On-line tools for sequence retrieval and multivariate statistics in molecular biology

Guy Perrière* and Jean Thioulouse

Laboratoire de Biométrie, Génétique et Biologie des Populations
URA CNRS nº 2055
Université Claude Bernard – Lyon 1
43, bd. du 11 Novembre 1918
69622 Villeurbanne Cedex (France)

Phone: +33 72-44-80-00 (ext. 33-14)
Fax: +33 78-89-27-19
Email: perriere@biomserv.univ-lyon1.fr, thioulou@biomserv.univ-lyon1.fr

Keywords: World-Wide Web; Sequence data banks; Retrieval system; Multivariate analysis; Sequence analysis.

* To whom reprint requests should be sent.

Abstract

We have developed a World-Wide Web server for browsing sequence collections structured under ACNUC format and for performing multivariate analyses on sequences. General collections (like GenBank or EMBL), as well as specialized data banks (like Hovergen and NRSub) can be accessed. This system allows to build complex queries, and the result of each query, represented by a list of sequences, is stored on the server. It is then possible to re-use this list to compute multivariate analyses on the sequences. Two examples of applications are shown. The first one consists in a study of codon usage with correspondence analysis on all the protein genes of *Haemophilus influenzae* Rd. This study allows to identify the highly expressed genes and the integral membrane proteins of this organism. The second one consists in an ordination of 70 aligned protein sequences of growth hormone with principal coordinate analysis. With this method, we are able to re-establish the patterns of relationships between the sequences previously determined with tree building programs.

Introduction

Among the multitude of methods available for sequence analysis, multivariate statistics have been extremely useful since fifteen years. For instance, they have been used for examining codon composition in protein genes (Grantham and Gautier, 1980; Grantham *et al.*, 1980; Holm, 1986; Sharp *et al.*, 1986; Shields and Sharp, 1987; Shields *et al.*, 1988; Médigue *et al.*, 1991), for predicting CoDing Sequences (CDS) localization (Fichant and Gautier, 1987), for studying the evolution of repetitive elements in primates and rodents (Quentin, 1988, 1989), for complementing phylogenies (Higgins, 1992), and for determining trends in amino-acid composition (Lobry and Gautier, 1994; Thioulouse and Lobry, 1995). Many software packages include multivariate analysis methods (*e.g.*, SAS, SYSTAT, BMDP, or S+), but none of them is really adapted for the study of biological sequences. Thus, use of these methods by molecular biologists is limited as the user often has to develop his own interfaces to link the multivariate analysis programs with sequence libraries and his data files. This is why we have decided to develop a World-Wide Web (WWW) server allowing to access easily multivariate methods that are suited for molecular biology studies. Mainly these methods are COrrespondence Analysis (COA) and Principal COordinate analysis (PCO).

As the goal of this server is the use of multivariate statistics for sequence analysis, it has to provide an access to the sequence data banks. To that purpose we have developed WWW-Query, a WWW version of the retrieval system Query (Gouy *et al.*, 1985). As its standalone predecessor, WWW-Query allows to make complex queries on various sequence banks structured under ACNUC format and to save the result of a query into a list. This list is stored on the server and can be re-used in further queries or by on-line programs. With some methods, the user also has the possibility to work on his own data set, as an interface for transferring sequences to the server has been implemented. All the steps of an analysis, from the selection of the sequences to the visualization of the graphics, can be realized on-line. In the same time, the user has always the possibility to download his results from the server to complete a study with his own software.

System and methods

ACNUC

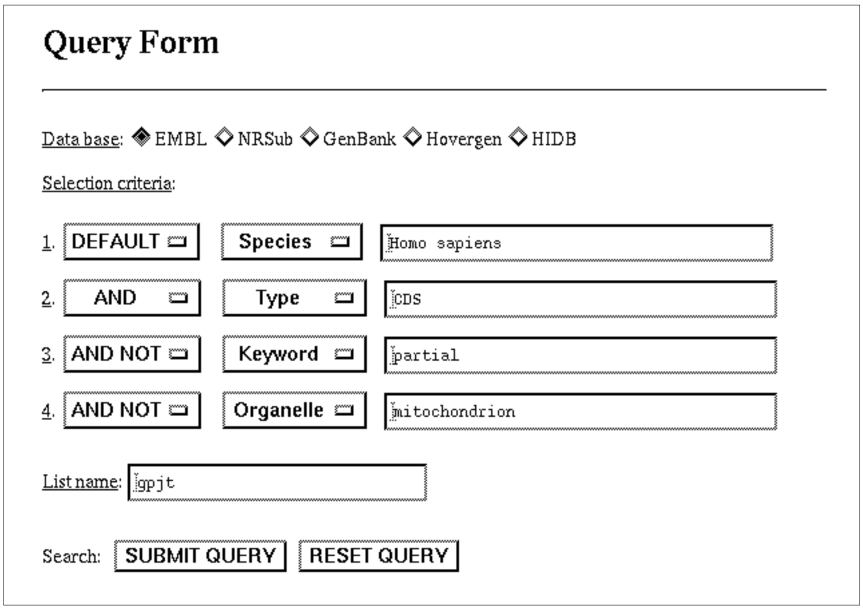
The first layer of the system is represented by a set of five sequence data banks indexed with ACNUC: EMBL (Emmert *et al.*, 1994), GenBank (Benson *et al.*, 1994), Hovergen (Duret *et al.*, 1994), NRSub (Perrière *et al.*, 1994) and HIDB (Fleischmann *et al.*, 1995). In fact, all the data banks in the EMBL, GenBank/DDBJ, or NBRF/PIR (George *et al.*, 1994) formats could be indexed with ACNUC and thereafter, put on the server. EMBL and GenBank are immediately updated after annoucement of the availability of new releases. Also, as Hovergen and NRSub are maintained by people working in our group, these data banks are permanently kept up-to-date.

Indexing with ACNUC allows to query the banks using many criteria to retrieve sequences. It is possible to use mnemonics, accession numbers, keywords, taxonomic data, bibliographic references, dates of insertion in the bank, the nature of the genome from which a sequence has been obtained (nuclear, mitochondrial, chloroplastic or kinetoplastic), the nature of the molecule sequenced (DNA, mRNA, tRNA, etc.) A great advantage provided by ACNUC is the notion of subsequence: each kind of genomic fragment described in the features can be defined as a subsequence and then retrieved independently from its parent sequence. Existence of subsequences allows to compose query for retrieving only CDS, tRNAs, rRNAs, snRNAs, and so on. ACNUC also includes the useful concept of lists. Each time a query is issued, its result is stored into a sequence list. It is then possible to use a previously defined list in further queries, and so to progressively refine the scope of a search.

A set of C and Fortran programs have been developed to interface ACNUC data bases with the WWW. These programs decode the queries sent by the users, retrieve the sequences matching a query in a given data base, and then write the pages containing hypertext links toward the sequences or the sequences themselves. Some routines write cross-references with other data banks as hypertext links. For instance cross-references with Medline, Swiss-Prot (Bairoch and Boeckmann, 1994), and Enzyme (Bairoch, 1994) are implemented. In the case of Hovergen, it is possible to use specialized helper applications to visualize the alignments and the phylogenetic trees that are important elements of this data base. Binaries ot these applications for Sun Sparc, IBM RS/6000, DEC Alpha and Macintosh computers are available at URL (Uniform Resource Locator) ftp://biom3.univ-lyon1.fr/pub/hovergen/helpers/.

User interface

The user interface is written in HTML (HyperText Markup Language). This language includes the definition of special tags that allow the creation of simple elements like buttons, selectors and editable text fields. The interface consists in a set of HTML pages allowing to build queries and to access the different functionalities of the server. The main entry point is represented by a form allowing to compose a query on a given data base, using up to four criteria (Figure 1). Each criterion is connected to the previous ones by a logical operator and four different operators are available: AND, OR, AND NOT, OR NOT. The user may give a name to the list that will be created after the submission of the query to the server. If he does not give a list name, the system automatically uses “list” by default. This list, as well as the other files created by the users are stored on our server at URL ftp://biom3.univ-lyon1.fr/pub/NetMul/data. This directory is cleaned up each Friday at 0.00 AM WET (Western European Time) and the limit set on the disk quota is 10 Mb.



Query Form

Data base: ◆ EMBL ◇ NRSub ◇ GenBank ◇ Hovergen ◇ HIDB

Selection criteria:

1. DEFAULT Species Homo sapiens
2. AND Type CDS
3. AND NOT Keyword partial
4. AND NOT Organelle mitochondrion

List name: gpjt

Figure 1. *WWW-Query HTML page for composing queries. The first field of the form allows to select a data base to query. The four following fields are used for criteria definition. For each of these fields, two selectors allow to define the logical operator used in the connection with preceding criterion (default: “AND”), and choose the nature of the criterion (default: “Keyword”). A line editor allows to type the criterion itself. The last field is another line editor in which the user enters the name of the list associated to the query (default name: “list”). In this example, the query built allows to retrieve all the non-partial, non-mitochondrial (i.e., nuclear) CDS from human in EMBL.*

Multivariate methods are accessed through HTML documents giving detailed explanations on how to perform an analysis. Realization of a COA requires four different steps: (i) the selection of a set of sequences using WWW-Query, (ii) the computation of the codons or amino-acids frequencies table, (iii) the computation of the COA itself, (iv) the visualization of the factor map. In the case of PCO, it is necessary to send to server a set of aligned nucleotide or protein sequences in the Mase format (Faulkner and Jurka, 1988). As our system does not provides tools for aligning sequences, the user has to align his sequences with a program like Clustal W (Thompson *et al.*, 1994). After that, he has to paste them onto a dedicated submission form. As not all WWW browsers are able to manage forms and associated copy/paste utilities, the user has the option to send his data using anonymous FTP. This kind of transfer is also more adapted for large data sets.

NetMul

The multivariate methods provided are taken from the NetMul library. This library is a subset of the ADE (Analysis of Environmental Data) package (Thioulouse *et al*., 1995). Two types of methods suited for sequence analysis are accessible through NetMul: COA for contingency tables (Williams, 1952; Benzécri, 1973), and PCO for distance matrices (Gower, 1966; Manly, 1994). In the case of COA, several interpretations help can be used after an analysis have been performed: inertia analysis for rows and columns (Greenacre, 1984), additional (or supplementary) rows and/or columns, and data reconstitution (Lebart *et al.*, 1984). Inertia analysis describes the importance of each element (row or column) in the construction of the factors, and reciprocally, the percentage of variability taken into account by each factor in the description of rows and columns. Additional elements are helpful when one wants to project into an analysis supplementary rows (or columns), *i.e.*, elements that were not used in the computations of the analysis. This can be used for confirmation, or as a validation step. The data reconstitution option allows to model the data table starting from the information taken into account by only some of the factors (the first one, the first two ones, etc.) It can be used to study which factor accounts for which feature in the data table.

Graphical programs

Two programs from public domain packages are used to draw graphics. These programs are gnuplot, from the GNU package, and ppmtogif, from the PBM (Portable BitMap) package. Gnuplot is able to draw plots in the PBM format, using coordinates files. The problem is that PBM is not usually recognized by WWW browsers, so we use ppmtogif to convert PBM files into GIF (Graphics Interchange Format) files, the format commonly used by WWW.

Algorithms

The algorithms used in the NetMul library are described in the ADE documentation (Thioulouse *et al.*, 1995). The source code of the computational parts of the programs are written in ANSI C, so they were just copied from the ADE package and re-compiled. In the case of PCO, the distances needed by this method are computed with a separate program. For computing these distances, the identity matrix is perfectly adequate for DNA or RNA sequences. However, with proteins, it is common to weight differentially the distance between pairs of amino-acids. To do so, we use the amino-acid distance matrix published by Smith and Smith (1990) for scoring the distances between sequences. This matrix has the peculiarity to be Euclidean, which is a required condition to produce distances usable with PCO (Higgins, 1992). Note that our distances computing program exclude all the positions with a gap in any sequence of the alignment, this in a way to produce distances that are perfectly Euclidean.

Implementation

The retrieval system WWW-Query and all the ACNUC data bases are set up on a Sun SparcServer 1000. The NetMul library itself is installed on a Sun SparcStation 2. The server can be reached at URL http://acnuc.univ-lyon1.fr/. All the sources of the programs and the scripts used are available upon request from the corresponding author by sending an Email message to the address: perriere@biomserv.univ-lyon1.fr.

System performance

As the sequence data banks are indexed with ACNUC, queries are fast, even when using collections like EMBL or GenBank. For example, a query on the server allowing to retrieve all the 8573 non-partial nuclear CDS from human in EMBL release 44 (September 1995) takes 34 sec. More difficult queries are the ones associating use of wildcards, parent sequences and subsequences. Multivariate analysis computations are also very fast, even with large data sets. With a 1680 × 61 table created with the codon frequencies of all the CDS from *Haemophilus influenzae* genome, COA computation takes around 25 sec. Note that these values are given only for information as they do not take into account lag time due to network transfer which is extremely variable.

Example of use of COA

It is well-known that, in unicellular organisms, protein coding sequences have a nonrandom usage of synonymous codons (see Andersson and Kurland, (1990) for review). Particularly, genes that are highly expressed have a strong codon bias (Gouy and Gautier, 1982; Holm, 1986; Sharp *et al*., 1986). An easy way to detect these genes is COA. Up to now, this approach has been used successfully for *Escherichia coli* (Holm, 1986; Médigue *et al.*, 1991), *Bacillus subtilis* (Shields and Sharp, 1987; Sharp *et al*., 1990; Perrière *et al*., 1994; Moszer *et al.*, 1995), *Saccharomyces cerevisiae* (Sharp *et al*., 1986), *Drosophila melanogaster* (Shields *et al*., 1988), and *Cænorhabditis elegans* (Stenico *et al.*, 1994). In this example, we have studied codon usage in *H. influenzae* Rd bacterium. Indeed, this organism is the first bacterium to have been completely sequenced (Fleischmann *et al.*, 1995). Due to that fact, it is the first time that it is possible to make a complete study on codon usage in a given organism.

As the complete genome of *H. influenza* is available as a single entry on our server, it is very easy to retrieve all its CDS with WWW-Query. On the 1680 genes found in this genome we have first determined the absolute codon frequencies, then we have computed the COA on the resulting table. The first factor of the analysis accounts for 11.5% of total variability while the second and third factor account respectively for 7% and 5.8% of the variability. The plot of the two firsts axis is given in Figure 2. The first factor discriminates the genes following their supposed expression level. In the leftmost part of the graph, we find genes coding mainly for ribosomal proteins, elements involved in translation and enzymes of the glycolysis (*i.e.*, putatively highly expressed genes). The second factor discriminates the genes following the hydropathic character of the encoded proteins. In the cloud of points located at the bottom of the graph, we find genes coding for proteins that have a very high Gravy score (Kyte and Doolittle, 1982). It means that these genes probably encode integral membrane proteins. At last, the third factor is highly correlated with the global G+C content of the genes.

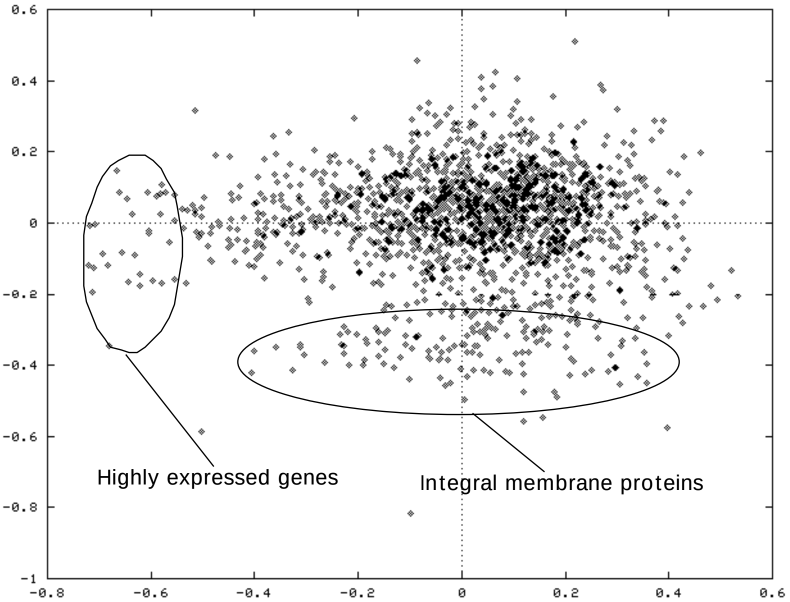


Figure 2. *Factor map of the two first factors of a COA realized on the codons of all the CDS from H. influenzae Rd genome. We have added to the graph return by the server two circles to show the location of the highly expressed genes and of the genes coding for integral membrane proteins.*

With the labelling option of our plotting tool, it is possible to number the dots on the graph and then, to identify which genes are the highly expressed ones. After that, we can establish the codon relative adaptiveness table (Sharp and Li, 1987) for *H. influenzae* (Table 1). For that purpose, we have used the codon frequencies of the 30 CDS with the lowest scores on the first axis of COA. Among these CDS, 29 correspond to well-identified genes: *eno*, *gapH*, *mopB*, *mopI*, *ompA*, *pal*, *pnp*, *rplA*, *rplB*, *rplC*, *rplG*, *rplI*, *rplK*, *rplM*, *rplO*, *rplP*, *rplS*, *rpmA*, *rpmF*, *rpmG*, *rpmH*, *rpsI*, *rpsK*, *rpsOA*, *rpsOB*, *strA*, *tufA*, *tufB*, and *tsf*. One corresponds to a still non-dentified gene which is labelled HI0965 by the authors. With the codon relative adaptiveness table it is then possible to compute the values of the Codon Adaptation Index (CAI) for all *H. influenzae* genes.

Table I. *Absolute frequencies (X_i) and relative adaptiveness (W_i) values for codons in H. influenzae. This table has been computed using the sequences of 30 putatively highly expressed genes from this organism. Note that some codons were not found in the sequences of our data set. In this case we have assigned a value of 0.5 to any X_i that would otherwise be zero.*

		X_i	W_i			X_i	W_i
Arg	AGA	5	0.0159744		GUC	14	0.0600858
	AGG	0	0.0015974		GUG	74	0.3175965
	CGA	3	0.0095846		GUU	233	1.0000000
	CGC	45	0.1437699				
	CGG	0.5	0.0015974	Lys	AAA	423	1.0000000
	CGU	313	1.0000000		AAG	23	0.0543735
Leu	CUA	16	0.0540540	Asn	AAC	156	1.0000000
	CUC	1	0.0033783		AAU	58	0.3717948
	CUG	1	0.0033783				
	CUU	55	0.1858108	Gln	CAA	175	1.0000000
	UUA	296	1.0000000		CAG	8	0.0457142
	UUG	21	0.0709459				
				His	CAC	93	1.0000000
Ser	AGC	30	0.2255639		CAU	27	0.2903225
	AGU	16	0.1203007				
	UCA	72	0.5413533	Glu	GAA	354	1.0000000
	UCC	1	0.0075187		GAG	35	0.0988700
	UCG	1	0.0075187				
	UCU	133	1.0000000	Asp	GAC	96	0.5581395
					GAU	172	1.0000000
Thr	ACA	79	0.3558558				
	ACC	39	0.1756756	Tyr	UAC	85	1.0000000
	ACG	17	0.0765765		UAU	49	0.5764705
	ACU	222	1.0000000				
				Cys	UGC	7	0.3500000
Pro	CCA	125	1.0000000		UGU	20	1.0000000
	CCC	0.5	0.0040000				
	CCG	26	0.2080000	Phe	UUC	119	1.0000000
	CCU	40	0.3200000		UUU	51	0.4285714
Ala	GCA	354	1.0000000	Ile	AUA	2	0.0089686
	GCC	12	0.0338983		AUC	223	1.0000000
	GCG	95	0.2683615		AUU	123	0.5515695
	GCU	180	0.5084745				
				Met	AUG	136	1.0000000
Gly	GGA	11	0.0257009				
	GGC	110	0.2570093	Trp	UGG	20	1.0000000
	GGG	1	0.0023364				
	GGU	428	1.0000000	Ter	UAA	-	-
					UAG	-	-
Val	GUA	223	0.9570815		UGA	-	-

Total number of codons: 5781

Example of use of PCO

PCO can be used to complement phylogenetic analysis conduced with more classical techniques (Higgins, 1992). To use this method on our server, it is necessary to have a set of pre-aligned sequences in the Mase format. In this example, we have directly used the Mase file from the Hovergen data base that contains the protein sequences belonging to the growth hormone family. In the release 13 of Hovergen (July 1995), this file contained 70 aligned protein sequences. After determination of the distances using the similarity matrix of Smith and Smith (1990), we have computed the PCO iself. In the resulting analysis, the first factor accounts for 17.2% of the variability and the second and third factors account respectively for only 5.4% and 4.3% of the variability.

The plot of the two first axis of the analysis is shown on Figure 3. We can see immediately that the first factor separates mammals from fishes and the second factor separates primates from other mammals. In the primates two sub-groups are distinguished: Human and Macaca. Between the different species of mammals, there is no clear separation. Inside the group of fishes, three sub-groups corresponding to the *Salmonidae*, *Cyprinidae* and *Percoidei* are found.

Two species of fishes are apart from the others on the map: *Anguilla japonica* and *Amia alva*. What is interesting is that the sequences of growth hormone in these two fishes are close to the ones of frog and xenopus and this relation is not apparent on the Hovergen tree. It is not surprising that the three main groups are well separated, as this information can be seen on the tree for this gene family. In fact, all the major trends of the phylogeny are visible on the map built with the two first axes. Less visible are the relationships between the closely related sequences as they are often hidden in the minor axes of the PCO.

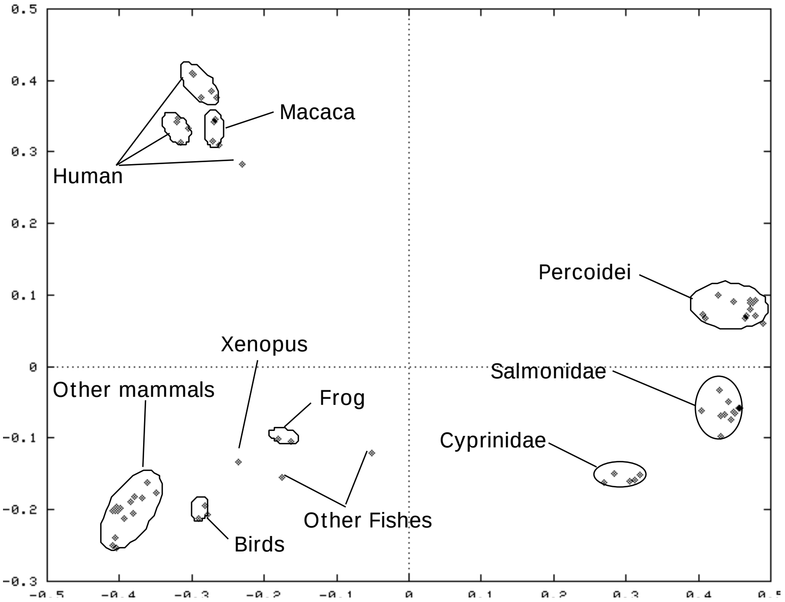


Figure 3. *Factor map of the two first factors of a PCO realized on 70 growth hormone protein sequences from the Hovergen data base. For illustration purpose we have added arrows and circles to indicate the groups that correspond to the observed clusters in the map.*

Discussion

The retrieval system WWW-Query is a convenient and flexible system for selecting sequences through WWW. Classical features of WWW sequence banks browsers like SRS (Sequence Retrieval System) (Etzold and Argos, 1993a, 1993b) are integrated: possibility to build multi-criteria queries, implementation of cross-references between different data banks as hypertext links, access not only to the parent sequence but also to the different fragments of biological interest, etc. It provides an original feature that is not available in other systems: the notion of permanent list. As each query result is stored on the server, a user can still re-use it until the weekly cleaning of the directory. Also, if a user wants to keep his list for a longer time, he can get it by FTP, and thereafter upload it back to the server. A limitation of WWW-Query is that it is not possible for a user to retrieve all the sequence entries that belong to a list. Each entry has to be transferred one by one using an HTML page containing links toward the sequences themselves. Indeed, in the case of selections containing tens of thousands of sequences, transfer of all the entries would led to extremely long delays and would contribute to the general overload of the network. This problem is also shared by the other WWW retrieval systems. Another limitation is that WWW-Query cannot query simultaneously many data bases as it is possible to do with SRS.

The main use of COA on our server is the identification of genes with high codon bias, but this method can be used for other purposes. For instance, it has been claimed to be able to identify possibly horizontally transferred genes in *E. coli* (Médigue *et al.*, 1991). When applied on amino-acid frequencies, COA allows to detect integral membrane proteins (Lobry and Gautier, 1994). The main problem, when using this method on sequences from the general repository collections, is that it is difficult to get rid of sequence redundancies. As a growing number of non-redundant data bases are developed, a first solution consist in the progressive installation of these data bases on our server. Also, as the sequencing projects produce now extremely long genomic fragments (*e.g.*, 1.6 Mb in the case of *E. coli*), another solution consists in the use of only one very large fragment to do the analysis. The last method, PCO, is useful for obtaining a view of a phylogeny which is more easy to read than a tree, particularly when there are many species considered. This method is powerful in the presence of long branches as it is often difficult to interpret these features. In the end, user has always the possibility to build a tree from the factorial coordinates as some reconstitution methods exist (Thioulouse *et al*., 1995).

An important problem lie in the fact that the interpretation of multivariate analysis is mainly graphical. The ADE package includes graphical modules for that purpose but these programs cannot be employed due to the present limitations of HTML and WWW. The solution of using programs like gnuplot and GIF files is only temporary as a new generation of WWW browsers is coming (see for example the HotJava™ system distributed by Sun Microsystems). Indeed these browsers provide the possibility to include not only text, pictures, movies and sounds in HTML documents, but also multi-platforms object code. With such browsers it will be soon possible to develop portable graphical applications immediately uploadable through WWW.

References

Benson, D.A., Boguski, M., Lipman, D.J. and Ostell, J. (1994) GenBank. *Nucleic Acids Res.*, **22**, 3441-3444.

Bairoch, A. and Boeckmann, B. (1994) The Swiss-Prot protein sequence data bank: current status. *Nucleic Acids Res.*, **22**, 3578-3580.

Bairoch, A. (1994) The Enzyme data bank. *Nucleic Acids Res.*, **22**, 3626-3627.

Benzécri, J.P. (1973) *L'analyse des Correspondances. L'analyse des Données, Vol. 2.* Bordas, Paris.

Duret, L., Mouchiroud, D. and Gouy, M. (1994) Hovergen: a database of homologuous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360-2365.

Emmert, D.B., Stoehr, P.J., Stoesser, G. and Cameron, G.N. (1994) The European Bioinformatic Institute (EBI) databases. *Nucleic Acids Res.*, **22**, 3445-3449.

Etzold, T. and Argos, P. (1993a) SRS – an indexing and retrieval tool for flat file data libraries. *Comput. Applic. Biosci.*, **9**, 49-57.

Etzold, T. and Argos, P. (1993b) Transforming a set of biological flat file libraries to a fast access network. *Comput. Applic. Biosci.*, **9**, 59-64.

Faulkner, D.V. and Jurka, J. (1988) Multiple aligned sequence editor (MASE). *Trends Biochem. Sci.*, **13**, 321-322.

Fichant, G. and Gautier, C. (1987) Statistical methods for predicting protein coding regions in nucleic acids sequences. *Comput. Applic. Biosci.*, **3**, 287-295.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C.A., Gocayne, J.D., Scott, J.D., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter, J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496-512.

George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1994) The PIR-international protein sequence database. *Nucleic Acids Res.*, **22**, 3569-3573.

Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055-7073.

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

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-328.

Grantham, R. and Gautier, C. (1980) Genetic distances from mRNA sequences. *Naturwissenschaften*, **67**, 93-94.

Grantham, R., Gautier, C. and Gouy, M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate base according to genome type. *Nucleic Acids Res.*, **8**, 1892-1912.

Greenacre, M. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Higgins, D.G. (1992) Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput. Applic. Biosci.*, **8**, 15-22.

Holm, L. (1986) Codon usage and gene expression. *Nucleic Acids Res.*, **14**, 3075-3087.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105-132.

Lebart, L., Morineau, L. and Warwick, K.M. (1984) *Multivariate Descriptive Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley and Sons, New York.

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.

Manly, B.F. (1994). *Multivariate Statistical Methods. A Primer*. Chapman and Hall, London.

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

Moszer, I., Glaser, P. and Danchin, A. (1995) SubtiList: a relational data base for the *Bacillus subtilis* genome. *Microbiology*, **141**, 261-268.

Perrière, G., Gouy, M. and Gojobori, T. (1994) NRSub: a non-redundant data base for the *Bacillus subtilis* genome. *Nucleic Acids Res.*, **25**, 5525-5529.

Quentin, Y. (1988) The Alu family developed through successive waves of fixation closely connected with primate lineage history. *J. Mol. Evol.*, **27**, 194-202.

Quentin, Y. (1989) Successive waves of fixation of B1 variants in rodent lineage history. *J. Mol. Evol.*, **28**, 299-305.

Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, 5125-5143.

Sharp, P.M. and Li, W.-H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281-1295.

Sharp, P.M., Higgins, D.G., Shields, D.C., Devine, K.M. and Hoch, J.A. (1990) *Bacillus subtilis* gene sequences. In Zukowski, M.M., Ganesan, A.T. and Hoch, J.A. (eds), *Genetics and Biotechnology of Bacilli*. Academic Press, San Diego, pp. 89-98.

Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.*, **15**, 8023-8040.

Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.*, **5**, 704-716.

Smith, R.F. and Smith, T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, **87**, 118-122.

Stenico, M., Lloyd, A.T. and Sharp, P.M. (1994) Codon usage in *Cænorhabditis elegans*: delineation of translational selection and mutation biases. *Nucleic Acids Res.*, **22**, 2437-2446.

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 editor, Bolonia, pp. 57-62.

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

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.

Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**, 274-289.