Nucleic Acids Research, 2004, Vol. 32, Database issue ***D101–D103***
DOI: 10.1093/nar/gkh065

The European ribosomal RNA database

Jan Wuyts, Guy Perrière[1] and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium and [1]Laboratoire de Biométrie et Biologie Évolutive—UMR CNRS no. 5558, Université Claude Bernard, Lyon 1, France

Received September 20, 2003; Accepted September 26, 2003

ABSTRACT

The European ribosomal RNA database aims to compile all complete or nearly complete ribosomal RNA sequences from both the small (SSU) and large (LSU) ribosomal subunits. All sequences are available in aligned format. Sequence alignment is based on the secondary structure of the molecules, as determined by comparative sequence analysis. Additional information about the sequences, such as taxonomic classification of the organism from which they have been obtained, and literature references are also provided. In order to identify the closest relatives to newly determined sequences, BLAST searches can be performed, after which the best matching sequences are aligned and a phylogenetic tree is inferred. As of 2003, the European ribosomal RNA database is maintained at Ghent University (Belgium). The database can be consulted at http://www.psb.ugent.be/rRNA/.

THE RIBOSOMAL RNA MOLECULES

Like most functional non-coding RNA molecules, rRNA molecules fold into a well-defined and evolutionarily conserved secondary structure. The structure of the complete bacterial ribosome, determined by X-ray crystallography, largely confirms the secondary structure models that were determined by comparative analysis (1,2). In the case of large subunit (LSU) rRNA, the secondary structure consists of one large multi-branched loop from which 10 substructures branch off. These individual substructures are lettered A–J. In the small subunit (SSU) molecule, an RNA pseudoknot forms the center of the molecule. To facilitate the discussion of individual helices or groups of helices we have defined four core structures with respect to which insertions and deletions can be observed (3). These core structures are the common core, the Bacterial core, the Archaeal core and the Eukaryotic core. The common core consists of those helices that are found in Bacteria, Archaea and Eukarya, although not necessarily in all organisms of these kingdoms. On the secondary structure models of the molecules these helices have been given separate numbers in the case of SSU or a letter–number combination in the case of LSU. Additional helices that are specific to one or two kingdoms are given the suffix a (for archaeal), b (for bacterial), e (for eukaryotic), ab, ae or be, corresponding to the kingdom(s) in which they are found. These helices are also given a sub-number to differentiate between the many helices that are sometimes inserted at one position in the common core. For example, helix G5/e1 is a eukaryote-specific helix that is inserted after helix G5 in the secondary structure model. One more uncertainty remains in this convention: a helix inserted after helix G5 could be inserted directly after the 5′ strand of helix G5 or directly after the 3′ strand. To differentiate these two cases the number and sub-number of the helix are separated by a slash (/) in the former case and a backslash (\) in the latter.

Overall, the secondary structure pattern of the rRNA molecules is well determined. However, some of the variable areas or expansion segments (4) show too little sequence conservation to infer a reliable alignment. Consequently, the secondary structure model of these areas is often not well determined. Fortunately, these areas of unknown secondary structure and questionable alignment are mostly restricted to sequences that are present in a limited number of organisms (3).

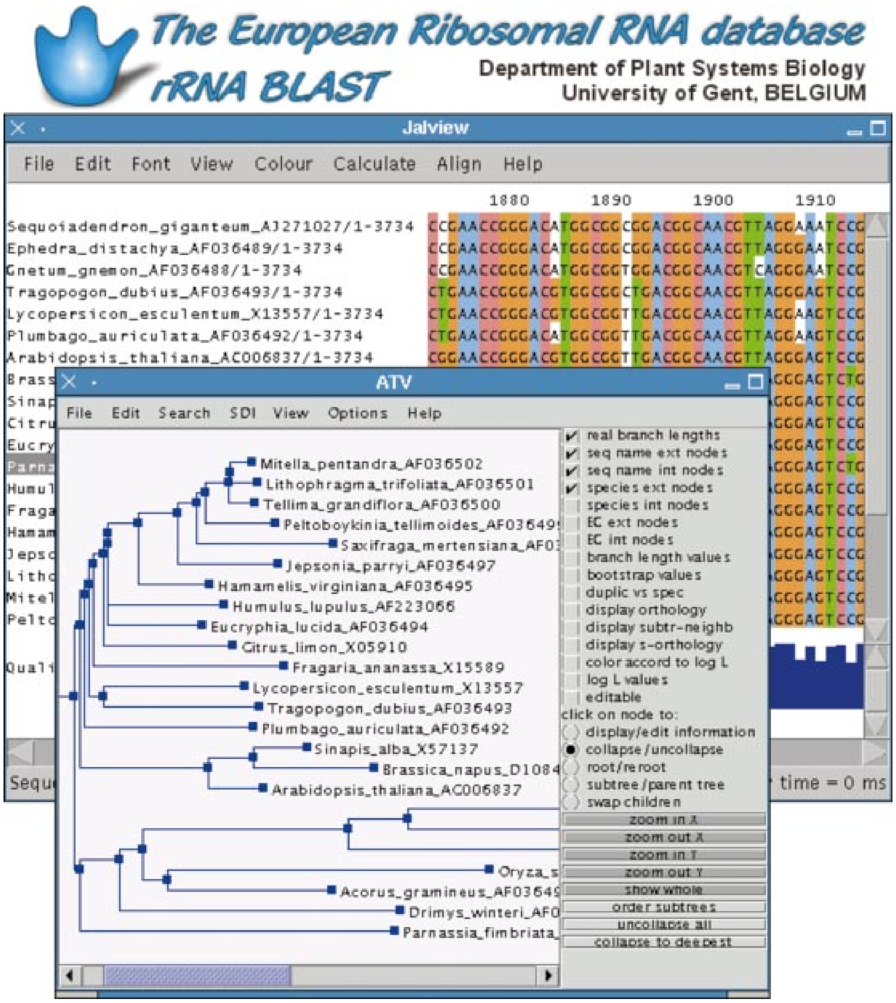
CONTENTS OF THE DATABASE

The European ribosomal RNA database is regularly updated by collecting all SSU and LSU rRNA database entries from the EMBL database (5) and now contains more than 13500/1100 prokaryotic (including plastids and mitochondria) and more than 6500/150 eukaryotic sequences for SSU/LSU, respectively. When adding new sequences to the database, the total length of the molecule is estimated by comparing it with the length of the rRNA of a close relative. If the rRNA of the new sequence is >70% of this length, the newly obtained sequence is added to the database. New and updated sequences are automatically aligned to the sequence of its closest relative. Afterwards, the secondary structure of the sequence of a close relative is superimposed on the newly determined sequence and corrected where necessary. Where appropriate, the primary structure alignment is adjusted to comply with the secondary structure.

ACCESSIBILITY

The European ribosomal RNA database was originally founded at the University of Antwerp (Belgium) in 1983 (6). From 2003 onwards, the database will be maintained at Ghent University (also in Belgium) and can be accessed at http://

*To whom correspondence should be addressed. Tel: +32 9 264 8756; Fax: +32 9 264 5349; Email: yves.vandepeer@psb.ugent.be



The European Ribosomal RNA database
rRNA BLAST
Department of Plant Systems Biology
University of Gent, BELGIUM
Jalview
File Edit Font View Colour Calculate Align Help
1880
1890
1900
1910
Sequoiadendron_giganteum_AJ271027/1-3734
Ephedra_distachya_AF036489/1-3734
Gnetum_gnemon_AF036488/1-3734
Tragopogon_dubius_AF036493/1-3734
Lycopersicon_esculentum_X13557/1-3734
Plumbago_auriculata_AF036492/1-3734
Arabidopsis_thaliana_AC006837/1-3734
time = 0 ms
ATV
File Edit Search SDI View Options Help
Mitella_pentandra_AF036502
Lithophragma_trifoliata_AF036501
Tellima_grandiflora_AF036500
Peltoboykinia_tellimoides_AF03649
Saxifraga_mertensiana_AF03
Jepsonia_parryi_AF036497
Hamamelis_virginiana_AF036495
Humulus_lupulus_AF223066
Eucryphia_lucida_AF036494
Citrus_limon_X05910
Fragaria_ananassa_X15589
Lycopersicon_esculentum_X13557
Tragopogon_dubius_AF036493
Plumbago_auriculata_AF036492
Sinapis_alba_X57137
Brassica_napus_D1084
Arabidopsis_thaliana_AC006837
Oryza_s
Acorus_gramineus_AF03649
Drimys_winteri_AF0
Parnassia_fimbriata
zoom in X
zoom out X
zoom in Y
zoom out Y
show whole
order subtrees
uncollapse all
collapse to deepest

☑ real branch lengths
☑ seq name ext nodes
☑ seq name int nodes
☑ species ext nodes
□ species int nodes
□ EC ext nodes
□ EC int nodes
□ branch length values
□ bootstrap values
□ duplic vs spec
□ display orthology
□ display subtr-neighb
□ display s-orthology
□ color accord to log L
□ log L values
□ editable

click on node to:
○ display/edit information
● collapse/uncollapse
○ root/reroot
○ subtree/parent tree
○ swap children

Figure 1. Example of the newly implemented tool to compare newly determined sequences with other sequences available in the rRNA databases. In this case, the LSU rRNA sequence of *Arabidopsis thaliana* (EMBL acc. no. AC006837) was BLAST-ed to the complete LSU rRNA database. Next, the 24 best scoring sequences were selected and an alignment was constructed and displayed in the Jalview applet (back). From this alignment, a neighbor-joining tree (11) was constructed and displayed by the use of the ATV applet [front (12)].

www.psb.ugent.be/rRNA/. All sequences in the database are stored in individual text files that contain both primary and secondary structure information as well as other information about the sequences such as literature references and taxonomic data. By user request, this format will automatically be translated into a number of other formats. Supported formats are: NBRF/PIR, a modified EMBL-based format, TREECON (7), DCSE (8), aligned FASTA and a printable (interleaved) alignment format. More formats will be added in the near future. To facilitate the retrieval of relevant sequences, three different interfaces are available to browse for sequences. Additionally, a BLAST (9) search can be performed after which matching sequences can be downloaded from the database.

A new service has been implemented for the identification of close relatives of a species for which the rRNA sequence is determined. The analysis of a newly determined sequence proceeds automatically and involves (i) a homology search for similar sequences by BLAST (9), (ii) sequence extraction and parsing, (iii) multiple sequence alignment by CLUSTALW (10) and (iv) tree construction by neighbor joining (11). In order to gain a significant increase in speed when aligning sequences, ClustalW starts from pre-aligned sequences that were produced by BLAST. Phylogenetic trees and alignments are displayed by the use of two Java applets, namely ATV (12) and Jalview (http://www2.ebi.ac.uk/~michele/jalview/). An example is shown in Figure 1. This option can be accessed from the main page by clicking ‘Quick phylogeny’ in the index column. Users can also browse the BLAST output in order to detect possible anomalies in the identification process.

Additional information available on the rRNA server includes:

(i) Detailed secondary structure models of prokaryotic, eukaryotic, mitochondrial and plastidial rRNA sequences.

(ii) Secondary structure variability maps of bacterial and eukaryotic rRNA sequences. The variabilities of individual nucleotides are calculated using the substitution rate calibration methods described previously (13,14).

(iii) Tertiary structure variability maps of bacterial rRNAs (3).

(iv) Software for sequence alignment (8), phylogenetic tree construction (7) and alignment format conversion (15).

(v) A database with information on PCR and sequencing primers.

ACKNOWLEDGEMENTS

We dedicate this paper to Professor E. M. Rupert De Wachter who founded the rRNA database in 1983 and supported it for almost two decades. We also acknowledge support from the Department of Plant Systems Biology at Ghent University and from the 'AlgaTerra' Project 01 LC 0026 of the BIOLOG program. J.W. is supported by a postdoc scholarship of the IWT (Belgium).

REFERENCES

1. Wimberly,B.T., Brodersen,D.E., Clemons,W.M.,Jr, Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
2. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
3. Wuyts,J., Van de Peer,Y. and De Wachter,R. (2001) Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Res.*, **29**, 5017–5028.
4. Hancock,J.M. and Dover,G.A. (1990) 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res.*, **18**, 5949–5954.
5. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
6. Erdmann,V.A., Huysmans,E., Vandenberghe,A. and De Wachter,R. (1983) Collection of published 5S and 5.8S ribosomal RNA sequences. *Nucleic Acids Res.*, **11**, r105–r133.
7. Van de Peer,Y. and De Wachter,R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.*, **10**, 569–570.
8. De Rijk,P. and De Wachter,R. (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. *Comput. Appl. Biosci.*, **9**, 735–740.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. 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.
11. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
12. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
13. Van de Peer,Y., Neefs,J.M., De Rijk,P. and De Wachter,R. (1993) Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. *J. Mol. Evol.*, **37**, 221–232.
14. Van de Peer,Y., Van der Auwera,G. and De Wachter,R. (1996) The evolution of stramenopiles and alveolates as derived by 'substitution rate calibration' of small ribosomal subunit RNA. *J. Mol. Evol.*, **42**, 201–210.
15. Raes,J. and Van de Peer,Y. (1999) ForCon: a software tool for the conversion of sequence alignments. EMBnet news, **6**, (http://www.hgmp.mrc.ac.uk/embnet.news/vol6_1/ForCon/forcon.html).