J. theor. Biol. (1986) **120**, 223-236

New Statistical Approach to Discriminate Between Protein Coding and Non-coding Regions in DNA Sequences and its Evaluation†

CHRISTIAN J. MICHEL‡

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, 11 rue Humann, 67085 Strasbourg, Cédex, France

(*Received* 5 *October* 1985, *and in revised form* 16 *December* 1985)

We propose a new approach to study protein coding and non-coding regions in DNA sequences, by making use of two complementary statistical methods. The principal component analysis (PCA) is a graphical method to represent DNA sequences which are characterized by some quantitative parameters: it is a help to the intuition. The discriminating analysis (DA) is a quantitative method which permits to classify the DNA sequences. It leads to an evaluation of the first method and to a decision. The value of this approach has been confirmed since we also have found some results which had been described recently in the literature. Furthermore, this general methodology has permitted us to show the existence of parameters which identify the nucleic acid sequence functional domains, without having to make use of the properties of the genetic code.

1. Introduction

Several methods have been developed in order to distinguish between protein coding (PCS) and non-coding (NCS) sequences. They can be classified into two categories:

(i) Type 1—methods. These permit one to locate the PCS and NCS exactly. They make use of similarity features which lead to the comparison of small number of sequences, such as: initiation and termination signals, ribosome binding sites, intron-exon junctions, homologies and symmetries of sequences, etc.

(ii) Type 2—methods. These lead one to show universal differences between the PCS and the NCS. They make use of the probabilities and of the statistics in order to treat large samples of sequences which are characterized by quantitative parameters, such as the percentage of bases. Of the type-2 methods, several probabilistic and statistical ones have been published, which permit processing of the information stored in each DNA sequence, with the four-base series. Staden & McLachlan (1982) compared, by testing the similarity of the codon usage strategy (Grantham *et al.*, 1981) a known PCS and the open reading frame. Shepherd (1981) determined, by correlation testing, which frame differs the least from a supposed original PCS, where the codons should have the form *RNY* (R = purine, Y = pyrimidine, N = purine or pyrimidine). Smith *et al.*, (1983) characterized the coding and non-coding

† The source code for the programs described in this paper is available free of charge upon request.

‡ Present address: Department of Microbiology, Biozentrum der Universität Basel, Klingelbergstrasse 70 CH-4056 Basel, Switzerland.

0022-5193/86/100223+14 $03.00/0

domains of vertebrate and non-vertebrate sequences, with measures of strand-pairing asymmetry, diad and triad nearest neighbour chi-square values, and cytosine-guanine suppression percentages. Tramontano *et al.* (1984) used two independent methods in order to evaluate the protein coding information content in different classes of complementary DNA strands: one is looking at the reading frame length distribution, based on the search for specific initiation and termination codons; the other is the testcode analysis developed by Fickett (1982).

Fickett's method (1982) assigns the probability of coding to a given sequence, taking into account the overall properties of the base-sequence itself. His methodology has a lot of features in common with ours. Both are general and do not depend on the judgement of the users. With large samples of sequences, our methods permit us to determine a combination of quantitative parameters—without having to lay down any hypothesis—leading to the best discrimination between the PCS and the NCS. Secondly, this information is tested with other samples in order to evaluate the reliability rate of the analysis. Eventually, these results allow the prediction of new coding and non-coding regions in published sequences.

2. Statistical Methodology

The use of statistics brings out the problem of samples and methods. We discuss these points of our approach.

(I) The sample of different taxonomic DNA sequences—obtained from the EMBL Nucleotide Sequence Data Library—constitutes the data of the two statistical methods. A representative sample is obtained with an identical number of sequences identified as PCS and NCS, NCS were all introns of eukaryotic genes. We carry out this choice *a priori*, because we suppose that a new sequence—i.e. a sequence with an unknown function—has the same chance of being a PCS or a NCS. There is an important restriction though: a representative sample from a data library does not necessarily reflect the features of the population. In order to minimize the errors which result from the choice of a single sample, the two statistical methods, for each analysis, have been tested with several representative samples.

(II) The principal component analysis (PCA) (Lebart *et al.*, 1979) is a graphical method which permits us to represent synthetically a data array without any statistical hypothesis.

FORMULATION OF THE PROBLEM

Given the number of individuals n and the number of variables p, given the individuals $W_1, \ldots, W_n$ and the variables $V_1, \ldots, V_p$, the data array is the matrix X with n lines and p columns

$$X = [x_{i,j}], \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, p$$

where $x_{i,j}$ is the value of the variable j with the individual i. Thus, to the individual

W_i is associated the ith line from X defined by the vector

$$x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}$$

Each line i is considered as an element of $\mathbb{R}^p$. $\mathbb{R}^p$ is called individual-space. The set of lines forms the individual-cloud. In the same way, to the variable V_j is associated the jth column from X defined by the vector

$$x_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$$

Each column j is considered as an element of $\mathbb{R}^n$. $\mathbb{R}^n$ is called variable-space. The set of columns forms the variable-cloud. In our study, the variables are quantitative parameters characterizing individuals which can be PCS or/and NCS. In order to get an invariable analysis for the measure unit, each element $x_{i,j}$ is divided by the standard deviation of its column. The matrix X leads then to the matrix Y

$$Y = [y_{i,j}], \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, p$$

with

$$y_{i,j} = \frac{x_{i,j}}{s_j}$$

and

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2$$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}.$$

The PCA on a space (individual-space or variable-space) consists of representing its cloud on an affined sub-space according to an orthogonal projection. This problem is equivalent to determining the straight line D which goes through the origin and which adjusts at best the cloud according to a least square approximation. We shall restrict ourselves to a demonstration pertaining to the individual-analysis.

INDIVIDUAL ANALYSIS

$\mathbb{R}^p$ is provided with the usual Euclidean metric

(i) Given two individuals W_i, $W_{i'}$ among the n individuals having $(y_{i,1}, \ldots, y_{i,p})$ and $(y_{i',1}, \ldots, y_{i',p})$ coordinates respectively, the distance between W_i and $W_{i'}$ is

$$d^2(W_i, W_{i'}) = W_i W_{i'}^2 = \sum_{j=1}^{n} (y_{i,j} - y_{i',j})^2.$$

(ii) Given their orthogonal projections H_i, $H_{i'}$ on a straight line D.

(iii) Given G the barycenter of the n individuals which has for coordinates $(g_1, \ldots, g_p)$ with $g_j = \bar{y}_j = 1/n \sum_{i=1}^{n} y_{i,j}$.

With the Euclidean metric we have the following relation

$$H_i H_{i'}^2 \leq W_i W_{i'}^2 \quad \forall i, \forall i'.$$

And with the n individuals

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} H_i H_{i'}^2 \leq \sum_{i=1}^{n} \sum_{i'=1}^{n} W_i W_{i'}^2.$$

The objective is to determine the best oriented straight line D which maximizes

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} H_i H_{i'}^2 \tag{1}$$

Drawing D through G, leads to:

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} H_i H_{i'}^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} (\overline{H_i G} + \overline{G H}_{i'})^2.$$

By developing (note that G is also the barycenter of the projected points) we obtain

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} H_i H_{i'}^2 = 2n \sum_{i=1}^{n} G H_i^2. \tag{2}$$

Moreover

$$\begin{aligned} \sum_{i=1}^{n} G W_i^2 &= \sum_{i=1}^{n} (\overline{GH}_i + \overline{H_i W}_i)^2 \\ &= \sum_{i=1}^{n} G H_i^2 + \sum_{i=1}^{n} H_i W_i^2 \qquad (3) \\ &= C^{st}. \end{aligned}$$

Then, to maximize $\sum_{i=1}^{n} \sum_{i'=1}^{n} H_i H_{i'}^2$ (1) is equivalent to maximizing $\sum_{i=1}^{n} G H_i^2$(2) which is equivalent to minimizing $\sum_{i=1}^{n} H_i W_i^2$ (3).

FACTOR PLANE

The chosen affined sub-space will have two dimensions. Its associated vectorial sub-space is called the factor plane. It is determined by two unitary eigenvectors of one dimension (factor axis or principal component) of the matrix ${}^tT.\, T$, where

$$T = [t_{i,j}], \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, p$$

with

$$t_{i,j} = \frac{1}{(n)^{1/2}} (y_{i,j} - g_j).$$

This transformation permits centering the barycenter G on the origin of the individual-space.

For each analysis we have always projected the clouds in the factor plane which retains the largest information from the numerical matrix (in general around 80%), the initial information being found again with all the factor planes.

GRAPHICAL REPRESENTATIONS

There are two spaces; therefore there will be two graphical representations.

The correlation circle-graph represents the projection of the variable-space. Each variable is represented in the space $\mathbb{R}^n$ by a point on the sphere whose center is the origin and whose radius is equal to 1. Each variable is projected inside a circle (correlation-circle) having the same center and radius. In our figures, the variables (parameters) are identified by alphabetical letters.

The square-graph represents the projection of the individual-space. In our figures, the individuals (sequences) are identified by a number followed by two letters. The number 1 represents the NCS, the number 2 the PCS. The variables are also projected with the individuals in the square-graph but the projection point of a given variable is not significant, with the exception of the line which goes through this point and the barycenter of the individual-space (which is the center of the square-graph). This straight line is called the variable axis. Therefore, a sequence with a high (respectively low) value for a given variable in the matrix, will have in the factor plane an orthogonal projection on the given variable axis nearest to (respectively far away from) the projection point of the given variable. In this case, the variables are identified by the number 0 followed by two numbers and located on the square perimeter.

(III) The discriminating analysis (DA) (Romeder, 1973) is a numerical method which permits to classify, according to a given combination of quantitative parameters, individuals which are divided into classes without any statistical hypothesis. The DA is a PCA about the barycenters of individual classes. This analysis is done with a metric (D^2 of Mahalanobis, 1936) on the individual-space for two purposes: (i) in order to keep the barycenters away from one another; (ii) in order to group the individuals of each class around their barycenter.

Taking the above mentioned definitions, then the distance of Mahalanobis between two individuals $W_i W_{i'}$ is

$$d^2(W_i, W_{i'}) = {}^t(x_i - x_{i'})\, V^{-1}(x_i - x_{i'})$$

where V is the variance-covariance matrix:

$$V = [v_{j,j'}], \qquad j = 1, \ldots, p, \qquad j' = 1, \ldots, p$$

with

$$v_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)(x_{i,j'} - \bar{x}_{j'}).$$

In the case of our study, this analysis is evident since there are only two classes: the class of PCS and the class of NCS. Therefore, there is only one factor axis which is the straight line joining the barycenters of the two classes. The decision

rule is a rule of assignment to the nearest neighbour and hence, the success-percentage of a sample is the percentage of sequences correctly classified with the DA. In actual fact, the representative sample is separated into two representative samples. For a given combination of parameters, the DA determines with a basis sample (7/10 of the initial sample sequences and with an equal number of PCS and NCS), the new metric and the success-percentage (which is called success-percentage of the basis sample). Next, this information is tested with a test sample which contains 3/10 of the initial sample sequences and also with an equal number of PCS and NCS. Obviously, in this case there is only one evaluation of the success-percentage (which is called success-percentage of the test sample). Now, to determine the best combination of parameters, a step-by-step method permits us to choose successively 1, 2, ... i variables, among the p variables, calling into question—at the i step—the choice of the variables at the $i-1$ step. For each step, this latter method keeps only the combination of parameters which has given the best success-percentage of the basis sample with the DA. Therefore, the best combination of parameters will be the combination of parameters which has the best success-percentages with the basis and test samples. Programs were written in FORTRAN 77 and run on a NORD 500 computer.

3. Results

(A) D_j PARAMETER ANALYSIS

The D_j parameter is defined as follows: $D_j = T_j / L$.

T_j: number of thymine-couples where the two thymines are separated by j bases in a given sequence.

L: number of bases in a given sequence which permits a normalization independent of the base distribution.

By varying j between 0 and 9, a given sequence is characterized by 10 parameters which may have the forms D_{0+3n}, D_{1+3n} and D_{2+3n} with $0 \leq n \leq 3$.

These D_j parameters are similar to Fickett's parameters called autocorrelation for thymine (cf. Fickett (1982), Fig. 1).

(*a*) *Sample of* 507 *PCS with lengths greater than* 200 *bases*

Principal component analysis on the D_j parameter-space. Figure 1 shows two groups of well separated parameters; on the one hand the group 2, 5, 8 (graphical symbols *C*, *F*, *I*) which have the form (D_{2+3n}, $n \geq 0$); on the other hand the group 0, 1, 3, 4, 6, 7, 9 (graphical symbols *A*, *B*, *D*, *E*, *G*, *H*, *J*) which have the form (D_{0+3n}, D_{1+3n}, $n \geq 0$). In order to get a readable graph, we have only represented 10 parameters. But this difference of periodicity between the two groups of parameters has been found again with the D_j parameters, j varying between 0 and 198. Furthermore, the same conclusions have been obtained with the analyses of the three other bases: adenine, cytosine and guanine.

Clearly, our statistical method allowed us to again produce Fickett's results (1982) of autocorrelation for thymine. Indeed, the top graph of Fickett's Fig. 1, shows that

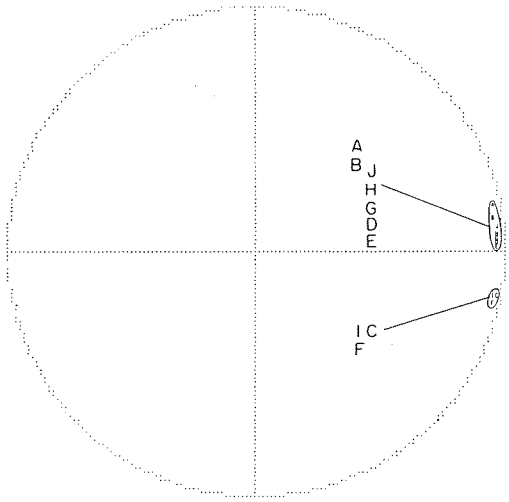

FIG. 1. D_j parameter analysis. Sample of 507 PCS with lengths greater than 200 bases. Principal component analysis on the D_j parameter-space (j varying between 0 and 9). Correlation circle with a radius equal to 1.

Graphical symbols	D_j parameters	Coordinate (Horizontal)	Coordinate (Vertical)
A	D_0	0·94	0·19
B	D_1	0·95	0·13
C	D_2	0·96	−0·18
D	D_3	0·97	0·03
E	D_4	0·97	0·03
F	D_5	0·95	−0·22
G	D_6	0·97	0·06
H	D_7	0·96	0·07
I	D_8	0·96	−0·19
J	D_9	0·97	0·09

there is a difference of percentages between the thymine-couples separated by $2+3n$ bases and those separated by $0+3n$, $1+3n$ bases, n varying between 0 and 66. $2+3n$ are represented by peaks, whereas $0+3n$, $1+3n$ are represented by troughs. Shulman *et al.* (1981) also showed this property by making use of statistical tests pertaining to the nucleotide sequences of the RNA phage MS2 and the DNA phase ØX.

Principal component analysis on the PCS-space. In order to simulate chance, we have generated a random sequence with the Monte Carlo method. This one is projected to the barycenter of the PCS. Hence, the mean value of the ten D_j parameters with all the PCS is near the mean value of the ten D_j parameters with the random sequence which has base frequencies equal to 0·25 with a variation smaller than 0·004.

This result still agrees with Fickett's results (1982). The top graph of Fickett's Fig. 1, shows that the mean percentage of the thymine-couples, separated by k bases,

is small, k varying between 0 and 198 and particularly for k varying between 0 and 9. This mean percentage value is near to zero which demonstrates the randomness.

(*b*) *Sample of* 90 *NCS with lengths greater than* 200 *bases*

For the purpose of characterizing specifically the PCS, we have applied the same analyses with the NCS.

Principal component analysis on the D_j parameter-space. Figure 2 does not allow to identify groups of D_j parameters (graphical symbols $A, \ldots, J$). We have never found any groups of D_j parameters with j varying between 0 to 198. Furthermore, the same conclusions have been obtained with the analyses of the three other bases.

Along the same lines, the bottom graph of Fickett's Fig. 1, (1982) shows that the percentages of the thymine-couples separated by $0+3n$, $1+3n$ and $2+3n$ bases, are almost identical, n varying between 0 and 66.

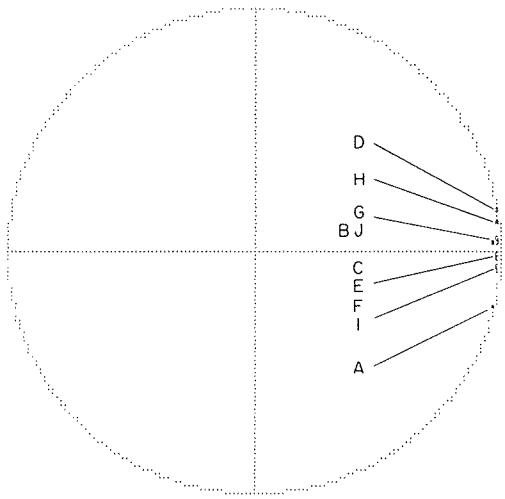

FIG. 2. D_j parameter analysis. Sample of 90 NCS with lengths greater than 200 bases. Principal component analysis on the D_j parameter-space (j varying between 0 and 9). Correlation circle with a radius equal to 1.

Graphical symbols	D_j parameters	Coordinate (Horizontal)	Coordinate (Vertical)
A	D_0	0·96	−0·23
B	D_1	0·97	0·03
C	D_2	0·98	−0·02
D	D_3	0·98	0·16
E	D_4	0·98	−0·03
F	D_5	0·98	−0·07
G	D_6	0·98	0·05
H	D_7	0·98	0·12
I	D_8	0·98	−0·07
J	D_9	0·99	0·05

Principal component analysis on the NCS-space. Contrary to the PCS, the random sequence has an orthogonal projection on the D_j parameter axes far away from the projection points of the D_j parameters, i.e. it has low values (below average) for the D_j parameters. Hence, the mean value of the ten D_j parameters with all the NCS is higher than the mean value of the ten D_j parameters with the random sequence.

This result still agrees with Fickett's results (1982). The bottom graph of Fickett's Fig. 1, shows that the mean percentage of the thymine-couples, separated by k bases, is higher than zero with the first values of k, in particular for k varying between 0 and 9.

(*c*) *Sample of* 58 *PCS and* 57 *NCS with lengths greater than* 200 *bases*

The ten D_j parameters permit to distinguish between the PCS and the NCS. As a matter of fact:

(i) only with the PCS, two groups of parameters (D_{2+3n}, $n \geq 0$) and (D_{0+3n}, D_{1+3n}, $n \geq 0$) have a difference of periodicity.

(ii) The mean value of the ten D_j parameters with the NCS is higher than the mean value of the ten D_j parameters with the random sequence and the PCS.

We want to test if these characteristics are sufficient to classify the sequences into coding ones or into non-coding ones.

Principal component analysis on the D_j parameter-space. The two groups of parameters are individualized, but not quite as readable as with the single sample of PCS. This is probably a consequence of the background noise introduced by the NCS.

Principal component analysis on the space of PCS and NCS. Figure 3 shows a separation between the PCS (letters preceded by number 2) and the NCS (letters preceded by number 1). We have characterized this separation by a straight line D. In future, we shall make use of automatic classification in order to draw the lines. As a rule, the NCS have orthogonal projections on the ten D_j parameter axes which are closer to the projection points of the ten D_j parameters than to the PCS and the random sequence (graphical symbol 2$$).

Discriminating analysis

Figure 4 shows that, at the sixth step, the best combination is obtained with the parameters D_2, D_3, D_4, D_7, D_8, D_9, which classify into coding or into non-coding sequences; the sequences of a basis sample (40 PCS and 40 NCS) and of a test sample (18 PCS and 17 NCS) with success-percentages of about 86%. The introduction of additional parameters does not increase the success-percentages, whereas the suppression of parameters may sometimes involve an important modification of them. The parameters D_3, D_9 having the form (D_{0+3n}, $n \geq 0$), D_4, D_7 having the form (D_{1+3n}, $n \geq 0$) and D_2, D_8 having the form (D_{2+3n}, $n \geq 0$), suggest that the discriminating parameters do not necessarily depend on the properties of the genetic code. This fact will be tested below with other samples and parameters.

Fickett's method (1982) misclassifies 5% of the regions tested and gives an answer of "no opinion" one-fifth of the time. If we consider that there is one chance out

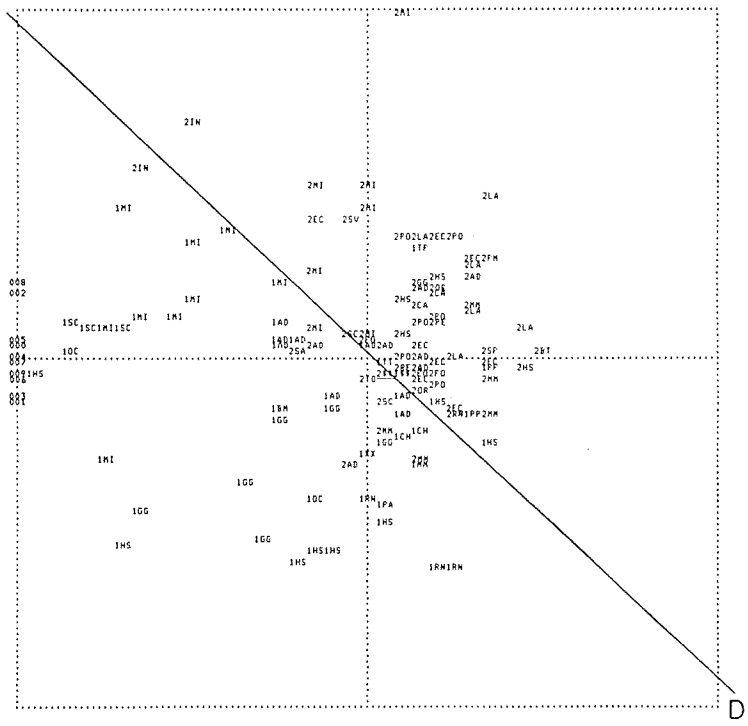

FIG. 3. D_j parameter analysis. Sample of 58 PCS and 57 NCS with lengths greater than 200 bases. Principal component analysis on the space of PCS and NCS. The letters having the form $00j$ represent the parameter D_j (j varying between 0 and 9). The letters preceded by number 1 represent the NCS, the ones preceded by number 2 the PCS. The graphical symbol 2$$ represents the random sequence. The visual separation limit of the two groups of sequences has been characterized with a straight line D.

of two to classify a sequence among the "no opinion", then the success-percentage if 85%. Therefore, Fickett's probabilistic method, using parameters related to the properties of the genetic code, and our statistical method—using parameters which are only based on the distribution of thymine—give success-percentages having similar magnitudes.

(B) DI_j PARAMETERS ANALYSIS

In order to carry on the investigations of the above hypothesis, we define the DI_j parameter as follows: $DI_j = TI_j / L$.

TI_j: Number of thymine-couples where two thymines are separated by j bases other than thymine in a given sequence.

L: Number of bases in a given sequence which permits a normalization independent of the base distribution.

By varying j between 0 and 15, a given sequence is characterized by 16 parameters. The DI_j parameters do not have the same chances to appear. But it is not necessary

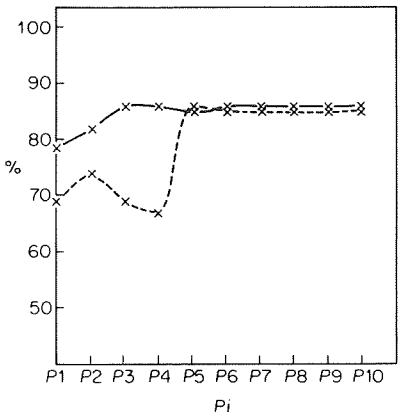

FIG. 4. D_j parameter analysis. Sample of 58 PCS and 57 NCS with lengths greater than 200 bases. Discriminating analysis: The horizontal axis represents the *Pi* which corresponds to *i* parameters D_j for the step *i* which has given the best success-percentage of the basis sample (*i* varying between 1 and 10, *j* varying between 0 and 9). The vertical axis represents the success-percentage. The full line represents the basis sample, the dashed line the test sample.

*P*1: D_3
*P*2: D_2, D_4
*P*3: D_0, D_6, D_8
*P*4: D_0, D_2, D_6, D_8
*P*5: D_3, D_4, D_6, D_7, D_8
*P*6: $D_2, D_3, D_4, D_7, D_8, D_9$
*P*7: $D_0, D_2, D_3, D_4, D_7, D_8, D_9$
*P*8: $D_1, D_2, D_3, D_4, D_6, D_7, D_8, D_9$
*P*9: $D_0, D_1, D_2, D_3, D_4, D_6, D_7, D_8, D_9$
*P*10: $D_0, D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9$

to do probabilitistic corrections because the numerical data are divided by the standard deviation of their columns.

The sample contains 57 PCS and 56 NCS with lengths greater than 200 bases.

Principal component analysis on the DI_j parameter-space

This analysis individualizes a group of three parameters: DI_0, DI_1, DI_2 which are strongly correlated. Hence, a sequence, being either a coding or non-coding one, which has a high (respectively low) value for one of the three parameters, will have high (respectively low) values for the two others.

Principal component analysis on the space of PCS and NCS

Figure 5 shows a separation (straight line D) between the PCS (letters preceded by number 2) and the NCS (letters preceded by number 1). As a rule, the IVS have orthogonal projections on the DI_0, DI_1, DI_2 parameters axes which are closer to the projection points of the DI_0, DI_1, DI_2 parameters than to the PCS. Therefore, the NCS have more doublets TT (thymine-thymine), triplets TXT, quadruplets TXXT (X represents any base but thymine) than the PCS.

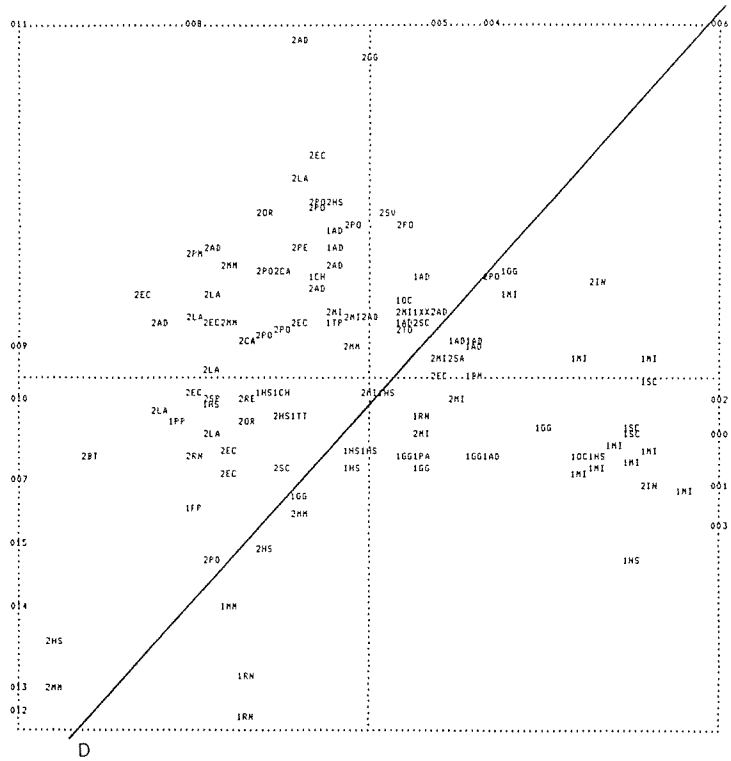

FIG. 5. DI_j parameter analysis. Sample of 57 PCS and 56 NCS with lengths greater than 200 bases. Principal component analysis on the space of PCS and NCS. The letters having the forms $00j$ or $0j$ represent the DI_j parameters (j varying between 0 and 15). The letters preceded by number 1 represent the NCS, the ones preceded by number 2 the PCS. The visual separation limit of the two groups of sequences has been characterized with a straight line D.

Discriminating analysis

Figure 6 shows that, at the fifth step, the best combination is obtained with the parameters DI_0, DI_1, DI_2, DI_3 and DI_5. The success-percentage is about 91% with the basis sample (40 PCS and 39 NCS) and about 84% with the test sample (17 PCS and 17 NCS).

Therefore, the DI_j parameters as the D_j parameters can discriminate the sequences without using any properties of the genetic code.

4. Discussion

The principal component analysis and the discriminating analysis without statistical hypothesis allow DNA sequences to be analysed. Indeed, these two complementary statistical methods permitted, on the one hand, reproduction of many results of Fickett's probabilistic method (1982); on the other hand, showing the existence

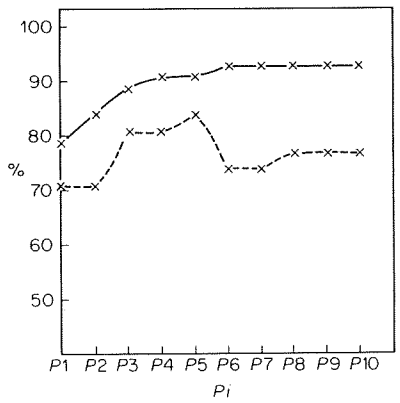

FIG. 6. DI_j parameter analysis. Sample of 57 PCS and 56 NCS with lengths greater than 200 bases. Discriminating analysis: The horizontal axis represents the *Pi* which corresponds to i parameters DI_j for the step i which has given the best success-percentage of the basis sample (i varying between 1 and 10, j varying between 0 and 15). The vertical axis represents the success-percentage. The full line represents the basis sample, the dashed line the test sample.

*P*1: DI_1
*P*2: DI_1, DI_7
*P*3: DI_1, DI_2, DI_3
*P*4: DI_1, DI_2, DI_3, DI_5
*P*5: DI_0, DI_1, DI_2, DI_3, DI_5
*P*6: DI_0, DI_1, DI_2, DI_3, DI_4, DI_7
*P*7: DI_0, DI_1, DI_2, DI_3, DI_4, DI_6, DI_7
*P*8: DI_0, DI_1, DI_2, DI_3, DI_4, DI_5, DI_7, DI_9
*P*9: DI_0, DI_1, DI_2, DI_3, DI_4, DI_5, DI_6, DI_7, DI_9
*P*10: DI_0, DI_1, DI_2, DI_3, DI_4, DI_5, DI_6, DI_7, DI_8, DI_9

of parameters which discriminate DNA sequences in coding or non-coding ones, without having to make use of the properties of the genetic code.

Nevertheless, there are differences with the data and with their treatment between the probabilistic and the statistical methods. Fickett uses a sample having a number of coding sequences which is greater than the number of non-coding sequences in a ratio of 1·3; this perceived bias was not corrected. The essential point is that the probabilistic method is losing information because the data cannot be retrieved from the results.

The principal component analysis is a graphical method which permits expression of a large data array. There is no waste of information because the study of all factor planes leads again to the initial information. This fast method allows the analysis of biological hypotheses: it is a help to the intuition. The discriminating analysis is a quantitative method to confirm the hypotheses: it is an aid to the decision.

The generality of this methodology allows testing of any quantitative parameters without limitation of number or of combination; this being valid on any set or any subset of sequences. Indeed, with an interactive communication, the sequences can be chosen according to taxonomic groups or intervals of length. Therefore, different groups of discriminating parameters can be used to classify sequences.

The principal component analysis and the discriminating analysis, associated with other methods and with the increase of known sequences, allow not only to distinguish between coding and non-coding sequences, but also in a more general way, to reveal genetic constraints which are specific to some species.

I would like to express my sincere thanks to Professor Thomas Bickle and to Dr John Shepherd, both from the Biozentrum of Basel; to Dr Didier Arquès, from the Institut des Sciences Exactes et Apliquées de Mulhouse, and to Dr Pierre Oudet, from the Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS de Strasbourg for their cooperation and their kind encouragement.

REFERENCES

FICKETT, J. W. (1982). *Nucleic Acids Res.* **10,** 5303.
GRANTHAM, R., GAUTIER, C., GOUY, M., JACOBZONE, M. & MERCIER, R. (1981). *Nucleic Acids Res.* **9,** r43.
LEBART, L., MORINEAU, A. & FÉNELON, J. P. (1979). In: *Traitement des Données Statistiques.* Paris: Dunod.
MAHALANOBIS, P. C. (1936). *Proc. natn. Inst. Science-India* **12,** 49.
ROMEDER, J. M. (1973). In: *Méthodes et Programmes d'Analyse Discriminante.* Paris: Dunod.
SHEPHERD, J. C. W. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 1596.
SHULMAN, M. J., STEINBERG, C. M. & WESTMORELAND, N. (1981). *J. theor. Biol.* **88,** 409.
SMITH, T. F., WATERMAN, M. S. & SADLER, J. R. (1983). *Nucleic Acids Res.* **11,** 2205.
STADEN, R. & MCLACHLAN, A. D. (1982). *Nucleic Acids Res.* **10,** 141.
TRAMONTANO, A., SCARLATO, V., BARNI, N., CIPPOLLARO, M., FRANZE, A., MACCHIATO, M. F. & CASCINO, A. (1984). *Nucleic Acids Res.* **12,** 5049.