Evolutionäre Optimierung von Molekülen

Von mathematischer Modellierung zur Bestätigung im Experiment

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

DMV-Jahrestagung 2002

Halle an der Saale, 16.– 21.09.2002

Das Darwinsche Optimierungsprinzip baut auf drei Voraussetzungen auf.

1. Reproduktion von Organismen durch Vermehrung der **Phänotypen**

   Die Reproduktionseffizienz wird gemessen als Zahl der fruchtbaren Nachkommen oder Fitness.

2. **Variation** der **Genotypen** durch Kopierfehler und Rekombination

   Die Genotypen oder Genome sind der Träger der genetischen Information.

3. Selektion durch Unterschiede in der Fitness der **Phänotypen**

Zwei zusätzlichen Voraussetzungen

4. Eine hinreichend große Zahl unterschiedlicher **Genotypen** und eine hinreichend große Vielfalt an **Phänotypen**
5. Eine für die Optimierung unterstützende Beziehung zwischen den **Genotypen** und den **Phänotypen**

   Die Beziehung zwischen Genotypen und Phänotypen wird als eine Abbildung von einem Raum der Genotypen in einen Raum der Phänotypen verstanden.

Die Ursache für den Erfolg und die universelle Anwendbarkeit des Darwinschen Optimierungsprinzips bildet gleichzeitig den Grund für seine einscheidende Beschränkung:

Die inneren Strukturen der sich reproduzierenden Einheiten gehen nur in Form der Fitnessparameter ein. Es ist gleichgültig, ob Moleküle, nicht-autonome oder autonome Organismen, Kolonien, Vielzeller oder Gesellschaften vermehrt werden.

In dieser Form bietet die biologische Evolutionstheorie nur eine rein ordnende makroskopische Beschreibung der beobachtbaren Phänomene an.

1. **Optimierung durch Variation und Selektion in Populationen**

2. **Neutrale Netzwerke in Genotype-Phänotyp-Abbildungen**

3. **Optimierung im RNA-Modell**

4. **Evolutionsexperimente mit Molekülen im Laboratorium**

Das Darwinsche Optimierungsprinzip ist im Fall von null verschiedener Mutationsraten ($q<1$ oder $p>0$) nur als eine **Optimierungsheuristik** zu verstehen. Es gilt nur in einem Teil des Simplex der relativen Konzentrationen. Mit steigender Mutationsrate p wird der Teil des Konzentrationsraumes, in welchem das Optimierungsprinzip gilt, immer kleiner.

Analog gilt für das Selektions-Rekombinationsmodell, dass das Fishersche Optimierungskriterium nur eingeschränkt auf das Ein-Gen-Modell (Single locus model) gültig ist.

Evolutionary Optimization of Molecules

From mathematical models to confirmation by experiment

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

DMV-Jahrestagung 2002

Halle an der Saale, 16.– 21.09.2002

The Darwinian principle of optimization is built on three prerequisites:

1. Reproduction of organisms through multiplication of **phenotypes**

   Efficiency of reproduction is measured as fitness being tantamount to the number of fertile descendants which are brought into the next generation.

2. **Variation** of **genotypes** though copying errors and recombination

   The genotypes or genomes are the carriers of genetic information.

3. Selection through differences in the fitness of **phenotypes**

Two additional prerequisites

4. A large enough number of **genotypes** and a sufficiently large reservoir of diversity of **phenotypes**
5. A relation between **genotypes** and **phenotypes** that supports optimization through variation and selection

   The relation between genotypes and phenotypes is understood as a mapping from a space of genotypes onto a space of phenotypes.

The basis for success and universal applicability of the Darwinian priciple of optimization represents, at the same time, also its most serious limitation:

The internal structures of the reproducing units are addressed only in terms of fitness parameters. Therefore, it does not matter whether multiplication concerns molecules, non-autonomous or autonomous cells, colonies, multicellular organisms or societies.

The theory of biological evolution in this form can provide only a macroscopic description and classification as well as ordering relations of the observed phenomena.

1. **Optimization through variation and selection in populations**

2. **Neutral networks in genotype-phenotype mappings**

3. **Optimization in the RNA model**

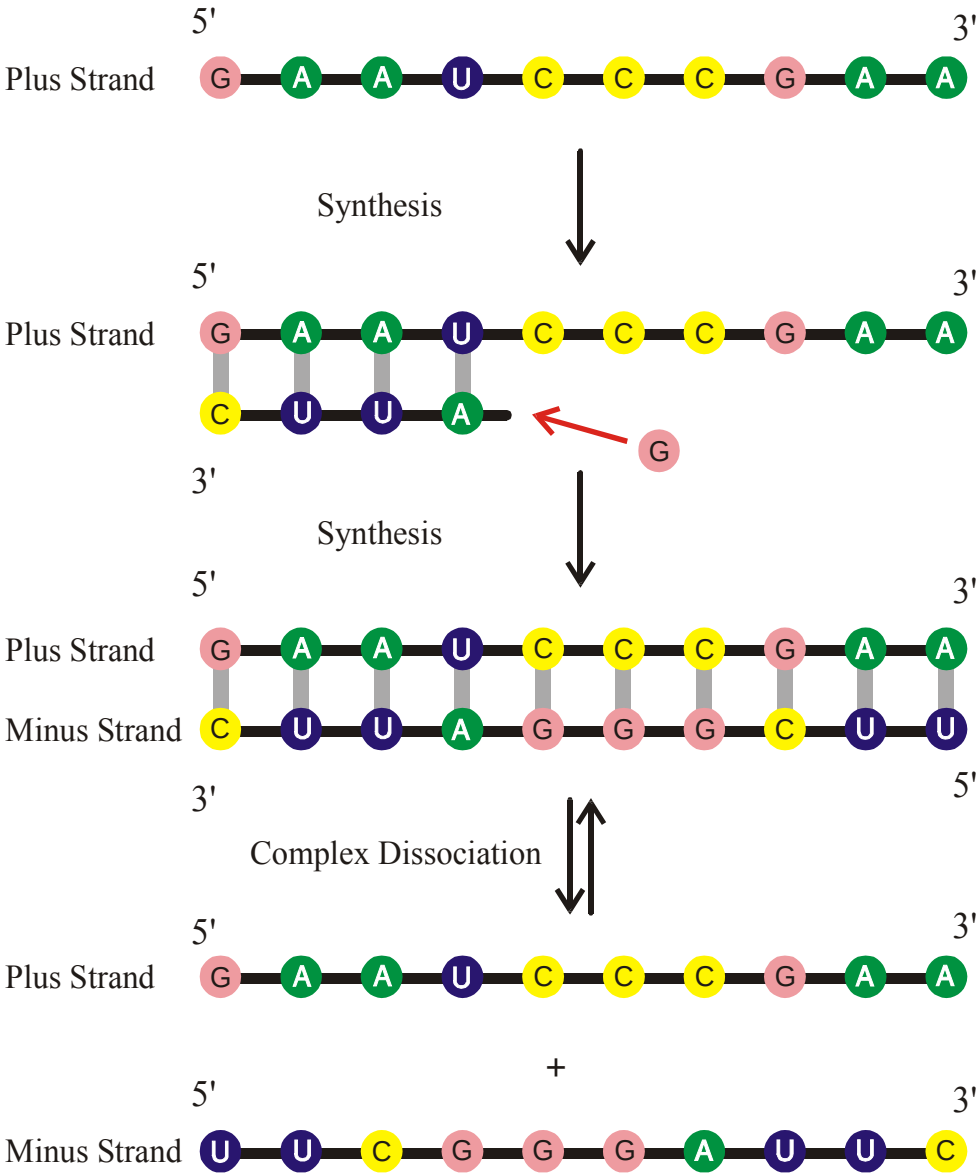
4. **Evolution experiments with molecules in the laboratory**

1. **Optimization through variation and selection** in populations

2. Neutral networks in genotype-phenotype mappings

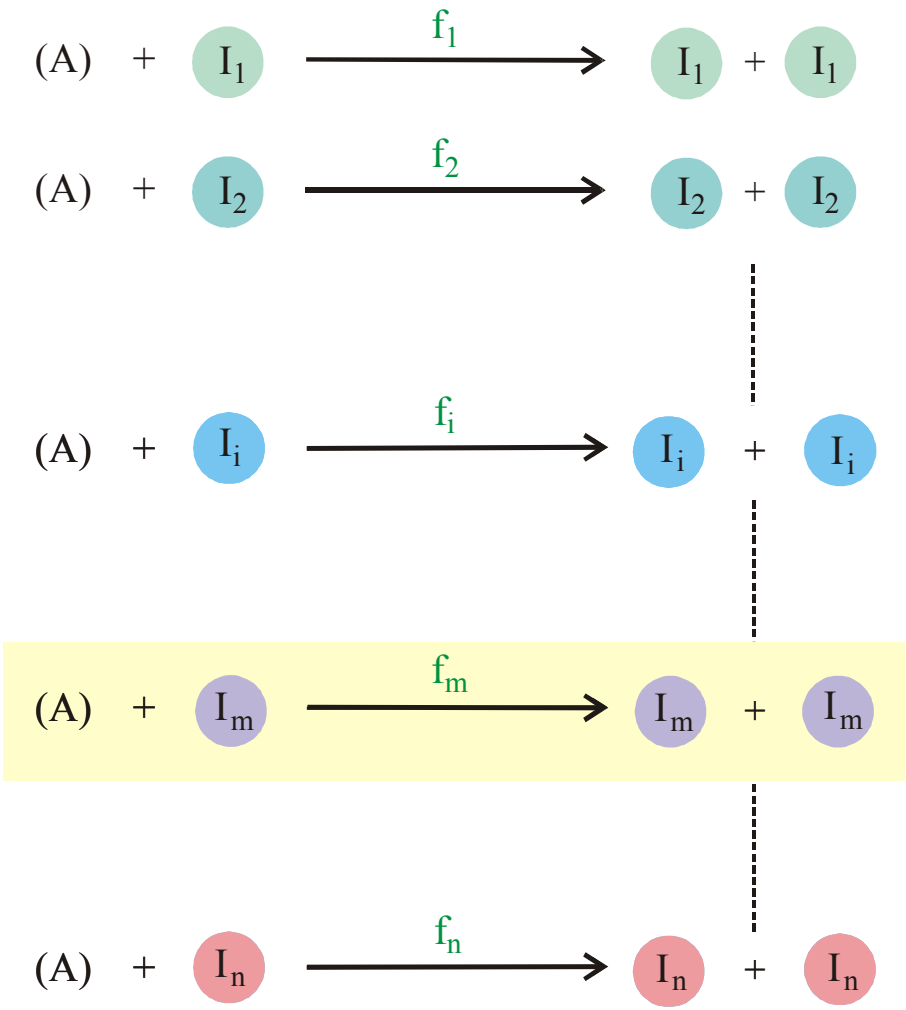
3. Optimization in the RNA model

4. Evolution experiments with molecules in the laboratory



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

G Í C and A=U



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

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

$[I_i] = x_i$ Æ 0 ; i =1,2,...,n ;

[A] = a = constant

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

$x_m(t)$ Á 1 for t Á ¸

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

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

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

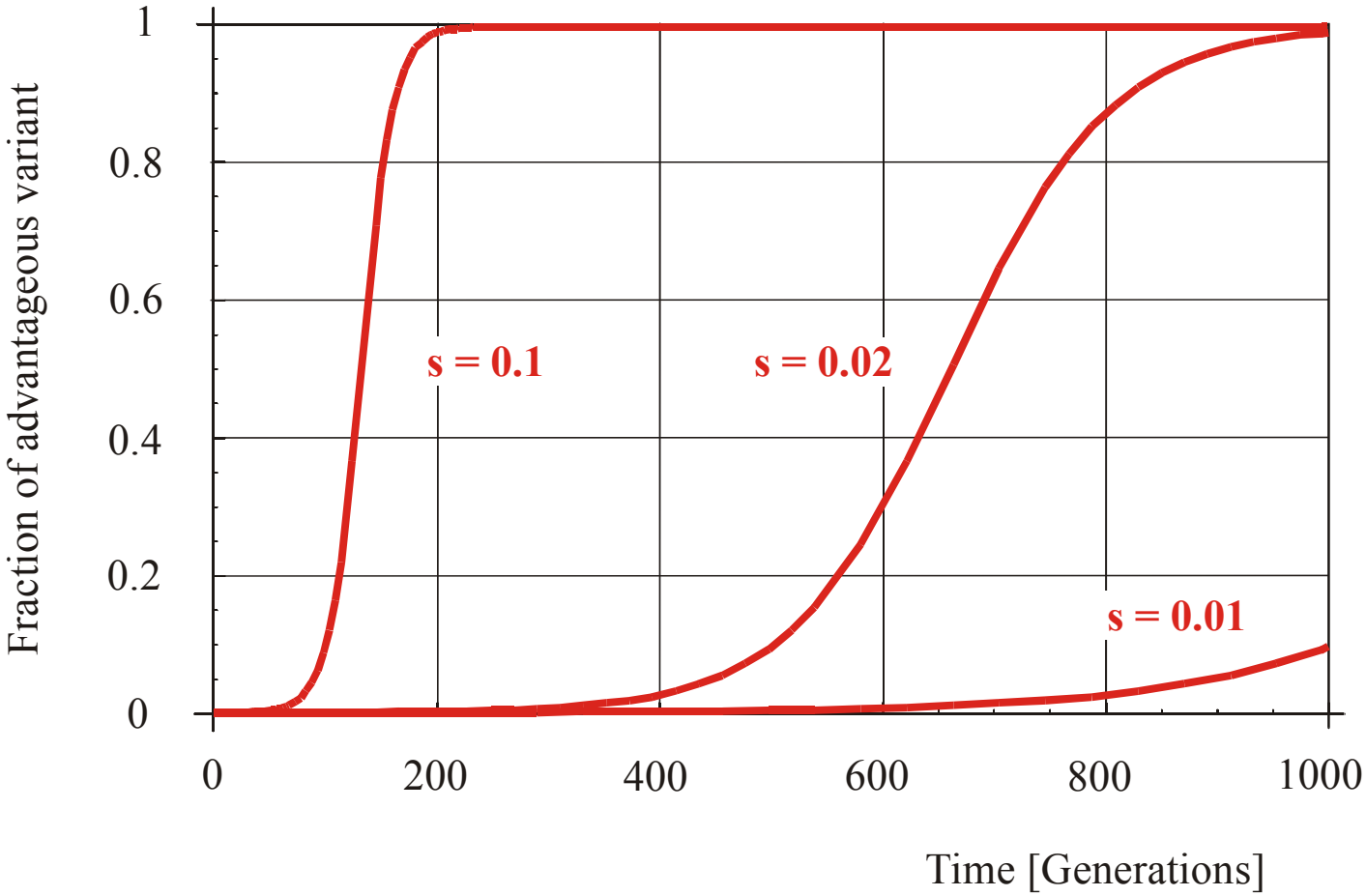
Mean fitness or dilution flux, ϕ(t), is a **non-decreasing function** of time,

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

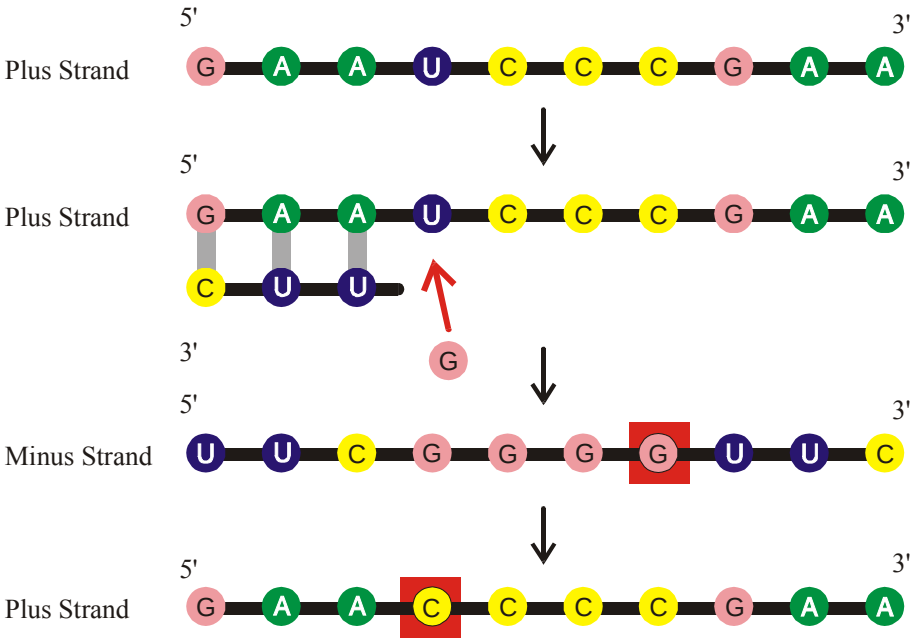
Solutions are obtained by integrating factor transformation

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

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



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



Point Mutation

GAAUCCCGAA → GAAUCCCGUCCCGAA

Insertion

GAAUCCCGAA → GAAUCCGA

Deletion

Mutations in nucleic acids represent the mechanism of **variation** of **genotypes**.

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

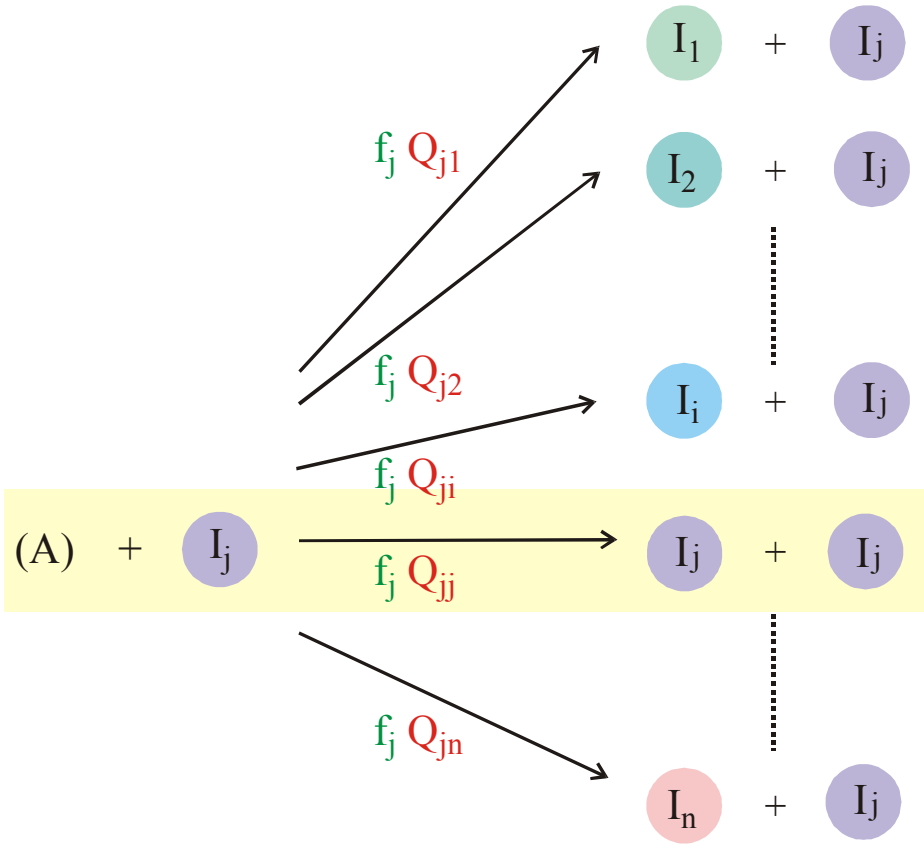
M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle***. Naturwissenschaften **65** (1978), 341-369

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



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

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

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

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

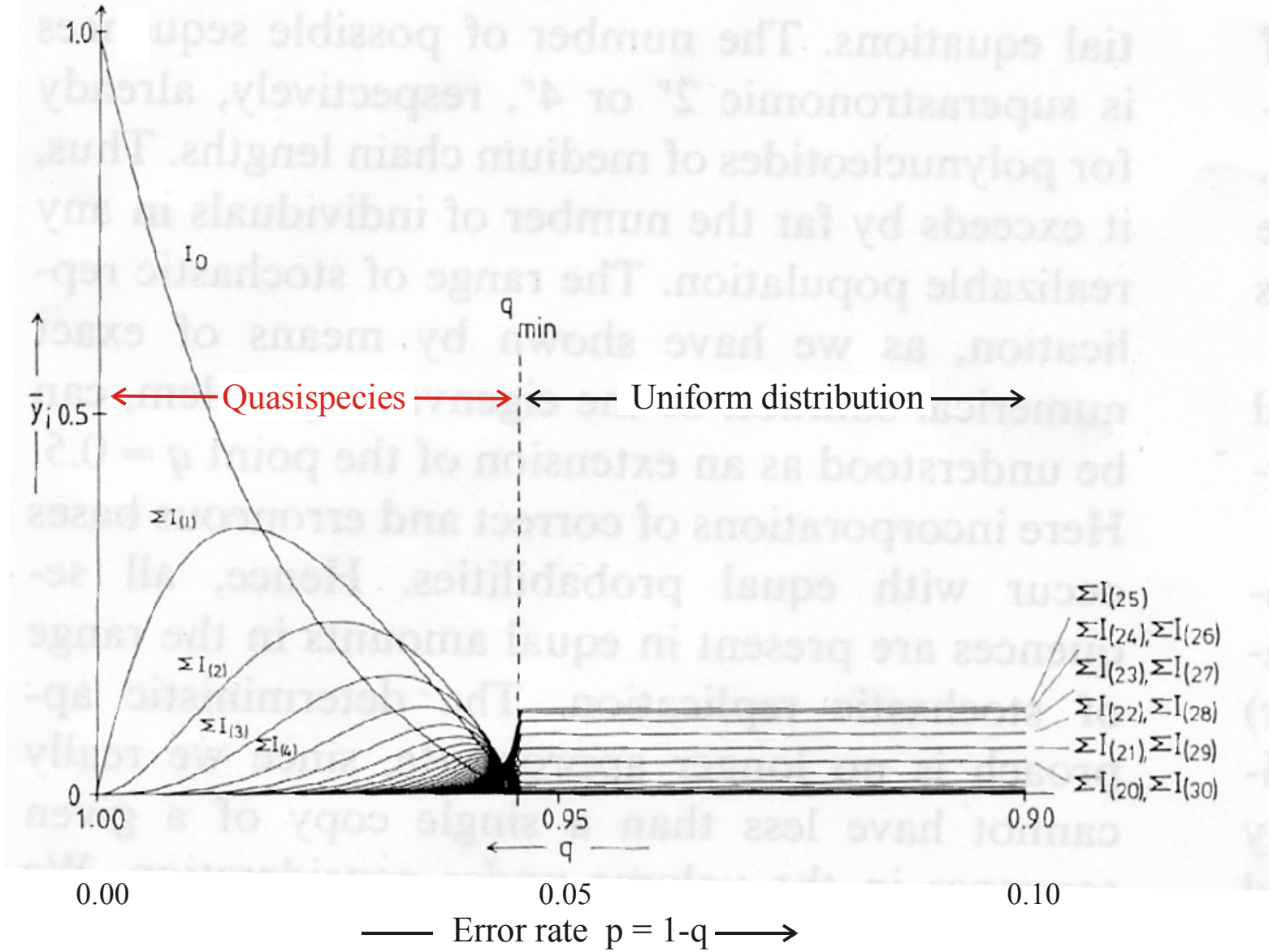
$$Q_{ij} = (1-p)^{l-d(i,j)} p^{d(i,j)}$$

p ..... Error rate per digit

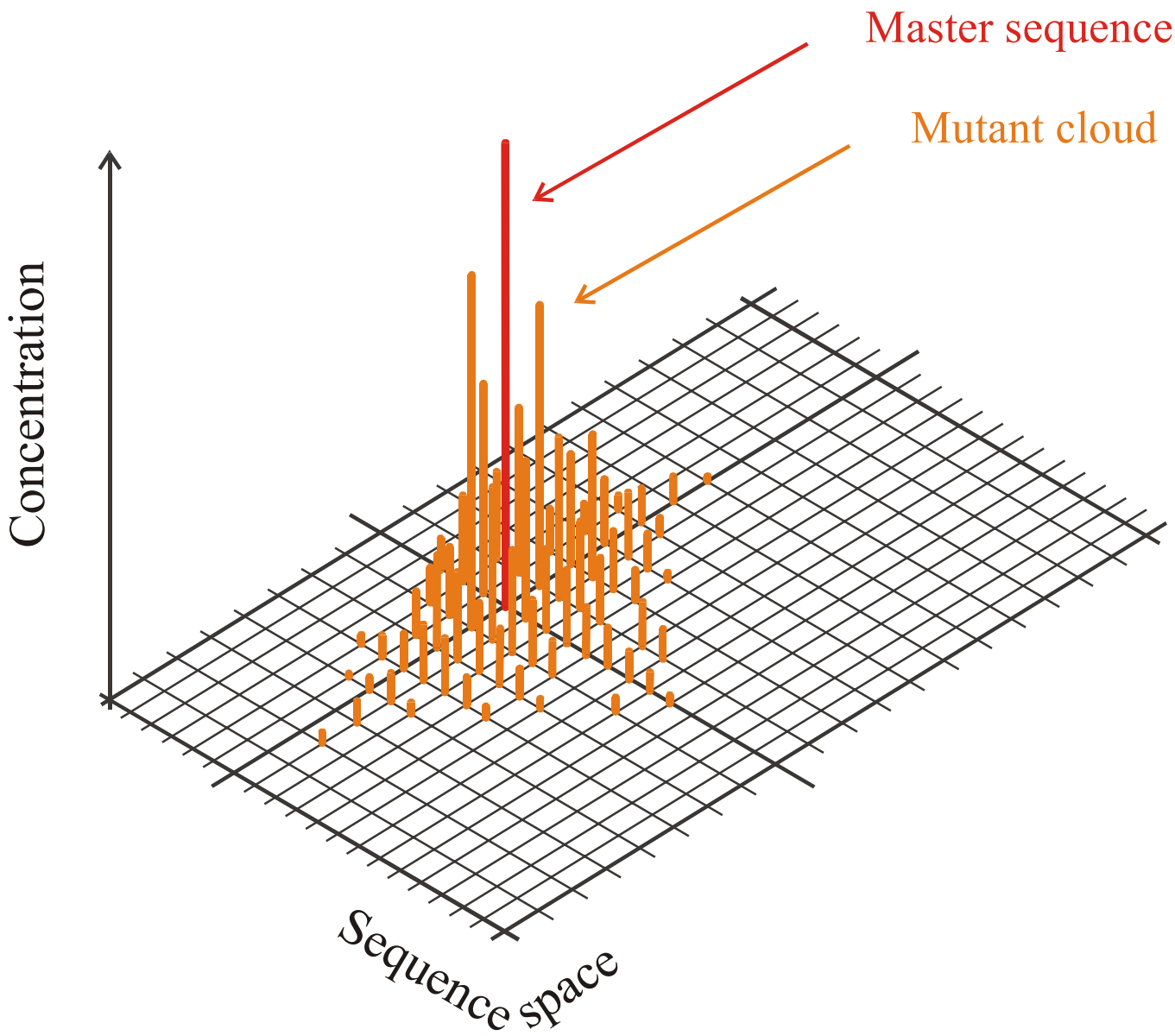
l ..... Chain length of the polynucleotide

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

Chemical kinetics of replication and mutation as parallel reactions



Quasispecies as a function of the replication accuracy q



The molecular quasispecies in sequence space

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

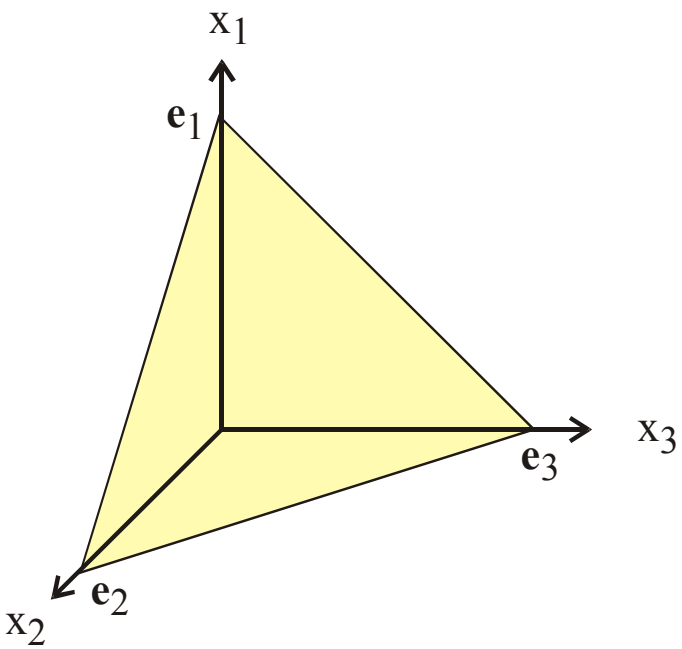
$$\frac{dx_i}{dt} = \sum_{j=1}^{n} f_j Q_{ji}\, x_j - x_i\, \phi, \quad i = 1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

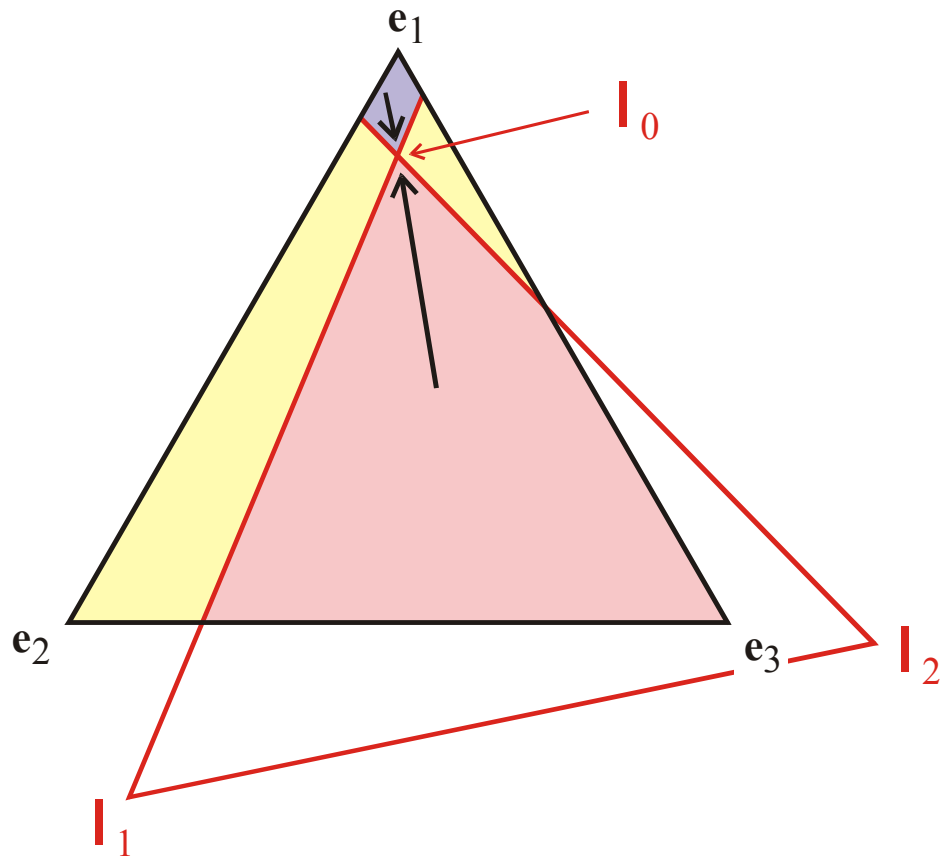
Solutions are obtained after integrating factor transformation by means of an eigenvalue problem

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

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

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





The quasispecies on the concentration simplex $S_3 = \left\{ x_i \geq 0,\ i = 1,2,3;\ \sum_{i=1}^{3} x_i = 1 \right\}$

In the case of non-zero mutation rates ($p>0$ or $q<1$) the Darwinian principle of optimization of mean fitness can be understood only as an **optimization heuristic**. It is valid only on part of the concentration simplex. There are other well defined areas were the mean fitness decreases monotonously or were it may show non-monotonous behavior. The volume of the part of the simplex where mean fitness is non-decreasing in the conventional sense decreases with inreasing mutation rate p.

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

1. Optimization through variation and selection in populations

2. Neutral networks in genotype-phenotype mappings

3. Optimization in the RNA model

4. Evolution experiments with molecules in the laboratory

Theory of genotype – phenotype mapping

P. Schuster, W.Fontana, P.F.Stadler, I.L.Hofacker, ***From sequences to shapes and back: A case study in RNA secondary structures***. Proc.Roy.Soc.London **B 255** (1994), 279-284

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks***. Mh.Chem. **127** (1996), 355-374

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering***. Mh.Chem. **127** (1996), 375-389

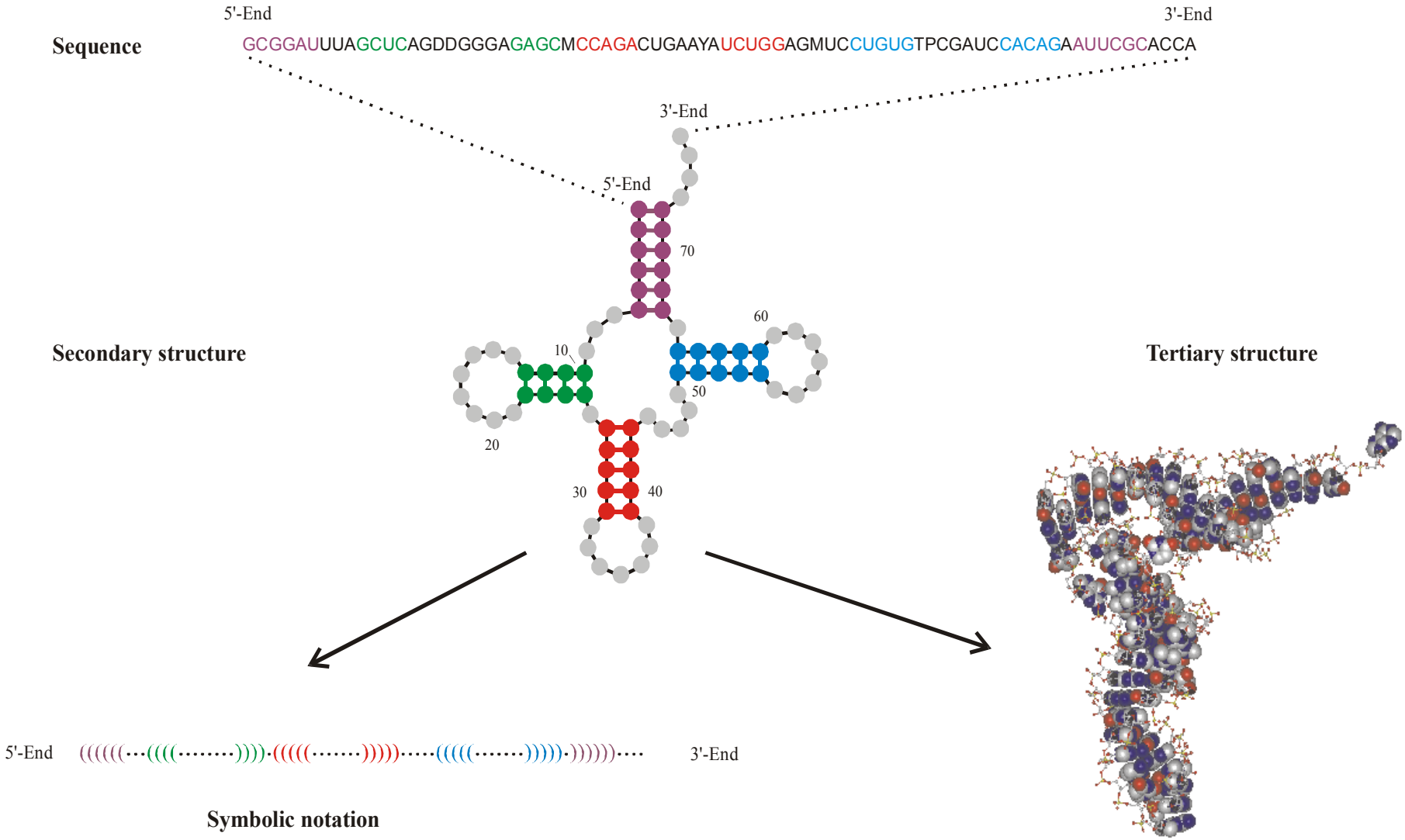
C.M.Reidys, P.F.Stadler, P.Schuster, ***Generic properties of combinatory maps***. Bull.Math.Biol. **59** (1997), 339-397

I.L.Hofacker, P. Schuster, P.F.Stadler, ***Combinatorics of RNA secondary structures***. Discr.Appl.Math. **89** (1998), 177-207

C.M.Reidys, P.F.Stadler, ***Combinatory landscapes***. SIAM Review **44** (2002), 3-54

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

The RNA genotype-phenotype relation is understood as a mapping from the space of RNA sequences into a space of RNA structures.



The **RNA secondary structure** is a listing of **GC**, **AU**, and **GU** base pairs. It is understood in contrast to the full 3D- or **tertiary structure** at the resolution of atomic coordinates. RNA secondary structures are biologically relevant. They are, for example, conserved in evolution.

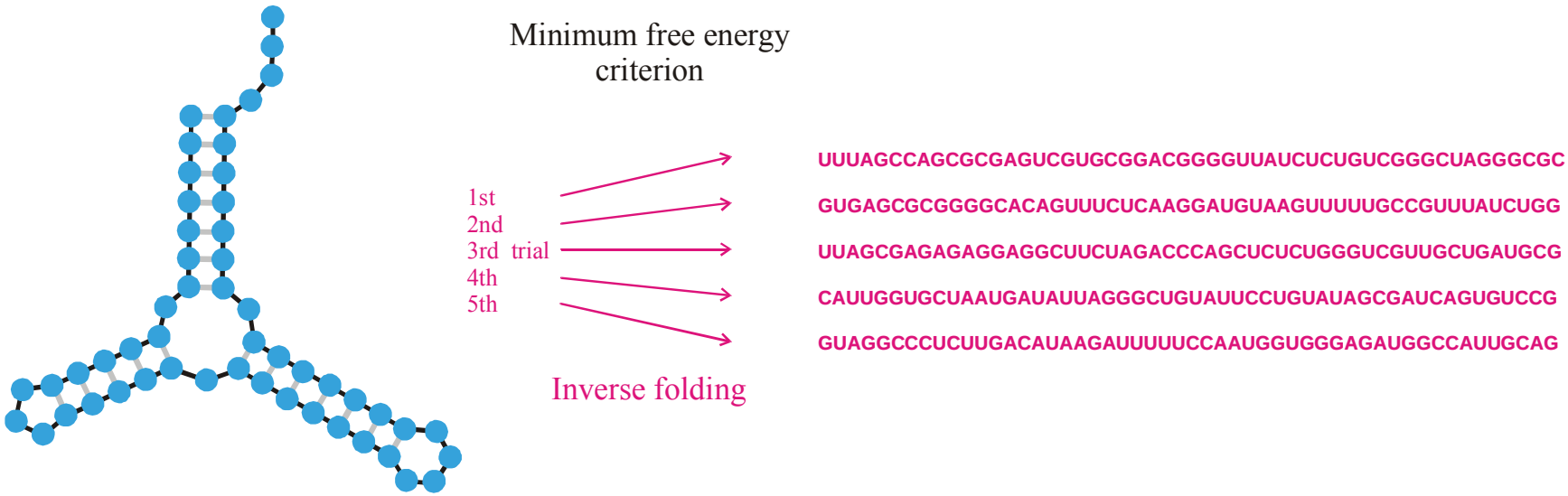
RNA Minimum Free Energy Structures

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

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

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

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



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

Criterion of
Minimum Free Energy

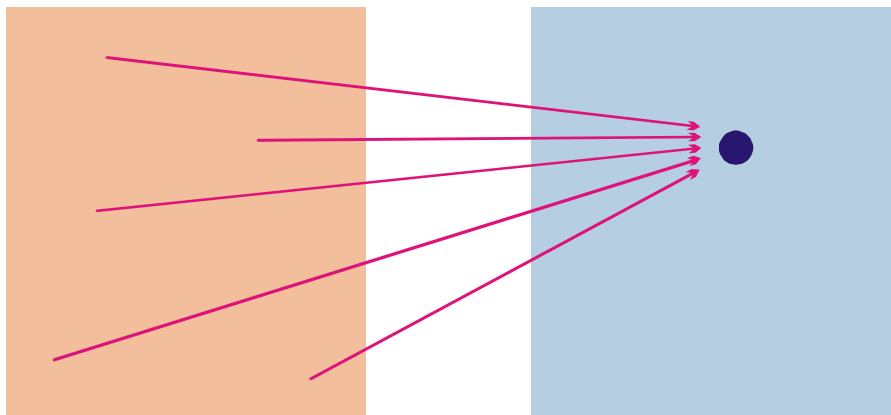
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

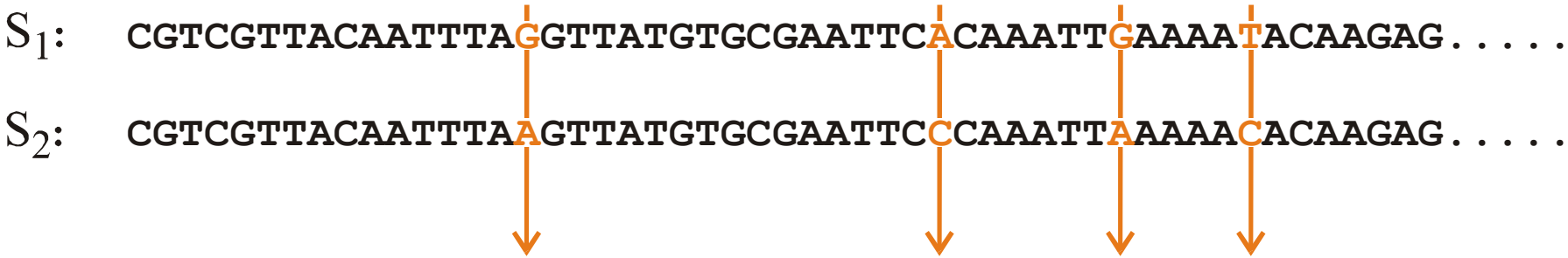
GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG



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

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

Genotypes, i.e. RNA sequences, which are mapped onto the same phenotype, i.e. the same RNA secondary structure, form **neutral networks**. Neutral networks are represented by graphs in sequence space.



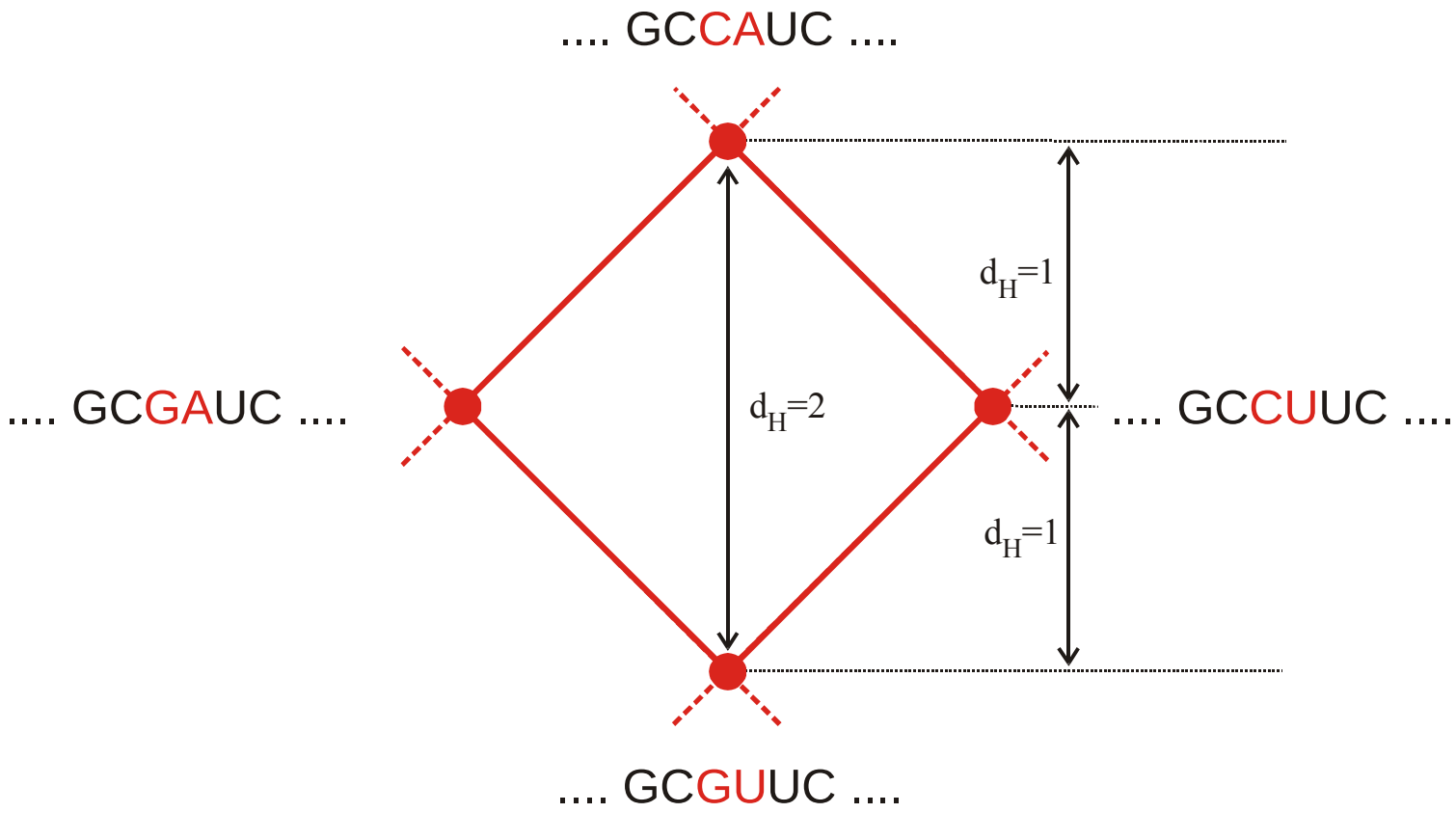
Hamming distance $d_H(S_1,S_2) = 4$

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

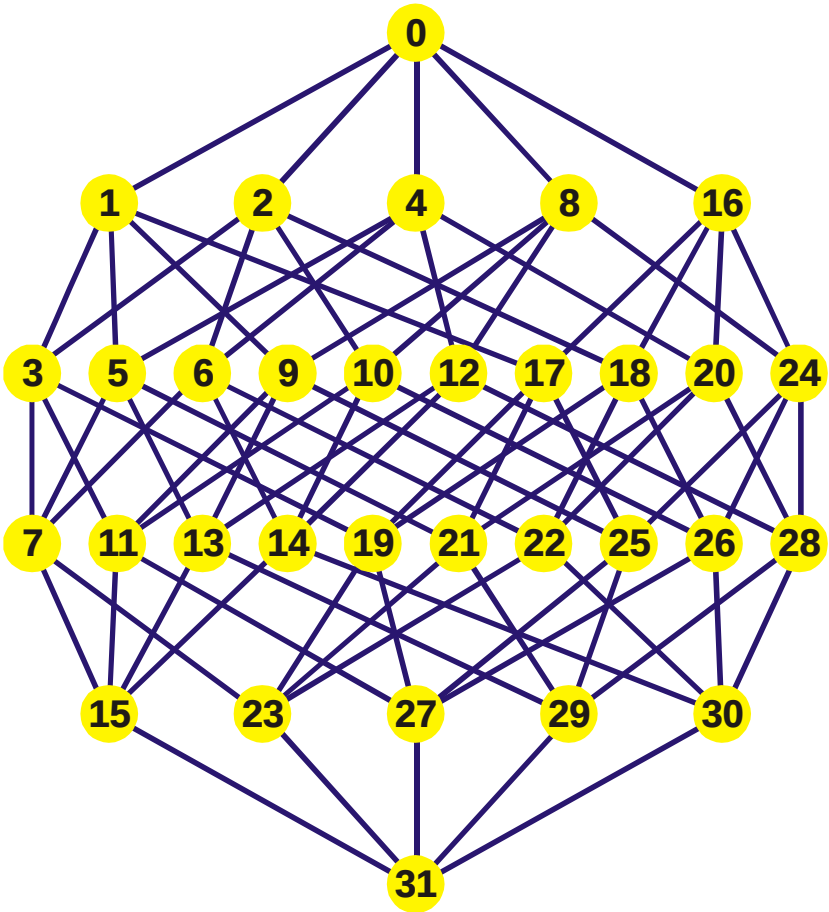
(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

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

The Hamming distance induces a metric in sequence space



Single point mutations as moves in sequence space



Mutant class

0

1

2

3

4

5

Binary sequences are encoded by their decimal equivalents:

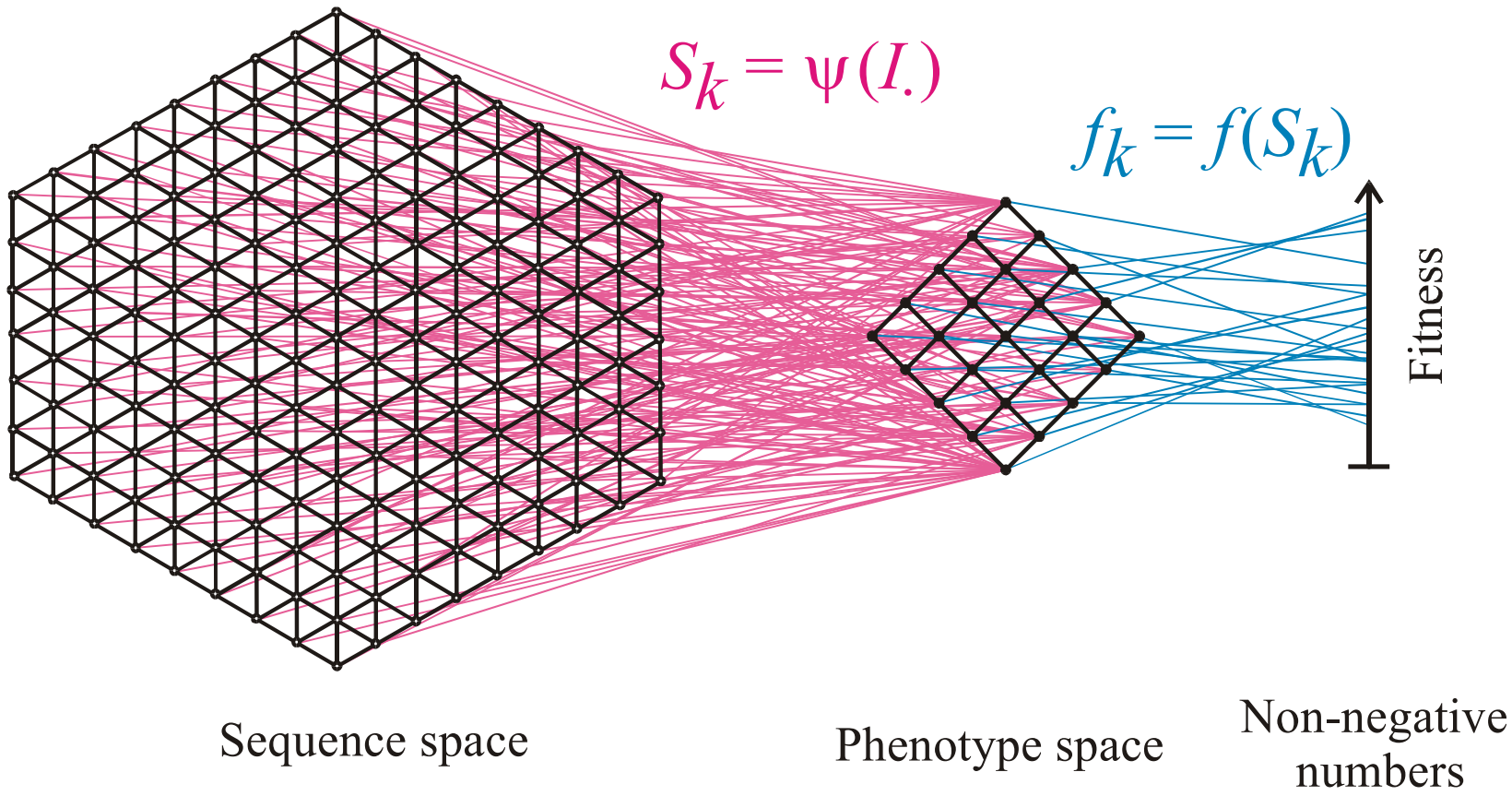
C = 0 and G = 1, for example,

"0" ≡ 00000 = CCCCC,

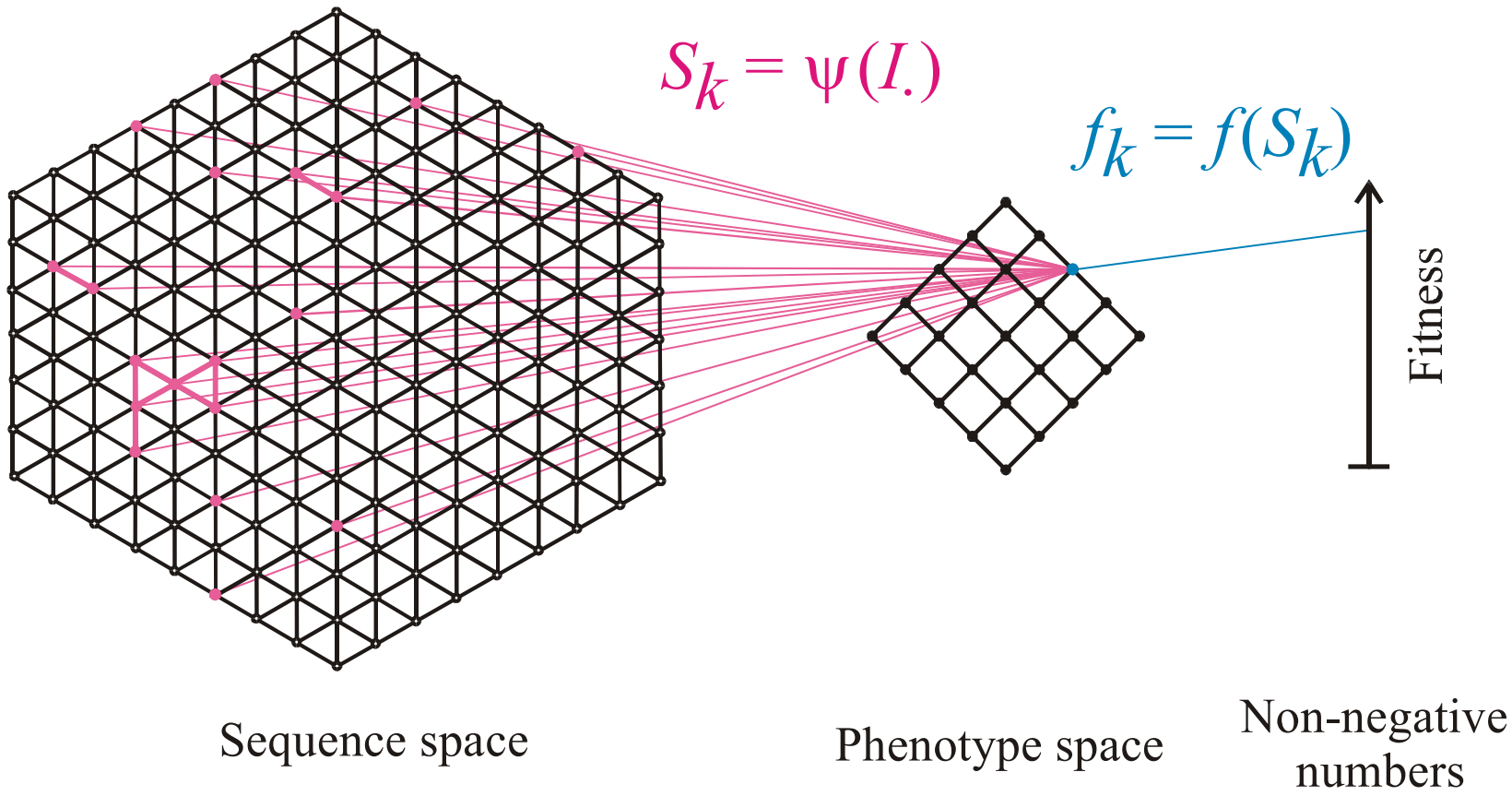
"14" ≡ 01110 = CGGGC,

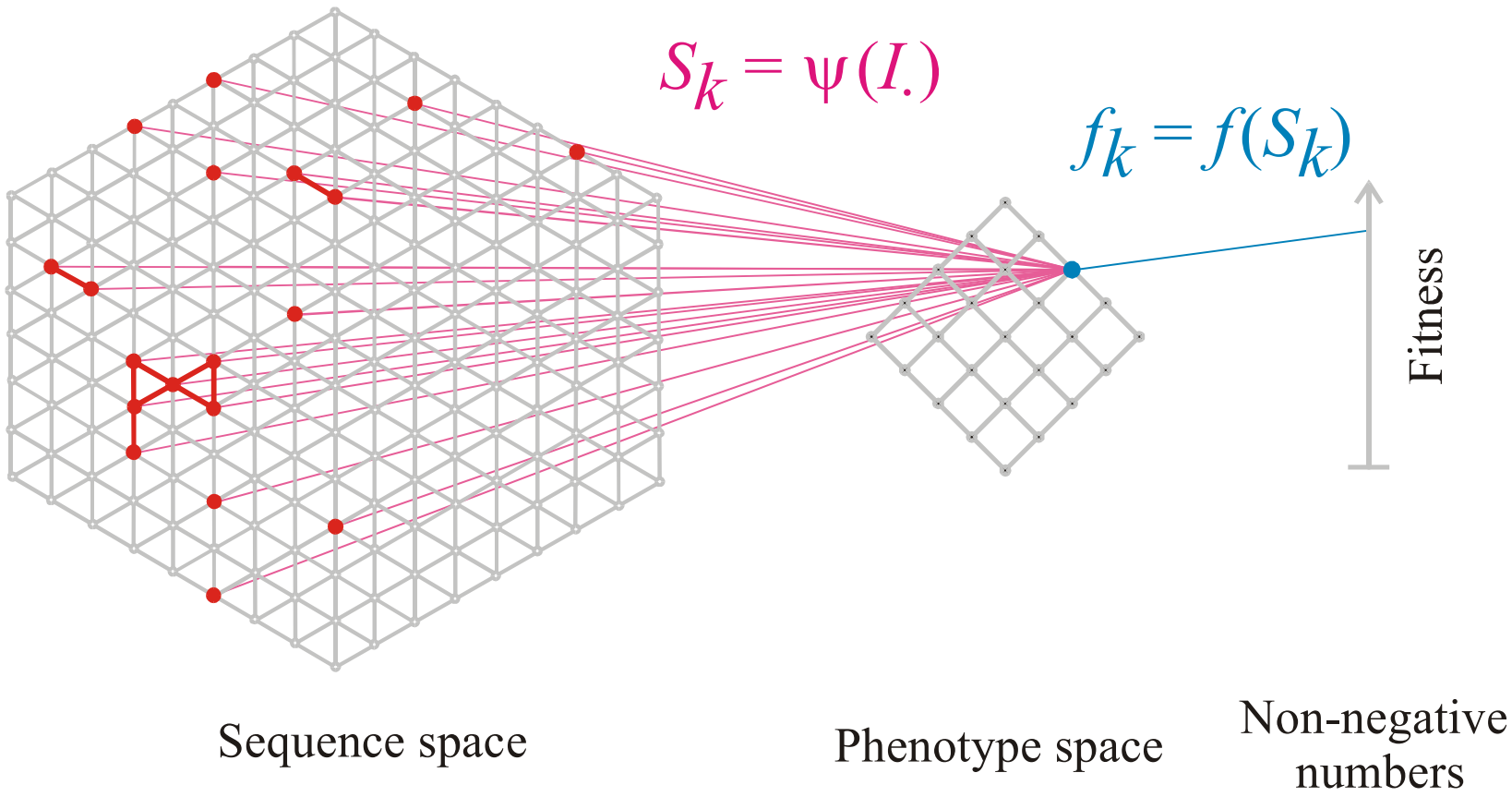
"29" ≡ 11101 = GGGCG, etc.

Sequence space of binary sequences of chain lenght n=5



Mapping from sequence space into phenotype space and into fitness values


$S_k = \psi(I_\cdot)$
$f_k = f(S_k)$
Fitness
Sequence space
Phenotype space
Non-negative 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. $\mathbf{G_k}$ is the pre-image of the structure S_k in sequence space:

$$\mathbf{G_k} = \psi^{-1}(S_k) \equiv \{\psi_j \mid \psi(I_j) = S_k\}$$

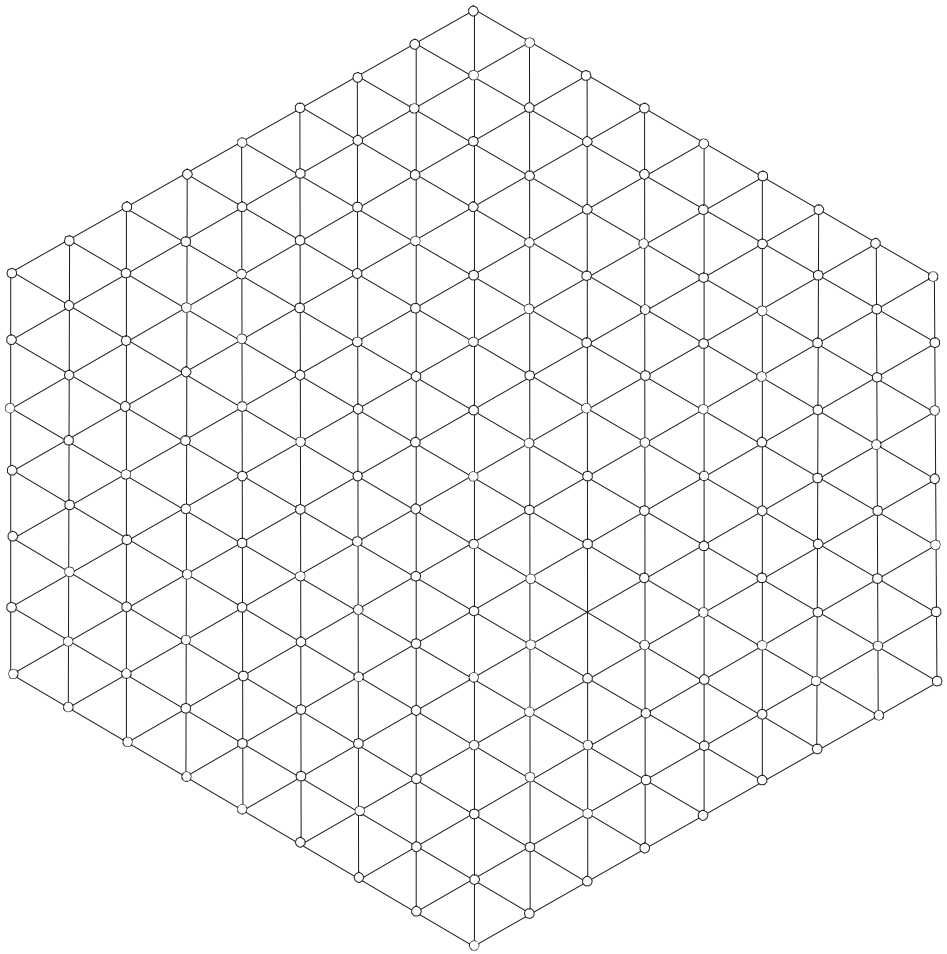
The set is converted into a graph by connecting all sequences of Hamming distance one.

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.

Step 00

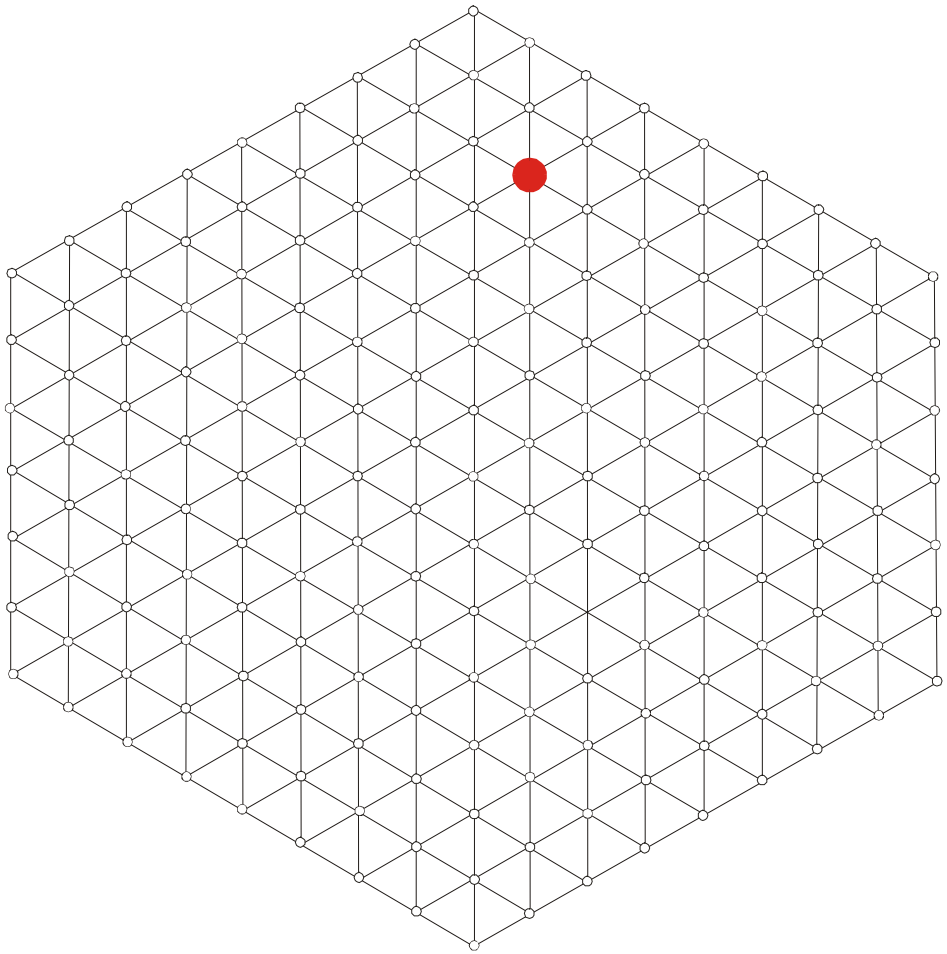
Sketch of sequence space



Random graph approach to neutral networks

Step 01

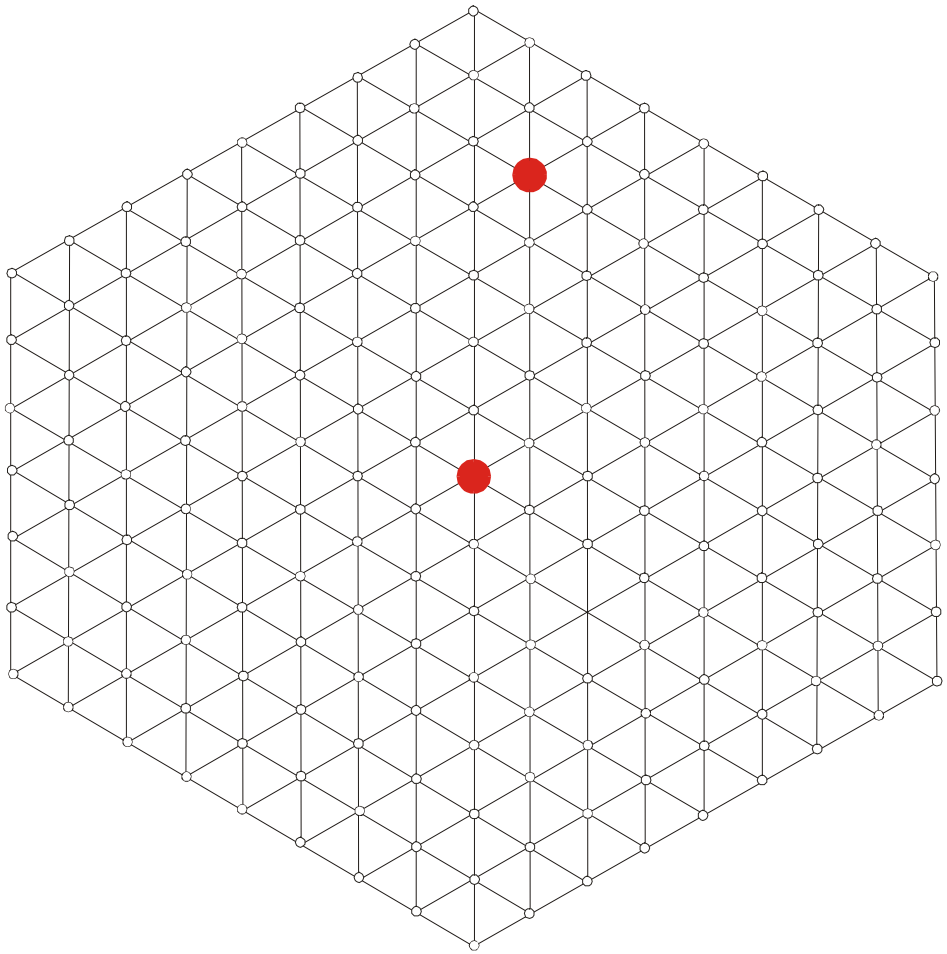
Sketch of sequence space



Random graph approach to neutral networks

Step 02

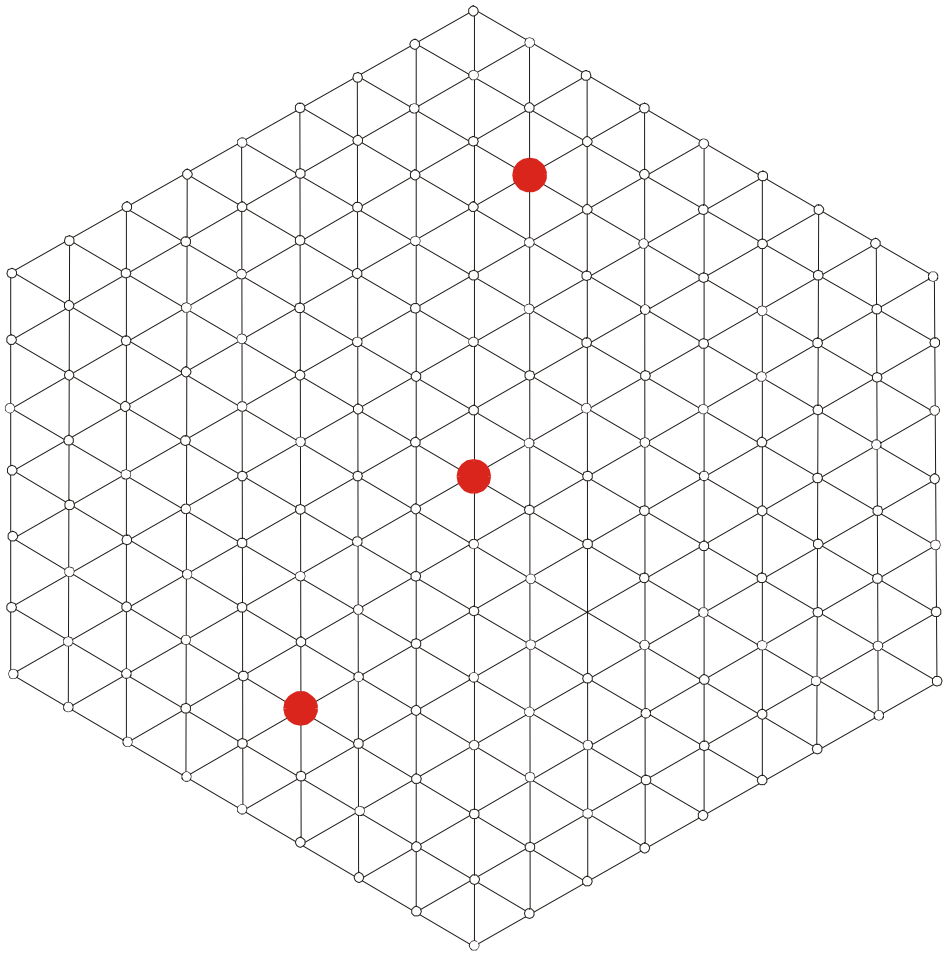
Sketch of sequence space



Random graph approach to neutral networks

Step 03

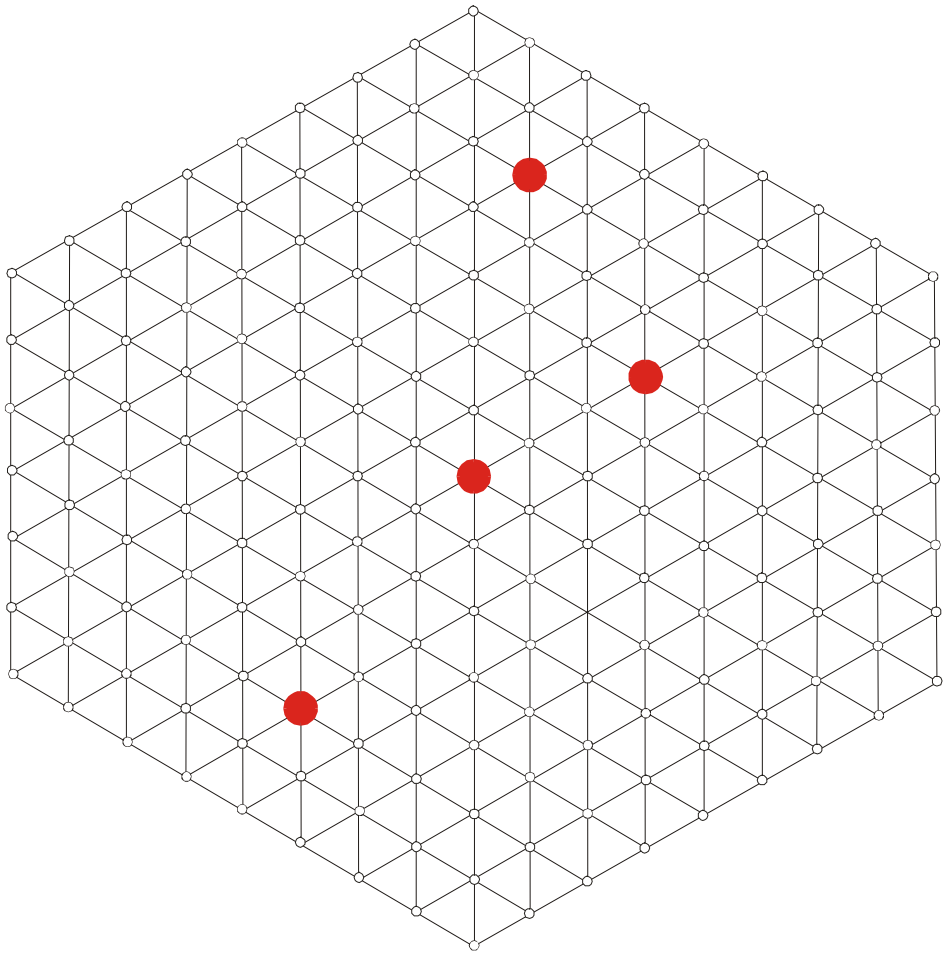
Sketch of sequence space



Random graph approach to neutral networks

Step 04

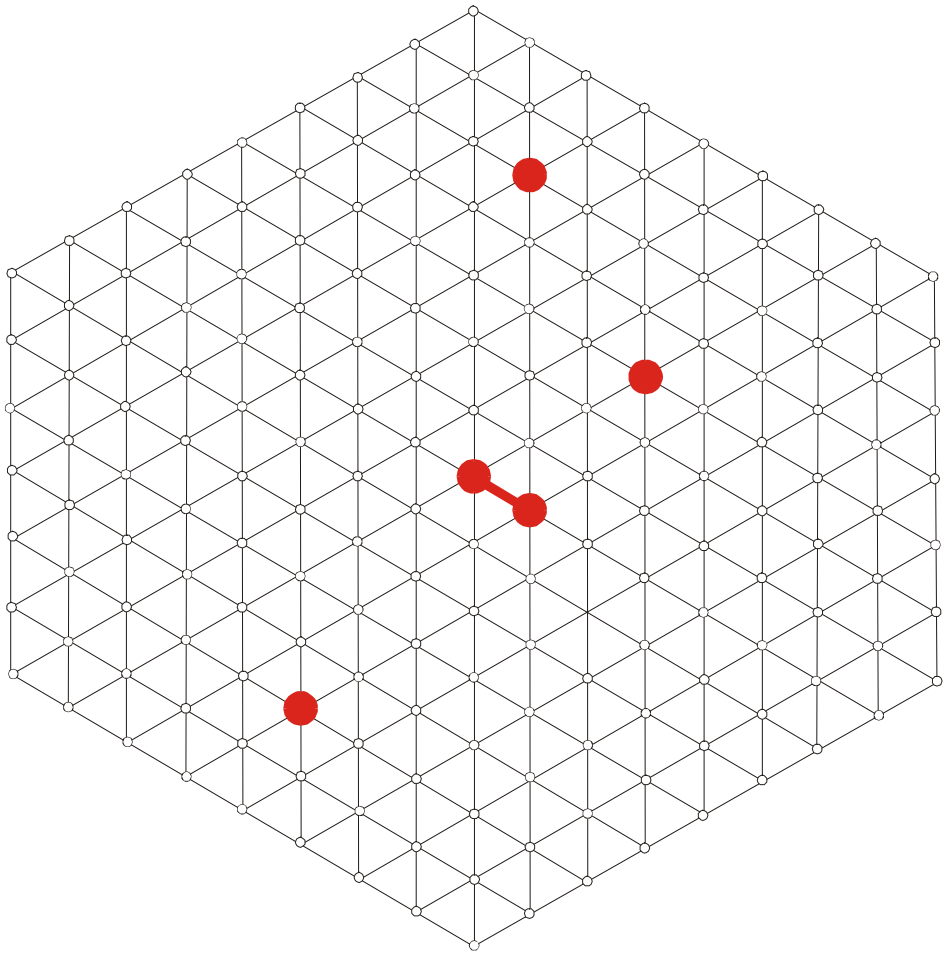
Sketch of sequence space



Random graph approach to neutral networks

Step 05

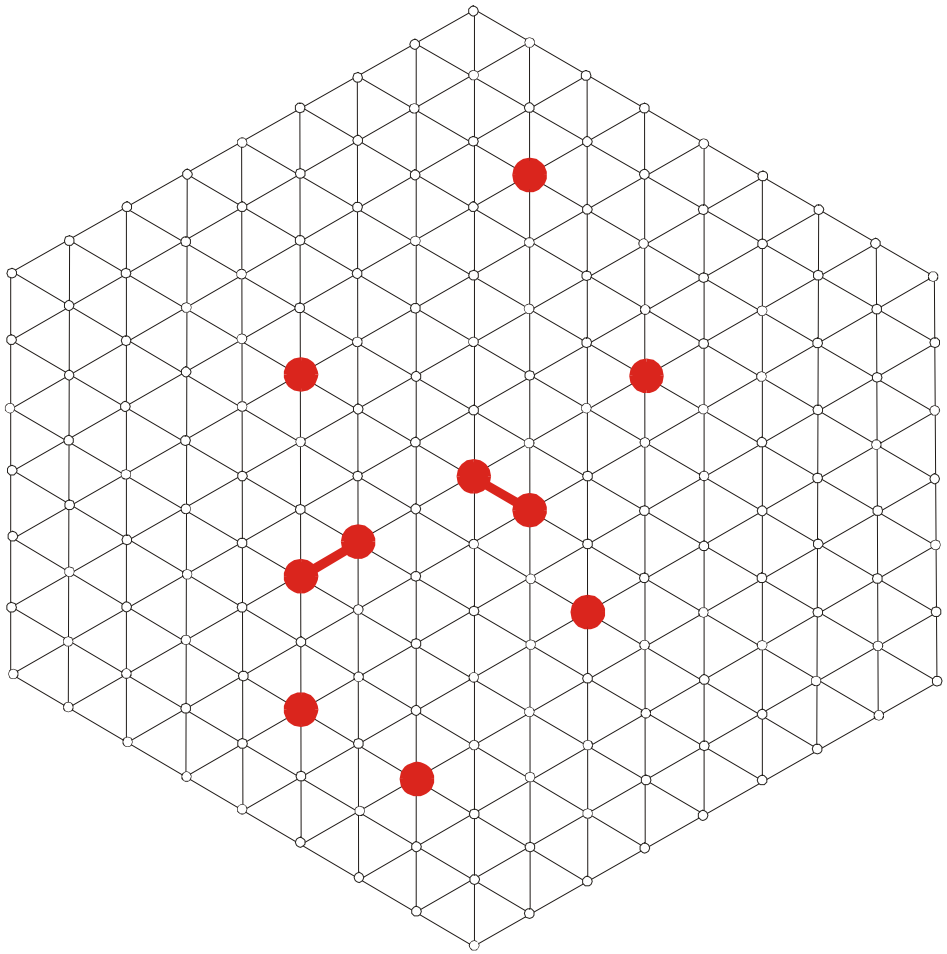
Sketch of sequence space



Random graph approach to neutral networks

Step 10

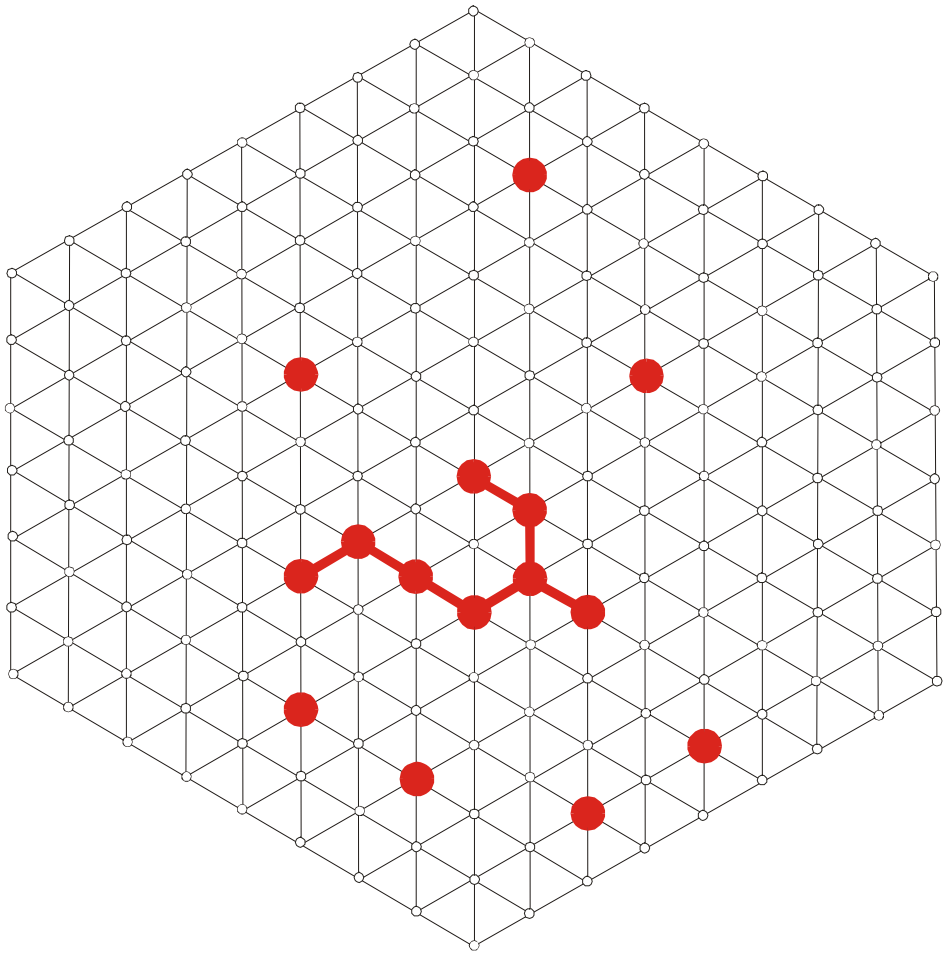
Sketch of sequence space



Random graph approach to neutral networks

Step 15

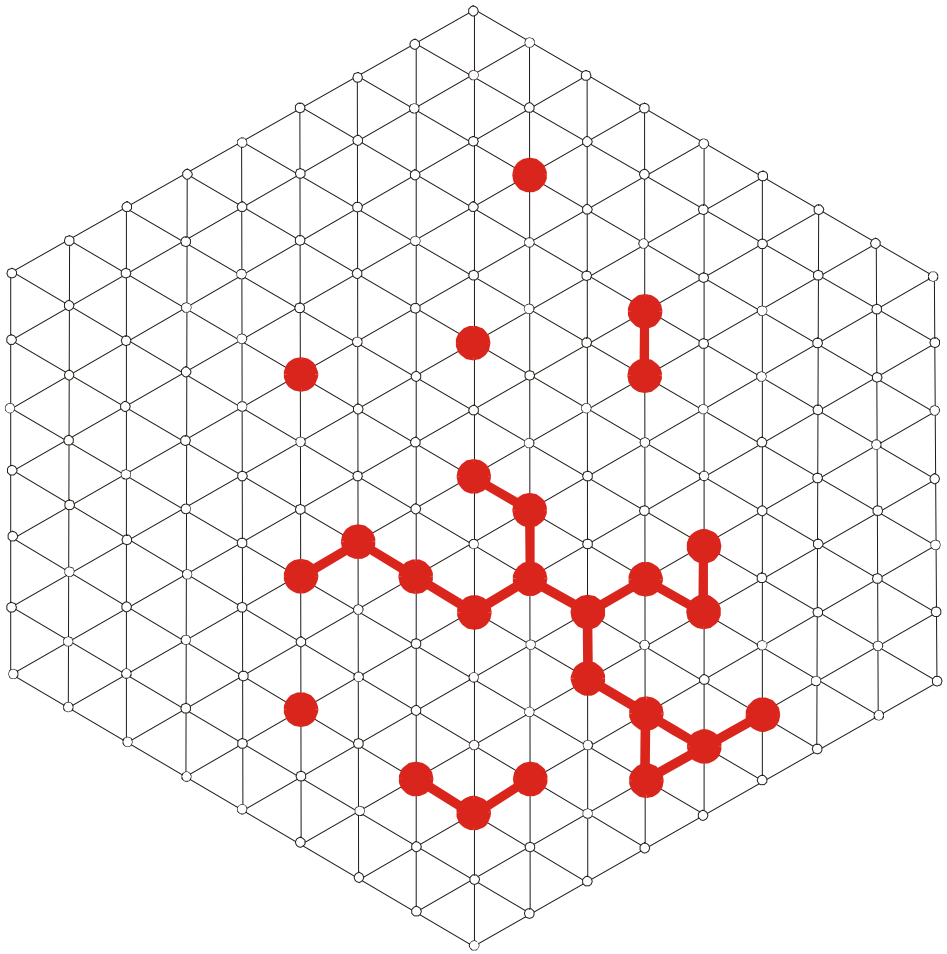
Sketch of sequence space



Random graph approach to neutral networks

Step 25

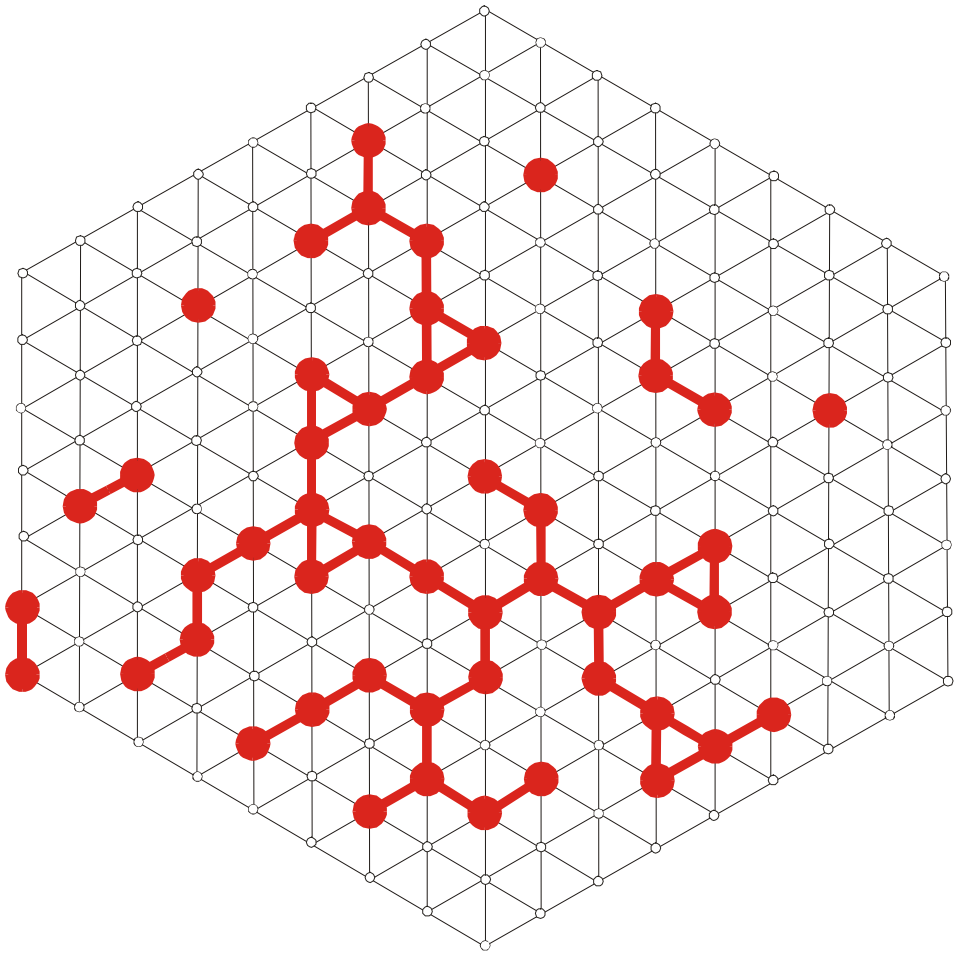
Sketch of sequence space



Random graph approach to neutral networks

Step 50

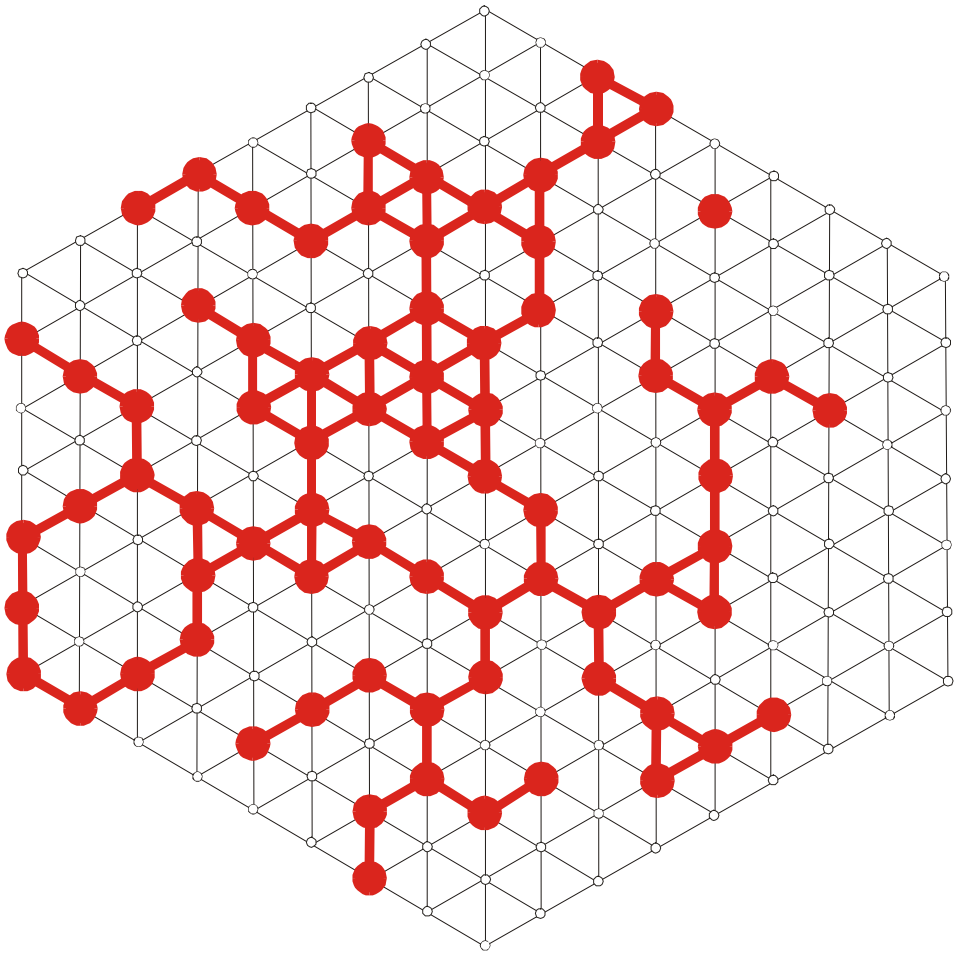
Sketch of sequence space



Random graph approach to neutral networks

Step 75

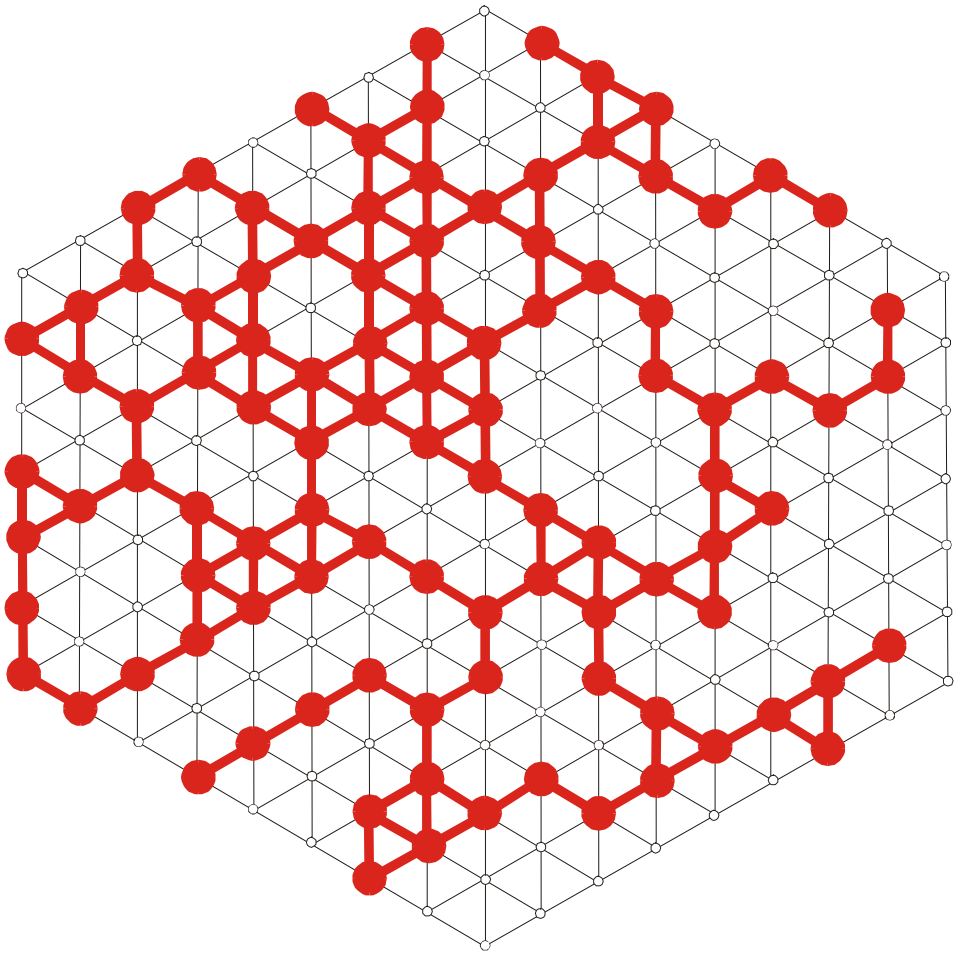
Sketch of sequence space



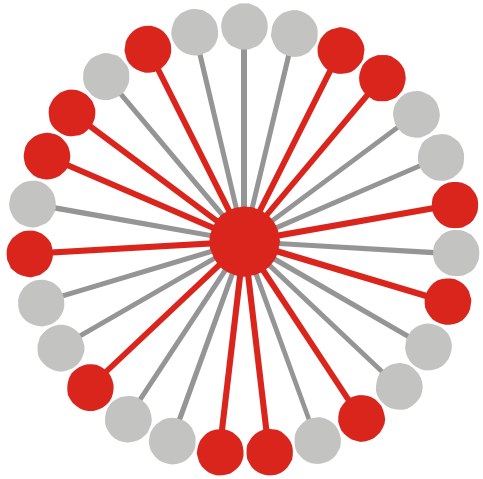
Random graph approach to neutral networks

Step 100

Sketch of sequence space



Random graph approach to neutral networks



$$G_k = y^{-1}(S_k) \cup \{ I_j \mid y(I_j) = S_k \}$$

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

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

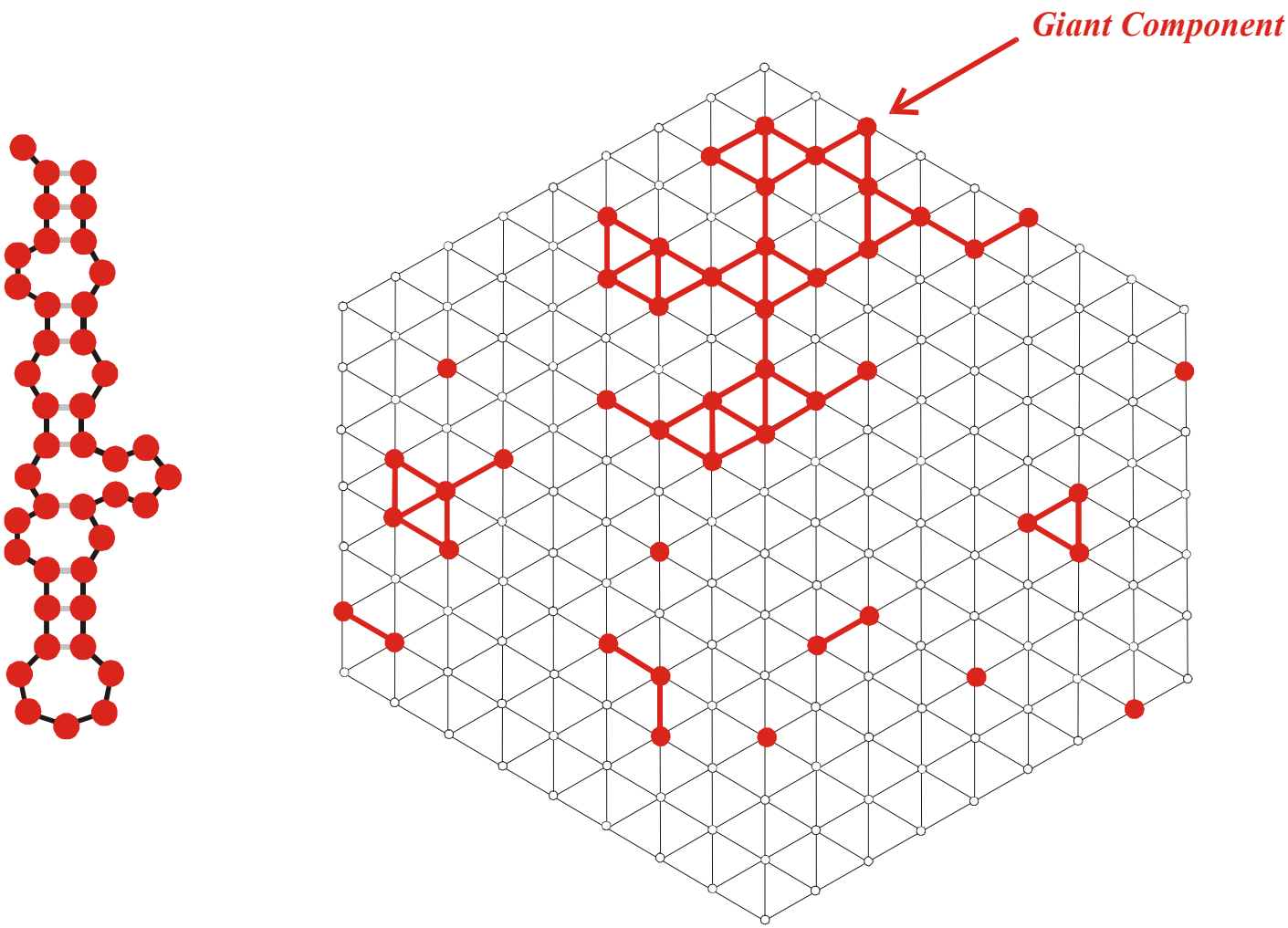
Alphabet size k : **AUGC** ñ k = 4

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

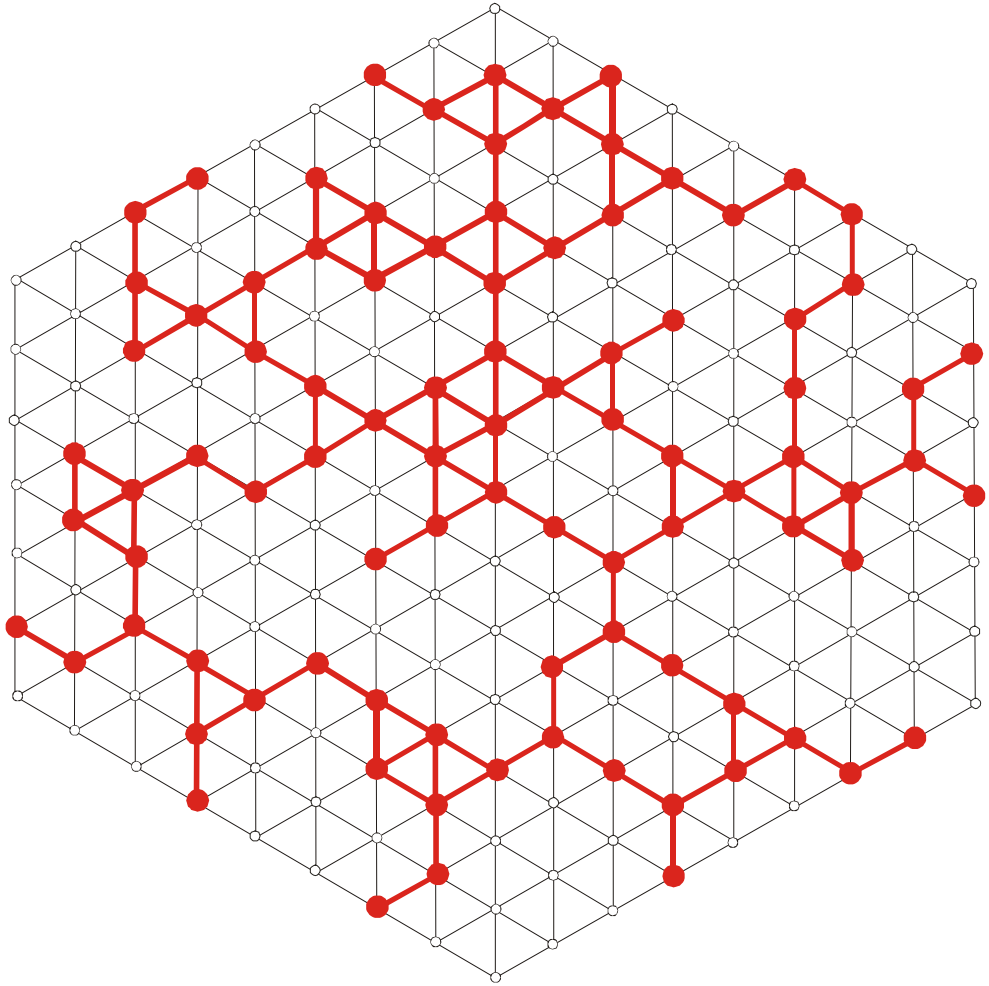
$\bar{\lambda}_k < \lambda_{cr}$ . . . . network G_k is **not** connected

k	l_{cr}
2	0.5
3	0.4226
4	0.3700

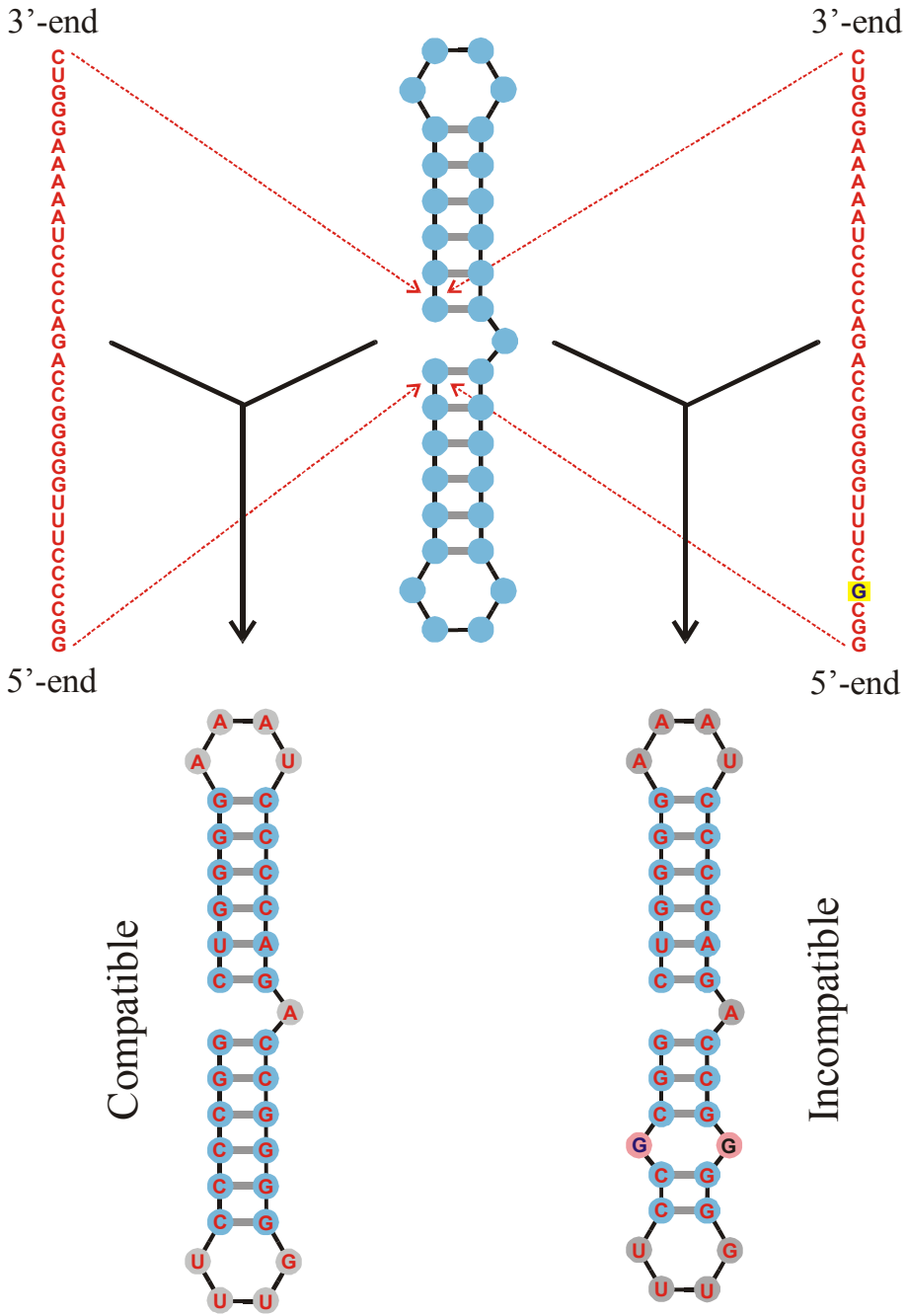
Mean degree of neutrality and connectivity of **neutral networks**



A multi-component neutral network



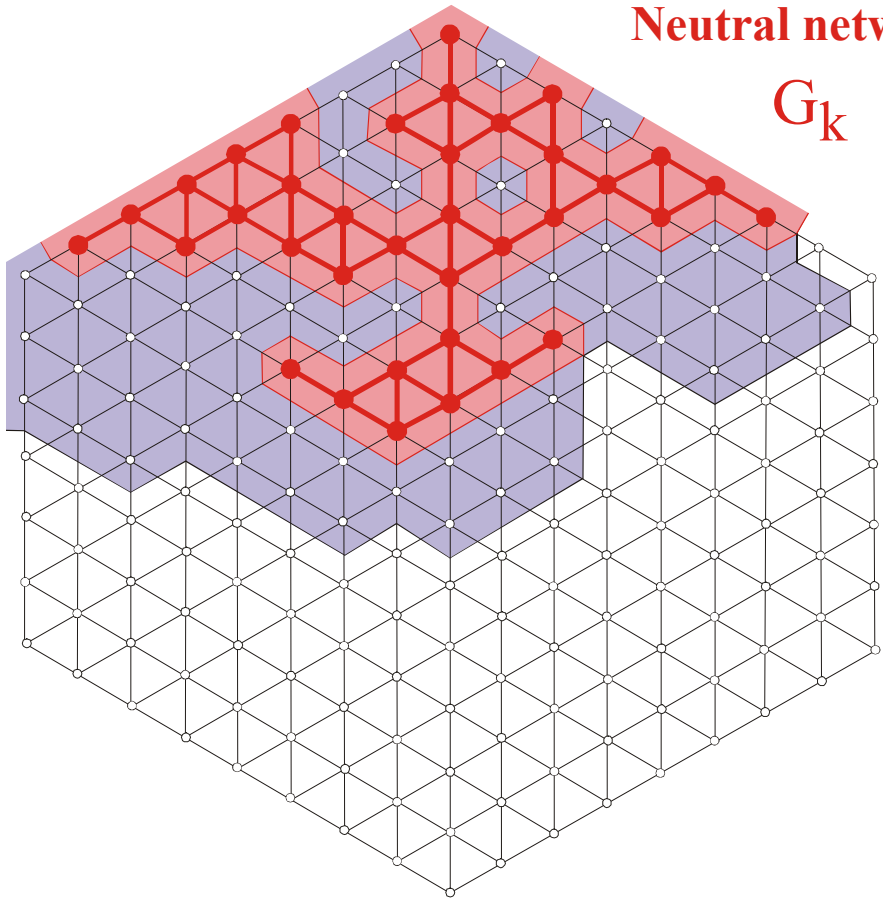
A connected neutral network



Compatibility of sequences with structures

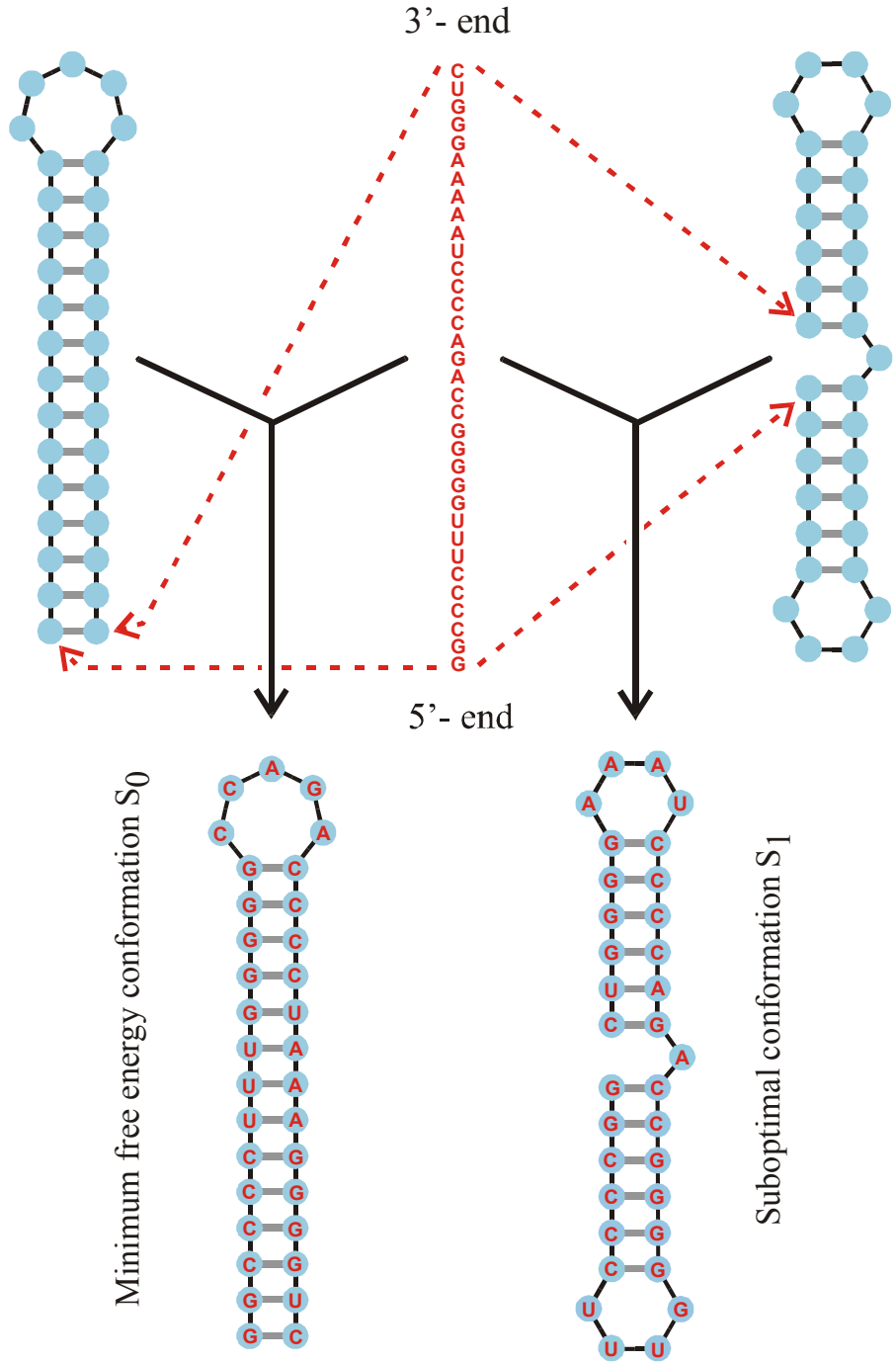
A sequence is compatible with its minimum free energy structure and all its suboptimal structures.

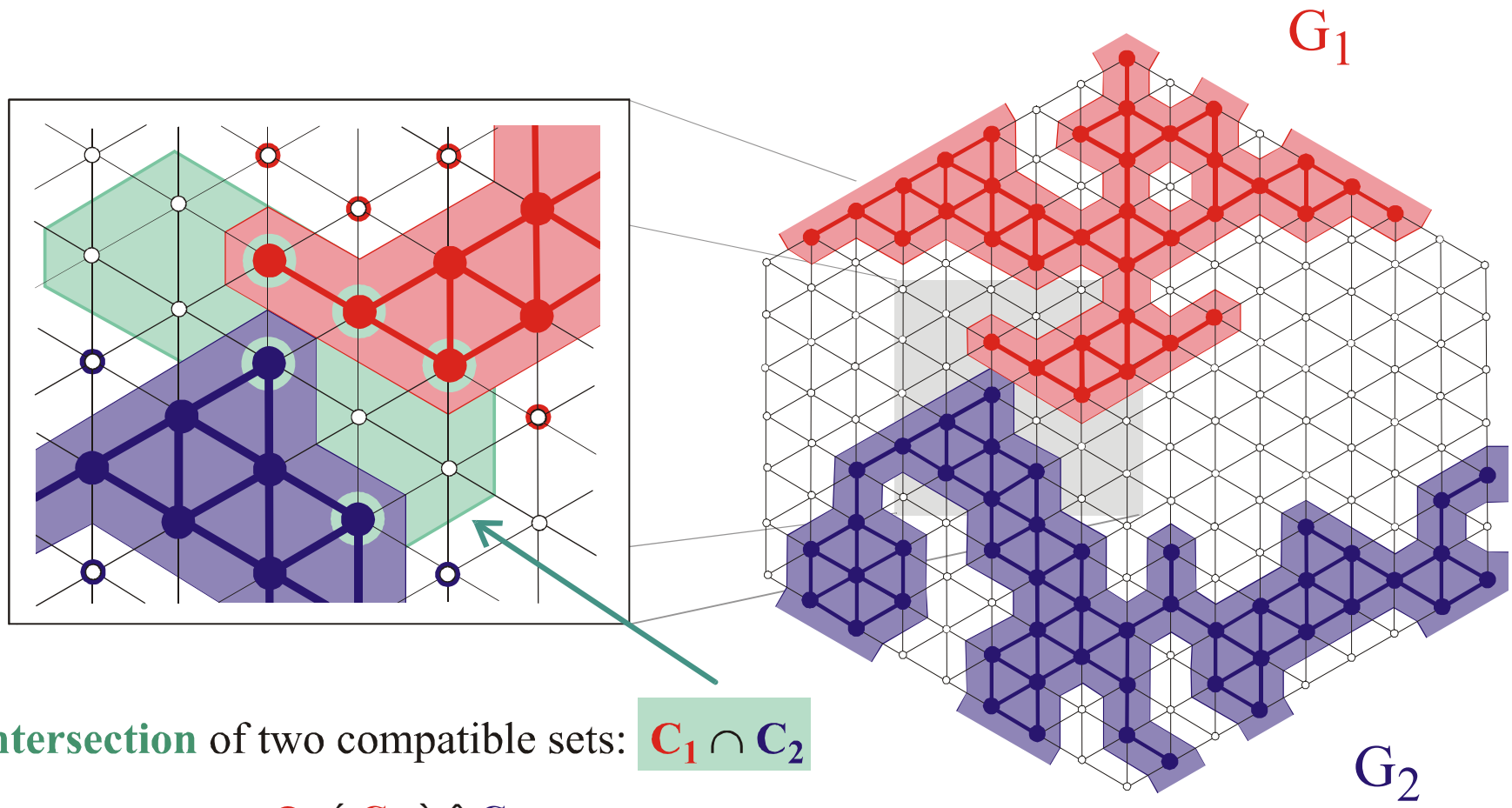
G_k à C_k



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

A sequence at the **intersection** of two neutral networks is compatible with both structures





Intersection of two compatible sets: $C_1 \cap C_2$

: á C_1 ì â C_2

: â C_1 ì á C_2

The intersection of two compatible sets is always non empty: C_1 Ú C_2 â Ù

1. Optimization through variation and selection in populations

2. Neutral networks in genotype-phenotype mappings

3. **Optimization** in the RNA model

4. Evolution experiments with molecules in the laboratory

Optimization of RNA molecules *in silico*

W.Fontana, P.Schuster, ***A computer model of evolutionary optimization***. Biophysical Chemistry **26** (1987), 123-147

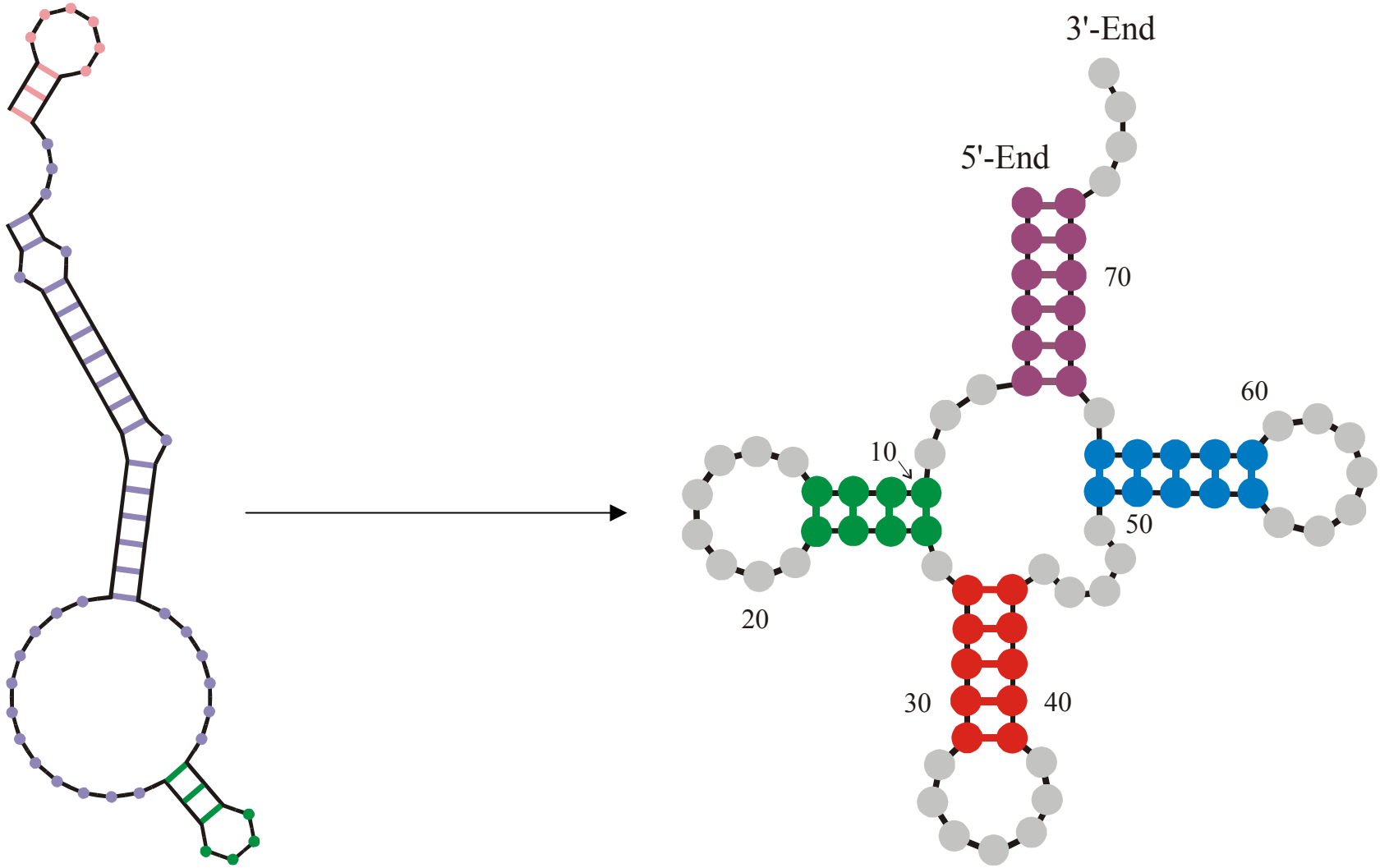
W.Fontana, W.Schnabl, P.Schuster, ***Physical aspects of evolutionary optimization and adaptation***. Phys.Rev.A **40** (1989), 3301-3321

M.A.Huynen, W.Fontana, P.F.Stadler, ***Smoothness within ruggedness. The role of neutrality in adaptation***. Proc.Natl.Acad.Sci.USA **93** (1996), 397-401

W.Fontana, P.Schuster, ***Continuity in evolution. On the nature of transitions***. Science **280** (1998), 1451-1455

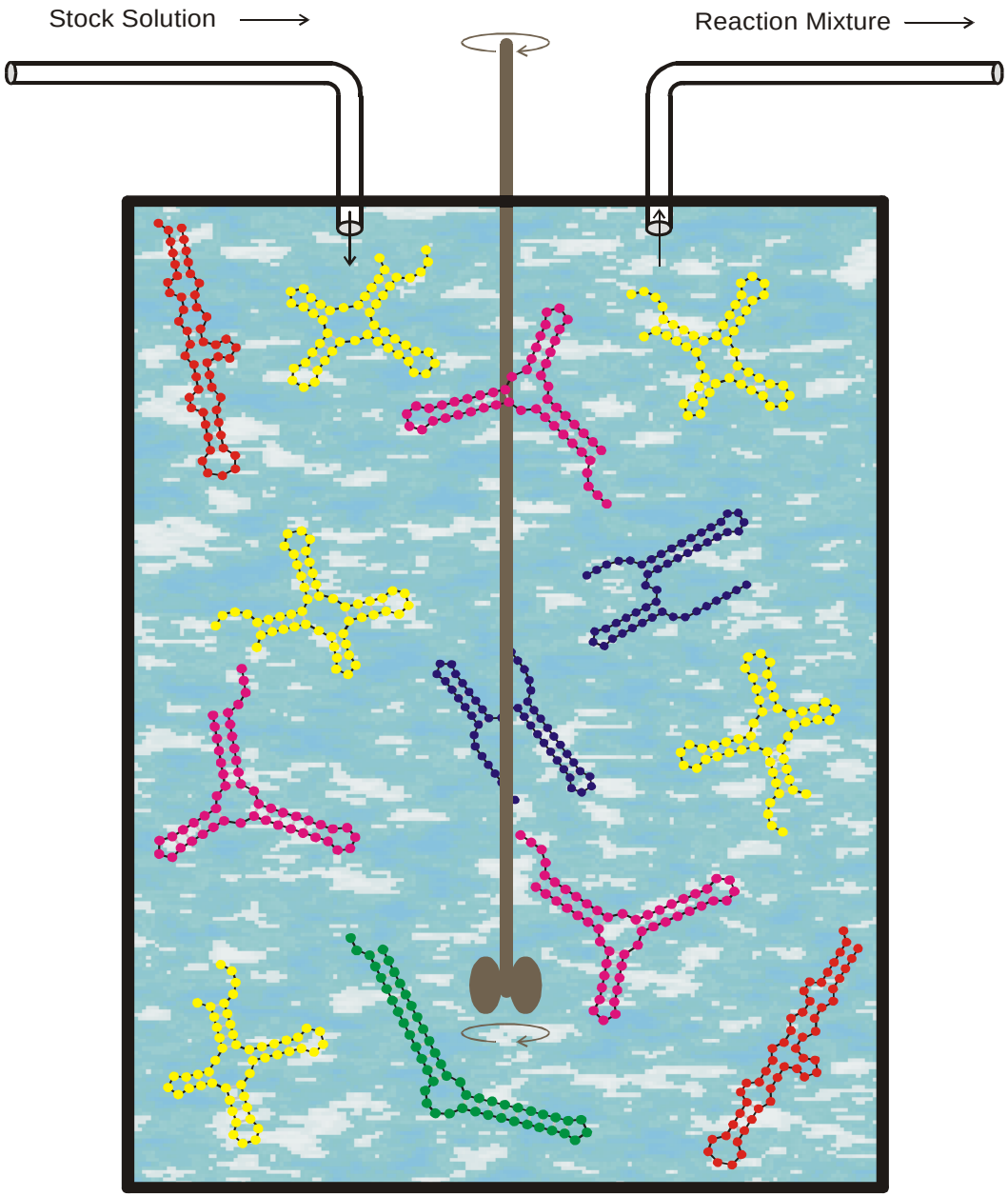
W.Fontana, P.Schuster, ***Shaping space. The possible and the attainable in RNA genotype-phenotype mapping***. J.Theor.Biol. **194** (1998), 491-515

B.M.R. Stadler, P.F. Stadler, G.P. Wagner, W. Fontana, ***The topology of the possible: Formal spaces underlying patterns of evolutionary change.*** J.Theor.Biol. **213** (2001), 241-274



Randomly chosen
initial structure

Phenylalanyl-tRNA as
target structure

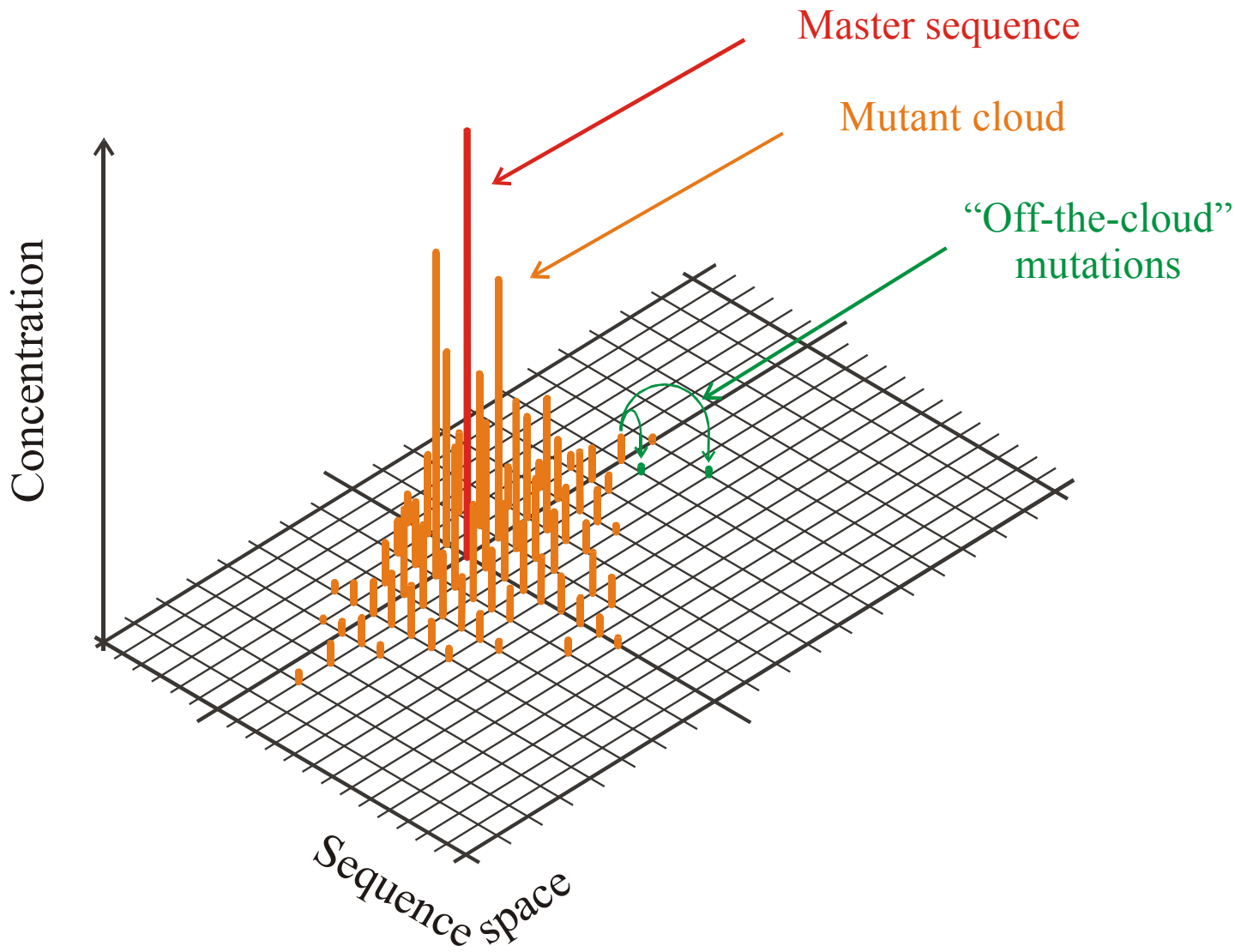


Fitness function:

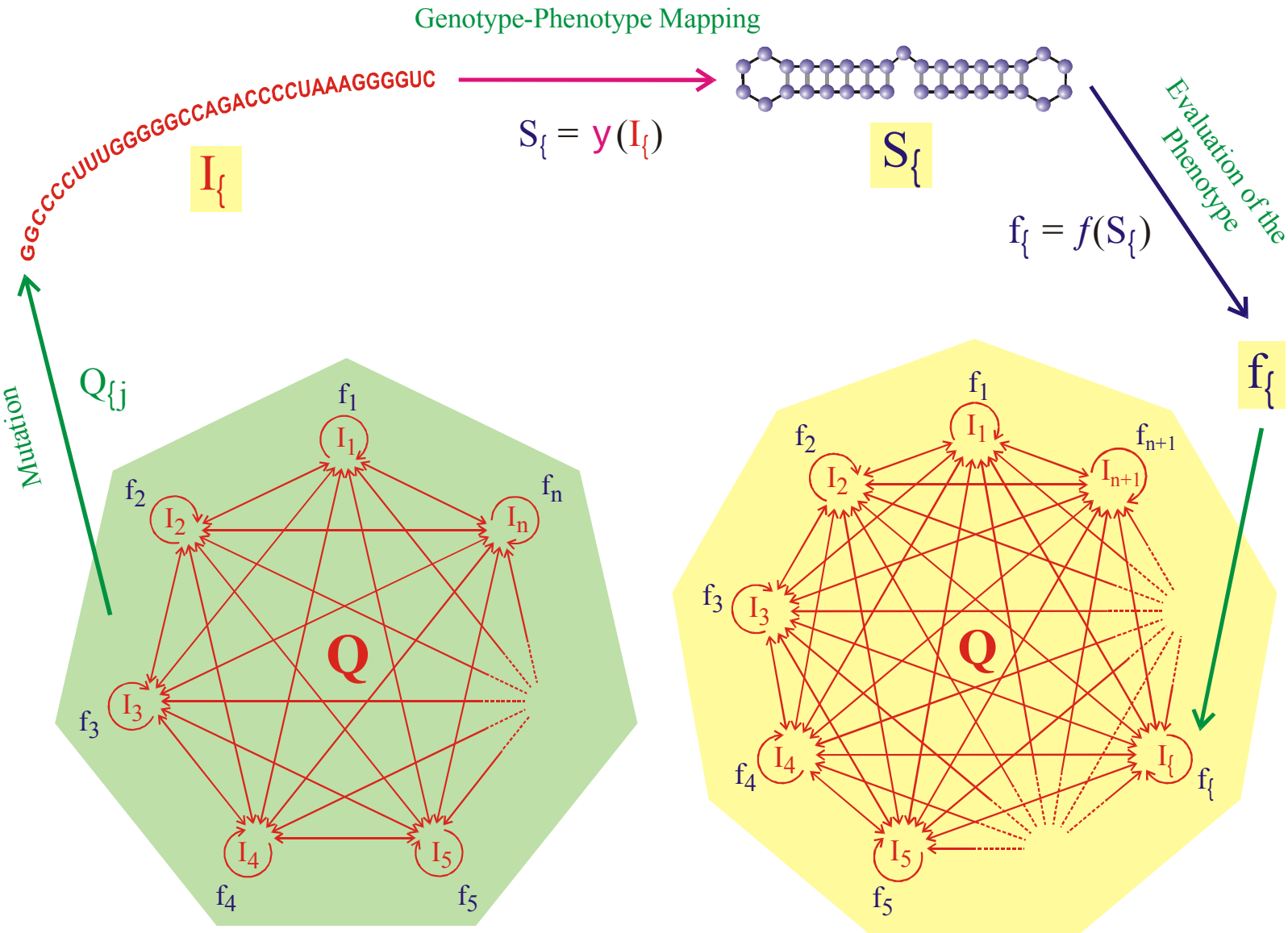
$f_k = g / [a + Dd_S^{(k)}]$

$Dd_S^{(k)} = d^s(I_k, I_t)$

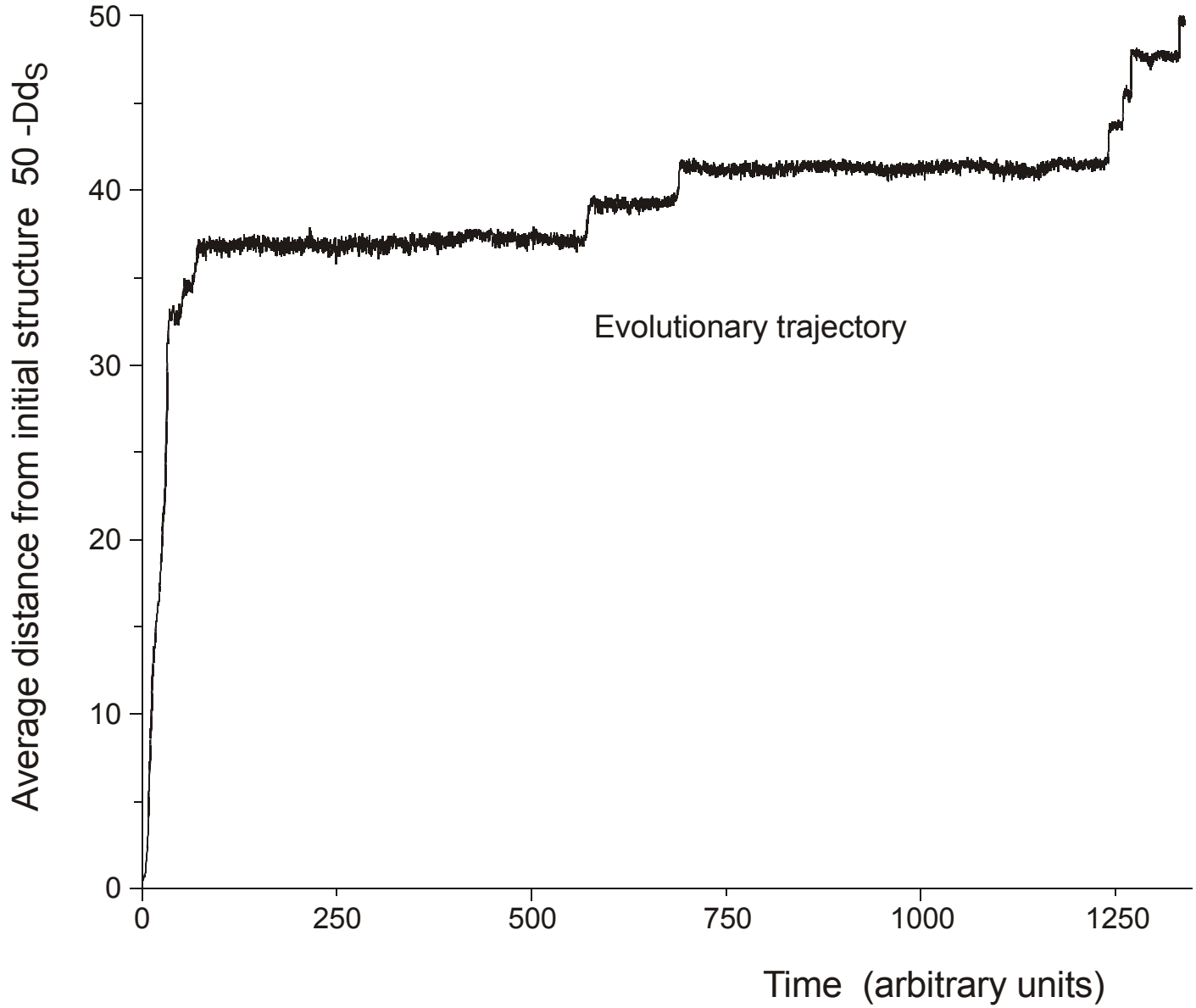
The flowreactor as a device for studies of evolution ***in vitro*** and ***in silico***



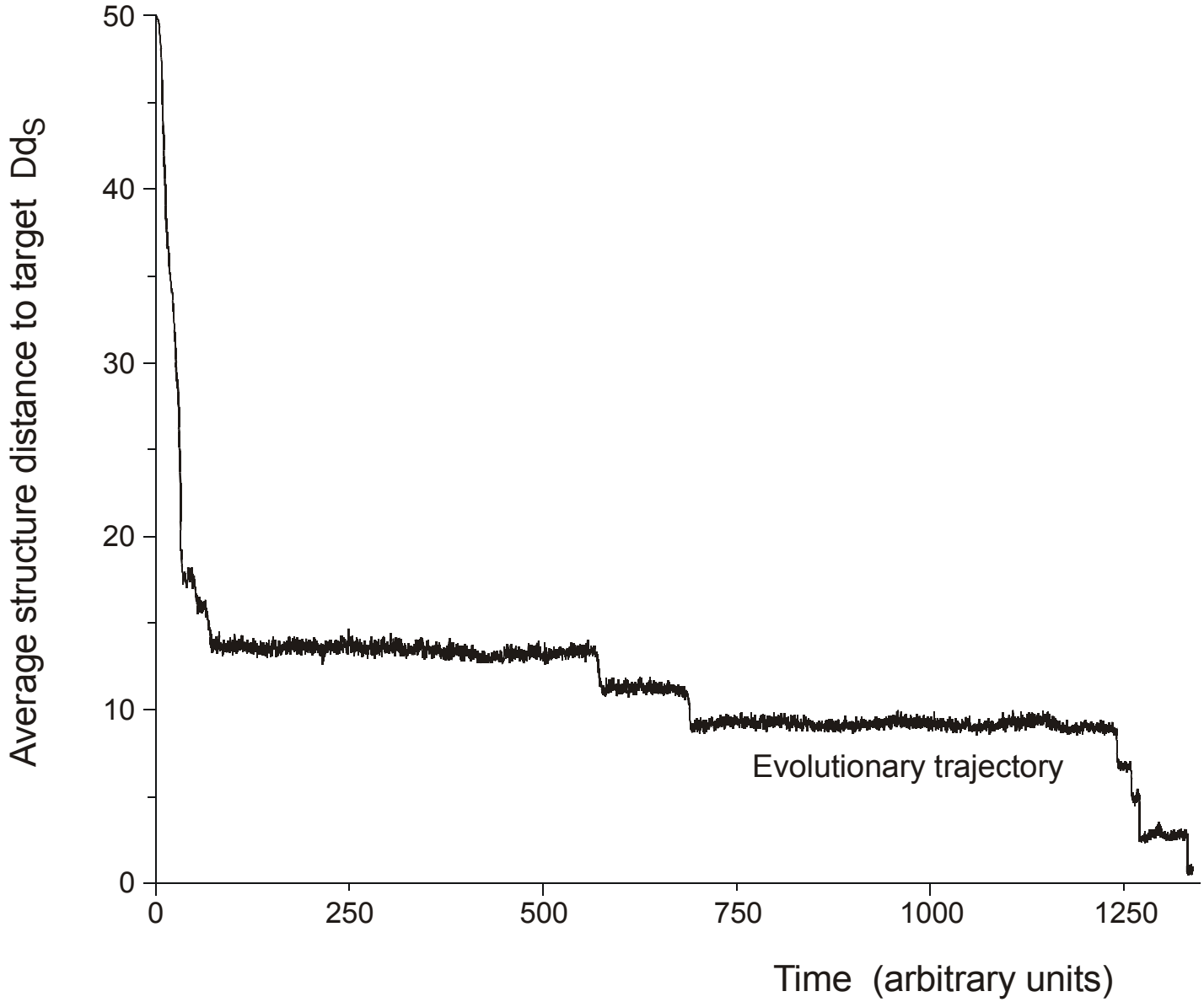
The molecular quasispecies in sequence space



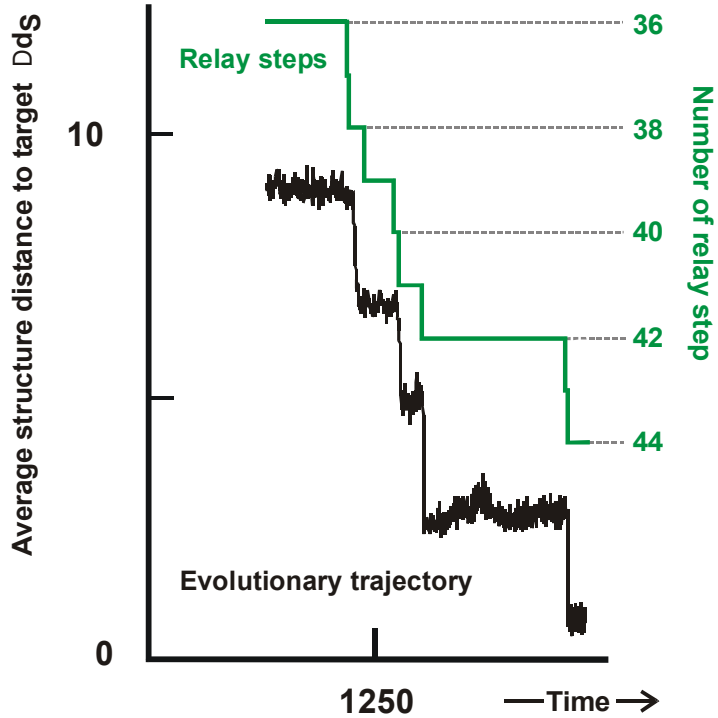
Evolutionary dynamics
including molecular phenotypes

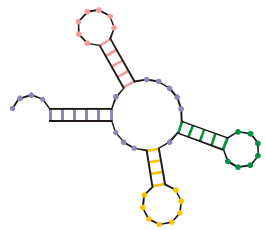


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)



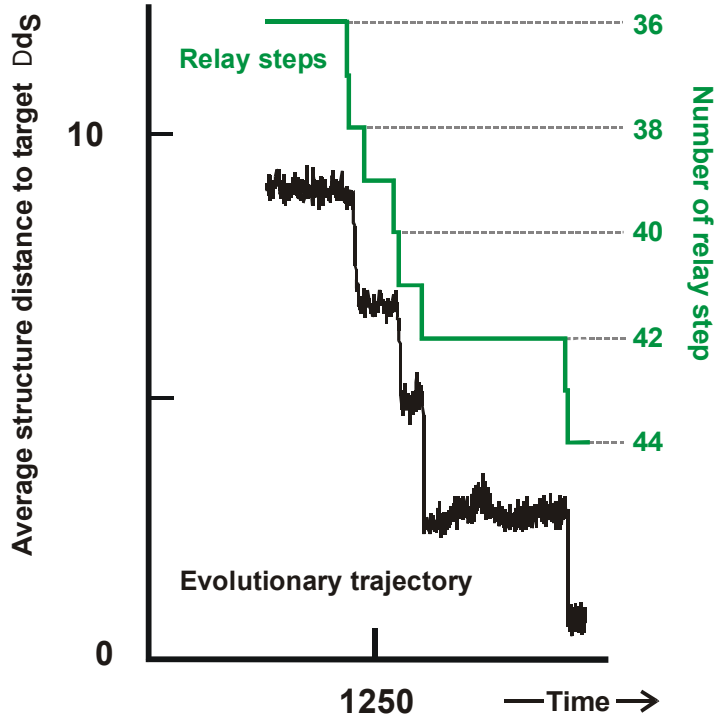
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)

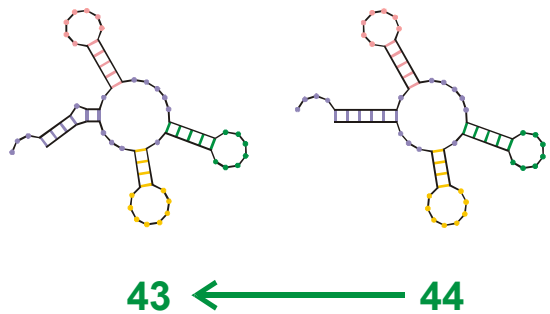




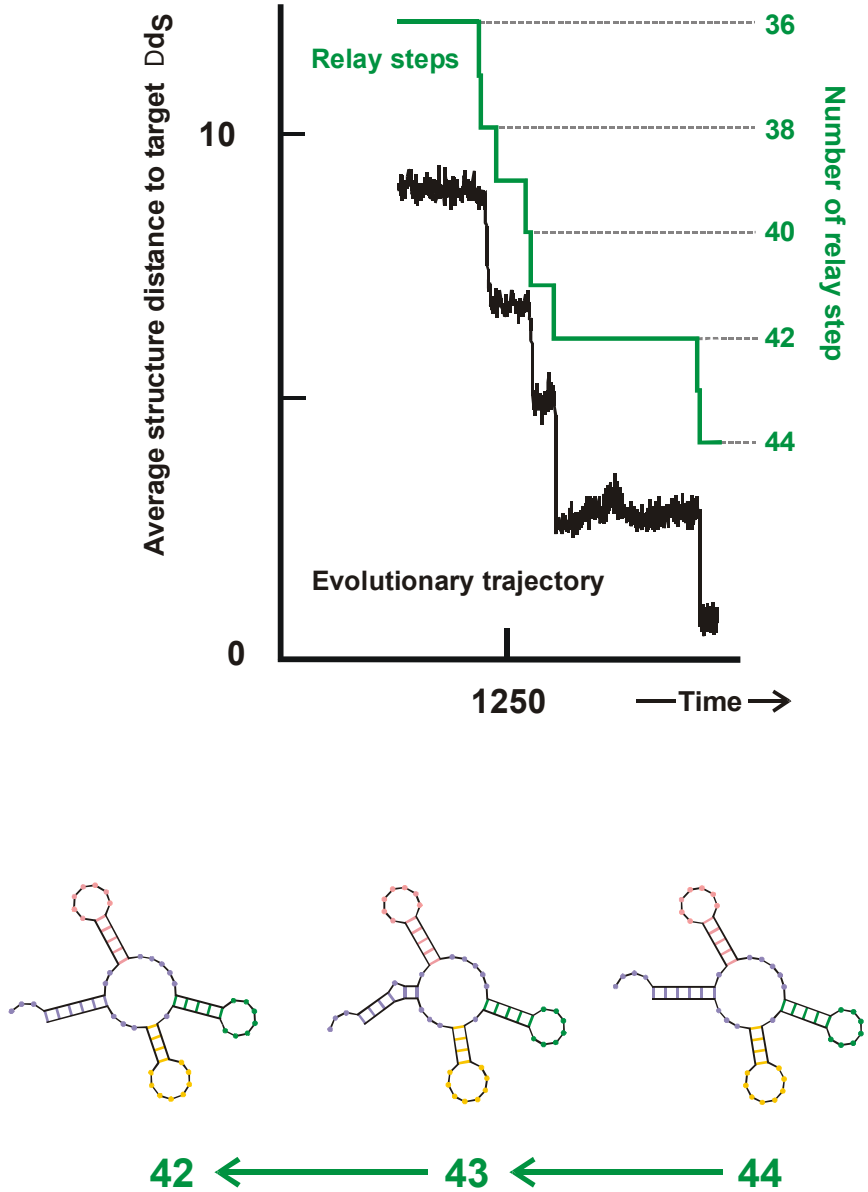
44

Endconformation of optimization

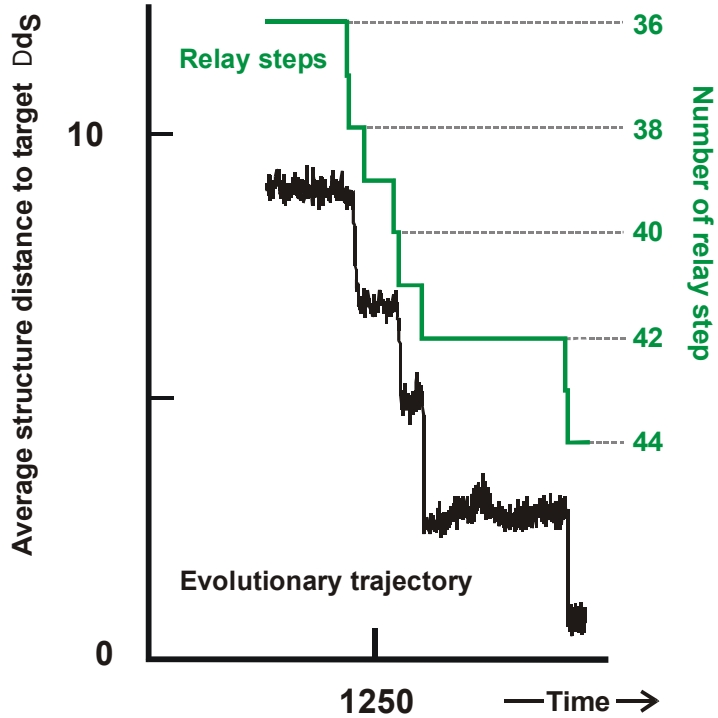


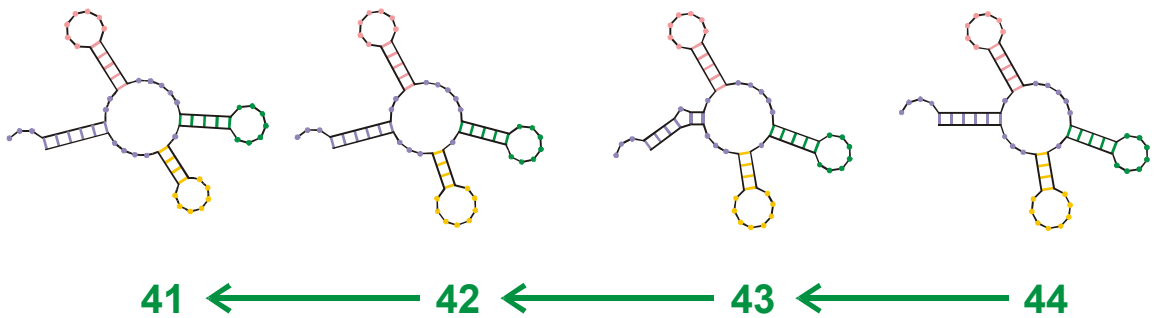


Reconstruction of the last step 43 Á 44

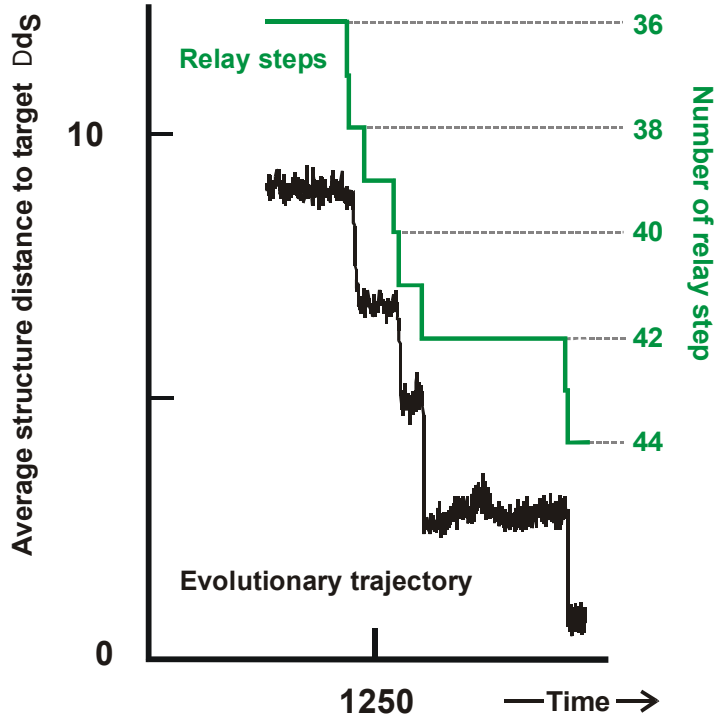


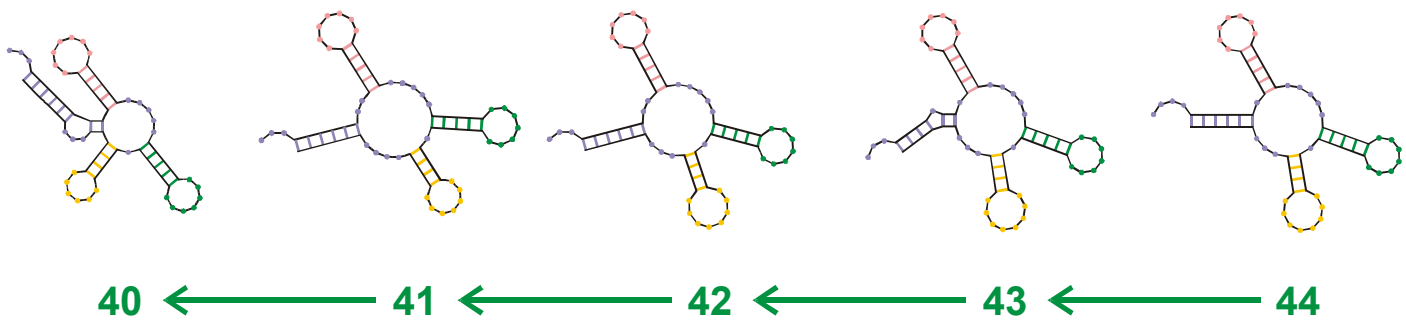
Reconstruction of last-but-one step 42 Á 43 (Á 44)



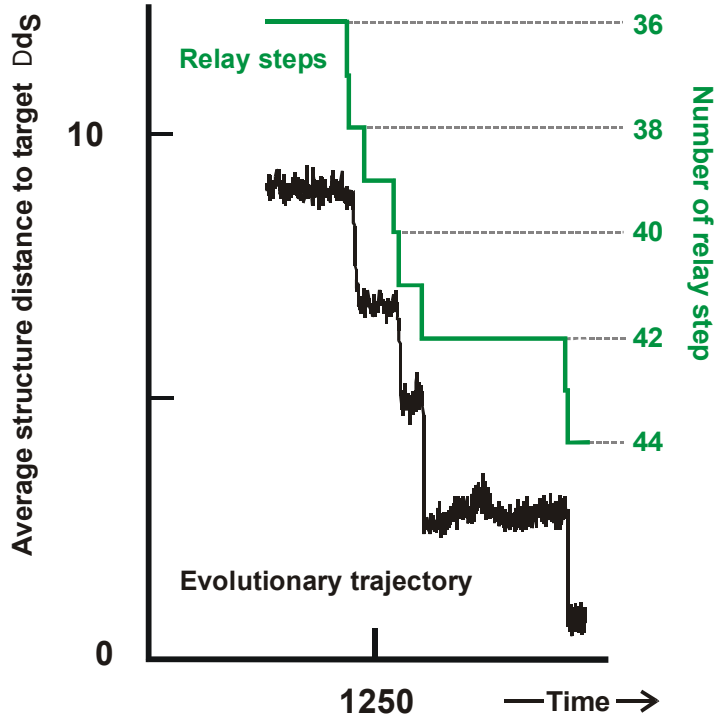


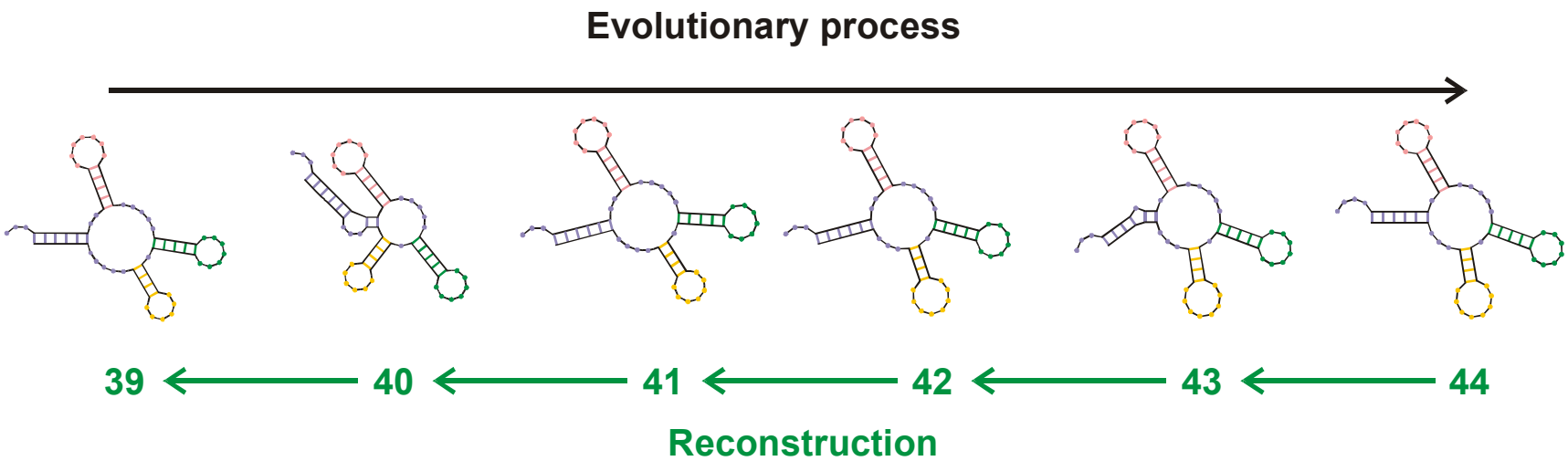
Reconstruction of step 41 Á 42 (Á 43 Á 44)



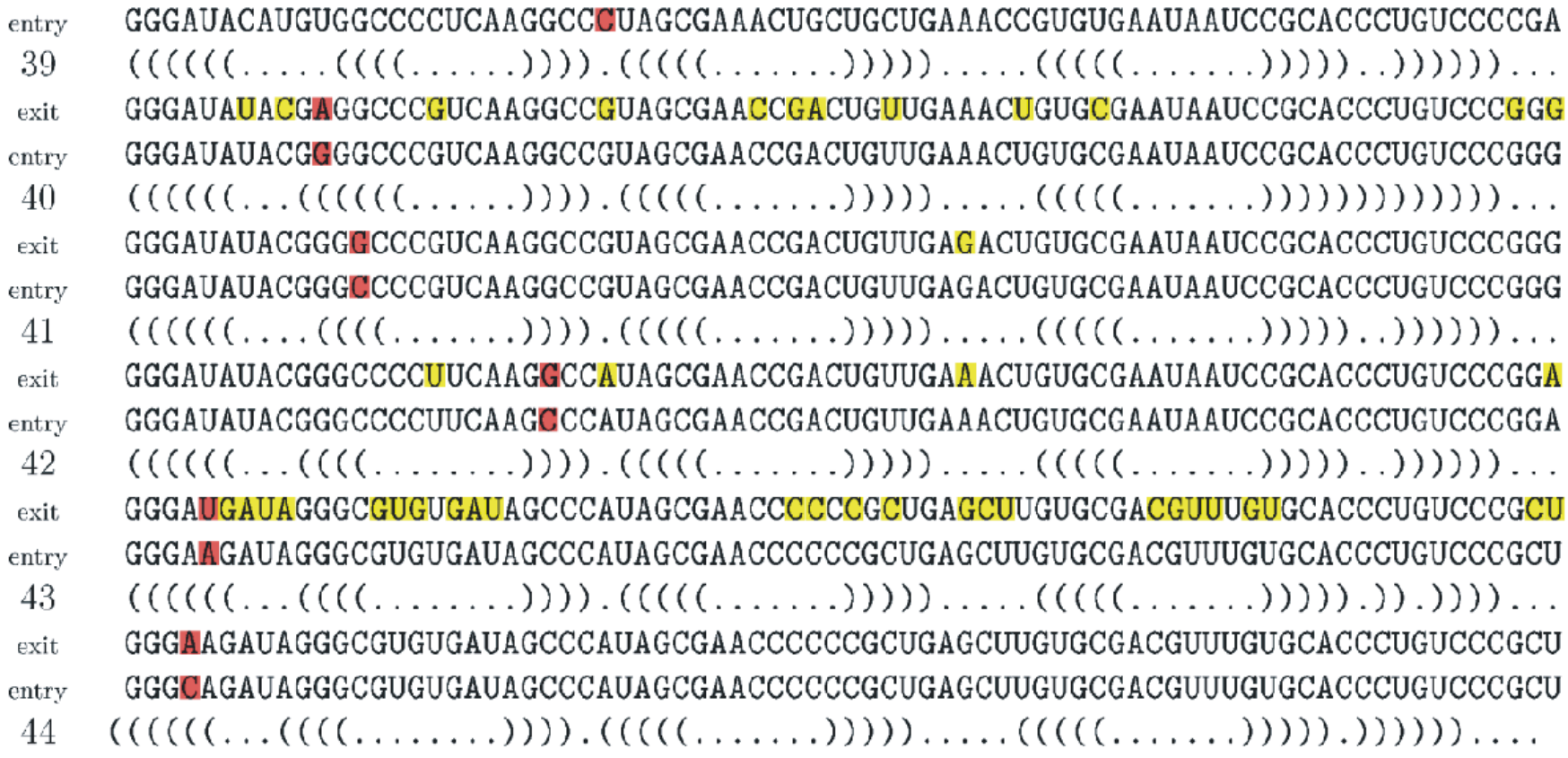


Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)



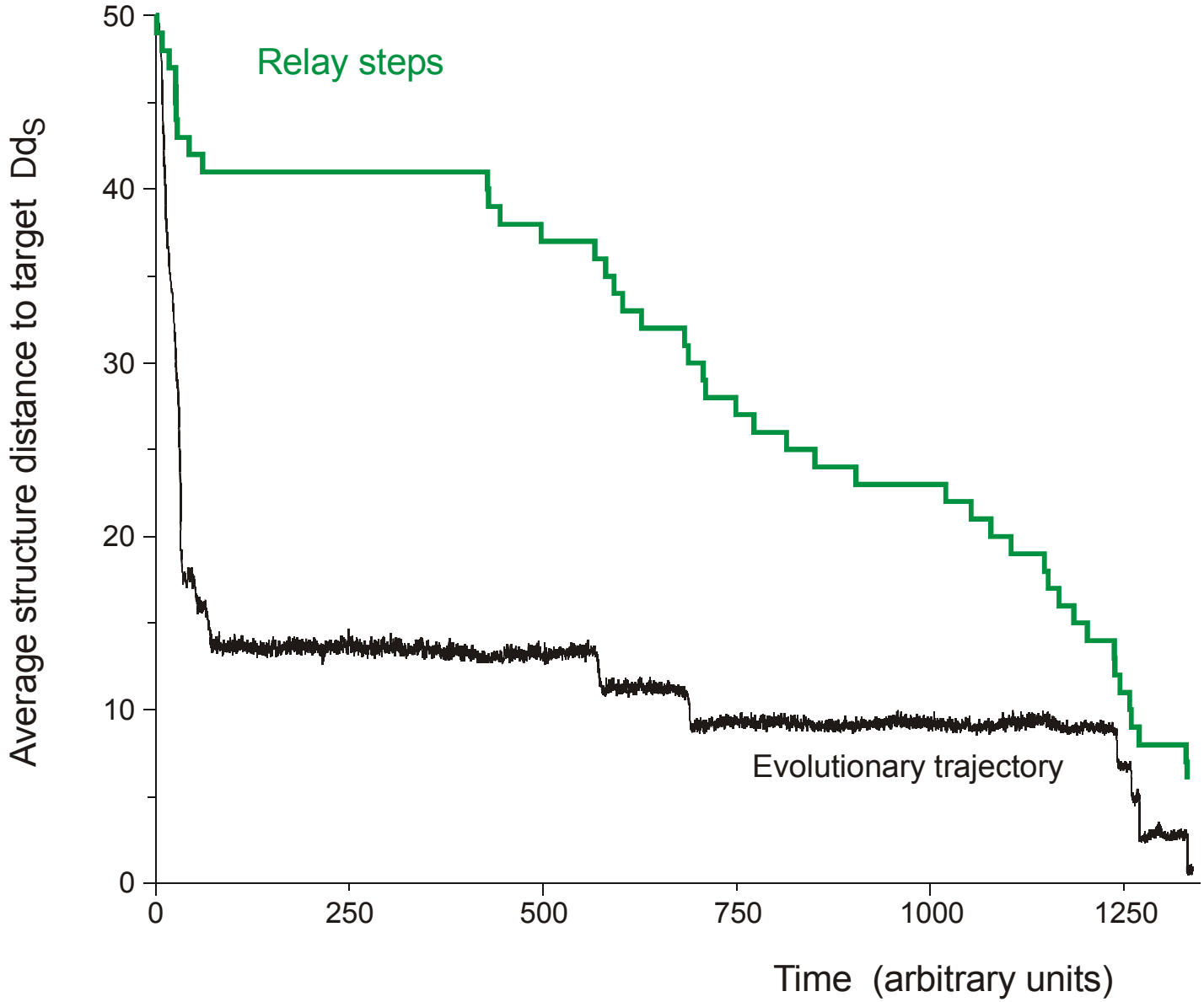


Reconstruction of the relay series

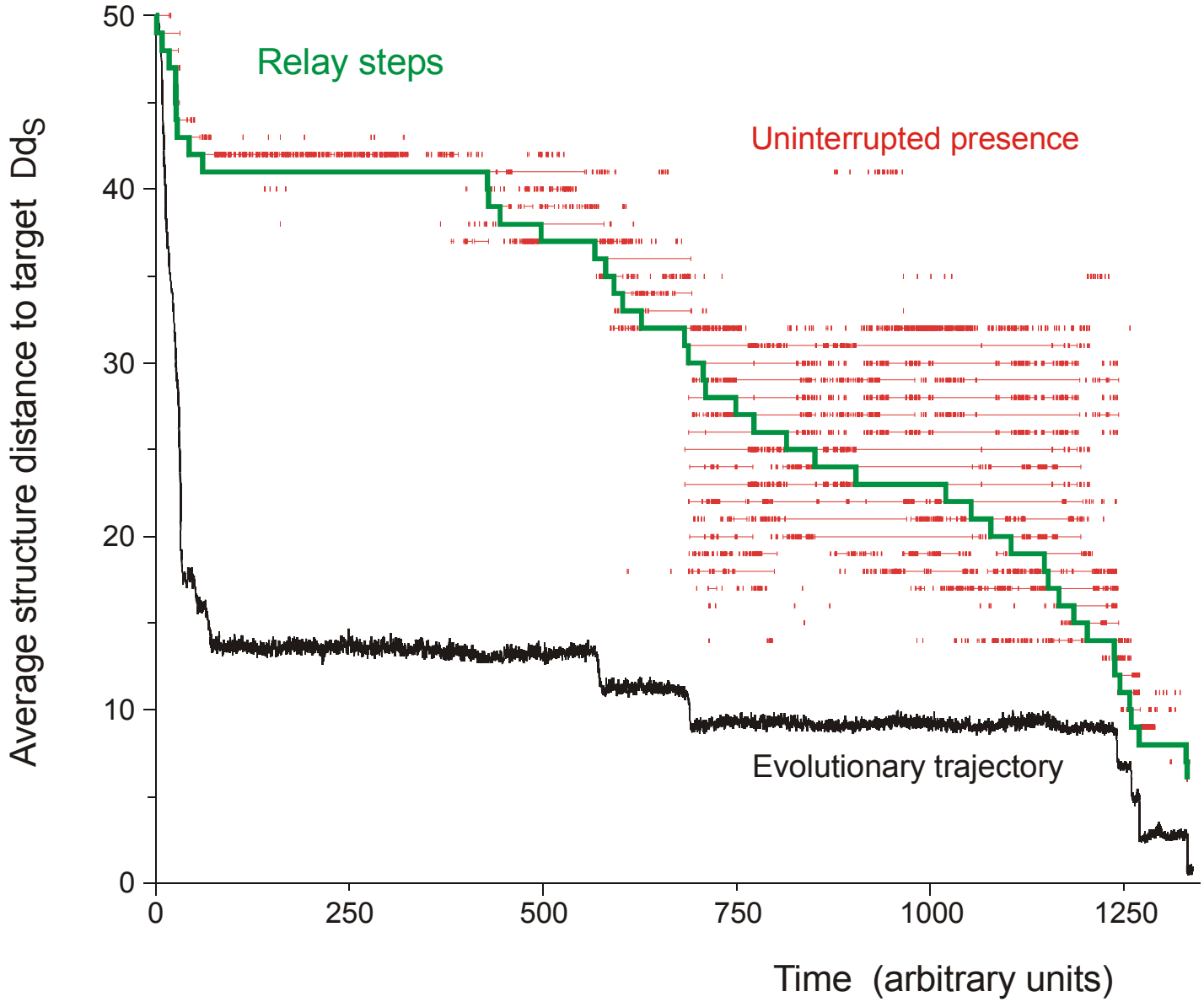


Transition inducing point mutations **Neutral point mutations**

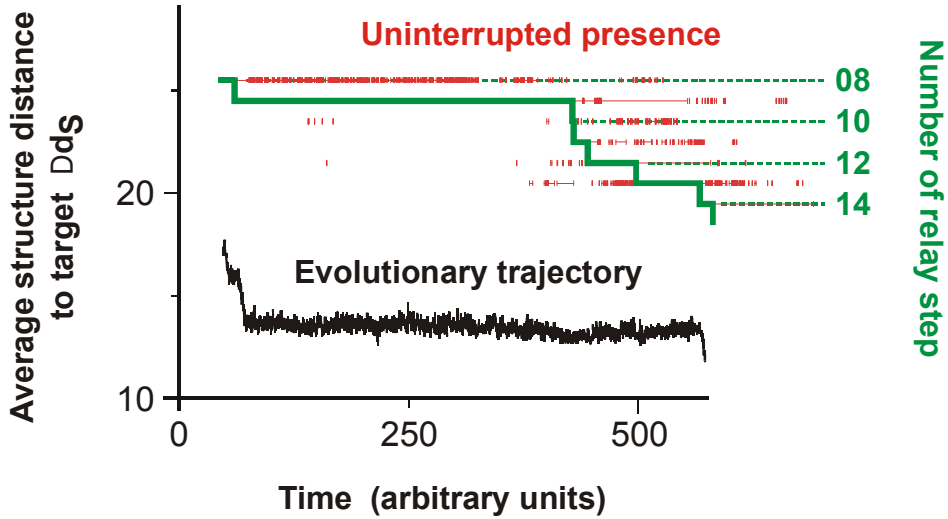
Change in RNA sequences during the final five relay steps 39 Á 44



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



In silico optimization in the flow reactor: Uninterrupted presence

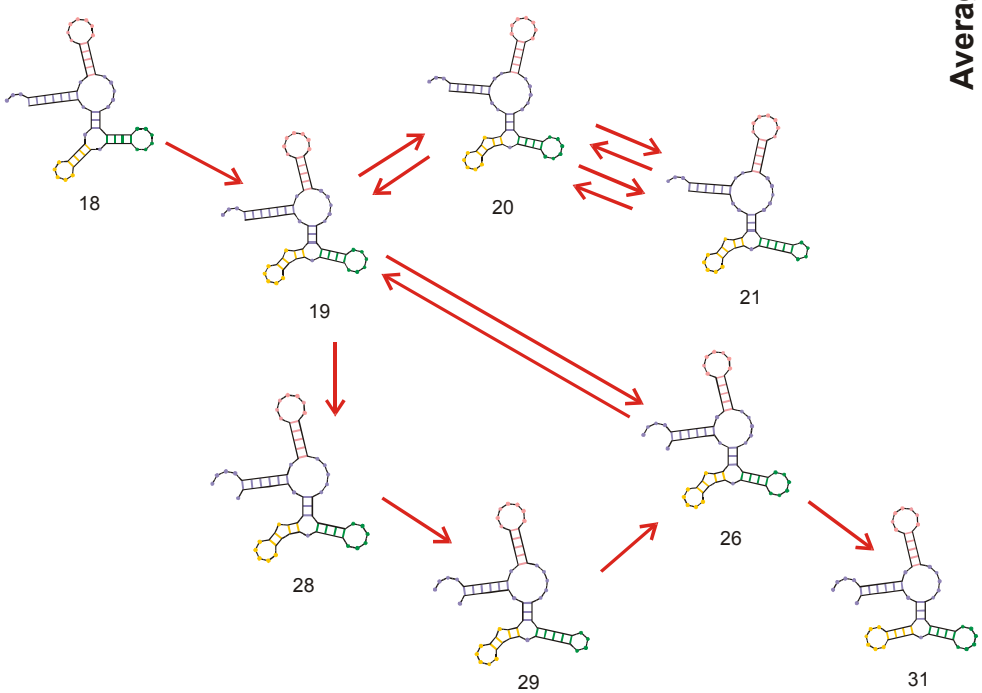


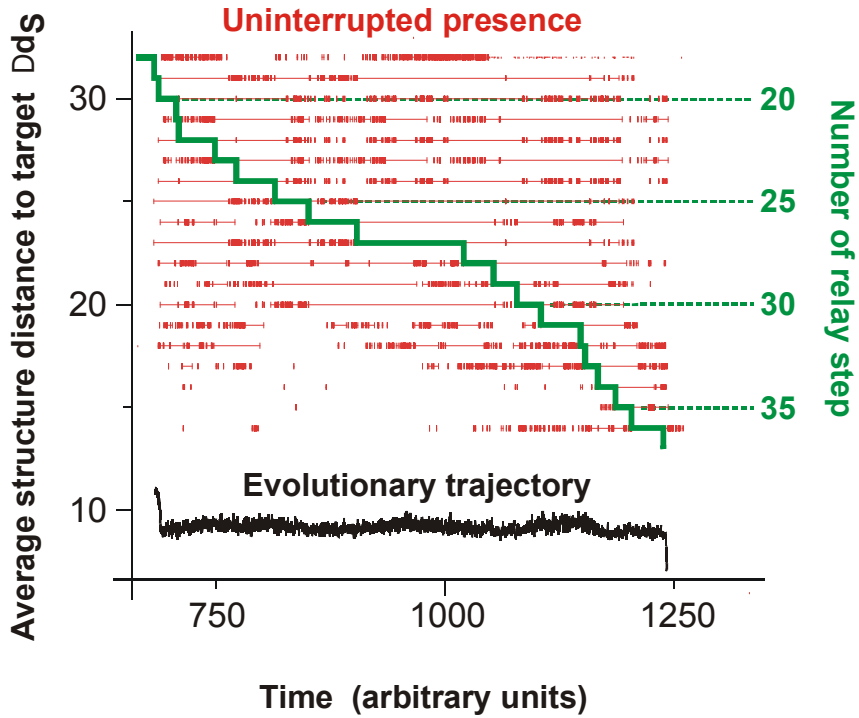

```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8      .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations

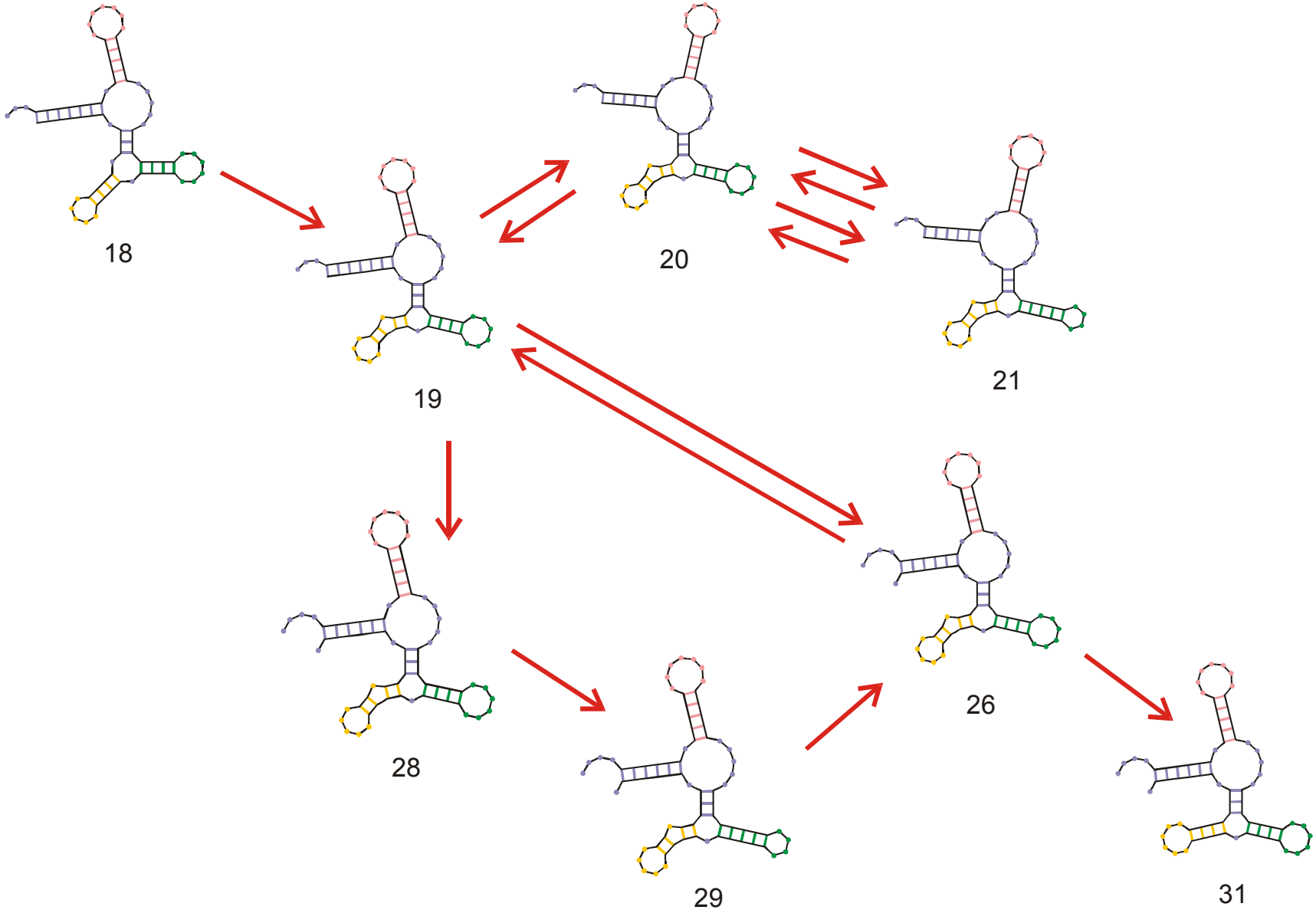
Neutral point mutations

Neutral genotype evolution during phenotypic stasis

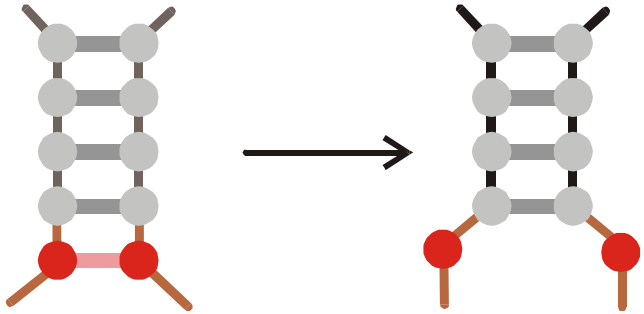




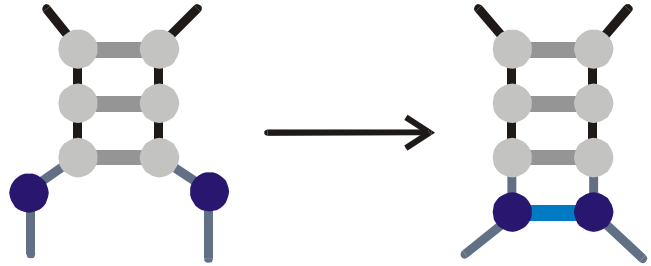
A random sequence of **minor** or continuous **transitions** in the relay series



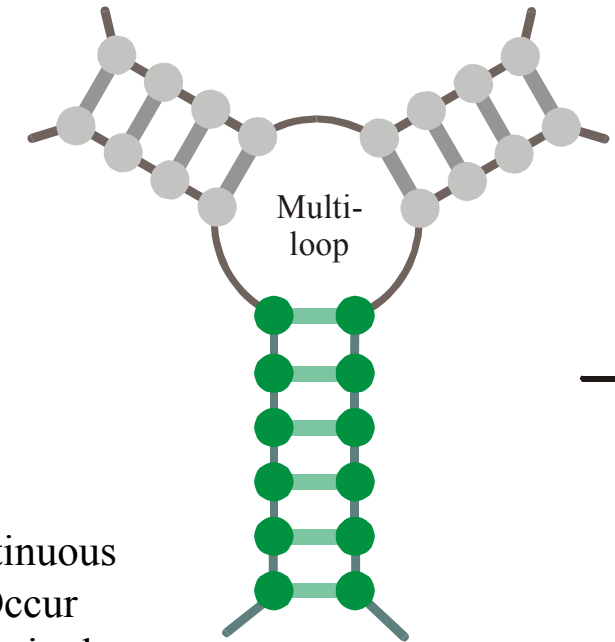
A random sequence of **minor** or continuous **transitions** in the relay series

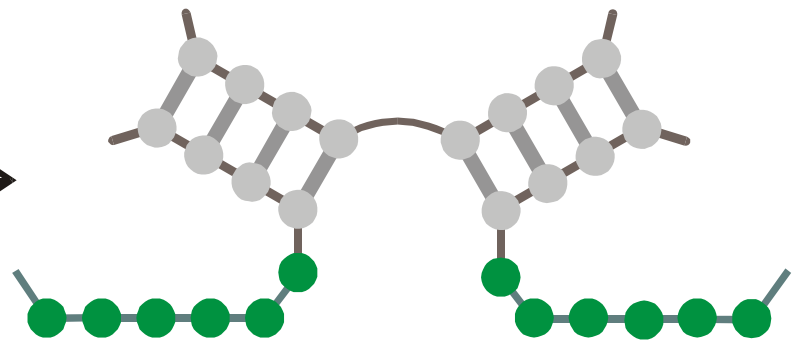


Shortening of Stacks



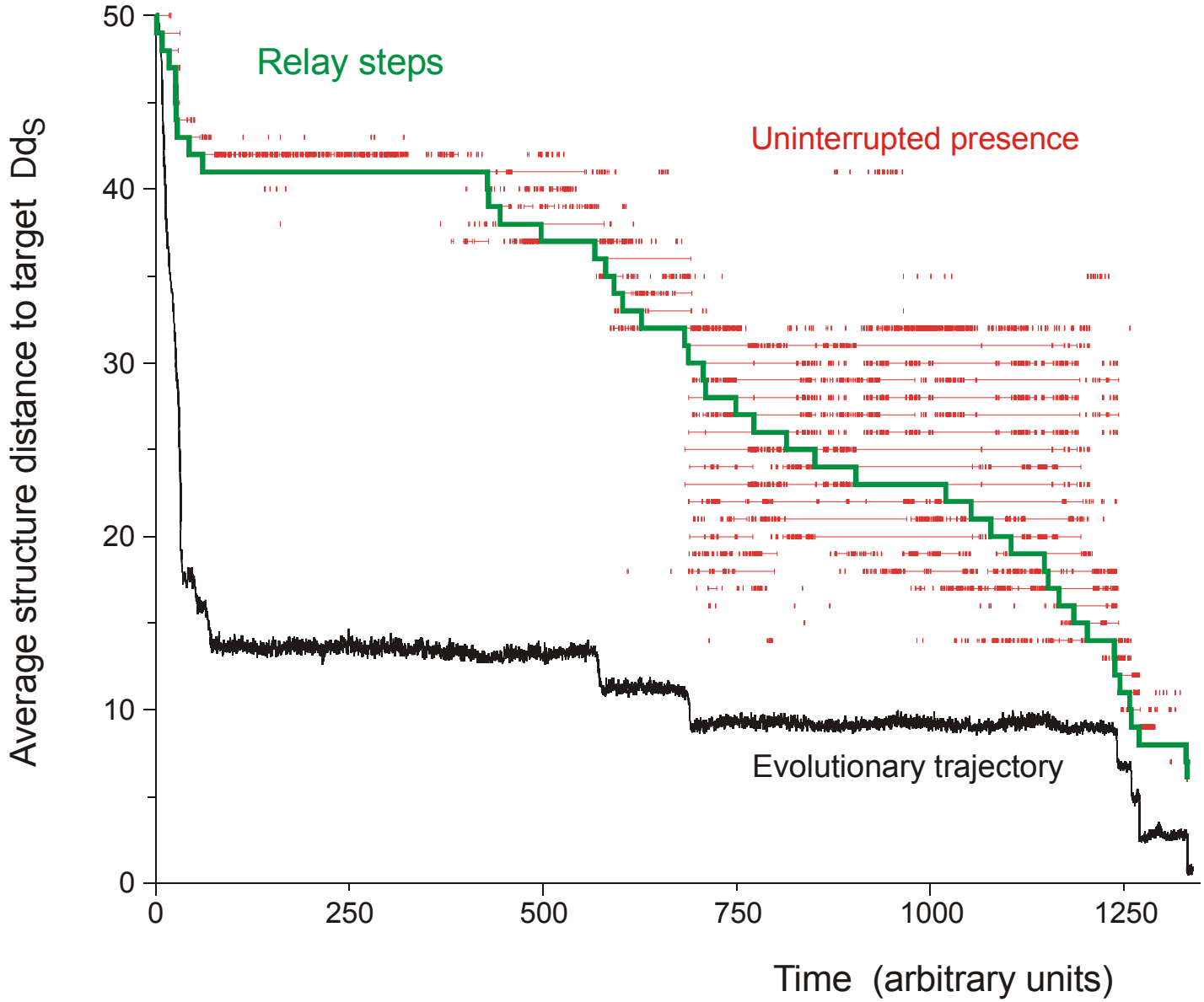
Elongation of Stacks



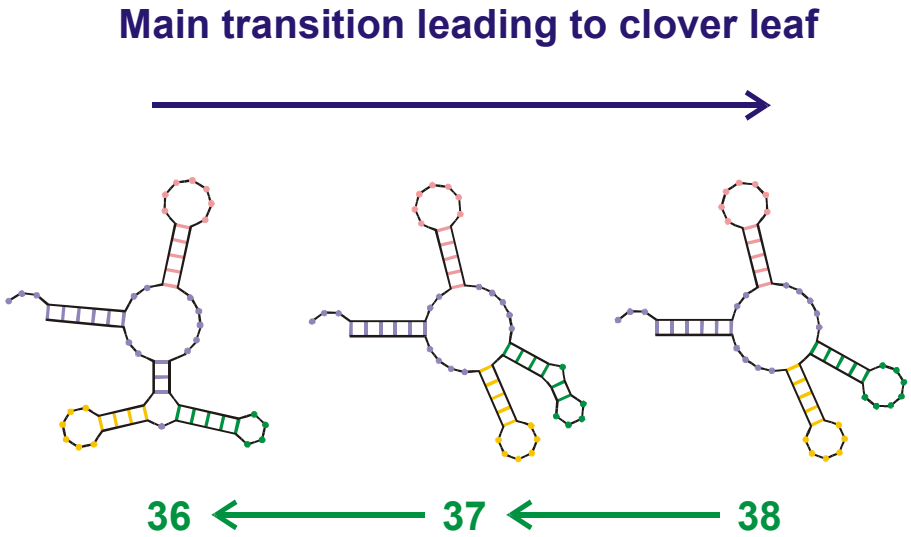


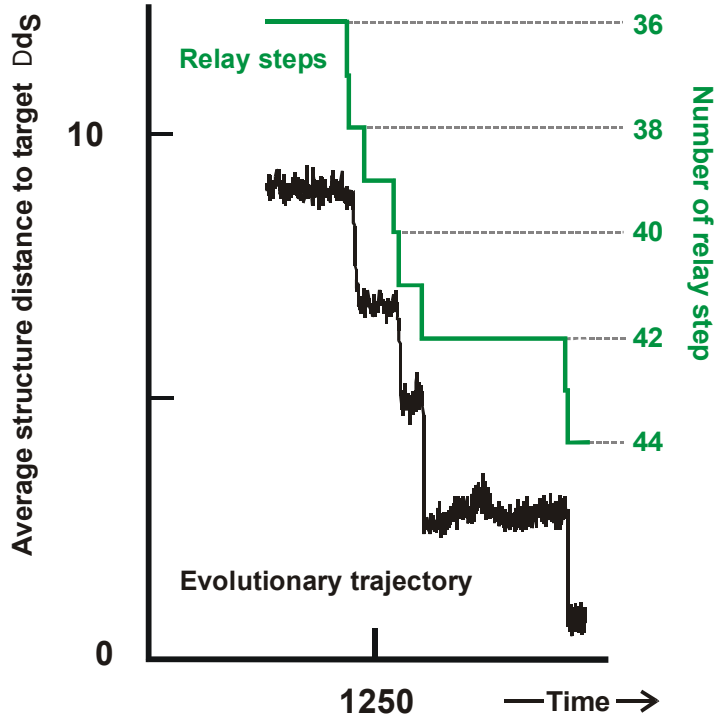
Opening of Constrained Stacks

Minor or continuous **transitions**: Occur **frequently** on single point mutations

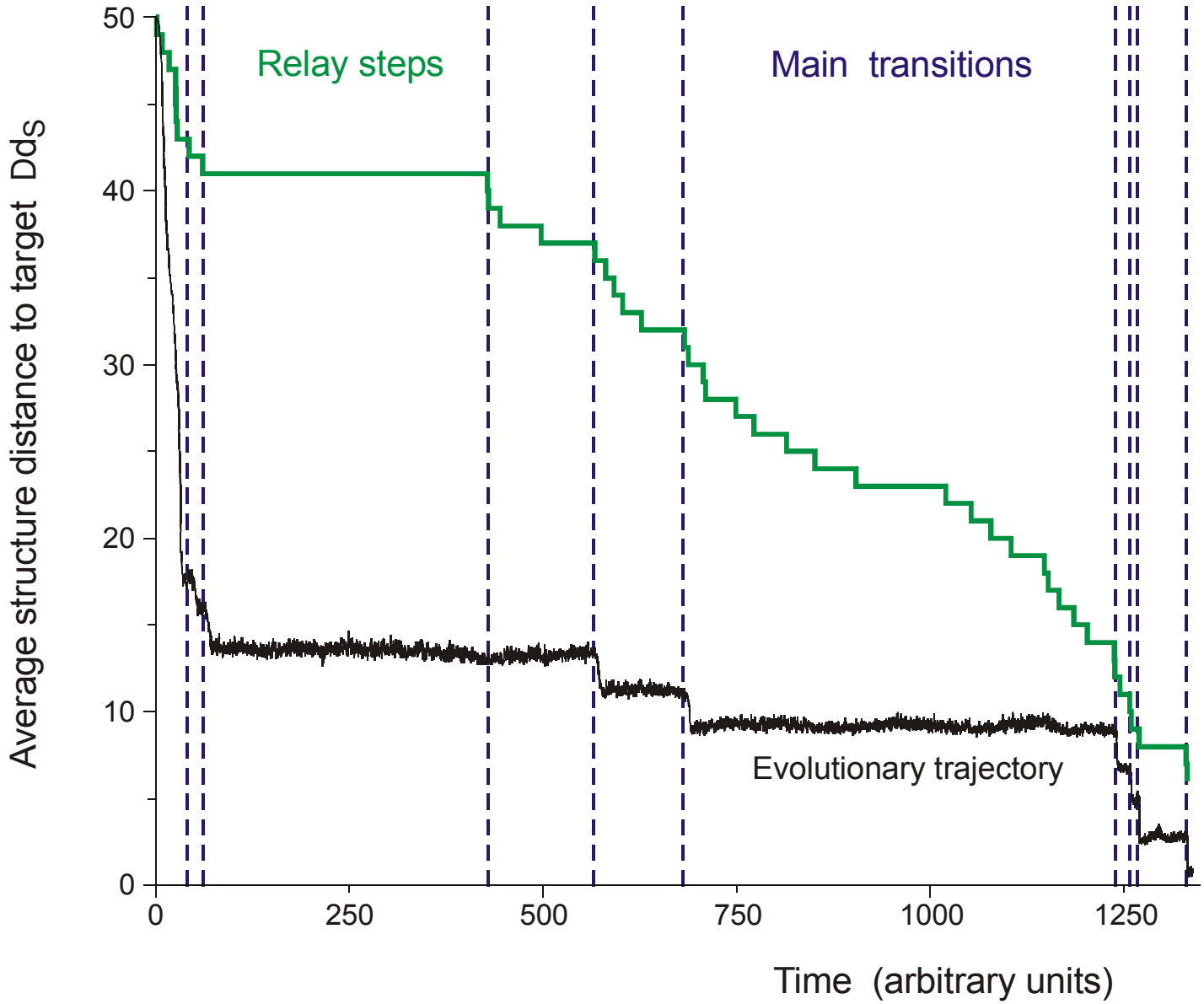


In silico optimization in the flow reactor: **Uninterrupted presence**

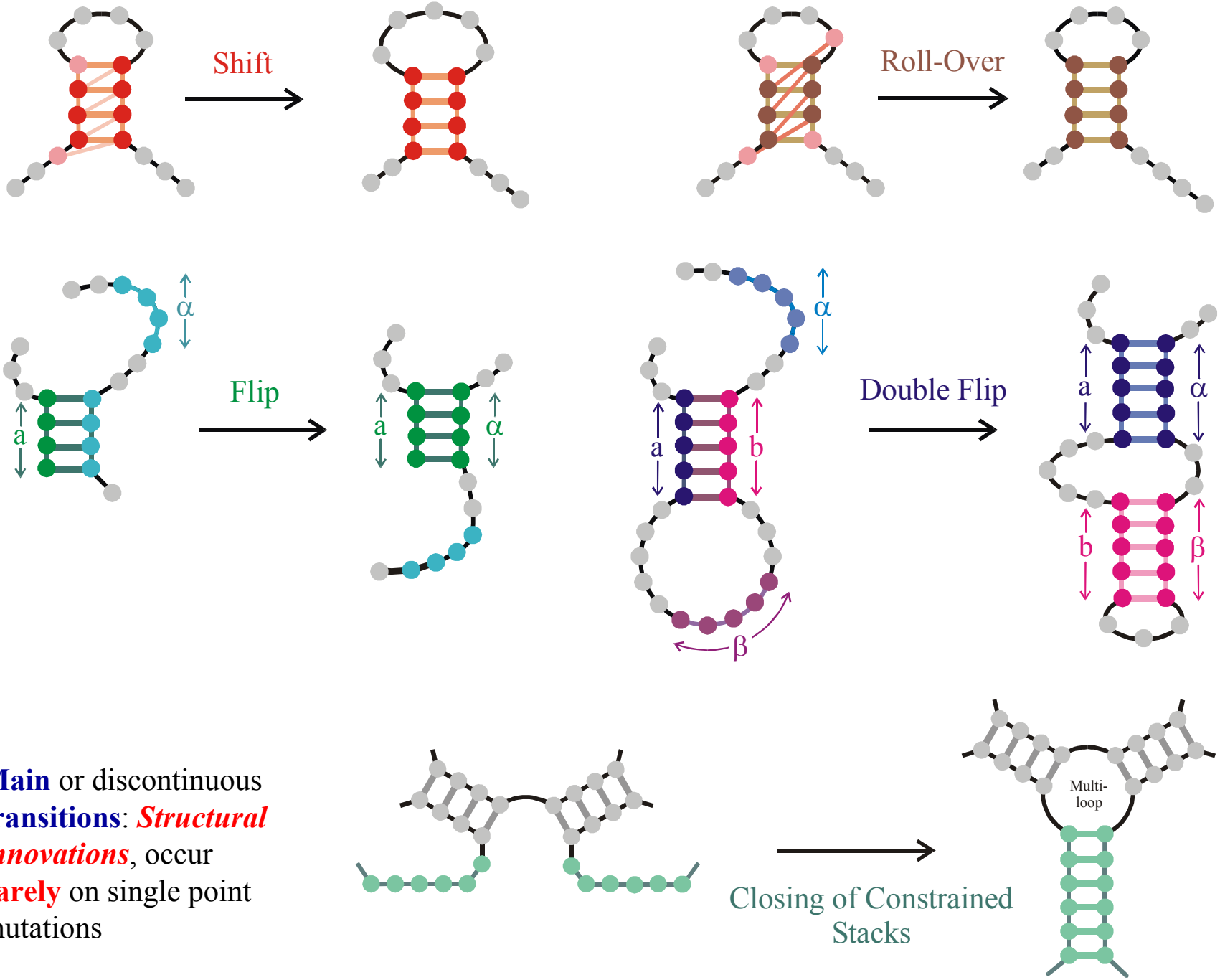


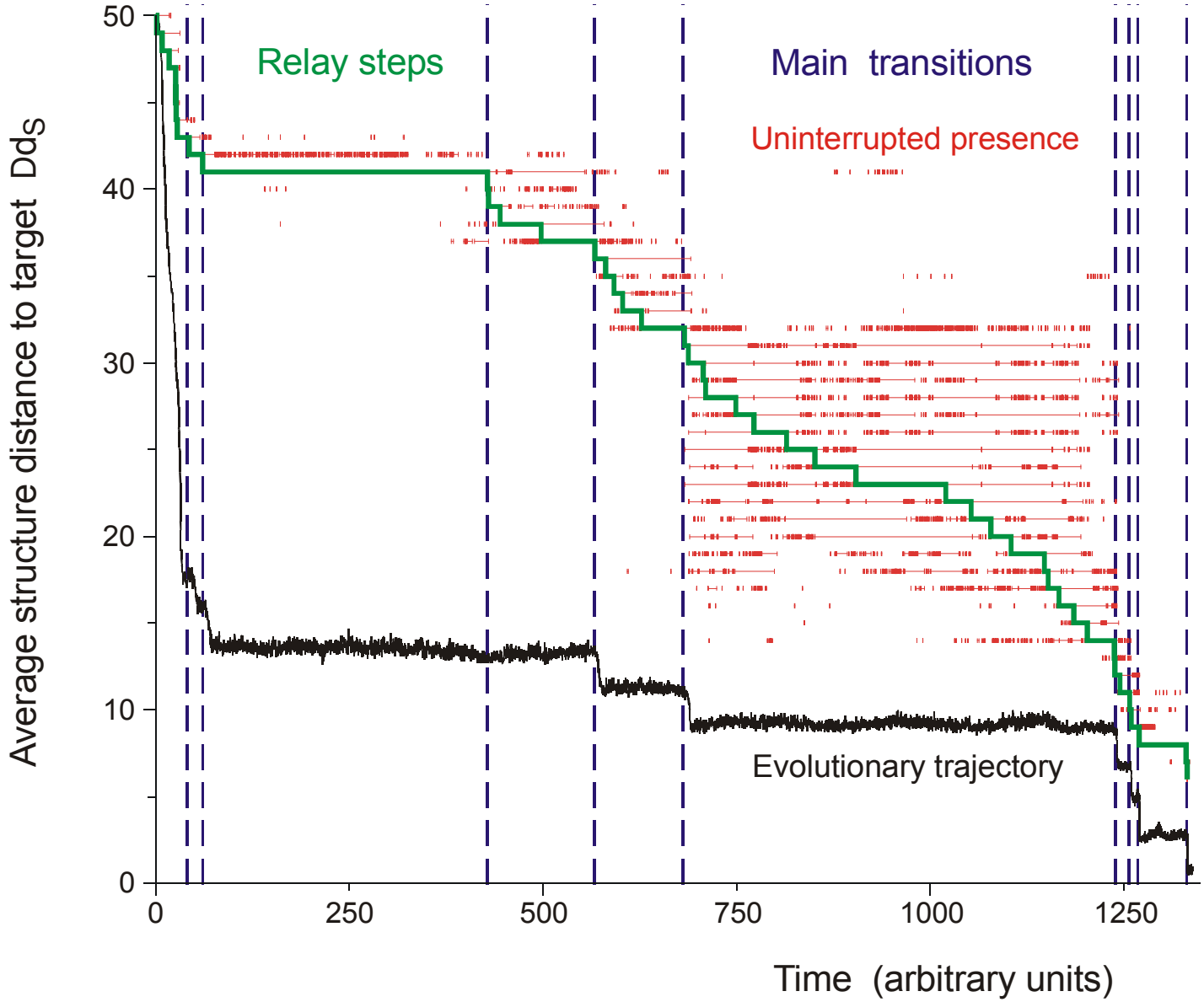


Reconstruction of a main transitions 36 Á 37 (Á 38)



In silico optimization in the flow reactor: Main transitions


Shift
Roll-Over
α
a
Flip
α
a
α
a
b
β
Double Flip
a
α
b
β
Main or discontinuous transitions: Structural innovations, occur rarely on single point mutations
Multi-loop
Closing of Constrained Stacks



In silico optimization in the flow reactor

The one-error neighborhood of the **neutral network G_k** corresponding to the structure S_k is defined by

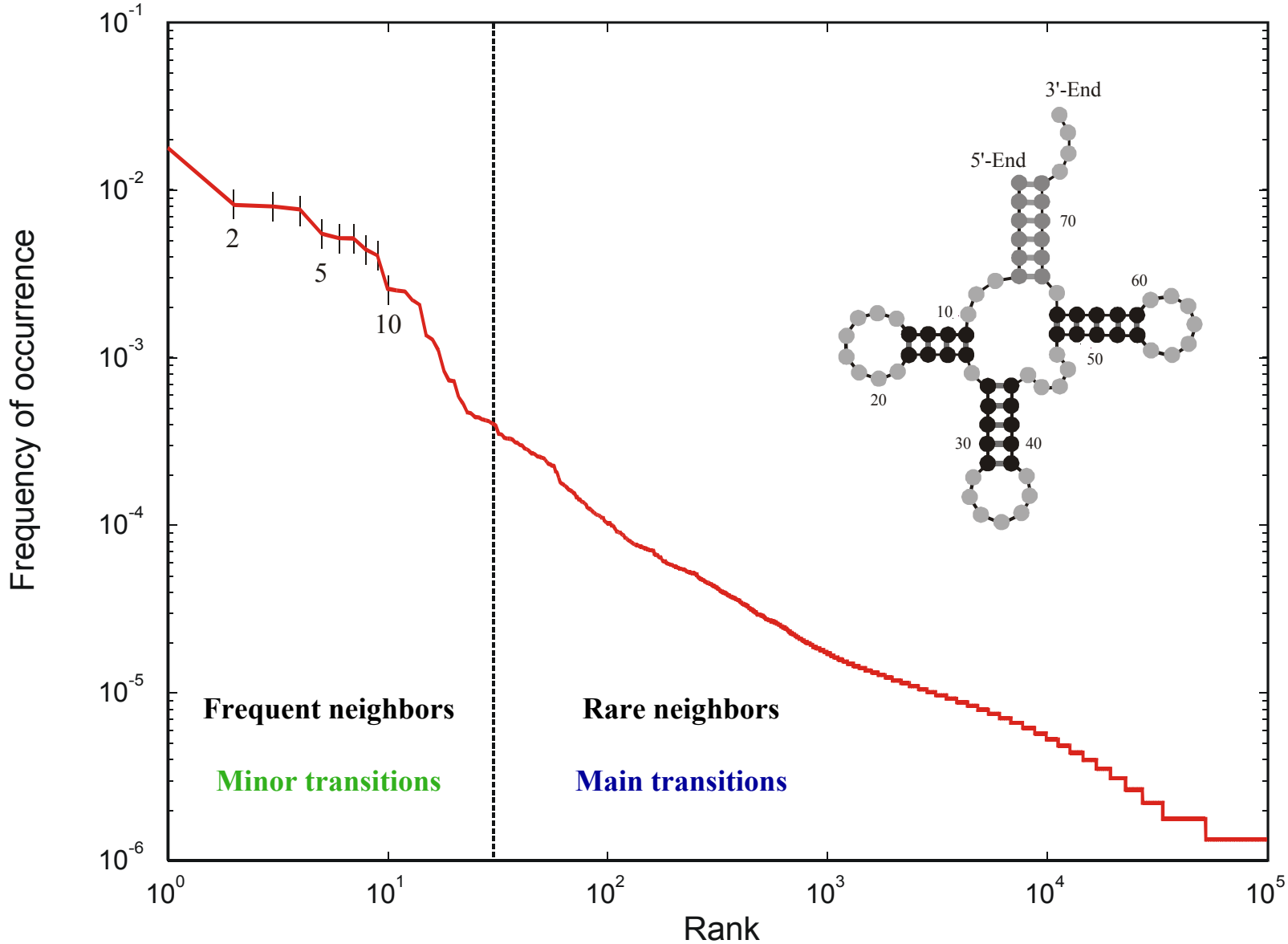
$$S(S_k) = \{S_j \mid S_j = y(I_i) \text{ ì } d^h(I_i,I_m)\ ,\ I_m \text{ á } \mathbf{G_k}\}$$

Let g_{jk} be the number of points, at which the two **neutral networks G_k** and **G_j** are in Hamming distance one contact, with $g_{jk} = g_{kj}$. The probability of occurrence of S_j in the neighbothood of S_k is then given by

$$r(S_j;S_k) \;=\; g_{jk} \;\text{⌫}\; (\text{lê}(k\text{-}1)\text{ê}|\mathbf{G_k}|)$$

We note that this probability is not symmetric, $r(S_j;S_k)$ Ì $r(S_k;S_j)$, except the two networks are of equal size, $|\mathbf{G_k}| = |\mathbf{G_j}|$. The definition of a **statistical e-neighborhood** of the structure S_k allows for precise distinction between frequent and rare neighbors. Frequent neighbors are contained in the **statistical neighborhood**

$$\text{´}_e(S_k) = \{S_j \text{ á } S(S_k) \mid r(S_j;S_k) \text{ Æ } e\}\ .$$

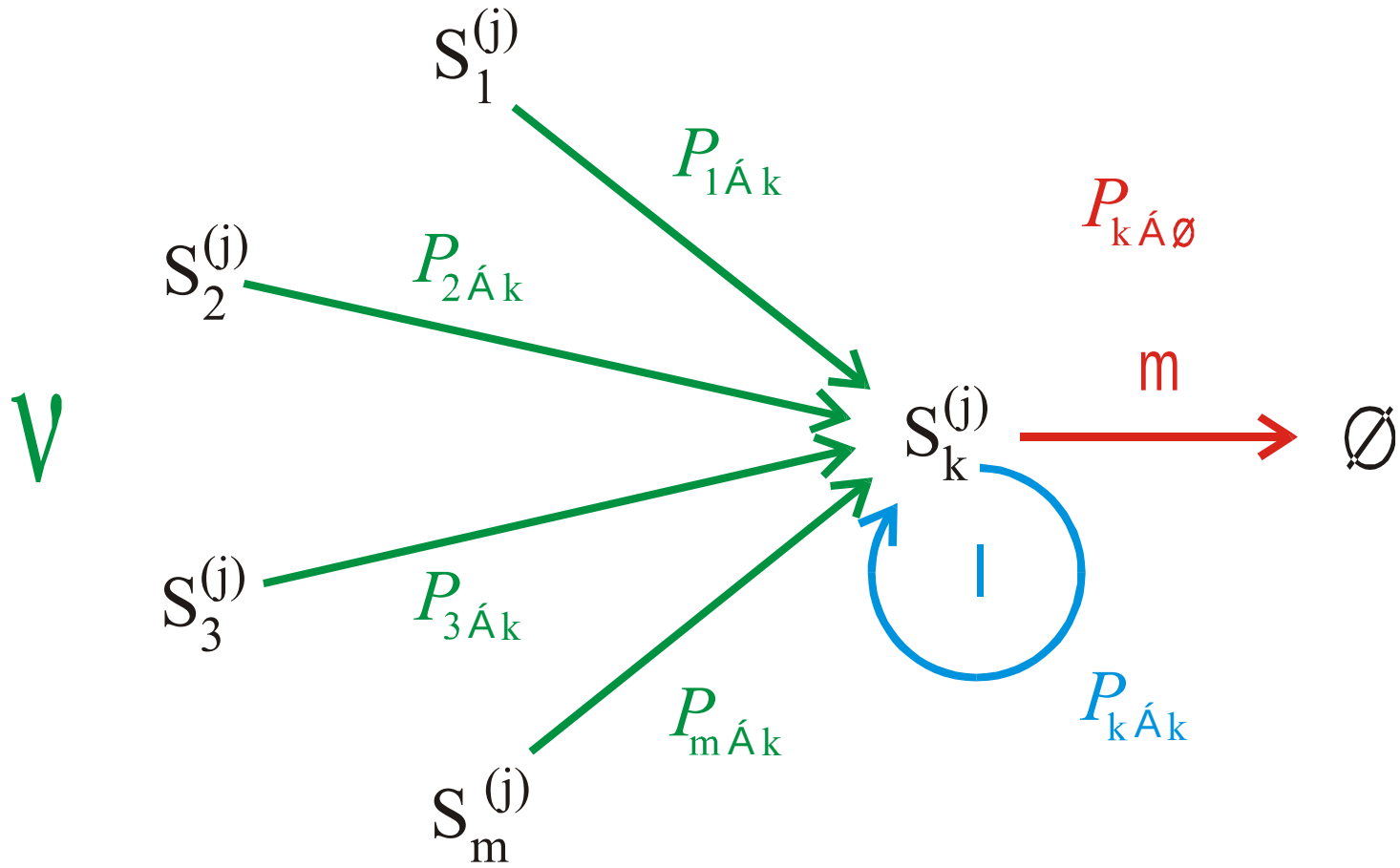


Probability of occurrence of different structures in the mutational neighborhood of $tRNA^{phe}$

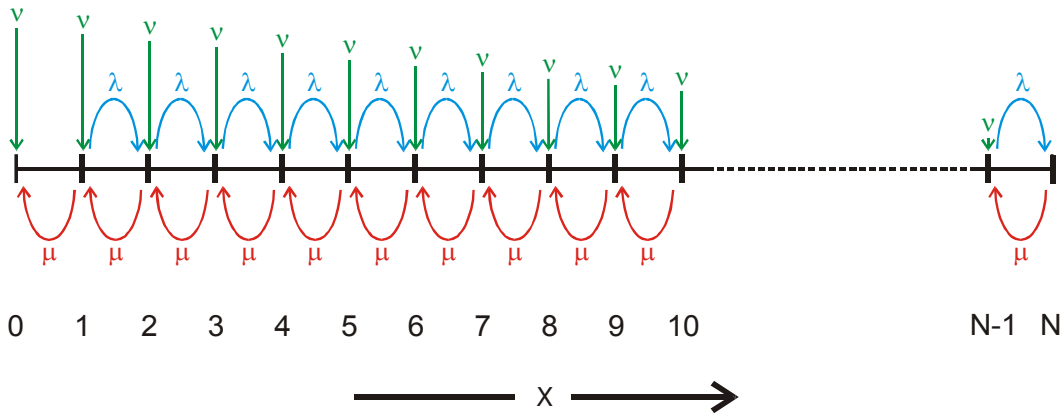
Statistics of evolutionary trajectories

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

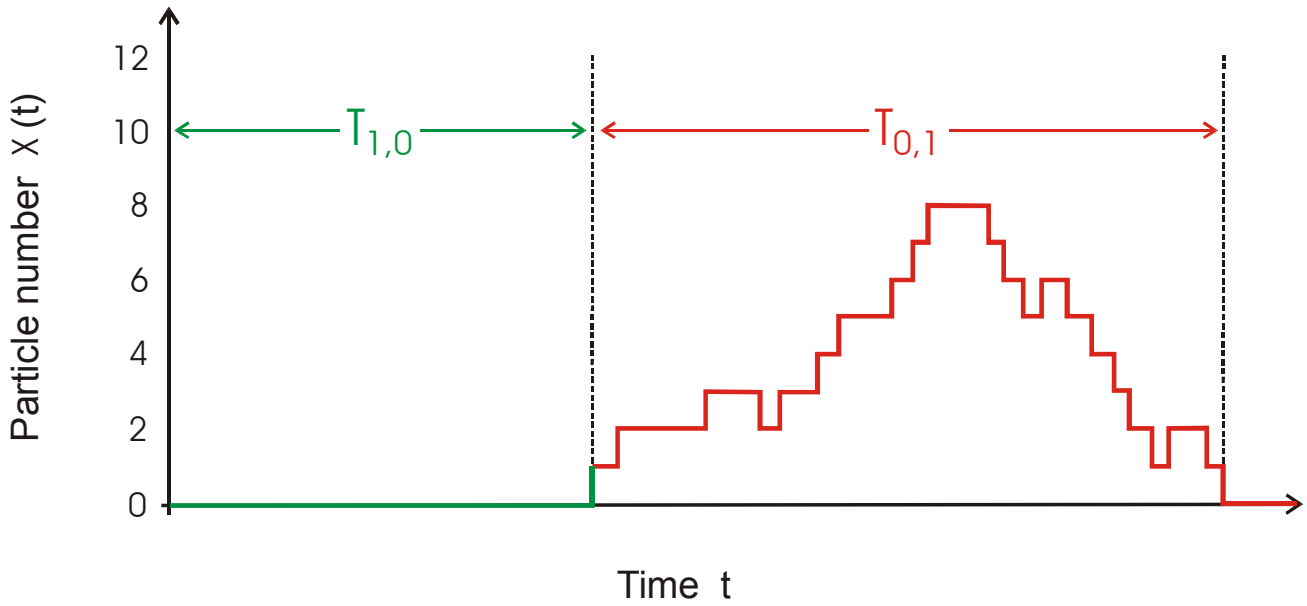
The number of **main transitions** or evolutionary innovations is constant.



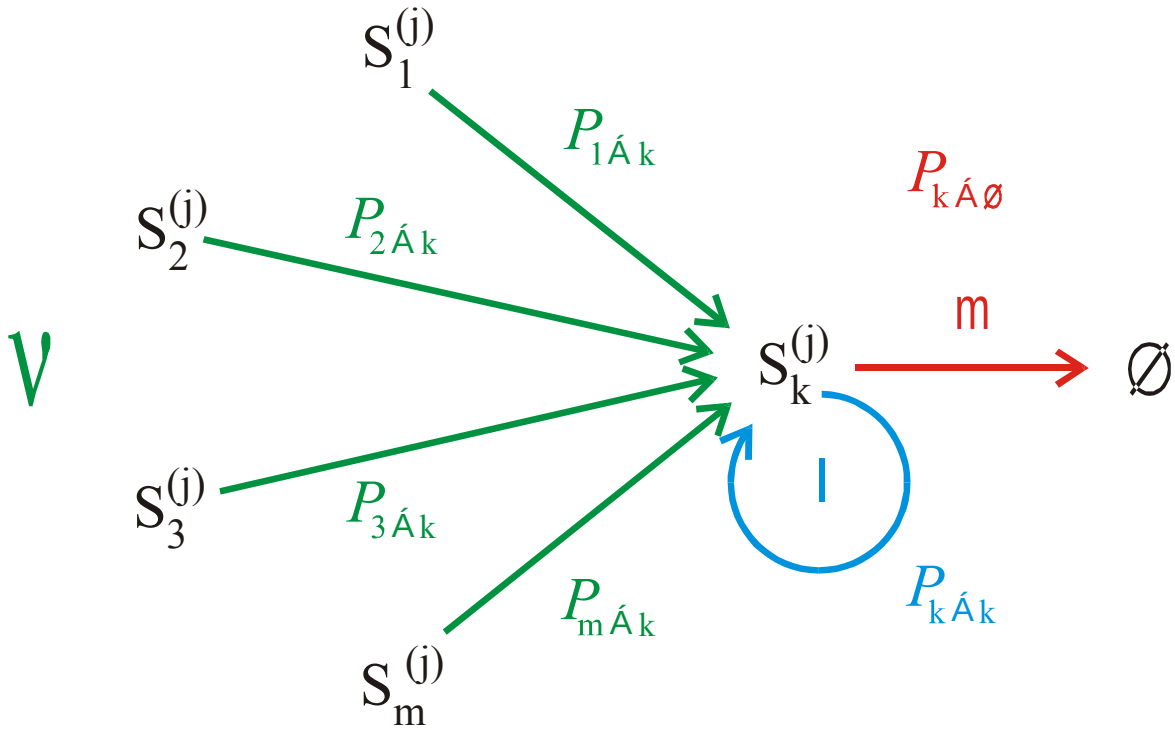
Transition probabilities determining the presence of phenotype $S_k^{(j)}$ in the population



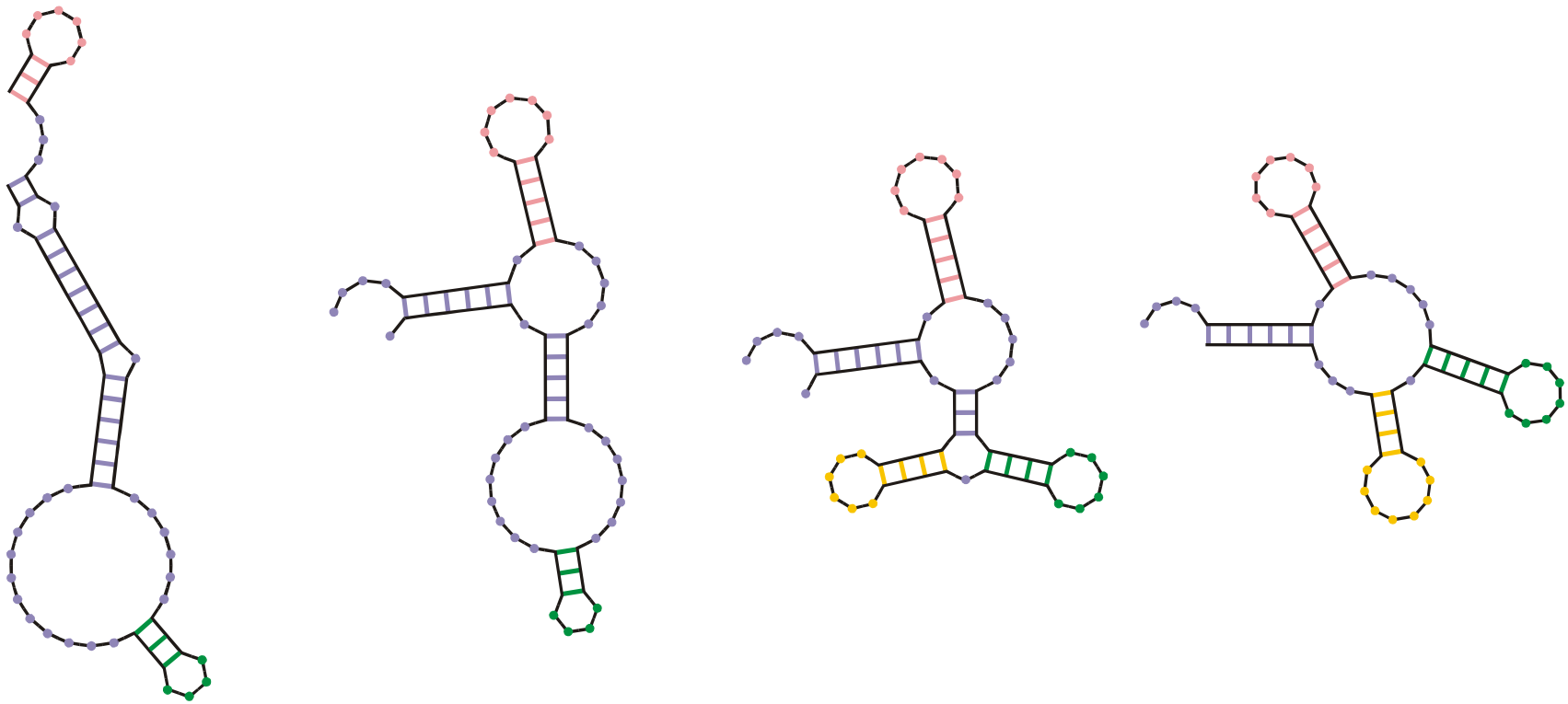
$$\lambda(x) = \lambda\, x + \nu\,(N\text{-}x) \qquad \mu(x) = \mu\, x$$



Calculation of transition probabilities by means of a birth-and-death process with immigration

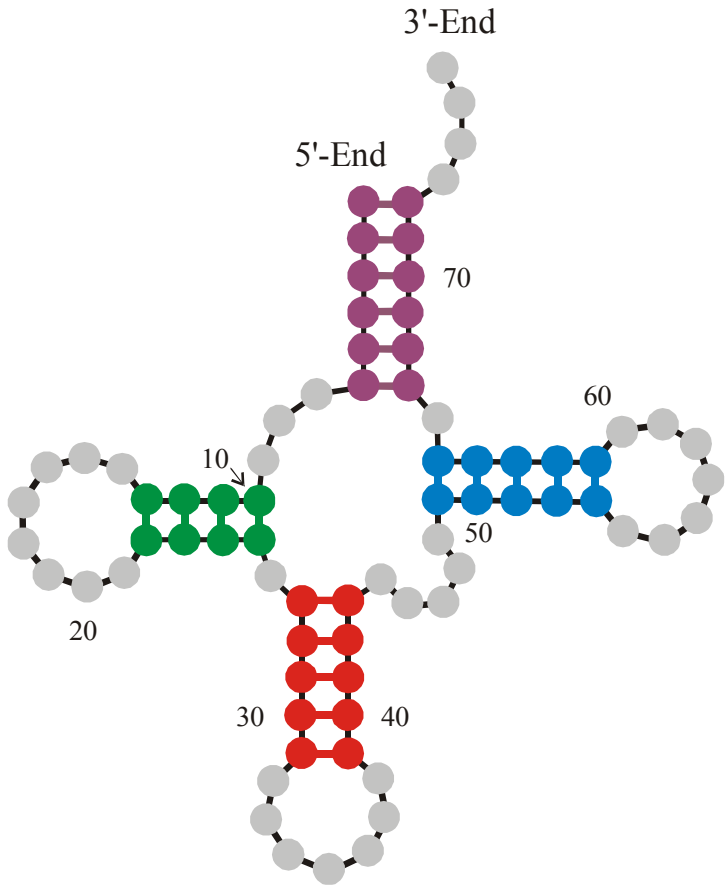


$$N_{sat}^{(j)} = \frac{1}{p \,.\, l \,.\, \langle r^{(j)} \rangle}$$



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

Stable tRNA clover leaf structures built from binary, **GC**-only, sequences exist. The corresponding sequences are readily found through inverse folding. Optimization by mutation and selection in the flow reactor has so far always been unsuccessful.



The neutral network of the tRNA clover leaf in **GC** sequence space is not connected, whereas to the corresponding neutral network in **AUGC** sequence space is very close to the critical connectivity threshold, I_{cr}. Here, both inverse folding and optimization in the flow reactor are successful.

The success of optimization depends on the connectivity of neutral networks.

Main results of computer simulations of molecular evolution

• No trajectory was reproducible in detail. Sequences of target structures were always different. Nevertheless **solutions of the same quality** are almost always achieved.

• Transitions between molecular phenotypes represented by RNA structures can be classified with respect to the induced structural changes. Highly probable **minor transitions** are opposed by **main transitions** with low probability of occurrence.

• **Main transitions** represent important **innovations** in the course of evolution.

• The number of **minor transitions** decreases with increasing population size.

• The number of **main transitions** or evolutionary innovations is approximately constant for given start and stop structures.

• **Not all known structures are accessible** through evolution in the flow reactor. An example is the tRNA clover leaf for GC-only sequences.

1. Optimization through variation and selection in populations

2. Neutral networks in genotype-phenotype mappings

3. Optimization in the RNA model

4. **Evolution experiments** with molecules in the laboratory

	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

Generation times and evolutionary timescales

Evolution of RNA molecules based on Qβ phage

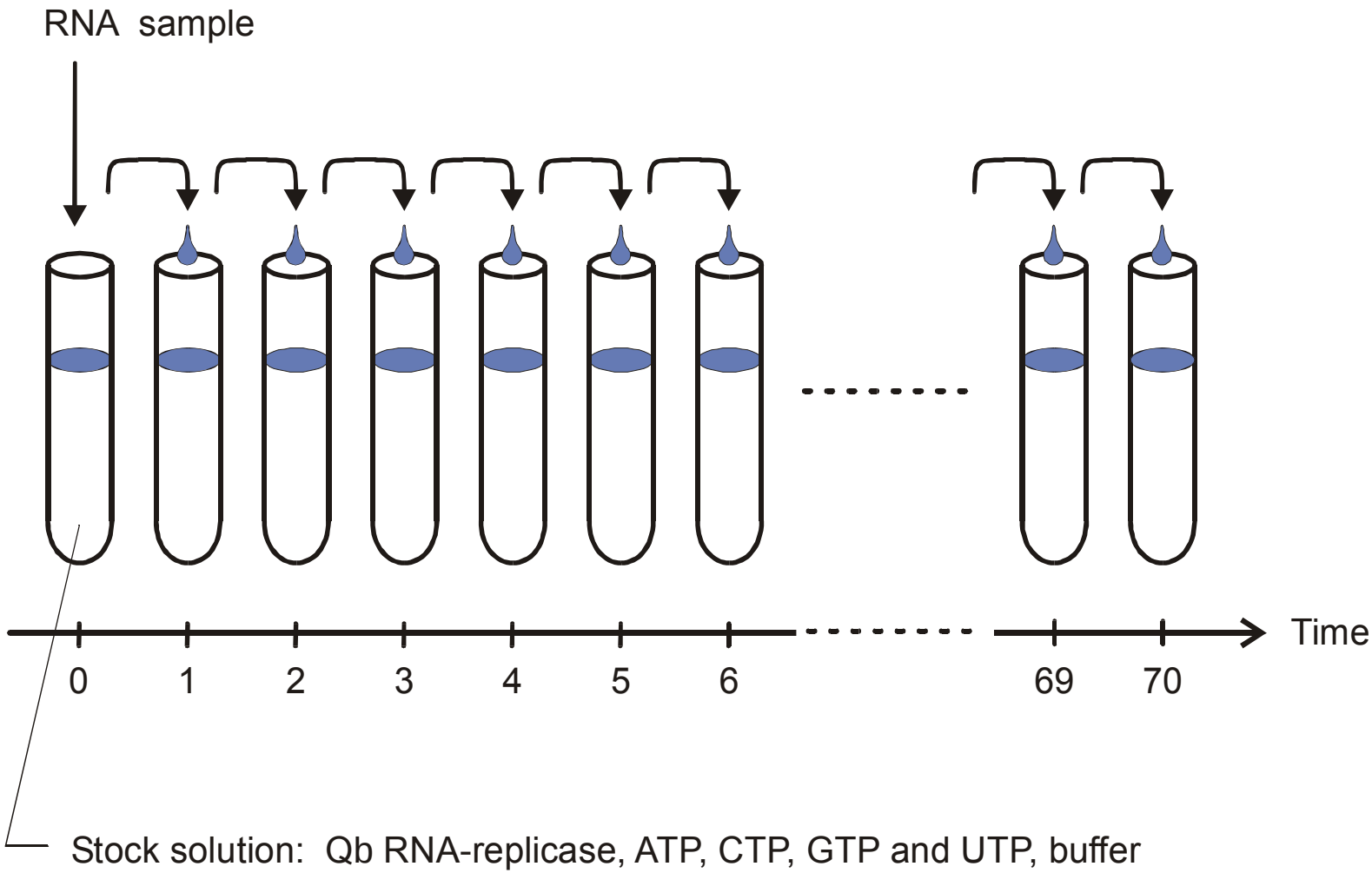
D.R.Mills, R,L,Peterson, S.Spiegelman, ***An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule***. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224

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

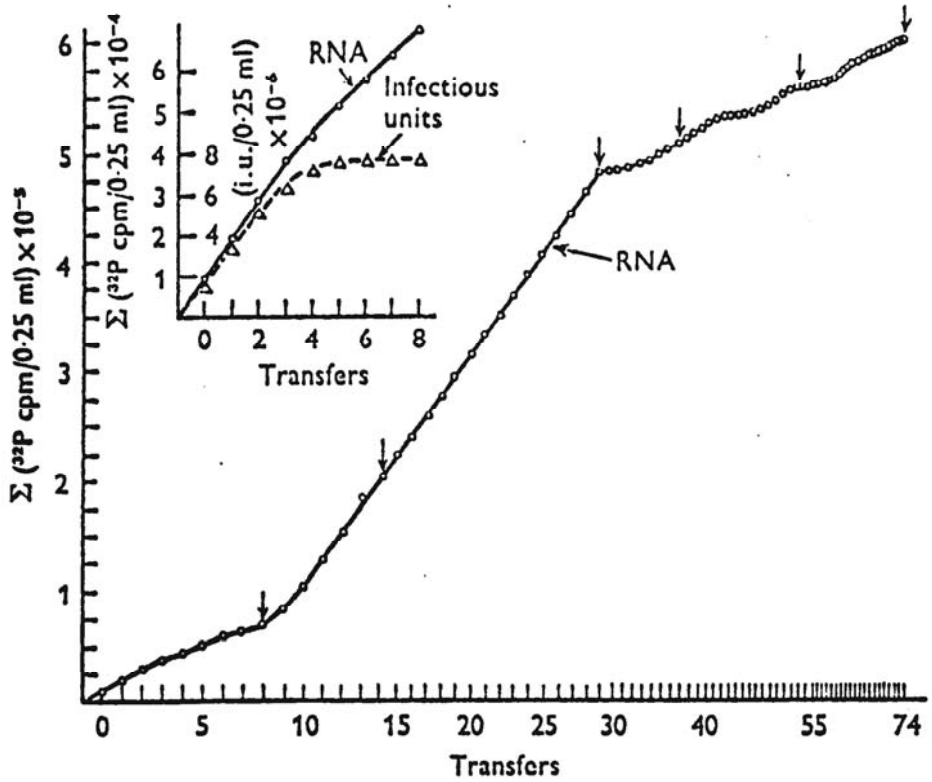
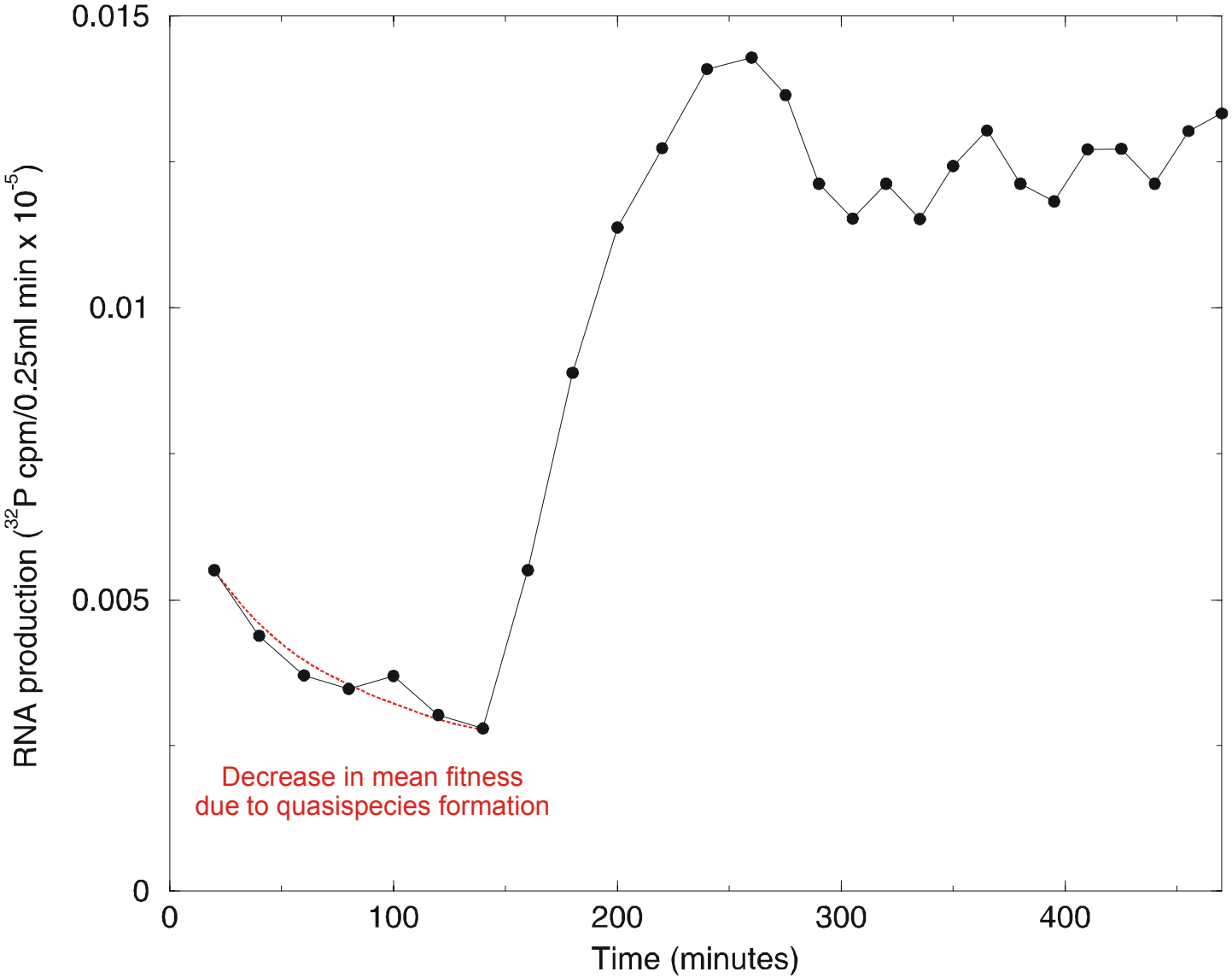


Fig. 9. Serial transfer experiment. Each 0·25 ml standard reaction mixture contained 40 μg of Qβ replicase and ^{32}P-UTP. The first reaction (0 transfer) was initiated by the addition of 0·2 μg ts-1 (temperature-sensitive RNA) and incubated at 35 °C for 20 min, whereupon 0·02 ml was drawn for counting 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).

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



The increase in RNA production rate during a serial transfer experiment

Evolutionary design of RNA molecules

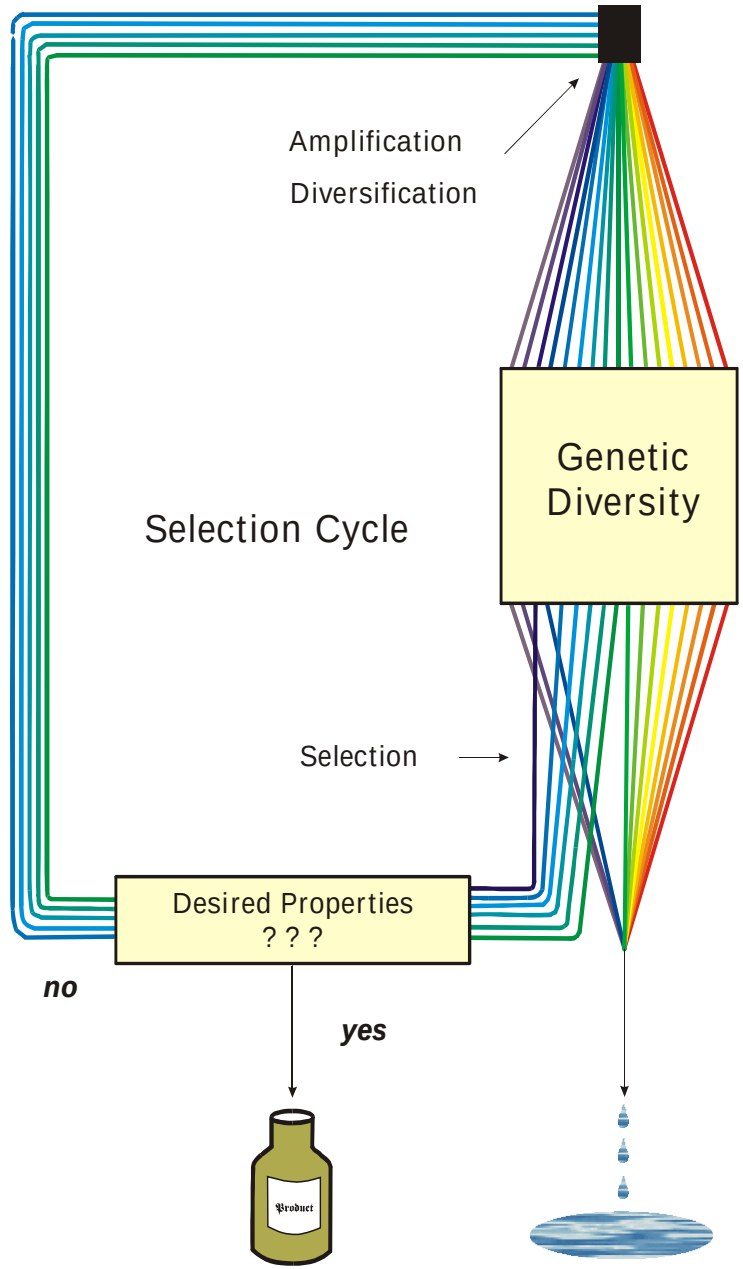
D.B.Bartel, J.W.Szostak, **In vitro** ***selection of RNA molecules that bind specific ligands***. Nature **346** (1990), 818-822

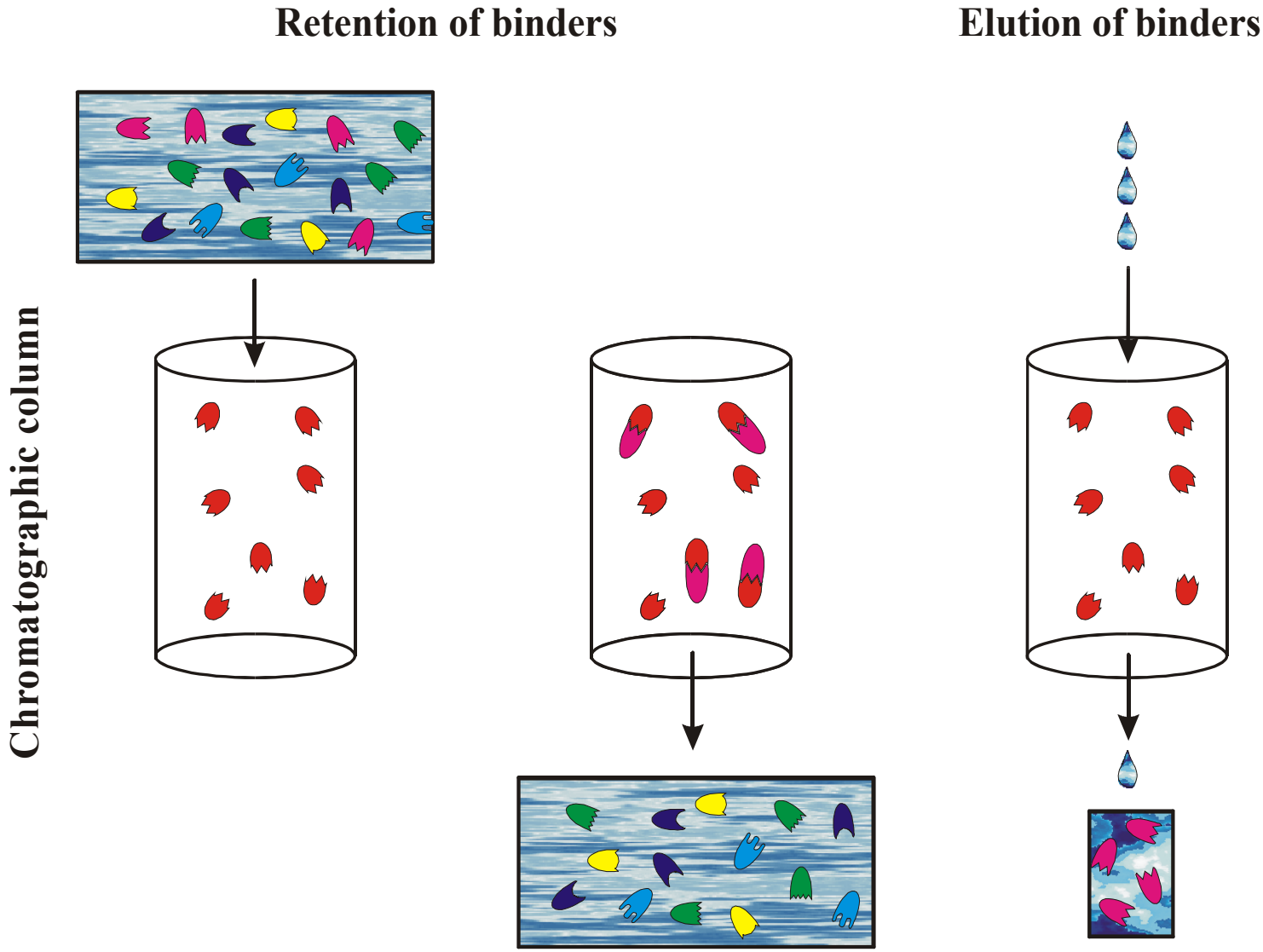
C.Tuerk, L.Gold, **SELEX** - ***Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage*** **T4** ***DNA polymerase***. Science **249** (1990), 505-510

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

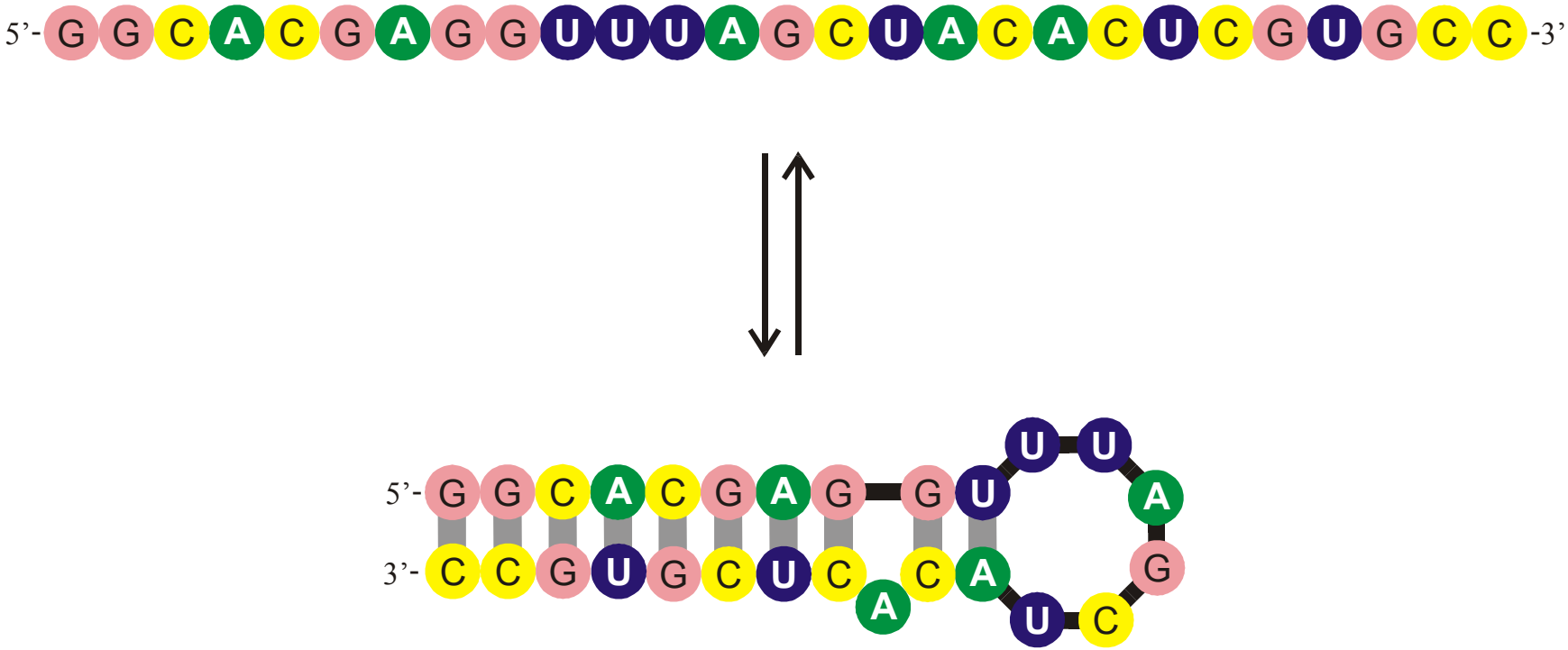
R.D.Jenison, S.C.Gill, A.Pardi, B.Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429

Selection cycle used in applied molecular evolution to design molecules with predefined properties





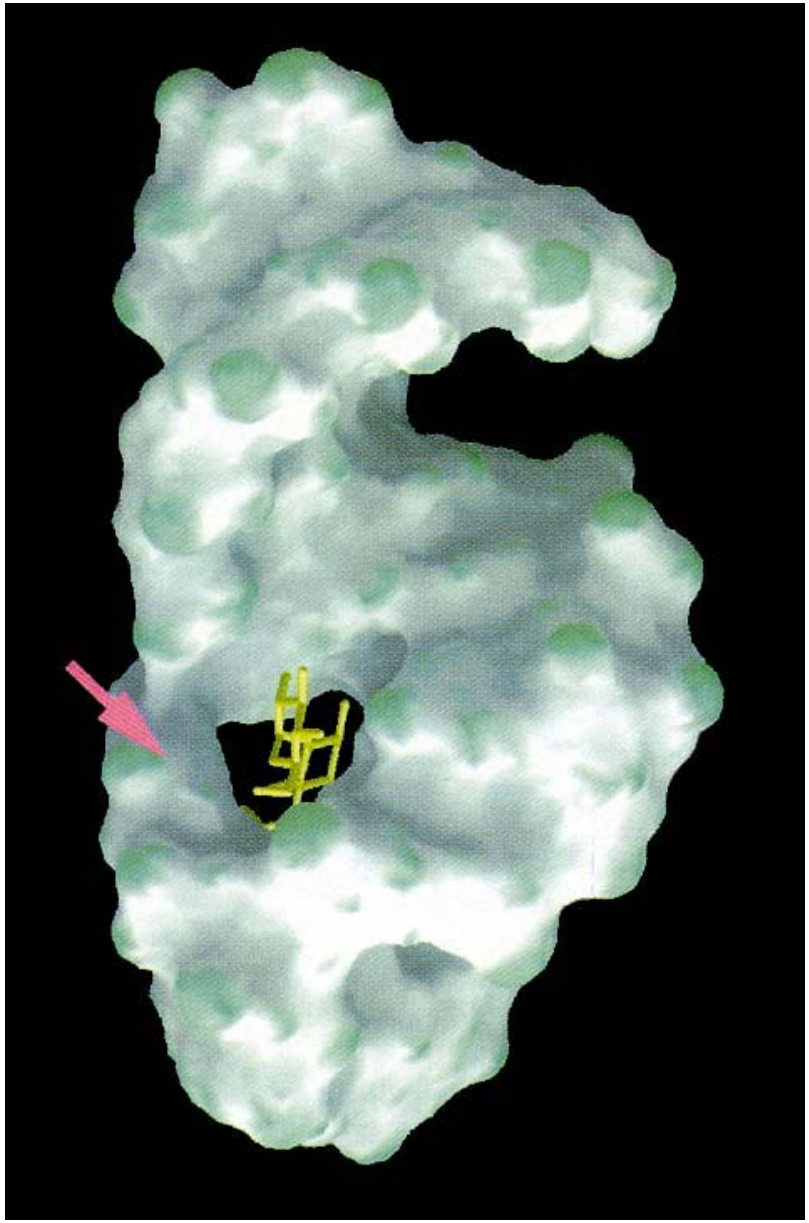
The SELEX technique for the evolutionary design of ***aptamers***



Formation of secondary structure of the tobramycin binding RNA aptamer

l = 27 ñ $4^{l} = 1.801 \text{ Ç } 10^{16}$ possible different sequences

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, 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)

A ribozyme switch

E.A.Schultes, D.B.Bartel, ***One sequence, two ribozymes: Implication for the emergence of new ribozyme folds***. Science **289** (2000), 448-452

Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.

0092-8240/97 $17.00 + 0.00

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,*

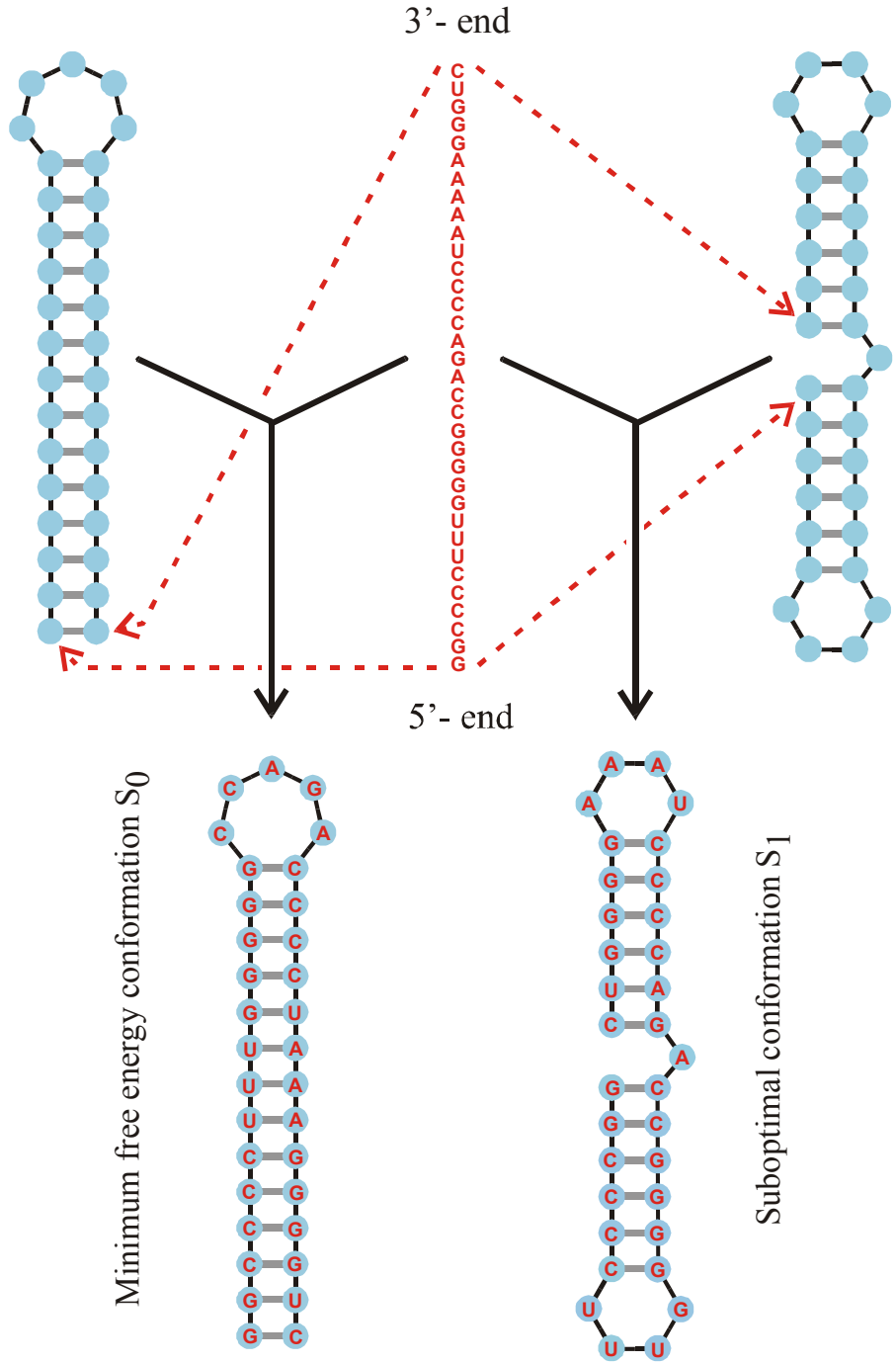
$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \varnothing.$$

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 $\gamma(s, s') \cong D_m$ 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. ■

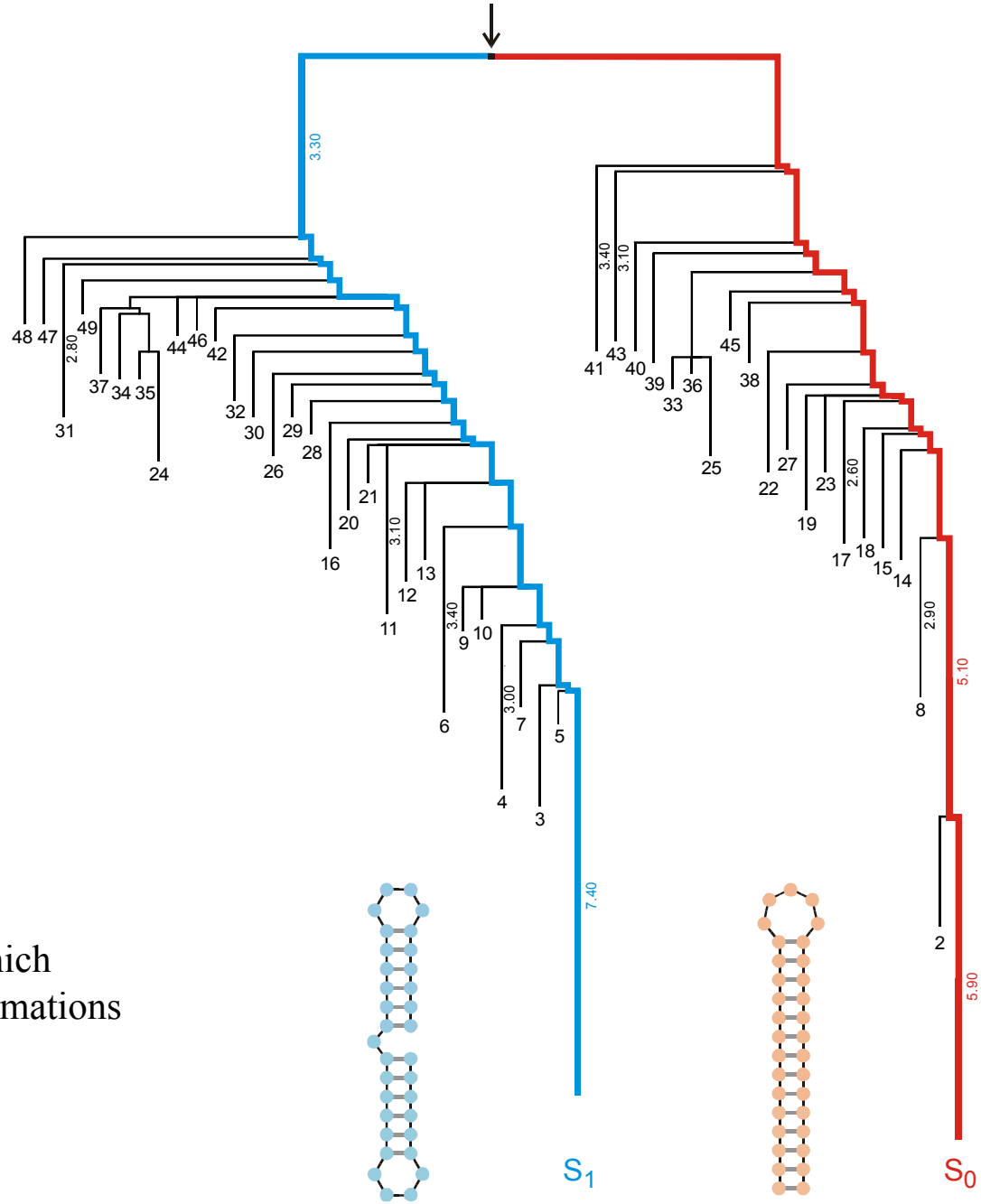
Remark. A generalization of the statement of theorem 5 to three different structures is false.

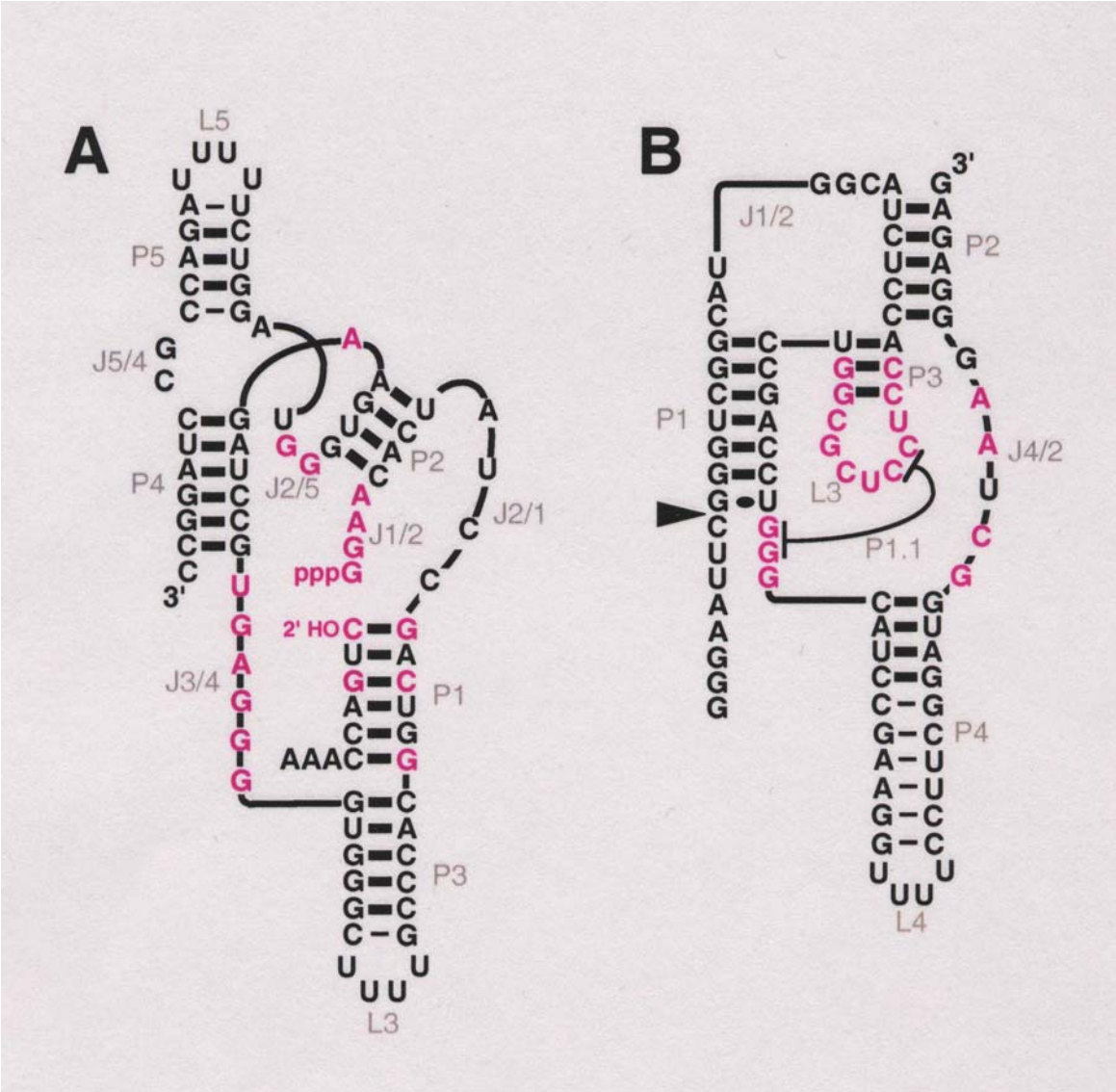
Reference for the definition of the intersection and the proof of the ***intersection theorem***

A sequence at the **intersection** of two neutral networks is compatible with both structures



Barrier tree of a sequence which
switches between two conformations

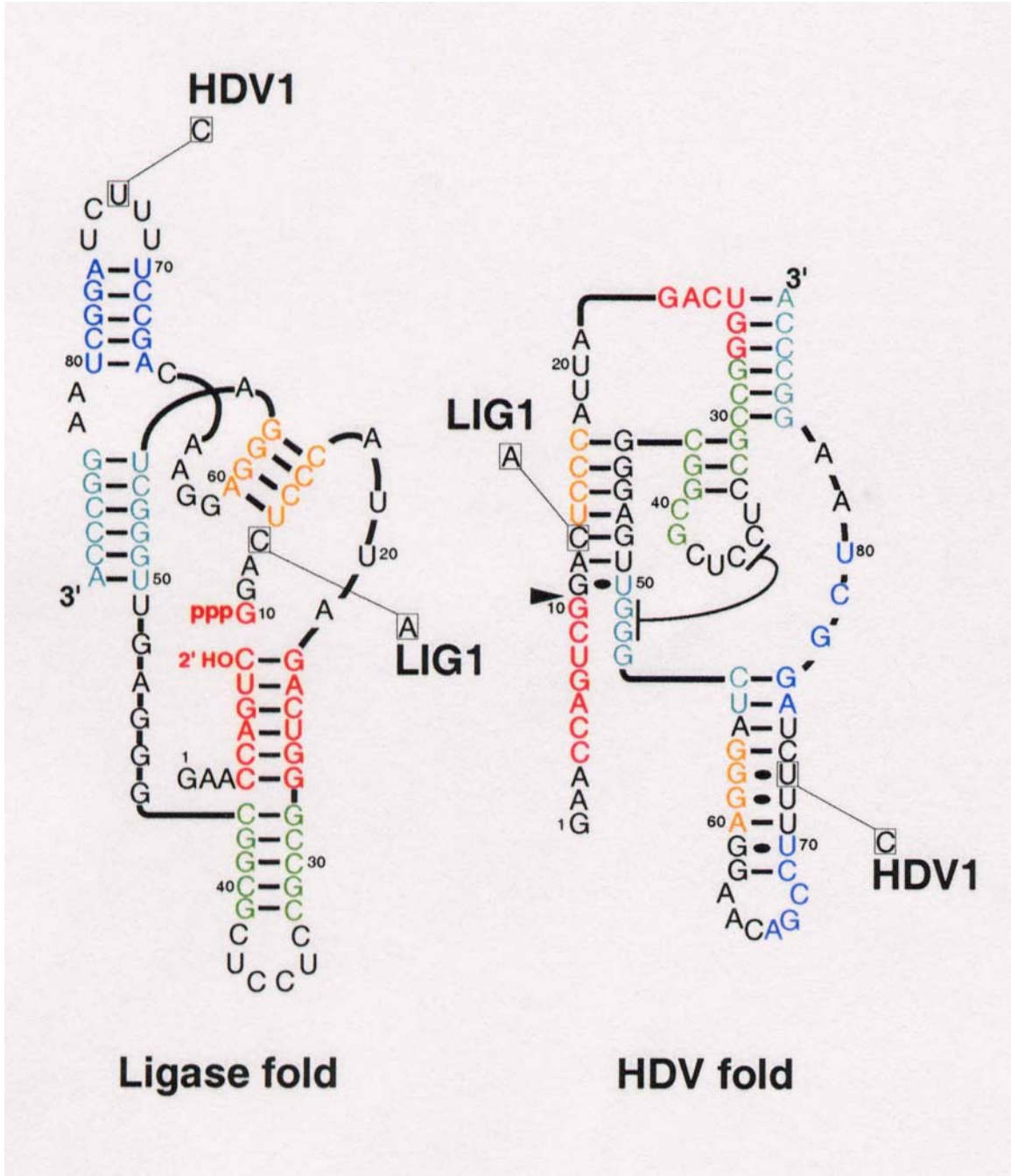


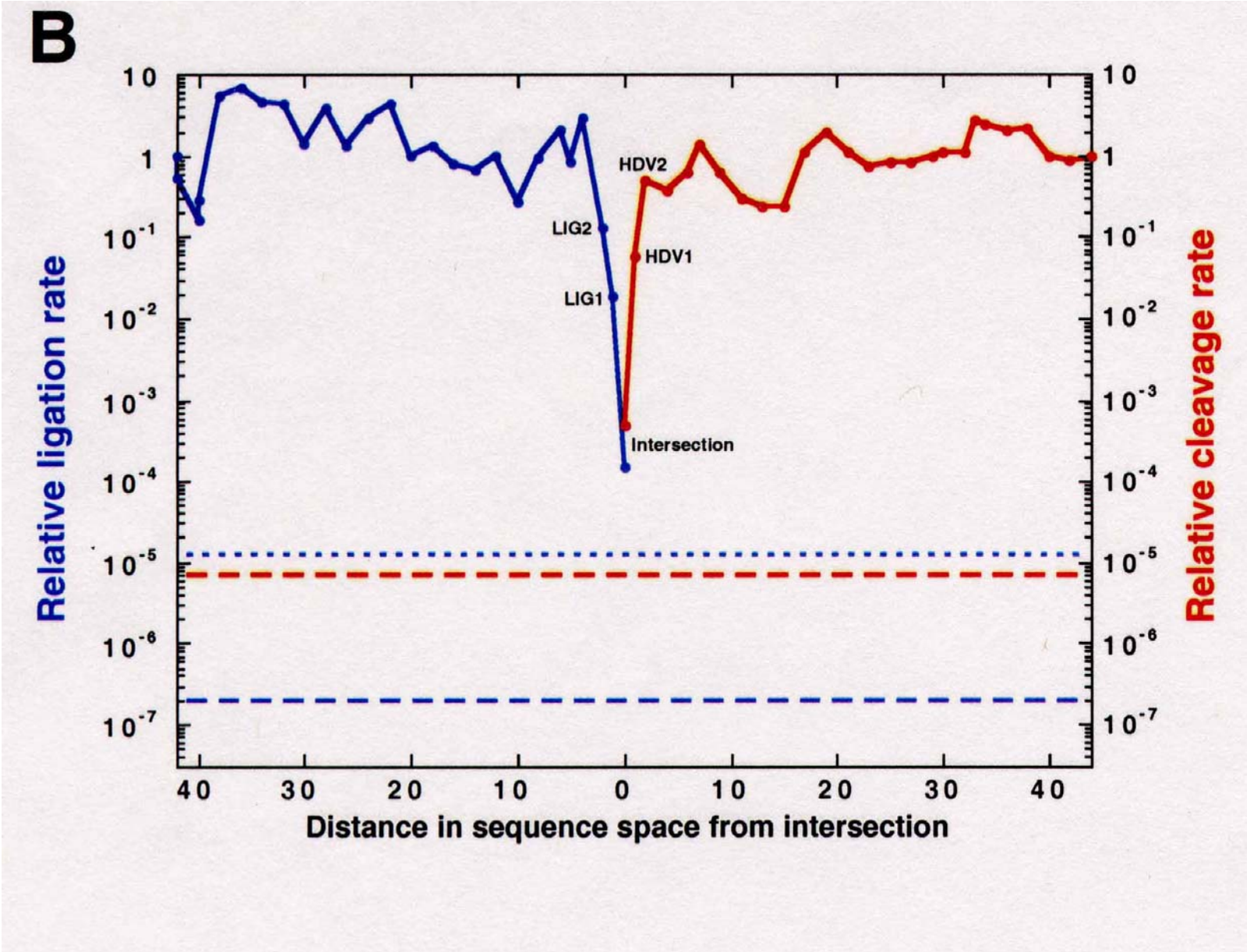


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)

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

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.

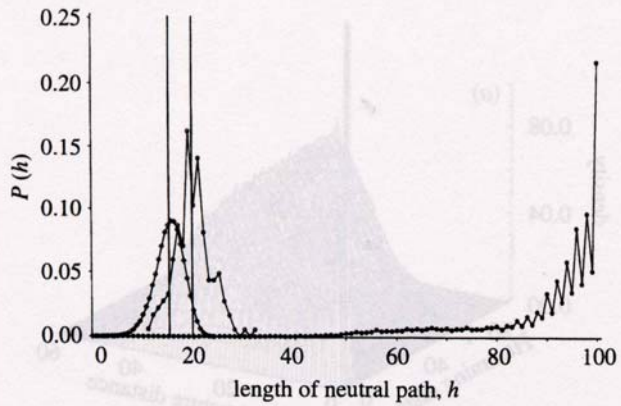


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and 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

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

Coworkers

Walter Fontana, Santa Fe Institute, NM

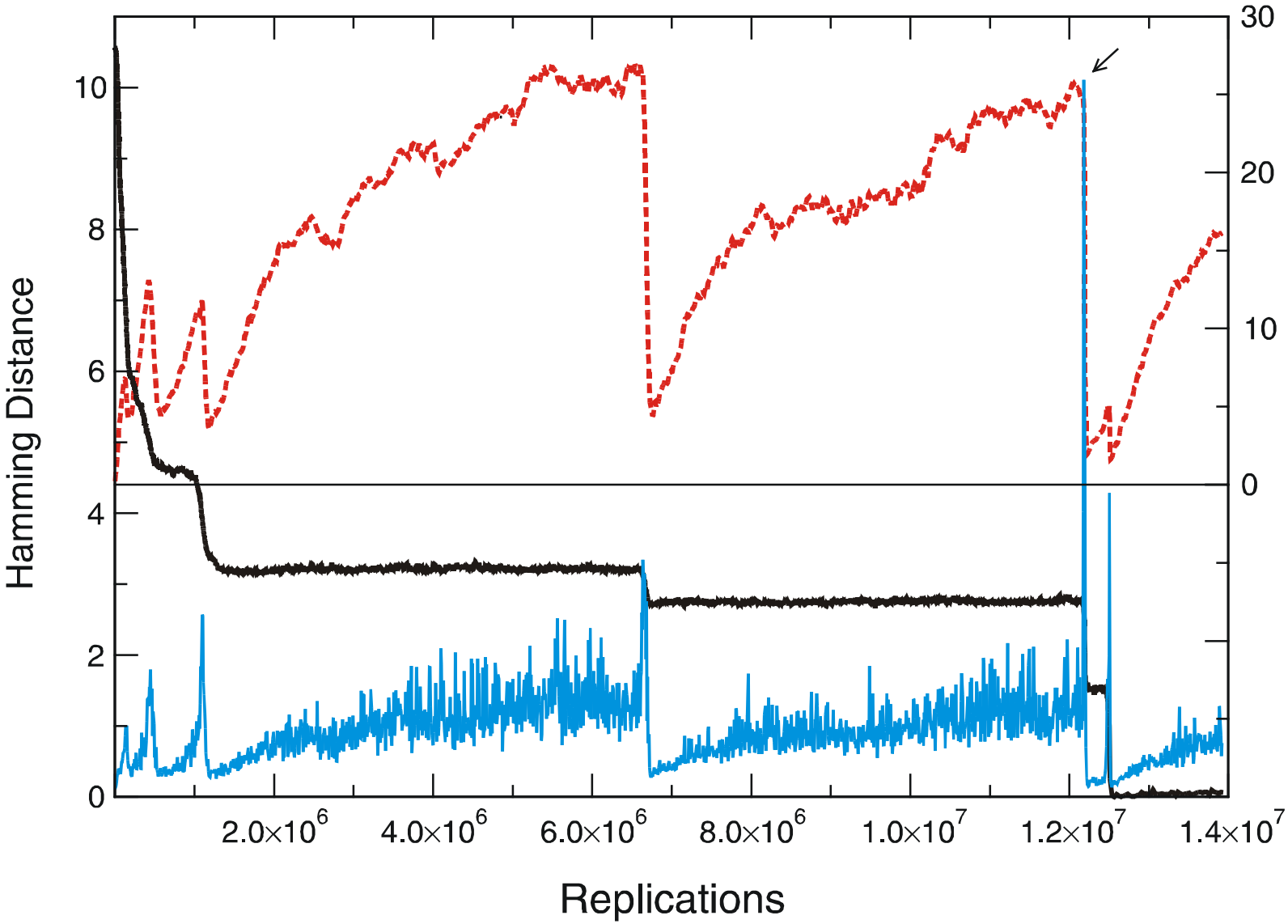
Christian Reidys, Christian Forst, Los Alamos National Laboratory, NM

Peter Stadler, Universität Leipzig, GE

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

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

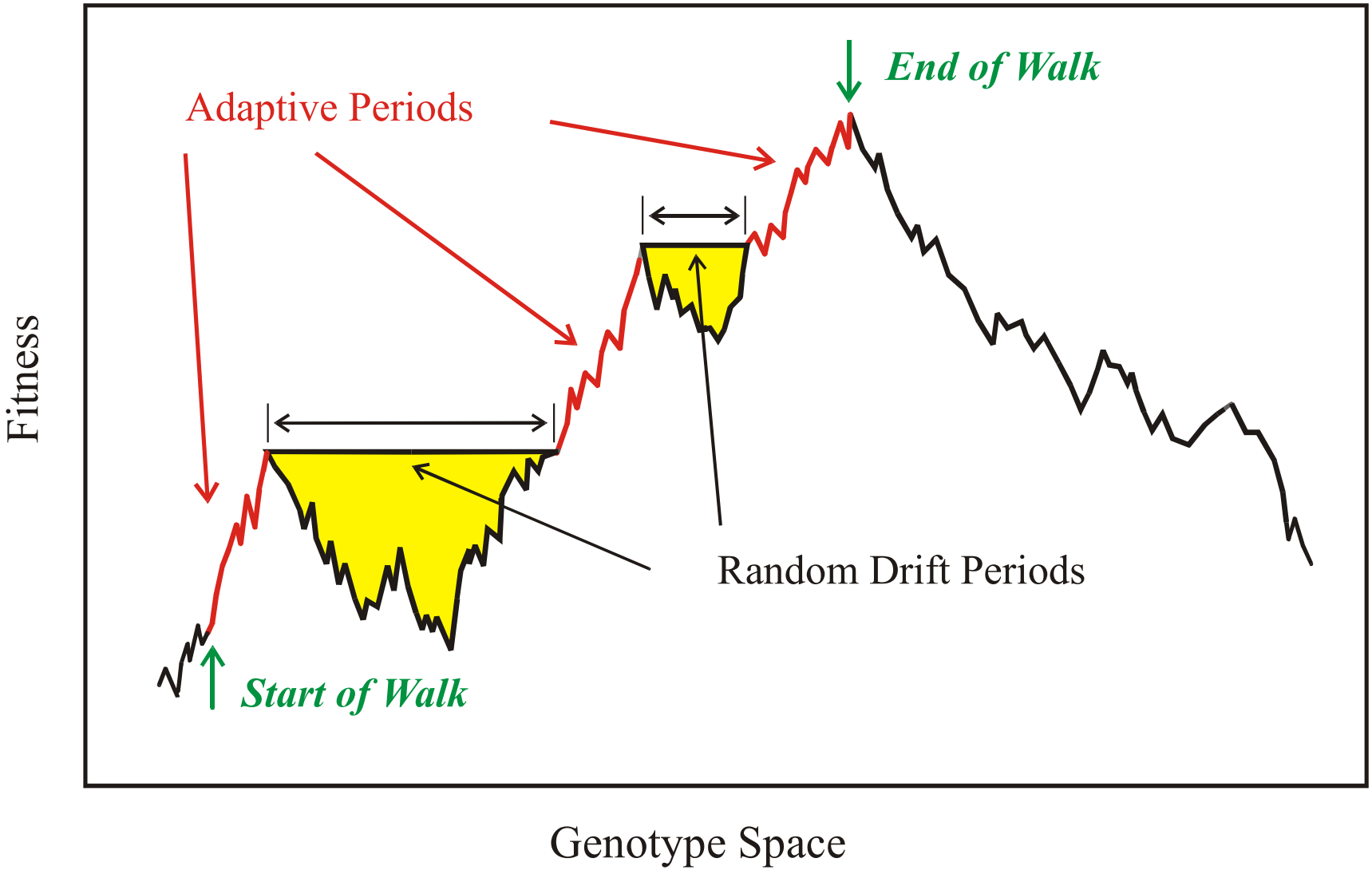
Ulrike Göbel, Institut für Molekulare Biotechnologie, Jena, GE
Walter Grüner, Stefan Kopp, Jaqueline Weber



Variation in genotype space during optimization of phenotypes

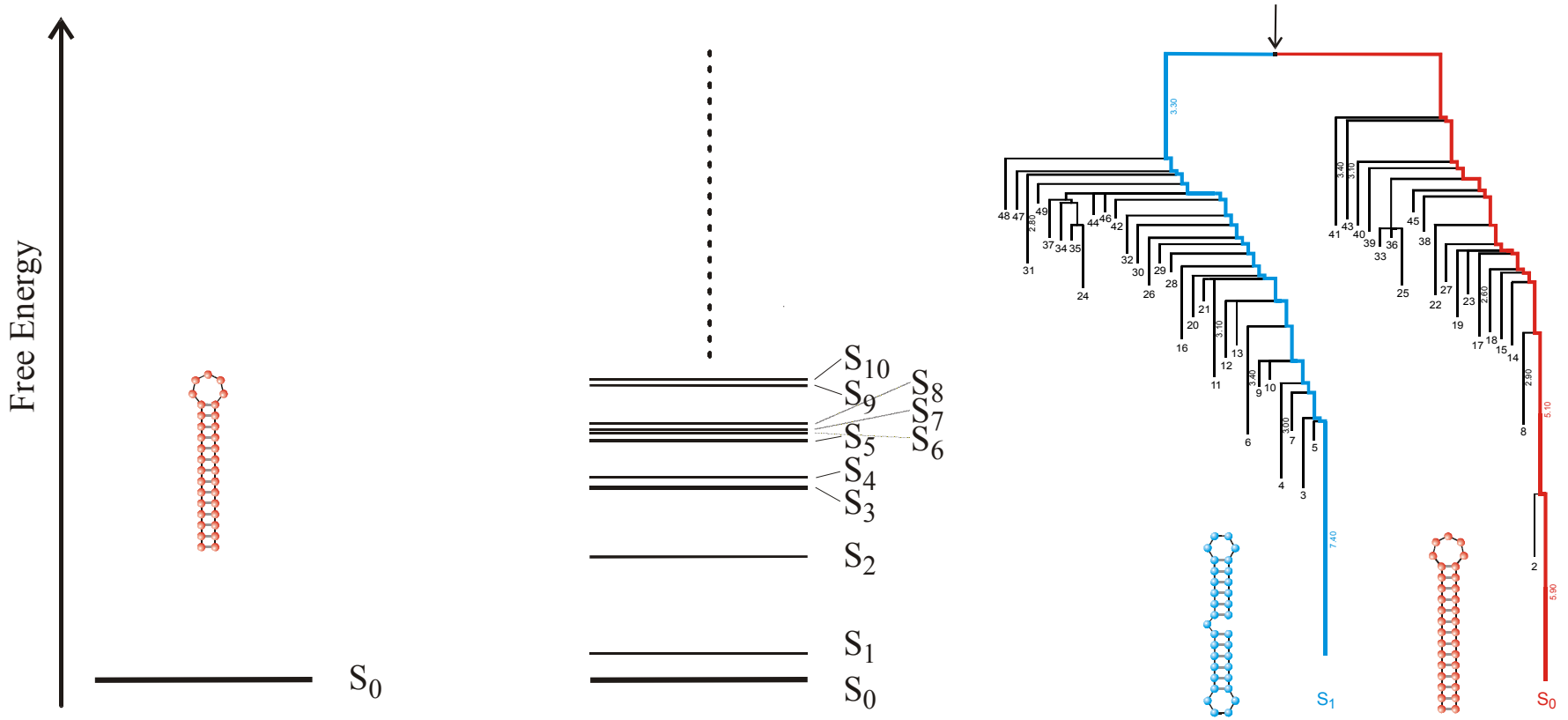
„...Variations neither useful not injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions. ...“

Charles Darwin, Origin of species (1859)



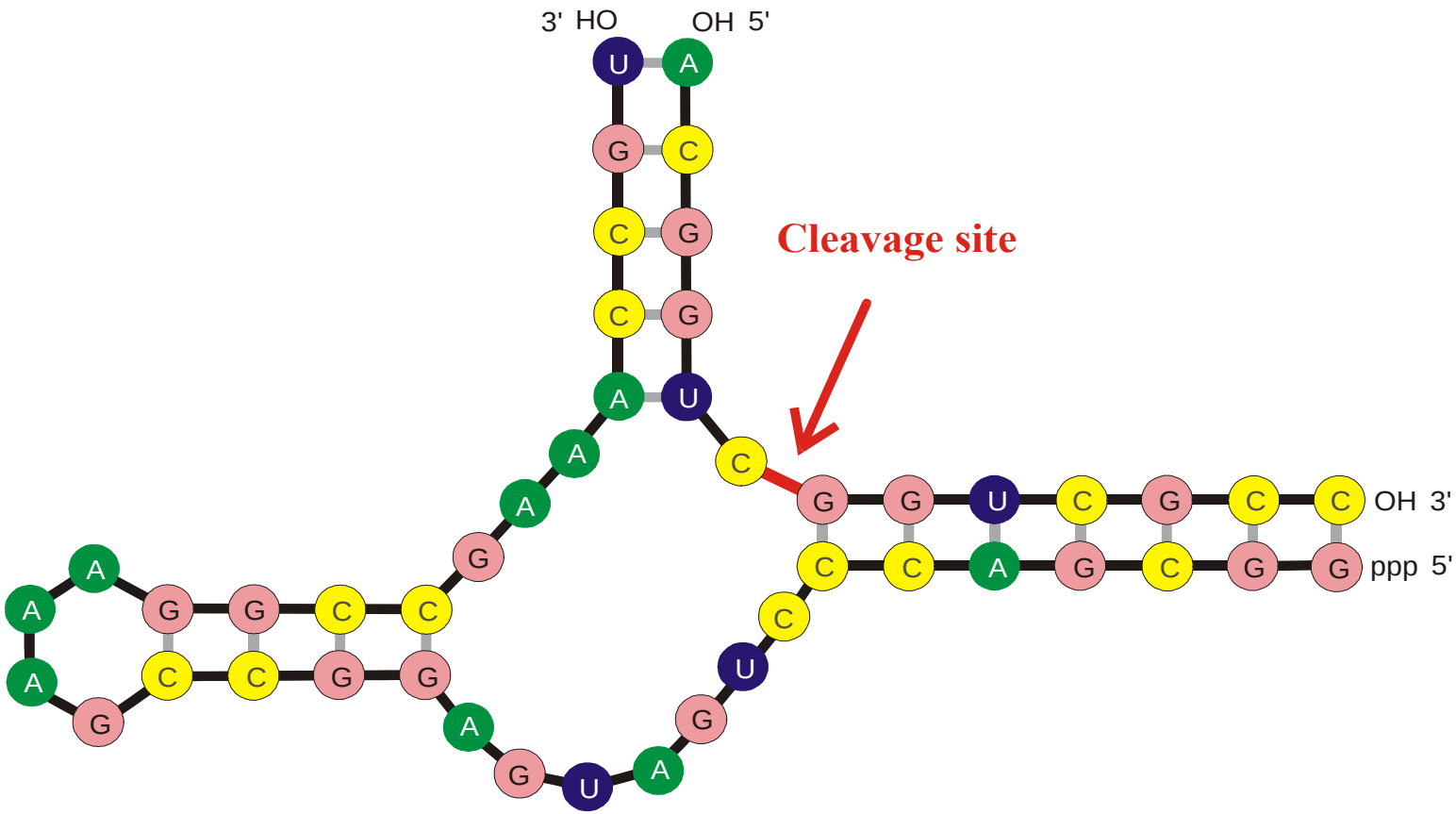
Evolution in genotype space sketched as a non-descending walk in a fitness landscape

T = 0 K , t Á ¸ T > 0 K , t Á ¸ T > 0 K , t finite



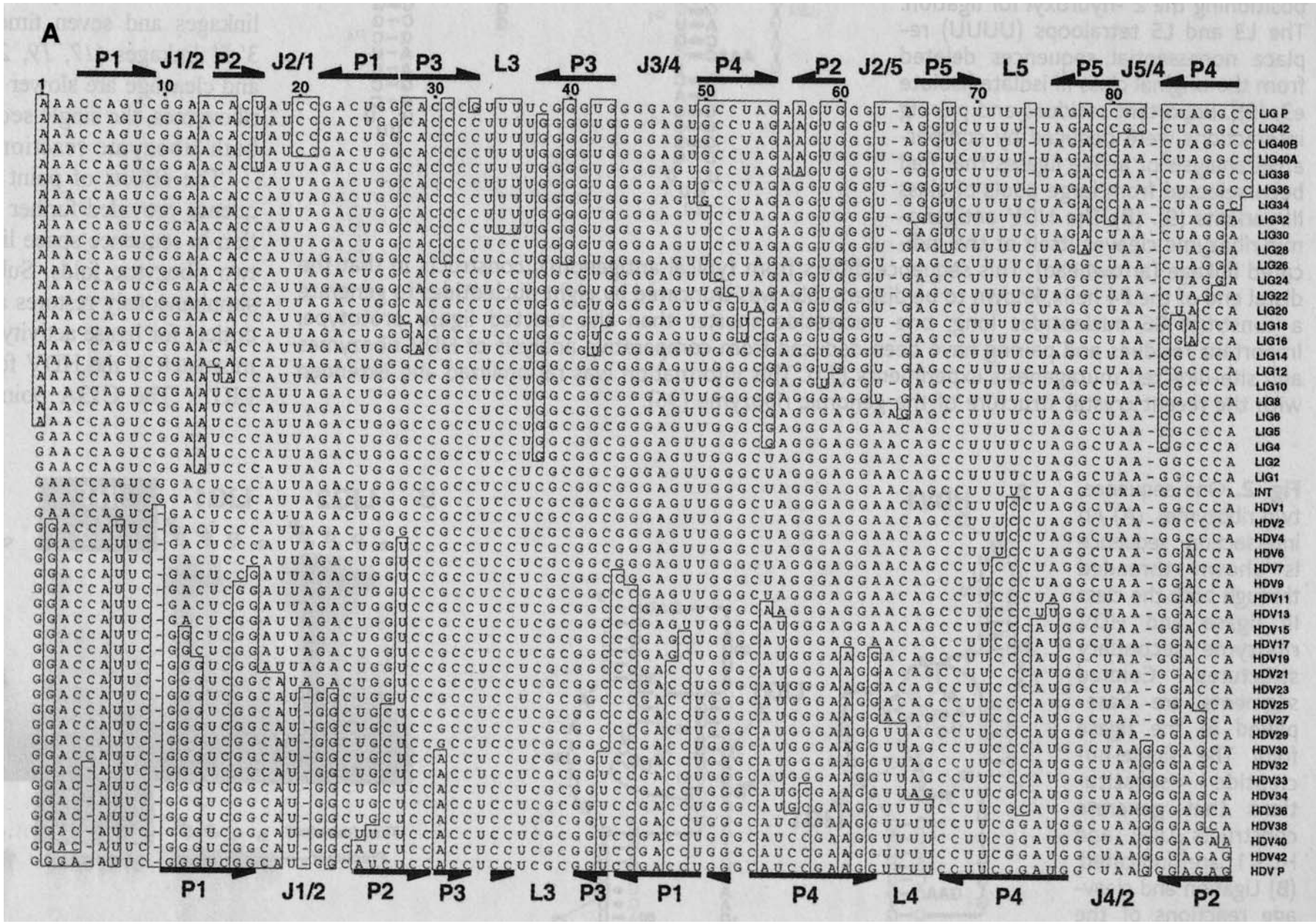
Minimum Free Energy Structure Suboptimal Structures Kinetic Structures

Different notions of RNA structure including suboptimal conformations



The "hammerhead" ribozyme

The smallest known catalytically active RNA molecule



Sequence of mutants from the intersection to both reference ribozymes