# BioInformatic approaches
# to detect and verify ncRNAs

Sven Findeiß[1], Maribel Hernandez Rosales[1], Peter F. Stadler[1]

[1]Bioinformatics, Inst. f. Informatik, University Leipzig, Germany

Tel: +49 341 97 16 705  Fax: ++49 341 97 16 679  Email: {sven, maribel, studla}@bioinf.uni-leipzig.de

WorldWideWeb: http://www.bioinf.uni-leipzig.de/

## NcDNAlign

Rose et al. 2007

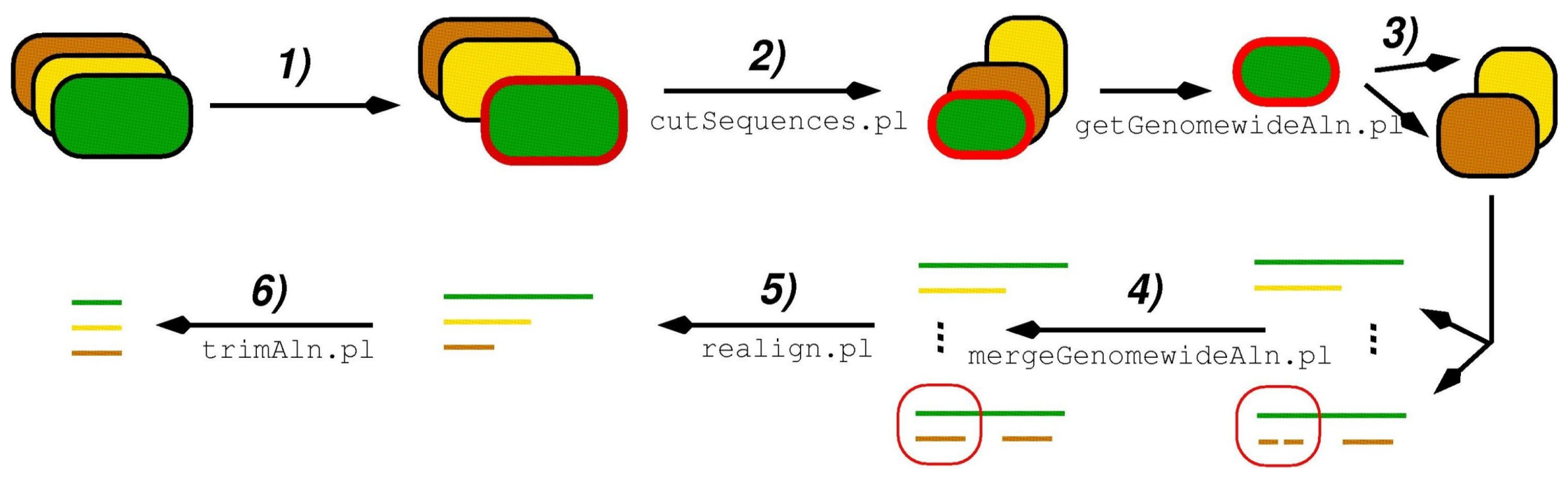
### Overall Goal:

Generate multiple sequence alignments (MSA) of non-protein-coding genomic sequences efficiently and customizably.

### Approach:

- Based on a pipeline concept, implemented in `Perl`, combining custom written and standard bioinformatic software tools.
- Only sequence data in common `FASTA` or `GeneBank` format required as input.
- Sequence regions not of interest can be discarded beforehand.
- Follows a progressive paradigm starting from pairwise `BLAST` alignments.
- Uses a configurable heuristics for alignment beautification.

### Workflow:



1. One species out of all given genomic sequences has to be selected as reference.
2. If selected, sequence data are ridded of potentially interfering or uninteresting sequence stretches, reducing the data set to genomic subsequences.
3. All sub-sequences of the reference are compared to all sub-sequences of all other species and local alignments are calculated heuristically (`BLAST`).
4. Adjacent compatible hits are combined.
5. The best hits (E-value) in each organism for each sub-sequence of the reference are aligned (`DIALIGN`).
6. Finally, the alignments are pruned, poorly aligned sequences are removed and the remaining sequences are optionally realigned to obtain an optimal alignment.

### Results:

- Displays comparable sensitivity and specificity as `TBA` (Blanchette et al., Genome Research, 2004) in RNA gene finding using `RNAz`.
- Aligns less DNA than `TBA`, albeit at higher alignment quality.
- Significantly outperforms `TBA` in computational effort.
- Appears to be a reasonable alternative to TBA for pilot studies, when infrastructure is limited, or when analysis pipelines should be rerun frequently to incorporate newly available data.

www.bioinf.uni-leipzig.de/Software/NcDNAlign/

UNIVERSITÄT LEIPZIG

Printed by the Universitätsrechenzentrum Leipzig.

Many thanks to Kristin, Jörg and Aileen for help.

## RNAz

Washietl et al. 2005

### Overall Goal:

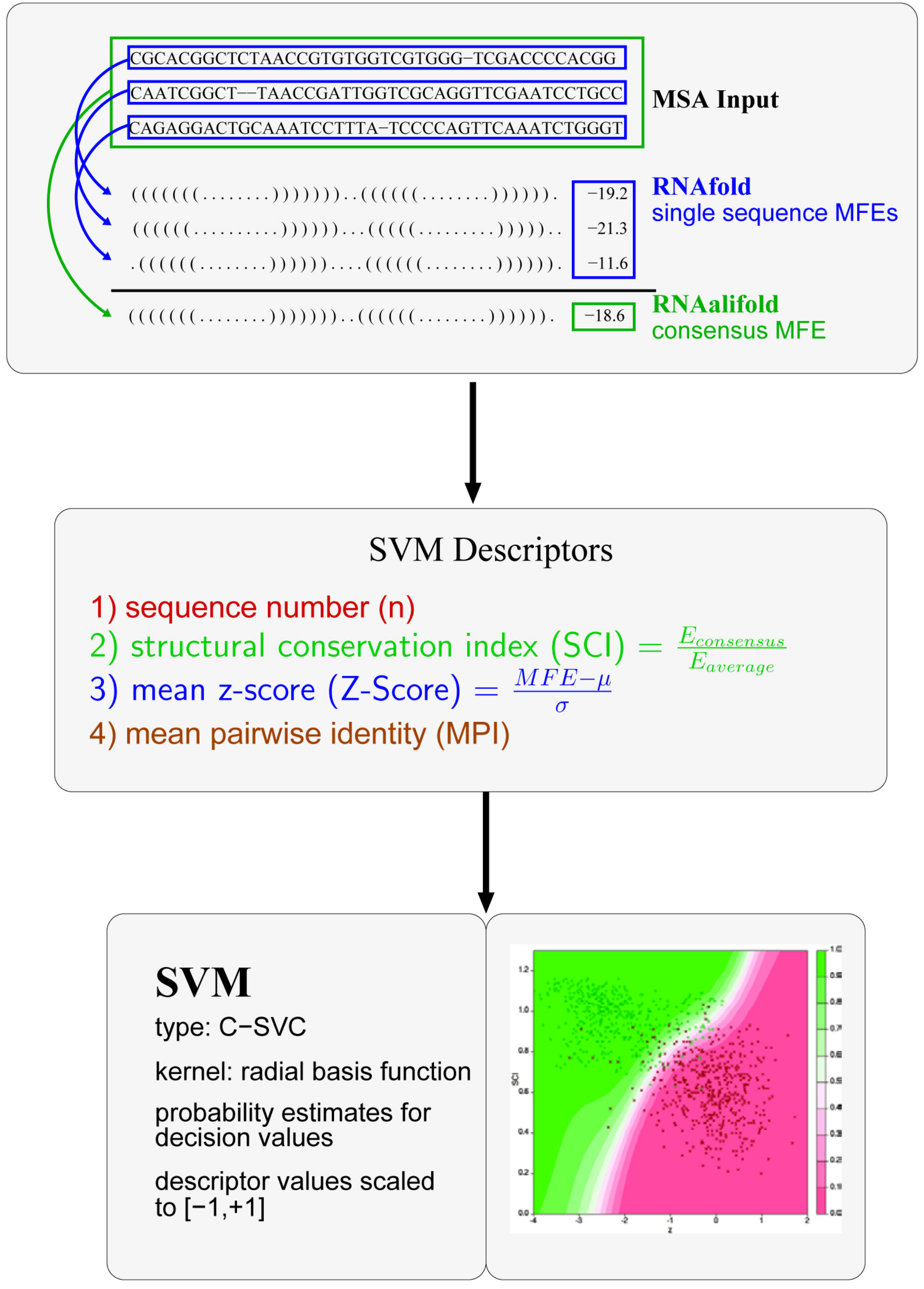
Using MSA for fast and reliable prediction of non coding RNAs.

### Approach:

- Structured ncRNAs (tRNA, rRNA, . . . ) have a defined and evolutionary conserved secondary structure which is of functional importance.
- Prediction of structurally conserved and thermodynamically stable RNA secondary structures.
- Based on a minimum free energy (MFE) structure prediction algorithm

### Workflow:

RNAz performs the classification by means of a support vector machine (SVM) that takes four descriptors into account.



SVM Descriptors

1) sequence number (n)

2) structural conservation index (SCI) $= \frac{E_{consensus}}{E_{average}}$

3) mean z-score (Z-Score) $= \frac{MFE-\mu}{\sigma}$

4) mean pairwise identity (MPI)

**SVM**

type: C−SVC

kernel: radial basis function

probability estimates for decision values

descriptor values scaled to [−1,+1]

### Results:

- Output interpretable by RNA-class probability (p-value) which indicates the SVM decision. A value close to one indicates a high significance for a structured RNA within the alignment.
- The current version of `RNAz` is limited to a maximum of six sequences and it is not able to score alignments longer than 400 columns.
- Within this SPP project we purpose to address the limitations of `RNAz` by a sensitive version for short RNA motifs and the ability to analyze MSA with more than six sequences.

user friendly web server
http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi

## LocARNA based Clustering

Will et al. 2007

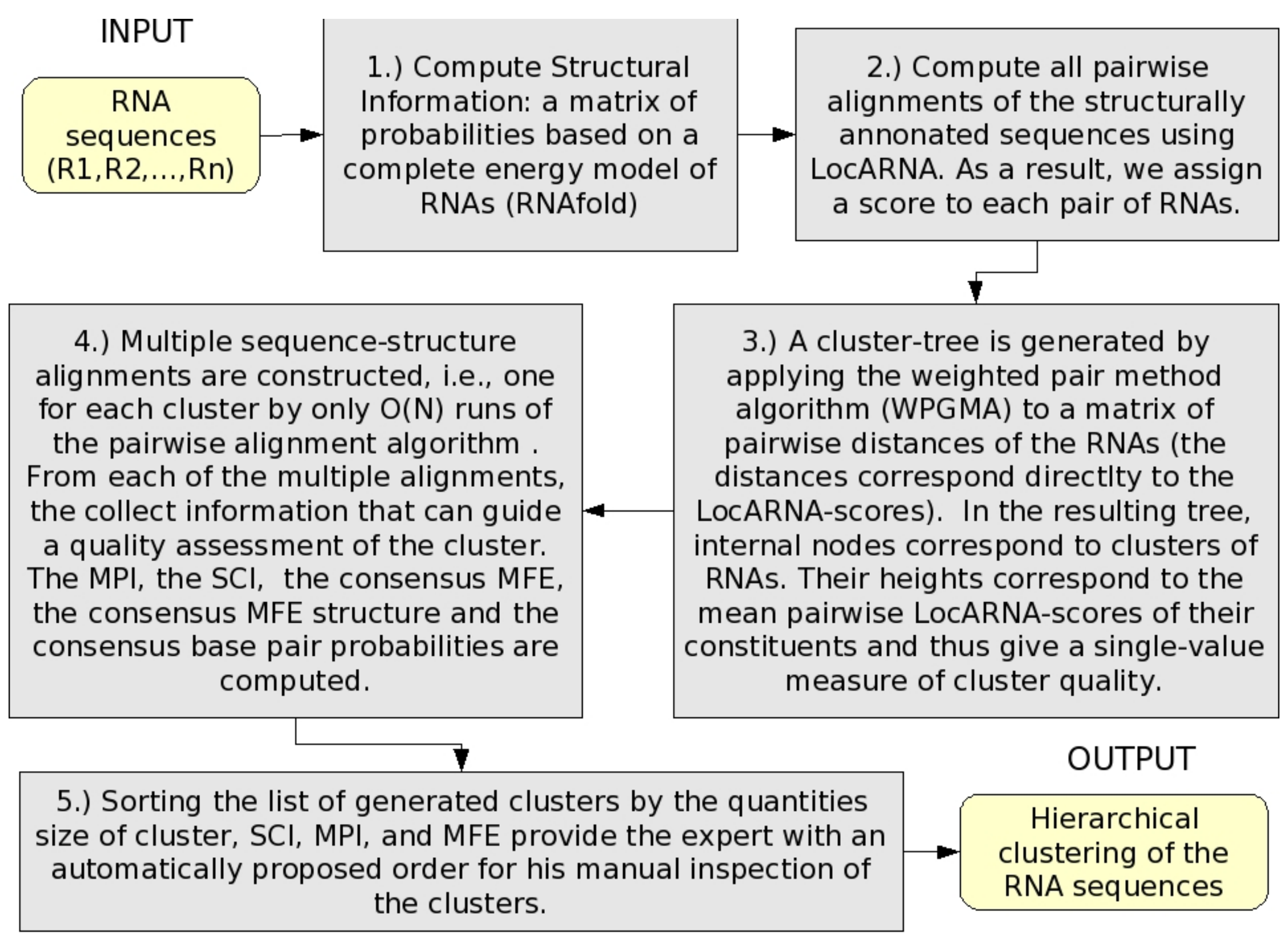
### Overall Goal:

Establish a structure-based clustering approach that is capable of extracting putative RNA classes from genome-wide surveys for structured RNAs.

### Approach:

- Computationally (`RNAz`) predicted RNA families can be grouped together, forming a ncRNA class whose members have no discernible homology at sequence level, but still share common structural and functional properties.
- Check evidence for novel families or classes for which we have not yet seen experimentally verified representatives.
- Utilizes `LocARNA` (local alignment of RNA), a local pairwise structural alignment algorithm for pseudoknot-free RNA secondary structures and its multiple version `mLocARNA`.

### Workflow:



INPUT

RNA sequences (R1,R2,...,Rn)

1.) Compute Structural Information: a matrix of probabilities based on a complete energy model of RNAs (RNAfold)

2.) Compute all pairwise alignments of the structurally annonated sequences using LocARNA. As a result, we assign a score to each pair of RNAs.

3.) A cluster-tree is generated by applying the weighted pair method algorithm (WPGMA) to a matrix of pairwise distances of the RNAs (the distances correspond directlty to the LocARNA-scores). In the resulting tree, internal nodes correspond to clusters of RNAs. Their heights correspond to the mean pairwise LocARNA-scores of their constituents and thus give a single-value measure of cluster quality.

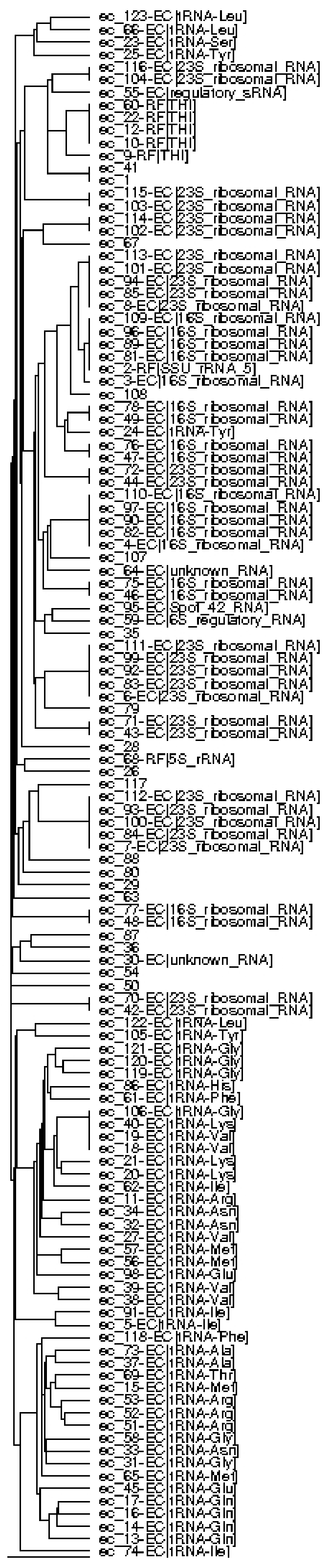
4.) Multiple sequence-structure alignments are constructed, i.e., one for each cluster by only O(N) runs of the pairwise alignment algorithm . From each of the multiple alignments, the collect information that can guide a quality assessment of the cluster. The MPI, the SCI, the consensus MFE, the consensus MFE structure and the consensus base pair probabilities are computed.

5.) Sorting the list of generated clusters by the quantities size of cluster, SCI, MPI, and MFE provide the expert with an automatically proposed order for his manual inspection of the clusters.

OUTPUT

Hierarchical clustering of the RNA sequences

### Results:



Six related gammaproteobacteria were screened with `RNAz`. The result is a ncRNA candidate set of 123 unique loci of the reference organism *Escherichia coli.* The cluster-tree on the right hand side depicts that the majority of all ncRNAs could be annotated with known *E. coli* ncRNAs (labeled with EC[. . . ]). If such an annotation was not available, a blast search against the `RFAM` database identified further homologue ncRNAs (labeled with RF[. . . ]). It can be seen that the tRNAs cluster well in subgroups. Even tRNAs which code for the same amino acid are clustered together. The fact that several 16S RNAs are divided in different sub-clusters are due to the limitations of `RNAz` (see Will et al, 2007).