are found, which leads to overprediction. Both cases would hurt performance. There is a broad range of window sizes and cutoff points for which the predictive accuracy is above 65%. Thus the method is relatively stable with respect to the window size and cutoff point.

We have determined empirically that a reference class of 55 proteins chosen by a global similarity metric is about optimal for including the few proteins whose segments will match any segment of the unknown well enough (over the cutoff) to make any prediction at all. On average, 3 proteins are all that contributed at the second level. If the initial similarity comparisons were more precise at level one, then, only about 3 reference proteins could be needed (on average) for prediction at this level of accuracy.

Different Similarity Matrices

As pointed out earlier, many similarity matrices can be found in the biology literature, some were based on protein statistical information, some based on chemical-physical properties, and some on structural or genetic information, etc. Sometimes, it is hard to decide which one to use. In this paper, we provide an empirical comparison of fourteen similarity matrices previously proposed, plus an identity matrix and a random one which we devised. Among these fourteen similarity matrices, six are the different versions of the Dayhoff matrix: PAM40, PAM80, PAM120, PAM200, PAM250, PAM320, (Dayhoff *et al.*, 1978; Schwartz & Dayhoff, 1978), the rest are PET91 matrix (Jones *et al.*, 1992), structure-genetic scoring matrix (McLachlin, 1972), properties scoring matrix, gonnet mutation matrix (Gonnet *et al.*, 1992), conformational similarity weight (CSW) matrix (Kolaskar & Kulkarni-Kale, 1992), EMPAR matrix (Rao, 1987), structure matrix (Risler *et al.*, 1988), and genetic matrix (Erickson & Sellers, 1983). All these matrices have been tested on Qian & Sejnowski's data set and the same experiment was conducted for each matrix. During the experiments, the number of reference proteins was tried from 1 to 55 (the result from the previous experiment indicates that at number 55, the predictive accuracy is getting stable, we stop at 55 to save some computation time), the window size was varied between 20 to 24, cutoff point varied according to the average similarity value along the main diagonal of a given matrix. For each matrix, the best result among all these parameter settings was selected as its performance. The results are given in Table 3 and discussed in the next section.

Table 3
Performance of different similarity matrices

Similarity Matrix	Q_3	C_α	C_β	C_{coil}
PAM200	69.3%	0.462	0.516	0.392
PAM40	69.2%	0.463	0.512	0.389
PET91	69.2%	0.464	0.507	0.388
PAM250	69.1%	0.463	0.523	0.396
STRUCTURE-GENETIC	68.8%	0.448	0.488	0.383
PAM320	68.7%	0.447	0.499	0.379
GENETIC	68.2%	0.459	0.477	0.379
PAM80	68.2%	0.457	0.496	0.385
STRUCTURE	68.2%	0.425	0.454	0.389
CSW	68.2%	0.461	0.458	0.386
PROPERTIES	68.0%	0.419	0.477	0.368
PAM120	67.8%	0.454	0.480	0.378
IDENTITY	67.5%	0.450	0.460	0.368
EMPAR	67.4%	0.435	0.424	0.350
GONNET MUTATION	65.4%	0.392	0.370	0.350
RANDOM	53.8%	0.230	0.137	0.119

These matrices were tested on Qian & Sejnowski's data set.

Discussions

Longer Range Interactions

It has long been realized by researchers in biology (Robson & Garnier, 1986; Qian & Sejnowski, 1988) that using local information alone, predictive accuracy cannot be improved over 70%, because long range interactions must be taken into consideration. we have taken three techniques - longer window size, overlapping window (overlapping segments) and grouping proteins into class(es) - to bring in global information. We now discuss them each in turn.

Longer window size: our method gives the best results with a window size of 22 amino acids, the largest window among the published prediction methods. The length statistics of α-helix, β-strand, β-sheet, and coil in Qian & Sejnowski's data set are given in Table 4.

Table 4
Length statistics (106 proteins)

structure	total.no	min	max	mean	sd	%<22AA's
α-helix	536	4	28	10	5	98%
β-strand	879	2	16	5	3	100%
β-sheet	103	8	333	100	68	11%
coil	1500	1	106	8	8	95%

The second column lists the total number of segments of each type. The column "% < 22 AA's" gives the percentage of the number of the segments with length less than 22 over the total number of the segments of each type.

If we look at the last three columns, it is clear that size 22 is long enough to cover most α-helices, β-strands and

coils. From the third column, we can see that the minimum lengths of helix and strand are relative short, thus, further enlarging the window size, would include more structural elements. With an average helix having a length of 10 a window of 22 is large enough in contain most of an "average" helix-turn-helix motif which constitutes an important fraction of the non-local interactions for those sequence elements. Similarly, the window size of 22 is large enough to contain two or three strands of an antiparallel β sheet connected by short turns or the strand-helix-strand motif frequently found in parallel β sheets. These are also important fractions of the non-local interactions for these motifs. Increasing the window size undoubtedly includes more amino acids that do not directly interact at all with the section being predicted. Thus we speculate that a window size of 22 represents a balance between including important non-local interactions and excluding noise from non-interacting regions of the sequence.

Overlapping window: incrementing the predictive weights for every amino acid in the window when a good match is found creates a quasi-variable window of up to 43 amino acids (21 in the N terminal direction, 21 in the C terminal direction, plus the predicted amino acid itself) which potentially influence the prediction for every amino acid in the protein sequence. Which of these amino acids have an influence and how much influence they have depends on which arrangements of the sequences and window show a high similarity score.

An added benefit of this quasi-variable window is that longer regions of high similarity have a greater influence on the prediction than do shorter regions of high similarity. We believe that this is important in the good performance of our method because our selection of reference sequence with similar compositions increases the statistical likelihood of short spurious good matches that result from similar composition rather than similar structure.

Class information: researchers have long realized (Garnier, et.al., 1978) that secondary structure prediction would be improved if a protein was known to fall into one of the four structural classes proposed by Levitt and Chothia, 1976 (α-helices and β-sheets: α, β, α/β and α+β). Knowledge of a protein's structural class is a major addition of non-local information to the predictive process. Some researchers (Cohen *et al.*, 1983; Cohen *et al.*, 1986; Taylor & Thornton, 1984) have shown that if the class type of a protein is known, then secondary structure prediction could achieve 90% accuracy.

Protein structural classes have been predicted with 75% accuracy from the amino acid composition of the protein (Nakashima *et al.*, 1986; Chou, 1989; Zhang & Chou, 1992). We do not explicitly predict the structural class of the protein. Rather, based on amino acid composition we select a reference set of known structures. This reference set will be strongly biased toward the most probable structural class but in most cases will include members of other classes as well since the classes show appreciable overlap of composition. This allows us to include the non-local information from the structural class while limiting the negative effects of misidentifying the structural class. In our experiments, the predictive accuracy made from a group of closely related proteins is higher than predictions made from a single closest protein or from all of the proteins in the data set, 69.3% vs. 64.8% and 68.2% respectively.

The Effects of Different Similarity Matrices

Much of the domain specific knowledge in our method is encoded in the matrix of amino acid similarities used to identify the best subsequences from which the secondary structure is predicted. Karlin and Altschul (1990) Altschul (1991) established, in the context of sequence alignment, that the most powerful similarity matrix is a log-odds matrix of target frequencies. Target frequencies in this case are the expected composition frequencies of amino acids after some interval of evolution. The four matrices that give the best performance are all matrices explicitly derived as log-odds matrices of such target frequencies. The performance difference among these four matrices is very nearly insignificant.

The identity matrix is a useful reference point in determining how accurately the target frequencies expressed in a specific matrix reflect the evolutionary processes of mutation and selection. The identity matrix embodies the assertion that our knowledge of the correct target frequencies is so limited and uncertain that only the presence of identical amino acids in two sequences can be taken as evidence of common evolution. In other words the identity matrix is an assertion of ignorance.

All of the matrices that performed better than the identity matrix were constructed from some experimental observations of the evolutionary processes of mutation and selection. Many were constructed from observations of target frequencies. The various PAM matrices Dayhoff et al. (1978), Schwartz & Dayhoff (1978) in many ways are the most carefully constructed target frequency matrices because of the amount of effort that went into insuring that amino acid substitutions that were counted were those between sequences with immediate common evolutionary ancestors in the data examined. The genetic matrix is based exclusively on observations of process of mutation while the structure and properties matrices focus on selection processes.

The random matrix shows that in this context the profession of ignorance embodied in the identity matrix is in fact better than random guesses about the target frequencies. The random matrix was constructed so that it does include the knowledge that any specific amino acid is

more likely to remain the same rather than change. Thus its deficiencies result mostly from the inaccuracies in the amino acid substitution frequencies rather than a gross misstatement of the overall rate of evolution.

It is noteworthy that one matrix, the EMPAR matrix (Rao, 1987), performs at just below the same level as the identity matrix and a second, the Gonnet mutation matrix (Gonnet et al., 1992), performs significantly worse than the identity matrix. The EMPAR matrix focuses not on evolutionary target frequencies but on the likelihood that amino acids appear in the same kind of secondary structure environment.

While the Gonnet mutation matrix was constructed from observed amino acid substitution data, like the PAM and PET91 Taylor (1992) matrices, there was much less editing of which substitutions to count. A specific evolutionary model is, in fact, implied by the selection of which amino acid substitutions to count in constructing the matrix. The Gonnet mutation method of selection implies an evolutionary model that can be described as a series of interconnected star burst as opposed to the binary tree model implied by the selection process used in constructing other matrices. It seems likely that this is the source of the poor performance of this matrix.

Related Work

Several other methods are related to our method: Salzberg & Cost (1992) took case-based approach, Zhang *et al.* (1992) mixed case-based, neural network, and statistical designing, however, they represented cases as segments only and used their own similarity matrices rather than the ones developed by biologists.

Levin *et al.* (1986, 1988), Sweet (1986), and Nishikawa & Ooi (1986) have taken what is called "homologous method" which is similar to the case-based approach. The main difference between theirs and ours is that we have designed a general architecture which explicitly reasons at both the object level and the internal structure level and takes of account the relationships between a protein and its segments, and relationships among segments. The experiments reported here demonstrated that the architecture is flexible in varying different domain knowledge, evidence gathering strategies, and system parameters.

Nakashima *et al.* (1986), Chou (1989), and Zhang & Chou (1992) have studied protein class prediction based on amino acid composition, however, they didn't use the class prediction to further guide the secondary structure prediction. Thus the top-level similarity of our method in used in a different purpose and results are not comparable.

Qian & Sejnowski (1988), Kneller *et al.* (1990), and Maclin & Shavlik (1992) all took neural network approach which is different from ours.

Conclusion

A general architecture was presented for protein secondary structure prediction. A series of experiments have been conducted on the architecture and the results are encouraging. Directions for future work include encoding more domain knowledge into the system and working on the interactions among the segments that are spatially far away from each other.

Acknowledgement

The ISL is supported in part by grants from the NLM (LM 05104) and the W.M.KECK FOUNDATION, the Pittsburgh Supercomputing Center is supported through grants from the NSF, (ASC - 8902826) and NIH (RR 06009).

Appendix

In this appendix, we give two examples shown in Figure 2.

Avian pancreatic polypeptide (1ppt)

Weights on helix	Weights on strand	Weights on coil	predicted	actual	sequence
0	0	0	-	-	G
0	0	0	-	-	P
0	0	0	-	-	S
0	0	0	-	-	Q
0	0	0	-	-	P
0	0	0	-	-	T
0	0	0	-	-	Y
0	0	0	-	-	P
0	0	0	-	-	G
0	0	0	-	-	D
0	0	0	-	-	D
0	0	0	-	-	A
0	0	0	-	-	P
0	0	0	-	h	V
256	0	0	h	h	E
256	0	0	h	h	D
256	0	0	h	h	L
256	0	0	h	h	I
256	0	0	h	h	R
256	0	0	h	h	F
256	0	0	h	h	Y
256	0	0	h	h	D
256	0	0	h	h	N
0	0	256	h	h	L
256	0	0	h	h	Q
256	0	0	h	h	Q
256	0	0	h	h	Y
256	0	0	h	h	L
0	0	256	-	h	N
0	0	256	-	h	V
0	0	256	-	h	V
0	0	256	-	-	T
0	0	256	-	-	R
0	0	256	-	-	H
0	0	256	-	-	R

Rat mast cell protease (3rp2)

Weights on helix	Weights on strand	Weights on coil	predicted	actual	sequence

0	0	0	-	-	I
0	0	0	-	-	I
0	0	0	-	-	G
0	0	0	-	-	G
0	0	0	-	e	V
0	0	0	-	e	E
0	0	0	-	-	S
0	0	0	-	-	I
0	0	0	-	-	P
0	0	0	-	-	H
0	0	0	-	-	S
0	0	0	-	-	R
0	0	0	-	-	P
0	0	0	-	-	Y
0	0	0	-	e	M
0	0	0	-	e	A
0	0	0	-	e	H
0	0	0	-	e	L
0	0	0	-	e	D
0	0	0	-	e	I
0	0	0	-	e	V
0	0	0	-	-	T
0	0	0	-	-	E
0	265	534	-	-	K
0	548	1081	-	-	G
0	825	1639	-	-	L
0	2224	1084	e	e	R
0	4162	0	e	e	V
0	5026	0	e	e	I
0	5891	0	e	e	C
0	6723	0	e	e	G
0	7547	0	e	e	G
0	8090	0	e	e	F
0	8913	0	e	e	L
0	9460	0	e	-	I
0	9996	0	e	-	S
0	0	10535	-	-	R
0	0	11078	-	-	Q
0	11619	0	e	e	F
0	12156	0	e	e	V
0	12425	0	e	e	L
0	12425	0	e	e	T
0	0	12689	-	-	A
0	0	12689	-	-	A
0	0	12953	-	-	H
0	0	12154	-	-	C
0	0	11324	-	-	K
0	0	10489	-	-	G
0	0	9645	-	-	R
0	0	8791	-	e	E
0	0	7927	-	e	I
0	3545	3517	e	e	T
0	6496	0	e	e	V
0	5672	0	e	e	I
0	5129	0	e	e	L
0	1599	2707	-	-	G
0	1327	2432	-	-	A
0	0	3223	-	-	H
0	0	2684	-	-	D
0	0	2141	-	-	V
0	0	1600	-	-	R
0	0	1063	-	-	K
0	0	794	-	-	A
0	0	794	-	-	E
0	0	530	-	-	S
0	0	530	-	-	T
0	0	266	-	-	Q
0	266	0	e	e	Q
0	266	0	e	e	K
0	266	0	e	e	I
0	266	0	e	e	K
0	266	0	e	e	V
0	266	0	e	e	E
0	530	0	e	e	K
0	264	0	e	e	Q
0	264	0	e	e	I
0	264	0	e	e	I
0	0	264	-	-	H
0	0	264	-	-	E
0	0	264	-	-	S
0	0	264	-	-	Y
0	0	264	-	-	N
0	0	264	-	-	S
0	0	264	-	-	A
0	0	529	-	-	P
0	0	799	-	-	N
0	0	1073	-	-	L
0	0	1349	-	-	H
0	0	2160	-	-	D
0	0	2977	-	-	I
0	3785	0	e	e	M
0	4591	0	e	e	L
0	4856	0	e	e	L
0	4856	0	e	e	K
0	4856	0	e	e	L
0	0	4592	-	-	E
0	0	4592	-	-	K
0	0	4592	-	-	K
0	0	4592	-	-	V
0	0	4592	-	-	E
0	0	4592	-	-	L
0	0	4592	-	-	T
0	0	4592	-	-	P
0	0	4592	-	-	A
0	0	4592	-	-	V
0	0	4592	-	-	N
0	0	4867	-	-	V
0	0	5145	-	-	V
0	0	5410	-	-	P
0	0	5683	-	-	L
0	0	5403	-	-	P
0	0	4586	-	-	S
0	0	3778	-	-	P
0	0	2972	-	-	S
0	0	2707	-	-	D
0	0	2707	-	-	F
0	0	2707	-	-	I
0	0	2707	-	-	H
0	0	2707	-	-	P
0	0	2707	-	-	G
0	0	2707	-	-	A
0	1362	1345	e	e	M
0	2971	0	e	e	C
0	2971	0	e	e	W
0	2971	0	e	e	A
0	2971	0	e	e	A
0	2971	0	e	e	G
0	0	2971	-	-	W
0	0	2431	-	-	G
0	0	1883	-	e	K
0	0	1344	-	e	T
0	0	795	-	e	G
0	0	264	-	-	V
0	0	264	-	-	R
0	0	264	-	e	D
0	0	264	-	e	P
0	0	264	-	e	T
0	0	264	-	-	S
267	0	264	-	-	Y
0	0	795	-	-	T
0	0	795	-	-	L
0	264	796	-	e	R
0	264	1060	-	e	E
0	264	1327	-	e	V

0	0	1594	-	e	E
0	0	1860	-	e	L
0	0	2127	-	e	R
0	0	2127	-	e	I
0	0	2127	-	e	M
0	0	2127	-	-	D
0	0	2127	-	-	E
0	0	2127	-	-	K
0	2127	0	e	-	A
0	2127	0	e	-	C
0	2127	0	e	-	V
0	2127	0	e	-	D
0	2127	0	e	-	Y
0	2127	0	e	-	G
0	2127	0	e	-	Y
0	2127	0	e	-	Y
0	1860	0	e	-	E
0	1596	0	e	-	Y
0	0	1596	-	-	K
0	0	1331	-	-	F
0	267	1067	-	e	Q
0	537	800	-	e	V
0	1082	533	e	e	C
0	1882	267	e	e	V
0	2421	0	e	-	G
0	267	2424	-	-	S
0	267	2691	-	-	P
0	0	3498	-	-	T
0	0	3776	-	-	T
0	0	4059	-	-	L
0	0	4346	-	-	R
0	0	4632	-	-	A
0	0	4918	-	-	A
0	0	5208	-	-	F
0	0	5508	-	-	M
0	0	5802	-	-	G
0	0	6090	-	-	D
0	0	6372	-	-	S
0	0	6650	-	-	G
0	0	6924	-	-	G
0	7194	0	e	e	P
0	7458	0	e	e	L
0	7191	0	e	e	L
0	6921	0	e	e	C
0	1065	5311	-	-	A
0	533	5043	-	-	G
0	4770	267	e	e	V
0	4500	267	e	e	A
0	4500	0	e	e	H
0	4225	0	e	e	G
0	3682	539	e	e	I
0	3399	810	e	e	V
0	3112	1629	e	e	S
0	2826	2462	e	e	Y
0	0	5847	-	-	G
0	0	6397	-	-	H
0	0	6948	-	-	P
0	0	7520	-	-	D
0	0	8100	-	-	A
0	0	7818	-	-	K
0	0	7540	-	-	P
0	1909	5357	-	-	P
0	6732	264	e	e	A
0	6732	0	e	e	I
0	6732	0	e	e	F
0	6732	0	e	e	T
0	6732	0	e	e	R
0	0	6732	-	h	V
0	0	6732	-	h	S
0	0	6732	-	h	T
0	0	6732	-	h	Y
6467	0	0	h	h	V
6193	0	0	h	h	P
5922	0	0	h	h	W
5103	0	0	h	h	I
4270	0	0	h	h	N
3425	0	0	h	h	A
2585	0	0	h	h	V
1734	0	0	h	h	V
868	0	0	h	-	N

Figure 2. The first three columns record the similarity weights of helix, sheet and coil respectively. The fourth column is the predicted structure, the fifth is the actual structure with "h", "e" and "-" represents helix, sheet and coil accordingly. The last column lists the unknown protein's sequence. The weights are initialized to 0, when a row with three weights being 0 indicates that there is no prediction for the amino acid in that row, the default structure is coil because we can confirm neither helix nor sheet. The performances of these two examples are Q_3 = 88.5%, C_α = 0.79, C_{coil} = 0.79 and Q_3 = 80.3%, C_α = 0.76, C_β = 0.61, C_{coil} = 0.59 respectively.

In the first example, after matching, three columns of weights are kept, one column for each secondary structure type predicted: helix, sheet, and coil. The weights are the sums of the similarity scores from matches with similar segments in the reference protein set. For each row, if a structure is the only one with weight greater than 0, then the amino acid in that row is predicted to be that structure. If no weight is greater than 0, then coil is predicted for that amino acid.

The second example shows that when there is more than one structure with weights greater than 0, the one with the highest weight will be predicted. It also shows that when a region of homogeneous secondary structure is longer than the window size, overlopping the window through the region effectively covers it.

References

Altschul, S.F. (1991).Amino acid substitution matrices from and information theoretic perspective, *J. Mol. Biol.*, **219**, 555-565.

Chou, P.Y. (1989).Prediction of protein structural classes from amino acid compositions, in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. G.D. Fasman, 549-586, Plenum Press, New York.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., and Fletterick, R.J. (1983).Secondary structure assignment for alpha/beta proteins by a combinatorial approach, *Biochemistry*, **22**, 4894-4904.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., and Fletterick, R.J. (1986).Turn prediction in proteins using a pattern-matching approach, *Biochemistry*, **25**, 266-275.

Daniel, W.W. (1987).*Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley & Sons.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978).*Atlas of Protein Sequence and Structure*, **5**, 345-352, National Biomedical Research Foundation, Washington, D.C..

Erickson, B.W. and Sellers, P.H. (1983).Recognition of Patterns in Genetic Sequences, in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. D. Sankoff and J.B. Kruskal, 55-91, Addison-Wesley.

Garnier, J., Osguthorge, D.J., and 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.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992).Exhaus-

tive Matching of the Entire Protein Sequence Database, *Science*, **256**, 1443-1445.

Hayward, S. and Collins, J.F. (1992).Limits on alpha-Helix Prediction With Neural Network Models, *PROTEIN: Structure, Function, and Genetics*, **14**, 372-381.

Hunter, L. (1992).Artificial Intelligence and Molecular Biology, *AAAI-92*, 866-868.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992).The rapid generation of mutation data matrices from protein sequences, *Computer Applications in the Biosciences*, **8**, 275-282.

Kabsch, W. and Sander, C. (1983).Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, **22**, 2577-2637.

Karlin, S. and Altschul, S.F. (1990).Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proceedings of the National Academy of Sciences*, 2264-2268.

Kneller, D.G., Cohen, F.E., and Langridge, R. (1990).Improvements in Protein Secondary Structure Prediction by An Enhanced Neural Network, *J. Mol. Biol.*, **214**, 171-182.

Kolaskar, A.S. and Kulkarni-Kale, U. (1992).Sequence Alignment Approach to Pick Up Conformationally Similar Portein Fragments, *J. Mol. Biol.*, **223**, 1053-1061.

Lachenbruch, P.A. and Mickey, M.R. (1968).Estimation of Error Rates in Discriminant Analysis, *Technometrics*, **10**, 1-11.

Levin, J.M. and Garnier, J. (1988).Improvement in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool, *Biochimica et Biophysica Acta*, **955**, 283-295.

Levin, J.M., Robson, B., and Garnier, J. (1986).An algorithm for secondary structure determination in proteins based on sequence similarity, *FEBS*, **205:2**, 303-308.

Levitt, M. and Chothia, C. (1976).Structural patterns in globular proteins, *Nature*, **261**, 552-557.

Maclin, R. and Shavlik, J.W. (1992).Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou-Fasman Algorithm for Protein Folding, *AAAI-92*, 165-170.

Matthews, B.W. (1975).Comparison on the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta*, **405**, 443-451.

McLachlin, A.D. (1972).Repeating sequences and gene duplication in proteins, *J. Mol. Biol.*, **64**, 417-437.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986).The folding type of a protein is relevant to the amino acid composition, *J. Biochem.*, **99**, 153-162.

Nishikawa, K. and Ooi, T. (1986).Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods, *Biochimica et Biophysica Acta*, **871**, 45-54.

Qian, N. and Sejnowski, T.J. (1988).Predicting the Secondary Structure of Globular Proteins Using Neural Network Models, *J. Mol. Biol.*, **202**, 865-884.

Rao, J.K.M. (1987).New Scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters, *Int. J. Peptide Protein Res.*, **29**, 276-281.

Rendell, L. and Cho, H. (1990).Empirical Learning as a Function of Concept Character, *Machine Learning*, **5**, 267-298.

Risler, J.L., Delorme, M.O., Delacroix, H., and Henaut, A. (1988).Amino Acid Substitutions in Structurally Related Proteins - A Pattern Recognition Approach, *J. Mol. Biol.*, **204**, 1019-1029.

Robson, B. and Garnier, J. (1986).*Introduction to Proteins and Protein Engineering*, Elsevier, Amsterdam.

Rooman, M.J. and Wodak, S.J. (1988).Identification of predictive sequence motifs limited by protein struture data base size, *Nature*, **335**, 45-49.

Salzberg, S. and Cost, S. (1992).Predicting Protein Secondary Structure with a Nearest-neighbor Algorithm, *J. Mol. Biol.*, **227**, 371-374.

Schwartz, R.M. and Dayhoff, M.O. (1978).*Atlas of Protein Sequence and Structure*, **5**, 353-358, National Biomedical Research Foundation, Washington, D.C..

Sweet, R.M. (1986).Evolutionary Similarity Among Peptide Segments Is a Basis for Prediction of Protein Folding, *Biopolymers*, **25**, 1566-1577.

Taylor, W.R. and Thornton, J.M. (1984).Recognition of super-secondary structure in proteins, *J. Mol. Biol.*, **173**, 487-514.

Zhang, C.T. and Chou, K.C. (1992).An Optimization Approach to Predicting Protein Structural Class from Amino Acid Composition, *Protein Science*, **1**, 401-408.

Zhang, X., Mesirov, J.P., and Waltz, D.L. (1992).Hybrid System for Protein Secondary Structure Prediction, *J. Mol. Biol.*, **225**, 1049-1063.