CABIOS

Vol. 12 no. 6 1996
Pages 519–524

Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences

G.Perrière, J.R.Lobry[1] and J.Thioulouse

Abstract

This report describes two applications of a multivariate method for studying classes of nucleotide or protein sequences: correspondence discriminant analysis (CDA). The first example is the discrimination between Escherichia coli proteins according to their subcellular location (membrane, cytoplasm and periplasm). The high resolution of the method made it possible to predict the subcellular location of E.coli proteins for whom this information is not known. The second example is discrimination between the coding sequences of leading and lagging strands in four bacteria: Mycoplasma genitalium, Haemophilus influenzae, E.coli and Bacillus subtilis. The programs used for computing the analysis are integrated in a publicly available package that runs on MacOS 7.x or Windows 95 operating systems (http://biomserv.univ-lyon1.fr/ADE-4.html). These programs are also accessible through our World Wide Web server (http://biomserv.univ-lyon1.fr/NetMul.html).

Introduction

Discriminant analysis (DA) is a multivariate technique that has been used in sequence analysis to study the properties of protein sequences (Klein *et al.*, 1984), predict splice junctions in mRNA (Nakata *et al.*, 1985; Iida, 1988), discriminate between protein secondary structural segments (Kanehisa, 1988), separate intracellular and extracellular proteins (Nakashima and Nishikawa, 1994), and locate internal exons in human DNA (Solovyev *et al.*, 1994). This report describes the use of an extension of DA, correspondence discriminant analysis (CDA). CDA was initially developed to analyse ecological data (Chessel and Thioulouse, 1996) and is appropriate for comparing individuals belonging to several groups. It is particularly useful for describing how groups differ in terms of variables.

The two examples described show how CDA can be used to study pre-defined groups of protein or nucleic acid sequences. The first example discriminates between proteins according to their subcellular location in *Escherichia coli.* Multivariate techniques have been employed previously to try to separate proteins according to their location. For example, a study using correspondence analysis (CA) has shown that it is possible to distinguish integral membrane proteins from the others in *E.coli* (Lobry and Gautier, 1994). Another study used DA to show that it is possible to differentiate intracellular from extracellular proteins on the basis of their amino acid composition (Nakashima and Nishikawa, 1994). However, in both cases, the methods used have only produced a partition of the proteins into two groups. This report shows how CDA can distinguish between three classes of proteins, membrane proteins (MP), cytoplasmic proteins (CP) and periplasmic proteins (PP), on a single factorial map. The second example discriminates between coding sequences (CDS) according to their location on the leading or lagging strands of the chromosome for species of bacteria, *Mycoplasma genitalium*, *Haemophilus influenzae*, *E.coli* and *Bacillus subtilis.* Leading CDS are defined by the fact they are transcribed divergently from the origin of replication, while lagging CDS are transcribed convergently.

Algorithm

CDA is a powerful tool because it combines the advantages of standard CA and DA. Like CA, CDA removes the effect of sequence length and takes into account the global amino acid or codon frequencies. Like DA, CDA takes into account difference in the variability of groups, and the row scores maximize the between-class variance. The method could be thought of as an extension of DA to deal with contingency tables, or as an extension of CA to take into account the belonging of individuals to different groups. CDA is a descriptive method in that it describes how individuals from different groups differ, but it is also a predictive method in that an individual from an unknown group can be tentatively assigned to a group. A complete description of the algorithm can be found in section 5.6 of the ADE-4 documentation (Chessel and Thioulouse, 1996) (ftp://biom3.univ-lyon1.fr/pub/mac/ADE/ADE4/DocThem/Thema-5.6.hqx).

The first step in a CDA was the computation of a standard CA on the table containing the absolute

Laboratoire de Biométrie, Génétique et Biologie des Populations, UMR CNRS n 5558, Université Claude Bernard - Lyon 1, 43, bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

[1] *To whom correspondence should be addressed*

E-mail: {perriere, lobry, thioulou}@biomserv.univ-lyon1.fr

frequencies of amino acids or codons. The individuals (proteins or CDS) were ordered in their pre-defined groups within the table. Then CDA itself was then computed. A Monte Carlo test was used to check the significance of the discrimination. This method consisted of repeated random permutations of rows between the pre-defined groups. Once the factor scores for the rows and the columns had been obtained, supplementary individuals were projected into the analysis to assign them to a group. The threshold on a given axis of the analysis T allowing the separation of two classes is equal to:

$$T = \frac{\hat{s}_1\bar{x}_2 + \hat{s}_2\bar{x}_1}{\hat{s}_1 + \hat{s}_2} \qquad (1)$$

where $\bar{x}_1$, $\bar{x}_2$ and $\hat{s}_1$, $\hat{s}_2$ are the averages and standard deviations of the factor scores for the two classes considered. The factor score F_j^k for an individual j on axis k is equal to:

$$F_j^k = \frac{N_{..}}{N_{.j}} \times \sum_{i=1}^{n} \frac{A_i^k \times N_{ij}}{N_{i.}} \qquad (2)$$

where N_{ij} is the number of amino acids or codons i in the individual j, and A_i^k is the factor score of the amino acid or codon i on axis k. The reliability of the assignment was estimated using a cross-validation test. The data set was split randomly into two parts, a training set and a test set. The CDA was computed on the training set and the assignment was carried out on the individuals belonging to the test set.

Implementation

CDA is one of the methods implemented in the ADE-4 package (Thioulouse *et al.*, 1995). This package runs on the MacOS 7.x and Windows 95 operating systems. Instructions on how to download and use it can be found at http://biomserv.univ-lyon1.fr/ADE-4.html. The ADE-4 modules required to perform CDA are: ADE-Trans, COA, CategVar and Discrimin. The method is also implemented on the NetMul World Wide Web server (Thioulouse and Chevenet, 1996). This server can be reached at http://biomserv.univ-lyon1.fr/NetMul.html. All the analyses presented here were computed on a Macintosh and on the Web server. Factorial map and Gauss curves were drawn on a Macintosh with the ADE graphical modules ScatterClass and Graph1DClass.

Data sets

The initial data set used to study the discrimination of proteins according to their subcellular location was 3472 *E.coli* proteins from SWISS-PROT release 32 (Bairoch and Apweiler, 1996). Any hypothetical proteins, plasmid proteins, proteins with <50 amino acids, proteins without any indication of their subcellular location in the features, and proteins for which the subcellular location was unsure ('putative', 'potential', 'probable' or 'obtained by similarity') were removed from the data set. Proteins simply anchored in the inner membrane were not considered. Indeed, in many cases it was not possible to determine whether the unanchored terminal region of the protein lay on the periplasmic or in the cytoplasmic side of the membrane. Of the 413 remaining proteins, 188 (45.5%) were MP, 162 (39.2%) were CP and 63 (15.3%) were PP. Because the subcellular location of MP is more often documented than for the other groups of proteins, MP were over-represented in the data set compared to their real percentage in *E.coli*. Lobry and Gautier (1994) estimated that they account for 10%. A total of 1077 proteins matching the conditions given above, except that no subcellular location was given in the SWISS-PROT features, were used as supplementary individuals in the analysis.

The differences in codon usage in leading and lagging strands in *M.genitalium*, *H.influenzae*, *E.coli* and *B.subtilis* were studied using only large sequences that included the replication origin. For each species, partial CDS and CDS of < 50 codons were removed. The complete genome of *M.genitalium* is available as a single sequence (Fraser *et al.*, 1995) and 466 CDS were extracted from this sequence; 374 (80%) were in the leading group and 92 (20%) in the lagging group. The deviation from the universal genetic code for this species was taken into account when computing codon frequencies. The complete genome of *H.influenzae* is also available as a single sequence (Fleischmann *et al.*, 1995) and 1647 CDS were extracted, including 883 (54%) in the leading group and 764 (46%) in the lagging group. The 1 616 174 bp fragment located between min 67.4 and min 4.1 of the *E.coli* chromosome was used (Daniels *et al.*, 1992; Yura *et al.*, 1992; Blattner *et al.*, 1993; Burland *et al.*, 1993; Plunkett *et al.*, 1993; Sofia *et al.*, 1994) and 1416 CDS were extracted; 774 (55%) were in the leading group and 642 (45%) in the lagging group. The 276 609 bp sequence located between degree 348 and degree 12 of the *B.subtilis* chromosome was used. This sequence (accession number BS0328) was taken from the NRSub database release 7 (Perrière *et al.*, 1996) and 246 CDS were extracted, including 180 (73%) in the leading group and 66 (27%) in the lagging group.

All the data sets presented in this section, as well as the detail of the results obtained in the analyses, can be downloaded through anonymous FTP at ftp://biom3.univ-lyon1.fr/pub/datasets/CABIOS96/.

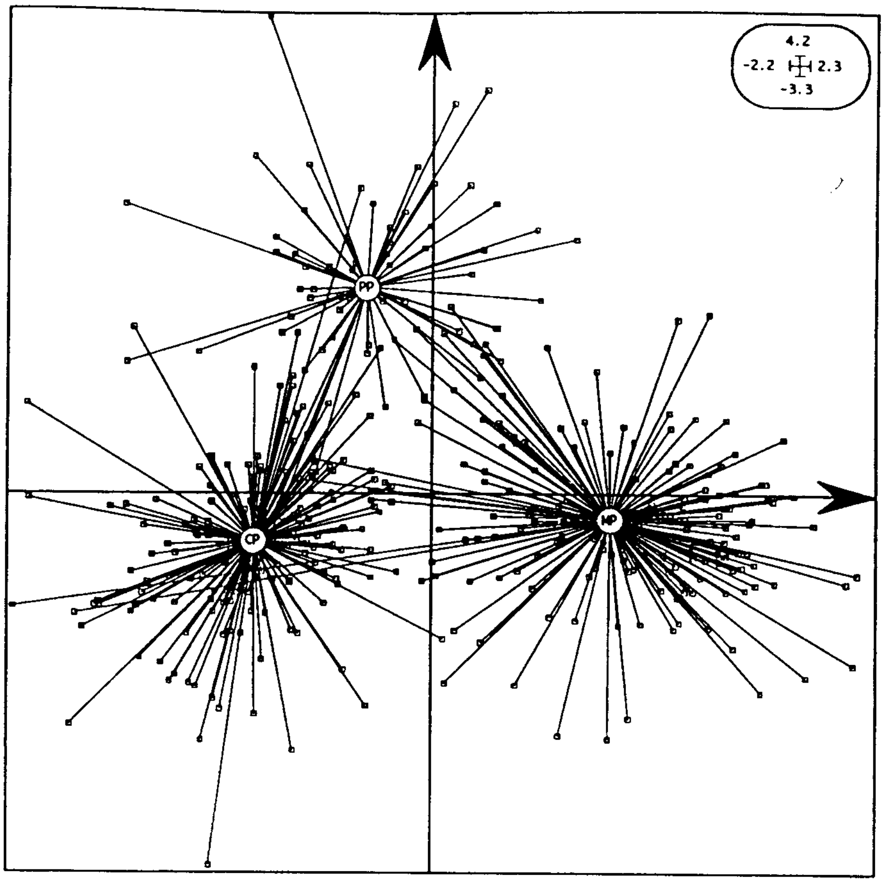

Fig. 1. Factorial map of the the two discriminant axes of the analysis on 413 *E.coli* proteins. Each protein is represented by a dot linked by a line to the gravity centre of the group it belongs to. The first axis discriminates membrane proteins (MP) from cytoplasmic proteins (CP) and periplasmic proteins (PP), while the second axis discriminates PP from CP and MP.

Results

The factorial map obtained by crossing the two discriminant axes for the proteins showed that MP, PP and CP were well separated from each other (Figure 1). The first axis separated MP from CP and PP. The amino acids with the highest factor scores on this axis (those identifying MP) were Ile, Met, Gln, Val, Leu and Trp (Table I). The second axis separated PP from MP and CP. The amino acids with the highest factor scores on this axis (those identifying PP) were Lys, Pro, Ser, Gln, Asn and Trp. CP were enriched in amino acids that had negative factor scores on the first axis and negative or slightly positive factor scores on the second axis (Cys, His, Glu and Asp). The reliability of the prediction between MP and non-PP (CP + PP) was 89%, and 84% between PP and non-PP (MP + CP), taking into account the number of false negatives and false positives (Table II).

The factor scores and thresholds obtained were used to assign putative subcellular locations to the 1077 proteins for which this information was not given in SWISS-PROT. This resulted in 44 proteins being attached to the MP group, 813 to the CP group and 150 to the PP group. It was not possible to assign a location to the 70 remaining proteins since their scores on both axes were very high, so that they would have been attached to both the MP and CP groups. An example of data reconstitution is given for protein AraJ (accession number P23910) in Table I. This protein is supposed to be involved in the transport of arabinose polymers, its high score on the first axis (1.627) and its low score on the second axis (−0.450) means that it probably belongs to the MP group.

The distribution of the CDS factor scores in the leading and lagging CDS discrimination analysis showed that the leading and lagging groups, for the four species, were separated (Figure 2). The reliability of the predictions, taking into account the number of false negatives and false positives, was 71% in *M.genitalium*, 66% in *H.influenzae*, 76% in *E.coli* and 80% in *B.subtilis* (Table III). When the discriminant power of synonymous codons was compared, leading CDS were found to be enriched in codons containing keto bases (G and T), while lagging ones were enriched in codons with amino bases (A and C) (Figure 3). There were few exceptions to this general trend, mainly Asn

Table I. Factor scores for the amino acids on the two axes of the discriminant analysis on 413 *E.coli* proteins and example of protein factor score computation

	A_i^1	A_i^2	$N_{i\cdot}$	N_{ij}	$A_i^1 N_{ij}/N_{i\cdot}$	$A_i^2 N_{ij}/N_{i\cdot}$
Arg	0.0340	−0.1593	7747	11	0.000048	−0.000226
Ala	0.0217	−0.0590	16 357	49	0.000065	−0.000177
Gln	0.2314	0.4897	6462	7	0.000251	
	0.000530					
Cys	−0.1180	−0.2125	1399	5	−0.000422	−0.000759
Leu	0.2204	−0.2785	17 445	57	0.000720	−0.000910
Gly	0.1629	−0.2156	13 134	44	0.000546	−0.000722
His	−0.4639	−0.0874	3206	5	−0.000723	−0.000136
Phe	0.1551	−0.4360	7608	29	0.000591	−0.001662
Ser	0.1495	0.5643	9341	28	0.000448	0.001691
Val	0.2205	−0.0015	12 591	28	0.000490	−0.000003
Glu	−0.7657	−1.4502	8648	7	−0.000620	−0.001174
Ile	0.4381	−0.8031	10 431	28	0.001176	−0.002156
Thr	0.1193	0.0581	8950	15	0.000200	0.000097
Lys	−0.1495	1.5349	7422	12	−0.000242	0.002482
Asp	−0.8567	−0.2267	7972	4	−0.000430	−0.000114
Met	0.2481	0.0760	5139	22	0.001062	0.000325
Pro	−0.1939	0.9525	7083	14	−0.000383	0.001883
Asn	0.1525	0.3277	6253	10	0.000244	0.000524
Tyr	0.1847	−0.2869	4909	15	0.000564	−0.000877
Trp	0.2095	0.2137	2782	4	0.000301	0.000307
Σ			164 879	394	0.003887	−0.001076
F_j^k					(1.627)	(−0.450)

Columns A_i^1 and A_i^2 contain the amino acid factor scores on the two discriminant axes, $N_{i\cdot}$ contains the absolute amino acid frequencies in the whole data set, and N_{ij} contains the absolute amino acid frequencies in protein AraJ (P23910). The factor score of AraJ on the two axes of the analysis is computed using equation (2), with $N_{\cdot\cdot}$ and $N_{\cdot j}$, respectively, equal to the sum of the $N_{i\cdot}$ and the N_{ij} columns of the table. The threshold value between MP/non-MP is equal to −0.024 and the threshold value between PP/non-PP is equal to 0.617.

codons in *M.genitalium*, Cys codons in *H.influenzae* and *M.genitalium*, and Phe and Ser codons in *B.subtilis*.

Discussion

Discrimination between proteins

The enrichment of MP in Ile, Met, Val, Leu was not surprising, as these proteins are known to be highly hydrophobic (Engelman *et al.*, 1986). The identification of CP by His, Glu and Asp is understandable as these amino acids are charged, and so are avoided in membrane and exported proteins. The relatively high number of Pro, Ser, Gln, Asn and Trp in PP could be, as proposed by Nakashima and Nishikawa (1994), because these amino acids slow the folding of proteins, and slow folding is required for exported proteins. Indeed, peptides can only be transported across the membrane in their unfolded state. More surprising is the large number of Lys in PP, as this amino acid is polar, and so should be avoided in proteins that are transported across the membrane (Engelman *et al.*, 1986).

Table II. Cross-validation test results for *E.coli* proteins

Real class	Predicted class	
	MP	Non-MP
MP	81 (86%)	13 (14%)
Non-MP	9 (8%)	104 (92%)
	PP	Non-PP
PP	30 (94%)	2 (6%)
Non-PP	30 (17%)	145 (83%)

The excellent discrimination between the groups of proteins on the two axes indicates that CDA can be used to predict the subcellular locations of proteins for which this information is not available. The best method available for discriminating MP from non-MP is probably the one published by Klein *et al.* (1985). The technique is a DA of protein sequence characteristics (such as maximum local hydrophobicity), and its accuracy is 95%. The resolution of our method is slightly lower (89%), and this small difference is probably due to the fact that we used only the amino acid composition to separate the proteins. Nakai and Kanehisa (1991) have developed an expert system for PP discrimination that uses amino acid composition data and which can identify these proteins with a reliability of 83%. Our analysis gives equivalent results (84%).The advantage of our approach is that discrimination between the three classes is done in a single analysis, while the preceding methods only separate the proteins into two groups.

Discrimination between CDS

CDA discriminated between leading and lagging CDS,

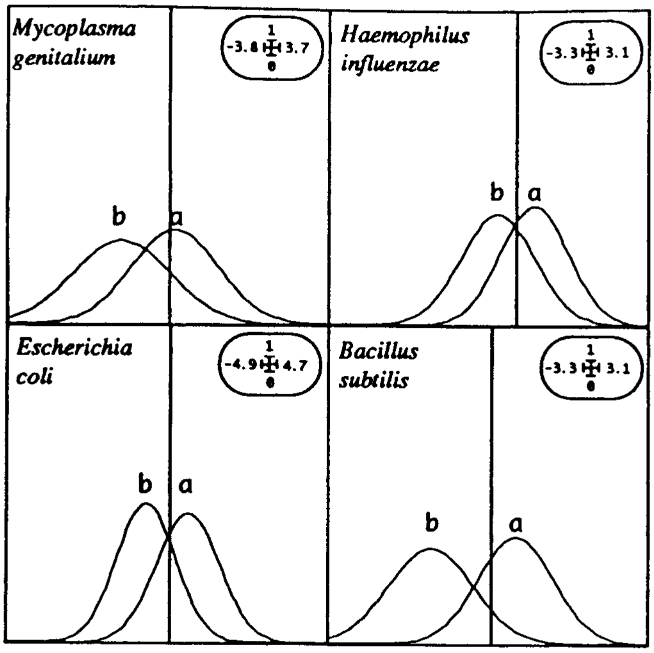

Fig. 2. Distribution of the factor scores on the discriminant axis of the coding sequences belonging to the leading (a) and lagging (b) groups.

despite the small difference in codon usage between the two groups (Lobry, 1996). The reason for this difference is not known, but it is probably the result of asymmetrical substitution patterns in the two DNA strands (Lobry, 1995). Asymmetries are generated by transcription-coupled repair in enterobacterial genes (Schaaper *et al.*, 1987; Francino *et al.*, 1996), but the reason why the effect on leading and lagging CDS should be different is still unclear. The salient feature emerging from the present result is that the relative enrichment of keto bases in leading CDS and amino bases in lagging CDS is common to the four bacterial species studied. This suggests that the underlying mechanism is ancient and has been conserved on an evolutionary time scale.

Table III. Cross-validation test results for *B.subtilis*, *E.coli*, *M.genitalium* and *H.influenzae* CDS

Real class	Predicted class Leading	Lagging
M.genitalium		
Leading	138 (74%)	49 (26%)
Lagging	19 (41%)	27 (59%)
H.influenzae		
Leading	320 (72%)	122 (28%)
Lagging	155 (41%)	227 (59%)
E.coli		
Leading	305 (79%)	82 (21%)
Lagging	89 (28%)	232 (72%)
B.subtilis		
Leading	73 (81%)	17 (19%)
Lagging	7 (21%)	26 (79%)

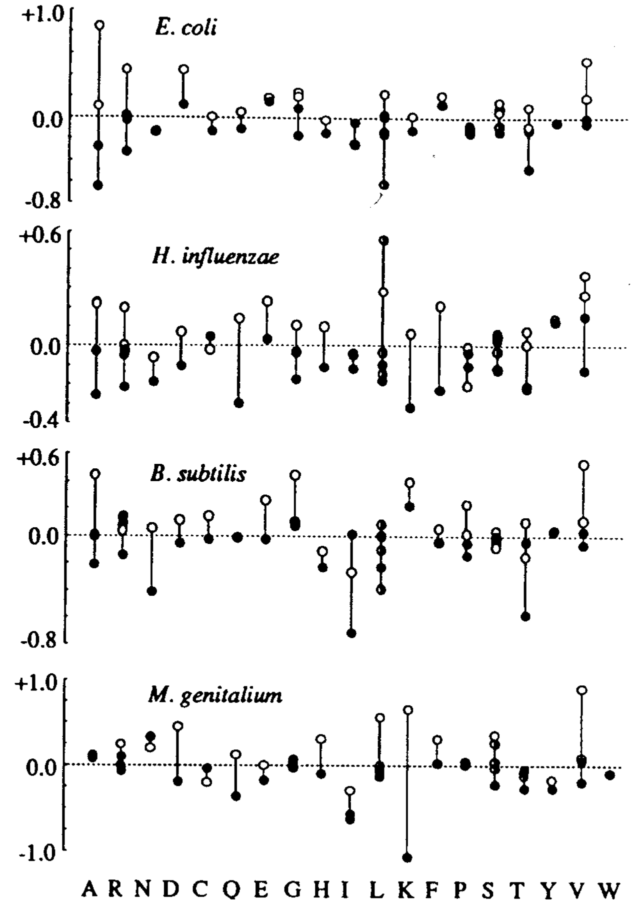

Fig. 3. Discriminant power of codons. Each point represents the discriminant score of one codon, a positive value means that the codon is more frequent in leading than in lagging coding sequences. Codons are grouped by amino acids according to the one-letter code at the bottom of the figure. White dots represent codons with a keto base (G or T) in their third synonymous position, while black dots represents codons with an amino base (A or C). When the first position is synonymous and keto or amino (Leu and Ser), only codons with two keto or two amino bases in the first and third position are depicted as such, the remainder are represented as black and white dots. For Arg, the first base is synonymous but always amino, so that only the third position is considered.

Other possible uses of CDA

The use of CDA is not limited to the examples given here. CDA is a good complement to CA in studies comparing codon usage or amino acid usage between different species. Codon usage in vertebrates is heterogeneous and varies greatly depending on the region of the genome studied (Bernardi *et al.*, 1985). Hence, CA cannot be used to compare codon usage in orthologous genes in these species, as the between-species differences will often be indistinguishable from the within-species differences. However, CDA takes advantage of the *a priori* knowledge of the species to find the linear combination of codons allowing discrimination between genes. Similarly, CDA can be used to compare amino acid usage in orthologous genes of different organisms.

Acknowledgements

We thank Daniel Chessel, University of Lyon, who developed CDA, and one of the anonymous referees for useful comments.

References

Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, **24**, 21–25.

Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.

Blattner,F.R., Burland,V., Plunkett,G., Sofia,H.J. and Daniels,D.L. (1993) Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. *Nucleic Acids Res.*, **21**, 5408–5417.

Burland,V., Plunkett,G., Daniels,D.L. and Blattner,F.R. (1993) DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics*, **16**, 551–561.

Chessel,D. and Thioulouse,J. (1996) *L'analyse discriminante des correspondances. Documentation Thématique ADE-4, Vol. 5, Section 6.* Université Claude Bernard – Lyon 1, Lyon.

Daniels,D.L., Plunkett,G., Burland,V. and Blattner,F.R. (1992) Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science*, **257**, 771–778.

Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Francino,M.P., Chao,L., Riley,M.A. and Ochman,H. (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**, 107–109.

Fraser,C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium. Science*, **270**, 397–403.

Iida,Y. (1988) Categorical discriminant analysis of 3′-splice site signals of mRNA precursors in higher eucaryote genes. *J. Theor. Biol.*, **135**, 109–118.

Kanehisa,M. (1988) A multivariate analysis method for discriminating protein secondary structural segments. *Protein Eng.*, **2**, 87–92.

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

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

Lobry,J.R. (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.*, **40**, 326–330.

Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Lobry,J.R. and Gautier,C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.*, **22**, 3174–3180.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110.

Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, **238**, 54–61.

Nakata,K., Kanehisa,M. and DeLisi,C. (1985) Prediction of splice junctions in mRNA sequences. *Nucleic Acids Res.*, **13**, 5327–5340.

Perrière,G., Moszer,I. and Gojobori,T. (1996) NRSub: a non-redundant database for *Bacillus subtilis. Nucleic Acids Res.*, **24**, 41–45.

Plunkett,G., Burland,V., Daniels,D.L. and Blattner,F.R. (1993) Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. *Nucleic Acids Res.*, **21**, 3391–3398.

Schaaper,R.M., Dunn,R.L. and Glickman,B.W. (1987) Mechanisms of ultraviolet-induced mutation. Mutational spectra in the *Escherichia coli lacI* gene for a wild-type and an excision-repair-deficient strain. *J. Mol. Biol.*, **198**, 187–202.

Sofia,H.J., Burland,V., Daniels,D.L., Plunkett,G. and Blattner,F.R. (1994) Analysis of the *Escherichia coli* genóme. V. DNA sequence of the region from 76.0 to 81.5 minutes. *Nucleic Acids Res.*, **22**, 2576–2586.

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156–5163.

Thioulouse,J. and Chevenet,F. (1996) NetMul, a World-Wide Web user interface for multivariate analysis software. *Comput. Stat. Data Anal.*, **21**, 369–372.

Thioulouse,J., Dolédec,S., Chessel,D. and Olivier,J.M. (1995) ADE software multivariate analysis and graphical display of environmental data. In Guariso,G. and Rizzoli,A. (eds), *Software per l'Ambiente.* Pàtron, Bolonia, pp. 57–62.

Yura,T., Mori,H., Nagai,H., Nagata,T., Ishihama,A., Fujita,N., Isono,K., Mizobuchi,K. and Nakata,A. (1992) Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. *Nucleic Acids Res.*, **20**, 3305–3308.

Received on May 2, 1996; revised and accepted on July 8, 1996