50–52 Nucleic Acids Research, 1996, Vol. 24, No. 1

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[2,*]

Biocomputing, Biozentrum, Klingelbergstraße 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 and [2]Dept. de Biochimie médicale, Centre Médicale Universitaire, 1, r. Michel Servet, 1211 Genève 4, Switzerland

Received November 6, 1995; Accepted November 15, 1995

ABSTRACT

We continued our effort to make a comprehensive database (LISTA) for the yeast *Saccharomyces cerevisiae*. As in previous editions [Dölz,R., Mossé,M.-O., Slonimski, P.P., Bairoch,A. and Linder,P. (1994) *Nucleic Acids Res.* 22, 3459–3461] the genetic names are consistently associated to each sequence with a known and confirmed ORF. If necessary, synonyms are given in the case of allelic duplicated sequences. Although the first publication of a sequence gives—according to our rules—the genetic name of a gene, in some instances more commonly used names are given to avoid nomenclature problems and the use of ancient designations which are no longer used. In these cases the old designation is given as synonym. 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. New entries will also contain the name from the systematic sequencing efforts. Since the release of LISTA4.1 we update the database continuously. 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. This release includes reports from full Smith and Watermann peptide-level searches against a non-redundant protein sequence database. The LISTA data base can be linked to the associated data sets or to nucleotide and protein banks by the Sequence Retrieval System (SRS). The database is available by FTP and on World Wide Web.

LISTA, a compilation of coding sequences

In view of the very rapid growth of sequence data from genetic and biochemical experiments and systematic sequencing we continued to compile a list of coding sequences from yeast (1–4). The database contains sequences from *Saccharomyces cerevisiae*, *Saccharomyces carlsbergiensis* and *Saccharomyces uvarum*, which are believed to constitute conspecific taxonomic species (5). 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 (LISTA4.1) contains 1461 sequences and 349 synonyms from 1386 individual protein sequences. It has been updated in 1994 to include cross-references to the SWISSPROT database. Since release LISTA4.1 the database is updated continuously.

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. New entries since LISTA4.1 contain a SN entry field with the ORF designation from the systematic sequencing. Unidentified open reading frames are not considered.

The database includes at present a gene name and a unique accession number. The date field (used for administrative purposes) is followed by a cross-reference to the SWISSPROT database and the chromosomal location (if known). Each known DNA sequence is listed afterwards, with 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 (6), the reference of the first publication of the sequence and the accession number of EMBL are listed for each entry. In the case of conflicting sequence data or nomenclature a commentary is given to point out the divergences. To increase the information content of databases starting with release 31 of the SWISSPROT database the entries therein will refer to the LISTA accession numbers to facilitate communication between the two.

* To whom correspondence should be addressed

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 (7). 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. In some cases an original gene name is no longer used in the current literature, but a new designation (sometimes more reasonable) has been adopted. In these cases we adopt the new gene name, but give the original name as synonym. An effort is undertaken to provide a nomenclature consistent with SWISSPROT and the SGD database.

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 which are listed in Table 1. An example has been shown previously (8). This arrangement of the database allows an easy integration with other data bank. Links between the LISTA database and the SWISSPROT and EMBL sequence data library were accomplished using the Sequence Retrieval System program (9). The database is also accessible on the World Wide Web (http://www.ch.embnet.org/) or by using SWISSPROT at Expasy (http://expasy.hcuge.ch/).

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	SN	Name from systematic sequencing (since release 4.1)
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

Data from sequence homology screening improved

The open reading frames collected in the LISTA4 database have been translated into protein sequences, and were screened against a non-redundant protein sequence database collections composed with the ‘nr’ program from NCBI (Gish, W. , National Center for Biotechnology Information, Bethesda, USA, software published on FTP server). In addition to the SWISSPROT and PIR databases, the automatically translated EMBL entries as compiled by the SRS program (9) were added, thus expanding the scope of the data used for comparison. The algorithm employed for protein searches was a full Smith and Watermann search as implemented in the MPSRCH program running on a MasPar supercomputer with 4096 processors. As the databases grow rapidly, we will update this homology screening database frequently.

Additionally, DNA sequences were subsequently screened against a non-redundant DNA sequence database collections composed with the same tool as described above, covering EMBL and GenBank DNA sequence databases including updates.

The blastn and MPsrch programs, respectively, were used to obtain top-scoring sequences with a significant homology (10,11). 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 for blast, probability < 10^{-100} for MPsrch) were applied to list only entries which are believed to be most significant. The entry codes (for a format description, see Table 2) 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. As LISTA and LISTA-HOP are plain flat files containing cross-references 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. 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.

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	DR	Reference to LISTA 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	HN	Name of top homologous sequence as derived from database
0 or more	XX	Placeholder
0 or more	CC	Additional comments

The much enhanced sensitivity of the MPsrch program, paired with statistically based threshold limitation, allows to list even remotely related sequences. 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 *bioftp.unibas.ch* (131.152.8.1) or are accessible on the World Wide Web (http://www.ch.embnet.org/) or by using SWISSPROT at Expasy (http://expasy.hcuge.ch/).

ACKNOWLEDGEMENTS

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

REFERENCES

1 Mossé, M.O., Brouillet, S., Risler, J.L., Lazowska, J. & Slonimski, P.P. (1988) *Curr. Genet.* **14**, 529–535.
2 Mossé, M.-O., Linder, P., Lazowska, J. & Slonimski, P.P. (1993) *Curr. Genet.* **23**, 66–91.
3 Mossé, M.O., Dölz, R., Lazowska, J., Slonimski, P.P. & Linder, P. (1995) In Wheals, A.E., Rose, A.H. & Harrison, S.E. (eds) *The Yeasts.* Vol. 6, Academic Press, London, pp. 499–582.
4 Dölz, R., Mossé, M.O., Slonimski, P.P., Bairoch, A. & Linder, P. (1994) *Nucleic Acids Res.* **22**, 3459–3461.
5 Barnett, J.A., Payne, R.W. & Yarrow, D., p.811 Cambridge University Press, Cambridge, 1983.
6 Bennetzen, J.L. & Hall, B.D. (1982) *J. Biol. Chem.* **257**, 3026–3031.
7 Mortimer, R.K., Contopoulou, C.R. & King, J.S. (1992) *Yeast* **8**, 817–902.
8 Linder, P., Dölz, R., Mossé, M.O., Lazowska, J. & Slonimski, P.P. (1993) *Nucleic Acids Res.* **21**, 3001–3002.
9 Etzold, T. and Argos, P. *CABIOS* **9**, 49–57 (1993).
10 Altschul, S.F., Gish, W.G., Miller, W., Myers, E.W. & Lipman, D.J. (1990) *J. Mol. Biol.* 215, 403–410.
11 Collins, J.F. and Coulson, A.F.W. (1990) *Methods Enzymol.* **183**, 474–487.