

The RNA World, Fitness Landscapes, and all that

Peter F. Stadler

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

RNomics Group, Fraunhofer Institute IZI, Leipzig
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

FOGA07, Cd. México, Jan 2007

Why RNA?

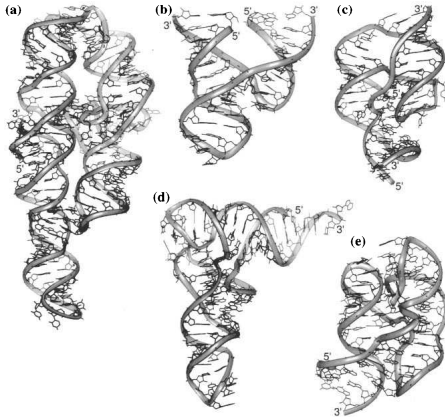
- ▶ until relatively recently:
Central Dogma of Molecular Biology
DNA → RNA → Protein
DNA = “genetic memory”, RNA = working copy, proteins do the work
- ▶ around 1980: discovery of catalytic RNAs (Nobelprize for Tom Cech and Sidney Altman)
nevertheless long considered “exotic” remnants from the ancient RNA world
⇒ **RNA World Hypothesis** for the Origin of Life
- ▶ around 2000: structure of the ribosome shows 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

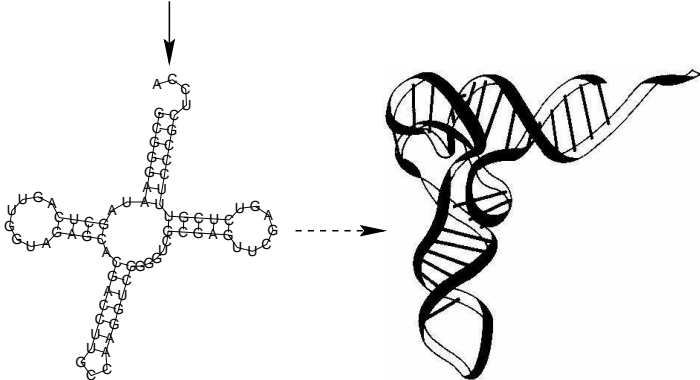


- (a) Group I intron P4-P6 domain
- (b) Hammerhead ribozyme
- (c) HDV ribozyme
- (d) Yeast tRNA^{Phe}
- (e) L1 domain of 23S rRNA

Hermann & Patel, JMB 294, 1999

The RNA Model

GCGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUCCCGCUCCA



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 $(i, k), (j, k) \in \Omega$ implies $i = j$.
- ▶ $(i, j) \in \Omega$ 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:

$$\bullet \text{xxxxxxx} + \sum (\text{xxxx})\text{xxxx}$$

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

RNA Folding in a nutshell



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

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

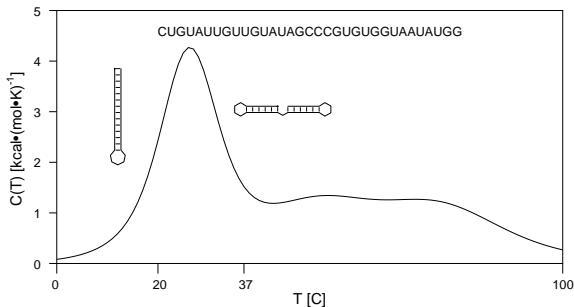
$$Z_{ij} = Z_{i+1,j} + \sum_{\substack{k \\ (i,k)\text{pair}}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\varepsilon_{ik}/RT)$$

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

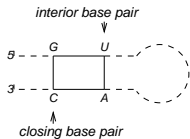
A word on the Partition Function

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.:

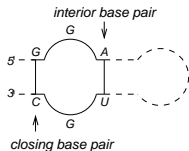
- ▶ Free energy $G = -RT \ln Z$
- ▶ Expected Energy $\langle E \rangle = RT^2 \frac{\partial \ln Z}{\partial T}$
- ▶ Heat Capacity $C_p = -T \frac{\partial^2 G}{\partial T^2}$



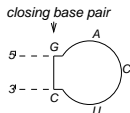
Realistic Energy Model



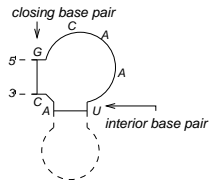
stacking pair



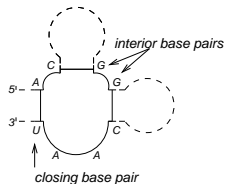
interior loop



hairpin loop



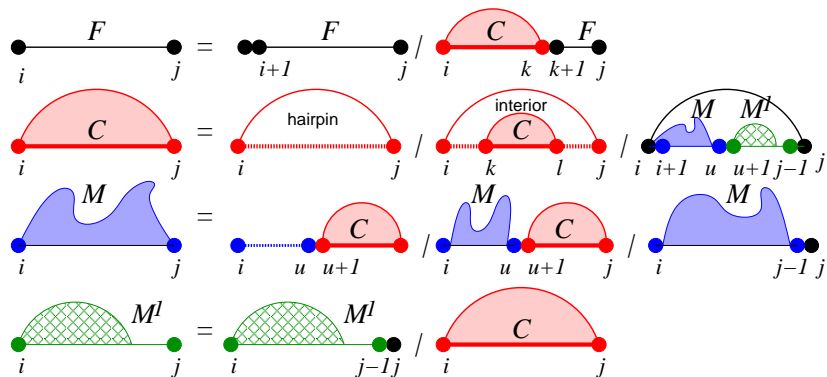
bulge



multi-loop

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

Recursions for Linear RNAs



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

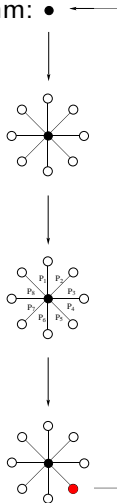
Simulate folding kinetics by a Monte-Carlo type algorithm:

- 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 $f : V \rightarrow \mathbb{R}$.

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)

Ruggedness



Bryce Canyon, UT
rugged



Capulin Volcano, NM
"smooth"

Measures of Ruggedness

- ▶ Number of Local Minima and Maxima

- ▶ Length of Adaptive Walks

- ▶ Correlation Functions:

Random walk on V as defined through \mathcal{X} : x_0, x_1, x_2, \dots
(Markov process on V).

\implies "Time Series" $f(x_0), f(x_1), f(x_2), \dots$

\implies Autocorrelation Function

$$r(s) = \frac{\left\langle (f(x_{t+s}) - \langle f \rangle) (f(x_t) - \langle f \rangle) \right\rangle_t}{\left\langle (f(x_t) - \langle f \rangle)^2 \right\rangle_t}$$

For simplicity:

$$f(x) \rightarrow f(x) - \bar{f} \quad \bar{f} = \frac{1}{|V|} \sum_{x \in V} f(x)$$

(Subtract the landscape average)

Random Walk = Markov process with transition matrix \mathbf{T}

Theorem. $r(s) = \langle f, \mathbf{T}^s f \rangle / \langle f, f \rangle$.

Theorem. $r(s)$ is an exponential function iff f is an eigenfunction of \mathbf{T} . We say (V, f, \mathbf{T}) is an *elementary* landscape.

For d -regular graphs: Relation with graph Laplacians

$$-\Delta = A - D = d(T - I)$$

Elementary landscapes are eigenfunctions of the Graph Laplacian

Amplitude Spectra

Let $\{\varphi_k\}$ be an orthonormal basis of eigenvectors of \mathbf{T} . We will be interested in the “Fourier Transform” $\{a_k\}$ where

$$f(x) = \sum_k a_k \varphi_k(x)$$

Theorem. $\text{var}[f] := \frac{1}{|V|} \sum_x f(x)^2 = \sum_{k \neq 0} |a_k|^2$.
Collect terms that belong to the *same eigenspace*:

$$B_p = \frac{1}{\text{var}[f]} \sum_{k: -\Delta\varphi_k = \Lambda_p \varphi_k} |a_k|^2$$

“Amplitude Spectrum of f ”.

Quantities derived from Amplitude Spectra

- ▶ Autocorrelation Function:

$$r(s) = \sum_{p \neq 0} B_p (1 - \Lambda_p / D)^s$$

- ▶ Correlation length

$$\ell = \sum_{s=0}^{\infty} r(s) = D \sum_{p \neq 0} B_p / \Lambda_p$$

- ▶ Elementary landscapes:

f is elementary if and only if $B_p = 1$ for a single p , and 0 otherwise

Some Elementary Landscapes

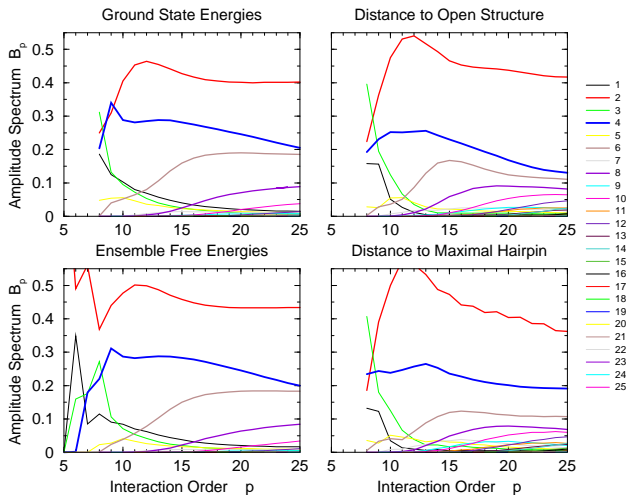
Problem	Move Set	D	Λ	ℓ/n
NAES	Hamming	n	4	1/4
p-spin	Hamming	n	$2p$	$1/(2p)$
WP	Hamming	n	4	1/4
GC	Hamming	$(\alpha - 1)n$	2α	$(1 - 1/\alpha)/2$
XY-Hamiltonian	Hamming	$(\alpha - 1)n$	2α	$(1 - 1/\alpha)/2$
	cyclic	$2n$	$8 \sin^2(\pi/\alpha)$	$1/[4 \sin^2(\pi/\alpha)]$
GBP	Exchange	$n^2/4$	$2(n - 1)$	$1/8 \cdot n/(n - 1)$
symmetric TSP	Transposition	$n(n - 1)/2$	$2(n - 1)$	1/4
	Inversions	$n(n - 1)/2$	n	$1 - 1/n)/4$
GMP	Transposition	$n(n - 1)/2$	$2(n - 1)$	$n/4$

NAES = Non-All-Equal-Satisfiability, WP = Weight Partition, GC = Graph Coloring with α colors, GBP = Graph

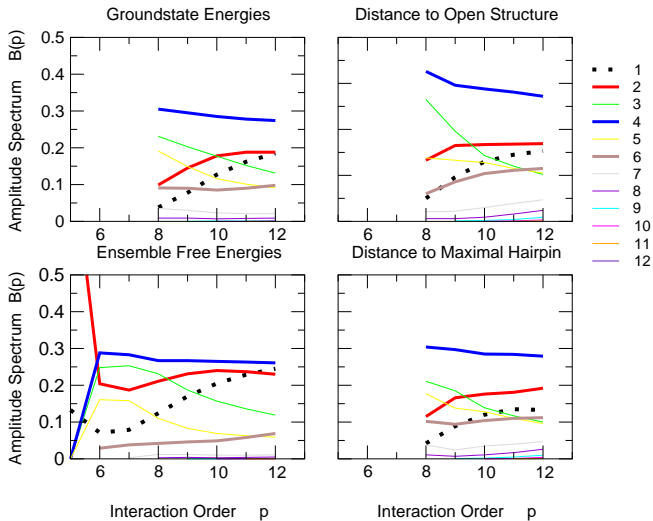
Bipartitioning, TSP = Traveling Salesman Problem, GMP = Graph Matching Problem XY-Hamiltonian:

$$\sum_{i < j} J_{ij} \cos\left(\frac{2\pi}{\alpha}(x_i - x_j)\right)$$

Amplitude Spectra of RNA Landscapes



Amplitude Spectra of RNA Landscapes



Neutrality

Degree of neutrality $\mathbb{E}[\nu_x]$

is the expected number of neutral neighbors of x .

Suppose $\mu_0 > 0$. Then

$$\mathbb{E}[\nu_x] = \sum_{y: y \in \partial x} \mu^{c_x(y)}$$

where

$$c_x(y) = |\{j | \theta_j(x) \neq \theta_j(y)\}| \quad y \in \partial x$$

Similar equations can be obtained for $\mathbb{E}[\nu_x \nu_y]$.

\implies Correlation length ℓ and degree of neutrality ν_x can be tuned independently of each other in short range p -spin models.

Energy barriers

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

Depth and Difficulty

(borrowed from simulated annealing theory)

$$D = \max \{ B(s) \mid s \text{ is not a global minimum} \}$$

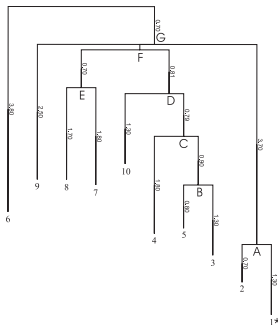
$$\psi = \max \left\{ \frac{B(s)}{f(s) - f(\min)} \mid s \text{ is not a global minimum} \right\}$$

Energy Barriers and Barrier Trees

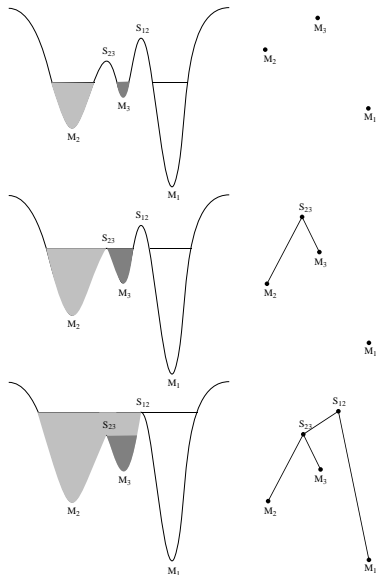
Some topological definitions:

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 conformation 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.

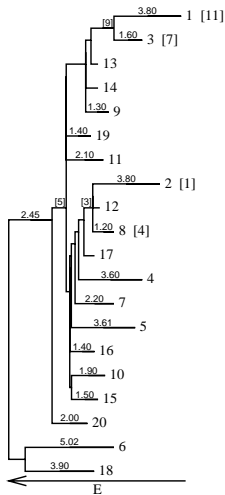
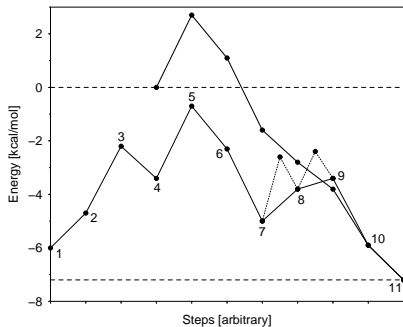
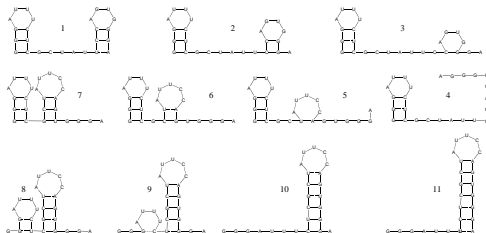
Basins $B(s_1), \dots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

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 :

$$r_{yx} = r_0 e^{-\frac{E_{yx}^\ddagger - E(x)}{RT}} \quad \text{for } x \neq y$$
$$r_{xx} = -\sum_{y \neq x} r_{yx}$$

Kinetics as a Markov process:

$$\frac{dp_x}{dt} = \sum_{y \in X} r_{xy} p_y(t).$$

Transition states:

$$E_{yx}^\ddagger = \max\{E(x), E(y)\}$$

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:

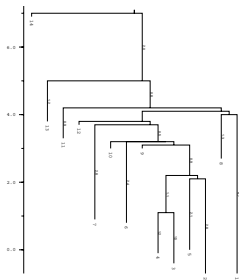
$$G(\alpha) = -RT \ln Z_{\alpha}$$

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

$r_{\beta\alpha}$ “on flight” while executing the barriers program.

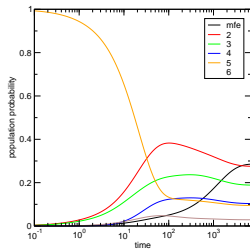
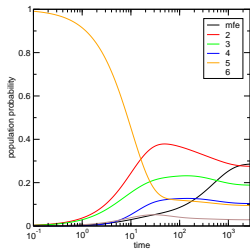
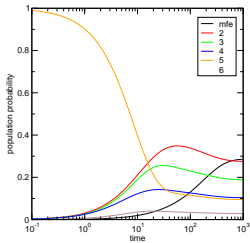
Transition state free energy:

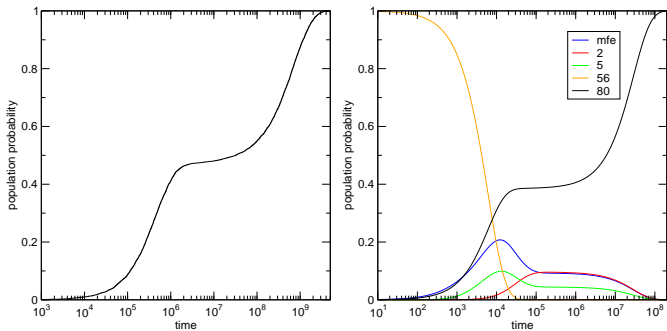
$$G_{\beta\alpha}^{\ddagger} = -RT \ln \sum_{y \in \beta} \sum_{x \in \alpha} e^{-\frac{E_{xy}^{\ddagger}}{RT}}$$



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

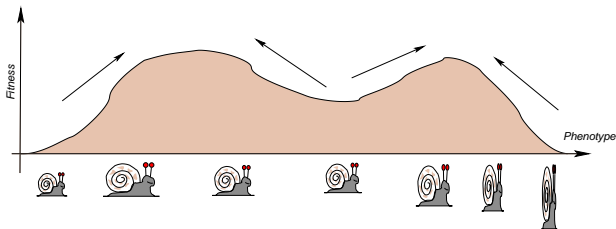
PART II: How Do RNAs Evolve

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:

genotype \mapsto phenotype \mapsto 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)

\implies less than 3^n structures

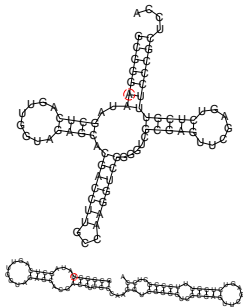
but 4^n sequences.

\implies **Redundancy**

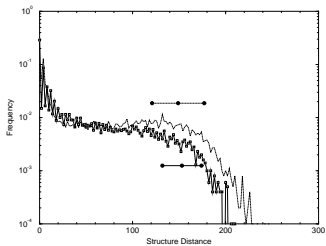
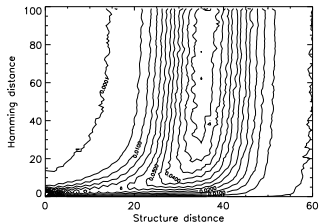
How are sequences folding into the same structure distributed in sequence space?

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

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

Paired bases: 5' and 3' side correlated:

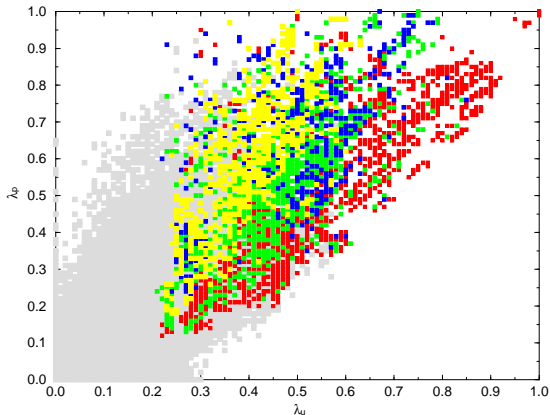
Alphabet: $\mathcal{B} = \{AU, UA, GC, CG, GU, UG, \}$.

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

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

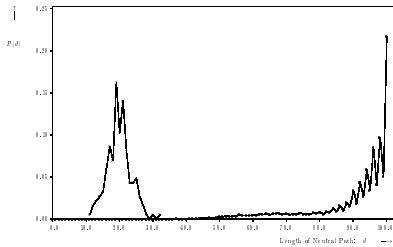
\implies Two neutrality parameters λ_u and λ_p

Connected Components of Neutral Networks

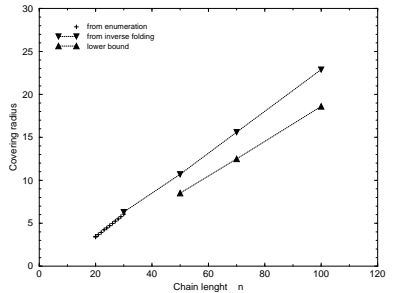


gray	many small components	red	1 connected component
green	2 equal sized components	yellow	3 components size 2:1
blue	4 equal sized components		

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

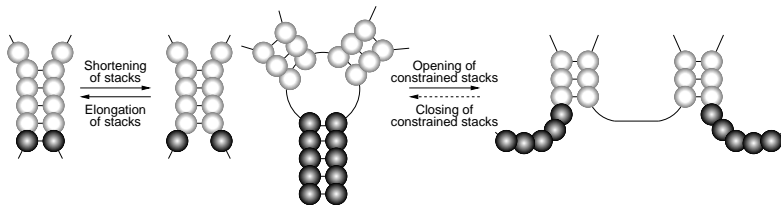
Accessibility

Fontana & Schuster 1998

Idea: The “interface” between two structures is large if 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:



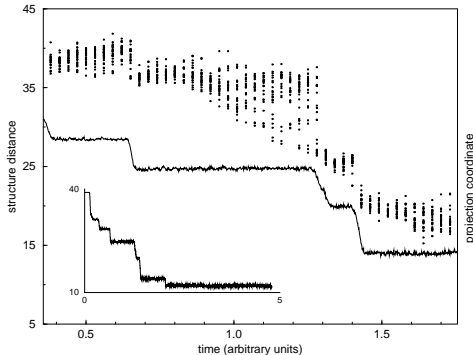
Summary II: Sequence-Structure Map of RNA

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

Diffusion Constant

... can be deduced from Moran model:

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

A ... replication rate

n ... sequence length

N ... population size

p ... mutation rate

$\bar{\lambda}$... 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:

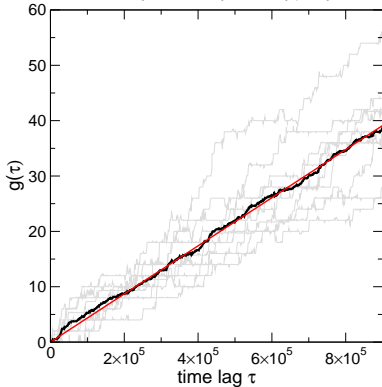
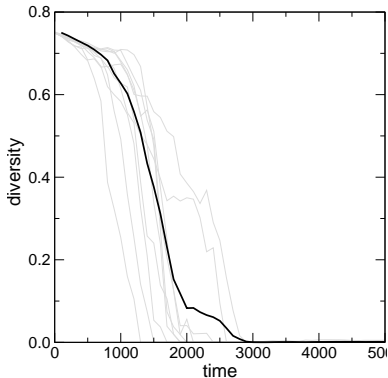
$$\dot{x}_k = x_k \left(\sum_j A_{kj} x_j - \sum_{i,j} A_{ij} x_i x_j \right) + \sum_{l,j} (Q_{kl} A_{lj} x_j x_l - Q_{lk} A_{kj} x_k x_j)$$

where

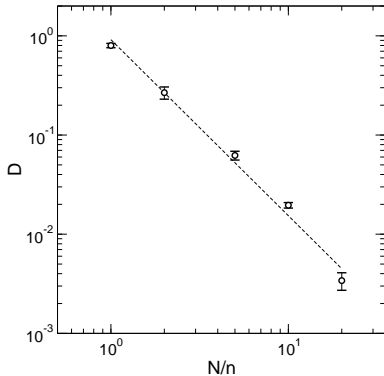
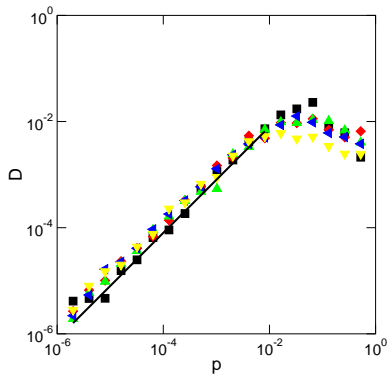
$$Q_{kl} = (1 - p)^{n-d(k,l)} \left(\frac{p}{\alpha - 1} \right)^{d(k,l)}$$

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

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



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

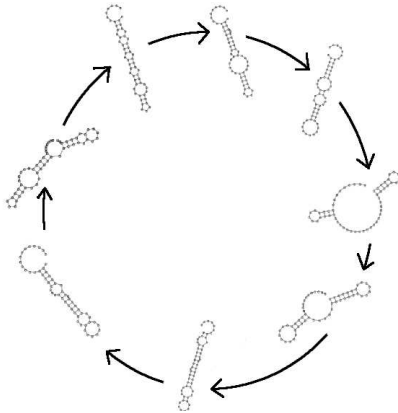


Left: Diffusion coefficient D as a function of the mutation rate for $N = 10, 20, 30, 40, 80$ and

$n = 10, 20, 30, 40, 80$ such that $N/n = 1$ after equilibration for 10^5 timesteps. Right: Dependence of the ratio

D/p on N/n .

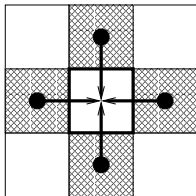
An RNA-Based Model in the Plane




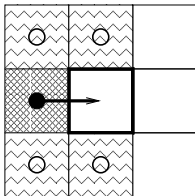
Target hypercycle with 8 members.


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

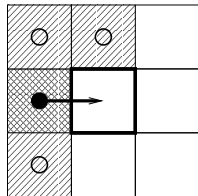
Spatial Extension: CA Model




 *Possible Replicators*

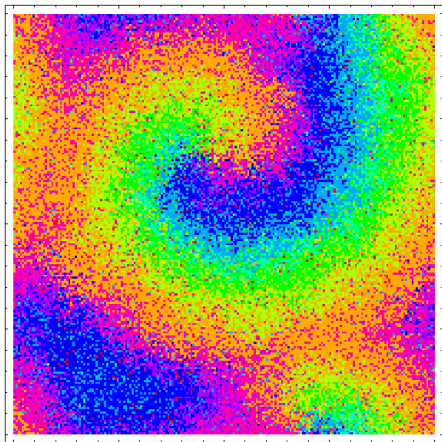


 *Possible Catalysts*



 *Actual Catalysts*

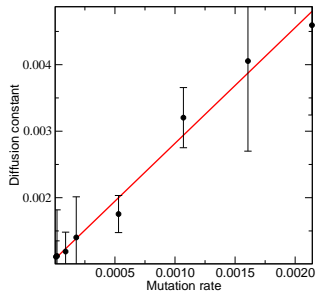
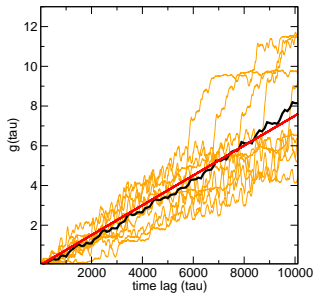
Rules of replication. For each of the neighbors (●) of the empty cell (marked by a bold outline) the replication rate ρ_z is computed taking into account their neighbors in the direction of the replication (○) as potential catalysts. The neighbor with the largest values of ρ_z invades the empty position. 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 III: Dynamics of RNA Evolution

- ▶ 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.
- ▶ \implies 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.

Acknowledgments: It's not my fault . . .

Peter Schuster, Walter Fontana, Wolfgang Schnabl

Christoph Flamm, Ivo L. Hofacker

Christian M. Reidys

Camille Stephan-Otto Attolini, Bärbel Stadler