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

Guy Perrière and Jean Thioulouse

Abstract

We have developed a World-Wide Web server for browsing sequence collections structured under the 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 complex queries to be constructed, and the result of each query, represented by a list of sequences, is stored on the server. It is then possible to reuse 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 the highly expressed genes and the integral membrane proteins of this organism to be identified. The second one consists in an ordering 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 for the last 15 years. For instance, these statistics 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 users often have to develop their own interfaces to link the multivariate analysis programs with sequence libraries and their data files. This is why we have decided to develop a World-Wide Web (WWW) server that allows users to access easily multivariate methods suited for molecular biology studies. These methods are mainly 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 access to sequence data banks. For that purpose we have developed WWW-Query, a WWW version of the retrieval system Query (Gouy et al., 1985). As with its standalone predecessor, WWW-Query allows users to make complex queries on various sequence banks structured under the ACNUC format and to save the result of a query into a list. This list is stored on the server and can be reused in further queries or by on-line programs. With some methods, users also have the possibility of working on their own data sets, 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. At the same time, users always have the option of downloading their results from the server to complete a study with their 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 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 the announcement of the availability of new releases. Furthermore, 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 users 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 kinetoplastid), or the nature of the molecule sequenced (DNA, mRNA, tRNA, etc.) A great advantage of 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. The existence of subsequences allows users to compose queries for retrieving only CDS, tRNAs, rRNAs, snRNAs, etc. ACNUC also includes the useful concept of lists. Each time a query is issued, its result is stored in a sequence list. It is then possible to use a previously defined list in further queries, and so to refine progressively the scope of a search.

A set of C and FORTRAN programs have been developed to interface ACNUC databases with the WWW. These programs decode the queries sent by the users, retrieve the sequences matching a query in a given database, and then write the pages containing hypertext links toward the sequences or the sequences themselves. Some routines write cross-references with other databases as hypertext links. For instance cross-references with Medline and SWISS-PROT (Bairoch and Boekmann, 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 database. Binaries of these applications for Sun SPARC, IBM RS/6000, DEC Alpha and Macintosh computers are available at URL ftp://biom3.univ-lyon1.fr/pub/hovergen/helpers/.

User interface

The user interface is written in HTML. 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 of a set of HTML pages that allow users to build queries and to access the different functionalities of the server. The main entry point is represented by a form that allows users to compose a query on a given database, 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. Users may give a name to the list that will be created after the submission of the query to the server. If a list name is not given, 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 cleared each Friday at 00.00 a.m. Western European Time and the limit set on the disk quota is 10 Mbytes.

![Query Form](http://bioinformatics.oxfordjournals.org/)
Multivariate methods are accessed through HTML documents that give detailed explanations on how to perform an analysis. Realization of a COA requires four different steps: (i) selection of a set of sequences using WWW-Query, (ii) computation of the codon or amino acid frequency table, (iii) computation of the COA itself, and (iv) visualization of the factor map. In the case of PCO, it is necessary to send to the server a set of aligned nucleotide or protein sequences in the Mase format (Faulkner and Jurka, 1988). As our system does not provide tools for aligning sequences, users have to align their sequences with a program like ClustalW (Thompson et al., 1994). After that, they have to paste them onto a dedicated submission form. As not all WWW browsers are able to manage forms and associated copy/paste utilities, users have the option of sending their data using anonymous FTP. This kind of transfer is also better suited 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 can be used after an analysis has 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 computation of the analysis. This can be used for confirmation, or as a validation step. The data reconstitution option allows one to model the data table starting from the information taken into account by only some of the factors (the first one, the first two, 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 package. gnuplot is able to draw plots in the PBM format, using files of coordinates. The problem is that PBM is not usually recognized by WWW browsers, so we use ppmtogif to convert PBM files into GIF files, the format commonly used on the 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, and hence were just copied from the ADE package and recompiled. 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 this, we use the amino acid distance matrix published by Smith and Smith (1990) for scoring the distances between sequences. This matrix has the peculiarity of being Euclidean, which is a required condition to produce distances usable with PCO (Higgins, 1992). Note that our distance computing program excludes all the positions with a gap in any sequence of the alignment, and hence produces distances that are perfectly Euclidean.

**Implementation**

The retrieval system WWW-Query and all the ACNUC databases are set up on a Sun SPARCStation 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 first author by sending an e-mail message to the address: perriere@biom-serv.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 to retrieve all the 8573 non-partial nuclear human CDSs in EMBL release 44 (September 1995) takes 34 s. Queries that use wildcards, parent sequences and subsequences are more difficult. 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 CDSs from *Haemophilus influenzae* genome, COA computation takes ~25 s. 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 non-random array of synonymous
codons (see Andersson and Kurland, 1990) for review). In particular, 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 *Caenorhabditis elegans* (Stenico et al., 1994). In this example, we have studied codon usage in *H. influenzae* bacterium. Indeed, this organism is the first bacterium to have been completely sequenced (Fleischmann et al., 1995), and because of this, it is the first time that it has been 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 CDSs with WWW-Query. For the 1680 genes found in this genome we have first determined the absolute codon frequencies, and have then 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 factors account for 7% and 5.8% respectively of the variability. The plot of the first two axes is given in Figure 2. The first factor discriminates the genes following their supposed expression level. On the leftmost part of the graph, we find genes coding mainly for ribosomal proteins, elements involved in translation and glycolytic enzymes (i.e. putatively highly expressed genes). The second factor discriminates the genes following the hydrophobic 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. Finally, the third factor is highly correlated with the global G + C content of the genes.

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 I). For that purpose, we have used the codon frequencies of the 30 CDSs with the lowest scores on the first axis of the COA. Among these CDSs, 29 correspond to well-identified genes: eno, gapH, mopB, mopI, mopA, pal, pnp, rplA, rplB, rplC, rplG, rplH, rplK, rplM, rplO, rpsL, rpmA, rpmF, rpmG, rpmH, rpsl, rpsK, rpsOA, rpsOB, strA, tufA, tufB and tsf. One corresponds to a still unidentified gene, which has been labelled H10965 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.

**Example of the use of PCO**

PCO can be used to complement phylogenetic analysis performed with more classical techniques (Higgins, 1992). To use this method on our server, it is necessary to have a set of prealigned sequences in the Mase format. In this example, we have directly used the Mase file from the
Table I. Absolute frequencies \( (X_i) \) and relative adaptiveness \( (W_i) \) values for codons in \( H.\text{influenzae} \)

<table>
<thead>
<tr>
<th></th>
<th>( X_i )</th>
<th>( W_i )</th>
<th></th>
<th>( X_i )</th>
<th>( W_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>AGA</td>
<td>5</td>
<td>0.0159744</td>
<td>GUC</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td>0</td>
<td>0.0015974</td>
<td>GUG</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>3</td>
<td>0.0095846</td>
<td>GUU</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>CGC</td>
<td>45</td>
<td>0.1437699</td>
<td>Lys</td>
<td>AAA</td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>0.5</td>
<td>0.0015974</td>
<td>AAG</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>CGU</td>
<td>313</td>
<td>1.0000000</td>
<td>Leu</td>
<td>CUA</td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>1</td>
<td>0.0033783</td>
<td>Asn</td>
<td>AAC</td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td>1</td>
<td>0.0033783</td>
<td>AAU</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>CUU</td>
<td>55</td>
<td>0.1858108</td>
<td>Gln</td>
<td>CAA</td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td>296</td>
<td>1.0000000</td>
<td>CAG</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>21</td>
<td>0.0709459</td>
<td>His</td>
<td>CAC</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
<td>30</td>
<td>0.2255639</td>
<td>Ser</td>
<td>CAU</td>
</tr>
<tr>
<td></td>
<td>AGU</td>
<td>16</td>
<td>0.1203007</td>
<td>UCG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>72</td>
<td>0.5413533</td>
<td>Gln</td>
<td>GAA</td>
</tr>
<tr>
<td></td>
<td>UCC</td>
<td>1</td>
<td>0.0075187</td>
<td>Glu</td>
<td>GAG</td>
</tr>
<tr>
<td></td>
<td>UCG</td>
<td>1</td>
<td>0.0075187</td>
<td>Asp</td>
<td>GAC</td>
</tr>
<tr>
<td></td>
<td>UCU</td>
<td>133</td>
<td>1.0000000</td>
<td>Met</td>
<td>AUG</td>
</tr>
<tr>
<td>Thr</td>
<td>ACA</td>
<td>79</td>
<td>0.3558558</td>
<td>Thr</td>
<td>UAC</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>39</td>
<td>0.1756576</td>
<td>Tyr</td>
<td>UAG</td>
</tr>
<tr>
<td></td>
<td>ACG</td>
<td>17</td>
<td>0.0763765</td>
<td>UAA</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>ACU</td>
<td>222</td>
<td>1.0000000</td>
<td>Pro</td>
<td>CUA</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.5</td>
<td>0.0040000</td>
<td>Pro</td>
<td>CGA</td>
</tr>
<tr>
<td></td>
<td>CCG</td>
<td>26</td>
<td>0.2080000</td>
<td>Phe</td>
<td>UUC</td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>40</td>
<td>0.3200000</td>
<td>Phe</td>
<td>UUU</td>
</tr>
<tr>
<td></td>
<td>Ala</td>
<td>GCA</td>
<td>125</td>
<td>1.0000000</td>
<td>Phe</td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td>12</td>
<td>0.0338983</td>
<td>Ile</td>
<td>AUA</td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>95</td>
<td>0.2683615</td>
<td>Ile</td>
<td>AUC</td>
</tr>
<tr>
<td></td>
<td>GCU</td>
<td>180</td>
<td>0.5084745</td>
<td>Val</td>
<td>GUA</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
<td>11</td>
<td>0.0251709</td>
<td>Gly</td>
<td>GAU</td>
</tr>
<tr>
<td></td>
<td>GGC</td>
<td>110</td>
<td>0.2570993</td>
<td>Trp</td>
<td>UGG</td>
</tr>
<tr>
<td></td>
<td>GGG</td>
<td>1</td>
<td>0.0023364</td>
<td>Trp</td>
<td>UGG</td>
</tr>
<tr>
<td></td>
<td>GUU</td>
<td>428</td>
<td>1.0000000</td>
<td>Gln</td>
<td>CAG</td>
</tr>
<tr>
<td></td>
<td>GGU</td>
<td>1</td>
<td>0.0023364</td>
<td>Gln</td>
<td>CAG</td>
</tr>
</tbody>
</table>

Total number of codons: 5781.
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.

Hovergen database that contains the protein sequences belonging to the growth hormone family. In 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 itself. In the resulting analysis, the first factor accounts for 17.2% of the variability and the second and third factors account for only 5.4% and 4.3% respectively of the variability.

The plot of the two first axes 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 subgroups are distinguished: human and macaca. Between the different species of mammals, there is no clear separation. Inside the group of fishes, three subgroups corresponding to the Salmonidae, Cyprinidae and Percoidae 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 those 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.

Discussion

The retrieval system WWW-Query is a convenient and flexible system for selecting sequences through the WWW. Classical features of WWW sequence bank browsers like SRS (Sequence Retrieval System) (Etzold and Argos, 1993ab) are integrated: the possibility of building multicriteria 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. WWW-Query provides an original feature that is not available in other systems: the notion of a permanent list. As each query result is stored on the server, a user can still reuse it until the weekly cleaning of the directory. Furthermore, if users want to keep their lists for longer, they can get them by FTP, and thereafter upload them 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 to the sequences themselves. Indeed, in the case of selections containing tens of thousands of sequences, transfer of all the entries would lead 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 as many databases 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 to amino acid frequencies, COA allows integral membrane proteins (Lobry and Gautier, 1994) to be detected. 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 databases are developed, a first solution is the progressive installation of these databases on our server. Furthermore, as sequencing projects now produce extremely long genomic fragments (e.g. 1.6 Mbases in the case of *E. coli*), another solution is 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 easier to read than a tree, particularly when there are many species to be considered. This method is powerful in the presence of long branches as it is often difficult to interpret these features. In the end, users always have the option to build a tree from factorial coordinates as some reconstitution methods exist (Thioulouse et al., 1995).

An important problem lies 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

Fig. 3. Factor map of the two first factors of a PCO realized on 70 growth hormone protein sequences from the Hovergen database. For illustration purposes we have added arrows and circles to indicate the groups that correspond to the observed clusters in the map.
the present limitations of HTML and the WWW. The solution of using programs like gnuplot and GIF files is only temporary as a new generation of WWW browsers is in the pipeline (see, for example, the HotJava™ system distributed by Sun Microsystems). Indeed, these new browsers offer the possibility of including not only text, pictures, movies and sounds in HTML documents, but also multi-platform object code. With such browsers it will be soon possible to develop portable graphical applications that are immediately uploadable through the WWW.

References