Modeling Molecular Evolution

The Origin of Information and Learning in Populations

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Agent Based Modeling and Simulation

Minneapolis, 03.– 06.11.2003

Web-Page for further information:

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

1. **RNA structure, replication kinetics, and origin of information**

2. **Evolution *in silico* and optimization of RNA structures**

3. **Random walks and ‚ensemble learning‘**

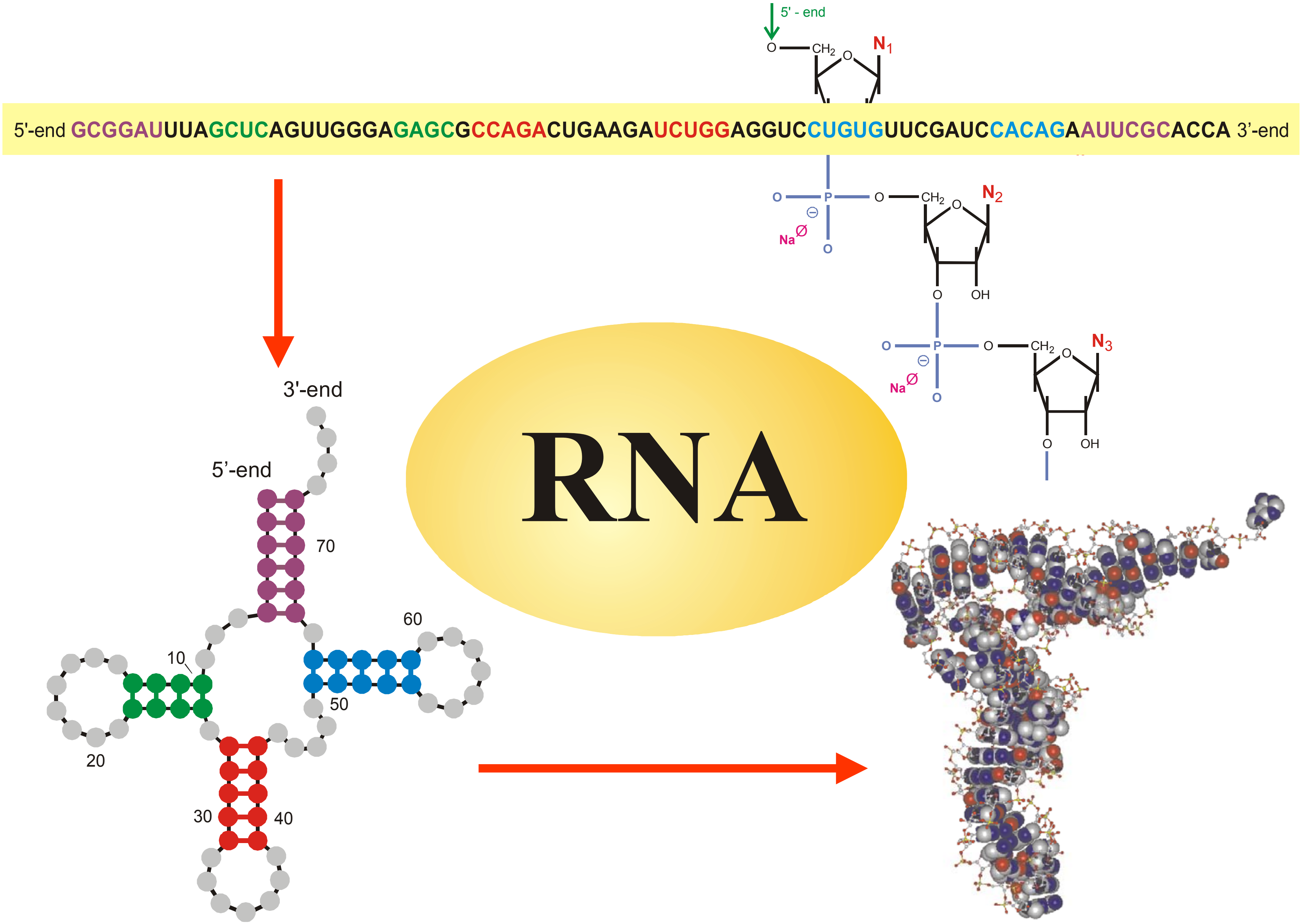
4. **Sequence-structure maps, neutral networks, and intersections**

1. **RNA structure, replication kinetics, and origin of information**

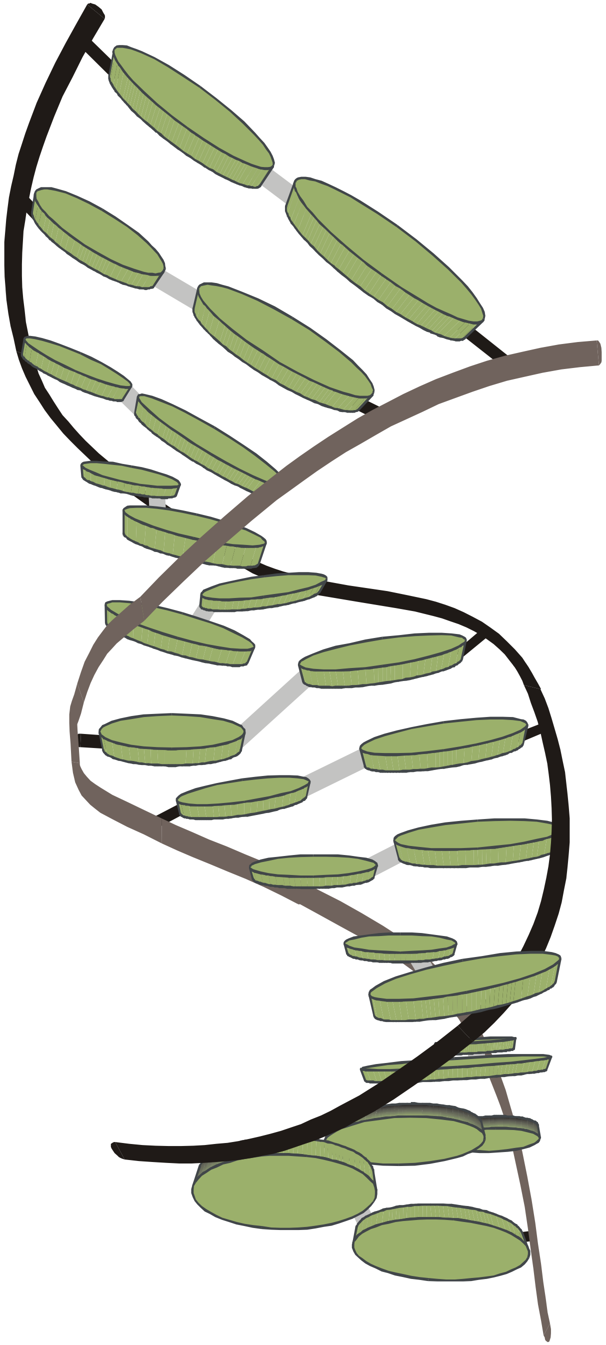
2. Evolution *in silico* and optimization of RNA structures

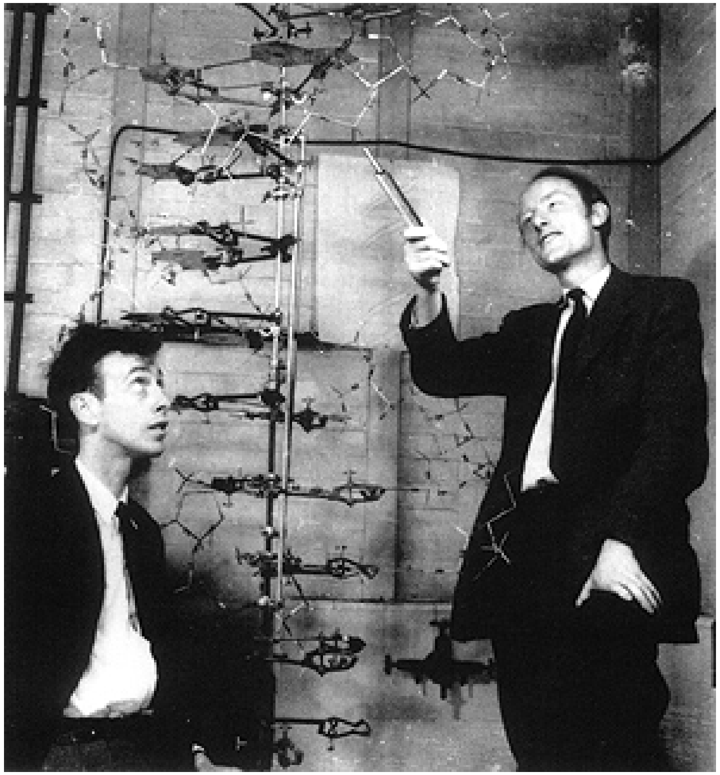
3. Random walks and ‚ensemble learning‘

4. Sequence-structure maps, neutral networks, and intersections


5' - end
N1
O
CH2
O
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end
O
P
O
Na
O
CH2
O
N2
O
OH
O
P
O
Na
O
CH2
O
N3
O
OH
RNA
3'-end
5'-end
70
60
10
50
20
30
40

Definition of RNA structure

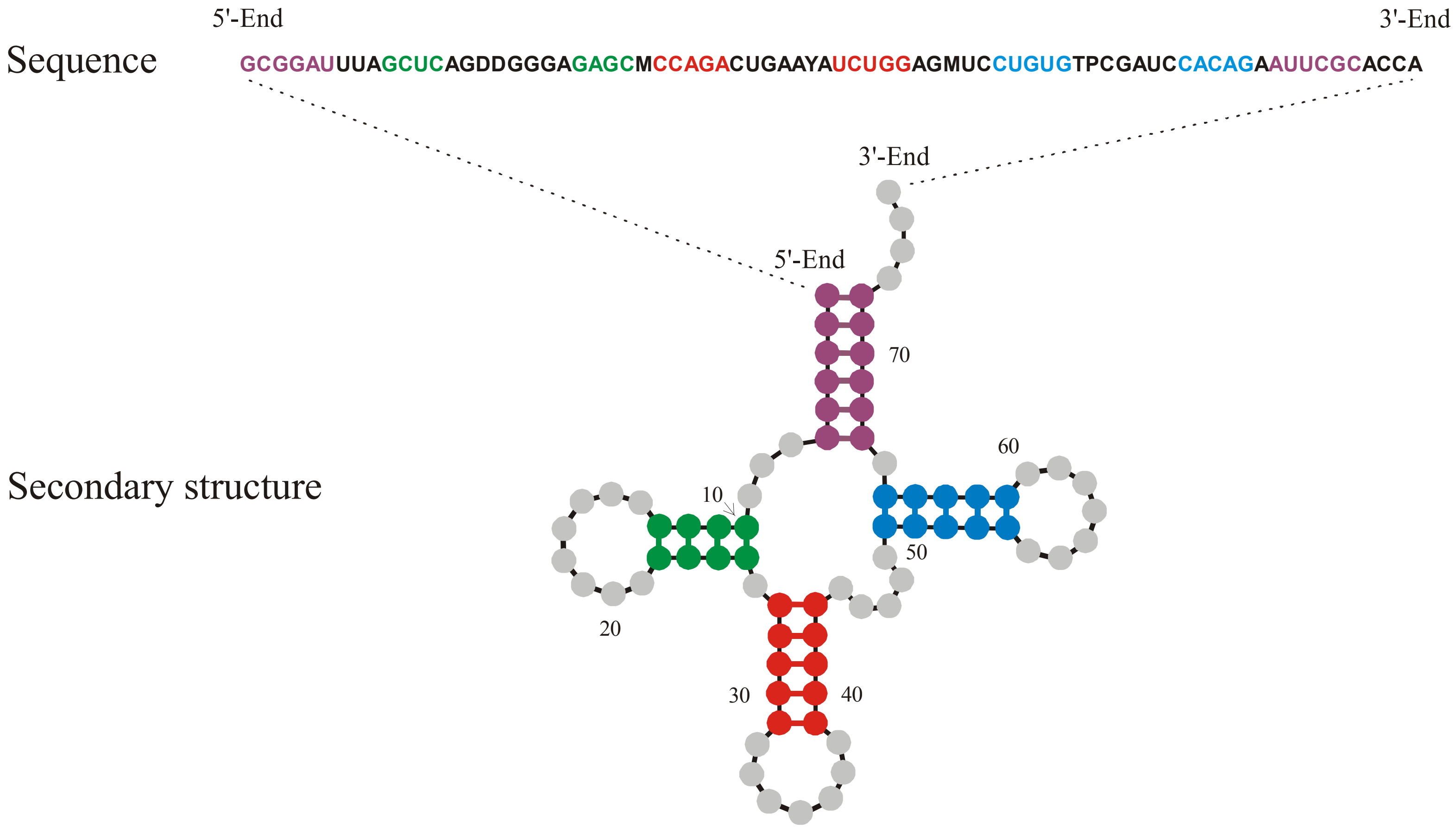


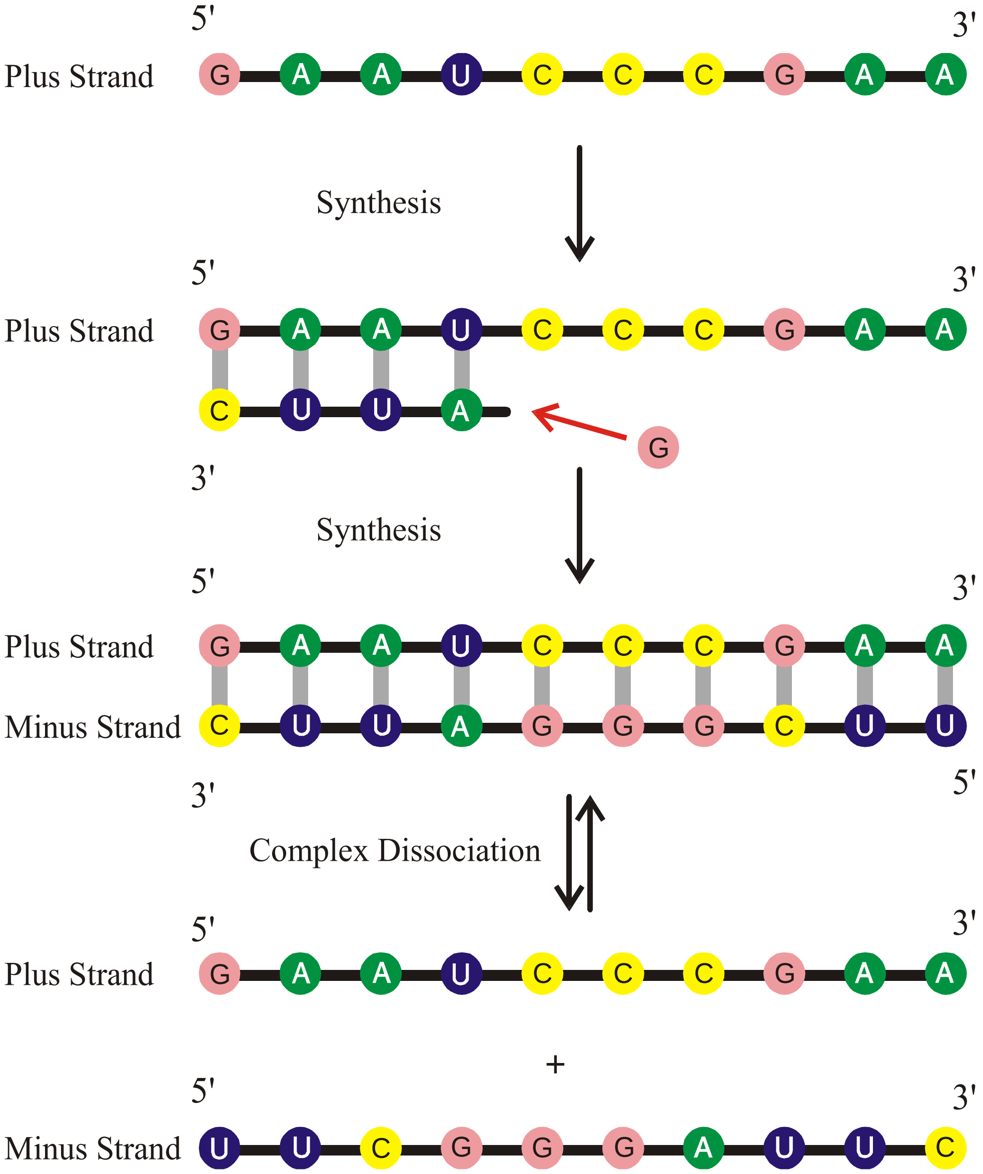


James D. Watson, 1928- , and Francis Crick, 1916- ,
Nobel Prize 1962

1953 – 2003 fifty years double helix

The three-dimensional structure of a short double helical stack of B-DNA


Sequence
5'-End
3'-End
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA
Secondary structure
3'-End
5'-End
70
60
10
50
20
30
40



Complementary replication as the simplest copying mechanism of RNA
Complementarity is determined by Watson-Crick base pairs:

GÍ**C** and **A**=**U**

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

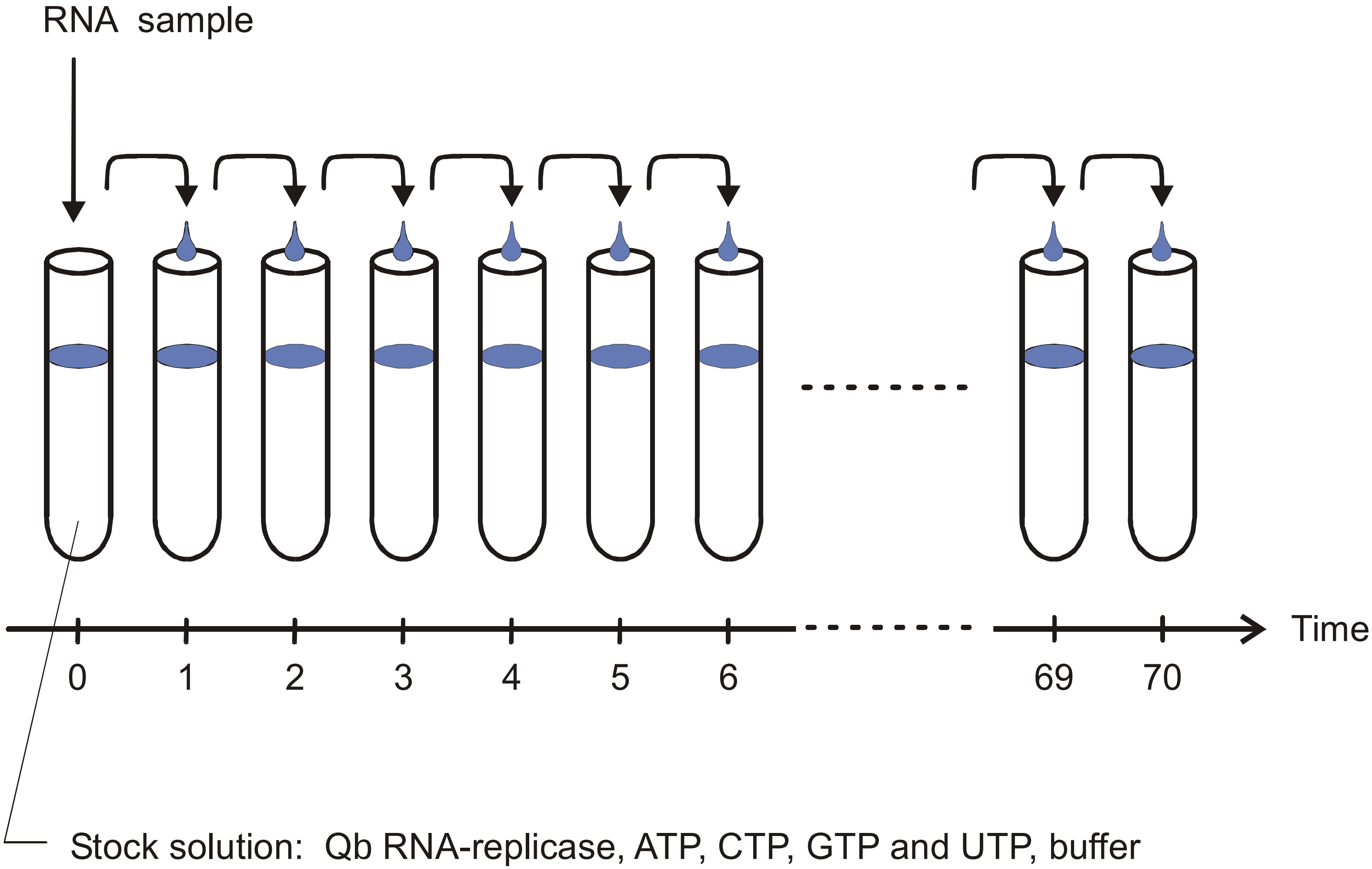
S.Spiegelman, ***An approach to the experimental analysis of precellular evolution***. Quart.Rev.Biophys. **4** (1971), 213-253

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

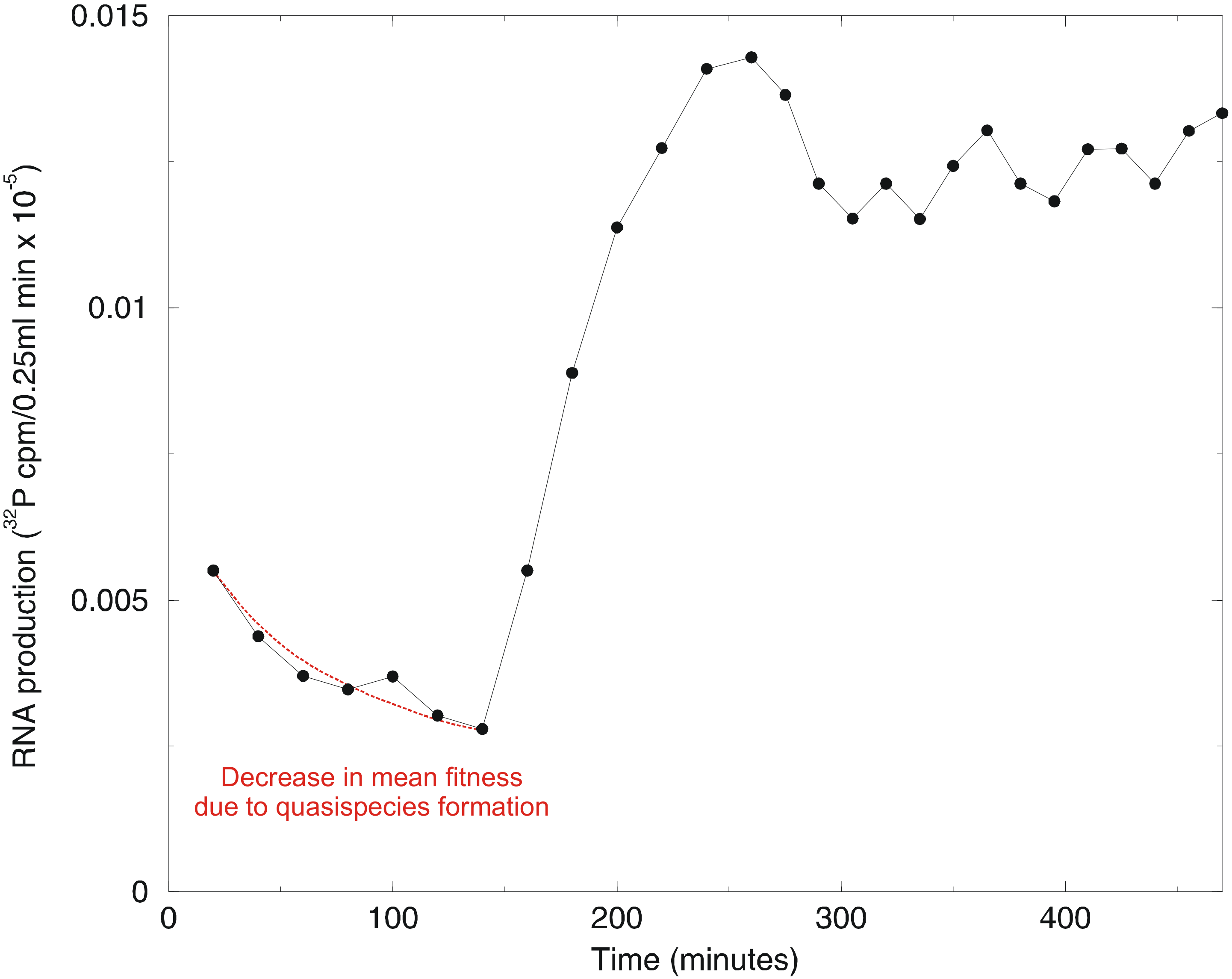
G.Bauer, H.Otten, J.S.McCaskill, ***Travelling waves of*** **in vitro** ***evolving RNA.*** *Proc.Natl.Acad.Sci.USA* **86** (1989), 7937-7941

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


RNA sample
0
1
2
3
4
5
6
69
70
Time
Stock solution: Qb RNA-replicase, ATP, CTP, GTP and UTP, buffer

The serial transfer technique applied to RNA evolution ***in vitro***



The increase in RNA production rate during a serial transfer experiment

No new principle will declare itself from below a heap of facts.

Sir Peter Medawar, 1985

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

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

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

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

$$(A) + I_n \xrightarrow{f_n} I_n + I_n$$

$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

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

$[I_i] = x_i$ Æ 0 ; $i = 1,2,\ldots,n$;

$[A] = a =$ constant

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

$x_m(t)$ Á 1 for t Á ¸

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

Selection equation: $[\mathrm{I}_i] = x_i$ Æ 0 , $f_i > 0$

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

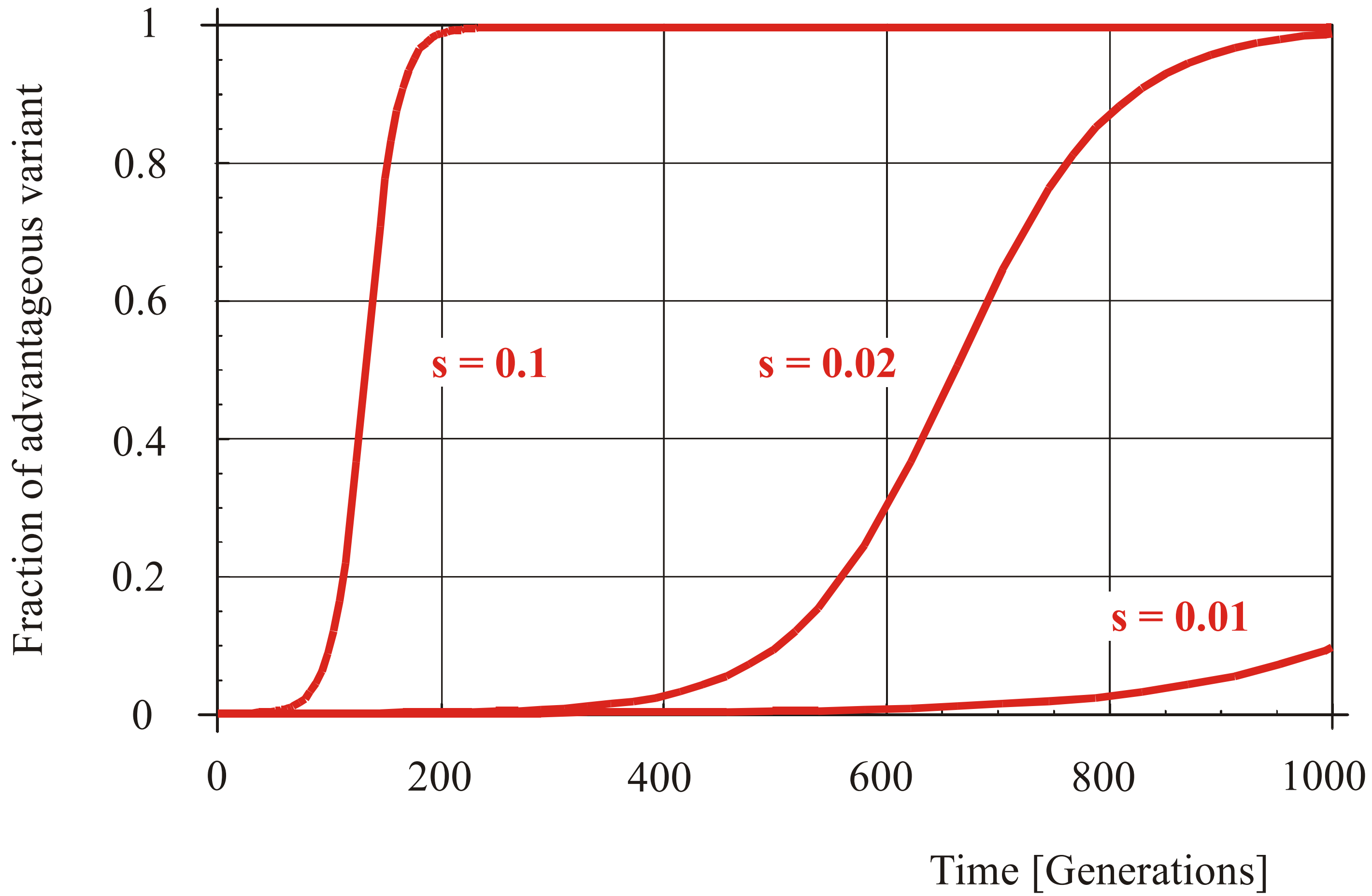
Mean fitness or dilution flux, ϕ (t), is a **non-decreasing function** of time,

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

Solutions are obtained by integrating factor transformation

$$x_i(t) = \frac{x_i(0)\cdot \exp\left(f_i t\right)}{\sum\nolimits_{j=1}^{n} x_j(0)\cdot \exp\left(f_j t\right)}; \quad i = 1,2,\cdots,n$$

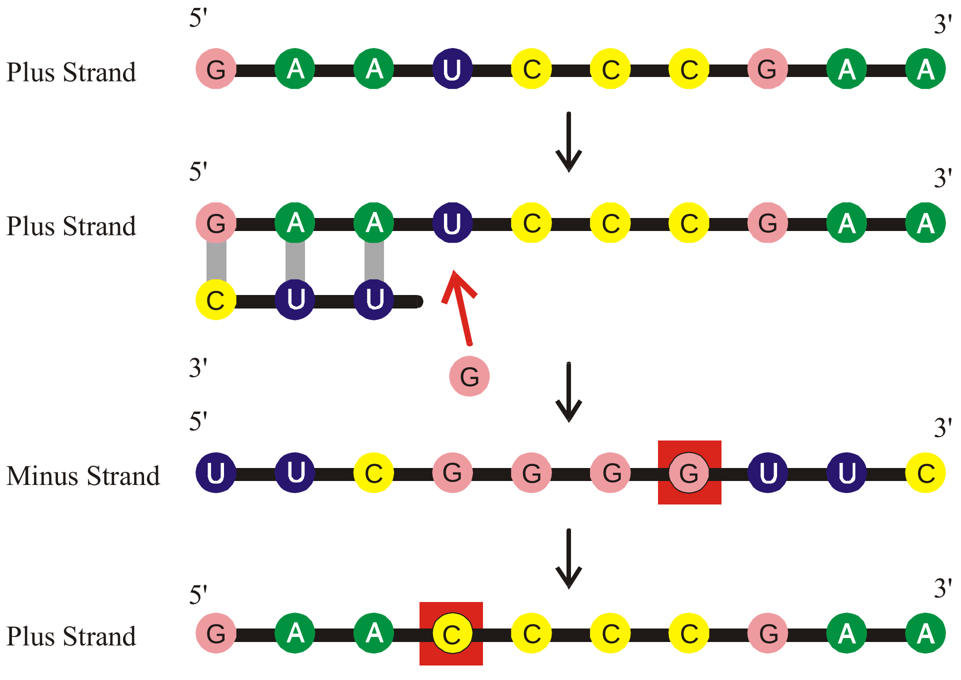
$\mathbf{s} = (f_2 - f_1) / f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/N\ ;\ x_2(0) = 1/N$



Selection of advantageous mutants in populations of N = 10 000 individuals

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

GAAUCCCGAA → GAAUCCCGUCCCGAA

Insertion

GAAUCCCGAA → GAAUCCA

Deletion

The origins of changes in RNA sequences are **replication errors** called **mutations**.

Theory of molecular evolution

M.Eigen, ***Self-organization of matter and the evolution of biological macromolecules***. Naturwissenschaften **58** (1971), 465-526

C.J. Thompson, J.L. McBride, ***On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules***. Math. Biosci. **21** (1974), 127-142

B.L. Jones, R.H. Enns, S.S. Rangnekar, ***On the theory of selection of coupled macromolecular systems.*** Bull.Math.Biol. **38** (1976), 15-28

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle***. Naturwissenschaften **58** (1977), 465-526

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle***. Naturwissenschaften **65** (1978), 7-41

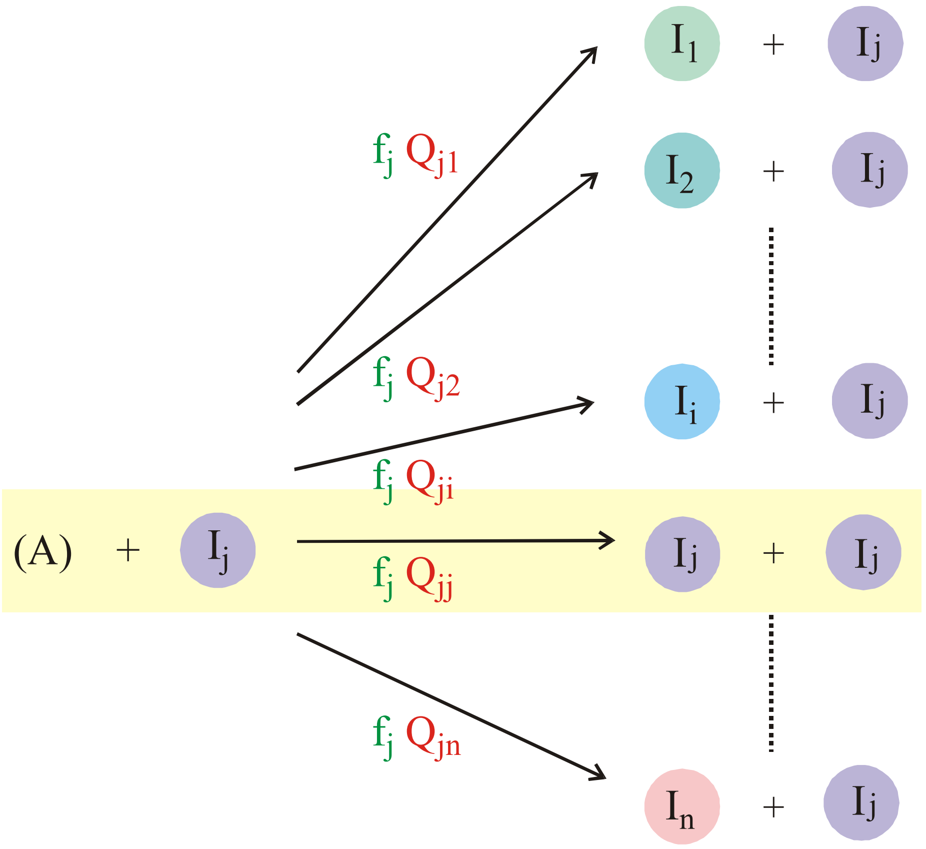
M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle***. Naturwissenschaften **65** (1978), 341-369

J. Swetina, P. Schuster, ***Self-replication with errors - A model for polynucleotide replication.*** Biophys.Chem. **16** (1982), 329-345

J.S. McCaskill, ***A localization threshold for macromolecular quasispecies from continuously distributed replication rates***. J.Chem.Phys. **80** (1984), 5194-5202

M.Eigen, J.McCaskill, P.Schuster, ***The molecular quasispecies***. Adv.Chem.Phys. **75** (1989), 149-263

C. Reidys, C.Forst, P.Schuster, ***Replication and mutation on neutral networks***. Bull.Math.Biol. **63** (2001), 57-94



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

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

$[I_i] = x_i$ Æ 0 ; $i = 1,2,\ldots,n$;

$[A] = a =$ constant

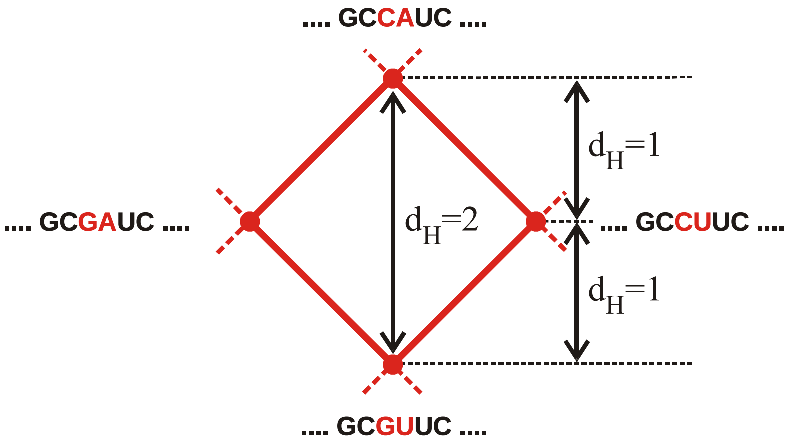
$$Q_{ij} = (1-p)^{\ell - d(i,j)}\, p^{d(i,j)}$$

p .... Error rate per digit

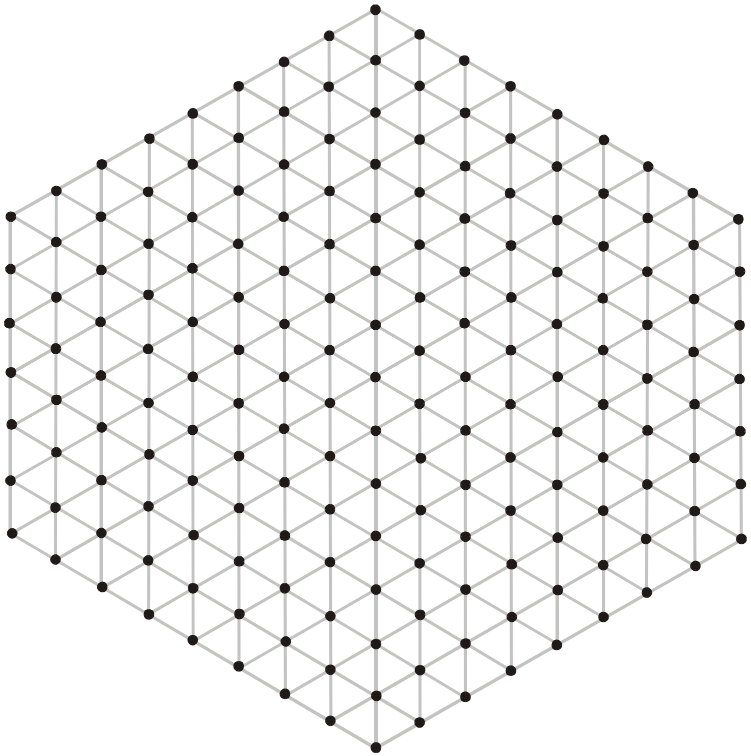
ℓ .... Chain length of the polynucleotide

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

Chemical kinetics of replication and mutation as parallel reactions

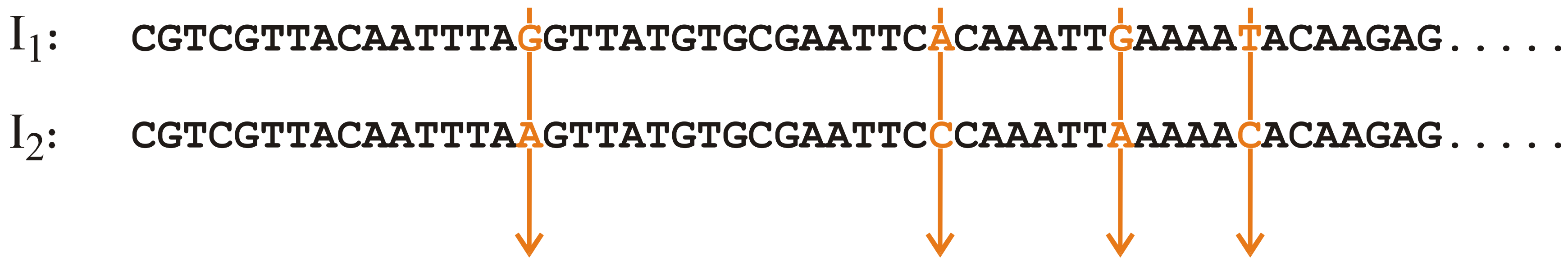


City-block distance in sequence space



2D Sketch of sequence space

Single point mutations as moves in sequence space



I_1: CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....

I_2: CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAAACACAAGAG.....

Hamming distance $d_H(I_1,I_2) = 4$

(i) $d_H(I_1,I_1) = 0$

(ii) $d_H(I_1,I_2) = d_H(I_2,I_1)$

(iii) $d_H(I_1,I_3)$ ¶ $d_H(I_1,I_2) + d_H(I_2,I_3)$

The Hamming distance between sequences induces a metric in sequence space

Mutation-selection equation: $[\mathrm{I}_i] = x_i$ Æ $0,\ f_i > 0,\ Q_{ij}$ Æ 0

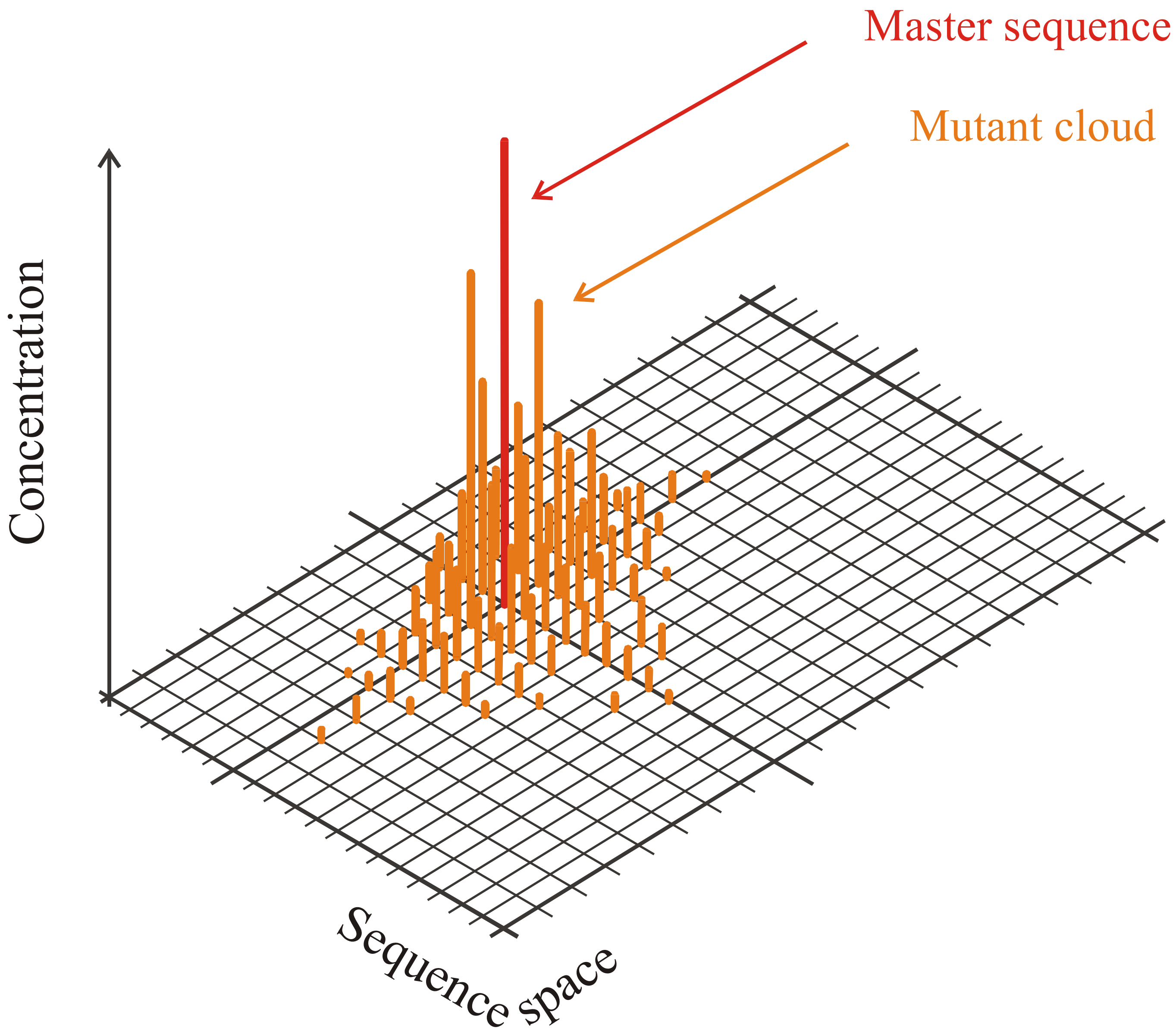
$$\frac{dx_i}{dt} = \sum\nolimits_{j=1}^{n} f_j Q_{ji}\, x_j - x_i\, \phi, \quad i = 1,2,\cdots,n; \quad \sum\nolimits_{i=1}^{n} x_i = 1; \quad \phi = \sum\nolimits_{j=1}^{n} f_j x_j = \overline{f}$$

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

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

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

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



The molecular quasispecies in sequence space

Information on the environment is created in the **population** during the **selection process** through autocatalytic self-enhancement of advantageous variants.

The **population** is visualized as a distribution of RNA molecules. In evolution the population carries a **temporary memory** on its recent history in terms of **previously selected variants** that are still present.

1. RNA structure, replication kinetics, and origin of information

2. **Evolution *in silico* and optimization of RNA structures**

3. Random walks and ‚ensemble learning‘

4. Sequence-structure maps, neutral networks, and intersections

In evolution **variation** occurs on **genotypes** but **selection** operates on the **phenotype**.

Mappings from genotypes into phenotypes are highly complex objects. The only computationally accessible case is in the evolution of RNA molecules.

The mapping from RNA sequences into secondary structures and function,

sequence ñ **structure** ñ **function**,

is used as a model for the complex relations between genotypes and phenotypes. Fertile progeny measured in terms of **fitness** in population biology is determined quantitatively by **replication rate constants** of RNA molecules.

Population biology	Molecular genetics	Evolution of RNA molecules
Genotype	**Genome**	**RNA sequence**
Phenotype	**Organism**	**RNA structure and function**
Fitness	**Reproductive success**	**Replication rate constant**

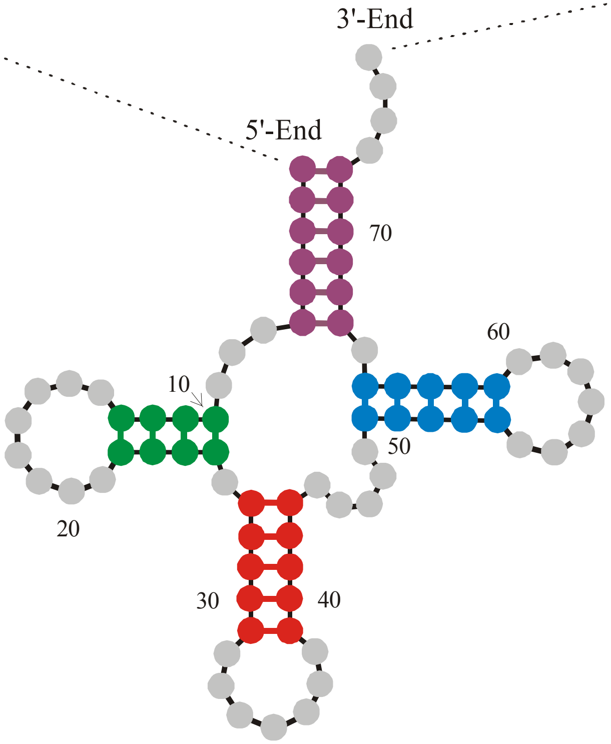
The RNA model

Sequence

5'-End 3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure



Symbolic notation

5'-End ((((((···((((········))))·(((((·······)))))·····(((((·······)))))·))))))···· 3'-End

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

How to compute RNA secondary structures

Efficient algorithms based on **dynamic programming** are available for computation of minimum free energy and **many** suboptimal secondary structures for given sequences.

M.Zuker and P.Stiegler. ***Nucleic Acids Res***. **9**:133-148 (1981)

M.Zuker, ***Science*** **244**: 48-52 (1989)

Equilibrium partition function and base pairing probabilities in Boltzmann ensembles of suboptimal structures.

J.S.McCaskill. ***Biopolymers*** **29**:1105-1190 (1990)

The **Vienna RNA Package** provides in addition: **inverse folding** (computing sequences for given secondary structures), computation of melting profiles from partition functions, **all** suboptimal structures within a given energy interval, barrier tress of suboptimal structures, **kinetic folding** of RNA sequences, RNA-hybridization and RNA/DNA-hybridization through **cofolding** of sequences, alignment, etc..

I.L.Hofacker, W. Fontana, P.F.Stadler, L.S.Bonhoeffer, M.Tacker, and P. Schuster. ***Mh.Chem.*** **125**:167-188 (1994)

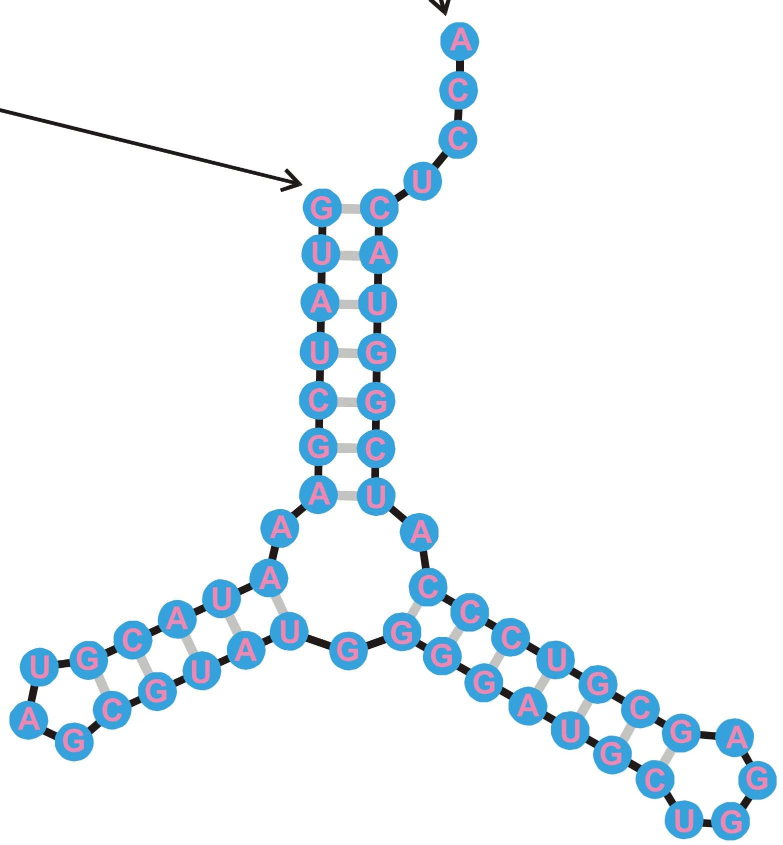
S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. ***Biopolymers*** **49**:145-165 (1999)

C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. ***RNA*** **6**:325-338 (1999)

Vienna RNA Package: http://www.tbi.univie.ac.at

5’-end 3’-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



RNAStudio.lnk

GGCGCGCCCGGCGCC

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

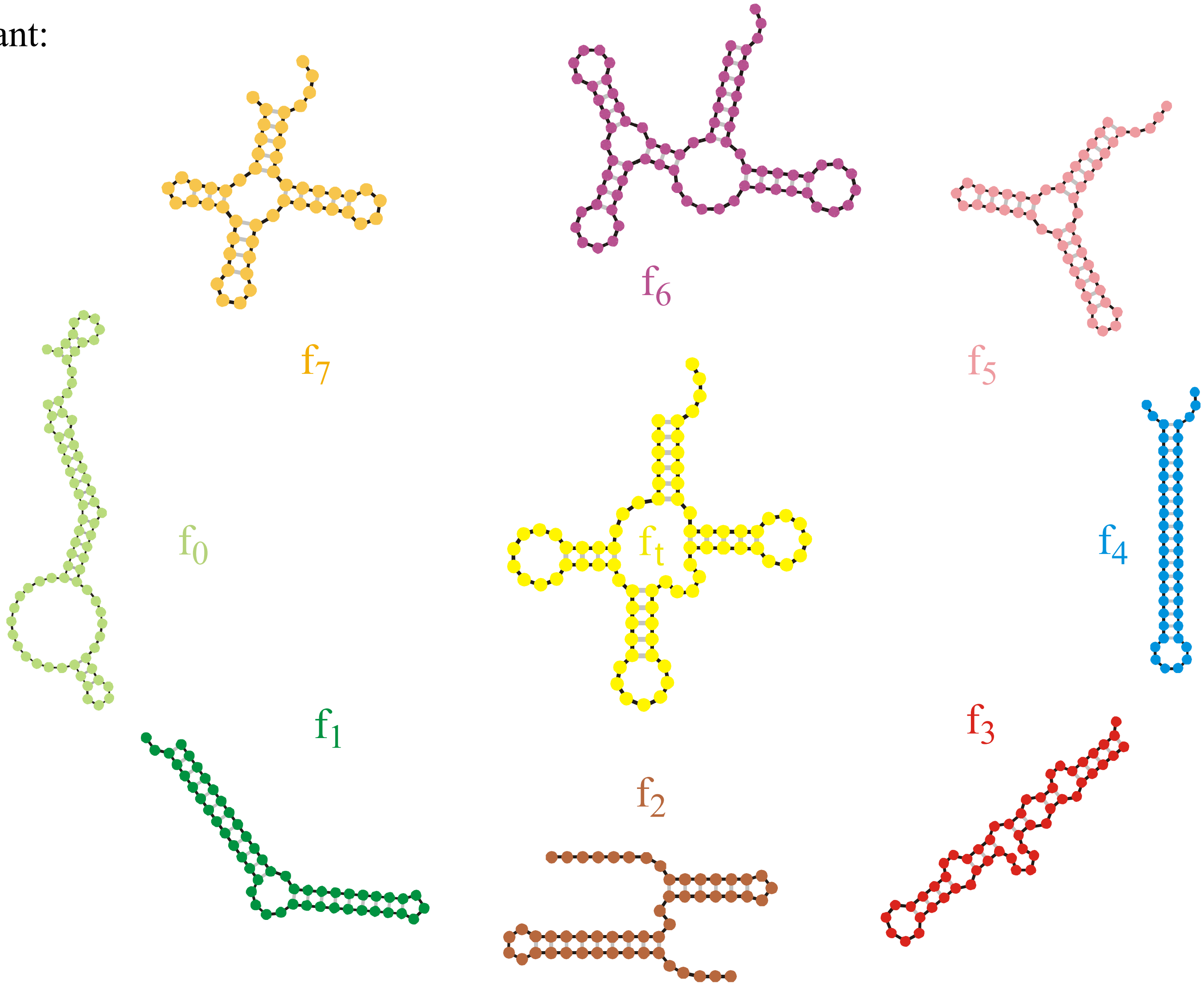
UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

Folding of RNA sequences into secondary structures of minimal free energy, DG_0^{300}

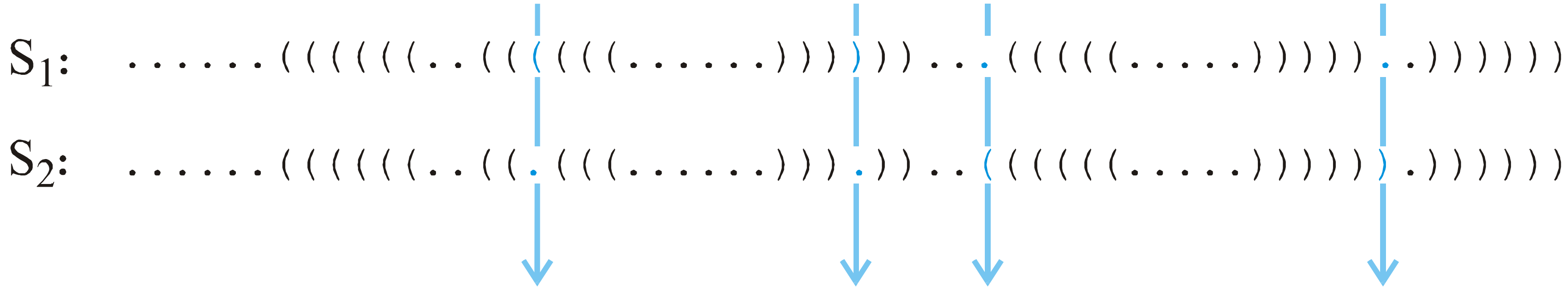
Replication rate constant:

$$f_k = g \,/\, [a + Dd_S^{\;(k)}]$$

$$Dd_S^{\;(k)} = d_H(S_k,S_t)$$



Evaluation of RNA secondary structures yields replication rate constants



S_1: ......((((((..((((((......)))))))...(((((.....)))))..))))))

S_2: ......((((((..((.(((......))).))..((((((.....))))))).))))))

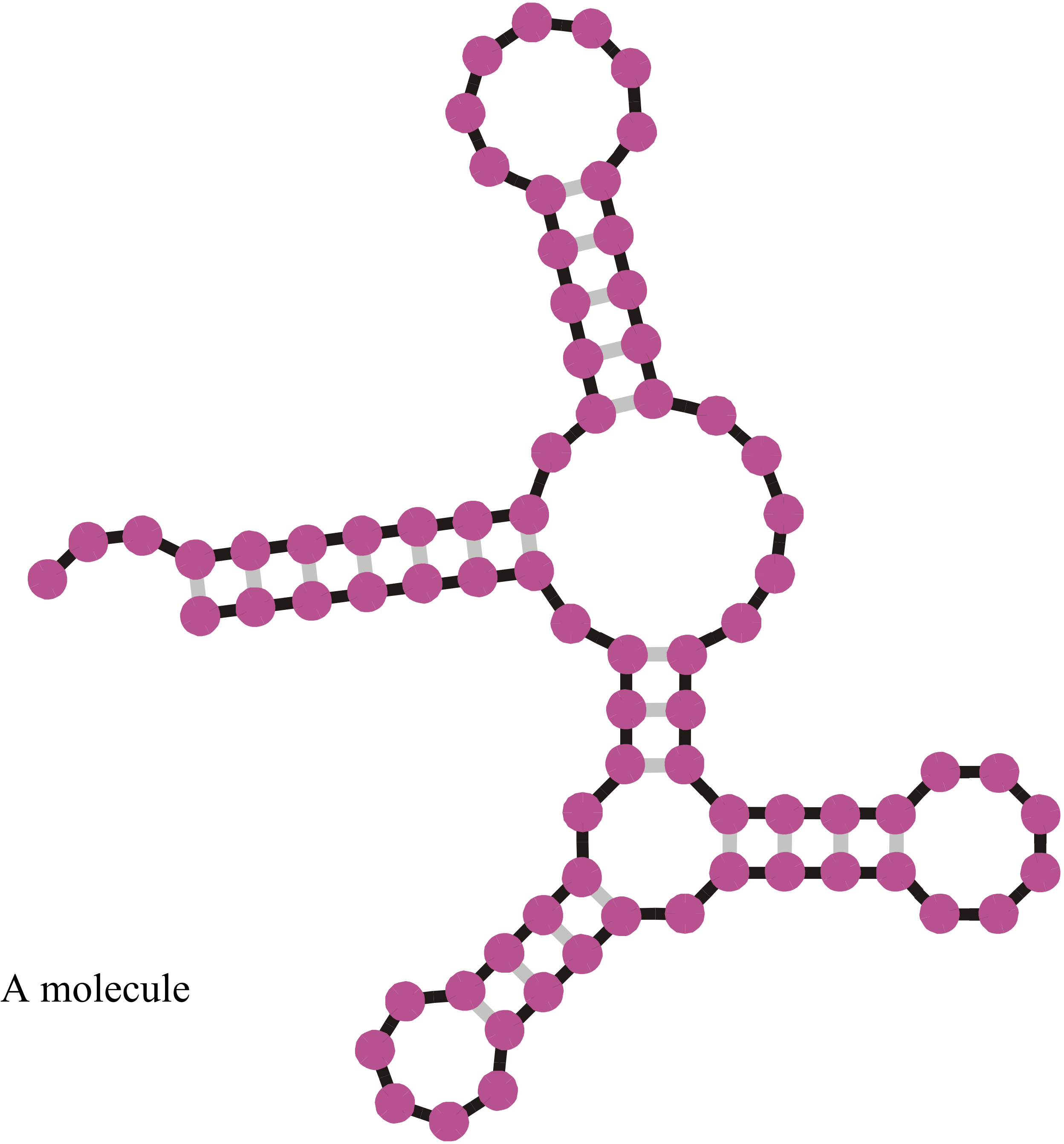
Hamming distance $d_H(S_1,S_2) = 4$

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

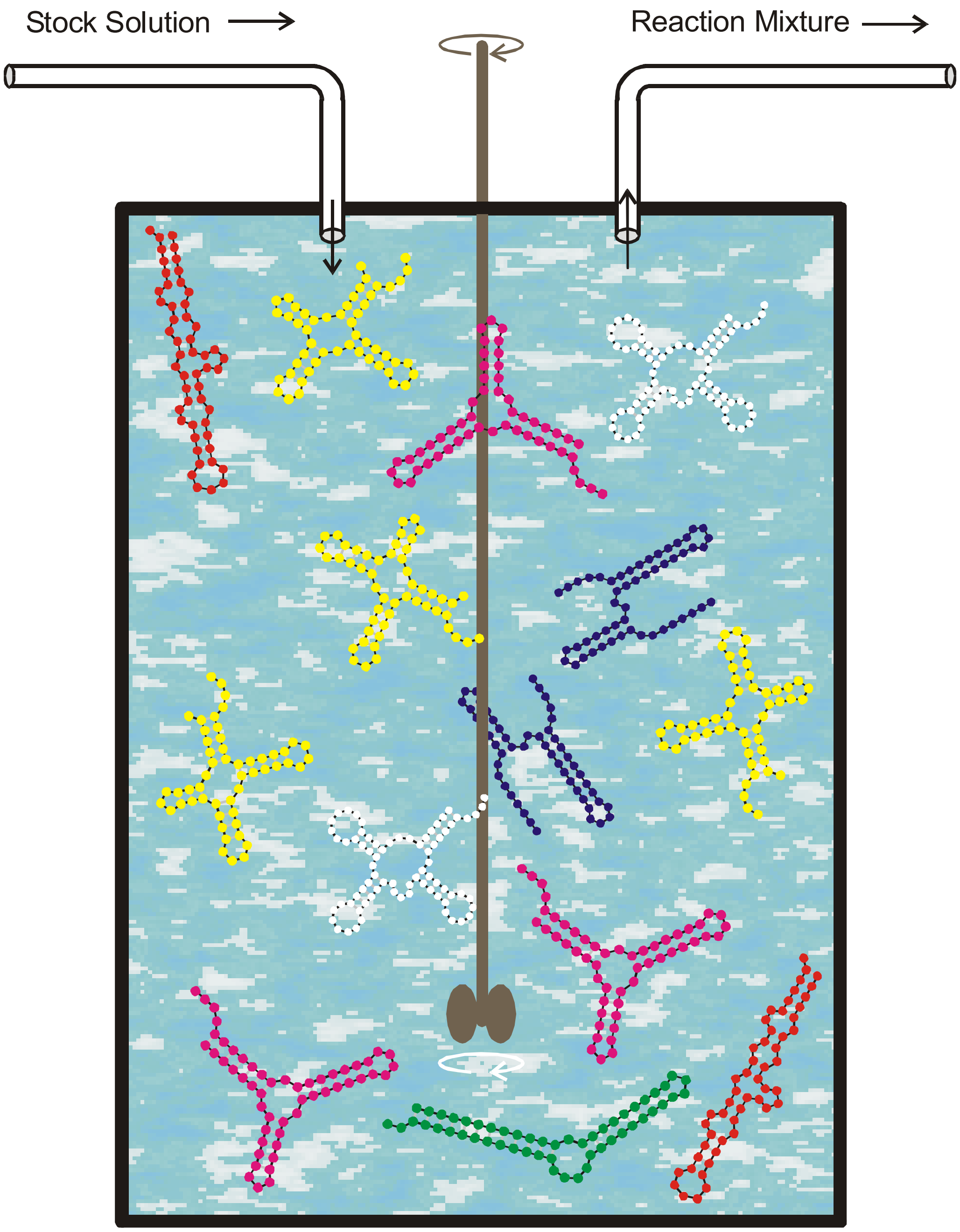
(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

(iii) $d_H(S_1,S_3)$ ¶ $d_H(S_1,S_2) + d_H(S_2,S_3)$

The Hamming distance between structures in parentheses notation forms a metric in structure space



Element class 1: The RNA molecule



Replication rate constant:

$$f_k = g / [a + Dd_S^{(k)}]$$

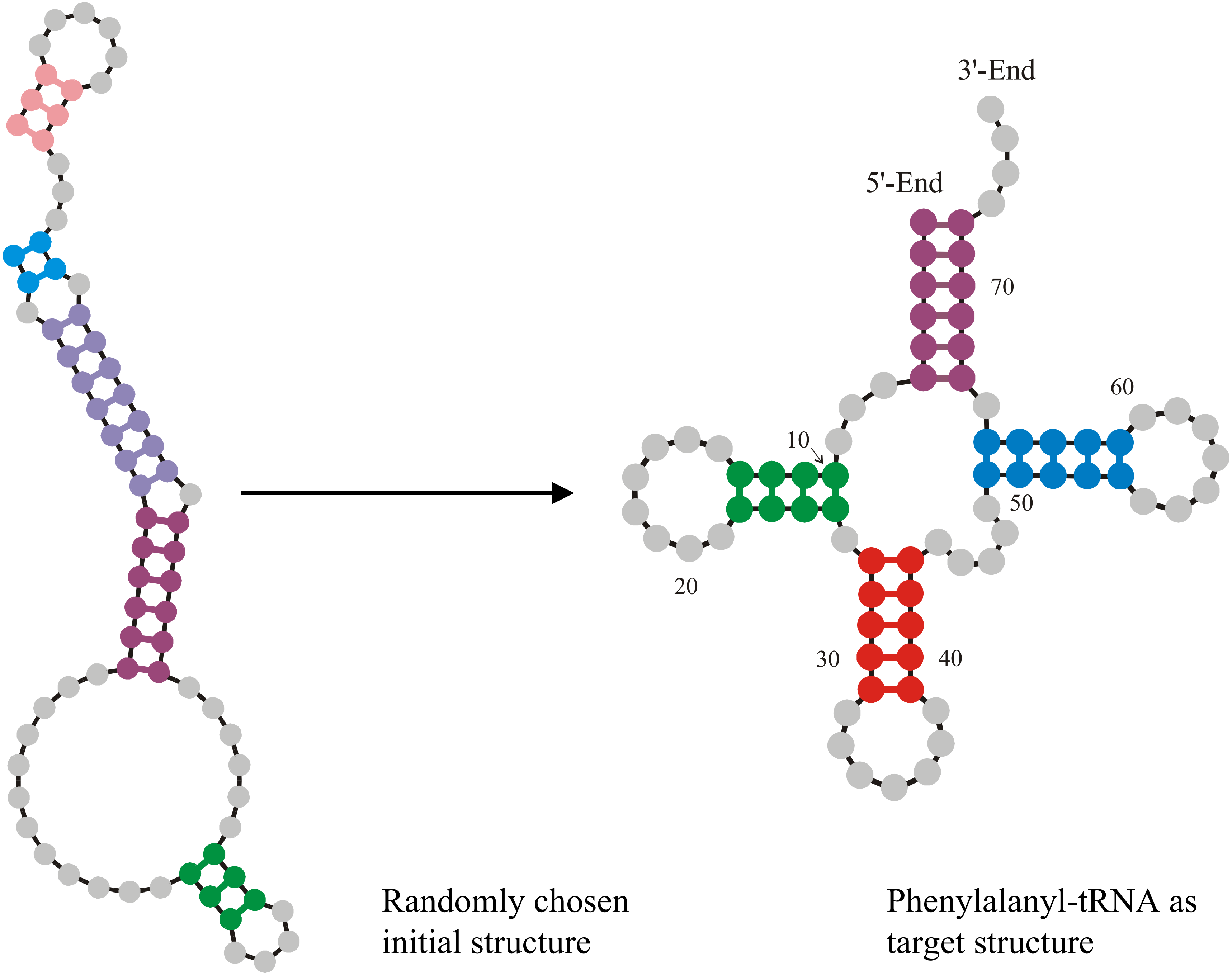
$$Dd_S^{(k)} = d_H(S_k, S_t)$$

Selection constraint:

# RNA molecules is controlled by the flow

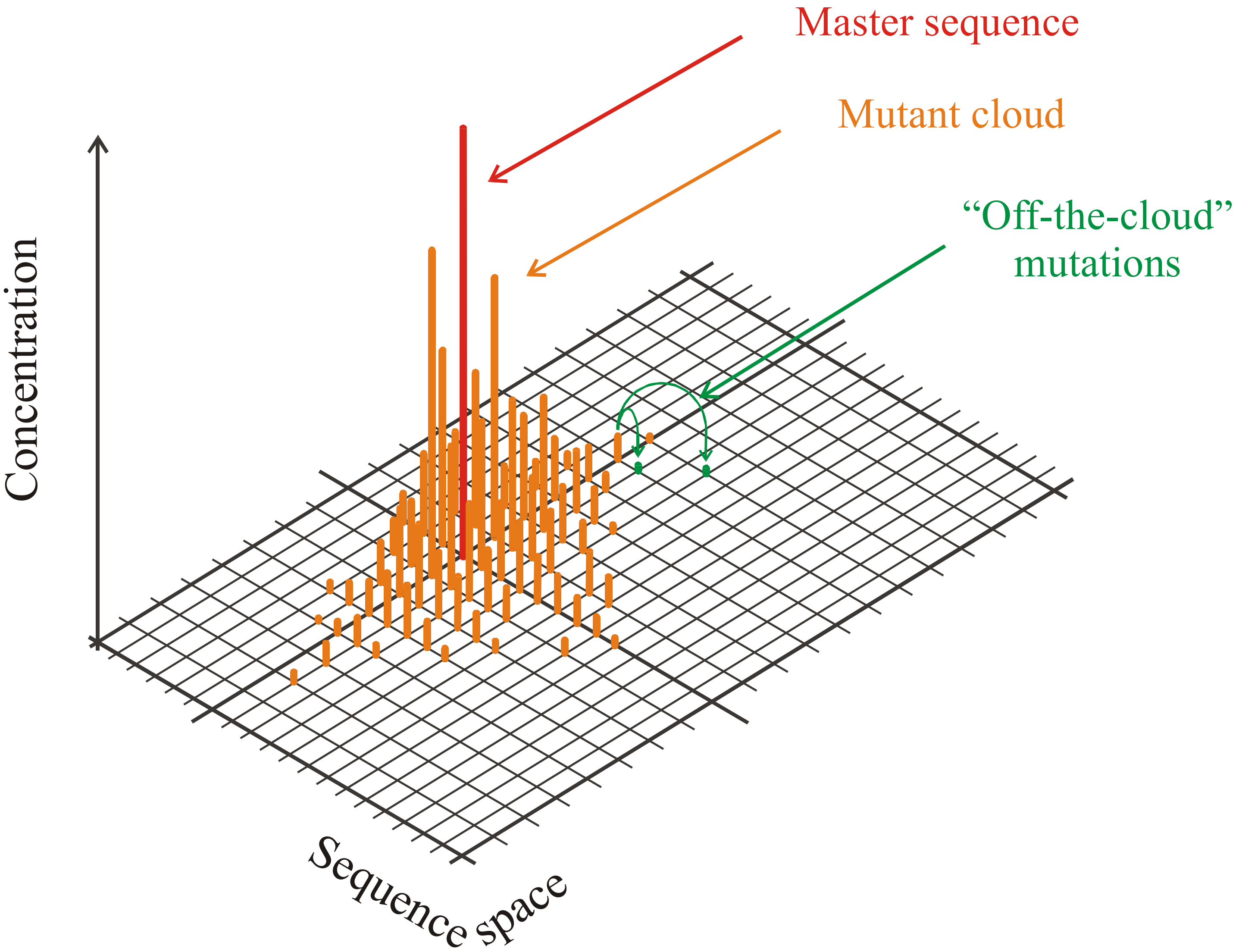
$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

The flowreactor as a device for studies of evolution ***in vitro*** and ***in silico***

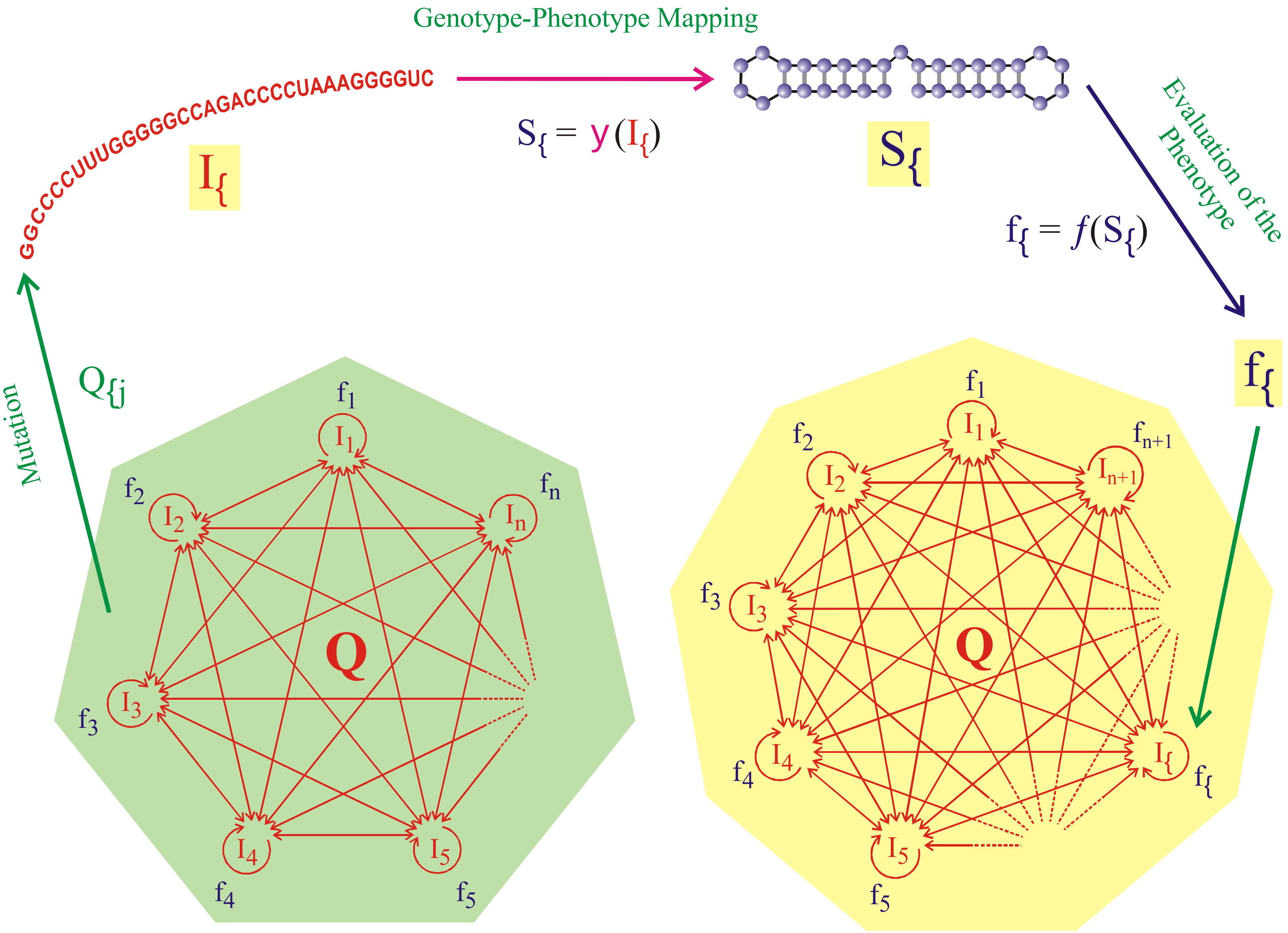


Randomly chosen
initial structure

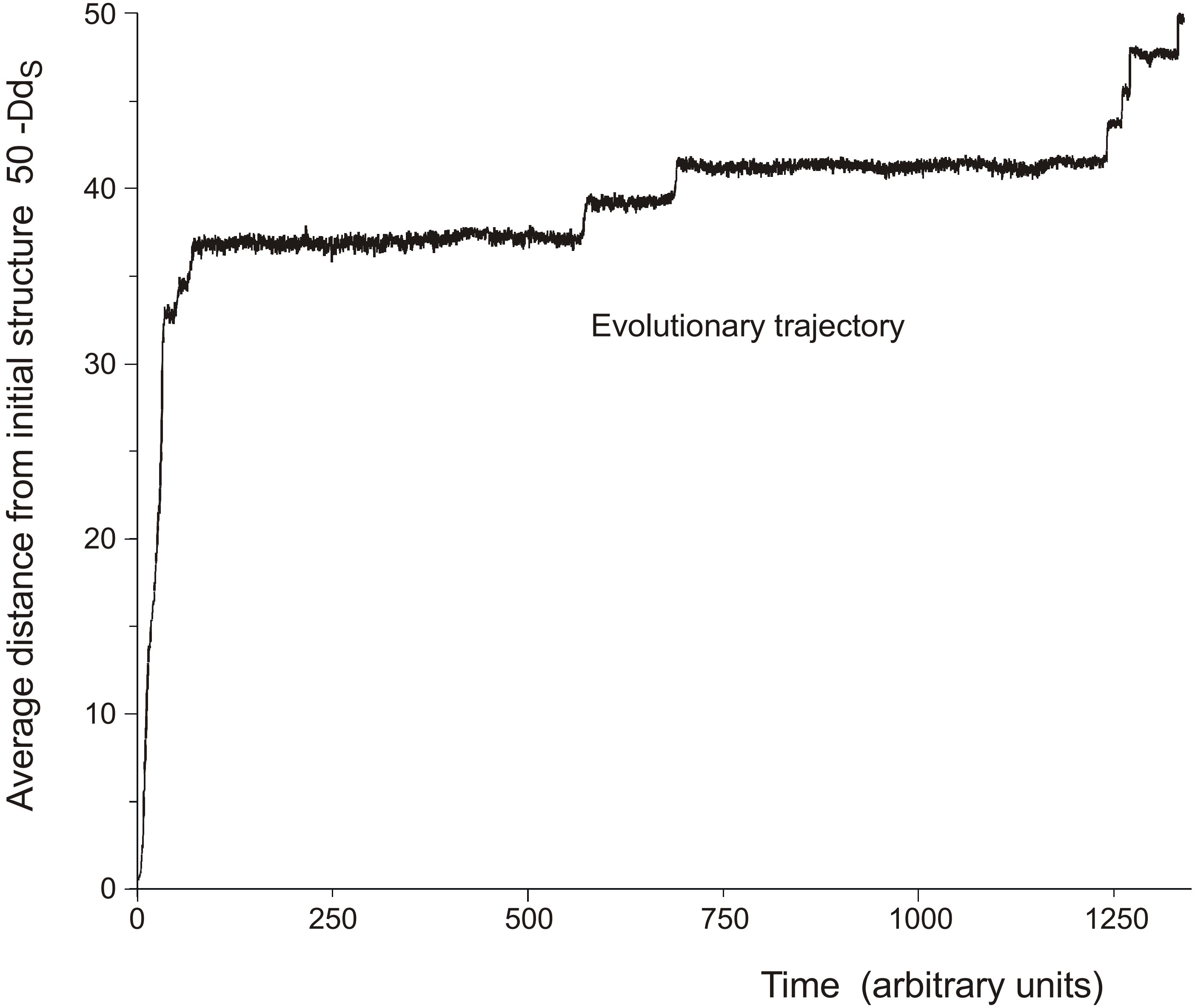
Phenylalanyl-tRNA as
target structure



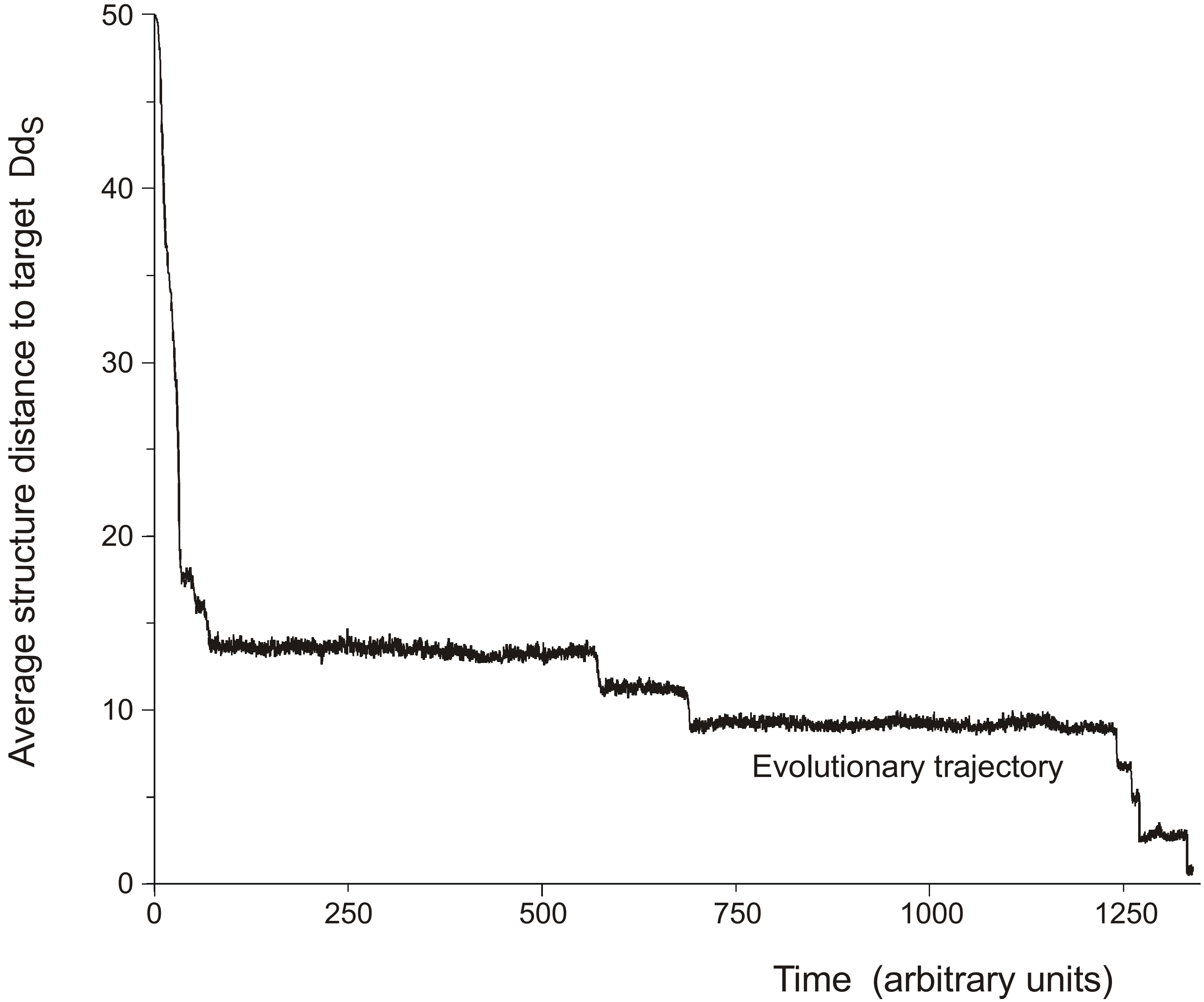
The molecular quasispecies
in sequence space



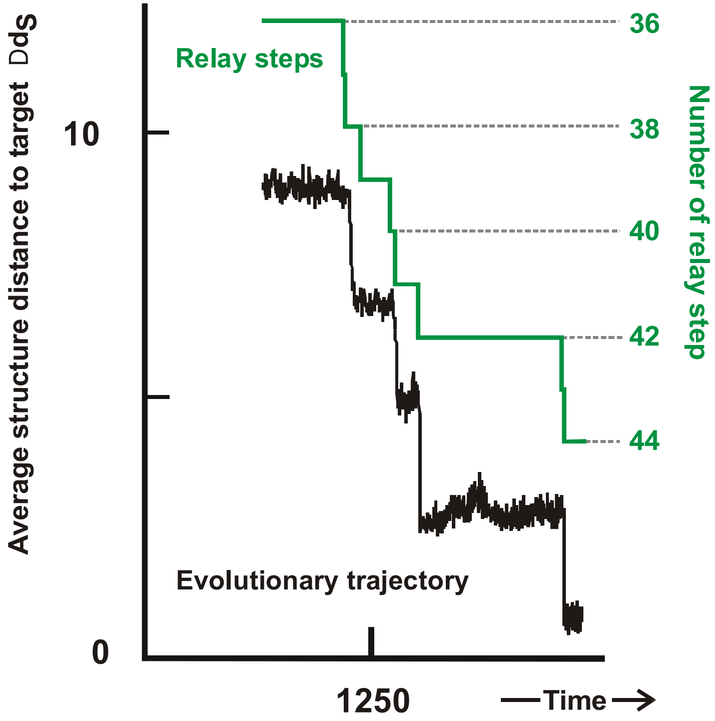
Evolutionary dynamics
including molecular phenotypes

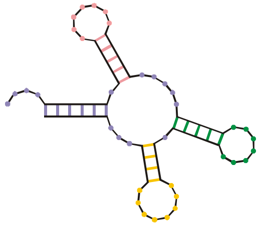


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)

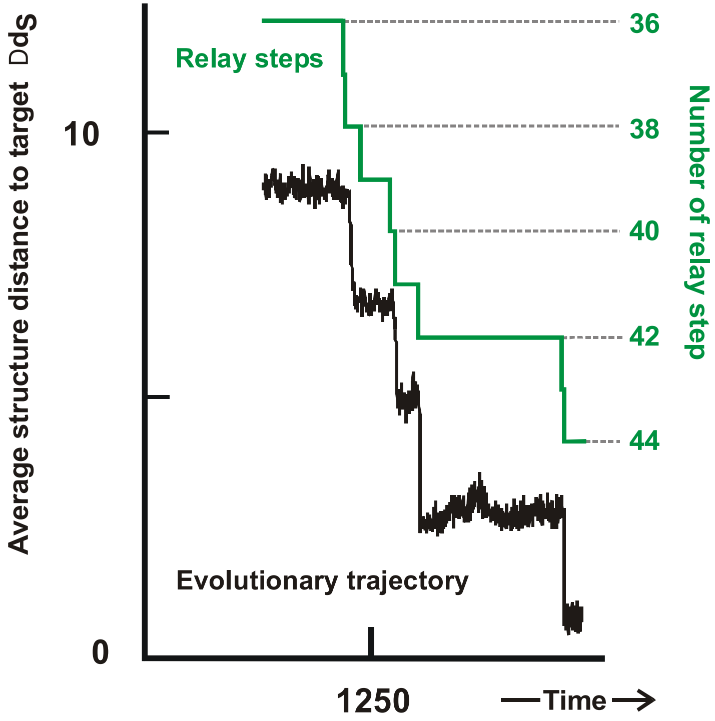


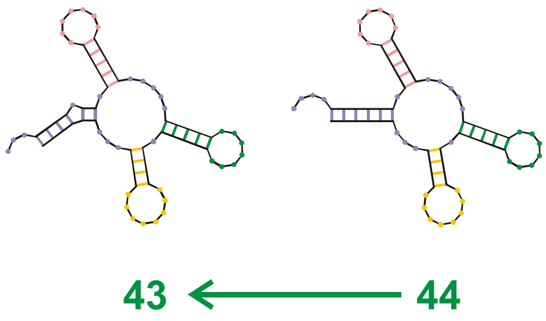
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)



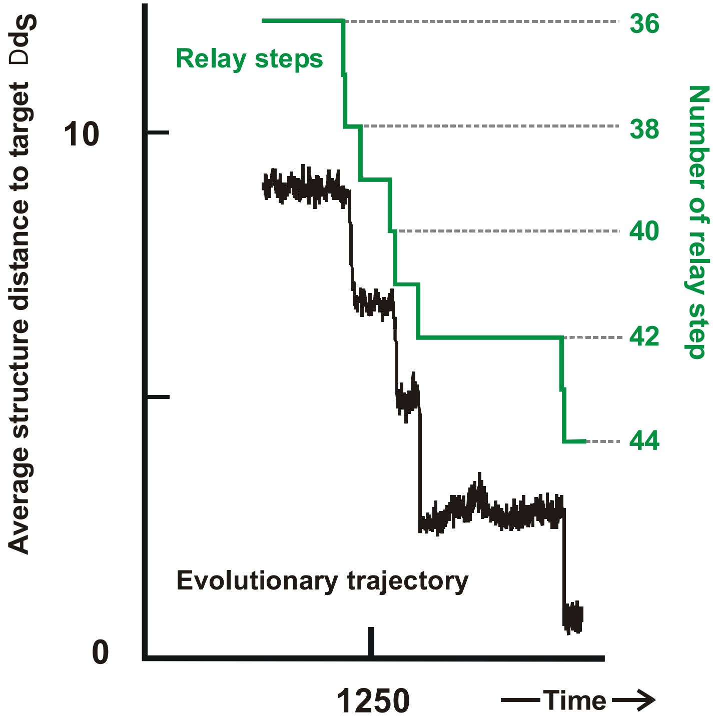


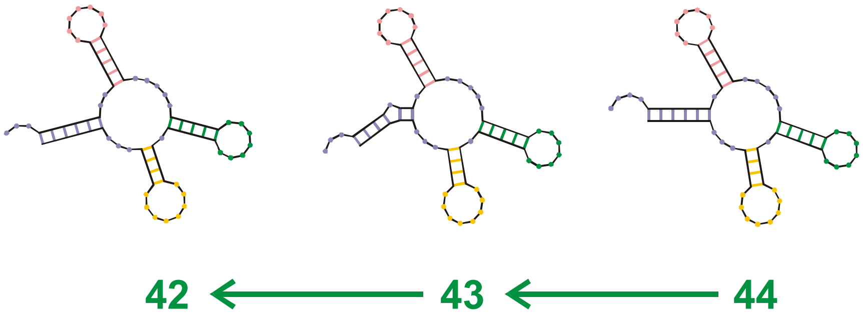
Endconformation of optimization



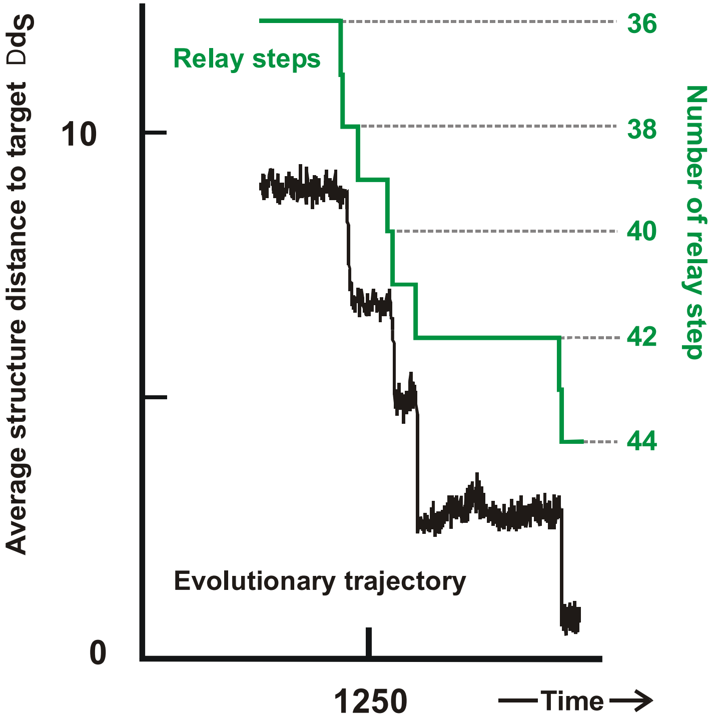


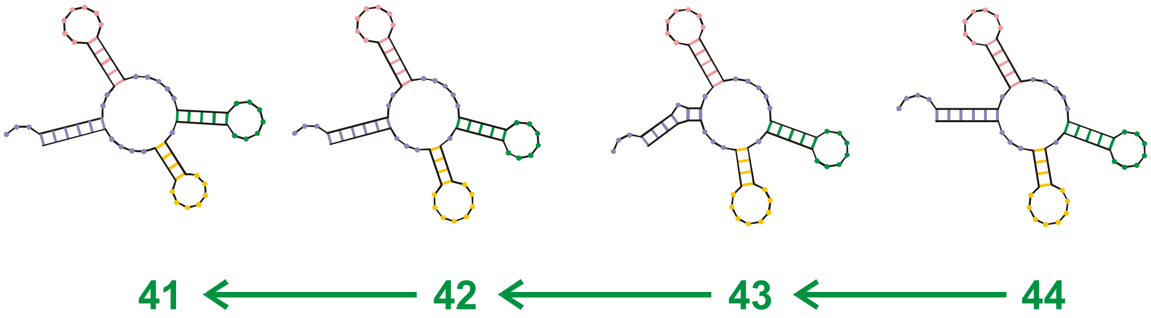
Reconstruction of the last step 43 Á 44



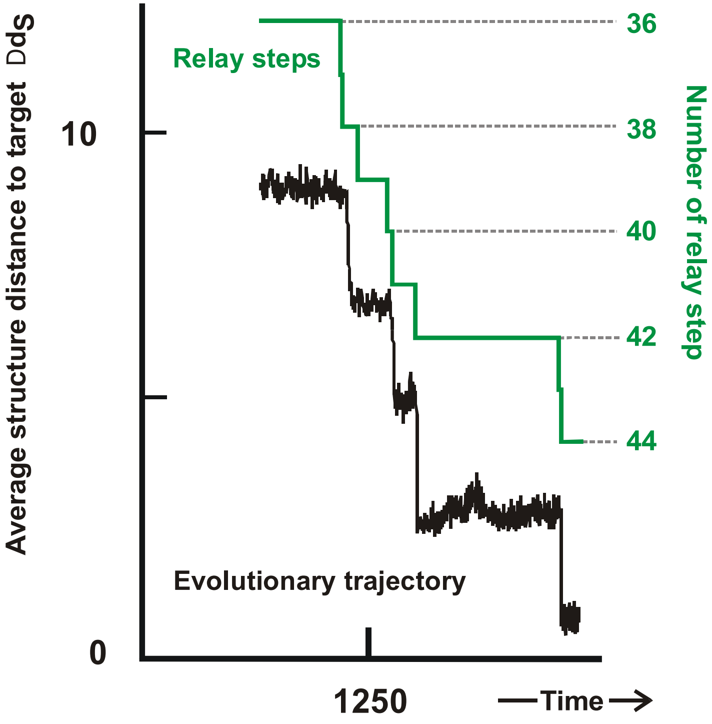


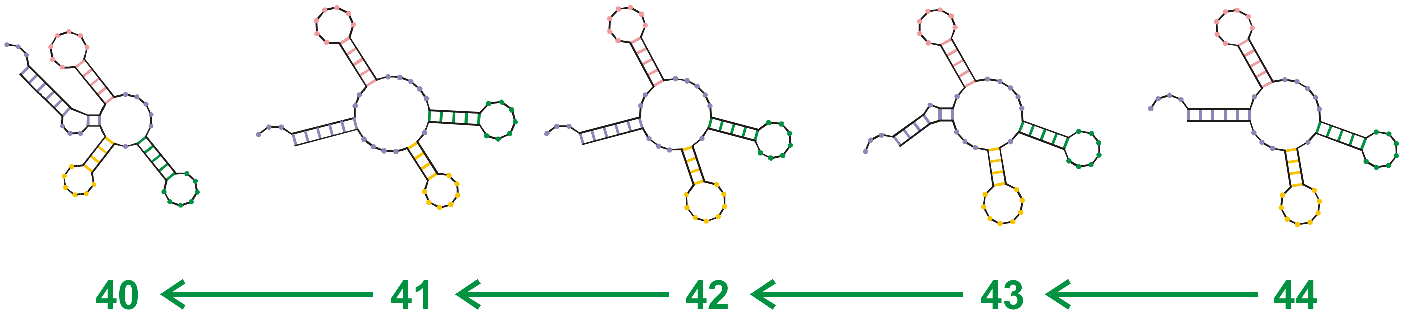
Reconstruction of last-but-one step 42 Á 43 (Á 44)



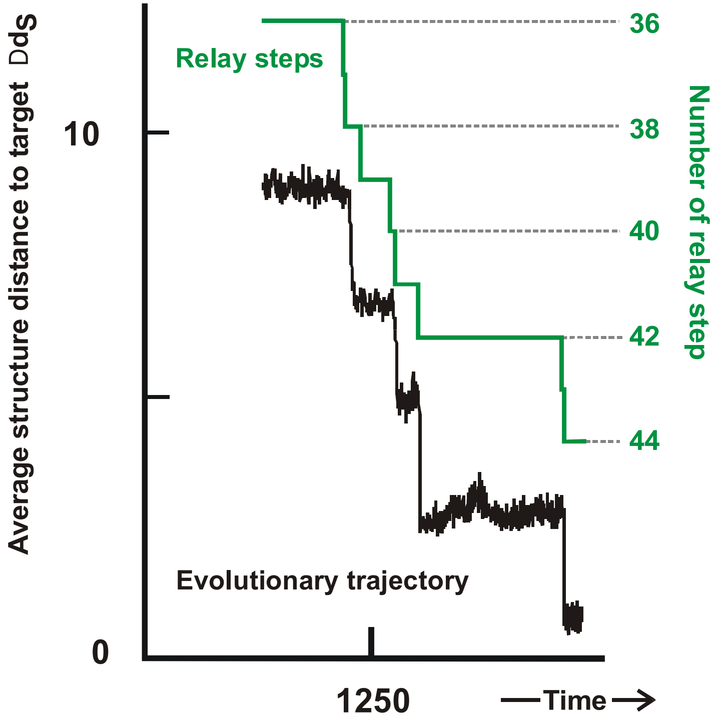


Reconstruction of step 41 Á 42 (Á 43 Á 44)

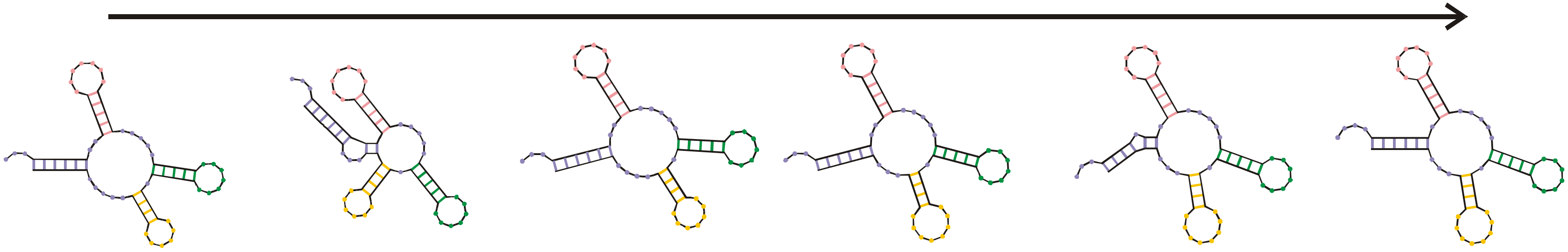




Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)



Evolutionary process



39 ← 40 ← 41 ← 42 ← 43 ← 44

Reconstruction

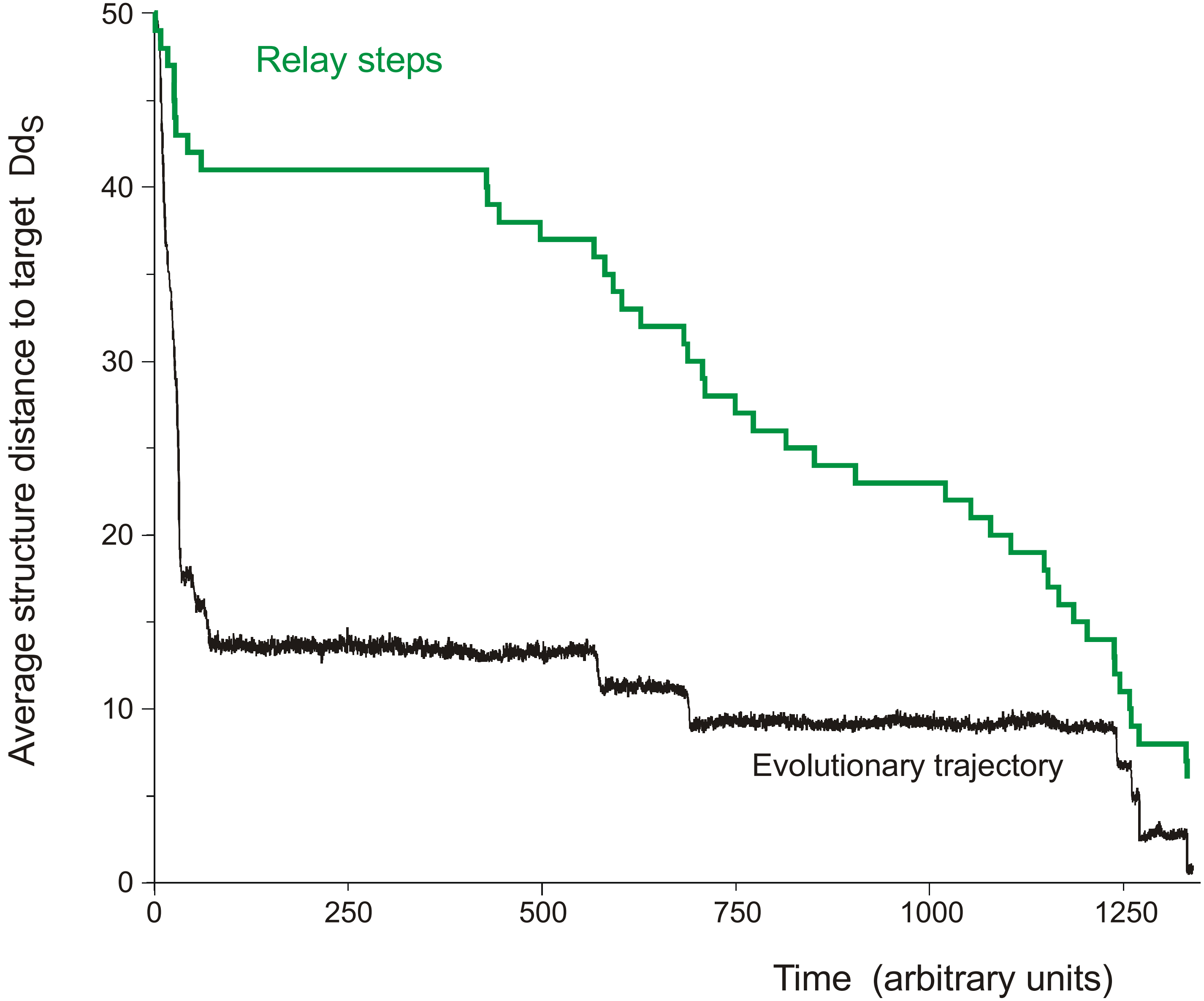
Reconstruction of the relay series

```
entry  GGGAUACAUGUGGCCCCUCAAGGCCCUAGCGAAACUGCUGCUGAAACCGUGUGAAUAAUCCGCACCCUGUCCCCGA
39     ((((((.....((((......)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUAUACGAGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
entry  GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
40     ((((((...((((((......)))).(((((.......))))).....(((((.......)))))))))))))...
exit   GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
entry  GGGAUAUACGGGCCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
41     ((((((....((((.......)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUAUACGGGCCCCUUCAAGGCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA
entry  GGGAUAUACGGGCCCCUUCAAGCCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA
42     ((((((...((((........)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
entry  GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
43     ((((((...((((........)))).(((((.......))))).....(((((.......))))).)).))))...
exit   GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
entry  GGGCAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
44     ((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....
```

Transition inducing point mutations

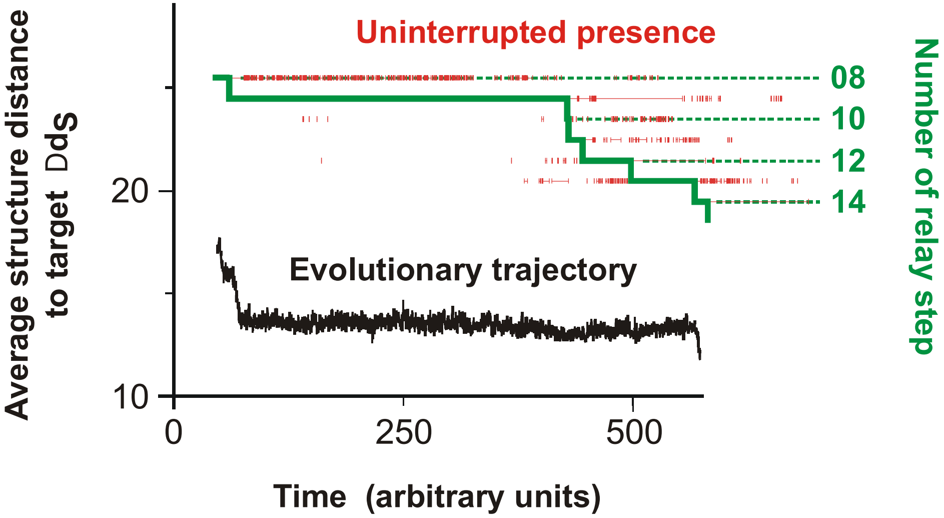
Neutral point mutations

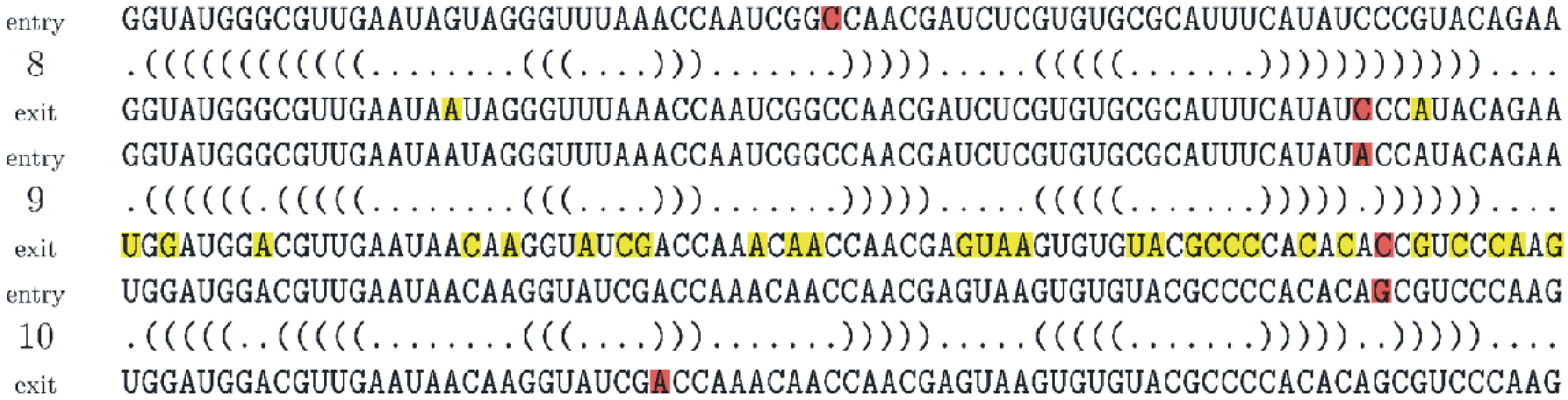
Change in RNA sequences during the final five relay steps 39 Á 44



In silico optimization in the flow reactor: Trajectory and relay steps

28 neutral point mutations during a long quasi-stationary epoch



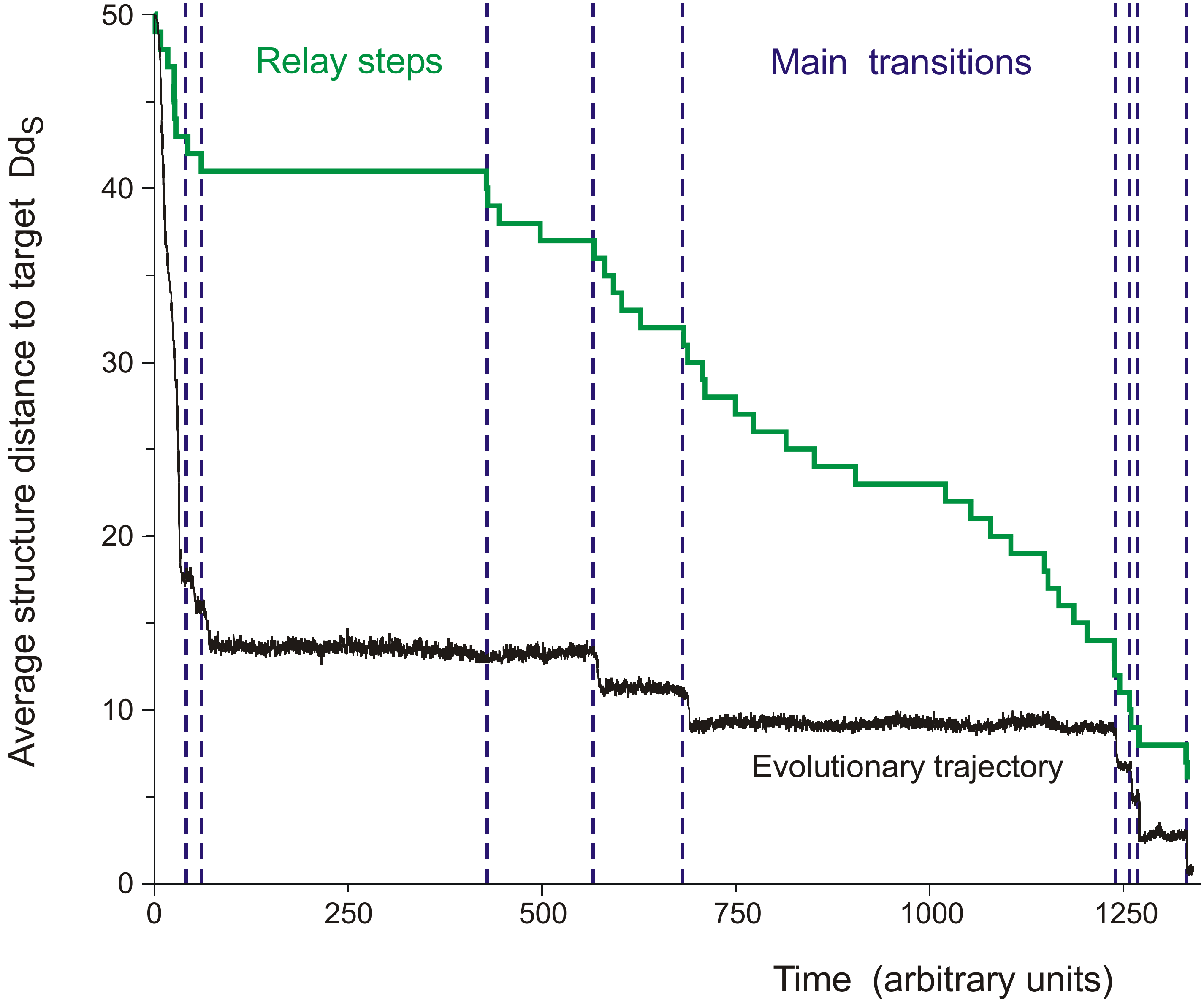


```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
  8    .((((((((((((........(((....))).......))))).....(((((.......))))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
  9    .((((((.(((((........(((....))).......))))).....(((((.......))))).))))))....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry  UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
 10    .(((((..(((((........(((....))).......))))).....(((((.......)))))..)))))....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

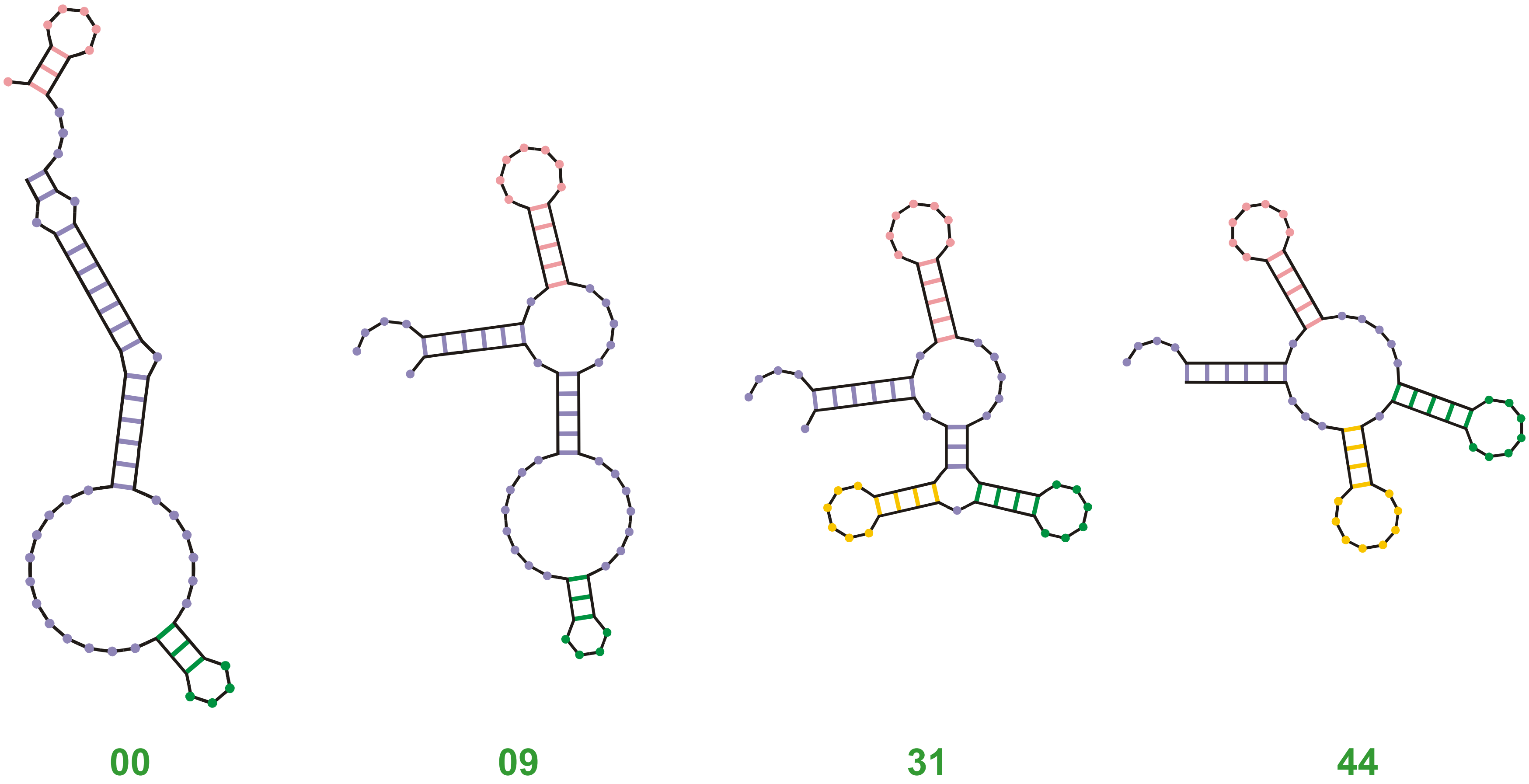
Transition inducing point mutations

Neutral point mutations

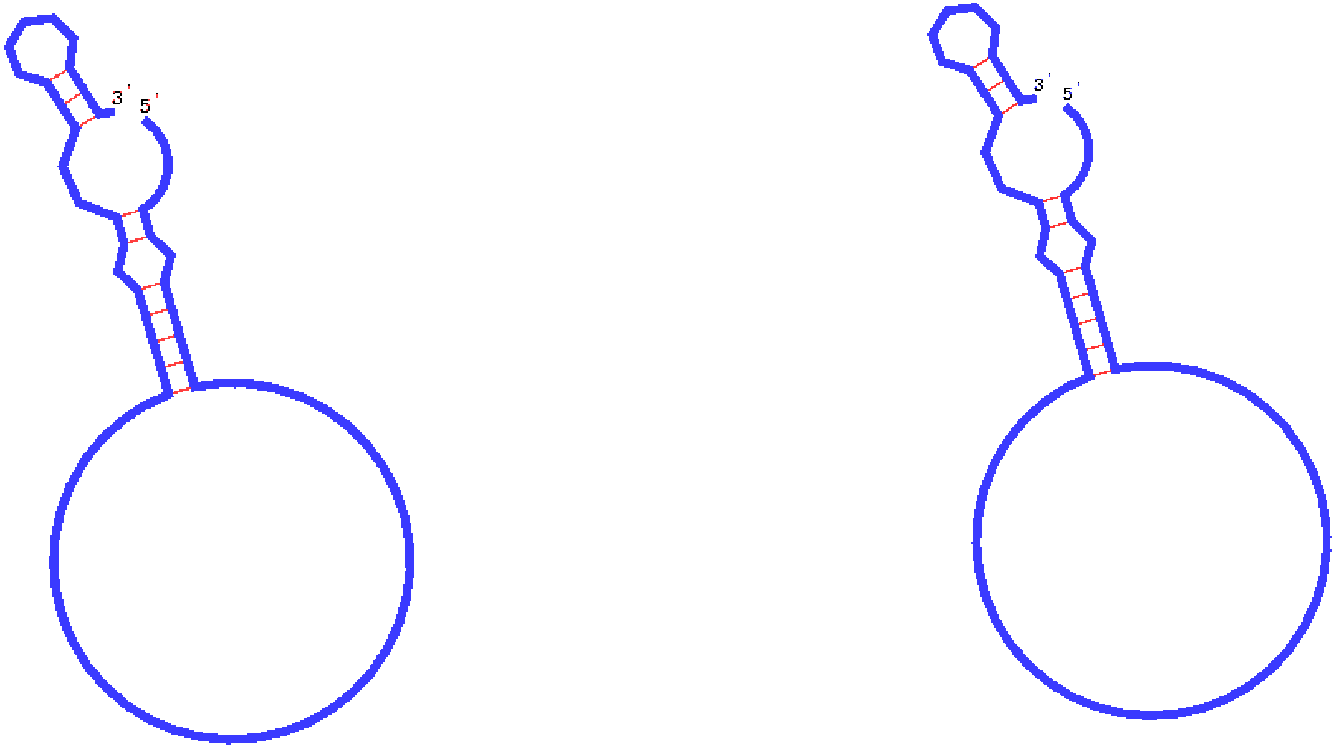
Neutral genotype evolution during phenotypic stasis



In silico optimization in the flow reactor: Main transitions



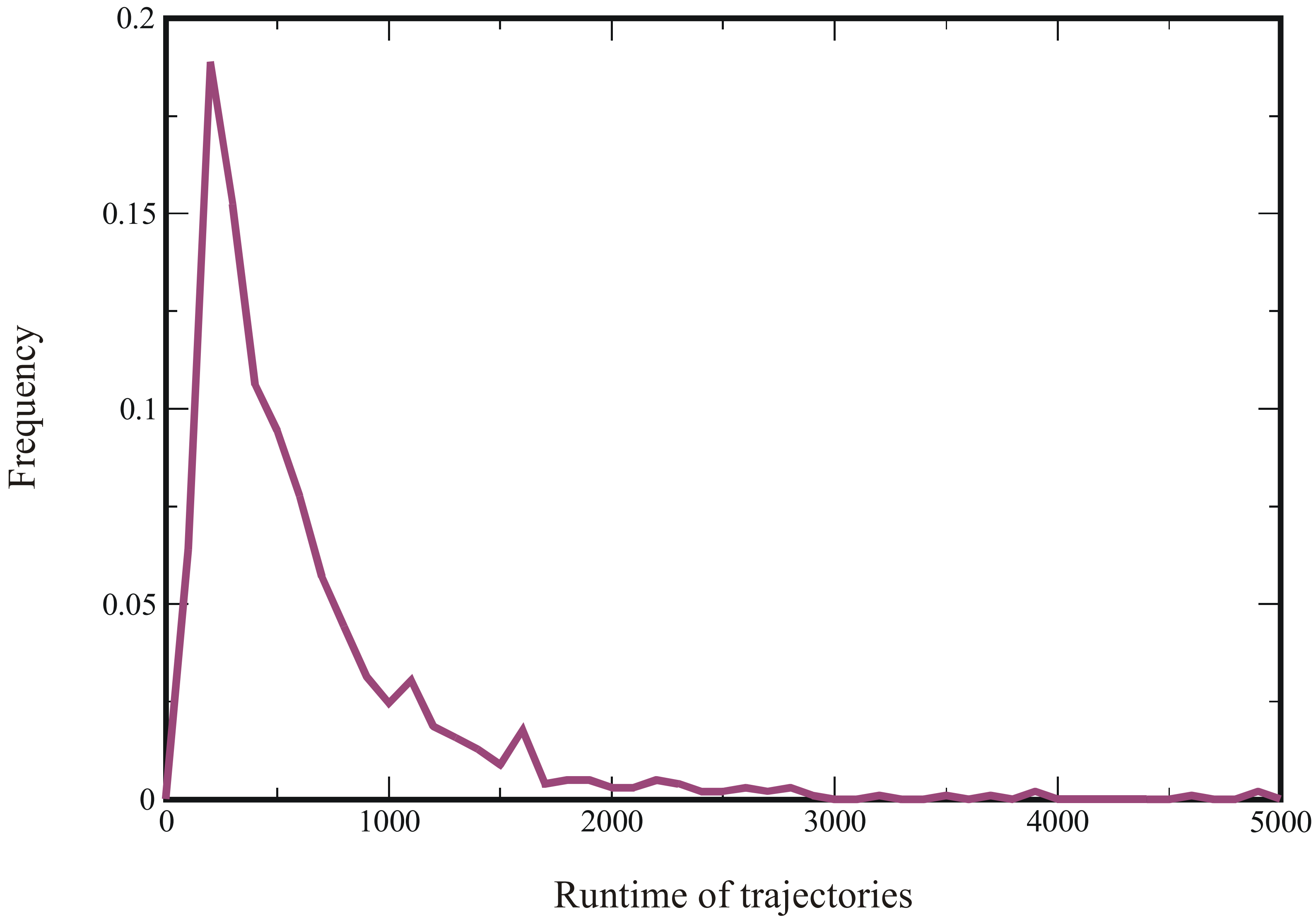
Three important steps in the formation of the tRNA clover leaf from a randomly chosen initial structure corresponding to three **main transitions**.



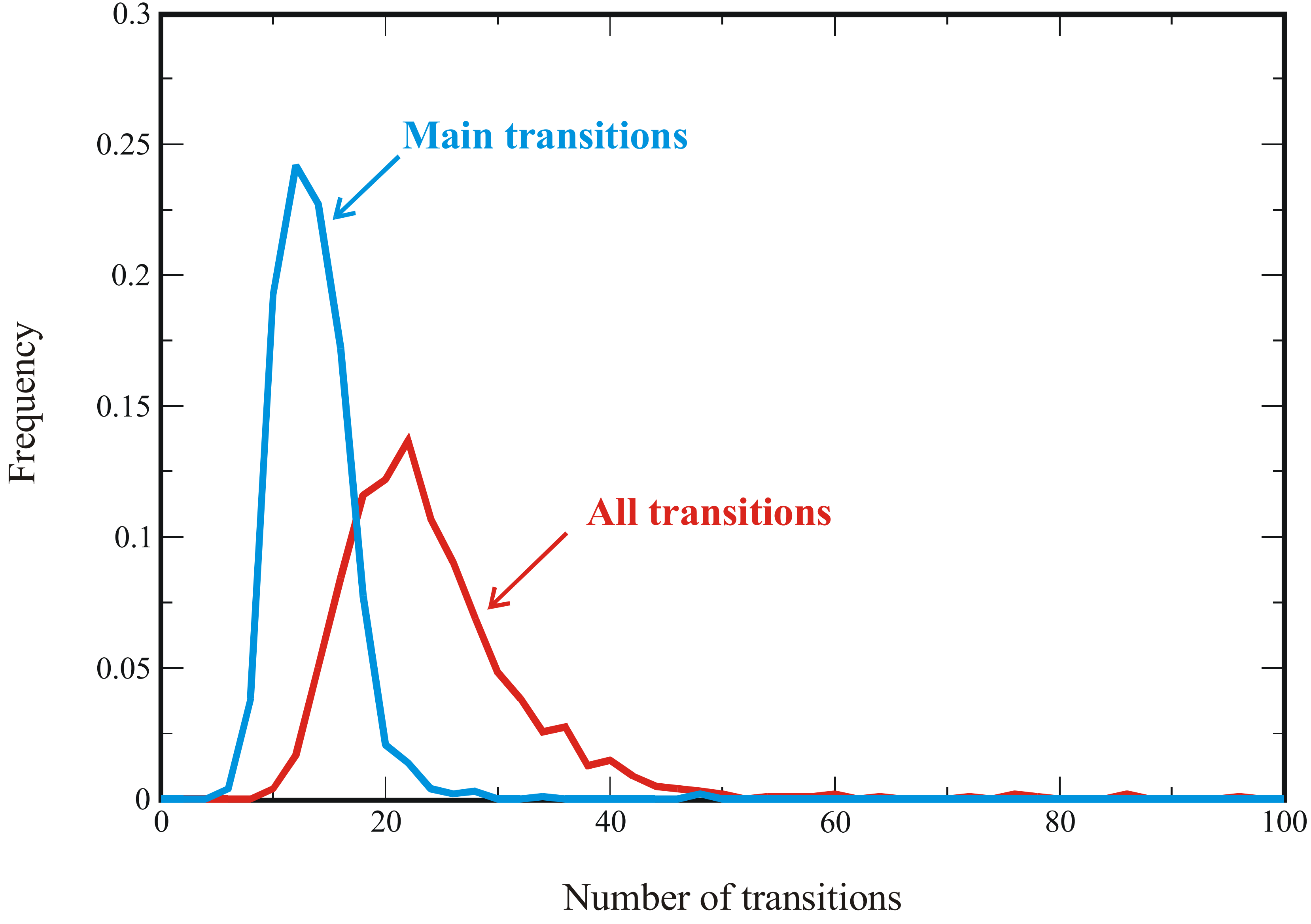
AUGC

GC

Movies of optimization trajectories over the **AUGC** and the **GC** alphabet



Statistics of the lengths of trajectories from initial structure to target (**AUGC**-sequences)



Statistics of the numbers of transitions from initial structure to target (**AUGC**-sequences)

Alphabet	Runtime	Transitions	Main transitions	No. of runs
AUGC	385.6	22.5	12.6	1017
GUC	448.9	30.5	16.5	611
GC	2188.3	40.0	20.6	107

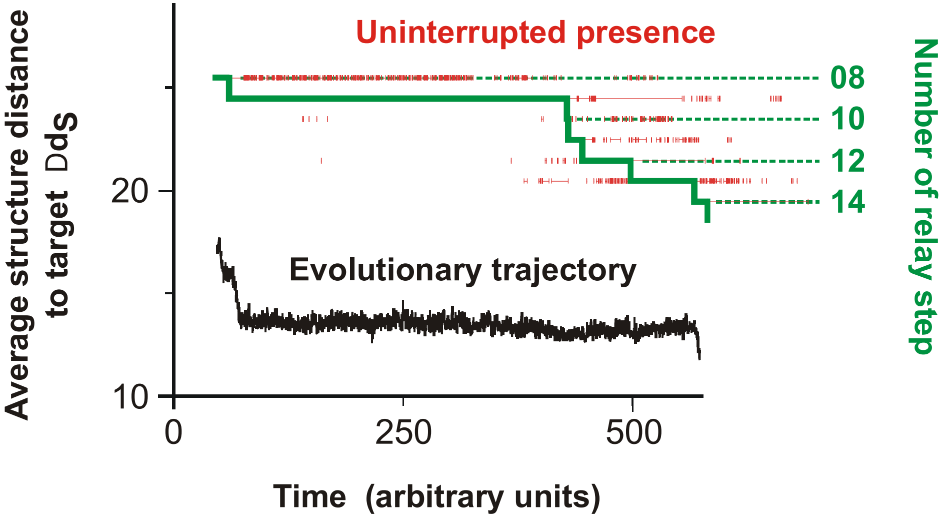
Statistics of trajectories and relay series (mean values of log-normal distributions)

1. RNA structure, replication kinetics, and origin of information

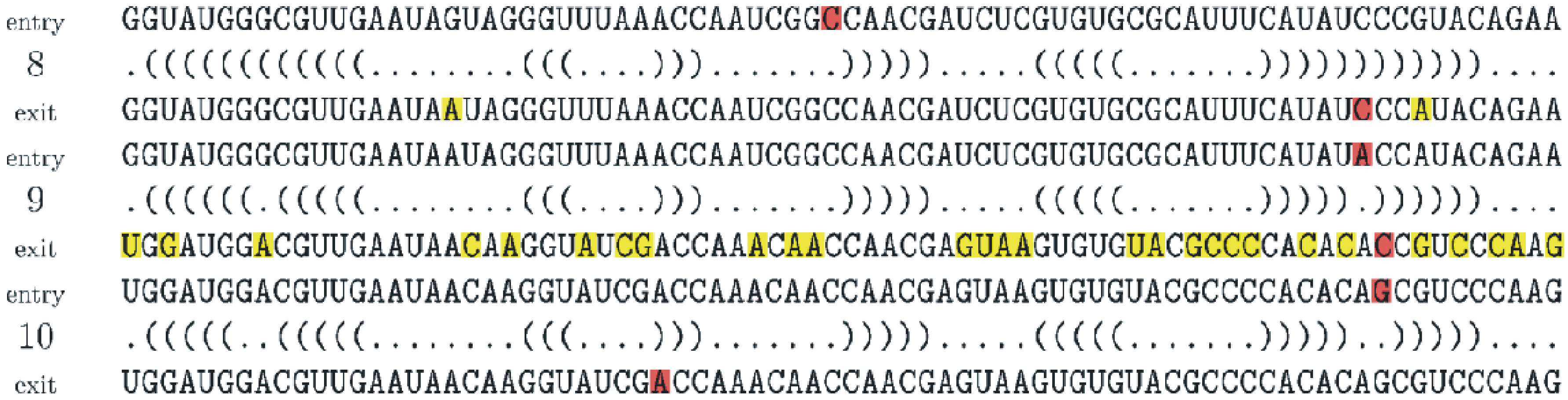
2. Evolution *in silico* and optimization of RNA structures

3. **Random walks and ‚ensemble learning‘**

4. Sequence-structure maps, neutral networks, and intersections



28 neutral point mutations during a long quasi-stationary epoch

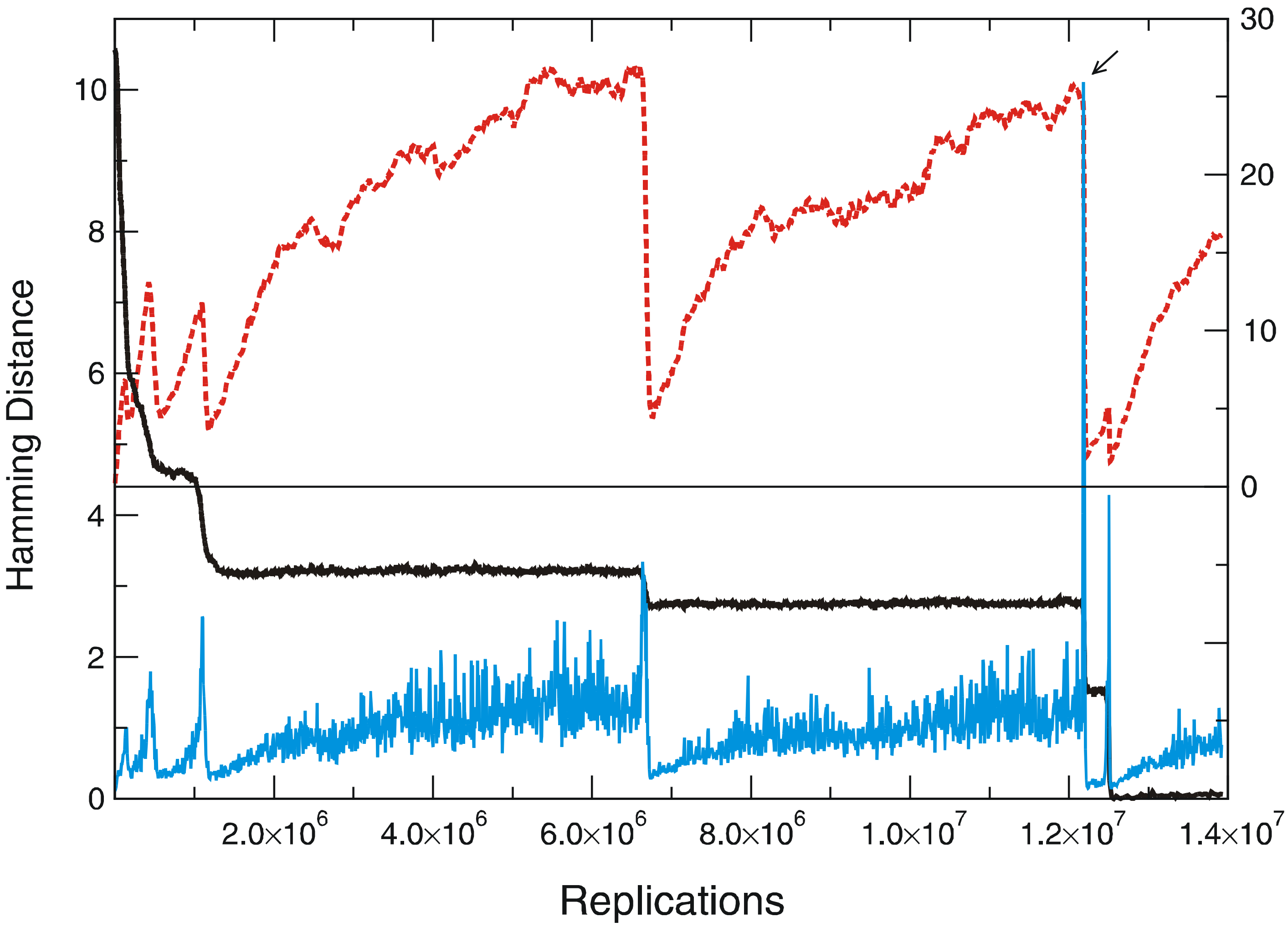


```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8      .((((((((((((........(((....))).......))))).....(((((.......))))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9      .((((((.(((((........(((....))).......))))).....(((((.......))))).))))))....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry  UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10     .(((((..(((((........(((....))).......))))).....(((((.......)))))..))))).....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

Transition inducing point mutations

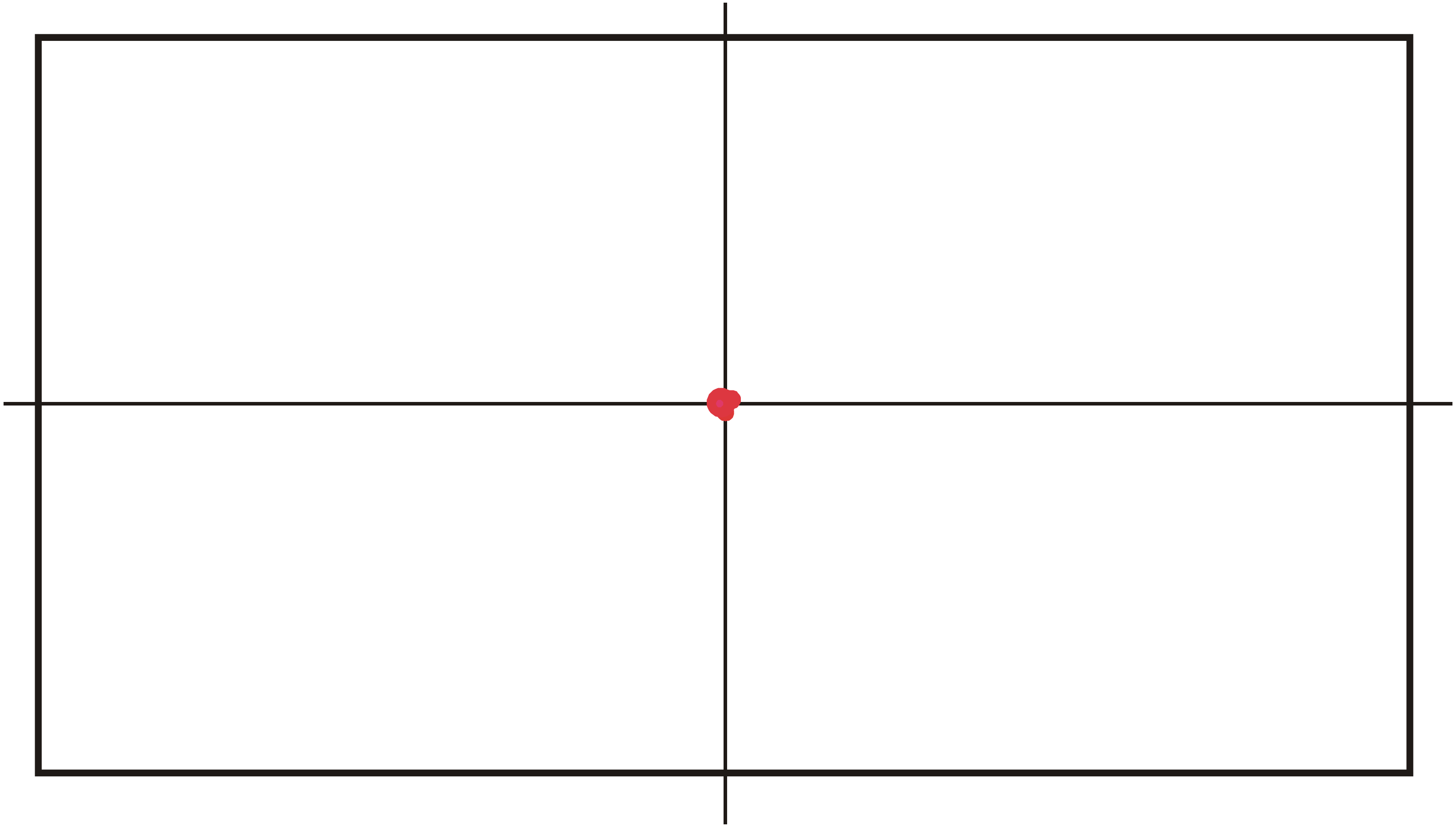
Neutral point mutations

Neutral genotype evolution during phenotypic stasis

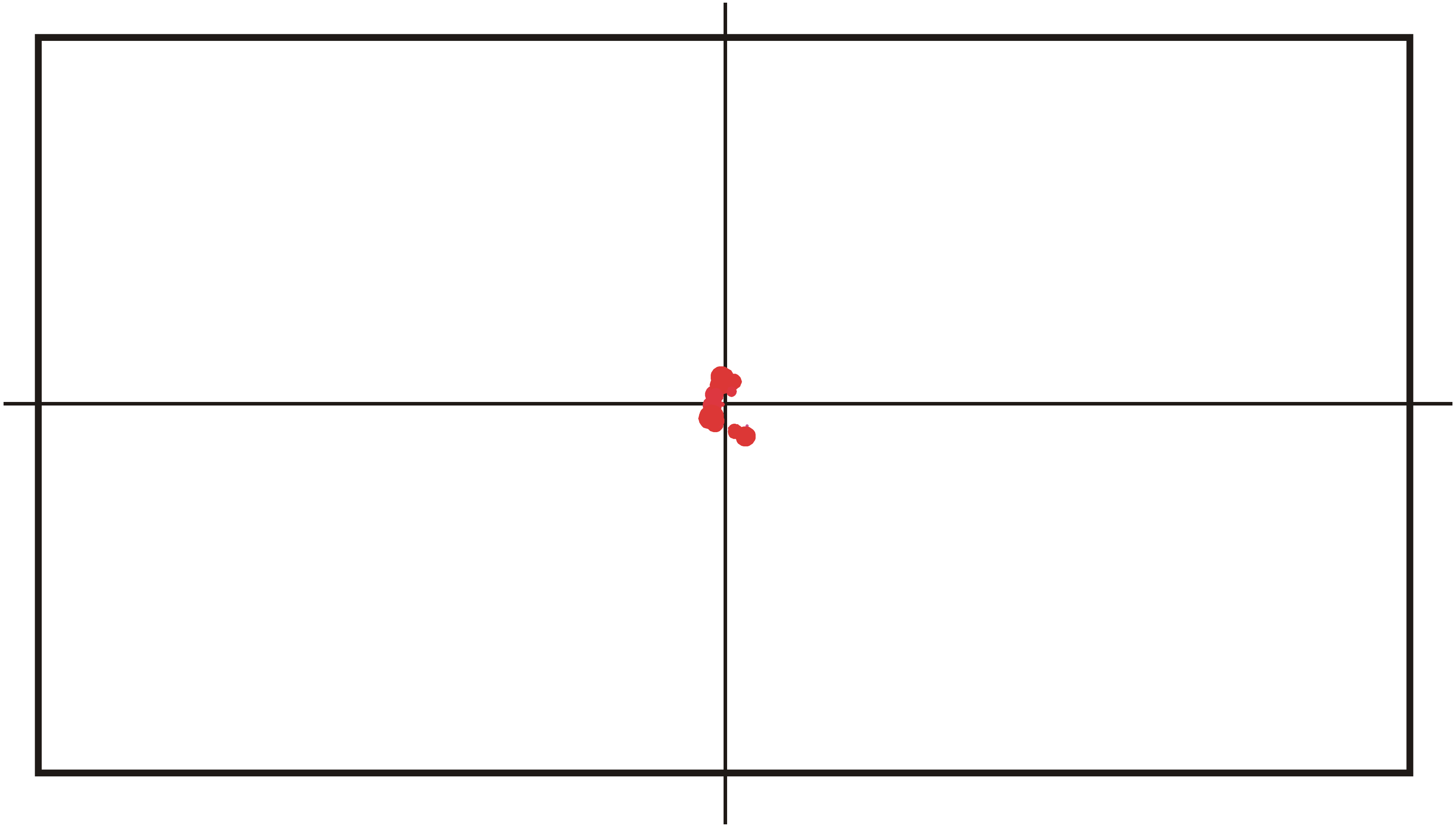


Variation in genotype space during optimization of phenotypes

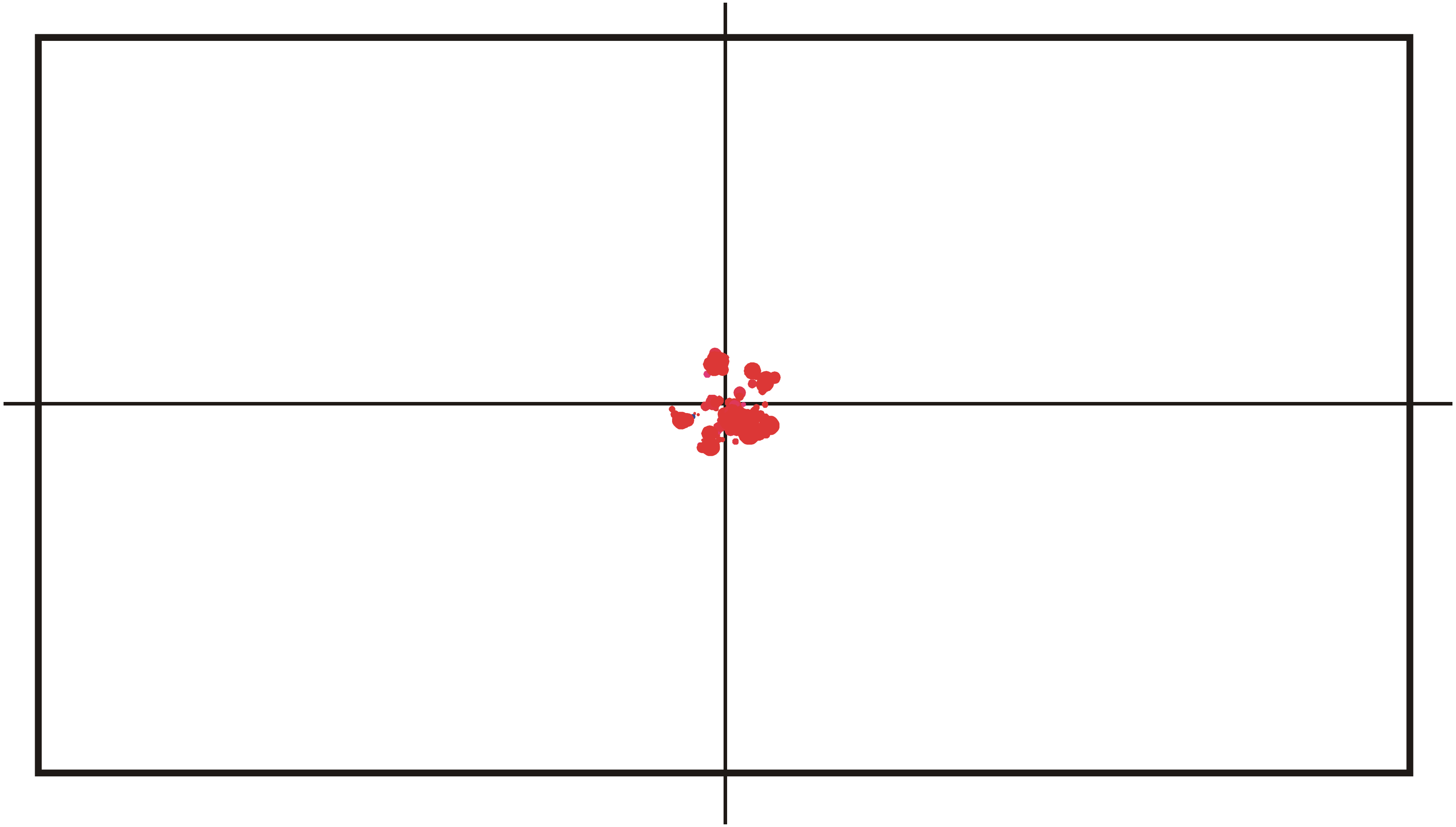
Mean Hamming distance within the population and **drift velocity of the population center** in sequence space.



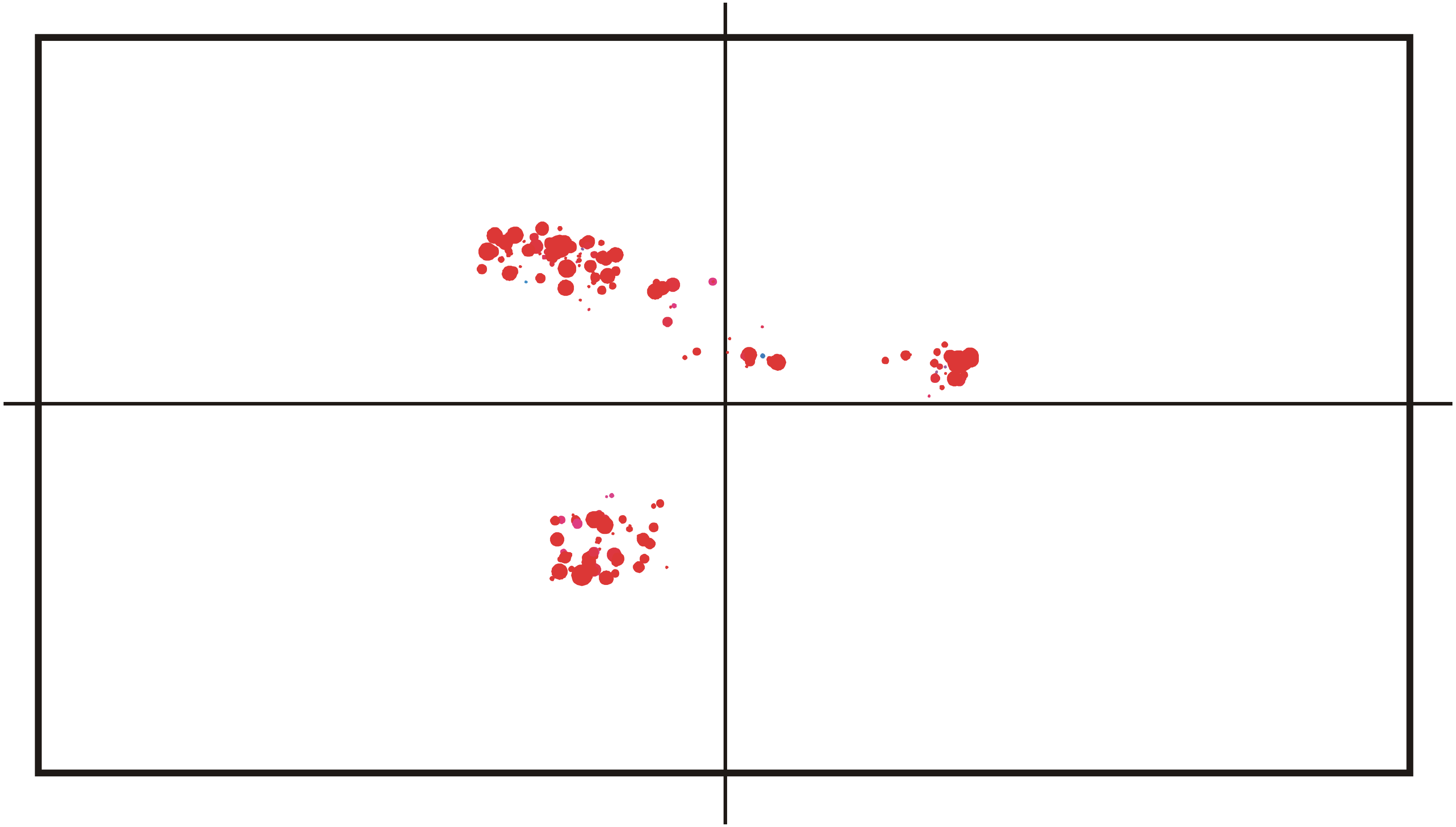
Spread of population in sequence space during a quasistationary epoch: t = 150



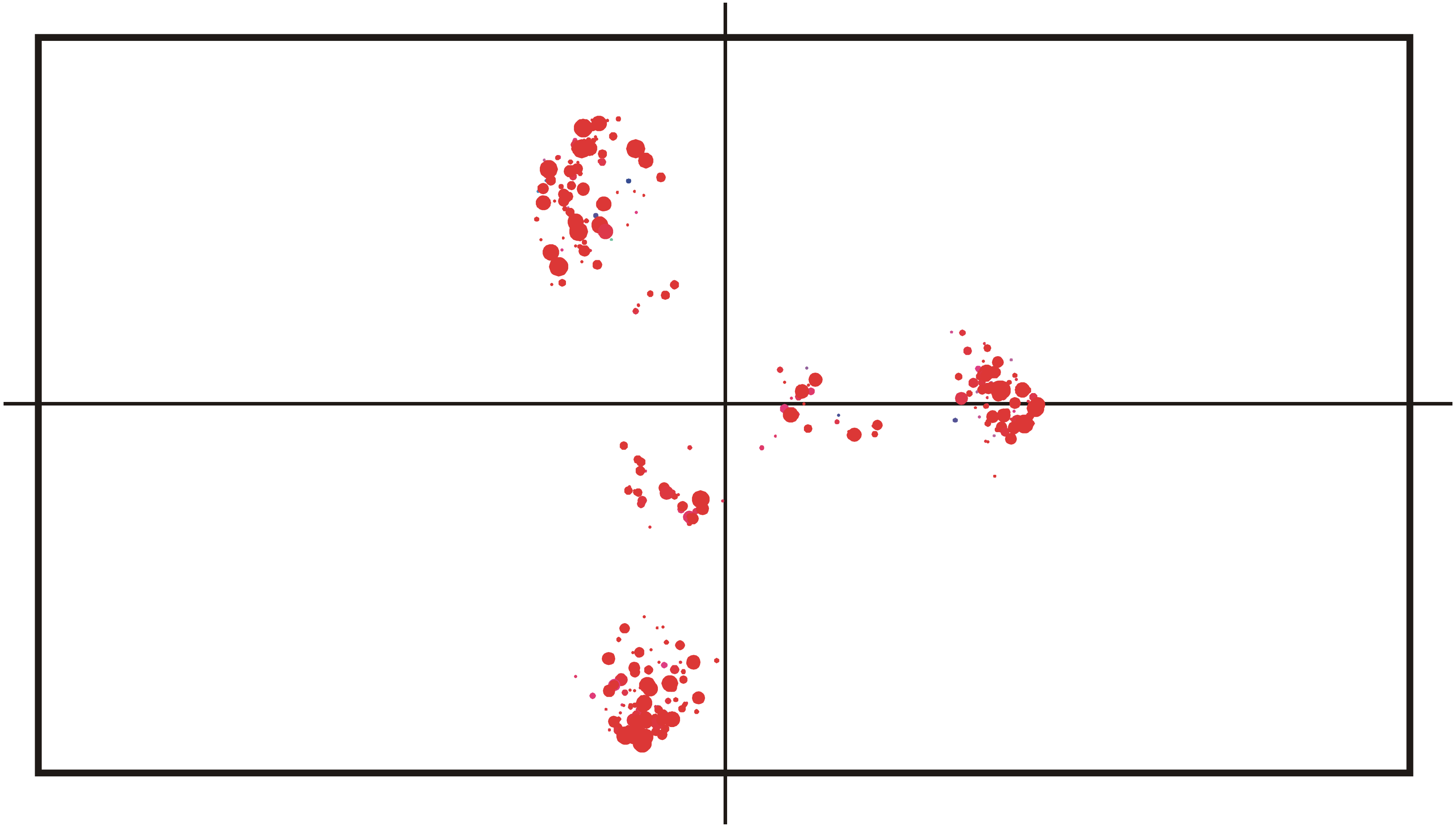
Spread of population in sequence space during a quasistationary epoch: t = 170



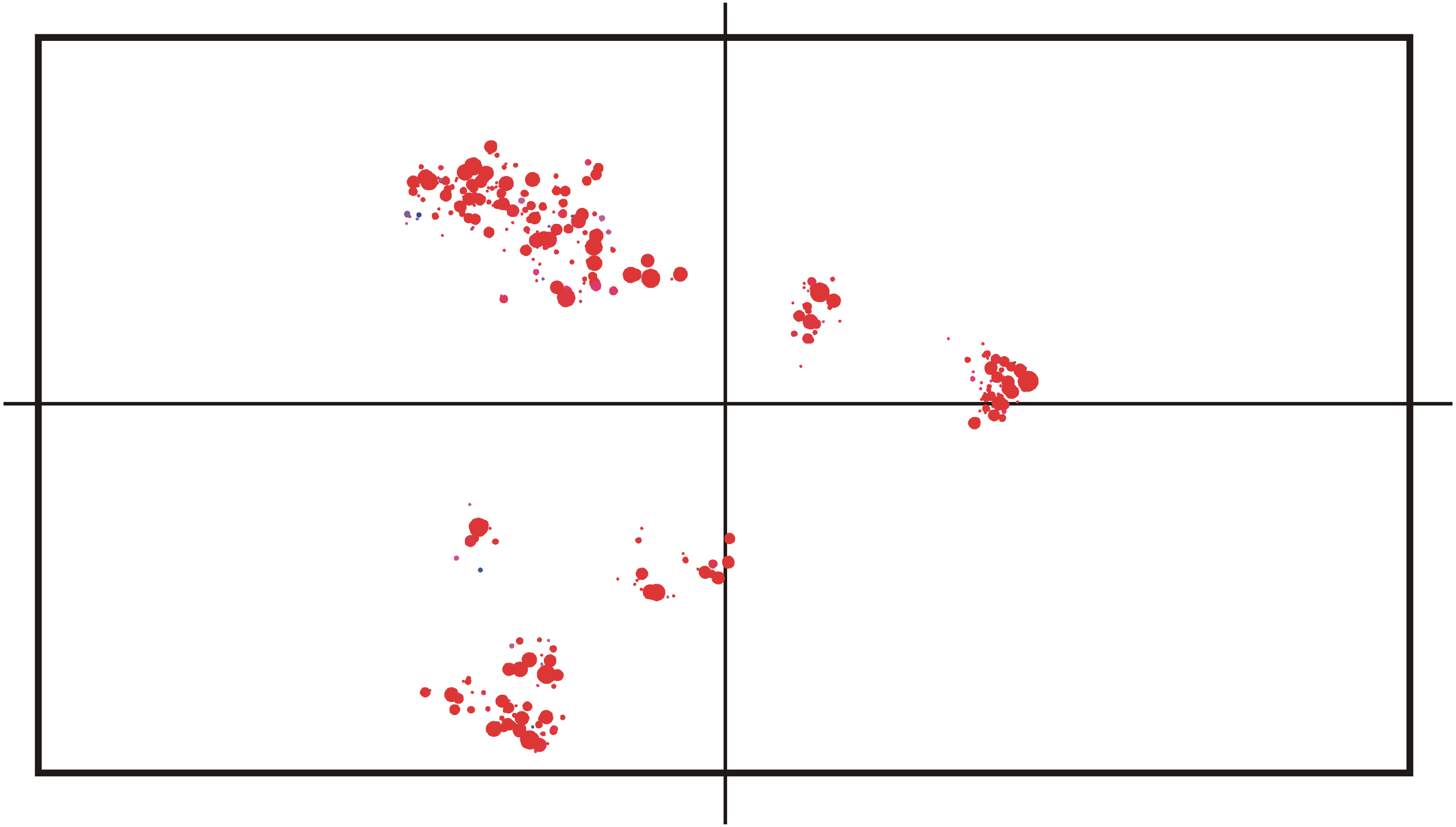
Spread of population in sequence space during a quasistationary epoch: t = 200



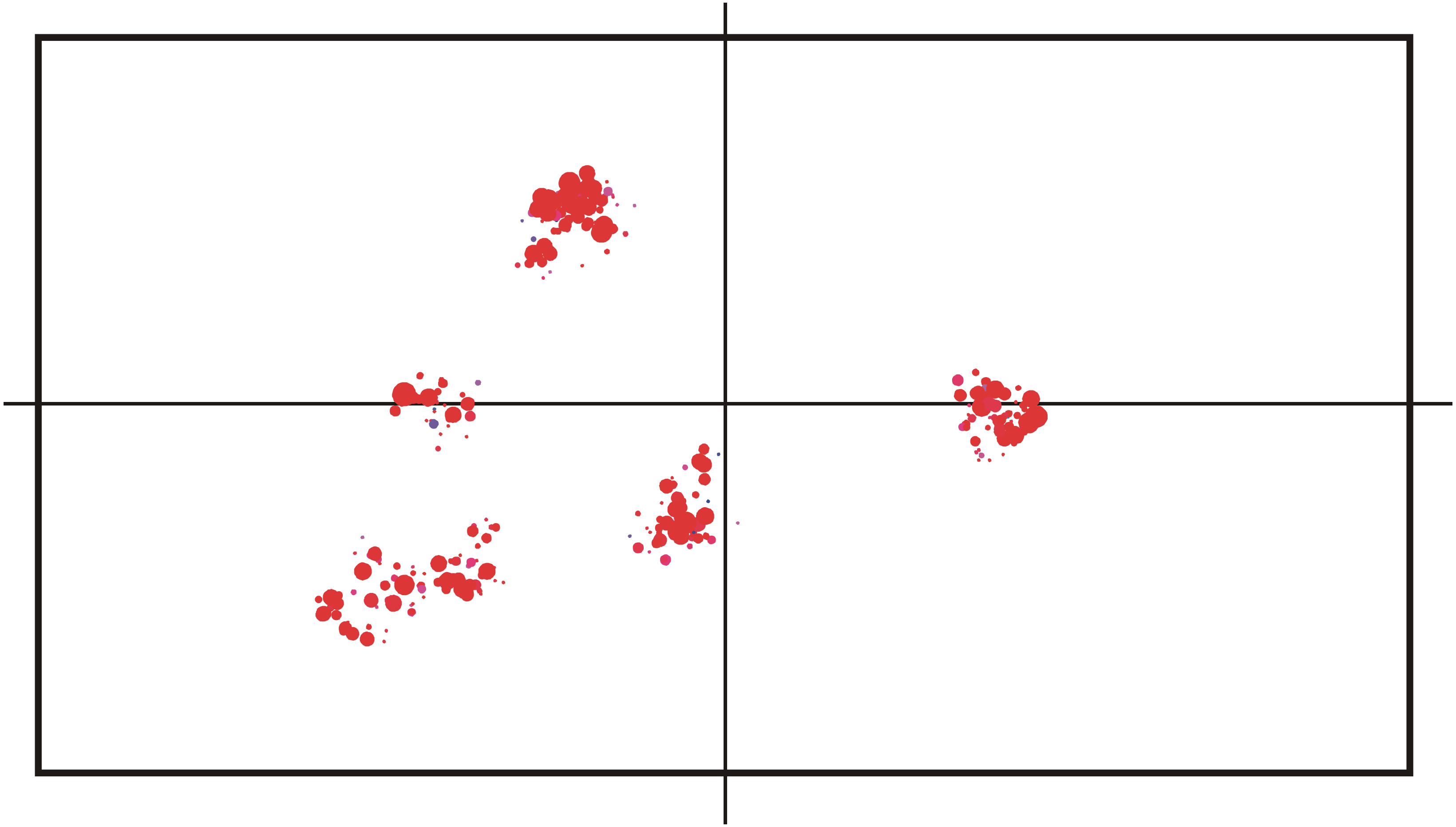
Spread of population in sequence space during a quasistationary epoch: t = 350



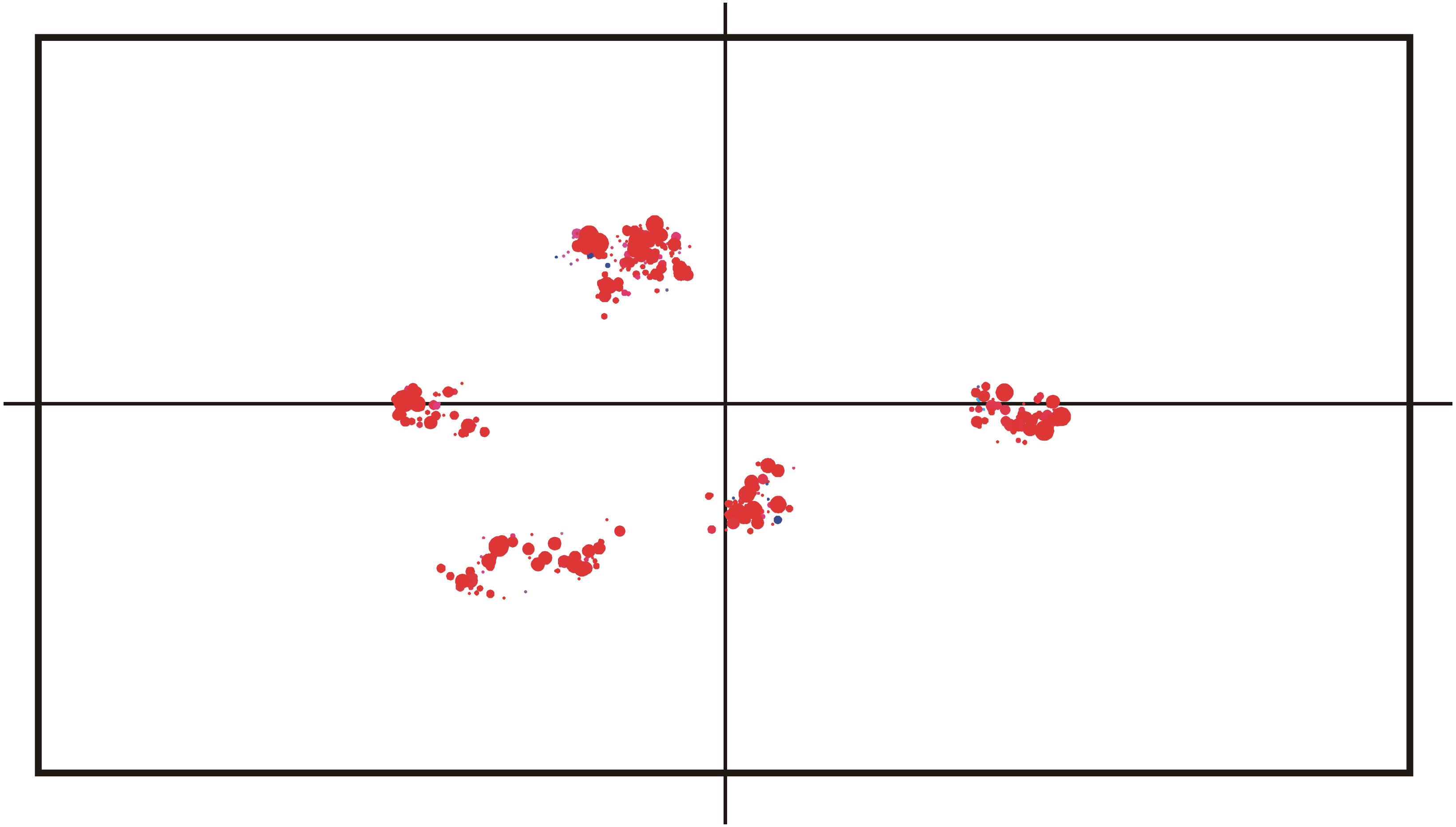
Spread of population in sequence space during a quasistationary epoch: t = 500



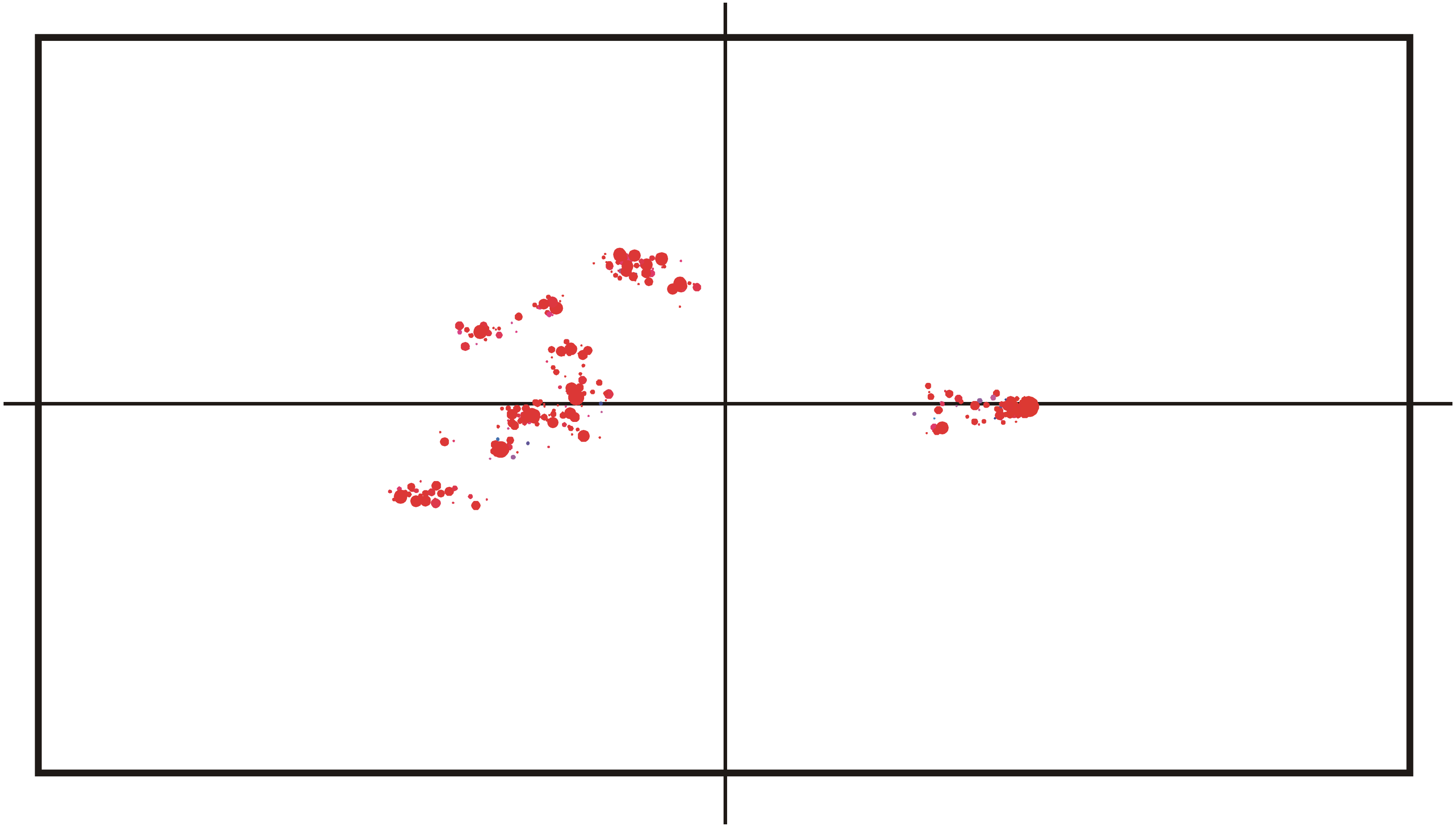
Spread of population in sequence space during a quasistationary epoch: t = 650



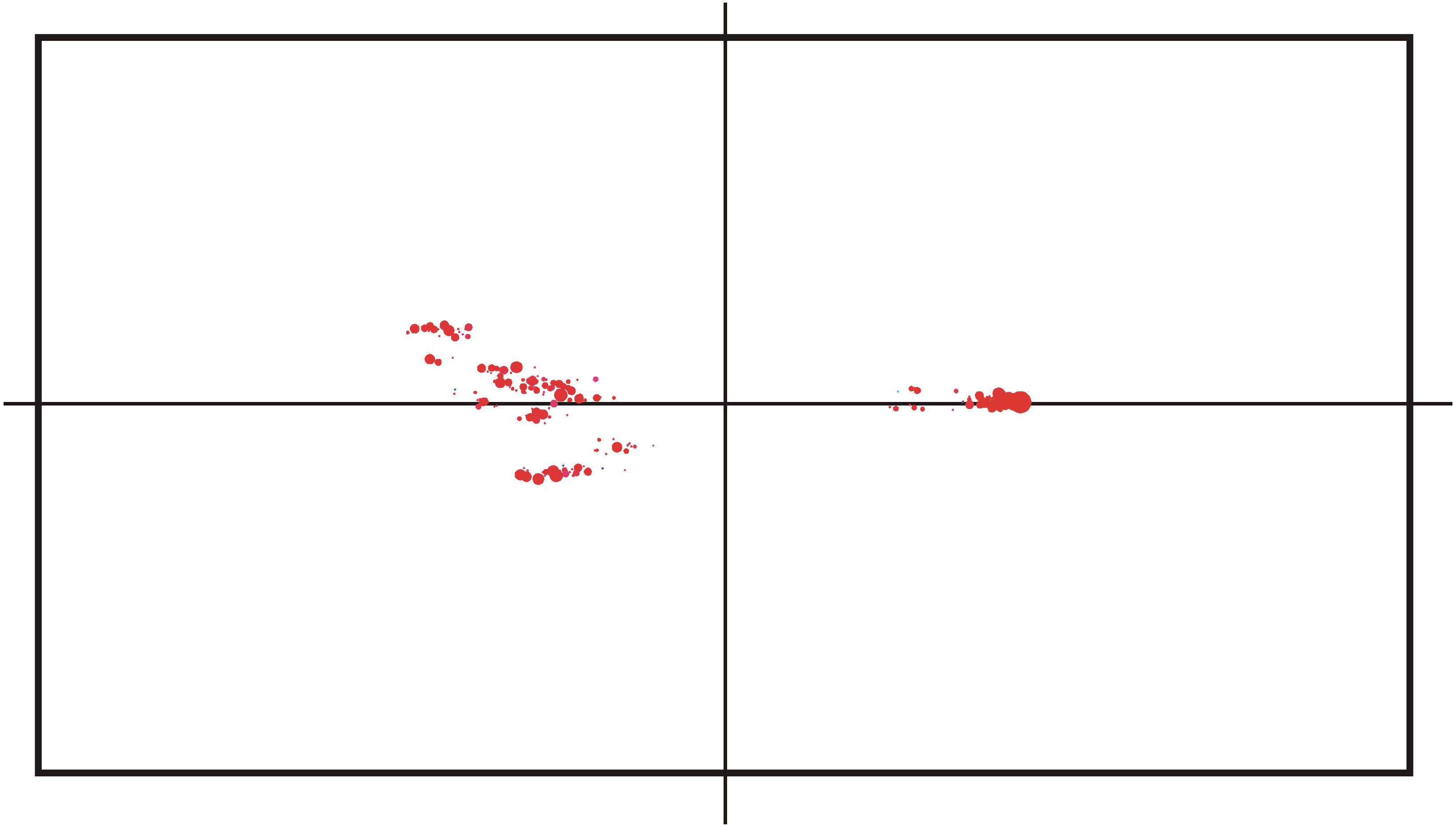
Spread of population in sequence space during a quasistationary epoch: t = 820



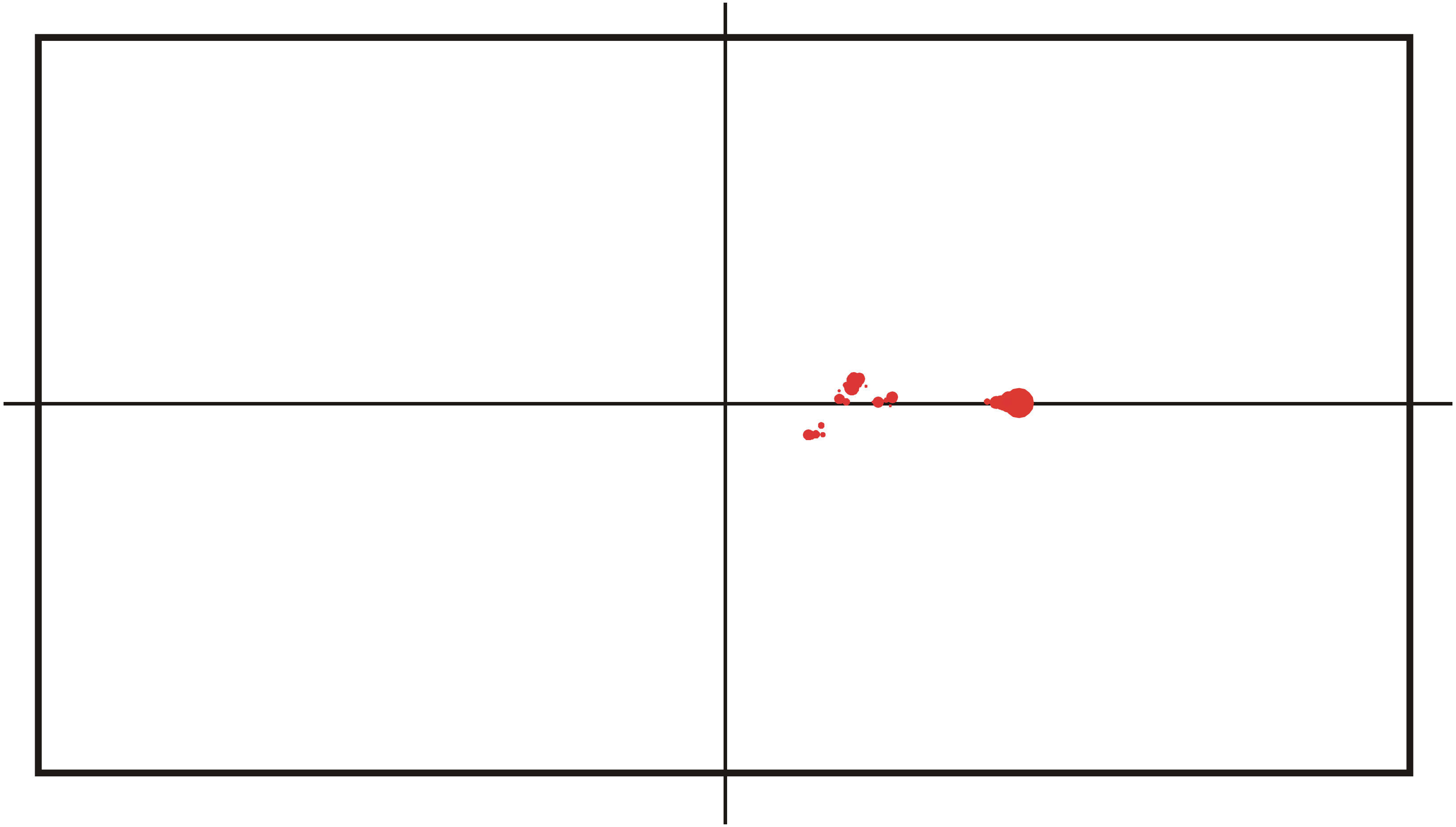
Spread of population in sequence space during a quasistationary epoch: t = 825



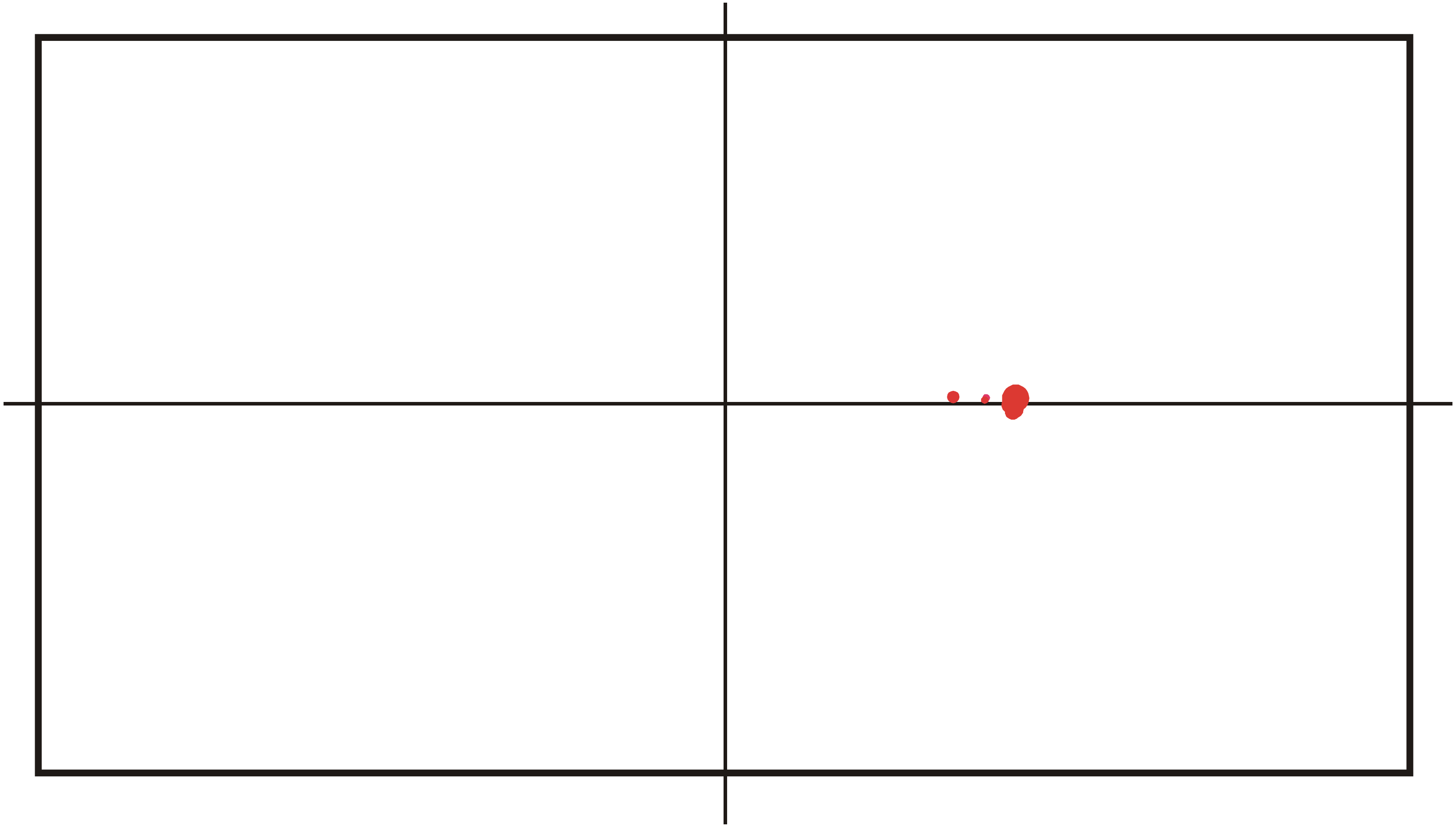
Spread of population in sequence space during a quasistationary epoch: t = 830



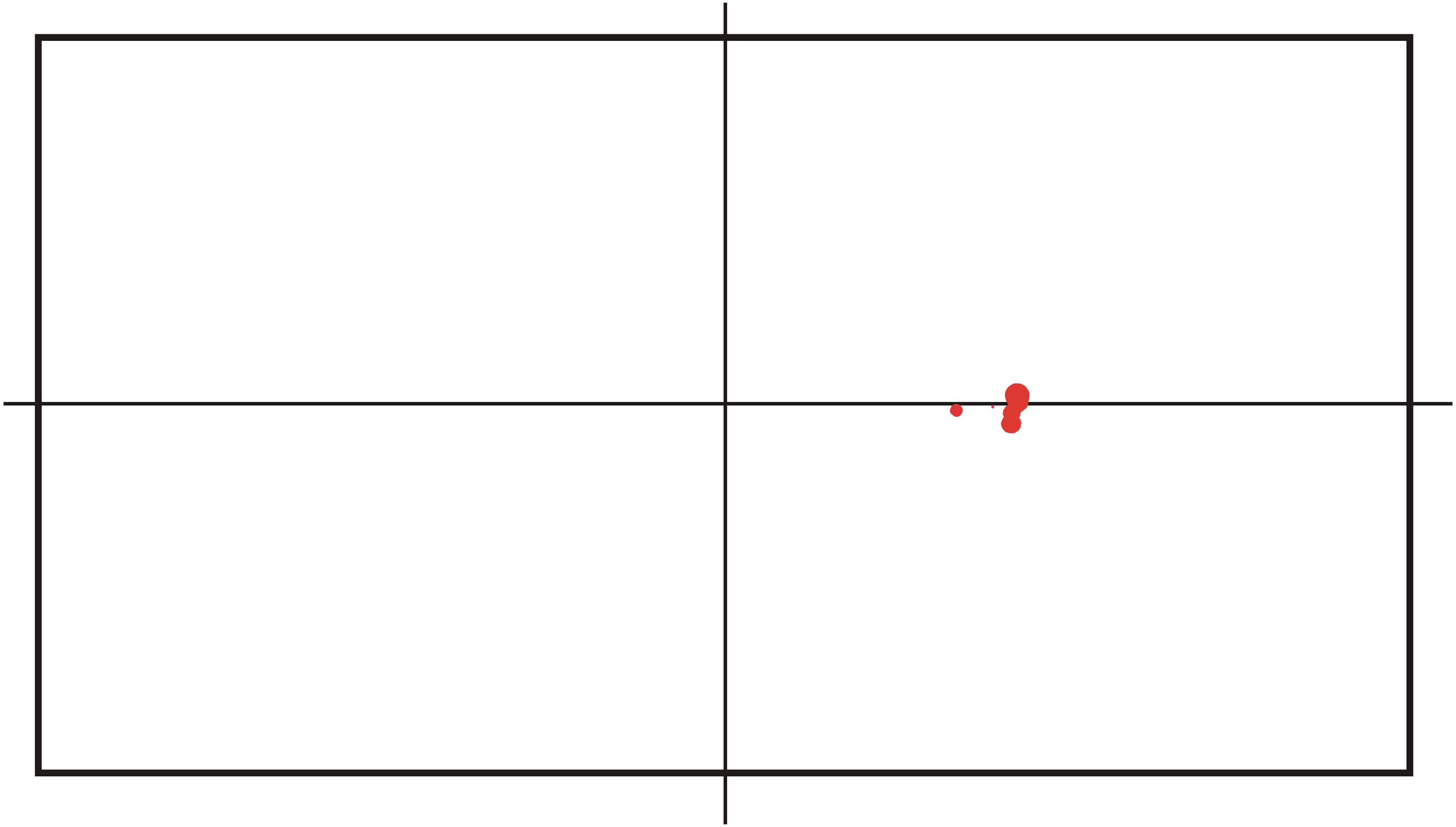
Spread of population in sequence space during a quasistationary epoch: t = 835



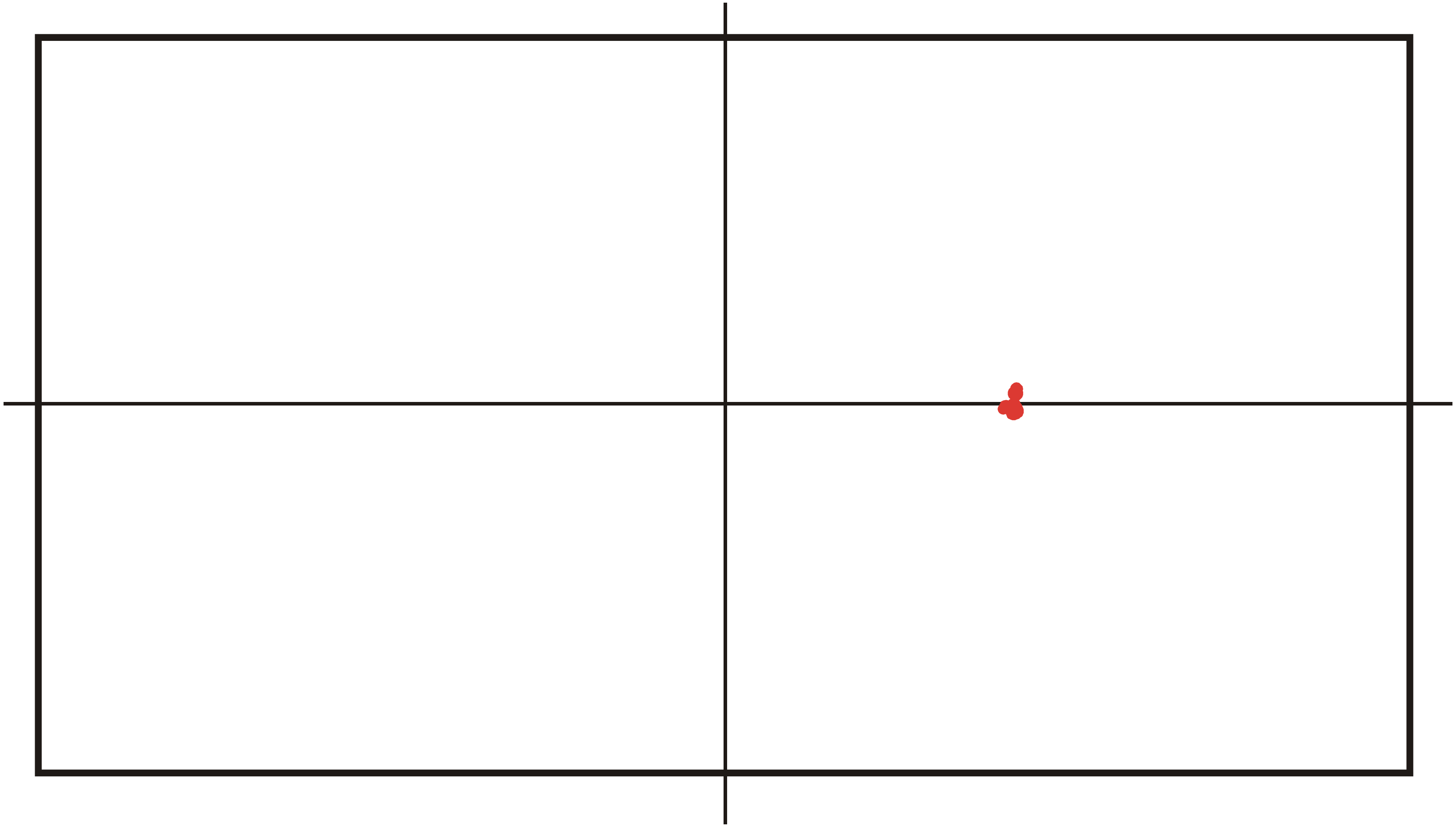
Spread of population in sequence space during a quasistationary epoch: t = 840



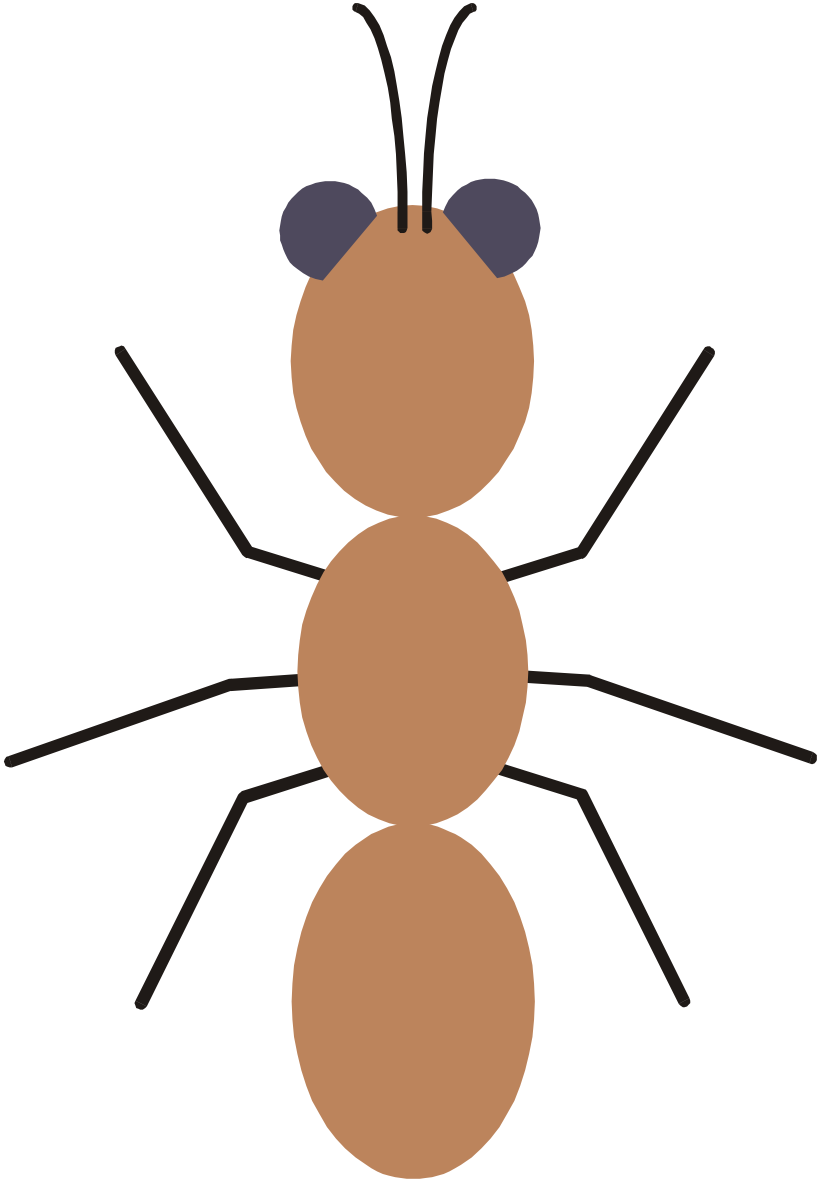
Spread of population in sequence space during a quasistationary epoch: t = 845



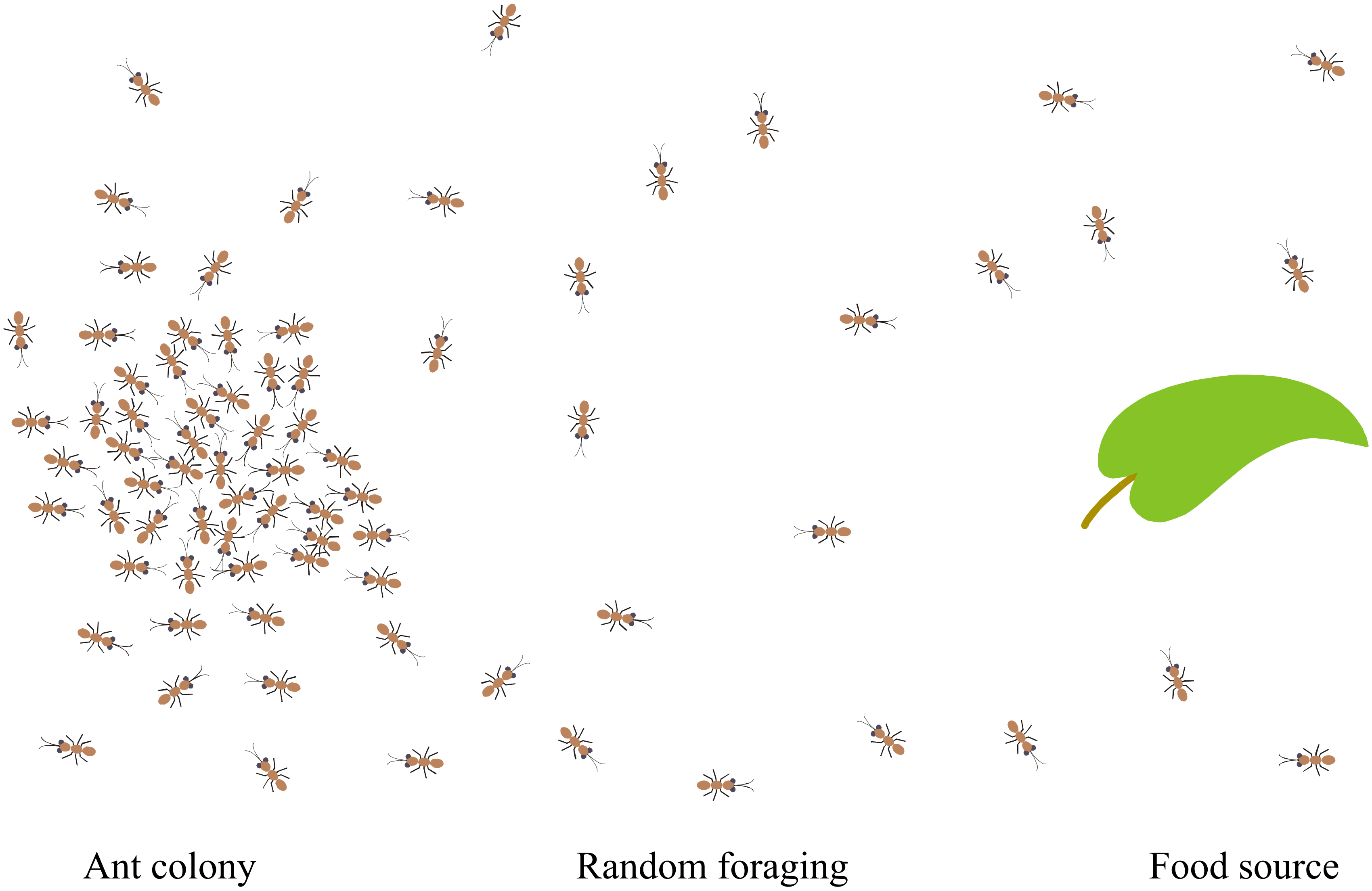
Spread of population in sequence space during a quasistationary epoch: t = 850



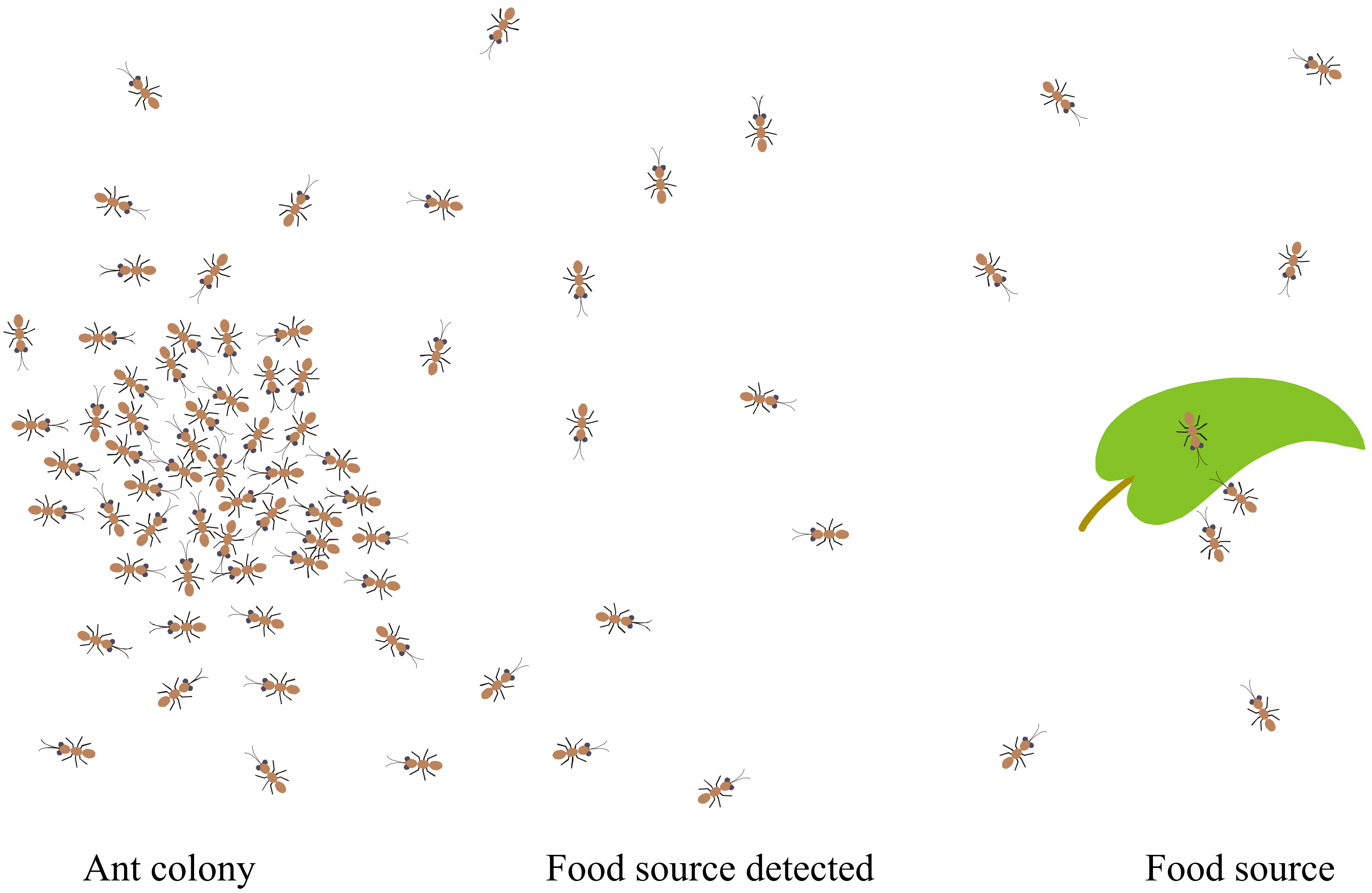
Spread of population in sequence space during a quasistationary epoch: t = 855



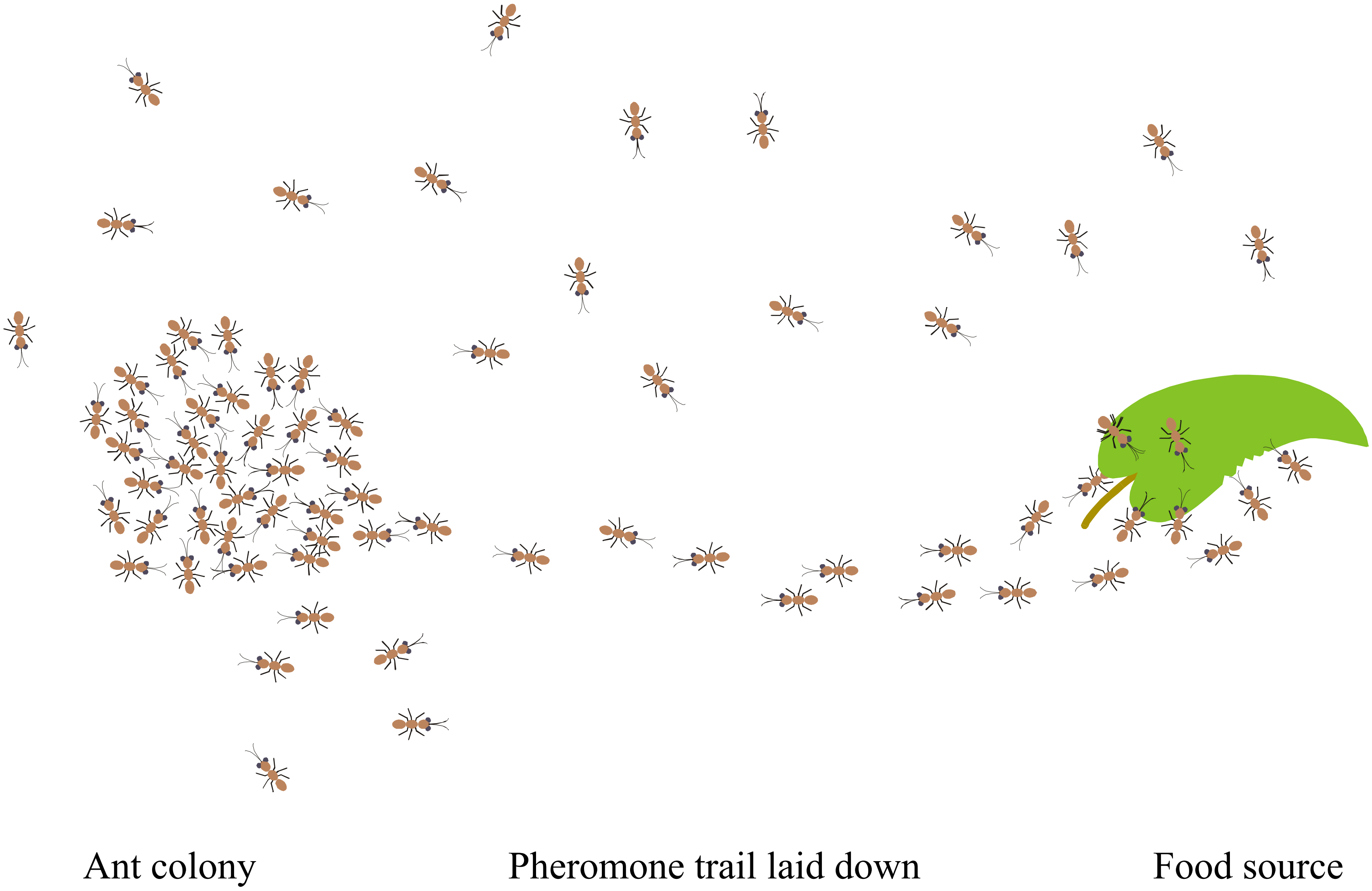
Element class 2: The ant worker



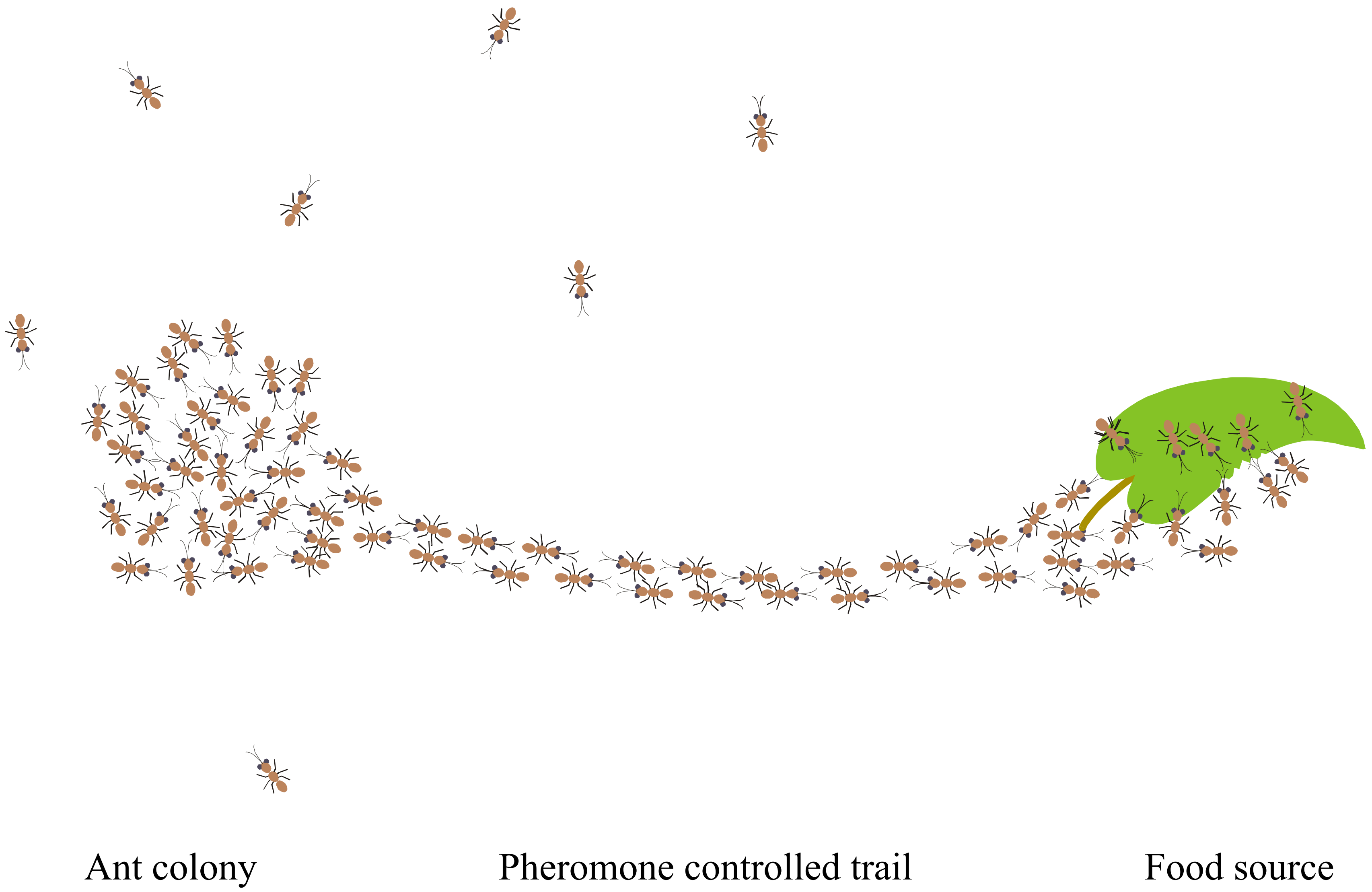
Foraging behavior of ant colonies



Foraging behavior of ant colonies



Foraging behavior of ant colonies



Foraging behavior of ant colonies

	RNA model	**Foraging behavior of ant colonies**
Element	**RNA molecule**	**Individual worker ant**
Mechanism relating elements	Mutation in quasi-species	Genetics of kinship
Search process	Optimization of RNA structure	Recruiting of food
Search space	Sequence space	Three-dimensional space
Random step	Mutation	Element of ant walk
Self-enhancing process	Replication	Secretion of pheromone
Interaction between elements	Mean replication rate	Mean pheromone concentration
Goal of the search	Target structure	Food source
Temporary memory	RNA sequences in population	Pheromone trail
‘Learning’ entity	**Population of molecules**	**Ant colony**

Learning at population or colony level by trial and error

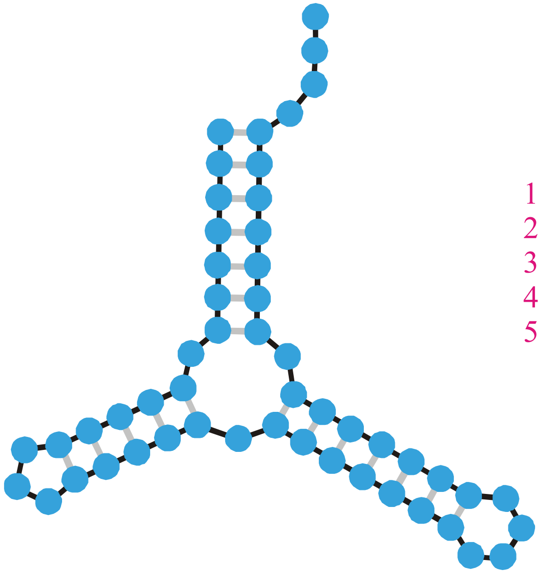
Two examples: (i) RNA model and (ii) ant colony

1. RNA structure, replication kinetics, and origin of information

2. Evolution *in silico* and optimization of RNA structures

3. Random walks and ‚ensemble learning‘

4. **Sequence-structure maps, neutral networks, and intersections**



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

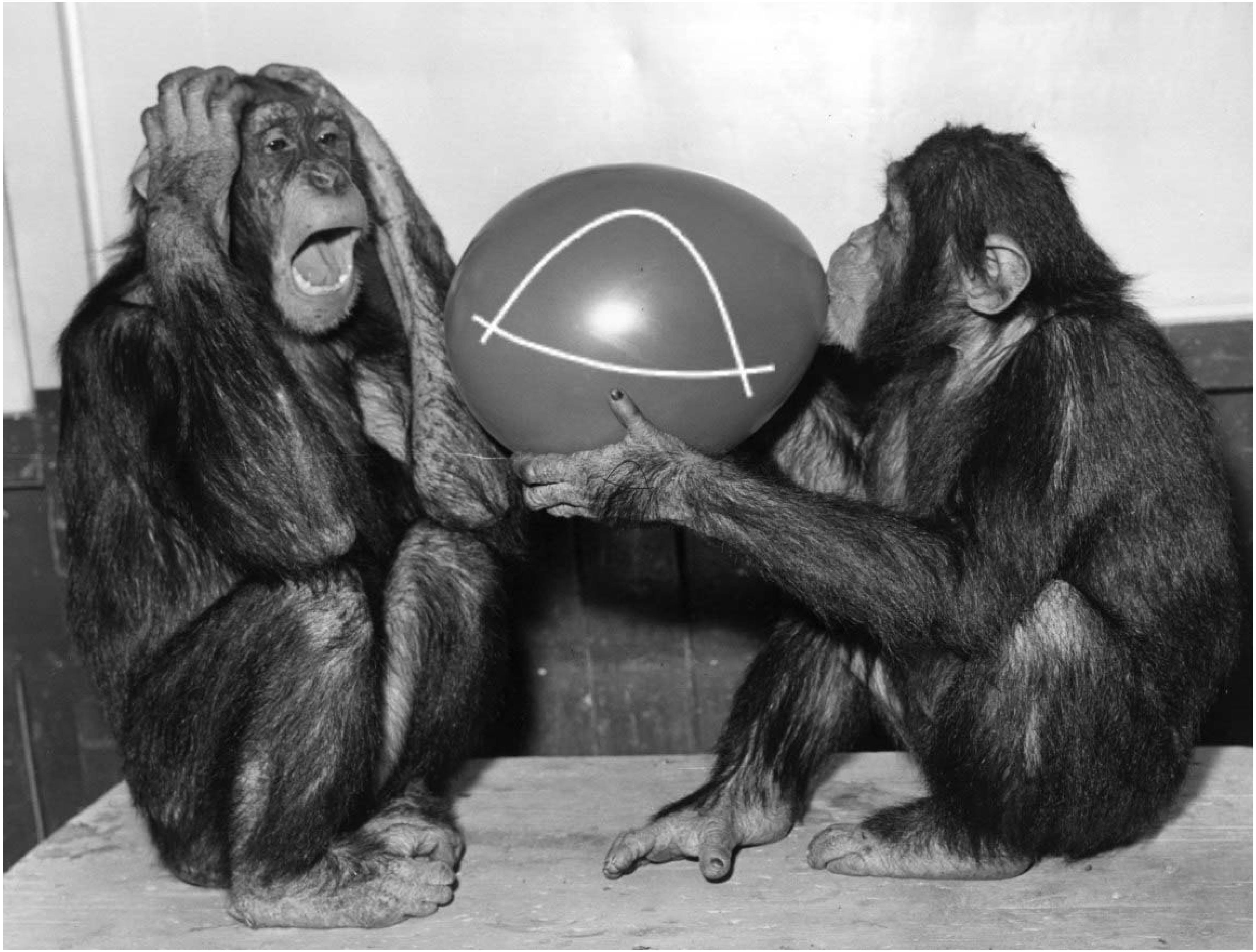
CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

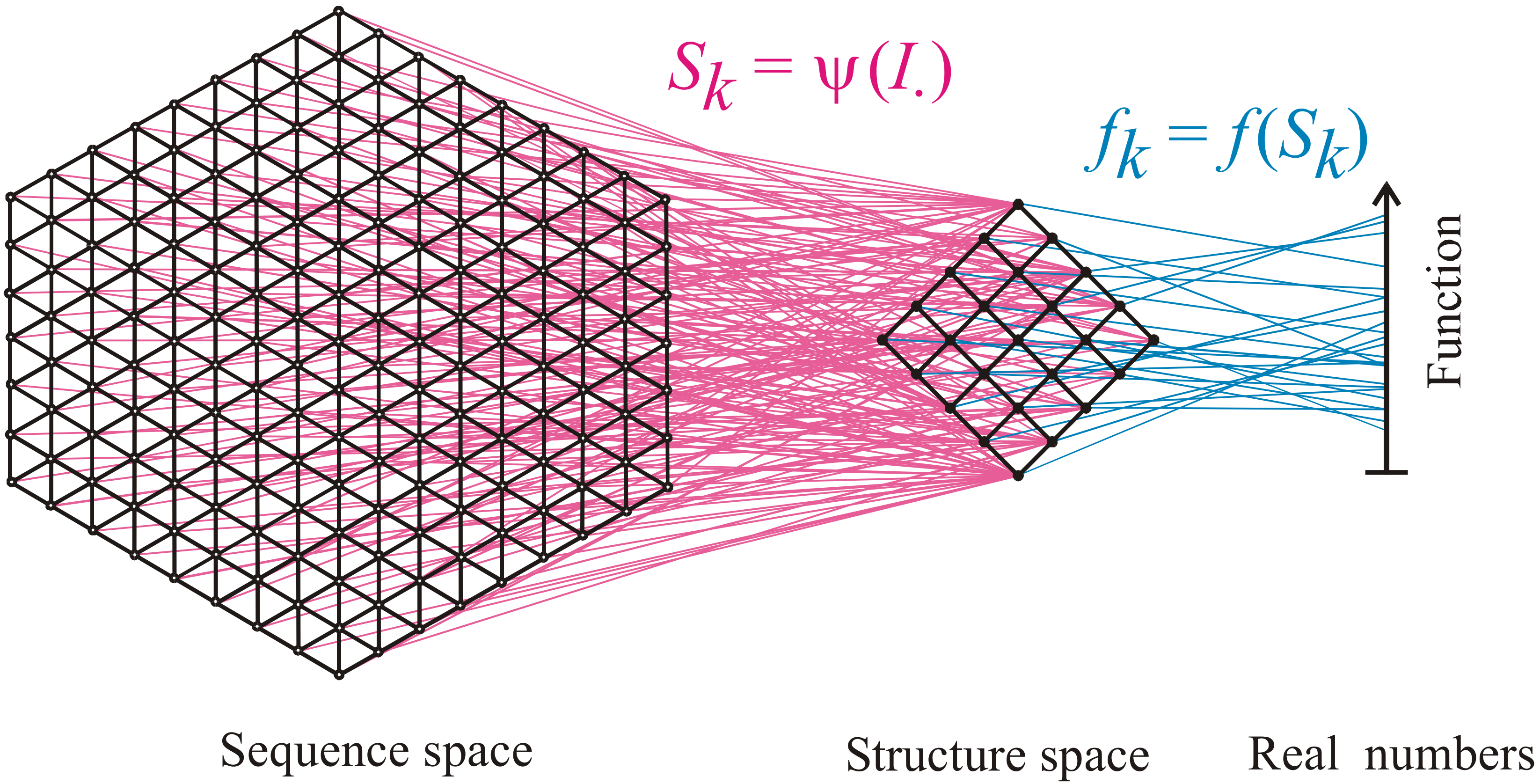
CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

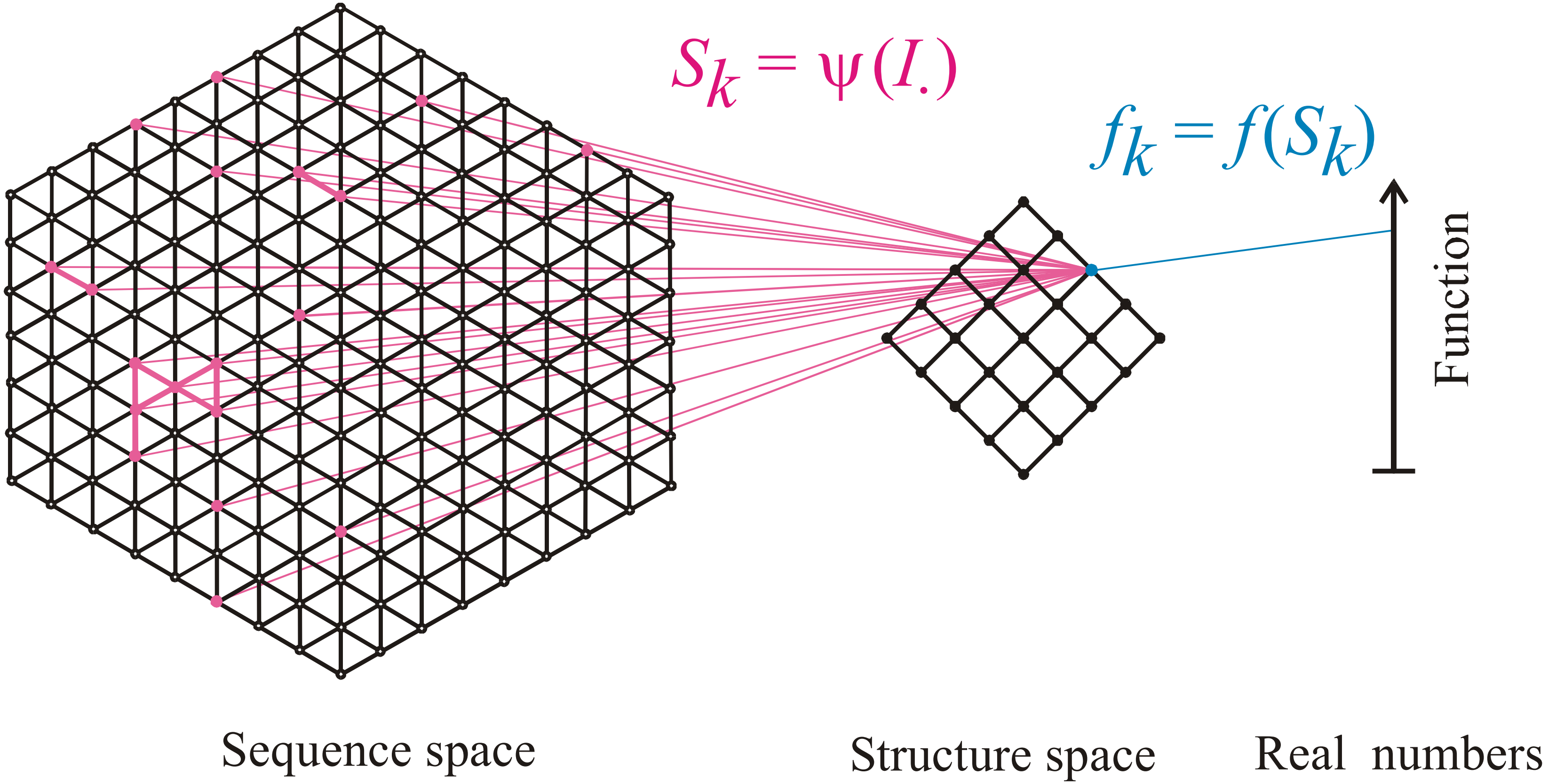
Inverse folding of RNA secondary structures

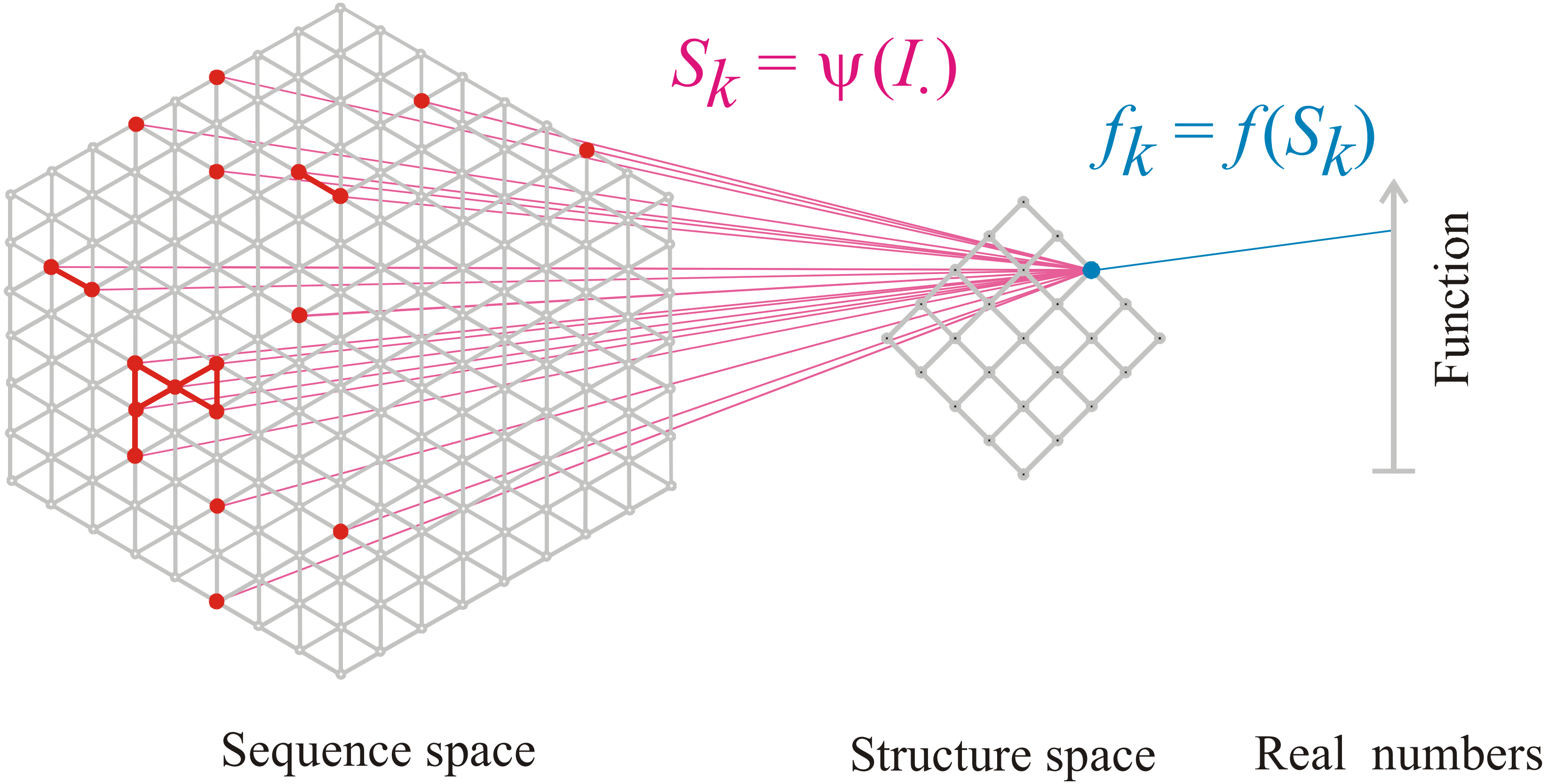
The inverse folding algorithm searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.





Mapping from sequence space into structure space and into function


$S_k = \psi(I.)$
$f_k = f(S_k)$
Function
Sequence space
Structure space
Real numbers



The pre-image of the structure S_k in sequence space is the **neutral network G_k**

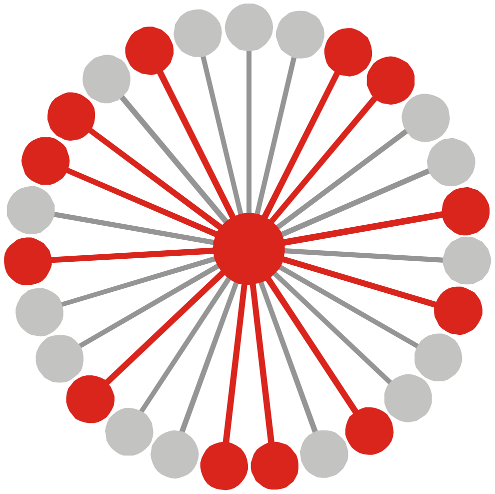
Neutral networks are sets of sequences forming the same structure. G_k is the pre-image of the structure S_k in sequence space:

$$G_k = y^{-1}(S_k) \, \pi \, \{y_j \mid y(I_j) = S_k\}$$

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

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

Neutral networks can be modelled by **random graphs** in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.



$$G_k = y^{-1}(S_k) \; \text{U} \; \{ \, I_j \mid y(I_j) = S_k \, \}$$

$$\lambda_j = 12 \,/\, 27 = 0.444 \;, \quad \bar{\lambda}_k = \frac{\underset{j \in |G_k|}{\text{ø}} \; l_j(k)}{|G_k|}$$

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

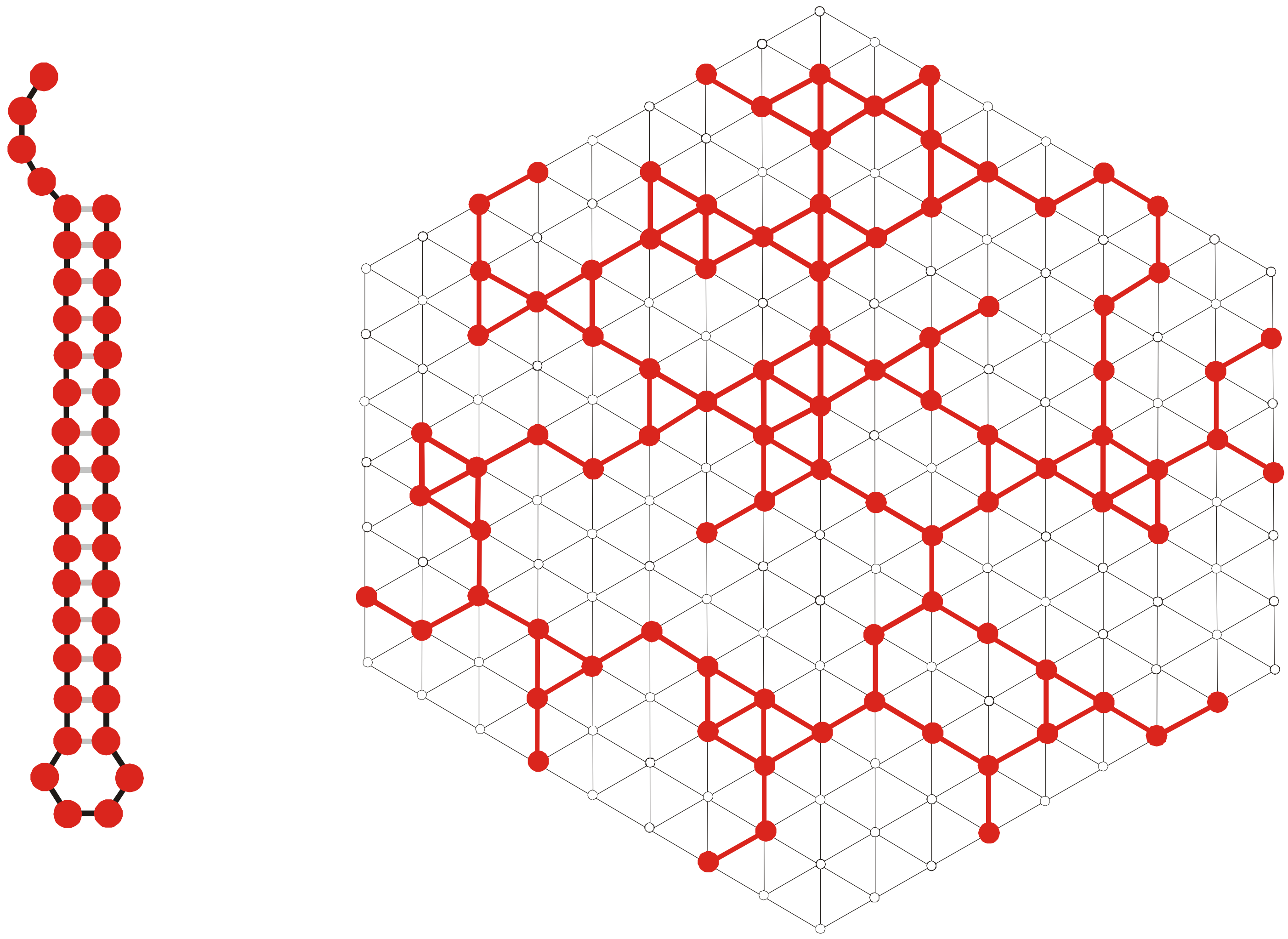
Alphabet size k : **AUGC** ñ k = 4

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

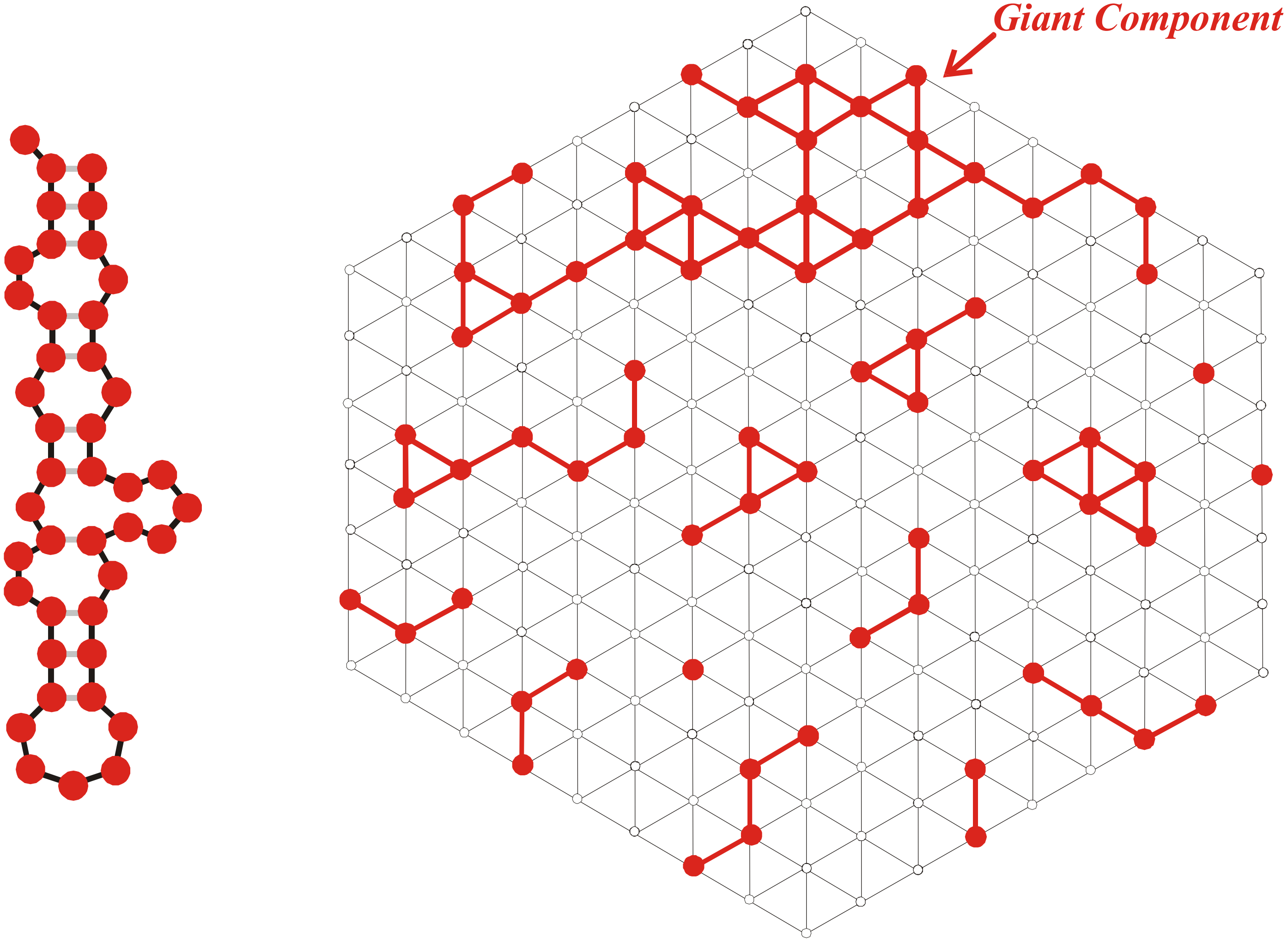
$\bar{\lambda}_k < \lambda_{cr}$ . . . . network G_k is **not** connected

k	l_{cr}	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

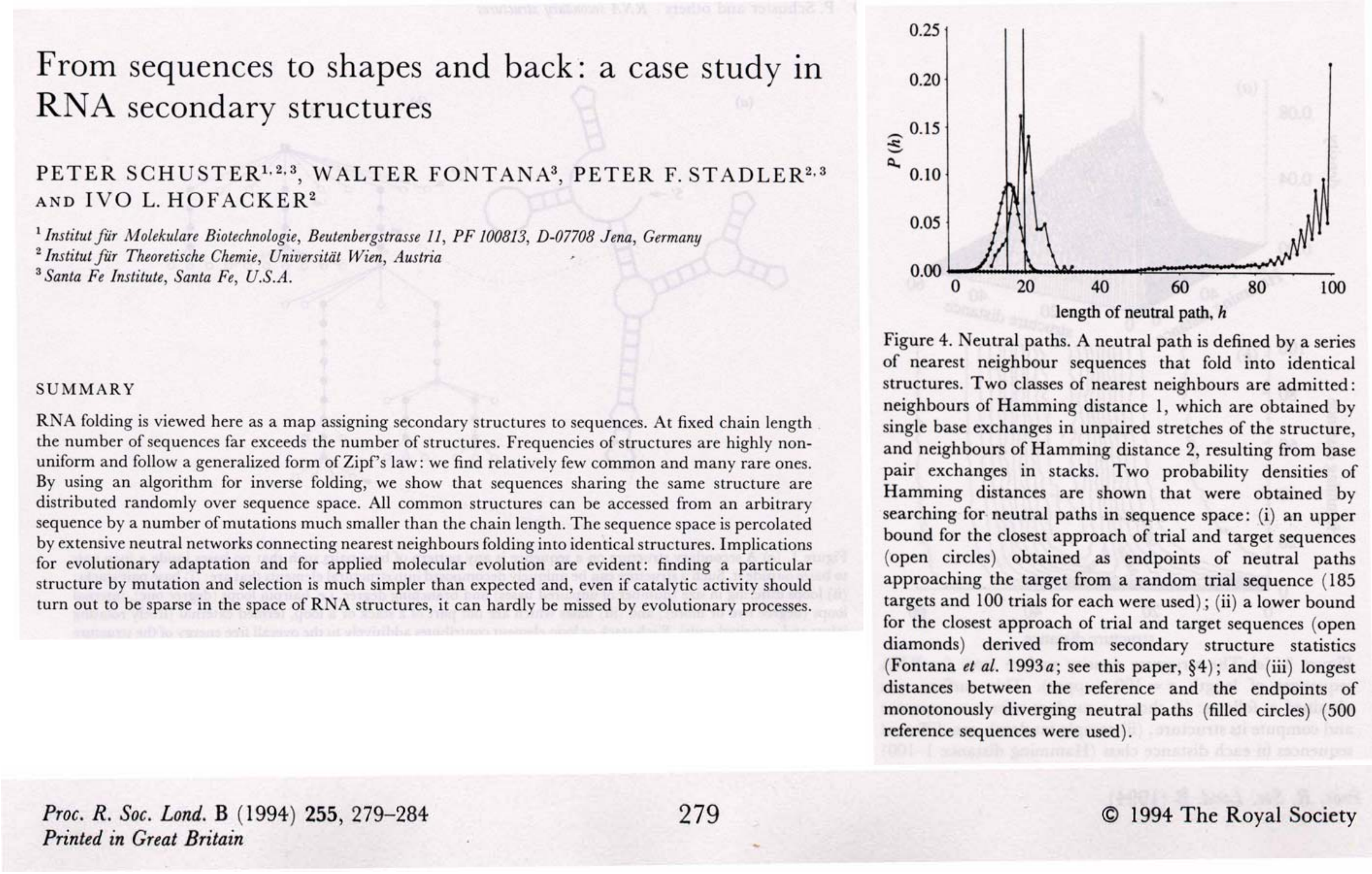
Mean degree of neutrality and connectivity of neutral networks



A connected neutral network



A multi-component neutral network



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.*

SUMMARY

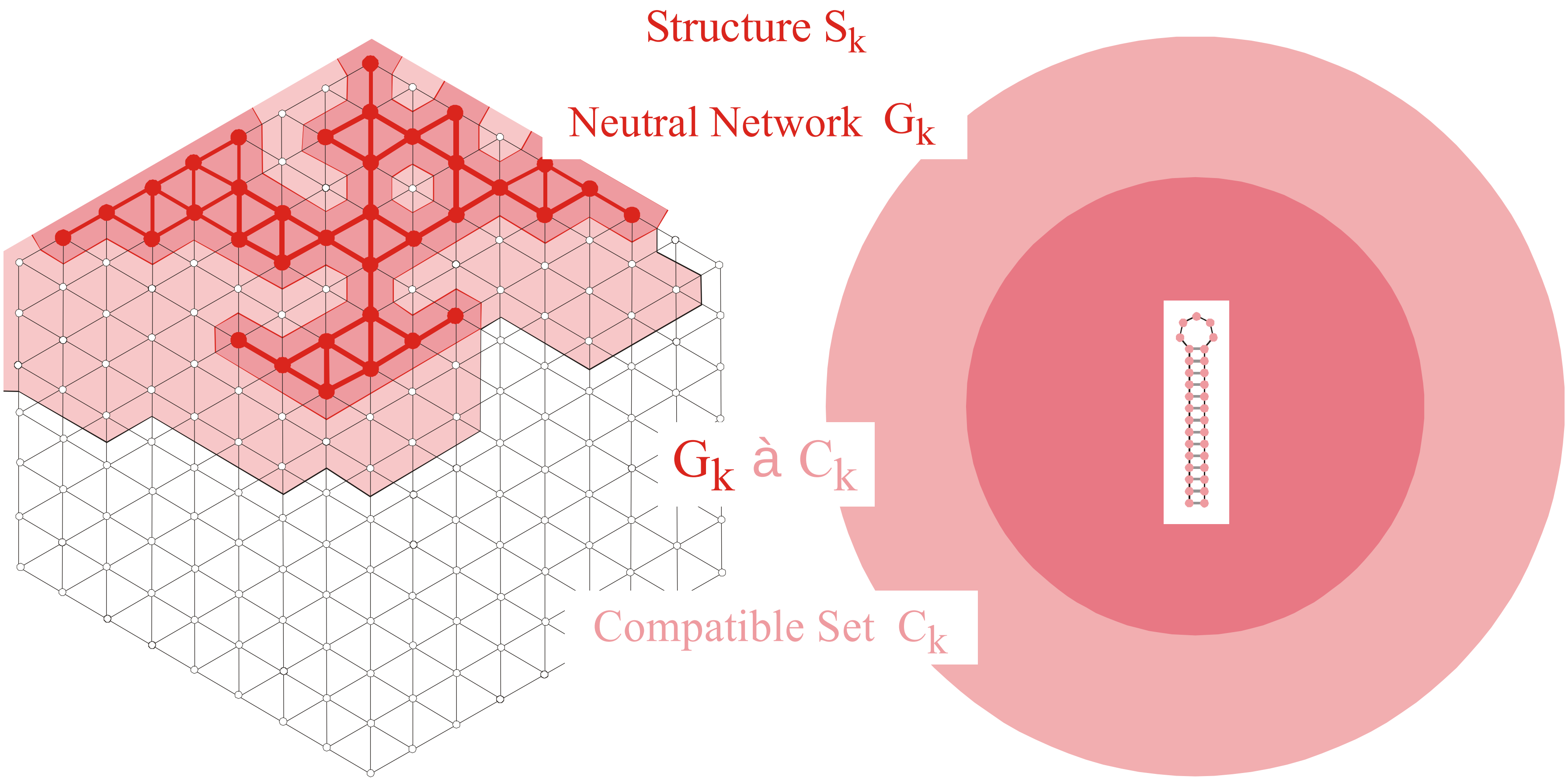
RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare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

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

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

279

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



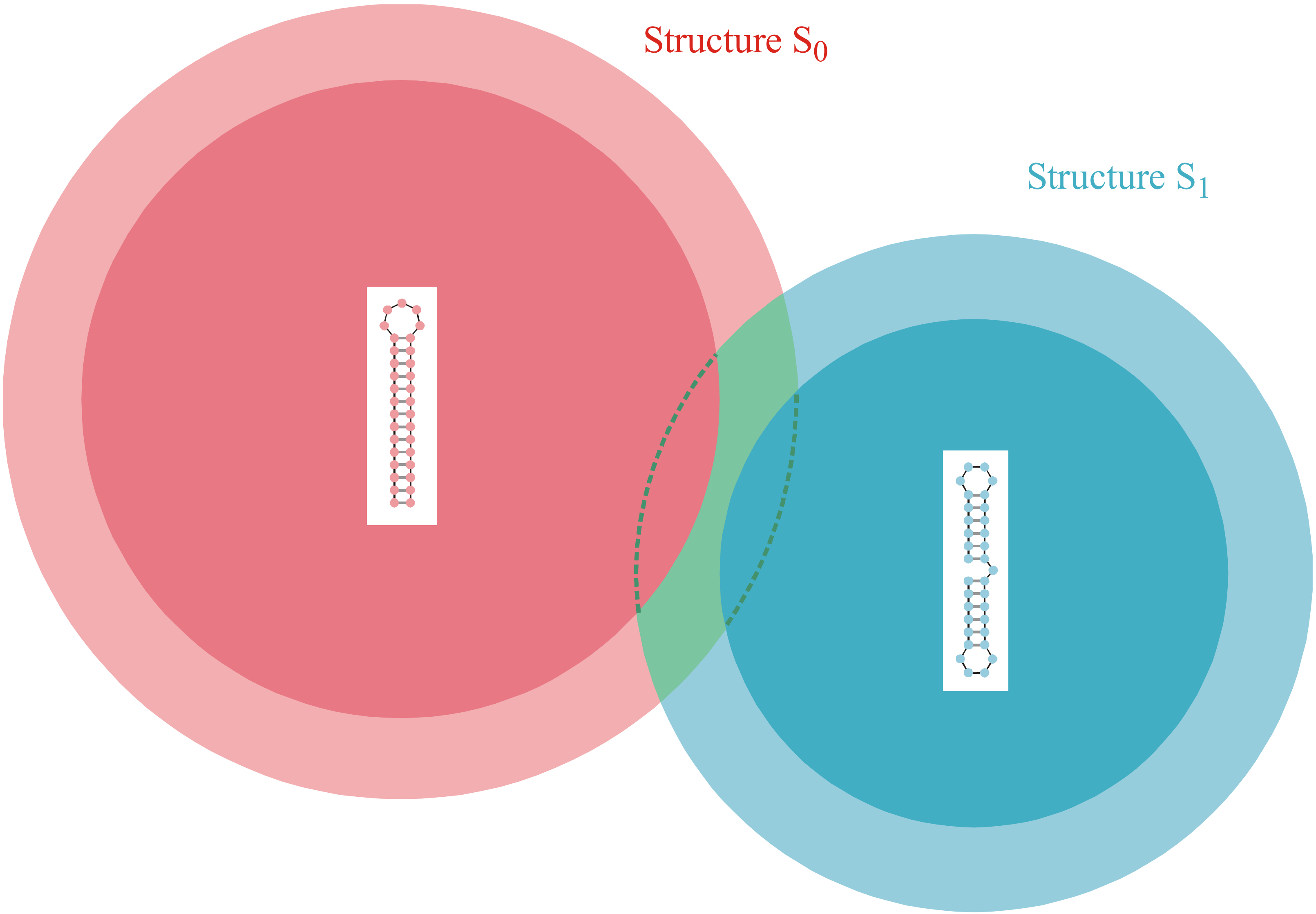
Structure S_k

Neutral Network G_k

G_k à C_k

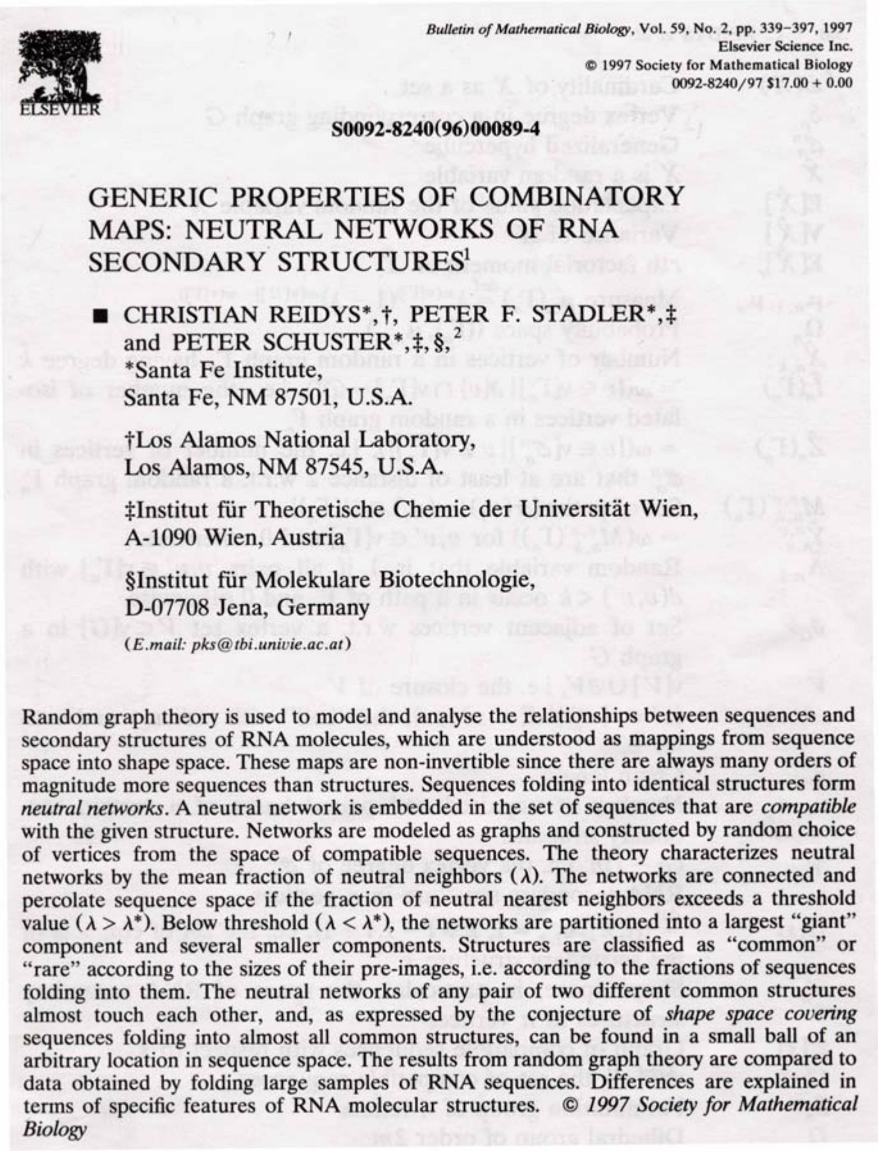
Compatible Set C_k

The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: C_0 Ú C_1 â Ù



Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.
© 1997 Society for Mathematical Biology
0092-8240/97 $17.00 + 0.00

ELSEVIER

S0092-8240(96)00089-4

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

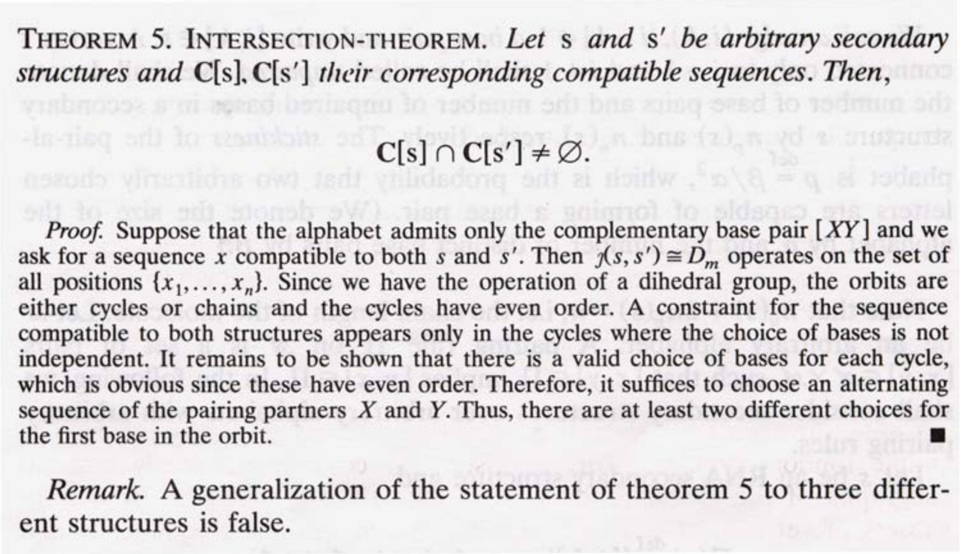
■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡ and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†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* C[s], C[s′] *their corresponding compatible sequences. Then,*

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

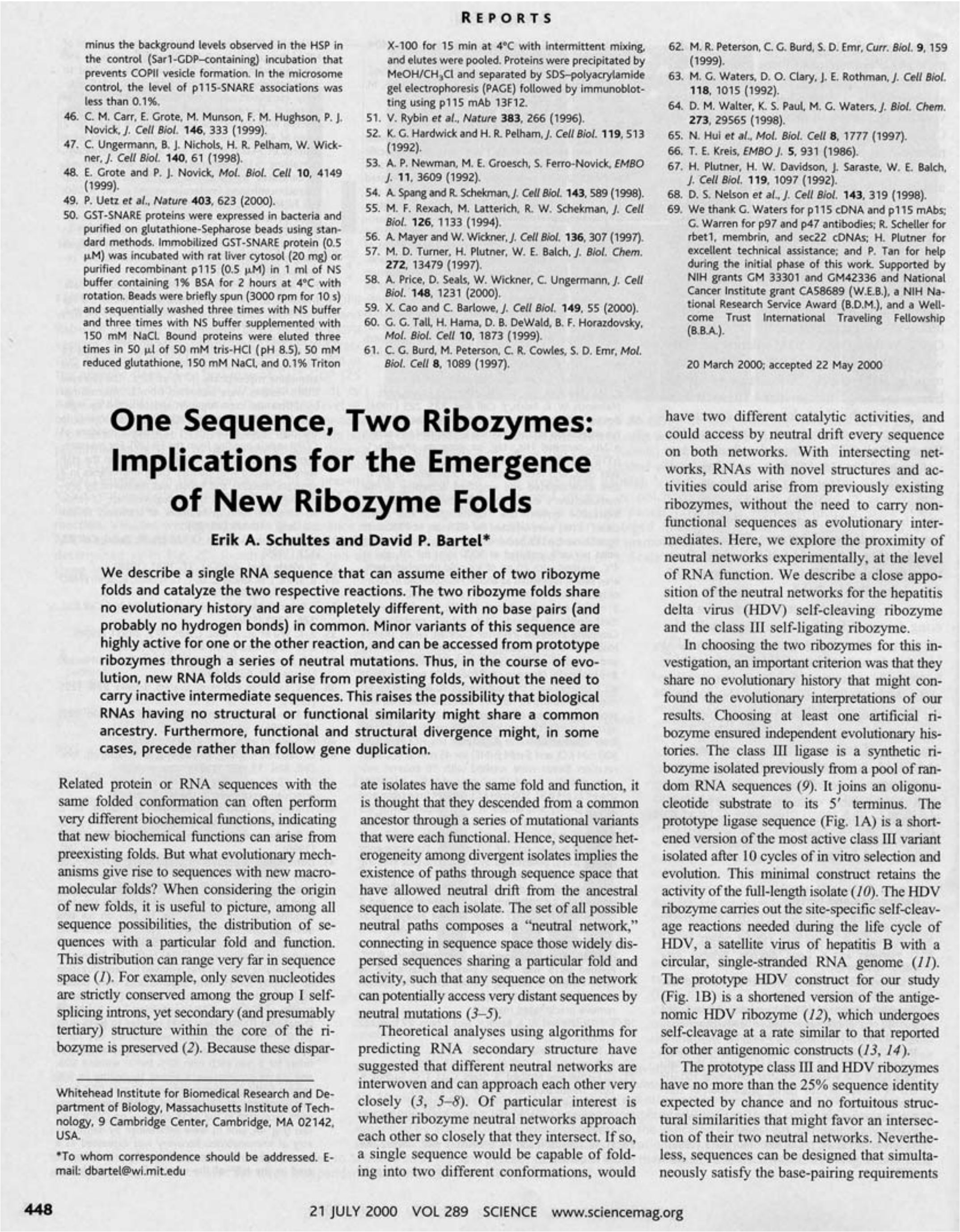
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\mathcal{J}(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \dots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either 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**

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/CH_3Cl 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*).

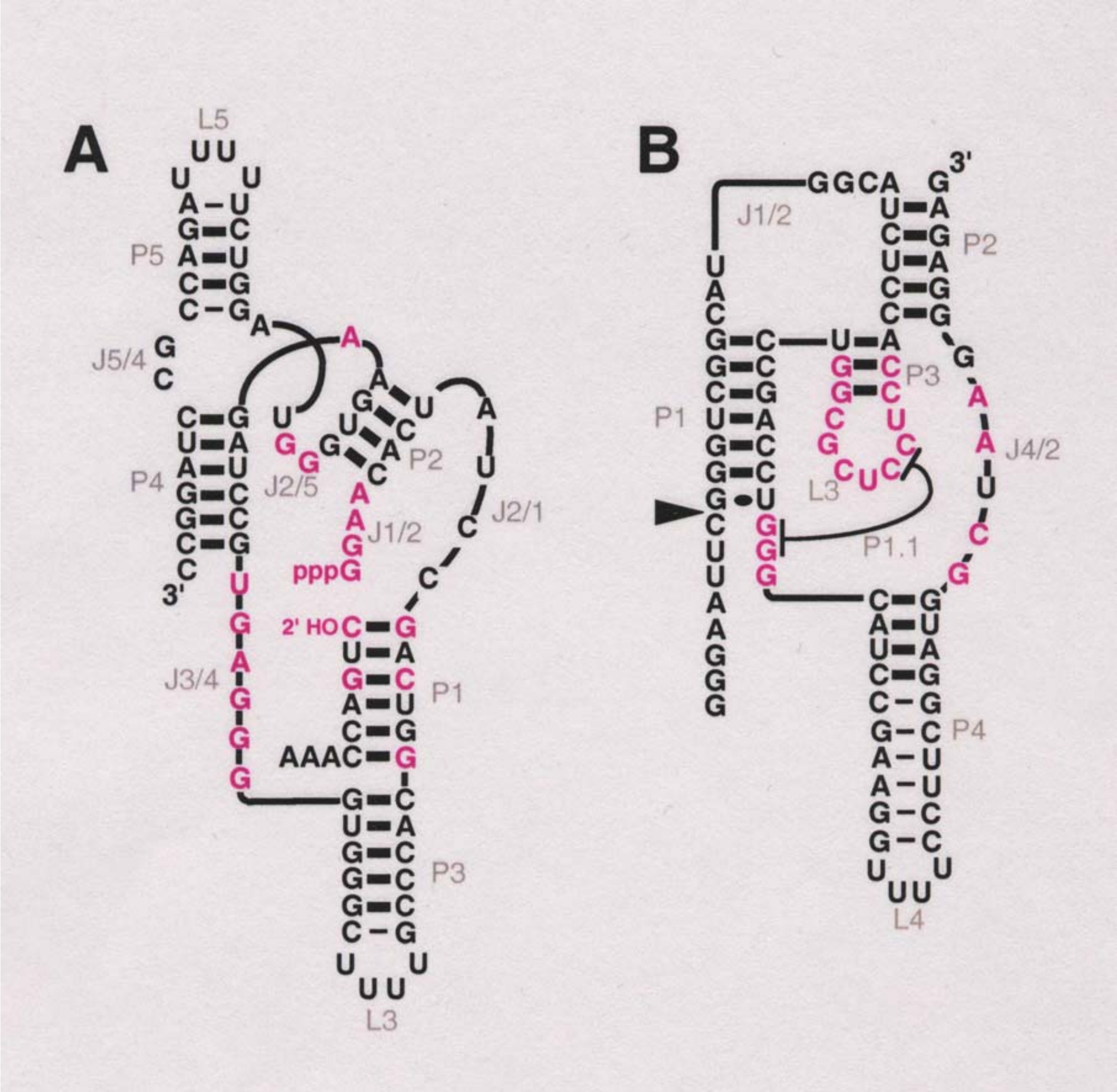
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. 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*).

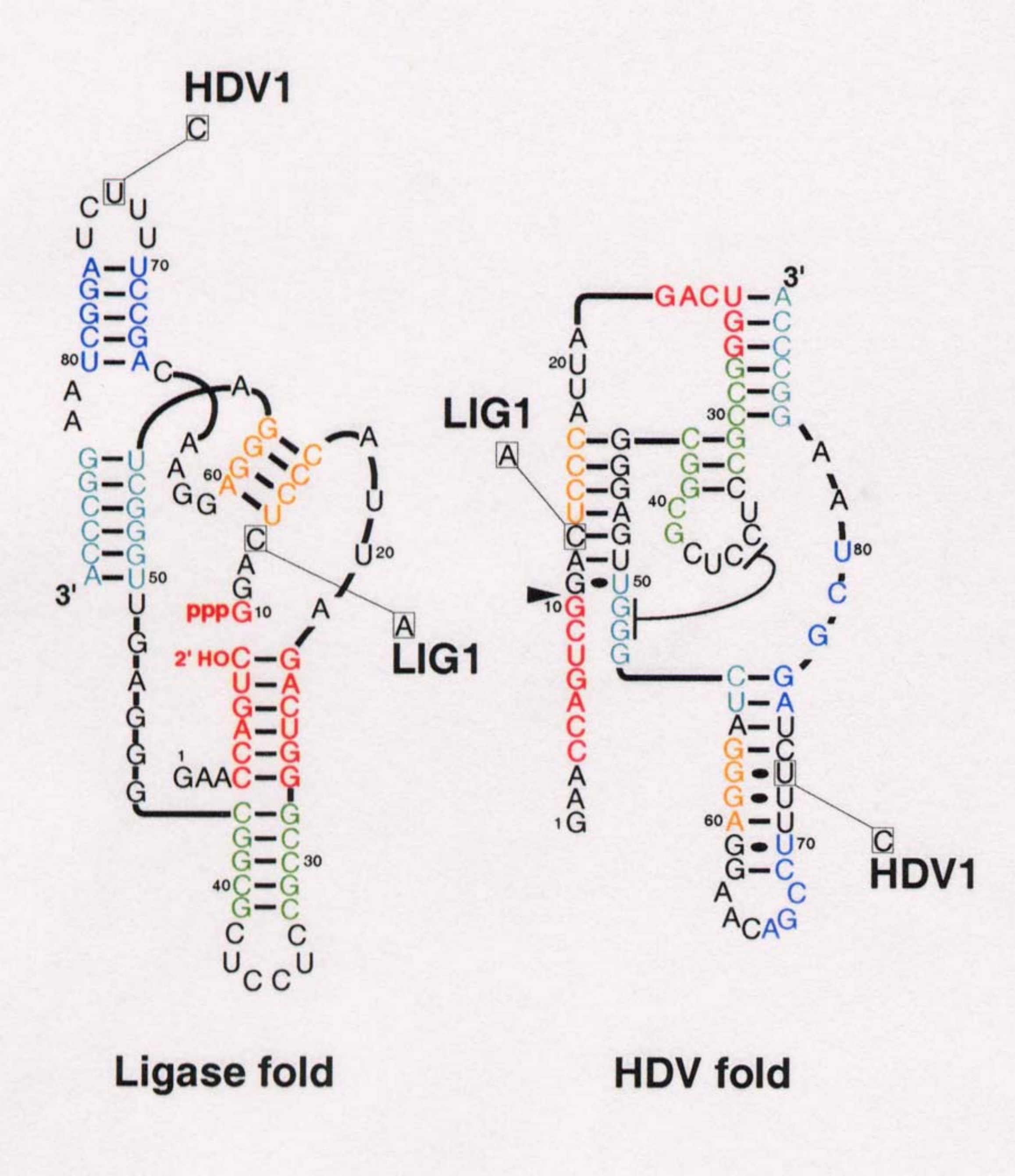
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

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

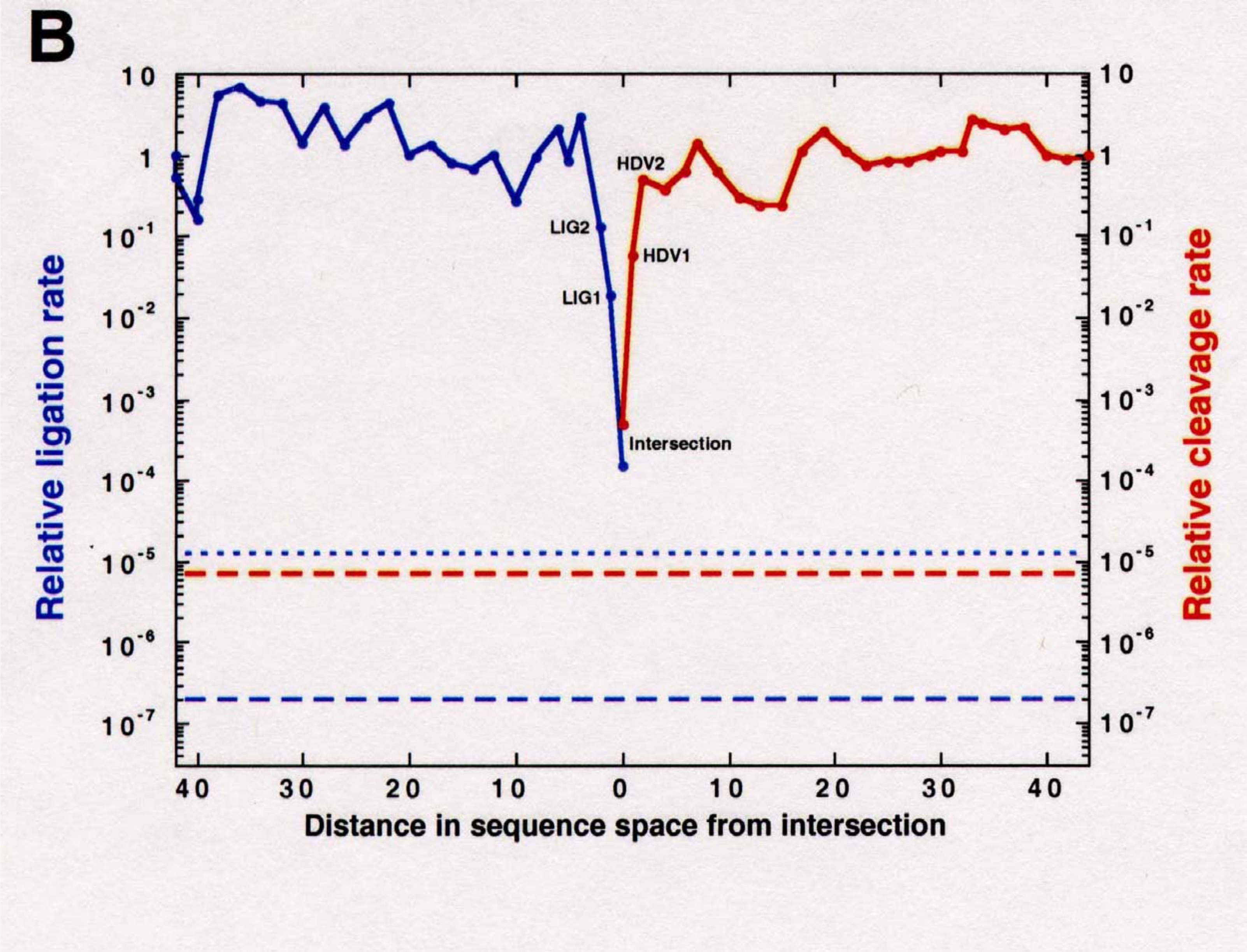


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)



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

Concluding remarks

(i) The RNA model allows for detailed insights into evolutionary optimization and experimental tests of predictions. Evolution occurs in steps: short adaptive phases are interrupted by long quasi-stationary epochs of neutral evolution.

(ii) RNA molecules share features with much more complex elements when they are subsumed in populations. The elements of a population are related by a genetic mechanism.

(iii) Creation of information and learning by trial and error occur at the level of populations although the individual elements are subjected to random processes.

(iv) In this sense the population is more than the sum of its elements. It carries a temporary memory of its past in the form of molecular species that had been selected in previous adaptive phases.

Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank

Project No. Nat-7813

European Commission: Project No. EU-980189

The Santa Fe Institute and the Universität Wien

The software for producing RNA movies was developed by Robert Giegerich and coworkers at the Universität Bielefeld

Universität Wien

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

Peter Stadler, **Bärbel Stadler,** Universität Leipzig, GE

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

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

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

Web-Page for further information:

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