Biomolecular Structure and Evolution

From Theory to the Design of RNA Molecules

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria

and

The Santa Fe Institute, Santa Fe, New Mexico, USA

EXYSTENCE Course on Complexity in Real-World Systems

ISI, Torino, 13.– 25.03.2006

Web-Page for further information:

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

Genotype, Genome

Collection of genes

Developmental program

Unfolding of the genotype

Highly specific environmental conditions

Phenotype

Evolution explains the origin of species and their interactions

Genotype, Genome

GCGGATTTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCCTGTGTTCGATCCACAGAATTCGCACCA

Omics

'the new biology is the chemistry of living matter'

**Biochemistry
molecular biology
structural biology
molecular evolution
molecular genetics
systems biology
bioinfomatics**

Unfolding of the genotype

Highly specific environmental conditions

Phenotype

John Kendrew
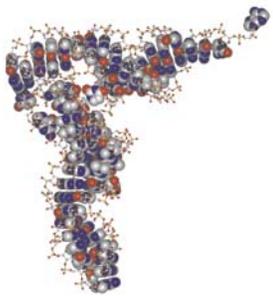
evolution of RNA molecules, ribozymes and splicing, the idea of an RNA world, selection of RNA molecules, RNA editing, the ribosome is a ribozyme, small RNAs and RNA switches.

The exciting RNA story
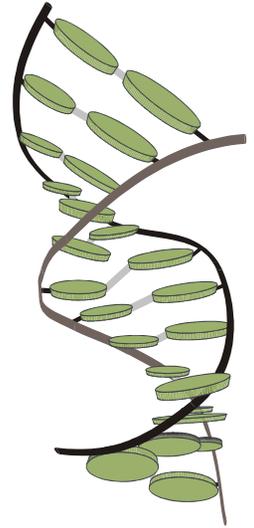
Manfred Eigen

Molecular evolution
Linus Pauling and
Emile Zuckerkandl

Hemoglobin sequence
Gerhard Braunitzer

Max Perutz
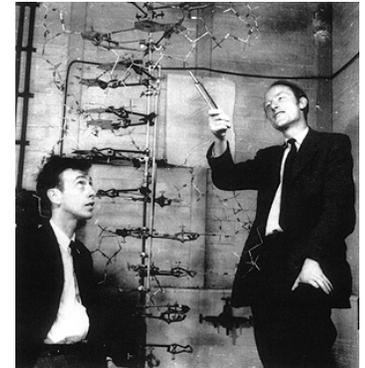
James D. Watson und
Francis H.C. Crick

Three necessary conditions for Darwinian evolution are:

1. **Multiplication**,

2. **Variation**, and

3. **Selection**.

Variation through mutation and recombination operates on the **genotype** whereas the **phenotype** is the target of **selection**.

One important property of the Darwinian scenario is that **variations** in the form of mutations or recombination events occur **uncorrelated** with their **effects on the selection process**.

All conditions can be fulfilled not only by cellular organisms but also by **nucleic acid molecules** in suitable **cell-free experimental assays**.

	Generation time	Selection and adaptation 10 000 generations	Genetic drift in small populations 10^6 generations	Genetic drift in large populations 10^7 generations
RNA molecules	10 sec	27.8 h = 1.16 d	115.7 d	3.17 a
	1 min	6.94 d	1.90 a	19.01 a
RNA viruses	5 min	34.72 d	9.51 a	95.13 a
Bacteria	20 min	138.9 d	38.03 a	380 a
	10 h	11.40 a	1 140 a	11 408 a
Multicelluar organisms	10 d	274 a	27 380 a	273 800 a
	20 a	200 000 a	2×10^7 a	2×10^8 a

1. Evolution experiments

2. Molecular evolution of RNA

3. Neutral networks

4. Evolutionary optimization of RNA structure

5. Kinetic structures and switching molecules

Bacterial Evolution

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants*. Science **272** (1996), 1802-1804

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812

lawn of E.coli

24 h

24 h

nutrient agar
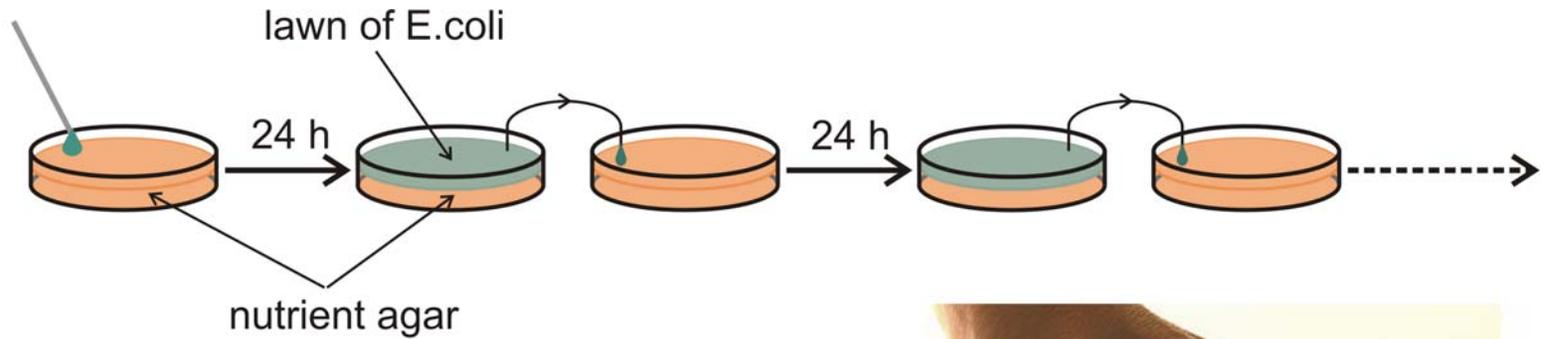
1 day   » 6.67 generations
1 month »   200 generations
1 year   » 2400 generations

Serial transfer of Escherichia coli
cultures in Petri dishes
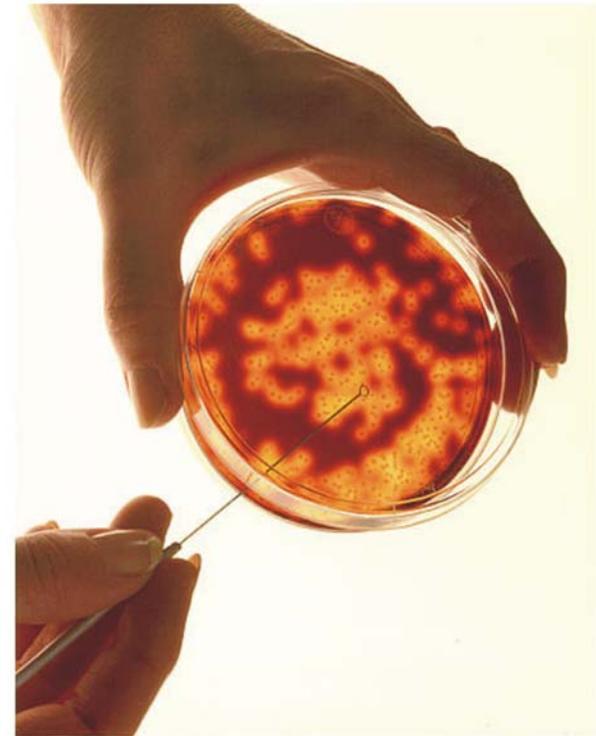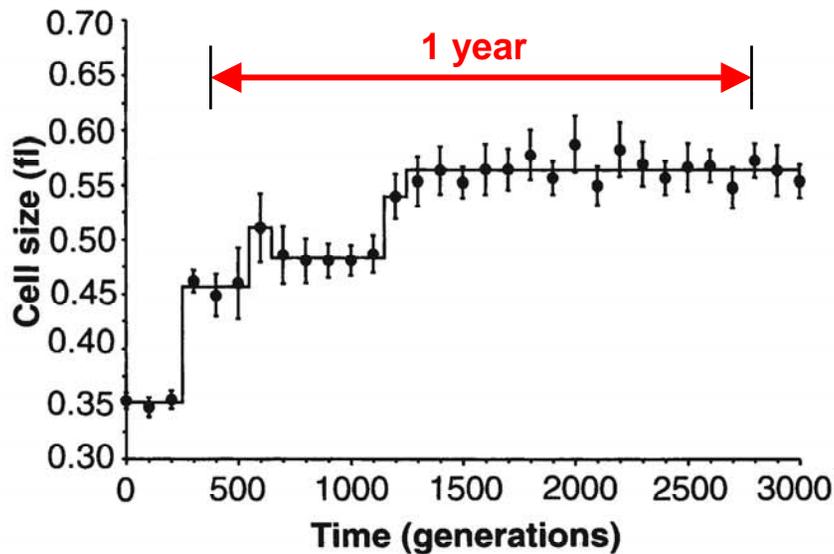
Fig. 1. Change in average cell size (1 fl = 10^{-15} L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (*22*). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these data (Table 1).
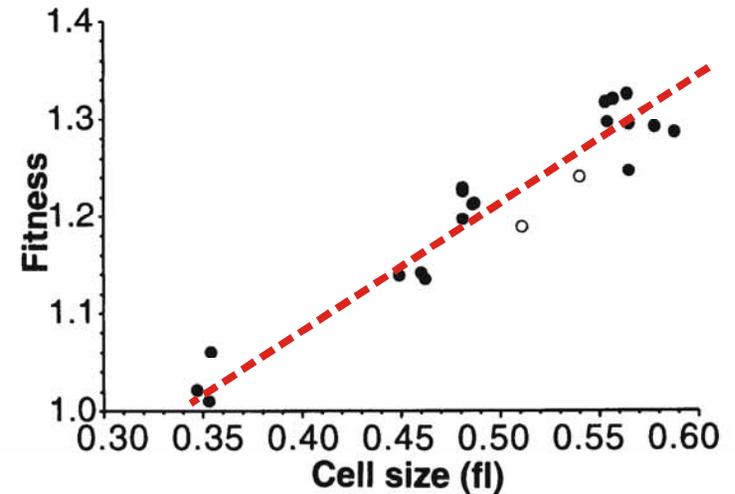
Fig. 2. Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (*6, 7*). The open symbols indicate 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

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants*. Science **272** (1996), 1802-1804
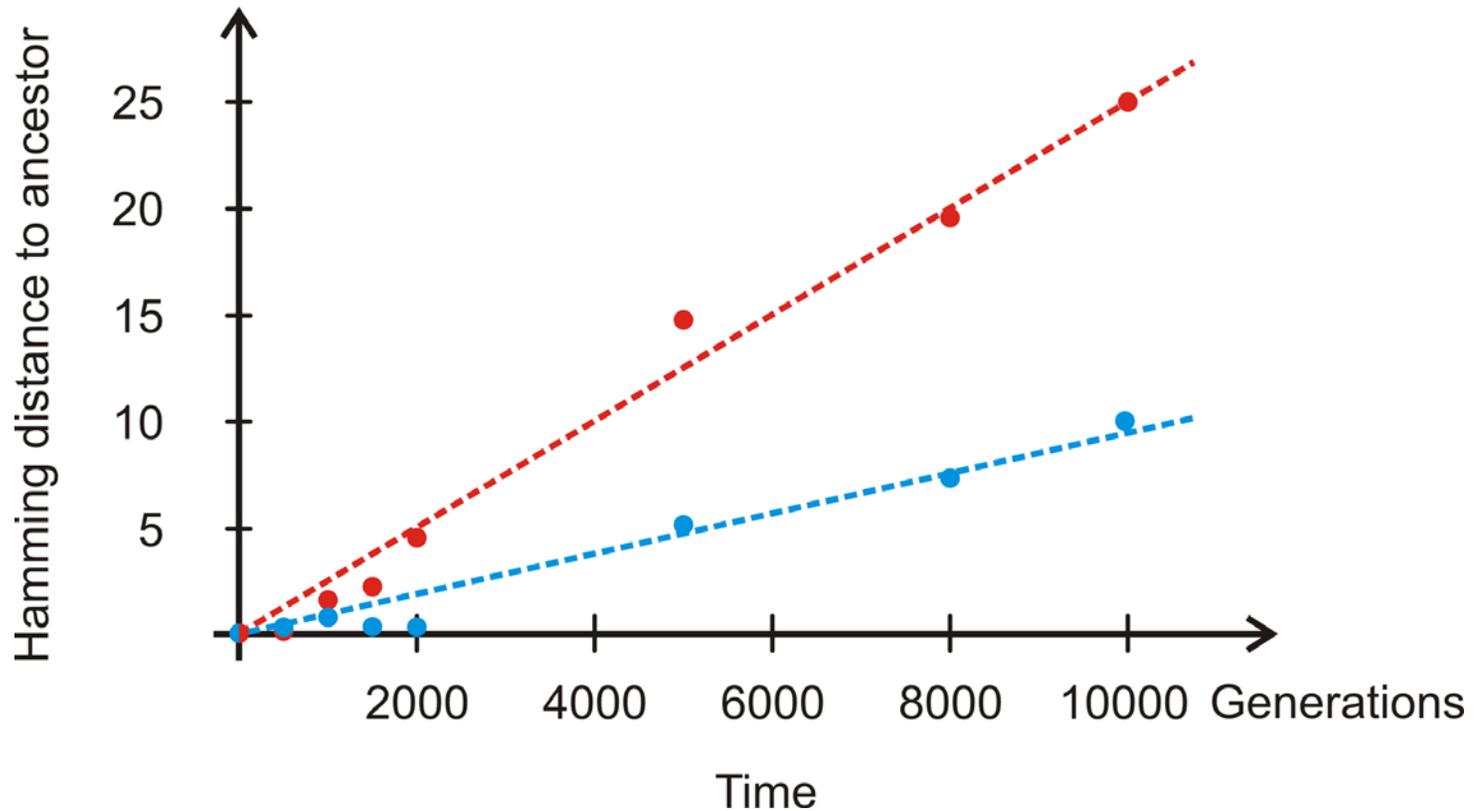
Variation of genotypes in a bacterial serial transfer experiment

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812

Evolution of RNA molecules based on Qβ phage

D.R.Mills, R,L,Peterson, S.Spiegelman, *An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule*. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224

S.Spiegelman, *An approach to the experimental analysis of precellular evolution*. Quart.Rev.Biophys. **4** (1971), 213-253

C.K.Biebricher, *Darwinian selection of self-replicating RNA molecules*. Evolutionary Biology **16** (1983), 1-52

C.K.Biebricher, W.C. Gardiner, *Molecular evolution of RNA* **in vitro**. Biophysical Chemistry **66** (1997), 179-192

G.Strunk, T. Ederhof, *Machines for automated evolution experiments* **in vitro** *based on the serial transfer concept*. Biophysical Chemistry **66** (1997), 193-202
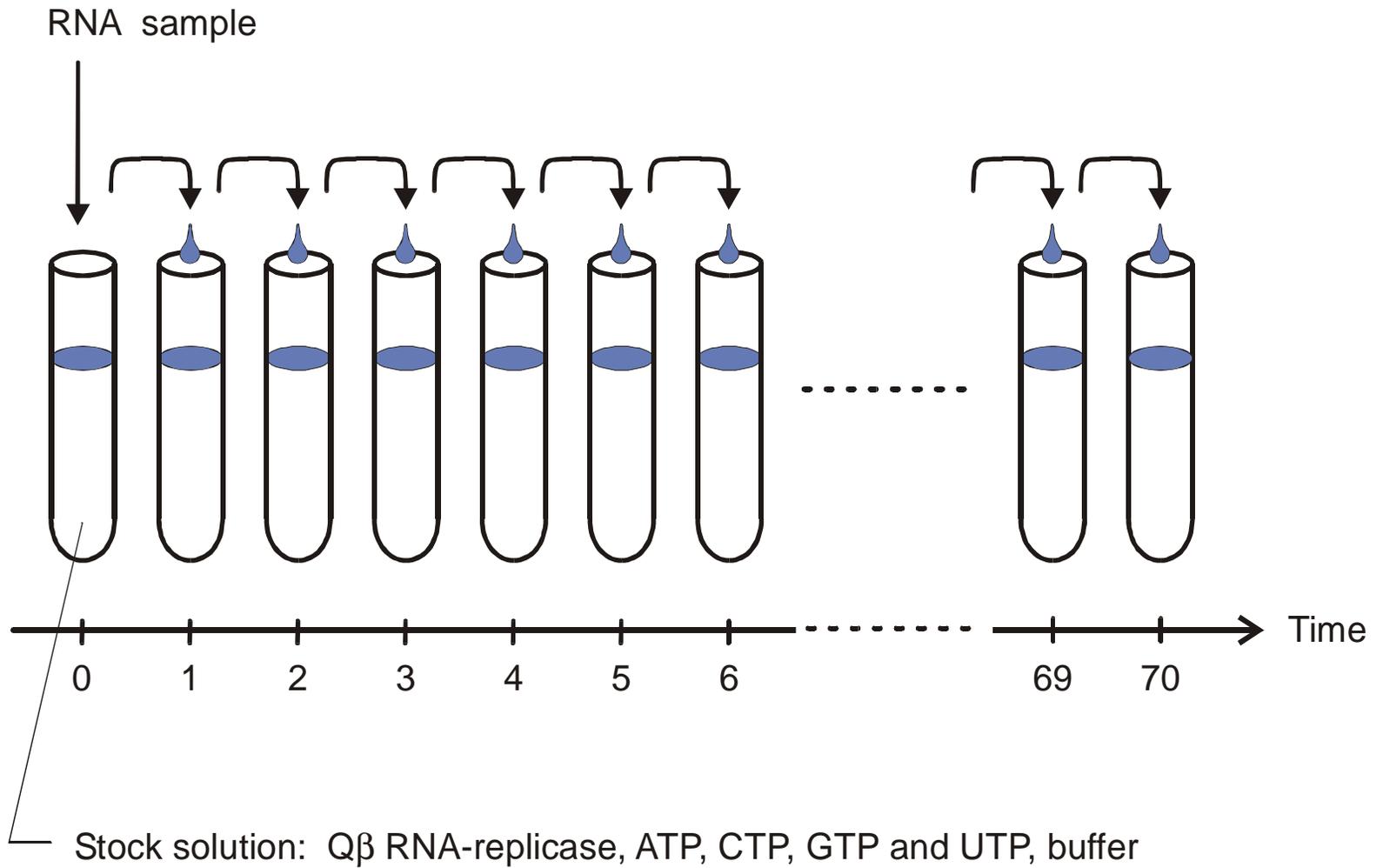
The serial transfer technique applied to RNA evolution *in vitro*

Reproduction of the original figure of the serial transfer experiment with Qβ RNA

D.R.Mills, R,L,Peterson, S.Spiegelman, *An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule*. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224
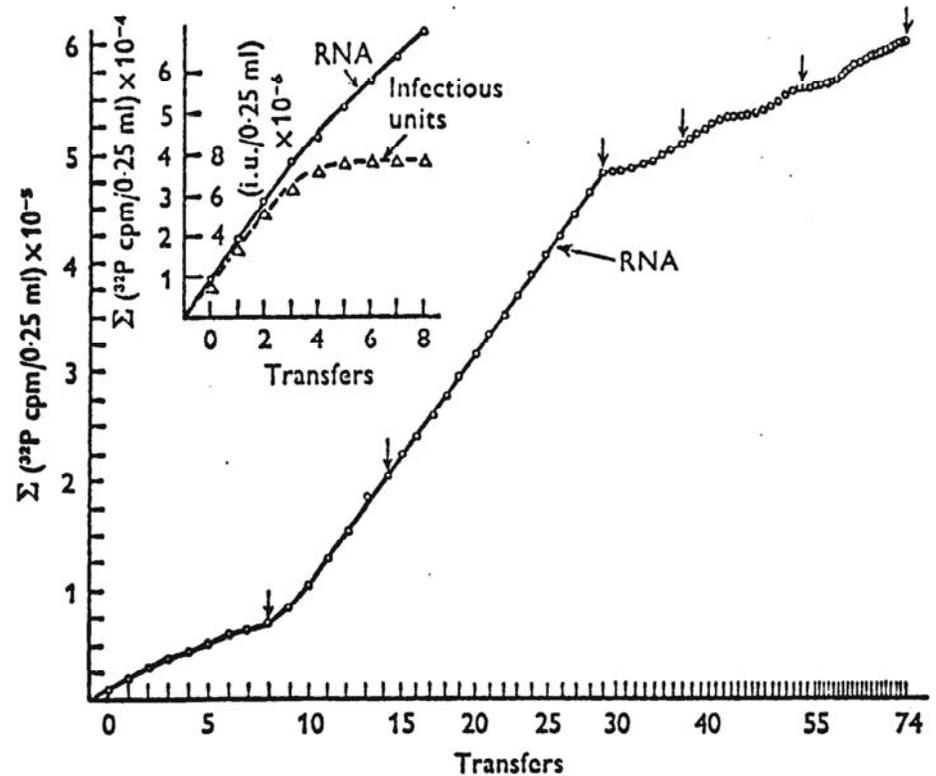
Fig. 9. Serial transfer experiment. Each 0·25 ml standard reaction mixture contained 40 μg of Qβ replicase and ³²P-UTP. The first reaction (0 transfer) was initiated by the addition of 0·2 μg ts-1 (temperature-sensitive RNA) and incubated at 35 °C for 20 min, whereupon 0·02 ml was drawn for counting 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 ceases to appear after the 4th transfer (Mills *et al.* 1967).
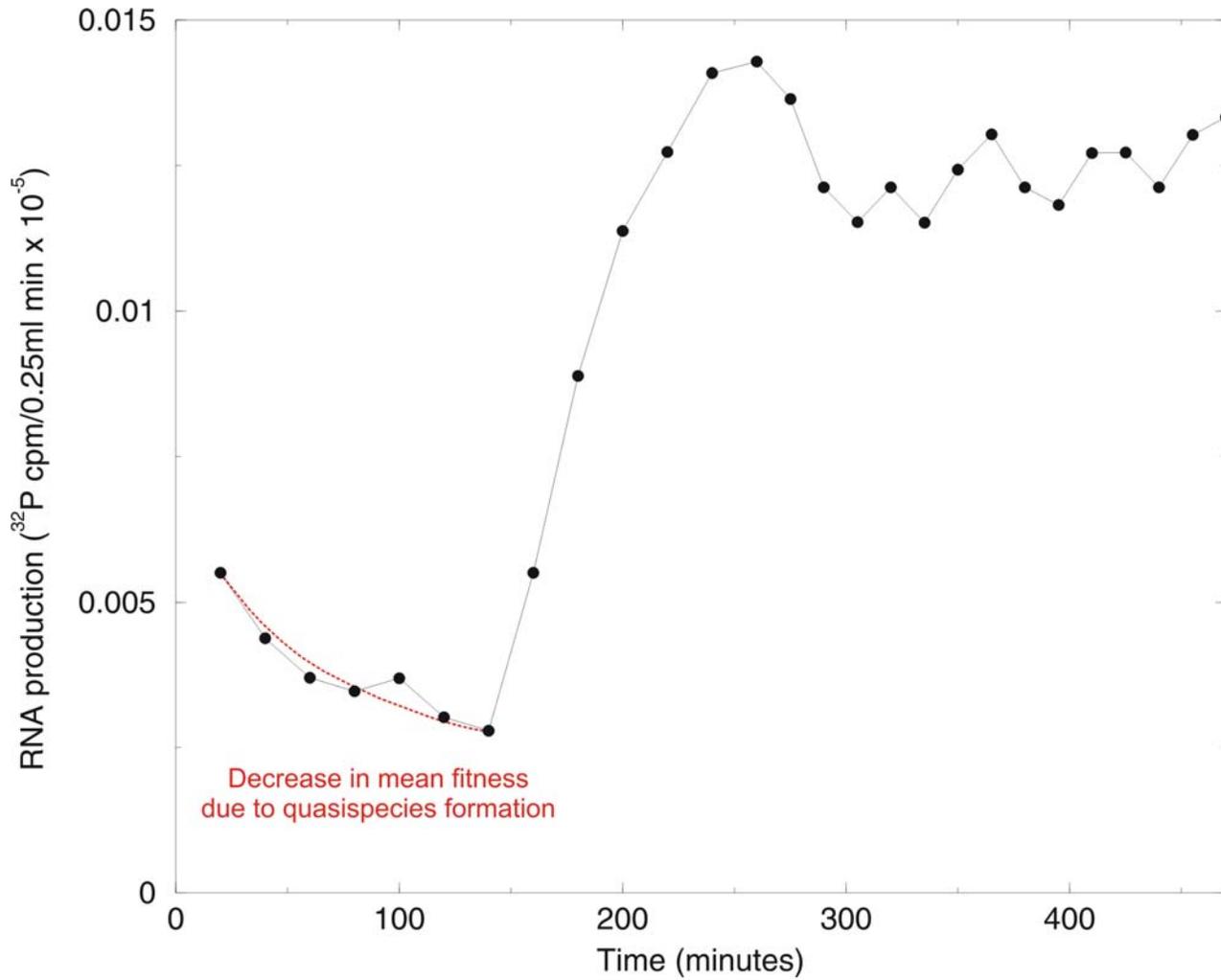
The increase in RNA production rate during a serial transfer experiment

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

R.D.Jenison, S.C.Gill, A.Pardi, B.Poliski, *High-resolution molecular discrimination by RNA*. Science **263** (1994), 1425-1429

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

Jiang, A. K. Suri, R. Fiala, D. J. Patel, *Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex*. Chemistry & Biology **4** (1997), 35-50

Amplification
Diversification

Selection cycle

Genetic
Diversity

Selection

Desired Propeties
? ? ?

No   Yes
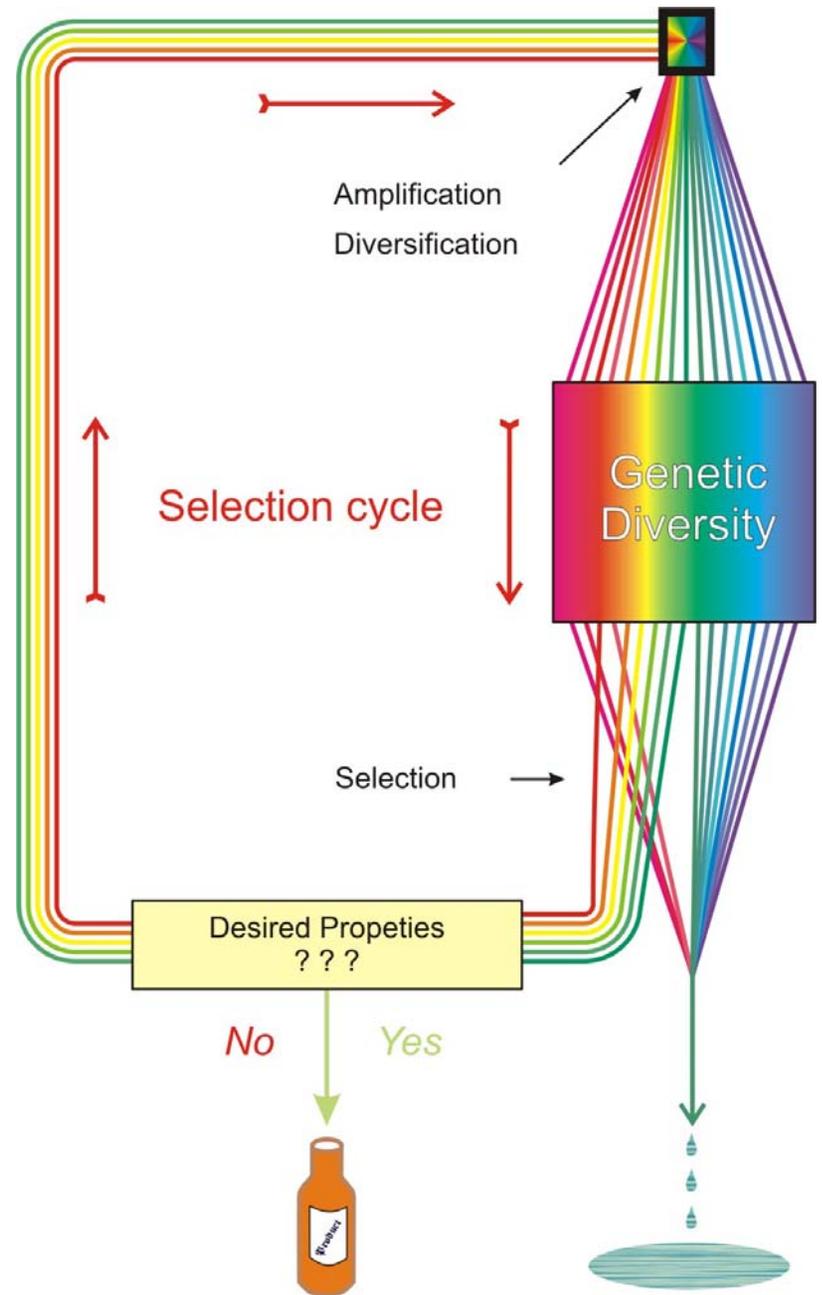
An example of 'artificial selection' with RNA molecules or 'breeding' of biomolecules

chromatographic column
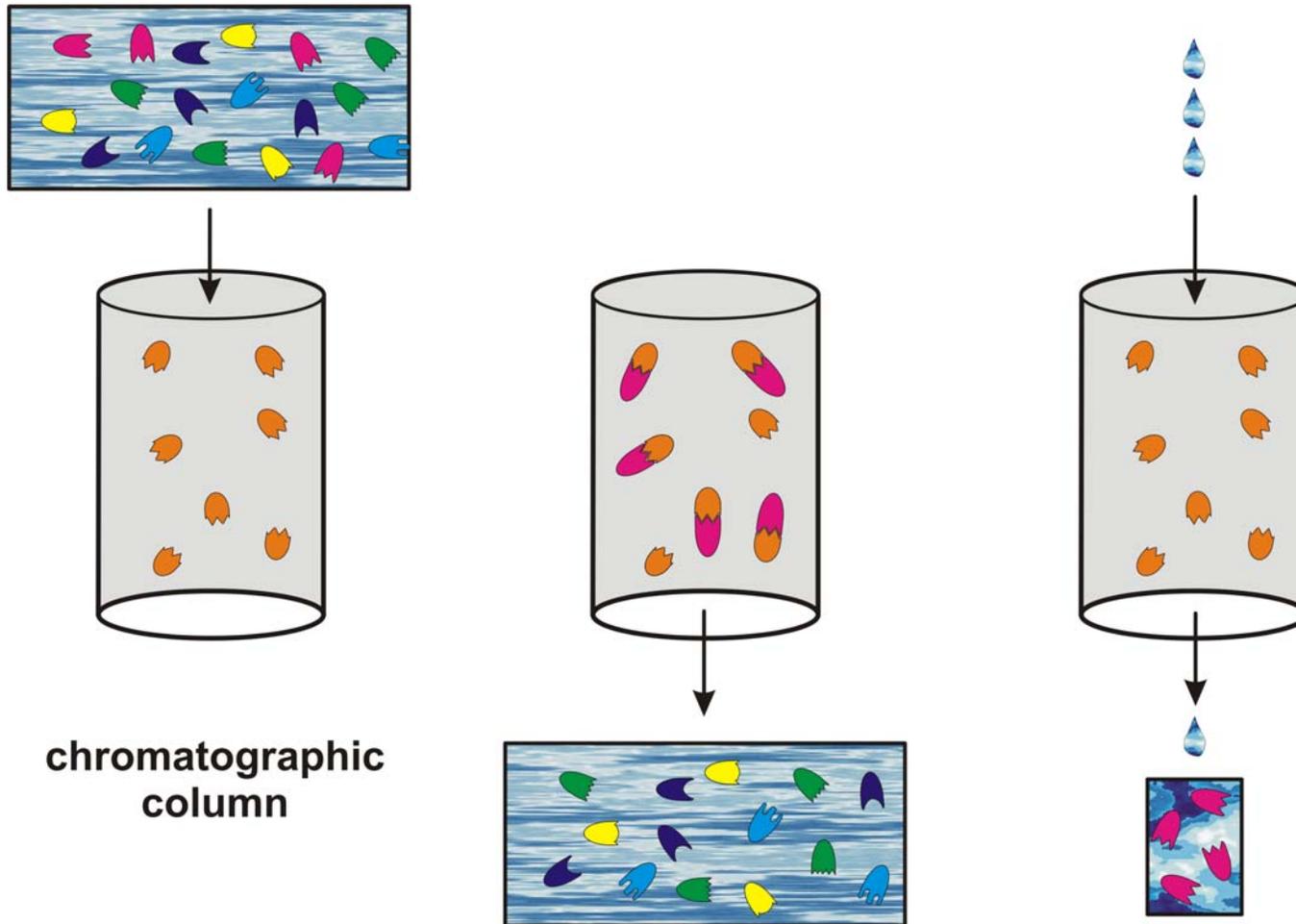
The SELEX technique for the evolutionary preparation of aptamers
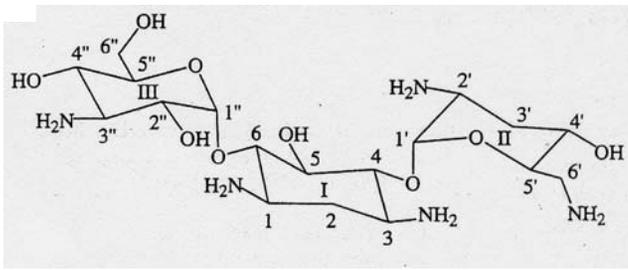
tobramycin
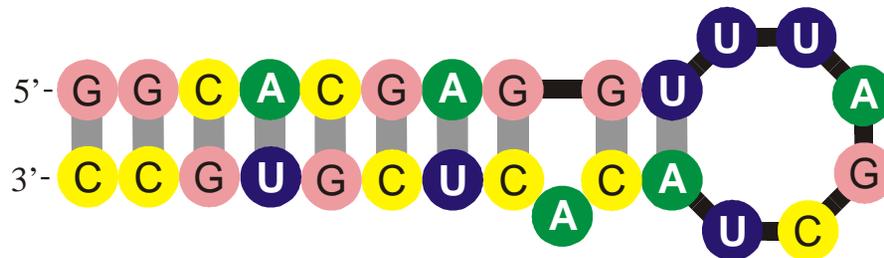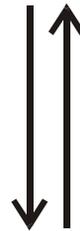
RNA aptamer

Formation of secondary structure of the tobramycin binding RNA aptamer with $K_D = 9$ nM

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, *Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex.* Chemistry & Biology **4**:35-50 (1997)
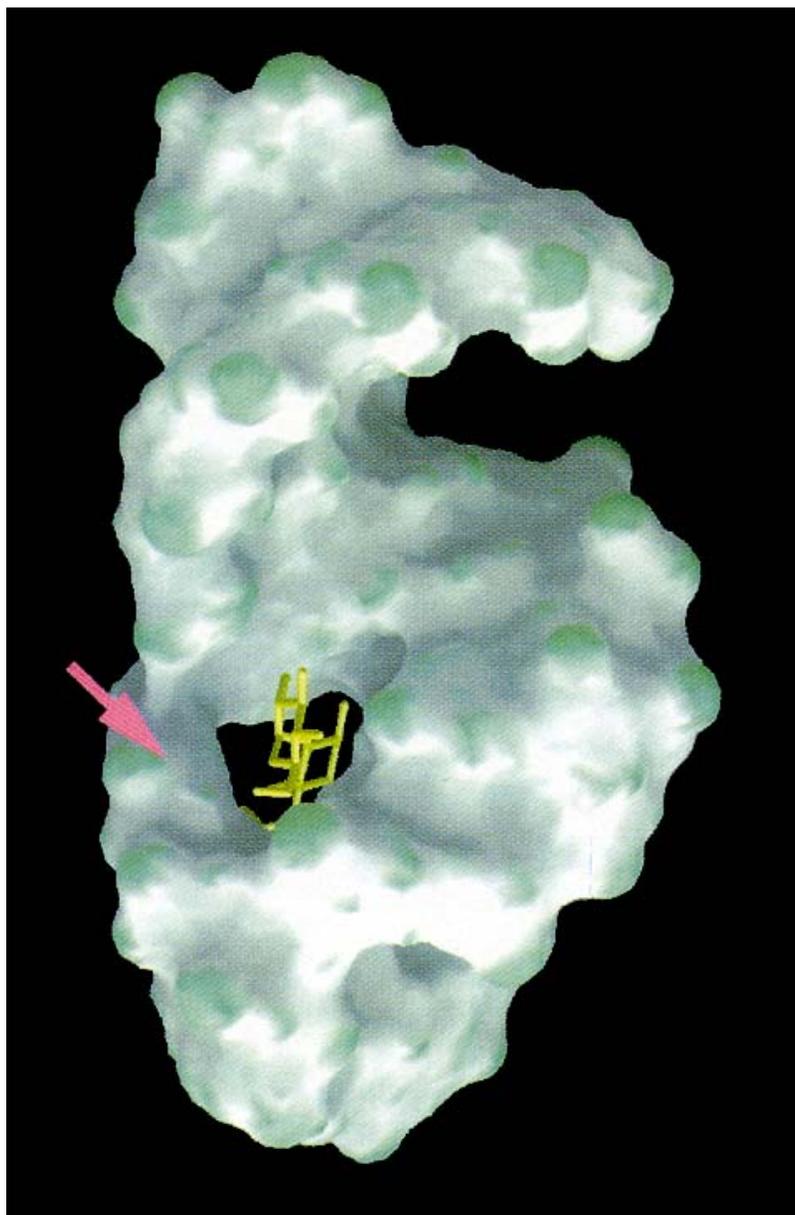
The three-dimensional structure of the tobramycin aptamer complex

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, Chemistry & Biology **4**:35-50 (1997)

Plus strand
A U G G U A C A U C A U G A   C U U G

Template induced synthesis

Plus strand
A U G G U A C A U C A U G A   C U U G
U A C C A U
Minus strand   G

Template induced synthesis

Plus strand
A U G G U A C A U C A U G A   C U U G
U A C C A U G U A G U A C U   G A A C
Minus strand

Complex dissociation

Plus strand
A U G G U A C A U C A U G A   C U U G
+
U A C C A U G U A G U A C U   G A A C
Minus strand
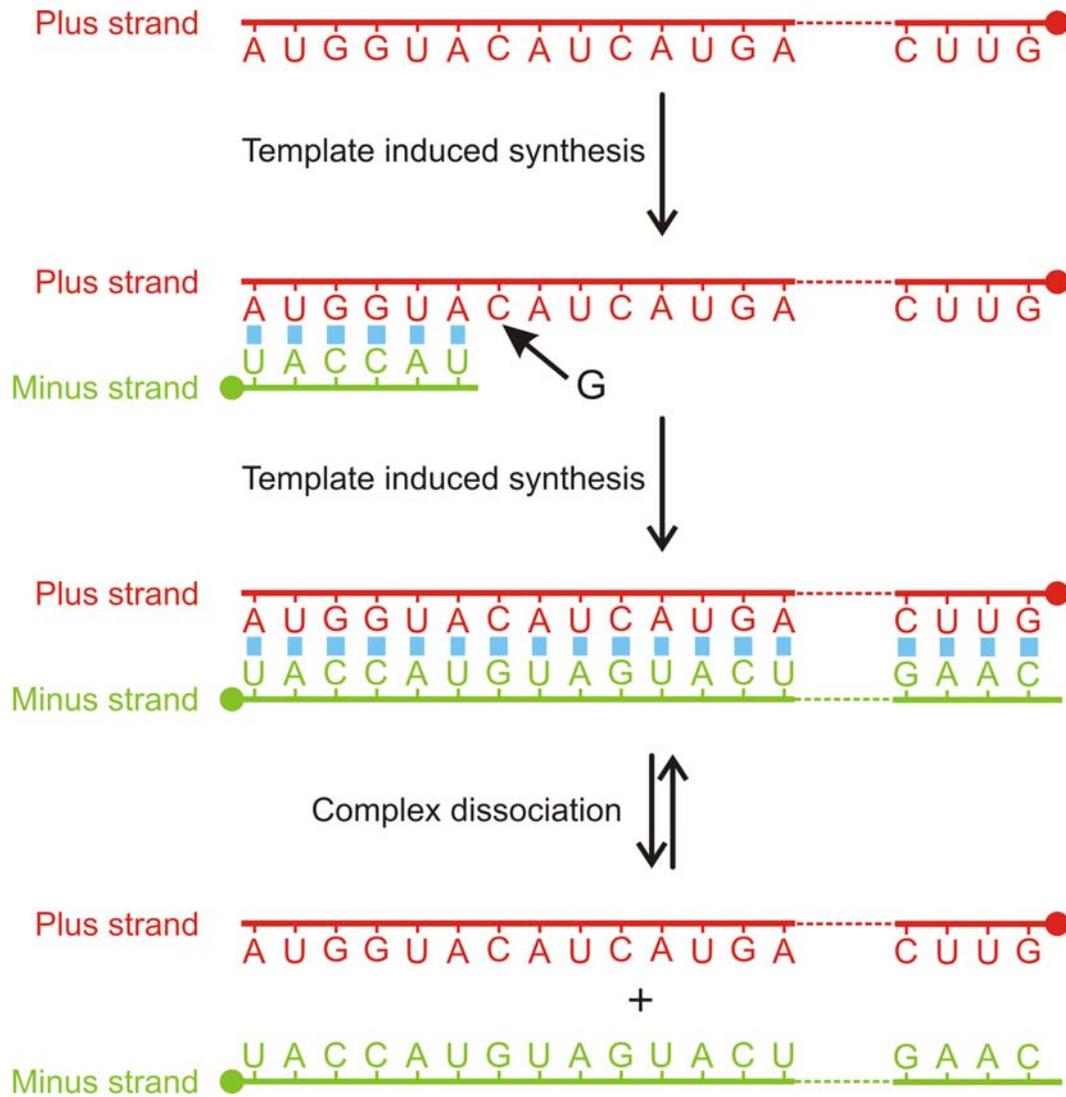
Complementary replication is the simplest copying mechanism of RNA.
Complementarity is determined by Watson-Crick base pairs:

G≡C and **A=U**

$$(A) + I_1 \xrightarrow{\ f_1\ } I_2 + I_1$$

$$(A) + I_2 \xrightarrow{\ f_2\ } I_1 + I_2$$
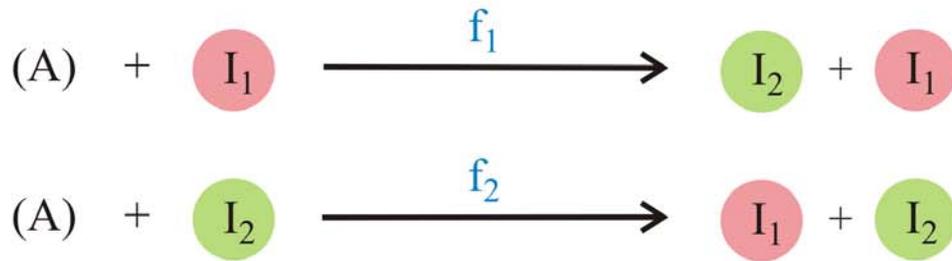
$$dx_1 / dt = f_2 x_2 - x_1 \Phi$$
$$dx_2 / dt = f_1 x_1 - x_2 \Phi$$

$$\Phi = \Sigma_i\, f_i\, x_i\ ;\quad \Sigma_i\, x_i = 1\ ;\quad i = 1,2$$

Complementary replication as the simplest molecular mechanism of reproduction

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

$$\frac{dx_1}{dt} = f_2 x_2 - x_1 \phi, \quad \frac{dx_2}{dt} = f_1 x_1 - x_2 \phi; \quad x_1 + x_2 = 1; \quad \phi = f_1 x_1 + f_2 x_2 = \overline{f}$$

Solutions are obtained by integrating factor transformation

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

$$\gamma_1(0) = \sqrt{f_1}\,x_1(0) + \sqrt{f_2}\,x_2(0), \gamma_2(0) = \sqrt{f_1}\,x_1(0) - \sqrt{f_2}\,x_2(0); \;\; f = \sqrt{f_1 f_2}$$

$$x_1(t) \rightarrow \frac{\sqrt{f_2}}{\sqrt{f_1} + \sqrt{f_2}} \;\; \text{and} \;\; x_2(t) \rightarrow \frac{\sqrt{f_1}}{\sqrt{f_1} + \sqrt{f_2}} \;\; \text{as} \;\; \exp(-ft) \rightarrow 0$$

$$(A) + I_1 \xrightarrow{f_1} I_1 + I_1$$

$$(A) + I_2 \xrightarrow{f_2} I_2 + I_2$$

$$(A) + I_i \xrightarrow{f_i} I_i + I_i$$

$$(A) + I_m \xrightarrow{f_m} I_m + I_m$$

$$(A) + I_n \xrightarrow{f_n} I_n + I_n$$
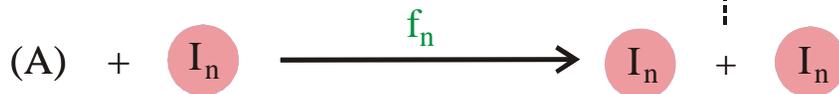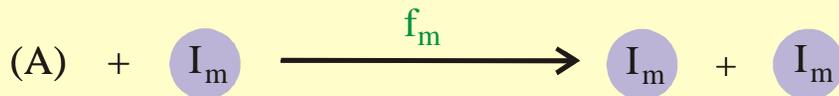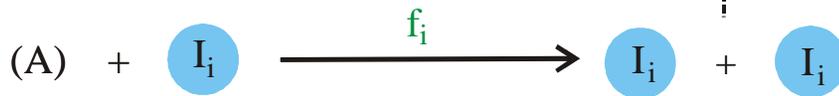
$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

$$\Phi = \Sigma_j f_j x_j ; \quad \Sigma_j x_j = 1 ; \quad i,j = 1,2,\dots,n$$

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

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

$$f_m = \max \{f_j; j=1,2,\dots,n\}$$

$$x_m(t) \to 1 \text{ for } t \to \infty$$

Reproduction of organisms **or replication** of molecules as the basis of selection

Selection equation: $[I_i] = x_i \geq 0$, $f_i > 0$

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

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

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

Solutions are obtained by integrating factor transformation

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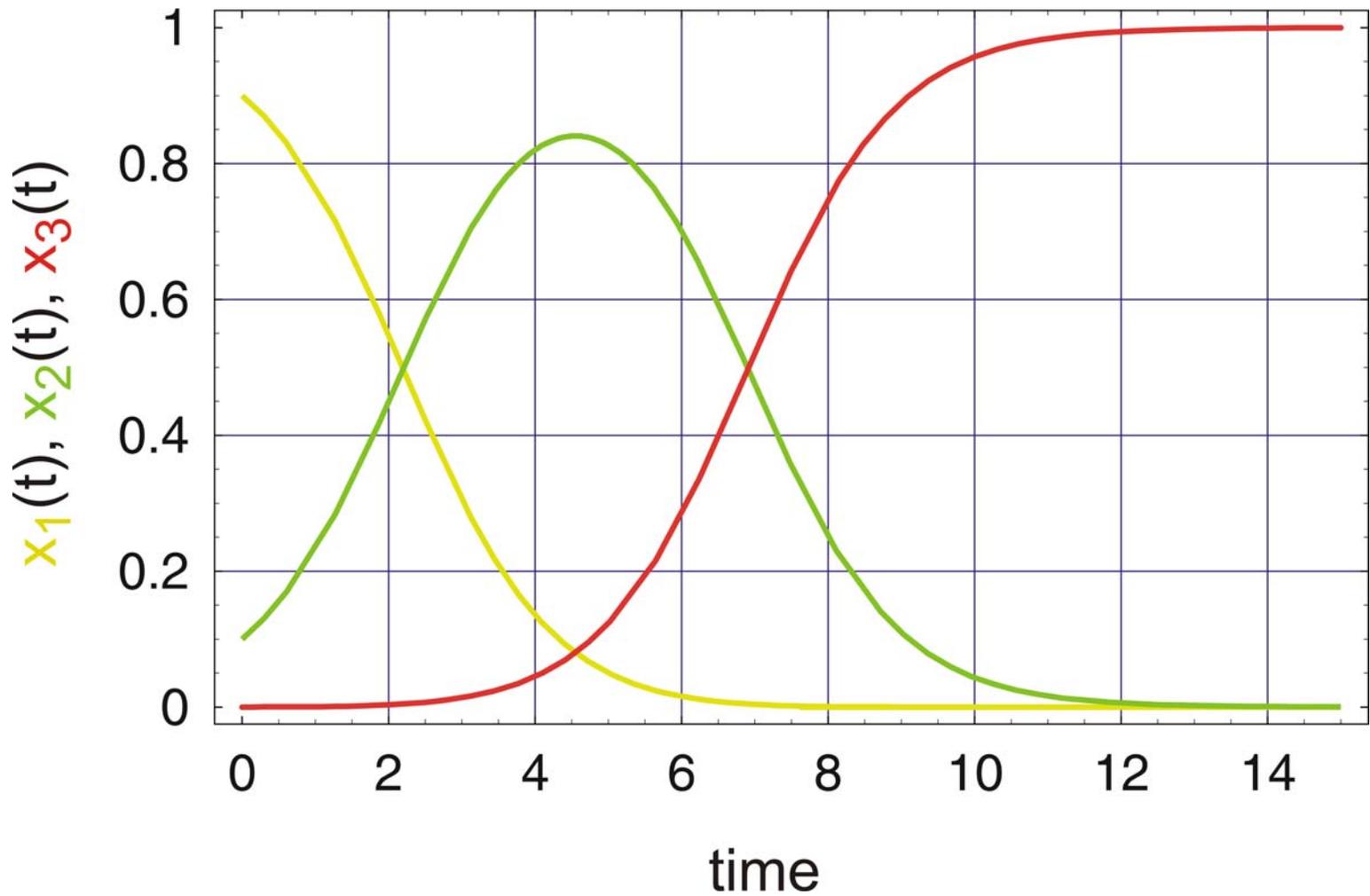
Selection between three species with $f_1 = 1, f_2 = 2$, and $f_3 = 3$

Changes in RNA sequences originate from replication errors called **mutations**.

Mutations occur uncorrelated to their consequences in the selection process and are, therefore, commonly characterized as **random elements** of evolution.

Point Mutation

GAA<mark>UCCCG</mark>AA → GAA<mark>UCCCGUCCCG</mark>AA

Insertion

GAAUCC<mark>CGA</mark>A → GAAUCCA
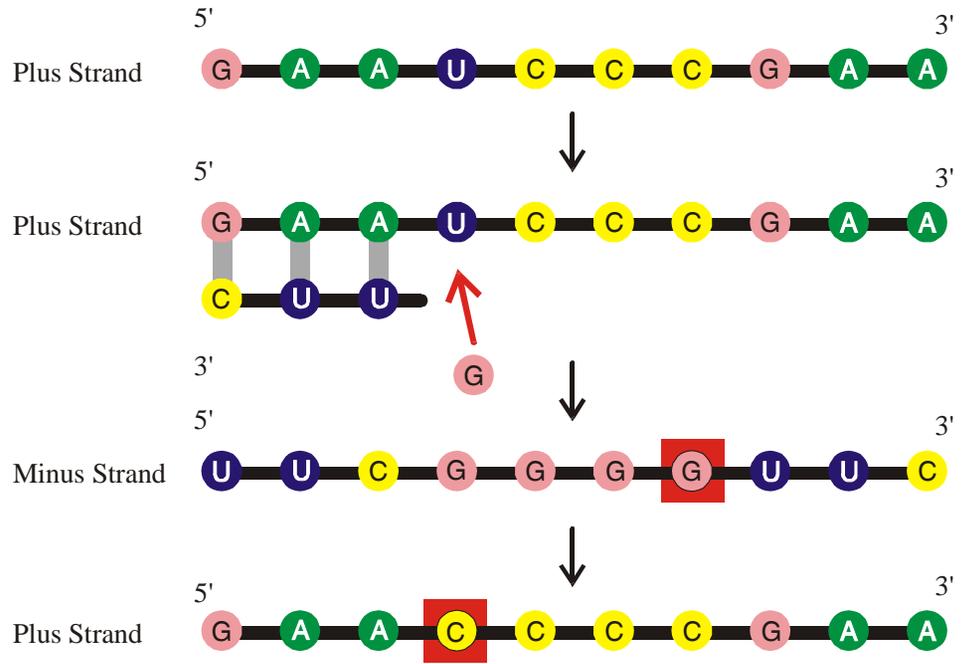
Deletion

The origins of changes in RNA sequences are **replication errors** called **mutations**.
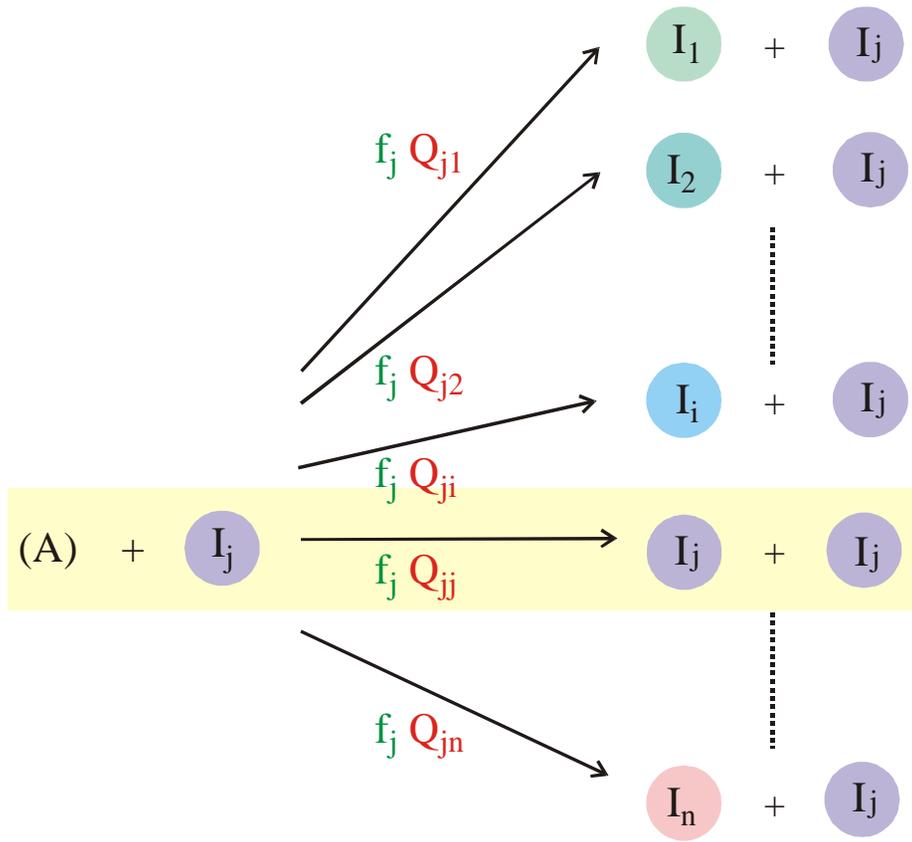
$$dx_i / dt = \Sigma_j\, f_j Q_{ji}\, x_j - x_i\, \Phi$$

$$\Phi = \Sigma_j\, f_j\, x_i\, ; \quad \Sigma_j\, x_j = 1\, ; \quad \Sigma_i\, Q_{ij} = 1$$

$[I_i] = x_i \geq 0\, ;\ i = 1, 2, ..., n\, ;$

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

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

p .... Error rate per digit

ℓ .... Chain length of the polynucleotide

d(i,j) .... Hamming distance between I_i and I_j

Chemical kinetics of replication and mutation as parallel reactions

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

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

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

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

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

$$L^{-1} \cdot W \cdot L \quad = \quad \Lambda \quad = \quad \left\{ \lambda_k;\ k = 0, 1, \cdots, n-1 \right\}$$

Master sequence

Concentration

Sequence

Space

Formation of a quasispecies
in sequence space

Master sequence

Mutant cloud

Concentration

Sequence

Space

Formation of a quasispecies
in sequence space

Concentration

Master sequence

Mutant cloud

Sequence

Space

Formation of a quasispecies
in sequence space

Master sequence

Mutant cloud

Concentration

Sequence

Space

Formation of a quasispecies
in sequence space

Concentration

Sequence

Space

Mutant cloud

Uniform distribution in sequence space

The error threshold in replication
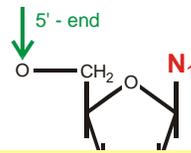
5'-end GCGGAU UUA GCUC AGUUGGGA GAGC GCCAGA CUGAAGA UCUGG AGGUC CUGUG UUCGAUC CACAG AAUUCGCACCA 3'-end
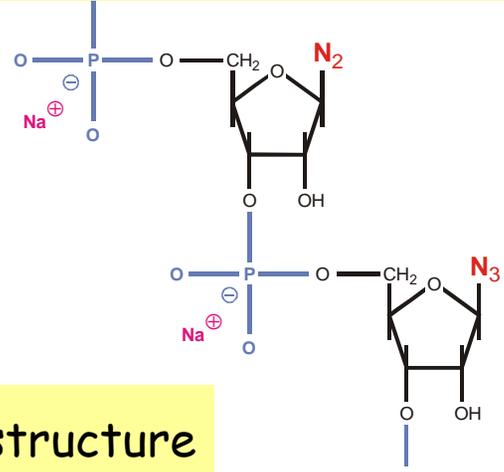
Definition of RNA structure

Sequence: 5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

$N = 4^n$

Secondary structure

3'-end

5'-end

70

60

50

10

20

30   40

Symbolic notation: 5'-End ((((((···((((········))))·(((((·······)))))····(((((·······))))))·)))))···· 3'-End

$N_S < 3^n$
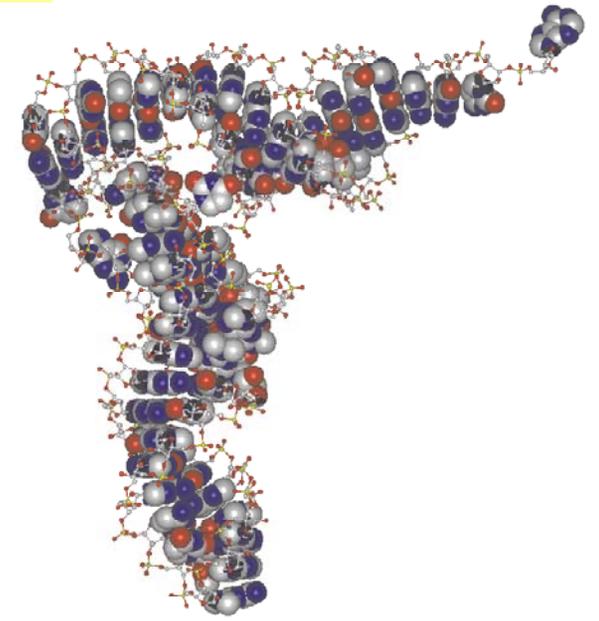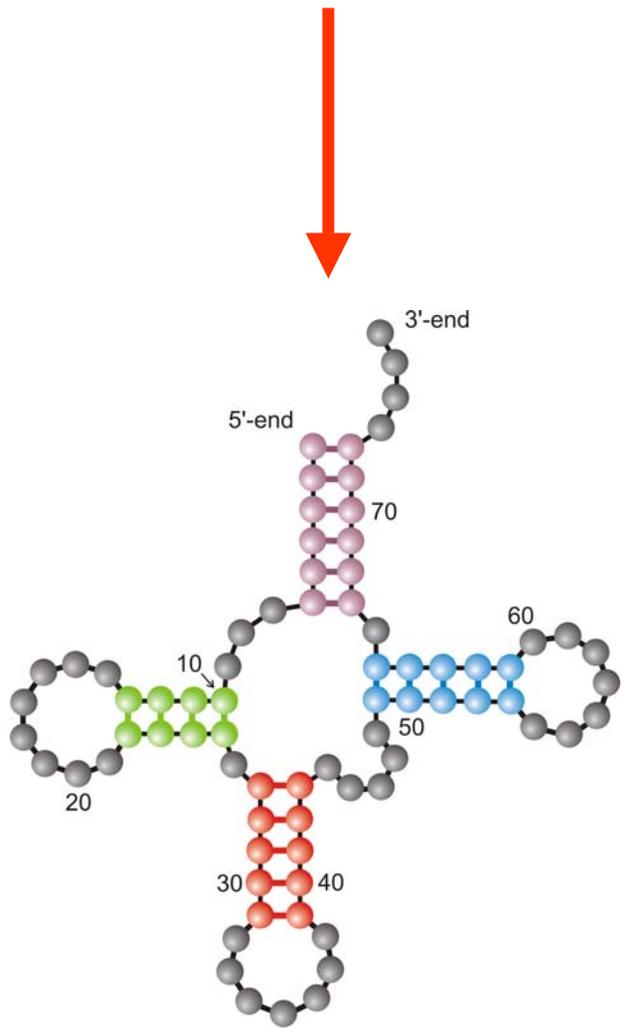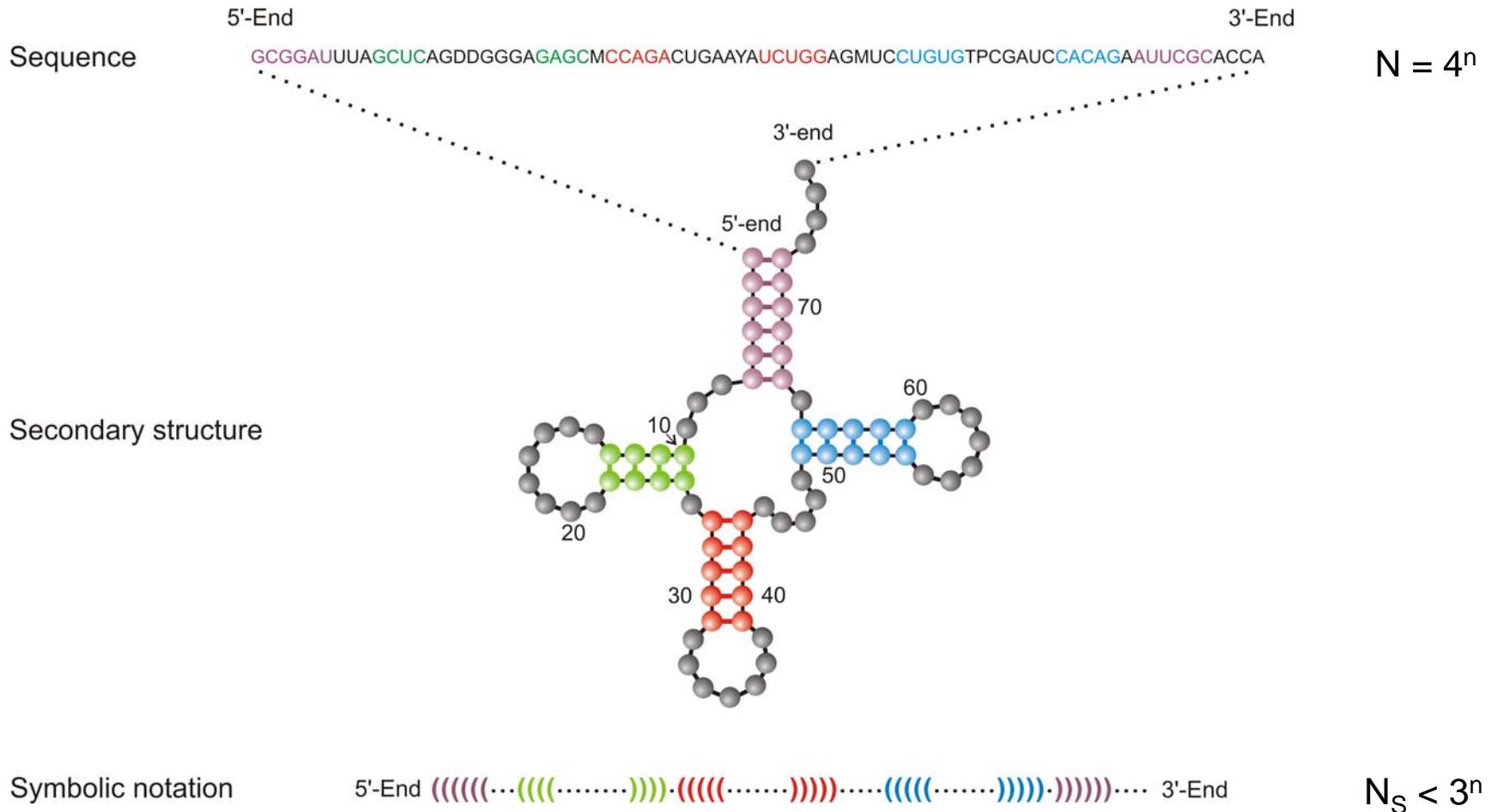
Criterion: Minimum free energy (mfe)

Rules: _ (_) _ ∈ {AU,CG,GC,GU,UA,UG}

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs

RNA sequence   GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

Structural biology,
spectroscopy of
biomolecules,
understanding
molecular function

Biophysical chemistry:
thermodynamics and
kinetics

Empirical parameters

RNA structure
of minimal free
energy
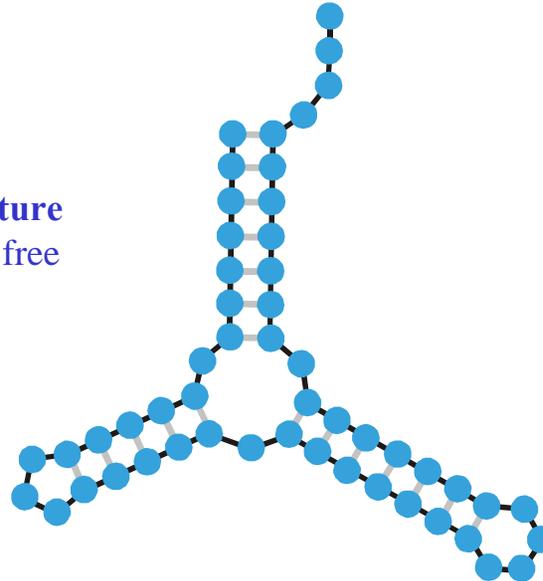
Sequence, structure, and design

5'-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

3'-end

$S_5^{(h)}$
$S_3^{(h)}$
$S_4^{(h)}$
$S_1^{(h)}$
$S_2^{(h)}$
$S_7^{(h)}$
$S_8^{(h)}$
$S_9^{(h)}$
$S_6^{(h)}$

Suboptimal conformations

Free energy ΔG^0

$S_0^{(h)}$

Minimum of free energy
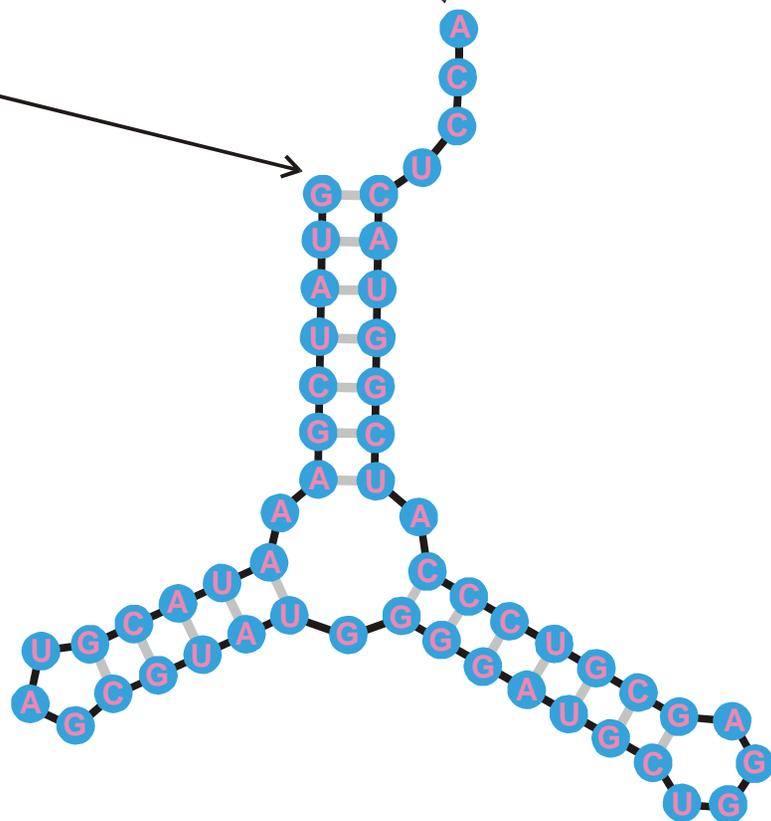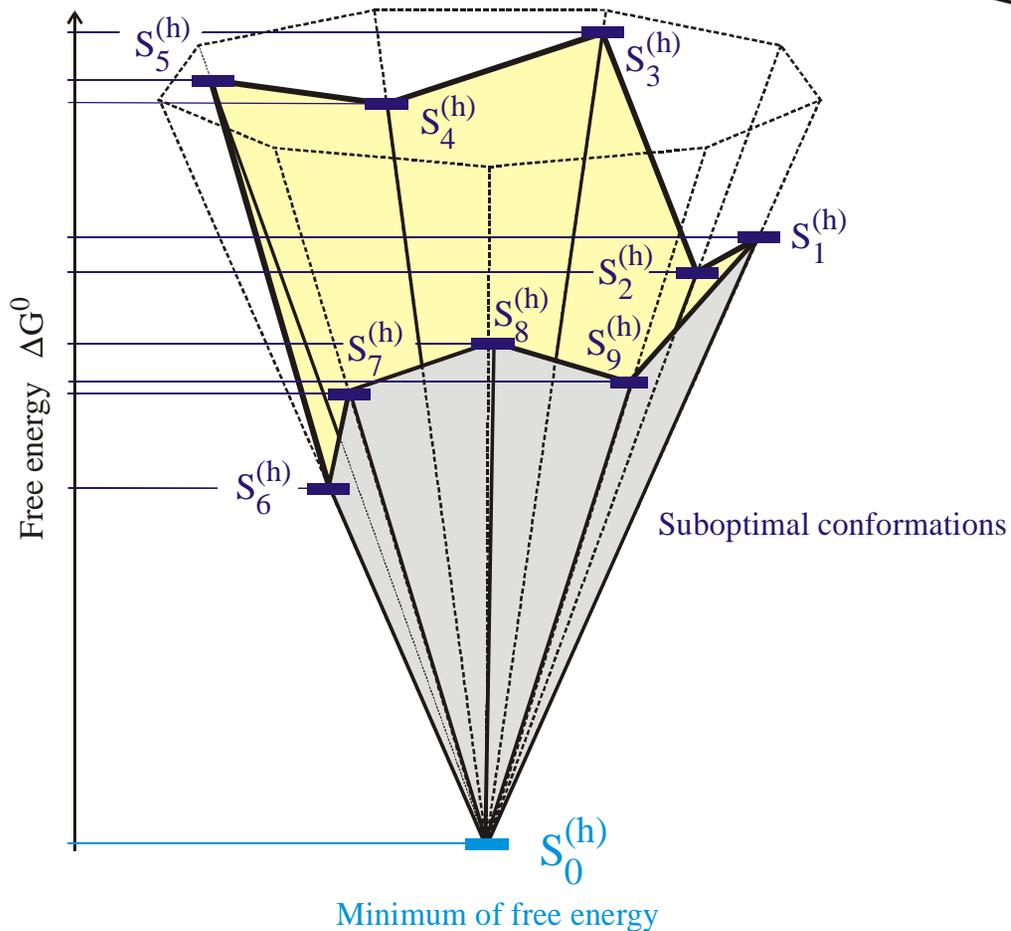
The minimum free energy structures on a discrete space of conformations

RNA sequence   GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

RNA folding:

Structural biology, spectroscopy of biomolecules, understanding **molecular function**

Inverse folding of RNA:

Biotechnology, **design of biomolecules** with predefined structures and functions

RNA structure of minimal free energy
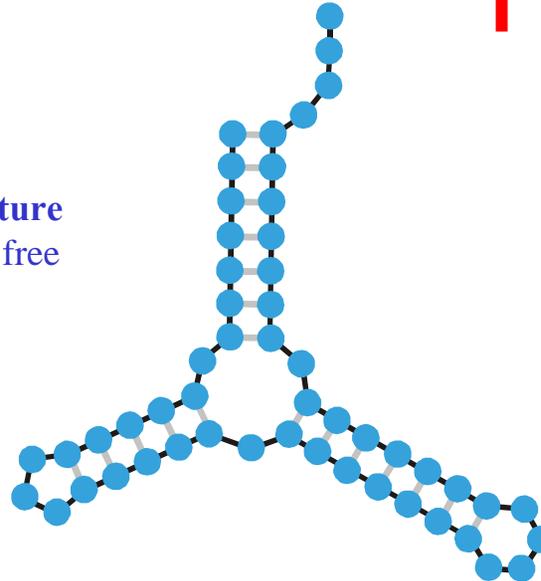
Sequence, structure, and design
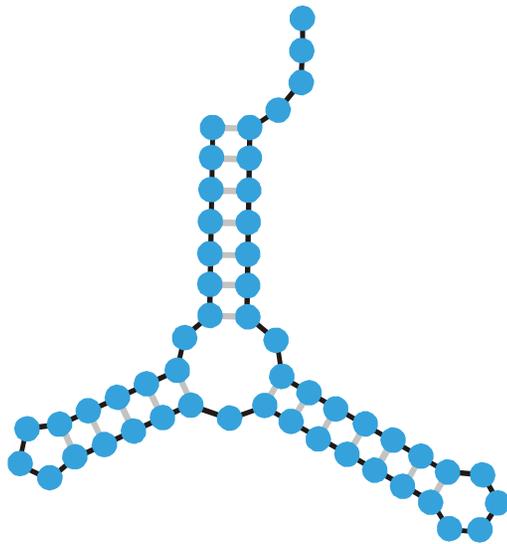
Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGGAGAUGGCCAUUGCAG

Inverse folding

The **inverse folding algorithm** searches for sequences that form a given
RNA secondary structure under the minimum free energy criterion.

Space of genotypes: $\mathcal{I} = \{I_1, I_2, I_3, I_4, \ldots, I_N\}$; Hamming metric

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

$$N \gg M$$

$$\psi(I_j) = S_k$$

$$\mathbf{G}_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

A mapping ψ and its inversion
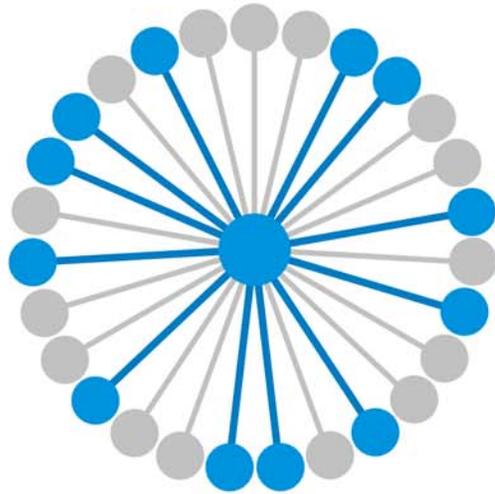
$$G_k = \psi^{-1}(S_k) \doteq \{\ I_j \mid \psi(I_j) = S_k\ \}$$

$$\bar{\lambda}_k = \frac{\underset{j \in |G_k|}{\Sigma}\ \lambda_j(k)}{|G_k|}$$

$$\lambda_j = \mathbf{12}\ /\ 27 = 0.444$$

Alphabet size κ :

κ	λ_{cr}	
2	0.5	**AU,GC,DU**
3	0.423	**AUG , UGC**
4	0.370	**AUGC**

$\bar{\lambda}_k > \lambda_{cr}$ .... network G_k is connected

$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

Connectivity threshold: $\lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)}$

Degree of neutrality of neutral networks and the connectivity threshold
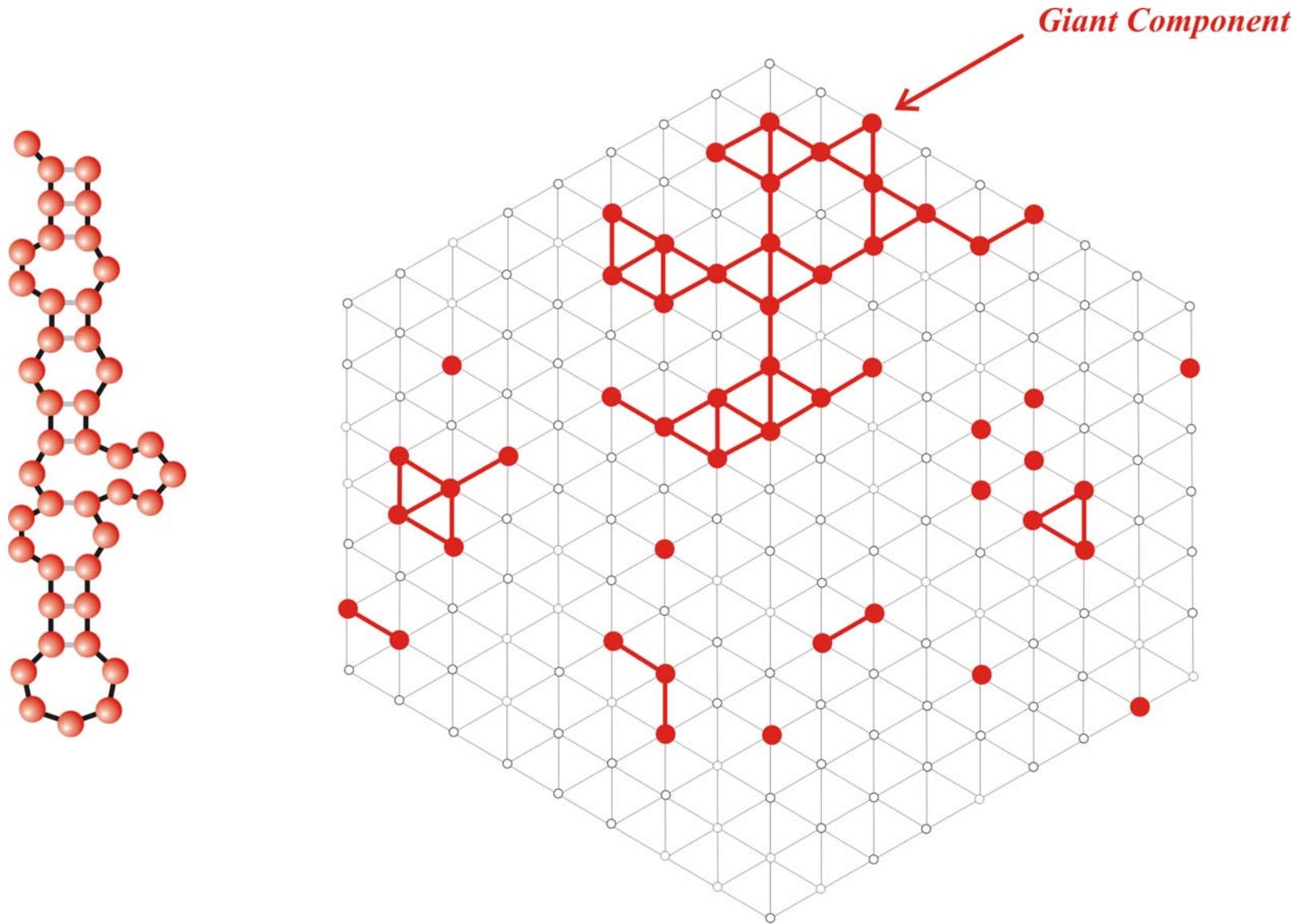
A multi-component neutral network formed by a rare structure: $\lambda < \lambda_{cr}$
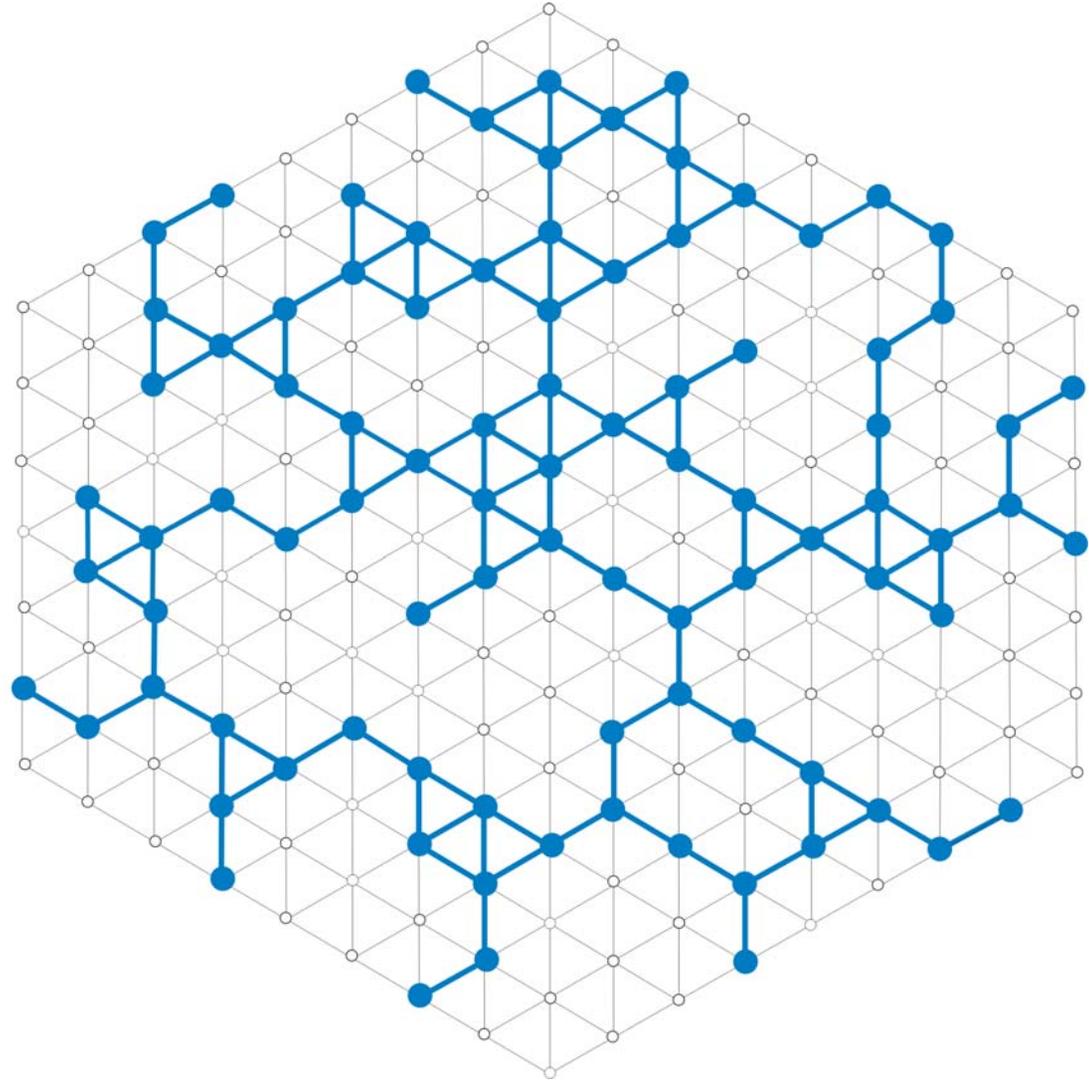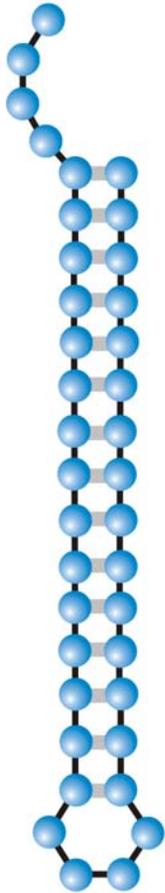
A connected neutral network formed by a common structure: $\lambda > \lambda_{cr}$

From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER[1,2,3], WALTER FONTANA[3], PETER F. STADLER[2,3] AND IVO L. HOFACKER[2]

[1] Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany
[2] Institut für Theoretische Chemie, Universität Wien, Austria
[3] Santa Fe Institute, Santa Fe, U.S.A.

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).

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.

Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

279

Reference for postulation and *in silico* verification of *neutral networks*

Computer simulation of RNA optimization

Walter Fontana and Peter Schuster,
Biophysical Chemistry 26:123-147, 1987

Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

A computer model of evolutionary optimization

Walter Fontana and Peter Schuster

Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

Molecular evolution is viewed as a typical combinatorial optimization problem. We analyse a chemical reaction model which considers RNA replication including correct copying and point mutations together with hydrolytic degradation and the dilution flux of a flow reactor. The corresponding stochastic reaction network is implemented on a computer in order to investigate some basic features of evolutionary optimization dynamics. Characteristic features of real molecular systems are mimicked by folding binary sequences into unknotted two-dimensional structures. Selective values are derived from these molecular 'phenotypes' by an evaluation procedure which assigns numerical values to different elements of the secondary structure. The fitness function obtained thereby contains nontrivial long-range interactions which are typical for real systems. The fitness landscape also reveals quite involved and bizarre local topologies which we consider also representative of polynucleotide replication in actually occurring systems. Optimization operates on an ensemble of sequences via mutation and natural selection. The strategy observed in the simulation experiments is fairly general and resembles closely a heuristic widely applied in operations research areas. Despite the relative smallness of the system – we study 2000 molecules of chain length $\nu = 70$ in a typical simulation experiment – features typical for the evolution of real populations are observed as there are error thresholds for replication, evolutionary steps and quasistationary sequence distributions. The relative importance of selectively neutral or almost neutral variants is discussed quantitatively. Four characteristic ensemble properties, entropy of the distribution, ensemble correlation, mean Hamming distance and diversity of the population, are computed and checked for their sensitivity in recording major optimization events during the simulation.

1. Molecular evolution and optimization

Conventional population genetics treats mutation as an external stochastic source. Moreover, mutations are considered as very rare events. In the absence of genetic recombination populations of haploid organisms are expected to be usually homogeneous. Experimental evidence on viral and bacterial populations is available now and it contradicts these expectations. Mutations appear much more frequently than was originally assumed.

Dedicated to Professor Manfred Eigen on the occasion of his 60th birthday.
Correspondence address: P. Schuster, Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria.

The molecular approach considers error-free replication and mutation as parallel reactions within the same mechanism. Detailed information on the molecular mechanisms of polynucleotide replication provides direct insight into the nature of mutations and their role in evolution. Several classes of mutations are properly distinguished: point mutations, deletions and insertions. Point mutations are of special importance: they represent the most frequent mutations and are easily incorporated into theoretical models of molecular evolution. This does not mean, however, that the other classes of mutations are not important in evolution. To give an example: there is a general belief that insertions leading to gene duplication played a major role in the development of present day enzyme families.

The first theoretical model of molecular evolu-

Physical aspects of evolutionary optimization and adaptation

Walter Fontana, Wolfgang Schnabl, and Peter Schuster*
Institut für Theoretische Chemie der Universität Wien, Währingerstrasse 17, A 1090 Wien, Austria
(Received 2 February 1989; revised manuscript received 5 May 1989)

A model of an objective function based on polynucleotide folding is used to investigate the dynamics of evolutionary adaptation in finite populations. Binary sequences are optimized with respect to their kinetic properties through a stochastic process involving mutation and selection. The objective function consists in a mapping from the set of all binary strings with given length into a set of two-dimensional structures. These structures then encode the kinetic properties, expressed in terms of parameters of reaction probability distributions. The objective function obtained thereby represents a realistic example of a highly "rugged landscape." Ensembles of molecular strings adapting to this landscape are studied by tracing their escape path from local optima and by applying multivariate analysis. Effects of small population numbers in the tail of the sequence distribution are discussed quantitatively. Close upper bounds to the number of distinct values produced by our objective function are given. The distribution of values is explored by means of simulated annealing and reveals a random scatter in the locations of optima in the space of all sequences. The genetic optimization protocol is applied to the "traveling salesman" problem.

Fig. 1. The reaction network. Synthesis on template I_k proceeds with the rate constant a_k and leads with frequency Q_{ik} to a new template I_i preserving the old copy. Materials A needed for polymerization are assumed to be buffered. Degradation to waste products B occurs with rate d_k and a controlled unspecific flux $\Phi(t)$ removes templates from the system.

Fig. 2. The evolution reactor. This kind of flow reactor consists of a reaction vessel which allows for temperature and pressure control. Its walls are impermeable to polynucleotides. Energy-rich material is poured from the environment into the reactor. The degradation products are removed steadily. Material transport is adjusted in such a way that the concentration of monomers is constant in the reactor. A dilution flux Φ is installed in order to remove excess of polynucleotides produced by replication. Thus, the sum of the numbers of individual particles $\sum_i X_i(t) = N(t)$ may be controlled by the flux Φ. Under 'constant organization' Φ is adjusted such that $N(t) = \Theta$ is essentially constant. By this we indicate that fluctuations with standard deviation $\sigma = \sqrt{N}$ occur regularly. The regulation of Φ requires internal control, which can be achieved by logistic coupling.

Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

Evolution *in silico*

W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455
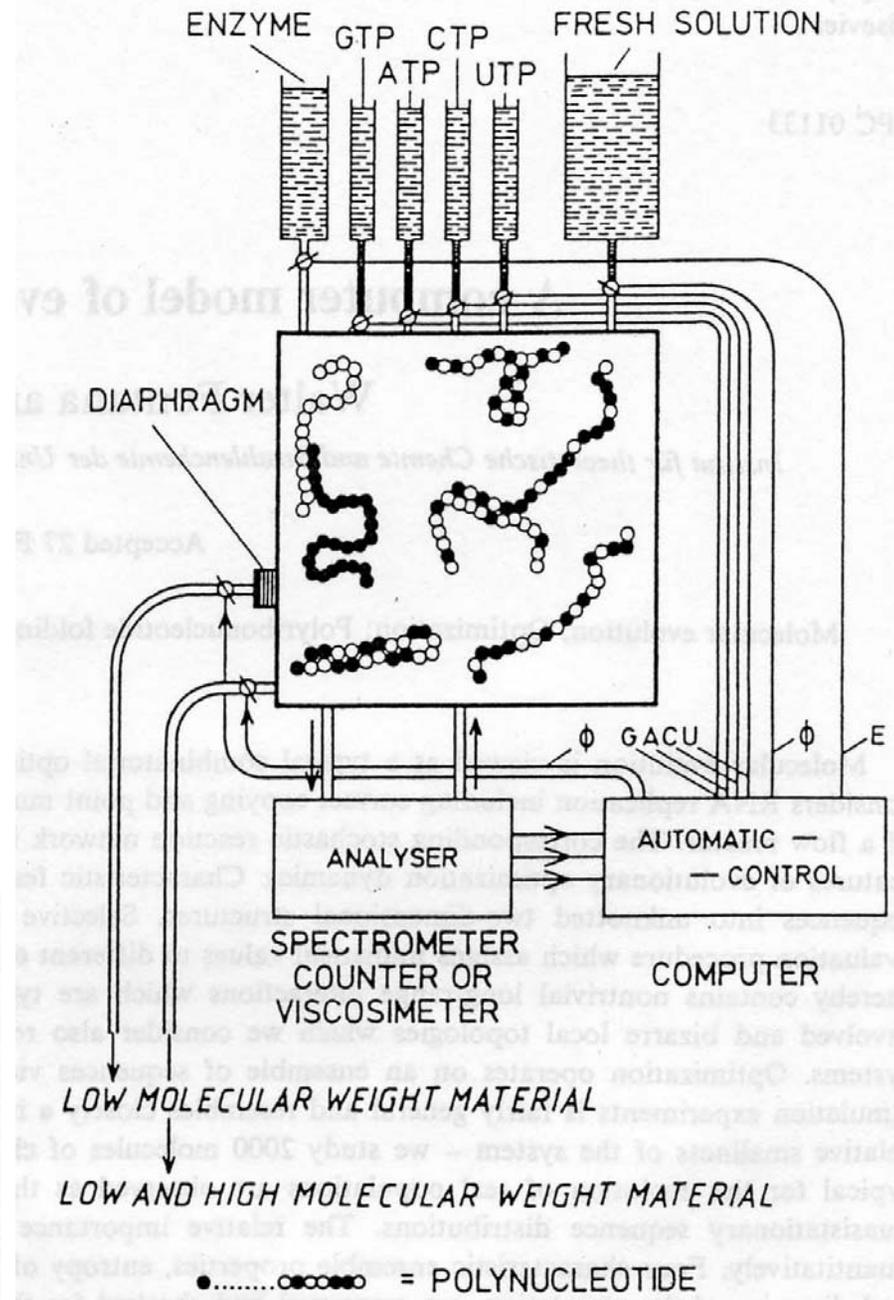
random individuals. The primer pair used for genomic DNA amplification is 5'-TCTCCCTGGATTCT-CATTTA-3' (forward) and 5'-TCTTTGTCTTCTGT-TCCACC-3' (reverse). Reactions were performed in 25 µl using 1 unit of Taq DNA polymerase with each primer at 0.4 µM; 200 µM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl$_2$,1.5 mM MgCl$_2$] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].

33. Data not shown; a dot blot with poly (A)$^+$ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (13).

34. Smith–Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [(6); K-S Chen, L. Potocki, J. R. Lupski, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS

phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.

35. R. A. Fridell, data not shown.

36. K. B. Avraham et al., *Nature Genet.* **11**, 369 (1995); X-Z. Liu et al., ibid. **17**, 268 (1997); F. Gibson et al., *Nature* **374**, 62 (1995); D. Weil et al., ibid., p. 60.

37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)$^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5'-GCATGACCTGCCGGCTAAT-GGG-3'; reverse, 5'-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3'). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR

product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengkala, Bali, and the two families involved in India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

To distinguish continuous 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. This nearness relation is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and provides 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). Our 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 level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicatable 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 an RNA aptamer 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 criterion. 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, here a canonical tRNA shape. The black curve traces the average distance to the target (inversely related to fitness) in the population against time. Aside from a short initial phase, the entire history is dominated by steps, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target structure (7). However, the dominant shapes in the population not only change at these marked events but undergo several fitness-neutral transformations 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 phenotypes itself can be called continuous or discontinuous.

A set of entities 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 issue arises because, in contrast to sequences, there are

Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.
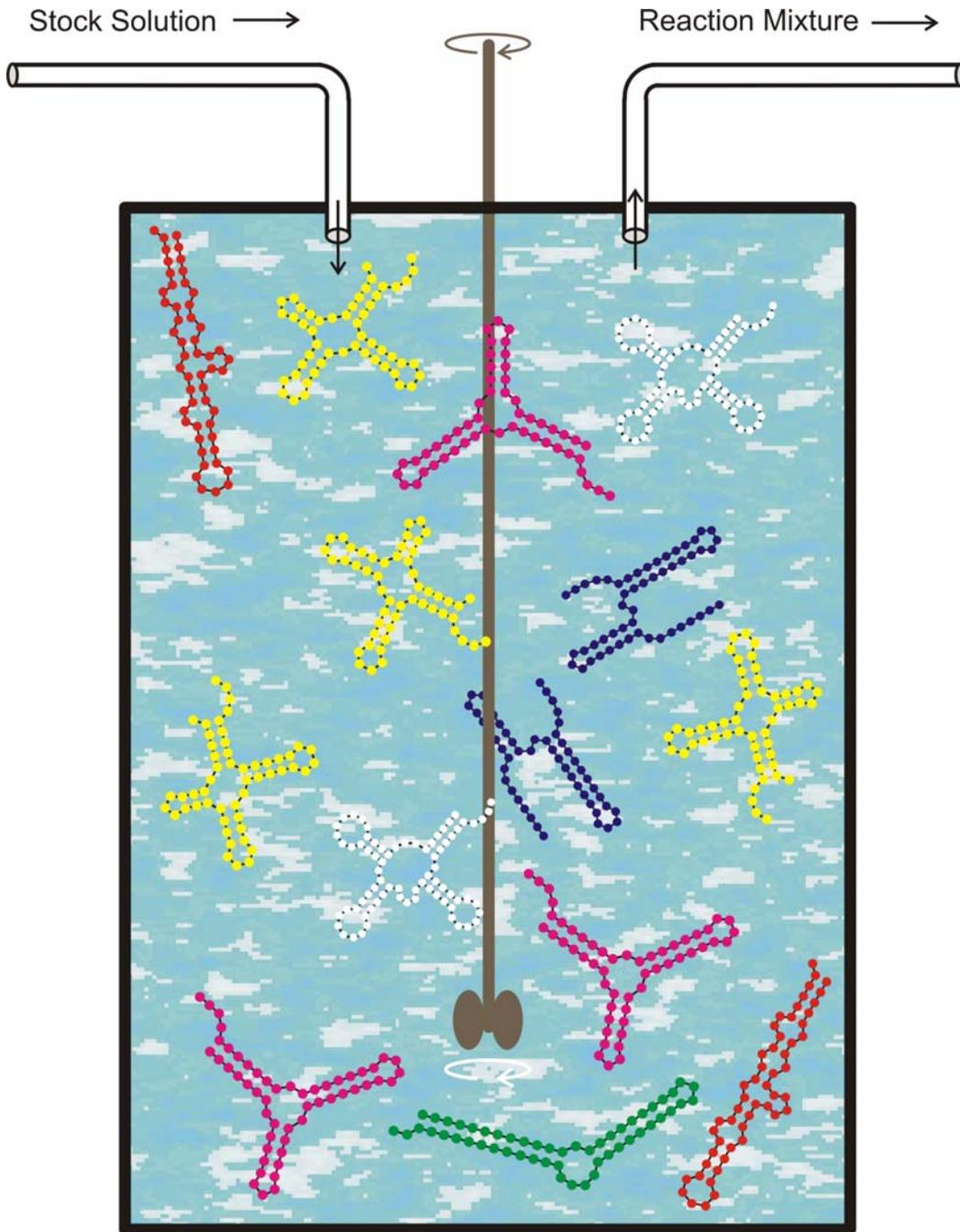
Replication rate constant:

$$f_k = \gamma \, / \, [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

Selection constraint:

Population size, $N = \#$ RNA molecules, is controlled by the flow

$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

Mutation rate:

$$p = 0.001 \, / \, \text{site} \times \text{replication}$$

The flowreactor as a device for **studies** of evolution *in vitro* and *in silico*
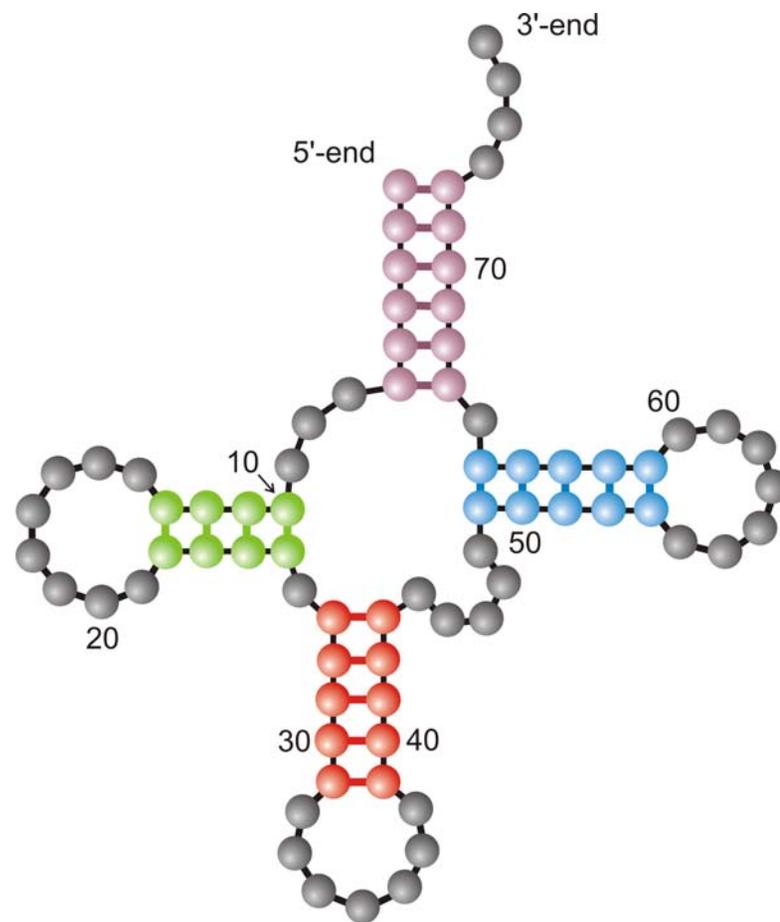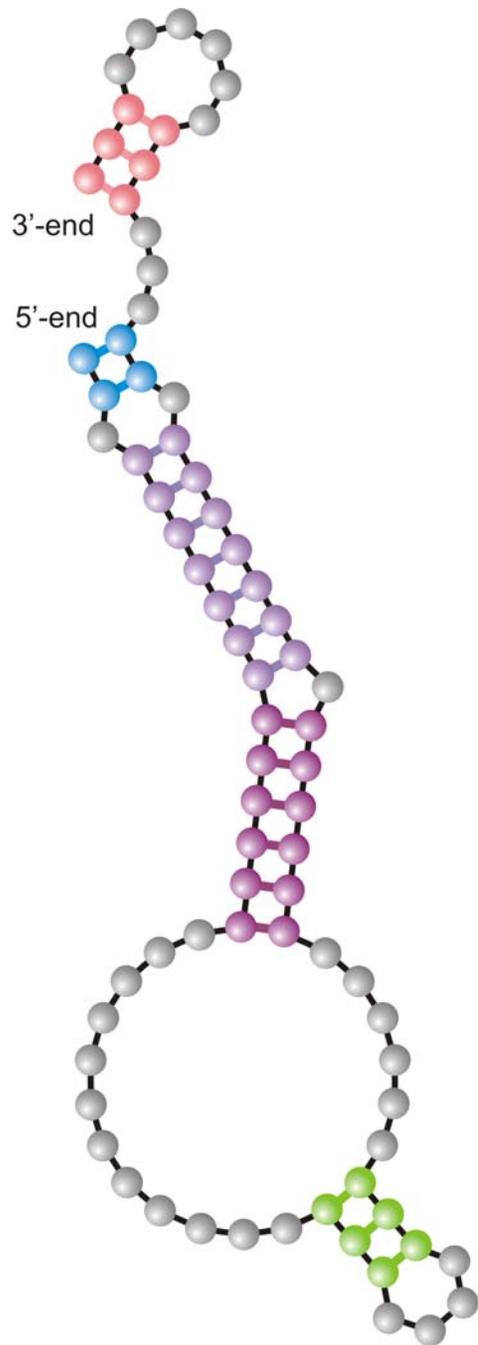
Randomly chosen
initial structure

Phenylalanyl-tRNA as
target structure

Genotype-Phenotype Mapping

$S_\{ = \psi(I_\{)$

$I_\{$

$S_\{$

GGCCCCUUUGGGGGGCCAGACCCCUAAAGGGGUC

Evaluation of the Phenotype

$f_\{ = f(S_\{)$

Mutation

$Q_{\{j}$

$f_\{$

Q

f_1
f_2
f_3
f_4
f_5
f_n
I_1
I_2
I_3
I_4
I_5
I_n

Q

f_1
f_2
f_3
f_4
f_5
f_{n+1}
$f_\{$
I_1
I_2
I_3
I_4
I_5
I_{n+1}
$I_\{$
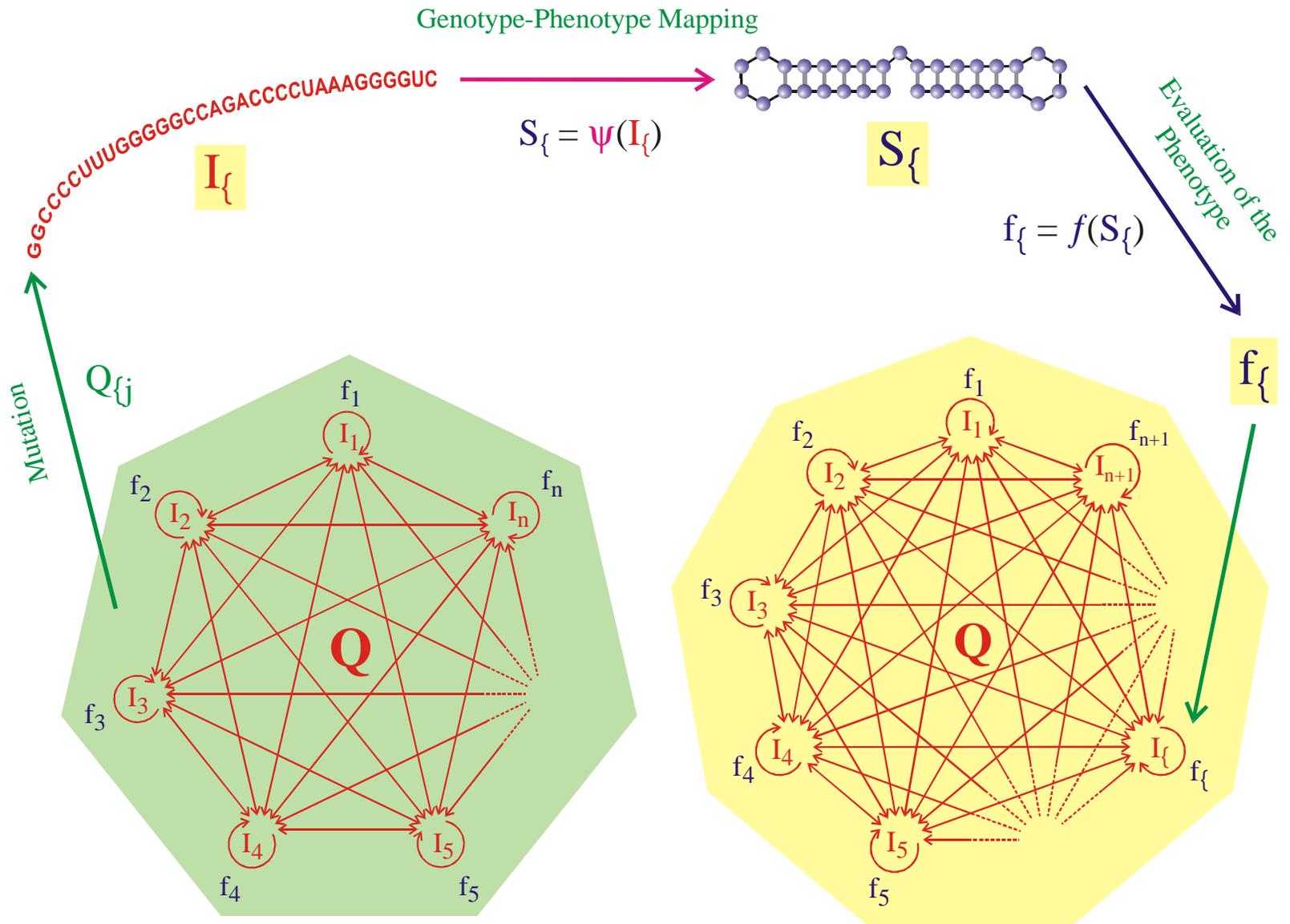
Evolutionary dynamics
including molecular phenotypes
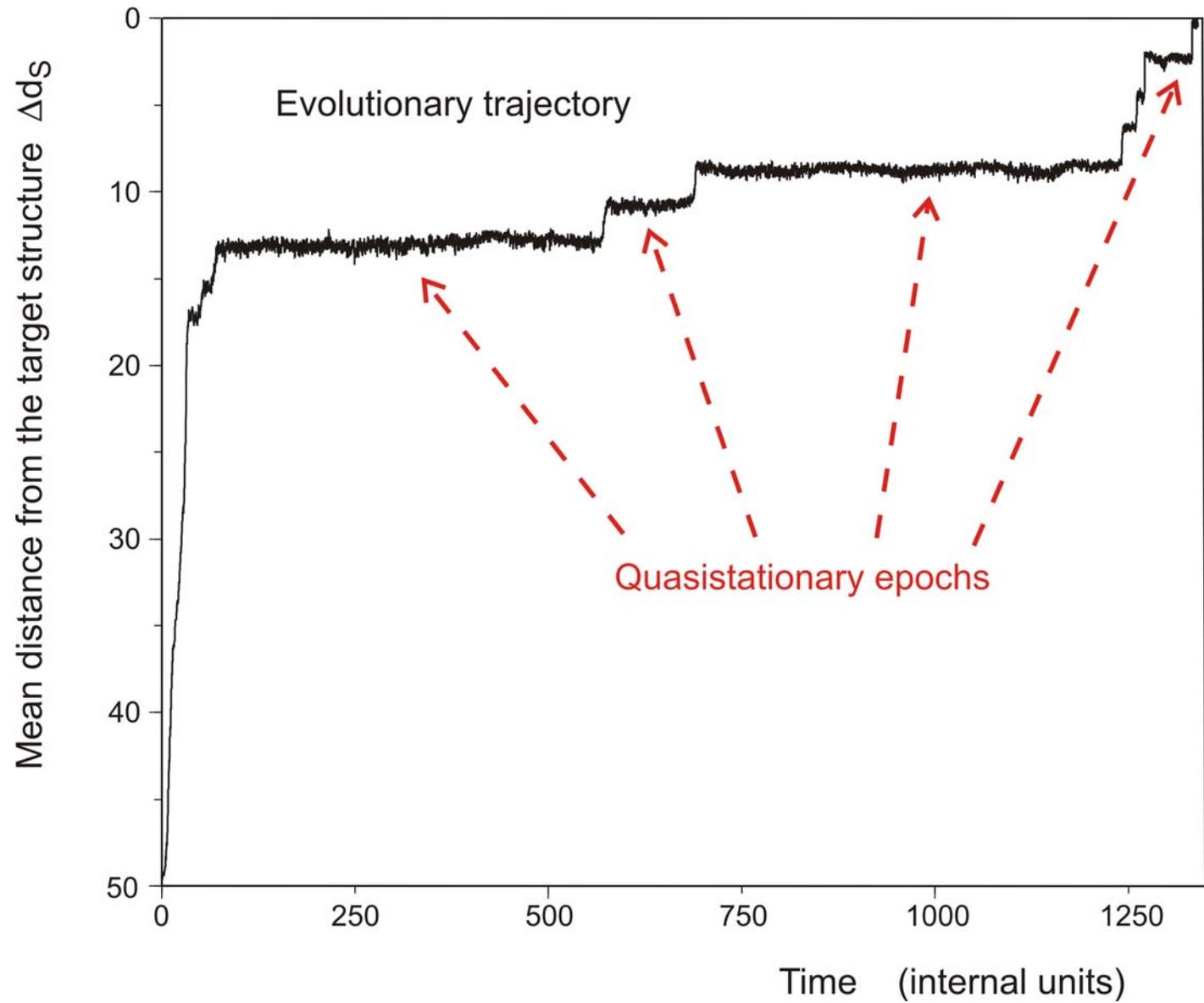
In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during
a long quasi-stationary epoch

Evolutionary trajectory

Average distance from target structure Δd̄S — Time (internal units)
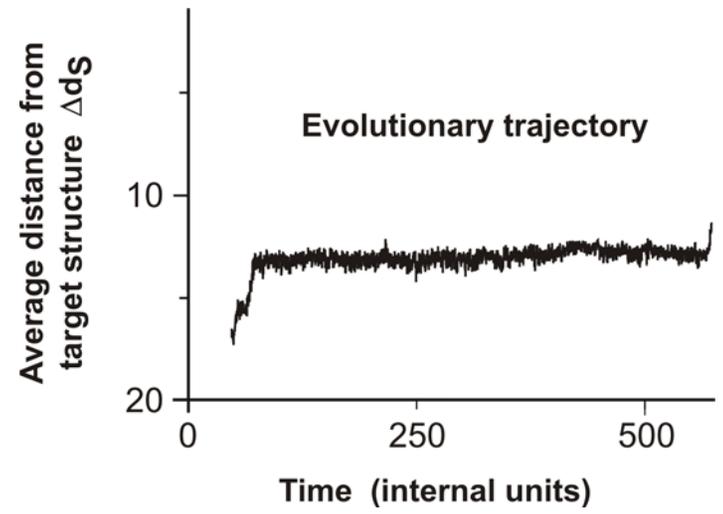
```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .(((((((((((........(((....)))......)))))....(((((......))))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9       .((((((.(((((........(((....)))......)))))....(((((......))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10      .(((((..(((((........(((....)))......)))))....(((((......))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

Transition inducing point mutations
change the molecular structure

Neutral point mutations leave the
molecular structure unchanged

Neutral genotype evolution during phenotypic stasis

Evolutionary trajectory

Spreading of the population
on neutral networks

Drift of the population center
in sequence space
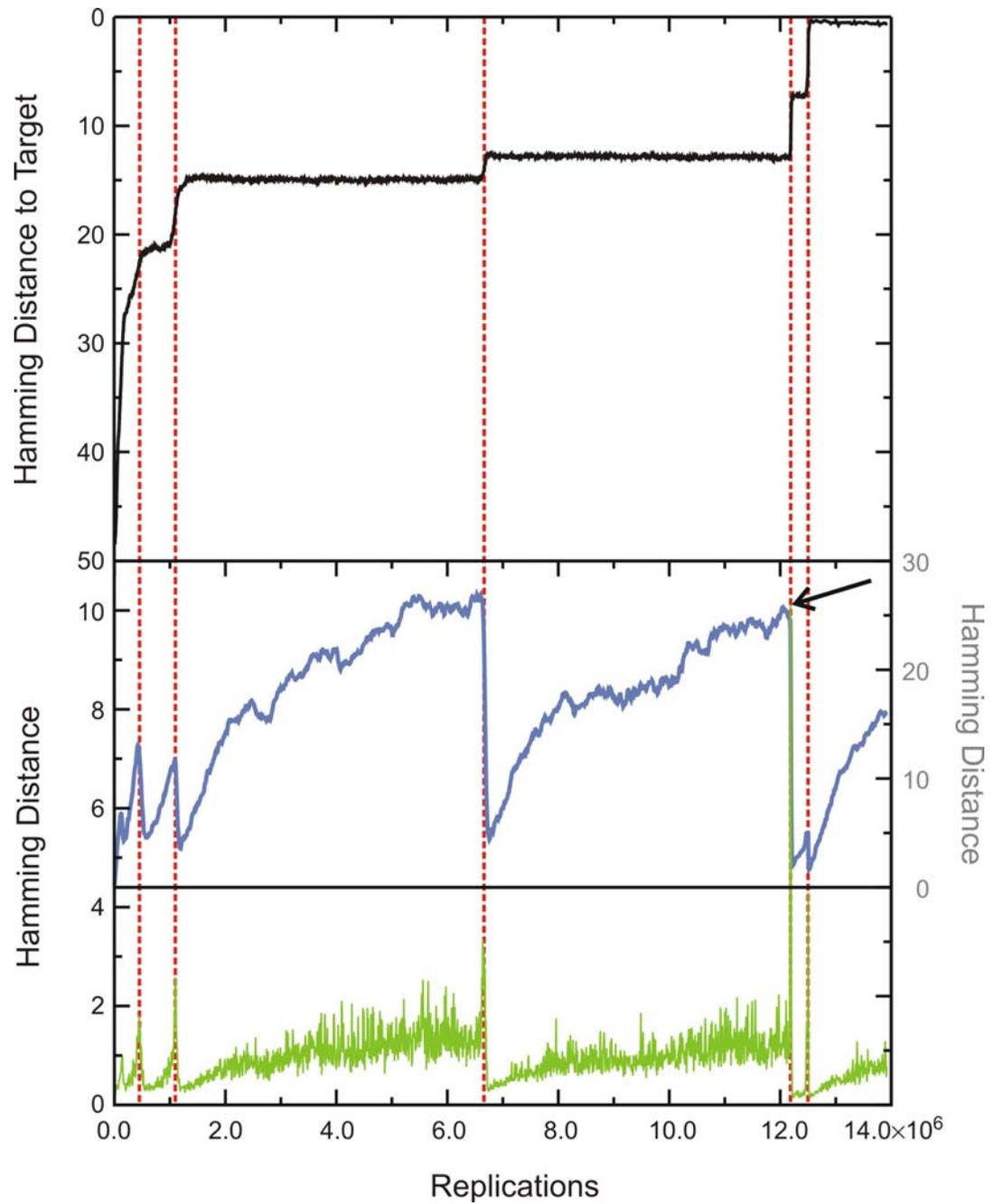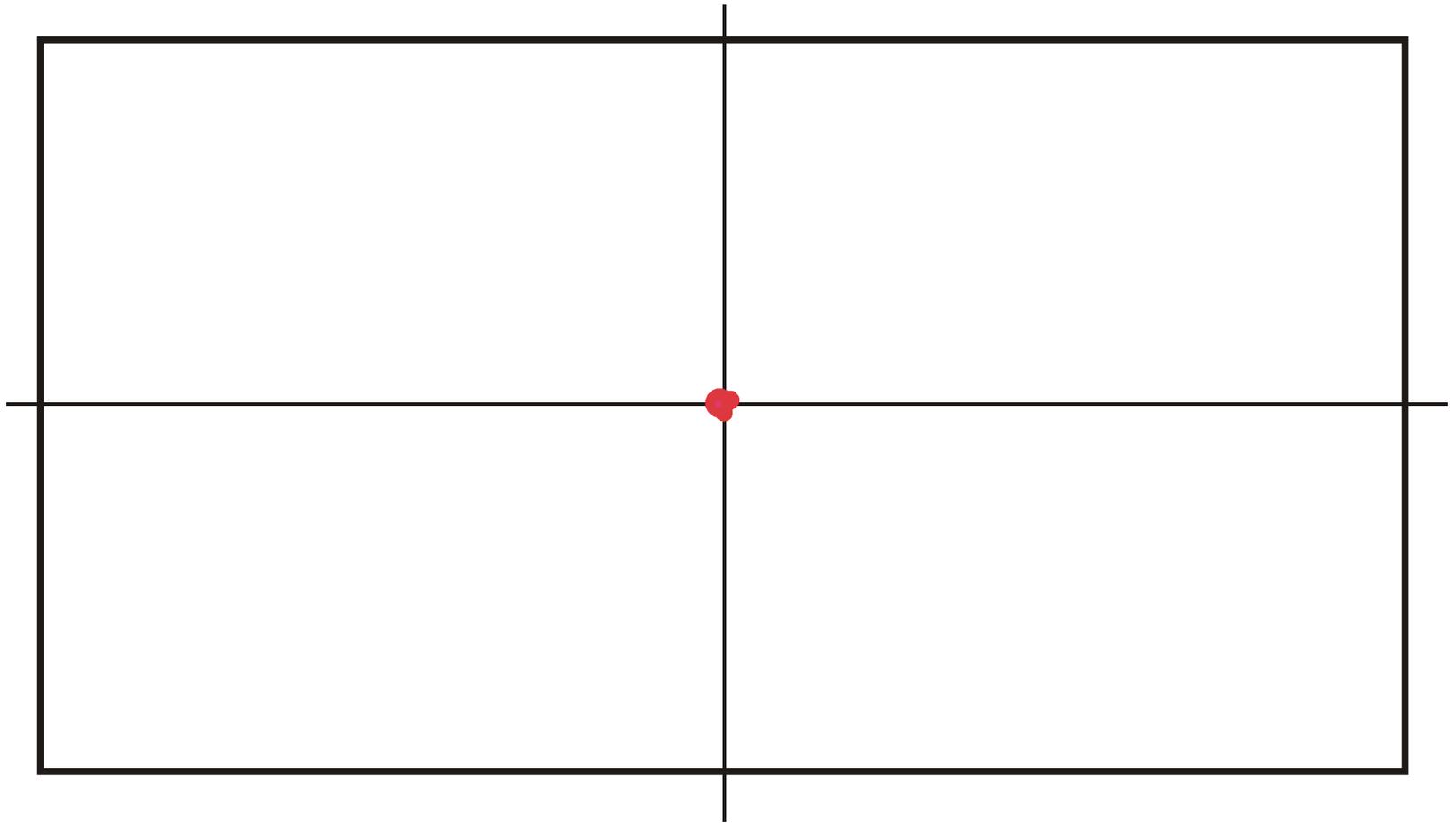
Spreading and evolution of a population on a neutral network: t = 150
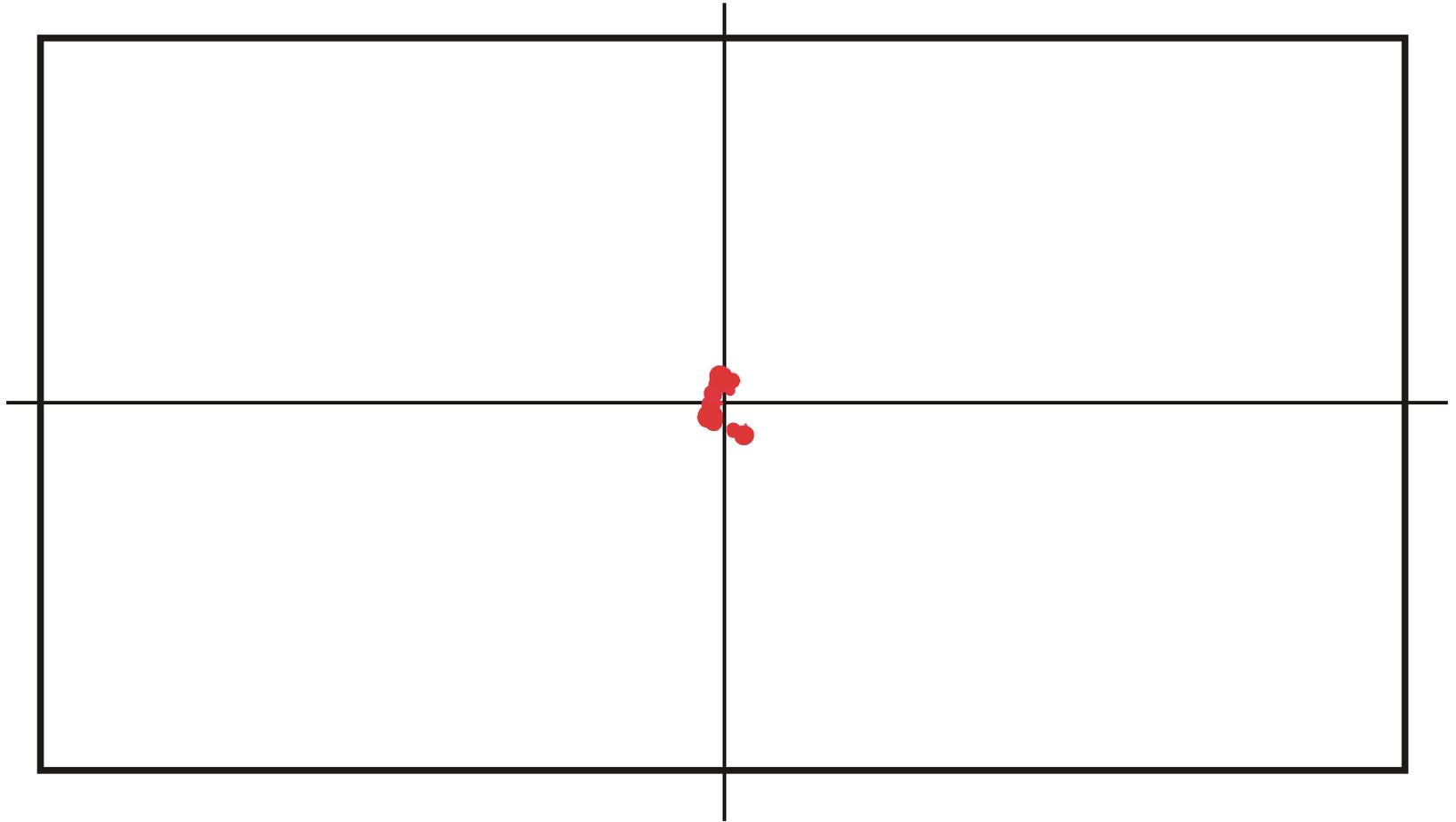
Spreading and evolution of a population on a neutral network : t = 170
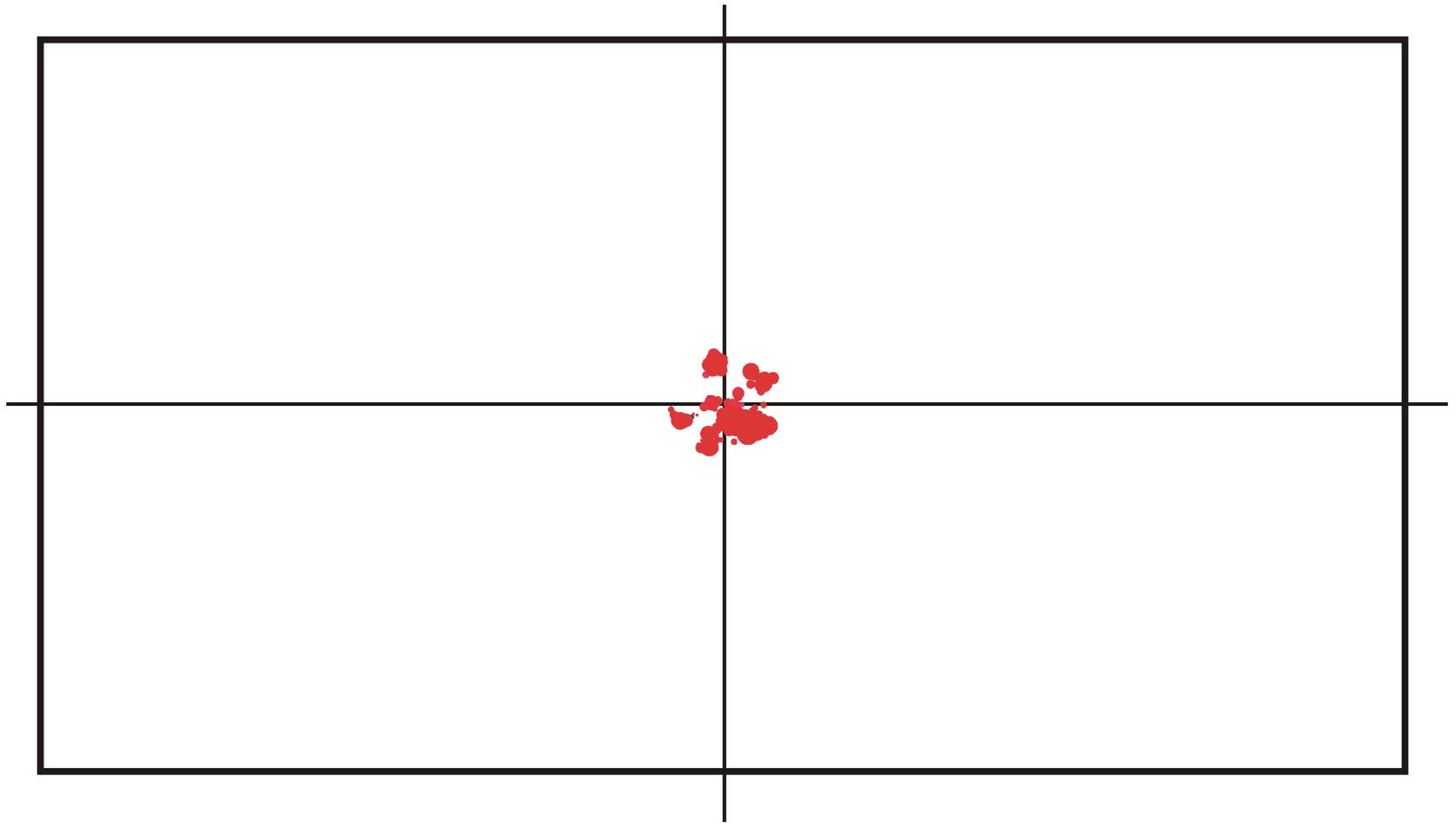
Spreading and evolution of a population on a neutral network : t = 200
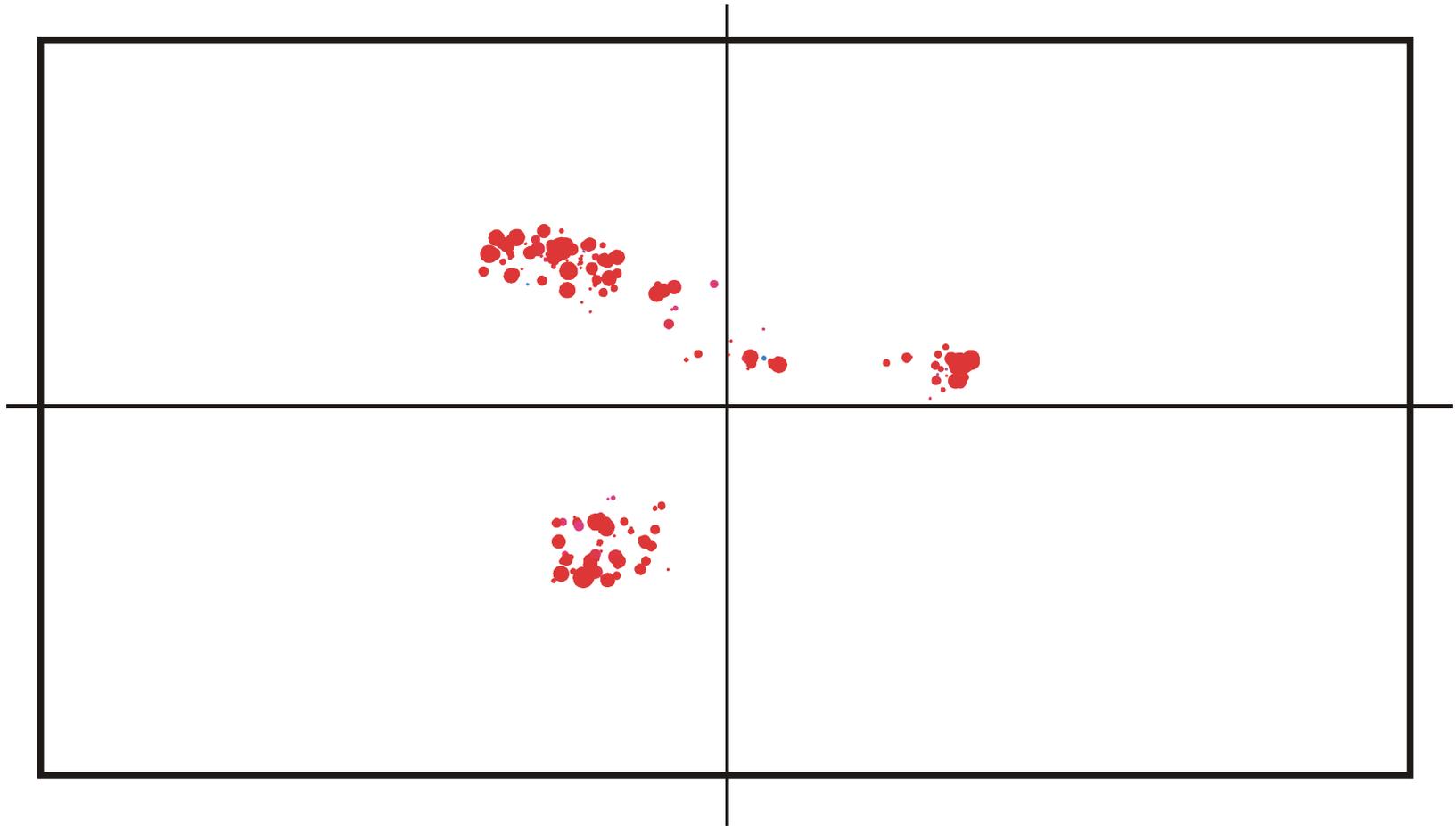
Spreading and evolution of a population on a neutral network : t = 350
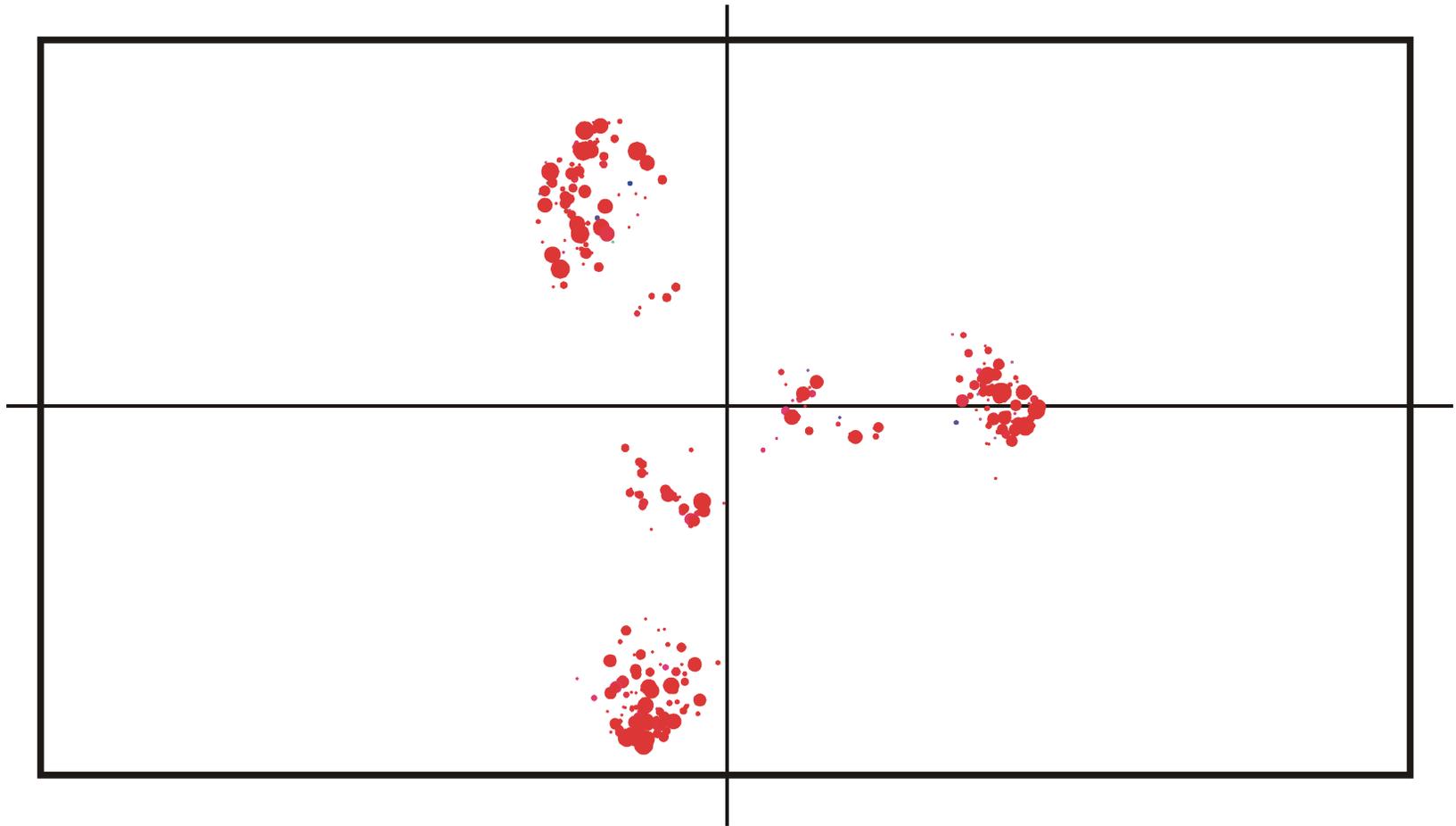
Spreading and evolution of a population on a neutral network : t = 500
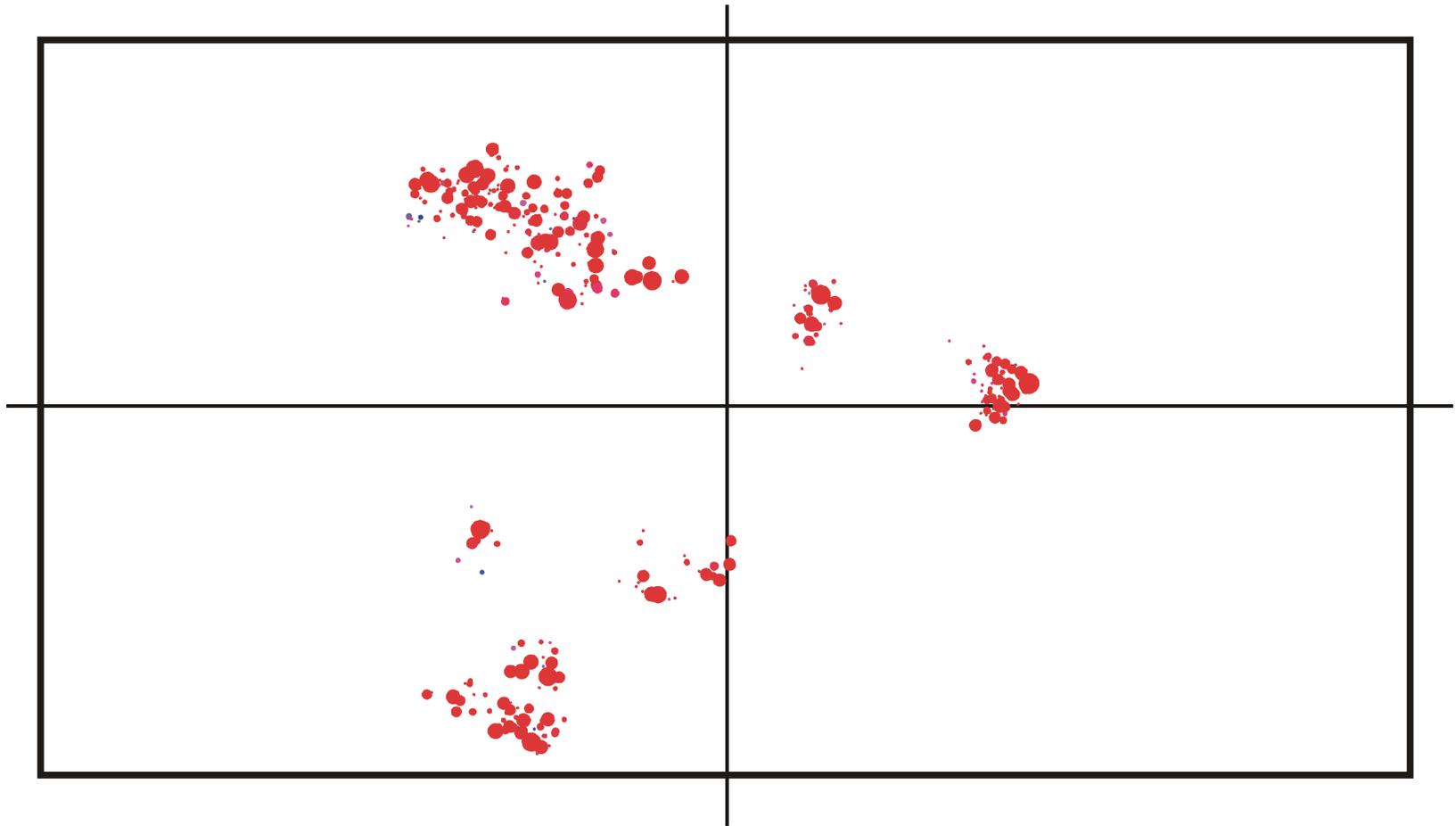
Spreading and evolution of a population on a neutral network : t = 650
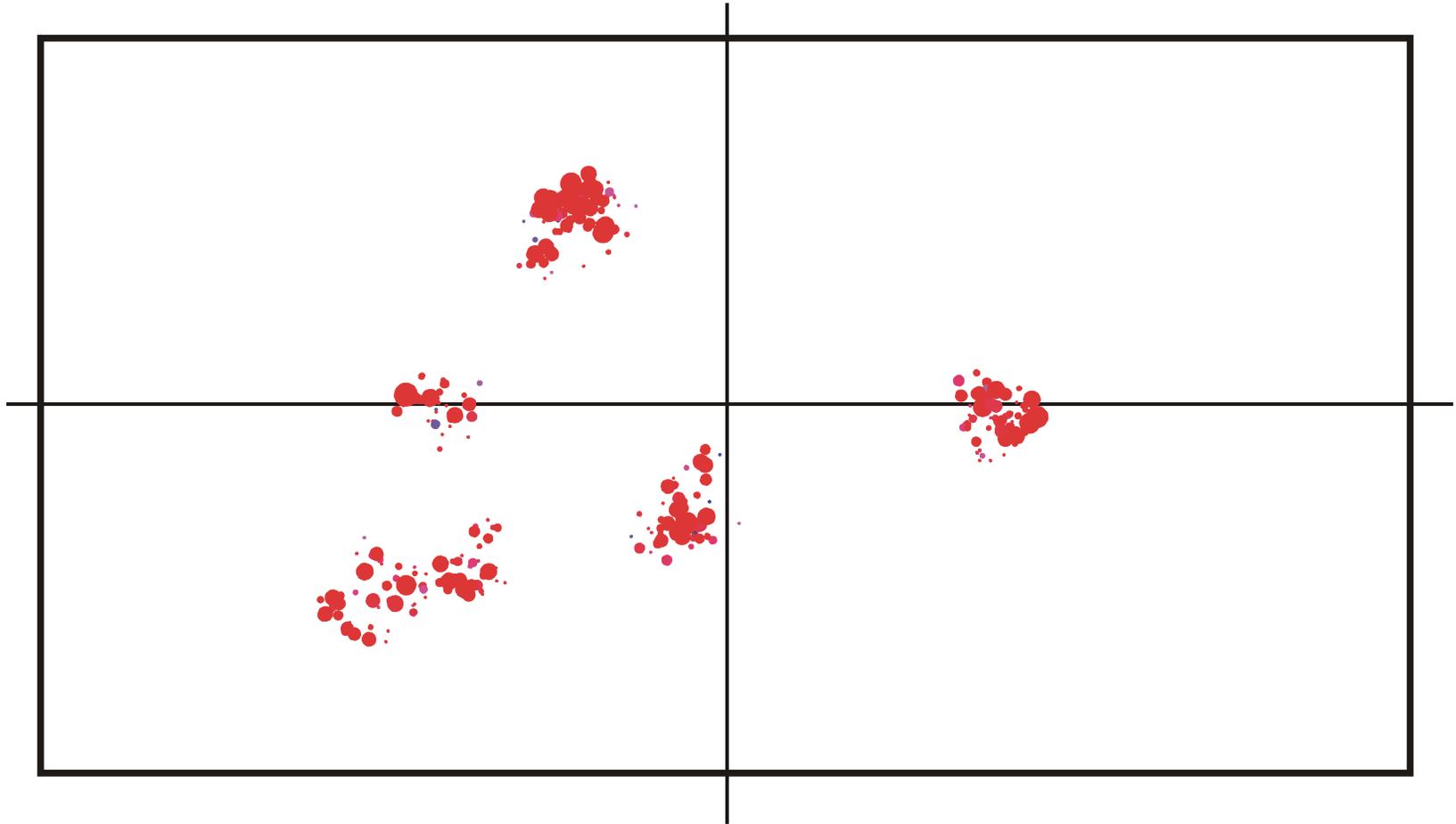
Spreading and evolution of a population on a neutral network : t = 820
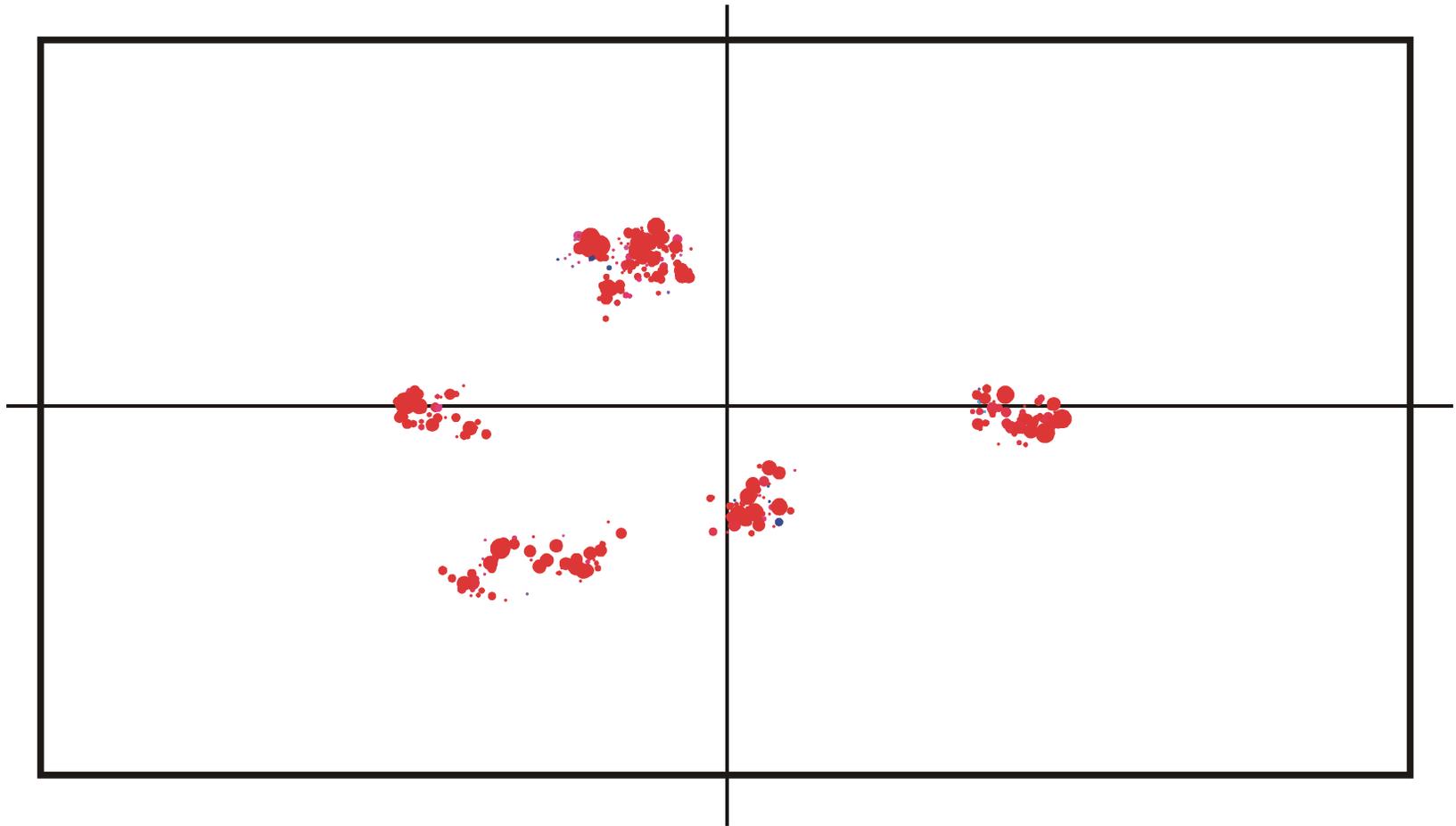
Spreading and evolution of a population on a neutral network : t = 825
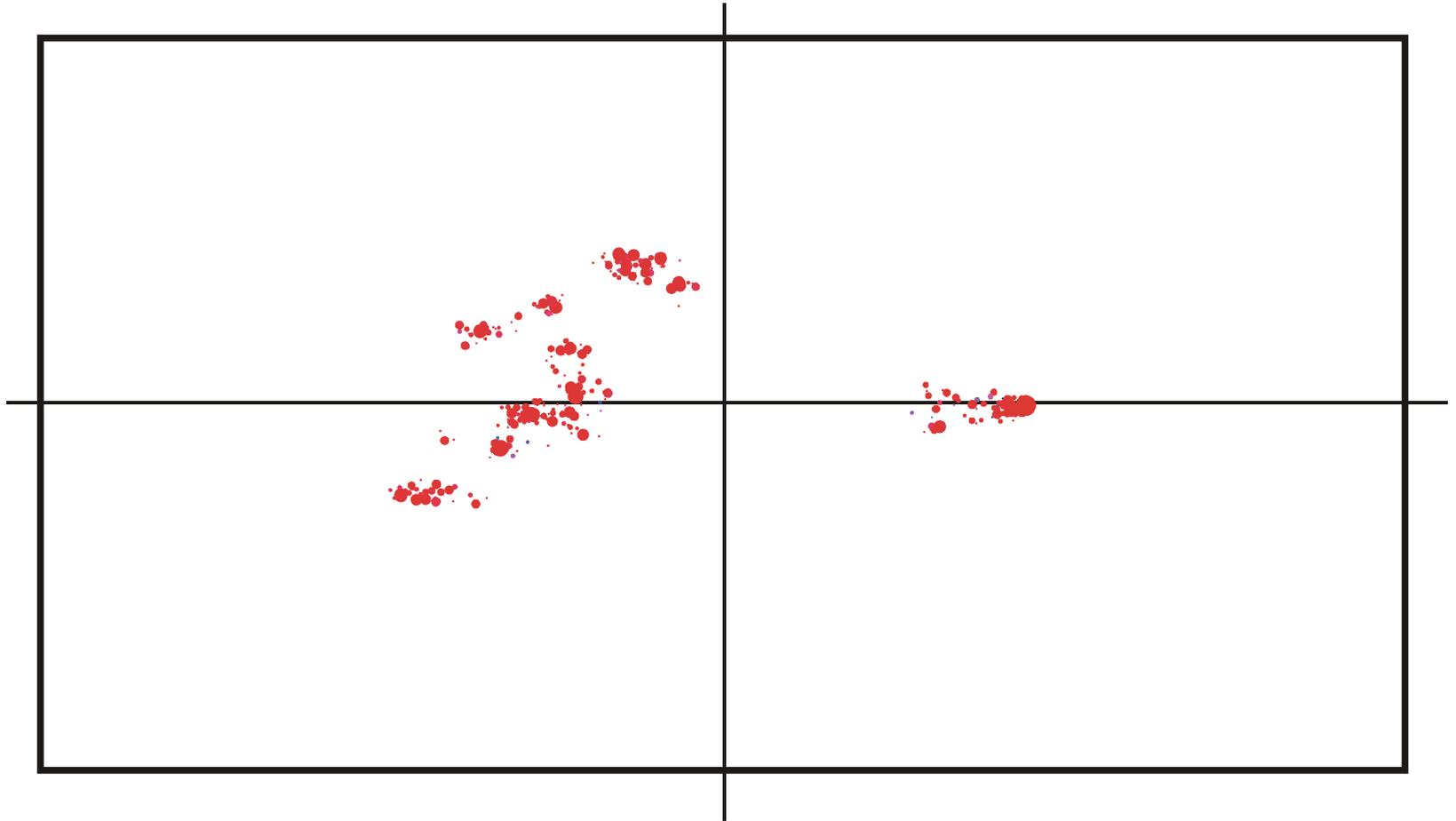
Spreading and evolution of a population on a neutral network : t = 830
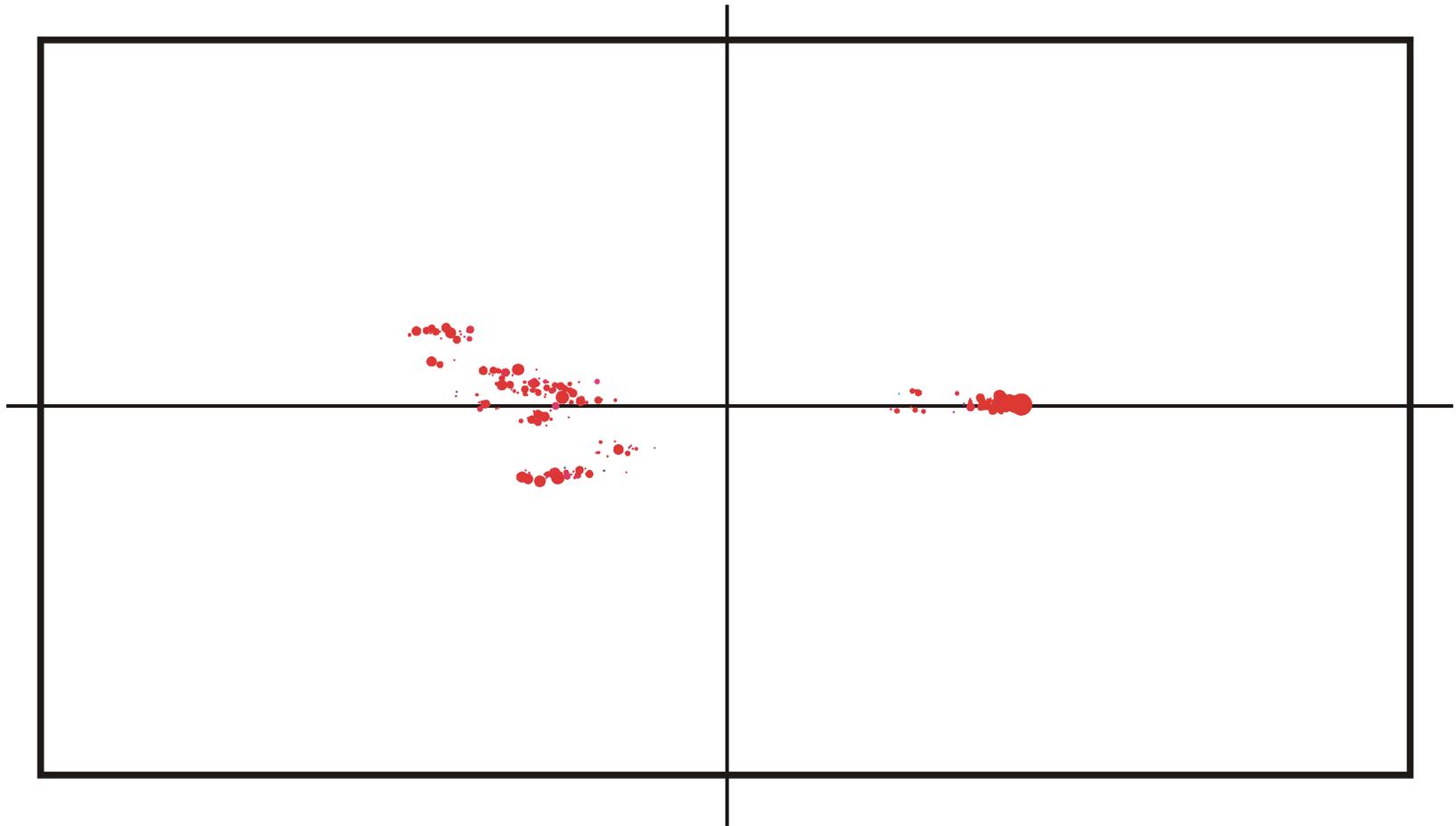
Spreading and evolution of a population on a neutral network : t = 835
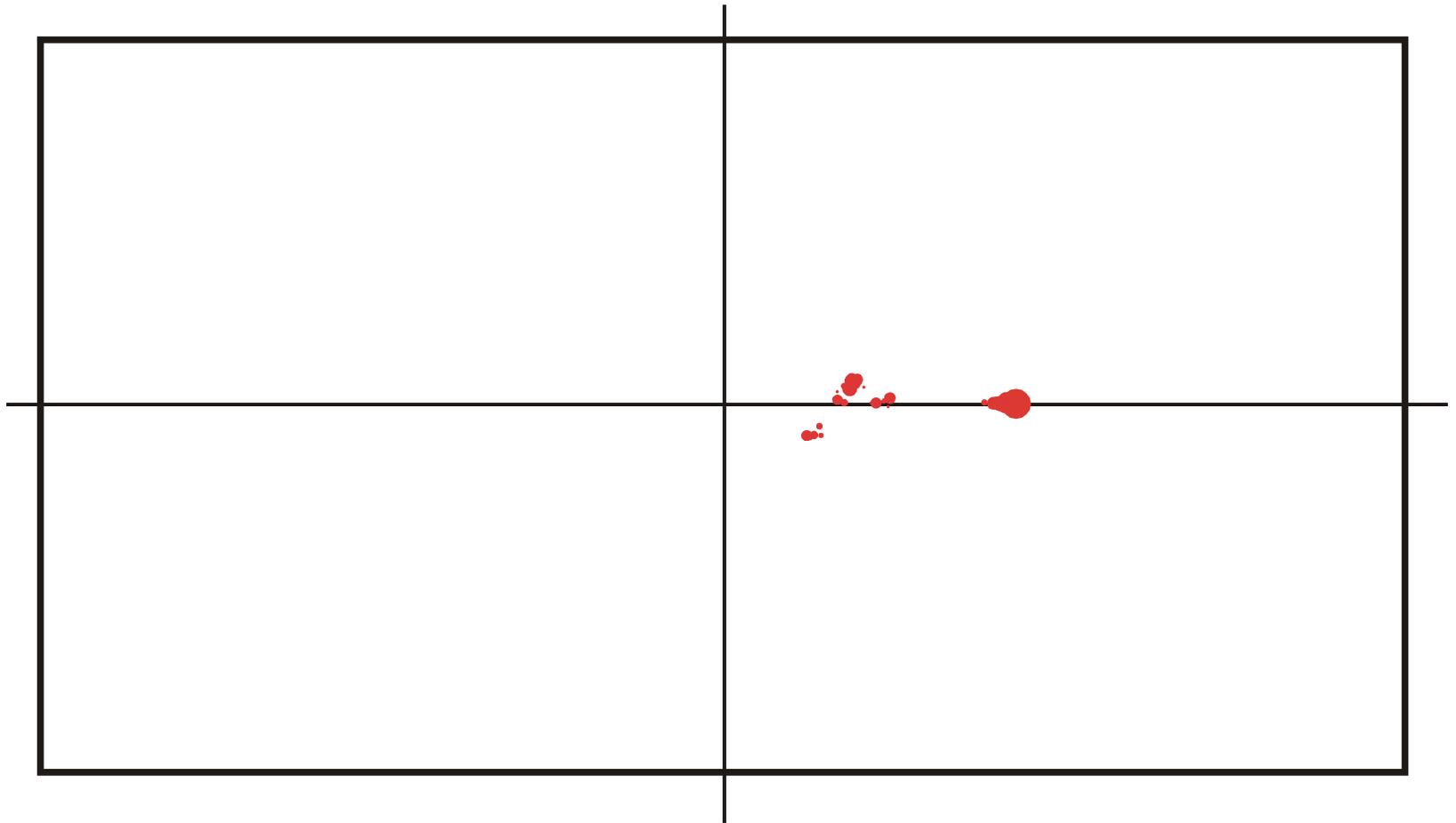
Spreading and evolution of a population on a neutral network : t = 840
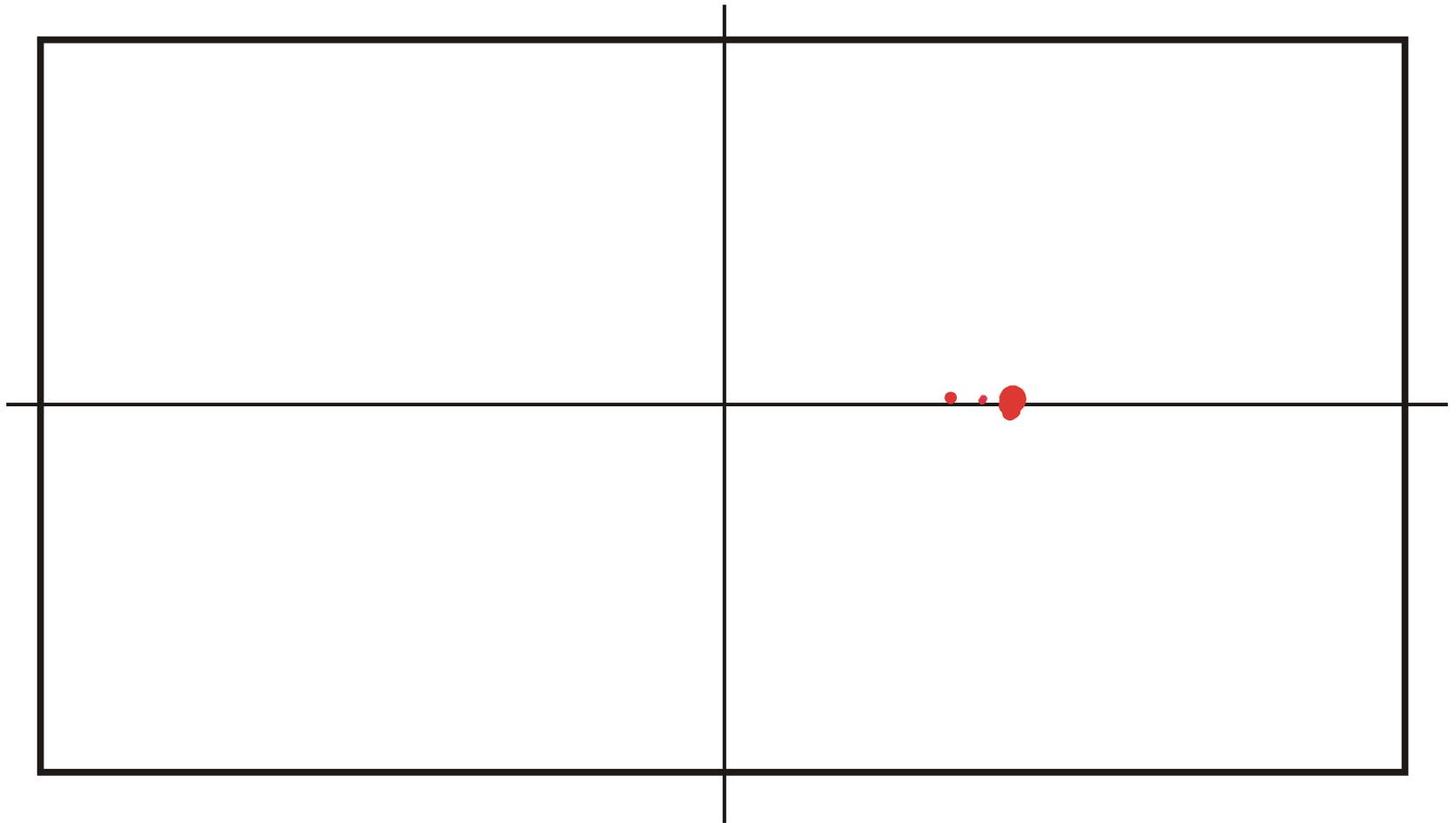
Spreading and evolution of a population on a neutral network : t = 845
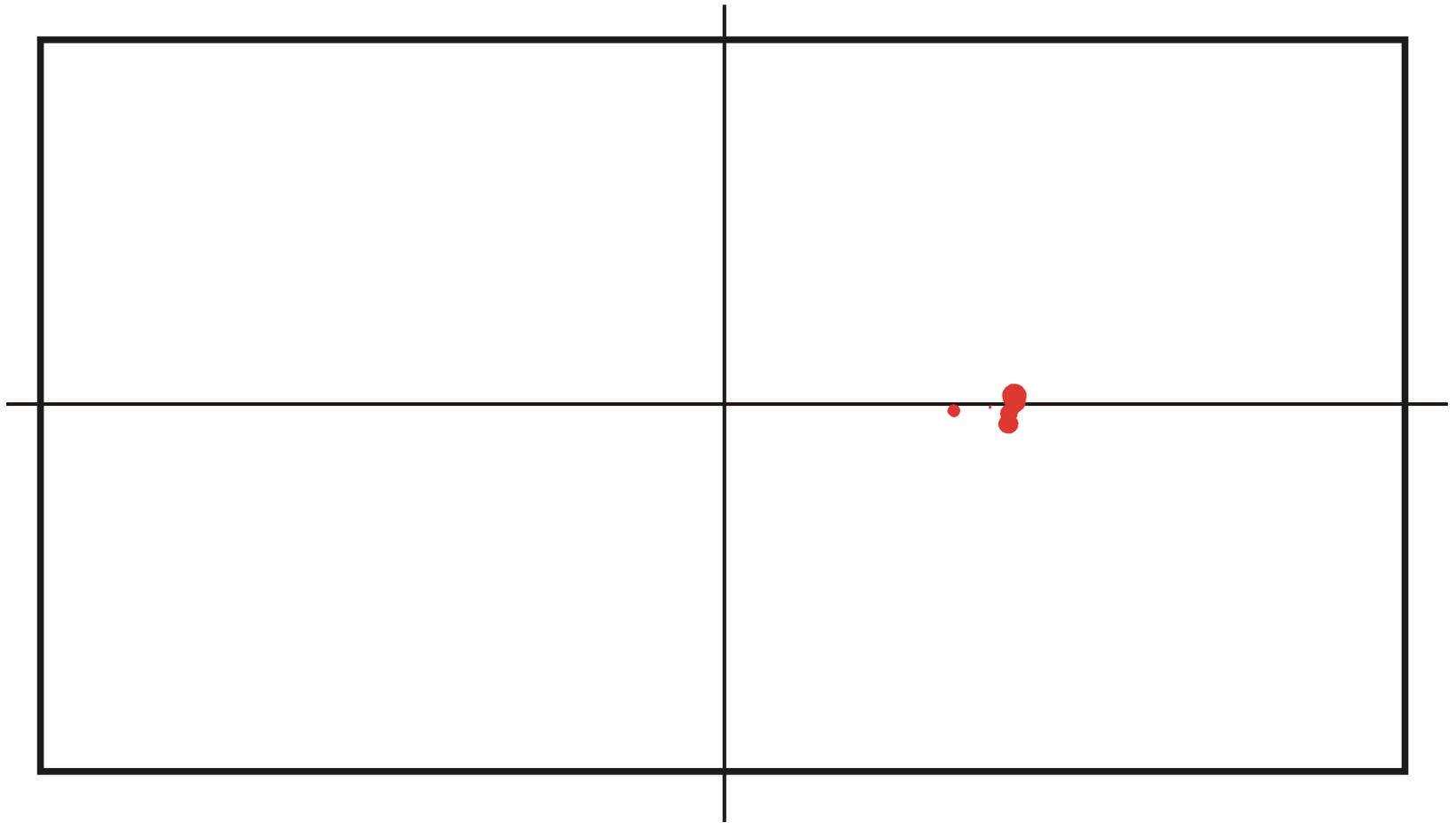
Spreading and evolution of a population on a neutral network : t = 850
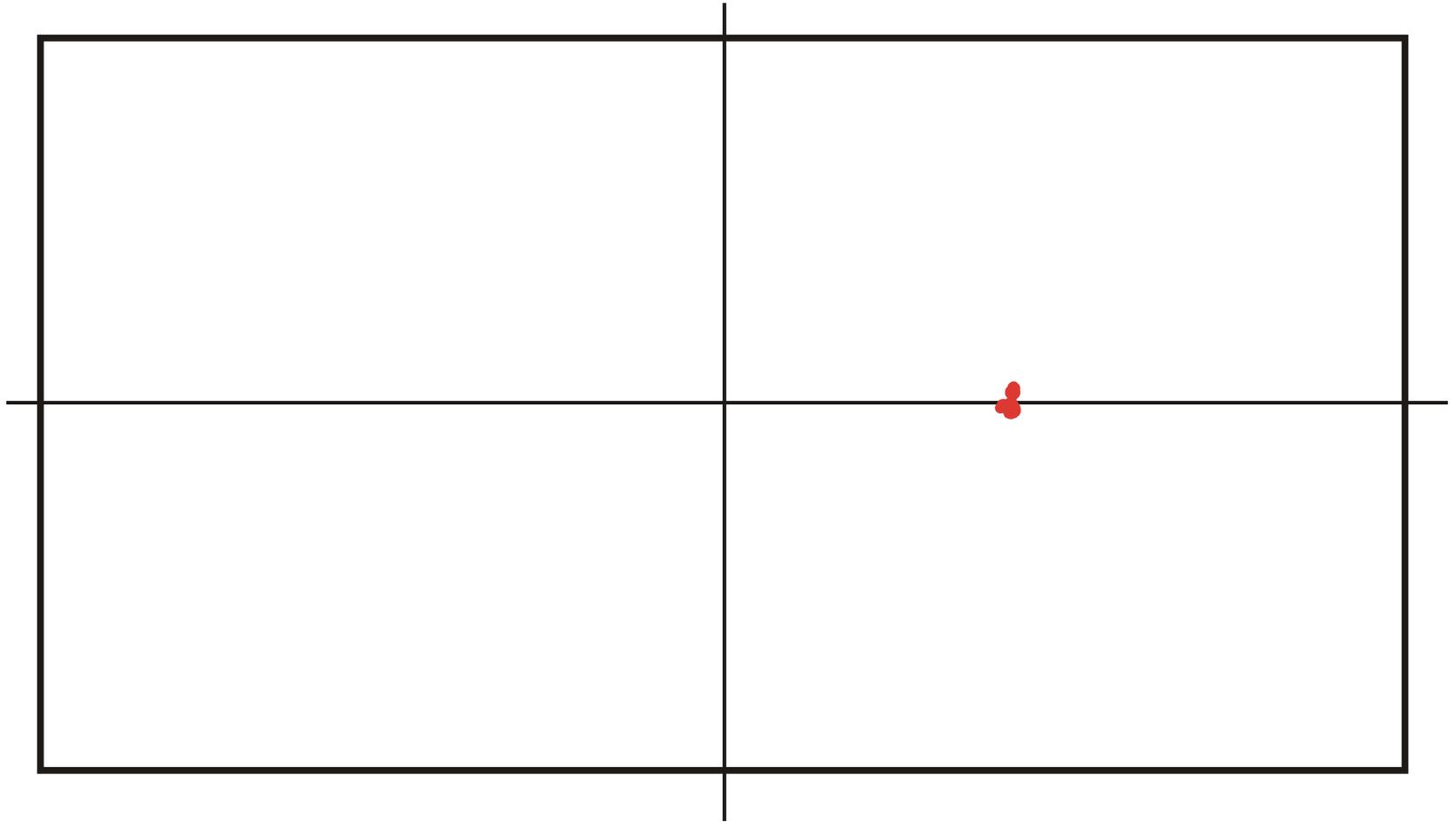
Spreading and evolution of a population on a neutral network : t = 855

Mount Fuji

Example of a smooth landscape on Earth

Dolomites
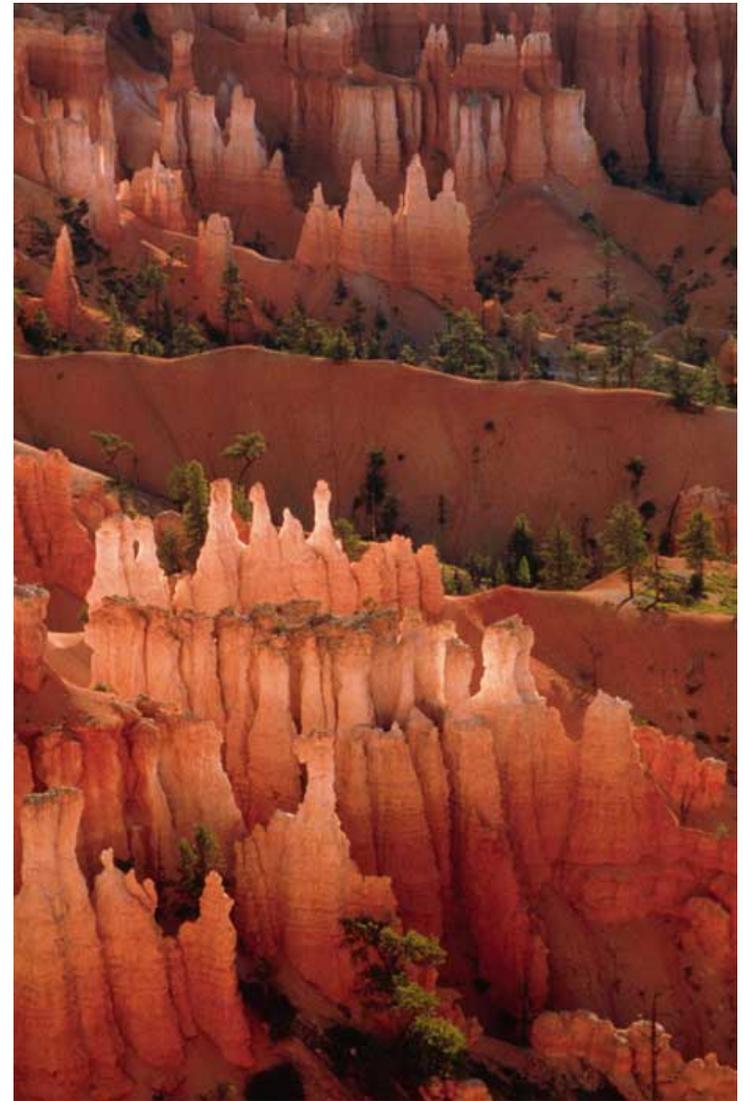
Bryce Canyon

Examples of rugged landscapes on Earth

Fitness

End of Walk

Start of Walk

Genotype Space
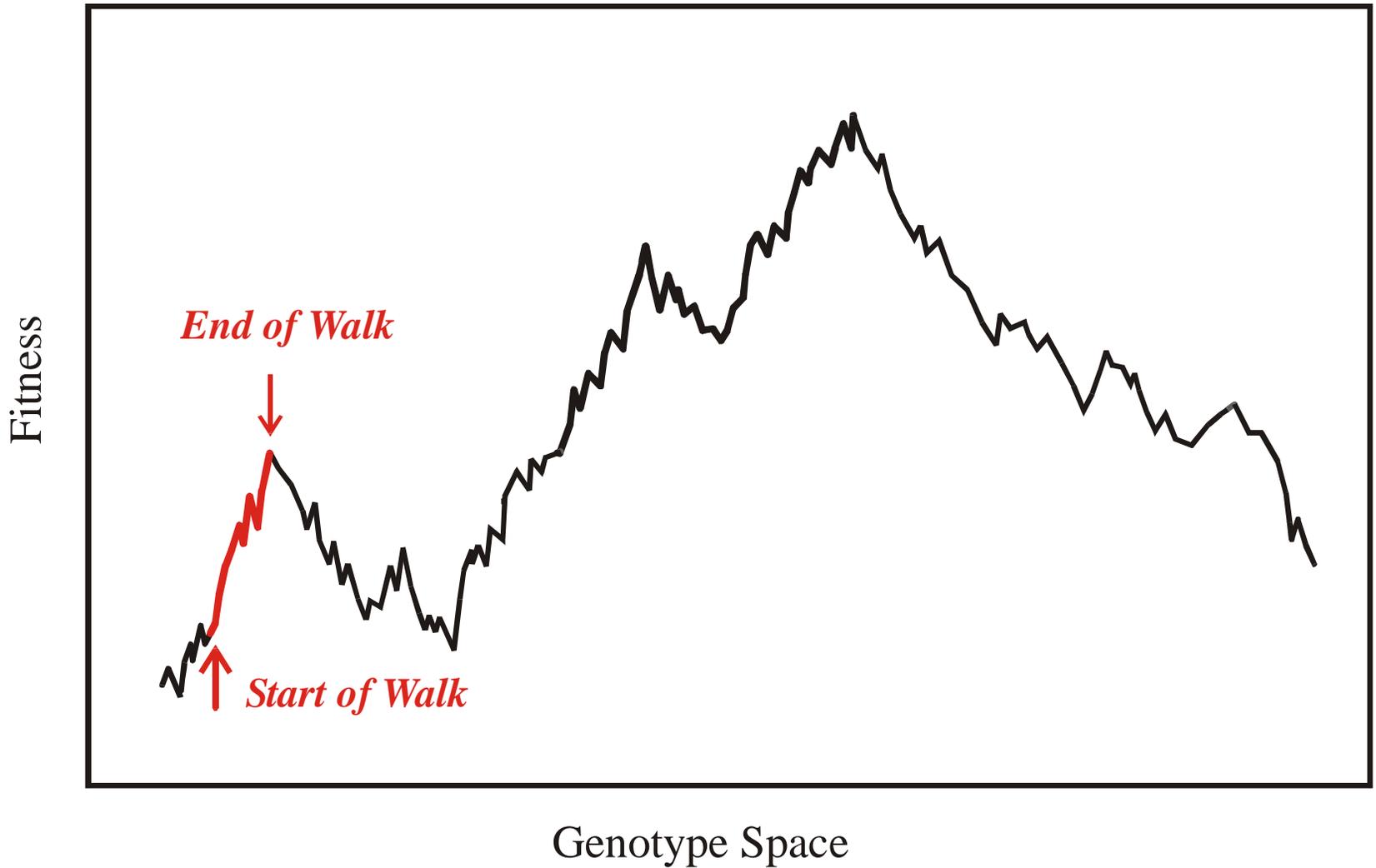
Evolutionary optimization in absence of neutral paths in sequence space
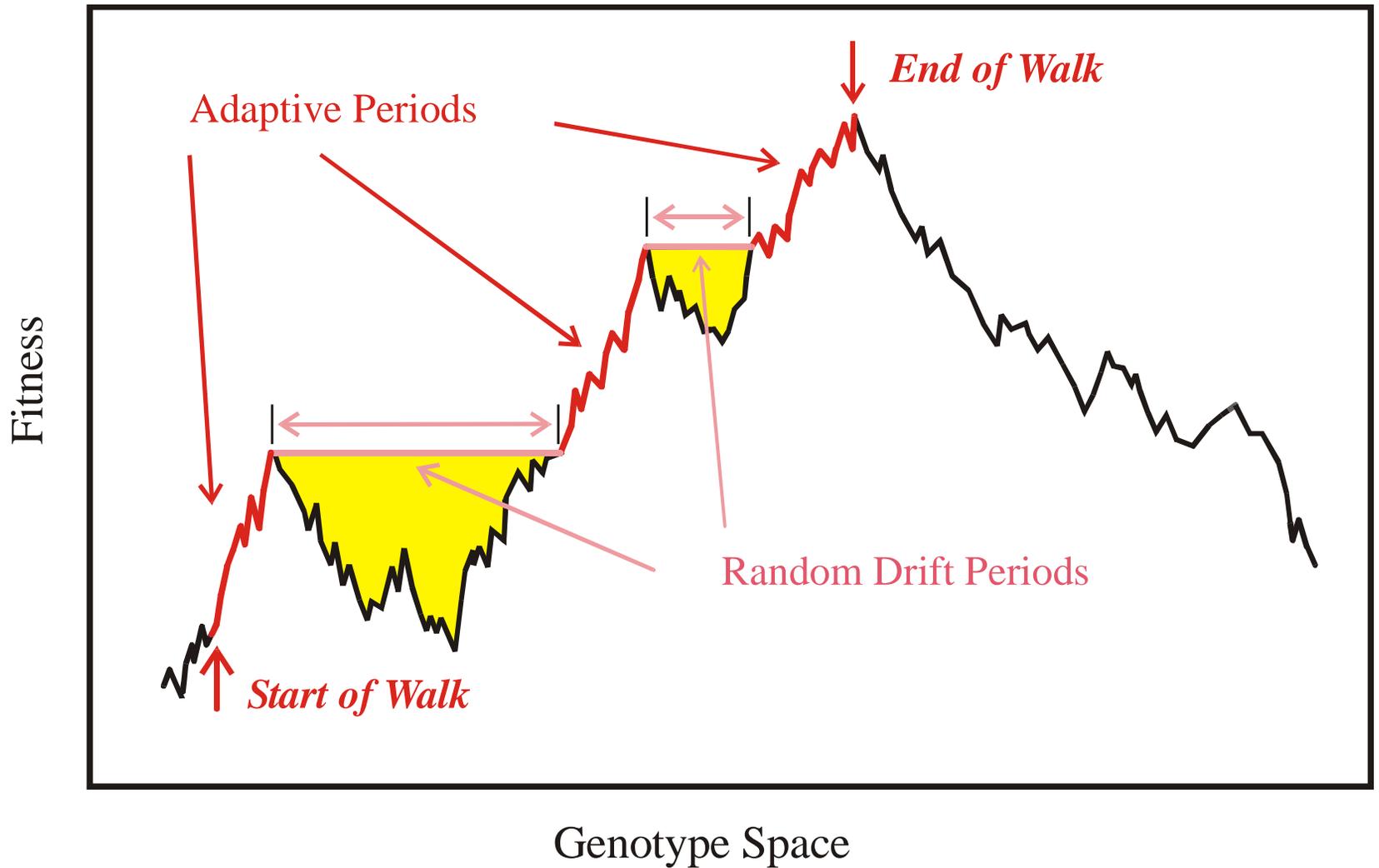
Evolutionary optimization including neutral paths in sequence space

Grand Canyon

Example of a landscape on Earth with 'neutral'
ridges and plateaus

1. Evolution experiments

2. Molecular evolution of RNA

3. Neutral networks

4. Evolutionary optimization of RNA structure

5. Kinetic structures and switching molecules

One sequence - one structure

Many suboptimal structures
Partition function

Metastable structures
Conformational switches

Free Energy

S_{10} S_8
S_9 S_7
S_5 S_6
S_4
S_3

S_2

S_1
S_0

S_0

S_1

S_0

Minimum free energy structure

Suboptimal structures

Kinetic structures

RNA secondary structures derived from a single sequence

The Folding Algorithm

A sequence I specifies an energy ordered set of compatible structures $\mathcal{S}(I)$:

$$\mathcal{S}(I) = \{S_0, S_1, \ldots, S_m, O\}$$

A trajectory $\mathfrak{T}_k(I)$ is a time ordered series of structures in $\mathcal{S}(I)$. A folding trajectory is defined by starting with the open chain O and ending with the global minimum free energy structure S_0 or a metastable structure S_k which represents a local energy minimum:

$$\mathfrak{T}_0(I) = \{O, S(1), \ldots, S(t-1), S(t), \\ S(t+1), \ldots, S_0\}$$
$$\mathfrak{T}_k(I) = \{O, S(1), \ldots, S(t-1), S(t), \\ S(t+1), \ldots, S_k\}$$

Master equation

$$\frac{dP_k}{dt} = \sum_{i=0}^{m+1}\left(P_{ik}(t) - P_{ki}(t)\right) = \sum_{i=0}^{m+1} k_{ik} P_i - P_k \sum_{i=0}^{m+1} k_{ki}$$

$$k = 0, 1, \ldots, m+1$$

Transition probabilities $P_{ij}(t) = \mathcal{Prob}\{S_i \rightarrow S_j\}$ are defined by

$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t)\, \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

$$P_{ji}(t) = P_j(t)\, k_{ji} = P_j(t)\, \exp(-\Delta G_{ji}/2RT) / \Sigma_j$$

$$\Sigma_k = \sum_{k=1, k\neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

The symmetric rule for transition rate parameters is due to Kawasaki (K. Kawasaki, *Diffusion constants near the critical point for time depen-dent Ising models*. Phys.Rev. **145**:224-230, 1966).

Formulation of kinetic RNA folding as a stochastic process

Nucleation

Base pair formation

Base pair cleavage

Elongation

Base pair formation
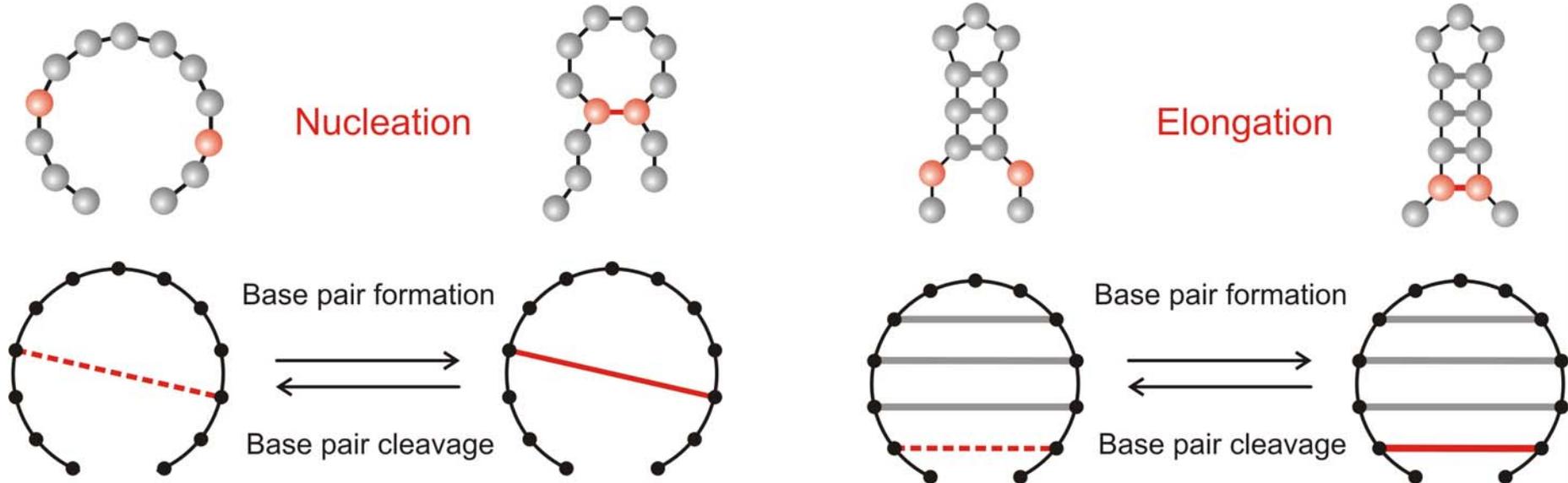
Base pair cleavage

Corresponds to base pair distance: $d_P(S_1, S_2)$

Base pair formation and base pair cleavage moves for nucleation and elongation of stacks
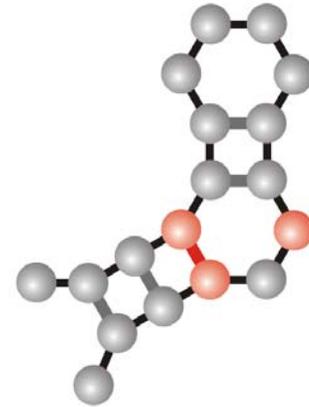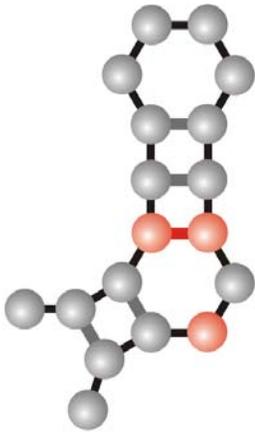
Base pair shift

Class 1
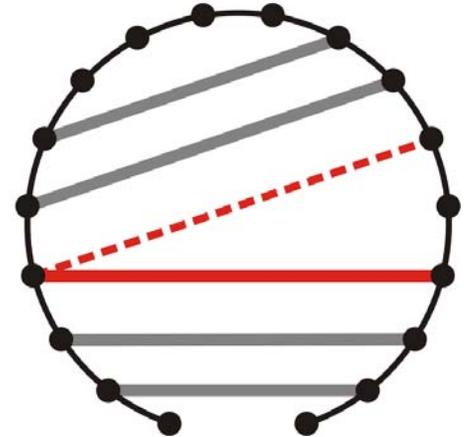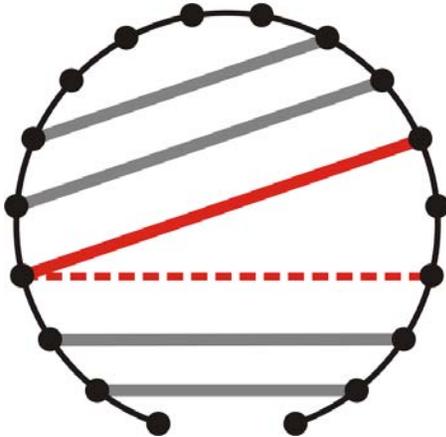
Base pair closure, opening and shift corresponds to Hamming distance: $d_H(S_1, S_2)$

Base pair shift move of class 1: Shift inside internal loops or bulges

Free energy ΔG^0

$S_5^{(h)}$

$S_4^{(h)}$

$S_3^{(h)}$

$S_1^{(h)}$

$S_2^{(h)}$

$S_8^{(h)}$

$S_7^{(h)}$

$S_9^{(h)}$

$S_6^{(h)}$

Suboptimal conformations
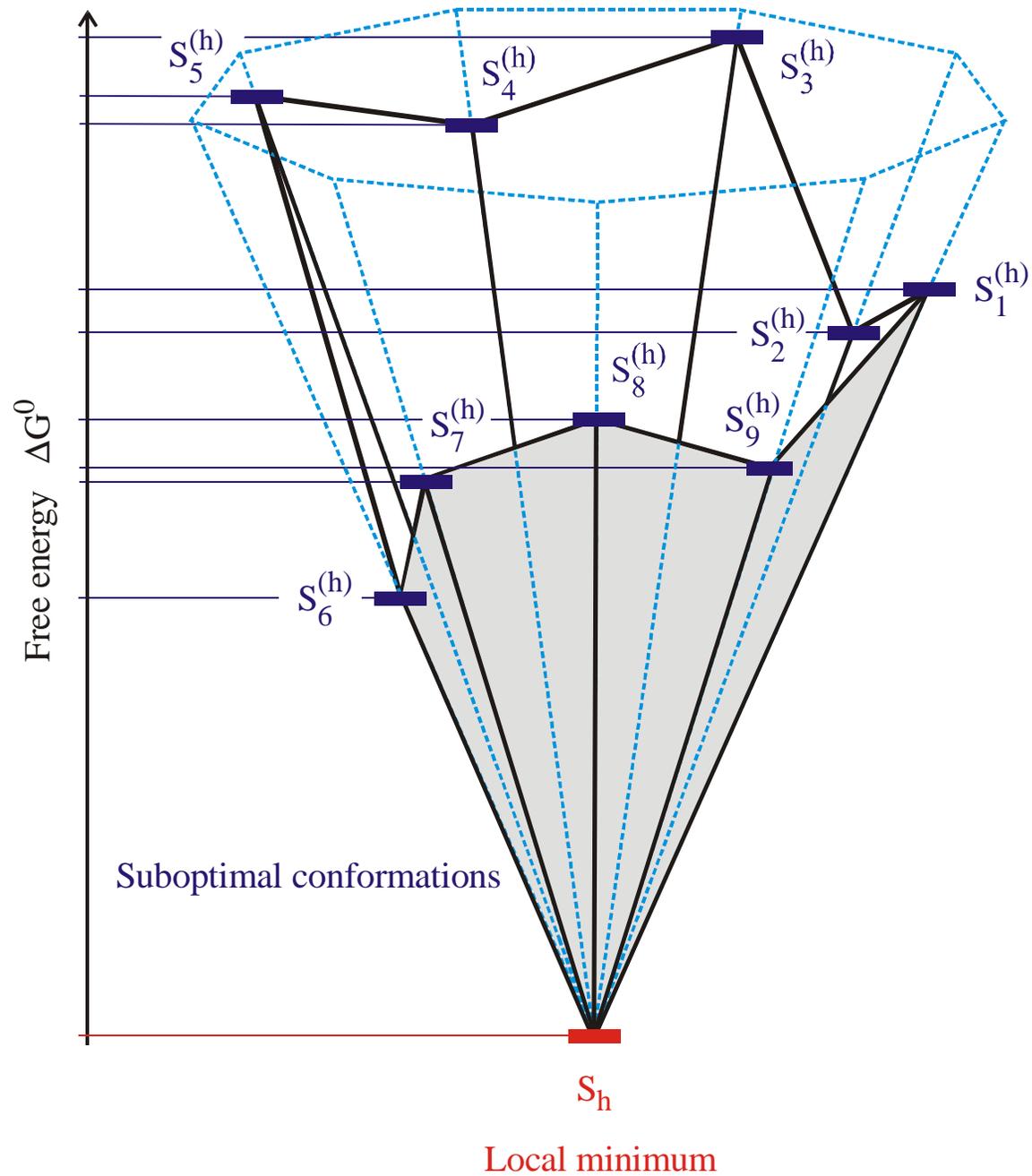
Search for local minima in conformation space

S_h

Local minimum

Free energy ΔG^0

Saddle point $T_{\{k}$

S_k

$S_\{$

"Reaction coordinate"

Free energy ΔG^0

$T_{\{k}$

$S_\{$

S_k
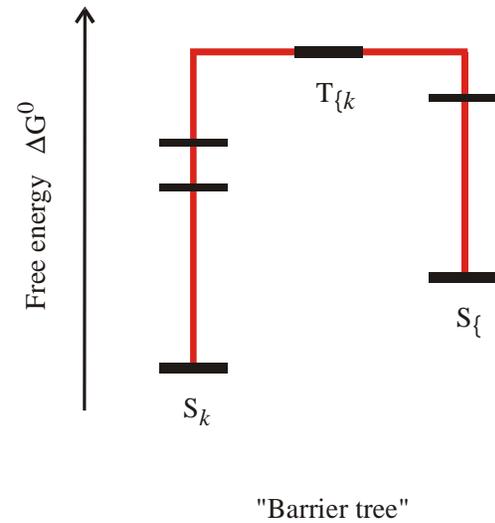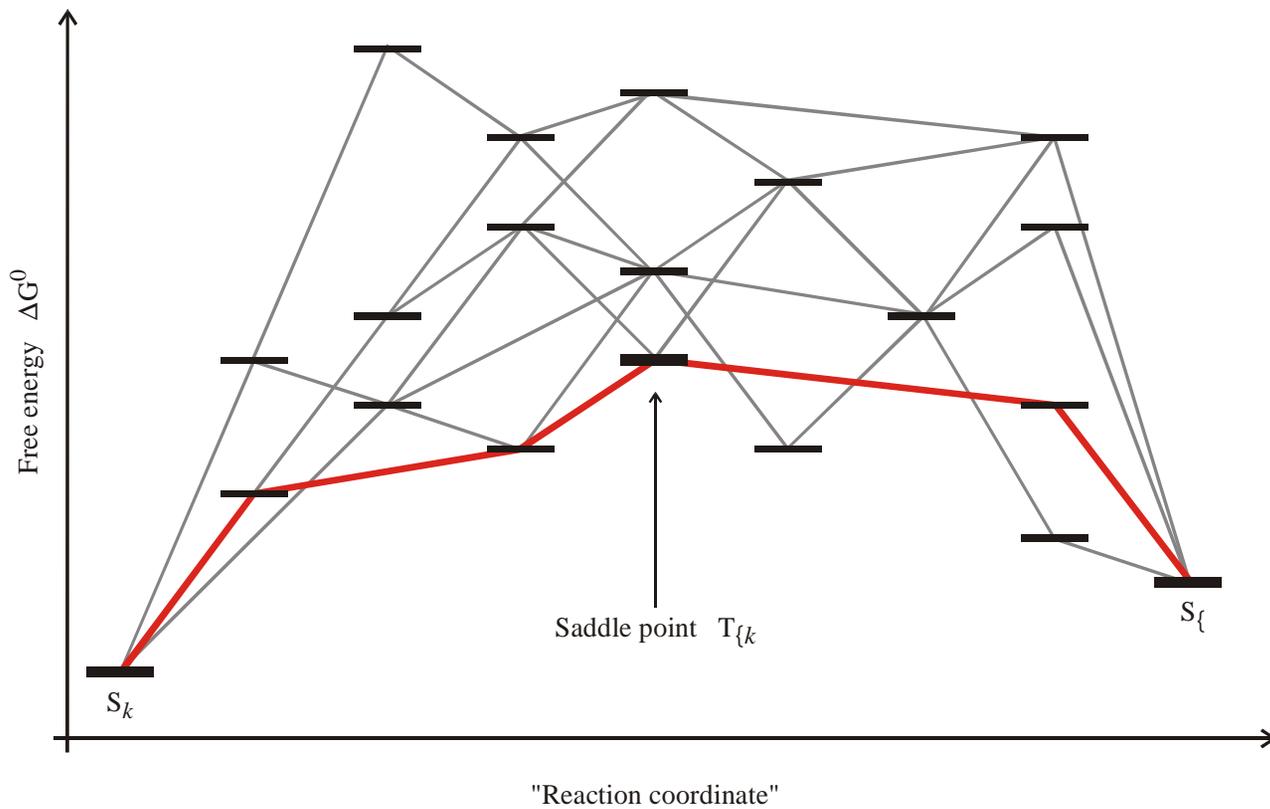
"Barrier tree"

Definition of a ‚barrier tree'

```
CUGCGGCUUUGGCUCUAGCC
....(((.........))))    -4.30
(((.(((....))).)))..    -3.50
(((..((....))..)))..    -3.10
...........(((....)))    -2.80
..((((....)))...)).     -2.20
....(((..........)))     -2.20
((..(((....)))..))..    -2.00
..((.((....))....)).     -1.60
....(((....)))......     -1.60
.....(((.........))).    -1.50
.((.(((....))).))...    -1.40
....(((((..(...).))))    -1.40
.((..((....))..))...    -1.00
(((.(((....)).))))..    -0.90
(((.((......)).)))..    -0.90
....(((((..(....)))))    -0.80
......((....))......     -0.80
..(.(((....))))....     -0.60
....(((....)).).....     -0.60
(((..(......)..)))..    -0.50
..(((((....)).).))).     -0.50
..(.(((....))).)....    -0.40
..((.......))......     -0.30
............((.....))    -0.30
............((....)).    -0.30
(((.(((....)))).))..    -0.20
....(((.(.......)))))    -0.20
....(((..((....)))))    -0.20
..(..((....))..)....     0.00
...................     0.00
.(..(((....)))..)...     0.10
```


M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:
Math.Gen.* **37**:4731-4741.

```
CUGCGGCUUUGGCUCUAGCC
....((((........)))) -4.30
(((.(((....))).))).. -3.50
(((..((....))..))).. -3.10
..........(((....))) -2.80
..(((((....)))...)). -2.20
....(((.........))) -2.20
((..((((....)))..)).. -2.00
..((.((....))....)). -1.60
....(((....)))...... -1.60
.....(((.......))). -1.50
.((.(((....))).))... -1.40
....(((((..(...).)))) -1.40
.((..((....))..))... -1.00
(((.(((...)).)))).. -0.90
(((.((......)).))).. -0.90
....((((..(....))))) -0.80
.....((....))...... -0.80
..(.(((....)))).. -0.60
....(((.....).)... -0.60
(((..(......)..))).. -0.50
..(((((....)).).)). -0.50
..(.(((....))).)... -0.40
..((......))...... -0.30
..........((......)) -0.30
...........((...)). -0.30
(((.(((....))).))).. -0.20
....(((.(......)))) -0.20
....(((..((....)))) -0.20
..(..((....))..)... 0.00
.................... 0.00
.(..(((....)))..)... 0.10
```

ΔG [kcal/mole]

M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:
Math.Gen.* **37**:4731-4741.

ΔG [kcal/mole]
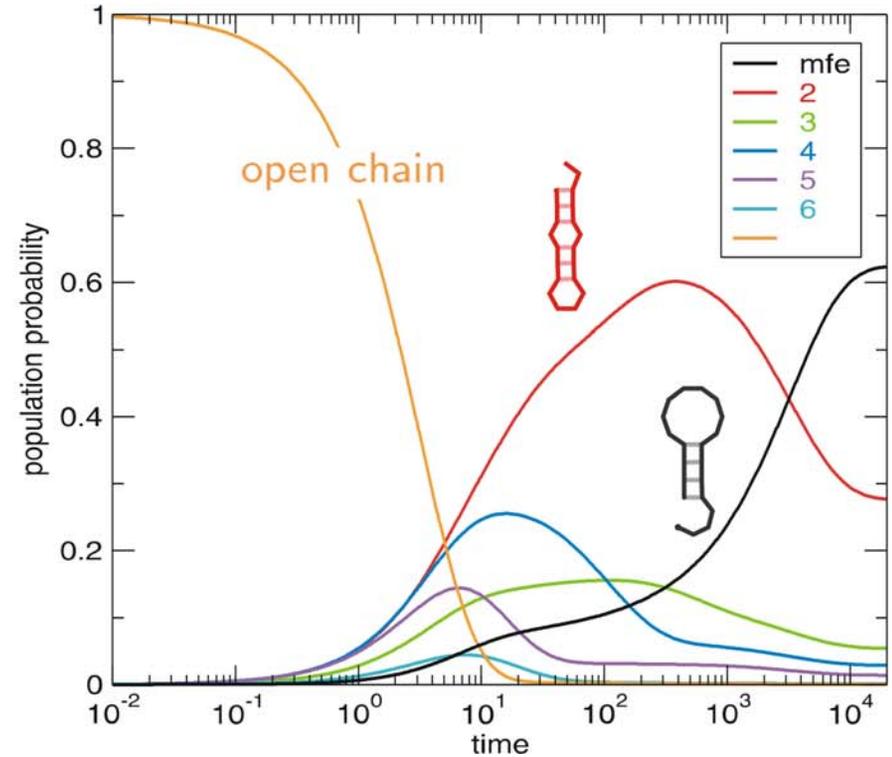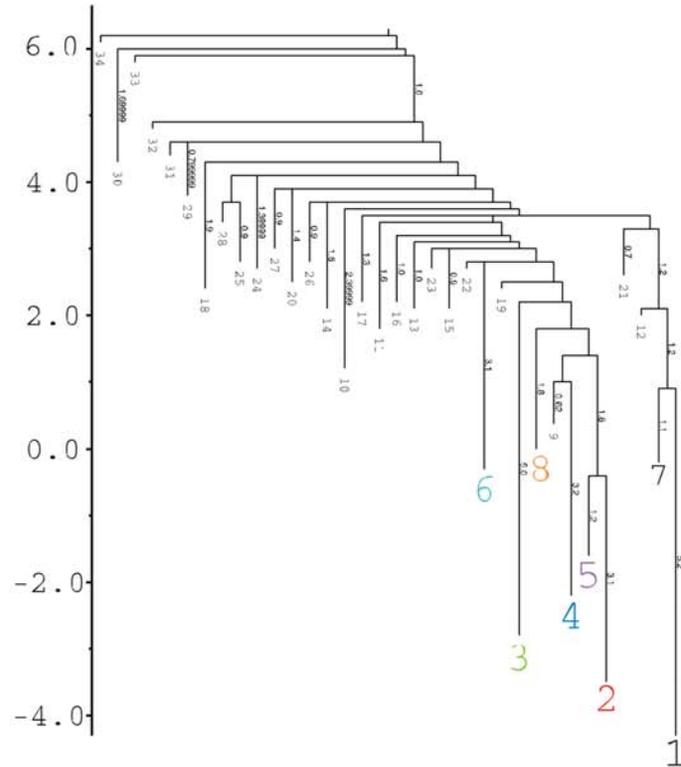
Arrhenius kinetics

M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:
Math.Gen.* **37**:4731-4741.
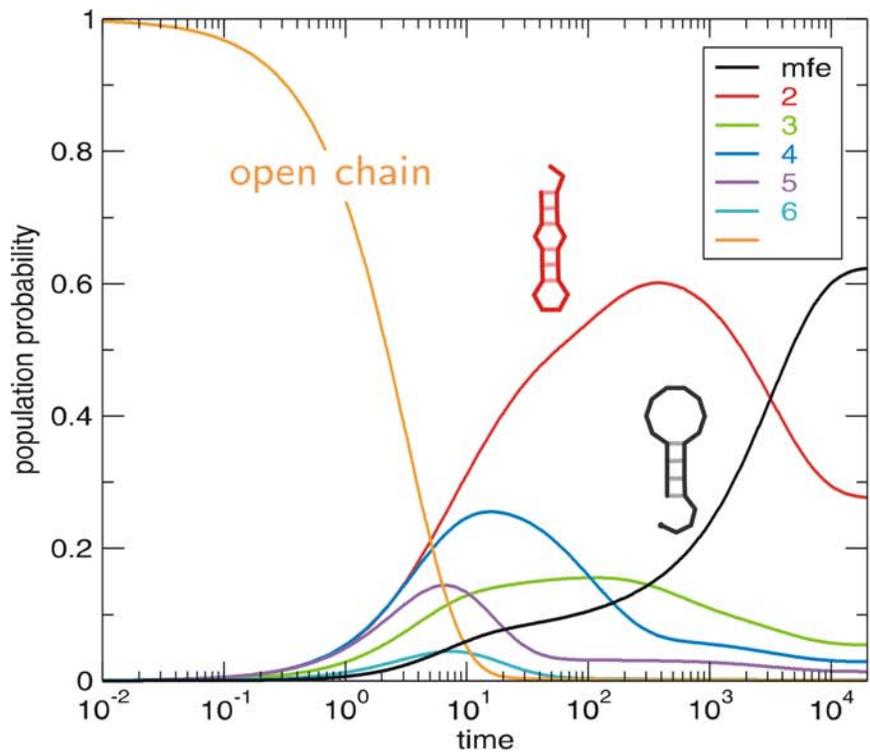
Arrhenius kinetic
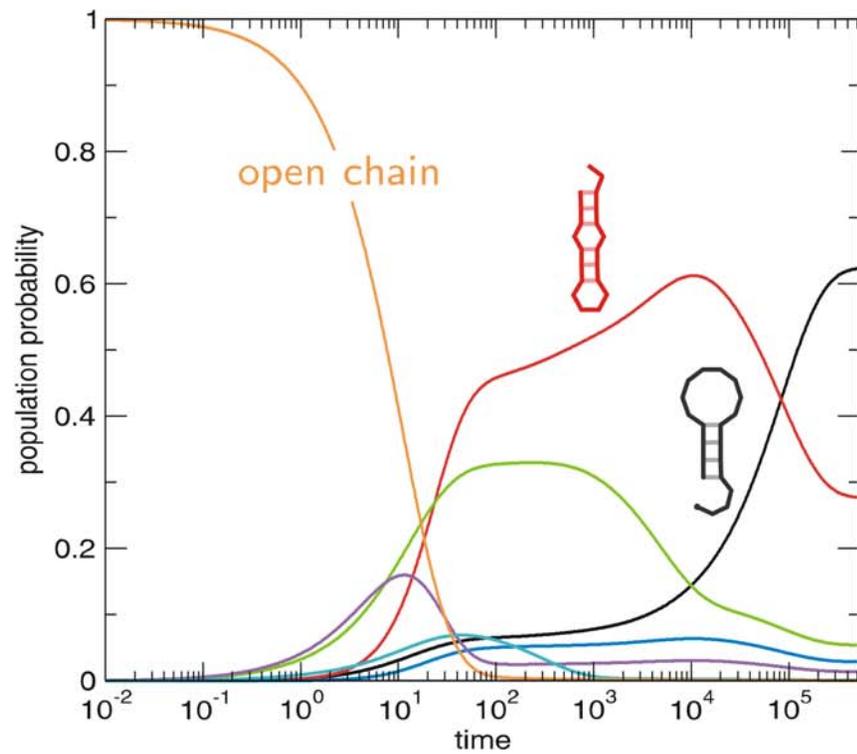
Exact solution of the master equation

M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:
Math.Gen.* **37**:4731-4741.

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES[1]

■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡
and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†Los Alamos National Laboratory,
Los Alamos, NM 87545, U.S.A.

‡Institut für Theoretische Chemie der Universität Wien,
A-1090 Wien, Austria

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

(*E.mail: pks@tbi.univie.ac.at*)

Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. Sequences folding into identical structures form *neutral networks*. A neutral network is embedded in the set of sequences that are *compatible* with the given structure. Networks are modeled as graphs and constructed by random choice of vertices from the space of compatible sequences. The theory characterizes neutral networks by the mean fraction of neutral neighbors (λ). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ($\lambda > \lambda^*$). Below threshold ($\lambda < \lambda^*$), the networks are partitioned into a largest "giant" component and several smaller components. Structures are classified as "common" or "rare" according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of *shape space covering* sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. © 1997 Society for Mathematical Biology

THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s' *be arbitrary secondary structures and* $\mathbf{C}[s], \mathbf{C}[s']$ *their corresponding compatible sequences. Then,*

$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \varnothing.$$

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 $\jmath(s, s') \cong D_m$ 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.

Reference for the definition of the intersection and the proof of the **intersection theorem**
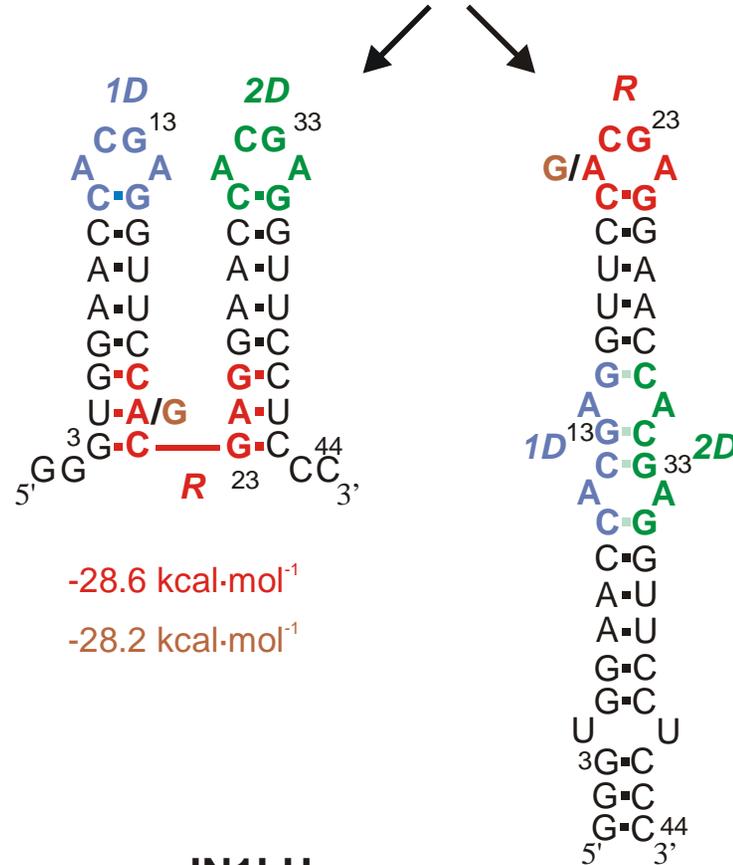
An RNA switch

JN1LH

J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke,
M.H. de Smit, P. Schuster, and C.W.A. Pleij.

J1LH barrier tree

A ribozyme switch

E.A.Schultes, D.B.Bartel, Science
289 (2000), 448-452
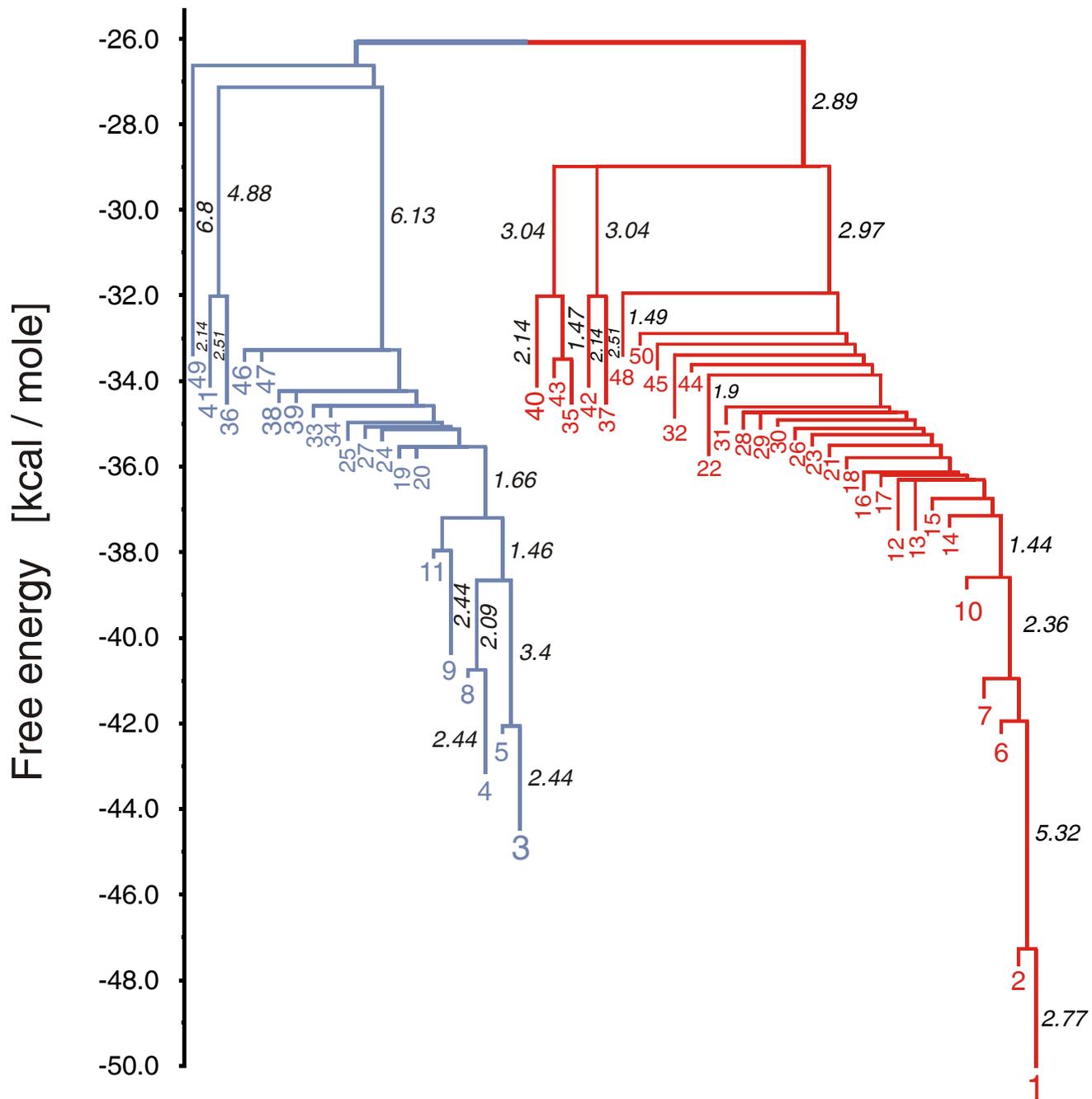
minus the background levels observed in the HSP in the control (Sar1-GDP–containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.

46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz et al., *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton

X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH₃Cl and separated by SDS–polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin et al., *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).

62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui et al., *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson et al., *J. Cell Biol.* **143**, 319 (1998).
69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

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 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 same folded 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 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 disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity 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 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*).

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, 5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would 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 could arise from previously existing ribozymes, without the need to carry 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 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 (*9*). 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 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (*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 antigenomic HDV ribozyme (*12*), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (*13, 14*).

The prototype class III and HDV ribozymes have no more than 25% sequence identity 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

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)
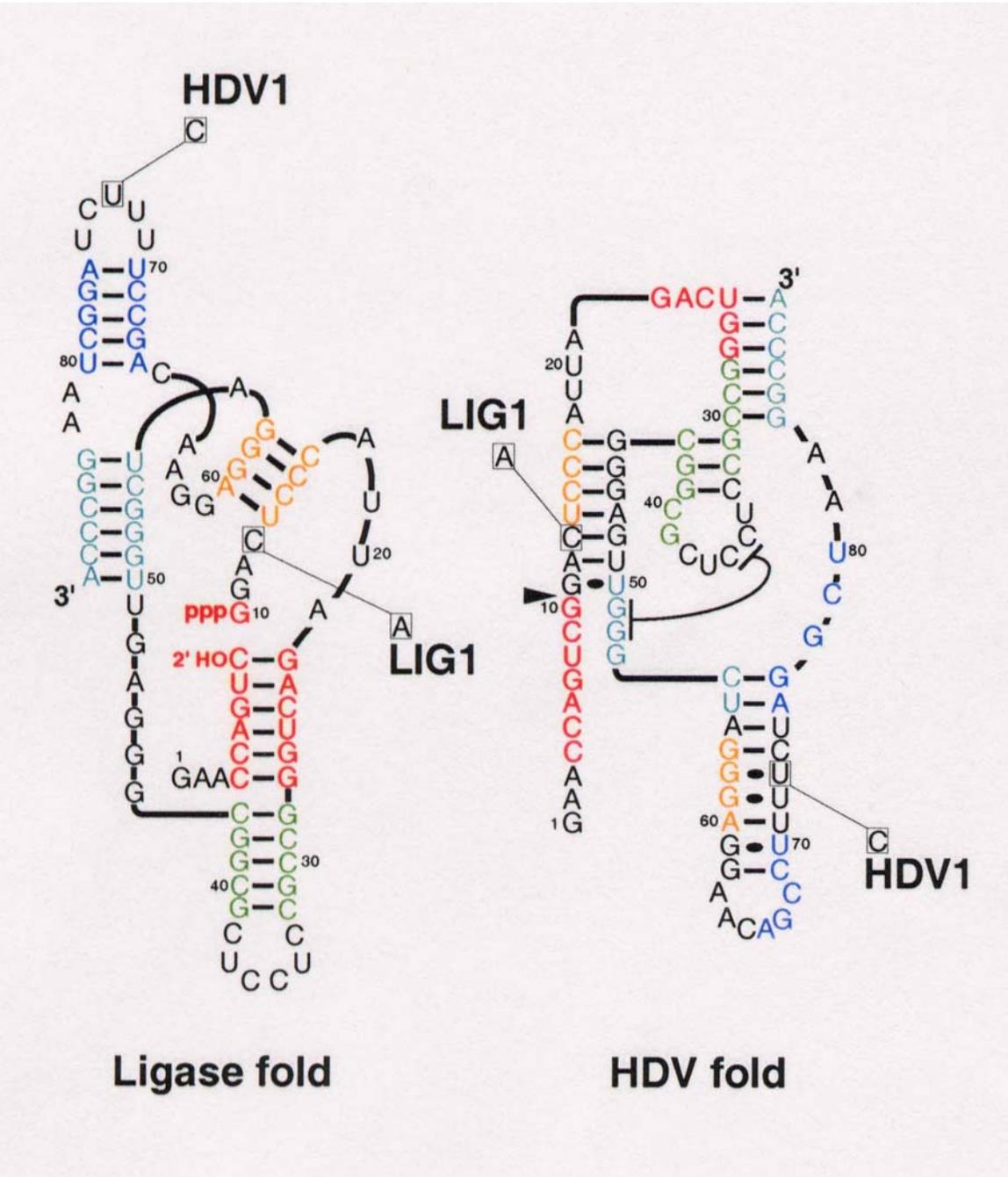
The sequence at the *intersection*:

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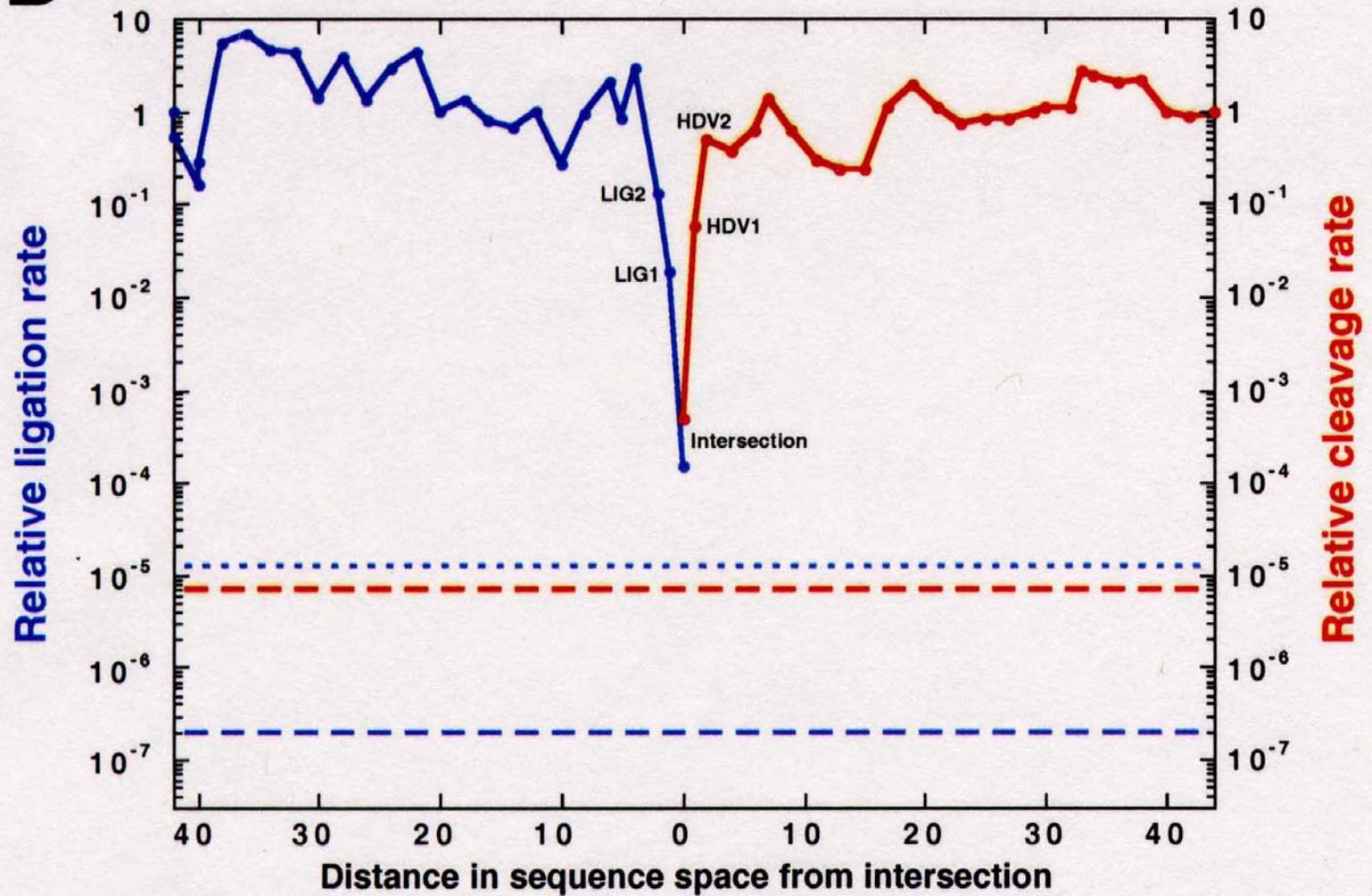
Two neutral walks through sequence space with conservation of structure and catalytic activity

Acknowledgement of support

Universität Wien

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 Machne,**
Ulrike Mückstein, Hakim Tafer, Thomas Taylor, Universität Wien, AT

Web-Page for further information:

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