Evolution of RNA Molecules
From Neutral Networks of Structures to Complex Interaction Patterns

Peter Schuster
Institut für Theoretische Chemie der Universität Wien, Austria
and the Santa Fe Institute, NM

Collectives formation and specialization in biological and social systems
Santa Fe, 20.– 22.04.2005
Web-Page for further information:

http://www.tbi.univie.ac.at/~pks
5'-End

GCGGAUUAGCUCAAGGCGAGACACCAUCAGAAYAUCUGGAGMUCUGUUGCACGACGACGACACGAUUUCGCACCA

Biochemical and chemical probing

Structure prediction

Crystallography

NMR, FRET, ....

3'-End

5'-End

3'-End

5'-End
1. Folding and inverse folding of RNA
2. Neutral networks
3. Darwinian evolution of RNA
4. Learning by the Darwinian mechanism
5. Folding kinetics and metastable structures
6. Intersections and conformational switches
1. Folding and inverse folding of RNA

2. Neutral networks

3. Darwinian evolution of RNA

4. Learning by the Darwinian mechanism

5. Folding kinetics and metastable structures

6. Intersections and conformational switches
RNA sequence: GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGGUACUCCA

RNA folding:
Structural biology, spectroscopy of biomolecules, understanding molecular function

Biophysical chemistry: thermodynamics and kinetics
Empirical parameters

RNA structure of minimal free energy

One sequence – one structure problem
The minimum free energy structures on a discrete space of conformation
RNA sequence: GUAUCGAAAUAUCGUCGUAGCGUAUUGGGGAUGCUUGGACGGUCCCAUCGGGUACUCCA

RNA folding:
Structural biology, spectroscopy of biomolecules, understanding molecular function

Inverse folding of RNA:
Biotechnology, design of biomolecules with predefined structures and functions

Inverse Folding Algorithm

Iterative determination of a sequence for the given secondary structure

RNA structure of minimal free energy

Sequence, structure, and design
1. Folding and inverse folding of RNA

2. Neutral networks

3. Darwinian evolution of RNA

4. Learning by the Darwinian mechanism

5. Folding kinetics and metastable structures

6. Intersections and conformational switches
The inverse folding algorithm searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.
Binary sequences are encoded by their decimal equivalents:

- "0" = 00000 = CCCCC,
- "14" = 01110 = CGGGC,
- "29" = 11101 = GGGCG, etc.

Mutant class

0

1

Binary sequences are encoded by their decimal equivalents:

- "0" = 00000 = CCCCC,
- "14" = 01110 = CGGGC,
- "29" = 11101 = GGGCG, etc.

2

3

4

5

Hypercube of dimension n = 5

Decimal coding of binary sequences

Sequence space of binary sequences of chain length n = 5
Hamming distance \( d_{H}(I_1, I_2) = 4 \)

(i) \( d_{H}(I_1, I_1) = 0 \)
(ii) \( d_{H}(I_1, I_2) = d_{H}(I_2, I_1) \)
(iii) \( d_{H}(I_1, I_3) \leq d_{H}(I_1, I_2) + d_{H}(I_2, I_3) \)

The Hamming distance between sequences induces a metric in sequence space
Mapping from sequence space into structure space and into function

\[ S_k = \psi(I.) \]

\[ f_k = f(S_k) \]
The Hamming distance between structures in parentheses notation forms a metric in structure space.
$S_k = \psi(I.)$

$f_k = f(S_k)$

Sequence space

Structure space

Real numbers
$S_k = \psi(I.)$
The pre-image of the structure $S_k$ in sequence space is the neutral network $G_k$. 

$$S_k = \psi(I.)$$
Space of sequences: \( I = \{I_1, I_2, I_3, I_4, \ldots, I_N\} \); Hamming metric

Space of structures: \( S = \{S_1, S_2, S_3, S_4, \ldots, S_M\} \); metric (not required)

\[ N \gg M \]

\[ \psi(I_j) = S_k \]

Neutral network: \( G_k = \psi^{-1}(S_k) \equiv \{ I_j | \psi(I_j) = S_k \} \)

A mapping \( \psi \) and its inversion
Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected
1. Folding and inverse folding of RNA
2. Neutral networks
3. **Darwinian evolution of RNA**
4. Learning by the Darwinian mechanism
5. Folding kinetics and metastable structures
6. Intersections and conformational switches
Copying of single-strand RNA-molecules:

**Plus-Minus-Replication**
Variation of the RNA sequence through copying errors
\[
\frac{dx_i}{dt} = \sum_j f_j Q_{ji} x_j - x_i \Phi
\]

\[
\Phi = \sum_j f_j x_j ; \quad \sum_j x_j = 1 ; \quad \sum_i Q_{ij} = 1
\]

\[
[I_i] = x_i \geq 0 ; \quad i = 1,2,\ldots,n ; \\
[A] = a = \text{constant}
\]

\[
Q_{ij} = (1-p)^{\ell-d(i,j)} p^{d(i,j)}
\]

- \(p\) \(\ldots\) Error rate per digit
- \(\ell\) \(\ldots\) Chain length of the polynucleotide
- \(d(i,j)\) \(\ldots\) Hamming distance between \(I_i\) and \(I_j\)

Chemical kinetics of replication and mutation as parallel reactions
Replication rate constant:

\[ f_k = \frac{\gamma}{\alpha + \Delta d_s^{(k)}} \]

\[ \Delta d_s^{(k)} = d_H(S_k, S_r) \]

Selection constraint:

Population size, \( N = \# \text{ RNA molecules} \), is controlled by the flow

\[ N(t) \approx \bar{N} \pm \sqrt{\bar{N}} \]

Mutation rate:

\[ p = 0.001 / \text{site} \times \text{replication} \]

The flowreactor as a device for studies of evolution \textit{in vitro} and \textit{in silico}
Randomly chosen initial structure

Phenylalanyl-tRNA as target structure
The molecular quasispecies in sequence space
*In silico* optimization in the flow reactor: Evolutionary trajectory
28 neutral point mutations during a long quasi-stationary epoch

Neutral genotype evolution during phenotypic stasis
1. Folding and inverse folding of RNA
2. Neutral networks
3. Darwinian evolution of RNA
4. **Learning by the Darwinian mechanism**
5. Folding kinetics and metastable structures
6. Intersections and conformational switches
Element in example 1: The RNA molecule
The molecular quasispecies in sequence space
Evolutionary trajectory

Spreading of the population through diffusion on a neutral network

Drift of the population center in sequence space
Spread of population in sequence space during a quasistationary epoch: $t = 150$
Spread of population in sequence space during a quasistationary epoch: $t = 170$
Spread of population in sequence space during a quasistationary epoch: $t = 200$
Spread of population in sequence space during a quasistationary epoch: $t = 350$
Spread of population in sequence space during a quasistationary epoch: $t = 500$
Spread of population in sequence space during a quasistationary epoch: $t = 650$
Spread of population in sequence space during a quasistationary epoch: $t = 820$
Spread of population in sequence space during a quasistationary epoch: $t = 825$
Spread of population in sequence space during a quasistationary epoch: $t = 830$
Spread of population in sequence space during a quasistationary epoch: $t = 835$
Spread of population in sequence space during a quasistationary epoch: $t = 840$
Spread of population in sequence space during a quasistationary epoch: $t = 845$
Spread of population in sequence space during a quasistationary epoch: $t = 850$
Spread of population in sequence space during a quasistationary epoch: $t = 855$
Element in example 2: The ant worker
Foraging behavior of ant colonies
Ant colony                              Food source detected                           Food source

Foraging behavior of ant colonies
Ant colony                          Pheromone trail laid down                        Food source

Foraging behavior of ant colonies
Ant colony                         Pheromone controlled trail                        Food source

Foraging behavior of ant colonies
<table>
<thead>
<tr>
<th>Element</th>
<th>Evolution of RNA</th>
<th>Foraging ants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>RNA nucleotide</td>
<td>Individual worker ant</td>
</tr>
<tr>
<td>Phenotype</td>
<td>RNA sequence</td>
<td>Worker ant collective</td>
</tr>
<tr>
<td>Learning entity</td>
<td>RNA structure</td>
<td>Foraging path</td>
</tr>
<tr>
<td>Relation between elements</td>
<td>Population of molecules</td>
<td>Ant colony</td>
</tr>
<tr>
<td>Search process</td>
<td>Mutation</td>
<td>Reorientation of path segment</td>
</tr>
<tr>
<td>Search space</td>
<td>Optimization of structure</td>
<td>Optimization of path</td>
</tr>
<tr>
<td>Random step</td>
<td>Sequence space</td>
<td>Three-dimensional space</td>
</tr>
<tr>
<td>Self-enhancing process</td>
<td>Mutation</td>
<td>Segment of ant walk</td>
</tr>
<tr>
<td>Measure of activity</td>
<td>Replication</td>
<td>Secretion of pheromone</td>
</tr>
<tr>
<td>Goal of the search</td>
<td>Mean replication rate</td>
<td>Mean pheromone concentration</td>
</tr>
<tr>
<td>Temporary memory</td>
<td>Target structure</td>
<td>Richest food source</td>
</tr>
<tr>
<td></td>
<td>Sequence distribution</td>
<td>Pheromone trail</td>
</tr>
</tbody>
</table>

**Learning** at population or colony level *by trial and error*

Two examples: (i) RNA model and (ii) ant colony
1. Folding and inverse folding of RNA
2. Neutral networks
3. Darwinian evolution of RNA
4. Learning by the Darwinian mechanism

5. **Folding kinetics and metastable structures**
6. Intersections and conformational switches
RNA secondary structures derived from a single sequence
Kinetic Folding of RNA Secondary Structures


Mean folding curves for three small RNA molecules with different folding behavior

I₁ = ACUGAU CU GU AG U CAC
I₂ = AUU GAG C AU AU U UC AC
I₃ = CGGGCU AU UU U AG CUG

S₀ = •• ( ( ( ( •••• ) ) ) ) •
Search for local minima in conformation space

Suboptimal conformations

Local minimum

Free energy $\Delta G^0$

$S_h$

$S_1^{(h)}$

$S_2^{(h)}$

$S_3^{(h)}$

$S_4^{(h)}$

$S_5^{(h)}$

$S_6^{(h)}$

$S_7^{(h)}$

$S_8^{(h)}$

$S_9^{(h)}$
Definition of a ‘barrier tree’
Example of an unefficiently folding small RNA molecule with \( n = 15 \)

\[ I_1 = \text{ACUGAU} \text{CGUAGUCAC} \]
Example of an easily folding small RNA molecule with $n = 15$

$I_2 = \text{AUUGAGCAUAUUUCAC}$

..(((((....)))))

$S_0$
Example of an easily folding and especially stable small RNA molecule with n = 15

\[ I_3 = \text{CGGGCUAUUUAGCUG} \]
A nucleic acid molecule folding in two dominant conformations
Folding dynamics of the sequence GGCCCUUUGGGGCGAGACCUCUAAAAAGGGUC
1. Folding and inverse folding of RNA
2. Neutral networks
3. Darwinian evolution of RNA
4. Learning by the Darwinian mechanism
5. Folding kinetics and metastable structures
6. Intersections and conformational switches
The **compatible set** $C_k$ of a structure $S_k$ consists of all sequences which form $S_k$ as its minimum free energy structure (the **neutral network** $G_k$) or one of its suboptimal structures.
**Intersection** of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \notin \emptyset$
The barrier tree connecting $S_1$ and $S_0$
A ribozyme switch

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds
Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototypic ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the samefolded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origins of new, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can vary far in sequence space (1). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (1). Because these disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of neutral variants that were functional. Hence, sequence homogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting sequence space just as those widely-dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (4-9).

Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (3, 4-9). Of particular interest is whether ribozyme neutral networks approach each other so closely that they interconvert. If so, a single sequence could be capable of folding into two different conformations, would have two different catalytic activities, and could access by neutral drift every sequence on both networks. With interacting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry non-functional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close approximation of the neutral networks for the hepatitis delta virus (HDV) self-activating ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (2). It joins an oligonucleotide substrate to its 5' terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 16 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isoform (10). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (11). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antiongenu- nor HDV ribozyme (12), which undergoes self-cleavage at a rate similar to that reported for other antiongenu- nor enzymes (13, 14).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no functional structural similarities that might favor an interaction of their two neutral networks. Nevertheless, sequences can be designed that simultan- eously satisfy the base-pairing requirements
Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (A) and a natural cleavage ribozyme of hepatitis-δ-virus (B)
The sequence at the intersection:

An RNA molecules which is 88 nucleotides long and can form both structures
Two neutral walks through sequence space with conservation of structure and catalytic activity


*Structural parameters affecting the kinetic competition of RNA hairpin formation*, in press 2004.

*Structural parameters affecting the kinetic competition of RNA hairpin formation*, in press 2004.
J1LH barrier tree

Free energy [kcal / mole]
Conclusions

I. The Darwinian mechanism of optimization through variation and selection operates equally well on simple and complex reproducing elements because only the number of fertile offspring counts.

II. Darwinian learning through trial and error takes place on the level of populations. It does not require sophisticated elements and occurs even with self-replicating molecules.

III. Even simple molecules have the capacity for a rich repertoire of properties and interactions. For example, they can have multiple structures and functions.
Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)
Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

European Commission: Project No. EU-980189

Austrian Genome Research Program – GEN-AU

Siemens AG, Austria

Universität Wien and the Santa Fe Institute
Coworkers

Walter Fontana, Harvard Medical School, MA

Christian Forst, Christian Reidys, Los Alamos National Laboratory, NM

Peter Stadler, Bärbel Stadler, Universität Leipzig, GE

Jord Nagel, Kees Pleij, Universiteit Leiden, NL

Ivo L.Hofacker, Christoph Flamm, Andreas Svrček-Seiler
Univiersität Wien, AT

Stefan Bernhart, Ulrike Langhammer, Ulrike Mückstein,
Universität Wien, AT

Ulrike Göbel, Walter Grüner, Stefan Kopp, Jaqueline Weber
Institut für Molekulare Biotechnologie, Jena, GE

Andreas Wernitznig, Michael Kospach, Kurt Grünberger, Stefan Wuchty
Web-Page for further information:

http://www.tbi.univie.ac.at/~pks