Published online 28 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D159–D162***
doi:10.1093/nar/gkn772

tRNAdb 2009: compilation of tRNA sequences and tRNA genes

Frank Jühling[1], Mario Mörl[2], Roland K. Hartmann[3], Mathias Sprinzl[4], Peter F. Stadler[1,5,6,7] and Joern Pütz[8,*]

[1]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, [2]Institute for Biochemistry, University of Leipzig, Brüderstrasse 34, 04103 Leipzig, [3]Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, D-35037 Marburg, [4]Laboratorium für Biochemie, Universität Bayreuth, Universitätsstrasse 30, D-95440 Bayreuth, [5]RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstrasse 1, D-04103 Leipzig, Germany, [6]Department of Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria, [7]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA and [8]Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg, France

Received September 11, 2008; Accepted October 7, 2008

ABSTRACT

One of the first specialized collections of nucleic acid sequences in life sciences was the 'compilation of tRNA sequences and sequences of tRNA genes' (http://www.trna.uni-bayreuth.de). Here, an updated and completely restructured version of this compilation is presented (http://trnadb.bioinf.uni-leipzig.de). The new database, tRNAdb, is hosted and maintained in cooperation between the universities of Leipzig, Marburg, and Strasbourg. Reimplemented as a relational database, tRNAdb will be updated periodically and is searchable in a highly flexible and user-friendly way. Currently, it contains more than 12 000 tRNA genes, classified into families according to amino acid specificity. Furthermore, the implementation of the NCBI taxonomy tree facilitates phylogeny-related queries. The database provides various services including graphical representations of tRNA secondary structures, a customizable output of aligned or un-aligned sequences with a variety of individual and combinable search criteria, as well as the construction of consensus sequences for any selected set of tRNAs.

INTRODUCTION

As a constantly increasing number of complete genomes is published, it becomes necessary to transfer existing sequence compilations to state-of-the-art IT infrastructure to cope with the challenges of the post-genomic era. One of the most abundant groups of nucleic acids are tRNA molecules, being present in all types of cells and organelles. Unique features of these molecules include (i) the high degree of structural conservation, (ii) the plethora of cellular factors tRNAs interact with, and (iii) the largest diversity and highest density of nucleoside modifications found in nature. Furthermore, with respect to phylogeny, tRNAs offer a unique complexity: since each cell and eukaryotic organelle has a cohort of related but distinct tRNA species, phylogenetic analyses permit an integrated view on an entire set of tRNAs that underwent coevolution, rather than being restricted to comparison of a single RNA species. Hence, a tRNA database must fulfil specific criteria in order to address these features. The tRNAdb, with one of the largest numbers of entries among RNA databases, is not only an excellent model system for implementing and validating algorithms and processes of automated nucleic acid data transfer, but also for the development of novel sequence analysis tools (1). In its re-structured version, the tRNAdb meets the demands of present-day web-based interfaces and provides a basis for integration of structure-function relationships and additional information on evolution and phylogeny.

DATABASE CONTENT AND ORGANIZATION

In the new tRNA database, sequences are stored on a MySQL database server (http://dev.mysql.com) and also as a BLAST database (2). The relational database management system implements a powerful search engine that allows access to all data and offers a high flexibility in queries. In particular, the opportunity of using the BLAST database provides highly efficient similarity searches. The database for mammalian mitochondrial tRNA sequences (Mamit-tRNA) with its user-friendly web interface (http://mamit-trna.u-strasbg.fr) served as

*To whom correspondence should be addressed. Tel: +33 3 88 41 70 48; Fax: +33 3 88 60 22 18; Email: j.puetz@ibmc.u-strasbg.fr

template for the new database. Accordingly, color codes and visualization styles were adopted from the Mamit-tRNA compilation (3).

The new version of tRNAdb is based on the 'Compilation of tRNA sequences and sequences of tRNA genes', distributed as a collection of MS Excel spread sheets (1). To integrate this original sequence collection into the new compilation, the complete dataset was retrieved and stored in indexed tables using several custom-made scripts. After the integrity of the individual sequences had been verified, the data were transferred into the relational database system. For detailed taxonomic queries, a tree provided by the NCBIs taxonomy section (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy) was implemented, now providing a complete set of individual taxon names and synonyms. Furthermore, all taxon names appearing in the original tRNA compilation were manually matched with the taxonomic tree. Several outdated entries, where the organisms have been renamed or reclassified in the meantime, were identified and adjusted according to current taxonomy. In addition, the bacterial sequences were subjected to manual proofreading, replacing the previous erroneous entry and being tagged as 'corrected' in the comment note. New ID management was implemented with prefixes 'tdbD' and 'tdbR' for DNA and RNA sequences, respectively. However, for compatibility reasons, the newly designed web interface supports both the former and the new ID format.

Besides the imported data sets, 255 new tRNA gene sequences retrieved from a series of completed archaeal genomes recently submitted to NCBI (*Methanococcus aeolicus Nankai-3*, *Methanosarcina acetivorans C2A*, *Methanospirillum hungatei JF-1*, *Nanoarchaeum equitans Kin4-M*, *Staphylothermus marinus F1* and *Sulfolobus acidocaldarius DSM 639*) were scanned by tRNAscan-SE (4) and imported using a new data input interface directly connected to the database.

For reasons of clarity and compatibility, all sequences are presented in the alignment format of the Mamit-tRNA database (3), including a structural annotation generated from the assignment of nucleotide position numbers. Currently, the data set contains 12 099 sequences from tRNA genes (577 organisms) and 623 tRNA sequences (104 organisms) (Table 1).

The database is organized in two independent and fully searchable parts. One part combines tRNA gene sequences which, in the previously published compilation (1), were divided into the sections 'Genomic tRNA Compilation' [mostly identified by the annotation of complete genomes using tRNAScan-SE (4)] and 'Compilation of tRNA Genes'. This part of the database also includes tRNA-like structures ('TLSs') encoded in viruses and phages. In the second part, sequences obtained from direct tRNA analysis [including identified nucleoside modifications (5,6)] are presented, corresponding to the former section 'Compilation of tRNA Sequences'.

SEQUENCE SEARCH TOOL

Using the advanced functionality of MySQL- and BLAST-based databases, the new compilation provides a powerful and fast search engine. Query results are stored on the server and are linked to the corresponding session object. Furthermore, the retrieved data can be edited manually. Queries can include DNA or RNA sequences, amino acid family, anticodon, references, Pubmed-ID of the reference, gene description as well as comments. Taxa can be identified by searching for specific names, strains, taxononomic IDs or even synonyms. In addition, individual searches concerning sequence and/or structural characteristics (e.g. conserved or semiconserved nucleotides) are possible. Besides that, the server accepts sequence IDs of the new and the previous tRNA database as queries, and can perform BLAST searches.

Query results are displayed in a clearly arranged list and can be adjusted concerning individual details. Since the 3′-CCA terminus is not included in the Mamit-tRNA color code (CCA ends are not encoded in mitochondrial tRNA genes), a new color was assigned to the CCA triplet. In addition, the list covers information related to each organism, the amino acid specificity and the primary sequence of the tRNA. Optionally, the secondary structure can be displayed for each kind of sequence (DNA or RNA). For convenience, a thumbnail presentation allows a fast preview of the secondary structures. To directly highlight the cloverleaf structure of a selected tRNA, an image generator has been implemented, supporting all tRNA domains including variable stem- and loop-sizes. Positions of nucleotides are numbered according to conventional rules (1,7). Furthermore, an additional module was implemented providing statistical information for each alignment output to allow an easy comparison of individual sequences. According to the Mamit-tRNA database, consensus and typical structures of selected sequences can be calculated and displayed (3).

Table 1. Actual entries of the updated version of tRNAdb

Taxon	Organisms		tRNA genes			tRNA sequences		
	tRNA genes	tRNA sequences	Cytoplasm	Mitochondria	Chloroplast	Cytoplasm	Mitochondria	Chloroplast
Root	577	104	9758	1965	376	474	111	38
Cellular organisms	571	99	9705	1965	376	457	111	38
Bacteria	235	19	6368	0	0	139	0	0
Archaea	49	9	1088	0	0	76	0	0
Eukaryota	287	71	2249	1965	376	242	111	38
Viruses	6	5	53	0	0	17	0	0

Most conveniently, the retrieved data can be downloaded in a variety of file formats for further investigation using other applications. Export of sequences in FASTA (8), ClustalW (9) and Vienna RNA Package (10) file formats facilitates further analysis.

The representation of tRNA sequences poses additional challenges compared to those of the tRNA genes. More than 90 modified nucleosides have been characterized in tRNAs from Bacteria, Archaea and Eukarya (http://library.med.utah.edu/RNAmods/). Most of the base modifications are faithfully represented in the tRNA database. However, further processing of this information is not trivial, as the majority of RNA bioinformatics software is unable to cope with non-standard nucleotides. Hence, retrieved RNA sequences can be transformed into compatible DNA sequences.

DISCUSSION AND CONCLUSION

Well-curated and up-to-date databases are a highly useful tool of molecular biology and genetics. While the first tRNA database edition was a valuable instrument for the tRNA research community, the overwhelming amount of newly available sequences released by the variety of different genome sequencing projects made it necessary to develop a modern relational database system. In the new edition, all sequences of the original Excel-based compilation (http://www.trna.uni-bayreuth.de) as well as complete sets of tRNA gene sequences of several recently published archaeal genomes have been included. Furthermore, the standardized NCBI taxonomy system has been implemented, leading to high compatibility with other sequence databases. The new versatile search engine allows complex query combinations concerning sequence, structure and taxonomy, thus meeting the demands of systematic investigations of tRNA sequence/structure relationships. For the next edition of this compilation, proofreading of the remaining sequences (Eukarya and Archaea) will be completed. In addition, newly published tRNA genes and tRNA sequences will be imported. Possible extensions of the database are (i) inclusion of 5′- and 3′-flanking nucleotides to extract information on tRNA maturation (11), (ii) indication of tRNA introns (12), (iii) tools to extract identity elements for aminoacylation (13), (iv) indication of anticodon editing (14), (v) display of pathological tRNA mutations (3,15), (vi) information on posttranscriptional modifications with known roles in fine-tuning tRNA structure and function (16), (vii) display of isoacceptors and isodecoders [tRNAs with identical anticodon but sequence deviations elsewhere (17)], or (viii) information on tRNA expression levels [e.g. tissue-specific differences in eukaryotes (18)].

ACCESS

tRNAdb is freely accessible at http://trnadb.bioinf.uni-leipzig.de. This article should be cited in research projects assisted by the use of the database. Comments, corrections and new entries are welcome.

ACKNOWLEDGEMENTS

We are grateful to Catherine Florentz and Richard Giegé for stimulating discussions and insightful comments on the manuscript.

FUNDING

This work was funded by the Centre National de la Recherche Scientifique (CNRS), Université Louis Pasteur Strasbourg 1, Association Française contre les Myopathies (AFM), Deutsche Forschungsgemeinschaft [DFG - MO-634/2, MO-634/3, HA 1672/7-3/4/5, and SPP-1174 ('Metazoan Deep Phylogeny') project STA 850/3-2] and by the French-German PROCOPE program (DAAD D/0628236, EGIDE PHC 14770PJ). Funding for open access charge: CNRS and DFG.

Conflict of interest statement. None declared.

REFERENCES

1. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
2. Altschul,S.F., Gish,W., Miller,W., Myers,EW. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Pütz,J., Dupuis,B., Sissler,M. and Florentz,C. (2007) Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. *RNA*, **13**, 1184–1190.
4. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
5. Rozenski,J., Crain,P.F. and McCloskey,J.A. (1999) The RNA Modification Database: 1999 update. *Nucleic Acids Res.*, **27**, 196–197.
6. Grosjean,H. and Benne,R. (eds) (1998) *Modification and Editing of RNA*, ASM Press, Washington, DC.
7. Schimmmel,P.R., Söll,D. and Abelson,J.N. (eds) (1979) *Proposed Numbering System in tRNAs Based on Yeast tRNA^{Phe} in Transfer-RNA: structure, Properties and Recognition*, Cold Spring Harbor Laboratory, New York, pp. 518–519.
8. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
9. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
10. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
11. Hartmann,R.K., Gößringer,M., Späth,B., Fischer,S. and Marchfelder,A. (2008) The making of tRNAs and more - RNase P and tRNase Z. *Progr. Nucleic Acid Res. Mol. Biol.*, **85** (in press).
12. Marck,C. and Grosjean,H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189–1232.
13. Giegé,R., Sissler,M. and Florentz,C. (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.*, **26**, 5017–35.
14. Mörl,M., Dörner,M. and Pääbo,S. (1995) C to U editing and modifications during the maturation of the mitochondrial tRNA^{Asp} in marsupials. *Nucleic Acids Res.*, **23**, 3380–3384.

15. Scaglia,F. and Wong,L.J. (2008) Human mitochondrial transfer RNAs: role of pathogenic mutation in disease. *Muscle Nerve*, **37**, 150–171.
16. Gustilo,E.M., Vendeix,F.A. and Agris,P.F. (2008) tRNA. *Curr. Opin. Microbiol.*, **11**, 134–140.
17. Goodenbour,J.M. and Pan,T. (2006) Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res.*, **34**, 6137–6146.
18. Dittmar,K.A., Goodenbour,J.M. and Pan,T. (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet.*, **2**, e221.