De novo prediction of structural noncoding RNAs

Stefan Washietl

18.417 - Fall 2011

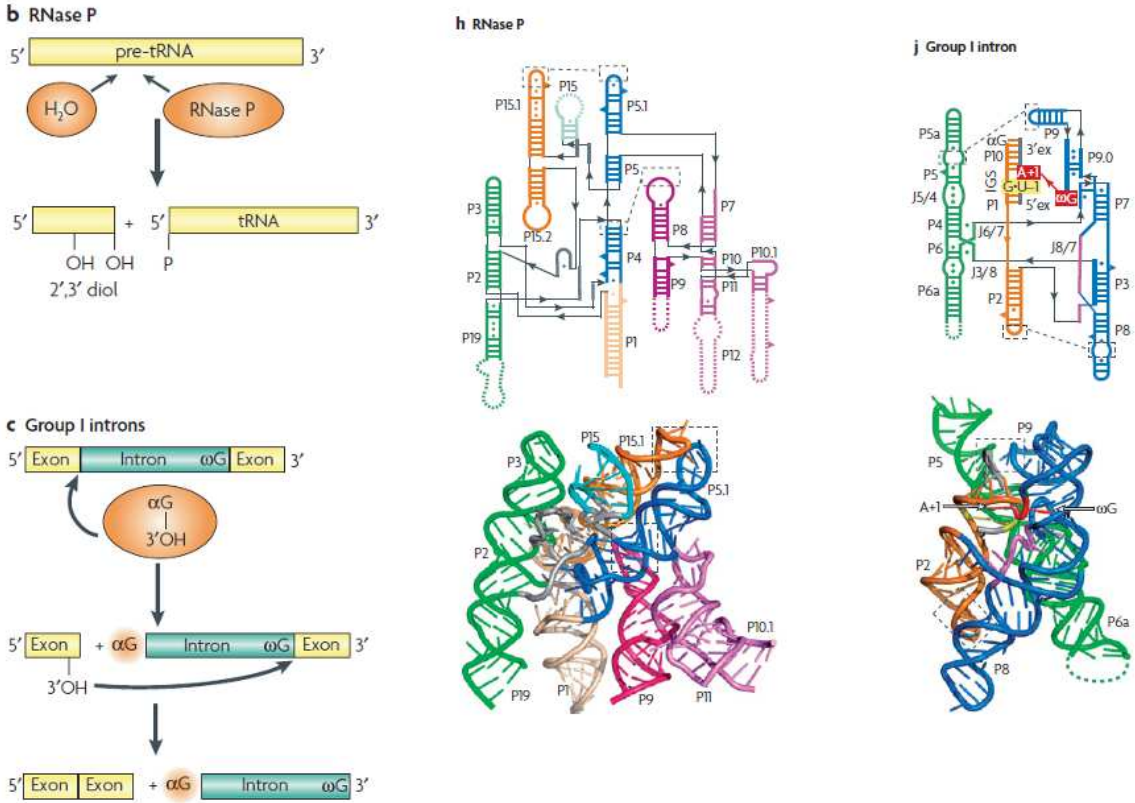
Outline

- Motivation: Biological importance of (noncoding) RNAs
- Algorithms to predict structural noncoding RNAs
  - RNAz: thermodynamical folding + phylogenetic information
  - EvoFold: phylogenetic stochastic context-free grammars
- A few applications of RNAz and Evofold

Essential biochemical functions of life

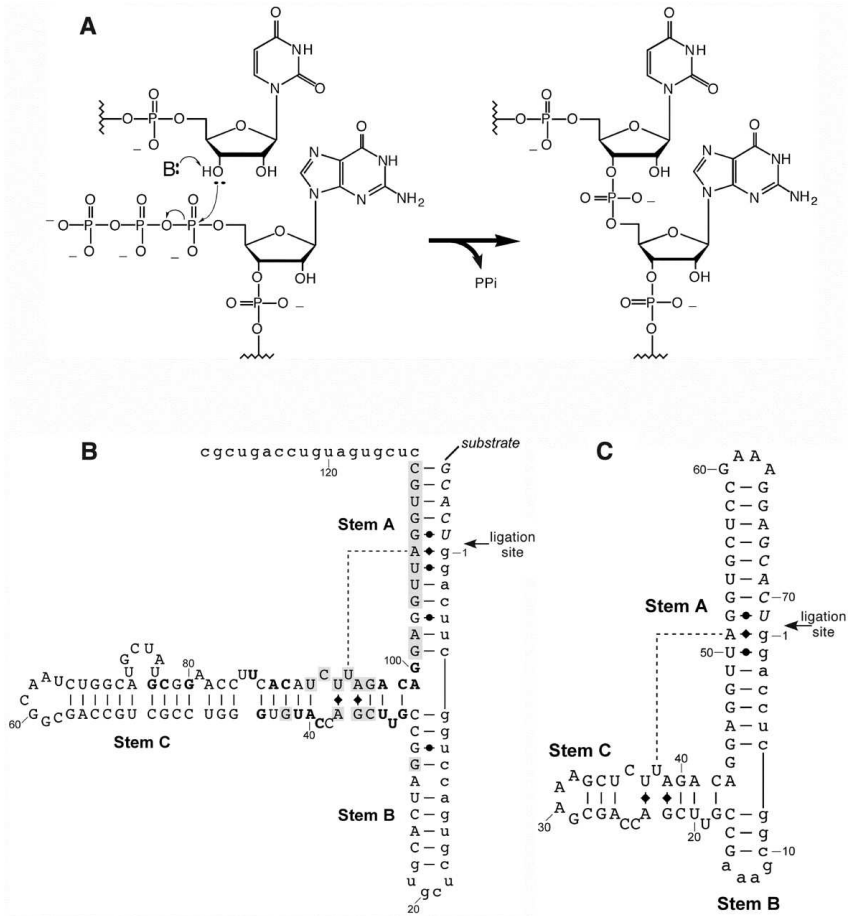
- Information storage and replication
- Enzymatic activity: catalyze biochemical reactions
- Regulator: sense and react to environment

Enzymatic activity: Ribozymes



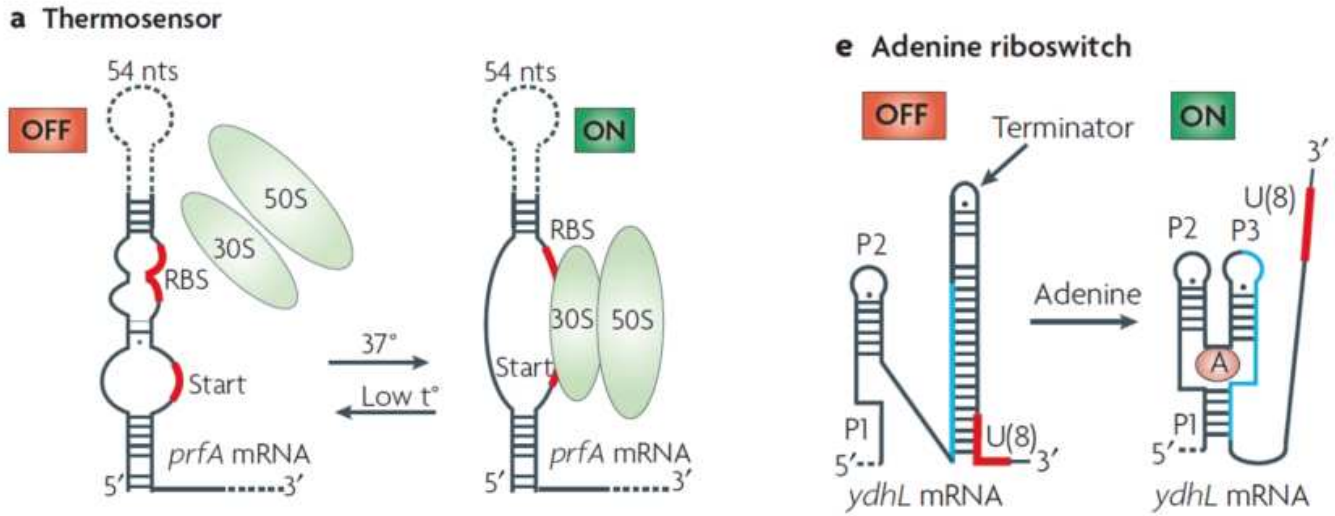
- Self splicing introns and RNAseP were the first examples of RNAs with catalytic activity. First discoverd by Sidney Altman and Thomas Cech.

Self duplication



- Ribozyme acting as RNA dependent RNA polymerase
- A chimeric construct of a natural ligase ribozyme with an *in vitro* selected template binding domain can replicate at least one turn of an RNA helix.

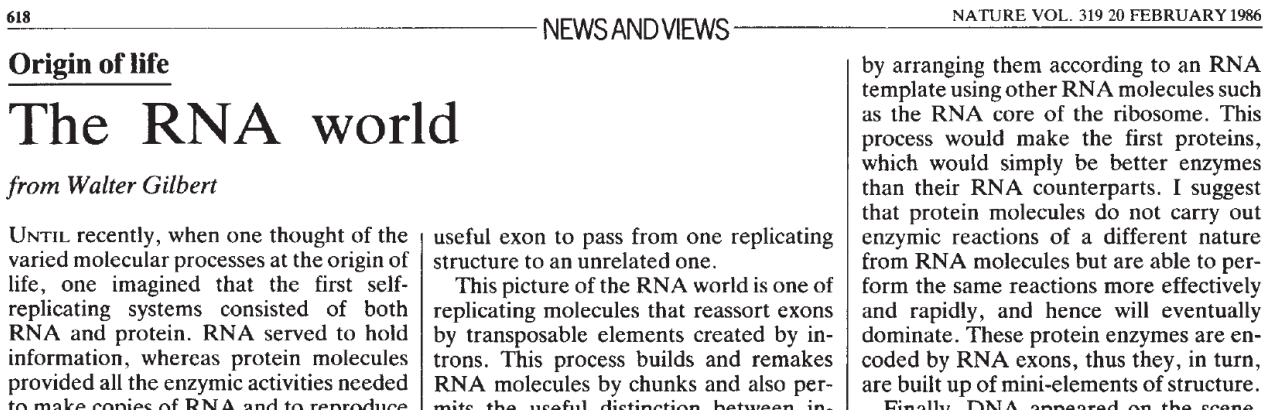
Regulation: Riboswitches



- Environmental stimuli change directly (without protein) the conformation of an RNA which affects gene activity.

Serganov A, Patel DJ, *Nat Rev Genet.* 2007 8:(10)776-90

Putting things together: RNA world hypothesis



618 NEWS AND VIEWS NATURE VOL. 319 20 FEBRUARY 1986

Origin of life

The RNA world

from Walter Gilbert

UNTIL recently, when one thought of the varied molecular processes at the origin of life, one imagined that the first self-replicating systems consisted of both RNA and protein. RNA served to hold information, whereas protein molecules provided all the enzymic activities needed to make copies of RNA and to reproduce

useful exon to pass from one replicating structure to an unrelated one.

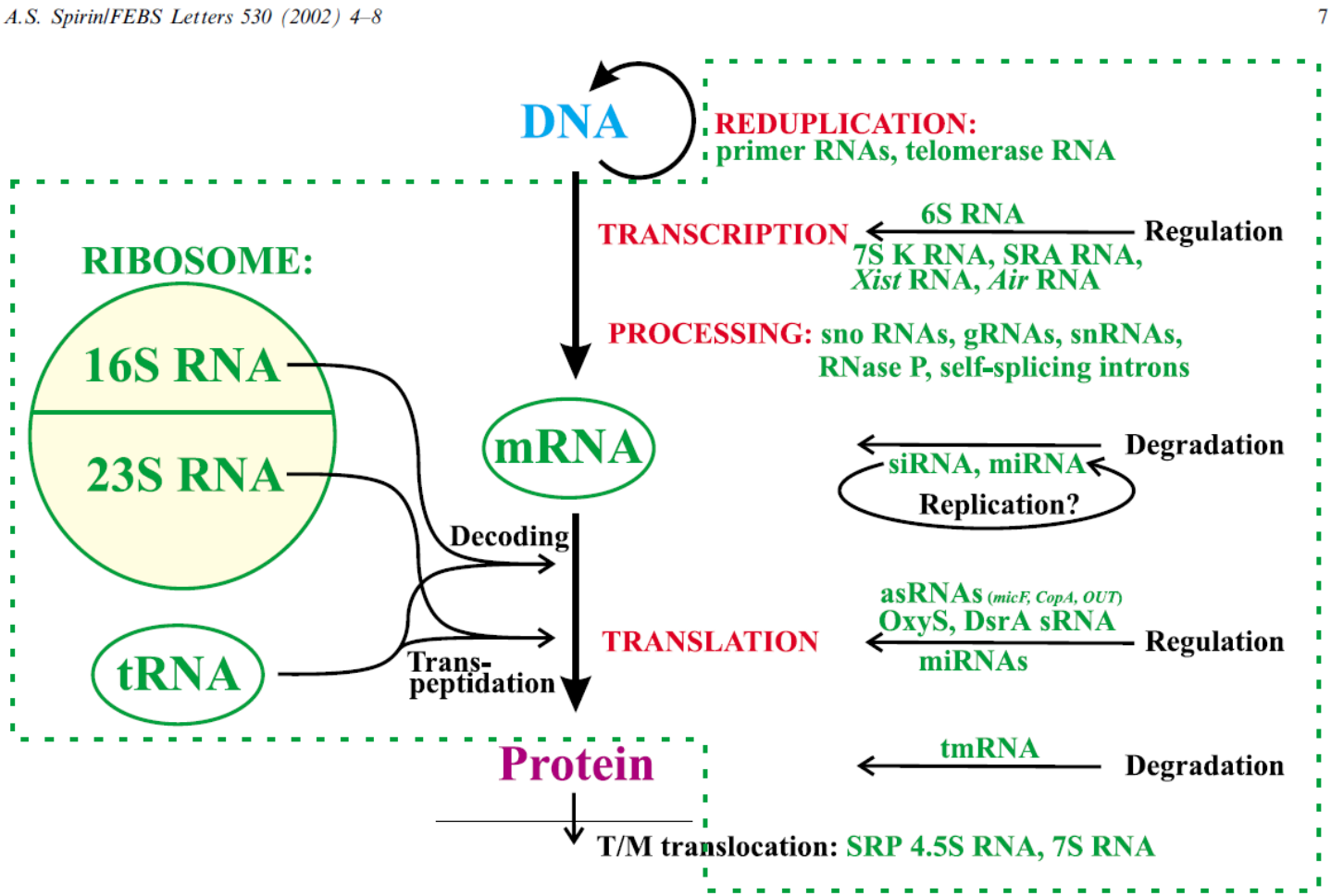
This picture of the RNA world is one of replicating molecules that reassort exons by transposable elements created by introns. This process builds and remakes RNA molecules by chunks and also permits the useful distinction between in-

by arranging them according to an RNA template using other RNA molecules such as the RNA core of the ribosome. This process would make the first proteins, which would simply be better enzymes than their RNA counterparts. I suggest that protein molecules do not carry out enzymic reactions of a different nature from RNA molecules but are able to perform the same reactions more effectively and rapidly, and hence will eventually dominate. These protein enzymes are encoded by RNA exons, thus they, in turn, are built up of mini-elements of structure.

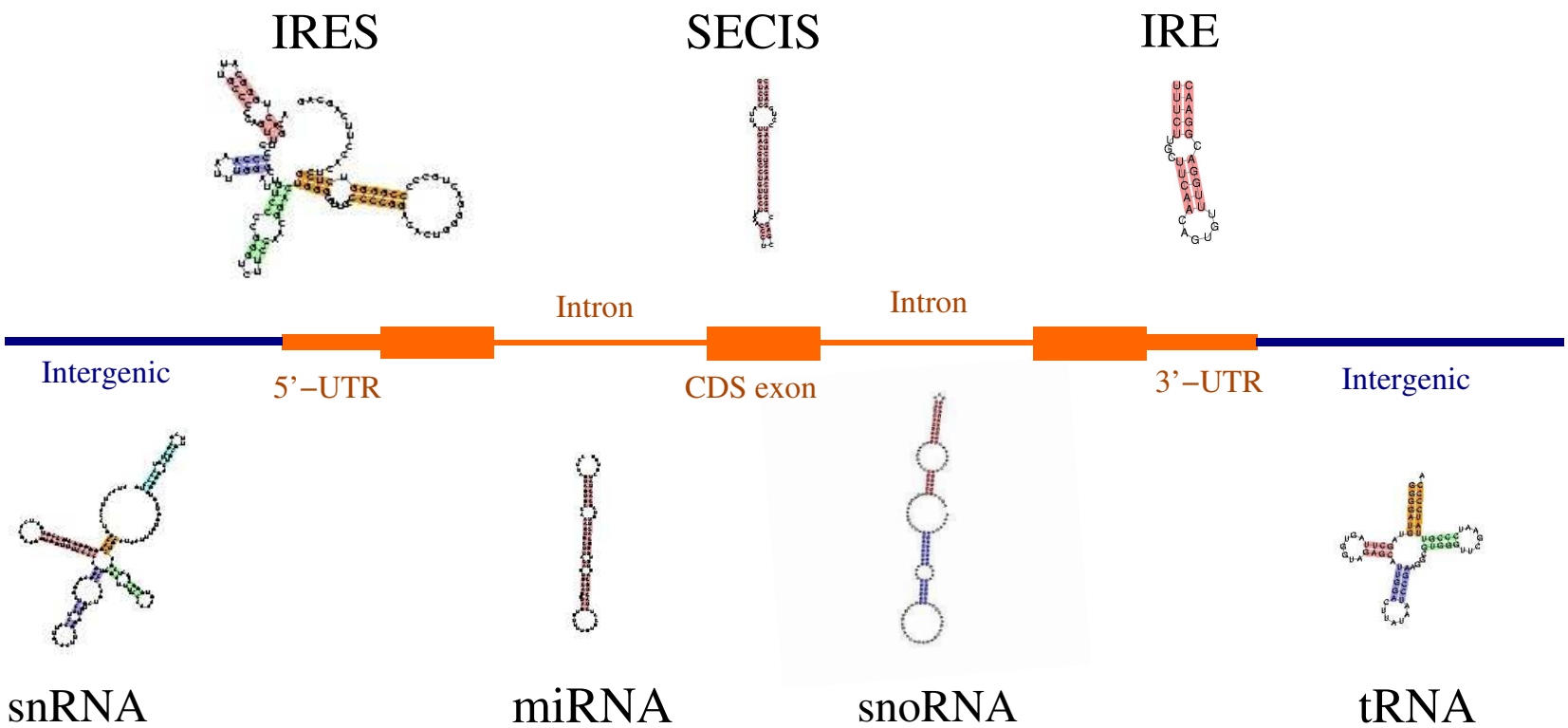
Finally, DNA appeared on the scene

- RNA or RNA-like molecules could have formed a pre-protein world.

Overview of RNA functions



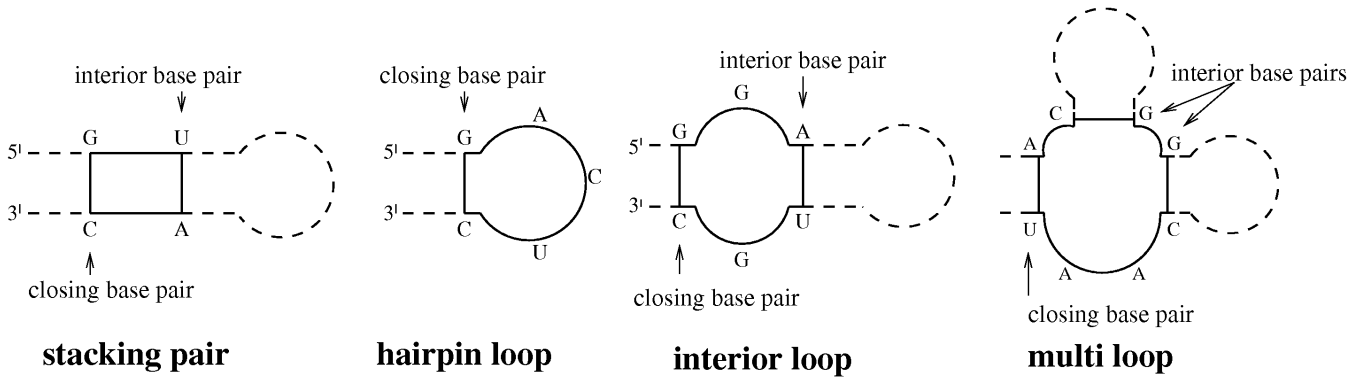
Examples of structured RNAs and their genomic context



Prediction of noncoding RNAs

- Compared to prediction of protein coding RNAs an extremely difficult problem:
  - No common strong statistical features in primary sequence such as start/stop codons, codon bias, open reading frame
  - ncRNAs are highly diverse (short, long, spliced, unspliced, processed, intron encoded, intergenic, antisense,...)
- Good progress in prediction for a subset of ncRNAs: structured ncRNAs

Prediction of RNA secondary structure



- The standard energy model expresses the free energy of a secondary structure $\mathcal{S}$ as the sum of the energies of its components L:

$$E(\mathcal{S}) = \sum_{L \in \mathcal{S}} E(L)$$

- The minimum free energy structure can be calculated by dynamic programming, e.g. by using `RNAfold`:

```
RNAfold < trna.fa
>AF041468
GGGGGUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGGUUCGAGUCCCCUUACCUCCA
(((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))). (-31.10)
```

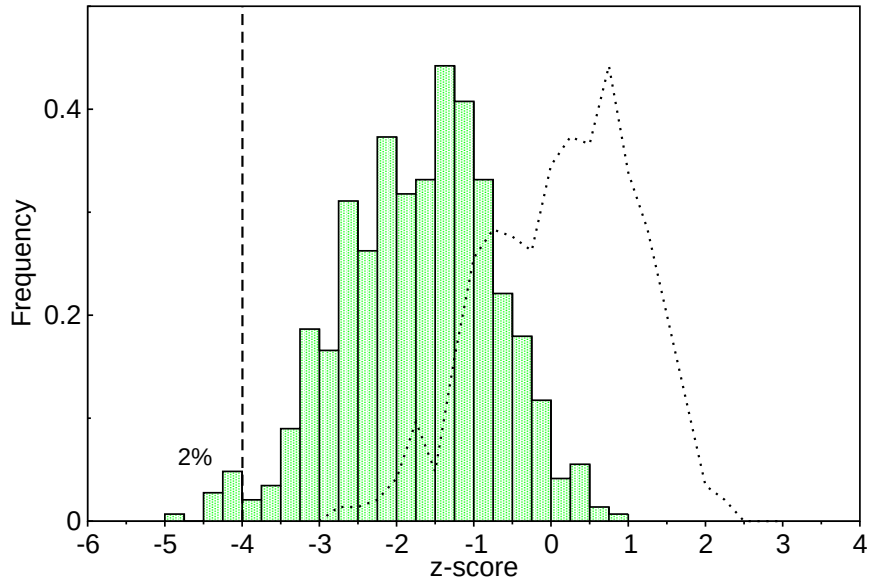
Significance of predicted RNA secondary structures: z-score statistics

- Has a natural occuring RNA sequence a lower minimum free energy (MFE) than random sequences of the same size and base composition?
  1. Calculate native MFE m.
  2. Calculate mean μ and standard deviation σ of MFEs of a large number of shuffled random sequences.
  3. Express significance in standard deviations from the mean as z-score

$$z = \frac{m - \mu}{\sigma}$$

- Negative z-scores indicate that the native RNA is more stable than the random RNAs.

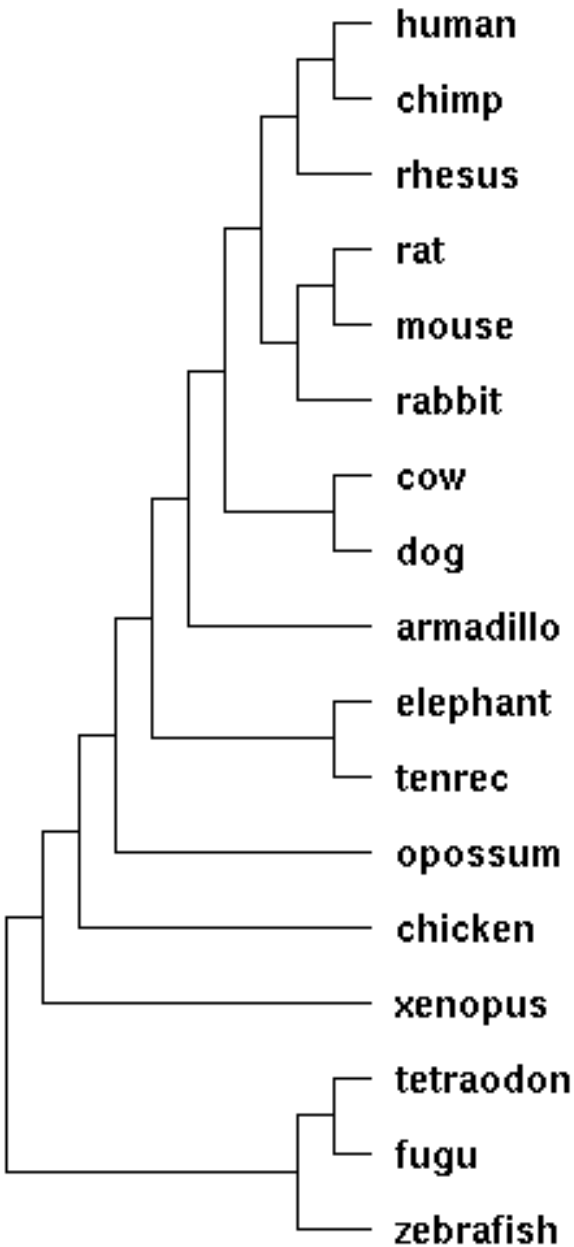
z-scores of structured RNAs



ncRNA Type	No. of Seqs.	Mean z-score
tRNA	579	−1.84
5S rRNA	606	−1.62
Hammerhead ribozyme III	251	−3.08
Group II catalytic intron	116	−3.88
SRP RNA	73	−3.37
U5 spliceosomal RNA	199	−2.73

Washietl & Hofacker, *J. Mol. Biol.* (2004) 342:19

Comparative genomics at our hands



- 30+ vertebrate genomes
- 12+ drosophila genomes
- 20+ yeast genomes
- and many more...

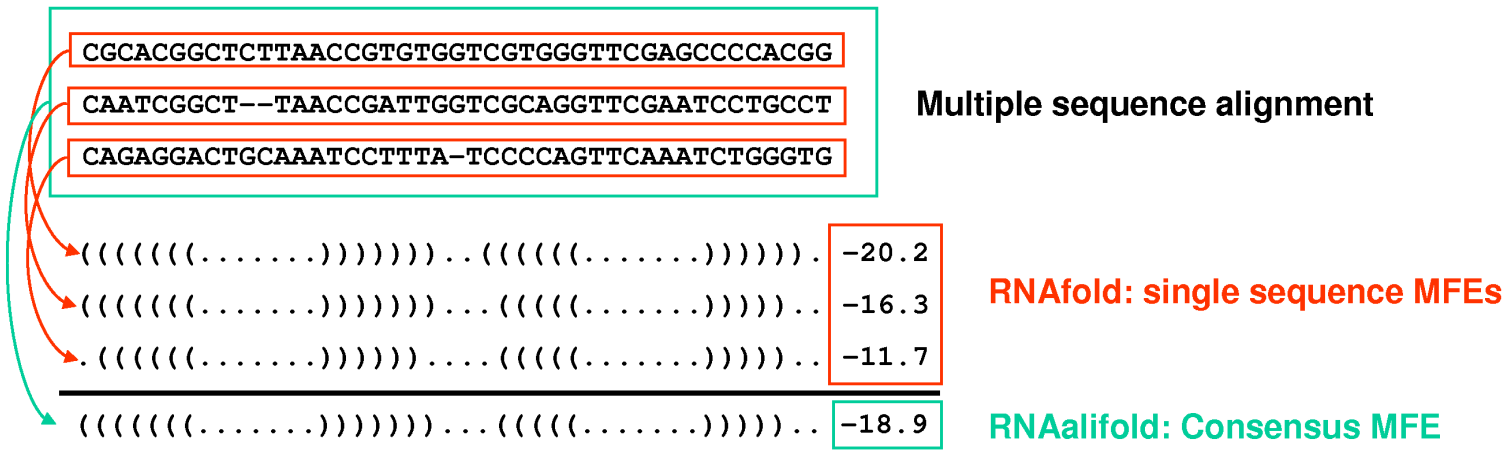
Consensus folding using RNAalifold

- `RNAalifold` uses the same algorithms and energy parameters as `RNAfold`
- Energy contributions of the single sequences are averaged
- Covariance information (e.g. compensatory mutations) is incorporated in the energy model.
- It calculates a consensus MFE consisting of an energy term and a covariance term:

```
((((((..(((.........))).(((((.......)))).....(((((.......)))))))))))).
GTTTCCGTAGTGTAGCGGTTATCACATTCGCCTCACACGCGAAAGGTCCCCGGTTCGATCCCGGGCGGAAACA
GTTTCCGTAGTGTAGTGGTTATCACGTTCGCCTAACACGCGAAAGGTCCCCGGTTCGAAACCGGGCGGAAACA
GTTTTCGTAGTGTAGTGGTTATCACGTGTGCTTCACACGCACAAGGTCCCCGGTTCGAACCCGGGCGAAAACA
**** ********** ********* *  ** * ******  ****************  ******* *****
(-24.76 = -23.43 +  -1.33)
```

Hofacker, Fekete & Stadler, *J. Mol. Biol.* (2002) 319:1059

The structure conservation index

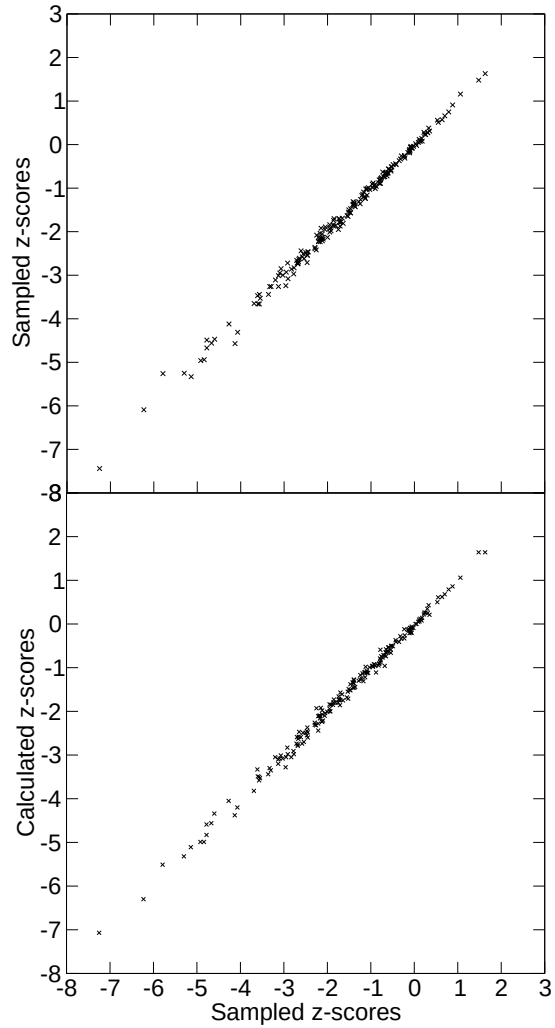


$$\text{SCI} = \frac{\text{Consensus MFE}}{\text{Mean single MFEs}}$$

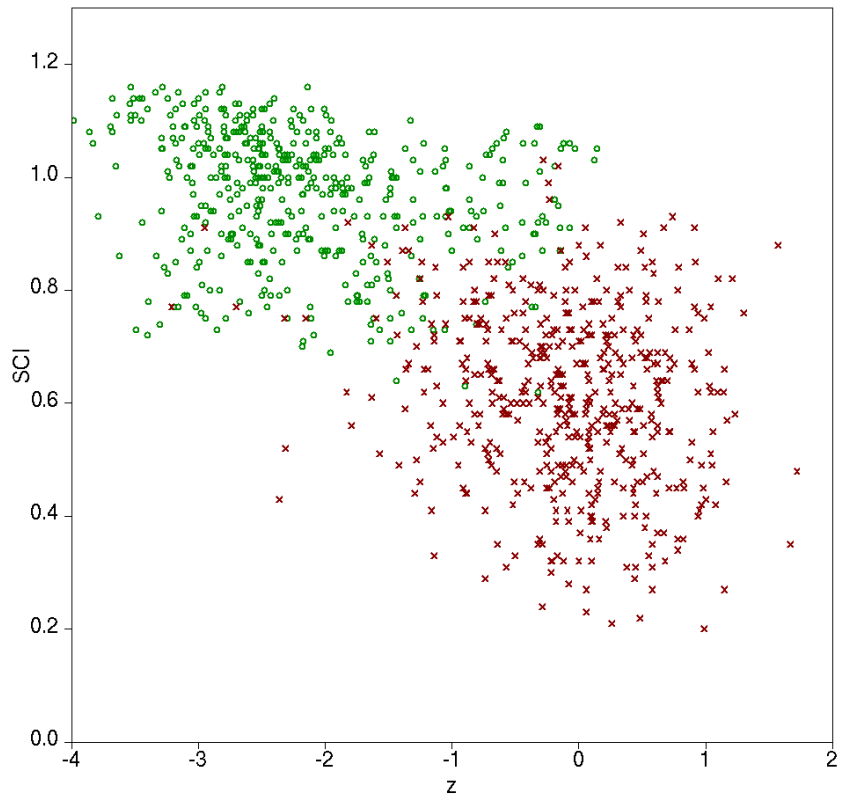
- The SCI is an efficient and convenient measure for secondary structure conservation.

Efficient calculation of stability z-scores

- The significance of a predicted MFE structure can be expressed as z-score which is normalized w.r.t. sequence length and base composition.
- Traditionally, z-scores are sampled by time-consuming random shuffling.
- The shuffling can be replaced by a regression calculation which is of the same accuracy.



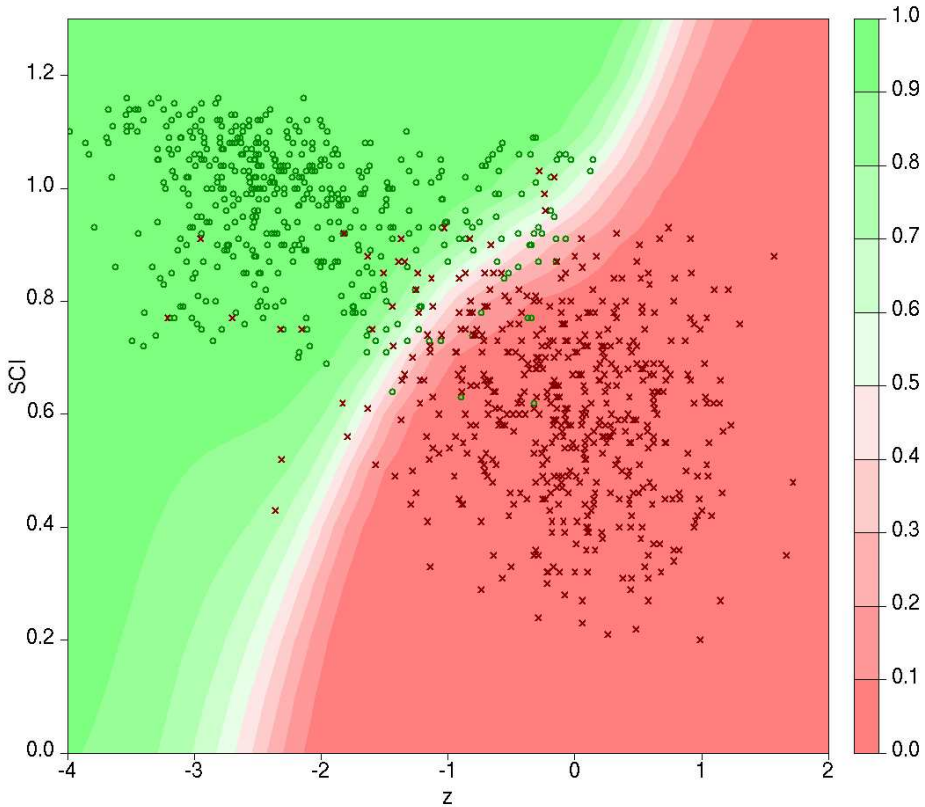
SVM classification based on both scores



- Both scores separate native ncRNAs from controls in two dimensions.

Washietl, Hofacker & Stadler, *Proc. Natl. Acad. Sci. USA* (2005) 33:2433

SVM classification based on both scores



- Both scores separate native ncRNAs from controls in two dimensions.
- A support vector machine is used for classification: `RNAz`.

Washietl, Hofacker & Stadler, *Proc. Natl. Acad. Sci. USA* (2005) 33:2433

Probabilistic approaches to fold RNA

- Hidden Markov Models are commonly used in computational biology to assign "states" to a sequence: e.g. exons in DNA sequence, conserved regions in alignments,
- Can we use a similar approach to parse a RNA sequence into structural states?

```
AGCUCUGAGGUGAUUUUCAUAUUGAAUUGCAAAUUCGAAGAAGCAGCUUCAAACCUGCCGGGGCUU
((((((((..((((...)))).((((((((...))))))))....(((((.......)))))))))))).
```

- The HMM framework needs to be extended to allow for nested correlations

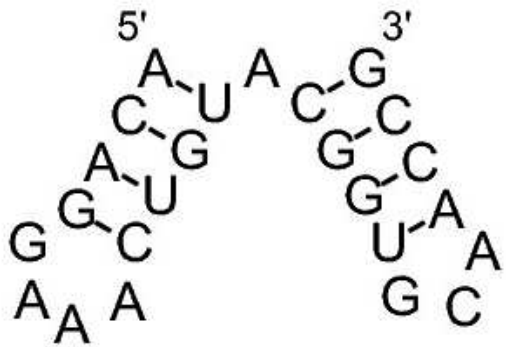
Context free grammars

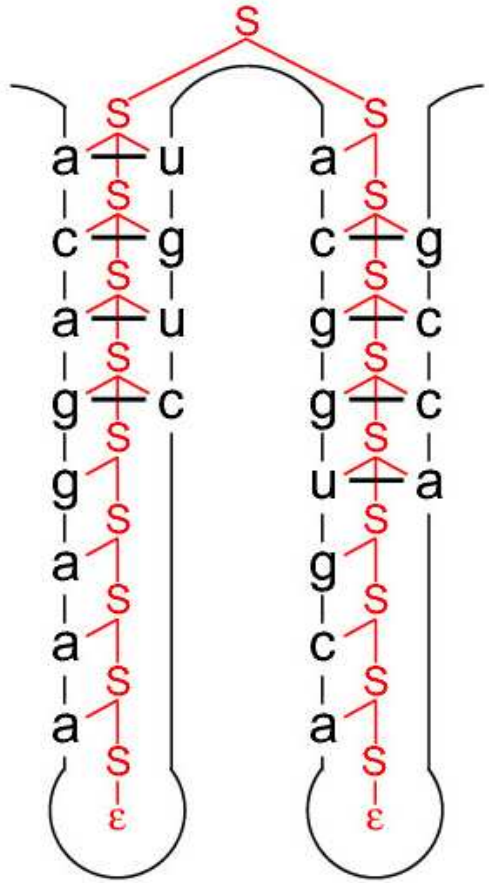
- A context-free grammar can be defined by $\mathcal{G}(V, T, P, S)$ where:
  - V is a finite set of nonterminal symbols ("states"),
  - T is a finite set of terminal symbols,
  - P is a finite set of production rules and
  - S is the initial (start) nonterminal ($S \in V$).
- A simple palindrome grammar: $V = \{S\}$, $T = \{a, b\}$, $P = \{S \rightarrow aSa, S \rightarrow bSb, S \rightarrow \epsilon\}$
  - Efficiently describes the set of all palindromes over the alphabet $\{a, b\}$.
  - Example production: $S \rightarrow aSa \rightarrow abSba \rightarrow abbSbba \rightarrow abbbba$
- Given the CFG $\mathcal{G}(V, T, P, S)$, we get a *stochastic* CFG (SCGF) by assigning each production rule $\alpha \in P$ a probability $Prob(\alpha)$ such that: $\sum_{\alpha} Prob(\alpha) = 1$

A simple RNA grammar

- $V = \{S\}$, $T = \{a, c, g, u\}$, $P =$
  - $S \rightarrow aSu|uSa|gSc|cSg|uSg|gSu$
  - $S \rightarrow aS|uS|gS|cS$
  - $S \rightarrow Sa|Su|Sa|Sc$
  - $S \rightarrow SS$
  - $S \rightarrow \epsilon$
- Shorthand $S \rightarrow aS\hat{a}|aS|Sa|SS|\epsilon$

Parse tree





- One possible parse tree Π of the string $x =$ ACAGGAAACUGUACGGUGCAACCG and its correspondence to a RNA secondary structure (nonterminals: red, terminals: black)

RNA folding using SCFG

- Find the parse tree of maximum probability using a Nussinov style recursion.
- $\gamma(i,j)$ is the maximum $log(Prob)$ for subsequence (i,j)
- Initialization: $\gamma(i,i-1) = \log p(S \to \epsilon)$

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j-1) + \log(Prob(S \to x_i S x_j) \\ \gamma(i+1,j) + \log(Prob(S \to x_i S) \\ \gamma(i,j-1) + \log(Prob(S \to S x_j) \\ \max_{i<k<j}\{\gamma(i,k) + \gamma(k+1,j) + \log(Prob(S \to SS)\} \end{cases}$$

Standard algorithms for SCFG

- Given a parameterized SCFG($\mathcal{G}, \Omega$) and a sequence x, the Cocke-Younger-Kasami (CYK) dynamic programming algorithm finds an optimal (maximum probability) parse tree $\hat{\pi}$:

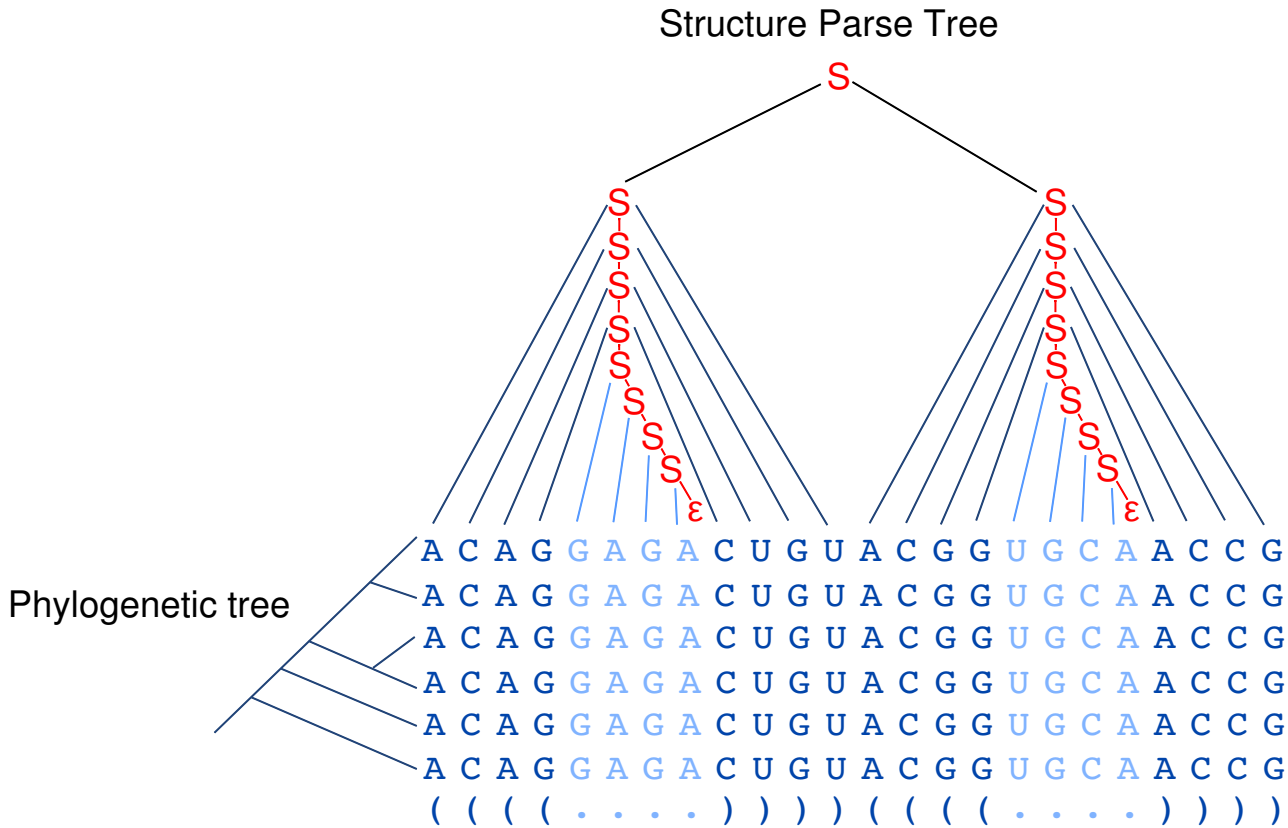
$$\hat{\pi} = \arg\max_{\pi} Prob(\pi, x|\mathcal{G}, \Omega)$$

- The *Inside algorithm*, is used to obtain the total probability of the sequence given the model summed over all parse trees,

$$Prob(x|\mathcal{G}, \Omega) = \sum_{\pi} Prob(x, \pi|\mathcal{G}, \Omega)$$

- Analogies to thermodynamic folding:
  - CYK $\leftrightarrow$ Minimum Free energy (Nussinov/Zuker)
  - Inside/outside algorithm $\leftrightarrow$ Partition functions (McCaskill)
- Analogies to Hidden Markov models:
  - CYK Minimum $\leftrightarrow$ Viterbi's algorithm
  - Inside/outside algorithm $\leftrightarrow$ Forward/backwards algorithm

Evofold: Phylo SCFGs



Single sequence:
Terminal symbols are bases or base-pairs
Emission probabilities are base frequencies in loops and paired regions

Phylo-SCFG:
Terminal symbols are single or paired alignment columns
Emission probabilities calculated from phylogenetic model and tree using Felsenstein's algorithm

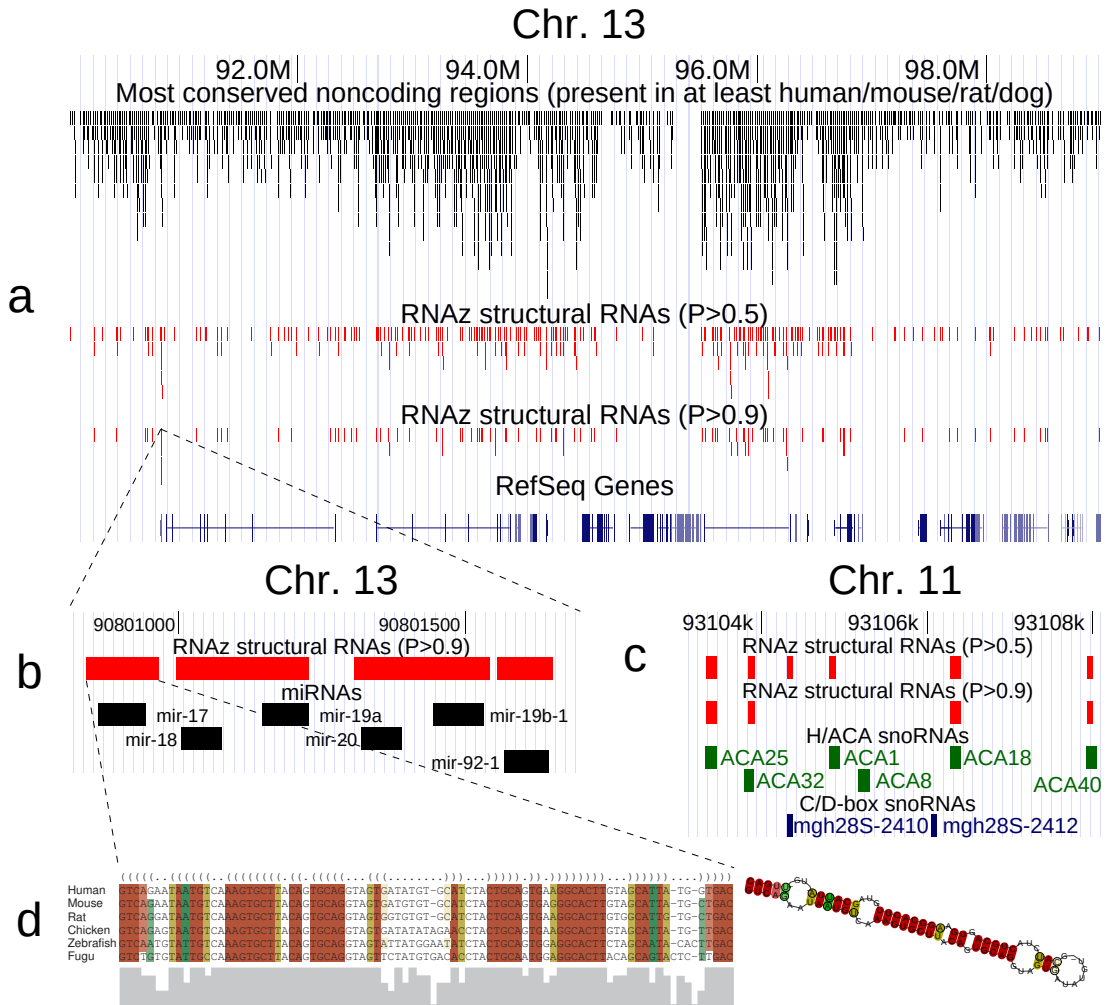
4x4 Matrix for single columns

16x16 Matrix for paired columns

EvoFold

- Structural RNA gene finding: EvoFold
  - Uses simple RNA grammar
  - Two competing models:
    - Non-structural model with all columns treated as evolving independently
    - Structural model with dependent and independent columns
  - Sophisticated parametrization

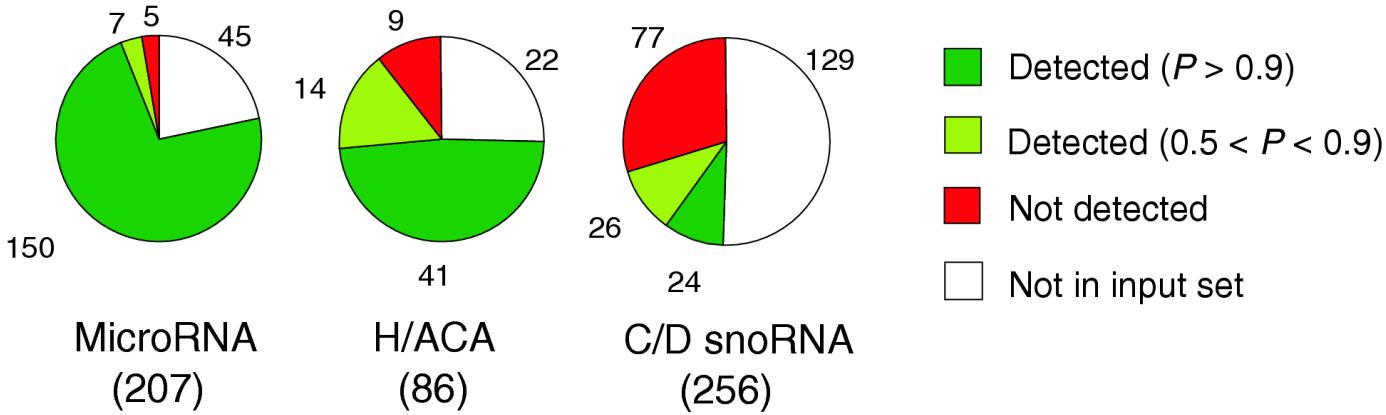
Screening the human genome with RNAz



- Large scale comparative screen of mammals/vertebrates
- $\approx$ 5% of the best conserved non-coding regions
- $\rightarrow$ 438,788 alignments covering 82.64 MB (2.88% of the genome)

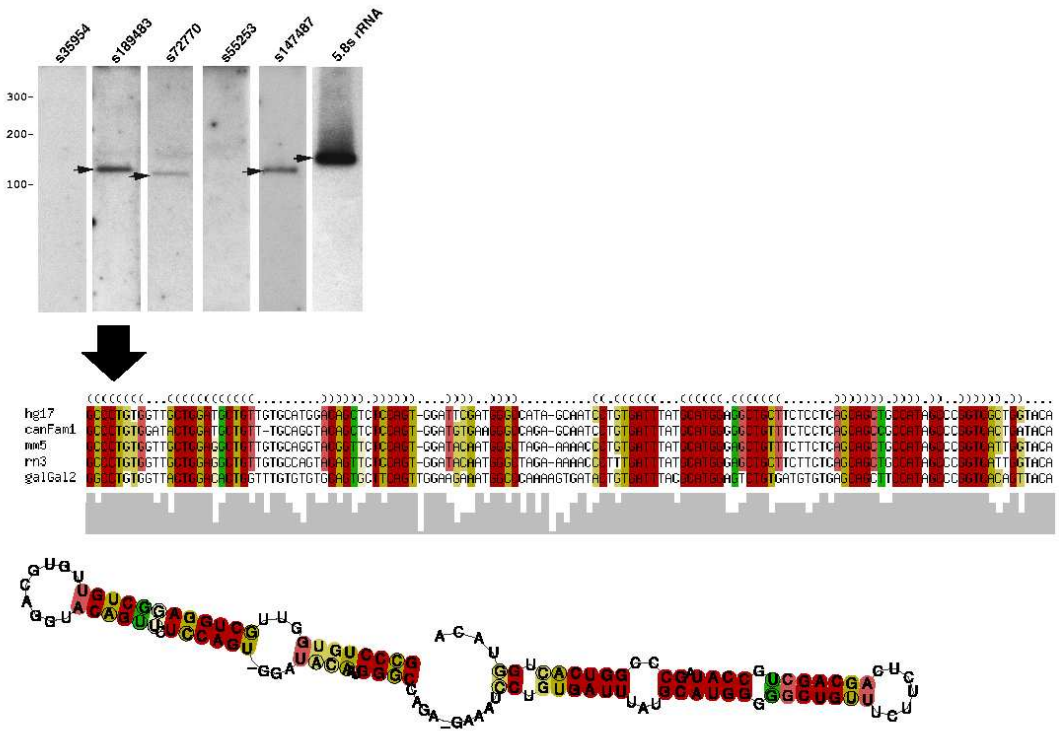
Washietl, Hofacker & Stadler, *Nat. Biotech.* (2005) 23:1383

Detection performance of well-known small ncRNAs



Washietl, Hofacker & Stadler, *Nat. Biotech.* (2005) 23:1383

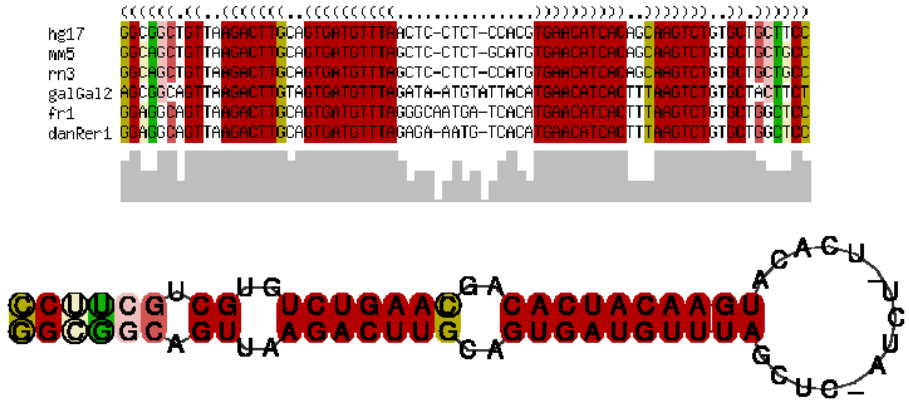
Searching for H/ACA snoRNAs



- Two stems of at least 15 pairs
- Unpaired hinge
- `ACA` in last 20 nucleotides
- $\rightarrow$ 137 candidates (28 known), 30-40 show typical structure upon visual inspection, 15 have canonical H-box motif `ANANNA`
- Five candidates were tested, 3 found on Northerns in HeLa cells

Washietl, Hofacker & Stadler, *Nat. Biotech.* (2005) 23:1383

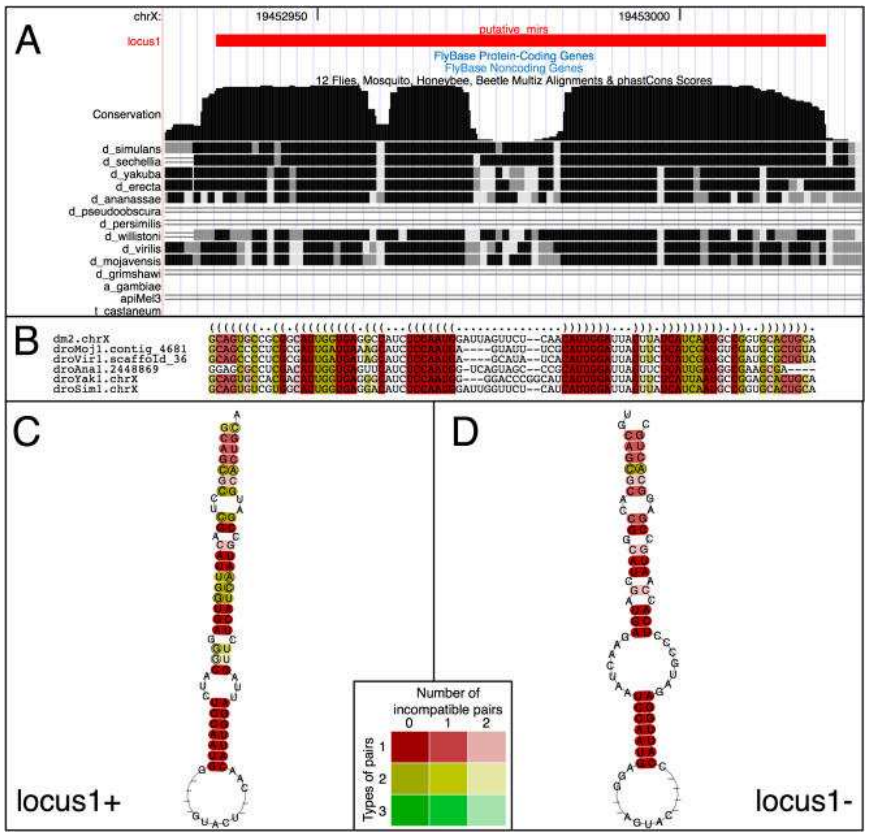
Searching for miRNA precursors

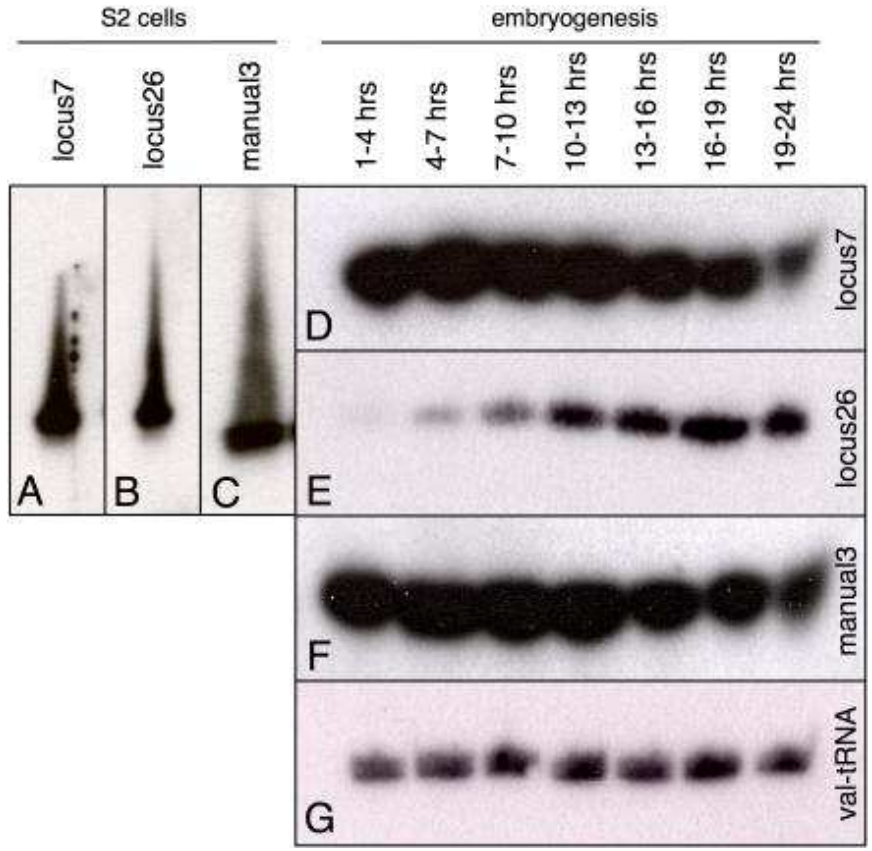


- Stem with at least 20 pairs
- Mean z-score < -3.5
- 22nt window with more than 95% identity
- $\rightarrow$ 312 candidates (109 known miRNAs)
- Automatized in `RNAmicro` (Hertel und Stadler, Bioinformatics 22:e197, 2006)

Washietl, Hofacker & Stadler, *Nat. Biotech.* (2005) 23:1383

miRNA precursors in Drosophila (Sandman & Cohen)

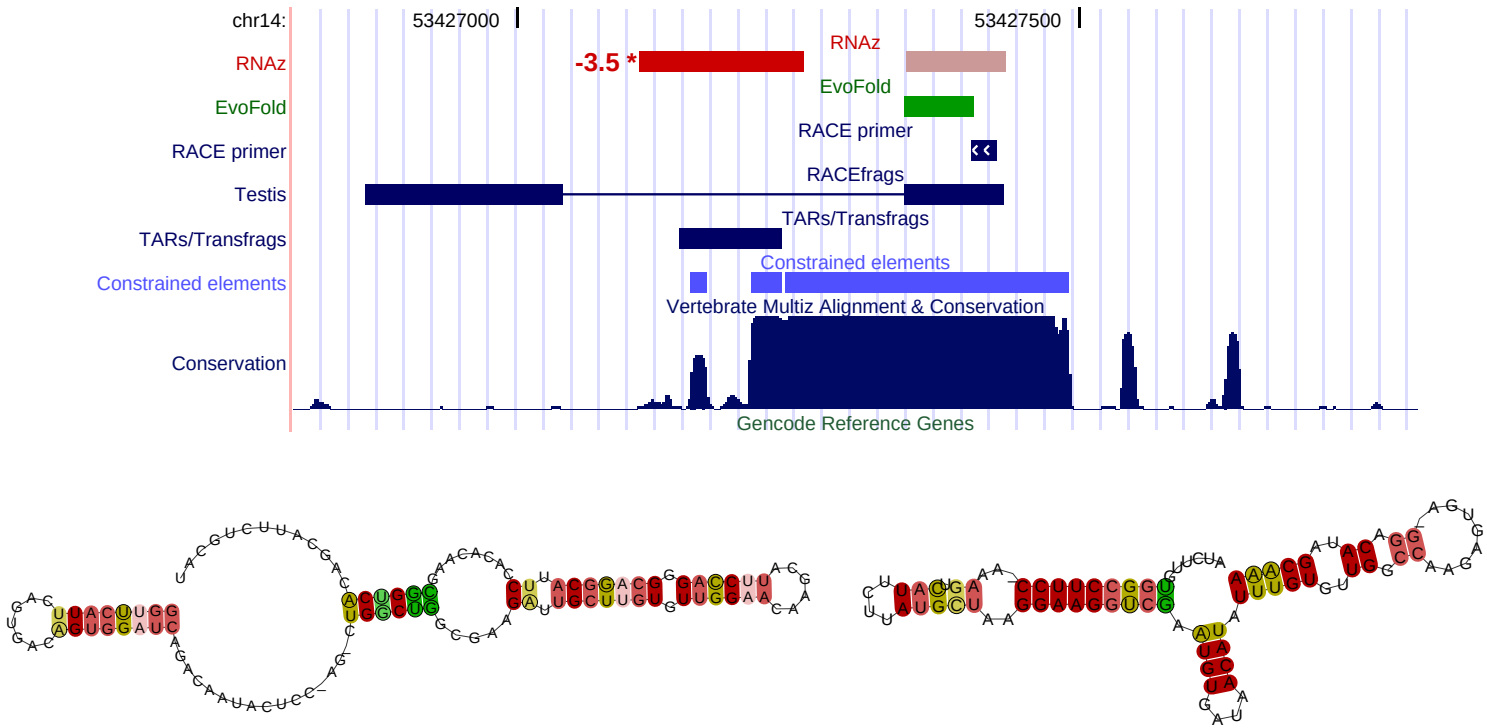




- 56 miRNAs predicted using `RNAz` and evolutionary patterns.
- 22 (39%) verified (16 Northern, 19 small RNA libraries, 13 both)

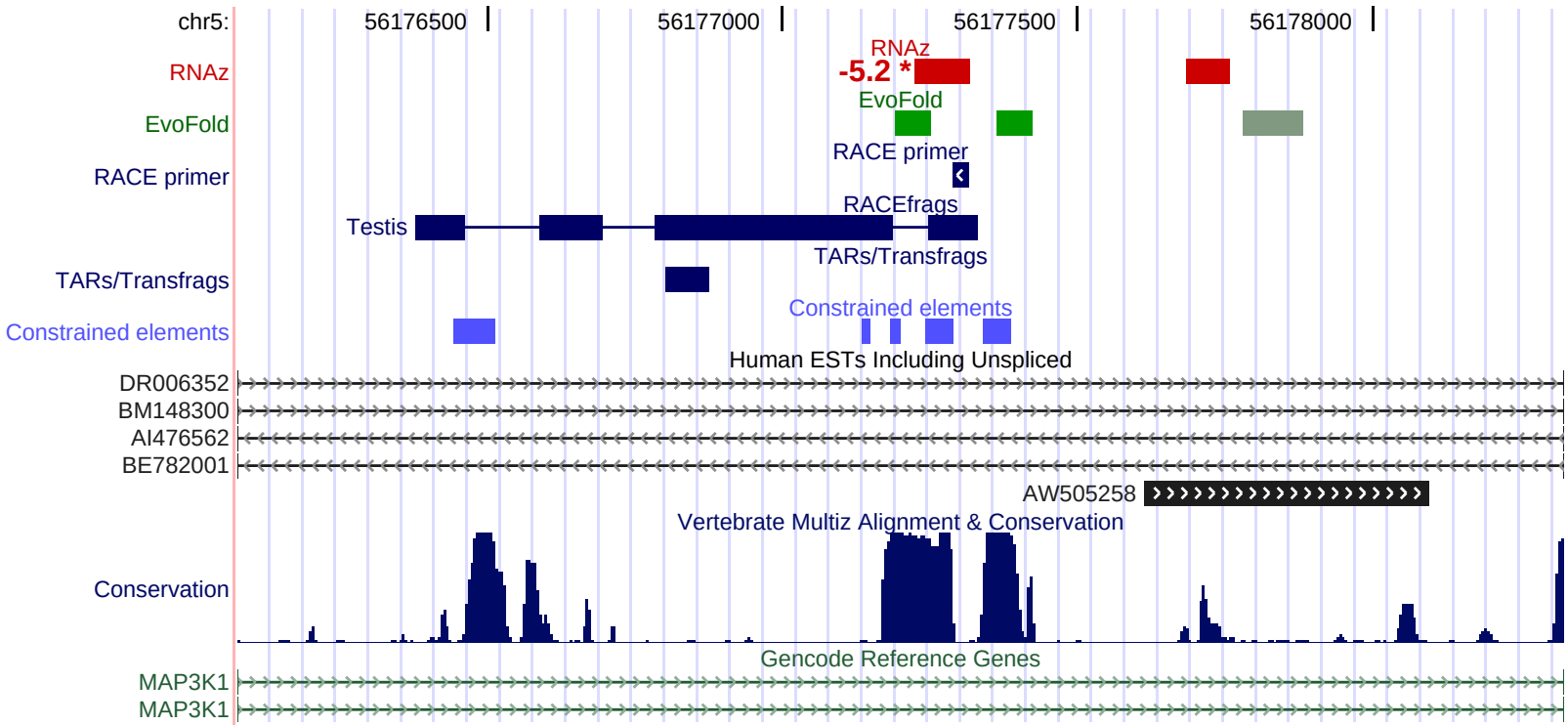
Sandman & Cohen, *PLoS One* (2007) 2:e1265

Intergenic RNAs



Washietl, Pedersen, Korbel *et al.*, *Genome Res.* (2007) 17:852

Intronic RNAs

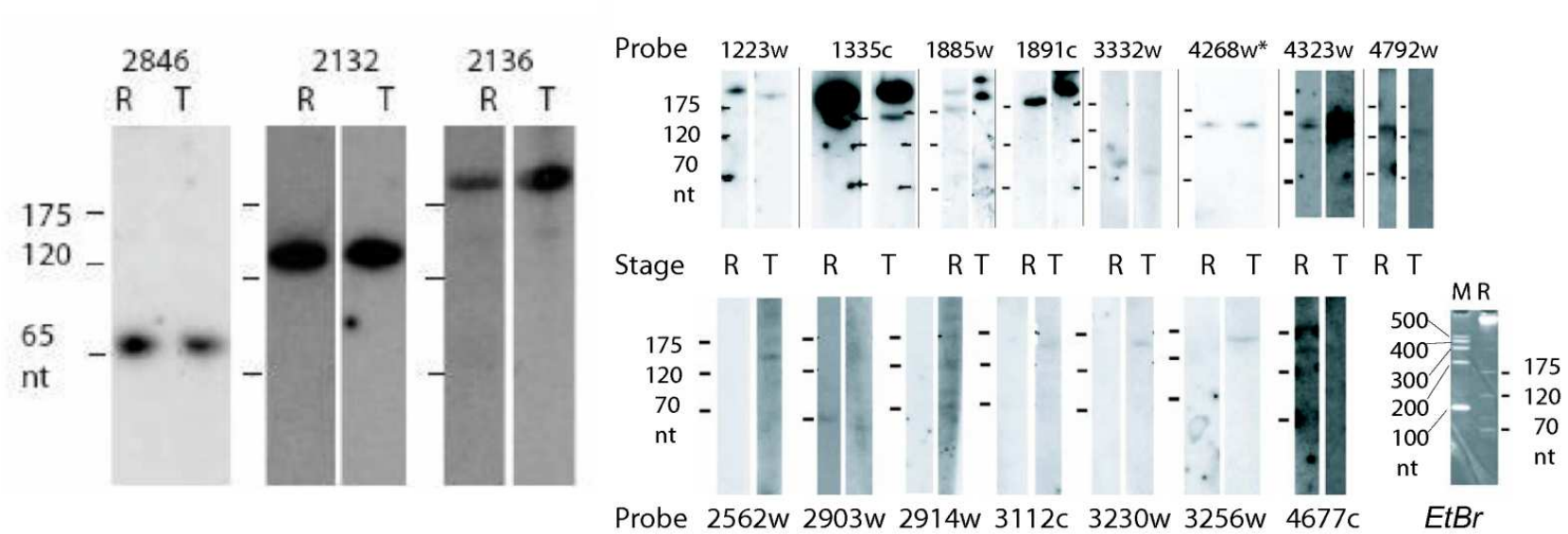


Washietl, Pedersen, Korbel *et al.*, *Genome Res.* (2007) 17:852

RNAz screen in other genomes

- **Drosophila melanogaster**: Rose *et al.*: *BMC Genomics* 2007, 8:406.
- **Ciana intestinalis**: Missal, Rose & Stadler: *Bioinformatics* 2005, 21 Suppl 2:77-78
- **Caenorhabditis elegans**: Missal *et al.*: *J Exp Zoolog B Mol Dev Evol* 2006, 306(4):379-392.
- **Saccharomyces cerevisiae**: Steigele *et al.*: *BMC Biol* 2007, 5:25-25.
- **Plasmodium falciparum**: Mourier *et al.*: *Genome Res.*, 2008

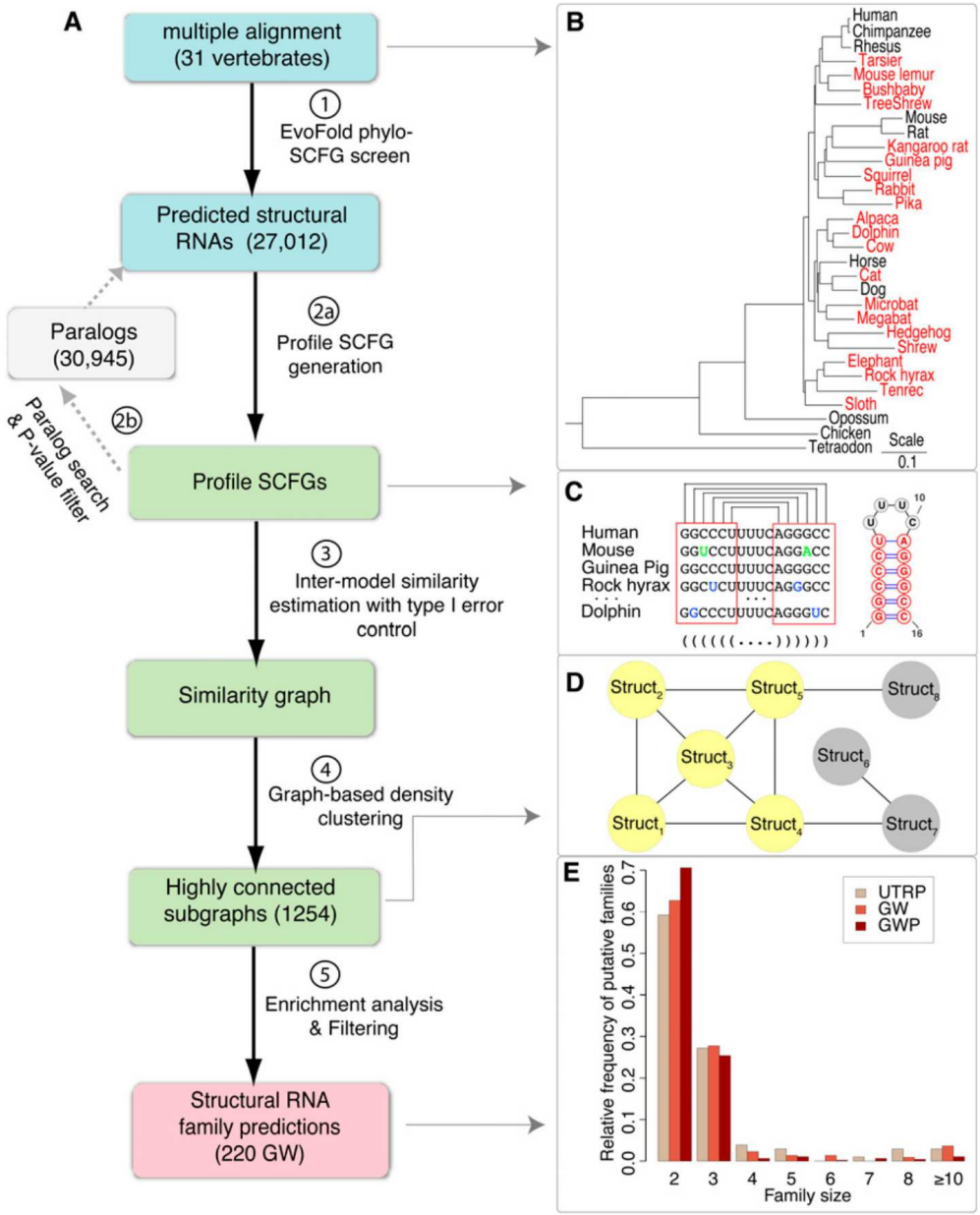
A RNAz screen in Plasmodium (Mourier et. al)



- 22 of 78 tested high scoring `RNAz` candidates (28%) were verified by Northern blot analysis.

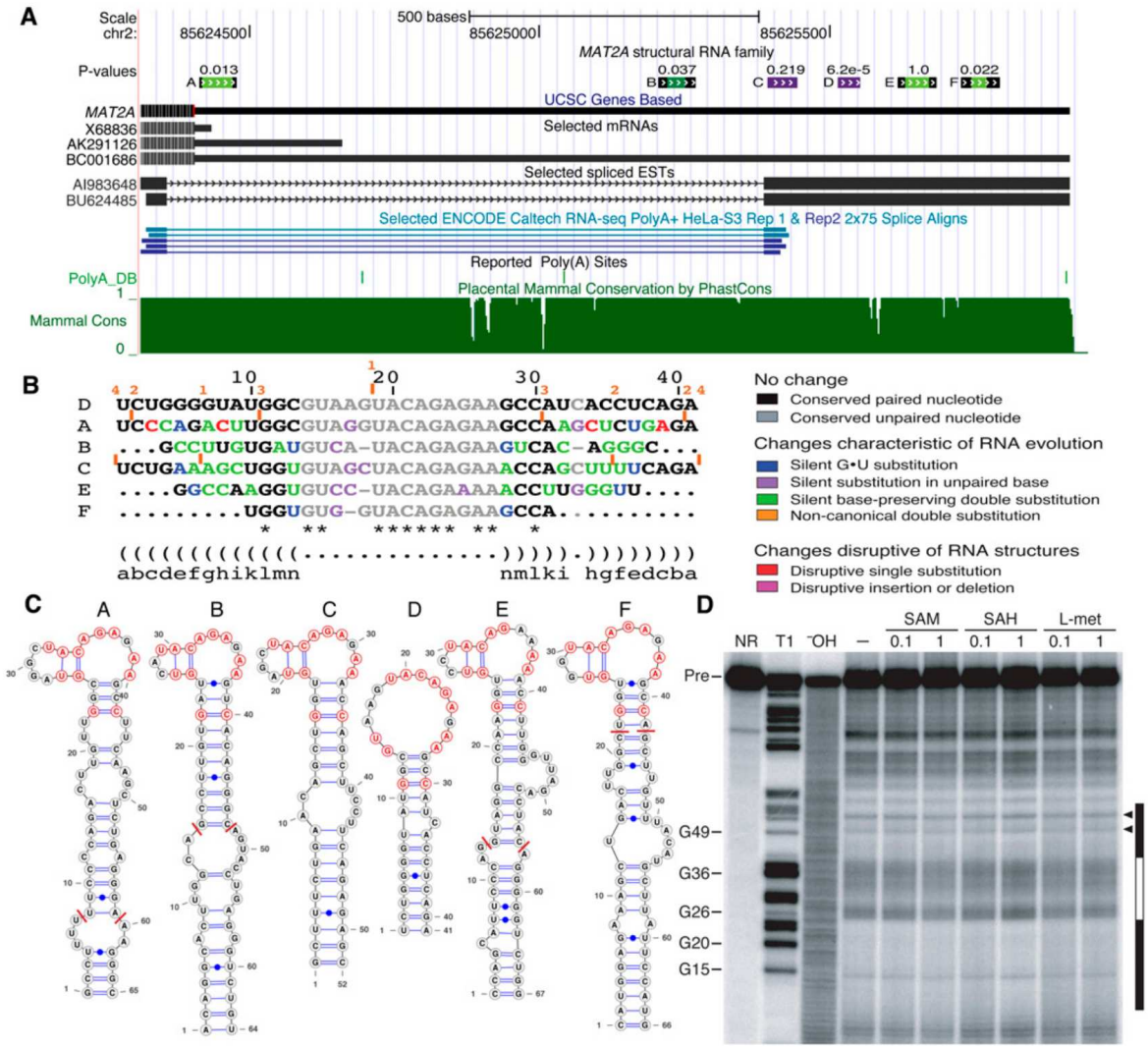
Mourier et al. *Genome Res.* 2008

Structure family identification using EvoFold+EvoFam



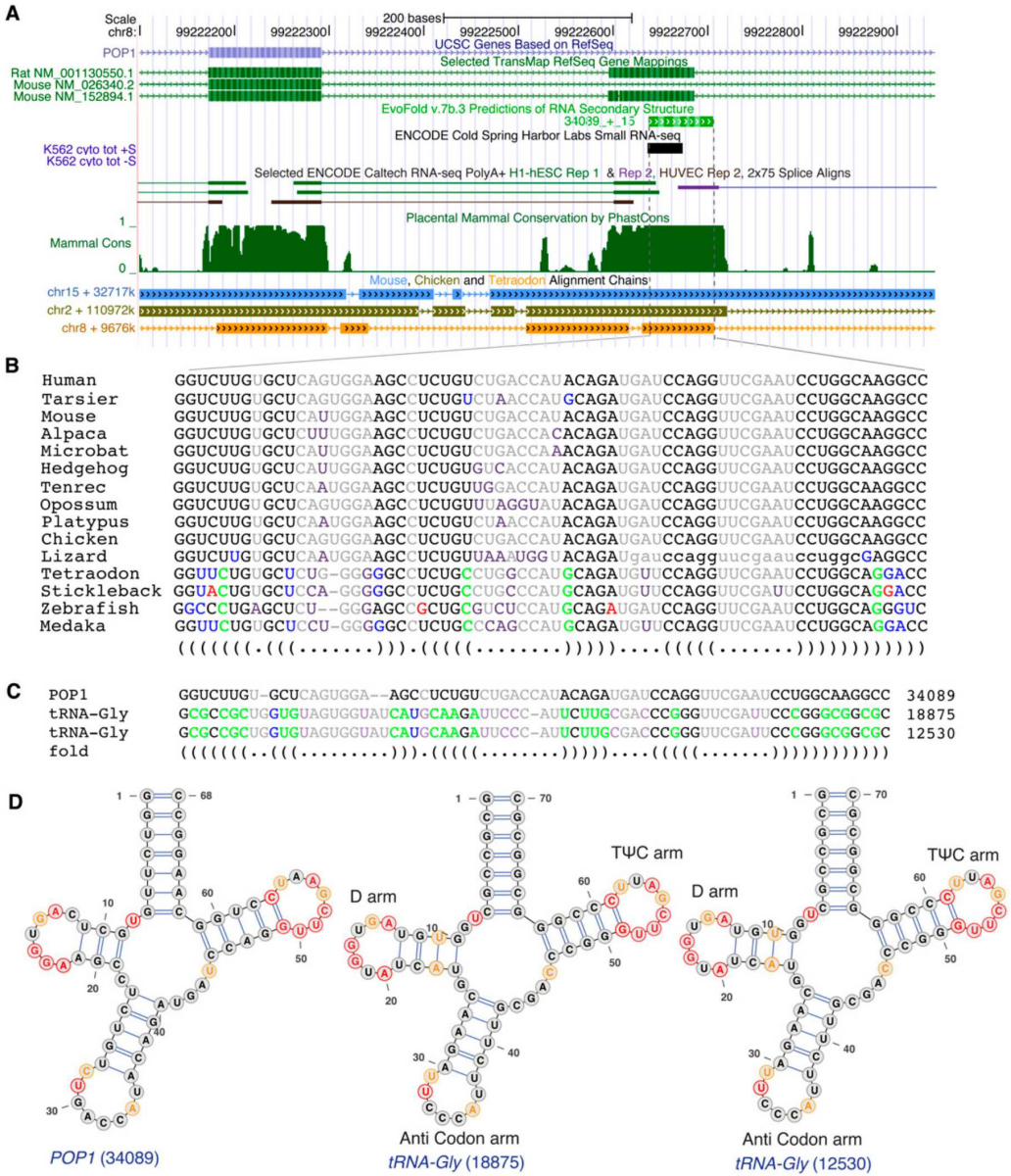
Parker et al. *Genome Res.* 2011

Family of hairpins in 39-UTR of MAT2A



Parker et al. *Genome Res.* 2011

tRNA like structures in intron of POP1



Parker et al. *Genome Res.* 2011