RNA – From Mathematical Models to Real Molecules

3. Optimization and Evolution of RNA Molecules

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

CIMPA – Genoma School

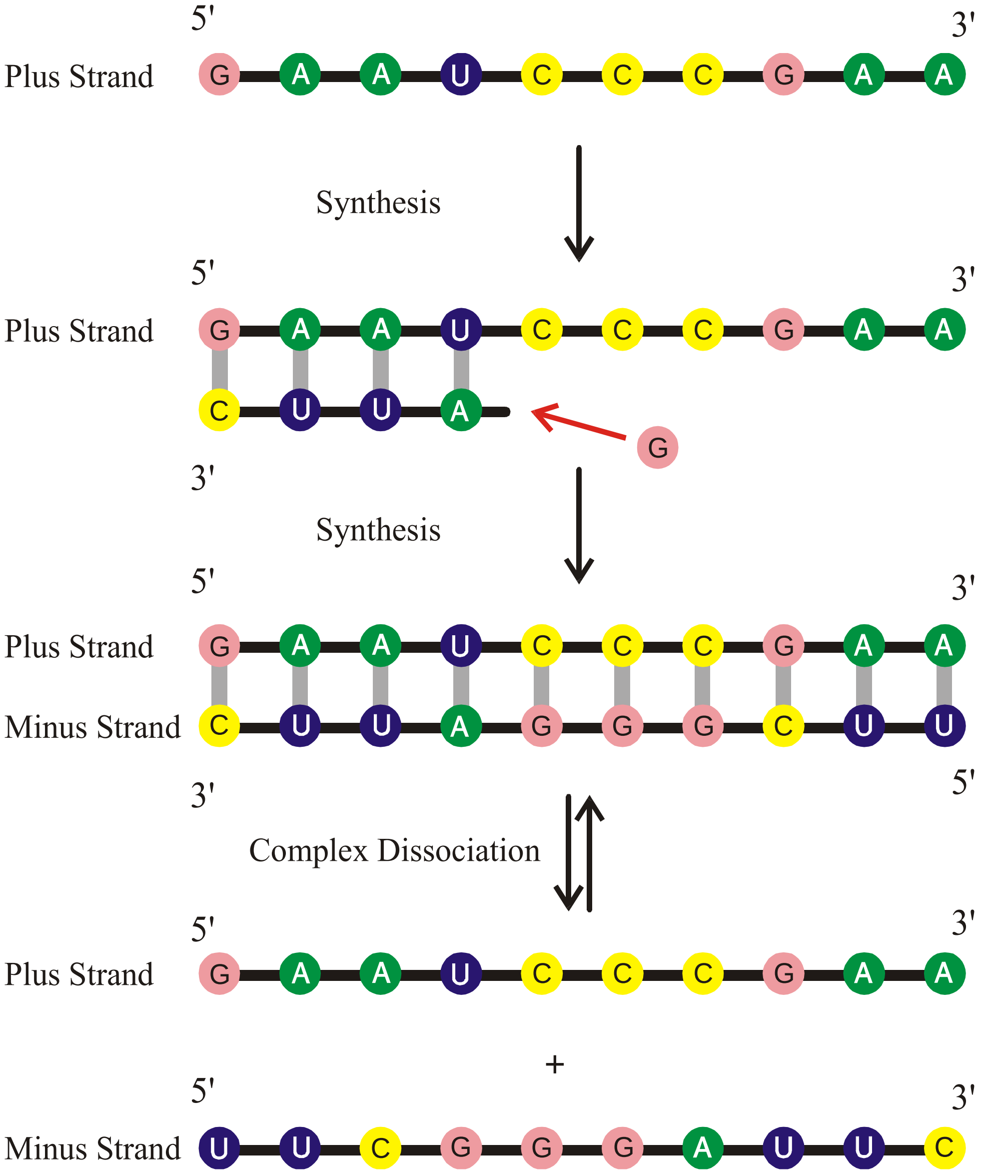
Valdivia, 12.– 16.01.2004

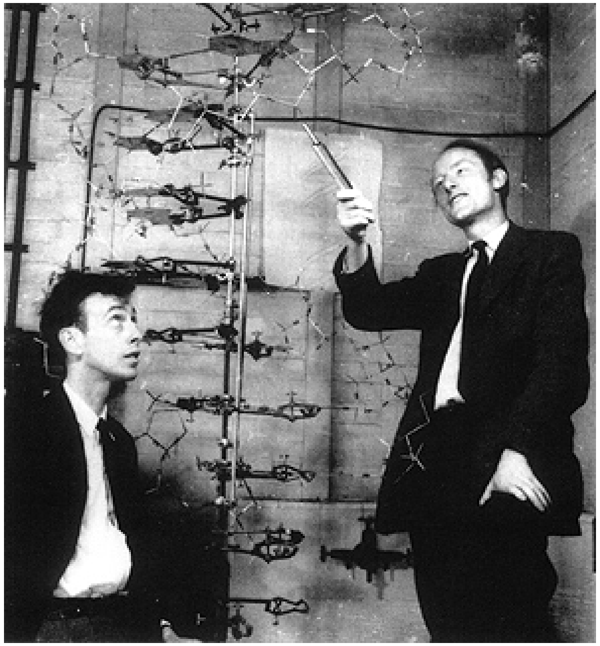
Web-Page for further information:

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

	Generation time	10 000 generations	10^6 generations	10^7 generations
RNA molecules	10 sec 1 min	27.8 h = 1.16 d 6.94 d	115.7 d 1.90 a	3.17 a 19.01 a
Bacteria	20 min 10 h	138.9 d 11.40 a	38.03 a 1 140 a	380 a 11 408 a
Higher multicelluar organisms	10 d 20 a	274 a 20 000 a	27 380 a 2×10^7 a	273 800 a 2×10^8 a

Time scales of evolutionary change

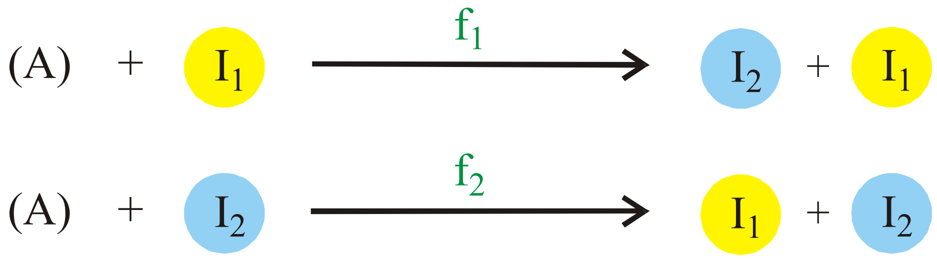




James Watson and Francis Crick, 1953

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

GÍC 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: $[\mathrm{I}_i] = x_i$ Æ $0\ ,\ f_i > 0$; i=1,2

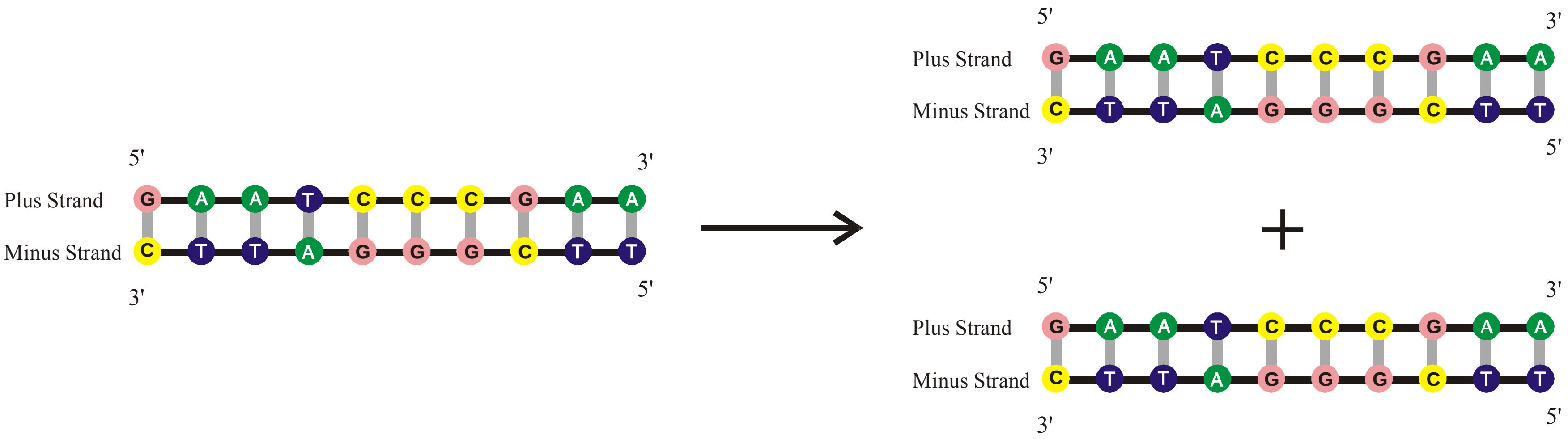
$$\frac{dx_1}{dt} = f_2 x_2 - x_1 \phi\,, \quad \frac{dx_2}{dt} = f_1 x_1 - x_2 \phi\,, \quad \phi = f_1 x_1 + f_2 x_2 = \overline{f}$$

Solutions are obtained by integrating factor transformation

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

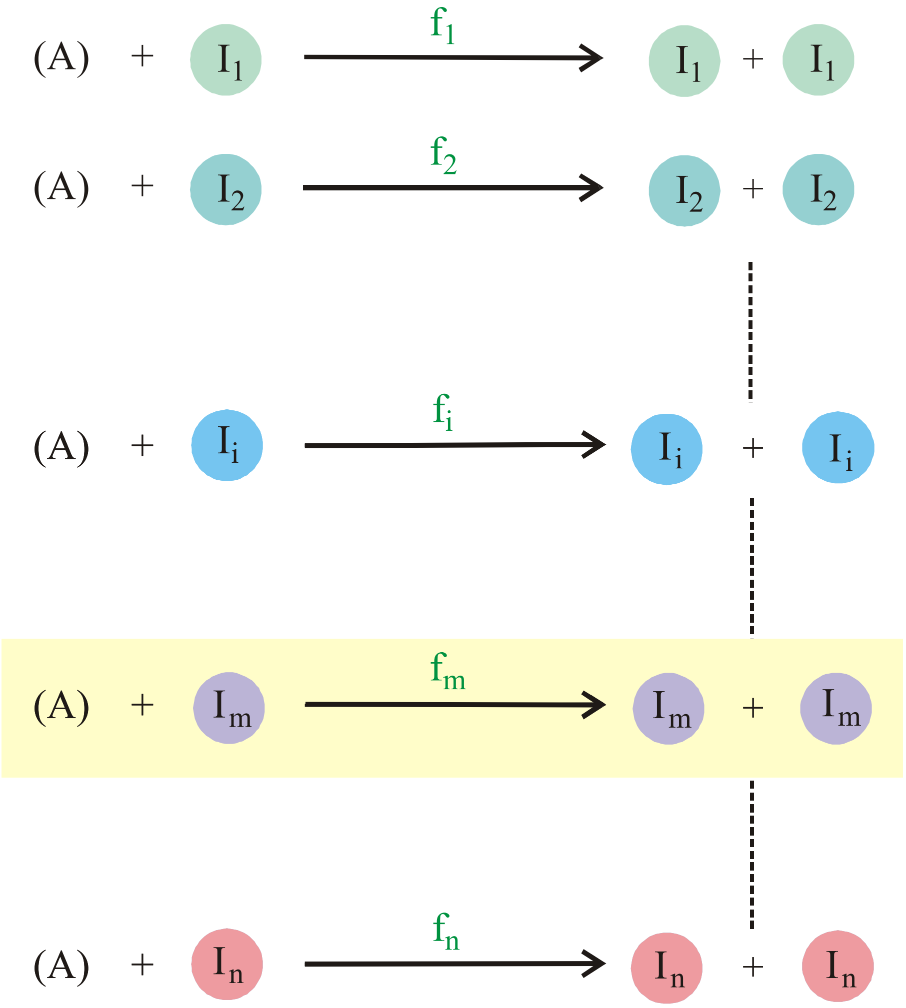
$$\gamma_1(0) = \sqrt{f_1}\,x_1(0) + \sqrt{f_2}\,x_2(0),\ \gamma_2(0) = \sqrt{f_1}\,x_1(0) - \sqrt{f_2}\,x_2(0),\ f = \sqrt{f_1 f_2}$$

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



Direct replication of DNA is a higly complex copying mechanism involving more than ten different protein molecules. Complementarity is determined by Watson-Crick base pairs:

GÍ**C** and **A**=**T**



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

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

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

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

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

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

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

$[I_i] = x_i$ 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 , $f_i > 0$

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

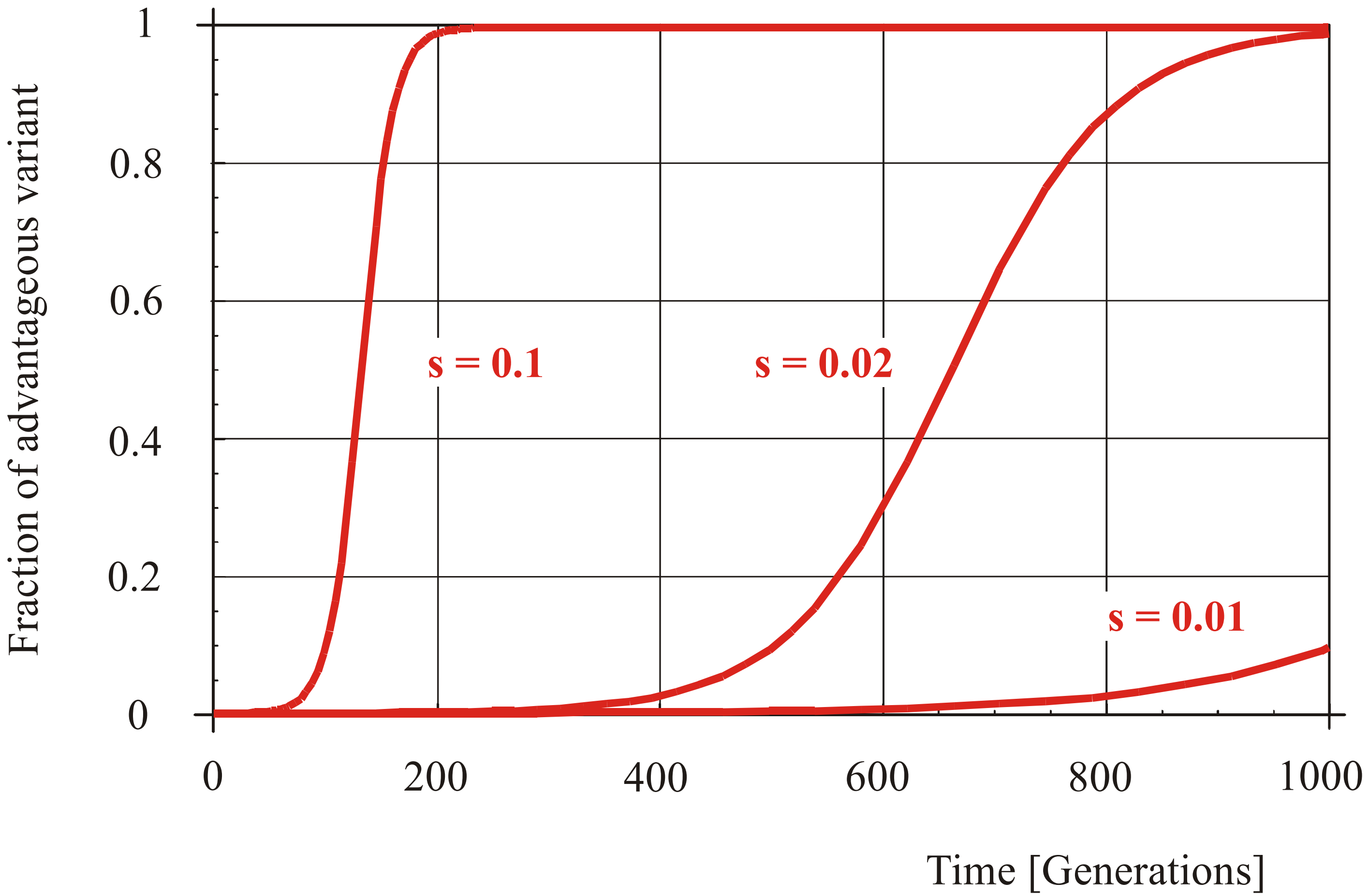
Mean fitness or dilution flux, ϕ (t), is a **non-decreasing function** of time,

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

Solutions are obtained by integrating factor transformation

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

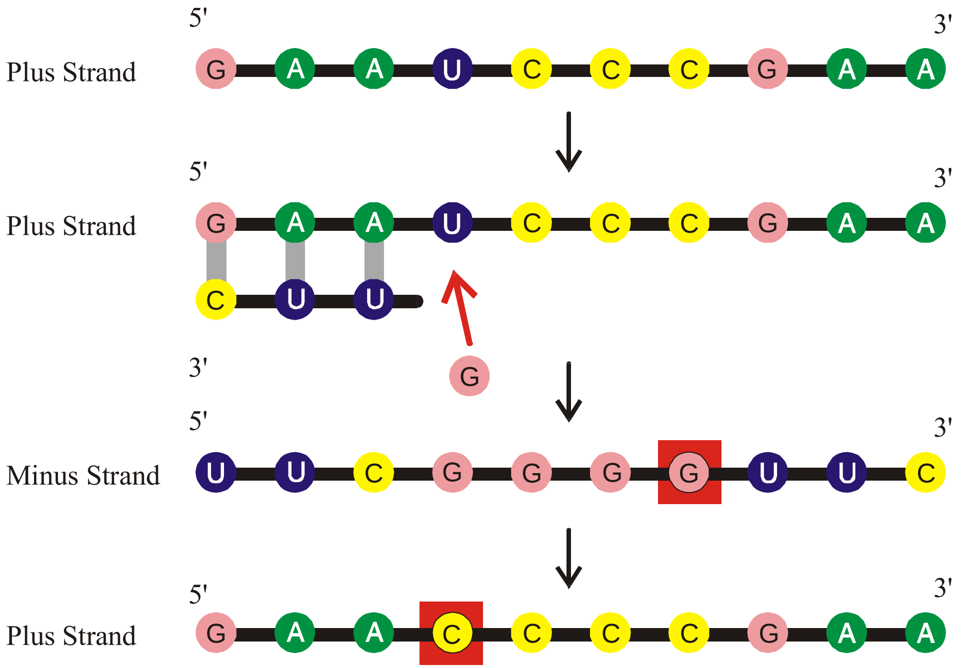
$\mathbf{s} = (f_2\text{-}f_1) / f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/N\ ;\ x_2(0) = 1/N$



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

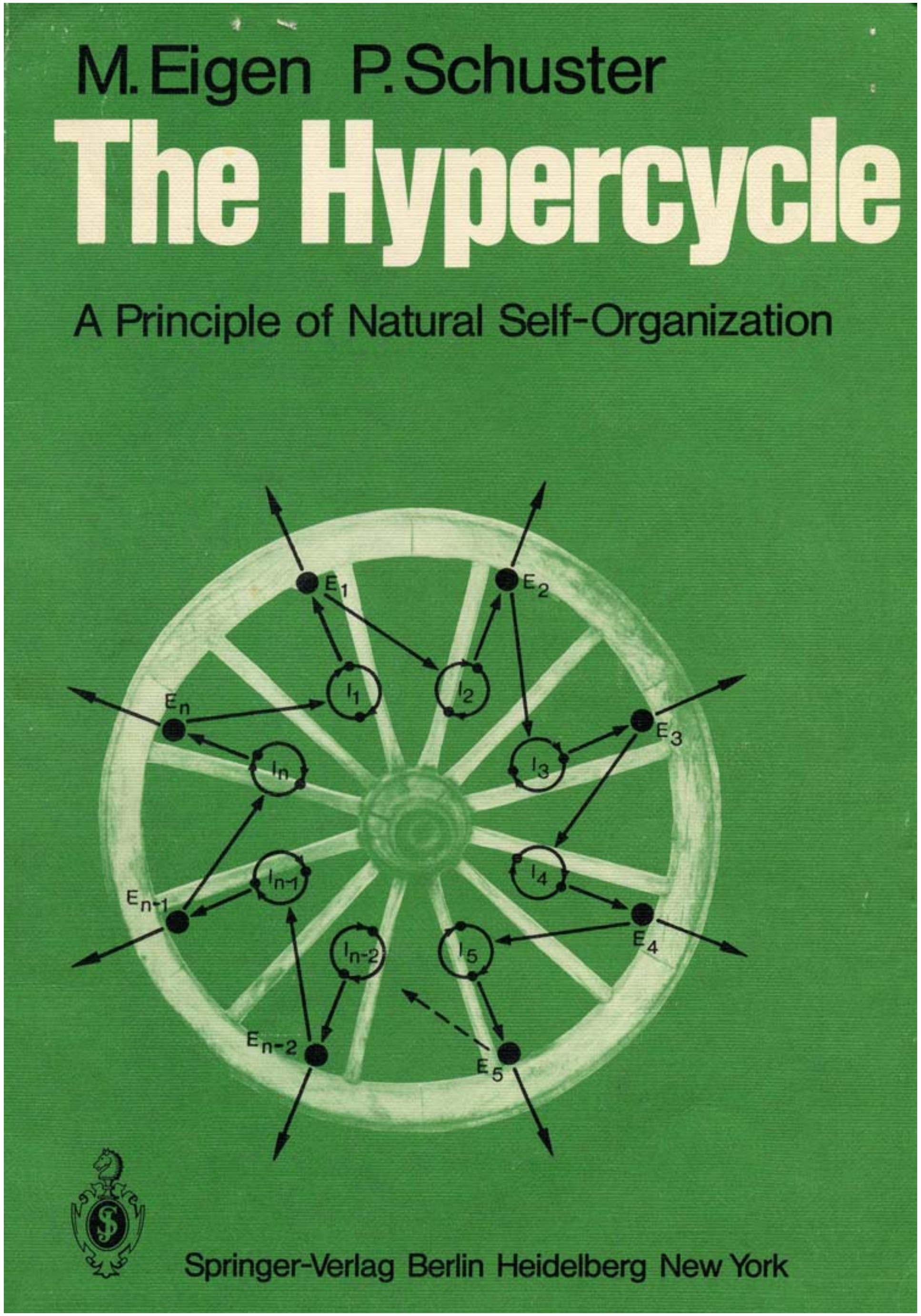
J.S. McCaskill, ***A localization threshold for macromolecular quasispecies from continuously distributed replication rates***. J.Chem.Phys. **80** (1984), 5194-5202

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

Chemical kinetics of molecular evolution

M. Eigen, P. Schuster, `The Hypercycle´, Springer-Verlag, Berlin 1979

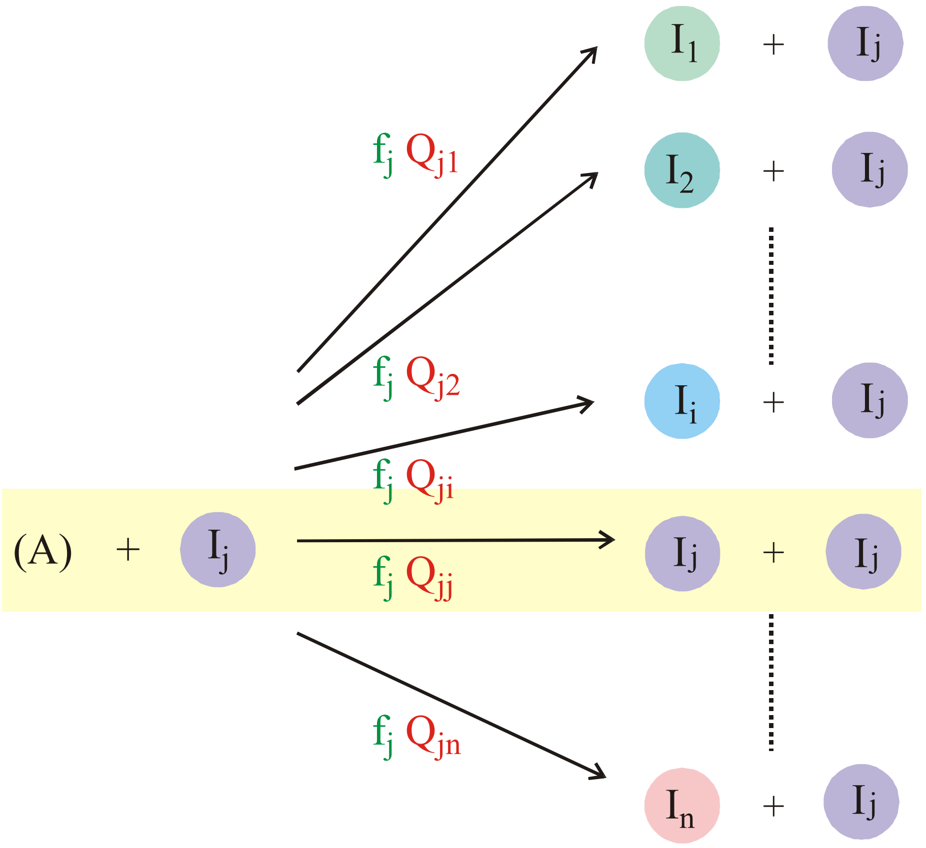


M. Eigen P. Schuster

The Hypercycle

A Principle of Natural Self-Organization

Springer-Verlag Berlin Heidelberg New York



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

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

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

$[A] = a =$ constant

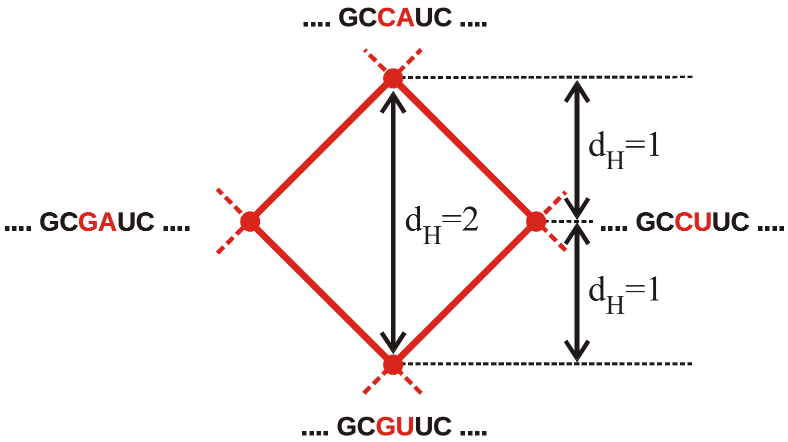
$$Q_{ij} = (1-p)^{\ell-d(i,j)}\ p^{d(i,j)}$$

p .... Error rate per digit

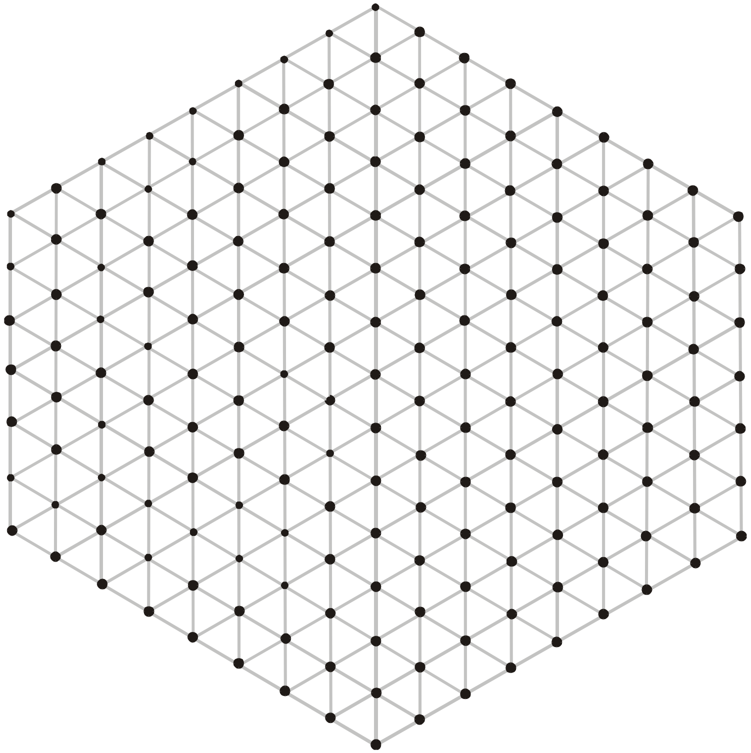
ℓ .... Chain length of the polynucleotide

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

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

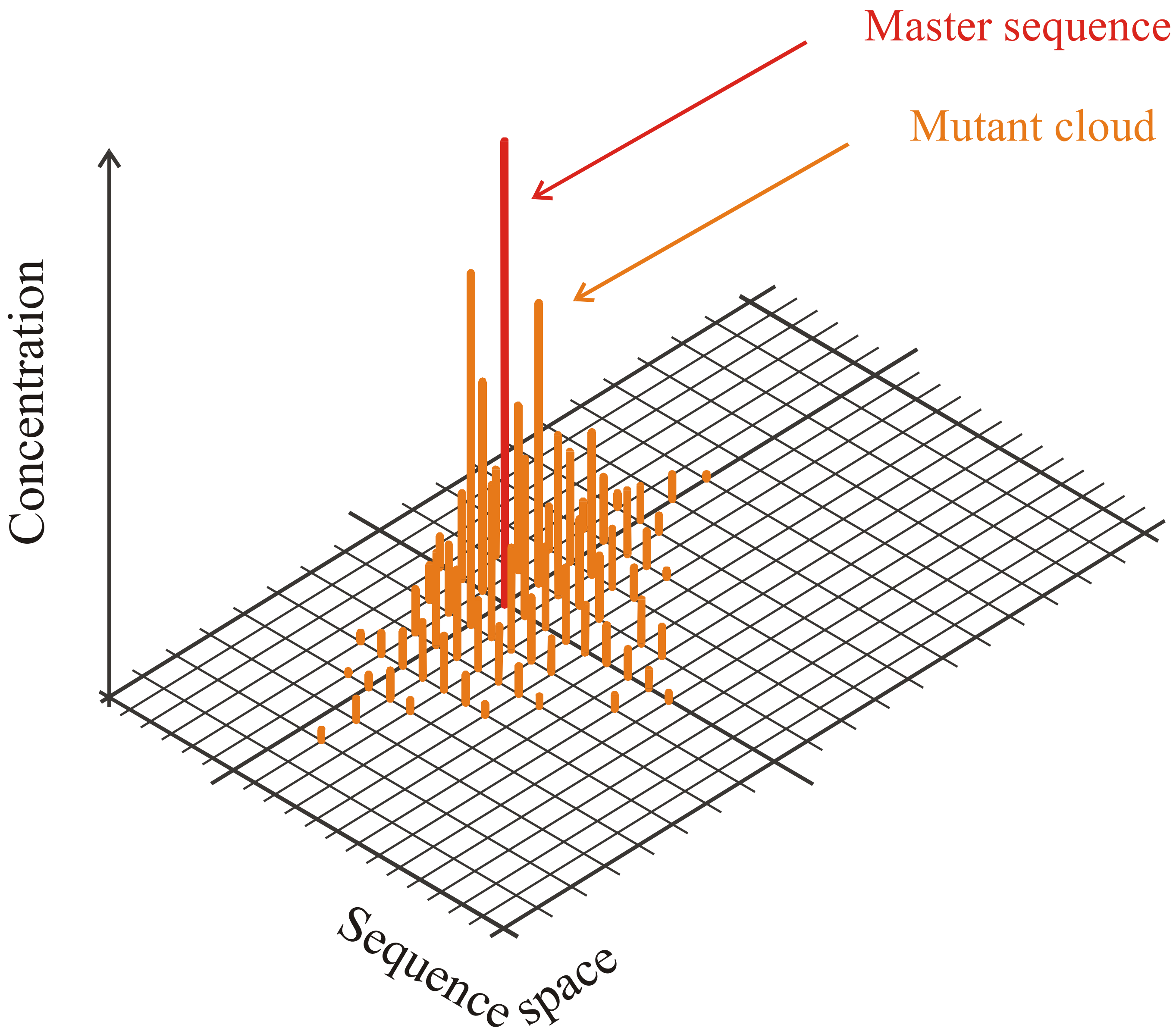
$$\frac{dx_i}{dt} = \sum\nolimits_{j=1}^{n} f_j Q_{ji}\, x_j - x_i\,\phi, \quad i = 1,2,\cdots,n; \quad \sum\nolimits_{i=1}^{n} x_i = 1; \quad \phi = \sum\nolimits_{j=1}^{n} f_j x_j = \overline{f}$$

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

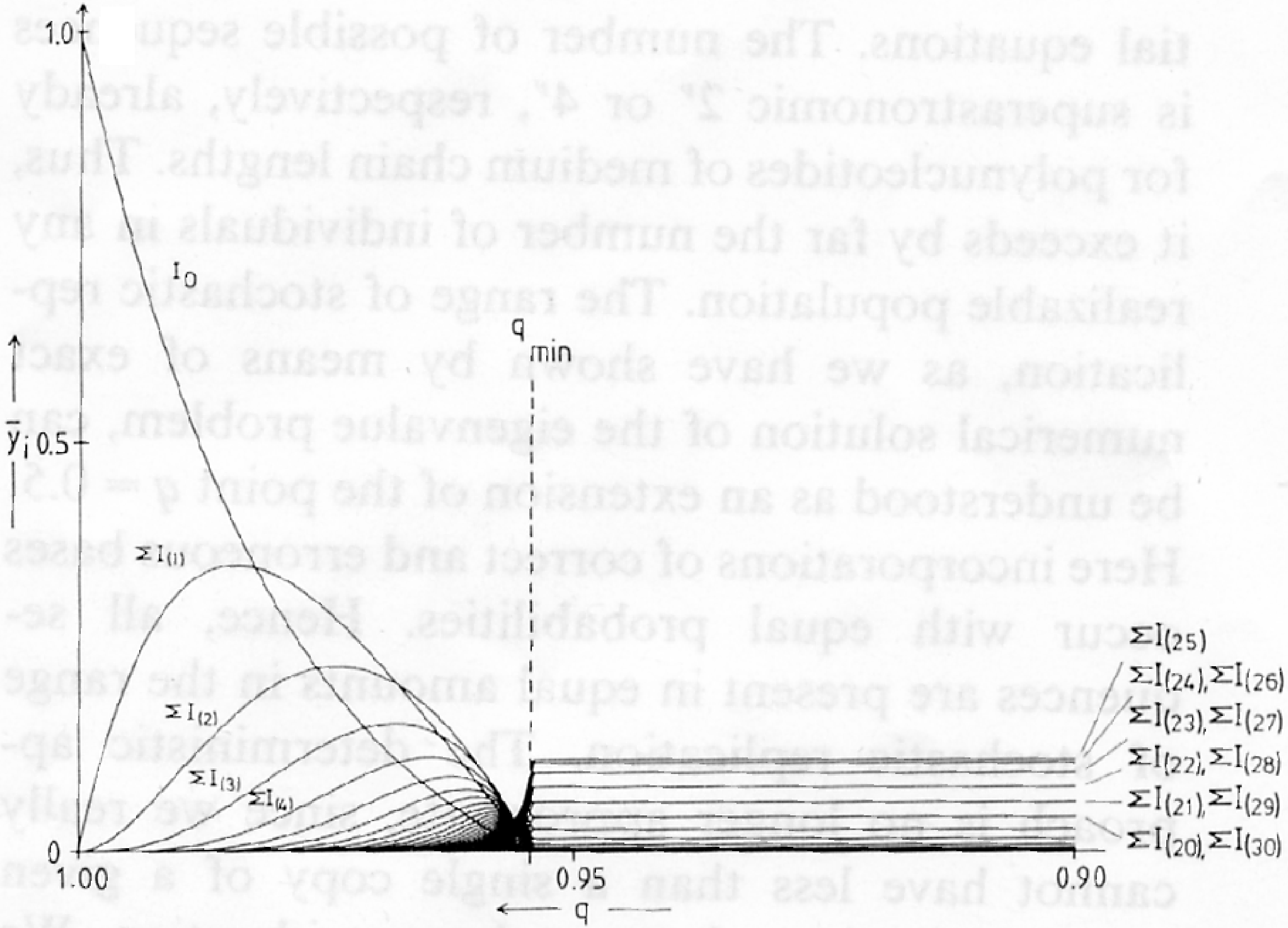
$$x_i(t) = \frac{\sum_{k=0}^{n-1} \ell_{ik} \cdot c_k(0) \cdot \exp(\lambda_k t)}{\sum_{j=1}^{n} \sum_{k=0}^{n-1} \ell_{jk} \cdot c_k(0) \cdot \exp(\lambda_k t)}; \quad i = 1,2,\cdots,n; \quad c_k(0) = \sum\nolimits_{i=1}^{n} h_{ki}\, x_i(0)$$

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

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



The molecular quasispecies in sequence space



Quasispecies as a function of the replication accuracy q

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.

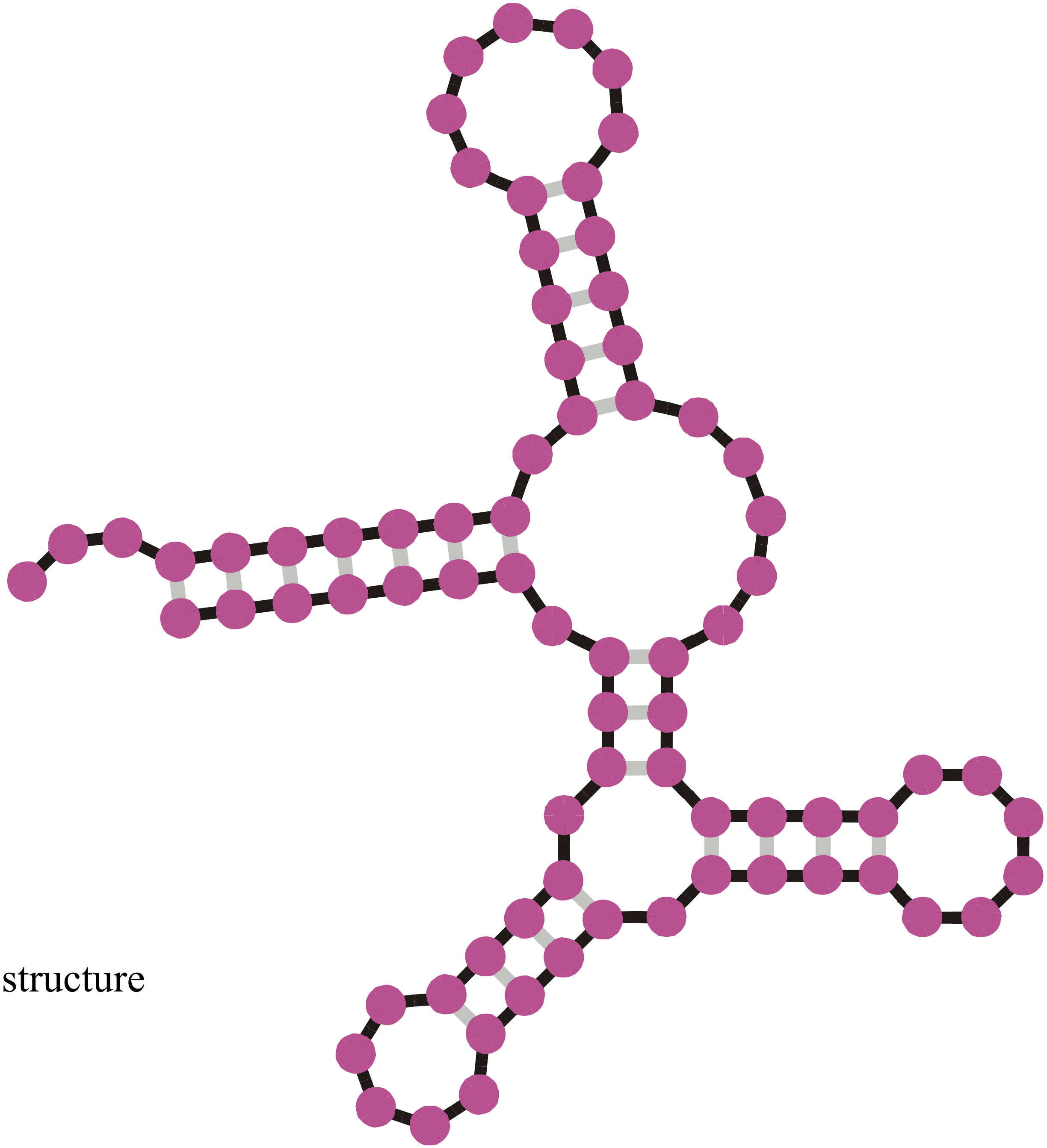
The mapping from RNA sequences into secondary structures and function,

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

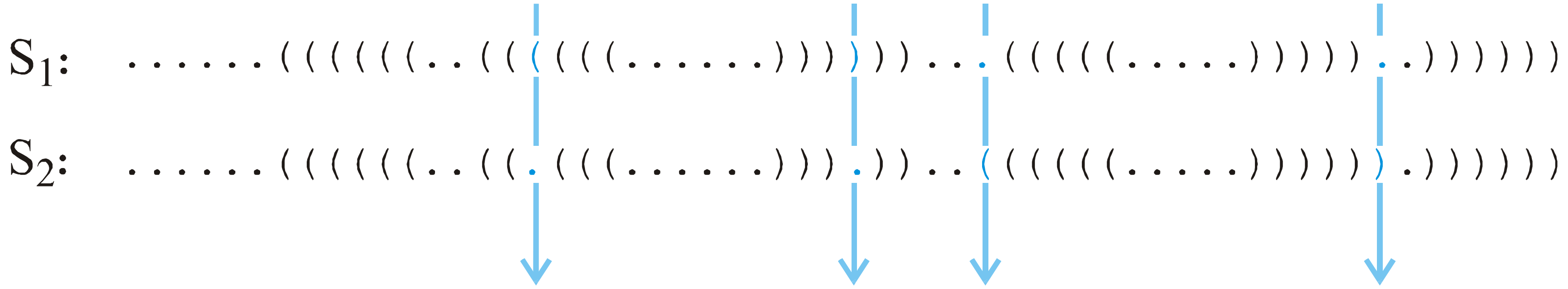
is used as a model for the complex relations between genotypes and phenotypes. Fertile progeny measured in terms of **fitness** in population biology is determined quantitatively by **replication rate constants** of RNA molecules.

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



Optimized element: RNA structure



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

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

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

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

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

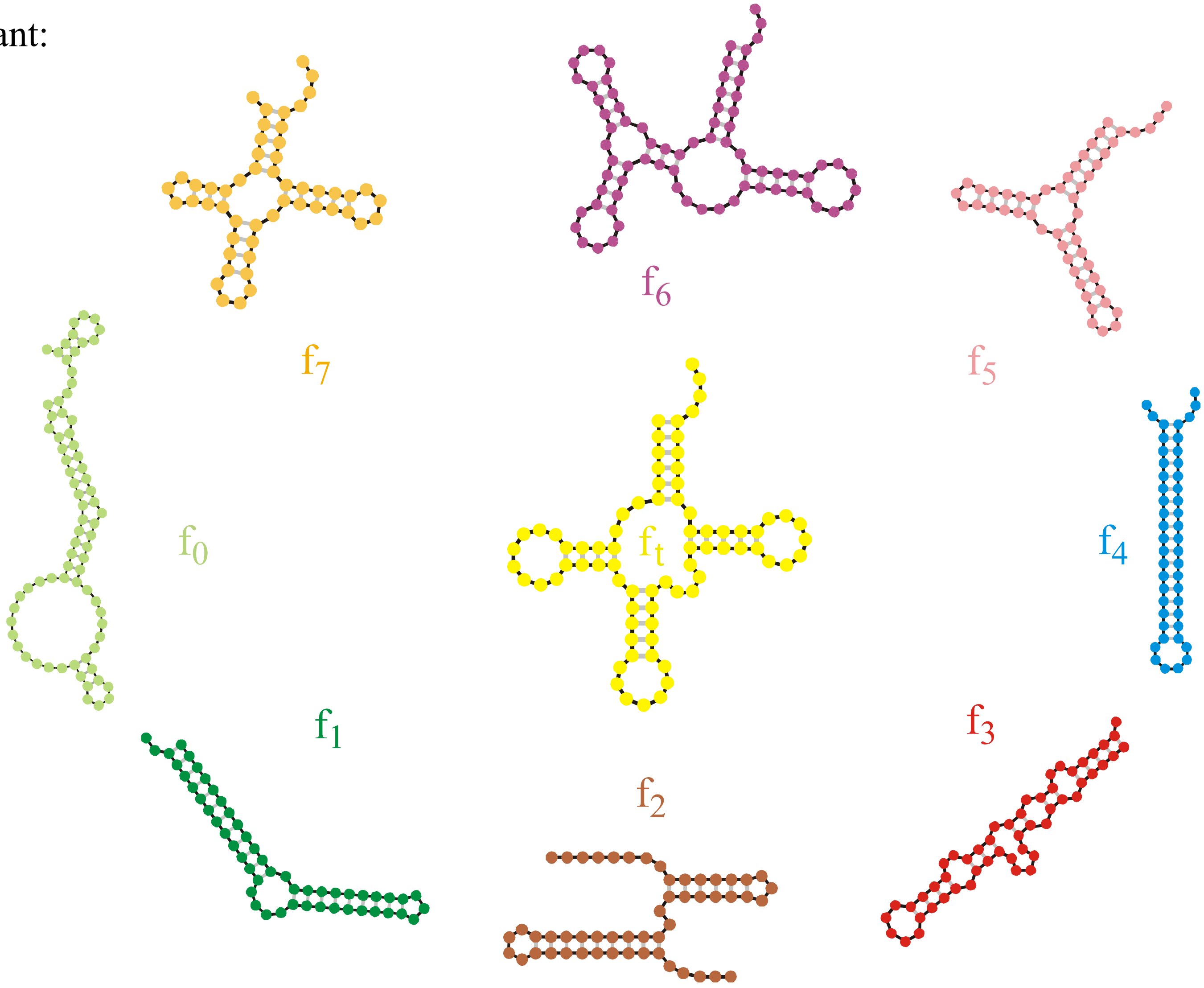
(iii) $d_H(S_1,S_3)$ ¶ $d_H(S_1,S_2) + d_H(S_2,S_3)$

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

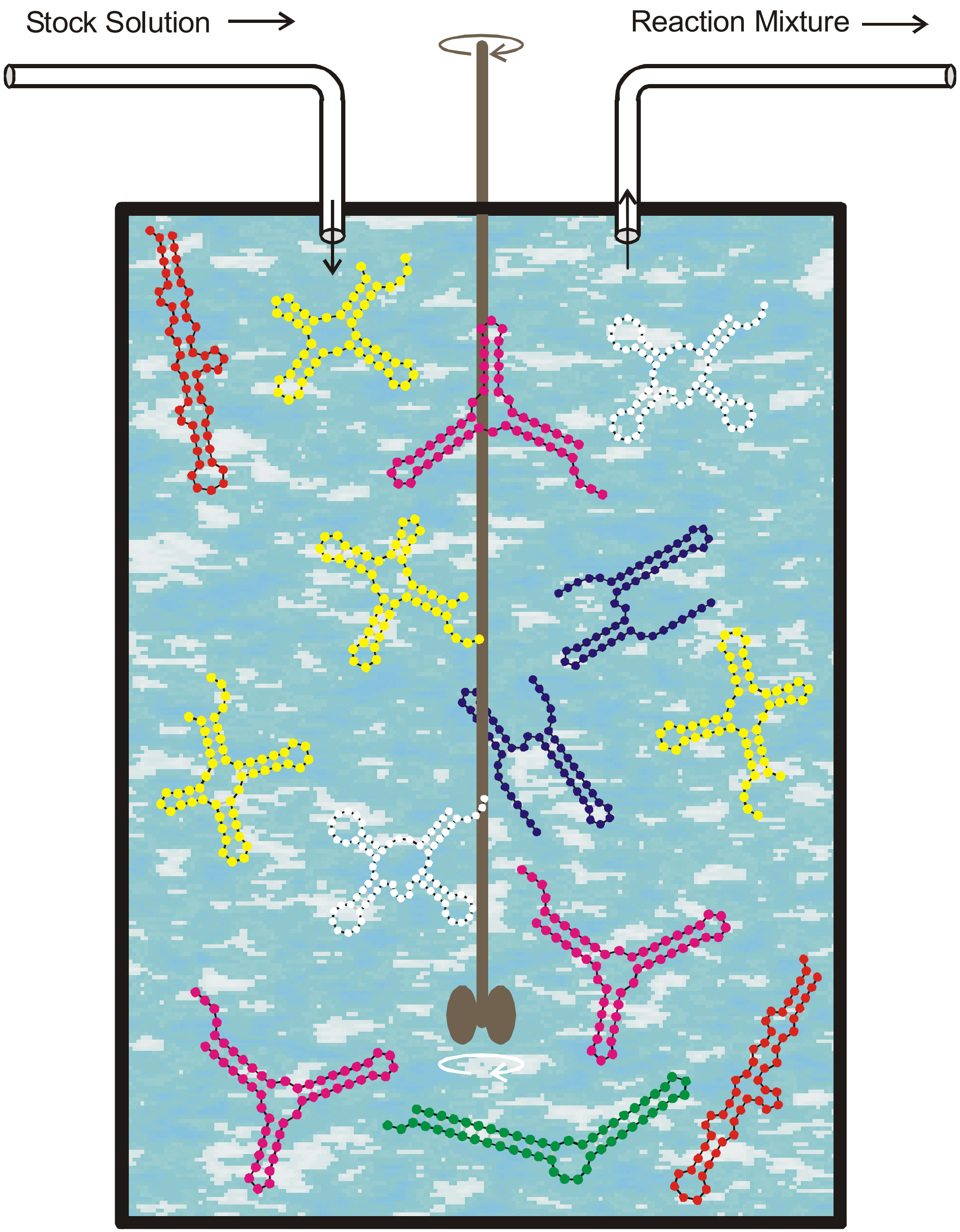
Replication rate constant:

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

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



Evaluation of RNA secondary structures yields replication rate constants



Replication rate constant:

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

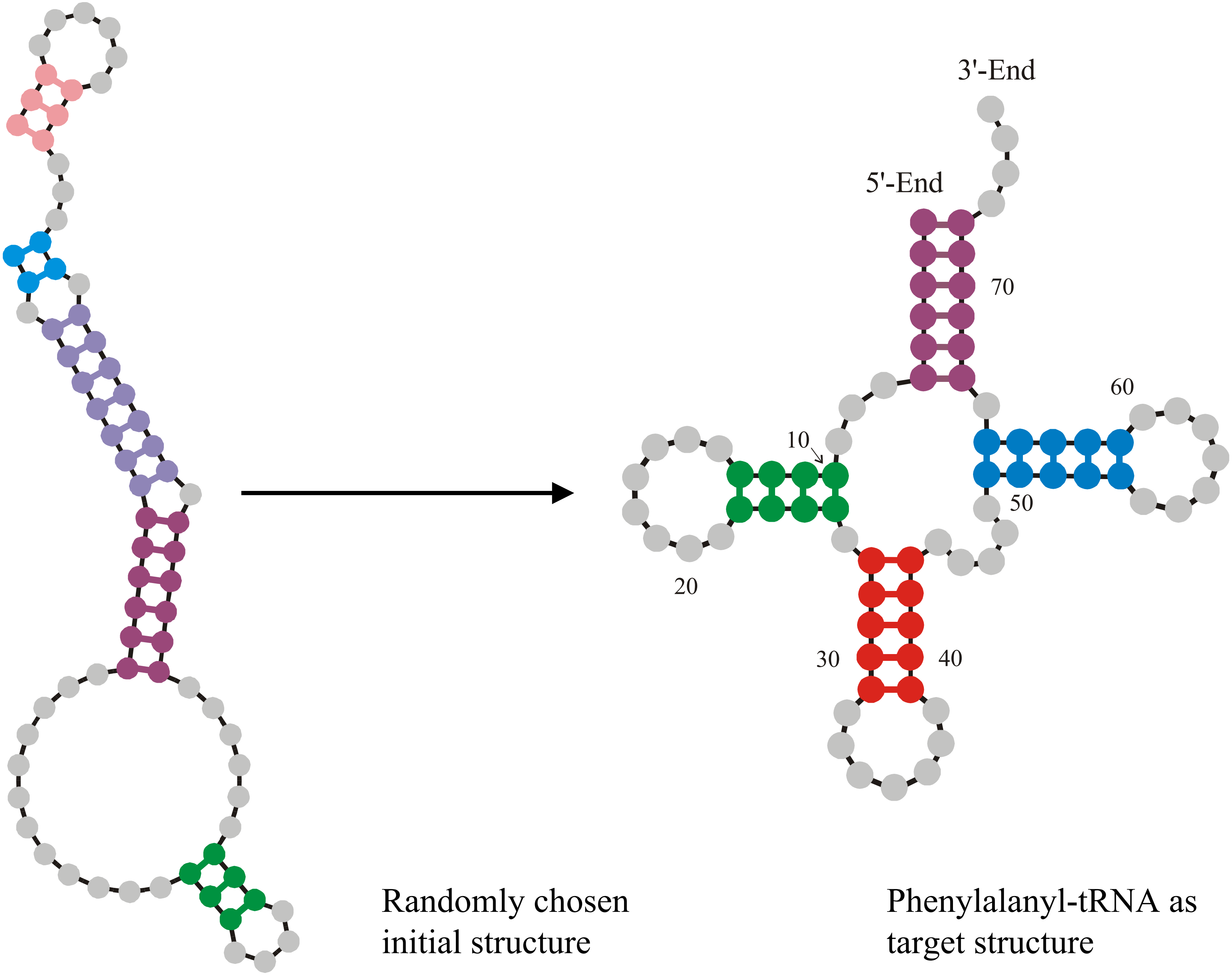
$$Dd_S^{(k)} = d_H(S_k, S_t)$$

Selection constraint:

RNA molecules is controlled by the flow

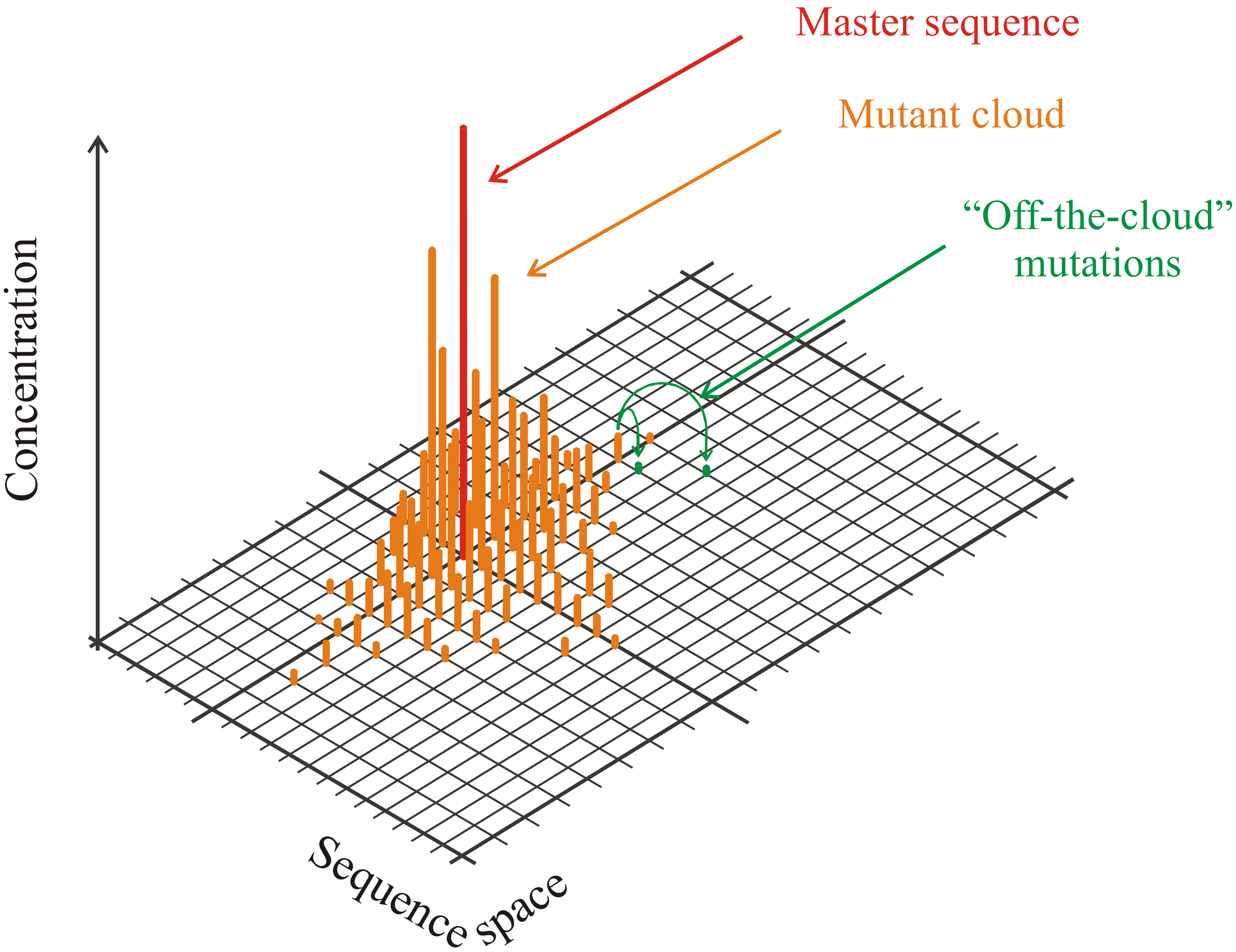
$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

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

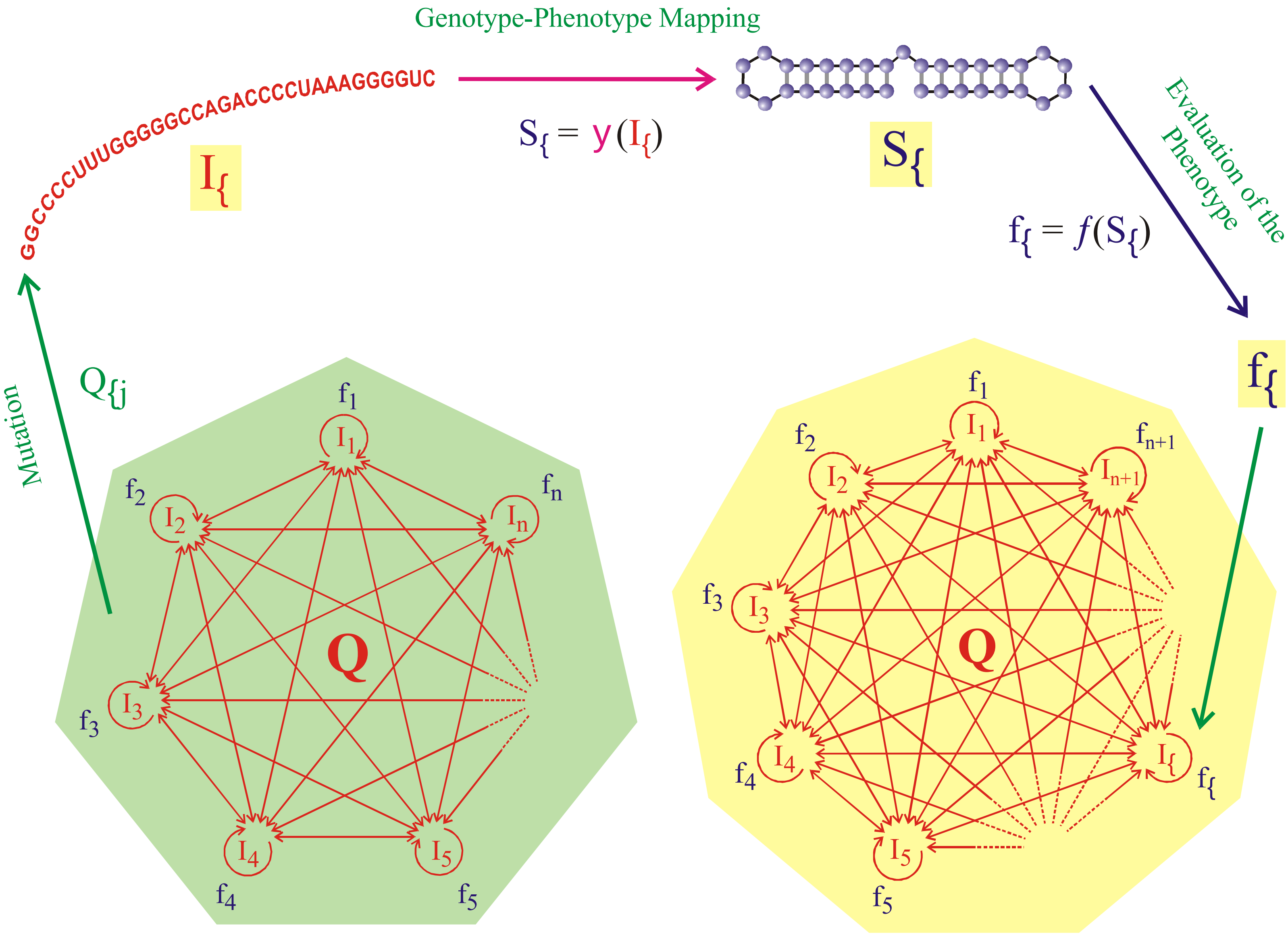


Randomly chosen initial structure

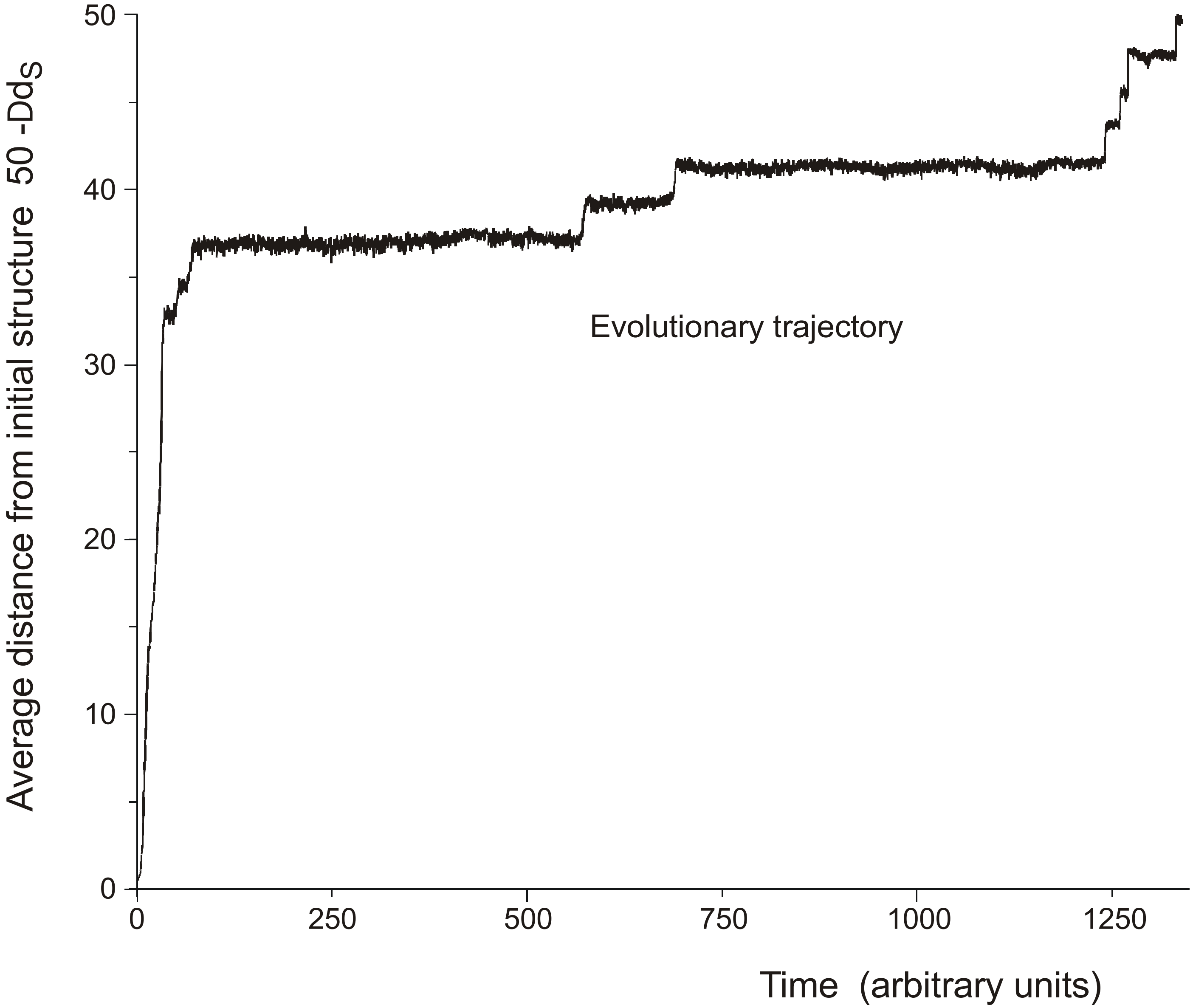
Phenylalanyl-tRNA as target structure



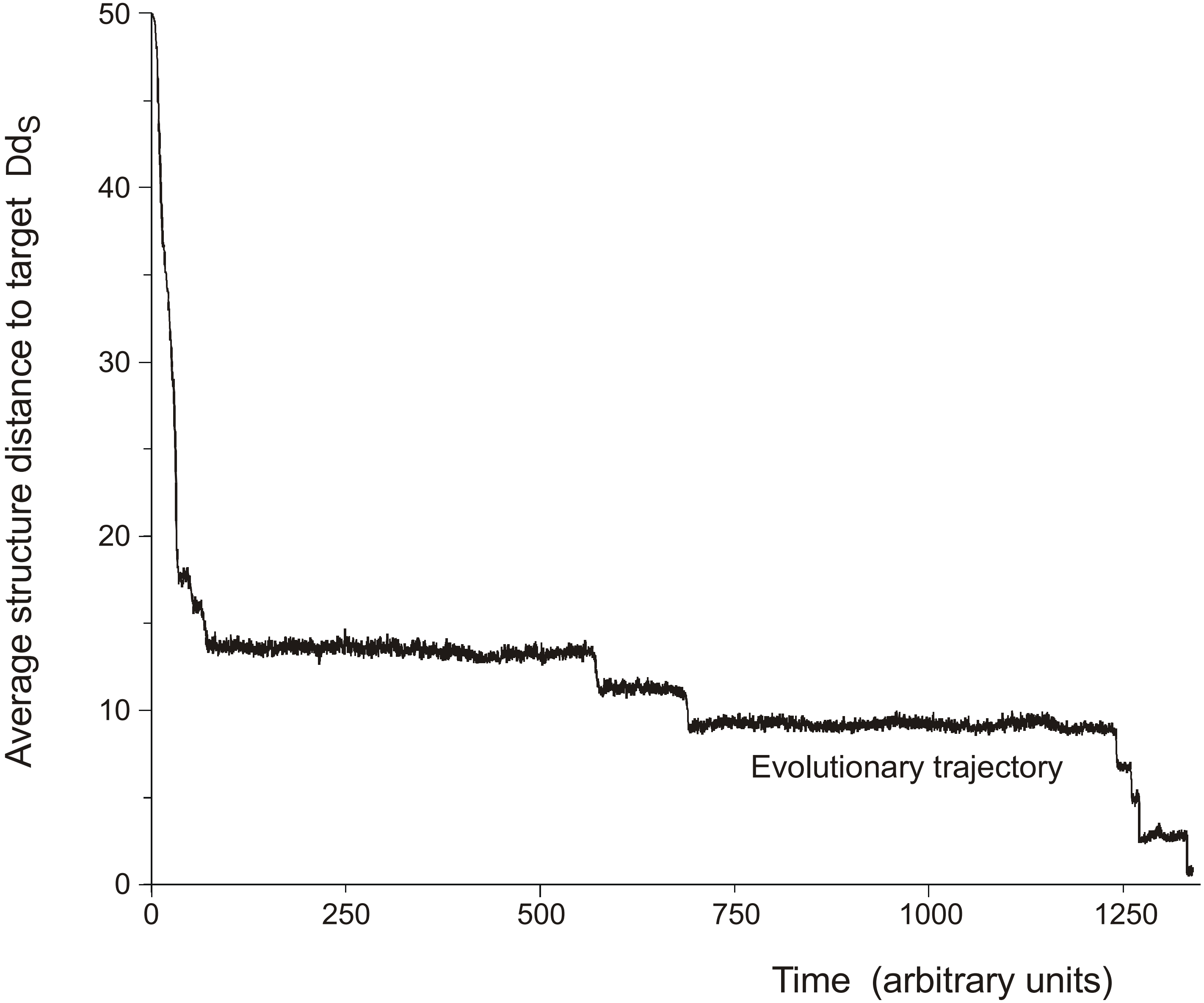
The molecular quasispecies
in sequence space



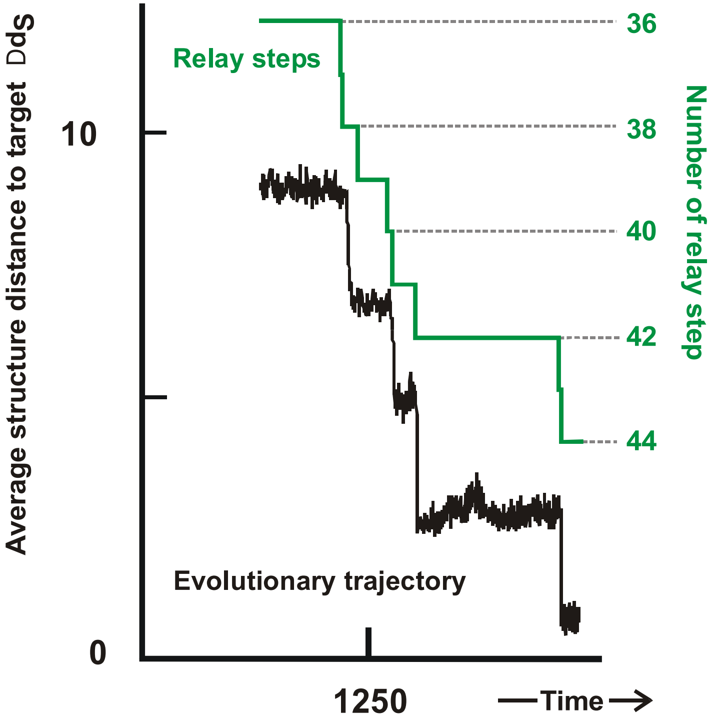
Evolutionary dynamics
including molecular phenotypes

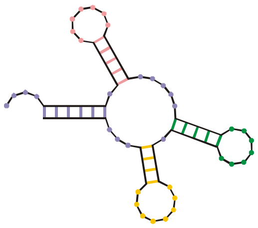


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)



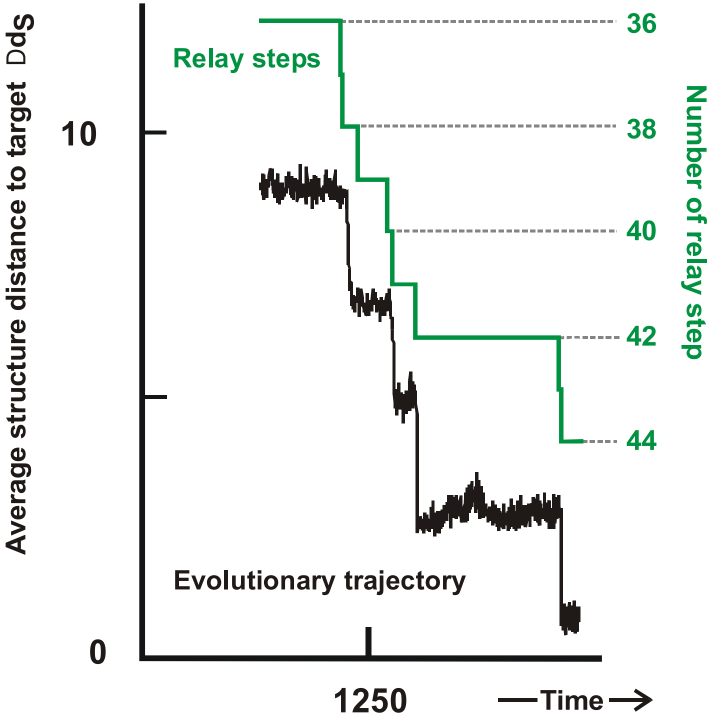
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)

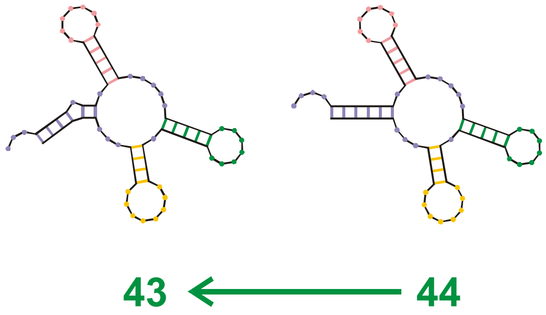




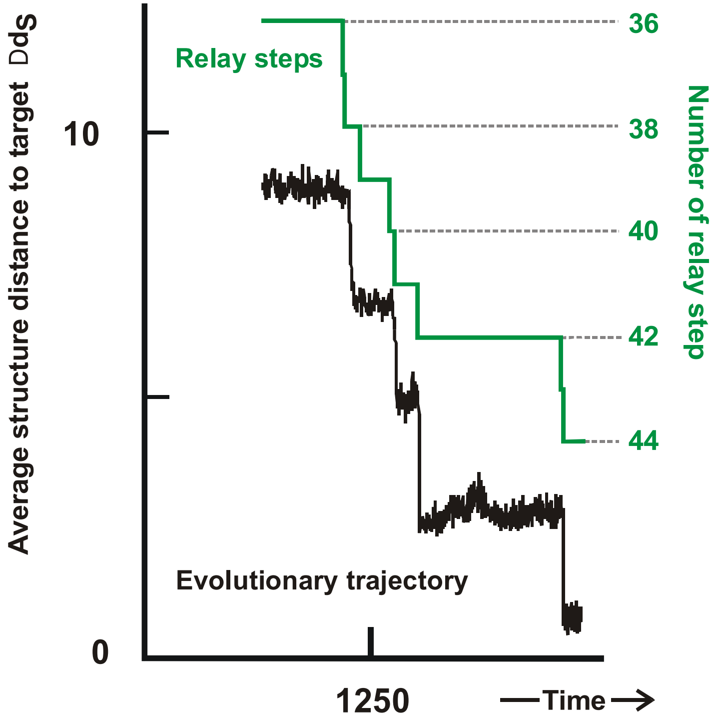
44

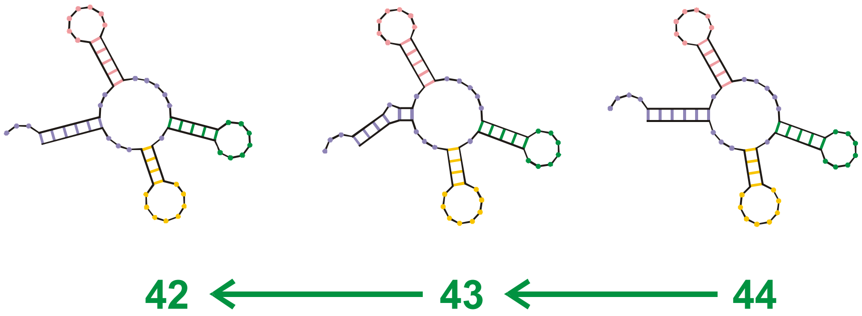
Endconformation of optimization



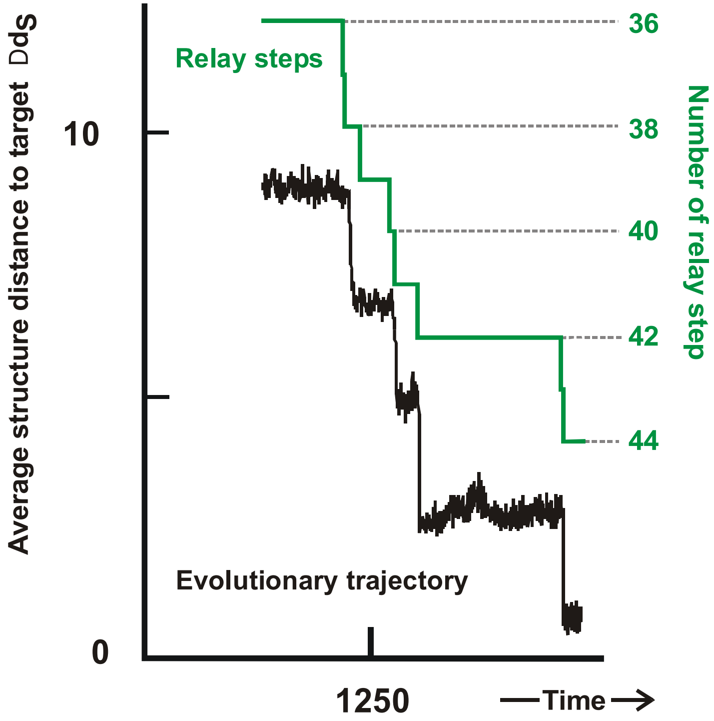


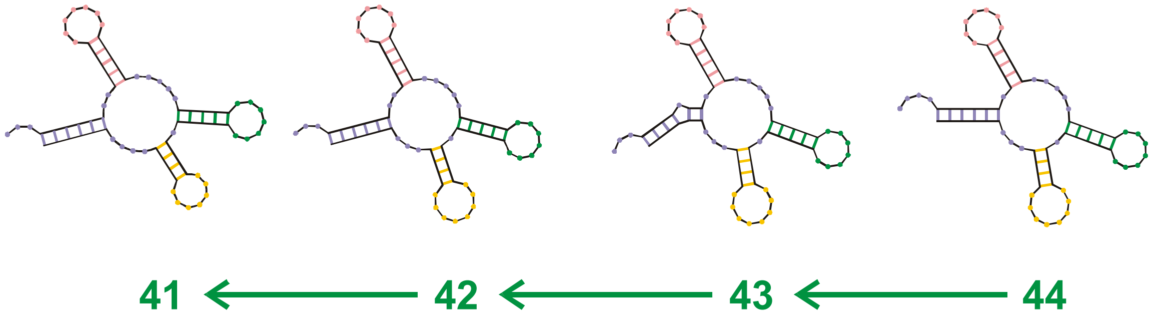
Reconstruction of the last step 43 Á 44



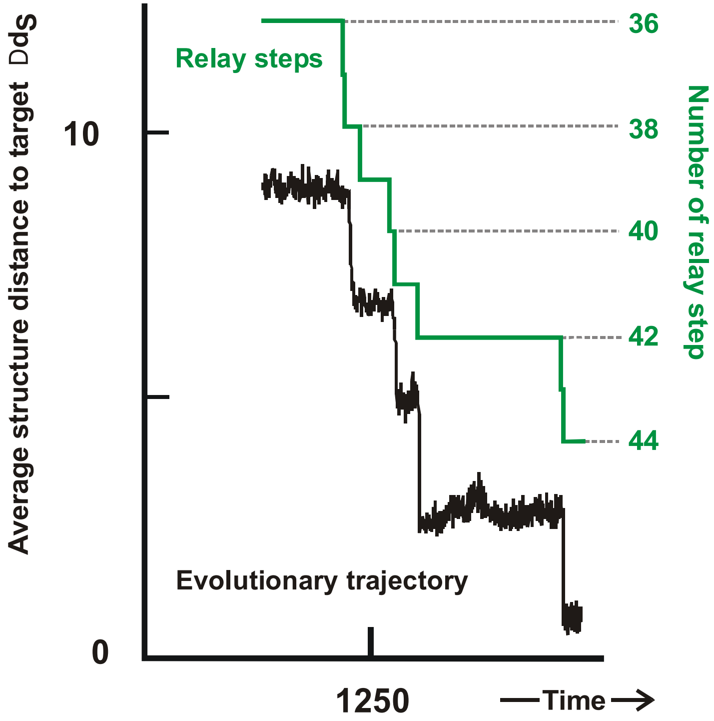


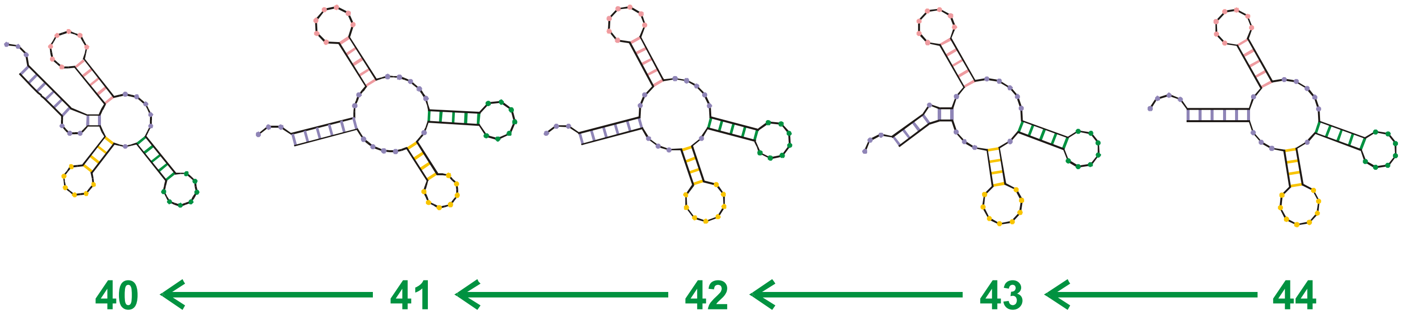
Reconstruction of last-but-one step 42 Á 43 (Á 44)



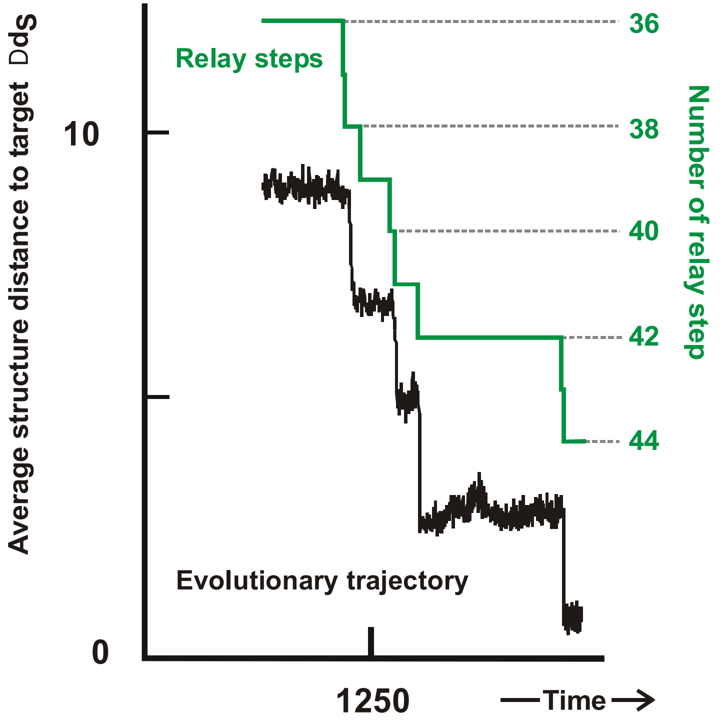


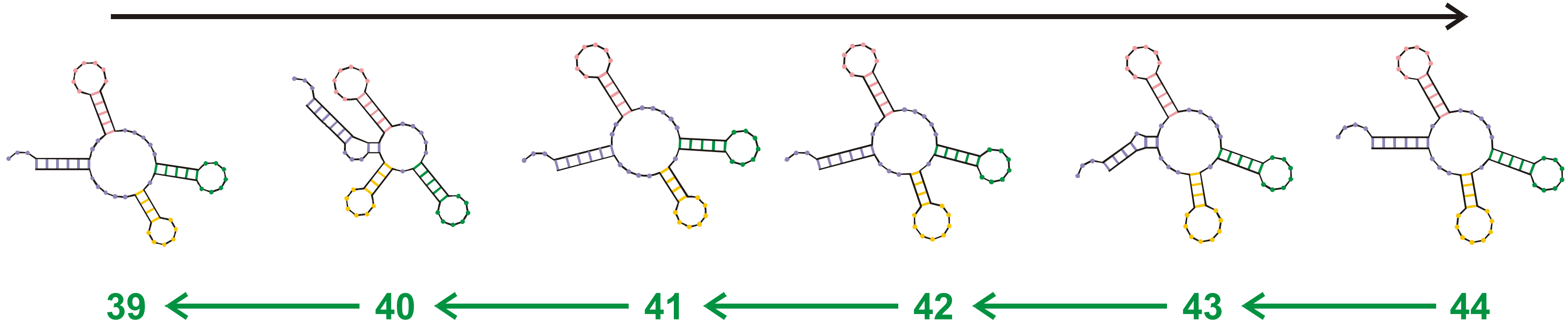
Reconstruction of step 41 Á 42 (Á 43 Á 44)





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



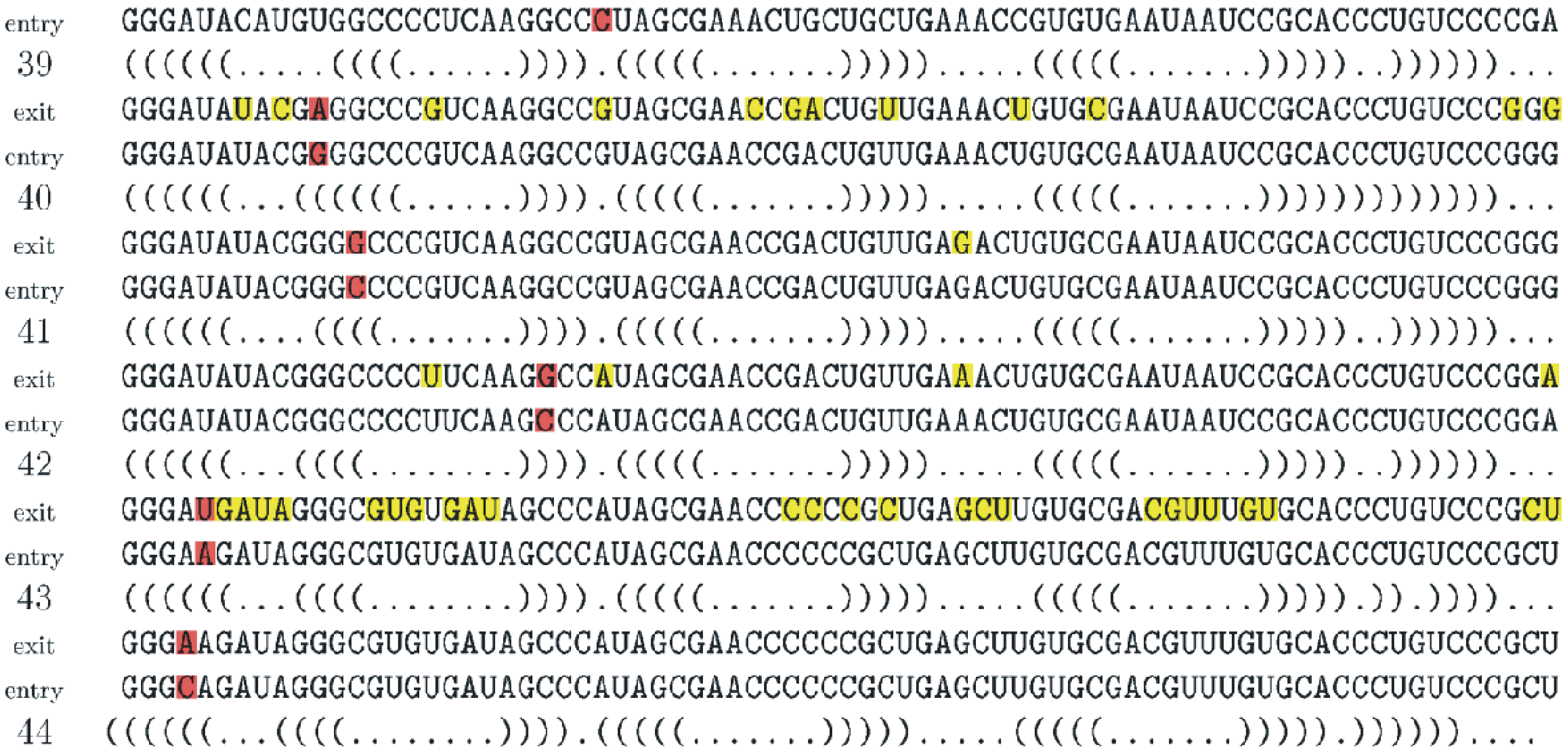


Evolutionary process

39 ← 40 ← 41 ← 42 ← 43 ← 44

Reconstruction

Reconstruction of the relay series

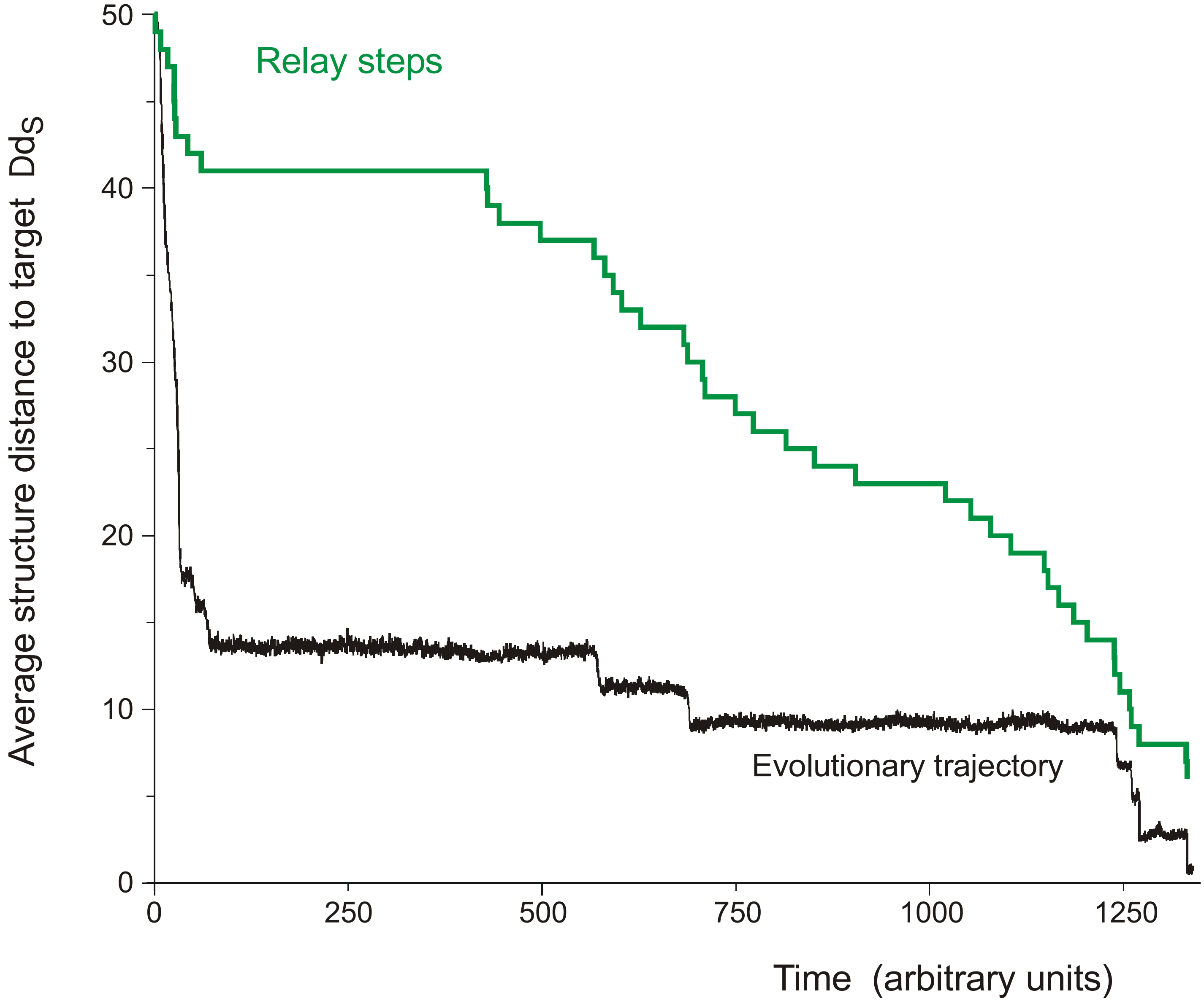


entry	GGGAUACAUGUGGCCCCUCAAGGCCCUAGCGAAACUGCUGCUGAAACCGUGUGAAUAAUCCGCACCCUGUCCCCGA
39	((((((.....((((......)))).(((((.......))))).....(((((.......)))))..))))))...
exit	GGGAUAUACGAGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
entry	GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG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

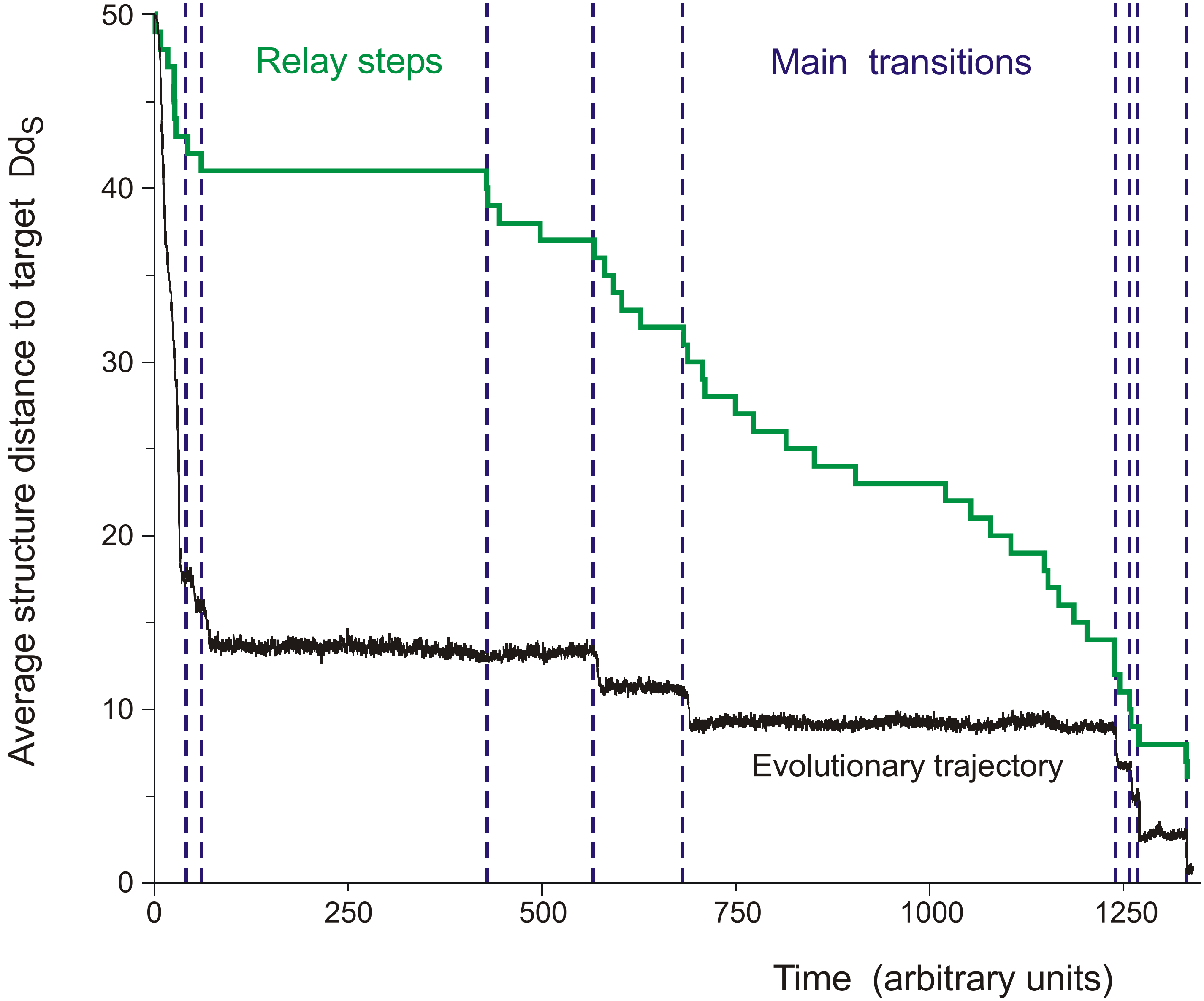
Transition inducing point mutations

Neutral point mutations

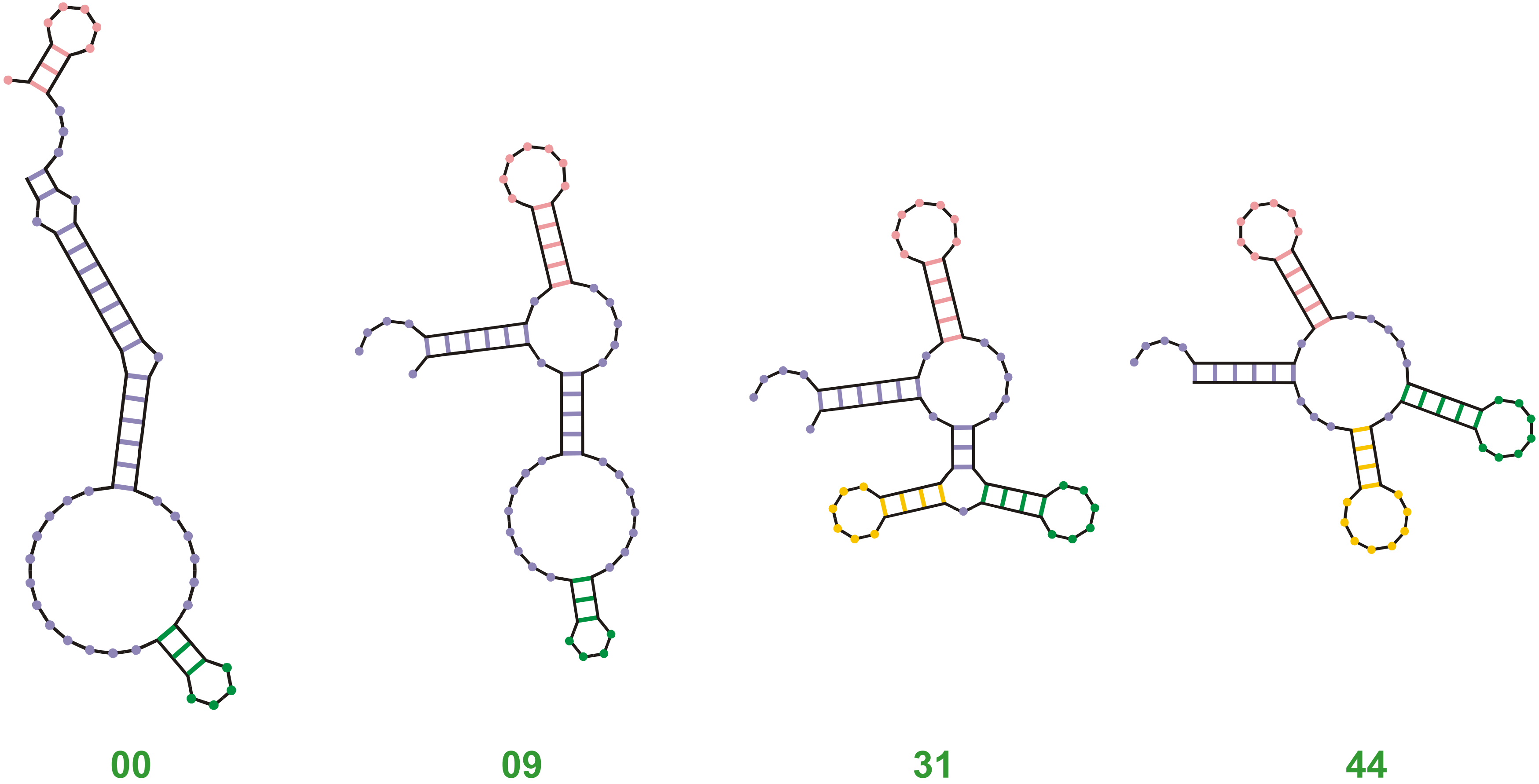
Change in RNA sequences during the final five relay steps 39 Á 44



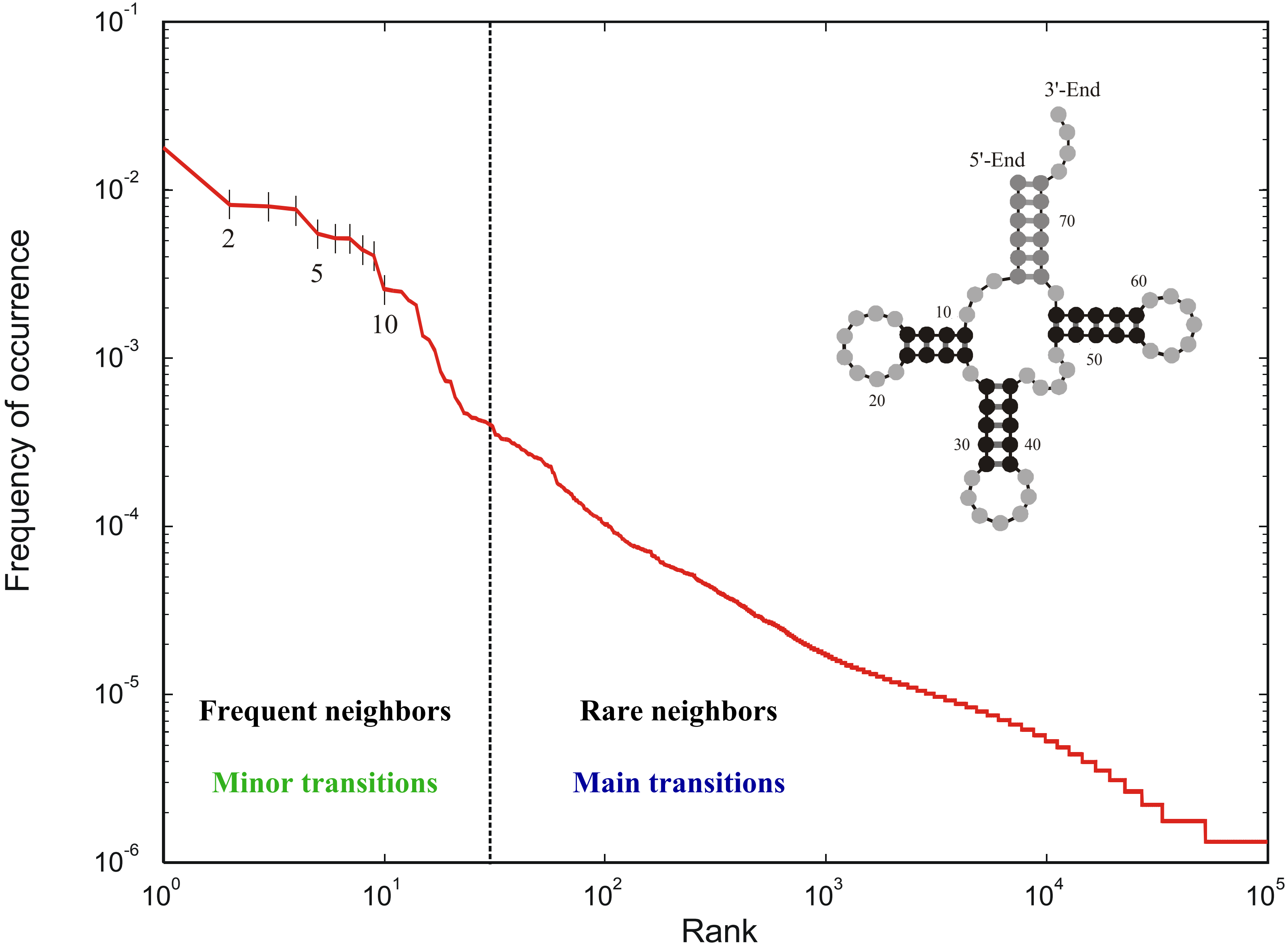
In silico optimization in the flow reactor: Trajectory and relay steps



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**.



Probability of occurrence of different structures in the mutational neighborhood of tRNA^{phe}

Definition of an e-neighborhood of structure S_k

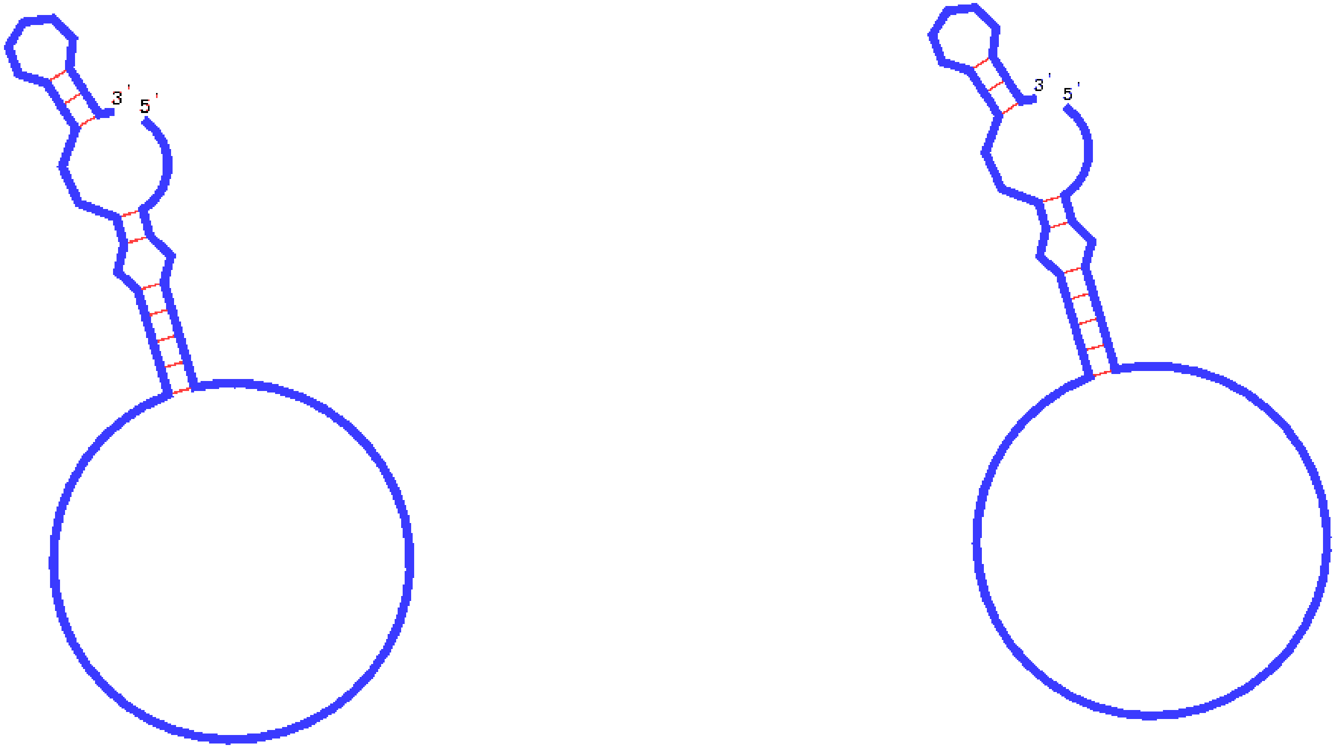
$Y(S_k)$ ... set of all structures occurring in the Hamming distance one neighborhood of the neutral network G_k of S_k

g_{jk} ... number of contacts between the two neutral networks G_j and G_k

$$g_{jk} = g_{kj}$$

Probability of occurrence: $\rho(S_j;S_k) = \dfrac{\gamma_{jk}}{n(\kappa-1)\left|G_k\right|}$; $\rho(S_k;S_j) \neq \rho(S_j;S_k)$

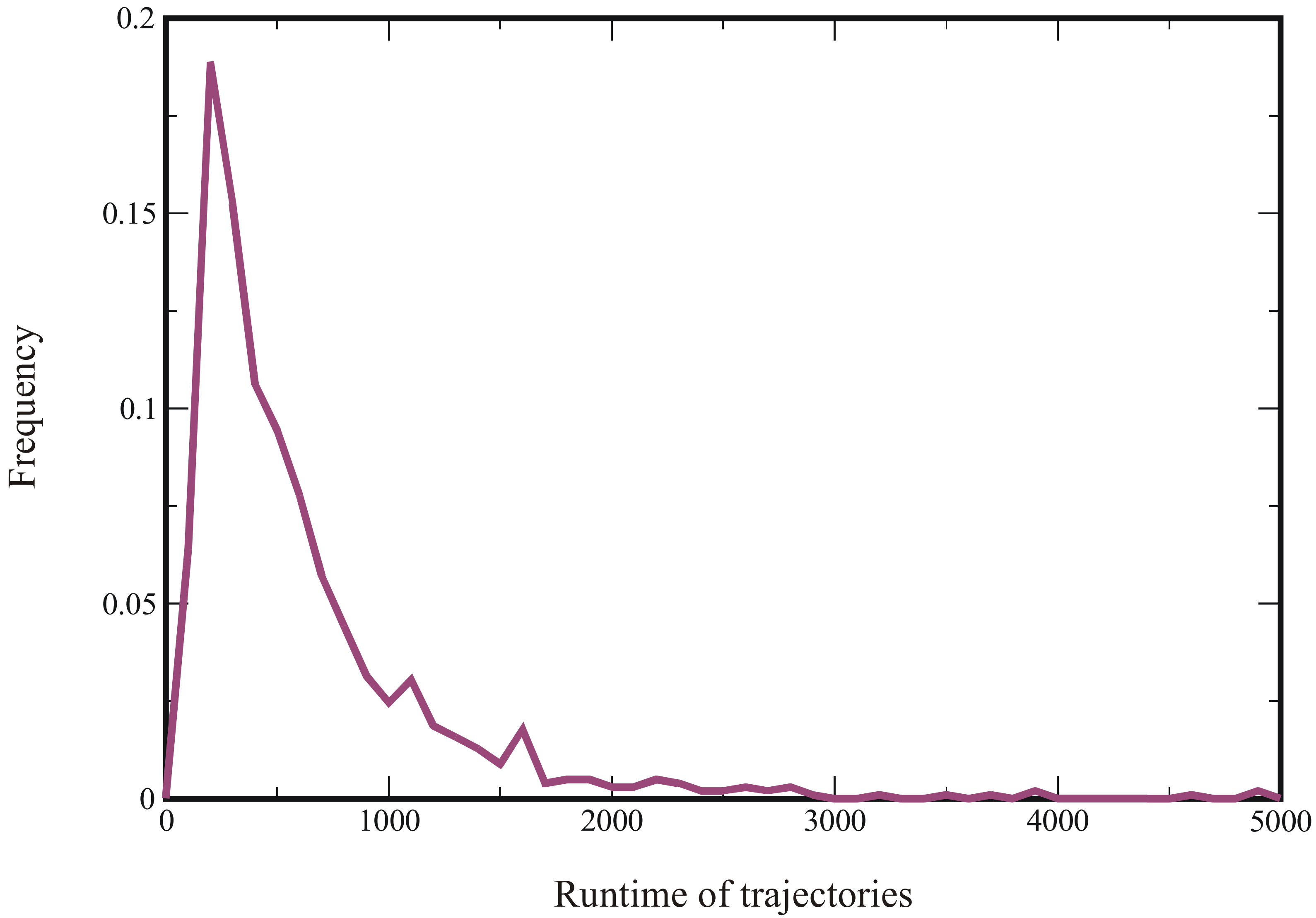
$$\varepsilon-\text{neighborhood of } S_k : \quad \Psi_\varepsilon(S_k) = \left\{ S_j \in Y(S_k) \,|\, \rho(S_j;S_k) > \varepsilon \right\}$$



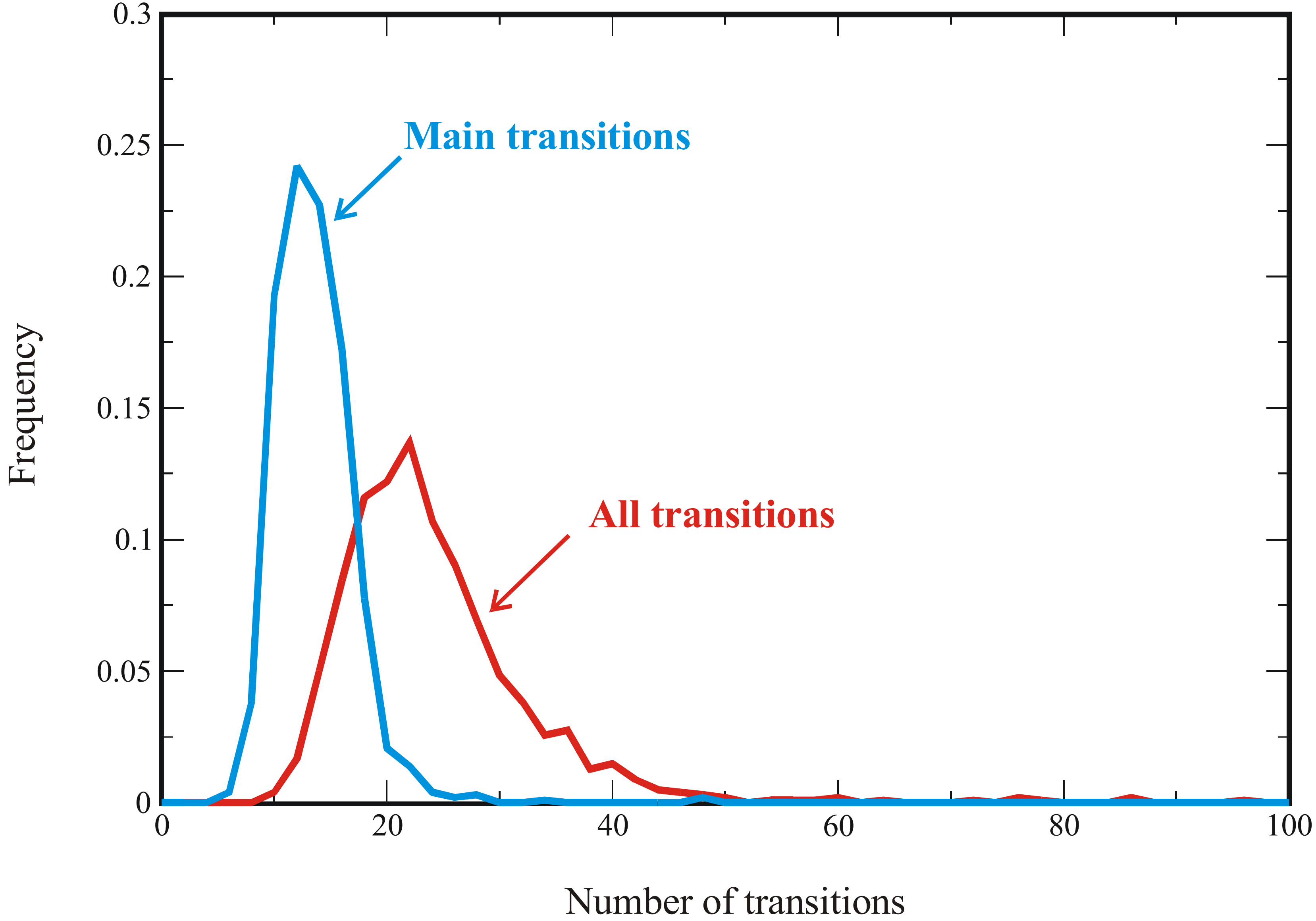
AUGC

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)

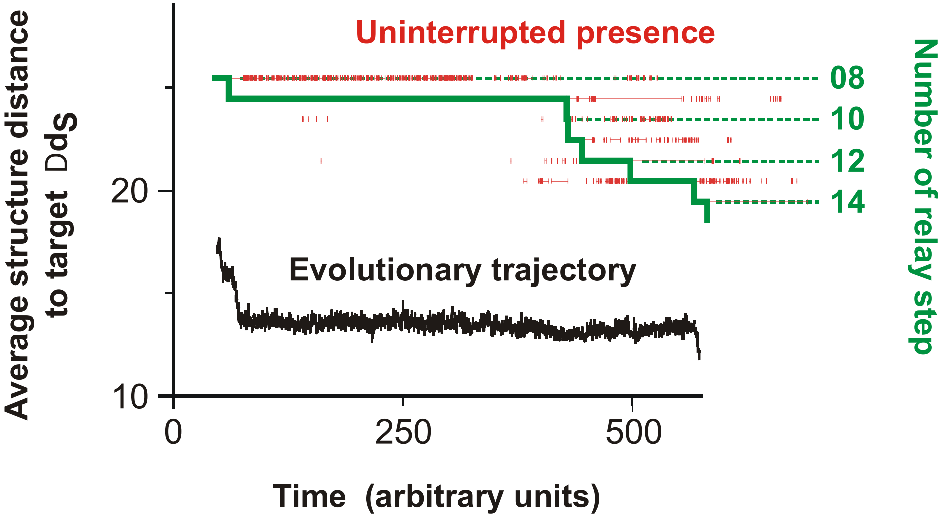


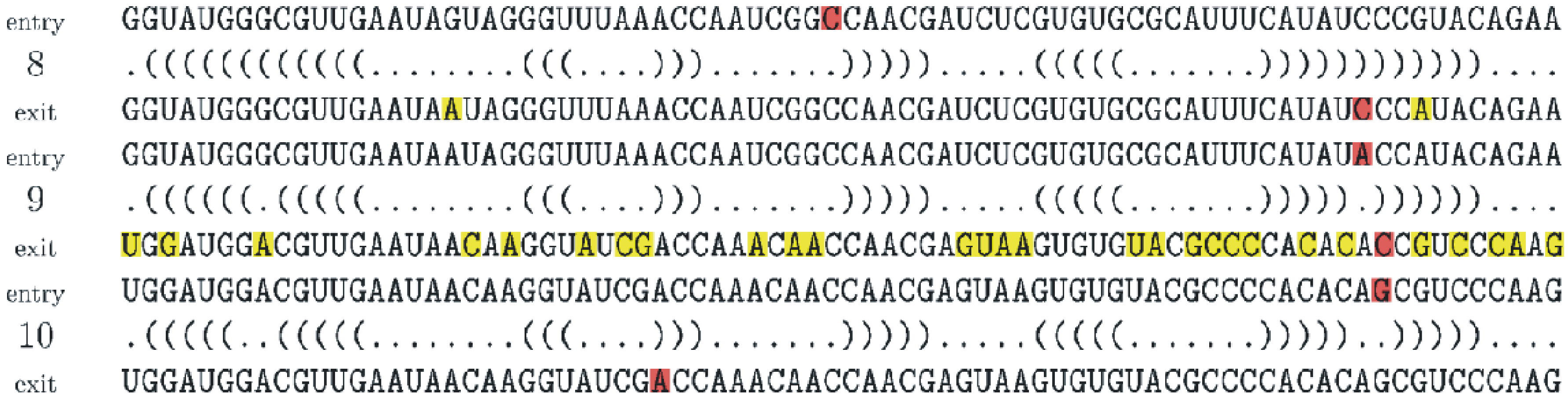
Statistics of the numbers of transitions from initial structure to target (**AUGC**-sequences)

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)

28 neutral point mutations during a long quasi-stationary epoch



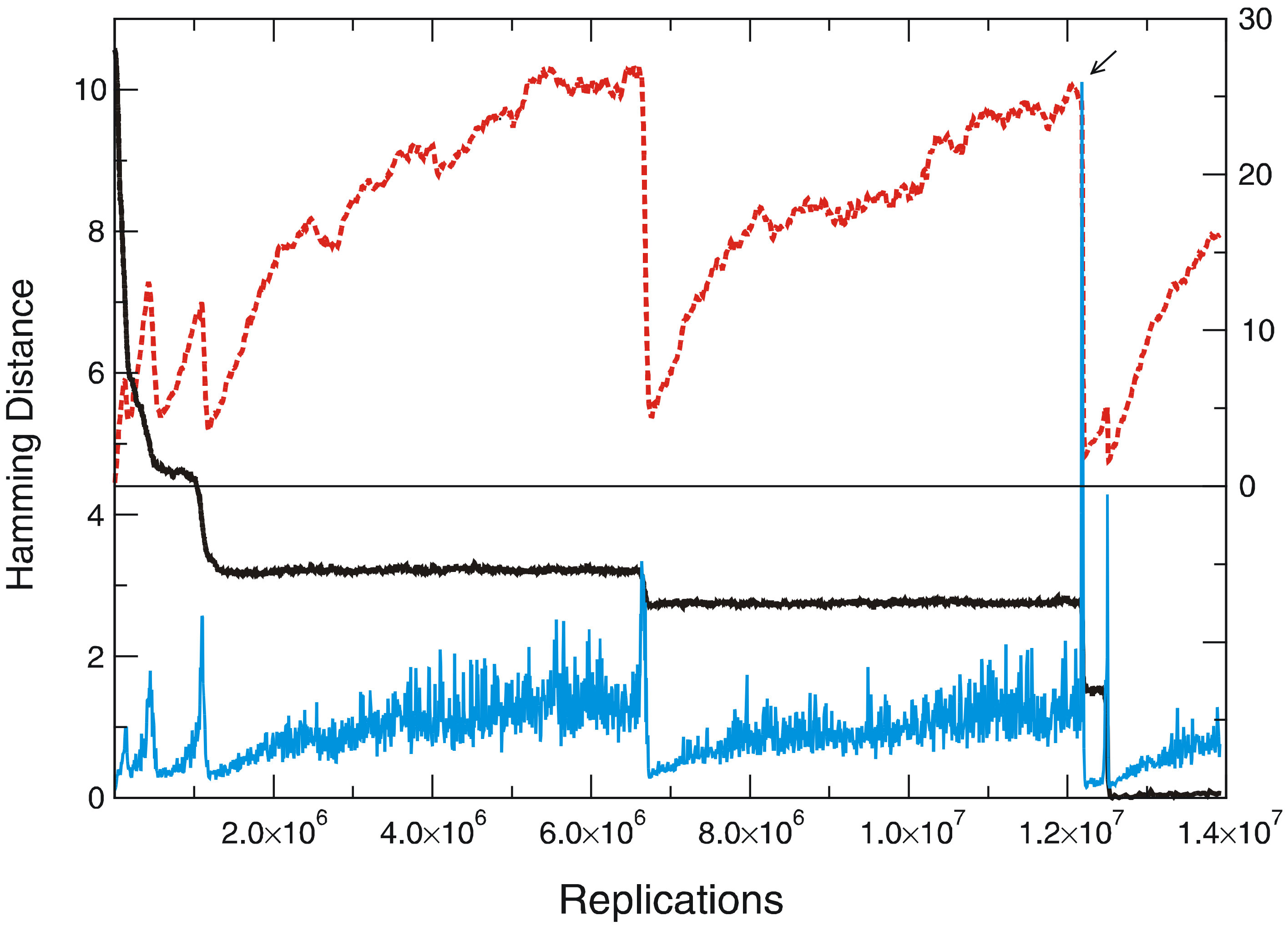


```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
  8    .((((((((((((........(((....))).......))))).....(((((.......))))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations

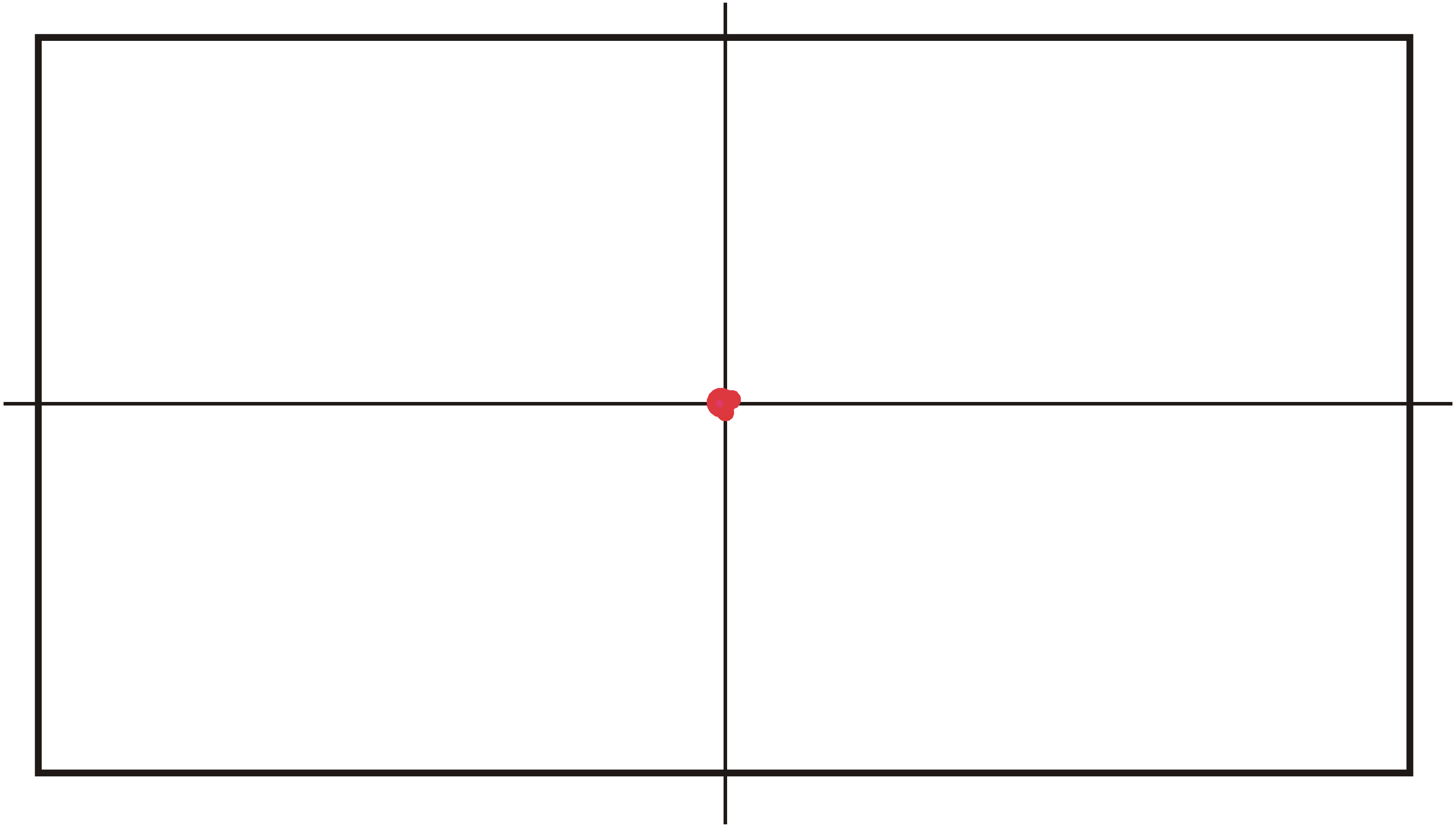
Neutral point mutations

Neutral genotype evolution during phenotypic stasis

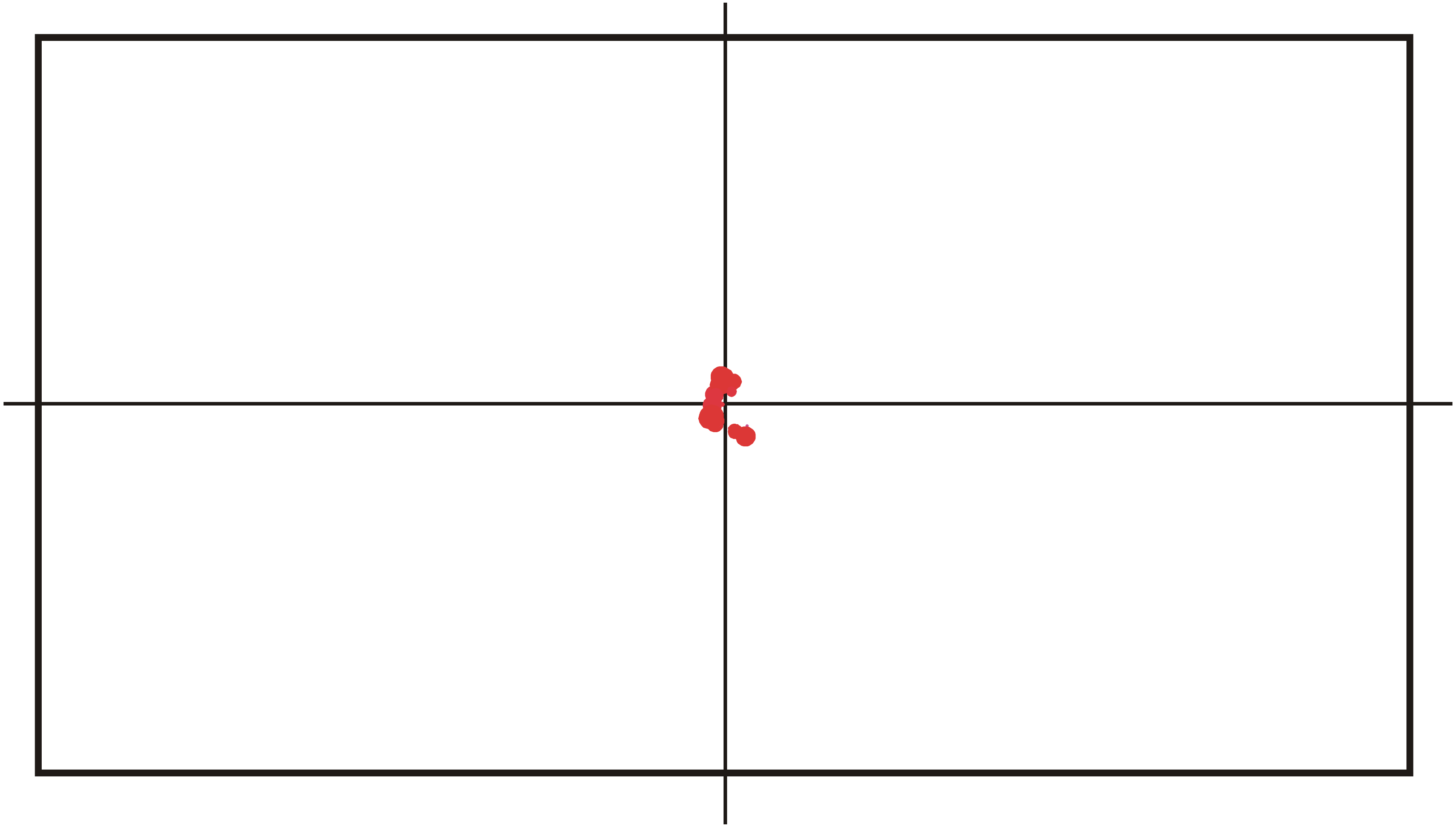


Variation in genotype space during optimization of phenotypes

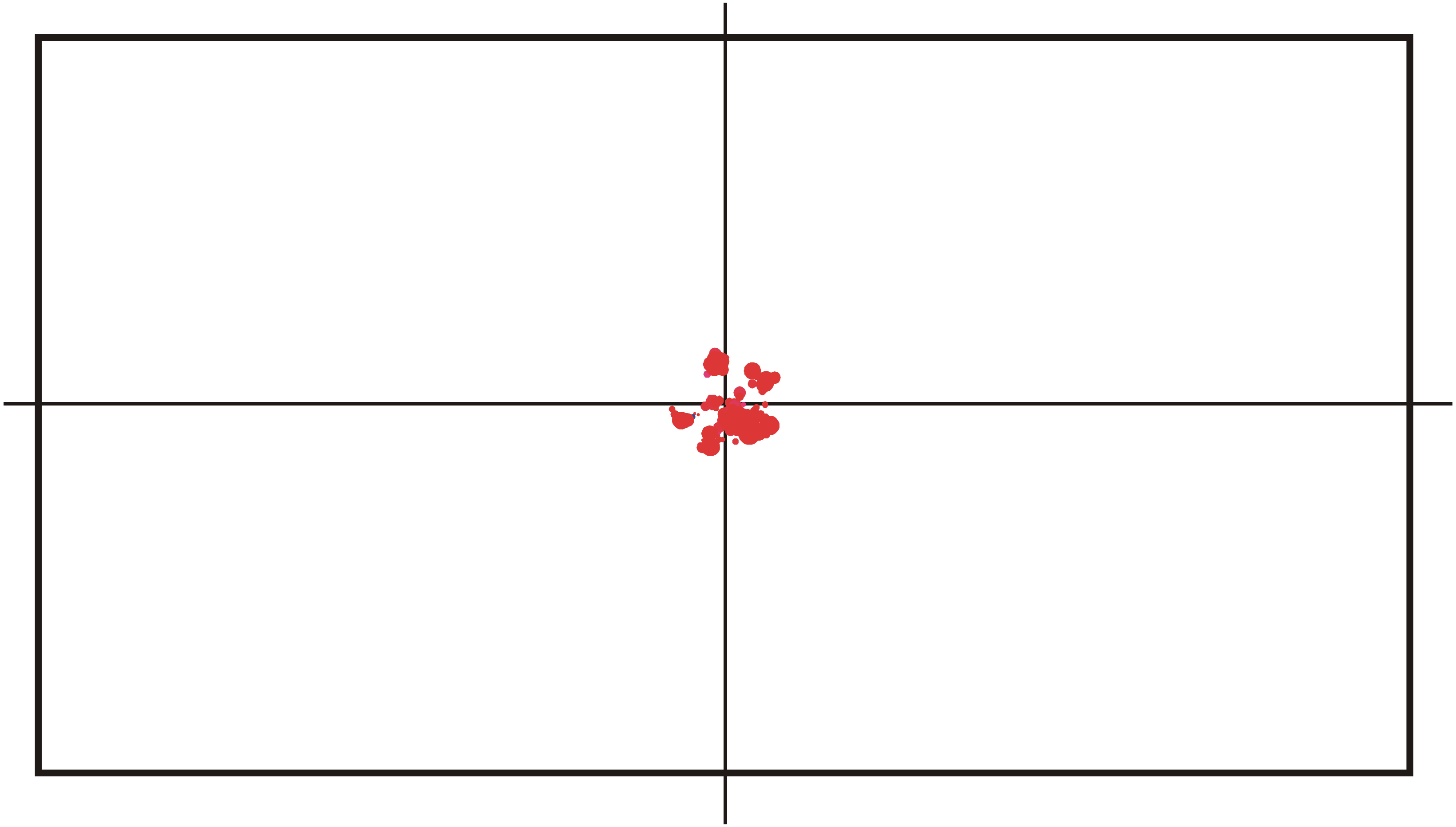
Mean Hamming distance within the population and **drift velocity of the population center** in sequence space.



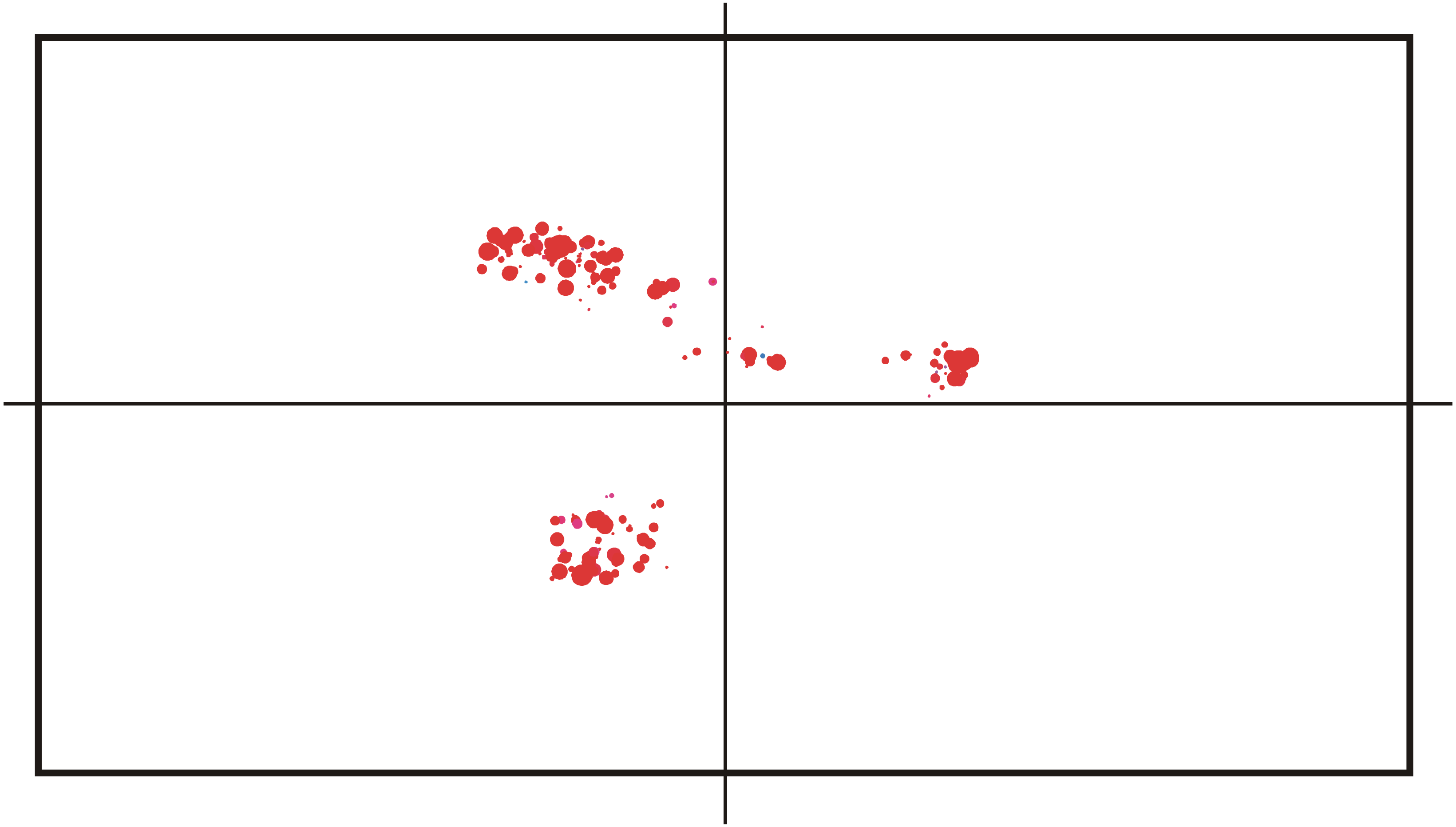
Spread of population in sequence space during a quasistationary epoch: t = 150



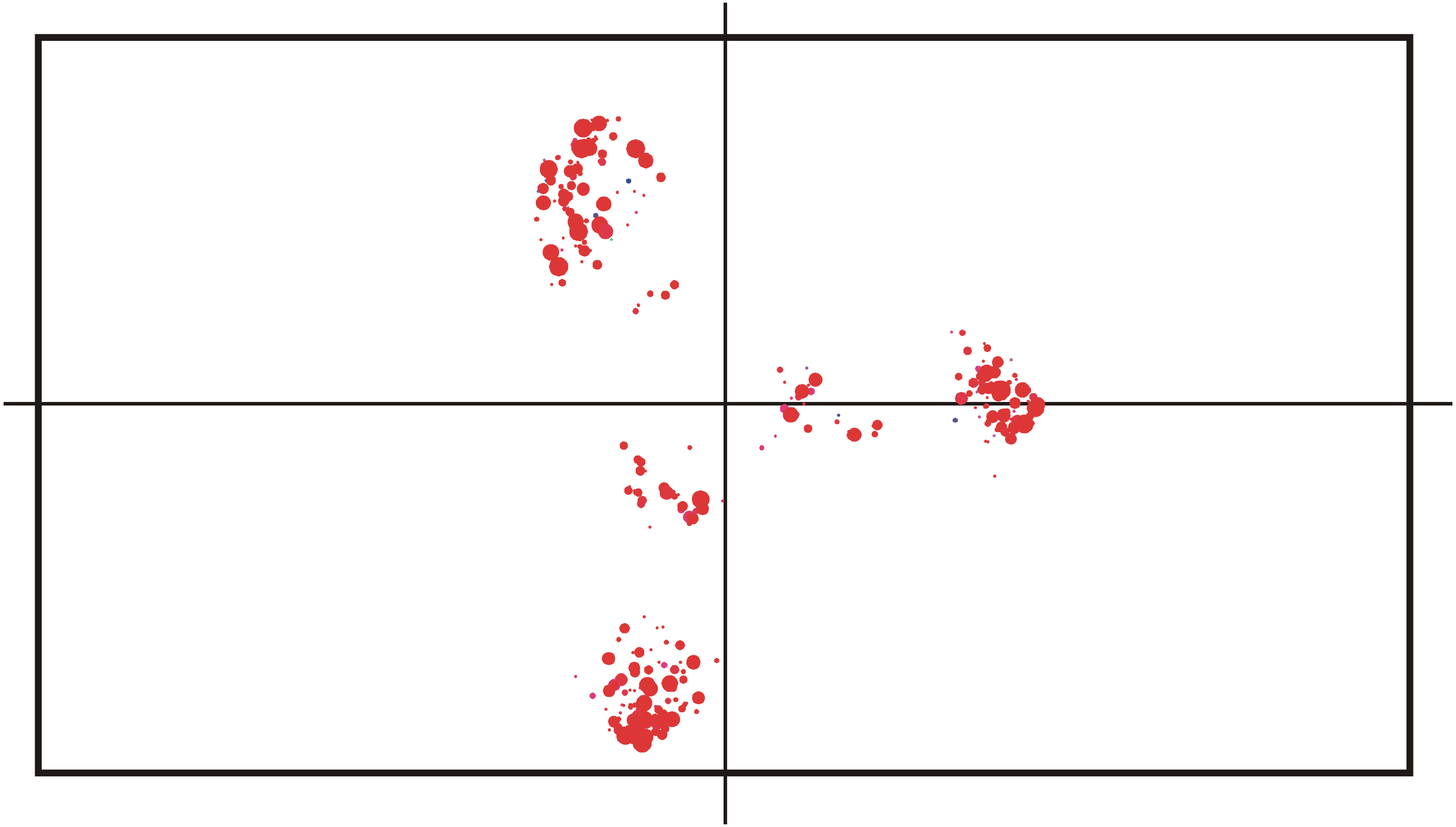
Spread of population in sequence space during a quasistationary epoch: t = 170



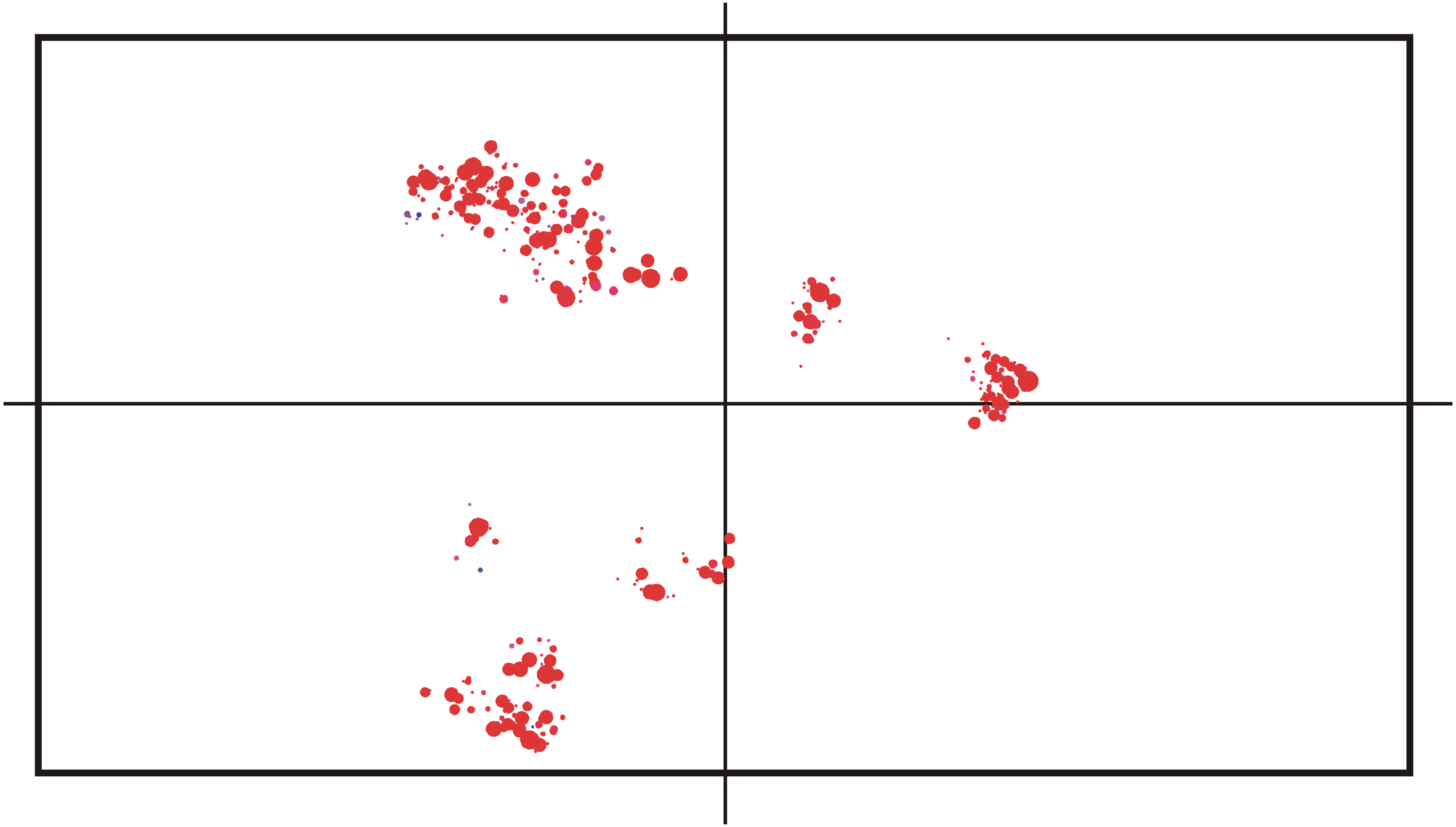
Spread of population in sequence space during a quasistationary epoch: t = 200



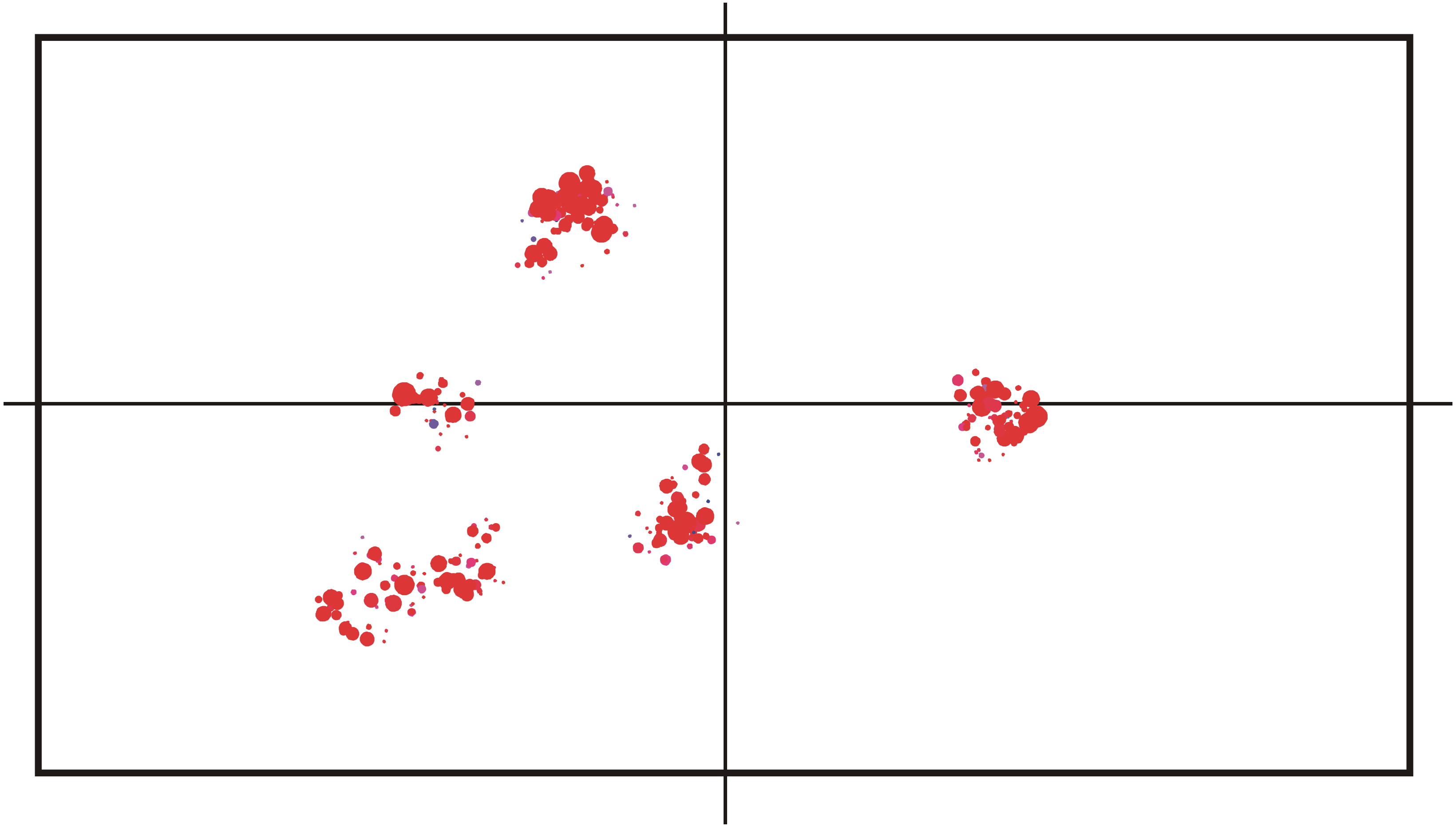
Spread of population in sequence space during a quasistationary epoch: t = 350



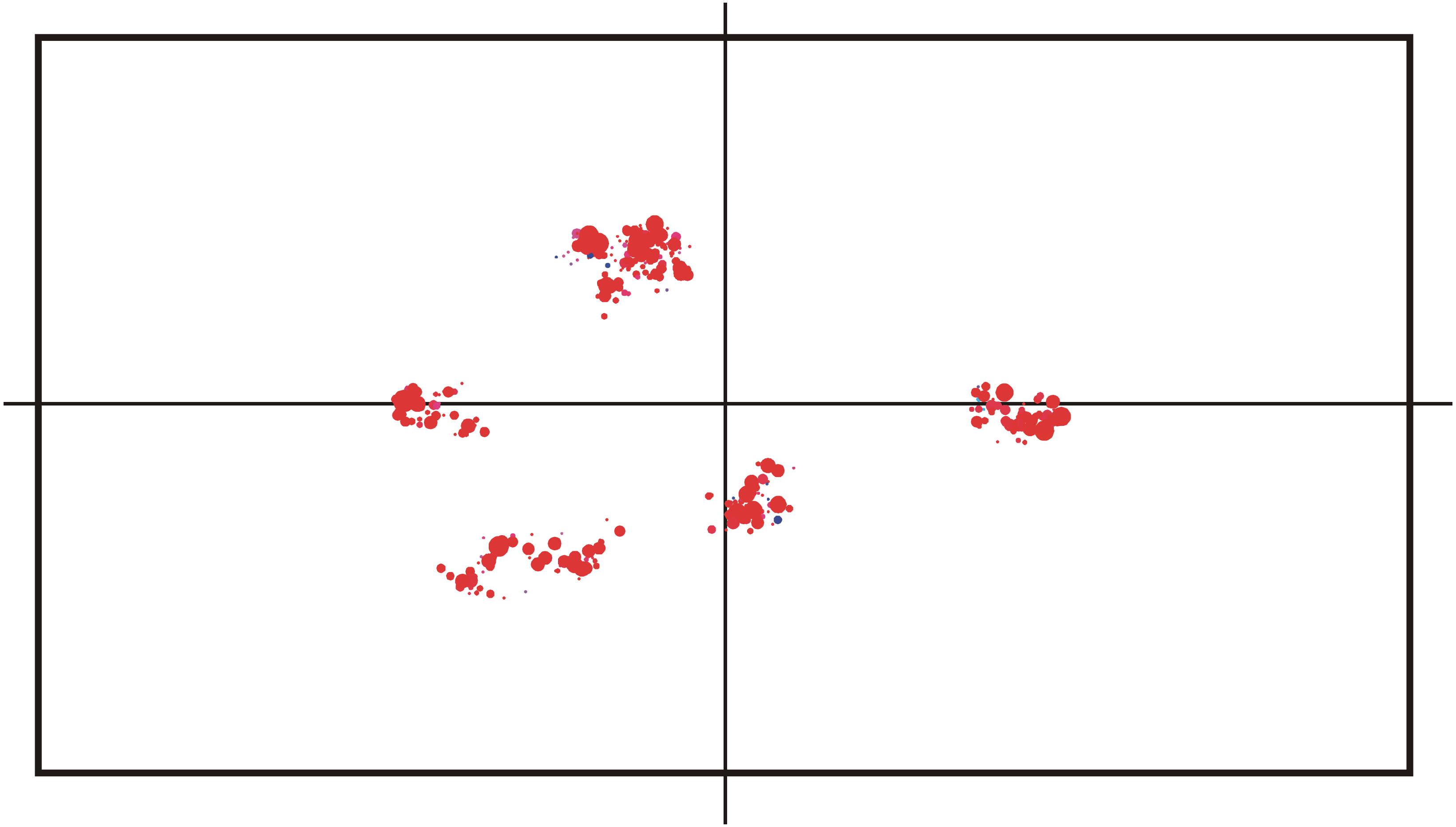
Spread of population in sequence space during a quasistationary epoch: t = 500



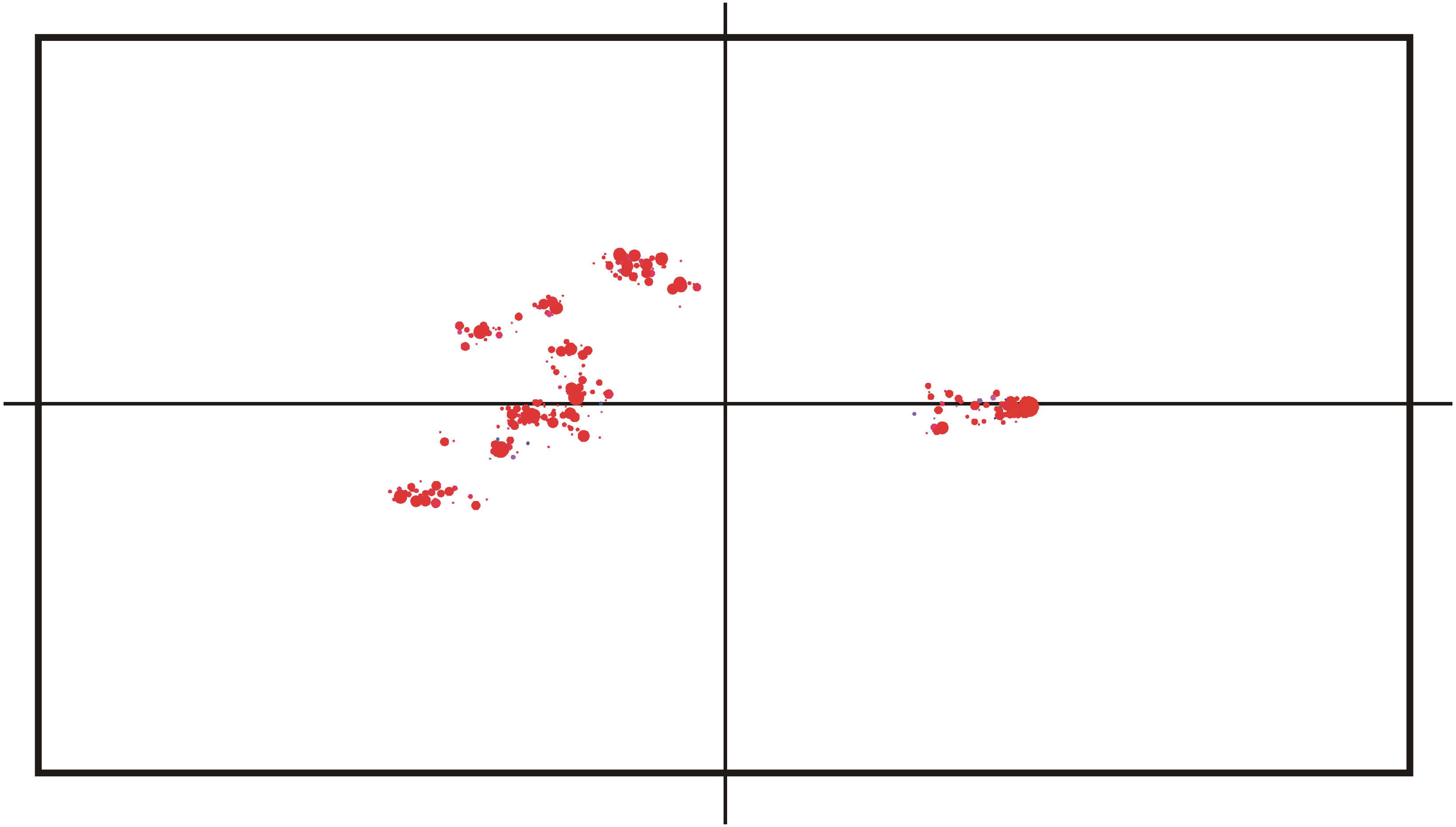
Spread of population in sequence space during a quasistationary epoch: t = 650



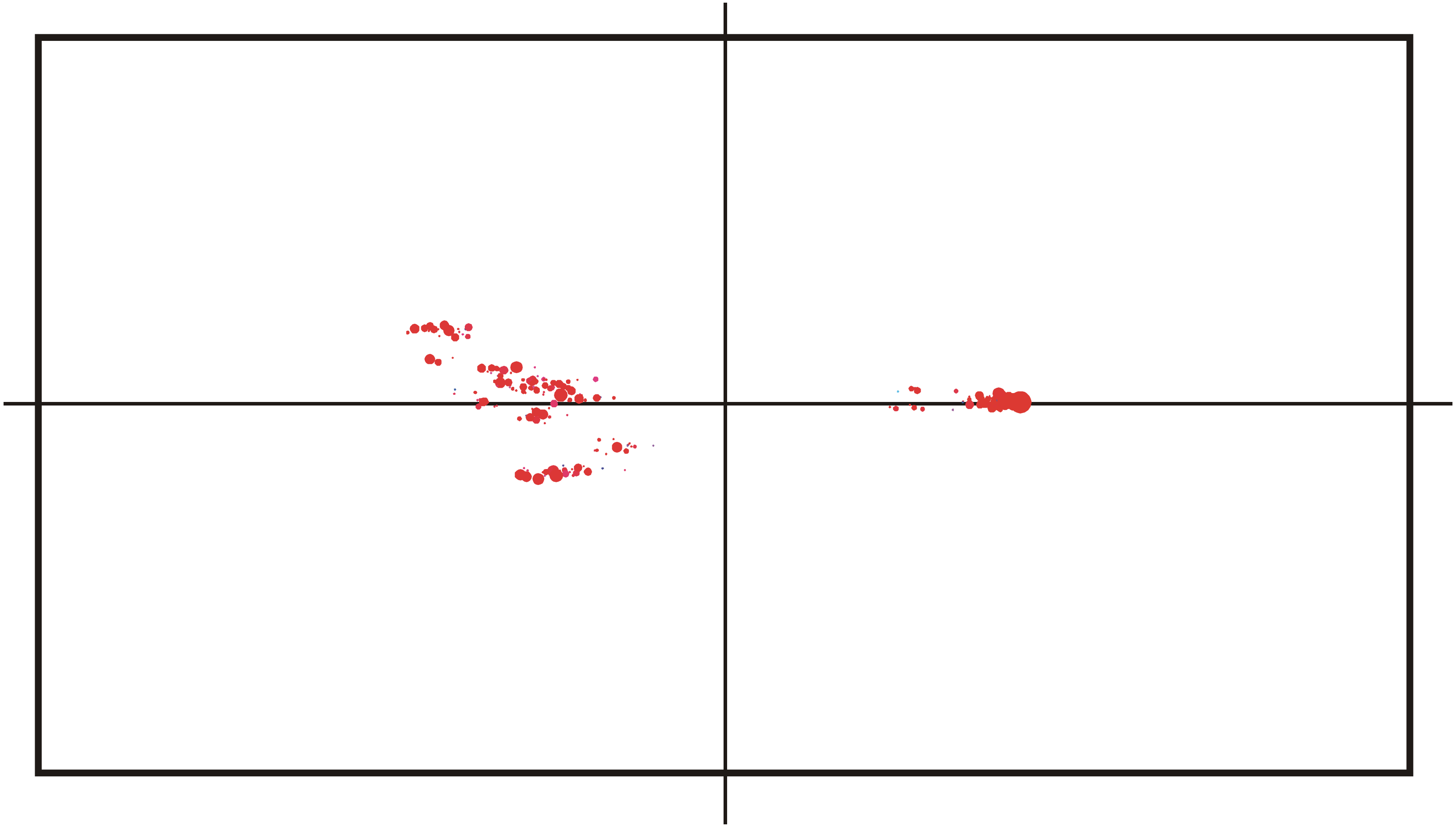
Spread of population in sequence space during a quasistationary epoch: t = 820



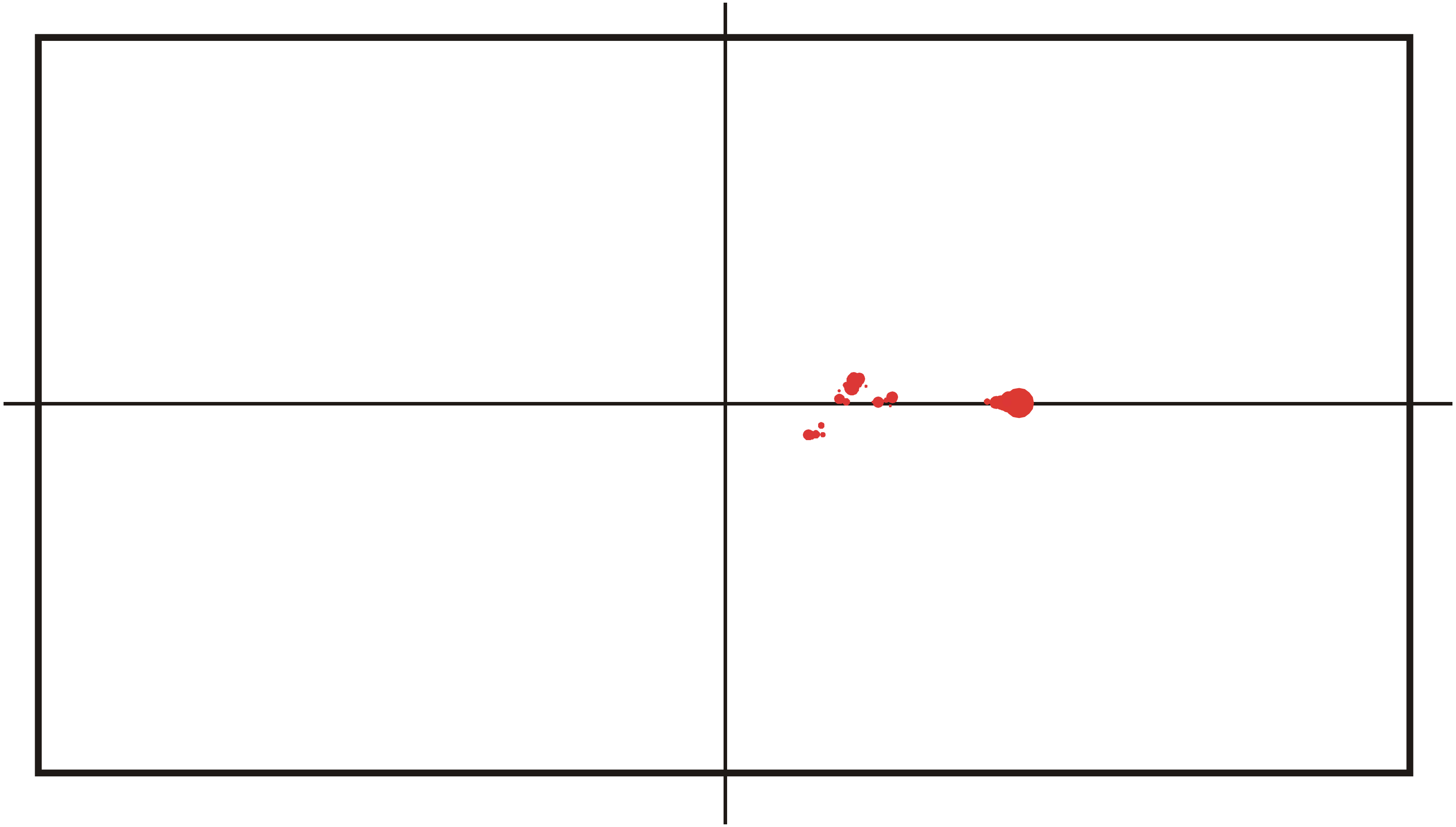
Spread of population in sequence space during a quasistationary epoch: t = 825



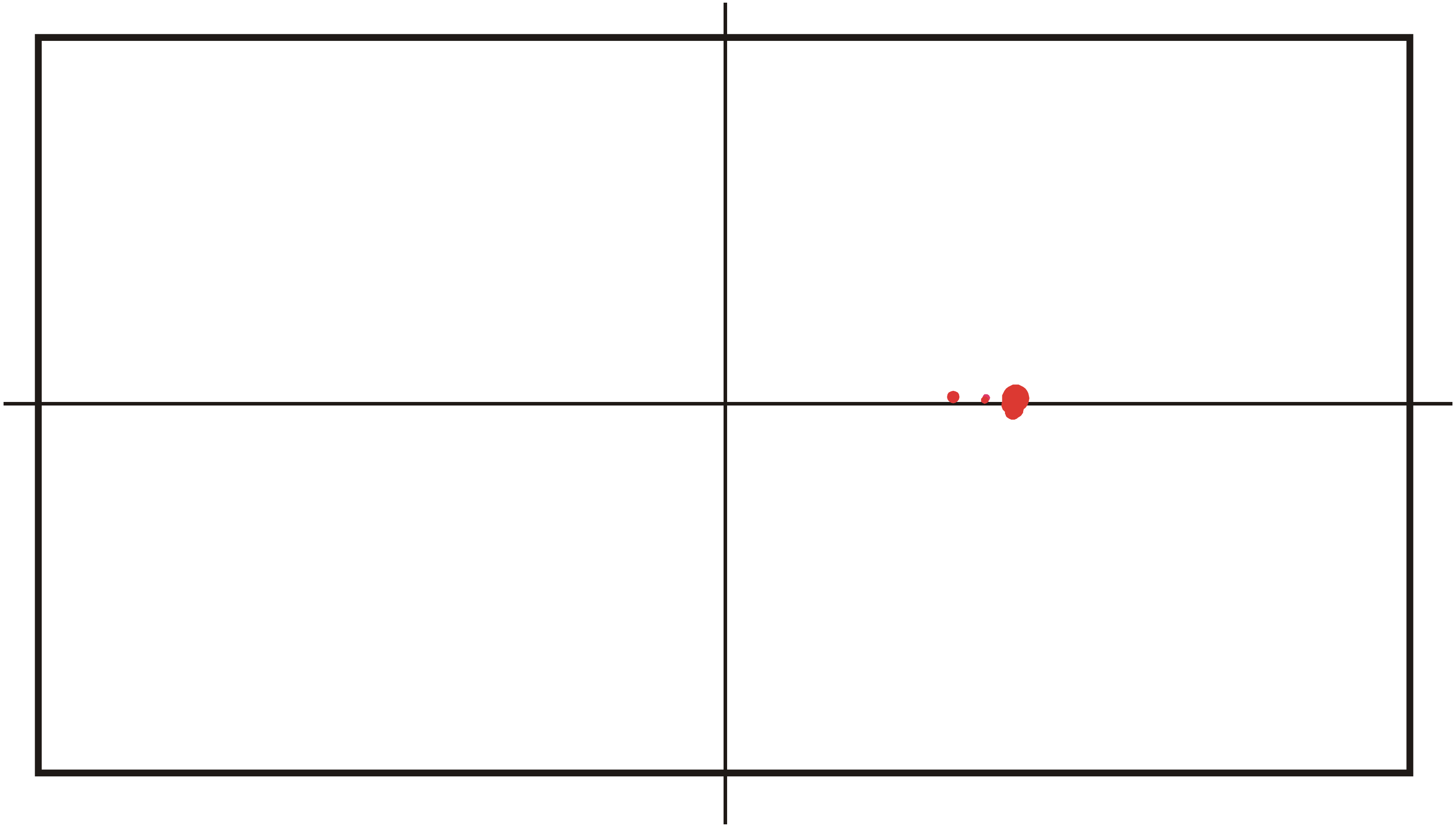
Spread of population in sequence space during a quasistationary epoch: t = 830



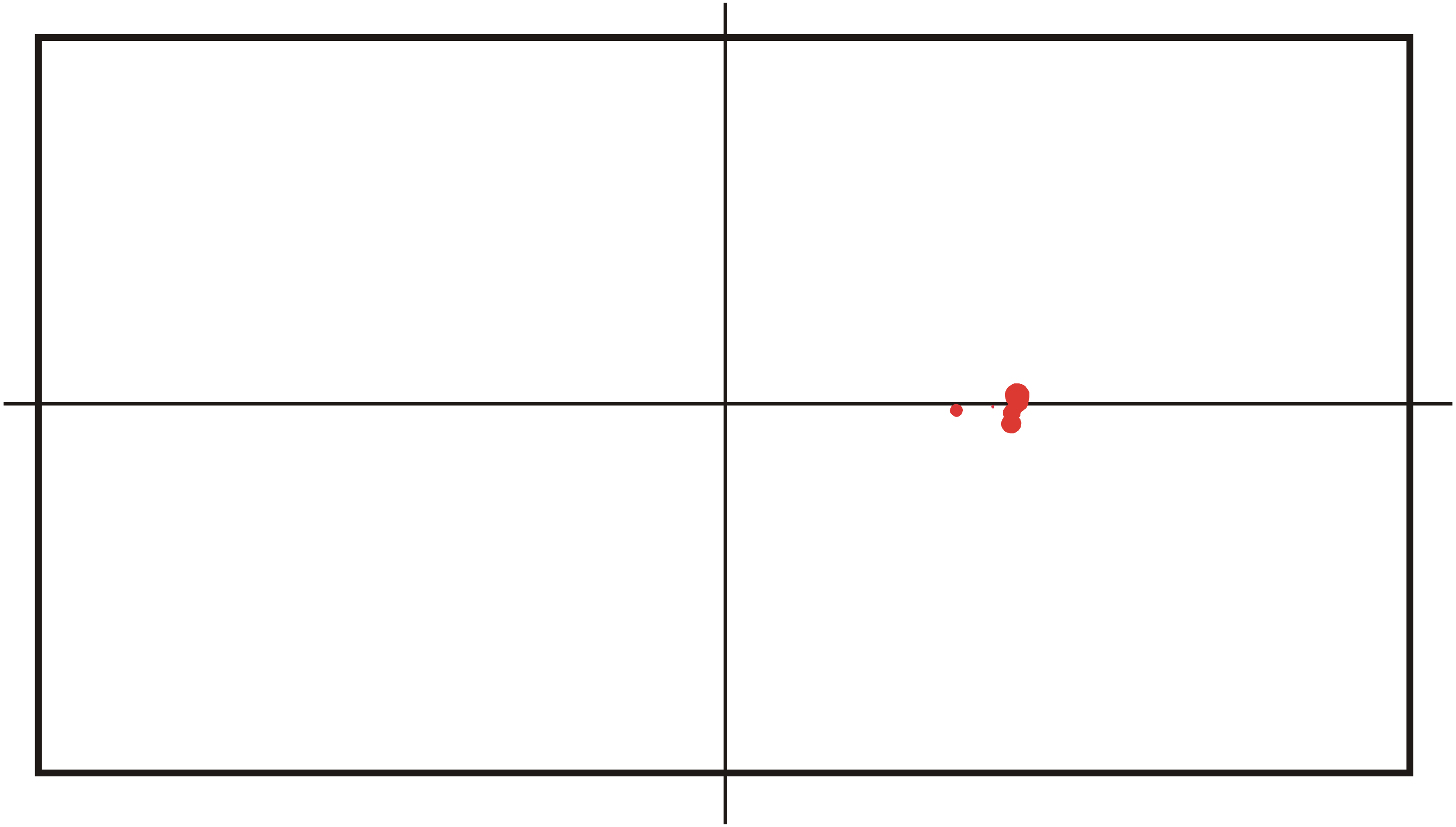
Spread of population in sequence space during a quasistationary epoch: t = 835



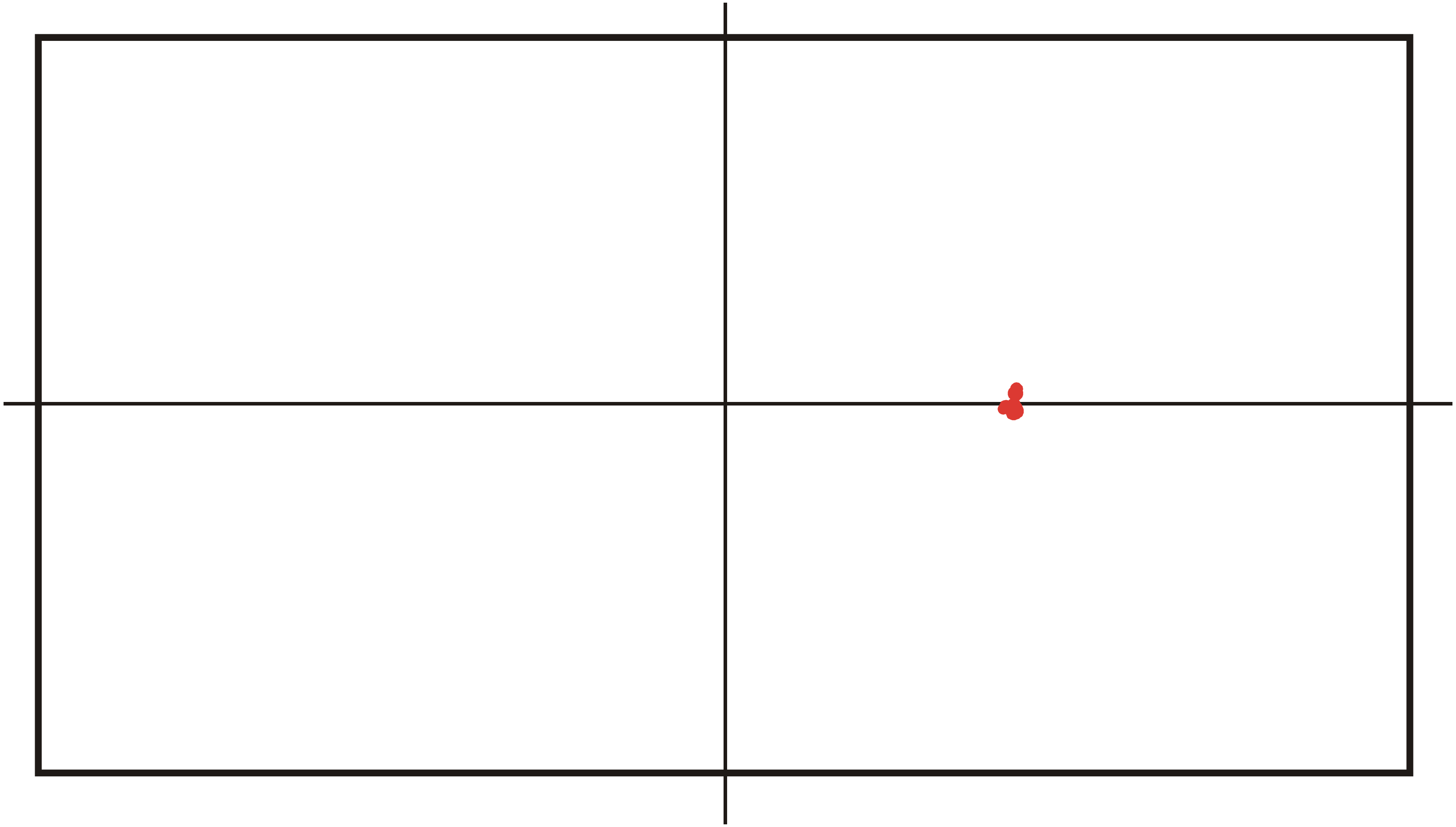
Spread of population in sequence space during a quasistationary epoch: t = 840



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

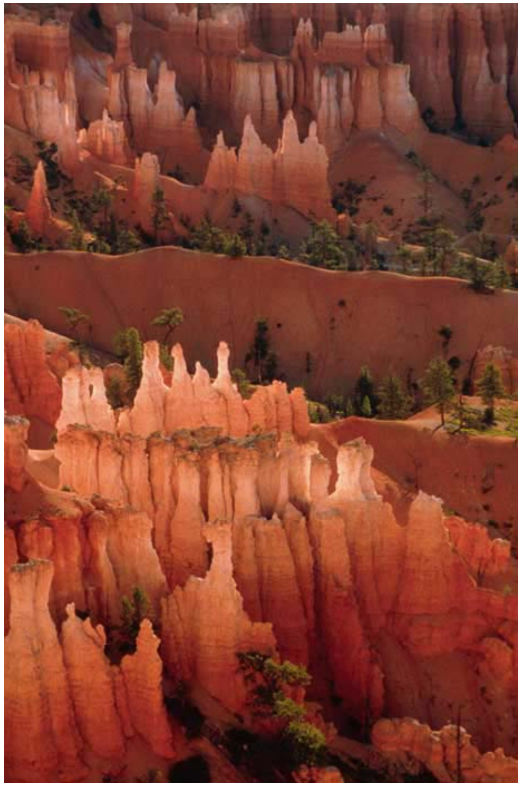
Massif Central

Mount Fuji

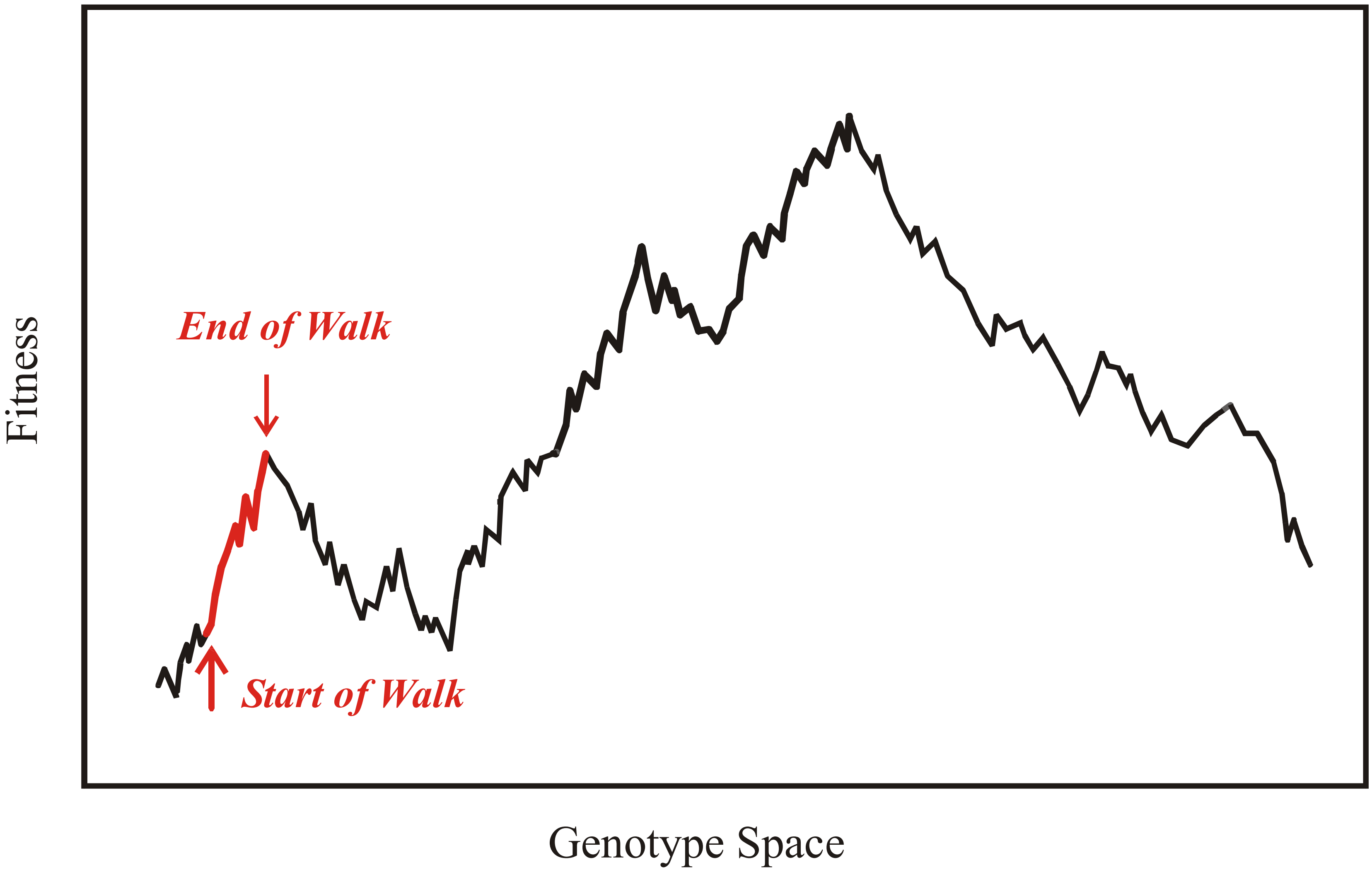
Examples of smooth landscapes on Earth

Examples of rugged landscapes on Earth

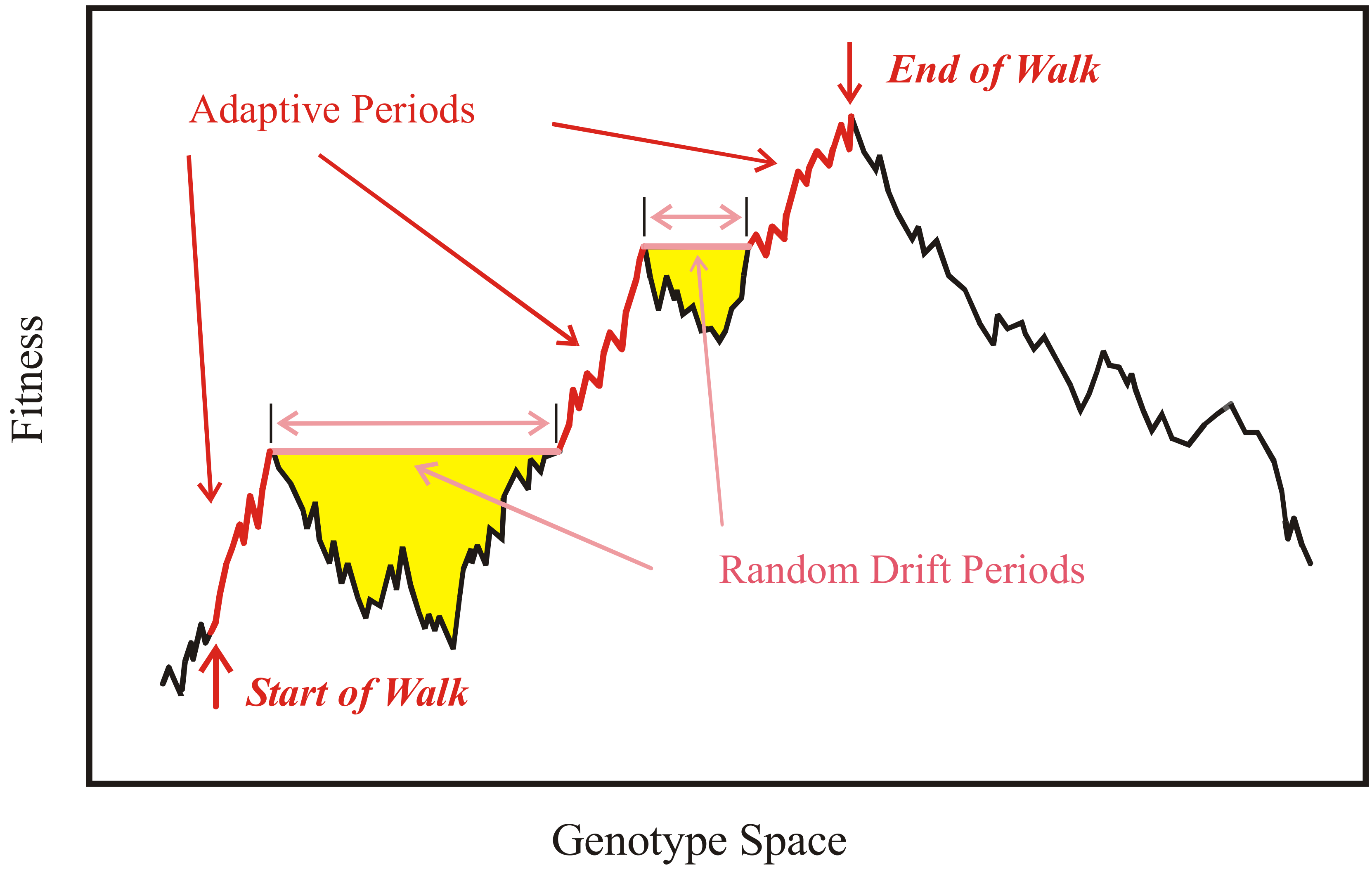
Dolomites



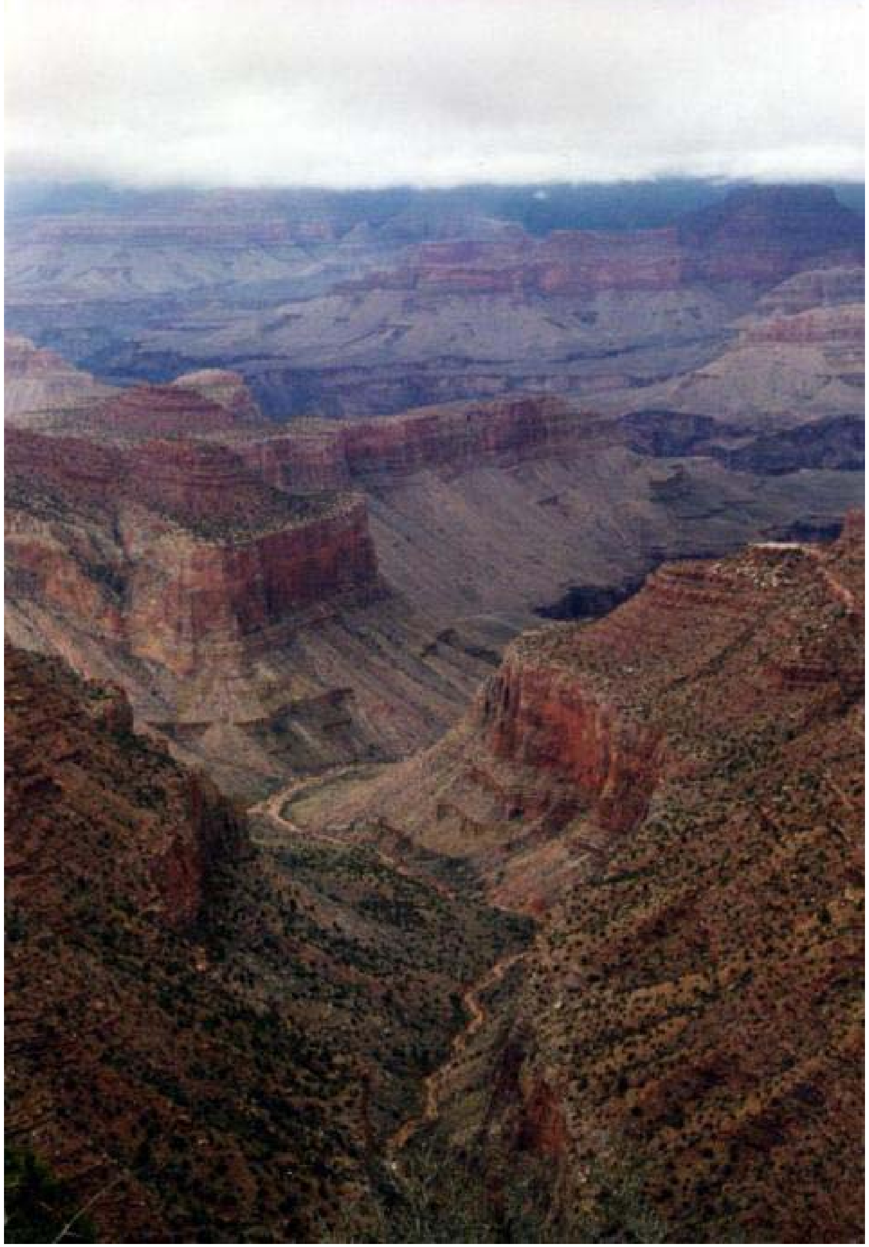
Bryce Canyon



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

Neutral ridges and plateus

Universität Wien

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

Siemens AG, Austria

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