Published online 11 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D377–D380***
doi:10.1093/nar/gkm827

The ITS2 Database II: homology modelling RNA structure for molecular systematics

Christian Selig, Matthias Wolf, Tobias Müller, Thomas Dandekar and Jörg Schultz*

Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland 97074 Würzburg, Germany

Received August 14, 2007; Revised and Accepted September 20, 2007

ABSTRACT

An increasing number of phylogenetic analyses are based on the internal transcribed spacer 2 (ITS2). They mainly use the fast evolving sequence for low-level analyses. When considering the highly conserved structure, the same marker could also be used for higher level phylogenies. Furthermore, structural features of the ITS2 allow distinguishing different species from each other. Despite its importance, the correct structure is only rarely found by standard RNA folding algorithms. To overcome this hindrance for a wider application of the ITS2, we have developed a homology modelling approach to predict the structure of RNA and present the results of modelling the ITS2 in the ITS2 Database. Here, we describe the database and the underlying algorithms which allowed us to predict the structure for 86 784 sequences, which is more than 55% of all GenBank entries concerning the ITS2. These are not equally distributed over all genera. There is a substantial amount of genera where the structure of nearly all sequences is predicted whereas for others no structure at all was found despite high sequence coverage. These genera might have evolved an ITS2 structure diverging from the standard one. The current version of the ITS2 Database can be accessed via http://its2.bioapps.biozentrum.uni-wuerzburg.de.

INTRODUCTION

The internal transcribed spacer 2 (ITS2) of the nuclear rRNA cistron is a widely used phylogenetic marker. As its sequence evolves comparably fast, it is mainly used for low-level analyses. Contrasting the sequence, the structure of the ITS2 is highly conserved. The hallmarks, namely four helices with the third as the longest, have been found in detailed exemplary studies (1) as well as in large-scale analyses (2). This lead to the suggestion to enlarge the application field to higher taxonomic levels (3). In addition to these phylogenetic analyses, a specific structural feature between two ITS2, a compensatory base change (CBC), can be used to distinguish two species from each other (4). This underlines the importance of considering not only the sequence but also the structure when performing any analysis based on the ITS2. But the proposed correct structure is only rarely automatically found by standard minimum free energy folding (MFE) (2). To overcome this hindrance for the wider application of the ITS2, we developed a homology-based structure modelling approach, which allowed predicting the structure for 20 000 sequences which were not found by RNAfold (5). As these can be used as a basis for any phylogenetic analysis, we have developed the ITS2 Database as a resource for sequence and structure information of the ITS2 (6). Here we report modifications and improvements of the database which allowed us to find structural information for 86 784 ITS2 sequences, which is 55% of all entries concerning ITS2 in GenBank.

RESULTS AND DISCUSSION

Rebuild and updates

In the first version of the database, every sequence whose correct structure could not be found by RNAfold was searched against the original set of 5 000 sequences with correct RNAfold based structures (2) to identify possible templates for homology modelling (models). As a first step in the development of the new version of the database, we checked whether there were additional novel sequences in GenBank whose structure could be determined directly by RNAfold. Indeed, we found a 2-fold increase in the amount of correctly predicted structures (Table 1, Method 1). We used this dataset as a starting point for a complete rebuild of the database. More importantly,

*To whom correspondence should be addressed. Tel: +49 0 931 888 4553; Fax: +49 0 931 888 4552;
Email: Joerg.Schultz@biozentrum.uni-wuerzburg.de
Correspondence may also be addressed to Matthias Wolf. Tel: +49 (0) 931 888 4562; Email: Matthias.Wolf@biozentrum.uni-wuerzburg.de
Correspondence may also be addressed to Tobias Müller. Tel: +49 (0) 931 888 4563; Email: Tobias.Mueller@biozentrum.uni-wuerzburg.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Methods used for ITS2 structure prediction and number of folded sequences.

Method	Description	Count
1	Direct RNAfold	10 667
2	Homology modelling, first iteration	27 044
3	Homology modelling, second iteration	11 306
4	Direct RNAfold, sequence discovery by BLAST	5 196
5	Homology modelling, first iteration, sequence discovery by BLAST	1 730
6	Homology modelling, second iteration, sequence discovery by BLAST	17 776
7	Partial structures from homology modelling, both iterations	13 065
	Total	86 784

this result led us to a change in the logic and therefore to a re-design of the update procedure. Each time the structure of an incoming sequence can be predicted directly by RNAfold or in the first round of homology modelling, it is added to the set of models. Thus, no core sequence/structure set (as before the set of 5 000) is existent any more but a dynamically growing set of possible structure models. In summary, this approach together with a second iteration of homology modelling allowed us to predict 38 350 structures (Table 1, Methods 2 and 3).

Reannotation of GenBank entries

A prerequisite for a phylogenetic analysis is the correct localization of the ITS2. If the boundaries are incorrect, missing or additional sequence fragments might be considered as a specific feature of an organism leading to a wrong phylogenetic classification. With the correct structure at hand, the boundaries of the ITS2 can be exactly determined, again underlining the importance of considering structure for phylogenetic analyses. Accordingly, already in the first version of the database, a CLUSTALW-based approach (7) was used to extend the sequence if the GenBank annotation missed the first or the last helix. As this approach was limited to cases where (i) there exists a feature annotation by GenBank and (ii) the homology modelling was of high quality for the other helices, we developed a novel, BLAST-based approach (8) for the re-annotation of GenBank entries. For the cases where no structure could be predicted for the ITS2 as annotated by GenBank, the whole GenBank entry is retrieved and searched against all sequences with known structures using BLAST. If a significant hit is found (E-value $\leq 10e^{-16}$), the homologous region of the query is cut. This fragment builds the basis for a second round of structure prediction. RNAfold is used to test whether this fragment can be folded in the correct structure. If not, homology modelling is used to find the correct structure. By this method, we were able to re-annotate the position of the ITS2 in 6901 GenBank entries. In most cases, this structure-based annotation lead to a slight shift of the 5' or/and the 3' end of the ITS2, but some entries were heavily shifted (Figure 1). For example the ITS2 of *Trifolium affine* (GI:85724133) is incorrectly annotated

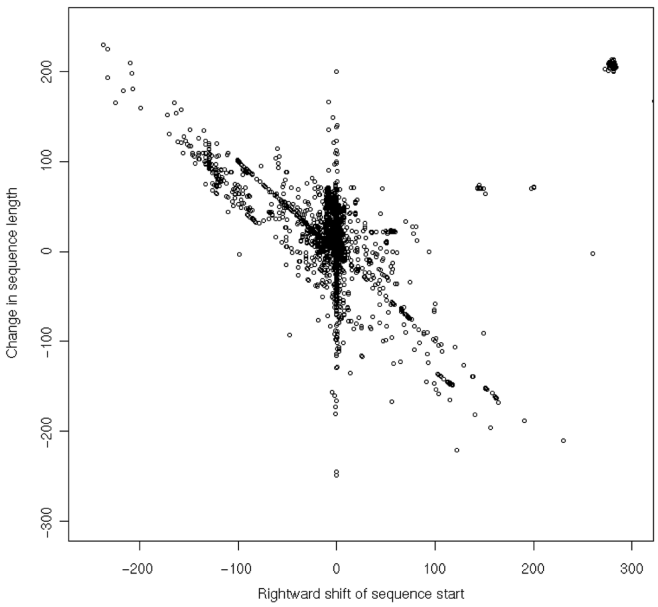

Figure 1. Re-annotated sequences, each dot representing a successfully predicted secondary structure—*X*-axis represents shift in the 5' end of the ITS2, *Y*-axis change of the length compared to the GenBank annotation. The cluster in the upper right corner consists of 206 sequences from *Trifolium spec*. Six outliers (GI: 5814072, 57999795, 2896060, 13507073, 4006937, 85724147) are not shown.

with a length of just 7 bp, its preceding 5.8S ribosomal RNA with 9 bp. Accordingly, length and position were re-annotated to 215 bp. These cases underline the advantage of the structure-based annotation compared to one based on sequence information alone.

In contrast to the method used in the previous version of the database, the BLAST-based approach is completely independent of any pre-annotated ITS2. This allowed us to locate the position of the ITS2 in any GenBank entry. Application to all entries containing the search term 'internal transcribed spacer 2' or 'ITS2' without a feature annotation lead to the new annotation of 17 801 ITS2 sequences.

Partial structures

Many of the sequences without predicted structure were fragments, i.e. they missed at least one helix of the structural hallmark and therefore did not fulfil the quality control of the standard homology modelling. Still, these sequences could increase the coverage of a systematic analysis. In contrast to the MFE approach, our homology-modelling algorithm is able to predict the structure of fragments. To assure a sufficient quality, only entries where at least two consecutive helices could be modelled with sufficient quality ($\geq$75%) were accepted. This method resulted in additional 13 065 ITS2 sequences with structural information (Table 1, Method 7).

ITS2-specific matrix

The existence of a large number of pairwise alignments allowed us to calculate ITS2-specific evolutionary models. Based on variants of the methods described in Müller and

Vingron (9) and Müller *et al*. (10), we were able to derive an ITS2-specific substitution model, which is an important ingredient for phylogenetic analyses. This model reflects nicely the special features of RNA and in particular, ITS2 sequence evolution. Based on this molecule-specific substitution model, an ITS2-specific scoring matrix is derived that strongly deviated from the unity matrix as used as default, for example in BLAST. To test the influence of this matrix compared to the standard identity matrix, we performed all calculations with the standard and the ITS2-specific matrix, respectively. In the GenBank version used in the test run, structural information was found for 57 680 sequences whereas the usage of the ITS2-specific matrix resulted in 76 721 structures. This underlines the importance of the correct evolutionary model in the homology modelling of ITS2 and presumably other RNA sequences. Accordingly, the ITS2-specific score matrix is now used in all calculations for the ITS2 database and can be downloaded from the web site as Supplementary Data.

Custom modelling

The process of homology modelling as described in Wolf *et al*. (5) is in principle applicable for any RNA sequence family. We therefore have added the possibility for 'Custom Modelling' to the web site. Here, the user provides an RNA sequence with a known structure and other, homologous sequences. For these, a homology model is calculated based on the known structure. When using this feature, it has to be taken into account that there is, in contrast to the modelling of ITS2, no quality measure for the model. Thus, it is the obligation of the user to check the validity of the results.

CONCLUSIONS

With the modifications of the ITS2 database outlined above, the structural features of 86 784 sequences were predicted, which was ~55% of all GenBank entries concerning the ITS2. As this number gives just an overall average, we tested the coverage of predicted structures within all genera. A clear separation was found between genera where the structure for nearly all sequences could be predicted and others, where no structure was found despite considerable sequence coverage (Figure 2). We suggest that in these genera the structure of the ITS2 deviates from the standard. This notion is supported by their length distribution being nearly equal to the length distribution of successfully folded sequences (data not shown). Furthermore, within those genera without any structural data, there is a strong bias towards metazoans (11). This is consistent with the observation that vertebrates have a more complex structure than the one described by Coleman (1). The latter one fits mostly for plants and fungi, taxa whose genera are strongly represented in the overall number of genera with structural data.

How could a user, who is interested in the phylogeny of a specific taxonomic group, use the ITS2 database? If he starts with an already known sequence, he can directly extract the corresponding structure from the database ('Search by GI/Accession/Taxon'). If he has sequenced his own organisms, he should first homology model the structure of this sequences ('Predict ITS2 Structure'). Second, he can extract ITS2 sequences and their structures for further organisms in the taxonomic group of interest ('Browse Taxonomy'). This will result in a set of ITS2 sequences with corresponding structures. In the third step, these have to be aligned. Here, an alignment program, which considers both sequence and structure, like 4SALE (12), will be suitable. Manual optimization of the sequence–structure alignment can be performed in the editor of this program. Finally, this sequence–structure-based alignment will be the input for standard phylogenetic analyses, e.g. in PAUP (13) or PHYLIP (14). Furthermore, one is now able to check for CBCs to distinguish possible different species in the dataset (4) or to calculate CBC trees (15).

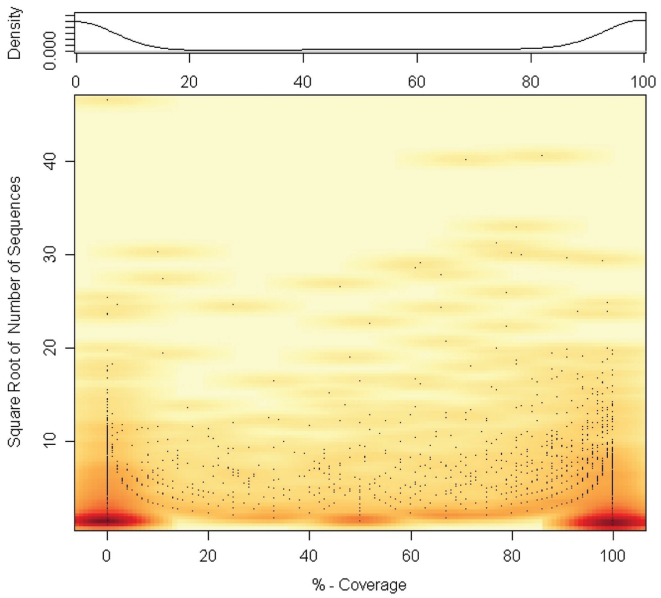

Figure 2. Structure coverage—each point indicates one genus. On the *Y*-axis, the square root of the number of sequences in the genus is indicated. On the *X*-axis, the percentage of correct structures for all sequences of the genus is plotted. Additionally on top of the scatter plot, a density plot is shown reflecting the coverage distribution over all genera. The colouring indicates the relative frequencies. A concentration of points at 50% is caused by genera containing only two sequences. A similar, less pronounced effect can be seen at 33.3% and 66.6% for genera with three sequences.

ACKNOWLEDGEMENTS

We would like to thank Philip Seibel for integration of 4SALE with the ITS2 Database. Parts of this work were funded by the Deutsche Forschungsgemeinschaft (DFG), grant Mu 2831/1-1 (Species phylogeny and the 'tree of life' based on an ITS2 sequence-structure Database and new algorithms).

Conflict of interest statement. None declared.

REFERENCES

1. Coleman,A.W. (2007) Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Res.*, **35**, 3322–3329.
2. Schultz,J., Maisel,S., Gerlach,D., Muller,T. and Wolf,M. (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, **11**, 361–364.
3. Coleman,A.W. (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.*, **19**, 370–375.
4. Muller,T., Philippi,N., Dandekar,T., Schultz,J. and Wolf,M. (2007) Distinguishing species. *RNA*, **13**, 1469–1472.
5. Wolf,M., Achtziger,M., Schultz,J., Dandekar,T. and Muller,T. (2005) Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, **11**, 1616–1623.
6. Schultz,J., Muller,T., Achtziger,M., Seibel,P.N., Dandekar,T. and Wolf,M. (2006) The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.*, **34**, W704–W707.
7. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
8. 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.
9. Muller,T. and Vingron,M. (2000) Modeling amino acid replacement. *J. Comput. Biol.*, **7**, 761–776.
10. Muller,T., Spang,R. and Vingron,M. (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, **19**, 8–13.
11. Joseph,N., Krauskopf,E., Vera,M.I. and Michot,B. (1999) Ribosomal internal transcribed spacer 2 (ITS2) exhibits a common core of secondary structure in vertebrates and yeast. *Nucleic Acids Res.*, **27**, 4533–4540.
12. Seibel,P.N., Muller,T., Dandekar,T., Schultz,J. and Wolf,M. (2006) 4SALE – a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, **7**, 498.
13. Swofford,D. (2002). *PAUP* Phylogenetic Analysis Using Parsimony (*and other methods) Version 4.0b10 win32*. Sinauer Associates, Sunderland.
14. Felsenstein,J. (2005). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
15. Wolf,M., Friedrich,J., Dandekar,T. and Muller,T. (2005) CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biol.*, **5**, 291–294.