Networks from Replicating Molecules

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria
and
The Santa Fe Institute, Santa Fe, New Mexico, USA

Workshop on Networks, Complexity, and Competition
Bled, 02.– 04.05.2008
Web-Page for further information:

http://www.tbi.univie.ac.at/~pks
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. Realistic fitness landscapes
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
1. Replication and selection

2. Mutation, quasispecies and error thresholds

3. Sequences, structures and neutrality

4. Realistic fitness landscapes

5. Replicating networks

6. RNA structure optimization

7. Experiments with RNA
James D. Watson, 1928-, and Francis H.C. Crick, 1916-2004
Nobel prize 1962

1953 – 2003 fifty years double helix

The three-dimensional structure of a short double helical stack of B-DNA
Base complementarity and conservation of genetic information
Replication fork in DNA replication

The mechanism of DNA replication is 'semi-conservative'
Complementary replication is the simplest copying mechanism of RNA. Complementarity is determined by Watson-Crick base pairs:

\[ G \equiv C \text{ and } A = U \]
Chemical kinetics of molecular evolution

Complementary replication as the simplest molecular mechanism of reproduction

\[
\begin{align*}
(A) + \text{I}_1 & \xrightarrow{f_1} \text{I}_2 + \text{I}_1 \\
(A) + \text{I}_2 & \xrightarrow{f_2} \text{I}_1 + \text{I}_2
\end{align*}
\]

\[
\frac{dx_1}{dt} = f_2 x_2 - x_1 \Phi \\
\frac{dx_2}{dt} = f_1 x_1 - x_2 \Phi
\]

\[
\Phi = \Sigma_i f_i x_i ; \quad \Sigma_i x_i = 1 ; \quad i = 1, 2
\]
Equation for complementary replication: \[ I_i = x_i \geq 0 , f_i > 0 \; ; i=1,2 \]

\[
\frac{dx_1}{dt} = f_2 x_2 - x_1 \phi , \quad \frac{dx_2}{dt} = f_1 x_1 - x_2 \phi , \quad \phi = f_1 x_1 + f_2 x_2 = f
\]

Solutions are obtained by integrating factor transformation

\[
x_{1,2}(t) = \frac{\sqrt{f_{2,1}} (\gamma_1(0) \cdot \exp(ft) + \gamma_2(0) \cdot \exp(-ft))}{(\sqrt{f_1} + \sqrt{f_2}) \gamma_1(0) \cdot \exp(ft) - (\sqrt{f_1} - \sqrt{f_2}) \gamma_1(0) \cdot \exp(-ft)}
\]

\[
\gamma_1(0) = \sqrt{f_1} x_1(0) + \sqrt{f_2} x_2(0) , \quad \gamma_2(0) = \sqrt{f_1} x_1(0) - \sqrt{f_2} x_2(0) , \quad f = \sqrt{f_1 \cdot f_2}
\]

\[
x_1(t) \rightarrow \frac{\sqrt{f_2}}{\sqrt{f_1} + \sqrt{f_2}} \quad \text{and} \quad x_2(t) \rightarrow \frac{\sqrt{f_1}}{\sqrt{f_1} + \sqrt{f_2}} \quad \text{as} \quad \exp(-ft) \rightarrow 0
\]
Kinetics of RNA replication

C.K. Biebricher, M. Eigen, W.C. Gardiner, Jr.
Biochemistry 22:2544-2559, 1983
Reproduction of organisms or replication of molecules as the basis of selection

\[ \frac{dx_i}{dt} = f_j x_i - x_i \Phi = x_i (f_i - \Phi) \]

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

\[ [I_i] = x_i \geq 0 ; \quad i = 1,2,\ldots,n \]

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

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

\[ x_m(t) \to 1 \text{ for } t \to \infty \]
Selection equation: \[ [I_i] = x_i \geq 0, \ f_i > 0 \]

\[ \frac{dx_i}{dt} = x_i \left( f_i - \phi \right), \quad i = 1, 2, \ldots, n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = f \]

Mean fitness or dilution flux, \( \phi(t) \), is a non-decreasing function of time,

\[ \frac{d\phi}{dt} = \sum_{i=1}^{n} f_i \frac{dx_i}{dt} = \overline{f}^2 - \left( \overline{f} \right)^2 = \text{var}\{f\} \geq 0 \]

Solutions are obtained by integrating factor transformation

\[ x_i(t) = \frac{x_i(0) \cdot \exp(f_i t)}{\sum_{j=1}^{n} x_j(0) \cdot \exp(f_j t)}; \quad i = 1, 2, \ldots, n \]
Selection between three species with $f_1 = 1$, $f_2 = 2$, and $f_3 = 3$
1. Replication and selection

2. Mutation, quasispecies and error thresholds

3. Sequences, structures and neutrality

4. Realistic fitness landscapes

5. Replicating networks

6. RNA structure optimization

7. Experiments with RNA
Variation of genotypes through mutation and recombination
Variation of genotypes through mutation

- Point mutation
- Insertion
- Deletion
Chemical kinetics of replication and mutation as parallel reactions
\[
\frac{dx_i}{dt} = \sum_{j=1}^{n} Q_{ij} f_j x_j - x_i \Phi
\]

with \[\Phi = \sum_{i=1}^{n} f_i x_i \text{ and } \sum_{i=1}^{n} x_i = 1\]

\[
\sum_{i=1}^{n} Q_{ij} = 1
\]

The replication-mutation equation
Mutation-selection equation: $[I_i] = x_i \geq 0, \ f_i > 0, \ Q_{ij} \geq 0$

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} Q_{ij} f_j x_j - x_i \Phi, \quad i=1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \Phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

Solutions are obtained after integrating factor transformation by means of an eigenvalue problem

$$x_i(t) = \frac{\sum_{k=0}^{n-1} \ell_{ik} \cdot c_k(0) \cdot \exp(\lambda_k t)}{\sum_{j=1}^{n} \sum_{k=0}^{n-1} \ell_{jk} \cdot c_k(0) \cdot \exp(\lambda_k t)}; \quad i=1,2,\cdots,n; \quad c_k(0) = \sum_{i=1}^{n} h_{ki} x_i(0)$$

$$W \div \{f_i Q_{ij}; \ i, j=1,2,\cdots,n\}; \quad L = \{\ell_{ij}; \ i, j=1,2,\cdots,n\}; \quad L^{-1} = H = \{h_{ij}; \ i, j=1,2,\cdots,n\}$$

$$L^{-1} \cdot W \cdot L = \Lambda = \{\lambda_k; k=0,1,\cdots,n-1\}$$
Variation of genotypes through point mutation
Uniform error rate model:

\[ Q_{ij} = p^{d_H(x_i, x_j)} (1 - p)^{(n - d_H(x_i, x_j))} \]

\[ d_H(x_i, x_j) \ldots \text{Hamming distance} \]
Formation of a quasispecies in sequence space
Formation of a quasispecies in sequence space
Formation of a quasispecies in sequence space
Formation of a quasispecies in sequence space
Uniform distribution in sequence space

Mutant cloud
The error threshold in replication
Antiviral strategy on the horizon

Error catastrophe laid its conceptual origins in the middle of the 20th century, when the consequences of mutations on an organism involved in proviral synthesis, as a theory of aging. In those times, biological processes were generally perceived differently than today. Infection diseases were regarded as a fleeting menace which would be eliminated through the use of antibiotics and antiviral agents. Microbial variation, although known in some cases, was not thought to be a significant problem for disease control. Variations in differentiated organisms were seen as residing essentially from exchanges of genetic material associated with sexual reproduction. The problem was to unravel the mechanisms of inheritance, expression of genetic information and metabolism. Few saw that genetic change was occurring at a速率 in all organisms, and still fewer recognized Darwinian principles as essential to the strategy of pathogenic viruses and cells. Population genetics was rarely used before or as experimental systems to define concepts in biological evolution. The discovery of genetic polymorphism among individuals of the same biological species came as a surprise when the first results on comparative electrophoretic mobility of enzymes were obtained. The advent of viral DNA and RNA, nucleoside antibiotics, and rapid nucleic acid sequencing techniques, molecular analyses of reagents reinforced the conclusion of extreme inter-individual genetic variation within the same species. Now, this largely to speculative progress in comparative genomics, we see cellular DNA, both prokaryotic and eukaryotic, as highly diverse. Most cellular processes, including such essential function, aging and transferring events as genetic replication, transcription and translation, are increasingly perceived as inherently inaccurate. Viruses, in particular RNA viruses, are among the most extreme examples of exploitation of replication accuracy for survival.

Error catastrophe, or the loss of meaningful genetic information through excess genetic variation, was formulated in qualitative terms as a consequence of an exponential theory, which was first developed to explain self-organization and adaptability of naturally occurring sequences in early stages of life. Recently, a conceptual extension of error catastrophe that could be defined as “induced genetic determinism” has emerged as a possible natural strategy. This is the topic of the current special issue of Virus Research.

Far would nowadays choose that one of the main strategies for the control of viral disease is short-term adaptability of viral pathogens. Adaptability of viruses follows the same Darwinian principles that have shaped biological evolution over eons, that is, repeated rounds of reproduction with genetic variation, competition, and selection, often perturbed by random events such as climatic fluctuations or population size. However, with viruses the consequences of the operation of these very same Darwinian principles are felt within very short times. Short-term evolution (within hours and days) can also be observed with some cellular pathogens, with subsets of normal cells, and cancer cells. The nature of RNA virus pathogenesis is unique for these reasons, and making the virus to cross the critical barrier towards transformation of genetic information is one of them.

The contributions to this volume have been chosen to reflect current lines of evidence (both experimental and theoretical) on which natural designs based on genetic determinism inflect upon viruses are being constructed. Theoretical studies have explored the copying error conditions that must be fulfilled by any information-bearing replicating system for the essential genetic information to be transmitted. Closely related to theoretical developments have been numerous experimental studies on quasispecies dynamics and their multiple biological manifestations. The latter can be summarized by saying that RNA virus, by virtue of accuracy in synthesis, replications rather than defined genetic entities, remarkably exploit their error-prone to overcome selective pressures unleashed by their replication instead. The base of survival strategies in clinical practice and the design of vaccines for a number of cancer RNA viruses-associated diseases, is currently provided by a sense of uncertainty. Another line of growing research is the understanding of copying fidelity by viral replication, aimed at understanding the molecular basis of quasispecies dynamic. Error catastrophe as a potential new antiviral strategy remains an important potential to be explored, that is the formation of natural and host-derived nucleic acids may be inserting, in some systems, its natural activity through enhanced integration.
Quasispecies as a function of the mutation rate $p$

$$f_0 = \sigma = 10$$

Single peak fitness landscape: $f_0 = f$ and $f_1 = f_2 = \ldots f_N = 1$

$$\sigma = \frac{f_0}{(1-x_0) \sum_{i=1}^{N} f_i x_i}$$

$I_0 \ldots$ master sequence; $N = \kappa^n$
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. Realistic fitness landscapes
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
$N = 4^n$

$N_S < 3^n$

Criterion: Minimum free energy (mfe)

Rules: \( \_ (\_ \_ ) \_ \in \{AU,CG,GC, GU, UA, UG\} \)

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs.
The inverse folding algorithm searches for sequences that form a given RNA structure.
Sequence space of binary sequences of chain length \( n = 5 \)
Sequence space of binary sequences of chain length $n = 5$

<table>
<thead>
<tr>
<th>Mutant class</th>
<th>Coding: $C = 0$ and $G = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$00000 = \text{CCCCC}$</td>
</tr>
<tr>
<td>1</td>
<td>$00001 = \text{CCCCG}$, ....</td>
</tr>
<tr>
<td>2</td>
<td>$00011 = \text{CCCGG}$, ....</td>
</tr>
<tr>
<td>3</td>
<td>$00111 = \text{CCGGG}$, ....</td>
</tr>
<tr>
<td>4</td>
<td>$01111 = \text{CGGGG}$, ....</td>
</tr>
<tr>
<td>5</td>
<td>$11111 = \text{GGGGG}$</td>
</tr>
</tbody>
</table>
Sequence space of binary sequences of chain length \( n = 5 \)

Binary sequences are encoded by their decimal equivalents:

\[ C = 0 \text{ and } G = 1 \]

"0" \( \equiv 00000 = \text{CCCCC,} \)

"14" \( \equiv 01110 = \text{CGGGC,} \)

"29" \( \equiv 11101 = \text{GGGCG,} \)

"31" \( \equiv 11111 = \text{GGGGG, etc.} \)
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
<table>
<thead>
<tr>
<th>Number</th>
<th>Mean Value</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52.31</td>
<td>85.30</td>
<td>9.24</td>
</tr>
</tbody>
</table>

Number of Structures: 1000

Shadow – Surrounding of an RNA structure in shape space – **AUGC** alphabet
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. **Realistic fitness landscapes**
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
Fitness landscapes showing error thresholds
\[ \sigma = \frac{f_0}{(1-x_0) \sum_{i=1}^{N} f_i x_i} \]

\( I_0 \) … master sequence; \( N = \kappa^n \)

Error threshold: Error classes and individual sequences

\( n = 10 \) and \( \sigma = 2 \)
Error threshold: Individual sequences

\( n = 10, \sigma = 2 \) and \( d = 0, 1.0, 1.85 \)
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. Realistic fitness landscapes
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
Neutral network

\[ \lambda = 0.01, \ s = 367 \]

Error threshold: Individual sequences

\[ n = 10, \ \sigma = 1.1, \ d = 1.0 \]
Error threshold: Individual sequences

\[ n = 10, \sigma = 1.1, d = 1.0 \]
STATIONARY MUTANT DISTRIBUTIONS AND EVOLUTIONARY OPTIMIZATION

PETER SCHUSTER and JÖRG SWETINA
Institut für theoretische Chemie
und Strahlenchemie der Universität Wien,
Währingerstraße 17,
A 1090 Wien,
Austria

Molecular evolution is modeled by erroneous replication of binary sequences. We show how the selection of two species of equal or almost equal selective value is influenced by its nearest neighbors in sequence space. In the case of perfect neutrality and sufficiently small error rates we find that the Hamming distance between the species determines selection. As the error rate increases the fitness parameters of neighboring species become more and more important. In the case of almost neutral sequences we observe a critical replication accuracy at which a drastic change in the "quasispecies", i.e., the stationary mutant distribution occurs. Thus, in frequently mutating populations fitness turns out to be an ensemble property rather than an attribute of the individual.

In addition we investigate the time dependence of the mean excess production as a function of initial conditions. Although it is optimized under most conditions, cases can be found which are characterized by decrease or non-monotonous change in mean excess productions.

1. Introduction. Recent data from populations of RNA viruses provided direct evidence for vast sequence heterogeneity (Domingo et al., 1987). The origin of this diversity is not yet completely known. It may be caused by the low replication accuracy of the polymerizing enzyme, commonly a virus-specific, RNA dependent RNA synthetase, or it may be the result of a high degree of selective neutrality of polynucleotide sequences. Eventually, both factors contribute to the heterogeneity observed. Indeed, mutations occur much more frequently than previously assumed in microbiology. They are by no means rare events and hence, neither the methods of conventional population genetics (Ewens, 1979) nor the neutral theory (Kimura, 1983) can be applied to these virus populations. Selectively neutral variants may be close with respect to Hamming distance and then the commonly made assumption that the mutation backflow from the mutants to the wild type is negligible does not apply.

A kinetic theory of polynucleotide evolution which was developed during the past 15 years (Eigen, 1971; 1985; Eigen and Schuster, 1979; Eigen et al., 1987; Schuster, 1986); Schuster and Sigmund, 1985) treats correct replication and mutation as parallel reactions within one and the same reaction network.
Neutral network

\[ \lambda = 0.10, \quad s = 367 \]

Error threshold: Individual sequences

\[ n = 10, \quad \sigma = 1.1, \quad d = 1.0 \]
Error threshold: Individual sequences

\[ n = 10, \sigma = 1.1, d = 1.0 \]
Neutral networks with increasing $\lambda$

$n = 10, \sigma = 1.1, d = 1.0$
Neutral networks with increasing $\lambda$

$N = 7$

Neutral network

$\lambda = 0.10, \ s = 229$
$N = 24$

Neutral networks with increasing $\lambda$
$N = 68$

Neutral networks with increasing $\lambda$

$\lambda = 0.20, \ s = 229$
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. Realistic fitness landscapes
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
Continuity in Evolution: On the Nature of Transitions
Water Fortana and Peter Schuster

To distinguish continuity from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. Our new approach is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and providing a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (1). One goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (2). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse to the level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and allow for the biological and biochemical importance from being a scaffold for the tertiary structure. For a goal of this type, we shall refer to secondary structure as "shapes." RNA crystals in a single molecule both genotype (replacement sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (3, 4).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (5, 6). In the laboratory, a goal might be to find a RNA sequence binding specifically to a molecule (4). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection environments. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target structure (3, 7). Here, a canonical DNA shape is used to represent the average distance to the target (visually related to fitness). Under these conditions, the population against time. Aside from a short initial phase, the entire history is dominated by a single shape, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target sequence (7). However, the dominant shapes in the population not only change at these marked events but undergo several changes in shape during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive shapes itself can be called continuous or discontinuous.

A set of entries is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The tree arises because, in contrast to sequences, there are

Continuity in Evolution: On the Nature of Transitions
Water Fortana and Peter Schuster

To distinguish continuity from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. Our new approach is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and providing a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (1). One goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (2). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse to the level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and allow for the biological and biochemical importance from being a scaffold for the tertiary structure. For a goal of this type, we shall refer to secondary structure as "shapes." RNA crystals in a single molecule both genotype (replacement sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (3, 4).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (5, 6). In the laboratory, a goal might be to find a RNA sequence binding specifically to a molecule (4). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection environments. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target structure (3, 7). Here, a canonical DNA shape is used to represent the average distance to the target (visually related to fitness). Under these conditions, the population against time. Aside from a short initial phase, the entire history is dominated by a single shape, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target sequence (7). However, the dominant shapes in the population not only change at these marked events but undergo several changes in shape during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive shapes itself can be called continuous or discontinuous.

A set of entries is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The tree arises because, in contrast to sequences, there are...
Replication rate constant:
\[ f_k = \gamma / [\alpha + \Delta d_S^{(k)}] \]
\[ \Delta d_S^{(k)} = d_H(S_k, S_\tau) \]

Selection constraint:
Population size, \( N = \# \) RNA molecules, is controlled by the flow
\[ N(t) \approx \overline{N} \pm \sqrt{N} \]

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
In silico optimization in the flow reactor: Evolutionary Trajectory
28 neutral point mutations during a long quasi-stationary epoch

Transition inducing point mutations change the molecular structure
Neutral point mutations leave the molecular structure unchanged

Neutral genotype evolution during phenotypic stasis
A sketch of optimization on neutral networks
1. Replication and selection
2. Mutation, quasispecies and error thresholds
3. Sequences, structures and neutrality
4. Realistic fitness landscapes
5. Replicating networks
6. RNA structure optimization
7. Experiments with RNA
Evolutionary design of RNA molecules

D.B.Bartel, J.W.Szostak, **In vitro selection of RNA molecules that bind specific ligands.** Nature **346** (1990), 818-822

C.Tuerk, L.Gold, **SELEX - Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.** Science **249** (1990), 505-510

D.P.Bartel, J.W.Szostak, **Isolation of new ribozymes from a large pool of random sequences.** Science **261** (1993), 1411-1418


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

An example of selection of molecules with predefined properties in laboratory experiments
The SELEX-technique for evolutionary design of strongly binding molecules called aptamers.
Formation of secondary structure of the tobramycin binding RNA aptamer with $K_D = 9$ nM

The three-dimensional structure of the tobramycin aptamer complex

Application of molecular evolution to problems in biotechnology
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 prototype 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 catalyze 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 same folded conformations can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to such sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very 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 (2). Because these disparities isolate the same fold and function, it is thought that they descended from a common ancestor through a series of neutral variants that were each 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 comprises a “neutral network,” connecting in sequence space 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 (3-5).

By using algorithms for predicting RNA secondary structure, we have suggested that different neutral networks are interconnected and can approach each other very closely (3, 5-9). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. 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 intersecting networks, RNAs with novel structures and activities would arise from previously existing ribozymes, without the need to catalyze nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing 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 (9). It joins an oligonucleotide substrate at its 3′ terminus. The prototype ligation sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 50 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length ligation (10). The HDV ribozyme carries out the site-specific self-cleaving 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 antigenic HDV ribozyme (12), which undergoes self-cleavage at a rate similar to that reported for other antigenic constructs (13, 14).

The prototype class III and HDV ribozymes have no more than the 35% sequence similarity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements.
Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (A) and a natural cleavage ribozyme of hepatitis-$\delta$-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
Acknowledgement of support

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

Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF)
Project No. Mat05

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

European Commission: Contracts No. 98-0189, 12835 (NEST)

Austrian Genome Research Program – GEN-AU: Bioinformatics
Network (BIN)

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute
Coworkers

Peter Stadler, Bärbel M. Stadler, Universität Leipzig, GE

Paul E. Phillipson, University of Colorado at Boulder, CO

Heinz Engl, Philipp Kügler, James Lu, Stefan Müller, RICAM Linz, AT

Jord Nagel, Kees Pleij, Universiteit Leiden, NL

Walter Fontana, Harvard Medical School, MA

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

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

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

Kurt Grünberger, Michael Kospach, Andreas Wernitznig, Stefanie Widder, Stefan Wuchty, Universität Wien, AT

Jan Cupal, Stefan Bernhart, Lukas Endler, Ulrike Langhammer, Rainer Machine, Ulrike Mückstein, Hakim Tafer, Thomas Taylor, Universität Wien, AT
Web-Page for further information:

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