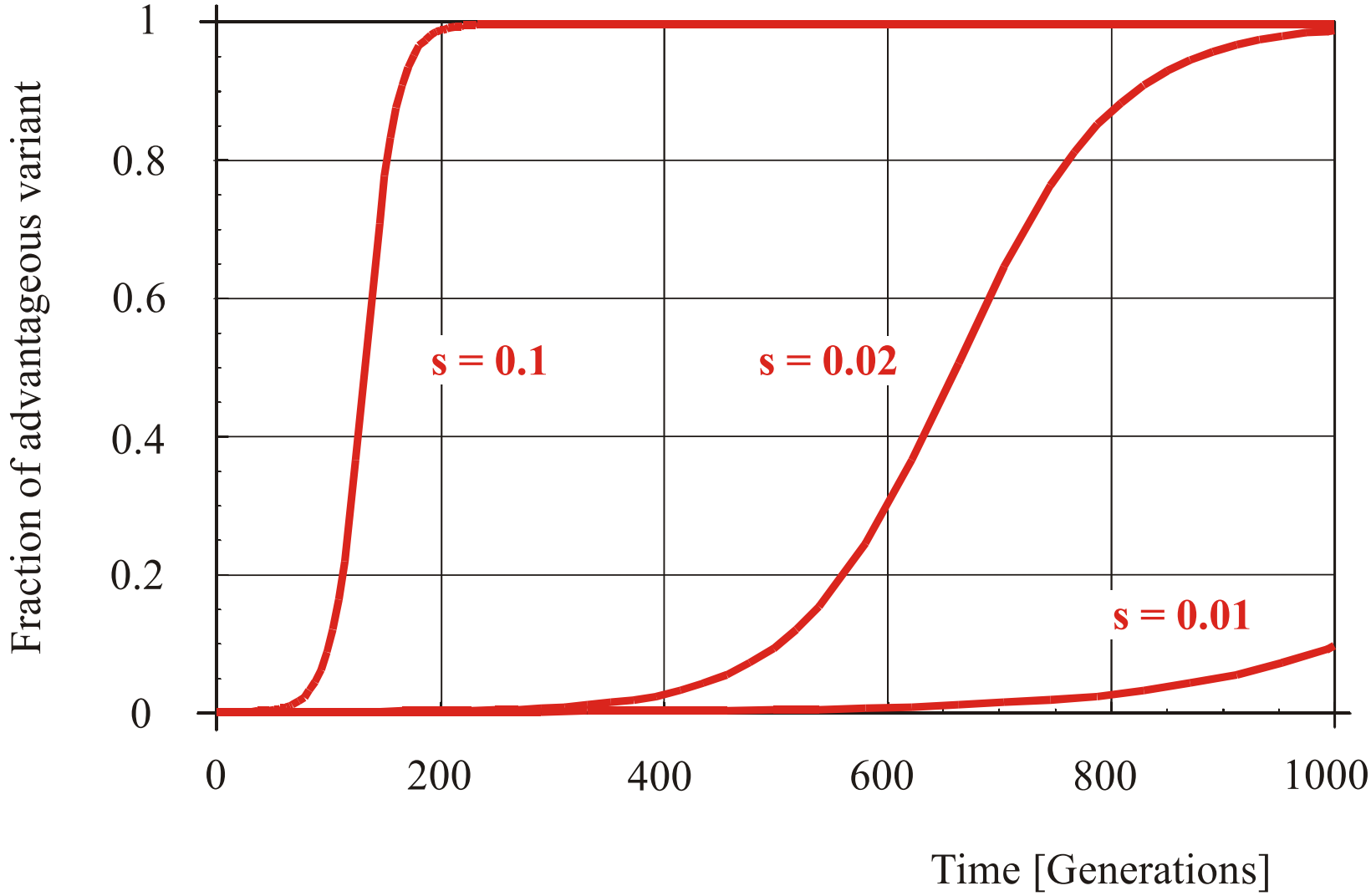
The role of genotype-phenotype mappings in adaptation

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Concepts for complex adaptive systems

Delmenhorst, 23.03.2002



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

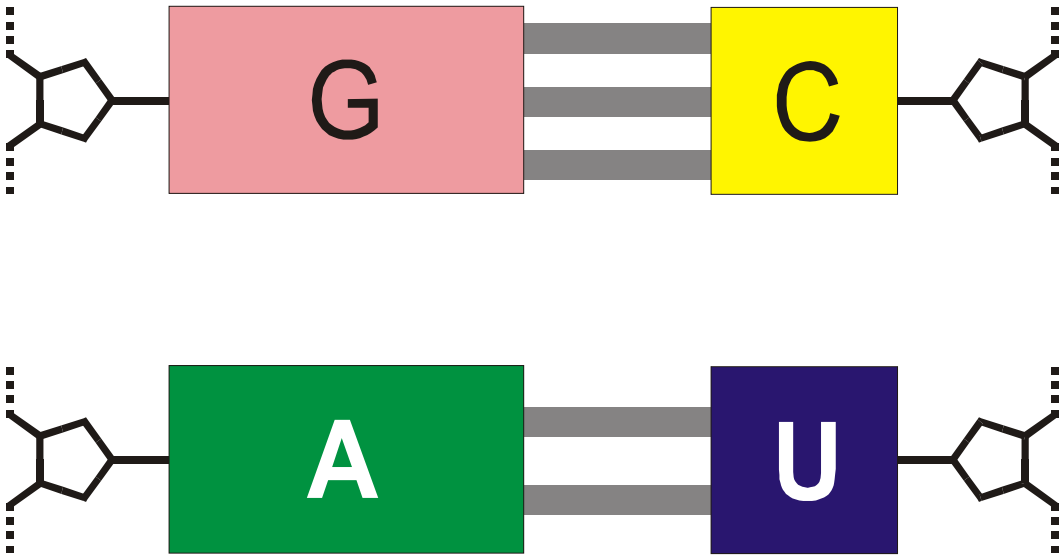
$4^{27} = 1.801 \text{ Ç } 10^{16}$ possible different sequences

Combinatorial diversity of sequences: $N = 4^{\{}$

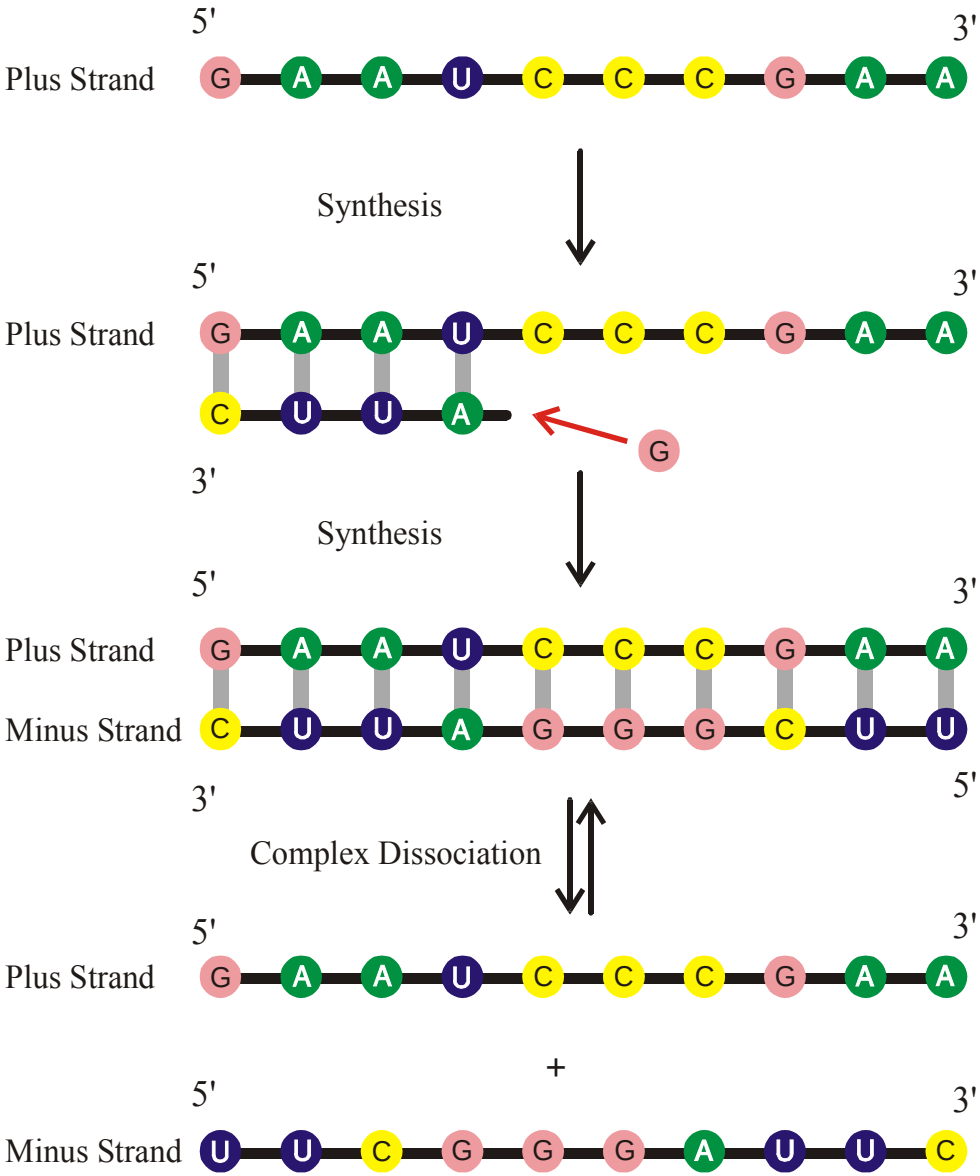
A = adenylate
U = uridylate
C = cytidylate
G = guanylate

Combinatorial diversity of heteropolymers illustrated by means of an RNA aptamer that binds to the antibiotic tobramycin

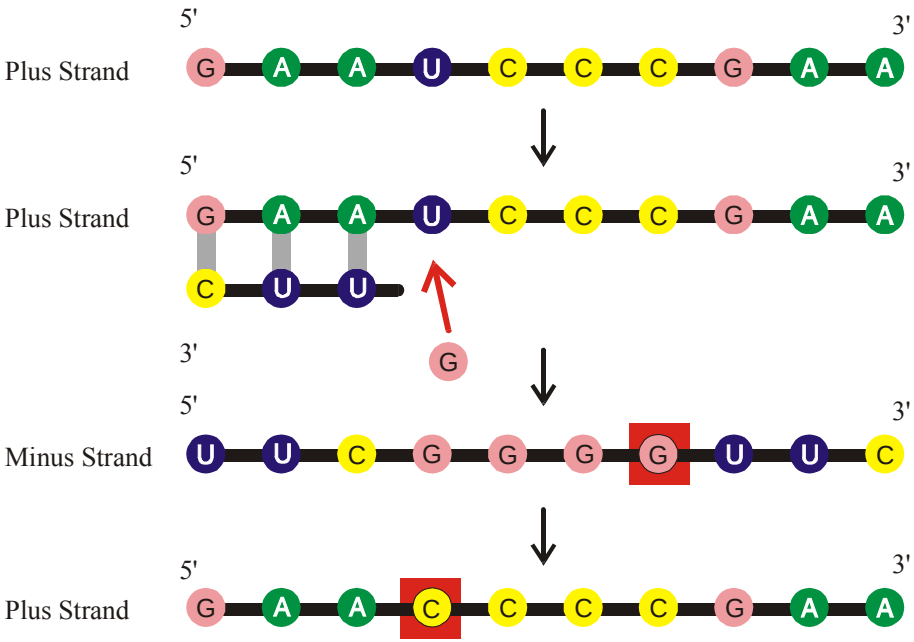
Hydrogen bonds



Hydrogen bonding between nucleotide bases is the principle of template action of RNA and DNA



Complementary replication as the simplest copying mechanism of RNA



Point Mutation

GAAUCCCGAA → GAAUCCCGUCCCGAA

Insertion

GAAUCCCGAA → GAAUCCGA

Deletion

Mutations represent the mechanism of variation in nucleic acids

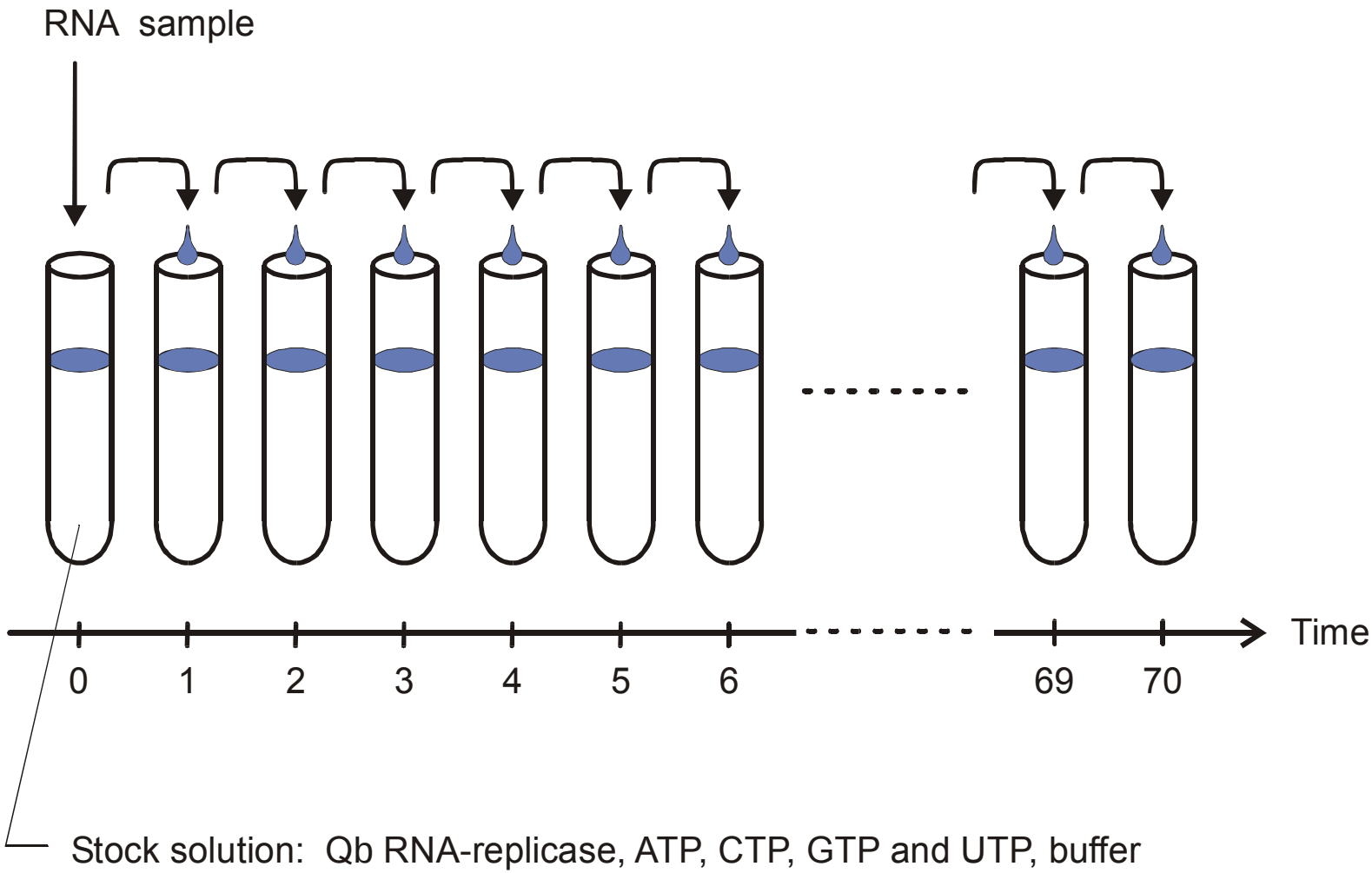
Evolution of RNA molecules based on Qβ phage

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

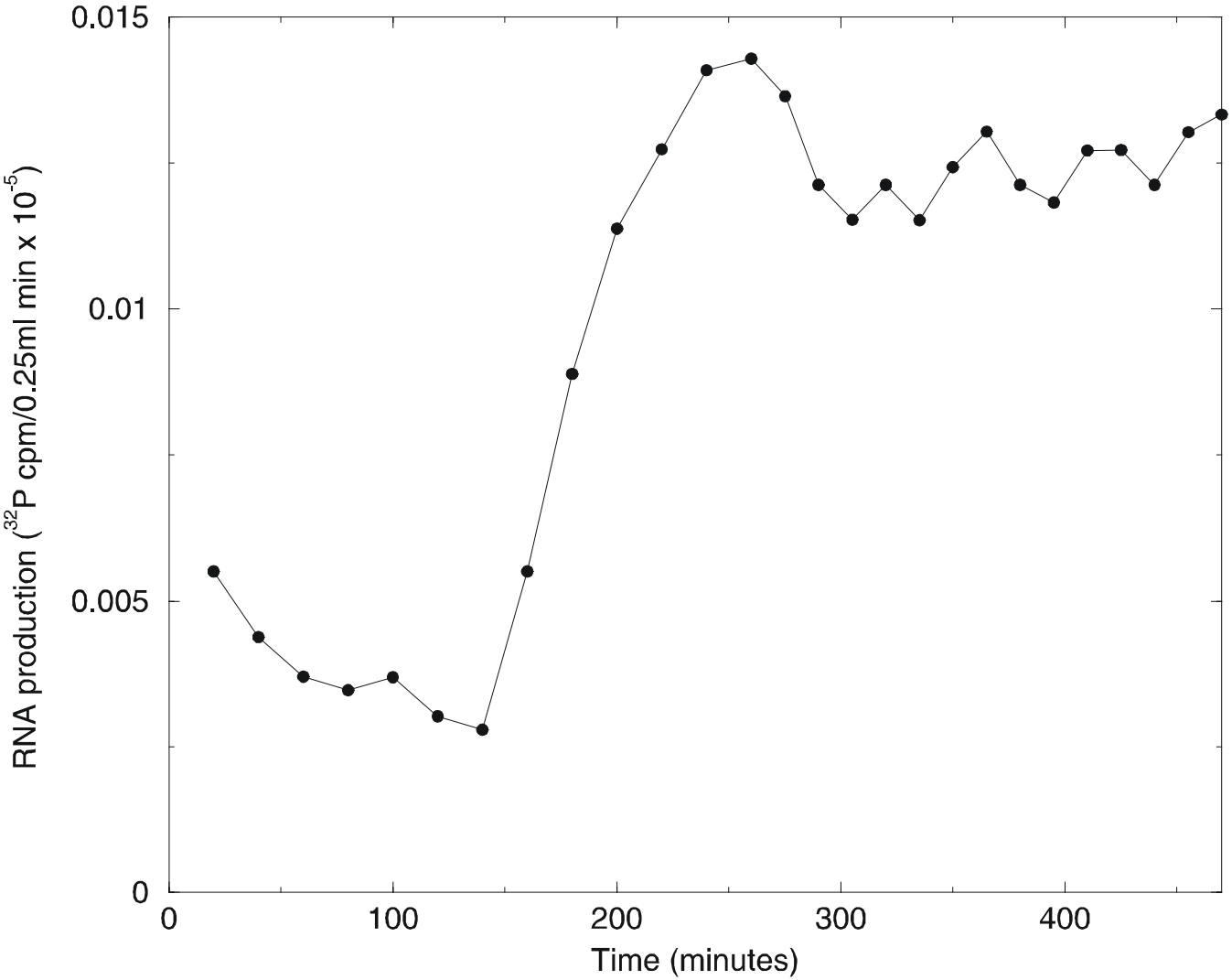
S.Spiegelman, ***An approach to the experimental analysis of precellular evolution***. Quart.Rev.Biophys. **4** (1971), 213-253

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



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



The increase in RNA production rate during a serial transfer experiment

Evolutionary design of RNA molecules

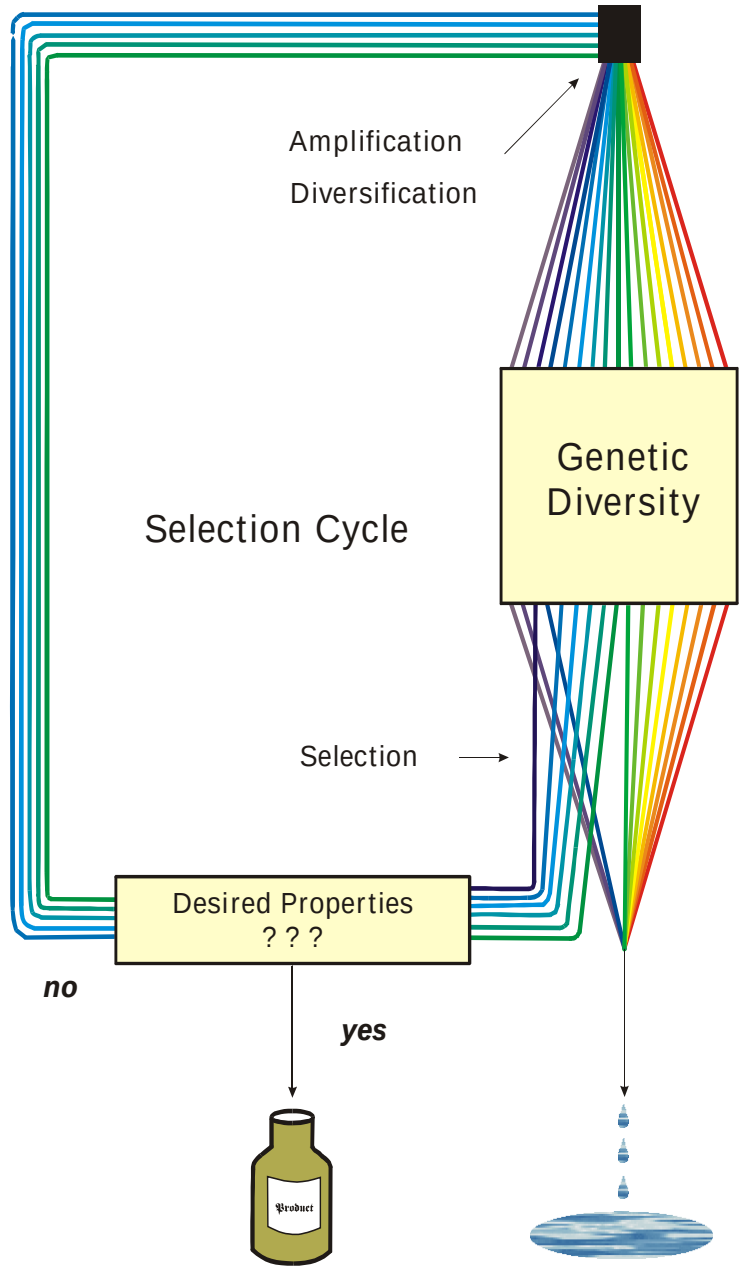
D.B.Bartel, J.W.Szostak, **In vitro** ***selection of RNA molecules that bind specific ligands***. Nature **346** (1990), 818-822

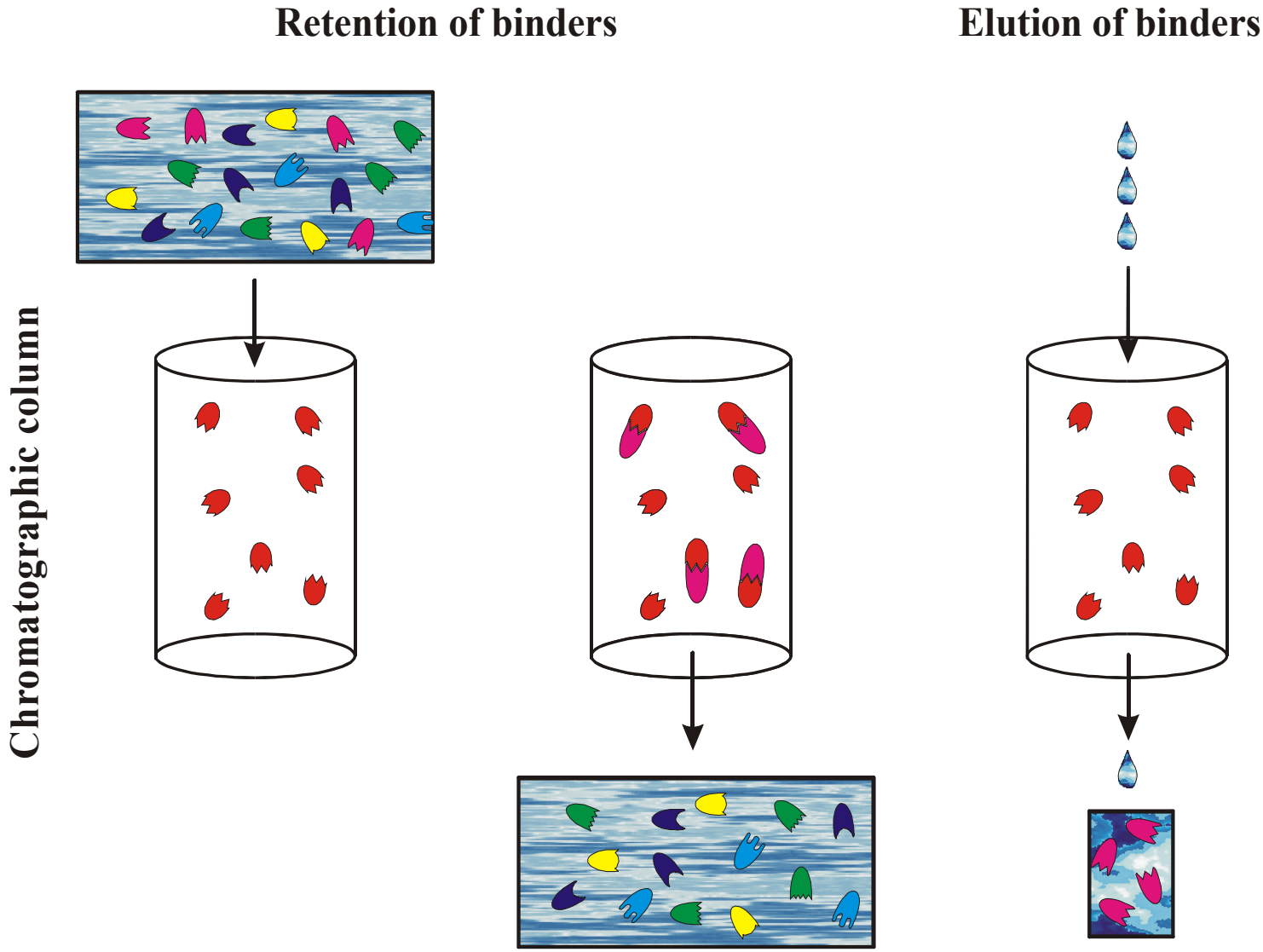
C.Tuerk, L.Gold, **SELEX** - ***Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage*** **T4** ***DNA polymerase***. Science **249** (1990), 505-510

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

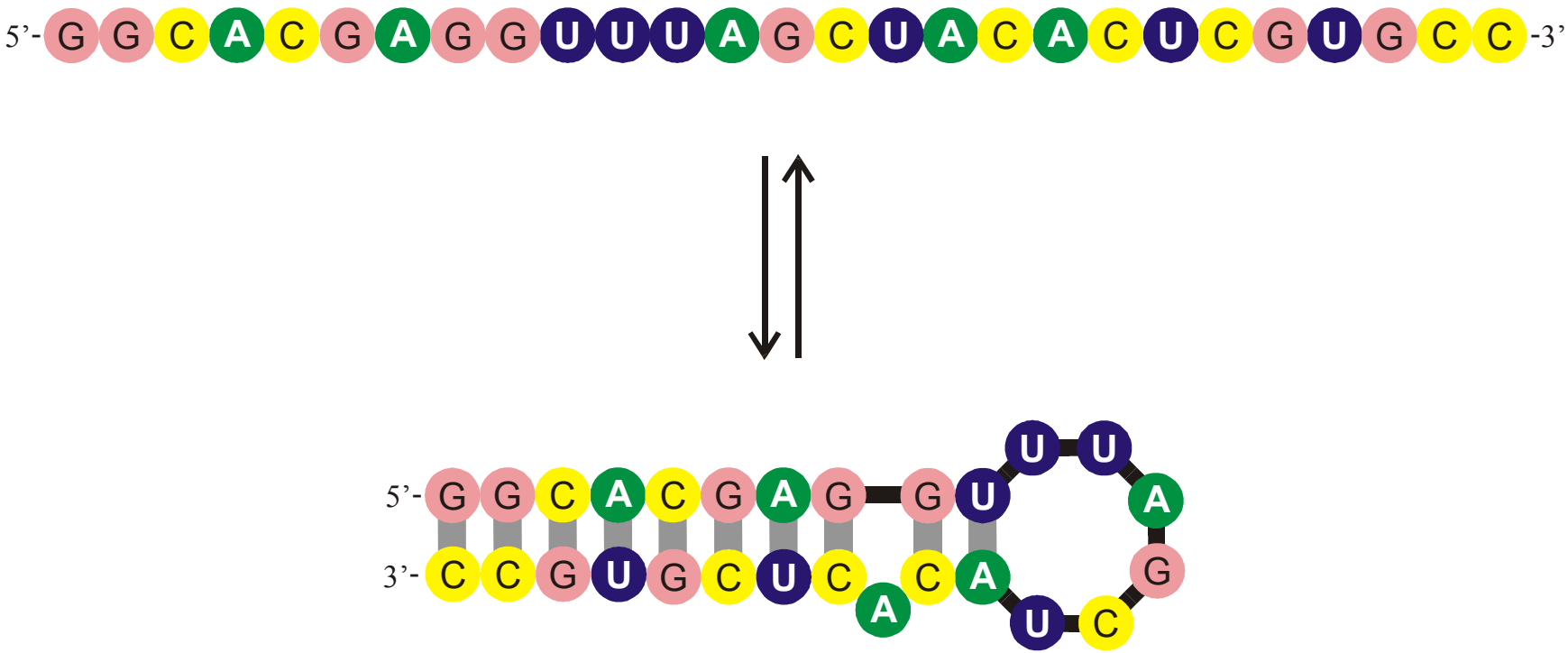
R.D.Jenison, S.C.Gill, A.Pardi, B.Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429

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



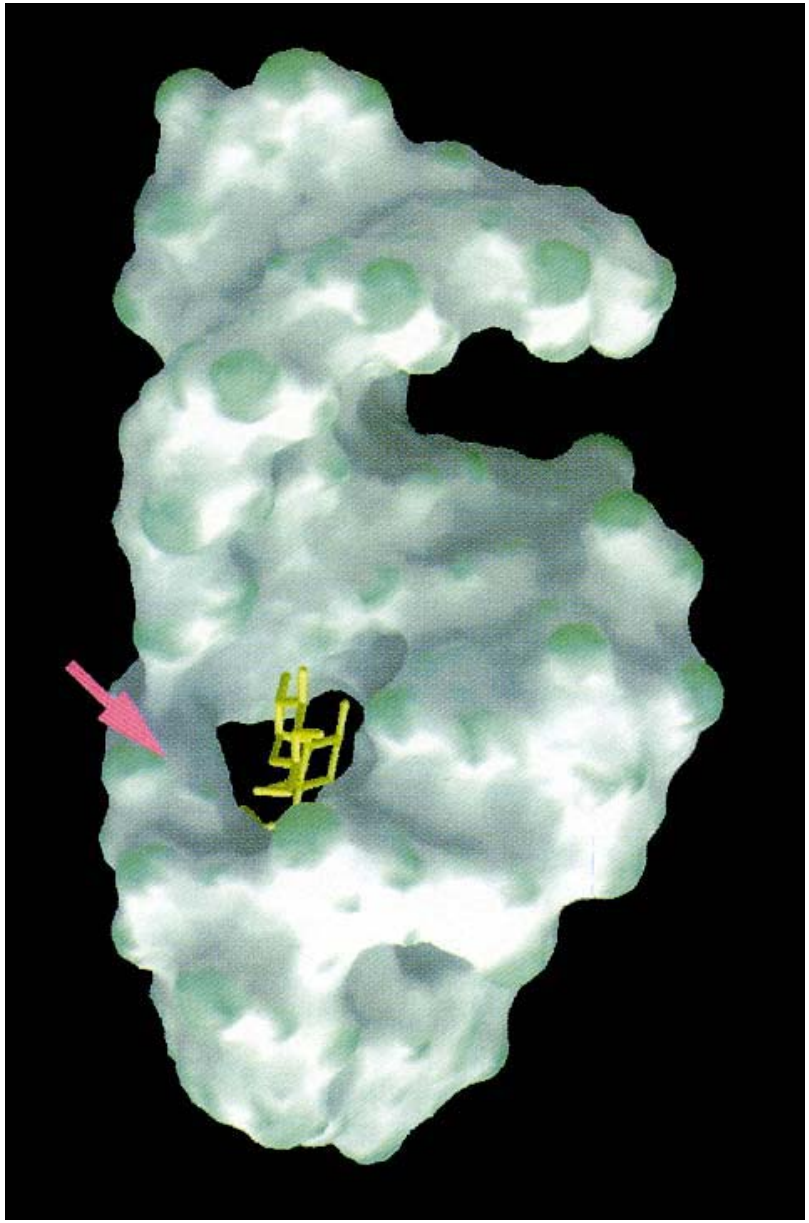


The SELEX technique for the evolutionary design of ***aptamers***



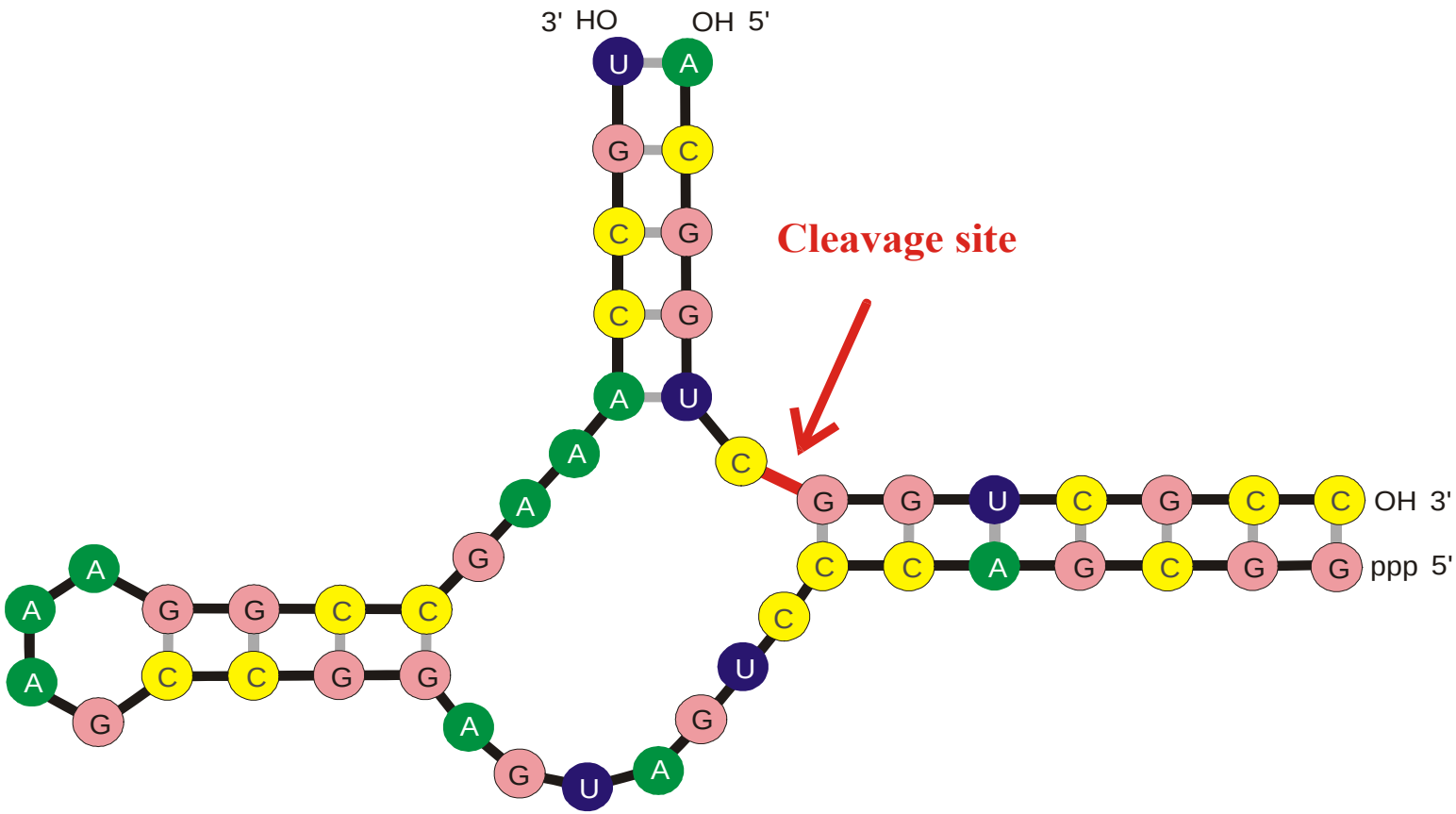
Formation of secondary structure of the tobramycin binding RNA aptamer

L. 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)

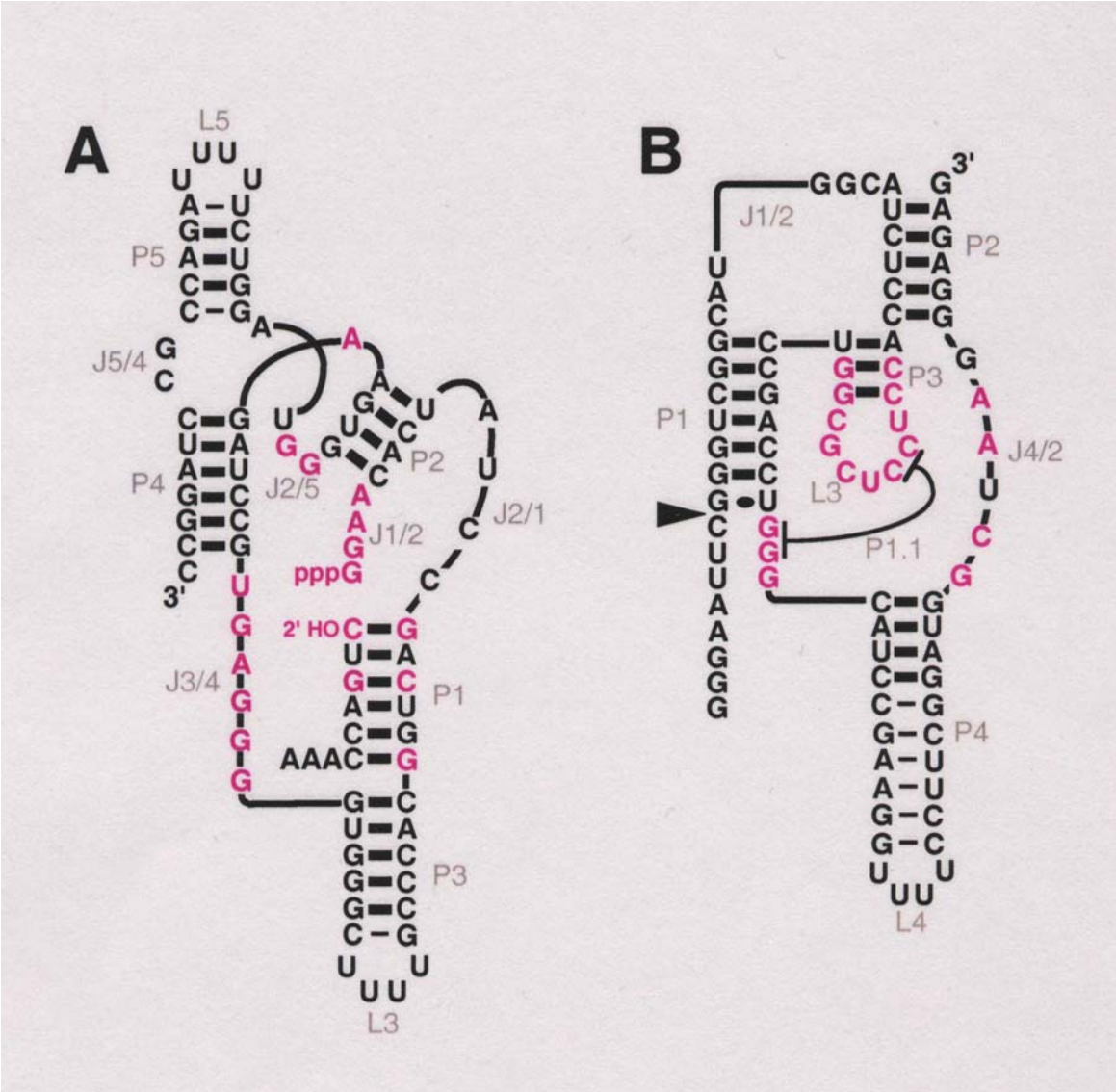


The "hammerhead" ribozyme

The smallest known catalytically active RNA molecule

A ribozyme switch

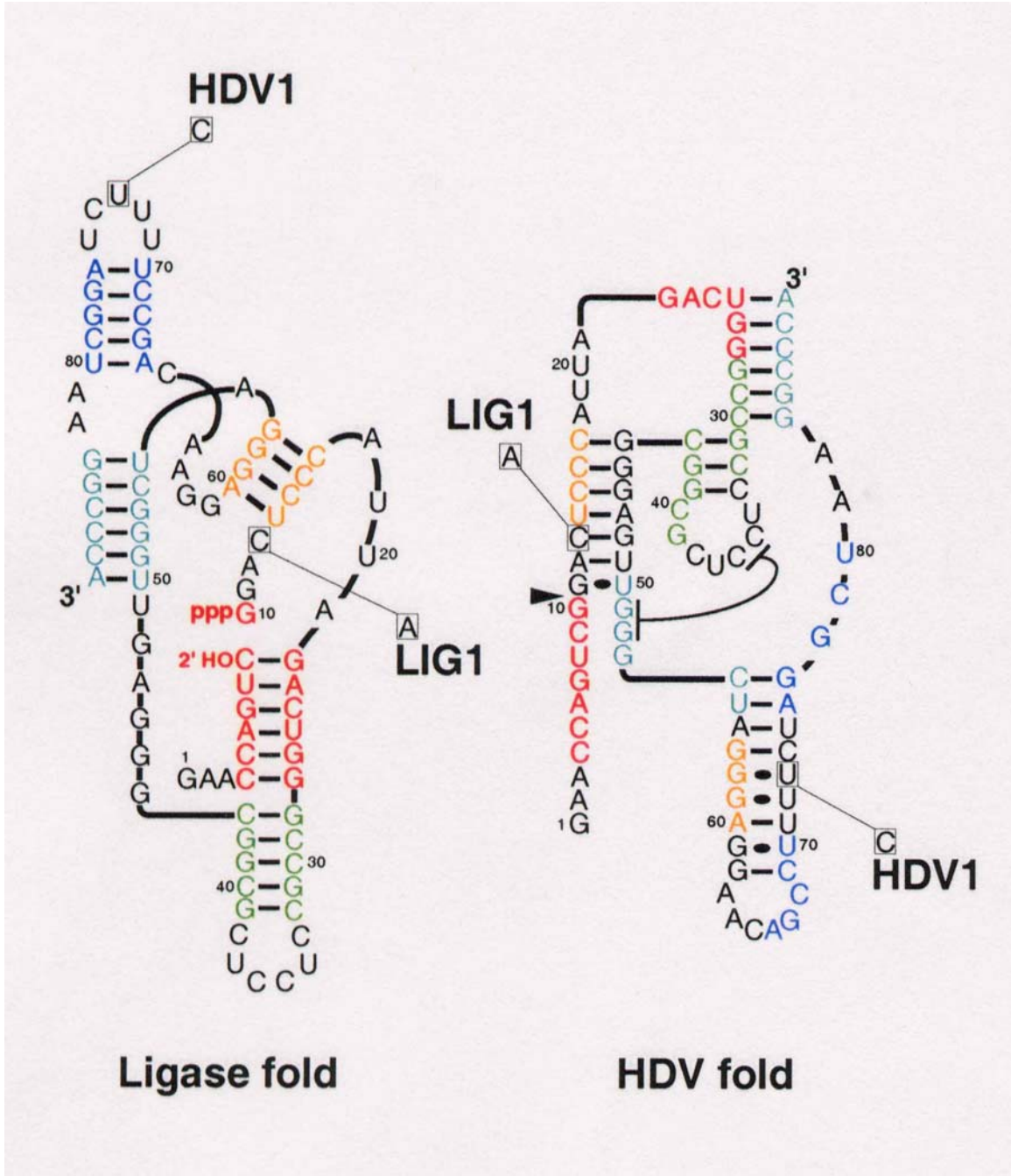
E.A.Schultes, D.B.Bartel, ***One sequence, two ribozymes: Implication for the emergence of new ribozyme folds***. Science **289** (2000), 448-452



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



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*

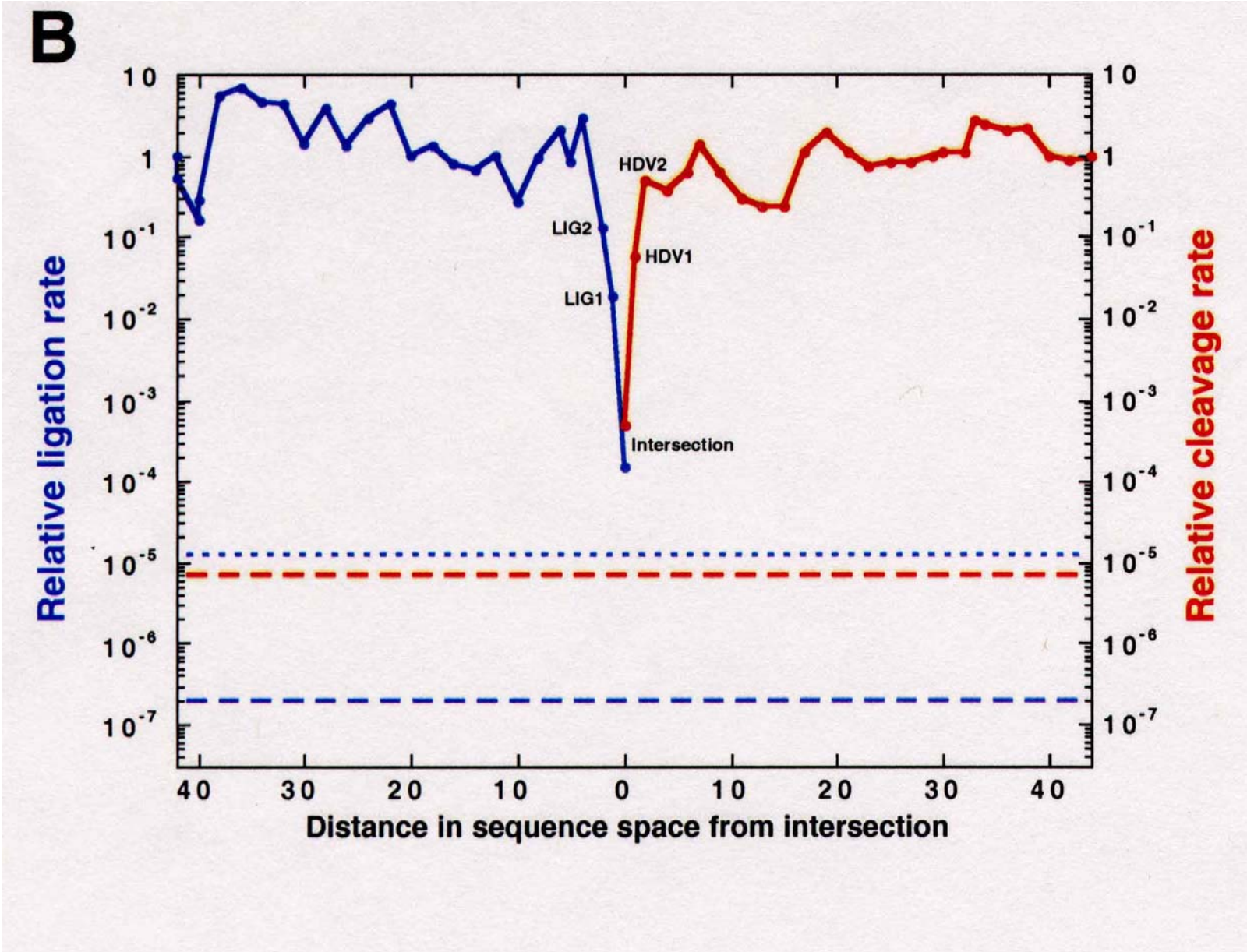
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* **C**[s], **C**[s′] *their corresponding compatible sequences. Then,*

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

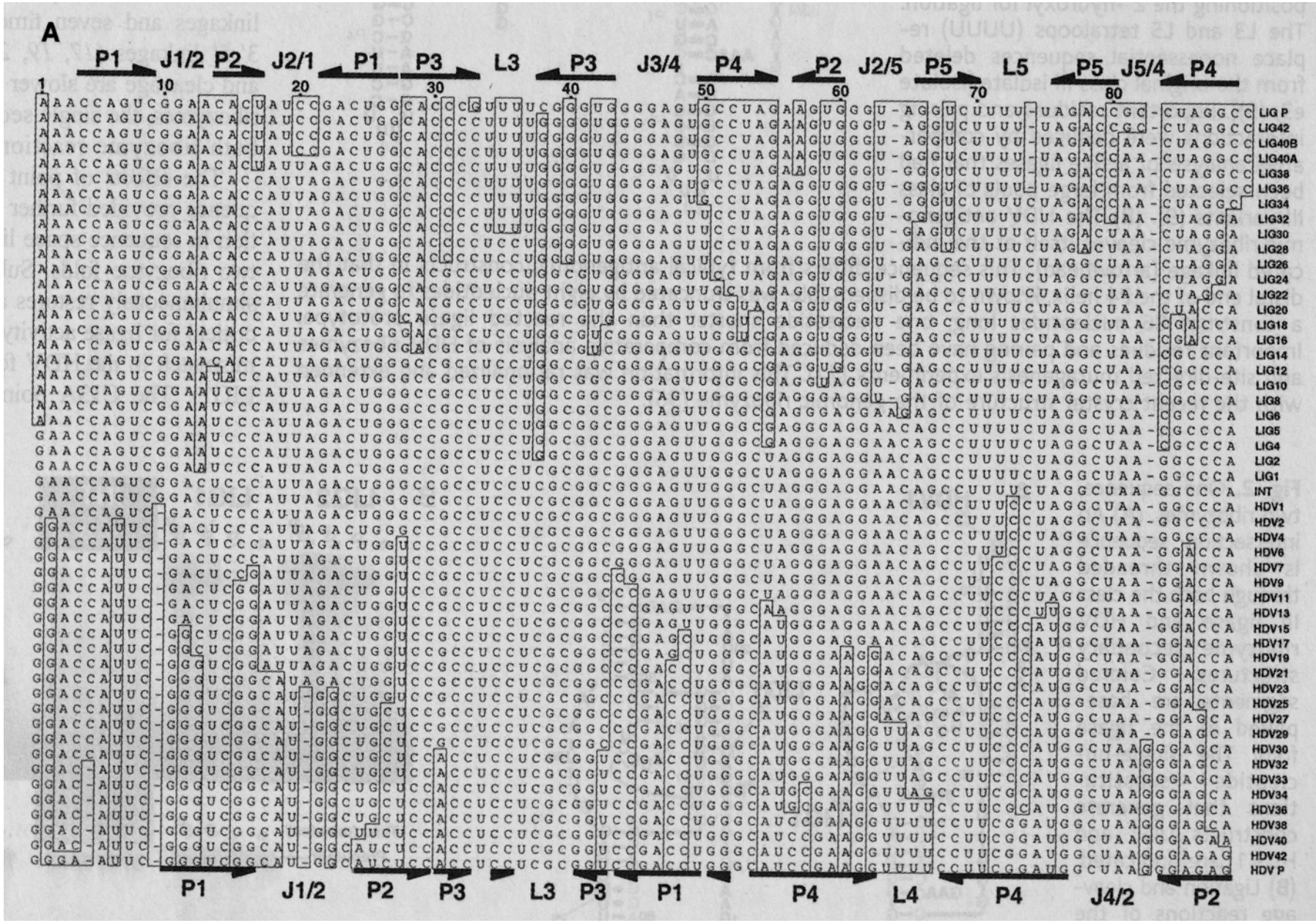
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\gamma(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles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



Two neutral walks through sequence space with conservation of structure and catalytic activity



Sequence of mutants from the intersection to both reference ribozymes

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.

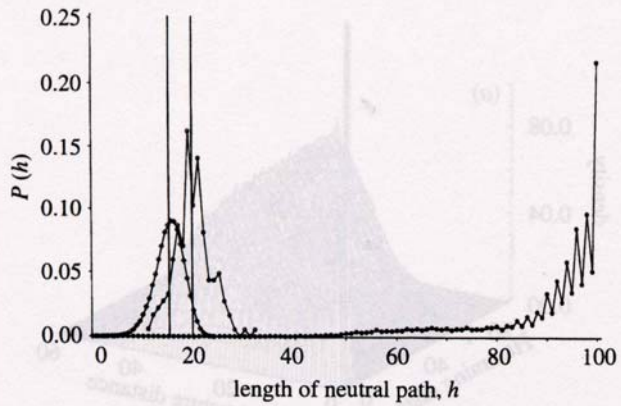


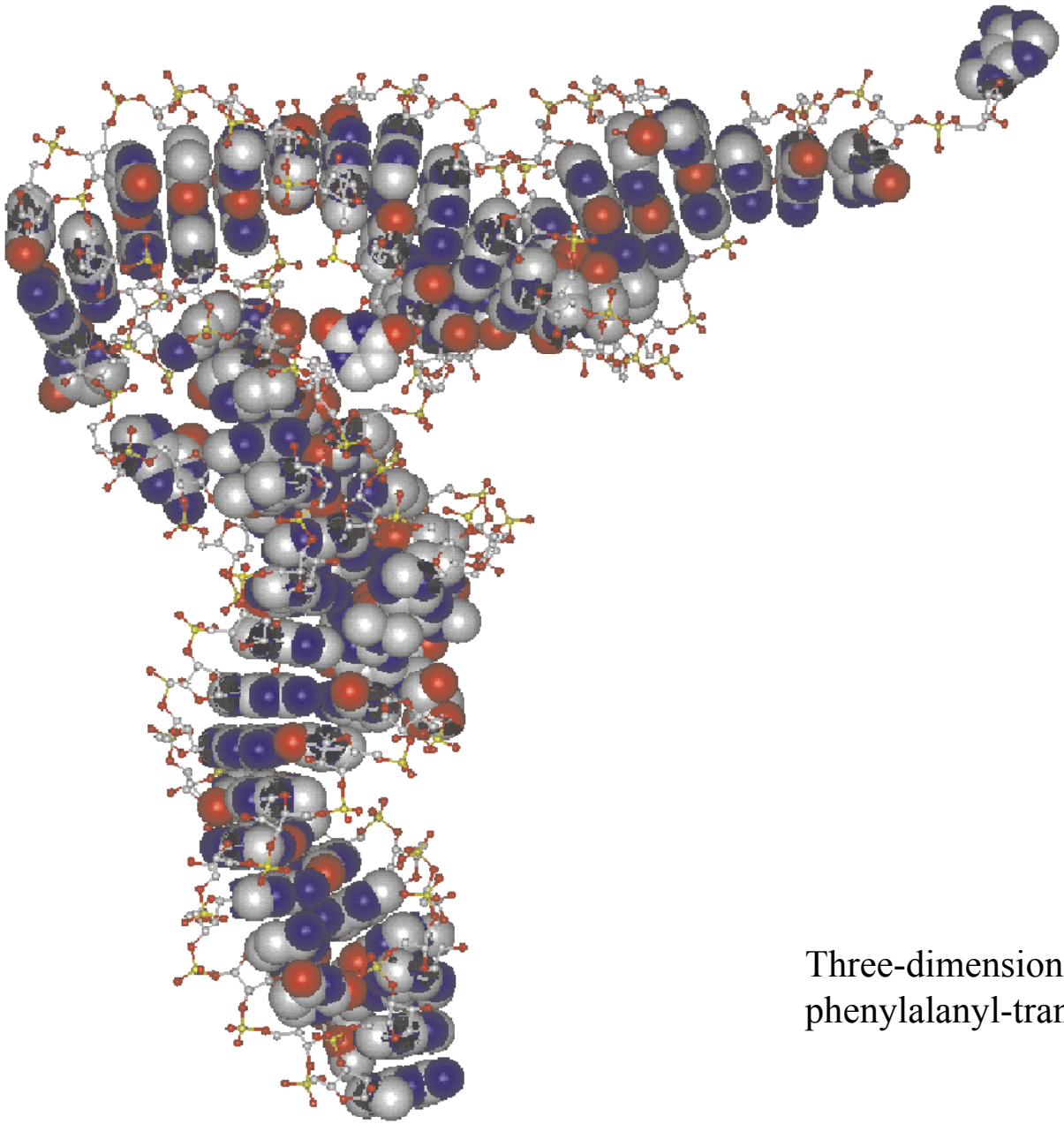
Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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

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

Sir Peter Medawar, 1985



Three-dimensional structure of
phenylalanyl-transfer-RNA

RNA Secondary Structures and their Properties

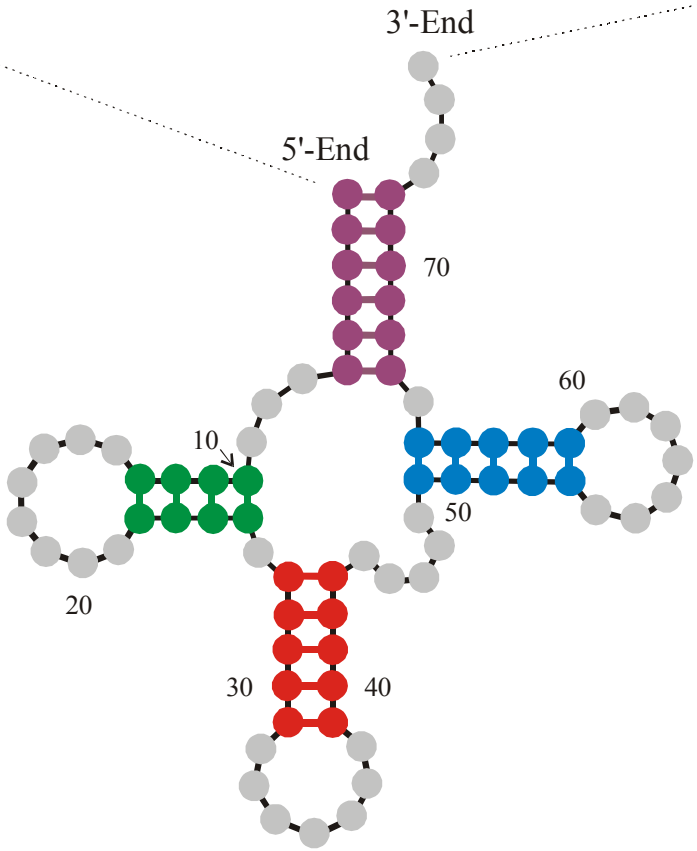
RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots. Secondary structures are folding intermediates in the formation of full three-dimensional structures.

D.Thirumalai, N.Lee, S.A.Woodson, and D.K.Klimov.
Annu.Rev.Phys.Chem. **52**:751-762 (2001)

Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

Secondary Structure



Symbolic Notation

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

Definition and formation of the secondary structure of phenylalanyl-tRNA

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)

Criterion of
Minimum Free Energy

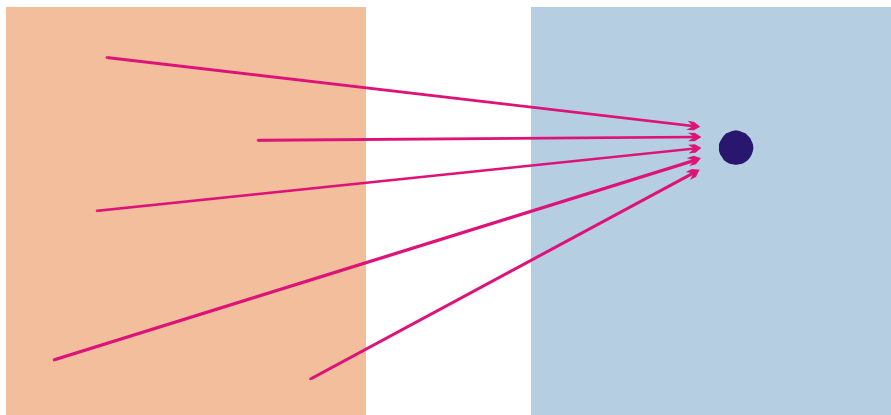
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

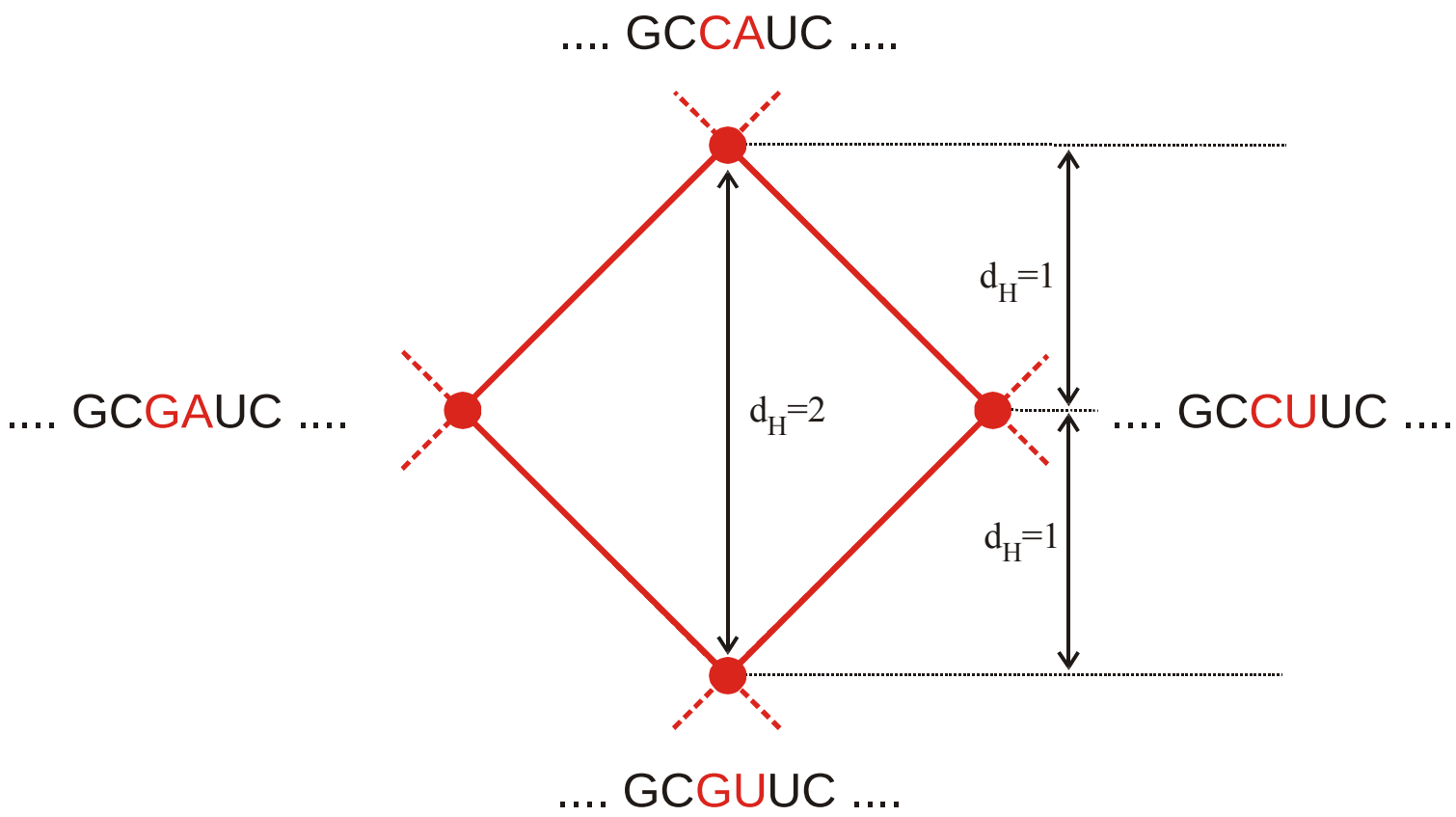
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

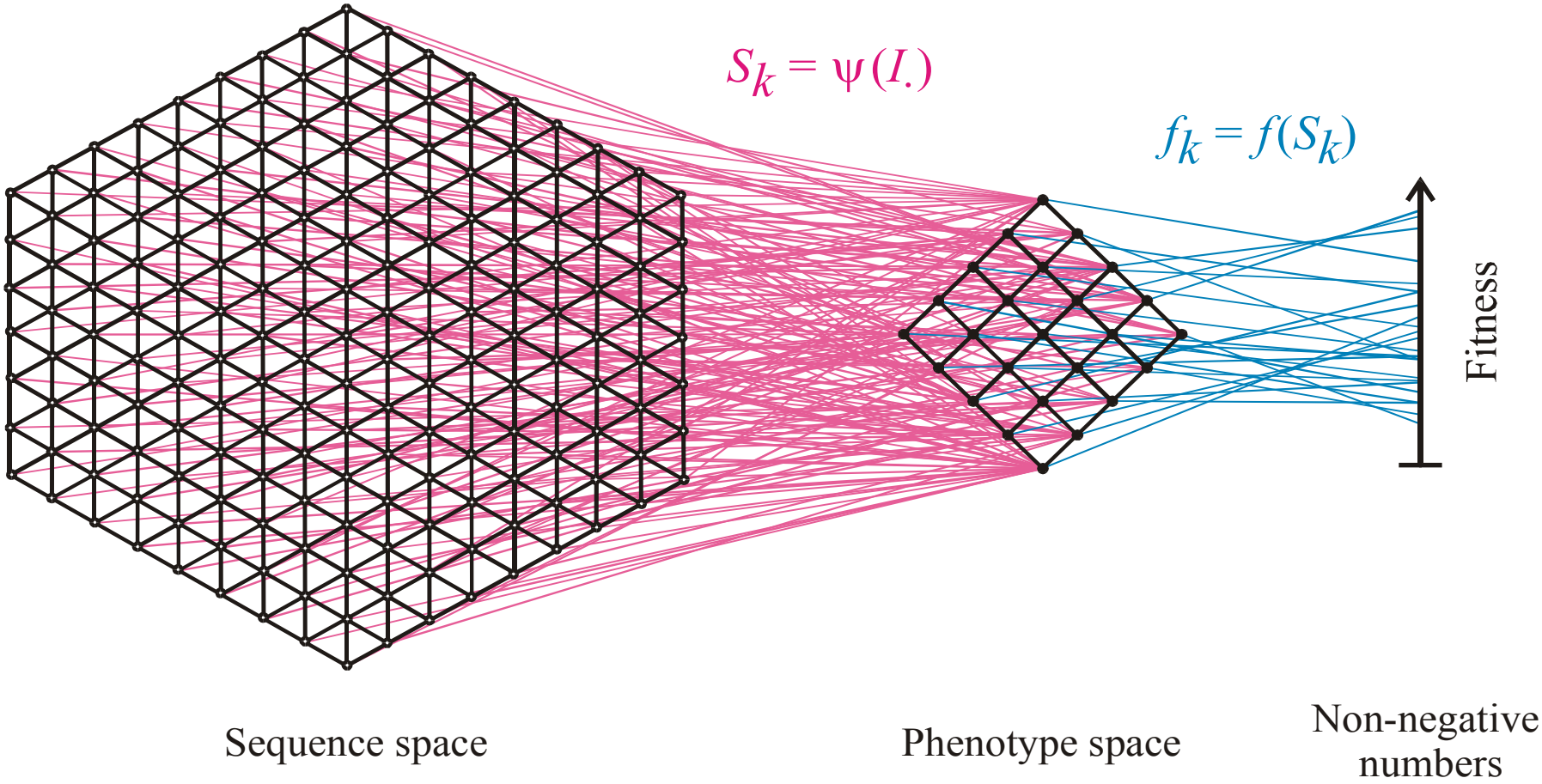
CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG

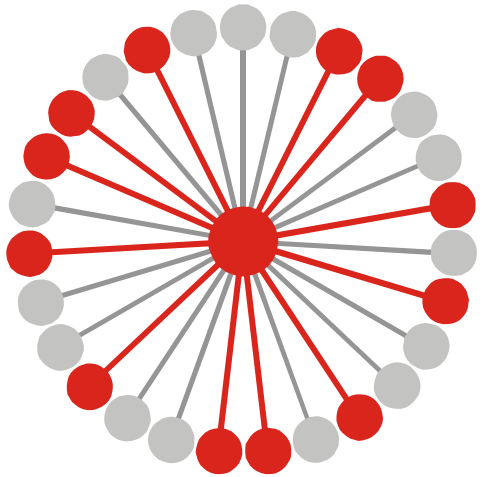




Point mutations as moves in sequence space



Mapping from sequence space into phenotype space and into fitness values



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

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

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

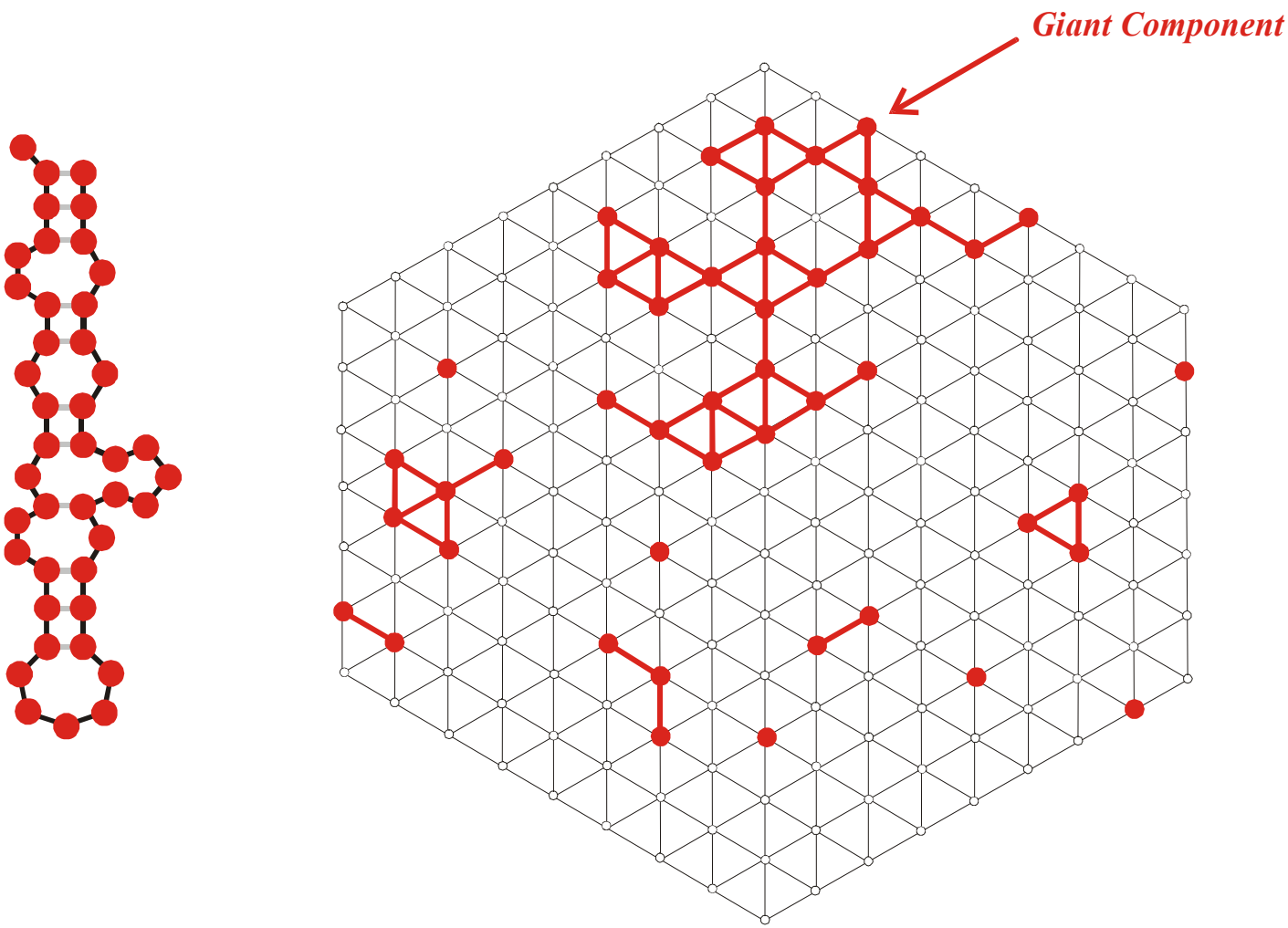
Alphabet Size k : **AUGC** ñ k = 4

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

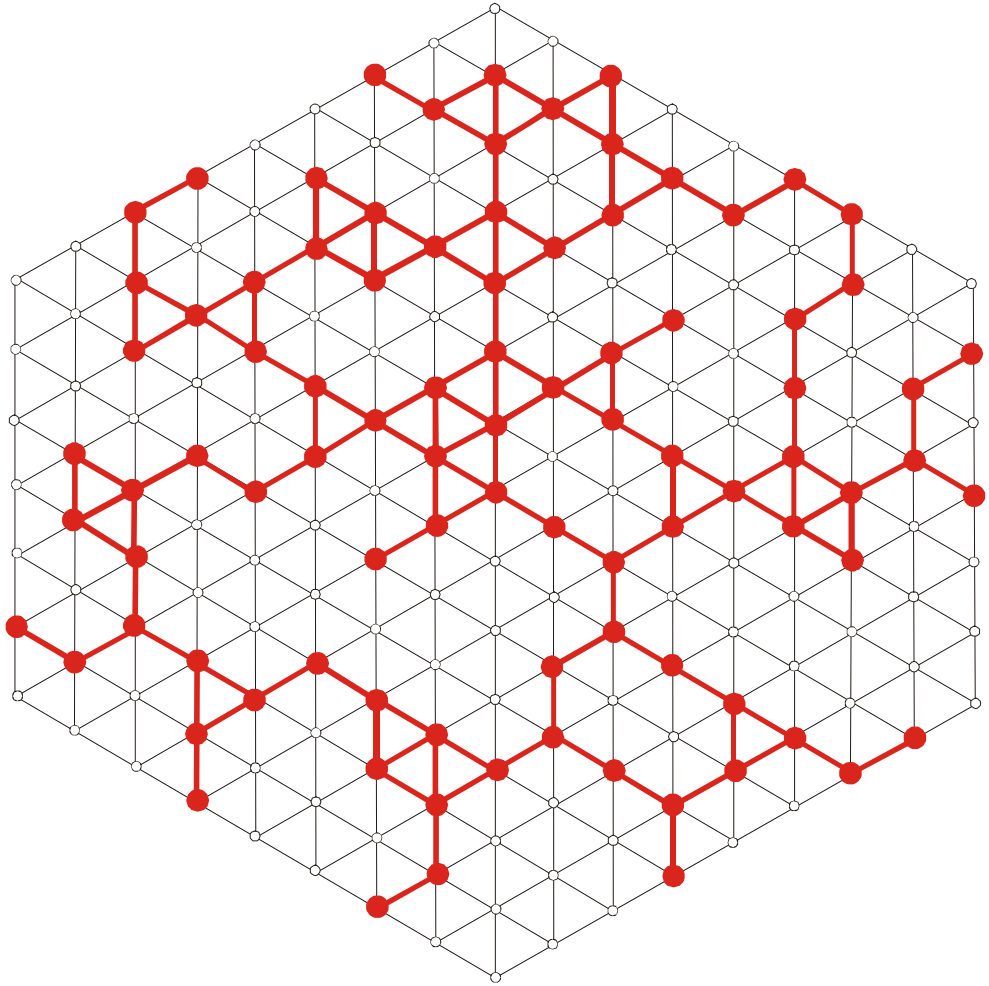
$\bar{\lambda}_k < \lambda_{cr}$ .... Network G_k is **not** connected

k	λ_{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

Theory of molecular evolution

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

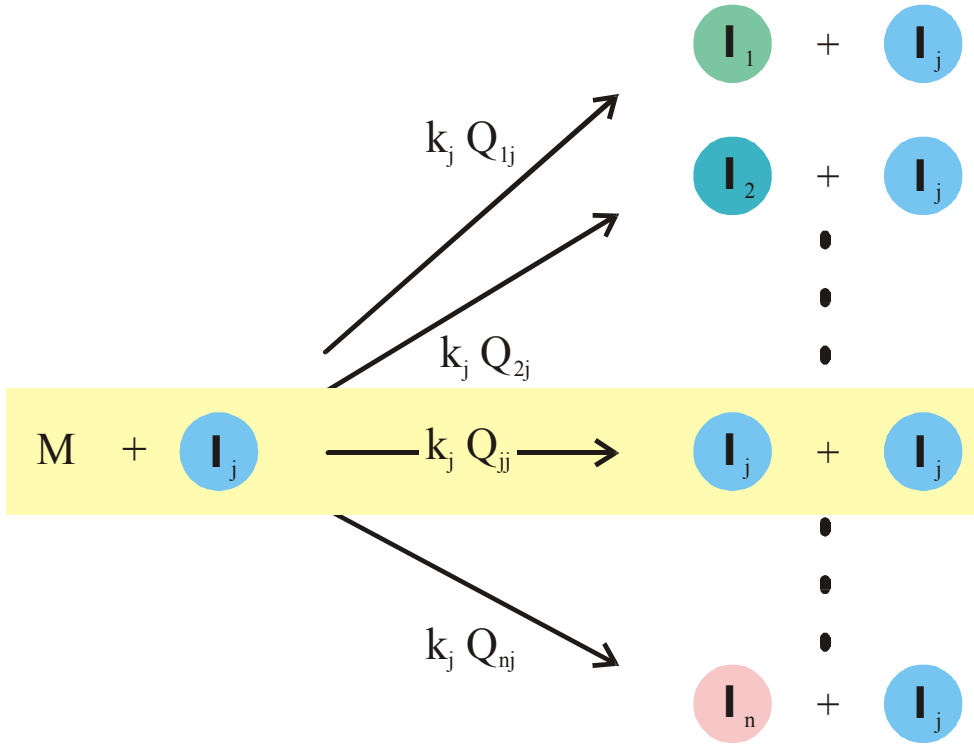
M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle***. Naturwissenschaften **58** (1977), 465-526

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

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



$$\Sigma_i Q_{ij} = 1$$

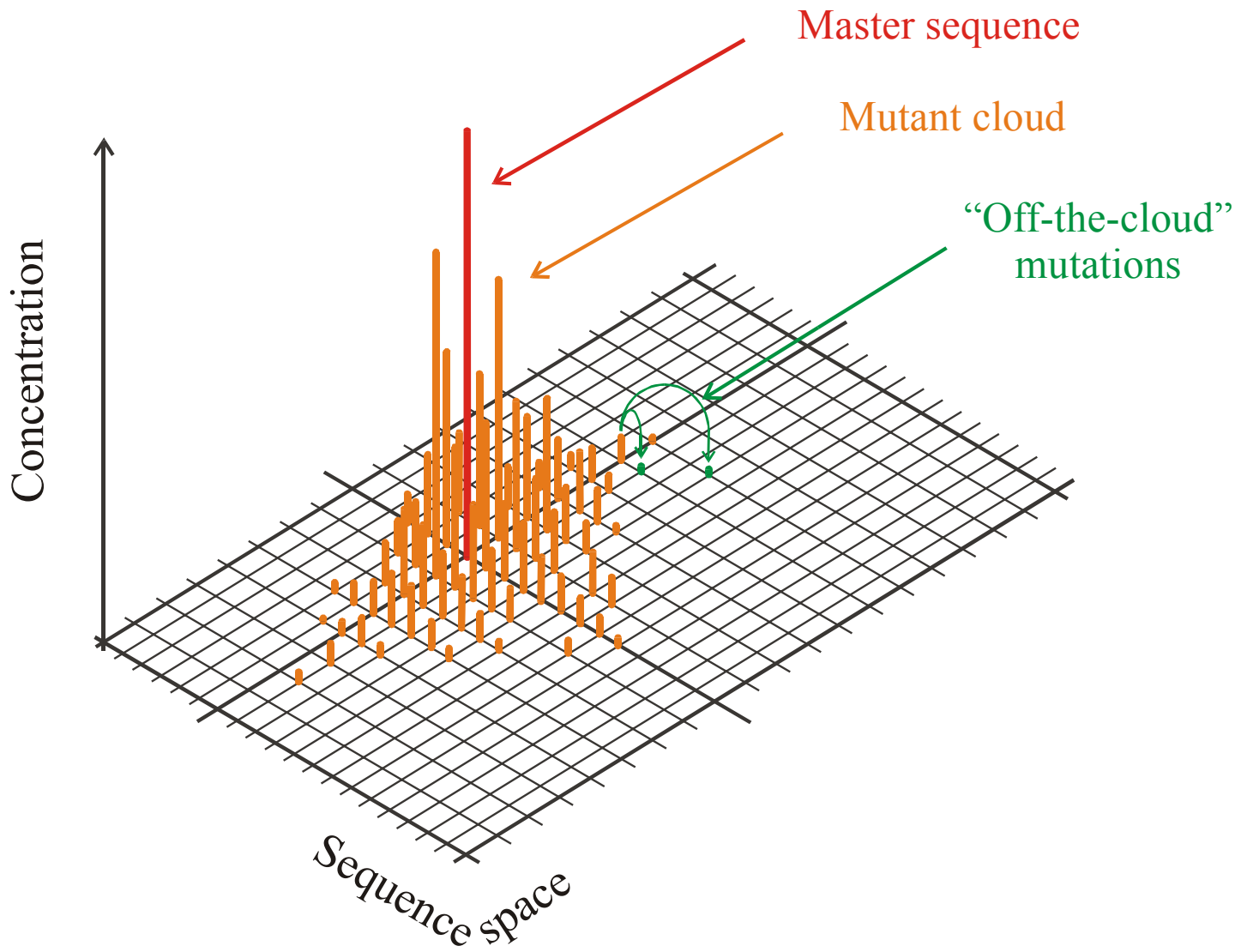
$Q_{ij} = (1-p)^{n-d(i,j)} p^{d(i,j)}$; p ..... error rate per digit

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

$$dx_j / dt = \Sigma_i k_i Q_{ji} x_i - x_j \Phi$$

$$\Phi = \Sigma_i k_i x_i ; \quad \Sigma_i x_i = 1$$

Chemical kinetics of replication and mutation as parallel reactions



The molecular quasispecies in sequence space

Optimization of RNA molecules *in silico*

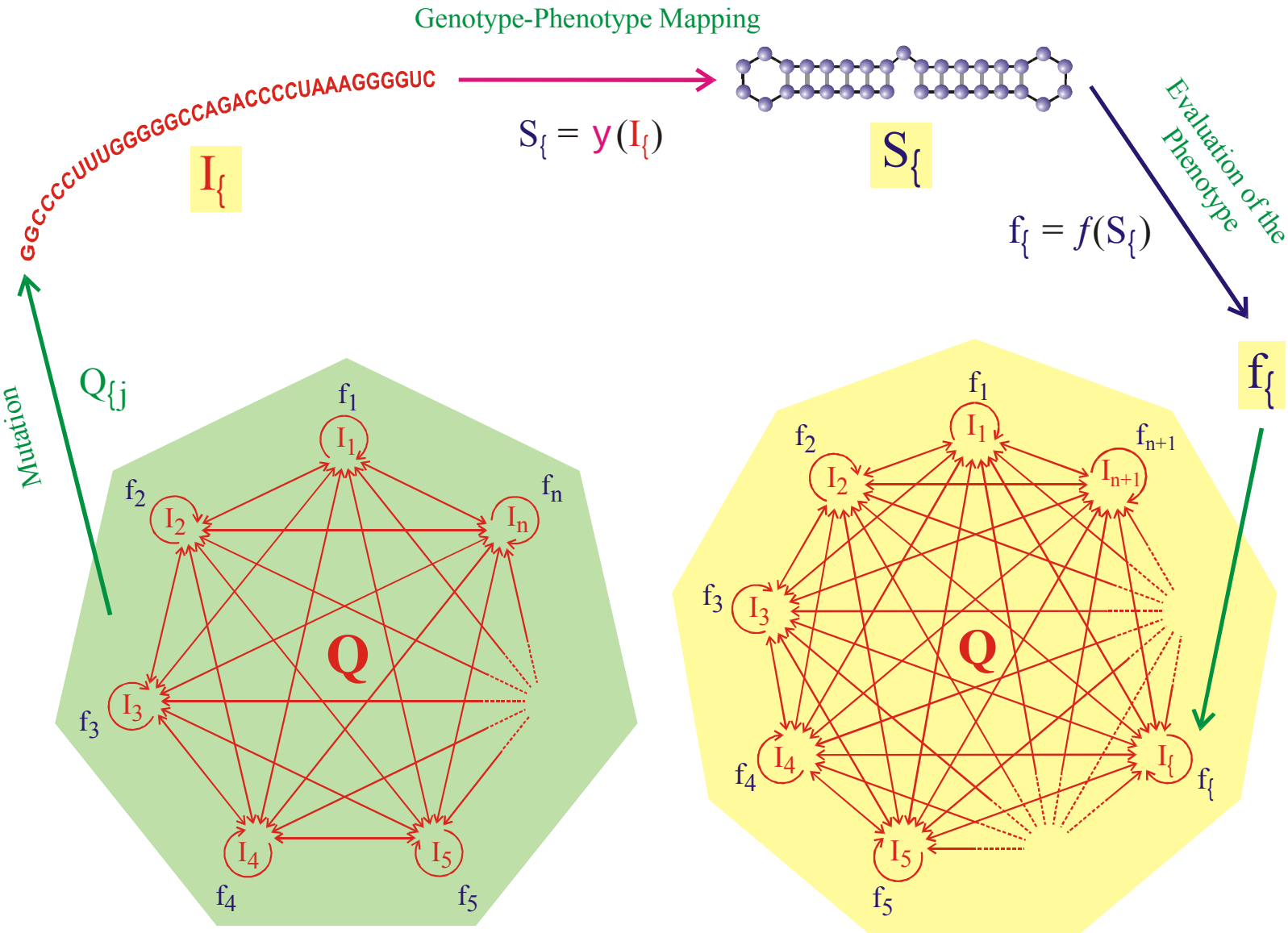
W.Fontana, P.Schuster, ***A computer model of evolutionary optimization***. Biophysical Chemistry **26** (1987), 123-147

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

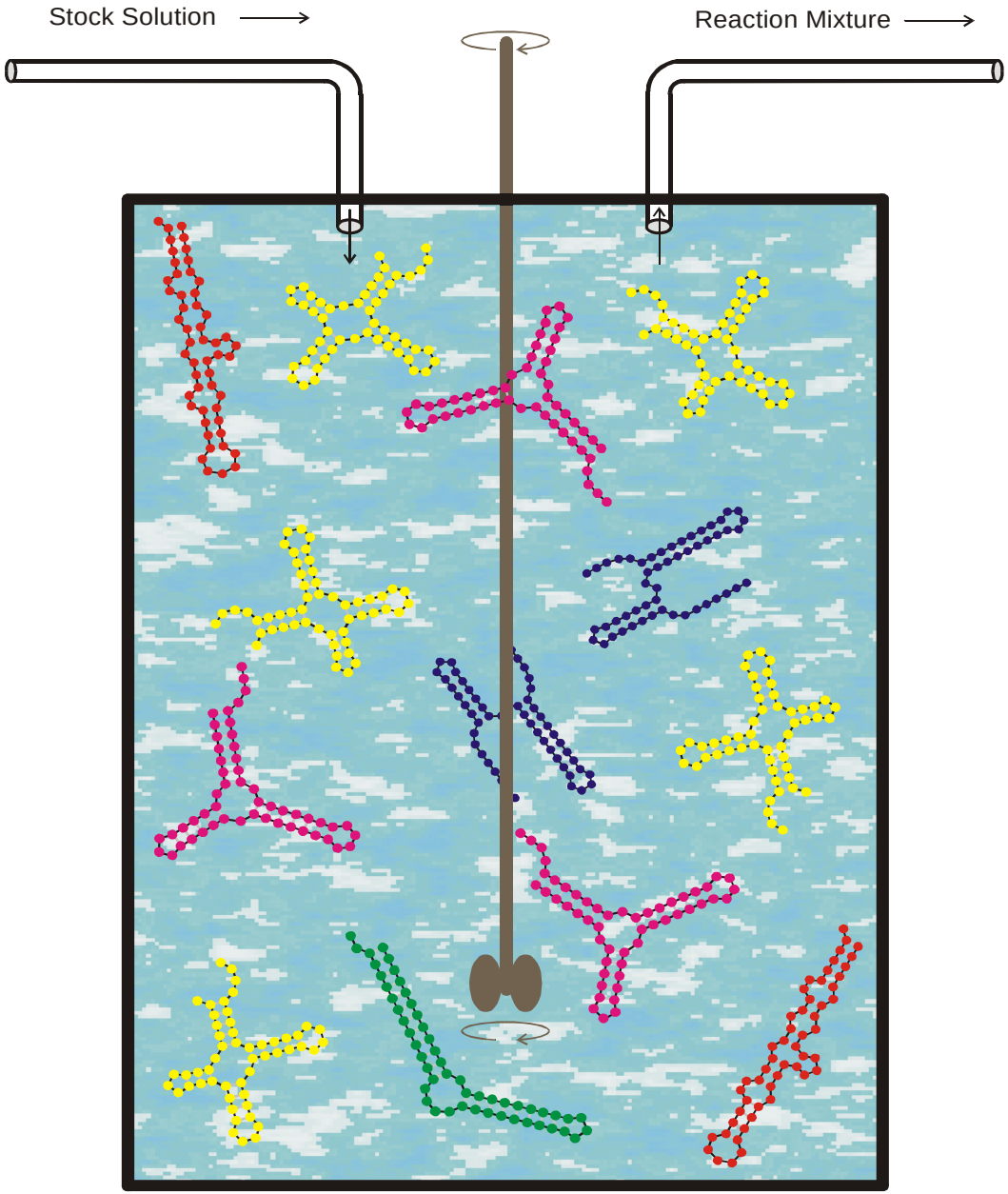
M.A.Huynen, W.Fontana, P.F.Stadler, ***Smoothness within ruggedness. The role of neutrality in adaptation***. Proc.Natl.Acad.Sci.USA **93** (1996), 397-401

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

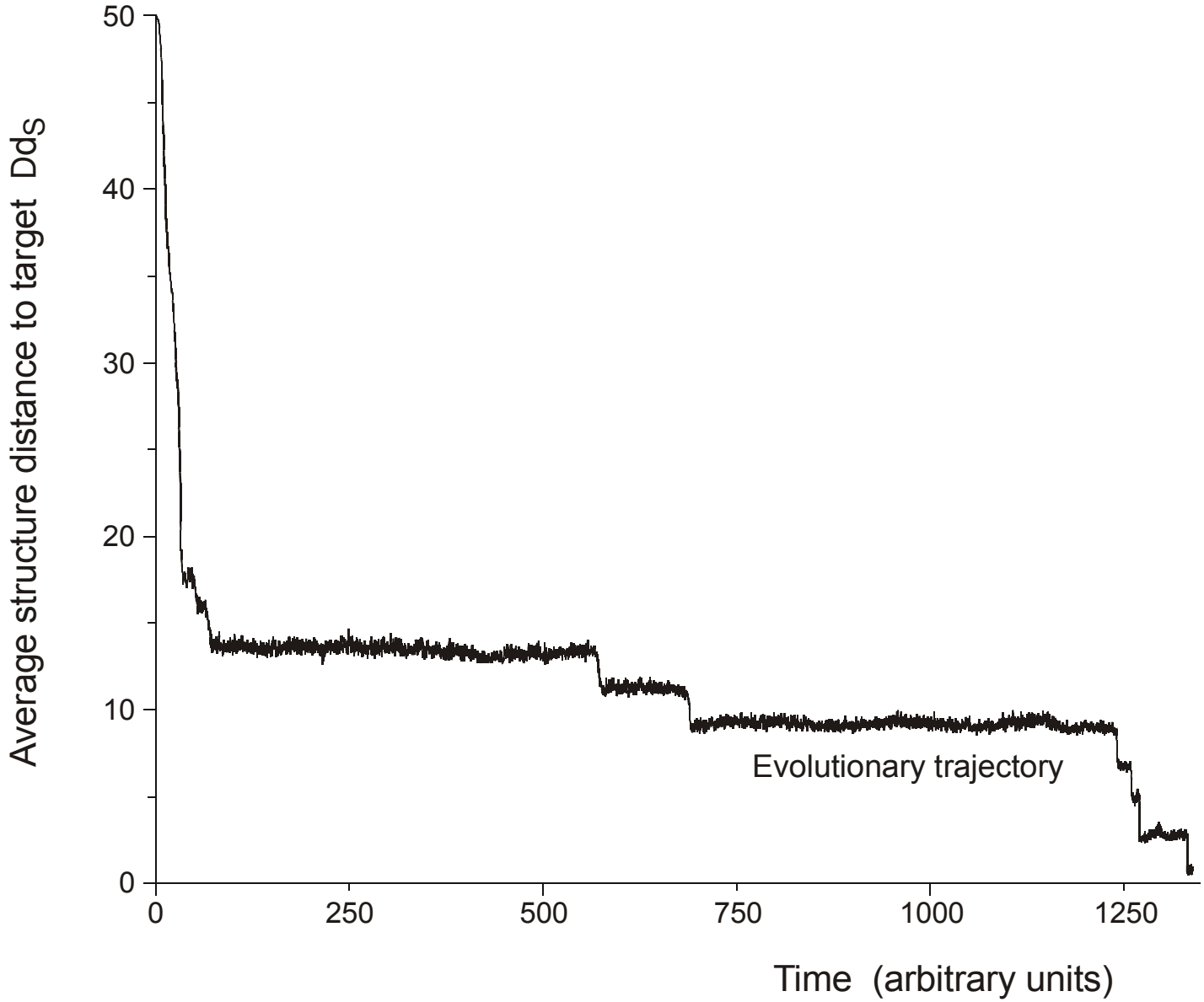
W.Fontana, P.Schuster, ***Shaping space. The possible and the attainable in RNA genotype-phenotype mapping***. J.Theor.Biol. **194** (1998), 491-515



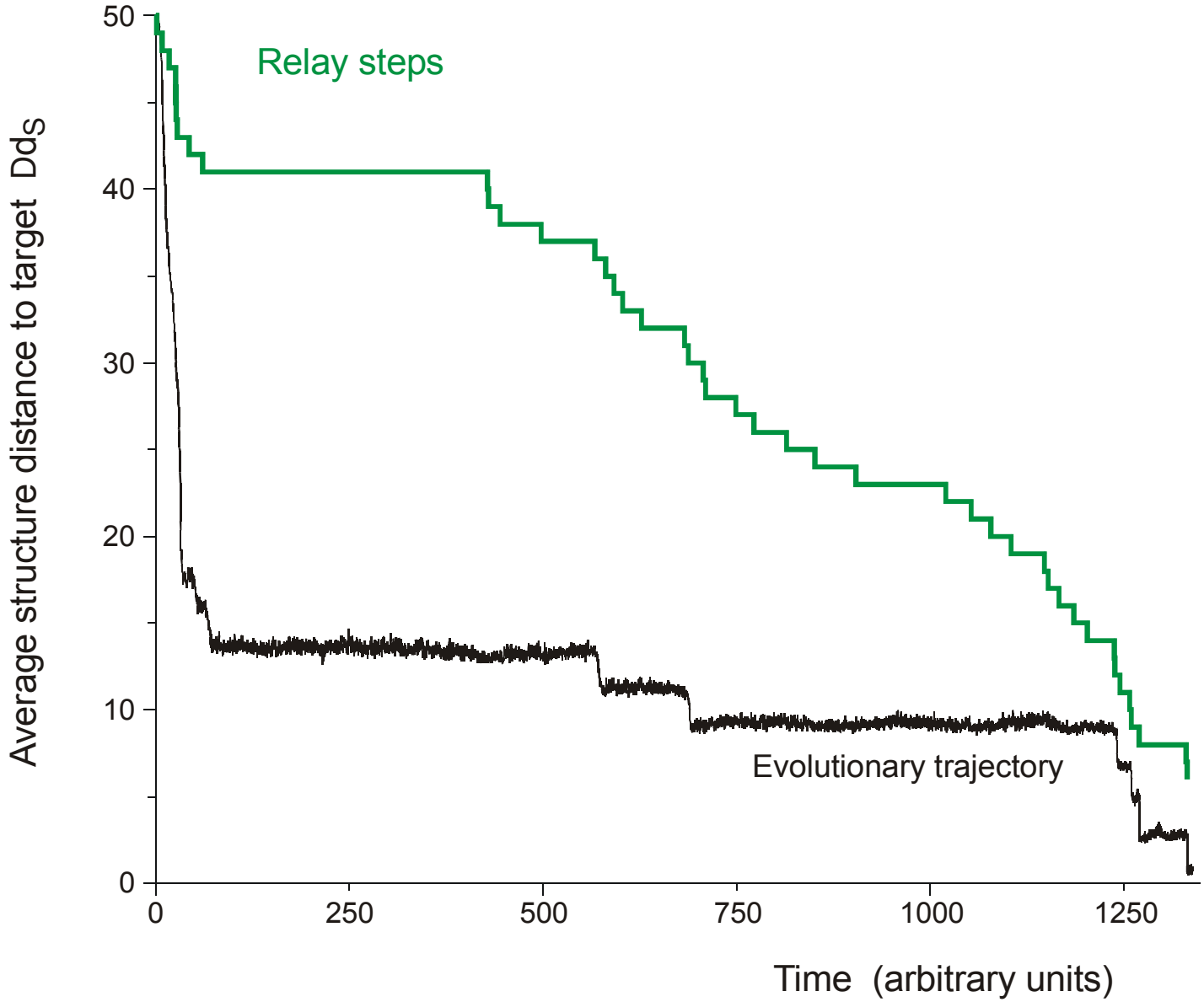
Evolutionary dynamics
including molecular phenotypes



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

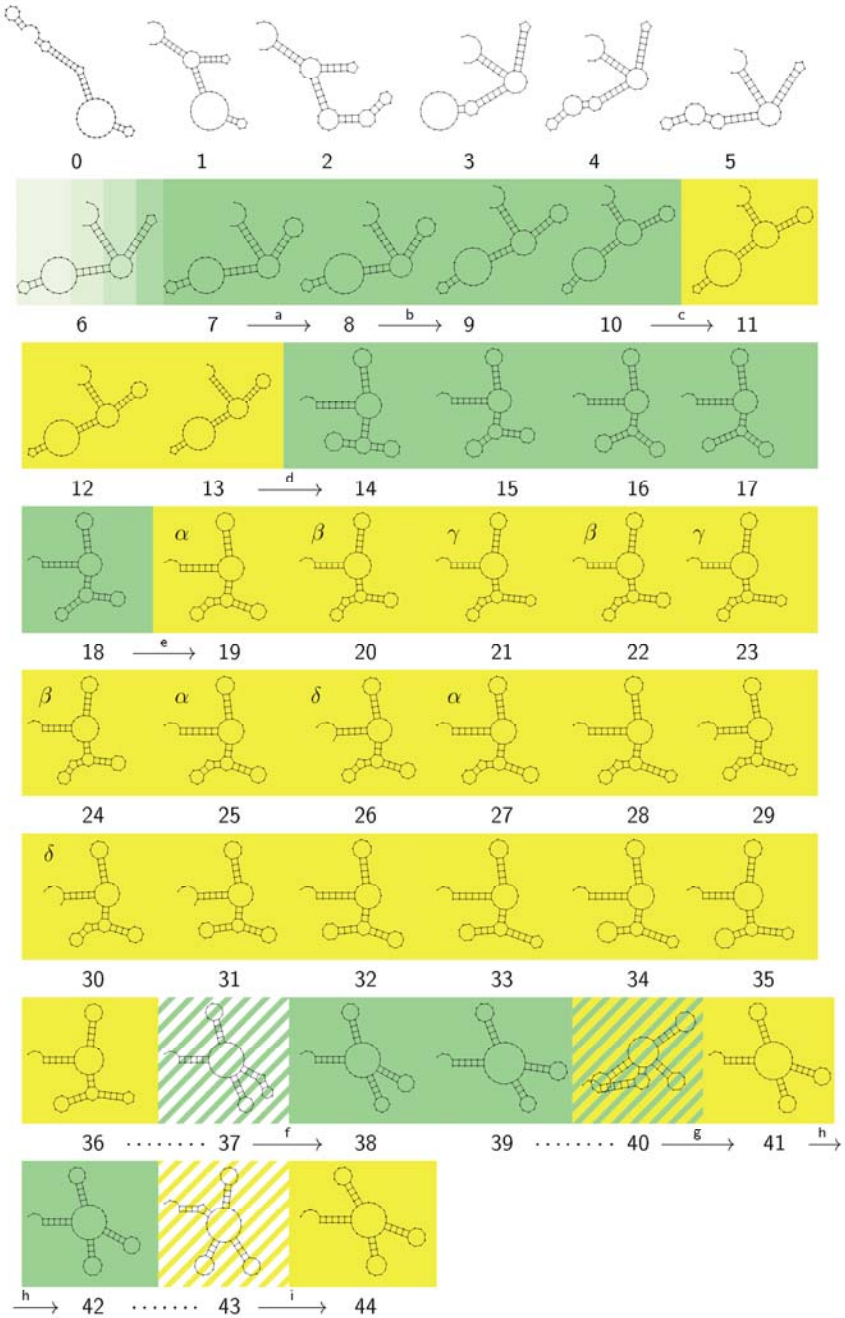


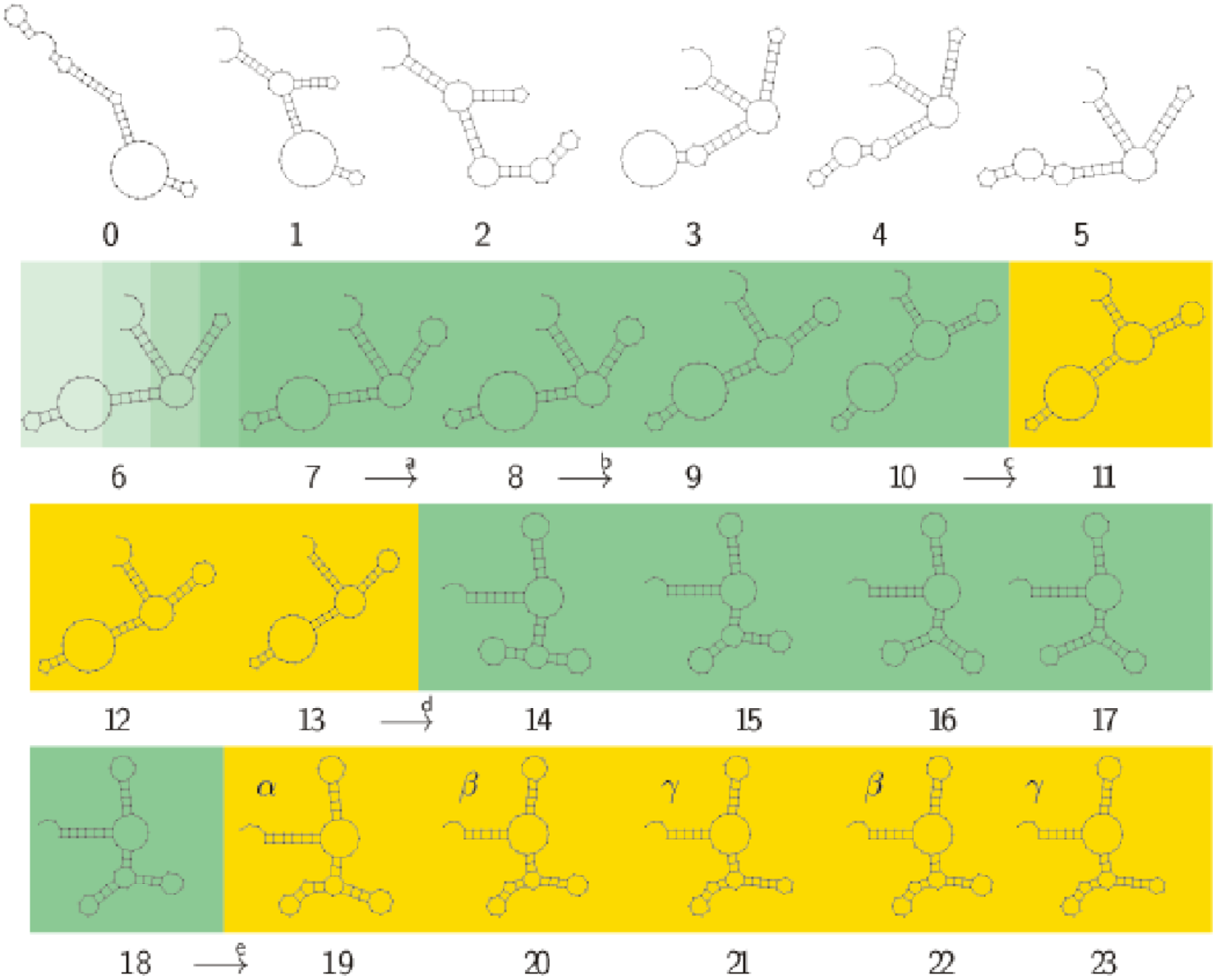
In silico optimization in the flow reactor: Trajectory



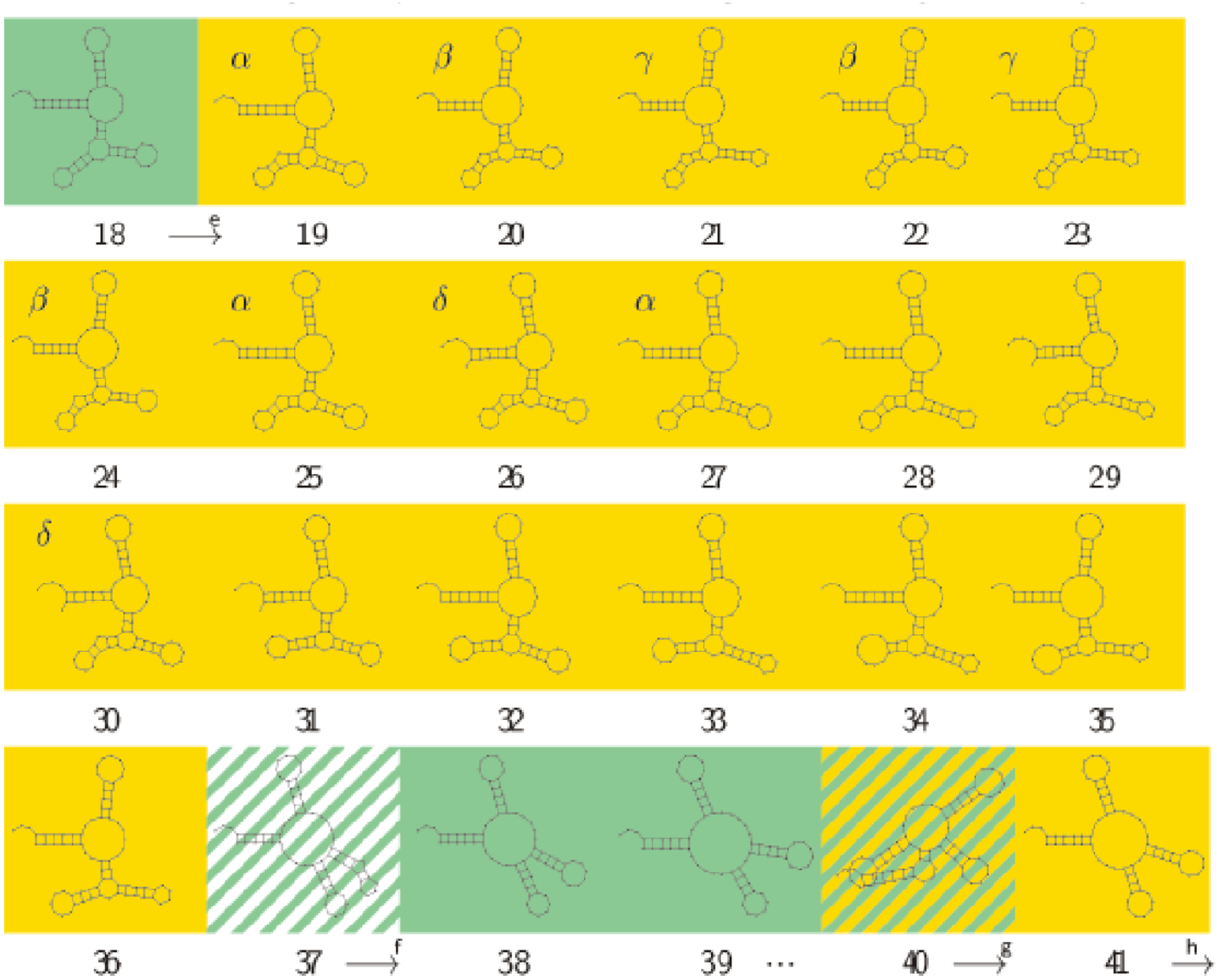
In silico optimization in the flow reactor: Trajectory and relay steps

Relay series of the trajectory leading from a randomly chosen initial structure to the clover-leaf of phenylalanyl-tRNA

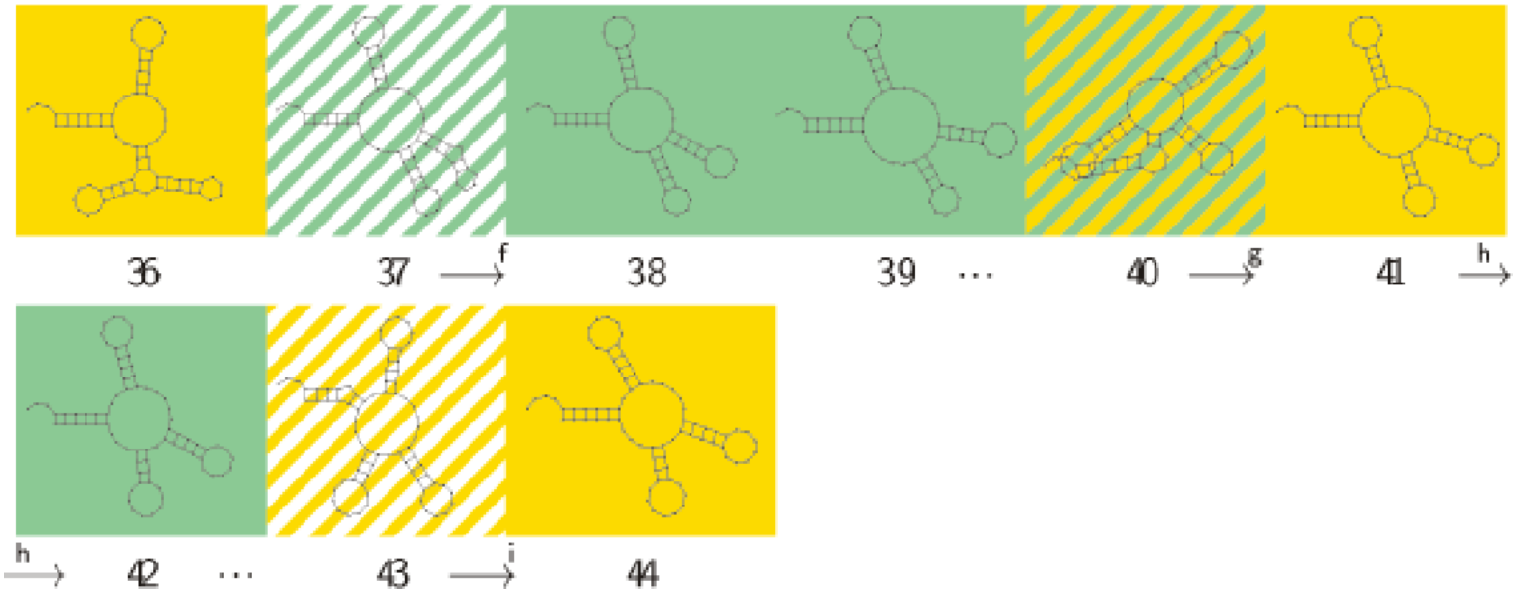




Relay series: initial sequence of events and long stasis at shape 9



Relay series: sequence of continuous transitions on a fitness plateau



Relay series: final section leading to the phenylalanyl-tRNA clover-leaf

Sequences Associated with Transitions

positions of structure changing mutations

positions of neutral mutations

Sequences involved in the transitions of the trajectory leading from a randomly chosen initial structure to the clover-leaf of phenylalanyl-tRNA

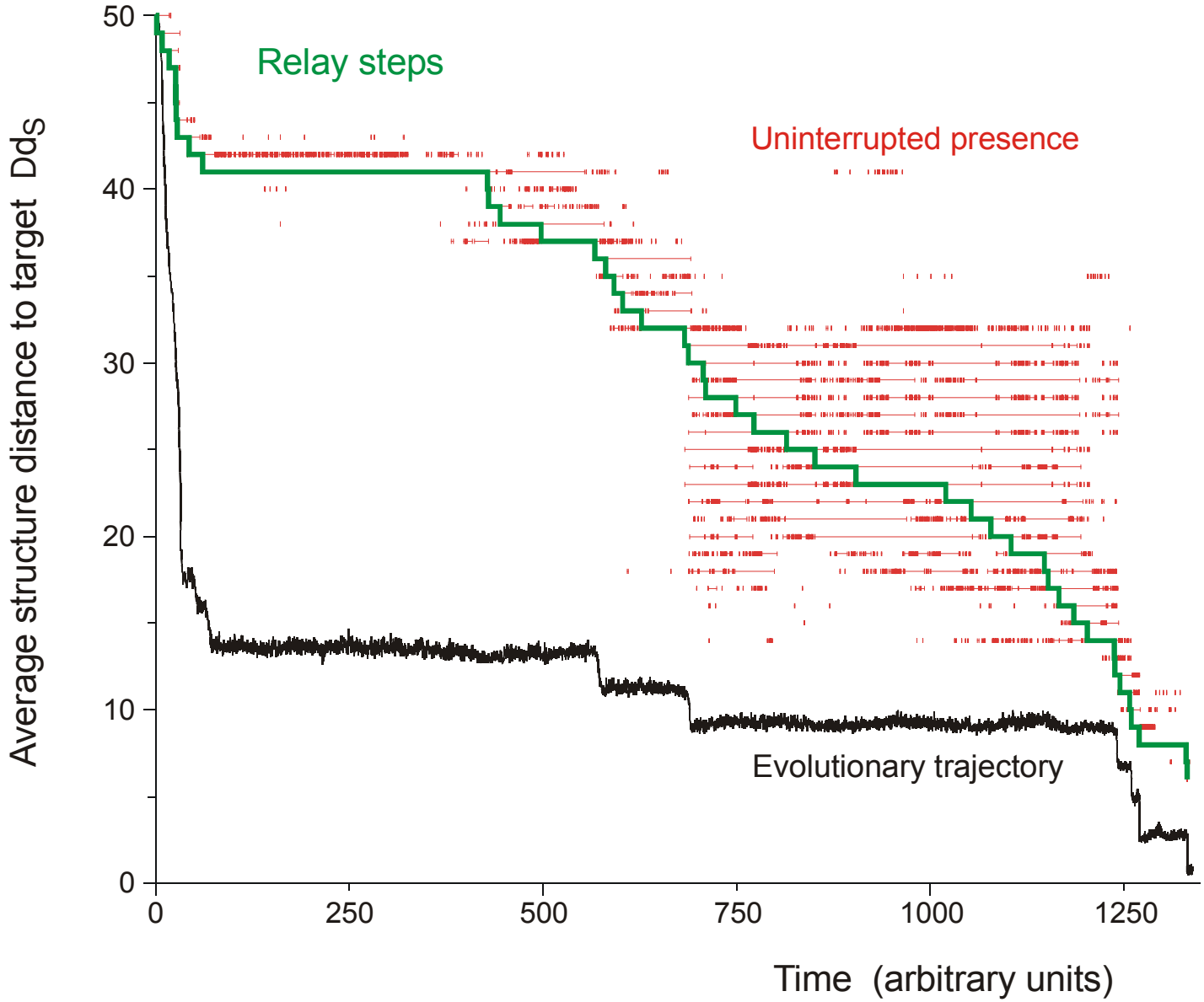
Part I

Sequences involved in the transitions of the trajectory leading from a randomly chosen initial structure to the clover-leaf of phenylalanyl-tRNA

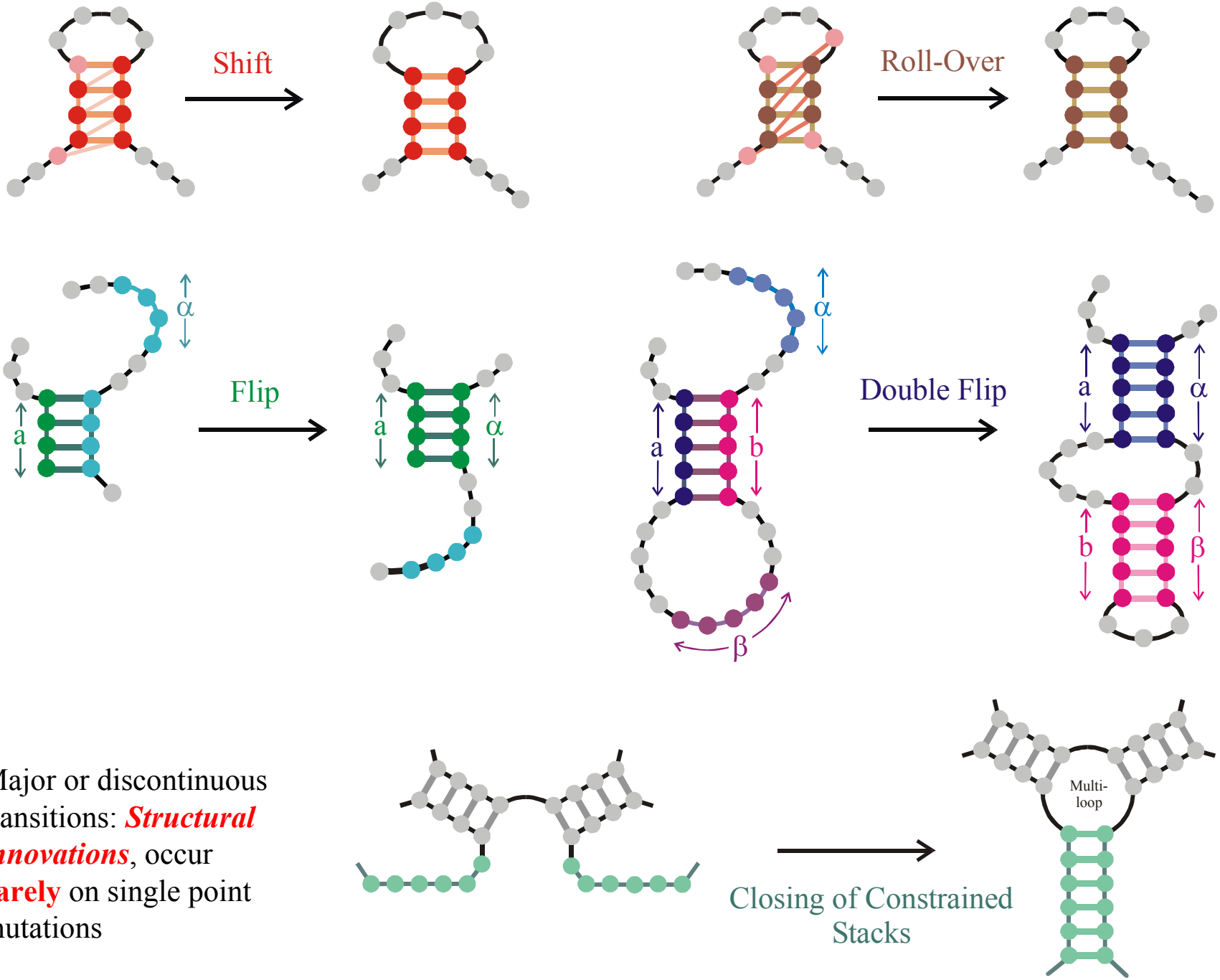
Part II

Sequences involved in the transitions of the trajectory leading from a randomly chosen initial structure to the clover-leaf of phenylalanyl-tRNA

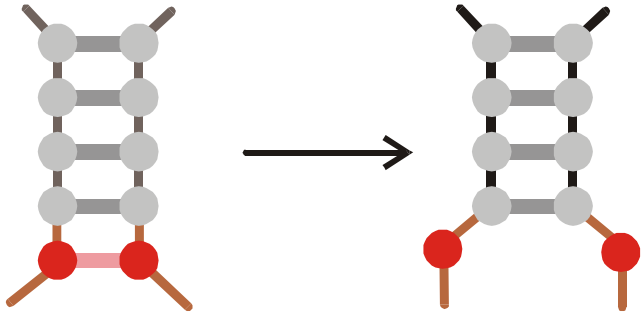
Part III



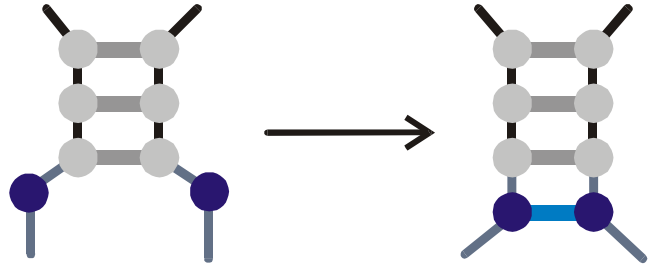
In silico optimization in the flow reactor: Uninterrupted presence



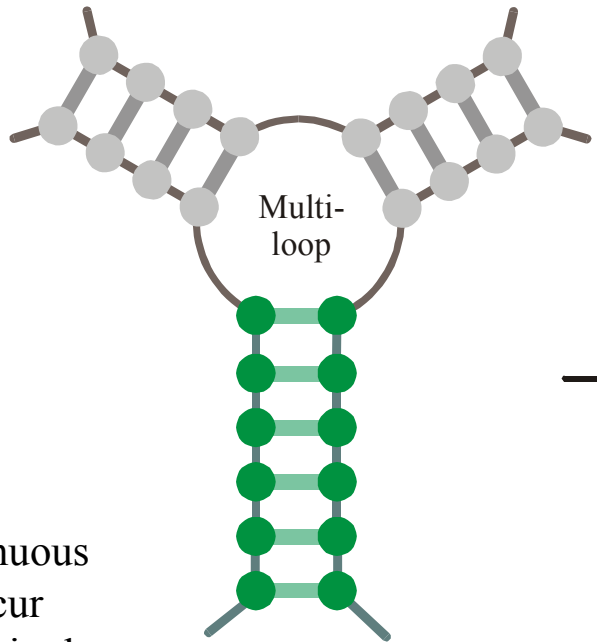
Major or discontinuous transitions: ***Structural innovations***, occur **rarely** on single point mutations

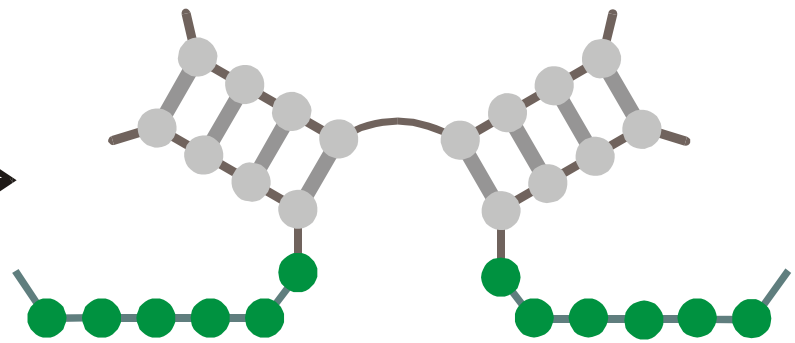


Shortening of Stacks



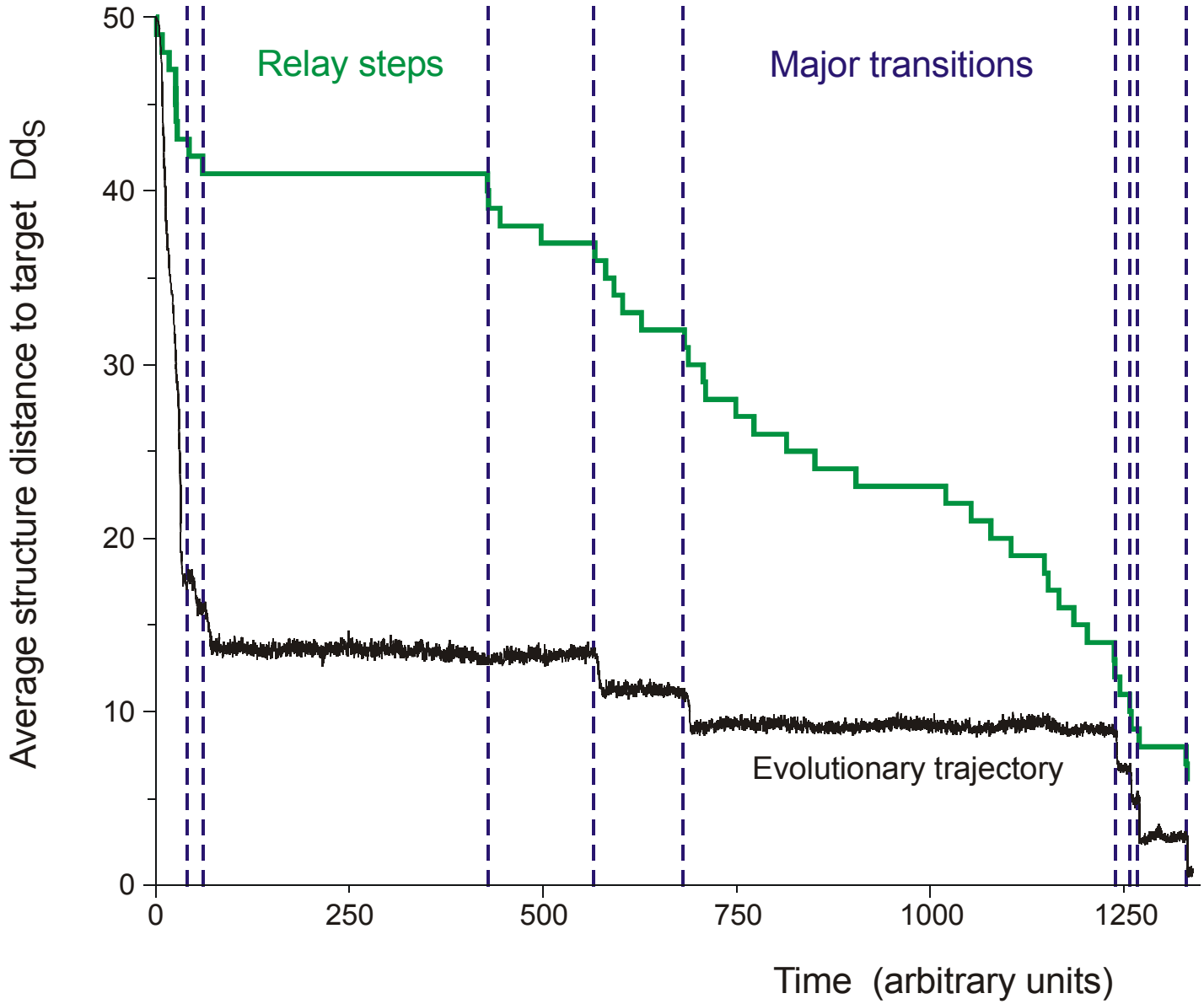
Elongation of Stacks



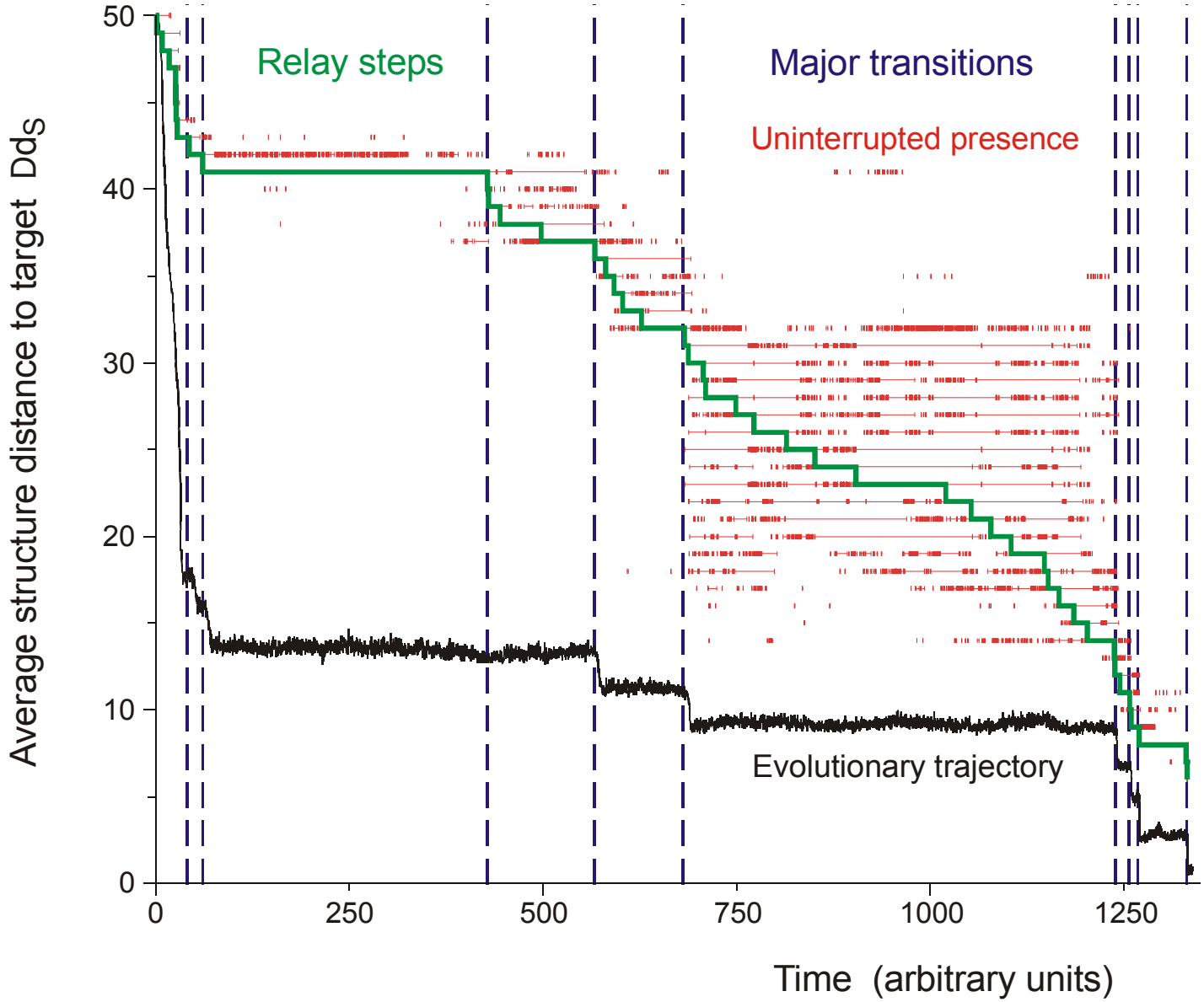


Opening of Constrained Stacks

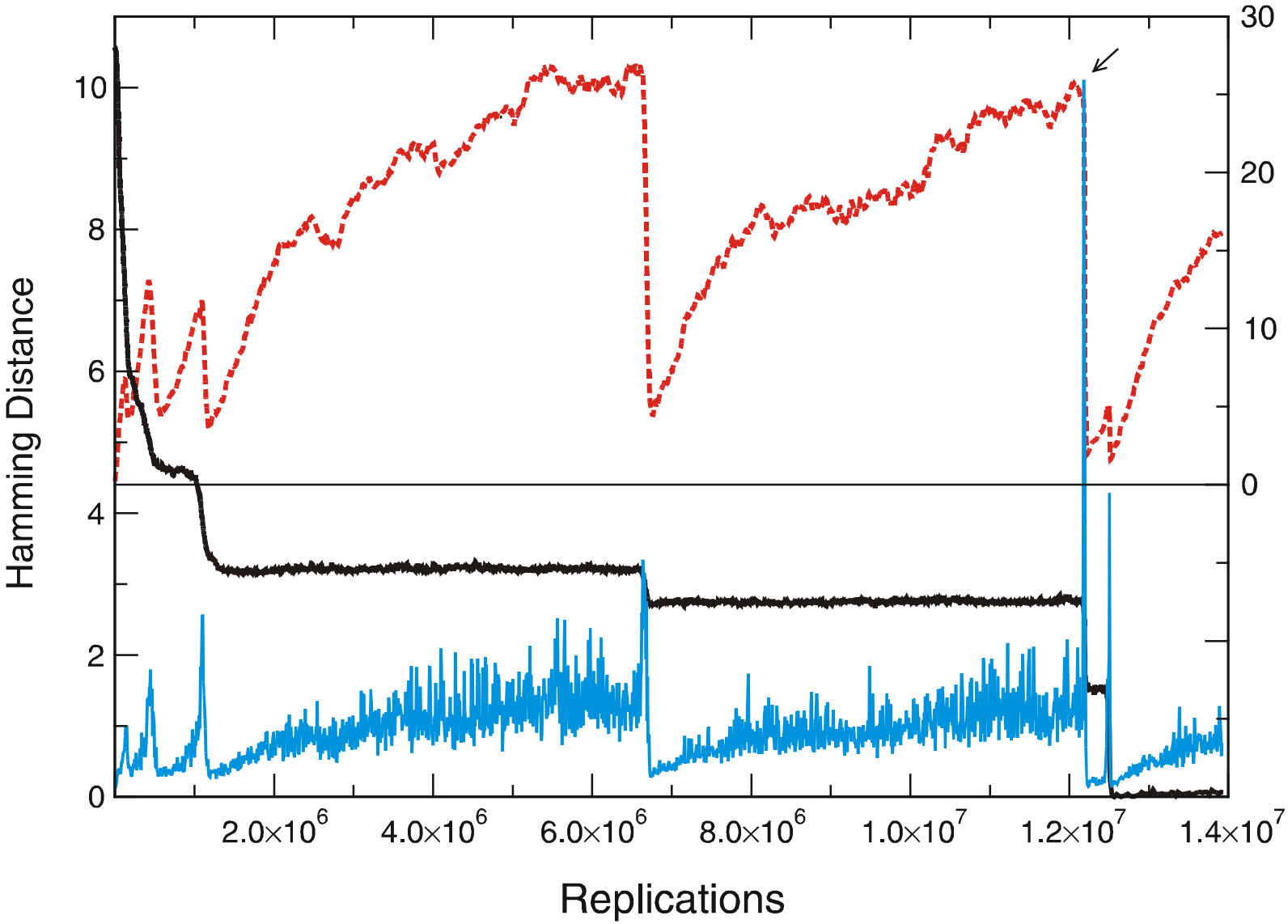
Minor or continuous transitions: Occur **frequently** on single point mutations



In silico optimization in the flow reactor: Major transitions



In silico optimization in the flow reactor



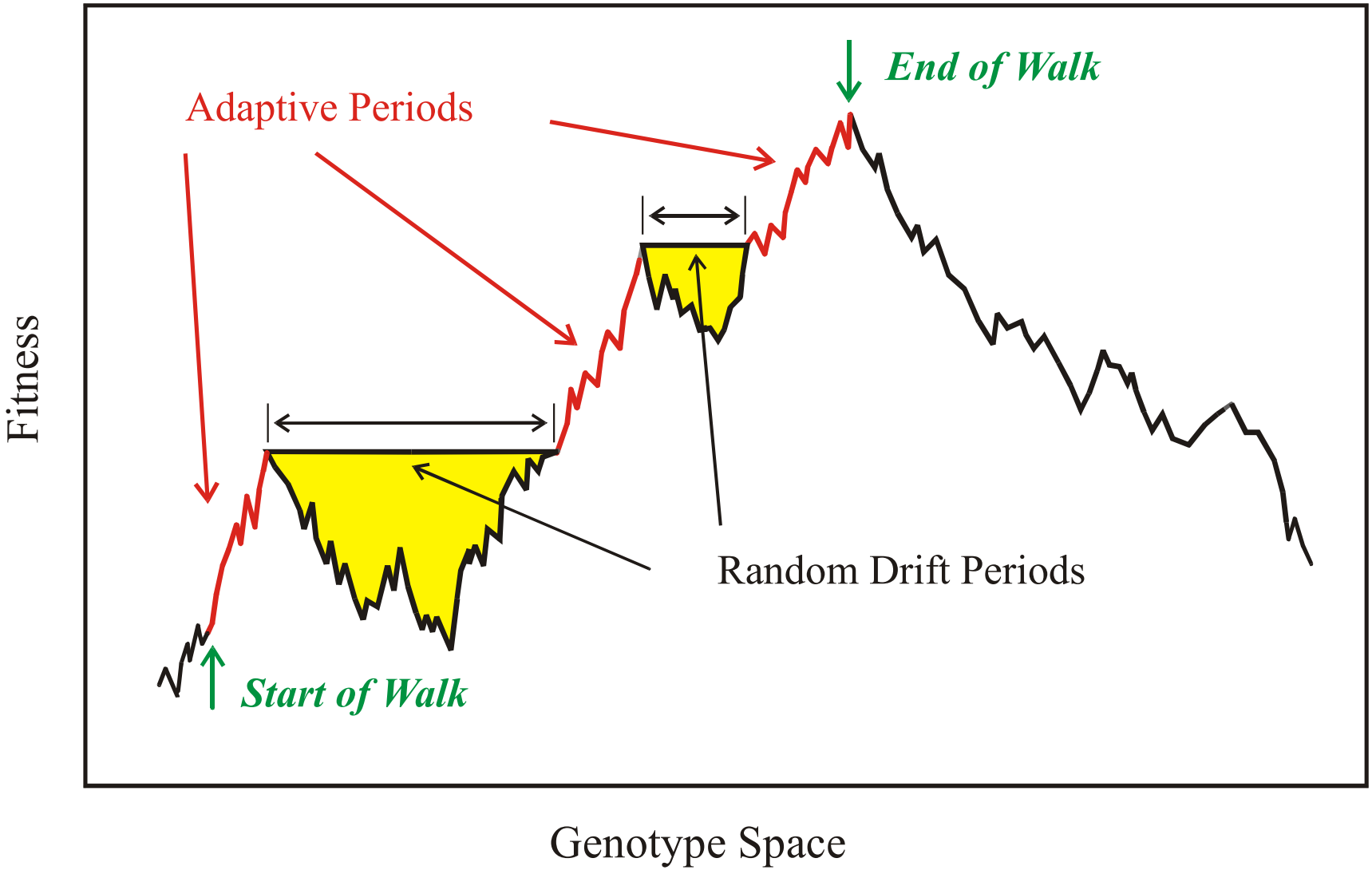
Variation in genotype space during optimization of phenotypes

Statistics of evolutionary trajectories

Population Size N	Number of Replications $< n_{\text{rep}} >$	Number of Transitions $< n_{\text{tr}} >$	Number of Major Transitions $< n_{\text{dtr}} >$	Epochal Phase $< d_{\tau}^{s}(t_{\text{ep}}) >$
1 000	$(5.5 \pm [6.9, 3.1]) \times 10^7$	$92.7 \pm [80.3, 43.0]$	$8.8 \pm [2.4, 1.9]$	$23.7 \pm [5.0, 4.1]$
2 000	$(6.0 \pm [11.1, 3.9]) \times 10^7$	$55.7 \pm [30.7, 19.8]$	$8.9 \pm [2.8, 2.1]$	$22.2 \pm [5.1, 4.2]$
3 000	$(6.6 \pm [21.0, 5.0]) \times 10^7$	$44.2 \pm [25.9, 16.3]$	$8.1 \pm [2.3, 1.8]$	$20.9 \pm [2.4, 2.2]$
10 000	$(1.2 \pm [1.3, 0.6]) \times 10^8$	$35.9 \pm [10.3, 8.0]$	$10.3 \pm [2.6, 2.1]$	$18.4 \pm [2.3, 2.1]$
20 000	$(1.5 \pm [1.4, 0.7]) \times 10^8$	$28.8 \pm [5.8, 4.8]$	$9.0 \pm [2.8, 2.2]$	$17.5 \pm [2.5, 2.2]$
30 000	$(2.2 \pm [3.1, 1.3]) \times 10^8$	$29.8 \pm [7.3, 5.9]$	$8.7 \pm [2.4, 1.9]$	$16.7 \pm [2.0, 1.8]$
100 000	$(3 \pm [2, 1]) \times 10^8$	$24 \pm [6, 5]$	9 ± 2	17 ± 1

„...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

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

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

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

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