 Nucleic Acids Research, 1999, Vol. 27, No. 1 365–367

KEYnet: a keywords database for biosequences functional organization

F. Licciulli[2], D. Catalano[2], D. D'Elia[2], V. Lorusso[3] and M. Attimonelli[1,*]

[1]Department of Biochemistry and Molecular Biology, Faculty of Sciences, University of Bari, 70126 Bari, Italy, [2]Area di Ricerca, CNR, 70126 Bari, Italy and [3]Department of Biochemistry, Faculty of Medicine, University of Bari, 70125 Bari, Italy

Received September 2, 1998; Revised September 21, 1998; Accepted October 23, 1998

ABSTRACT

KEYnet is a database where gene and protein names are hierarchically structured. Particular care has been devoted to the search and organisation of synonyms. The structuring is based on biological criteria in order to assist the user in the data search and to minimise the risk of loss of information. Links to the EMBL data library by the entry name and the accession number have been implemented. KEYnet is available through the World Wide Web at the following site: http://www.ba.cnr.it/keynet.html . Recently KEYnet has incorporated specific gene name classifications, which can be browsed starting from the above-mentioned KEYnet home page: the Mitochondrial Gene Names classification and the Rat Gene Names classification. KEYnet database has also been structured in a flatfile format and can be queried through SRS (http://bio-www.ba.cnr.t:8000/srs).

INTRODUCTION

The use of nucleic acid sequence databases is often made clumsy by the presence of inconsistencies, errors and redundancies. The most common interrogation criteria for databases are keywords. In order to have a targeted retrieval using such criteria, keywords need to be correctly coded. Since the present paper refers to EMBL data library keywords, the problems encountered in dealing with its keyword system will be mostly discussed.

In an EMBL entry, the keywords line describes the properties of the sequence, i.e., associated phenotype, biological and/or enzymatic activity of its product, general and functional classification of the gene and/or gene product. It also reports the macromolecules and substrates the gene product can bind, e.g., DNA, calcium or other proteins, the sub-cellular location of the gene product and any other information relevant to the entry (1).

The keywords chosen for each entry are, therefore, a reference for the sequence and provide information that can be used to extract sequence lists according to functional and/or structural criteria. Users should be given the opportunity to extract sequences with a known biological function by applying such criteria.

The assignment of keywords to entries is, however, often defective or inconsistent and left to the choice of the researcher submitting the sequence to the database. Such a situation, together with the recent explosion in the number of nucleic acid and protein sequences, has created many problems due to redundancy and inconsistency of data, which greatly reduce the usefulness of the EMBL data library. Indeed, the database usefulness is strictly connected to the availability of an efficient interrogation system, but even the best retrieval system fails if it is not supported by a correctly structured database containing consistent information.

EMBL keywords lack of organization is due to two main problems: biological and lexical.

The biological problem derives from the lack of standardization in the names associated to proteins and genes. Consequently, the same protein or gene can be named differently according to the context where it acts.

The lexical issue is related to the fact that for the same keyword different spellings or abbreviations are used.

As a consequence of such biological and lexical inconsistencies at the level of keyword codification, data retrieval often gives false results. Also noteworthy is the problem of errors in keywords format and of typos.

Consequently, an entry associated with a 'wrong' keyword can no longer be retrieved, unless an approximated search is carried out. Therefore, whoever performs a search by keywords should know all the names, spellings, abbreviations and short names used in the annotation of a given sequence in order to obtain correct and complete information.

To solve such a problem, in 1989 our group undertook the first tree structuring of the keywords for the GenBank (2) and EMBL databases (3), organising them into a hierarchical structure (4). In this structure, each keyword was classified according to the biological function of the associated sequence and was linked to other keywords by functional relationship. Links among lexical or biological synonyms were defined and implemented.

Recently, two parallel networks have been implemented: the RAT Gene Names Tree databases and the Mitochondrial Gene Names Tree database.

DATA SOURCES

Keywords, i.e., gene and protein names have been extracted from the EMBL data library.

*To whom correspondence should be addressed. Tel: +39 080 5482130; Fax: +39 080 5484467; Email: marcella@area.ba.cnr.it

The authors wish it to be known that, in their opinion, all authors should be regarded as joint First Authors

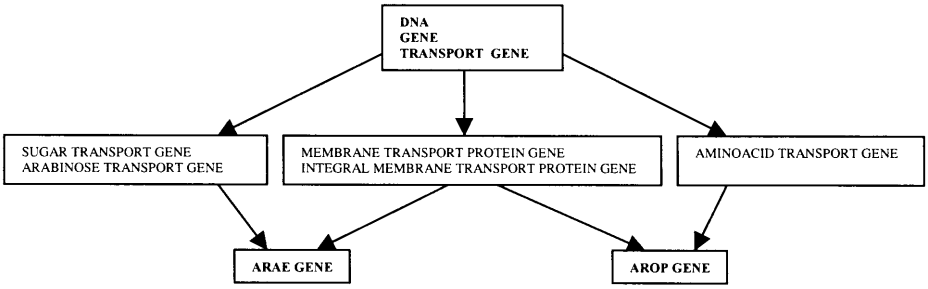

Figure 1. A schematic representation of two-transport protein gene classifications: ARAE and AROP. These genes, coding for membrane transport proteins, are classified under the 'integral membrane transport protein gene' node, which is descendant of 'membrane transport protein gene' node, on the basis of their cellular localization and under the 'Aminoacid Transport Gene' branch (AROP gene) and the 'Arabinose Transport Gene' branch (ARAE gene) depending upon their substrate (functional classification).

All the biological information about sequence associated with them have been extracted from the same primary databases (EMBL data library and GenBank) and from specialized databases such as SWISS-PROT (5), ENZYME (6) or any other suitable database. MEDLINE has been also consulted when the previous databases did not contain the necessary information for the keywords classification. As far as the Rat Gene Names branch is concerned, the data source is the Rat Locus List http://ratmap.gen.gu.se/lassolite/example/listsearch.html

KEYnet database is updated at each EMBL data library release, and at this time the link among the keywords in KEYnet and the EMBL data library entry names is established.

KEYnet STRUCTURE

KEYnet structure is made up of a set of elements, nodes, linked to form a father–son relation. At the highest level there is an element, the root, which links all the branches in the tree. The most important branches are the nodes protein, DNA and RNA (see figure 2 of ref. 4), all direct descendants of the tree root and ancestors to the keywords in the database. Namely, the structure is organized to allow branches to be linked in a network, which is quite different from the simple tree structure. Synonymous keywords are linked among them in a chain where one of the synonyms is chosen as the principal and all others are defined as secondary. In the structure each keyword is associated to the EMBL data library entries where the keyword itself and/or its synonymous and/or its descendants are reported in the KW line.

The implementation of the RAT Gene Names Tree and the Mitochondrial Gene Names Tree in two separate structures is due to their organism and cellular localization specificity, respectively. This avoids linking the RAT and Mitochondrial Gene Names to EMBL data library entries from different organisms and cellular location.

The structuring of the Rat Gene Names branch is performed starting from the RAT Locus List available at the above cited Web site taking into account the information associated with the gene both in the RATMAP site and in KEYnet. The Mitochondrion Gene Names classification has been structured as a contribution to the MitBASE project (7).

The present content of KEYnet database is shown in Table 1.

One of the major problems encountered during data classification is related with the gene names branch. The gene naming problem is recognized worldwide as difficult to solve, due to the freedom with which users assign names to a gene whenever it is discovered. Several attempts to address this problem are in progress (8,9, FlyBase Nomenclature Document Version 4.1 June 3, 1997; Genetic nomenclature for *Drosophila melanogaster* http://www.ebi.ac.uk:7081/docs/nomenclature/ and http://www.gene.ucl.ac.uk/nomenclature).

We have organized gene names by establishing a starting set of main ancestor keywords relevant to their primary biological functions. As most genes are indicated with an acronym, several sub-branches have been added to the structure of the GENE branch in the KEYnet database.

An example of the criteria adopted for the keyword classification in the GENE branch is the structuring of the two transport genes 'ARAE' and 'AROP' (Fig. 1).

Moreover the greatest advantage of KEYnet database, that is its speed and complete information, is almost invalidated by biological synonyms used to define gene names. One of the several cases where different genes have been given the same name is the ARCA gene. This gene in *Clostridium perfrigens* (accession nos X97768 and X97684) codes for arginine deiminase, an important enzyme in arginine metabolism, whereas in *Escherichia coli* (accession nos L34010 and L20873) another gene with the same name codes for a superoxide dismutase regulator under anaerobic conditions. Therefore, we have been forced to classify this gene both in the 'arginine deiminase gene' and in the 'regulatory gene' branches. Thus searching in KEYnetWWW (see below) for ARCA gene produces a sequence list related to heterogeneous data.

The problem with biological synonyms will be only eliminated when standard rules are set to guide gene nomenclature. A good alternative might be instead to consult KEYnet database whenever a new gene has to be named.

On the contrary the availability in KEYnet of synonym chains allows the end-user to retrieve a set of entries as complete as possible in a single query.

Table 1. KEYnet database content at August 1998

Keywords in KEYnet DNA branch	18 091
Keywords in KEYnet RNA branch	992
Keywords in KEYnet PROTEIN branch	34 805
Total keywords in KEYnet database	53 888
Keywords in Rat Gene Names database	2066
Keywords in Mitochondrial Gene Names database	889

KEYnet FLATFILE

A flatfile (ff) format for the KEYnet database has been designed. Each entry in the flatfile is identified by the principal synonym in the structure. At present 28 598 KEYnet ff entries have been generated of which 9067 contain secondary synonyms. The KEYnet ff can be distributed worldwide and downloaded independently from the computer system. Through the KEYnet ff, the database has been implemented in the SRS (10) system and linked dynamically to the EMBL data library. Moreover, links to any other biological database where gene and protein names are coded according to well defined rules (e.g. KW lines in the EMBL data library and in the SWISS-PROT database, Features lines in the GenBank database) could be implemented. At present, 23 100 KEYnet ff entries have been linked to 2 073 529 EMBL data library entries (Release 56).

KEYnet QUERY SYSTEMS

Different systems for querying KEYnet database have been developed. The RETKEY program, written in FORTRAN and C, is available at the CNR Research Area of the Bari server, while a slightly different version has been implemented in the World Wide Web, KEYnetWWW (http://www.ba.cnr.it/keynet.html). Moreover, KEYnet can be queried through the SRS server of the CNR Research Area of Bari (Italy) (http://bio-www.ba.cnr.t:8000/srs).

As far as the performance and easy usage of the KEYnet query systems are concerned, KEYnetWWW is the better system both because it can be accessed worldwide and because the retrievable information is the most complete.

KEYnetWWW usage

Starting from the KEYnet home page, clicking on the option 'KEYnet tree browsing' it is possible to navigate through the network either clicking on one of the three principal ancestors (DNA, RNA or PROTEINS) or by typing the complete or approximate keyword name to be searched. In the latter case the level of the max depth of the tree can be chosen. After this request the network relevant to the query is displayed. The button 'Sequence list' allows the retrieval of the list of the EMBL data library nucleotide sequences associated with the searched keyword and with its synonyms and descendants. Each EMBL data library entry of the Sequence list can be managed using the view, save and link options of the SRS system. The options 'RAT Genes Tree Browsing' and 'Mitochondrial Genes Tree Browsing' work in a similar way and the links to the EMBL data library sequences will soon be implemented.

KEYnet by SRS

The usage of KEYnet database by SRS is based on the KEYnet ff and it is possible to search data asking for a given keyword, for an ascendant, for a descendant or for synonyms. It is also possible to select keywords which are leaves in the tree or keywords of internal nodes or simply the root. It is also possible to guide the query by limiting the selected data on the basis of the number of descendants. The selected data are displayed in the KEYnet ff and by clicking on the Name or on the Synonymous lines, the list of the EMBL data library sequences associated with them is reported. By clicking on the ascendant name or on one of the descendant names the relevant KEYnet entry is displayed. The limit of the usage of KEYnet by SRS consists of the fact that it is not possible, while KEYnetWWW, to obtain the complete list of the EMBL data library entries related to descendants and synonyms.

As an example of the advantages of KEYnet for the retrieval of EMBL nucleotide sequences the results are reported here below of a search for 'arylesterase'.

When searching with KEYnetWWW, 66 EMBL data library entries are retrieved (list A); whereas searching with SRS in the EMBL data library, according to the 'Keywords' criteria or the 'All text' criteria for 'Arylesterase', 4 (list B) and 169 (list C) entries are extracted, respectively. The comparison between list A and list C shows 109 entries (list D) not retrieved by KEYnetWWW because in the majority they are related to genes 'similar' to arylesterase; only four entries of list D code for arylesterase but they are not correctly annotated. Searching for 'arylesterase' through SRS applied to the ENZYME database and linking the resulting list to the SWISS-PROT and EMBL databases, 36 entries are retrieved of which only seven are not retrieved by KEYnetWWW because the sequences here referred are related to genes coding for multifunctional enzymes. Once again, the incomplete annotation of these entries does not allow KEYnetWWW to retrieve them. On the other hand, 45 entries retrieved through KEYnetWWW (part of list A) are not extracted starting from the ENZYME database because they do not contain the cross-referencing line to the SWISS-PROT database.

Users of KEYnet are kindly invited to cite the present article.

ACKNOWLEDGEMENTS

The authors have contributed to KEYnet to the same extent: F. Licciulli has cared for the computer developments, D. Catalano, D. D'Elia and V. Lorusso for the biological classification of keywords and M. Attimonelli has coordinated and designed the database. This work has been partially supported by MPI (Italy), by the EU-Biotechnology Programme (Contract n. BIO4-CT95-0037) and by CNR Research Area of Bari (Italy).

REFERENCES

1 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
2 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
3 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
4 Tullo,A., Liuni,S.and Attimonelli,M. (1990) *Protein Seq. Data Anal.*, **3**, 327–334.
5 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
6 Bairoch,A. (1996) *Nucleic Acids Res.*, **24**, 221–222.
7 Attimonelli,M., Altamura,N., Benne,R., Boyen,C., Brennicke,A., Carone,A., Cooper,J.M., D'Elia,D., De Montalvo,A., de Pinto,B., De Robertis,M., Golik,P., Grienenberger,J.M., Knoop,V., Lanave,C., Lazowska,J., Lemagnen,A., Malladi,B.S., Memeo,F., Monnerot,M., Pilbout,S., Schapira,A.H.V., Sloof,P., Slonimski,P., Stevens,K. and Saccone,C. (1999) *Nucleic Acids Res.*, **27**, 128–133.
8 Lonsdale,D.M. and Leaver,C.J. (1988) *Plant Mol. Biol.*, **6**, 14–21.
9 Hallick,R.B. (1989) *Plant Mol. Biol.*, **7**, 266–275.
10 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.