Evolution with RNA Molecules: From experiment to theory and back

Peter Schuster
Institut für Theoretische Chemie und Molekulare Strukturiobiologie der Universität Wien

Workshop Biotechnologie

Heidelberg, 30.09.2002
<table>
<thead>
<tr>
<th></th>
<th>Generation time</th>
<th>10,000 generations</th>
<th>10⁶ generations</th>
<th>10⁷ generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA molecules</td>
<td>10 sec</td>
<td>27.8 h = 1.16 d</td>
<td>115.7 d</td>
<td>3.17 a</td>
</tr>
<tr>
<td></td>
<td>1 min</td>
<td>6.94 d</td>
<td>1.90 a</td>
<td>19.01 a</td>
</tr>
<tr>
<td>Bacteria</td>
<td>20 min</td>
<td>138.9 d</td>
<td>38.03 a</td>
<td>380 a</td>
</tr>
<tr>
<td></td>
<td>10 h</td>
<td>11.40 a</td>
<td>1 140 a</td>
<td>11 408 a</td>
</tr>
<tr>
<td>Higher multicellular</td>
<td>10 d</td>
<td>274 a</td>
<td>27 380 a</td>
<td>273 800 a</td>
</tr>
<tr>
<td>organisms</td>
<td>20 a</td>
<td>20 000 a</td>
<td>2 × 10⁷ a</td>
<td>2 × 10⁸ a</td>
</tr>
</tbody>
</table>

Generation times and evolutionary timescales
Evolution of RNA molecules based on Qβ phage


The serial transfer technique applied to RNA evolution *in vitro*
Reproduction of the original figure of the serial transfer experiment with Qβ RNA


Fig. 9. Serial transfer experiment. Each 0.25 ml standard reaction mixture contained 40 μg of Qβ replicase and 32P-UTP. The first reaction (0 transfer) was initiated by the addition of 0.2 μg ts-1 (temperature-sensitive RNA) and incubated at 35°C for 20 min, whereupon 0.02 ml was drawn for counting and 0.02 ml was used to prime the second reaction (first transfer), and so on. After the first 13 reactions, the incubation periods were reduced to 15 min (transfers 14-29). Transfers 30-38 were incubated for 10 min. Transfers 39-52 were incubated for 7 min, and transfers 53-74 were incubated for 5 min. The arrows above certain transfers (0, 8, 14, 29, 37, 53, and 73) indicate where 0.001-0.1 ml of product was removed and used to prime reactions for sedimentation analysis on sucrose. The inset examines both infectious and total RNA. The results show that biologically competent RNA censuses to appear after the 4th transfer (Mills et al. 1967).
Decrease in mean fitness due to quasispecies formation

The increase in RNA production rate during a serial transfer experiment
Ronald Fisher‘s conjecture of optimization of mean fitness in populations does not hold in general for replication-mutation systems: In general evolutionary dynamics the mean fitness of populations may also decrease monotonously or even go through a maximum or minimum. It does also not hold in general for recombination of many alleles and general multi-locus systems in population genetics.

Optimization of fitness is, nevertheless, fulfilled in most cases, and can be understood as a useful heuristic.
Selection of Qb-RNA through replication in a capillary

G. Bauer, H. Otten, J. S. McCaskill, 
No new principle will declare itself from below a heap of facts.

Sir Peter Medawar, 1985
Complementary replication as the simplest copying mechanism of RNA. Complementarity is determined by Watson-Crick base pairs: $\text{G} \rightleftharpoons \text{C}$ and $\text{A} = \text{U}$.
\[ \frac{dx_i}{dt} = f_i x_i - x_i \Phi = x_i (f_i - \Phi) \]

\[ \Phi = \sum_j f_j x_j \; ; \; \sum_j x_j = 1 \; ; \; i,j = 1,2,\ldots,n \]

\[ [I_i] = x_i \Phi \; ; \; i = 1,2,\ldots,n \; ; \]

\[ [A] = a = \text{constant} \]

\[ f_m = \max \{f_j; j=1,2,\ldots,n\} \]

\[ x_m(t) \leq 1 \; \text{for} \; t \leq \tilde{t} . \]

**Reproduction** of organisms or replication of molecules as the basis of selection.
\[ s = \frac{f_2 - f_1}{f_1}; \quad f_2 > f_1; \quad x_1(0) = 1 - 1/N; \quad x_2(0) = 1/N \]

Selection of advantageous mutants in populations of \( N = 10000 \) individuals
Mutations in nucleic acids represent the mechanism of variation of genotypes.
Theory of molecular evolution


(A) + $I_j \rightarrow [A] = a = \text{constant}$

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

$[I_i] = x_i \in \mathbb{R} ; \quad i = 1, 2, \ldots, n ;$

$dx_i / dt = \sum_j f_j Q_{ji} x_j - x_i \Phi$

$Q_{ij} = (1-p)^{l-d(i,j)} p^{d(i,j)}$

$p \quad \text{Error rate per digit}$

$l \quad \text{Chain length of the polynucleotide}$

$d(i,j) \quad \text{Hamming distance between} \ I_i \text{ and } I_j$

Chemical kinetics of replication and mutation as parallel reactions
Quasispecies as a function of the replication accuracy $q$
The molecular quasispecies in sequence space
In the case of non-zero mutation rates (p>0 or q<1) the Darwinian principle of optimization of mean fitness can be understood only as an optimization heuristic. It is valid only on part of the concentration simplex. There are other well defined areas were the mean fitness decreases monotonously or were it may show non-monotonous behavior. The volume of the part of the simplex where mean fitness is non-decreasing in the conventional sense decreases with increasing mutation rate p.

In systems with recombination a similar restriction holds for Fisher‘s „universal selection equation“. Its global validity is restricted to the one-gene (single locus) model.
Theory of genotype – phenotype mapping


Genotype-phenotype relations are highly complex and only the most simple cases can be studied. One example is the folding of RNA sequences into RNA structures represented in course-grained form as secondary structures.

The RNA genotype-phenotype relation is understood as a mapping from the space of RNA sequences into a space of RNA structures.
The RNA secondary structure is a listing of GC, AU, and GU base pairs. It is understood in contrast to the full 3D-or tertiary structure at the resolution of atomic coordinates. RNA secondary structures are biologically relevant. They are, for example, conserved in evolution.
RNA Minimum Free Energy Structures

Efficient algorithms based on dynamical programming are available for computation of secondary structures for given sequences. Inverse folding algorithms compute sequences for given secondary structures.


**Vienna RNA Package**: http://www.tbi.univie.ac.at (includes inverse folding, suboptimal structures, kinetic folding, etc.)

The inverse folding algorithm searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.
Criterion of Minimum Free Energy

Sequence Space

Shape Space
The **RNA model** considers RNA sequences as genotypes and simplified RNA structures, called secondary structures, as phenotypes.

The **mapping** from genotypes into phenotypes is many-to-one. Hence, it is redundant and not invertible.

Genotypes, i.e. RNA sequences, which are mapped onto the same phenotype, i.e. the same RNA secondary structure, form **neutral networks**. Neutral networks are represented by graphs in sequence space.
Hamming distance $d_H(S_1,S_2) = 4$

(i) $d_H(S_1,S_1) = 0$

(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

(iii) $d_H(S_1,S_3) \leq d_H(S_1,S_2) + d_H(S_2,S_3)$

The Hamming distance induces a metric in sequence space
Single point mutations as moves in sequence space
\[ S_k = \psi(I.) \]
\[ f_k = f(S_k) \]

Mapping from sequence space into phenotype space and into fitness values.
$S_k = \psi(I.)$

$f_k = f(S_k)$

Sequence space
Phenotype space
Non-negative numbers
The pre-image of the structure $S_k$ in sequence space is the neutral network $G_k$. 

$$S_k = \psi(I.)$$ 

$$f_k = f(S_k)$$
**Neutral networks** are sets of sequences forming the same structure. $G_k$ is the pre-image of the structure $S_k$ in sequence space:

$$G_k = y^{-1}(S_k) \oplus \{ y_j | y(I_j) = S_k \}$$

The set is converted into a graph by connecting all sequences of Hamming distance one.

**Neutral networks** of small RNA molecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, $N=4^n$, becomes very large with increasing length, and is prohibitive for numerical computations.

**Neutral networks** can be modelled by random graphs in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.
\[ G_k = y^{-1}(S_k) \cup \{I_j \mid y(I_j) = S_k\} \]

\[ \lambda_j = \frac{12}{27}, \quad \tilde{\lambda}_k = \frac{\emptyset \cup \{j(k)\}}{|G_k|} \]

Connectivity threshold:

\[ \lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)} \]

Alphabet size \( k \):  **AUGC**  \( \tilde{n} \)  \( k = 4 \)

<table>
<thead>
<tr>
<th></th>
<th>( k )</th>
<th>( \lambda_{cr} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.4226</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.3700</td>
<td></td>
</tr>
</tbody>
</table>

\( \tilde{\lambda}_k > \lambda_{cr} \) .... network \( G_k \) is connected

\( \tilde{\lambda}_k < \lambda_{cr} \) .... network \( G_k \) is **not** connected

Mean degree of neutrality and connectivity of **neutral networks**
A multi-component neutral network
A connected neutral network
Compatibility of sequences with structures

A sequence is compatible with its minimum free energy structure and all its suboptimal structures.
The **compatible set** $C_k$ of a structure $S_k$ consists of all sequences which form $S_k$ as its minimum free energy structure (neutral network $G_k$) or one of its suboptimal structures.
A sequence at the intersection of two neutral networks is compatible with both structures
The intersection of two compatible sets is always non empty: \( C_1 \cap C_2 \neq \emptyset \)
Optimization of RNA molecules in silico


Randomly chosen initial structure

Phenylalanyl-tRNA as target structure
Fitness function:
\[ f_k = \frac{g}{a + Dd_S^{(k)}} \]
\[ Dd_S^{(k)} = d_s(I_k, I_t) \]

The flowreactor as a device for studies of evolution \textit{in vitro} and \textit{in silico}
The molecular quasispecies in sequence space
Evolutionary dynamics including molecular phenotypes
In silico optimization in the flow reactor: Trajectory (biologists’ view)
In silico optimization in the flow reactor: Trajectory (physicists' view)
In silico optimization in the flow reactor: Main transitions

In silico optimization in the flow reactor: Main transitions
Main or discontinuous transitions: Structural innovations, occur rarely on single point mutations

Closing of Constrained Stacks
In silico optimization in the flow reactor
Elongation of Stacks

Shortening of Stacks

Opening of Constrained Stacks

Minor or continuous transitions: Occur frequently on single point mutations

Opening of Constrained Stacks
### Statistics of evolutionary trajectories

<table>
<thead>
<tr>
<th>Population size N</th>
<th>Number of replications $&lt;n_{\text{rep}}&gt;$</th>
<th>Number of transitions $&lt;n_{\text{tr}}&gt;$</th>
<th>Number of main transitions $&lt;n_{\text{dtr}}&gt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 000</td>
<td>$(5.5 \pm [6.9,3.1]) \times 10^7$</td>
<td>$92.7 \pm [80.3,43.0]$</td>
<td>$8.8 \pm [2.4,1.9]$</td>
</tr>
<tr>
<td>2 000</td>
<td>$(6.0 \pm [11.1,3.9]) \times 10^7$</td>
<td>$55.7 \pm [30.7,19.8]$</td>
<td>$8.9 \pm [2.8,2.1]$</td>
</tr>
<tr>
<td>3 000</td>
<td>$(6.6 \pm [21.0,5.0]) \times 10^7$</td>
<td>$44.2 \pm [25.9,16.3]$</td>
<td>$8.1 \pm [2.3,1.8]$</td>
</tr>
<tr>
<td>10 000</td>
<td>$(1.2 \pm [1.3,0.6]) \times 10^8$</td>
<td>$35.9 \pm [10.3,8.0]$</td>
<td>$10.3 \pm [2.6,2.1]$</td>
</tr>
<tr>
<td>20 000</td>
<td>$(1.5 \pm [1.4,0.7]) \times 10^8$</td>
<td>$28.8 \pm [5.8,4.8]$</td>
<td>$9.0 \pm [2.8,2.2]$</td>
</tr>
<tr>
<td>30 000</td>
<td>$(2.2 \pm [3.1,1.3]) \times 10^8$</td>
<td>$29.8 \pm [7.3,5.9]$</td>
<td>$8.7 \pm [2.4,1.9]$</td>
</tr>
<tr>
<td>100 000</td>
<td>$(3 \pm [2,1]) \times 10^8$</td>
<td>$24 \pm [6,5]$</td>
<td>$9 \pm 2$</td>
</tr>
</tbody>
</table>

The number of **main transitions** or evolutionary innovations is constant.
„...Variations neither useful not injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions.

...“

Charles Darwin, Origin of species (1859)
Evolution in genotype space sketched as a non-descending walk in a fitness landscape
Bacterial Evolution


Fig. 1. Change in average cell size (1 fl = $10^{-15}$ L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (22). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these data (Table 1).

Fig. 2. Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (6, 7). The open symbols indicate the only two samples assigned to different steps by the cell size and fitness data.

Epochal evolution of bacteria in serial transfer experiments under constant conditions

Variation of genotypes in a bacterial serial transfer experiment

Evolutionary design of RNA molecules


Y. Wang, R.R.Rando, Specific binding of aminoglycoside antibiotics to RNA. Chemistry & Biology 2 (1995), 281-290

Selection cycle used in applied molecular evolution to design molecules with predefined properties.
The SELEX technique for the evolutionary design of *aptamers*
Sequences of aptamers binding theophyllin, caffeine, and related compounds

Secondary structures of aptamers binding theophyllin, caffeine, and related compounds
Dissociation constants and specificity of theophylline, caffeine, and related derivatives of uric acid for binding to a discriminating aptamer TCT8-4.

Table 1. Competition binding analysis with TCT8-4 RNA. The chemical structures are shown for a series of derivatives used in competitive binding experiments with TCT8-4 RNA (Fig. 2) (20). The right column represents the affinity of the competitor relative to theophylline, $K_d(c)/K_d(t)$, where $K_d(c)$ is the individual competitor dissociation constant and $K_d(t)$ is the competitive dissociation constant of theophylline. Certain data (denoted by $>\$) are minimum values that were limited by the solubility of the competitor. Each experiment was carried out in duplicate. The average error is shown.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Structure</th>
<th>$K_d(c)$ (µM)</th>
<th>$K_d(c)/K_d(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theophylline</td>
<td></td>
<td>0.32 ± 0.13</td>
<td>1</td>
</tr>
<tr>
<td>CP-theophylline</td>
<td></td>
<td>0.93 ± 0.20</td>
<td>2.9</td>
</tr>
<tr>
<td>Xanthine</td>
<td></td>
<td>8.5 ± 0.40</td>
<td>27</td>
</tr>
<tr>
<td>1-Methylxanthine</td>
<td></td>
<td>9.0 ± 0.30</td>
<td>28</td>
</tr>
<tr>
<td>3-Methylxanthine</td>
<td></td>
<td>2.0 ± 0.7</td>
<td>6.3</td>
</tr>
<tr>
<td>7-Methylxanthine</td>
<td></td>
<td>&gt; 500</td>
<td>&gt; 1500</td>
</tr>
<tr>
<td>3,7-Dimethylxanthine</td>
<td></td>
<td>&gt; 500</td>
<td>&gt; 1500</td>
</tr>
<tr>
<td>1,3-Dimethyluric acid</td>
<td></td>
<td>&gt; 1000</td>
<td>&gt; 3100</td>
</tr>
<tr>
<td>Hypoxanthine</td>
<td></td>
<td>49 ± 10</td>
<td>153</td>
</tr>
<tr>
<td>Caffeine</td>
<td></td>
<td>3500 ± 1500</td>
<td>10,900</td>
</tr>
</tbody>
</table>
Fig. 3. Schematic representation of the RNA (purple) binding site for theophylline (blue).

Schematic drawing of the aptamer binding site for the theophylline molecule
Aptamer binding to aminoglycosid antibiotics: Structure of ligands

Aptamer binding to aminoglycosid antibiotics: Sequence of low affinity RNA aptamers

Aptamer binding to aminoglycosid antibiotics: Sequence of high affinity RNA aptamers and dissociation constants of RNA-antibiotic complexes

Secondary structure of RNA aptamers binding to tobramycin and analogues
Formation of secondary structure of the tobramycin binding RNA aptamer

The three-dimensional structure of the tobramycin aptamer complex

Hammerhead ribozyme – The smallest based catalyst


Hammerhead ribozyme: The smallest known catalytically active RNA molecule
Allosteric effectors:

FMN = flavine mononucleotide

H10 – H12

theophylline

H14

Self-splicing allosteric ribozyme

H13

Hammerhead ribozymes with allosteric effectors
A ribozyme switch

Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (A) and a natural cleavage ribozyme of hepatitis-$d$-virus (B)
An RNA molecules which is 88 nucleotides long and can form both structures
THEOREM 5. INTERSECTION-THEOREM. Let s and s' be arbitrary secondary structures and C[s], C[s'] their corresponding compatible sequences. Then,

\[ C[s] \cap C[s'] \neq \emptyset. \]

Proof. Suppose that the alphabet admits only the complementary base pair [XY] and we ask for a sequence x compatible to both s and s'. Then \( \chi(s, s') \equiv D_n \) operates on the set of all positions \( \{x_1, \ldots, x_n\} \). Since we have the operation of a dihedral group, the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore, it suffices to choose an alternating sequence of the pairing partners X and Y. Thus, there are at least two different choices for the first base in the orbit.

Remark. A generalization of the statement of theorem 5 to three different structures is false.
Two neutral walks through sequence space with conservation of structure and catalytic activity.
Sequence of mutants from the intersection to both reference ribozymes
From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER\textsuperscript{1,2,3}, WALTER FONTANA\textsuperscript{3}, PETER F. STADLER\textsuperscript{2,3} AND IVO L. HOFACKER\textsuperscript{2}

\textsuperscript{1} Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07701 Jena, Germany
\textsuperscript{2} Institut für Theoretische Chemie, Universität Wien, Austria
\textsuperscript{3} Santa Fe Institute, Santa Fe, U.S.A.

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law; we find relatively few common and many rare ones. By using an algorithm for inverse folding, we show that sequences sharing the same structure are distributed randomly over sequence space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected and, even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and neighbours of Hamming distance 2, resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana et al. 1993a; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

Reference for postulation and \textit{in silico} verification of \textit{neutral networks}
Coworkers

Walter Fontana, Santa Fe Institute, NM

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

Peter Stadler, Universität Leipzig, GE

Ivo L.Hofacker, Christoph Flamm, Universität Wien, AT

Bärbel Stadler, Andreas Wernitznig, Universität Wien, AT
Michael Kospach, Ulrike Langhammer, Ulrike Mückstein, Stefanie Widder
Jan Cupal, Kurt Grünberger, Andreas Svrček-Seiler, Stefan Wuchty

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