Remote access to ACNUC nucleotide and protein sequence databases at PBIL

Manolo Gouy, Stephane Delmotte

Remote access to ACNUC nucleotide and protein sequence databases at PBIL

Manolo Gouy[*1] and Stéphane Delmotte[1]

[1]Laboratoire de Biométrie et Biologie Evolutive ; UMR CNRS 5558 ; Université de Lyon ; Université Lyon 1 ; 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

[*]Corresponding author: Manolo Gouy; Tel: +33 4 72 43 12 87; Fax: +33 4 72 43 13 88; mgouy@biomserv.univ-lyon1.fr

Abstract

The ACNUC biological sequence database system provides powerful and fast query and extraction capabilities to a variety of nucleotide and protein sequence databases. The collection of ACNUC databases served by the Pôle Bio-Informatique Lyonnais includes the EMBL, GenBank, RefSeq and UniProt nucleotide and protein sequence databases and a series of other sequence databases that support comparative genomics analyses: HOVERGEN and HOGENOM containing families of homologous protein–coding genes from vertebrate and prokaryotic genomes, respectively; Ensembl and Genome Reviews for analyses of prokaryotic and of selected eukaryotic genomes. This report describes the main features of the ACNUC system and the access to ACNUC databases from any internet-connected computer. Such access was made possible by the definition of a remote ACNUC access protocol and the implementation of Application Programming Interfaces between the C, Python and R languages and this communication protocol. Two retrieval programs for ACNUC databases, Query_win, with a graphical user interface and raa_query, with a command line interface, are also described. Altogether, these bioinformatics tools provide users with either ready-to-use means of querying remote sequence databases through a variety of selection criteria, or a simple way to endow application programs with an extensive access to these databases. Remote access to ACNUC databases is open to all and fully documented (http://pbil.univ-lyon1.fr/databases/acnuc/acnuc.html).

Keywords: ACNUC; bioinformatics; nucleotide sequence database; database retrieval.

Abbreviations: Application Programming Interface (API); European Bioinformatics Institute (EBI); National Center for Biotechnology Information (NCBI); Pôle Bio-Informatique Lyonnais (PBIL).

1. **Introduction**

Nucleotide and protein sequence databases are key resources for today's biological sciences. Generalist databases (EMBL, GenBank, UniProt) gather all public biological sequence data. In addition, a number of more specialized databases have been constructed by various institutions and laboratories that supplement the annotations present in generalist databases or organize data differently, for example by grouping homologous sequences, or by presenting uninterrupted sequences at the chromosome scale. All of these databases can be queried using a web browser connected to various web sites, including EBI, NCBI, ExPASy. Several programmable interfaces for more elaborate uses than what is possible through a browser are also available (e.g., Entrez's E-Utilities [1]). However, accessing collectively all of these resources is often a difficult task because each tends to require specific software tools. A conspicuous exception to this difficulty is the SRS system, which provides a unified model and interface to hundreds of databases [2]. We present here a simple and efficient bioinformatics resource that both structures a variety of sequence databases in a single model, the ACNUC model, and allows very fast access to all of them using the same software tools from any internet-connected computer. Remote access to ACNUC databases is open to all and fully documented (http://pbil.univ-lyon1.fr/databases/acnuc/acnuc.html).

2. Results

2.1 Logical and physical structures

The ACNUC biological sequence database system has been designed in order to allow most structured fields of sequence annotations to be used as potential entry points in the database and to be combined in complex queries (Fig. 1). A single database model was conceived to accommodate both nucleotide and protein sequences and the three flat file formats they use, namely the EMBL, UniProt and GenBank formats. Nucleotide sequences generally contain regions having markedly distinct biological function (e.g., protein coding regions, introns, transposable elements, regulatory elements, untranscribed regions). The ACNUC nucleotide database structure was therefore organized around two major entities, parent sequences and sub-sequences. Parent sequences are contiguous pieces of sequenced DNA; sub-sequences are fragments, or series of fragments, of one or more parent sequences or of their complementary strands, that have a defined biological function (e.g., protein-coding region, ribosomal RNA gene). Sub-sequences thus model the organization of the genetic information encoded in a nucleotide sequence and allow direct access to sequence fragments of a given function. Practically, sub-sequences are created by entries of sequence

feature tables. The list of feature keys that create a sub-sequence is a parameter of an ACNUC database. For the large, generalist databases EMBL and GenBank this list is currently set to CDS, rRNA, tRNA, snRNA, scRNA and misc_RNA, that is, to regions that encode proteins and structural RNAs. Other ACNUC databases may create sub-sequences differently. In HOVERGEN and HOGENOM, two databases of homologous protein-coding genes that use the ACNUC system (see 2.3 below), the list of sub-sequence-creating feature keys contains in addition 5'NCR, 3'NCR, 5'INT, INT_INT, and 3'INT, to allow direct access to 5' and 3' non coding transcribed regions and to introns, divided in first, internal and last introns. Each sub-sequence is given a name obtained by adding an extension to the name of the parent sequence whose feature table creates it. The extension uses the value of the '/gene=' feature qualifier when present or is automatically generated when absent (Fig. 1). Sub-sequences are not implemented in the ACNUC protein sequence database model.

ACNUC retrieval criteria match either parent or sub-sequences depending on whether a criterion value applies to all of a parent sequence or to a specific sub-sequence. For example, taxonomic criteria operate at the parent sequence level; keywords generally operate at the sub-sequence level because they apply to a specific feature entry; the ACNUC 'type' criterion allows specifying a desired feature table key, CDS for example, directs the query to the sub-sequence level and matches only sub-sequences associated to the specified feature key.

Several pieces of information occur in diverse parts of sequence annotations besides KW/KEYWORD lines, which can adequately be modelled as sequence keywords. For example, the value of the /EC_number= feature qualifier fits as a keyword attached to the sub-sequence created by the corresponding feature key. Table I describes all additional keywords that ACNUC databases attach to parent or sub-sequences. In UniProt, DE lines are both information-rich and loosely structured. ACNUC indexing extracts from these lines several keywords, each attached to the processed sequence (Fig. 2).

ACNUC databases contain two tree structures. Species names are tree-organized according to the biological classification of species developed by the National Center for Biotechnology Information (NCBI) [1]. Taxa of any level can thus be used as selection criteria. Keywords are also tree organized, but in a very sparse way, that is, the depth of the keyword tree rarely exceeds two, and only a small number of keywords get assigned a parent tree node. This feature allows to logically connect series of keywords sharing a common property. Table I lists all such cases. For example, the ACNUC indexing procedure creates for each Enzyme Commission number a keyword placed in the keyword tree under parent node

'EC_NUMBERS' and links it to all corresponding sequences. The query language (see below) can thus be used to list all EC number keywords attached to a protein. The ACNUC logical structure has been previously presented [3] and has remained unchanged.

An *ad hoc* physical structure was defined to implement the ACNUC logical structure. This structure is in essence as previously presented [3], although several changes were applied to accommodate the huge size increase of sequence databases since 1985. ACNUC databases are now implemented through 14 index files that supplement the flat files distributed by the database producing institutions, e.g., the European Bioinformatics Institute (EBI) for EMBL and UniProt, the National Center for Biotechnology Information (NCBI) for GenBank. ACNUC index files are efficient in terms of volume: they account for 7 % of the volume of the EMBL nucleotide sequence database and for 27 % of the smaller and with higher annotation/sequence ratio UniProt protein sequence database. This structure has been highly used for a number of years and has proven to perform well in terms of query possibilities and data access speed. The structure of index files, useful for advanced uses of ACNUC Application Programming Interfaces, is fully documented (http://pbil.univ-lyon1.fr/databases/acnuc/structure.html).

2.2 The ACNUC query language

The ACNUC query language allows expression and logical combination of retrieval criteria to elaborate arbitrarily complex queries. Each query creates the list of all matching database elements. Most retrieval criteria produce lists of sequences, but the query language also allows creation of lists of taxa and of keywords. Each list is given a name, either automatically or by the user. The query language allows handling a large number of lists, each of any size. The full series of ACNUC retrieval criteria appears in Table II.

Query language operators (Table III) are used to logically combine elementary retrieval criteria, and to derive useful lists of taxa or of keywords from sequence lists or vice versa. Parentheses are used to specify operator ranges. Three operators (AND, OR, NOT) can, in rare instances, be ambiguous because they can also occur within valid criterion values; bracketing elementary selection criteria between double quotes is used to solve these ambiguities. For example,

```
"sp=Beak and feather disease virus" and "au=ritchie"
```

specifies which 'and' belongs to a criterion value and which one is an operator. The query language is case insensitive, except where filenames occur.

2.3 Publicly available ACNUC sequence databases

The collection of ACNUC databases served by the Pôle Bio-Informatique Lyonnais (PBIL) [4] includes the EMBL [5], GenBank [6], RefSeq [7], Genome Reviews [8] and UniProt [9] nucleotide and protein sequence databases, and other sequence databases that support comparative genomics analyses: HOVERGEN [10] and HOGENOM [11] containing families of homologous protein–coding genes from vertebrate and prokaryotic genomes, respectively; Ensembl [12] for analyses of selected eukaryotic genomes. The 'Whole Genome Shotgun' section of the EMBL sequence library is served as a distinct ACNUC database due to its extremely large size. All of these are structured following the single data model described above; the software tools described below can query all of them. The EMBL and GenBank generalist nucleotide sequence databases are indexed at each full release and nightly supplemented with updates distributed by the EBI and the NCBI. UniProt is weekly updated, that is, as often as possible. Ensembl is indexed at each full release. The PBIL aims at delivering yearly releases of HOVERGEN and HOGENOM, databases requiring very large computation for each release. Other databases (e.g., RefSeq) do not have a fixed updating policy at PBIL.

2.4 Architecture of the client/server system

Network access to ACNUC databases has been achieved by the definition and implementation of a remote ACNUC access protocol that governs information exchanges between the PBIL and remote clients. This protocol uses a TCP-IP socket connection to a server running at PBIL and makes retrieval operations to the above described collection of ACNUC databases possible, with usual academic internet connections, at a speed very close to that obtained when locally working. Therefore, any internet-connected computer can run as an ACNUC client. The remote ACNUC server can simultaneously handle several clients querying the same or distinct databases. A series of commands and their arguments were defined that allow 1) database opening, 2) query execution, 3) annotation and sequence display, 4) annotation, species and keywords browsing, and 5) sequence extraction. The protocol also allows clients to receive the list of served databases and to password-protect any database with a challenge-reply encryption mechanism that does not transfer password data on the network. The sequence extraction command can optionally send all data in gzip-compressed format to increase transfer speed. Several potentially time-consuming operations have been defined as client-interruptible, a useful property for implementation of the remote

ACNUC retrieval programs described below. The server software, a single C program, can also be installed on any UNIX system that would serve copies of PBIL's ACNUC databases. The remote ACNUC access protocol is documented (http://pbil.univ-lyon1.fr/databases/acnuc/ remote_acnuc.html). All bioinformatics resources presented below make use of this protocol.

2.5 The Query_win and raa_query retrieval programs

Two client ACNUC retrieval programs have been developed. Query_win uses the Vibrant toolkit [13] to perform remote ACNUC retrieval operations with a graphical user interface; raa_query performs the same operations with a line interface, and is therefore useful in a scripting context, possibly to repeatedly execute fixed retrieval operations. The Query_win program is freely available as executable files for the five most common computer operating systems (http://pbil.univ-lyon1.fr/software/query_win.html). In all cases, a single program is installed, so that remote ACNUC access can be started in a few seconds. The raa_query program, written in standard C, is portable to most computers from its freely available source code and is also available as an MSWindows executable program (http://pbil.univ-lyon1.fr/software/query.html). Both programs provide extensive on-line help. The rest of this section describes several Query_win capabilities, all equally possible with the non-graphical raa_query program.

Program Query_win first displays the list of available ACNUC databases and waits for its user to select one (Fig. 3). The major program display is shown in Fig. 4. A species tree browser (menu Browse) facilitates identification of taxonomic names of any level among the over 400,000 names of the species classification. A keywords browser allows pattern matching through the very large keyword lists of sequence databases. Selected sequence data can be extracted to local files in either flat, FASTA or text formats. Five kinds of sequence extractions are possible: 1) selected parent or sub-sequences; 2) protein translation of selected protein-coding sub-sequences using all genetic code and initiation codon information present in annotations; 3) any region flanking the start or the end of selected sub-sequences by specifying coordinates whose meaning is relative to sub-sequence endpoints; 4) any feature key from the feature tables of selected parent sequences; 5) combination of the last two operations to extract regions flanking the endpoints of any feature key. In addition to retrieval criteria presented above, sequence lists can be further refined by pattern matching in user-specified annotation items. A user could, for example, refine a previously defined sequence list by retaining only sequences containing the string '/tissue_type="liver' in the 'source'

items of their feature tables, or only those containing 'IFREMER' in their RL annotation lines.

2.6 The C, Python and R Application Programming Interfaces

The remote ACNUC access protocol has been interfaced with three programming languages, C, Python, and the widely used statistical computing environment R [14]. These language bindings provide application programmers with a simple, complete and fast network access to biological sequence databases. In the C binding, all network operations are handled within the API, so that programming can be conceptually done as if querying a local database. The Python binding was written on top of the C API, to achieve the same effect, and to provide identical capabilities. Both APIs are usable after downloading a single archive file, are fully documented (http://pbil.univ-lyon1.fr/databases/acnuc/raa_acnuc.html, http://pbil.univ-lyon1.fr/cgi-bin/raapythonhelp.csh) and provide all querying, extraction, and browsing capabilities of the ACNUC database model. The Query_win and raa_query retrieval programs use indeed the C API. Fig. 5 gives a simple example of remote ACNUC access with the C API. The C binding also allows concurrent access to several databases, say EMBL and UniProt, although queries operate on a single database at a time.

An R package called seqinR that offers a number of tools for statistical and evolutionary analyses of nucleotide sequences and also access to sequence databases was recently published [15, 16]. This last feature was implemented through use of the remote ACNUC access protocol. In terms of sequence database access, seqinR noticeably contains functions that perform all ACNUC queries (except those involving local file names), that extract through the network annotation and primary sequence data from matching database sequences, and that convert these into R objects suitable for analyses by other functions of the large collection of R packages.

3. Discussion

Access to ACNUC databases has long been possible in two ways, remotely with the WWW-QUERY web resource [17], and locally at the PBIL, with functionalities identical to that of Query_win and of the C API described above. Local access has also been used in various laboratories after full copying of PBIL's ACNUC databases. The bioinformatics tools introduced here allow access to all ACNUC databases from any internet computer with the same capabilities as, and a speed very similar to what is possible through local access. Heavy

copy operations of full databases from PBIL to client laboratories have thus become unnecessary without loss of function. Programs Query_win and raa_query allow extended retrieval operations in comparison to what is possible through the browser-based WWW-QUERY tool. These include all feature table-based extractions, and annotation, keyword, and taxon pattern-matching operations. These tools are also much better suited to simultaneously handle several potentially large sequence lists. Query_win can also be used essentially as easily as a web browser connected to NCBI, EBI, or PBIL web sites to access sequence databases through elementary retrieval criteria because it offers extremely fast access to relevant sequences from any accession number, keyword, author name, or species starting point. The only requirement is to download once the Query_win executable file.

An important feature of the ACNUC model is its coverage of the three major models of biological sequence databases, EMBL, GenBank, and UniProt. The remote ACNUC access thus differs from what is offered by the Entrez system [18], which does not cover EBI-specific resources, e.g., Ensembl and Genome Reviews.

Nightly updates of the EMBL and GenBank databases, and weekly for UniProt, ensure that remote ACNUC users always access comprehensive sequence data sets. Noticeably, these updates are performed at PBIL without service interruptions.

The remote ACNUC access covers fewer databases than the SRS system [2], which is remarkable by the number of databases that it can index and query. However, the SRS system is rather complex in its usage, is limited by the time required to index data volumes of the size of today's nucleotide databases, and has presently uncertain long-term commercial support.

For a single database, the remote ACNUC access protocol is comparable in retrieval capacity to Entrez's E-utilities [1] and to SRS. Access to PBIL's ACNUC databases is thus an alternative possibility opened to developers of sequence analysis software who can use the C, Python, or R bindings presented here. Users of other languages can also access ACNUC databases because the remote ACNUC access protocol is usable by any language able to open TCP-IP sockets; e.g., Perl scripts are currently employed at PBIL to access ACNUC databases. Importantly, the seqinR package [16] that interfaces the wide collection of R packages for statistical analyses with all of ACNUC databases is unique among bioinformatics tools for remote sequence database access.

Acknowledgements

We thank Delphine Charif and Jean R. Lobry for development of the R interface to the remote ACNUC access protocol.

References

[1] Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V., Church D.M., DiCuccio M., Edgar R., Federhen S., Geer L.Y., Kapustin Y., Khovayko O., Landsman D., Lipman D.J., Madden T.L., Maglott D.R., Ostell J., Miller V., Pruitt K.D., Schuler G.D., Sequeira E., Sherry S.T., Sirotkin K., Souvorov A., Starchenko G., Tatusov R.L., Tatusova T.A., Wagner L., Yaschenko E., Database resources of the National Center for Biotechnology Information, Nucleic Acids Research 35 (2007) D5-D12.

[2] Etzold T., Ulyanov A., Argos P., SRS: information retrieval system for molecular biology data banks, Methods Enzymology 266 (1996) 114-128.

[3] Gouy M., Gautier C., Attimonelli M., Lanave C., di Paola G., ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage, Computer Applications Biosciences 1 (1985) 167-172.

[4] Perrière G., Combet C., Penel S., Blanchet C., Thioulouse J., Geourjon C., Grassot J., Charavay C., Gouy M., Duret L., Deleage G., Integrated databanks access and sequence/structure analysis services at the PBIL, Nucleic Acids Res. 31 (2003) 3393-3399.

[5] Kulikova T., Akhtar R., Aldebert P., Althorpe N., Andersson M., Baldwin A., Bates K., Bhattacharyya S., Bower L., Browne P., Castro M., Cochrane G., Duggan K., Eberhardt R., Faruque N., Hoad G., Kanz C., Lee C., Leinonen R., Lin Q., Lombard V., Lopez R., Lorenc D., McWilliam H., Mukherjee G., Nardone F., Garcia-Pastor M.P., Plaister S., Sobhany S., Stoehr P., Vaughan R., Wu D., Zhu W., Apweiler R., EMBL Nucleotide Sequence Database in 2006, Nucleic Acids Res. 35 (2007) D16-D20.

[6] Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L., GenBank, Nucleic Acids Res. 35 (2007) D21-D25.

[7] Pruitt K.D., Tatusova T., Maglott D.R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 35 (2007) D61-D65.

[8] Sterk P., Kersey P.J., Apweiler R., Genome Reviews: standardizing content and representation of information about complete genomes, OMICS 10 (2006) 114-118.

[9] Wu C.H., Apweiler R., Bairoch A., Natale D.A., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Mazumder R., O'Donovan C., Redaschi N., Suzek B., The Universal Protein Resource (UniProt): an expanding universe of protein information, Nucleic Acids Res. 34 (2006) D187-D191.

[10] Duret L., Mouchiroud D., Gouy M., HOVERGEN: a database of homologous vertebrate genes, Nucleic Acids Res. 22 (1994) 2360-2365.

[11] Penel S., Perrière G., Gouy M., Duret L., HOGENOM: Database of Homologous Sequences from Complete Genomes, http://pbil.univ-lyon1.fr/databases/hogenom.html

[12] Hubbard T.J., Aken B.L., Beal K., Ballester B., Caccamo M., Chen Y., Clarke L., Coates G., Cunningham F., Cutts T., Down T., Dyer S.C., Fitzgerald S., Fernandez-Banet J., Graf S., Haider S., Hammond M., Herrero J., Holland R., Howe K., Howe K., Johnson N., Kahari A., Keefe D., Kokocinski F., Kulesha E., Lawson D., Longden I., Melsopp C., Megy K., Meidl P., Ouverdin B., Parker A., Prlic A., Rice S., Rios D., Schuster M., Sealy I., Severin J., Slater G., Smedley D., Spudich G., Trevanion S., Vilella A., Vogel J., White S., Wood M., Cox T., Curwen V., Durbin R., Fernandez-Suarez X.M., Flicek P., Kasprzyk

A., Proctor G., Searle S., Smith J., Ureta-Vidal A., Birney E., Ensembl 2007, Nucleic Acids Res. 35 (2007) D610-D617.

[13] Kans J.A., VIBRANT : Virtual Interface for Biological Research and Technology, (1991) http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/VIBRANT.HTML

[14] Ihaka R.,Gentleman R., R: A language for data analysis and graphics. J. Comp. Graph. Stat. 5 (1996) 299–314.

[15] Charif D., Thioulouse J., Lobry J.R., Perrière G., Online synonymous codon usage analyses with the ade4 and seqinR packages, Bioinformatics 21 (2005) 545-547.

[16] Charif D., Lobry J.R., SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, in: Bastolla U., Porto M., Roman H.E., Vendruscolo M. (Eds.), Structural approaches to sequence evolution: Molecules, networks, populations, Springer Verlag, New York, 2007, in press.

[17] Perrière G., Gouy M., WWW-query: an on-line retrieval system for biological sequence banks, Biochimie 78 (1996) 364-369.

[18] Schuler G.D., Epstein J.A., Ohkawa H., Kans J.A., Entrez: molecular biology database and retrieval system, Methods Enzymology 266 (1996) 141-162.

Table I

Additional keywords created during ACNUC indexing and their place in the keywords tree

Parent keyword	*Keyword*	*Note*
DIVISION NAMES	DIVISION *name*	From ID or LOCUS division information
RELEASE NUMBERS	RELEASE *number*	From EMBL DT lines or *a priori* data for GenBank
CHROMOSOMES	CHROMOSOME *name*	From GenBank ORIGIN lines or /chromosome= feature qualifier
GENETIC NAMES	*name*	From /gene=, /standard_name= feature qualifiers, or UniProt GN lines
EC_NUMBERS	*number*	From /EC_number= feature qualifier, or UniProt DE lines
PRODUCTS	*product_name*	From /product= feature qualifier
PROTEIN IDS	*pid_value*	From /protein_id= feature qualifier
PARTIAL	5'-PARTIAL, 3'-PARTIAL	When < or > present in feature location; from /partial feature qualifier; when '(fragment)' present in UniProt DE lines
SUBCELLULAR LOCATION	*location*	From UniProt 'CC -!- SUBCELLULAR LOCATION:' annotations
MISC_FEATURE	*feature_key*	Any feature key of any feature table
none	*strain_name*	From /strain= feature qualifier
none	*evidence_value*	From /evidence= feature qualifier
none	PSEUDO	When /pseudo feature qualifier present

Table II

Retrieval criteria of the ACNUC query language

Criterion	*Syntax*	*Resulting list*
Taxonomy	SP=*taxon*	Parent sequences attached to taxon of any level from NCBI's taxonomy (possible use of @ as wildcard)
Taxonomy	TID=*id*	Parent sequences from given taxonomic ID number according to NCBI's taxonomy
Host	H=*taxon*	Parent sequences whose host species belongs to taxon (OH lines in EMBL databases; possible use of @ as wildcard)
Keyword	K=*keyword*	Sub- or parent sequences attached to keyword (possible use of @ as wildcard)
Type	T=*type*	Sub-sequences created by feature table entries whose key is equal to given type
Journal	J=*j_name*	Parent sequences from named journal
Reference	R=*reference*	Parent sequences from given reference specified by *jcode/volume/page* (e.g., JMB/13/5432)
Author	AU=*name*	Parent sequences from given author (last name only)
Acc. No.	AC=*access*	Parent sequences from given accession number
Name	N=*name*	Sub- or parent sequence of given name (possible use of @ as wildcard)
L-Taxa	NS=taxon	List of taxa containing given taxon (possible use of @ as wildcard)
L-Keywords	NK=*keyword*	List of keywords containing given keyword (possible use of @ as wildcard)
Year	Y=*year*	Parent sequences published in given year; Y>*year* or Y<*year* can also be used
Organelle	O=*organelle*	Parent sequences from given organelle
Molecule	M=*molecule*	Parent sequences annotated with given molecule in ID/LOCUS
Status	ST=*class*	Parent sequences from specified data class (EMBL) or review level (UniProt)
File of names	F=*filename*	Sub- or parent sequences named in given file, one per line
File of acc nos	FA=*filename*	Parent sequences whose accession nos. are in given file, one per line
File of keywords	FK=*filename*	List of keywords named in given file, one per line
File of taxa	FS=*filename*	List of taxa named in given file, one per line
Previous list	*list_name*	Previously-built named list of sequences, taxa or keywords

Table III

Operations of the ACNUC query language

Operation	*Resulting list*
list1 AND *list2*	Elements common to *list1* and *list2*
list1 OR *list2*	Union of *list1* and *list2* elements
NOT *list*	All database elements not in *list*
PAR *list*	Sub-sequences from *list* are replaced by their parent sequences
SUB *list*	Sub-sequences of parent sequences present in *list* are added to resulting list
PS *list*	List of taxa attached to sequences present in *list*
PK *list*	List of keywords attached to sequences present in *list*
UN *list*	Sequences attached to elements of *list* that are either taxon names or keywords
SD *list*	All species placed in the tree below members of the taxon list operand
KD *list*	All keywords placed in the tree below members of the keyword list operand

Figure legends

Figure 1

Right panel: a typical EMBL nucleotide sequence whose annotation elements usable as ACNUC retrieval criterion are highlighted in red colour. The left panel illustrates uses of the ACNUC query language that would match this sequence by retrieving some of these annotation elements. The single CDS feature entry therein corresponds to an ACNUC sub-sequence named D13109.RPOB.

Figure 2

Example of processing of UniProt DE lines. Resulting keywords are highlighted in bold characters.

Figure 3

List of publicly available ACNUC databases as displayed by program Query_win. Other database names are accepted, among which those with password-protected access.

Figure 4

Major Query_win window organized around 4 panels. 'Current lists' gives, for each defined list, its name, content and origin, and contains buttons that perform various list-related operations. 'List content' displays names of sequences in currently selected list, allows selection of one such name, and allows choice of what information related to the currently selected sequence is to be displayed. Four displays are possible that provide selected or summarized annotations or primary sequence data. The small floating window shows how desired annotation items are selected. 'Text data output' displays such information and allows copy-to-clipboard. Central panel 'Query' receives queries either typed in by user of progressively built using the 'Select' and 'Oper' menus. This panel also allows direct sequence access from names or accession numbers.

Figure 5

Toy program using the C remote ACNUC access API to extract all encoded protein sequences from the Ensembl database [12] in FASTA format. API-specific program elements are italicized.

N=D13109	ID D13109; SV 1; linear; genomic DNA; STD; PLN; 1639 BP.
AC=D13109	AC D13109;
	DT 26-JAN-1993 (Rel. 34, Created)
K=release 83	DT 17-APR-2005 (Rel. 83, Last updated, Version 8)
	DE Oryza sativa (japonica cultivar-group) mitochondrial gene for
K=rpob protein	DE rpoB protein, partial cds.
SP=oryza sativa	KW rpoB protein.
	OS Oryza sativa (japonica cultivar-group)
	OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
SP=poales	OC Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae;
	OC BEP clade; Ehrhartoideae; Oryzeae; Oryza.
O=mitochondrion	OG Mitochondrion
	RN [1]
	RP 1-1639
AU=hirai	RA Hirai A.;
	RL Submitted (28-AUG-1992) to the EMBL/GenBank/DDBJ databases.
	RL Atsushi Hirai, Faculty of Agriculture, The University of Tokyo
	RN [2]
	RX DOI; 10.1007/BF00277131.
	RX PUBMED; 8437578.
AU=nakazono	RA Nakazono M., Hirai A.;
AU=hirai	RT "Identification of the entire set of transferred chloroplast
	RT DNA sequences in the mitochondrial genome of rice";
R=mgg/236/341	RL Mol. Gen. Genet. 236(2-3):341-346(1993).
J=MGG	FH Key Location/Qualifiers
Y=1993	FH
K=source	FT source 1..1639
SP=oryza	FT /organism="Oryza sativa (japonica cultivar-group)"
O=mitochondrion	FT /organelle="mitochondrion"
	FT /cultivar="Nipponbare"
M=genomic dna	FT /mol_type="genomic DNA"
TID=39947	FT /db_xref="taxon:39947"
K=misc_feature	FT misc_feature 830..1639
	FT /note="homologous to ctDNA"
N=D13109.RPOB	FT .RPOB CDS 1059..>1639
T=CDS	FT /codon_start=1
K=rpoB	FT /gene="rpoB"
K=rpoB protein	FT /product="rpoB protein"
	FT /db_xref="GOA:Q04936"
	FT /db_xref="InterPro:IPR007642"
	FT /db_xref="UniProtKB/TrEMBL:Q04936"
K=BAA02417	FT /protein_id="BAA02417.1"
	FT /translation="MLRNGNEGMSTIPGFSQIQFEGFCRFINQGLA..
	FT .....WARVSRKQKISVLVLSSAMGSNLKEIL"
	SQ Sequence 1639 BP; 482 A; 363 C; 333 G; 461 T; 0 other;

Fig. 1 - Gouy

```
ID   3MG1_ECOLI              Reviewed;         187 AA.
DE   DNA-3-methyladenine glycosylase 1 (EC 3.2.2.20) (DNA-3-methyladenine
DE   glycosylase I) (3-methyladenine-DNA glycosylase I, constitutive)
DE   (TAG I) (DNA-3-methyladenine glycosidase I).
```


Fig. 2 - Gouy

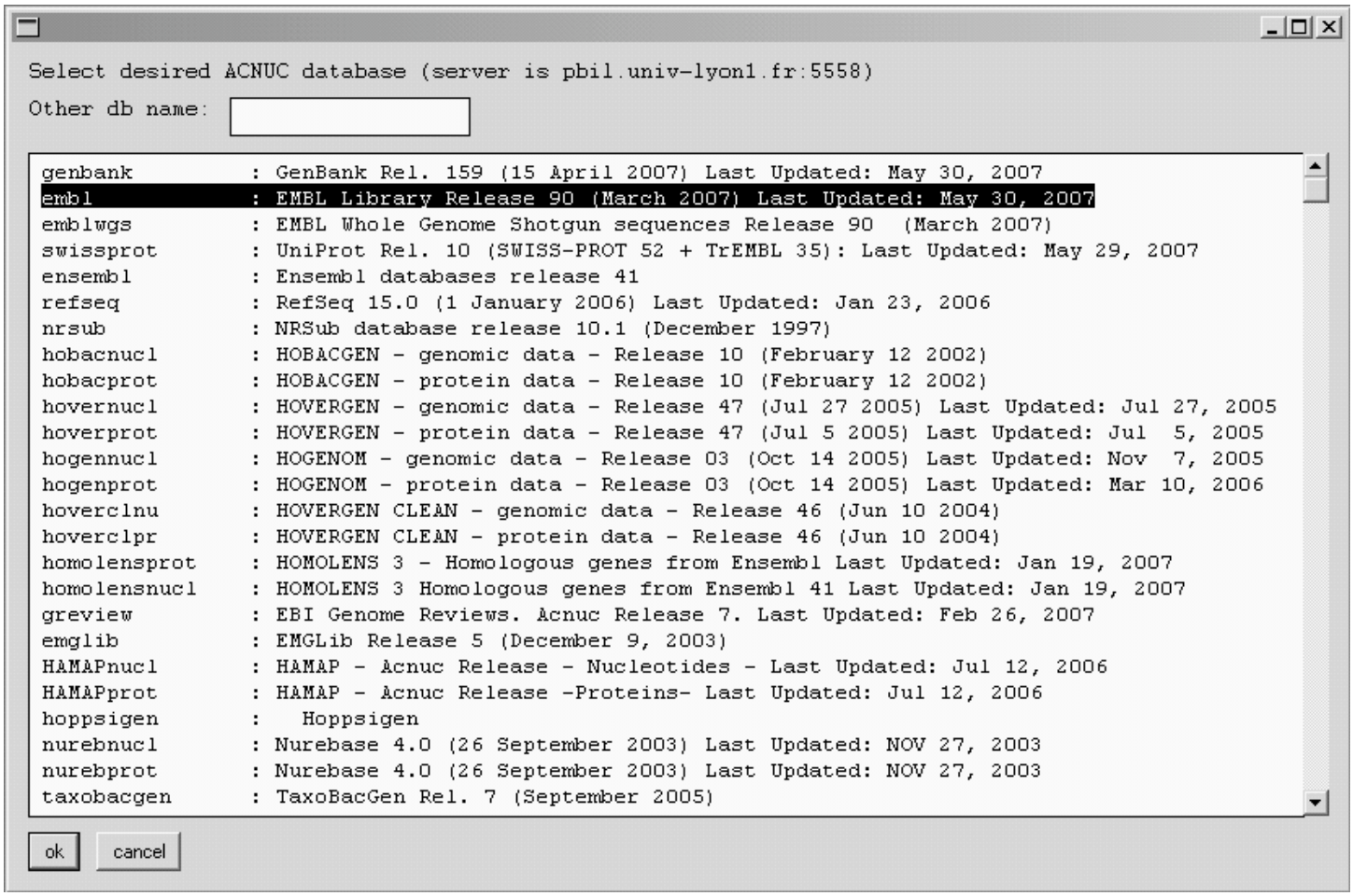
Select desired ACNUC database (server is pbil.univ-lyon1.fr:5558)
Other db name:
genbank : GenBank Rel. 159 (15 April 2007) Last Updated: May 30, 2007
embl : EMBL Library Release 90 (March 2007) Last Updated: May 30, 2007
emblwgs : EMBL Whole Genome Shotgun sequences Release 90 (March 2007)
swissprot : UniProt Rel. 10 (SWISS-PROT 52 + TrEMBL 35): Last Updated: May 29, 2007
ensembl : Ensembl databases release 41
refseq : RefSeq 15.0 (1 January 2006) Last Updated: Jan 23, 2006
nrsub : NRSub database release 10.1 (December 1997)
hobacnucl : HOBACGEN - genomic data - Release 10 (February 12 2002)
hobacprot : HOBACGEN - protein data - Release 10 (February 12 2002)
hovernucl : HOVERGEN - genomic data - Release 47 (Jul 27 2005) Last Updated: Jul 27, 2005
hoverprot : HOVERGEN - protein data - Release 47 (Jul 5 2005) Last Updated: Jul 5, 2005
hogennucl : HOGENOM - genomic data - Release 03 (Oct 14 2005) Last Updated: Nov 7, 2005
hogenprot : HOGENOM - protein data - Release 03 (Oct 14 2005) Last Updated: Mar 10, 2006
hoverclnu : HOVERGEN CLEAN - genomic data - Release 46 (Jun 10 2004)
hoverclpr : HOVERGEN CLEAN - protein data - Release 46 (Jun 10 2004)
homolensprot : HOMOLENS 3 - Homologous genes from Ensembl Last Updated: Jan 19, 2007
homolensnucl : HOMOLENS 3 Homologous genes from Ensembl 41 Last Updated: Jan 19, 2007
greview : EBI Genome Reviews. Acnuc Release 7. Last Updated: Feb 26, 2007
emglib : EMGLib Release 5 (December 9, 2003)
HAMAPnucl : HAMAP - Acnuc Release - Nucleotides - Last Updated: Jul 12, 2006
HAMAPprot : HAMAP - Acnuc Release -Proteins- Last Updated: Jul 12, 2006
hoppsigen : Hoppsigen
nurebnucl : Nurebase 4.0 (26 September 2003) Last Updated: NOV 27, 2003
nurebprot : Nurebase 4.0 (26 September 2003) Last Updated: NOV 27, 2003
taxobacgen : TaxoBacGen Rel. 7 (September 2005)
ok
cancel

Fig. 3 - Gouy


acnuc retrieval of EMBL Library Release 90 (March 2007) Last Updated: May 26, 2007
Quit Edit Select Oper Browse Log Help
Current lists
list1 18509 sp=cyanobacteria
list2 5835210 k=release 91
list3 1766 list1 and list2
save list
extract seqs. to file
edit list
scan annots.
replace by parents
add subseqs.
del. list
del. all lists
Query
clear
exec
list1 and list2
by name
Status
Query Ok.
List content
DQ225055
DQ225056
DQ225057
DQ225058
DQ225059
DQ225060
DQ225061
DQ225062
DQ225063
DQ225064
DQ225065
DQ225066
DQ225067
choose annots
Choose shown
Annotations
Annotations
Sequence
Short descript.
Name + length
Text data output
ID DQ225060; SV 1; linear; ger
DT 24-OCT-2005 (Rel. 85, Creat
DT 20-APR-2007 (Rel. 91, Last
DE Uncultured cyanobacterium c
DE partial sequence.
KW ENV.
OS uncultured cyanobacterium
OC Bacteria; Cyanobacteria; er
RA Yannarell A.C., Steppe T.F.
RL Environ. Microbiol. 9(3):57
RA Yannarell A.C., Steppe T.F.
RL Submitted (28-SEP-2005) to
RL Institute of Marine Science
RL 3431 Arendell St., Morehead
ggcaaccctg acggagcaac gccgc
ttctcaggga agacggactg acggt
cagccgcggt aagacggagg aggca
caggcggtta tttaagtctg ctgtt
actgaataac ttgagagcgg taggg
agatatcggg aagaacacca gtggc
79 BP.
16S ribosomal RNA gene,
databases.
h Carolina at Chapel Hill,
gggttg taaaccactt 60
gctaac tccgtgccag 120
gggcgt aaagcgtccg 180
cggatc ggcagtggaa 240
tagcgg tgaaatgcgt 300
gcgact gacgctgagg 360
Choose annotations
ALL
AC
PR
DT
KW
OS
OC
OG
RN
RC
RP
RX
RG
RA
RT
RL
DR
CC
AH
AS
FH
FT
SQ
SEQ
ok
cancel

Fig. 4 - Gouy

```
#include "raa_acnuc.h"
int main(int argc, char **argv)
{
int err, numlist, count, loop, length;
char *message, *name, *protseq;
err = raa_acnucopen_alt("pbil.univ-lyon1.fr", 5558,
        "ensembl", "testp");
err = raa_proc_requete("sp=rattus and t=cds", &message,
        "ratcds", &numlist, NULL, NULL, NULL);
loop = 1;
while( (loop = raa_nexteltinlist(
                loop, numlist, &name, &length)) != 0) {
        protseq = raa_translate_cds(loop);
        fprintf(stdout, ">%s %d\n%s\n",
                name, length, protseq);
        }
raa_acnucclose();
return 0;
}
```


Fig. 5 - Gouy