ELSEVIER

Computer Methods and Programs in Biomedicine 70 (2003) 99–105

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins

Guy Perrière *, Jean Thioulouse

Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS no. 5558, Université Claude Bernard–Lyon 1, 43, bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received 1 June 2001; received in revised form 12 October 2001; accepted 24 January 2002

Abstract

Correspondence discriminant analysis (CDA) is a multivariate statistical method derived from discriminant analysis which can be used on contingency tables. We have used CDA to separate Gram negative bacteria proteins according to their subcellular location. The high resolution of the discrimination obtained makes this method a good tool to predict subcellular location when this information is not known. The main advantage of this technique is its simplicity. Indeed, by computing two linear formulae on amino acid composition, it is possible to classify a protein into one of the three classes of subcellular location we have defined. The CDA itself can be computed with the ADE-4 software package that can be downloaded, as well as the data set used in this study, from the Pôle Bio-Informatique Lyonnais (PBIL) server at http://pbil.univ-lyon1.fr.

Keywords: Correspondence discriminant analysis; Gram negative bacteria; Protein subcellular location

1. Introduction

In Gram negative bacteria, after being synthetized by the translation apparatus, a protein can stay in the cytoplasm or be exported. In the case of exported (but not secreted) proteins, four possible subcellular locations exist: the inner (plasmidic) membrane, the periplasmic space, the cell wall and the outer membrane. In sequence databases, information on the subcellular location is available for some proteins only. In the case of Gram negative bacteria, this information is given for 5325 proteins sequences among 16 561 (32%) in SWISS-PROT 38 [1]. It is then interesting to have a general and simple method able to predict the location when this information is not known.

Multivariate statistics are particularly adapted to study compositional data in proteins. For example, correspondence analysis (CA) was employed to determine trends in amino acids usage in *Escherichia coli* [2]. Co-inertia analysis has been used to examine amino-acid physico-chemical properties and protein composition [3]. Discriminant analysis (DA) has served to determinate protein secondary structural segments [4], to differentiate intracellular and extracellular proteins [5], and to detect membrane-spanning proteins [6].

* Corresponding author. Tel.: +33-472-44-62-96; fax: +33-478-89-27-19

E-mail address: perriere@biomserv.univ-lyon1.fr (G. Perrière).

0169-2607/02/$ - see front matter
PII: S0169-2607(02)00011-1

Correspondence discriminant analysis (CDA) is a method that can be used on frequency tables while the classical DA is limited to quantitative variables. So CDA can easily be employed with sequence data such as codon or amino acid frequencies tables. In a previous study, we used CDA to predict the subcellular location of *E. coli* proteins divided into three classes: cytoplasmic, periplasmic, and integral membrane proteins [7]. The good results obtained convinced us to extend the use of this method to all Gram negative bacteria.

2. Correspondence discriminant analysis

2.1. General presentation

CDA is a peculiar case of the duality diagram [8,9]: a triplet (Z, M, N) is made of an n by p data table Z, a matrix M defining an Euclidean metric in the subject space $E = \mathbb{R}^p$, and a matrix N defining an Euclidean metric in the variable space $F = \mathbb{R}^n$. From this we deduce by matrix diagonalization four families of vectors with several optimality properties (Fig. 1). We use this diagram in the following peculiar case: $X = [x_{ij}]$ is a contingency table with q (proteins) lines and p (amino acids) columns. Notations employed are the classical ones of CA:

$$x_{\cdot\cdot} = \sum_{i=1}^{q} \sum_{j=1}^{p} x_{ij} \tag{1}$$

$$f_{ij} = \frac{x_{ij}}{x_{\cdot\cdot}} \tag{2}$$

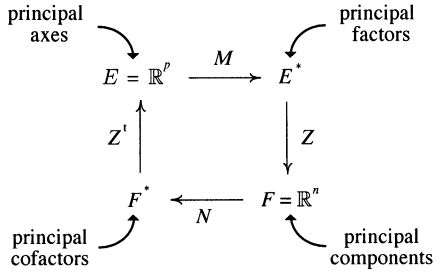

Fig. 1. Duality diagram of the triplet (Z, M, N). Here E^* and F^* are the dual spaces of E and F.

$$F = [f_{ij}] \tag{3}$$

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij} \tag{4}$$

$$f_{\cdot j} = \sum_{i=1}^{q} f_{ij} \tag{5}$$

$$D_q = \mathrm{Diag}(f_{1\cdot}, \ldots, f_{q\cdot}) \tag{6}$$

$$D_p = \mathrm{Diag}(f_{\cdot 1}, \ldots, f_{\cdot p}) \tag{7}$$

The q lines are split into n classes (here $n = 3$) and we have the contingency table Y derived by adding the lines from X by classes:

$$Y = [y_{kj}]_{1 \le k \le n} \quad \text{with} \quad y_{kj} = \sum_{i/Cl(i)=k} x_{ij} \tag{8}$$

The notation $i/Cl(i) = k$ means that the line (protein) i belongs to the class k. We have:

$$y_{\cdot\cdot} = \sum_{k=1}^{n} \sum_{j=1}^{p} y_{kj} = x_{\cdot\cdot} \tag{9}$$

$$g_{kj} = \frac{y_{kj}}{y_{\cdot\cdot}} \tag{10}$$

$$G = [g_{kj}] \tag{11}$$

$$g_{k\cdot} = \sum_{j=1}^{p} g_{kj} \tag{12}$$

$$g_{\cdot j} = \sum_{k=1}^{n} g_{kj} = f_{\cdot j} \tag{13}$$

$$D_n = \mathrm{Diag}(g_{1\cdot}, \ldots, g_{n\cdot}) \tag{14}$$

The triplet (Z, M, N) is then defined by:

$$Z = D_n^{-1} G - U_{np} D_p \quad \text{with} \quad U_{np} = [1] \tag{15}$$

At line k and column j the general term of Z is:

$$Z_{kj} = \frac{g_{kj}}{g_{k\cdot}} - g_{\cdot j} \tag{16}$$

which is the difference between the frequency of amino acid j in the class k and its total frequency.

$$M = (F^t D_q^{-1} F)^{-1} \tag{17}$$

$$N = D_n \tag{18}$$

The duality diagram corresponding to this triplet is shown on Fig. 2. The principal factors define a M^{-1}-orthonormal basis of eigenvectors from matrix MZ^tNZ. Vector 1_p is in the kernel of Z because:

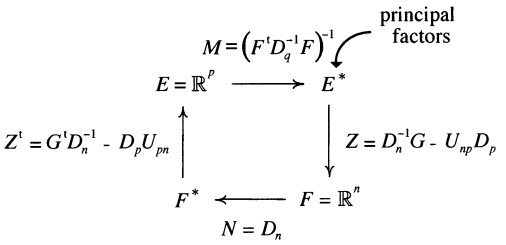

Fig. 2. Duality diagram of the triplet $(D_n^{-1}G - U_{np}D_p, (F^tD_q^{-1}F)^{-1}, D_n)$.

$$Z1_p = D_n^{-1}G1_p - U_{np}D_p1_p = 1_n - 1_n = 0 \tag{19}$$

Principal factors (u) are numerical scores of F columns that are orthogonal to 1_p. They are centered because:

$$u^t((F^tD_q^{-1}F)^{-1})^{-1}1_p = u^tF^tD_q^{-1}F1_p = u^tF^t1_q = u^t\begin{bmatrix} g_{\cdot 1} \\ \vdots \\ g_{\cdot p} \end{bmatrix} = 0 \tag{20}$$

They are also M^{-1}-normed, so we have:

$$u^t((F^tD_q^{-1}F)^{-1})^{-1}u = 1 = (u^tF^tD_q^{-1})D_q(D_q^{-1}Fu) \tag{21}$$

They are numerical scores, centered by amino acids, for which the means by lines of F are normed and, as they are centered, we have:

$$(u^tF^tD_q^{-1})D_q1_q = u^tF^t1_q = 0 \tag{22}$$

Their variances are equal to 1 and they maximize:

$$\|Zu\|_N^2 = \|(D_n^{-1}G - U_{np}D_p)u\|_{D_n}^2 \tag{23}$$

Due to the centering:

$$U_{np}D_pu = 0 \tag{24}$$

$$\|Zu\|_N^2 = \|D_n^{-1}Gu\|_{D_n}^2 = (u^tG^tD_n^{-1})D_n(D_n^{-1}Gu) \tag{25}$$

$D_n^{-1}Gu$ contains the means by classes of factor scores. This vector is centered because:

$$u^tG^tD_n^{-1}D_n1_n = u^tG^t1_n = u^t\begin{bmatrix} g_{\cdot 1} \\ \vdots \\ g_{\cdot p} \end{bmatrix} = 0 \tag{26}$$

and the square of its norm is its variance for D_n weighting. As a consequence, the first principal factor verifies the relationship presented in Fig. 3.

To perform the computations we can note that:

$$F^tD_q^{-1}F = W\Lambda W^t \tag{27}$$

$$(F^tD_q^{-1}F)^{-1} = W\Lambda^{-1}W^t \tag{28}$$

We diagonalize:

$$\Lambda^{-1/2}W^tZ^tD_nZW\Lambda^{-1/2} = V\Lambda V^t \tag{29}$$

and then the factors are obtained by $W\Lambda^{-1/2}V$.

2.2. Practical use

CDA can be computed with modules included in the ADE-4 package devoted to multivariate statistics [10]. This package runs on micro-computers under MacOS (7.1 or higher) and Windows (95 or higher) operating systems. It may be downloaded from the Pôle Bio-Informatique Lyonnais (PBIL) World-Wide Web server at http://pbil.univ-lyon1.fr/ADE-4. The modules required to perform CDA are ADETrans, COA, CategVar and Discrimin. Note that an online version is also implemented on the server [11].

The first step is the computation of a CA on the table containing the amino acid absolute frequencies with the COA module. After that, the discrimination itself is computed and tested with the Discrimin module. To estimate the significance of the discrimination the program computes a Monte-Carlo test. This test consists of repeated random permutations of lines between the classes followed by a recomputation of the CDA. Once the factor scores for the lines and the columns have been obtained, the Discrimin module can also be used to introduce supplementary individu-

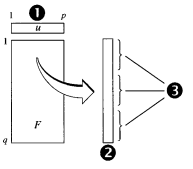

Fig. 3. Examination scheme of the analysis. 1, Columns (amino acids) scores centered for D_p. 2, Mean by lines from F with variance equal to 1. 3, Mean by classes with maximized variance.

als into the analysis by averaging. To do this, we use the relationship:

$$v_i(k) = \frac{x_{..}}{x_{i.}} \sum_{j=1}^{p} \frac{u_j(k) x_{ij}}{x_{.j}} \quad (30)$$

where $v_i(k)$ and $u_j(k)$ are respectively the score for an individual (protein) i and the score for a column (amino acid) j on factor k. Thanks to this option it is possible to estimate the reliability of the assignment using a test set for which the belonging of the individuals to the predefined classes is known. The proportion of the individuals from the test set that are correctly classified will give the accuracy of the prediction.

3. Data set

To establish our data set we have used the release 38 of SWISS-PROT structured with the ACNUC sequence database management system [12]. The advantage provided by SWISS-PROT is the fact that almost all exact redundancies have been removed so that there is no risk to introduce biases due to sequence duplications. In our data set we have discarded hypothetical proteins, partial proteins, proteins with less than 50 amino acids, proteins without any indication of their subcellular location and proteins for which the subcellular location was unsure (potential, putative or obtained by similarity).

To study the subcellular location of proteins from Gram negative bacteria we have refined our selection by retaining only cytoplasmic proteins, integral membrane proteins and periplasmic proteins. We have not integrated proteins from the cell wall because their number was too low (only nine in SWISS-PROT 38). We have not distinguished inner membrane proteins from outer membrane proteins because this information was almost never available. Another problem would have been the fact that amino acid composition of these two kinds of proteins was extremely similar and so, CDA would have been unable to distinguish them. Proteins anchored in the inner membrane were not considered because it was often not possible to determine if the unanchored region of the protein laid on the periplasmic or in the cytoplasmic side of the membrane.

Table 1
Breakdown between species for the data set

Species	*N*	%
Escherichia coli	509	28.2
Salmonella typhimurium	121	6.7
Haemophilus influenzae	107	5.9
Synechocystis sp.	70	3.9
Helicobacter pylori	53	2.9
Pseudomonas aeruginosa	45	2.5
Treponema pallidum	35	1.9
Thermus aquaticus	33	1.8
Paracoccus denitrificans	27	1.5
Klebsiella pneumoniae	26	1.4
Others	782	43.3
Total	1808	100.0

Our selection contained 1808 proteins with 850 (47%) cytoplasmic proteins, 665 (36.8%) integral membrane proteins, and 293 (16.2%) periplasmic proteins. These 1808 proteins came from 201 different species, the top ten species representing 56.7% of the total (Table 1). The percentage of integral membrane proteins was higher than expected (10%, following an estimation in *E. coli* [2]). This over-representation was due to the fact that the location of this kind of proteins is more often documented in the database than the others. The final selection was randomly split into two parts, one analysis set (with a total of 362 114 amino acids) and one test set, with equal numbers (± 1) of individuals for each of the three classes. The CDA itself was performed on the analysis set and the measure of the discrimination accuracy was computed on the test set. Both sets can be downloaded at ftp://pbil.univ-lyon1.fr/pub/datasets/CMPB2001/.

4. Results

The map obtained by crossing the two factors of the CDA performed on our analysis set showed that the first factor separates the integral membrane proteins from the cytoplasmic and periplasmic proteins, while the second factor separates the periplasmic proteins from the cytoplasmic and integral membrane proteins (Fig. 4). On the first

factor, the mean of the scores obtained by integral membrane proteins was −1.060 (S.D. = 0.754) and the mean of the scores obtained by the other proteins was 0.577 (S.D. = 0.629). Also, when compared to the two other classes, integral membrane proteins showed a higher variance for their factor scores (0.569 instead of 0.310 for cytoplasmic proteins and 0.399 for periplasmic proteins). The resulting cutoff value to separate the two groups was −0.167. On the second axis mean of the scores obtained by periplasmic proteins was −1.699 (S.D. = 0.937) and the mean of the factor scores obtained by the other proteins was 0.268 (S.D. = 0.829). The corresponding cutoff value was −0.655.

Looking at the discriminant power of amino acids (Table 2) we can say that the amino acids with positive values on both factors discriminate cytoplasmic proteins (Arg, Glu, His). Amino acids with negative values on the first factor and positive scores on the second factor discriminate integral membrane proteins (Phe, Leu, Ile). Finally, amino acids with slightly positive values on the first factor and negative on the second factor discriminate periplasmic proteins (Asn, Pro, Gln, Thr).

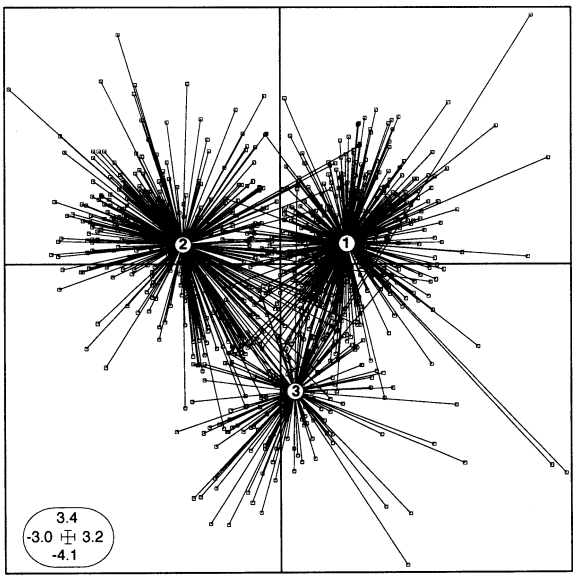

Fig. 4. Factor map of the two discriminant axes of the analysis on the 904 proteins from our analysis set. Each protein is represented by a square linked by a line to the gravity center of the group it belongs to (1, cytoplasmic; 2, integral membrane; 3, periplasmic).

With the averaging formula given in Section 2.2, it is possible to project supplementary individuals in the analysis and, using the thresholds shown above, to predict the subcellular location of a given protein. Factor scores for the amino acids on the two axes and an example of projection of supplementary individual are shown in Table 2. For the separation between integral membrane proteins and the group containing cytoplasmic and periplasmic proteins, accuracy of the prediction on our test set was 88.2%. For the separation between cytoplasmic proteins and the group containing periplasmic and integral membrane proteins, accuracy of the prediction on our test set was 85.2%.

5. Discussion

The results obtained in the discrimination of proteins from Gram negative bacteria following their subcellular location confirm and extend our previous results on *E. coli* [7]. The discrimination of integral membrane proteins by amino acids like Phe, Leu and Ile is not surprising as these amino acids are known to be hydrophobic. Also, discrimination of the cytoplasmic proteins by Arg, Glu and His can be easily explained as these three amino acids are charged and hydrophilic and so are required in soluble proteins. At last, discrimination of periplasmic proteins by Asn, Pro, Gln and Thr, Asn and Gln can be explained by the fact these amino acids are known to slow protein folding [5], and slow folding is required in exported proteins.

The higher variance observed for the factor scores of the integral membrane proteins on the first axis means that amino acid composition is more variable in this class of proteins. This is probably due to the fact that these proteins have regions of variable length not inserted in the membrane and these regions contain non-hydrophobic amino acids.

The good separation between the different classes on the two axes indicates that CDA can be

Table 2
Factor scores and discriminant power for the amino acids and example of averaging for a test protein from *E. coli*

	$u_j(1)$	$u_j(2)$	$d_j(1)$	$d_j(2)$	x_{ij}	$x_{\cdot j}$	$v_{ij}(1)$	$v_{ij}(2)$
Arg	0.1502	0.2423	0.1848	0.1714	17	17 932	0.1302	0.2100
Leu	−0.3233	0.5793	−0.1270	0.1084	45	36 479	−0.3646	0.6534
Ser	−0.3494	−0.3040	−0.1134	−0.0222	32	20 168	−0.5070	−0.4411
Thr	−0.2155	−0.0992	−0.0322	−0.0683	17	19 367	−0.1730	−0.0796
Pro	0.3583	−1.3684	0.0649	−0.0961	17	15 857	0.3512	−1.3415
Ala	−0.3243	−0.1922	−0.0791	−0.0399	52	35 627	−0.4328	−0.2565
Gly	−0.1870	0.0249	−0.0766	−0.0138	28	29 376	−0.1630	0.0217
Val	−0.3930	0.2492	−0.0373	0.0300	29	27 493	−0.3791	0.2403
Lys	0.2215	−1.3552	0.2367	−0.2289	6	17 933	0.0678	−0.4146
Asn	−0.1117	−0.2529	−0.0018	−0.1595	10	13 279	−0.0769	−0.1742
Gln	−0.1558	−0.3920	0.0395	−0.0895	7	13 733	−0.0726	−0.1827
His	0.3227	0.0111	0.1489	0.0485	7	7794	0.2651	0.0091
Glu	0.9552	1.5223	0.3345	0.1326	7	21 208	0.2883	0.4595
Asp	0.8754	0.0362	0.2429	−0.1032	5	18 600	0.2152	0.0089
Tyr	−0.1700	0.1507	−0.0530	−0.0550	10	10 728	−0.1449	0.1285
Cys	0.0815	0.0671	0.1789	−0.0308	4	3473	0.0858	0.0706
Phe	−0.1738	0.4322	−0.2063	0.0758	37	15 910	−0.3696	0.9191
Ile	−0.0813	1.0229	−0.0858	0.1307	34	22 037	−0.1148	1.4431
Met	−0.2522	−0.2954	−0.1236	0.0245	23	9886	−0.5366	−0.6284
Trp	−0.2274	−0.0787	−0.3626	−0.0336	9	5234	−0.3576	−0.1237
Total					396	362 114	−2.2890	0.5220

In this table column $u_j(k)$ contains the factor scores, $d_j(k)$ the discriminant power values, x_{ij} the amino acid frequencies of *E. coli* protein BCR_ECOLI (P28246), $x_{\cdot j}$ the amino acid frequencies in the whole analysis set. For the last two columns we have $v_{ij}(k) = \frac{x_{\cdot\cdot} u_j(k) x_{ij}}{x_{i\cdot} x_{\cdot j}}$.

So the respective sum of these columns gives the score of the protein on the two factors of the analysis. Due to these scores, and taking into account the thresholds obtained, we can classify BCR_ECOLI as an integral membrane protein.

used to predict the subcellular locations of proteins in Gram negative bacteria when this information is not known. The best method available for discriminating integral membrane proteins from globular soluble proteins is a DA of protein sequence characteristics (such as maximum local hydrophobicity), and its accuracy is 95% [6]. The resolution of our method is a bit lower (88.2%), and this difference is partly due to the fact that we used only amino acid frequencies to discriminate these proteins. In Gram negative bacteria, the best method already published to identify periplasmic proteins is an expert system using amino acid composition [13]. This system can identify these proteins with a reliability of 83% and our analysis gives equivalent results (85.2%) using a much simpler computation.

Even if our method gives results with an accuracy only equivalent to what was published before it has the advantage of producing a discrimination between the three classes in a single analysis, as the preceding methods only separate proteins into two groups (integral membrane proteins versus soluble globular proteins and periplasmic proteins versus other proteins). So it is difficult to directly compare the results obtained by these methods and those obtained by CDA. Another point is that the analyses previously published [6,13] were done on much smaller sets of proteins (respectively 102 and 106 proteins instead of 904), and it is not sure that these methods would produce similar accuracies with present day protein data.

Lastly, CDA is a much more general method than the two cited above which are limited to very

specific protein studies. Indeed CDA could be used in any study for which genes or proteins are classified into predefined groups. For instance, it could be used to see which codons are specific of putatively horizontally transferred genes in some bacterial species, as it has been proposed that these genes have a codon usage that differs from the average use in these organisms [14,15]. Also, as codon usage in vertebrates is heterogeneous and varies greatly depending on the region of the genome studied [16], CDA can be used to find the linear combination of codons discriminating genes belonging to different classes. Indeed 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. Similarly CDA could be used to compare amino acid usage in orthologous genes of different organisms.

Acknowledgements

Thanks are due to Manolo Gouy for his helpful comments and careful reading of the manuscript and to Daniel Chessel for his help on CDA mathematical basis.

References

[1] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.

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

[3] J. Thioulouse, J.R. Lobry, Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package, Comput. Appl. Biosci. 11 (1995) 321–329.

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

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

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

[7] G. Perrière, J.R. Lobry, J. Thioulouse, Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences, Comput. Appl. Biosci. 12 (1996) 519–524.

[8] M. Tenhenhaus, F.W. Young, An analysis and synthesis of multiple correspondence analysis, optimal scaling and other methods for quantifying categorical multivariate data, Psychometrika 50 (1985) 91–119.

[9] Y. Escoufier, The duality diagram: a means of better practical applications, in: P. Legendre, L. Legendre (Eds.), Developments in Numerical Ecology, NATO Advanced Institute, Springer Verlag, Berlin, 1987, pp. 139–156.

[10] J. Thioulouse, D. Chessel, S. Dolédec, J.M. Olivier, ADE-4: a multivariate analysis and graphical display software, Stat. Comput. 7 (1997) 75–83.

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

[12] M. Gouy, C. Gautier, M. Attimonelli, C. Lanave, G. di Paola, ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage, Comput. Appl. Biosci. 1 (1985) 167–172.

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

[14] C. Médigue, T. Rouxel, P. Vigier, A. Hénaut, A. Danchin, Evidence for horizontal gene transfer in *Escherichia coli* speciation, J. Mol. Biol. 222 (1991) 851–856.

[15] J.G. Lawrence, H. Ochman, Molecular archaeology of the *Escherichia coli* genome, Proc. Natl. Acad. Sci. USA 95 (1998) 9413–9417.

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