Nucleic Acids Research, 1993, Vol. 21, No. 13 3001–3002

LISTA, a comprehensive compilation of nucleotide sequences encoding proteins from the yeast *Saccharomyces*

Patrick Linder, Reinhard Dölz[1], Marie-Odile Mossé[2], Jaga Lazowska[2] and Piotr P.Slonimski[2]
Department of Microbiology, Biozentrum, [1]Biocomputing, Biozentrum, Klingelbergstr. 70, 4056 Basel, Switzerland and [2]Centre de Génétique Moléculaire, Laboratoire propre du CNRS associé à l'Université Pierre et Marie Curie, F-91190 Gif sur Yvette, France

ABSTRACT

The amount of nucleotide sequence data is increasing exponentially. We therefore made an effort to make a comprehensive database (LISTA) for the yeast *Saccharomyces cerevisiae*. Each sequence has been attributed a single genetic name and in the case of allelic duplicated sequences, synonyms are given, if necessary. For the nomenclature we have introduced a standard principle for naming gene sequences based on priority rules. We have also applied a simple method to distinguish duplicated sequences of one and the same gene from non-allelic sequences of duplicated genes. By using these principles we have sorted out a lot of confusion in the literature and databanks. Along with the genetic name, the mnemonic from the EMBL databank, the codon bias, reference of the publication of the sequence and the EMBL accession numbers are included in each entry.

In view of the very rapid growth of sequence data we started to compile a list of coding sequences from the yeast [1, 2]. The database contains sequences from *Saccharomyces cerevisiae*, *Saccharomyces carlsbergiensis* and *Saccharomyces uvarum*, which are believed to constitute conspecific taxonomic species [3]. Sequences from the unrelated Schizosaccharomyces *pombe*, *Candida*, *Hansenula* and others are not included. We also exclude sequences from extragenomic elements like the 2-micron (2μm) plasmid, mitochondrial DNA, killer sequences and from Ty elements. The actual list contains 1001 sequences from 818 individual genes. It is currently updated and will be published this year.

The database includes at present a gene name, a synonym in the case the same sequence has been published more than once under different names, the mnemonic, the length of the coding sequence without the stop codon, the codon bias according to [4], the reference of the first publication of the sequence, the accession number and if necessary a commentary. Other items such as the chromosomal localization, description of the gene product, cross-reference to other databases and adjacent genes will be included in the future.

A major problem in establishing such a database is the nomenclature. We tried whenever possible to follow the genetic nomenclature and use the glossary compiled by [5]. In many cases, however, no or incorrect gene designations have been given to published sequences. Moreover, the same name was given to different sequences or different names have been given to the same sequence. To sort out this problem of nomenclature we adopted a *priority rule* for naming genes in the present database [2]. According to this rule the name of the first published sequence (date of acceptance of the publication) is used in the list, provided it is in accordance with the standard genetic nomenclature. Other names are included as synonyms. In some cases four letter designations (*ARGR1*, *MRPL20*) or gene names followed by a letter (*RPL4A*, *TIF51A*) have also been used. In the case of historically well established gene designations such as *HO*, it was self-evident that they should be retained.

Sequences of open reading frames which occur more than once may represent allelic sequences originating from the same gene or non allelic sequences from duplicated genes. We distinguished in this database between these two cases by comparing the 5′ and 3′ non coding sequences, which in general diverge considerably in non allelic duplicated genes but are highly similar or identical in allelic sequences. Exceptions have been discussed [2]. In both cases, the results of the comparisons are included in the commentary.

Differences in duplicated sequences may be due to polymorphisms or to sequencing errors. In general we do not distinguish between these two possibilities, but include in the commentary the percentages of identity of nucleotide and protein sequences. In some cases obvious frameshifts in the sequences lead to modifications of the length of the predicted protein, which are also outlined in the commentary.

Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data which make up the entry. Note that each line begins with a two-character line code, which indicates the type of information contained in the line. The currently used line types, along with their respective line codes, are listed in table 1. This arrangement of the database allows an easy integration with other databanks. For example, links between the LISTA database and the EMBL sequence database were accomplished using the Sequence Retrieval System program [6].

The LISTA database is available either on diskettes (M.-O-Mossé, Centre de Génétique Moléculaire, CNRS, F-91190 Gif sur Yvette; mosse@frcgm51.bitnet) or by anonymous FTP from

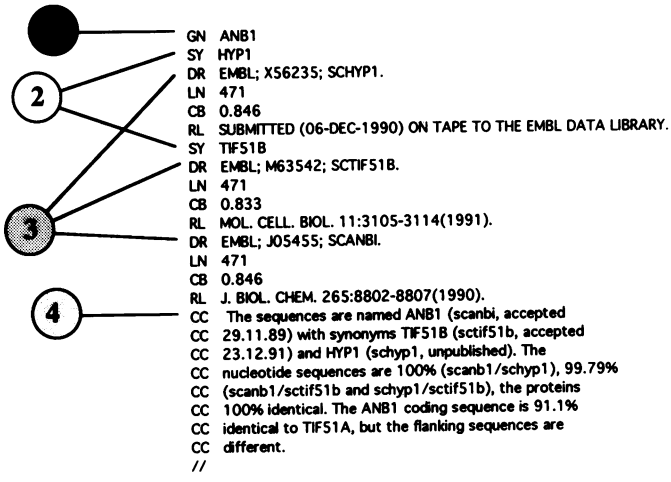

Figure 1. Detailed view of a single entry. The identifier of the entry is a gene name (1), with several synonyms if appropriate (2). Each of the synonyms, and the sequence attributed to the gene names, has a separate reference (3). The comment lines (4) refer to detailed information and, if needed, refer to the printed version of the database if needed in order to explain complex families.

Table 1. Format of the LISTA database in electronic form

Number of fields	Key	Description
always 1 (begins each entry)	GN	gene name
0 or more	SY	synonym
1 or more per GN or SY	DR	EMBL accession number and Mnemonic
1 per DR	LN	length of sequence
1 per DR	CB	codon bias
1 per DR	RL	Literature reference
0 or more	CC	additional comments
1 per entry	//	end of entry

bioftp.unibas.ch [131.152.8.1] on the internet. New sequences and comments on the existing database may be sent to P. Linder (linder@urz.unibas.ch).

Figure 1 outlines some specific entries from the LISTA database. Note that each entry has one or more sequence references, indicating the relationship between sequence data with respect to their common gene name.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministère de l'Education Nationale, the Ligue Nationale contre le Cancer and E.E.C. (to P.S.) and by the Swiss National Science Foundation and Kanton Basel-Stadt (to P.L. and R.D.).

REFERENCES

1. Mossé, M.O., Brouillet, S., Risler, J.L., Lazowska, J. and Slonimski, P.P. (1988) *Curr. Genet.* **14**, 529–535.
2. Mossé, M.-O., Linder, P., Lazowska, J. and Slonimski, P.P. (1993) *Curr. Genet.* **23**, 66–91.
3. Barnett, J.A., Payne, R.W. and Yarrow, D. (1983) (Cambridge University Press, Cambridge) 811.
4. Bennetzen, J.L. and Hall, B.D. (1982) *J. Biol. Chem.* **257**, 3026–3031.
5. Mortimer, R.K., Contopoulou, C.R. and King, J.S. (1992) *Yeast* **8**, 817–902.
6. Etzold, T. and Argos, P. (1993) *CABIOS* **9**, 49–57.