Using RNA Secondary Structure in Phylogeny

Alexander Donath[1], Steve Hoffmann[1], and Peter F. Stadler[1,2,3,4]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
[2]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI), Leipzig, Germany
[3]Department of Theoretical Chemistry, University of Vienna, Austria
[4]Santa Fe Institute, Santa Fe, New Mexico, USA

Email: {alex,steve,studla}@bioinf.uni-leipzig.de WorldWideWeb: http://www.bioinf.uni-leipzig.de/

Application

Overall Goal:
Utilization of RNA secondary structure to infer phylogeny.

Approach:

- Usage of clade-specific features (e.g. large insertions) as molecular morphological characters.
- Given a sequence/structure alignment and a proposed phylogeny:
- Is it possible to obtain motifs in primary and/or secondary structure that either support or reject the given phylogeny?

Workflow:

1. Motif discovery and evaluation in primary and secondary structure (e.g. RNAsalsa, noisy, Vienna RNA package, MEME).
2. Evaluation of the tree given these patterns.
3. If phylogeny is rejected - find the best tree that supports the pattern distribution.

Problems:

- Reliable structure prediction of large RNA molecules (RNAsalsa).
- Definition and weighting of patterns.
- Pattern evolution (e.g. 7SK 5' stems).

noisy

(Dress et al., 2008)

Overall Goal:
Identification of phylogenetically uninformative sites in a multiple alignment.

Approach:

- Detection of homoplastic characters using circular orderings.

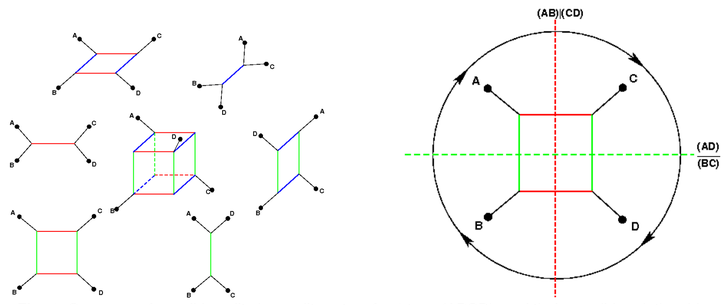


Fig. 1. Summary of possible splits in a split system for 4 taxa (ABCD), and their possible relationship.

Workflow:

1. Construction of the circular ordering.
2. Calculation of a reliability score q for each character.
3. Removal of characters smaller than a given threshold.

Results:

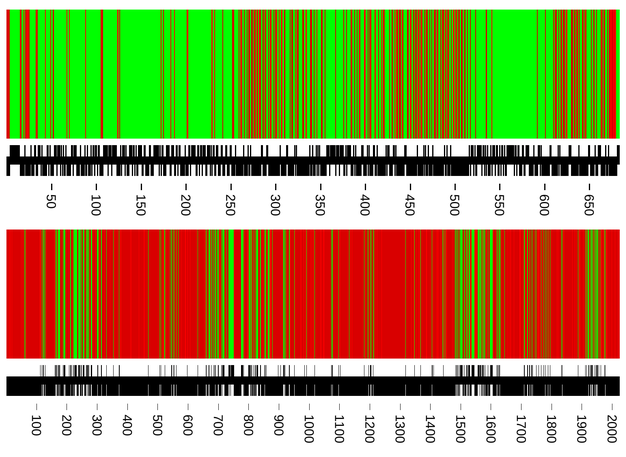
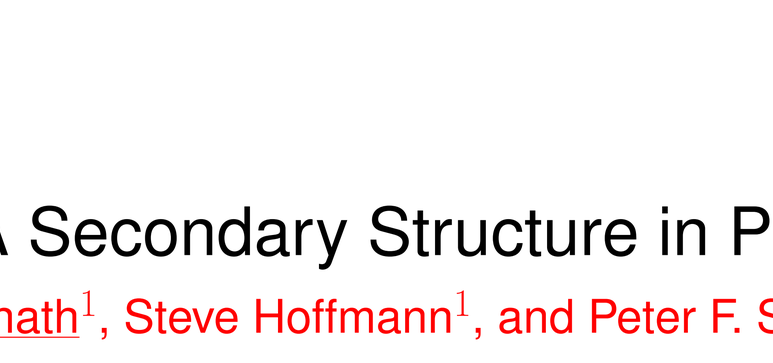


Fig. 2. Distribution of homoplastic sites for the mitochondrial atp6 gene of squamata (top) and for 18S RNA of Coleoptera (Korte et al., 2004) (bottom). The quality of the two data sets is very different. While the majority of sites in atp6 are parsimony informative (green; at least two different character states occur in at least two taxa) and approximately one third of the sites have a reliability score above q_{cutoff} = 0.8, this is clearly not the case for the data set by Korte et al. where most of the sites are constant (dark red) or unreliable (yellow (missing data), red (singleton site)). The black bar below the alignment displays the sites with $q \geq q_{cutoff}$ (upper half) and $q < q_{cutoff}$ (lower half).



	raw	Stockley	noisy
length	2528	2227	2076
PI−sites	543	465	260
HI	0.54	0.59	0.50
RI	0.41	0.44	0.56
RC	0.19	0.20	0.28

Fig. 3. MP trees of spatangoid sea urchins from combined 28S rRNA, 16S rRNA, and mitochondrial COI sequences (Stockley et al., 2005). L.h.s. from original data, r.h.s. from a reduced alignment with cutoff q = 0.8. The latter tree matches the biological expectation and fits very well with those reported by Stockley et al. that were obtained from a manually reduced alignment. (HI = homoplasy index, RC = rescaled consistency index, RI = retention index.)

noisy is available under the GNU Public Licence from:
http://www.bioinf.uni-leipzig.de/Software/noisy/

RNAsalsa

(Stocsits et al., 2008 (in press))

Overall Goal:
Performing improved RNA structure predictions as well as alignments of structural RNA sequences. Utilization of the slower evolution of structural features as a reproducible source of information.

Approach:

- Utilization of prior knowledge about structural patterns.
- Direct thermodynamic folding via adapted constraints.
- Using of structure information for adjusting and refining of the sequence alignment and vice versa.

Workflow:

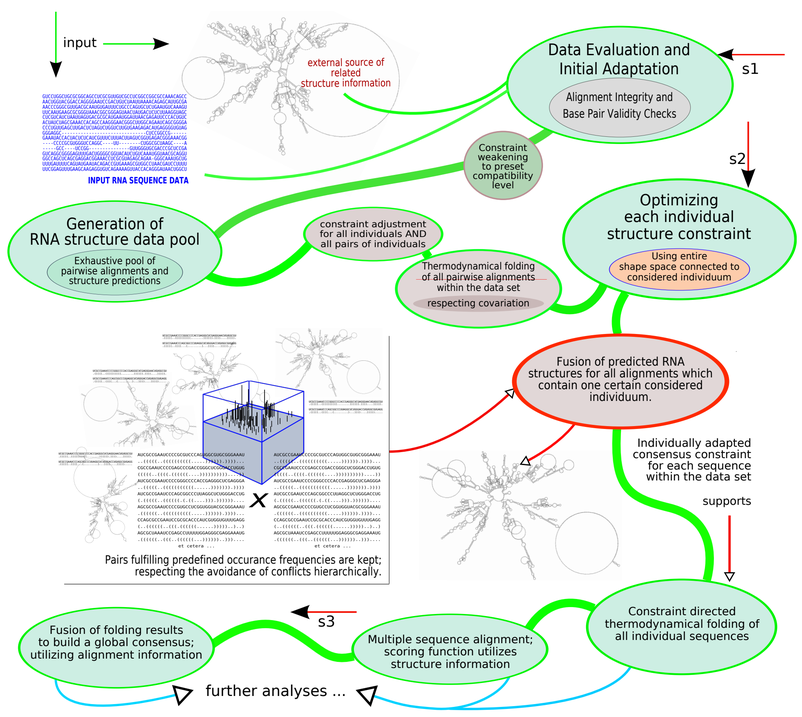


Fig. 4. Algorithmic concepts throughout the workflow of RNAsalsa.

Results:

- Complete individual secondary structures for each sequence.
- A final alignment by taking both structure and sequence information of each position into account.
- A sequence alignment with a consensus structure that allows the application of RNA substitution models in the conserved paired regions of the considered RNA genes.

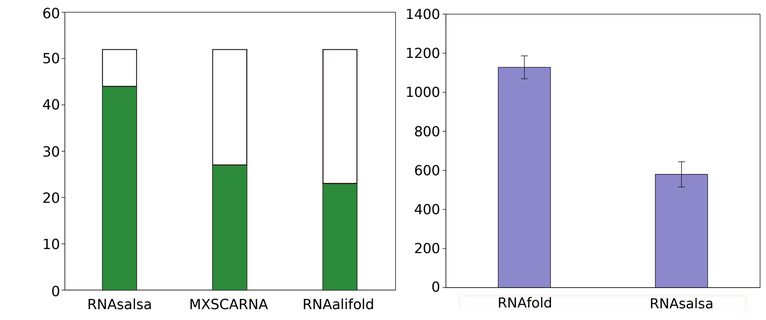


Fig. 5. Accuracy of structure prediction. L.h.s. fraction of correctly predicted helices (green bars) compared to the mammalian 16S rRNA consensus models (3-sample test for equality of proportions without continuity correction; χ^2 = 19.96, df = 2, $p < 0.0001$). R.h.s. average tree-edit distance between predicted individual structures and the mammalian 16S rRNA reference model (paired sampled t-Test; t = 33.46, df = 1, $p < 0.0001$; N = 26.)

version 0.7.4 beta:
http://www.rnasalsa.zfmk.de/

Biological Examples

7SK (Gruber et al., 2008a, 2008b)
The 7SK snRNA is a key player in the regulation of polymerase II transcription and highly conserved across vertebrates. Homologs in basal deuterostomes, a few lophotrochozoans, and arthropods were only recently reported.

Patterns:

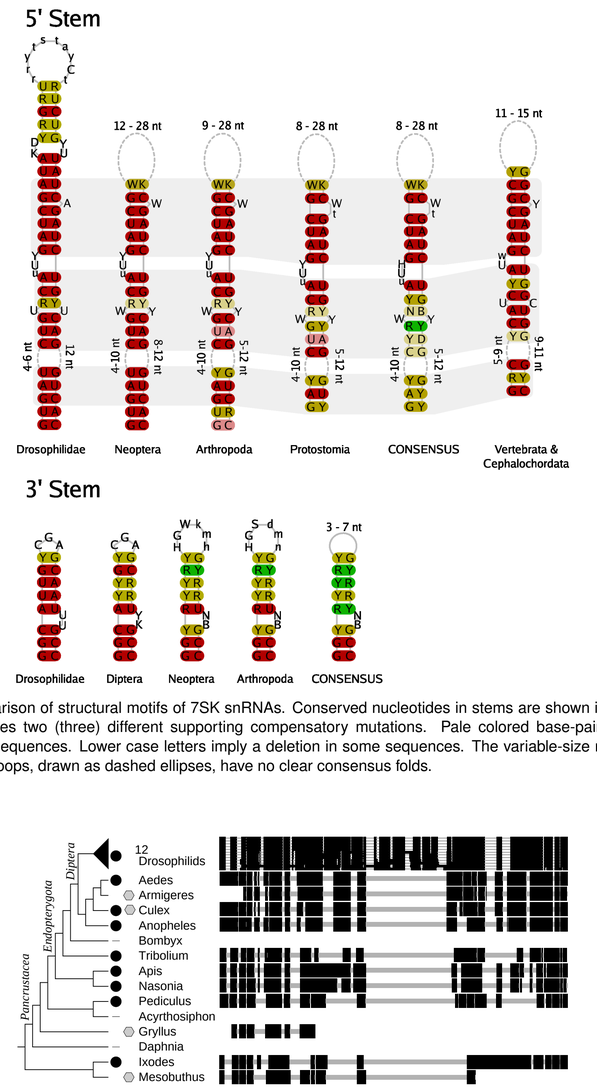


Fig. 6. Comparison of structural motifs of 7SK snRNAs. Conserved nucleotides in stems are shown in red. Ocher (green) indicates two (three) different supporting compensatory mutations. Pale colored base-pairs cannot be formed by all sequences. Lower case letters imply a deletion in some sequences. The variable-size regions close to the hairpin loops, drawn as dashed ellipses, have no clear consensus folds.

Fig. 7. Phylogenetic distribution of 7SK candidate sequences. Bullets indicate a match in the genomic sequence, hexagons refer to partial ESTs. Aligned blocks are shown in black, gray bars indicate alignment gaps, missing sequence data appears white. The underlying tree is composed from the genome-wide near intron positions (Krauss et al., 2008), a mosquitoe phylogeny (Harbach and Kitching, 1998) and two recent studies of arthropod phylogeny (Cameron et al., 2004; Kjer, 2004).

Other Examples:

- 16S rRNA
- 5S, 18S RNA
- RNAse P RNA
- other SSU and LSU rRNAs

Acknowledgements:
Financial support by the DFG SPP 1174 "Deep Metazoan Phylogeny".

UNIVERSITÄT LEIPZIG

Printed at the Universitätsrechenzentrum Leipzig.