Video Article

The ITS2 Database

Benjamin Merget[1,2], Christian Koetschan[1], Thomas Hackl[1], Frank Förster[1], Thomas Dandekar[1], Tobias Müller[1], Jörg Schultz[1], Matthias Wolf[1]

[1]Department of Bioinformatics, Biocenter, University of Würzburg

[2]Institute of Pharmacy and Food Chemistry, University of Würzburg

Correspondence to: Matthias Wolf at Matthias.Wolf@biozentrum.uni-wuerzburg.de

URL: http://www.jove.com/video/3806
DOI: doi:10.3791/3806

Keywords: Genetics, Issue 61, alignment, internal transcribed spacer 2, molecular systematics, secondary structure, ribosomal RNA, phylogenetic tree, homology modeling, phylogeny

Date Published: 3/12/2012

Citation: Merget, B., Koetschan, C., Hackl, T., Förster, F., Dandekar, T., Müller, T., Schultz, J., Wolf, M. The ITS2 Database. *J. Vis. Exp.* (61), e3806, doi:10.3791/3806 (2012).

Abstract

The internal transcribed spacer 2 (ITS2) has been used as a phylogenetic marker for more than two decades. As ITS2 research mainly focused on the very variable ITS2 sequence, it confined this marker to low-level phylogenetics only. However, the combination of the ITS2 sequence and its highly conserved secondary structure improves the phylogenetic resolution[1] and allows phylogenetic inference at multiple taxonomic ranks, including species delimitation[2-8].

The ITS2 Database[9] presents an exhaustive dataset of internal transcribed spacer 2 sequences from NCBI GenBank[11] accurately reannotated[10]. Following an annotation by profile Hidden Markov Models (HMMs), the secondary structure of each sequence is predicted. First, it is tested whether a minimum energy based fold[12] (direct fold) results in a correct, four helix conformation. If this is not the case, the structure is predicted by homology modeling[13]. In homology modeling, an already known secondary structure is transferred to another ITS2 sequence, whose secondary structure was not able to fold correctly in a direct fold.

The ITS2 Database is not only a database for storage and retrieval of ITS2 sequence-structures. It also provides several tools to process your own ITS2 sequences, including annotation, structural prediction, motif detection and BLAST[14] search on the combined sequence-structure information. Moreover, it integrates trimmed versions of 4SALE[15,16] and ProfDistS[17] for multiple sequence-structure alignment calculation and Neighbor Joining[18] tree reconstruction. Together they form a coherent analysis pipeline from an initial set of sequences to a phylogeny based on sequence and secondary structure.

In a nutshell, this workbench simplifies first phylogenetic analyses to only a few mouse-clicks, while additionally providing tools and data for comprehensive large-scale analyses.

Video Link

The video component of this article can be found at http://www.jove.com/video/3806/

Protocol

1. Correct Annotation of ITS2 Sequence

1. Access the ITS2 Database phylogeny workbench here: http://its2.bioapps.biozentrum.uni-wuerzburg.de
2. Begin your analysis by clicking the "Annotate" icon in the section "Tools." Then, type or paste your sequence into the sequence editor at the top of the website. The sequence editor automatically checks, whether your ITS2 sequences are valid.
3. Choose an HMM model suitable for your sequences (e.g. Viridiplantae for plants).
4. Start the process by clicking "Annotate."
5. By hovering over the "Hybridize" icon you can view an image of the 5.8S and 28S rRNA hybrid as a confirmation of the HMM annotation's accuracy.
6. Click on the green plus sign of the resulting ITS2 sequence to select your way of secondary structure prediction: To predict the structure without a known template, click on "Predict structure." If you want to use your own template for the Homology Modeling, click "Model structure."

2. Secondary Structure Prediction

1. Predict
    1. The annotated ITS2 sequence is automatically pasted into the sequence editor.
    2. To start the secondary structure prediction with default settings, click the "Predict structures" button.

3. Save the resulting ITS2 sequence including the modeled secondary structure into the data pool by clicking on the green plus sign and then "Add to pool." Alternatively, you can add it to your data pool via drag and drop (Figure 1).
4. If the sequence could not fold directly, the best results of the homology modeling are shown. Save the most suitable sequence-structure via drag and drop to the data pool. Alternatively, save the sequence-structure into the data pool with a right click and then a click on "Add to pool."

2. Custom Modeling
    1. Type or paste one or multiple templates (with known structure) into the upper sequence editor.
    2. Type or paste one or multiple target sequences (without structure) into the lower sequence editor.
    3. Click on "Predict best template(s)" to start the Homology Modeling with default settings.
    4. The best template-target combinations are shown in the resulting list.
    5. Save the modeled sequence-structure(s) of your choice either via drag and drop to the data pool or by a right click and a click on "Add to pool."

3. Motif Search

1. Type or paste your query sequence(s) into the sequence editor at the top of the website.
2. Choose the correct HMM model (e.g. Viridiplantae for plants). 3.3. Click on "Motif search" to start the process.
3. ITS2 sequences with highlighted motifs are illustrated at the bottom of the website.
4. Click on the icon beside the sequence header to display the motifs highlighted in the secondary structure.

4. Search and Browse

1. Search
    1. Type either a taxon name or a GenBank Identifier (GI) into the search field at the top of the website.
    2. A search by taxon name is supported by an appearing live-search box.
    3. You can perform a multiple search by comma-separating your queries.
    4. Click the "Search" button to execute the search.
    5. Your results appear listed in a new tab.
    6. Click on a column name to sort your results according to the particular column. You can also add or remove columns of your choice with the column menu. The column menu can be entered with a click on the appearing arrow icon within a column name.
    7. Click on "Show details" to view the details of a sequence-structure.
    8. Save the sequence-structure(s) of your choice either via drag and drop to the data pool or by a right click and a click on "Add to pool."
    9. To save your results to an external file, click on "Save selection" or "Save all."

2. Browse
    1. Browse the ITS2 Database by navigating through the tree-like structure at the left of the website.
    2. Click on a plus-sign to view the taxa one level lower.
    3. Click on a taxon name to open a new tab containing each sequence-structure of the taxon.
    4. Click on "Show details" to view the details of a sequence-structure pair.
    5. Save the sequence-structure(s) of your choice either via drag and drop to the data pool or by a right click and a click on "Add to pool."
    6. To save your results to an external file, click on "Save selection" or "Save all."

5. ITS2 Blast

1. Type or paste one or multiple query sequences into the sequence editor. Your sequences may either be plain nucleotide sequences or sequence-structure pairs. You can also type several secondary structures below one sequence. By checking the box "Serialize XXFASTA sequences" these structures are used subsequently as individual queries.
2. To start BLAST with default settings, click on "Blast." Depending on the nature of your query, either a common BLASTN or the ITS2 sequence-structure BLAST is performed.
3. A sub-tab is opened for each query sequence within the appearing tab "BLAST Results," as well as an overview of the executed searches.
4. Click on "Show Alignments" to view the calculated BLAST alignments.
5. Save the BLAST hits of your choice either via drag and drop to the data pool or by a right click and a click on "Add to pool."
6. To save your results to an external file, click on "Save selection" or "Save all."

6. Multiple Sequence-structure Alignment

1. Take a look at your data pool by clicking "Manage dataset" and then the magnifying glass symbol right next to the number of sequences in your pool. Alternatively, you can click on the data pool sign at the bottom left of the website.
2. Click on a sequence-structure pair in your data pool to view its details.
3. To create a multiple sequence-structure alignment of all sequence-structure pairs in your pool, click on "Analyze dataset" and then "Sequence & Structure."
4. Now you are asked to select the graphic mode of your alignment. If your alignment contains only a few sequences, decline the slim mode by clicking "No." Otherwise choose the slim graphic mode by clicking "Yes."
5. In a few moments, your alignment is shown in a new tab (Figure 2). Moreover, it is automatically saved to the data pool.

6. To save your alignment to an external file, click on "Save alignment."

7. Phylogenetic Tree

1. To calculate a sequence-structure based Neighbor Joining tree of your multiple alignment, click on "Analyze Dataset" and then "Neighbor Joining."
2. The resulting tree is illustrated in a new tab (Figure 3).
3. Scale your tree freely with the scroll bar "Zoom tree."
4. Reroot your tree by clicking on a node or leaf of the tree and then "Reroot at this node."
5. If you want to remove a taxon from your data pool, click on the leaf and choose "Remove this node from pool." Now you can recalculate your alignment and tree with the reduced taxon sampling.
6. Click on "Save tree" to save your phylogenetic tree as a final result of your analysis to an external NEWICK file.

8. Additional Software

1. Click on "About this website"-"Tools" to find additional information about the stand-alone tools 4SALE and ProfDistS.
2. Beside the alignment and Neighbor Joining function provided by the ITS2 Database web interface, you can now access several new functions, e.g. species delimitation based on compensatory base changes (CBCs).

9. Representative Results

The workflow as described above has successfully been applied in several open access surveys[3,4]. Examples can be viewed through the following links:

- http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0016931
- http://www.biomedcentral.com/1756-0500/3/320

In these large scale studies, we were able to resolve the phylogeny of Chlorophyta as well as Hypnales (Bryophyta) with high resolution. In both cases, an exhaustive taxon sampling was gathered from the ITS2 Database[9], automatically aligned with 4SALE[15,16] and lastly processed by ProfDistS[17] into a phylogenetic tree. In all these steps, sequence and structure information were used simultaneously. Bootstrap support for the phylogenetic backbone was achieved using Profile Neighbor Joining (PNJ)[19], which is available in the stand-alone version of ProfDistS.

For a smaller set of sequence-structure pairs, figures 1 to 3 describe the key steps of this automated workflow[5] directly on the new ITS2 Database workbench: taxon sampling, the multiple sequence-structure alignment and eventually the phylogenetic tree calculation.

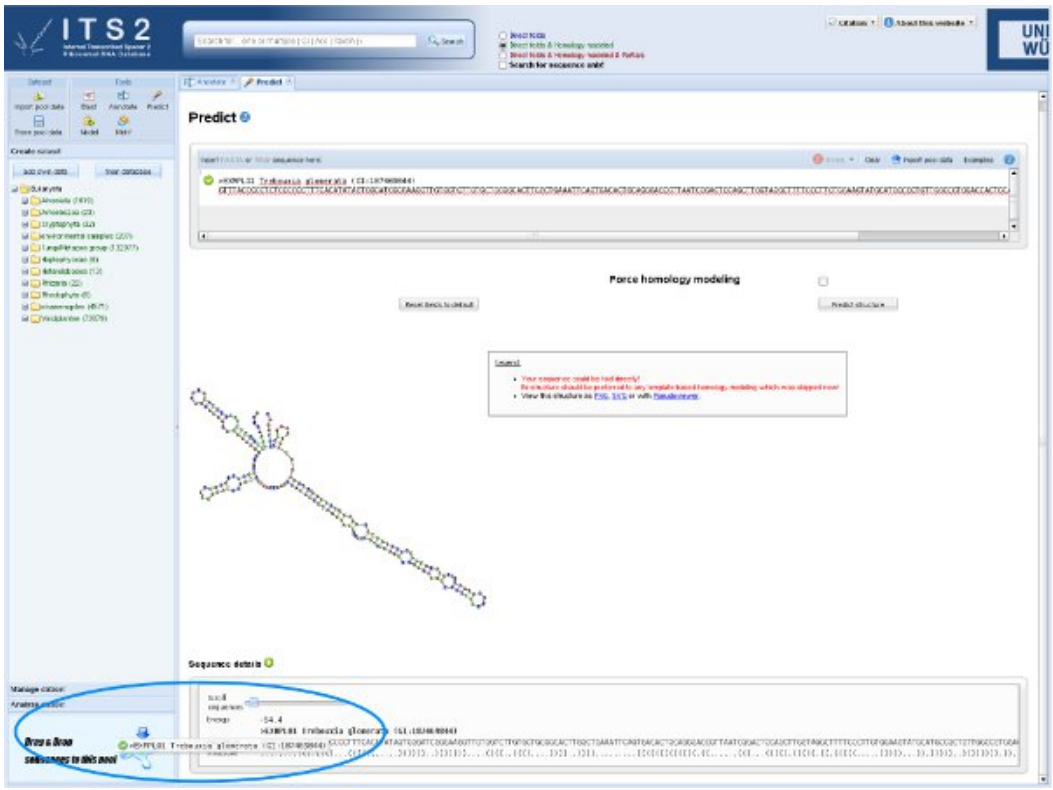

Figure 1. Taxon sampling per drag and drop. At any time sequences or sequence-structure pairs can be added to the data pool, for instance via drag and drop. Here a sequence-structure is added using drag and drop after secondary structure prediction. The blue ellipse marks the area where the sequence-structure is dropped into the data pool. Click here to view the full-sized version of this image.

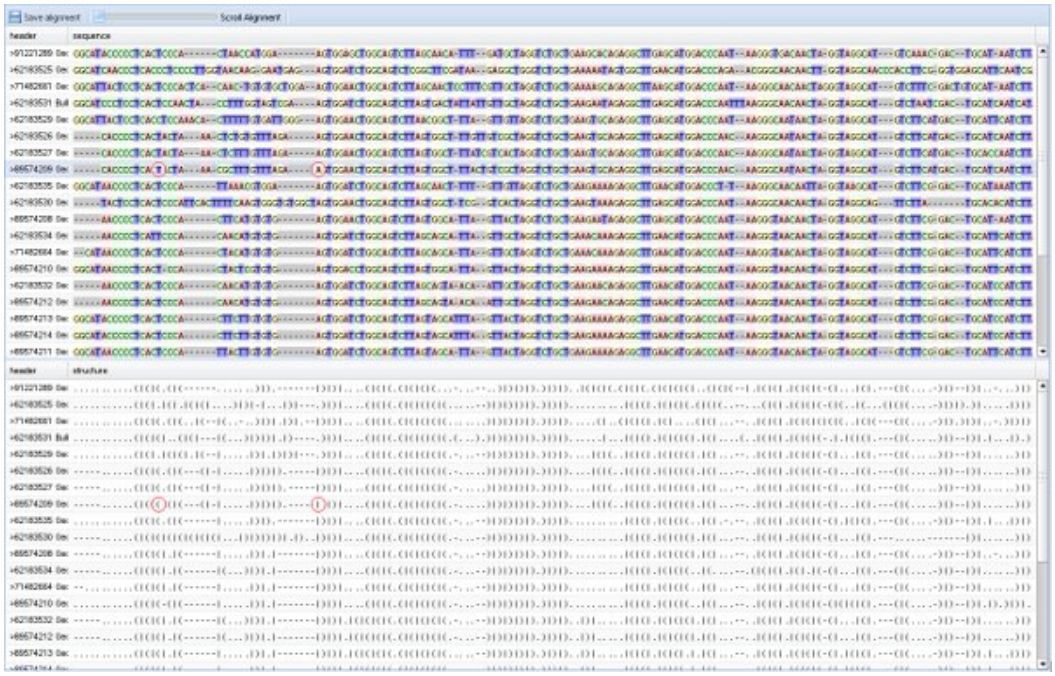

Figure 2. Multiple sequence-structure alignment in full graphic mode. For the few sequences in the data pool, the full graphic mode was chosen. Bases are colored; base pairs can be highlighted with red circles by clicking on one base or bracket of a base pair. Click here to view the full-sized version of this image.

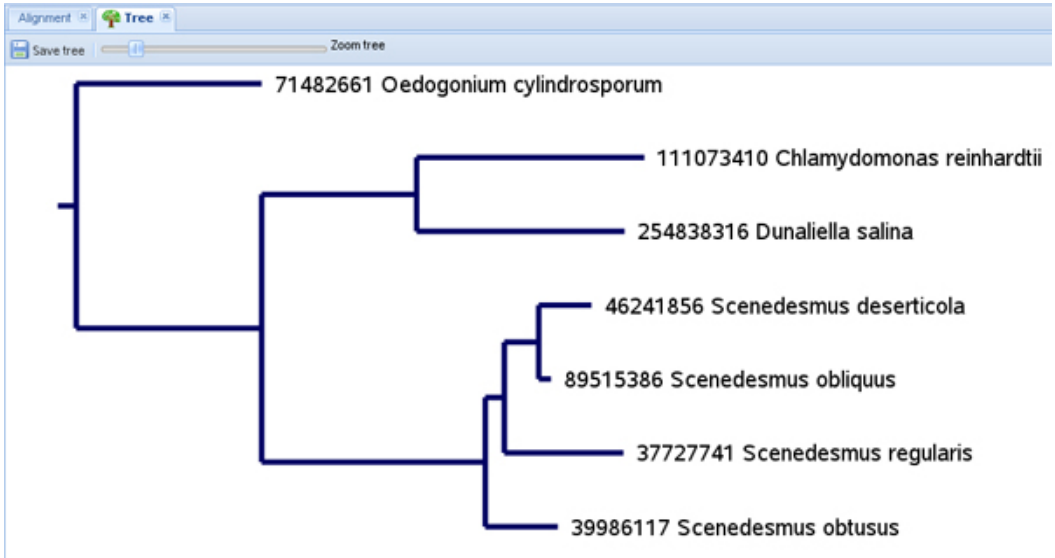

Figure 3. Sequence-structure Neighbor Joining tree. The freely scalable tree calculated of a seven taxa multiple sequence-structure alignment can be saved in the NEWICK format.

Discussion

The ITS2 Database is a complete and fully functional workbench for internal transcribed spacer 2 sequence-structure-based phylogenetics. The website can be operated very fast and intuitively. While other web-based phylogeny workbenches like ARB[20] or Mobyle[21] are only able to work on sequence and/or consensus structure information, the ITS2 Database[9] considers sequences and individual secondary structures for each taxon simultaneously. However, due to limitations in the computational capacity of the web server, it is highly recommended to use the stand-alone tools for multiple alignment and Neighbor Joining[18] calculation, 4SALE[15,16] and ProfDistS[17], respectively, for large datasets. Beside the basic ITS2 sequence-structure phylogeny workflow[5], these tools feature several additional functions, like calculating bootstrap replicates, Profile Neighbor Joining (PNJ)[19] or species delimitation based on compensatory base changes (CBCs)[8]. They can be accessed through the "About this website"-"Tools" section for download and detailed information. To use 4SALE and ProfDistS, it is necessary to always bring files into the correct

format. A taxon sampling to be processed by 4SALE must have the ending .fasta or .txt, whereas the sequence-structure alignment as an input for ProfDistS must end with .xfasta.

We are currently implementing alternative methods for phylogenetic tree reconstruction in the ITS2 database as well as in the related tools. Thus, methods like sequence-structure-based Maximum Parsimony[22] and/or Maximum Likelihood[23] will be accessible in the future.

Disclosures

No conflicts of interest declared.

Acknowledgements

We cordially thank the ITS2 group, Biocenter, University of Würzburg, for rich and valuable feedback. We also thank the Deutsche Forschungsgemeinschaft (DFG; grant Mu-2831/1-1) for funding.

References

1. Keller, A., *et al*. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct.* **5**, 4 (2010).
2. Schultz, J., Maisel, S., Gerlach, D., Müller, T., & Wolf, M. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA.* **11**, 361-364 (2005).
3. Buchheim, M., *et al*. Internal Transcribed Spacer 2 (nu ITS2 rRNA) Sequence-Structure Phylogenetics: Towards an Automated Reconstruction of the Green Algal Tree of Life. *PLoS ONE.* **6**, e16931 (2011).
4. Merget, B. & Wolf, M. A molecular phylogeny of Hypnales (Bryophyta) inferred from ITS2 sequence-structure data. *BMC Research Notes.* **3**, 320 (2010).
5. Schultz, J. & Wolf, M. ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics. *Molecular Phylogenetics and Evolution.* **52**, 520-523 (2009).
6. Coleman, A. ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics.* **19**, 370-375 (2003).
7. Coleman, A. The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. *Protist.* **151**, 1-9 (2000).
8. Müller, T., Philippi, N., Dandekar, T., Schultz, J., & Wolf, M. Distinguishing species. *RNA.* **13**, 1469-1472 (2007).
9. Koetschan, C., *et al*. The ITS2 Database III-sequences and structures for phylogeny. *Nucleic Acids Research.* **38**, D275-279 (2010).
10. Keller, A., *et al*. 5.8 S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene.* **430**, 50-57 (2009).
11. Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., & Sayers, E. GenBank. *Nucleic Acids Research.* **39**, D32-37 (2011).
12. Markham, N. & Zuker, M. Software for nucleic acid folding and hybridization. *Methods in Molecular Biology.* **453**, 3-31 (2008).
13. Wolf, M., Achtziger, M., Schultz, J., Dandekar, T., & Müller, T. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA.* **11**, 1616-1623 (2005).
14. Altschul, S., *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* **25**, 3389-3402 (1997).
15. Seibel, P., Müller, T., Dandekar, T., & Wolf, M. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4 SALE. *BMC Research Notes.* **1**, 91 (2008).
16. Seibel, P., Müller, T., Dandekar, T., Schultz, J., & Wolf, M. 4 SALE - A tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics.* **7**, 498 (2006).
17. Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J., & Müller, T. ProfDistS:(profile-) distance based phylogeny on sequence-structure alignments. *Bioinformatics.* **24**, 2401-2402 (2008).
18. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution.* **4**, 406-425 (1987).
19. Müller, T., Rahmann, S., Dandekar, T., & Wolf, M. Accurate and robust phylogeny estimation based on profile distances: a study of the Chlorophyceae (Chlorophyta). *BMC Evolutionary Biology.* **4**, 20 (2004).
20. Wolfgang Ludwig, *et al*. ARB: a software environment for sequence data. *Nucleic Acids Research.* **32**, 1363-1371 (2004).
21. Néron, B., *et al*. Mobyle: a new full web bioinformatics framework. *Bioinformatics.* **25**, 3005-3011 (2009).
22. Camin, J.H. & Sokal, R.R. A method for deducing branching sequences in phylogeny. *Evolution.* **19**, 311-326 (1965).
23. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution.* **17**, 368-376 (1981).