Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures

MATTHIAS WOLF, MARCO ACHTZIGER, JÖRG SCHULTZ, THOMAS DANDEKAR, and TOBIAS MÜLLER

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

ABSTRACT

Structural genomics meets phylogenetics and vice versa: Knowing rRNA secondary structures is a prerequisite for constructing rRNA alignments for inferring phylogenies, and inferring phylogenies is a precondition to understand the evolution of such rRNA secondary structures. Here, both scientific worlds go together. The rRNA internal transcribed spacer 2 (ITS2) region is a widely used phylogenetic marker. Because of its high variability at the sequence level, correct alignments have to take into account structural information. In this study, we examine the extent of the conservation in structure. We present (1) the homology modeled secondary structure of more than 20,000 ITS2 covering about 14,000 species; (2) a computational approach for homology modeling of rRNA structures, which additionally can be applied to other RNA families; and (3) a database providing about 25,000 ITS2 sequences with their associated secondary structures, a refined ITS2 specific general time reversible (GTR) substitution model, and a scoring matrix, available at <http://its2.bioapps.biozentrum.uni-wuerzburg.de>.

Keywords: homology modeling; internal transcribed spacer 2 (ITS2); phylogeny; RNA; secondary structure

INTRODUCTION

The internal transcribed spacer 2 (ITS2) region is a standard marker in phylogenetics. However, because of its high sequence variability with an assumed conservation in structure, it is a double-edged tool for eukaryote evolutionary comparisons (Alvarez and Wendel 2003; Coleman 2003). Utilizing its high mutation rate acting on the primary sequence, it is currently used for species level classifications. However, its conservation in structure—four helices with helix III as the longest—might boost its value as a marker for megasystematics (Coleman 2003; Schultz et al. 2005). For both applications, it is absolutely necessary to consider both—sequence and structure—when calculating align-

ments and resulting phylogenies. To date, there are about 90,000 ITS2 sequences in GenBank (NCBI). But, for only about 5000 (from at that time 70,000 ITS2 sequences) was the common core secondary structure predicted with standard structure prediction algorithms (Schultz et al. 2005).

On the one hand, the remaining 85,000 ITS2 structures could, in principle, be rather different and have species or organism specific solutions for ITS2 processing. On the other hand, all available data on rRNA structure assembled to date support that rRNA processing is a very conserved process. To show and verify this for the ITS2 region requires detecting or rejecting conserved common RNA secondary structures in thousands of sequences over a large divergent phylogenetic range. We show that, in fact, 20,000 ITS2 rRNA sequences have a conserved RNA processing structure. Therefore, the data set derived in this study should constitute an optimal starting point for any ITS2 based phylogenetic analysis and clearly for ITS2 structure prediction.

RESULTS AND DISCUSSION

Homology modeling

At the time of the study, 91,873 ITS2 sequences were annotated in GenBank (NCBI) but the common core secondary

Reprint requests to: Matthias Wolf, Department of Bioinformatics, University of Würzburg, Biocenter, Am Hubland, D-97074 Würzburg, Germany; e-mail: matthias.wolf@biozentrum.uni-wuerzburg.de; fax: +49 931 888 4552; Jörg Schultz, Department of Bioinformatics, University of Würzburg, Biocenter, Am Hubland, D-97074 Würzburg, Germany; e-mail: joerg.schultz@biozentrum.uni-wuerzburg.de; fax: +49 931 888 4552; Thomas Dandekar, Department of Bioinformatics, University of Würzburg, Biocenter, Am Hubland, D-97074 Würzburg, Germany; e-mail: dandekar@biozentrum.uni-wuerzburg.de; fax: +49 931 888 4552; or Tobias Müller, Department of Bioinformatics, University of Würzburg, Biocenter, Am Hubland, D-97074 Würzburg, Germany; e-mail: tobias.mueller@biozentrum.uni-wuerzburg.de; fax: +49 931 888 4552.

Article and publication are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.2144205>.

structure could hitherto be predicted only for about 5000 ITS2 sequences with standard RNA folding programs (Schultz et al. 2005). Even discarding fragments and spurious sequences leaves 47,683 ITS2 sequences with unknown secondary structures. Here, we aim for the prediction of these structures by exploiting possible homology to sequences with already assigned, correct structures following the algorithm outlined in Figure 1. Due to the high variability at the sequence level, one is not able to construct correct multiple alignments that constitute the training data

for SCFG (stochastic context free grammar) (Eddy and Durbin 1994). In particular, the SCFG approach would yield a consensus structure, whereas we want to derive a concrete secondary structure. To this end we clustered all 47,683 sequences with unknown structures together with the 5003 sequences of known structures and checked each cluster for the presence of sequences with assigned structures. The best alignment hit (minimal p value) for the template sequence with its assigned structure was used to transfer the structure to the homologous sequences. To ensure high prediction

performance, the structure was transferred only if >75% of all base pairs for each helix could be adopted (Fig. 2A). To improve the quality of the model, post-processing of the predicted structure was performed. First, free base-pairings were closed following the Nussinov algorithm (Nussinov and Jacobsen 1980; Fig. 2B). Second, the structure was scanned for structural elements that are not physically possible like closed helix loops, which were fixed (Fig. 2C). Third, too long sequences were cut, leading to a reannotation of the GenBank entry (Fig. 2D). This process resulted in 16,214 predicted novel structures. An additional 21,807 sequences did show a significant sequence identity to at least one template sequence, but structure adoption did not satisfy our criteria. In a subset, the first or the last helix was totally missing, whereas for the other helices the structure could be predicted for >75% of the bases. As this could be indicative for false annotations in GenBank, the corresponding database entry was retrieved and the full sequence was aligned. For a total of 794 sequences, this approach not only allowed the prediction of the correct structure, but it also improved the original annotation (Fig. 2E). To exploit these novel structures, the whole algorithm was repeated, adding the novel structures to the templates, resulting in 5533 further predicted structures. These consist of 4745 structures that could be predicted directly from a template structure plus 788 sequences for which the annotation of the ITS2 region had to be modified. To ensure the quality and to avoid error propagation in this repeated prediction, we decided to add only structures found before the post-processing to the template database. Additionally, none of the structures derived after reannotation was used.

FIGURE 1. A flowchart illustrating our strategy in homology modeling. At each step a set of sequences was rejected and the remaining ones in their homology prediction refined. Further explanations are given in the text (see also Fig. 2).

FIGURE 2. A set of figures describing various refinements and modeling steps in the algorithm. (A) Homology modeling. (I) One of the 5000 correctly folded ITS2 secondary structures described in Schultz et al. (2005). (II) Two homologous hits according to I showing wrongly predicted ITS2 secondary structures derived by RNAfold. (III) The second one after homology modeling—i.e., the structure of I was partially adopted. (B) Closing free pairings. (IV) Clearly, there are some free base-pairings left after structure adoption. (V) These are closed by the Nussinov algorithm (Nussinov and Jakobsen 1980). (C) Opening helix loops. (VI) Local folding results in structural elements that are not physically possible such as closed helix loops. (VII) Such structural errors are fixed. (D) Cutting ends. (VIII) Homology modeling yields due to additional information (sequence alignment and structure adoption) a correctly folded ITS2 secondary structure, although the annotated sequence is too long. (IX) The free ends of sequence VIII were cut to correct the annotation errors described in VIII. (E) Enlarging sequence. (X) An annotated ITS2 sequence that is too short (either at the beginning or at the end); the homology modeling works well, but only on three of the four helices. (XI) Going back to GenBank the ITS2 sequence is reannotated based on the flanking 5.8S or 28S rRNA sequences.

In summary, examining homology-based helix and structure modeling of 50,000 sequences revealed more than 20,000 homologous ITS2 secondary structures that, by definition, could not be modeled by standard RNA folding programs like RNAfold or mfold (Zuker and Stiegler 1981; Hofacker et al. 1994), but would instead have yielded wrongly folded predictions that would not match the common ITS2 pattern. Recently, new algorithms such as Alifold (Hofacker et al. 2002) have been published to predict a consensus secondary structure for a set of pre-aligned RNA sequences. However, this method would fail on our task, because it cannot exploit template structures to check for correct structure transfer.

A further 30,000 ITS2 sequences were rejected as they do not match any template in our database.

A database providing a total of about 25,000 ITS2 sequences and their predicted secondary structures (5000 templates and 20,000 new models) is available at <http://its2.bioapps.biozentrum.uni-wuerzburg.de>.

These available structures have been modeled as accurately as possible but represent only template-based predictions and thus may not be as accurate as experimentally verified RNA structures.

However, to test the accuracy of the prediction method we took one of the 5092 known template sequences (Schultz et

al. 2005) and repredicted its structure by the proposed homology modeling algorithm. Then the predicted structure was compared to the “true” structure of the template sequence by counting position-wise structural matches. This procedure was repeated for all 5092 available template sequences. According to our filters (p value and base pair adoption) for 4716 sequences (93%), a reprediction was possible. For those we got on average 97% structural identities between the “true” and the modeled structure. This represents a position-wise 97% prediction accuracy.

Characterization of predicted ITS2 structures

Having derived a data set of about 25,000 homologous ITS2 sequences together with their secondary structures, we are now able to evaluate the quality of the models and to describe ITS2 specific features. Beside the universally conserved U-U mismatch in helix II and the UGGU motif near the 5' site apex of helix III (Table 1), these are (1) the percentage of adopted base pairs per helix (homology modeling), (2) the free energy of the predicted structures, and (3) the helix length distribution.

Box plots in Figure 3A elucidate the homology modeling prediction performance per helix. Focusing on the medians,

TABLE 1. Taxa distribution of total number of the ITSs and consensus folds (updated according to Schultz et al. 2005)

Taxon	Total ITS2	Consensus fold (Schultz et al. 2005)	Consensus fold (this study)	Consensus fold (all)	UGGU motif	U-U mismatch	UGGU and U-U
Rhodophyta	318	0	1	1	0	0	0
-Florideophyceae	274	0	1	1	0	0	0
Viridiplantae	33,969	3962	15,036	18,998	16,883	14,248	13,383
-Chlorophyta	1427	90	144	234	35	102	29
-Streptophyta	32,542	3872	14,892	18,764	16,848	14,146	13,354
Fungi/Metazoa Group	31,858	1046	6053	7099	2243	3490	1601
-Metazoa	6974	53	141	194	158	151	129
--Eumetazoa	6626	52	131	183	157	151	129
--Bilateria	5678	46	128	174	154	143	127
--Cnidaria	931	3	3	6	0	6	0
-Fungi	24,884	993	5912	6905	2085	3339	1472
--Glomeromycota	776	63	340	403	106	332	101
--Zygomycota	396	8	1	9	0	5	0
--Ascomycota	16,312	551	3598	4149	636	1261	399
--Basidiomycota	7070	360	1901	2261	1318	1719	965
Alveolata	1032	60	58	118	13	94	10
-Dinophyceae	636	59	58	117	12	93	9
-Apicomplexa	132	1	0	1	1	1	1
-Ciliophora	165	0	0	0	0	0	0
Stramenopiles	1784	21	0	21	20	20	19
-Phaeophyceae	496	0	0	0	0	0	0
-Oomycetes	1142	21	0	21	20	20	19
Heterolobosea	102	0	0	0	0	0	0

Only taxa with more than 100 ITS2 sequences are shown. Subgroups are indicated by a – in the taxon column, that is Bilateria and Cnidaria are subgroups of the Eumetazoa. Additional columns count the number of sequences with two ITS2 specific motifs (a U-U mismatch in helix II and a UGGU sequence pattern in helix III (cf. Coleman 2003)).

FIGURE 3. ITS2 statistics. (A) Box plots of the percentage of adopted base pairs per helix. In a box plot, the box shows the interquartile range (IQR) of the data, whereas the line through the box shows the median. The IQR is defined as the difference between the 75th percentile and the 25th percentile and is therefore a robust measure of the variance of the data. Focusing on the medians, helices II and IV show near optimal prediction performance (medians at 100%, as more than 50% of the structures could be assigned), whereas helices I and III are slightly worse (median at 94% and 95%, respectively). However, the variances of the prediction favor helices II and III. (B) Box plots of the helix-wise length distribution of the predicted 20,000 ITS2 structures. The lengths' medians fit the well-known four-fingered hand structure of the ITS2, with helix III as the longest (Coleman 2003). On average, the total length of the ITS2 was 208 bases. (C) This figure correlates predicted secondary structure energies as derived by RNAfold to folding energies derived by homology-based structure prediction. Note that we examine here only the structures in which the folding predicted by RNAfold was wrong with respect to the core structure of the ITS2. Therefore, this picture demonstrates the degree of deviance between optimal template free energy folding according to RNAfold and the energy of more likely structures given the known 5000 homologous template secondary structures that take biological constraints of the ITS2 such as rRNA and protein interactions into account (as an example, one ITS2 secondary structure is given in the background; helices are numbered I–IV). To clarify the correlation between both energy levels, illustrated with thin lines for the complete data set, we embedded smooth local regression lines (LOWESS as implemented in the R statistical computing environment, from R Foundation for Statistical Computing, Vienna, Austria) showing the following trend: The lower the energy, the more both methods coincide (red indicates first iteration, green indicates second iteration). The ITS2 structure is critical for rRNA processing and, hence, is extremely well-conserved. Focusing on the structure energy according to ITS2 base-pairings alone is not sufficient, while careful template-based modeling captures many structural constraints resulting in a sixfold increase in the number of available predicted structures (cf. database).

helices II and IV show near optimal prediction performance (medians at 100%, i.e., >50% of structures were completely assigned), whereas helices I and III are more difficult to model and perform slightly worse (median at 94% and 95%, respectively). However, the variances of the prediction favor helices II and III. Notches show a small 95% confidence interval for the medians estimate.

The box plots in Figure 3B show that the medians for the identified helix lengths fit the well-known four-fingered hand structure of the ITS2, with helix III as the longest (cf. Coleman 2003).

There are structures in which the folding predicted by RNAfold was wrong with respect to the core structure of the ITS2. The correlation of predicted secondary structure energies as derived by RNAfold to folding energies derived by homology-based structure prediction is shown in a scatterplot (Fig. 3C). It demonstrates the degree of deviance between optimal template free energy folding according to RNAfold and the energy of the homology modeled structures. Obviously, the energy of homology-based structures has to be higher than that of the optimal structure, and structures found in the first iteration have in general lower energies than those found in the second iteration (Fig. 3C). The cloud of secondary structures in the lower right corner (Fig. 3C) that shows a large difference (~ 80 kcal/mol) in the optimal free energy by usual RNA folding tools and the energy according to homology modeling was checked by looking at an adequate sample ($n = 20$) of such structures. Even those were confirmed to fit the expected biological constraints—that is, four helices with helix III as the longest. This indicates that the homology-based predicted structures represent the biologically active form, as the ITS2 secondary structure is critical for rRNA processing, involved in several RNA–RNA and RNA–protein interactions, and hence extremely well conserved. Focusing on the structure energy by standard base-pairing alone is not sufficient. Careful template-based modeling captures many structural constraints including complex interactions required for processing. This results in a sixfold increase in the number of structures that now can be assigned to their template.

For the wide application of ITS2 as a phylogenetic marker, a broad taxonomic distribution is crucial. Therefore, we tested the prevalence of the novel predicted structures for different taxonomic groups and compared it with the distribution of the energy-based predicted structures used as templates (Table 1). We observe that the sixfold increase in successfully predicted secondary structures is consistently reflected in the homogenous increase of structures per taxon (Fig. 4). This suggests a sixfold higher prediction power of the homology modeling-based approach, independent of the taxonomic grouping.

Now having a large number of homologous correctly pairwise aligned sequences at hand, we are able to derive a ITS2 specific general time reversible (GTR) substitution model

FIGURE 4. A scatter plot of energy-based predicted structures per taxonomic group versus homology modeling-based structures per taxonomic group. Regression analysis reveals that a quadratic curve fits the data, whereas omitting the largest groups—even a linear function with a slope of six—fits the data well, as is demonstrated by diagnostic plots (data not shown).

(Lanave et al. 1984; Müller and Vingron 2000) and scoring matrices. The substitution model is one of the most important ingredients for phylogenetic studies. Typically, such a matrix is derived on small data sets and, therefore, it cannot really reflect the typical ITS2 specific substitution behavior.

However, a refined substitution model can be derived from the large data set that is available now, that is, beside all sequences with their predicted secondary structures; a substitution model and a scoring matrix is also available at <http://its2.bioapps.biozentrum.uni-wuerzburg.de>.

CONCLUSIONS

Structural genomics concentrates often only on protein structures. Here, RNA structure modeling starting from a carefully designed template database (Schultz et al. 2005) is performed. Five-thousand known sequences and secondary structures as well as structure selection filters allow us to reveal more than 20,000 ITS2 secondary structures in about 14,000 organisms. Such obtained structures provide a basis for large-scale phylogenetic studies.

Obviously, for the homology modeling-based approach, an ab initio correctly folded secondary structure template is essential. This means that both methods are important for themselves but have to go with each other to achieve the tremendous increase in structure prediction. For instance, combining both we can now predict about 3534 ITS2 structures in the Asteraceae, from which recently 340 have been described by Goertzen et al. (2003). The consensus structure obtained by their comparative approach (Fig. 1 of Goertzen et al. 2003) matches well our predictions for the Asteraceae ITS2.

Having biodiversity in mind, we expect and know there are additional variants (today about 30,000) deviating from our structural template. As already discussed in Schultz et al. (2005), there are known ITS2 structures having, for example, only three helices, branching ones, or more than four. Pseudogenes (Alvarez and Wendel 2003), with their rapid rate of sequence decay, may also deviate from our structure template, and similarly ITS2 sequences of >150 nt yet only partially available in the database will also be rejected by the algorithm. According to the used templates, such structures cannot and should not be found by our homology modeling approach. Indeed, these structures were correctly rejected during the template-based modeling. Mapping the structure variations to the tree of life would dramatically elucidate our knowledge about structure evolution and RNA processing. The method presented here can accommodate further templates as they became available.

From a practical point of view, we present in this article a data set that can and should be used as a starting point for any ITS2-based phylogenetic analysis and ITS2 structure prediction. It yields related ITS2 sequences and secondary structures and therefore constitutes a guideline for the alignment construction, which is essential for a good phylogenetic study.

From a theoretical point of view, we propose a powerful method that is actually independent of the considered RNA molecule and can therefore be easily transferred to any other RNA family, other than rRNA. So, it would be highly desirable to achieve a tree-of-life-wide database of other important RNA structures.

MATERIALS AND METHODS

Data ascertainment

All entries with the annotation “internal transcribed spacer 2” or “ITS2” were retrieved from GenBank and ITS2 subsequences were extracted as annotated. We discarded fragments and spurious sequences by assuming typical sequence lengths known from the already described structures (Schultz et al. 2005).

Sequence alignments

As these ITS2 sequences are globally homologous, global all-against-all pair-wise alignments were calculated according to the Needleman and Wunsch algorithm (Needleman and Wunsch 1970) using a GeneMatcher2 system (Paracel, Inc.) with default settings. Alignments with a very conservative p value of $<10^{-16}$ were extracted.

Clustering

Aligned sequences showing global sequence similarity and finding each other directly or indirectly by a cascade were clustered.

Folding

(Cf. Fig. 1 for further details and steps.) Within a cluster we searched for one of the 5003 sequences with known structures to iteratively adopt that structure throughout the cascade of homologous sequences. If a cluster contained more than one sequence of known structure, the prediction for a specific sequence within the cluster was based on the most similar structural template (according to the p value). Generally, structures are adopted according to possible Watson–Crick base pairs (Watson and Crick 1953) and G–U pairings. More deviant noncanonical base pairs (Leontis et al. 2002) are not considered in the energy function or in the structure transfer. If the majority of base pairs—i.e., >75% per helix—could be adopted, structures are post-processed according to the algorithm of Nussinov and Jacobson (1980).

Iteration

Structures predicted in the first iteration (16,214) were used as templates for a second round of predictions (5533).

Visualization

Energies and structure images are derived by use of the Vienna Package (Hofacker et al. 1994).

ACKNOWLEDGMENTS

We thank Daniel Gerlach, Stefanie Maisel, and Stefan Pinkert (all of the University of Würzburg, Germany) for providing parts of the data and the DFG for support (Bo-1099/5-3).

Received June 21, 2005; accepted August 5, 2005.

REFERENCES

- Alvarez, I. and Wendel, J.F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.* **29**: 417–434.
- Coleman, A.W. 2003. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.* **19**: 370–375.
- Eddy, S.R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**: 2079–2088.
- Goertzen, L.R., Cannone, J.J., Gutell, R.R., and Jansen, R.K. 2003. ITS secondary structure derived from comparative analysis: Implications for sequence alignment and phylogeny of the Asteraceae. *Mol. Phylogenet. Evol.* **29**: 216–234.
- 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.
- Hofacker, I.L., Fekete, M., and Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**: 1059–1066.
- Lanave, C., Preparata, G., Saccone, C., and Serio, G. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**: 86–93.
- Leontis, N.B., Stombaugh, J., and Westhof, E. 2002. The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**: 3497–3531.
- Müller, T. and Vingron, M. 2000. Modeling amino acid replacement. *J. Comput. Biol.* **7**: 761–776.
- Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**: 443–453.

- Nussinov, R. and Jacobson, A.B. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.* **77**: 6309–6313.
- 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.
- Watson, J.D. and Crick, F.H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.
- Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**: 133–148.

RNA

A PUBLICATION OF THE RNA SOCIETY

Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures

MATTHIAS WOLF, MARCO ACHTZIGER, JÖRG SCHULTZ, et al.

RNA 2005 11: 1616-1623

References

This article cites 14 articles, 2 of which can be accessed free at:

<http://rnajournal.cshlp.org/content/11/11/1616.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>