RNA Bioinformatics

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science & Interdisciplinary Center for Bioinformatics, **University of Leipzig**

Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

CSSS, June 2006

Overview

- PART 1: RNA Structures and How to Compute Them
- PART 2: RNA Landscapes
- PART 3: The Modern RNA World

PART1

Why RNA?

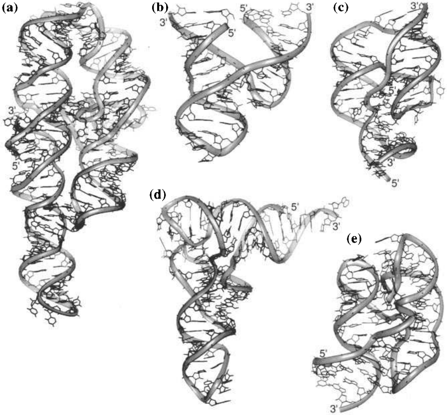
- until relatively recently:
  Central Dogma of Molecular Biology
  DNA → RNA → Protein
  DNA = "genetic memory", RNA = working copy, proteins do the work
- around 1980: discovery of catalytic RNAs (Nobelprize for Tom Cech and Sidney Altman)
  nevertheless long considered "exotic" remnants from the ancient RNA world
- around 2000: structure of the ribosome showns that the ribosome is an "RNA enzyme"
- around 2000: microRNAs are discovered as a large class of regulatory RNAs that inhibit translation of proteins
- 2006: the ENCODE project shows that human gene expression is quite different from textbook knowledge

RNA Bioinformatics

RNA Secondary Structures are an appropriate level of description

- explain the thermodynamics of RNA Structures
- often highly conserved in evolution
- can be computed efficiently

Many Functional RNAs are Structured

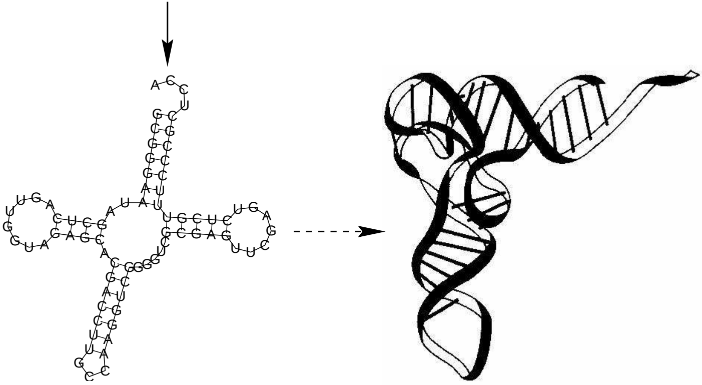


(a) Group I intron P4–P6 domain
(b) Hammerhead ribozyme
(c) HDV ribozyme
(d) Yeast tRNA^{phe}
(e) L1 domain of 23S rRNA

Hermann & Patel, JMB 294, 1999

The RNA Model

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA



Formal Definition

A secondary structure on a sequence s is a collection of pairs (i, j) with $i < j$ such that

- Base pairing rules are respected, i.e., $(i, j) \in \Omega$ implies (s_i, s_j) form an allowed pair (**GC, CG, AU, UA, GU, UG**)
- Each base is involved in at most one pair, i.e., Ω is a matching, $(i, j), (i, k) \in \Omega$ implies $j = k$ and $(i, k), (j, k) \in \Omega$ in implies $i = j$.
- $(i, j)\Omega$ implies $|j - i| > 3$ (sterical constraint)
- No-crossing rule: $(i, j), (k, l) \in \Omega$ and $i < k$ implies either $i < k < l < j$ or $i < j < k < l$.
  This excludes so-called pseudoknots

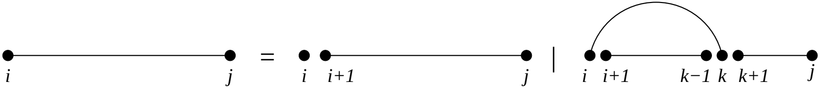
Let's count the structures . . .

Counting secondary structures. Given a sequence of length n.
$\Pi_{kl} = 1$ if sequence positions k, l **can** form a pair **GC, CG, AU, UA, GU, UG** and $\Pi_{kl} = 0$ otherwise.
N_{kl} = number of structures of the *subsequence* from k to l.
Basic recursion:

$$\bullet\, xxxxxxx + \sum (xxxx)xxxx$$

$$N_{kl} = N_{k+1,l} + \sum_{j=k+m}^{l} \Pi_{kj} N_{k+1,j-1} N_{j+1,l}$$

RNA Folding in a nutshell



$$N_{ij} = N_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} N_{i+1,k-1} N_{k+1,j}$$

$$E_{ij} = \min \left\{ E_{i+1,j} + \min_{\substack{k \\ (i,k)\text{pair}}} \left(E_{i+1,k-1} + E_{k+1,j} + \varepsilon_{ik} \right) \right\}$$

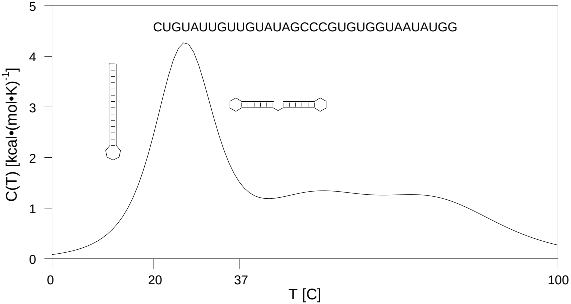
$$Z_{ij} = Z_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\varepsilon_{ik}/RT)$$

Partition function: $Z = \sum_{\Omega} \exp(-E(\Omega)/RT)$

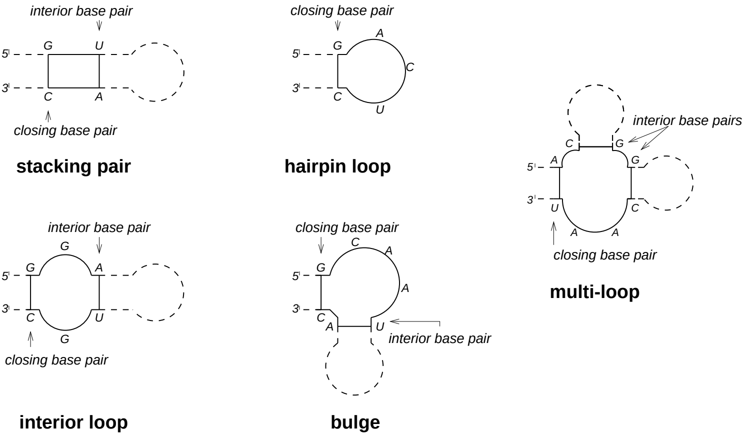
A word on the Partition Function

The partition function is the link between the combinatorics of the structures (in general: states in an ensemble) and the thermodynamic properties of the physical ensemble, e.g.:

- Free energy $G = -RT \ln Z$
- Expected Energy $\langle E \rangle = RT^2 \frac{\partial \ln Z}{\partial T}$
- Heat Capacity $C_p = -T \frac{\partial^2 G}{\partial T^2}$



Realistic Energy Model



interior base pair

closing base pair

stacking pair

closing base pair

hairpin loop

interior base pairs

closing base pair

multi-loop

interior base pair

closing base pair

interior loop

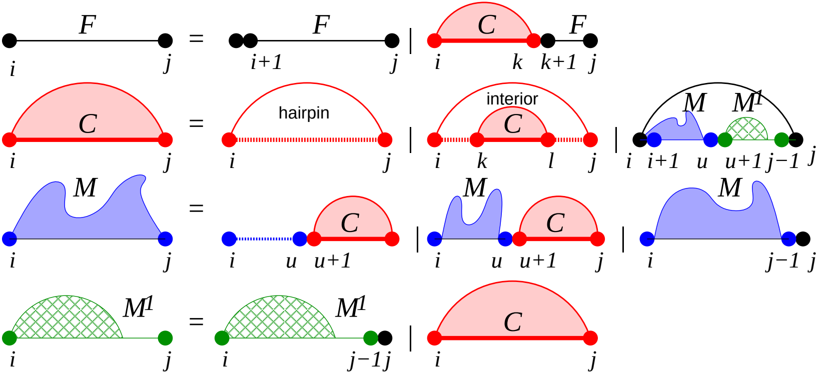
closing base pair

interior base pair

bulge

Parameters from large number of melting experiments by Douglas Turner, David Matthews, John Santa Lucia, and others

Recursions for Linear RNAs



Recursions for Linear RNAs

F_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$.

C_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that i and j form a base pair.

M_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that that $x[i,j]$ is part of a multiloop and has at least one component, i.e., a sub-sequence that is enclosed by a base pair.

M^1_{ij} free energy of the optimal substructure on the subsequence $x[i,j]$ subject to the constraint that that $x[i,j]$ is part of a multiloop and has exactly one component, which has the closing pair i,h for some h satisfying $i \leq h < j$.

Recursions for Linear RNAs

$$
\begin{aligned}
F_{ij} &= \min\left\{F_{i+1,j},\ \ \min_{i<k\leq j} C_{ik} + F_{k+1,j}\right\} \\
C_{ij} &= \min\Big\{\mathcal{H}(i,j),\ \ \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l), \\
&\qquad\quad \min_{i<u<j} M_{i+1,u} + M^{1}_{u+1,j-1} + a\Big\} \\
M_{ij} &= \min\Big\{\min_{i<u<j} (u-i-1)c + C_{u+1,j} + b, \\
&\qquad\quad \min_{i<u<j} M_{i,u} + C_{u+1,j} + b,\ \ M_{i,j-1} + c\Big\} \\
M^{1}_{ij} &= \min\left\{M^{1}_{i,j-1} + c,\ \ C_{ij} + b\right\}
\end{aligned}
$$

Backward Recursion: Base Pairing Probabilities

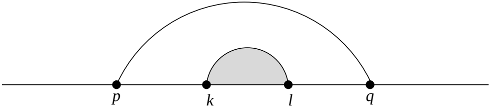
$$p_{ij} = \frac{Z_{1,i-1}\widehat{Z}_{i,j}Z_{j+1,n}}{Z_{1,n}} + \sum_{k<i}\sum_{l>j} p_{kl}\Xi_{ij,kl}\,.$$

$\Xi_{ij,kl}$ is a ratio of the two partition functions:
$\widehat{Z}_{ij,kl}$ ... both i, j and k, l pair
$\widehat{Z}_{kl}$ ... k, l pair.
Simplest case:
$\widehat{Z}_{ij,kl} = Z_{k+1,i-1}\widehat{Z}_{ij}Z_{j+1,l-1}\zeta_{kl}$ where $\zeta_{kl} = \exp(-\beta_{kl}/RT)$ is the Boltzman factor of the pairing energy

Backward recursion: full model

Backward recursion:

$$\begin{aligned} P_{kl} = P^{\circ}_{kl} + \sum_{p<k;q>l} P_{pq} \frac{Z^{B}_{k,l}}{Z^{B}_{p,q}} \Bigg\{ & e^{-\mathcal{I}(p,q,k,l)} \\ & + \left(\sum_{p<u<k} Z^{M}_{p+1,u} Z^{M1}_{u+1,k-1}) \right) e^{-(a+(q-l-1)c)} \\ & + \left(\sum_{l<u<q} Z^{M}_{l+1,u} Z^{M1}_{v+1,q-1}) \right) e^{-(a+(k-p-1)c)} \\ & + Z^{M}_{p+1,k-1} Z^{M}_{l+1,q-1} \Bigg\} \end{aligned}$$

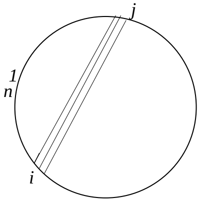


Single-Stranded Circular RNAs

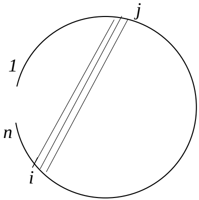
- Viroid RNA
- Hepatitis Delta Virus Genome
- Cryptic by-products of splicing formed intronic sequence
- Circularized C/D box snoRNAs were recently reported in *Pyrococcus furiosus*
- Synthetic constructs for *in vitro* selection

Circular, Linear, and Interacting RNAs

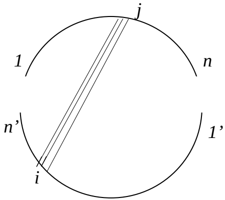
In the maximum matching case
$\Longrightarrow$ same algorithm for all three cases



CIRCULAR FOLDING



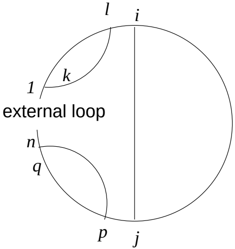
LINEAR FOLDING



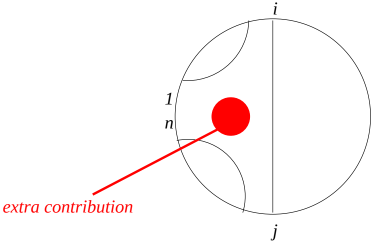
BINARY COFOLDING

Linear versus Circular Folding

Linear folding: energy contributions *inside* a pair (i, j) *only*.
Co-folding: additional contribution for loop spanning $[n, 1]$.

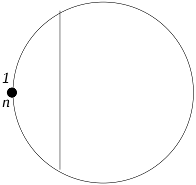


no energy contribution for external loop

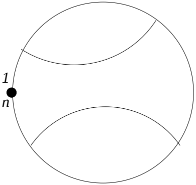


no external loop

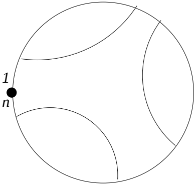
- Strategy 1 (e.g. Michael Zucker's `mfold`)
  For each pair (i, j): compute energy both inside and outside the pair
  $\Rightarrow$ doubles memory requirements
- Strategy 2 (`Vienna RNA Package`)
  First compute linear folding energies. Then compute energies for the loop spanning $[n, 1]$.



hairpin loop



interior loop or bulge

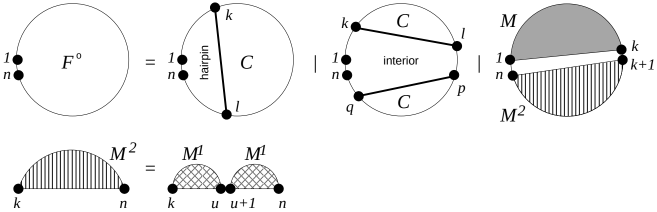


multi−branch loop

Implementing Circular Folding

Relative to linear folding, only the loop containing the cut has to be re-evaluated.
Three cases: cut in Hairpin, Interior-, or Multi-loop

$$F^{\circ} = \min\{F^{\circ}_{H}, F^{\circ}_{I}, F^{\circ}_{M}\}$$



- **Exterior Hairpin.**

$$F^{\circ}_{H} = \min_{p<q} \{ C_{pq} + \mathcal{H}(q, p) \}$$

- **Exterior Interior Loop.**

$$F^{\circ}_{I} = \min_{k<l<p<q} \{ C_{pq} + C_{kl} + \mathcal{I}(q, p, l, k) \}$$

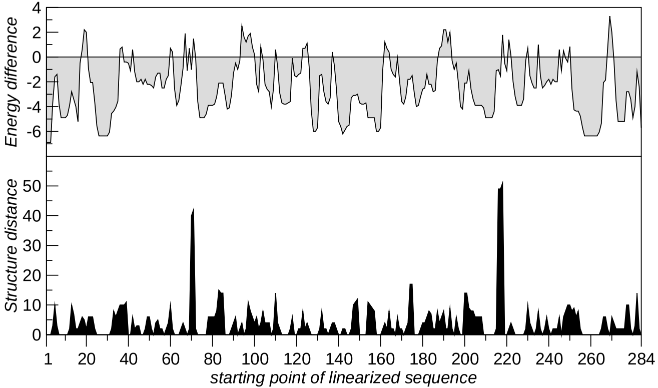
- **Exterior Multi-Loop.**
  Modified decomposition: one or more components $M_{1,k}$ + exactly two components $M^2_{k+1,n}$

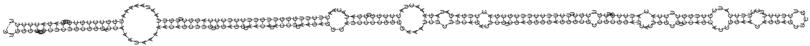
$$M^2_{kn} = \min_{k<u<n} \left(M^1_{ku} + M^1_{u+1,n} \right)$$
$$F^{\circ}_{M} = \min_{1<k<n} \left\{ M_{1,k} + M^2_{k+1,n} + a \right\}$$

- **Folding energy:** $F^{\circ} = \min\{F^{\circ}_{H}, F^{\circ}_{I}, F^{\circ}_{M}\}$

Applications of Circular Folding

It does make a difference





Citrus viroid IV

`RNAfold -circ` in the `Vienna RNA Package`

Local structures

Idea: Restrict Recursion to base pairs (i,j) with $j-i<L$.
Special interest in robust structures:
$Z_{ij}^{u,L}$ … partition function of sub-sequence $[i,j]$ when sequence window $[u,u+L]$ is folded
$p_{ij}^{u,L}$ … probability that i and j form a base pair when window $[u,u+L]$ is folded.

$$Z_{ij}^{u,L}=\begin{cases} Z_{ij} & \text{if} \quad [i,j]\subseteq[u,u+L] \\ 0 & \qquad\qquad \text{otherwise}\end{cases}$$

$$\begin{aligned} p_{ij}^{u,L} &= \frac{Z_{1,i-1}^{u,L}\widehat{Z}_{i,j}^{u,L}Z_{j+1,n}^{u,L}}{Z_{u,u+L}^{u,L}}+\sum_{k<i}\sum_{l>j}p_{kl}^{u,L}\Xi_{ij,kl}^{u,L} \\ &= \frac{Z_{u,i-1}\widehat{Z}_{i,j}Z_{j+1,u+L}}{Z_{u,u+L}}+\sum_{k<i}\sum_{l>j}p_{kl}^{u,L}\Xi_{ij,kl}\,. \end{aligned}$$

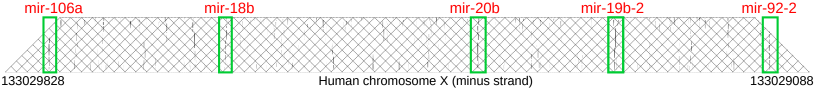
Robust local structures

Average probability of an (i, j) pair over all folding windows containing the sequence interval $[i, j]$

$$\pi^L_{ij} = \frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i} p^{u,L}_{ij} \,.$$

Direct Recursion:

$$\begin{aligned}\pi^L_{ij} &= \underbrace{\frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i}\frac{Z^{u,L}_{1,i-1}\widehat{Z}^{u,L}_{i,j}Z^{u,L}_{j+1,n}}{Z^{u,L}_{1,n}}}_{\pi^{*L}_{ij}} + \frac{1}{L-(j-i)+1}\sum_{u=j-L}^{i}\sum_{k<i}\sum_{l>j} p^{u,L}_{kl}\Xi_{ij,kl} \\ &= \pi^{*L}_{ij} + \sum_{k=j-L}^{i-1}\sum_{l=j+1}^{i+L}\sum_{u=l-L}^{k}\frac{p^{u,L}_{kl}\Xi_{ij,kl}}{L-(j-i)+1} = \pi^{*L}_{ij} + \sum_{k=j-L}^{i-1}\sum_{l=j+1}^{i+L}\frac{L-(k-l)+1}{L-(j-i)+1}\pi^L_{kl}\Xi_{ij,kl} \,.\end{aligned} \tag{1}$$

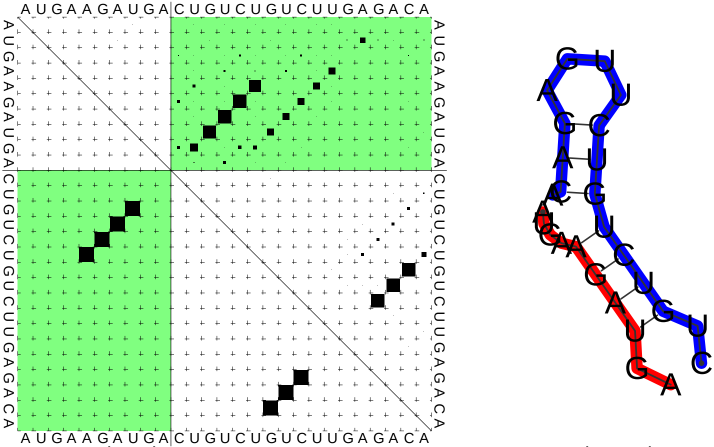


Local structures (L=100) in a 740 nt region of human X chromosome

Cofold: How to deal with Concentration?

- Algorithmically that same as linear folding
  special energy contribution for “loop with the cut”
- Additional energy contribution for forming duplex
- At least 5 molecular species need to be taken into account (Dmitrov & Zuker, 2005): A, B, A_2, B_2, AB.
- Their folding energies and partition functions are easily computed

Cofold


A U G A A G A U G A C U G U C U G U C U U G A G A C A
A U G A A G A U G A C U G U C U G U C U U G A G A C A

Dot plot (left) and mfe structure representation (right) of the cofolding structure of the two RNA molecules AUGAAGAUGA (red) and CUGUCUGUCUUGAGACA.

Cofold: Concentration dependencies

$$\mathcal{Q} = V^n \frac{a!b! \times (Z'^A)^{n_A}(Z'^{AA})^{n_{AA}}(Z'^{AB})^{n_{AB}}(Z'^{BB})^{n_{BB}}(Z'^B)^{n_B}}{n_A!n_B!\,2\,n_{AA}!\,2\,n_{BB}!n_{AB}!}$$

where $a = n_A + 2n_{AA} + n_{AB}$. The system minimizes the free energy $-kT \ln \mathcal{Q}$.

solving this optimization problem yields the equilibria:

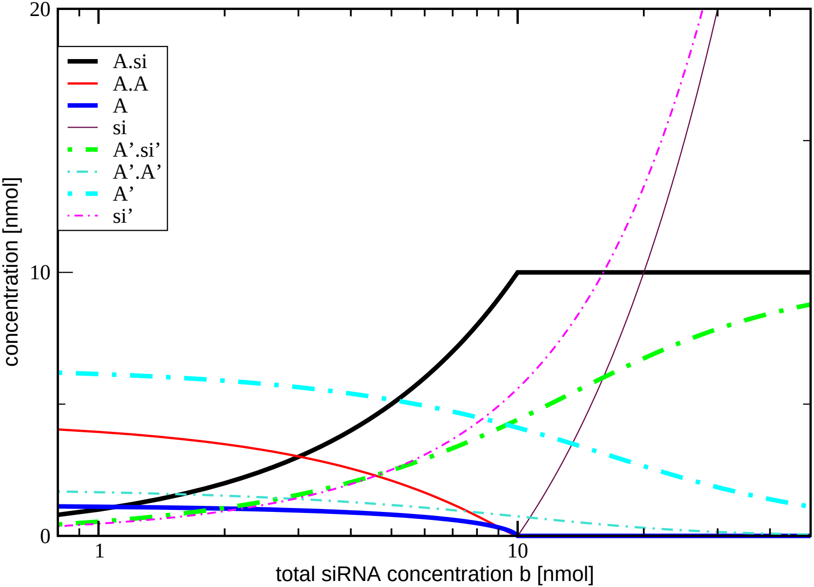
$[AA] = K_{AA}\,[A]^2\,, \qquad [BB] = K_{BB}\,[B]^2\,. \qquad [AB] = K_{AB}\,[A]\,[B].$

with $[A] = 6.023 \times 10^{23} n_A$, etc., and

$$K_{AA} = \frac{Z'^{AA}}{(Z^A)^2} = \frac{(Z^{AA} - (Z^A)^2)e^{-\Theta_I/RT}/2}{(Z_A)^2} = \frac{1}{2}\,e^{-\Theta_I/RT}\left(\frac{Z^{AA}}{(Z^A)^2} - 1\right)$$

$$K_{BB} = \frac{1}{2}\,e^{-\Theta_I/RT}\left(\frac{Z^{BB}}{(Z^B)^2} - 1\right)$$

$$K_{AB} = e^{-\Theta_I/RT}\left(\frac{Z^{AB}}{Z^A Z^B} - 1\right)$$



Example for the concentration dependency for two mRNA-siRNA binding experiments.

RNAup: Small RNAs Binding to Large Ones

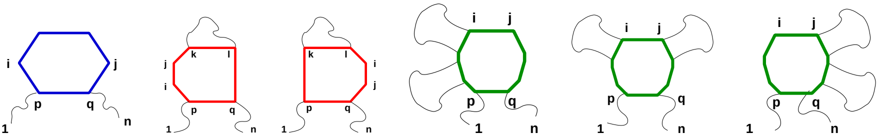
- RNA folding excludes pseudoknots, i.e., non outerplanar graphs
- `cofold` thus does not allow small RNA binding into loop regions of large ones
- ... but this happens in reality

Remedy: Compute energy/partition function

$$P_u[i,j] = \underbrace{\frac{Z[1,i-1] \times 1 \times Z[j+1,N]}{Z}}_{exterior} + \sum_{\substack{p,q \\ p<i\leq j<q}} \underbrace{P_{pq} \times \frac{Z_{pq}[i,j]}{Z^b[p,q]}}_{enclosed}$$

that subsequence $[i,j]$ is unpaired and the energy of binding a short molecule in this location

RNAup



(a) (b') (b") (c) (d) (e)

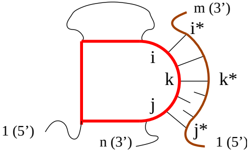
$$
\begin{aligned}
Z_{pq}[i,j] = & \underbrace{\exp(-\beta H(p,q))}_{(a)} \\
& + \sum_{\substack{p\ <\ i\ \le\ j\ <\ k \text{ or} \\ l\ <\ i\ \le\ j\ <\ q}} \underbrace{Z^b[k,l]e^{-\beta I(p,q;k,l)}}_{(b)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[p+1,i-1]e^{-\beta c(q-i)}}_{(c)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m}[p+1,i-1]Z^{m}[j+1,q-1]e^{-\beta c(j-i+1)}}_{(d)} \\
& + \sum_{p<i\le j<q} \underbrace{Z^{m2}[j+1,q-1]e^{-\beta c(j-p)}}_{(e)}
\end{aligned}
$$

RNAup: Interaction part

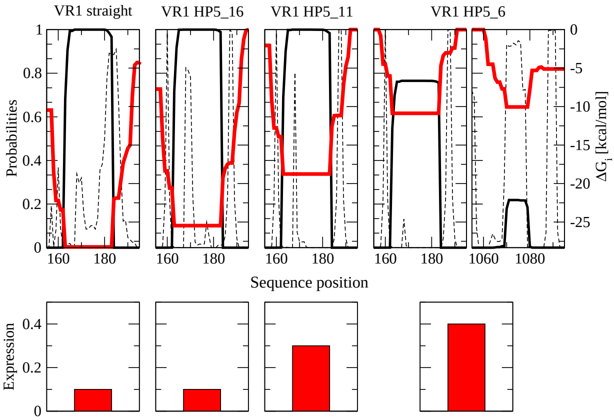
$$Z^I[i,j,i^*,j^*] = \sum_{\substack{i<k<j\\ i^*>k^*>j^*}} Z^I[i,k,i^*,k^*]e^{-\beta I(k,k^*;j,j^*)}$$

$$Z^*[i,j] = P_u[i,j] \sum_{i^*>j^*} Z^I[i,j,i^*,j^*];$$

$$P^*[i,j] = Z^*[i,j] / \sum_{k<l} Z^*[k,l]$$



RNAup: Application



Binding of siRNAs to VR mRNA.
$P_u[i, i]$ (dashed line), P_i^* (thick black line), ΔG_i (thick red line).
Below: activity of siRNA

Alternative Approach

Consider RNA Folding as a Machine Learning Problem
Context Free Grammar + probabilities for production rules
$\Rightarrow$ Stochastic Context Free Grammars
see work by Sean Eddy, Jotun Hein, and collaborators

Folding Kinetics

RNA molecules may have kinetic traps which prevent them from reaching equilibrium within the lifetime of the molecule. Long molecules are often trapped in such meta-stable states during transcription.
Possible solutions are

- Stochastic folding simulations can predict folding pathways and final structures. Computationally expensive, few programs available.
- Predicting structures for growing fragments of the sequence can show whether large scale re-folding will occur during transcription. Cheap but inaccurate.
- Analysis of the energy landscape based on complete suboptimal folding can identify possible traps (local minima).

Kinetic Folding Algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:

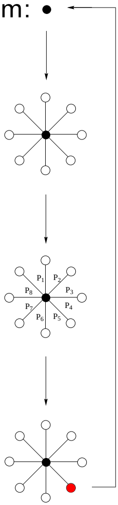
Generate all neighbors using the move-set

Assign rates to each move, e.g.

$$P_i = \min\left\{1, \exp\left(-\frac{\Delta E}{kT}\right)\right\}$$

Select a move with probability proportional to its rate

Advance clock $1/\sum_i P_i$.



Characterization of Landscapes

A landscape consists of a configuration space V, a move set within that configuration space and an energy function $f : V \to \mathbb{R}$.
Simplest move set for secondary structure: opening and closing of base pairs.
Speed of optimization depends on the *roughness* of the Landscape.
Measures of roughness suggested in the literature:

- Number of local optima
- Correlation lengths (e.g. along a random walk)
- Lengths of adaptive walks
- Folding temperature vs. glass temperature T_f/T_g
- Energy barriers between the local optima. Especially, the maximum barrier height ("depth" in SA literature)

Energy barriers

$$E[s,w] = \min\left\{ \max\left[f(z) \middle| z \in \mathbf{p} \right] \;\middle|\; \mathbf{p} : \text{path from } s \text{ to } w \right\},$$
$$B(s) = \min\left\{ E[s,w] - f(s) \middle| w : f(w) < f(s) \right\}$$

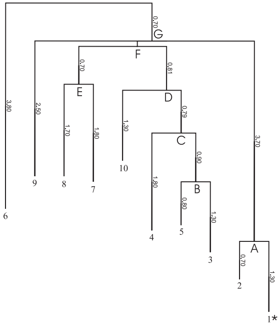
Depth and Difficulty
(borrowed from simulated annealing theory)

$$\mathrm{D} = \max\left\{ B(s) \middle| s \text{ is not a global minimum } \right\}$$
$$\psi = \max\left\{ \frac{B(s)}{f(s) - f(\min)} \middle| s \text{ is not a global minimum} \right\}$$

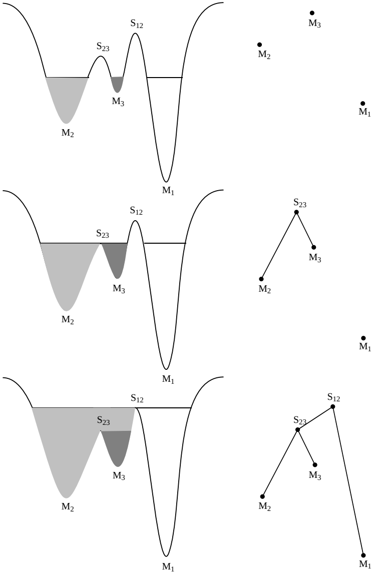
Energy Barriers and Barrier Trees

Some topological definitions:
A structure is a

- *local minimum* if its energy is lower than the energy of **all** neighbors
- *local maximum* if its energy is higher than the energy of **all** neighbors
- *saddle point* if there are at least two local minima that can be reached by a downhill walk starting at this point



Calculating barrier trees



The flooding algorithm:
Read conformations in energy sorted order.
For each confirmation x we have three cases:

(a) x is a local minimum if it has no neighbors we've already seen

(b) x belongs to basin $B(s)$, if all known neighbors belong to $B(s)$

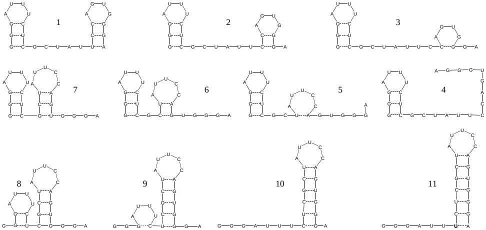
(c) if x has neighbors in several basins $B(s_1) \dots B(s_k)$ then it's a saddle point that *merges* these basins.
Basins $B(s_1), \dots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

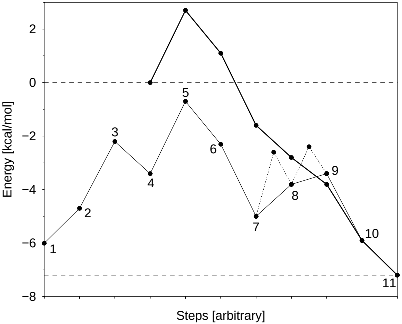
Information from the Barrier Trees

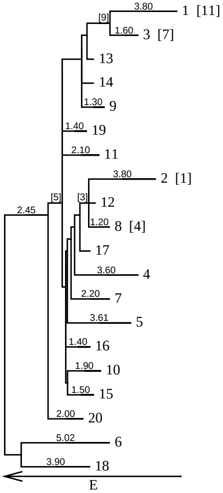
- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions and free energies of (gradient) basins
- Depth and Difficulty of the landscape

N.B.: A *gradient basin* is the set of all initial points from which a gradient walk (steepest descent) ends in the same local minimum.

Energy Landscape of a Toy Sequence







Folding Kinetics

Transition rates from x to y:

$$\begin{aligned} r_{yx} &= r_0 e^{-\frac{E^{\neq}_{yx} - E(x)}{RT}} \quad \text{for } x \neq y \\ r_{xx} &= -\sum_{y \neq x} r_{yx} \end{aligned}$$

Kinetics as a Markov process:

$$\frac{\mathrm{d}p_x}{\mathrm{d}t} = \sum_{y \in X} r_{xy} \, p_y(t) \, .$$

Transition states:

$$E^{\neq}_{yx} = \max\{E(x), E(y)\}$$

or more complex models (Tacker et al 1994, Schmitz et al 1996)

Reduced Description of the Folding Dynamics

Macrostates = Classes of a partition of the state space.
Partition function for a macro state:

$$Z_\alpha = \sum_{x\in\alpha} \exp(-E(x)/RT)$$

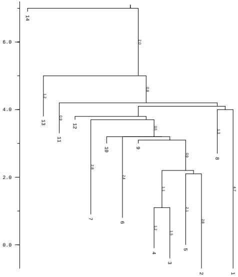
Free energy of a macro state:

$$G(\alpha) = -RT \ln Z_\alpha$$

$$\begin{aligned} r_{\beta\alpha} &= \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx} \mathrm{Prob}[x|\alpha] \quad \text{for } \alpha \neq \beta \\ &= \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx} e^{-E(x)/RT} \end{aligned}$$

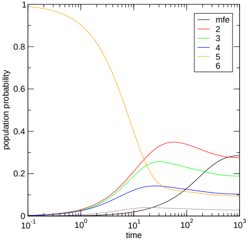
$r_{\beta\alpha}$ “on flight” while executing the `barriers` program.
Transition state free energy:

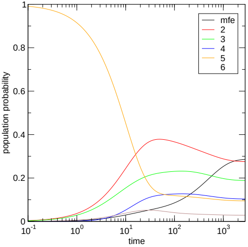
$$G^{\neq}_{\beta\alpha} = -RT \ln \sum_{y\in\beta}\sum_{x\in\alpha} e^{-\frac{E^{\neq}_{xy}}{RT}}$$

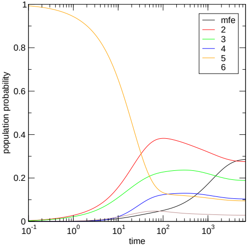


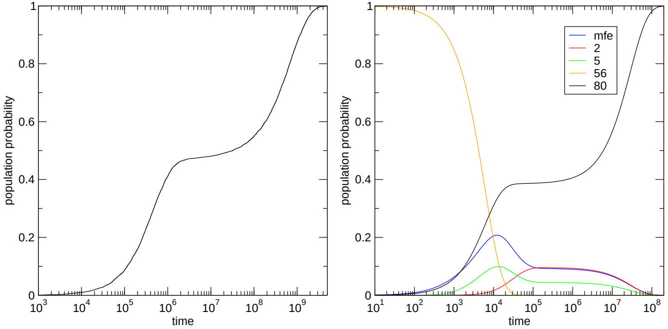
lilly

A simple model sequence









Refolding of a tRNA molecule.

Summary I:

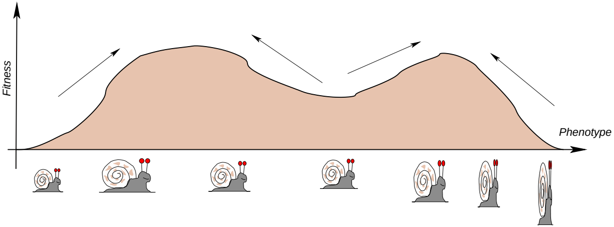
- RNA structures can be computed efficiently by means of dynamic programming
- Computations are based on a set of carefully measures energy parameters and an additive energy model
- Algorithms exist for ground state energy and structure, full partition functions, density of states, interacting structures, . . .
- The folding kinetics of a given RNA Sequence can also be investigated as the level of secondary structures
- VIENNA RNA PACKAGE

PART II: How Do RNAs Evolve

Basic Assumption

Selection Acts on Secondary Structures, Mutations acts on the underlying sequences
⇒ We need to understand the sequence-to-structure map of RNAs
(hang on, we'll discuss the empirical evidence for that a bit later)

Sewall Wright's Fitness Landscapes



How do realistic fitness landscapes look like?

Biological Landscapes

The RNA case is a special case of a very general paradigm:

$$\text{genotype} \mapsto \text{phenotype} \mapsto \text{fitness}$$

What is the relationship between Genotyp and Phenotype?

Central topic in any theory of evolution

because:

* Selection acts on the Phenotype

* Mutation/Recombination acts on the Genotype

Biopolymers as the simplest model:

The molecule is **both** genotype (sequence) and phenotype (structure).

The map from genotype to genotype is determined by physical chemistry:

$\Longleftrightarrow$ *folding problem*

Computational Analysis of the RNA Map

There are many more sequences than structures.

(.)-string: 3-letters (with constraints)

$\Longrightarrow$ less than 3^n structures

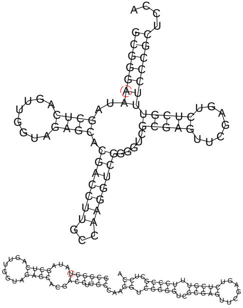
but 4^n sequences.

$\Longrightarrow$ **Redundancy**

How are sequences folding into the same structure distributed in sequence space?

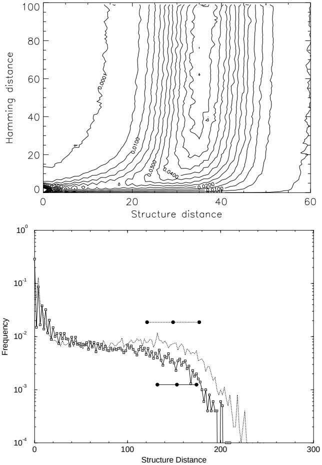
Neutral Set $S(\psi) = \{x \in \mathcal{Q}_\alpha^n | f(x) = \psi\}$

Sensitivity and Neutrality



Effect of a single

point mutation



Distribution of structure distances

The Random Graph Model

Approach:
Model $S(\psi)$ as a *random induced subgraph* Γ with a given value

$$\lambda = \frac{\langle \#\text{neutral neighbors} \rangle}{(\alpha - 1)n}$$

Threshold value:

$$\lambda^* = 1 - \left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}}$$

Theorem. [Reidys, Stadler, Schuster]
If $\lambda > \lambda^*$ then Γ is *a.s.* dense and connected,
if $\lambda < \lambda^*$ then Γ is *a.s.* neither dense nor connected

A complication: Base Pairing Rules

Unpaired bases:

Alphabet $\mathcal{A} = \{\mathsf{A}, \mathsf{U}, \mathsf{G}, \mathsf{C}\}$

Paired bases: 5' and 3' side correlated:

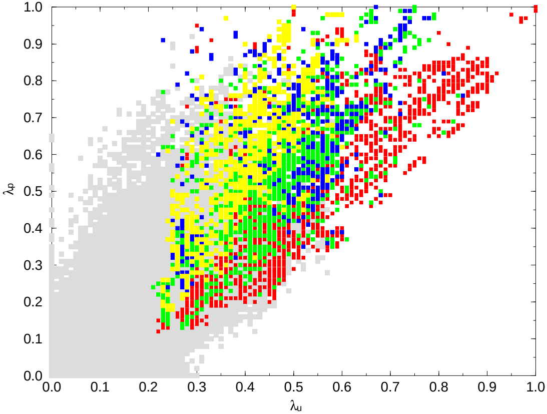
Alphabet: $\mathcal{B} = \{\mathsf{AU}, \mathsf{UA}, \mathsf{GC}, \mathsf{CG}, \mathsf{GU}, \mathsf{UG}, \}$.

Thus consider only the set of *compatible sequences* $C(\psi)$:

$S(\psi) \subseteq C(\psi) \equiv \mathcal{Q}_4^{n_u} \times \mathcal{Q}_6^{n_p}$.

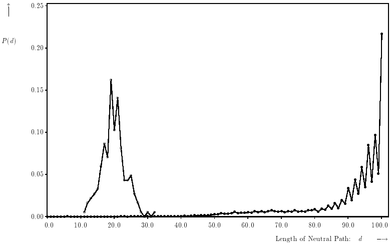
$\Longrightarrow$ Two neutrality parameters λ_u and λ_p

Connected Components of Neutral Networks

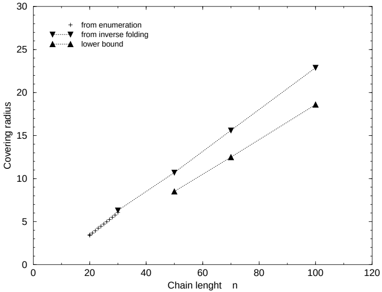


gray	many small components	red	1 connected component
green	2 equal sized components	yellow	3 components size 2:1
blue	4 equal sized components		

Explanation: for this deviation from the random graph model in terms of the energy model. Some structures can be made only with a significant bias in the G/C ratio.



Distance to Target structure
length neutral paths



Covering radius

Closest Approach

Intersection Theorem. For any two secondary structures ϕ, and ψ holds

$$C(\phi) \cap C(\psi) \neq \emptyset$$

What is the distance of neutral networks

$$\delta(\phi, \psi) = \min\{d(x, y) | f(x) = \phi \text{ and } f(y) = \psi\}$$

Random graph Theory: If $\lambda > \lambda^*$ then $\delta(\phi, \psi) \approx 2$.
Computer simulations: upper bounds on $\delta(\phi, \psi)$:

n	GC	AU	AUGC
50	5.6	2.6	2.1
70	9.3	4.6	3.4
100	13.0	7.8	5.6

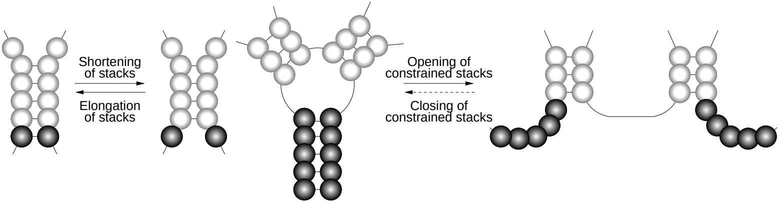
Accessibility

Fontana & Schuster 1998

Idea: The "interface" between two structures is large is they are "similar".

More precisely: Structure ψ is *accessible* for ϕ if $x \in S(\phi)$ is like to have neighbor (mutant) $x' \in S(\psi)$.

Structural characterization of "easy" (*continuous*) transitions:



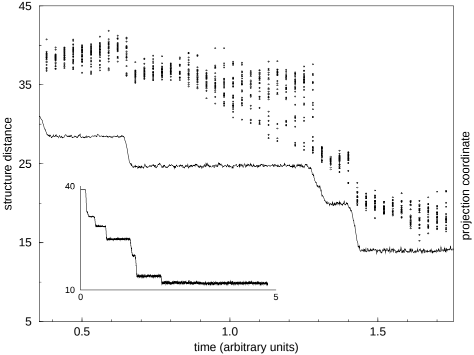
SUMMARY: Sequence-Structure Map of RNA

1. *Redundancy:* Many more sequences than structures
2. *Sensitivity:* Small changes in the sequences may lead to large changes in the structure
3. *Neutrality:* A substantial fraction of mutations does not alter the structure.
4. Isotropy: $S(\psi)$ is "randomly" embedded in $C(\psi)$.

Implications:

1. *Neutral Networks:* $S(\psi)$ forms a connected "percolating" network in sequence space for all "common" structures.
2. *Shape Space Covering:* Almost all structures can be found in a relatively small neighborhood of almost every sequence.
3. *Mutual Accessibility:* The neutral networks of any two structures almost touch each other somewhere in sequence space.

Simulated Trajectories



Punctuated equilibra = diffusion of neutral networks +
constant rate of innovation +
exponential selection of rare mutants

Proc.Natl.Acad.Sci. **93**: 397-401 (1996)

Diffusion Constant

…can be deduced from Moran model:

$$D = \overline{\lambda}\frac{6Anp}{3+4Np}(1+1/N) \sim \begin{cases} (3/2)A(n/N) & p \gg 0 \qquad \mathrm{or} N \gg 1 \\ 2Anp & p \ll 1 \end{cases}$$

A …replication rate
n …sequence length
N …population size
p …mutation rate
$\overline{\lambda}$ …neutrality of network

Dynamics of Interacting Replicators

$$\mathbb{I}_k + \mathbb{I}_j \longrightarrow \mathbb{I}_l + \mathbb{I}_k + \mathbb{I}_j$$

With mutation:

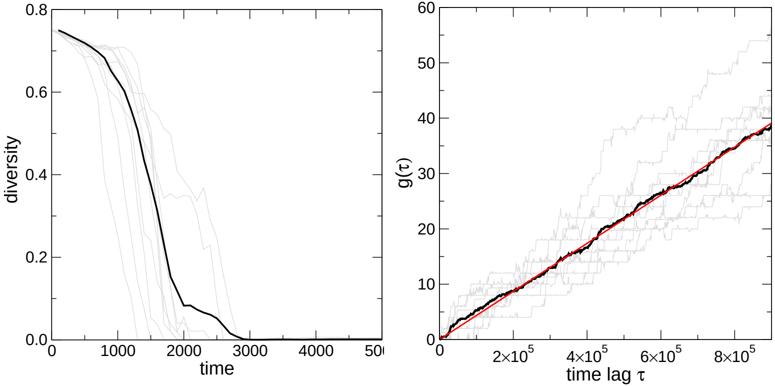
$$\dot{x}_k = x_k \left(\sum_j A_{kj} x_j - \sum_{i,j} A_{ij} x_i x_j \right) + \sum_{l,j} \left(Q_{kl} A_{lj} x_j x_l - Q_{lk} A_{kj} x_k x_j \right)$$

where

$$Q_{kl} = (1-p)^{n-d(k,l)} \left(\frac{p}{\alpha - 1} \right)^{d(k,l)}$$

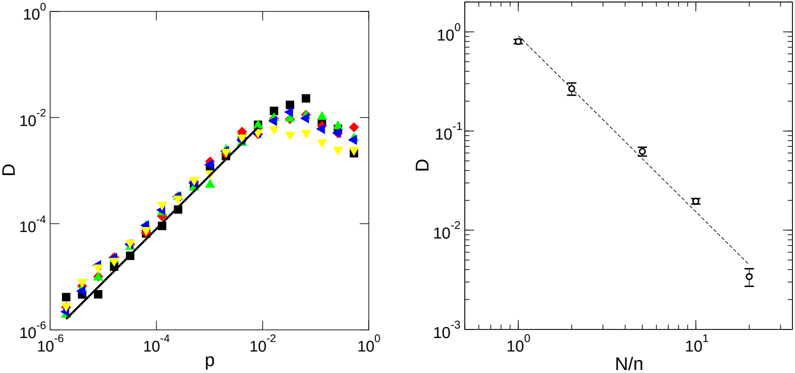
How does this behave in **sequence space?**

Simplest case: Simplest case: $A_{kl} = A_0(1 - d(\mathbb{I}_k, \mathbb{I}_l)/n)$:



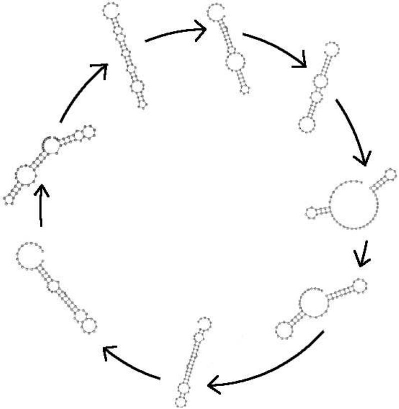
$$g(\tau) = \frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} \|\mathbf{p}(t+\tau) - \mathbf{p}(t)\|^2$$

B.M.R. Stadler, *Adv. Complex Syst.* (2003)



Left: Diffusion coefficient D as a function of the mutation rate for $N = 10, 20, 30, 40, 80$ and $n = 10, 20, 30, 40, 80$ such that $N/n = 1$ after equilibration for 10^5 timesteps. Right: Dependence of the ratio D/p on N/n.

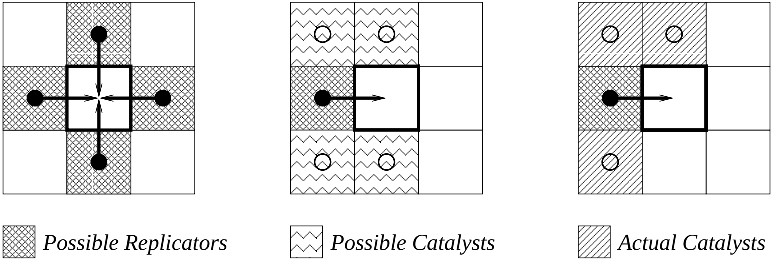
An RNA-Based Model in the Plane



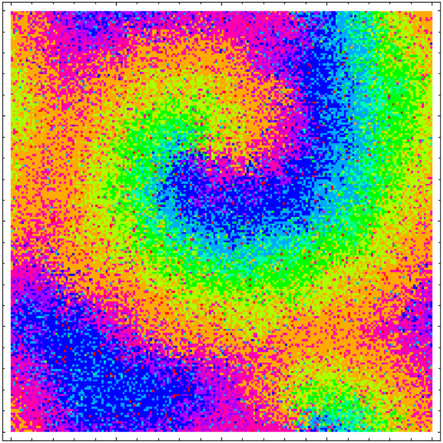
Target hypercycle with 8 members.

Model:
Hypercyclically coupled species, each sequence has a *function* that depends on its structure.

Spatial Extension: CA Model



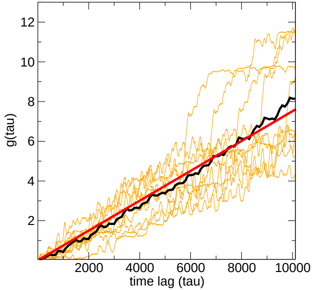
Rules of replication. For each of the neighbors (●) of the empty cell (marked by a bold outline) the replication rate ρ_z is computed taking into account their neighbors in the direction of the replication (○) as potential catalysts. The neighbor with the largest values of ρ_z invades the empty position. In this example, for the chosen replicator, only three of its neighbours are catalysts according to the hypercycle topology.

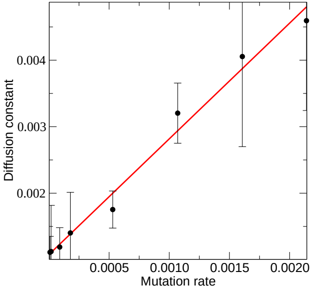


Spirals formed after 3000 generations in an evolution experiment started with 300 random sequences in the absence of parasites.

see also Borlijst & Hogeweg (1993)

Diffusion in Sequence Space





Summary

- Neutrality of the Sequence-Structure Map implies diffusion/drift-like motion in sequence independent of details of the selection/mutation mechanisms and whether spatial extension is taken into account or not.
- $\Longrightarrow$ The basic assumption of molecular phylogenetics, namely a dominating influence of drift in **sequence** evolution, holds true even when **phenotypic** evolution is dominated by interactions
(co-evolution).
- **TODO** Development of a rigorous mathematical theory describing the motion in sequence space of a population with strong interactions.

Evolutionary histories of some structured RNAs

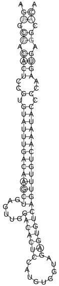
Ribosomal RNAs (rRNAs) are the most frequently used sequence data for reconstructing phylogenies from molecular data
How does that work:
In a nutshell:
(1) compute evolutionary distances from the sequence data
(2) “fit” an additive tree to the distances
(In reality, there are other methods such as maximum parsimony and maximum likelihood approaches, but the basic idea is the same)
Observation: all tRNAs have more or less the same clover-leaf structure.

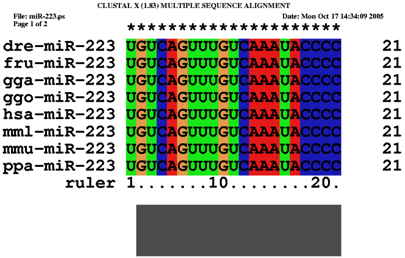
MicroRNAs

- processed from precursor hairpins
- short (∼ 22nt) RNA molecules
- highly conserved

Function

- bind to 3'UTRs of mRNA targets
  - supress expression of this mRNA
  - mark mRNA molecule for degradation
  - in plants involved in DNA methylation





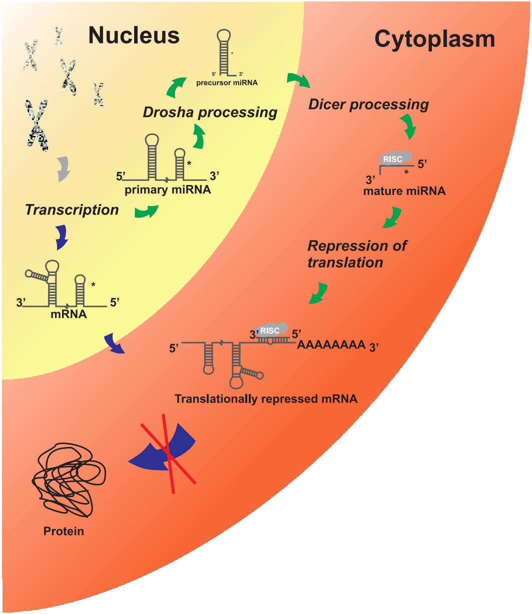
CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

File: miR-223.ps
Page 1 of 2

Date: Mon Oct 17 14:34:09 2005

```
             *********************
dre-miR-223  UGUCAGUUUGUCAAAUACCCC      21
fru-miR-223  UGUCAGUUUGUCAAAUACCCC      21
gga-miR-223  UGUCAGUUUGUCAAAUACCCC      21
ggo-miR-223  UGUCAGUUUGUCAAAUACCCC      21
hsa-miR-223  UGUCAGUUUGUCAAAUACCCC      21
mml-miR-223  UGUCAGUUUGUCAAAUACCCC      21
mmu-miR-223  UGUCAGUUUGUCAAAUACCCC      21
ppa-miR-223  UGUCAGUUUGUCAAAUACCCC      21
       ruler 1.......10.........20.
```

MicroRNAs — processing and function

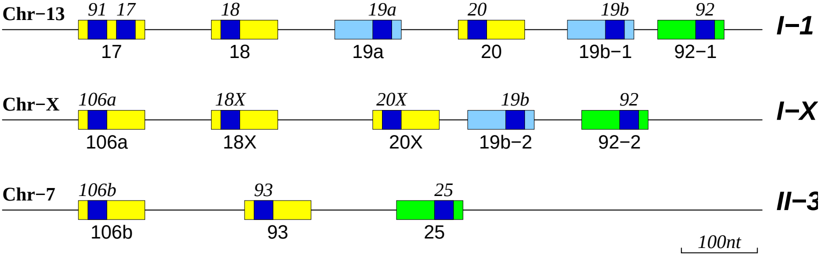


MicroRNAs ...

- transcribed in longer transcripts (*primary-miRNA*)
- in some cases: *polycistronic* "clusters"
- Drosha processing → *precursor miRNA*
- export to cytoplasm *Exportin-5 pathway*
- Dicer processing → *mature miRNA*

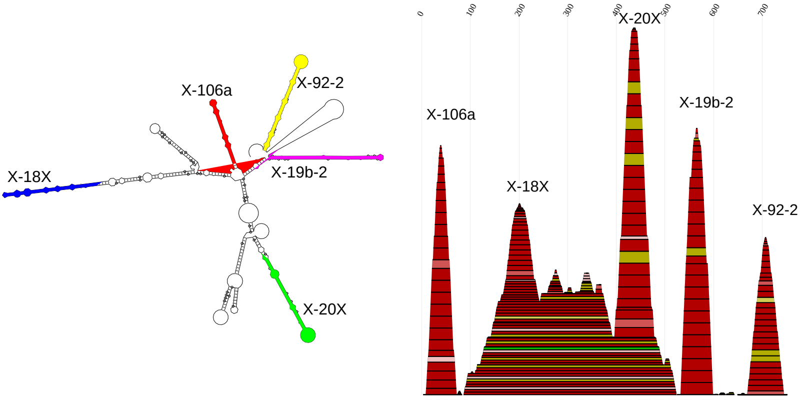
Evolution of microRNA Families: **mir-17** clusters

Many miRNAs are transcribed from polycistronic transcripts
Most spectacular example: Human **mir-17** clusters



J. Mol. Biol. **339**: 327-335 (2004)

Case Study: **mir-17** clusters

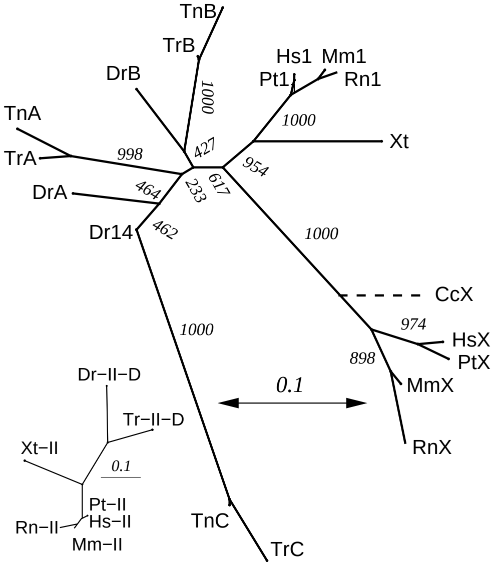


(a) (b)

Structure of the pri-pre-mir-17 at the human X chromosome.

Construction of Gene Trees

from concatenated sequences in the cluster



Distant Homologies with unreliable Alignments

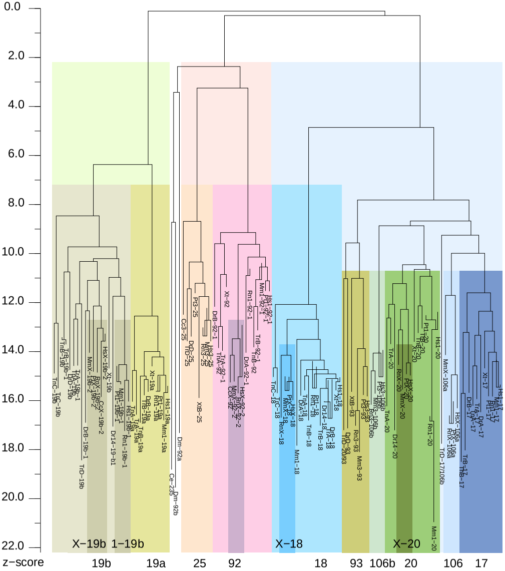
How to quantify sequence similarity when we cannot get a good alignment?

- measure pairwise sequence similarity $s(x, y)$
- compare to the distribution of similarity values of alignments of shuffled sequences
- define a z-score

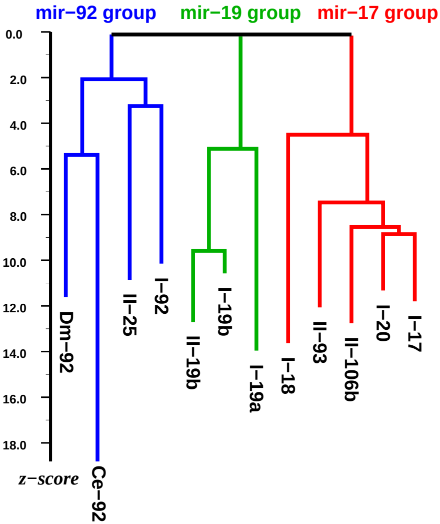
$$z(x, y) = \frac{s(m, y) - \langle s(\pi(x), \pi'(y)) \rangle_{\pi,\pi'}}{\sqrt{\mathrm{var}_{\pi,\pi'}(s(\pi(x), \pi'(y)))}}$$

- use $z(x, y)$ as similarity measure in WPGMA clustering

Gene Tree of **mir-17** cluster members



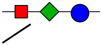
Collapsed tree of microRNA subgroups

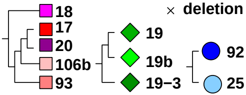


- obtained by collapsing vertebrate, insect, and nematode species trees to single vertices
- next step: combine gene trees and synteny information to a duplication history

Scenario for the evolution of the mir17 family

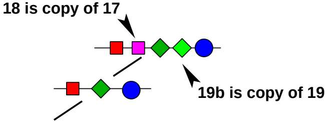
ancestral mir17 cluster probably contained
mir-17, mir-19 and mir-92

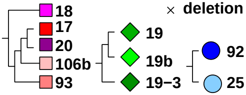




Scenario for the evolution of the mir17 family

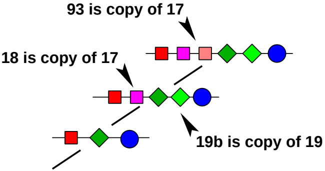
first detectable duplication event:
branch mir-17 and mir-18

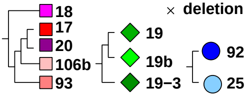




Scenario for the evolution of the mir17 family

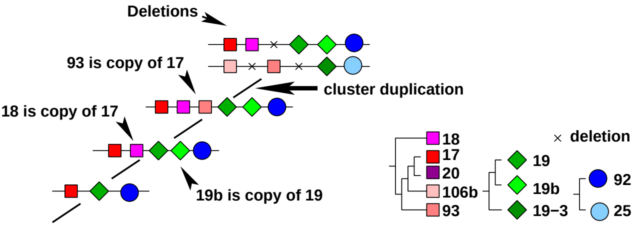
series of duplications:
branch mir-19 and mir19b, mir-17 and mir-93





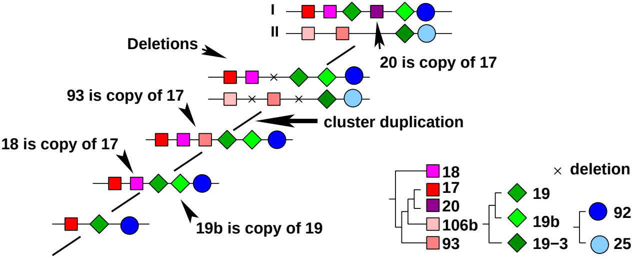
Scenario for the evolution of the mir17 family

genome wide duplication:
duplication of whole cluster and loss of individual miRNAs



Scenario for the evolution of the mir17 family

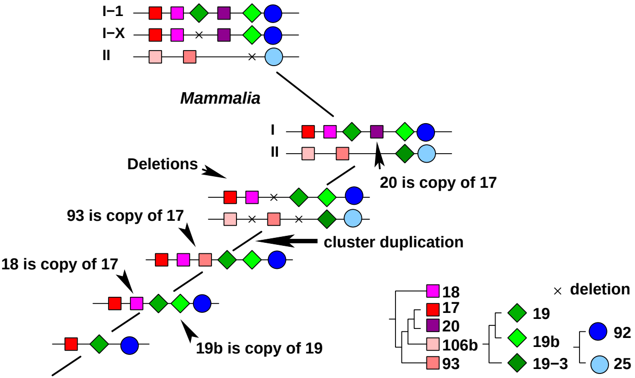
independent miRNA duplications in type I cluster



Scenario for the evolution of the mir17 family

split of teleosts and mammalia

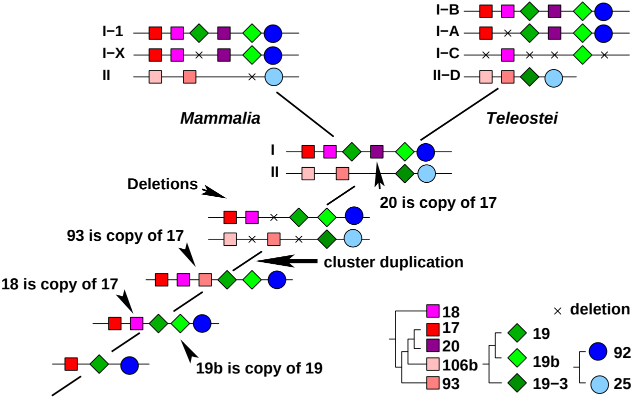
teleost specific genome duplication



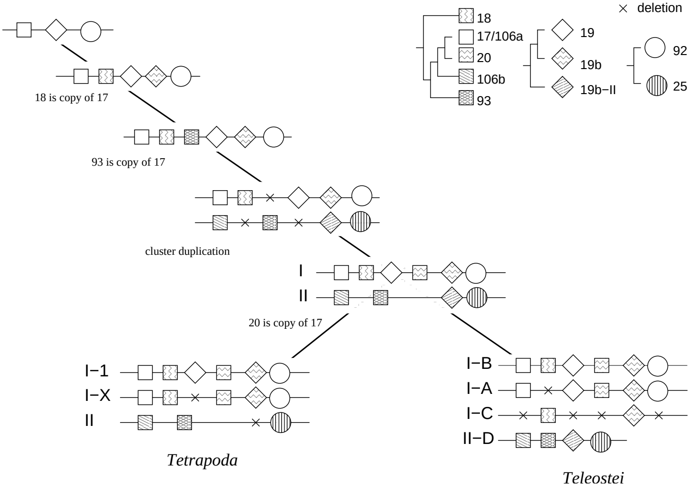
Scenario for the evolution of the mir17 family

split of teleosts and mammalia

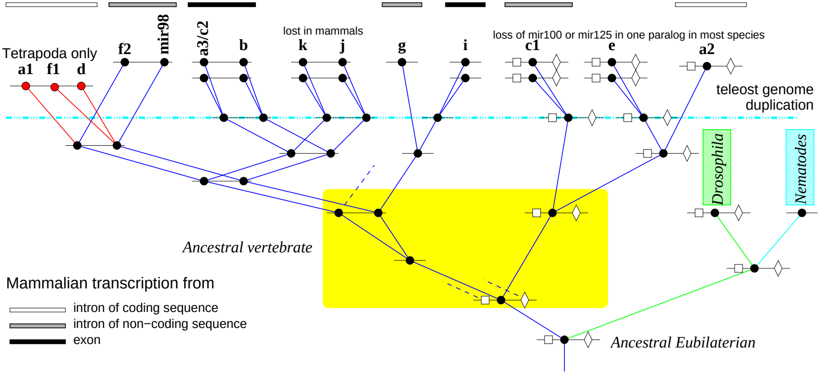
teleost specific genome duplication



History of the **mir-17** cluster: updated data



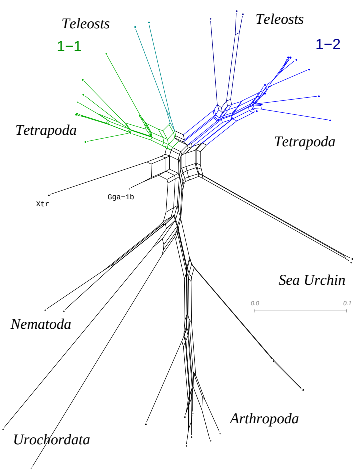
Further Examples: *let-7 family*



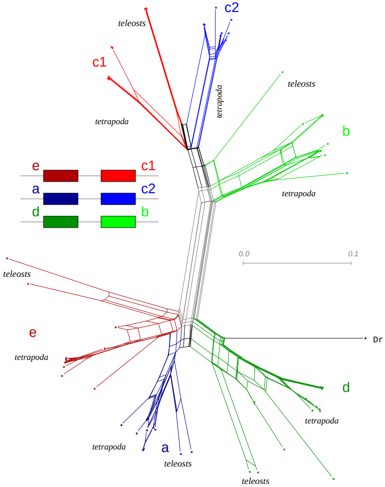
Mammalian transcription from

- ▭ intron of coding sequence
- ▨ intron of non−coding sequence
- ▬ exon

Further Examples: *mir-1* and *mir-30*

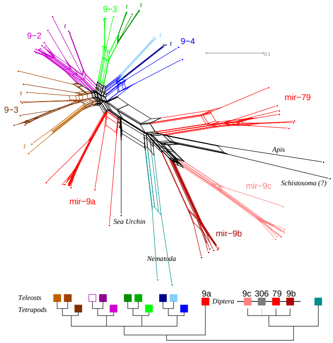


mir-1

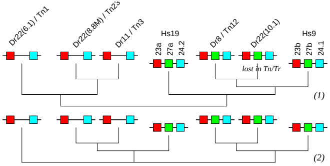


mir-30

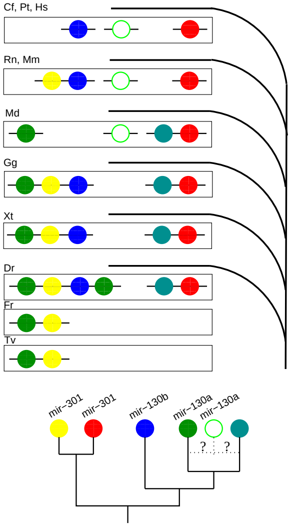
Further Examples: *mir-9*, *mir-23*, *mir130/301*



mir-9

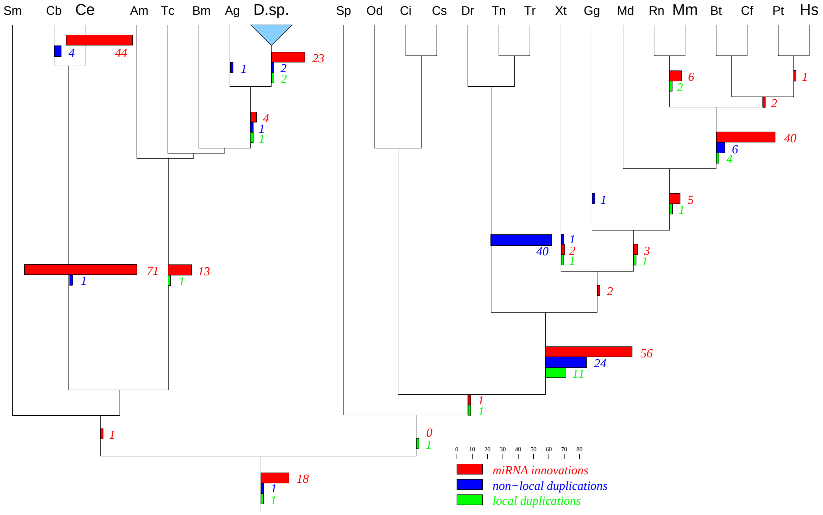


mir-23 cluster



mir-130 cluster

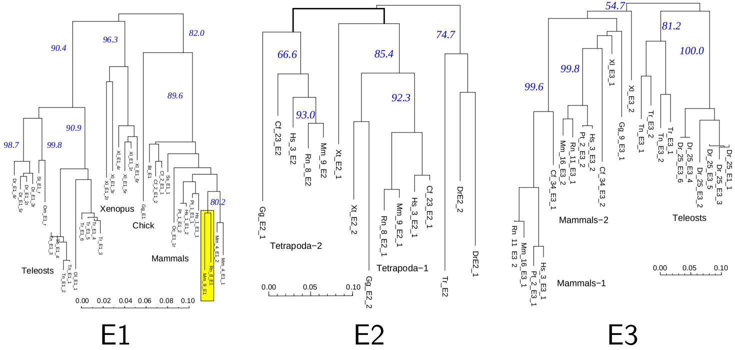
Expansion of the Metazoan MicroRNA Repertoire



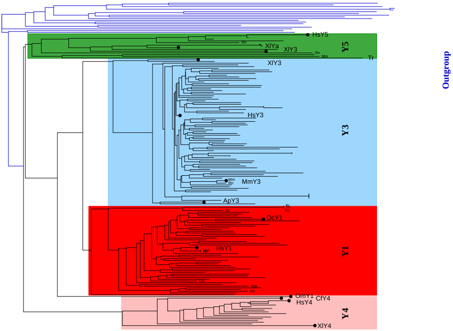
Similar Situation: snoRNA

- snoRNAs direct chemical modification of other RNAs (mostly rRNA, snRNA, and (some?) messenger RNAs
- two classes: box-H/ACA and box-C/D
- known in eukaryotes and archea, not in eubacteria

H/ACA box snoRNAs in Vertebrates



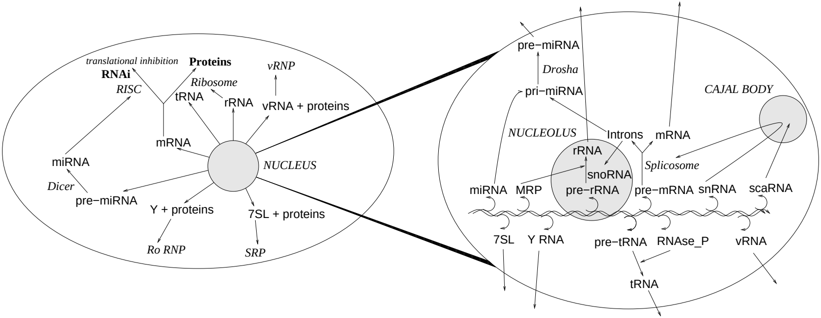
Vertebrate Y RNAs



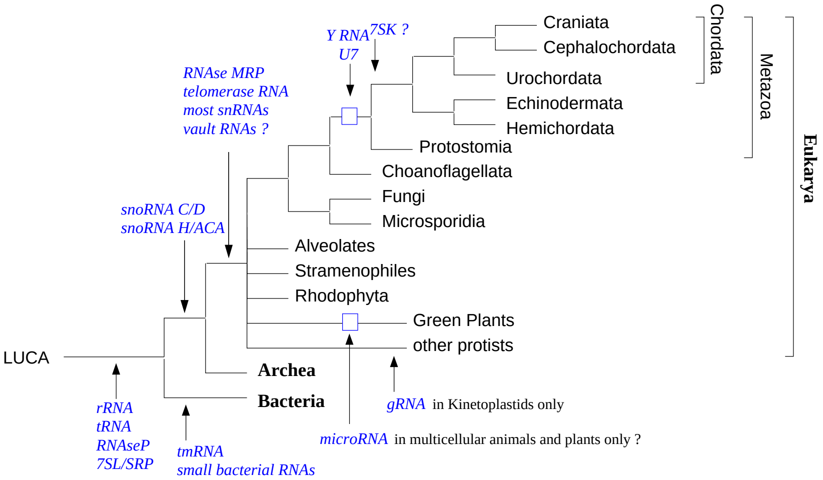
Summary

- The genotype-phenotype map of RNA is charcterized by an interplay of “ruggedness” and neutrality
- Selection plus drift results in diffusion on neutral networks
- Many non-coding RNAs have highly constrained (i.e., evolutionarily very well conserved) structures but fairly rapidly evolving sequences
- Drift of sequences is independent of the details of the selection mechanism
- Ongoing research: elucidate the evolutionary histories of structured ncRNAs

PART III: The Modern RNA World



Multiple Origins of ncRNAs



Surveys for noncoding RNAs

- $>$ 5% of the human genome is under stabilizing selection (from man/mouse comparison), less than 1/3 of this codes for protein
- Virtually the entire genome is transcribed as primary nuclear transcripts in at least one direction
  (ENCODE Genes&Transcripts group, unpublished data)
- $\sim$ 80% of the ENCODE regions are transcribed in as parts of protein coding transcripts including introns and UTRs
- Only a tiny part of the primary transcripts is protein coding
- Large fraction of apparently non-protein-coding cDNAs
- The functions of most of these transcripts are unclear.

"There is need for reliable experimental and computational methods for comprehensive identification of non-coding RNAs."

–International Human Genome Sequencing Consortium, Nature 431, p.943, October 2004

The ENCODE Project

ENCyclopedia Of DNA Elements

- Public research consortium launched by NHGRI in 2003
- Purpose: “testing and comparing existing methods to rigorously analyze a defined portion of the human genome sequence”.
- Focus: specified 30 megabases (1% of genome) in more than 20 species
- Informally organized in subgroups: Sequencing Technology, Comparative Genomics, Genes and Transcripts, Genetic Variation, ...
- Results from 1st phase currently under review
- Phase 2: scale-up to complete genome

Highlights from ENCODE Genes and Transcripts Analysis Group

(Data presented by Tom Gingeras in Bethesda, Jan 12 2006)

- Only a fraction of processed RNA transcripts correspond to `GeneCode` annotated transcripts:
  70% correlated with annotated (m)RNAs
  52% correlate with annotated protein coding sequences
- Substantial fraction of transcription is specific of cellular conditions
  only 2.6% of transfrags are common to all 11 cell-lines.
- The same genomic sequence may be processed into multiple RNA sequences with different fates
- Virtually the entire genome is transcribed as primary nuclear transcript in at least one direction.

Transcriptional output is MUCH more extensive AND much more complex than previously thought.

Recall: Sequence-Structure Map of RNA

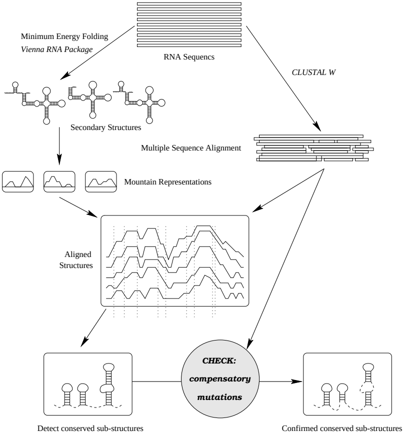
1. *Redundancy:* Many more sequences than structures
2. *Sensitivity:* Small changes in the sequences may lead to large changes in the structure
3. *Neutrality:* A substantial fraction of mutations does not alter the structure.
4. Isotropy: $S(\psi)$ is "randomly" embedded in $C(\psi)$.

Implications:

1. *Neutral Networks:* $S(\psi)$ forms a connected "percolating" network in sequence space for all "common" structures.
2. *Shape Space Covering:* Almost all structures can be found in a relatively small neighborhood of almost every sequence.
3. *Mutual Accessibility:* The neutral networks of any two structures almost touch each other somewhere in sequence space.

Proc.Roy.Soc.B **255** 279-284 (1994), Proc. Natl. Acad. Sci. USA **93**, 397-401 (1996),

Bull. Math. Biol. **59**, 339-397 (1997), RNA **7**: 254-265 (2000).



Minimum Energy Folding

Vienna RNA Package

RNA Sequencs

CLUSTAL W

Secondary Structures

Multiple Sequence Alignment

Mountain Representations

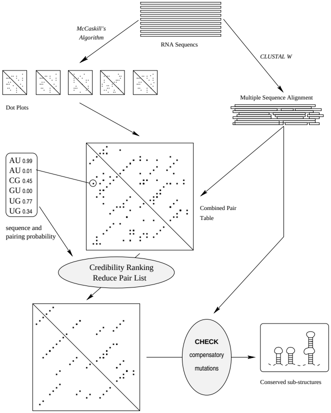
Aligned Structures

CHECK: compensatory mutations

Detect conserved sub-structures

Confirmed conserved sub-structures

Minimum Energy



McCaskill's Algorithm

RNA Sequencs

CLUSTAL W

Dot Plots

Multiple Sequence Alignment

AU 0.99
AU 0.01
CG 0.45
GU 0.00
UG 0.77
UG 0.34

sequence and pairing probability

Combined Pair Table

Credibility Ranking
Reduce Pair List

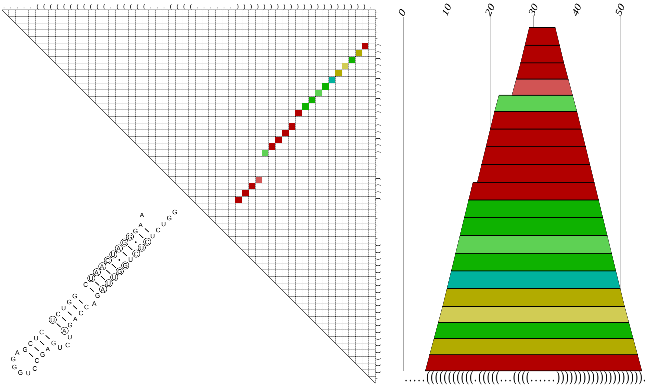
CHECK
compensatory
mutations

Conserved sub-structures

Base Pairing Probabilities

Nucl. Acids Res. **26**: 3825-3836 (1998), Comp. & Chem. **23**: 401-414 (1999)

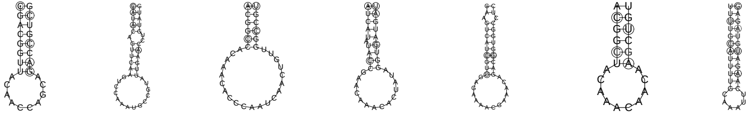
Examples: HIV-1 TAR-hairpin



Flaviviridae: Nucl. Acids Res. **29**: 5079-5089 (2001), Picornaviridae: J. Gen. Virol. **85**: 1113-1124 (2004), Broad survey: Bioinformatics **20**: 1495-1499 (2004)

Examples: Picornaviridae: Cis-acting-Replication Element (CRE)

The function of the CRE probably involves the initiation of the synthesis of the negative-sense strand template RNA during virus replication.



Aphthovirus	Enterovirus	Cardiovirus	HRV-A	HRV-B	Teschov.	Hepatov.
region:2C	2C	1B	2A	1B	2C	2C

```
Aphto      ~~~~CGAC-GGUU------ACA-CCAAGCA------GACCGUCG~~~~~
Entero     CAUACAGU-UCAAG--------UCCAAAU-GCCGUAUUGAACCUGUAUG
Cardio     ~~~~~ACG-GCCA---CAAACACCCAAUCAACUGU-UGGCCGU~~~~~~
HRV-A      ~~~AUCAUAUACCGAACAAACA---------CUAUAGGUGAUGAU~~~~
HRV-B      GAAGUCAU-CGUUGAGAAAACG---AAACA------GACGGUGGCCUC~
Tescho     ~~~~~~AC-GGCU--ACAAACA-----ACA------AGCUGU~~~~~~~
Hepato     UUUUGCAU-UUUG---CAAA--------------UUCAAGAUGUAGAG~
           ~~~(((((-((((.......................)))))))))~~~~
           1.......10........20........30........40.........
```

predicted in *Nucl. Acids Res.* **29** 5079-5089 (2001),

experimentally detected by Gerber, Wimmer Paul *J.Virol.* **75** 10979-10990 (2001).

A Method for Large Genomes: RNAz

∗ Two ingredients: Thermodynamic Stability & Structure Conservation

Measuring thermodynamic stability of ncRNAs

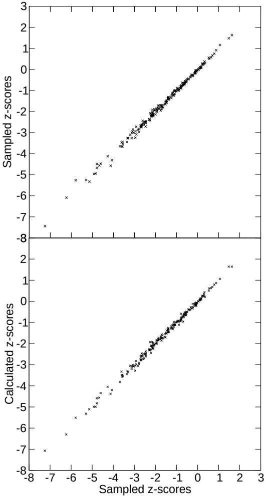
- Naturally occurring structured RNAs have a lower folding energy compared to random sequences of the same size and base composition?
  1. Calculate native MFE m.
  2. Calculate mean μ and standard deviation σ of MFEs of a large number of shuffled random sequences.
  3. Express significance in standard 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.

Efficient calculation of stability *z*-scores

- The mean μ and standard deviation σ of random samples of a given sequence are functions of the length and the base composition:

$$\mu, \sigma(length, \frac{GC}{AT}, \frac{G}{C}, \frac{A}{T})$$

- **Calculating** *z*-scores is thus a 5 dimensional regression problem.
- The regression problem is solved using a Support Vector Machine regression algorithm.
- The SVM was trained on 10,000 synthetic sequences spaced evenly in the variable space.
- The regression calculation is of the same accuracy as the sampling procedure.



z-scores of known ncRNAs

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

- Functional RNAs are clearly more stable than random sequences.
- However: The scores are too small to discriminate reliably in a genome-wide screens since the z-score distributions have heavy tails.

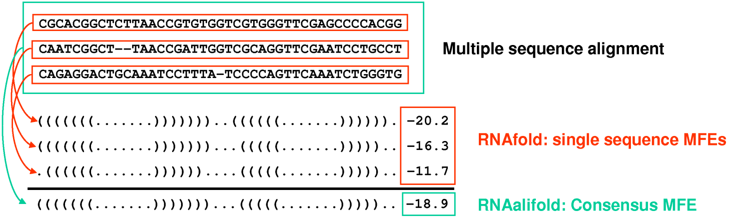
Consensus folding using RNAalifold

- 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)
```

J.Mol.Biol. **319**:1059-1066 (2002)

The structure conservation index



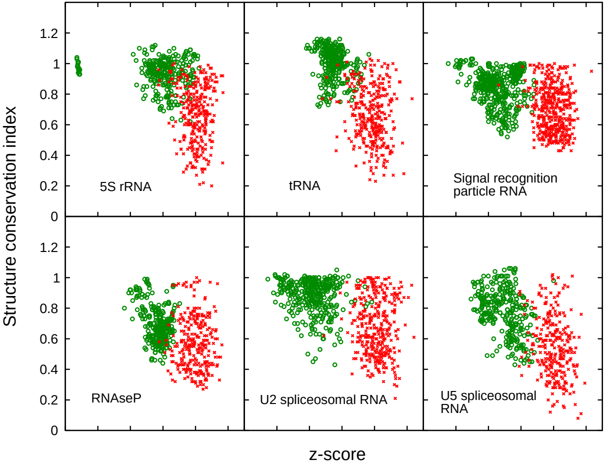
```
CGCACGGCTCTTAACCGTGTGGTCGTGGGTTCGAGCCCCACGG
CAATCGGCT--TAACCGATTGGTCGCAGGTTCGAATCCTGCCT
CAGAGGACTGCAAATCCTTTA-TCCCCAGTTCAAATCTGGGTG

(((((((.......)))))))..((((((.......)))))).  -20.2
(((((((.......)))))))...(((((.......)))))..  -16.3
.((((((.......))))))....(((((.......)))))..  -11.7
(((((((.......)))))))...(((((.......)))))..  -18.9
```

Multiple sequence alignment
RNAfold: single sequence MFEs
RNAalifold: Consensus MFE

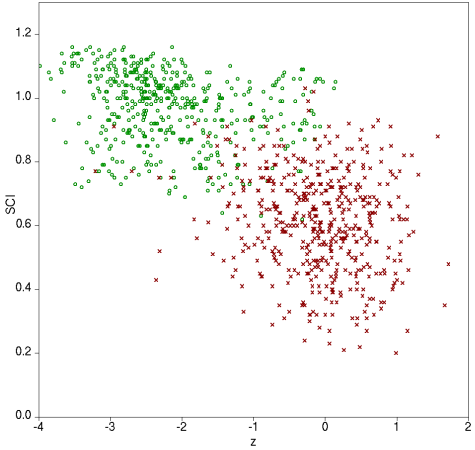
$$\text{SCI} = \frac{\text{Consensus MFE}}{\text{Mean single MFEs}}$$

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

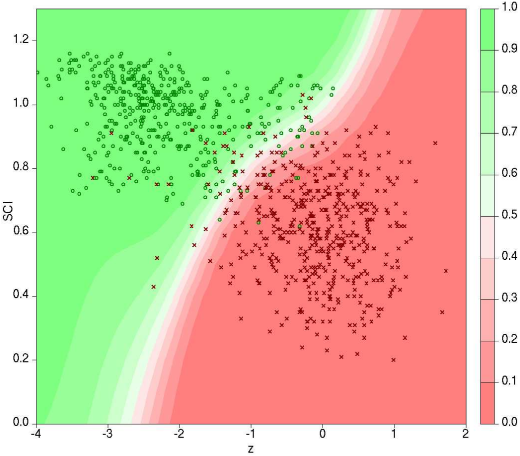
Separation of native ncRNAs from random controls in two dimensions



Classification based on both scores



Classification based on both scores



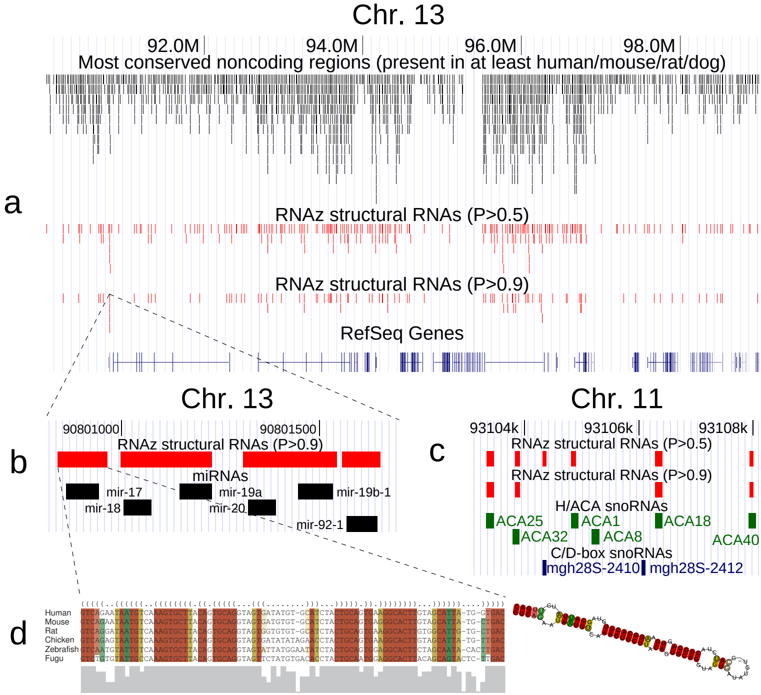
Implementation and availability

- The approach is implemented in ANSI `C` in the program `RNAz`.
- The z-score regression is limited to 400 nucleotides.
- The classification model is currently limited to alignments of six sequences.
- At least an order of magnitude faster than other programs.
- `RNAz` is freely available:
  Download from `www.tbi.univie.ac.at/~wash/RNAz`

Proc. Natl. Acad. Sci. USA **102**: 2454-2459 (2005)

Screening the human genome

- Large scale comparative screen including:
  - human, mouse, rat, dog
  - chicken
  - fugu, zebrafish
- Reduction of the $\approx$ 3.095 MB human genome:
  - Take $\approx$ 5% of the best conserved regions
  - Remove all annotated coding exons
  - Only take alignments strictly conserved in all 4 mammals.
- $\rightarrow$ 438,788 alignments alignments covering 82.64 MB

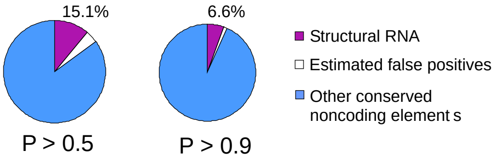

Chr. 13
92.0M
94.0M
96.0M
98.0M
Most conserved noncoding regions (present in at least human/mouse/rat/dog)
a
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
RefSeq Genes
Chr. 13
b
90801000
90801500
RNAz structural RNAs (P>0.9)
miRNAs
mir-17
mir-18
mir-19a
mir-20
mir-19b-1
mir-92-1
Chr. 11
c
93104k
93106k
93108k
RNAz structural RNAs (P>0.5)
RNAz structural RNAs (P>0.9)
H/ACA snoRNAs
ACA25
ACA32
ACA1
ACA8
ACA18
ACA40
C/D-box snoRNAs
mgh28S-2410
mgh28S-2412
d
(((((..((((((..((((((((.((.(((((...(((.........)))...))))).)).))))))))...))))))....)))))
Human GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATGT-GCATCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-GTGAC
Mouse GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATGTGT-GCATCTACTGCAGTGAGGGCACTTGTAGCATTA-TG-CTGAC
Rat GTCAGGATAATGTCAAAGTGCTTACAGTGCAGGTAGTGGTGTGT-GCATCTACTGCAGTGAAGGCACTTGTGGCATTG-TG-CTGAC
Chicken GTCAGAGTAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATATAGAACCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-TTGAC
Zebrafish GTCAATGTATTGTCAAAGTGCTTACAGTGCAGGTAGTATTATGGAATATCTACTGCAGTGGAGGCACTTCTAGCAATA-CACTTGAC
Fugu GTCTGTGTATTGCCAAAGTGCTTACAGTGCAGGTAGTTCTATGTGACACCTACTGCAATGGAGGCACTTACAGCAGTACTC-TTGAC

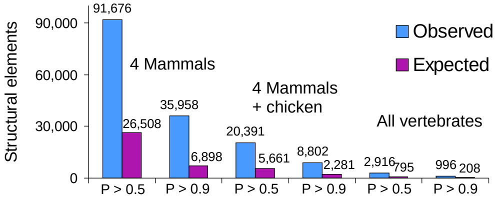
Results of Human Genome Screen

	Genome Coverage		Alignments	RNAz hits $p > 0.9$		
	Size	Fraction	Number	Size	Fraction of	Number
	(MB)	(%)		(MB)	input (%)	
Human genome	3,095.02	100.00	–			
PhastCons most conserved	137.85	4.81	1,601,903			
without coding regions	110.04	3.84	1,291,385			
without alignments $< 50nt$	103.83	3.33	564,455			
Set 1: 4 Mammals	82.64	2.88	438,788	5.46	6.62	35,985
Set 2: + Chicken	24.00	0.85	104,266	1.34	5.50	8,802
Set 3: + Fugu or zebrafish	6.86	0.24	30,896	0.14	2.03	996

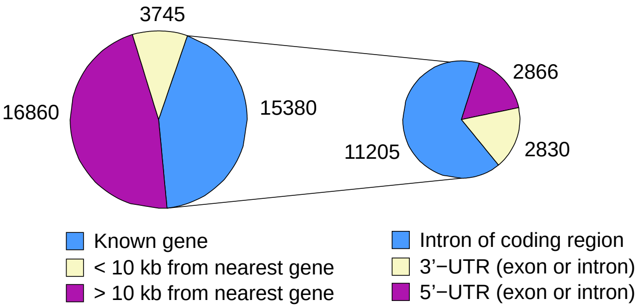
Nature Biotechn. **23**: 1383-1390 (2005)

Pictures instead of Numbers

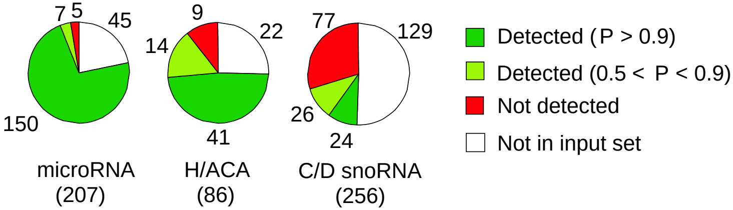




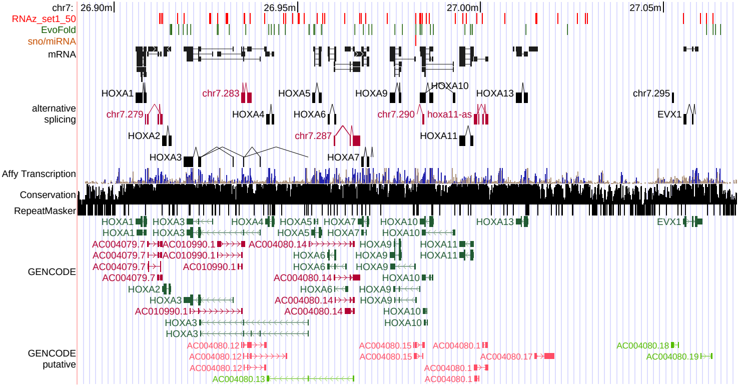
Distribution related to known protein gene annotation



Sensitivity on known classes of ncRNAs



Not all ncRNAs have conserved secondary structures!



Other RNAz Screens

- Urochordates: *Ciona intestinalis* & *Ciona savignyi*
  only a few conserved RNA with Oikopleura dioica

  Bioinformatics **21**(S2): i77-i78

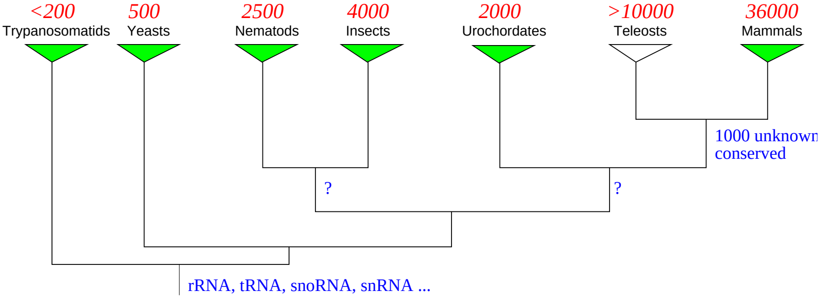
- Nematodes: *Caernorhabditis elegans* & *Caenorhabditis briggsae*

  JEZ:MDE 2006 epub

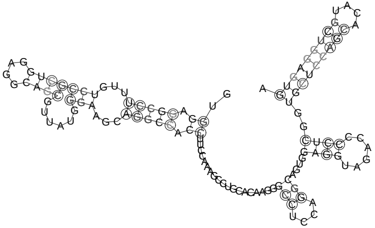
- Teleost fishes: *Danio rerio*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Oryzias latipes* (partial)
  (in progress)
- Trypanosomatids: Trypananosoma and Leishmania species
- Yeasts. (joint work with Kay Nieselt and Stephan Steigele)

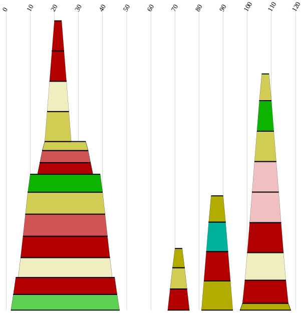
Summary

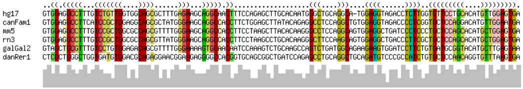
Predicted structured RNAs (RNAz predictions, $p > 0.9$)



Novel Human ncRNA Candidates

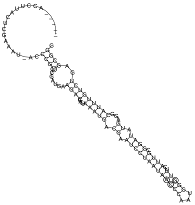




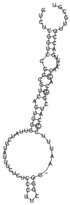


```
        ..((.(((((...(((((((....))))......)))...))))).))...................(((....))).....((((......))))..(((((((((....))))))))).
hg17    GTGGAGGCCTTTGTCTGTTGGTGGCAGCGCTTTGAGAAGCAGGCAACTTTCCAGAGCTTGCACAATGGGCTGCAGGCA-TGGAGGTAGACCTCTTGGTGTTCCTGCACATGTTGGAGTGA
canFam1 GTGGAGGCCTTCATCCGCTGGAGGCAGCGCTATGGGAAGCAGGCCACCTTCTGGAGCTTATACAGAGGCCTCCAGGCTGTGGAGGTAGACCCCTCGGTGCTCCAGGACATGTTGGAATGA
mm5     GTGGAGGCCTTTGTCCGCTGGCGCCAGCGTTTTGGGAAGCAGGCCACCTTCCTAAGCTTACACAAGGGCCTCCAGGCAGTGGAGGCTGATCCCTCCCTGCTCCAGCACATGCTGGAGTGA
rn3     GTGGAGGCCTTTGTCCGCTGGCGCCAGCGTTATGGGAAGCAGGCCACCTTCCTAAGCCTGCACAAGGGCTTCCAAGCAGTGGAGGCTGACCCTTCGCTGCTCCAGCACATGCTGGAGTGA
galGal2 GTAGCTGCGTTTGTCCTGTGGAGACAGCGTTTTGGGAAAAGTGCAACAATCCAAAGTCTGCAAGCCAGTCTGATGGCAGAAGAAGTGGATCCTTCTGTGATGCGGTACATGCTTGAGTAA
danRer1 CTCGCTGGGCTGGTGATGTGGACGCAGAGGAACGGAAGAGGGGCCACGGTGCAGCGGCTGATCCAGAGCCTGCAGGCTGCAGATGTCCCGCCATCTGTGCTCCAACAGGTGTTTAAGTGA
```

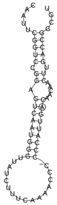
Novel ncRNA Candidates in *Caenorhabditis*



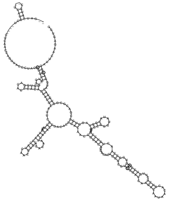
CeN23 (UM1)
unknown



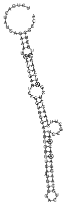
CeN74 (UM3)
sb-RNA



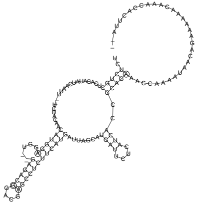
CeN77 (UM3)
sb-RNA



513253
(UM2)



515948
(UM3)



513590
(UM1)

Efforts to Annotate the RNAz Results

ongoing effort

- Large number of microRNA candidates
- approximately 30-40 good H/ACA-box snoRNAs
- only 6% of hits (comparable to estimated false positive rate) overlaps with predicted coding regions
- few clusters of signals with high sequence-similarity work in progress: structure-based clustering (joint work with Rolf Backofen's lab in Freiburg)

BOTTOM LINE: most signals still unclassified.
We need MUCH better methods to recognize members of known RNA classes

`RNAmicro`: A classificator for microRNA Precursors

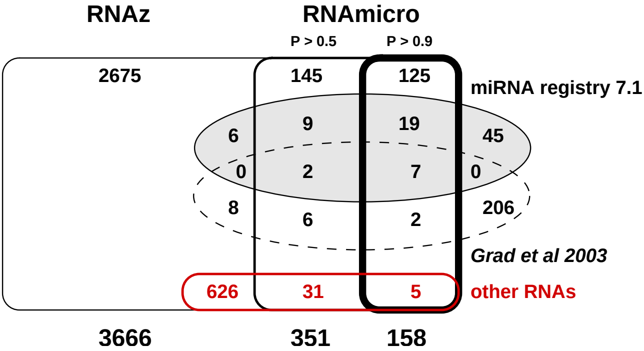
- Input: Multiple Sequence alignment
- Preprocessing: non-restrictive check for almost-hairpin structure

  Some known microRNA precursors, notably some `let-7` family members have small branches!
- SVM Classification with few descriptors:

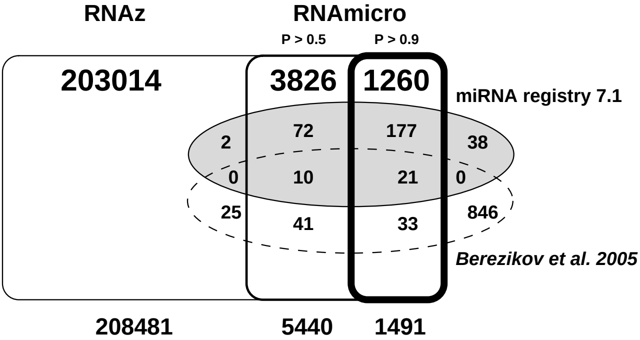
Property	#	Descriptors
Structure	2	l_s, l_h
Sequence composition	1	G+C
Sequence conservation	4	$S_{5'}$, $S_{3'}$, S_0, $S_{\min}$
Thermodynamic stability	4	$\bar{E}$, $\bar{\epsilon}$, $\bar{\eta}$, $\bar{z}$
Structure conservation	1	E_{cons}

ISMB 2006, in press

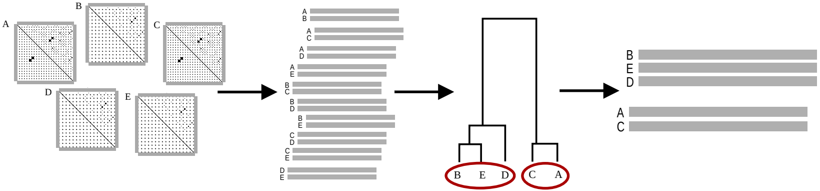
Results: *Caenorhabditis*



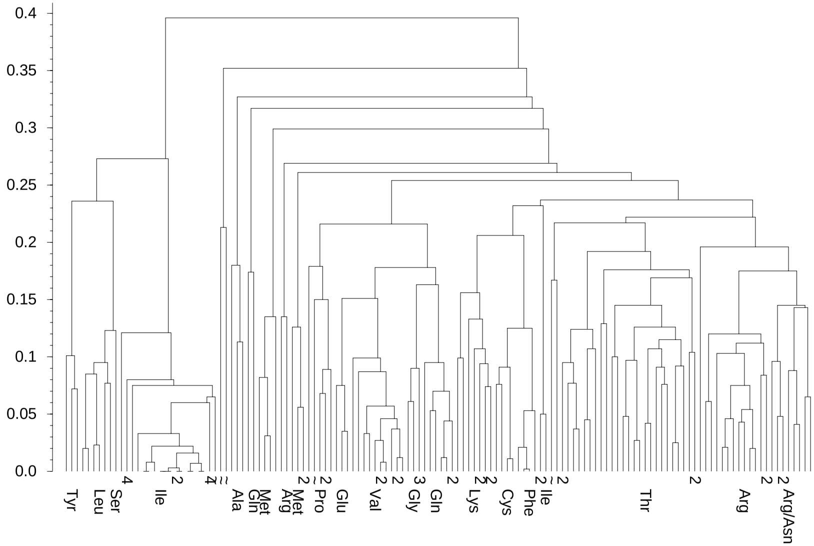
Results: *Mammals*



Clustering



Proof of Concept: tRNAs in *Ciona intestinalis*



Summary

- Some classes of ncRNAs, namely the structures ones, can be found efficiently by means of comparative genomics
- There are Tens of Thousands of structured RNAs of unknown function in the human genome
- Some of them probably act, like microRNA and snoRNAs by binding to other RNAs. These could be investigated using RNA cofolding approaches (ongoing research).
- *So far, we know only of the proverbial tip of the iceberg of the complexity of cellular regulation*
- & RNA bioinformatics is a really cool research topic ...

Acknowledgments: It's not my fault . . .

- Leipzig: Kristin Missal, Dominic Rose, Jana Hertel, Manja Lindemeyer, Matthias Kruspe, Sonja J. Prohaska, Claudia Fried, Roman Stocsits, Axel Mosig Bettina Müller (FH Weihenstephan), Katrin Sameith (U Jena)
- Vienna: Stefan Washietl, Ivo L. Hofacker, Christoph Flamm, Andrea Tanzer, Stefan Bernhart
  Susanne Rauscher, Caroline Thurner, Christina Witwer, Ingrid Abfalter, and many others
- Havard: Walter Fontana
- Beijing: Xiaopeng Zhu, Wei Deng, Geir Skogerbø, Runsheng Chen
- Tübingen: Stephan Steigele, Kay Nieselt,
- Copenhagen: Jan Gorodkin, Stefan Seemann
- Freiburg: Rolf Backofen, Sebastian Will