Prediction of Protein Function by Discriminant Analysis*

PETR KLEIN
Division of Biological Sciences, National Research Council, Ottawa, Ontario, Canada K1A 0R6

JOHN A. JACQUEZ
Department of Physiology, The University of Michigan, Ann Arbor, Michigan 48109

AND

CHARLES DELISI†
Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205

Received 5 May 1986; revised 15 May 1986

ABSTRACT

Approximately 53% of the protein sequences in the National Biomedical Research Foundation (NBRF) database can be allocated to one of 26 functional classes, each of which can be characterized by the joint occurrence of four or fewer attributes. The attributes reflect collective physicochemical properties of the sequences in a class, ranging from simple characteristics of composition, such as average hydrophobicity and net charge, to amphipathicity and the propensities of various residues to be in certain preferred configurations. In some, though not all instances, these variables can be related in a general way to topological or other structural features of the particular class they characterize. We show that the attributes permit 17 of the 26 groups to be filtered from all other proteins in the database with a misclassification error of less than 2%, and that the remaining 9 groups can be filtered with errors not exceeding 13%. Thus for a given functional class, the results point to the existence of relatively few characteristic variables which capture most of the intraclass similarity and interclass variability that is common and peculiar to members of that class.

1. INTRODUCTION

Because most nucleic acids can be translated—either by direct application of the genetic code, or when necessary with the aid of auxiliary

*NRCC No. 25894.

†Present address: Office of Health and Environmental Research, Department of Energy, Washington, DC 20545.

information on exon-intron boundaries—the recent rapid increase in the number of gene sequences is spawning a similar increase in the products of those sequences. The collection of these deduced protein sequences potentially contains an enormous and rapidly growing body of information on function. It also contains, perhaps more directly, a large amount of information on higher order structure and cellular location, both of which are more tightly coupled to function than sequence alone. Extracting this information is, however, not fully realizable at present, since the principles governing the translation of sequence to higher order structure, function or location, are not adequately developed.

In earlier papers we reported the development and application of statistical methods for positional classification of membrane proteins [8], for structure prediction of beta rich molecules [4] and folding type [9], and for functional classification of a sequence within a limited number (six) of functional categories [7]. In this paper we report a major improvement in the accuracy and number of categories in the functional classification problem. In particular, with the NBRF database divided into 26 functional classes, we have been able to find, for each class, three or four characteristic parameters that capture the properties of that class. These parameters enable allocation of a sequence to one of the classes with a minimum reliability of nearly 90%, and in many cases with a reliability close to 100%.

2. METHODS

2.1. *THE DATABASE*

The February 1983 version of the NBRF protein sequence database [3, 13] consists of 2145 sequences divided into 663 superfamilies. After exclusion of miscellaneous and hypothetical proteins, and of sequences of less than 12 residues, the remaining 1603 sequences were segmented into 27 *groups*: the 26 listed in Table 1, plus a 27th containing all the remaining proteins.

2.2. *DISCRIMINANT ANALYSIS*

We use the generic term "attribute" to denote any compositional or physicochemical property of an amino acid sequence. A large number of attributes, some of which are listed in the Appendix, can be used to characterize protein sequences. The observation that certain attributes will be more pronounced in some protein families than in others serves as the basis for interfamily discrimination. Before presenting the method in full generality, a specific example will help fix the main idea. Consider the magnitude of 3.6 residue per turn amphipathicity, i.e. the extent to which the hydrophobicity of residues along the sequence increases and then decreases with an average period of 100 degrees. A precise numerical value can be

attached to this property for every sequence in every group of proteins in the database. For each group, this property will have some characteristic *distribution of values*. Thus, by analyzing the database we can construct $P(x|G_i)$, the probability that a sequence picked at random from the ith group has amphipathicity value x. We refer to $P(x|G_i)$ as the probability distribution of property x in group G_i. Specifically, in the globins the interval in amphipathicity from 3.59 to 4.09 covers one standard deviation on either side of the mean, whereas for the other proteins the corresponding interval is from 2.12 to 3.32. As these numbers suggest, relatively few proteins from other groups have amphipathicities as high as the mean value for globins; hence the odds are good that a sequence having a high value of this property will be a globin. This generally high value of amphipathicity reflects the overall structural architecture of the globin family, which consists of eight alpha helical segments in a globular arrangement. The 3.6 residue per turn amphipathicity tells us that the alpha segments have one face relatively polar and the opposite face relatively apolar, the former presumably facing outward in a direction for favorable interactions with water, and the latter facing inward, protected against unfavorable interactions with water.

The relatively few errors that are made in allocating highly amphipathic structures to the globin category are often the result of classifying hormones as globins, since they too sometimes have high values of amphipathicity. Evidently, if one wanted to do better, a second variable could be introduced which is high in globins and lower in hormones. The frequency of histidine is such a variable. Thus we would have a two variable distribution function, with amphipathicity as one variable and frequency of histidine as another (x would now be a two component vector).

More generally, suppose the objective is to distinguish some group of proteins (globins, cytochromes, toxins, and so forth) from all the rest. We then have two groups: G_1, the group to be identified; and G_2, all the remaining proteins. These are to be distinguished from one another on the basis of N attributes, $x = (x_1, \ldots, x_N)$, the values allowable to each such vector being distributed according to some multivariate function. Given a sequence with attribute vector x_0, we would like to determine the group to which it is most likely to belong. This is done by evaluating the posterior probability distributions $P(G_i|x)$ for $x = x_0$ (i.e. the probability that a protein with attribute vector x_0 belongs to group i) for each group, and choosing the larger of the two. Such a procedure is known to minimize the probability of incorrect allocation [10]. Thus if $P(G_1|x_0) > P(G_2|x_0)$, the protein is allocated to G_1. To calculate the *posterior* probabilities, the Bayes formula is used [1]:

$$P(G_i|x) = \frac{P(x|G_i)\,P(G_i)}{P(x)}, \qquad i = 1,2, \tag{1}$$

TABLE 1

Groups, Attributes, Errors of Discrimination, and Means and Standard Deviations for the Best Attributes in Groups[a]

No.[b]	Name of group, G_1	Size n_1	Attributes[c]	No. of misclassifications	Probab. of correct classification, p_0	Mean (SD) in G_1	Mean (SD) in G_2
1	Collagen	12	$A_{gly}(8,3)$	0	1.00	2.41 (0.22)	0.51 (0.19)
2	α-Crystallin	7	phe, H, log L	0	1.00	0.08 (0.003)	0.04 (0.02)
3	Nitrogenase	5	R_{gly}, gly, log L	0	1.00	5.20 (1.10)	0.87 (1.23)
4	Cytochrome b	5	H, max R_H, β	0	1.00	1.04 (0.02)	0.63 (0.13)
5	Cytochrome c_3	6	$R_{\text{his}L}$, his	0	1.00	0.09 (0.0004)	0.0007 (0.002)
6	Cytochrome c′	16	ala, H, log L	1	1.00	0.21 (0.04)	0.08 (0.04)
7	Carbonic anhydrase	7	FHFHW peptide	0	1.00		
8	Mammal ribonuclease	20	$R_{\text{met}L}$, met, H	0	1.00	0.008 (0.0001)	0.004 (0.002)
9	Nonhistone chromosomal	22	max $H(12)$, H	1	1.00	0.30 (0.09)	1.15 (0.24)
10	ATPase	11	β_t, $A_C(3,3.6)$	1	1.00	0.80 (0.01)	0.93 (0.07)
11	Pancreatic hormones	6	tyr, pro	2	1.00	0.12 (0.01)	0.03 (0.02)
12	Keratins	9	CCC peptide	2	1.00		
13	Anthranilate synthase	5	PGPG peptide	2	0.99		
14	Phospholipase	24	cys, H, log L	7	0.99	0.11 (0.01)	0.04 (0.05)
15	Other hormones	21	R_{cys}, cys, log L	4	0.99	1.90 (0.44)	0.17 (0.67)

16	Actin, tubulin	9	$R_H(3), H, \log L$	7	0.98	29.67 (2.83)	9.12 (8.49)
17	Prolactin, somatotropin	12	$R_{\text{leu}\,L}$, leu, $\log L$	8	0.98	0.03 (0.01)	0.005 (0.01)
18	Dehydrogenase	16	R_{val}, val, $\log L$	11	0.97	4.62 (1.58)	0.59 (0.99)
19	Ferredoxin	48	C, asp, H, $\log L$	17	0.96	−13.33 (3.94)	1.00 (8.89)
20	Globins	169	$A_H(12,3.6)$, α, ala, his	21	0.95	3.84 (0.25)	2.72 (0.60)
21	Other contractile	45	α, β, C_L, H	22	0.93	1.06 (0.04)	0.95 (0.06)
22	Cytochrome c	54	R_{lys}, $R_{\text{lys}\,L}$, lys	11	0.93	2.78 (0.96)	0.71 (1.36)
23	Toxins	101	$R_{\text{cys}\,L}$, α, $\log L$	34	0.91	0.02 (0.01)	0.001 (0.01)
24	Ig variable regions	166	ER_H, $A_C(3,3.6)$, var $H(4)$, ser	29	0.89	3.48 (0.42)	2.89 (0.50)
25	Histones	46	max $C(12)$, H, lys	26	0.87	6.54 (1.50)	3.13 (1.40)
26	IgG constant regions	12	PKPK or QTQT peptide	1	0.87		

[a] If more than one attribute is listed, means and standard deviations are those of the best one (listed first).

[b] Groups 1–26 contain the following superfamilies from the NBRF database [3], respectively: 332; 333; 46; 5; 3 without cytochrome c7; 4; 146; 98; 263; 266–268; 130–131; 211; 328–329; 145; 88; 195–196, 201; 337, 340; 203; 21–24; 11; 256; 335–336, 338–339, 341–344; 1 (first 54 proteins); 231–232, 238, 241–243, 249; 253; 258–262, 264–265; part of 254.

[c] General definitions for the attributes are given in the Appendix. Examples are: $A_{\text{gly}}(8,3)$ = periodicity in the appearance of glycine with period 3 in segments of 8 residues; phe = frequency of phenylalanine; R_{gly} = number of consecutive occurrences of at least two glycines; FHFHW peptide = occurrence of pentapeptide phe-his-phe-his-trp; max $H(12)$ = maximum hydrophobicity in segments of 12 residues; α = average propensity to form α-helix.

where $P(x) = P(x|G_1)P(G_1) + P(x|G_2)P(G_2)$. Here $P(x|G_i)$, $i = 1,2$, are determined from empirical distributions of attributes in the two groups, as indicated above, and $P(G_i)$, $i = 1,2$, are *prior* probabilities that an unknown protein belongs to G_i (i.e. the probabilities before the values of attributes are determined). These latter probabilities can be estimated from group sizes. As the denominator in (1) is the same for both groups, the rule—called the Bayes discriminant rule [10]—is to allocate a protein to the group for which $P(x|G)P(G)$ has the larger value.

The distributions obtained from an analysis of the database are of course discrete, and the vector of values x_0 of a protein to be classified will very likely not match the vectors for which the probability distributions have known values unless the data are very dense. Two alternatives for circumventing this problem exist: one is to estimate the values of the probability distribution corresponding to x_0 by interpolation; the other is to assume the distribution is of some analytic form, usually Gaussian, and to use standard techniques to find parameters characterizing it (the mean vector and covariance matrix in the case of a multivariate Gaussian). In this paper we use the latter parametric procedure and refer to the derived density functions as $f(x|G_i)$.

Generally we are not told that a protein belongs either to one group or to all the rest: we are given a sequence and a number of possible categories and we want to know the most likely category. The greater the number of categories, the better the resolution of the classification. There are at least two ways to proceed. The most direct way is to evaluate the density function $f(x_0|G_i)$ for each category. A second method begins by numbering the categories (1–26 in this case) and then performing seriatim two-category discrimination: category 1 against all the rest; category 2 against all the rest, and so on. The former procedure was used previously [7] for distinguishing six categories. Here we use the latter, primarily because additional discrimination procedures can be included and changes can be made in the procedure for identifying a particular group without affecting the procedure for identifying the remaining groups.

The disadvantage of the second approach is that it can, if not properly implemented, lead to substantial error accumulation. Each time two groups are compared with an allocation to one group or the other, an error is made; our objective is to minimize the overall cumulative error after all groups have been compared. To this end we accepted certain constraints on individual discriminations, and only those discriminations satisfying these constraints have been included in the series.

Let m_{ij} be the number of proteins in G_i classified as belonging to G_j, where i and j can be either 1 or 2 independently of one another; let n_i be the size of group G_i; and let p_{ij} be m_{ij}/n_i. We require that (i) $n_1 \geq 5$, (ii) $m_{12} + m_{21} < n_1$, and (iii) $p_{11} > 0.5$. Requirement (ii) means that our proce-

dure is better than a simple rule "allocate everything to the bigger group G_2," which makes exactly n_1 mistakes; (iii) means that we detect at least 50% of the proteins in the smaller group G_1.

Candidates for groups were selected by examining one dimensional histograms of attribute distributions in groups of superfamilies from the NBRF database. Superfamilies (or their groups) with attribute distributions well separated from those of other proteins were then screened by discriminant analysis using one or more attributes for each candidate group. If the above constraints were satisfied, the group was confirmed as separable from other proteins and included in the series of discriminations.

The program we have developed uses nonlinear discriminant analysis (i.e. no assumptions about equality of variances). It reads values of attributes for all proteins in different groups from a file, estimates parameters (means and covariances) of multivariate normal probability density functions, and estimates error probabilities by allocating each protein and counting the numbers of those correctly and incorrectly allocated. The program used is DISCRDV. It was written in FORTRAN and run on a VAX 11/780. This program, as well as the program ALOF for allocation of additional proteins to one of the functional groups (see below), can be obtained by writing to P. Klein.

3. RESULTS

We found that three or four attributes were generally sufficient to distinguish each of 26 functional categories from the remainder of the database. The attributes ranged from simple characteristics of composition, such as average hydrophobicity and net charge, to attributes describing structural features of the sequence (see Appendix). If several such attributes were found for one group, we used discriminant analysis to identify the set of attributes giving the best discrimination for that group of proteins. Table 1 contains a collection of 22 clusters of proteins which can be distinguished by discriminant analyses satisfying (i)–(iii), and another 4 groups which can be distinguished on the basis of signature peptides (i.e., peptides of three to five amino acids appearing, ideally, only in the given group). In each of these analyses, the posterior probability of correct allocation

$$p = P(G_1)\,p_{11} + P(G_2)\,p_{22} \tag{2}$$

was at least 0.98. If we estimate the prior probabilities from group sizes, $P(G_i) = n_i/n$, $i = 1,2$, then

$$p = \frac{m_{11} + m_{22}}{n}, \tag{3}$$

where n is the total number of proteins in the database we are using.

We ordered these 26 groups so that the error increased with the number of procedures used, and allocated all proteins in the database. We found that only 29 proteins from the whole database were not allocated uniquely: 28 were allocated to two groups, and 1 (a cytochrome c3) was allocated to three groups (cytochromes c3, cytochromes c, and histones). The simplest rule, and the one minimizing error in these 29 cases, is to place the protein in the first group to which it is allocated (Table 1). The last column of Table 1 shows how the overall probability of correct allocation, p_0, gradually decreases as more procedures are included. The first 15 groups, for which the overall correct classification probability p_0 is 0.99, include 11% of the database; 28% of the database is covered at $p_0 = 0.95$, 40% at $p_0 = 0.91$, and all 26 groups include 53% of the database with $p_0 = 0.87$. If we exclude from the remaining 47% those superfamilies with less than five proteins (which would be too small for any reliable discrimination), this series of 27 discriminations covers 68% of the database.

A test of significance of the success rate in predictions is provided by comparing $p = 0.87$ with random assignment. Suppose we have a set of proteins in m groups distributed as follows: fraction f_1 in group 1, fraction f_2 in group 2, and so on. Suppose these proteins are assigned to groups only on the basis of prior probabilities $p_1, \ldots, p_m$ (i.e. without attributes). The expected fraction of correctly assigned proteins will be $p_1 f_1 + \cdots + p_m f_m$. If the distribution of proteins in groups were the same as the prior probabilities, the fraction correctly assigned would be $p_1^2 + \cdots + p_m^2$. For the 26 groups, for which $p_0 = 0.87$, this formula gives a value of 0.25 for the fraction that would be correctly assigned by random allocation.

When the discrimination error is estimated from the data used to derive the discrimination rule ("training data"), the actual error is usually somewhat underestimated (especially for small sized groups) [10]. It is therefore customary to check the reliability of the estimates either by the jackknife method or by using another set of data. We allocated 108 sequences added to the database after February 1983. Of these, 93 (86%) were correctly allocated; this is very close to the expected $p = 0.87$. In all, 34 were allocated to one of 26 functional groups (24 correctly) and 74 were not (69 correctly).

When classifying a test set or any other set of proteins different from the training set, we should bear in mind that for best discrimination the prior probabilities should reflect the distribution in groups of this set of proteins rather than of the training set. This is difficult to do in practice, although we can expect that prior probabilities will change as the database grows. Table 2 gives the distribution in groups in the test set and corresponding priors from the training set. To see how this might influence reliability, we performed another classification of the test set, this time using priors corresponding to that set. There was only one change: one of the misclassified proteins would now be classified correctly (increase in reliability from 0.86 to 0.87).

TABLE 2

Distribution of Proteins in the Test Set and Corresponding Prior Probabilities from the Training Set

Group No.	Probability in test set	Prior probability
10	0.009	0.007
13	0.009	0.003
14	0.019	0.015
16	0.009	0.006
19	0.074	0.030
20	0.046	0.105
21	0.056	0.028
22	0.009	0.034
24	0.019	0.104
26	0.019	0.007
27	0.731	0.459

Certain proteins, those classified as miscellaneous by Dayhoff et al. [3], were not used in the training set. Some are obviously related to one of the 26 functional categories, whereas others properly belong to group 27. An application of discriminant analysis to these miscellaneous proteins provides another estimate on the reliability of the method. Those belonging to the functional categories were all correctly classified as such (hemerythrin, complement anaphylatoxin). Of the remaining proteins, 18% were misclassified as belonging to one of the 26 functional groups, so the error here too is approximately the same as estimated above.

Table 3 is the overall performance matrix Q for the series of procedures. Group 27 contains all proteins which do not belong to groups 1–26. The element q_{ij} is the number of proteins from group i allocated to group j. The rows of this matrix can be used to estimate all probabilities of correct and incorrect allocation in individual discriminations. The overall probability of correct allocation, p_0, is the sum of the diagonal elements divided by the sum of all elements. Columns of this matrix can be used to estimate empirically the probabilities that a protein allocated to a certain group really belongs to that group or other groups. For instance (column 24), of the 155 proteins that were allocated to immunoglobulin (Ig) variable regions, 146 really were Ig variable regions, 2 were IgG constant regions, and 7 were from group 27. So if an unknown protein is allocated to Ig variable regions, we estimate the probability that it is classified correctly as $\frac{146}{155}$, and the probability it is classified incorrectly as $\frac{9}{155}$: a $\frac{2}{155}$ chance that it belongs to an Ig constant region, and a $\frac{7}{155}$ chance that it belongs to group 27. The probability that it belongs to any one of the other 24 groups is estimated at 0.

TABLE 3

Overall Performance Matrix Q^a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	12																										
2		7																									
3			5																								
4				5																							
5					6																						
6						15														1							
7							7																				
8								20																			
9									22																		
10										11																	
11											6																
12												7			1												1
13													4														1
14														22													2
15															18								1				2
16																8											1
17																	7										5
18																1		13									2
19																			39								9
20																				156							13
21																					28						17
22																						47					7
23																							100				1
24																								146			20
25																						1	1		24	1	19
26																		1						2		9	
27									1	1	2		1	5	1	5	3	8	8	8	5	3	32	7	6	1	650

[a] q_{ij} is the number of elements in group i allocated to group j by the method.

4. DISCUSSION

This paper describes a method which uses a series of discriminant analyses to allocate a protein sequence of unknown function to one of 26 functional categories. Allocation is based on attributes (characteristics) of the sequence, which are compared with probability distributions of the same attributes in the functional groups. For some groups the attributes clearly reflect the known structure (periodicity in glycine for collagen, hydrophobic periodicity in globins) or function (positive charge in some DNA binding proteins), and the probabilities of correct allocation are good measures of the extent to which these attributes characterize the proteins. For other

groups, the correct choice of discriminant variable is not intuitively obvious (low propensity to form beta turns in ATPases, high variance in hydrophobicity in Ig variable regions), and those that were found are generally difficult to interpret in terms of what is known about the function of the group.

Related work has been reported by Nishikawa et al. [11, 12], who used amino acid composition to distinguish extra- and intracellular enzymes and nonenzymes. Classification rules in 18 dimensional space were derived empirically and allowed 66% correct allocation into one of the four groups.

Although our method can predict, with a high degree of success, broad functional categories to which a protein of unknown function might belong, such prediction is not, at present, the primary significance of these results. What is of greater potential importance, we believe, is that the surprising success in classification using just a few attributes points to the existence of consensus properties that, collectively, are common and peculiar to a functional class and hence characterize the class with high reliability. The performance matrix (Table 3) indicates that in spite of the enormous sequence complexity characterizing a functional family, three or four variables can capture most of the information about the class that discriminates it from the other classes. As a practical matter, one might imagine that, with more insight into correlations between structure and function, such variables could be potentially significant as design parameters. Evidently what we report here is still in an early phase of development, and can profit by greater resolution (an increased number of functional groups) and greater precision (fewer misclassification errors). It nevertheless indicates that characteristic variables, probably capturing topological (rather than geometrical) properties [4] of functional families, exist that enable us to allocate sequences reliably to those families.

APPENDIX. DEFINITION OF ATTRIBUTES

(1) *Average hydrophobicity H.* Let H_i be the hydrophobicity value (see [6] for a review) of the ith residue in the sequence of length L. Then

$$H = \frac{1}{L} \sum_{i=1}^{L} H_i.$$

(2) *Maximum hydrophobicity, variance in hydrophobicity.* We define

$$\max H(l) = \max_{i=1,\ldots,L-l+1} H(i,l),$$

where $H(i,l)$ is the average hydrophobicity of a segment of length l starting

at position i, and

$$\operatorname{var} H(l) = \frac{1}{L-l+1} \sum_{i=1}^{L-l+1} [H(i,l) - H_{\text{av}}]^2,$$

where

$$H_{\text{av}} = \frac{1}{L-l+1} \sum_{i=1}^{L-l+1} H(i,l).$$

(3) *Frequencies of occurrence of the 20 amino acid residues* are denoted by their three letter codes.

(4) Let C denote the *net charge*,

$$C = (\text{arg} + \text{lys} - \text{asp} - \text{glu})\, L,$$

$\max C(l)$ the *maximum charge* in all segments of length l (defined analogously to maximum hydrophobicity), and C_L the *frequency of occurrence of charged residues*,

$$C_L = \text{arg} + \text{lys} + \text{asp} + \text{glu}.$$

(5) Define the *hydrophobic periodicity with period* λ in a segment of length l begining with the kth residue as

$$A(l,\lambda,k) = \left[\left(\sum_{i=k}^{k+l-1} [H_i - H(k,l)] \cos \frac{2\pi i}{\lambda} \right)^2 + \left(\sum_{i=k}^{k+l-1} [H_i - H(k,l)] \sin \frac{2\pi i}{\lambda} \right)^2 \right]^{1/2}.$$

Then $A_H(l,\lambda)$, an average taken over all $L-l+1$ such segments, is a measure of hydrophobic periodicity.

(6) If we put $H_i = 1$ for all charged residues and 0 for other residues in the formula for hydrophobic periodicity, we can define $A_C(l,\lambda)$—a *periodicity in charge*.

(7) Similarly, if $H_i = 1$ for a specific amino acid residue R and 0 otherwise, then $A_R(l,\lambda)$ is a measure of the *periodicity of appearance* of this residue.

(8) We can look for consecutive occurrences (runs) of the same (type of) residues. Denote by $\max R_H$ the length of the *longest run* of hydrophobic residues in the sequence; by $R_H(l)$ the *number of* hydrophobic *runs* at least l residues long (if $l = 2$ the argument will be omitted); by $R_{HL}(l)$ the ratio

$R_H(l)/L$; and by $ER_H(l)$ the *mean length of* a hydrophobic *run*. The number of runs of a specific residue R will be denoted analogously.

(9) In an analogy to the definition of average hydrophobicity (1), we can define the *average propensity to form an α-helix, β-sheet, or β-turn* if instead of H_i we use the values derived by Chou and Fasman [2] for each amino acid and take their geometric means (see [5]). We shall denote these quantities as α, β, and β_t.

(10) We also examined the database for the appearance of short characteristic patterns (signature peptides) that would, ideally, fully distinguish some groups of proteins. In this case, we do not need discriminant analysis for discrimination; an unknown protein is allocated to a particular group only if it has the signature peptide. Signature peptides are denoted by a one letter code for amino acids.

REFERENCES

1 T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York, 1984.

2 P. Y. Chou and G. D. Fasman, Empirical predictions of protein conformation, *Ann. Rev. Biochem.* 47:251–276 (1978).

3 M. O. Dayhoff, L. T. Hunt, W. C. Barker, B. C. Orcutt, L. S. Yeh, H. R. Chen, D. G. George, M. C. Blomquist, and G. C. Johnson, *Protein Sequence Database*, Nat. Biomed. Res. Foundation, Washington, D.C., 1983.

4 C. DeLisi, P. Klein, and M. Kanehisa, Some comments on protein taxonomy: Procedures for functional and structural classification, in *Molecular Basis of Cancer, Part A* (R. Rein, Ed.), Liss, New York, 1985, pp. 431–441.

5 M. J. Dufton and R. C. Hider, Snake toxin secondary structure predictions. Structure activity relationship, *J. Mol. Biol.* 115:177–193 (1977).

6 D. Eisenberg, Three-dimensional structure of membrane and surface proteins, *Ann. Rev. Biochem.* 53:595–623 (1984).

7 P. Klein, M. Kanehisa, and C. DeLisi, Prediction of protein function from sequence properties. Discriminant analysis of a database, *Biochim. Biophys. Acta* 787:221–226 (1984).

8 P. Klein, M. Kanehisa, and C. DeLisi, The detection and classification of membrane spanning proteins, *Biochim. Biophys. Acta* 815:468–476 (1985).

9 P. Klein and C. DeLisi, Prediction of protein structural class from the amino acid sequence, *Biopolymers*, to appear.

10 K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic, London, 1979.

11 K. Nishikawa, Y. Kubota, and T. Ooi, Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution, *J. Biochem.* 94:981–995 (1983).

12 K. Nishikawa, Y. Kubota, and T. Ooi, Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types, *J. Biochem.* 94:997–1007 (1983).

13 B. C. Orcutt, D. G. George, and M. O. Dayhoff, Protein and nucleic acid sequence database systems, *Ann. Rev. Biophys. Bioengrg.* 12:419–441 (1983).