© 1994 Oxford University Press
Nucleic Acids Research, 1994, Vol. 22, No. 17 3459–3461

LISTA, LISTA-HOP and LISTA-HON: a comprehensive compilation of protein encoding sequences and its associated homology databases from the yeast *Saccharomyces*

Reinhard Dölz, Marie-Odile Mossé[1], Piotr P.Slonimski[1], Amos Bairoch[2] and Patrick Linder[3]*
Biocomputing, Biozentrum, Klingelbergstr. 70, 4056 Basel, Switzerland, [1]Centre de Génétique Moléculaire, Laboratoire propre du CNRS associé à l'Université Pierre et Marie Curie, F-91190 Gif sur Yvette, France, [2]Departement de Biochimie Médicale, Centre Médicale Universitaire, 9 Ave Champel, 1211 Genève 4 and [3]Department of Microbiology, Biozentrum Klingelbergstr. 70, 4056 Basel, Switzerland

ABSTRACT

We continued our effort to make a comprehensive database (LISTA) for the yeast *Saccharomyces cerevisiae*. In this database each sequence has been attributed a single genetic name. In the case of duplicated sequences a simple method has been applied to distinguish between sequences of one and the same gene from non-allelic sequences of duplicated genes. If necessary, synonyms are given in the case of allelic duplicated sequences. Thus sequences can be found either by the name or by synonyms given in LISTA. Each entry contains the genetic name, the mnemonic from the EMBL data bank, the codon bias, reference of the publication of the sequence, Chromosomal location as far as known, Swissprot and EMBL accession numbers. To obtain more information on the included sequences, each entry has been screened against non-redundant nucleotide and protein data bank collections resulting in LISTA-HON and LISTA-HOP. The LISTA data base can be linked to the associated data sets or to nucleotide and protein banks by the Sequence Retrieval System (SRS).

LISTA, A COMPILATION OF CODING SEQUENCES

In view of the very rapid growth of sequence data we started to compile a list of coding sequences from the yeast [1–3] The database contains sequences from *Saccharomyces cerevisiae*, *Saccharomyces carlsbergiensis* and *Saccharomyces uvarum*, which are believed to constitute conspecific taxonomic species [4]. Likewise, sequences from *Schizosaccharomyces pombe*, *Candida*, *Hansenula* and others are not included. Not included are sequences from extragenomic nuclear elements, mitochondria and Ty elements. The actual list (LISTA3) contains 1400 sequences from 1138 individual genes. It is currently updated and a new release (LISTA4) will be available in summer 1994.

Data from the systematic sequencing of the yeast genome are only included if they have a defined function or a homologous sequence in the data banks. Unidentified open reading frames are not considered.

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 [5], the reference of the first publication of the sequence, the accession number of EMBL. In the case of conflicting sequence data or nomenclature a commentary is given to point out the divergences. The chromosomal localization as far as indicated in the sequence data banks or deduced from homologies will be included in the new version of LISTA4. To increase the information content of databases future releases will contain Swiss-Prot accession numbers and Swiss-Prot entries will refer to the LISTA accession numbers to facilitate communication between the two.

A major problem in establishing such a database is the nomenclature. We tried whenever possible to follow the genetic nomenclature and follow the glossary compiled by [6]. 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 use the name of the first published sequence (date of acceptance of the publication), provided it is in accordance with the standard genetic nomenclature [2]. In the case of historically well established gene designations such as *HO*, it was self-evident that they should be retained.

Duplicated sequences from the same gene or non allelic sequences from duplicated genes can be distinguished 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.

Each entry in the database is composed of lines. Different types of lines, each with their own format, are used to record the

*To whom correspondence should be addressed

Table 1. Format of the LISTA database in electronic form

Number of fields	Key	Description
always 1 (begins each entry)	GN	gene name
always 1	AC	LISTA accession number
0 or more	SY	synonym
1 or more per GN or SY	DR	Data references to either EMBL, Swissprot, LISTA-HON or LISTA-HOP
1 per DR	LN	length of sequence
1 per DR	CB	codon bias
1 per DR	RL	Literature reference
1 or more	DT	Date information for maintenance
0 or more	CC	additional comments
1 per entry	//	end of entry

Table 2. Format of the LISTA-HON and LISTA-HOP database, respectively, in electronic form

Number of fields	Key	Description
always 1 (begins each entry)	ID	gene name
1 per entry	DE	Description
1 per entry	AC	LISTA accession number
1 per entry	RL	Reference line from Sequence database
1 per entry	DT	Last change of entry
1 per entry	GN	Gene name
1 or more per entry	HY	Homology found
1 or more per HY	HD	Description of Homology
1 per entry	HT	Name of top scoring sequence in full length
0 or more per entry	HA	Alignment of top scoring sequence
0 or more	CC	additional comments
0 or more	XX	placeholder
1 per entry	SQ	Sequence entries (LISTA-HON only).

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. An example has been shown previously [7]. This arrangement of the database allows an easy integration with other data bank. Links between the LISTA database and the EMBL sequence datalibrary were accomplished using the Sequence Retrieval System program [8].

DATA ADDED FROM SEQUENCE HOMOLOGY SCREENING

The open reading frames collected in the LISTA4 database have been translated into protein sequences, and were subsequently screened against a non-redundant DNA sequence database collections composed with the 'nr' program from NCBI (Gish, W., National Center for Biotechnology Information, Bethesda, USA, software published on FTP server). Similarly, the open reading frames were screened against the same database on nucleotide level. The blastn and tblastn programs, respectively, were used to obtain top-scoring sequences with a significant homology [9]. To make the output more versatile, the output of the screening process was post-processed to give one line of description per sequence found, and one line per matching segment pair. Arbitrary but reasonable cut-offs (Probability < 10 e^{-30} or $>60\%$ Identity) were applied to list only entries which are believed to be most significant. The entry codes of LISTA-HON, and LISTA-HOP, respectively, reflect the fact that multiple reading frames can occur in the same sequence. If this is the case, the number of the reading frame is part of the ID, e.g. ENTRY-1 would designate the ORF 1 of the entry ENTRY.

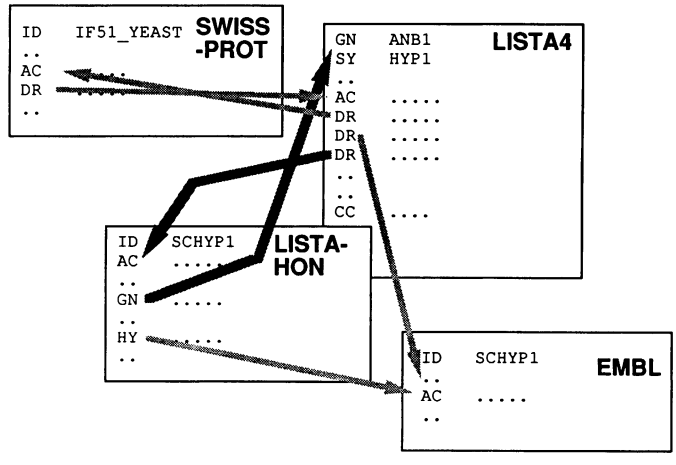

Figure 1. Schematic view of the interconnection of the LISTA4 database and the LISTA-HON database. The LISTA-HOP is connected in identical fashion. Additional pointers from SWISS-PROT to LISTA4 and from LISTA4 to SWISS-PROT and EMBL, resp., allow easy 'travelling' in between the different sources. Similarly, data in the homology database cross-reference to the EMBL data library. For a full LISTA entry example, see [7].

As LISTA and LISTA-HOP are plain flat files containing crossreferences to the sequence databases, the linkage between LISTA, LISTA-HON, and LISTA-HOP can be achieved using a sequence retrieval program which is capable of utilizing this information (see also Figure 1). We have successfully used the Sequence Retrieval System SRS (Etzold *et al.*), which is available in the public domain, and also accessible on public servers on the Internet. Whereas LISTA-HON is useful for checking sequence homologies on the DNA level, which are commonly found in yeast sequences, but rarely branch to other organisms, the LISTA-HOP database entries point to homologies of yeast genes found in other organisms. This is in particular useful to classify families of genes with respect to interspecies homology. In combination with the SRS Program, it is possible to reverse the usage of the link: Given a vertebrate sequence, it is possible to look up this entry in either LISTA-HON or LISTA-HOP to query LISTA for a similar gene in *Saccharomyces cerevisiae* based on sequence homology.

The LISTA, LISTA-HOP and LISTA-HON databases are available by anonymous FTP from *biofip.unibas.ch* [131.152.8.1].

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministère de la Recherche et de l'Espace (program GREG) (P.S.), from the Swiss National Science Foundation (R.D. and A.B) and the University of Basel (R.D. and P.L). We are very grateful to S.Brouillet, J.L.Risler and the Rechenzentrum of the University of Basel for help.

REFERENCES

1. Mossé, M.O., Brouillet, S., Risler, J.L., Lazowska, J. & Slonimski, P.P. *Curr. Genet.* 14, 529–535 (1988).
2. Mossé, M.-O., Linder, P., Lazowska, J. & Slonimski, P.P. *Curr. Genet.* 23, 66–91 (1993).
3. Mossé, M.O., Dölz, R., Lazowska, J., Slonimski, P.P. & Linder, P. in *The Yeasts* (eds. Wheals, A.E., Rose, A.H. & Harrison, S.E.) Academic Press, London, in press).
4. Barnett, J.A., Payne, R.W. & Yarrow, D. 811 (Cambridge University Press, Cambridge, 1983).
5. Bennetzen, J.L. & Hall, B.D. *J. Biol. Chem.* 257, 3026–3031 (1982).
6. Mortimer, R.K., Contopoulou, C.R. & King, J.S. *Yeast* 8, 817–902 (1992).
7. Linder, P., Dölz, R., Mossé, M.O., Lazowska, J. & Slonimski, P.P. *Nucleic Acids Res.* 21, 3001–3002 (1993).
8. Etzold, T. & Argos, P. *CABIOS* 9, 49–57 (1993).
9. Altschul, S.F., Gish, W.G., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* 215, 403–410 (1990).