Neutral Networks of RNA Genotypes and RNA Evolution *in silico*

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

RNA Secondary Structures in Dijon

Dijon, 24.– 26.06.2002

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

Sir Peter Medawar, 1985

The **genotypes** or **genomes** of individuals and species, being reproductively related ensembles of individuals, are DNA or RNA sequences. They are changing from generation to generation through mutation and recombination.

Genotypes unfold into **phenotypes** or organisms, which are the targets of the evolutionary selection process.

Point mutations are single nucleotide exchanges. The **Hamming distance** of two sequences is the minimal number of single nucleotide exchanges that mutually converts the two sequence into each other.

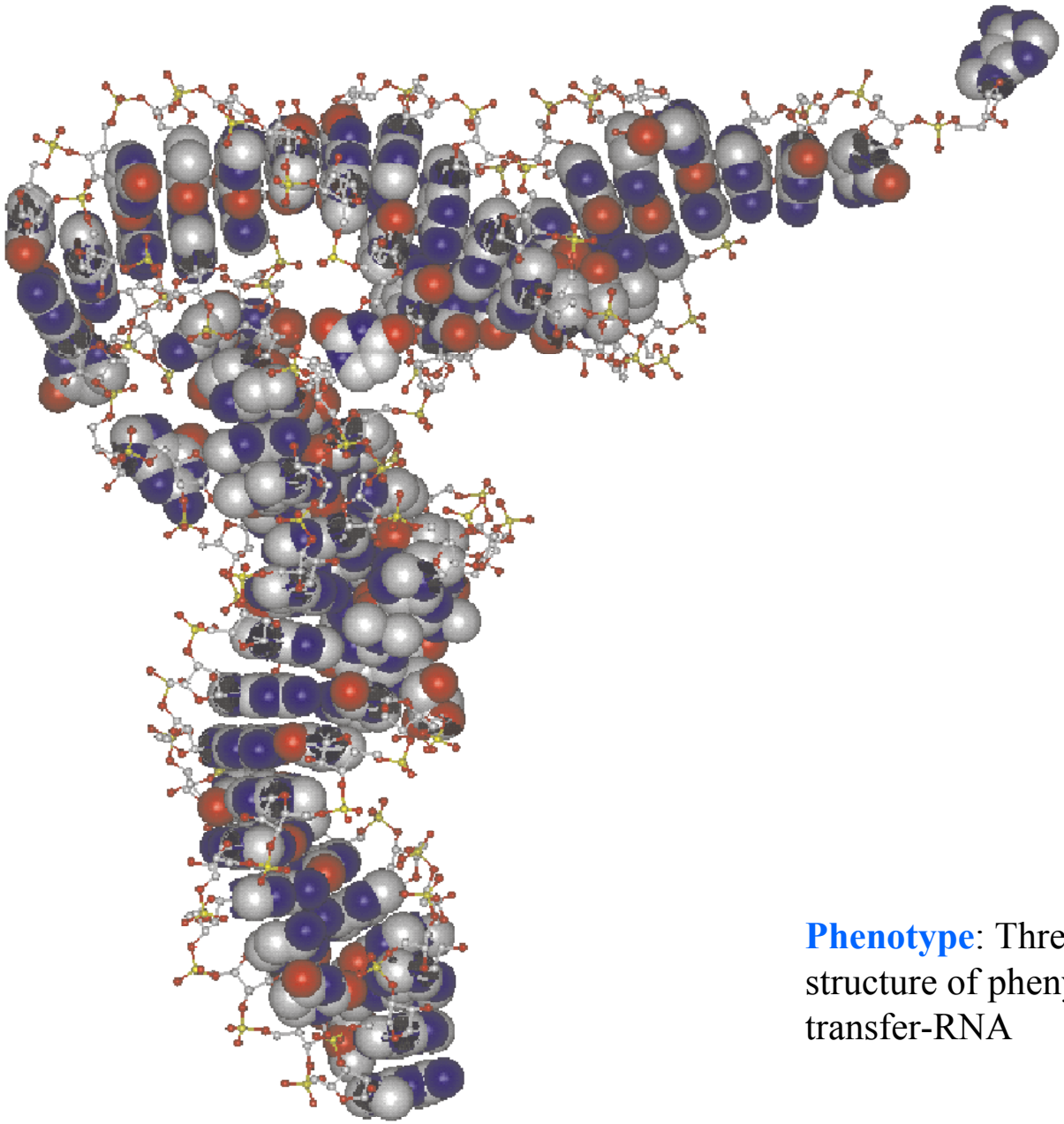
A = adenylate

C = cytidylate

U = uridylate

G = guanylate

Genotype: The sequence of an RNA molecule consisting of monomers chosen from four classes.

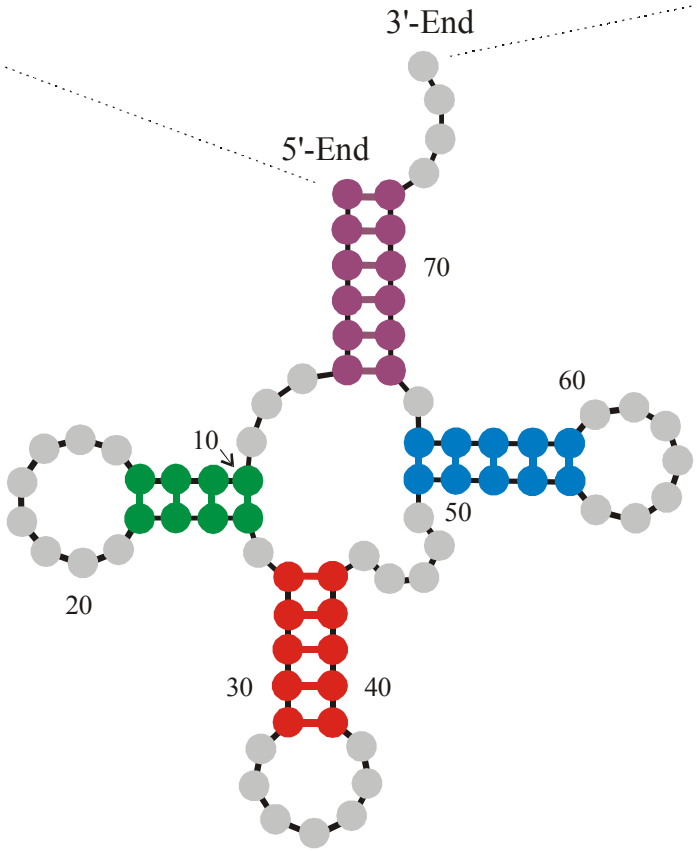


Phenotype: Three-dimensional structure of phenylalanyl transfer-RNA

Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

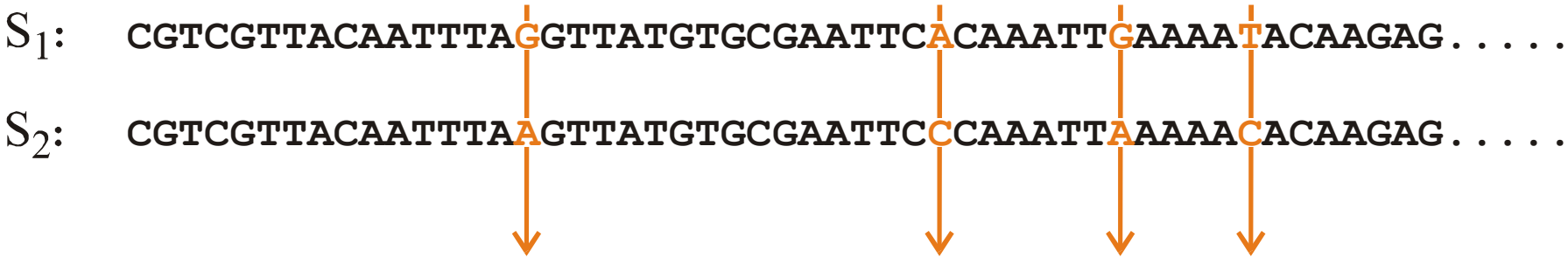
Secondary Structure



Symbolic Notation

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

Definition and formation of the secondary structure of phenylalanyl-tRNA



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

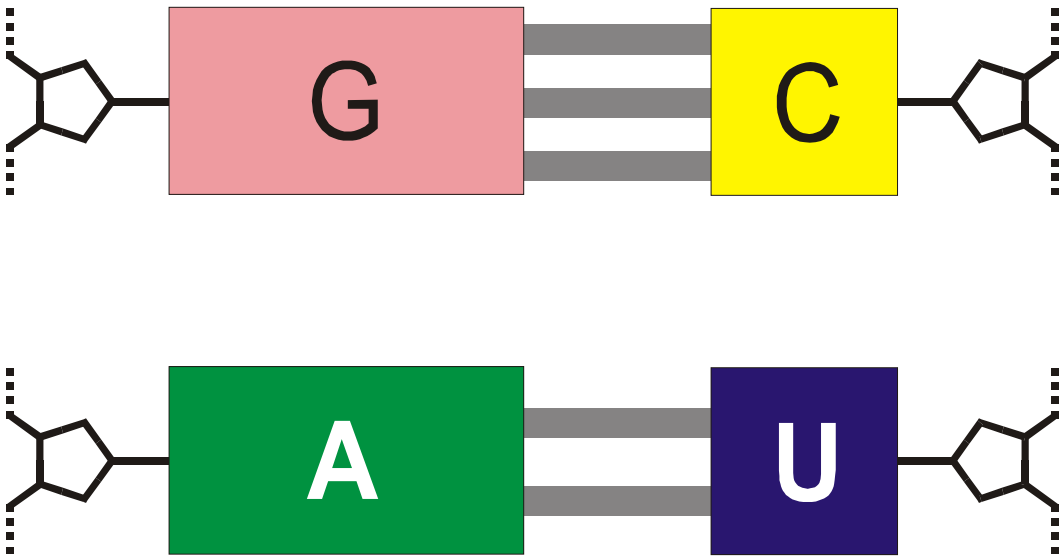
(i) $d_H(S_1,S_1) = 0$

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

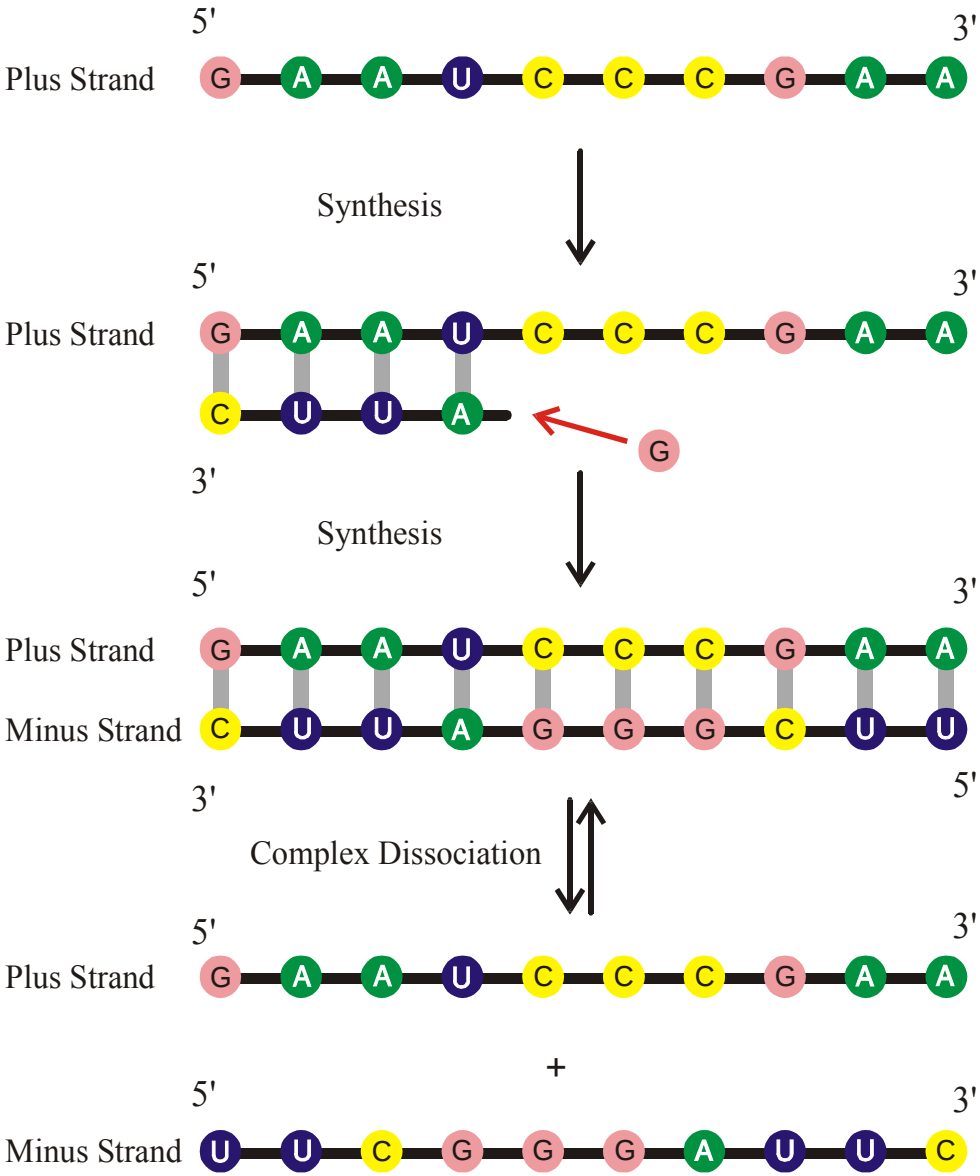
(iii) $d_H(S_1,S_3) \P d_H(S_1,S_2) + d_H(S_2,S_3)$

The Hamming distance induces a metric in sequence space

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

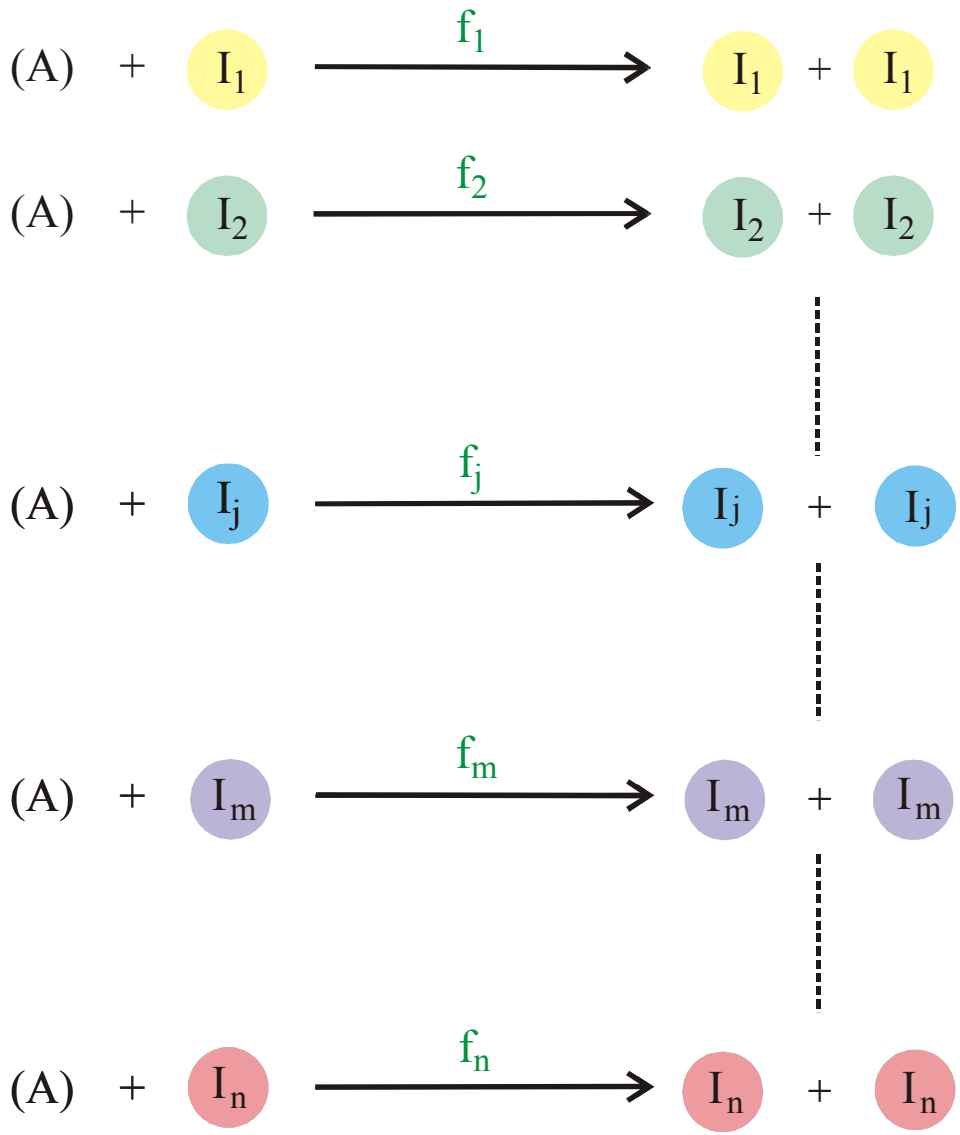
$$dx_j / dt = f_j x_j - x_j \Phi = (f_j - \Phi) x_j$$

$\Phi = \Sigma_i f_i x_i$; $\Sigma_i x_i = 1$; i,j =1,2,...,n

[A] = a = constant

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

$x_m(t)$ Á 1 for t Á ¸



$s = (f_{m+1}-f_m)/f_m$

succession of temporarily fittest variants:

m Á m+1 Á ...

Selection of the „fittest“ or fastest replicating species I_m

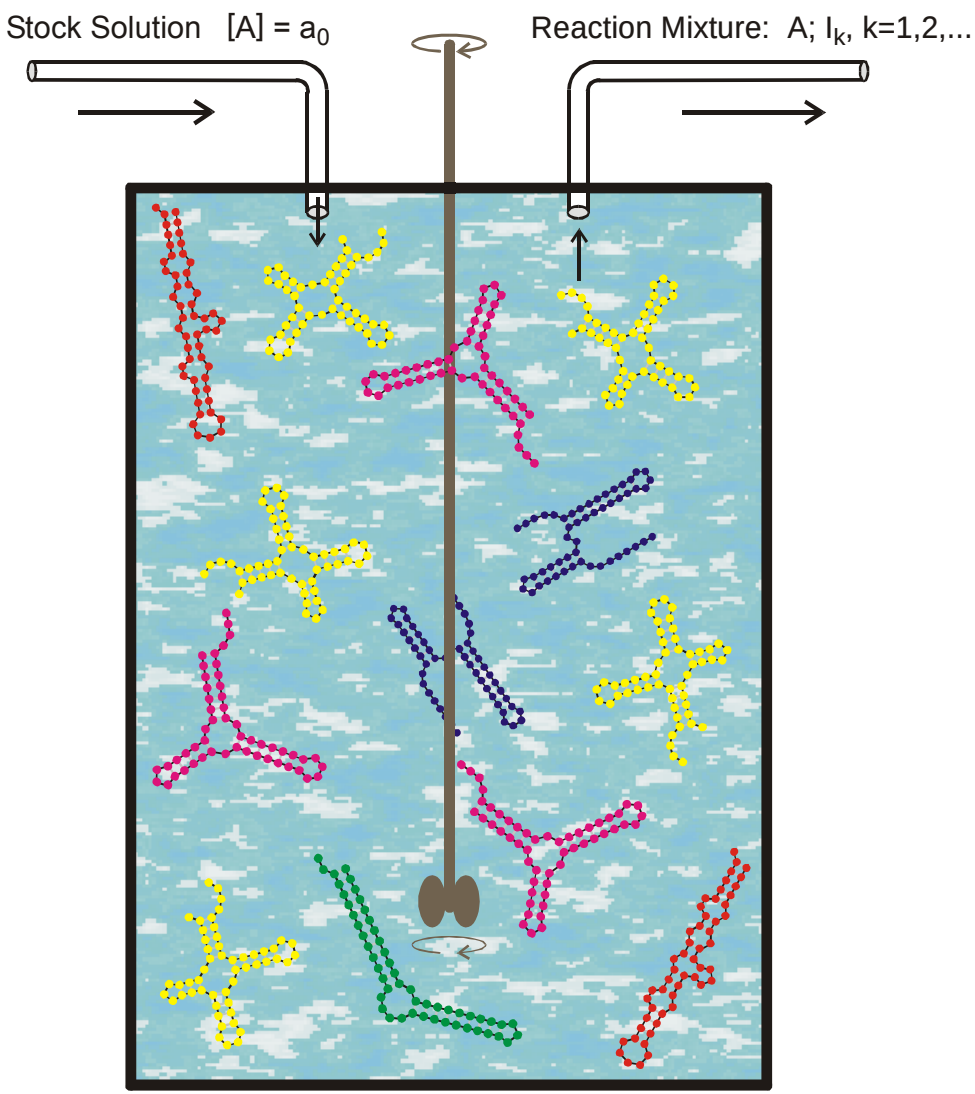
$$A + I_1 \underset{d_1}{\overset{k_1}{\rightleftharpoons}} 2\,I_1$$

$$A + I_2 \underset{d_2}{\overset{k_2}{\rightleftharpoons}} 2\,I_2$$

$$A + I_3 \underset{d_3}{\overset{k_3}{\rightleftharpoons}} 2\,I_3$$

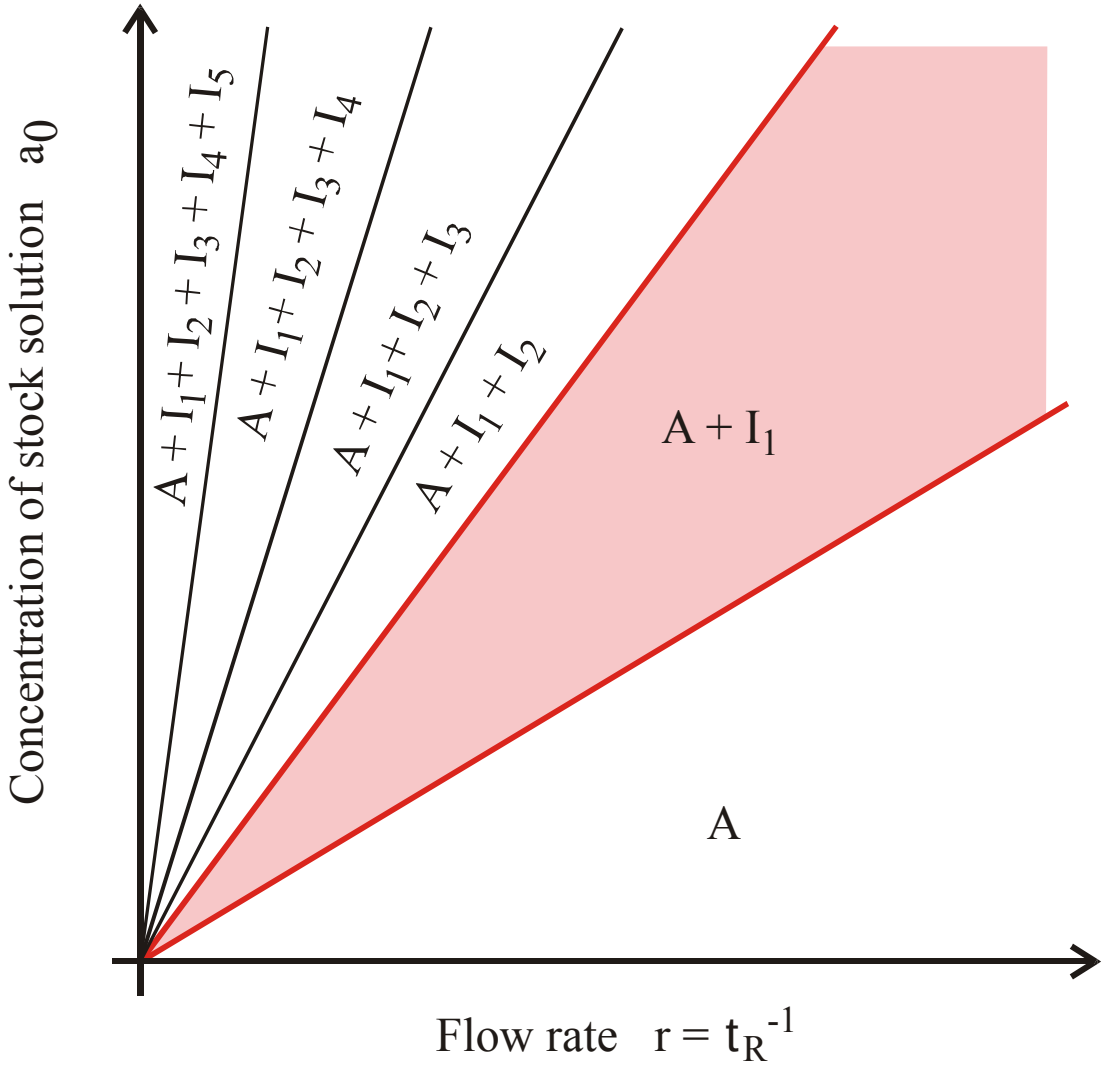
$$A + I_4 \underset{d_4}{\overset{k_4}{\rightleftharpoons}} 2\,I_4$$

$$A + I_5 \underset{d_5}{\overset{k_5}{\rightleftharpoons}} 2\,I_5$$



Replication in the flow reactor

P.Schuster & K.Sigmund, Dynamics of evolutionary optimization, *Ber.Bunsenges.Phys.Chem.* **89**: 668-682 (1985)



$$A + I_1 \underset{d_1}{\overset{k_1}{\rightleftarrows}} 2\,I_1$$

$$A + I_2 \underset{d_2}{\overset{k_2}{\rightleftarrows}} 2\,I_2$$

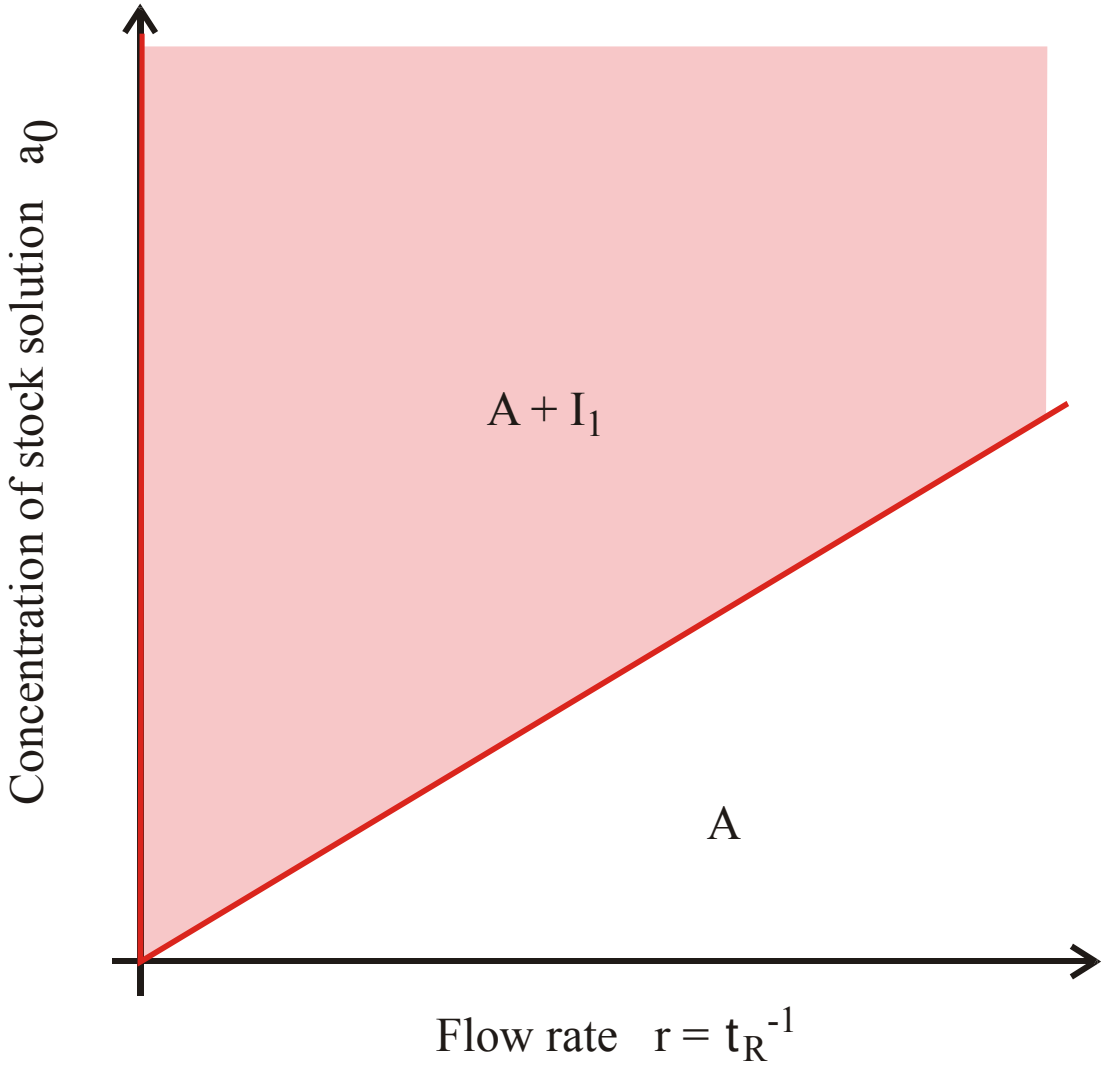
$$A + I_3 \underset{d_3}{\overset{k_3}{\rightleftarrows}} 2\,I_3$$

$$A + I_4 \underset{d_4}{\overset{k_4}{\rightleftarrows}} 2\,I_4$$

$$A + I_5 \underset{d_5}{\overset{k_5}{\rightleftarrows}} 2\,I_5$$

$$k_1 > k_2 > k_3 > k_4 > k_5$$

Selection in the flow reactor: Reversible replication reactions



$$A + I_1 \xrightarrow{f_1} 2\,I_1$$

$$A + I_2 \xrightarrow{f_2} 2\,I_2$$

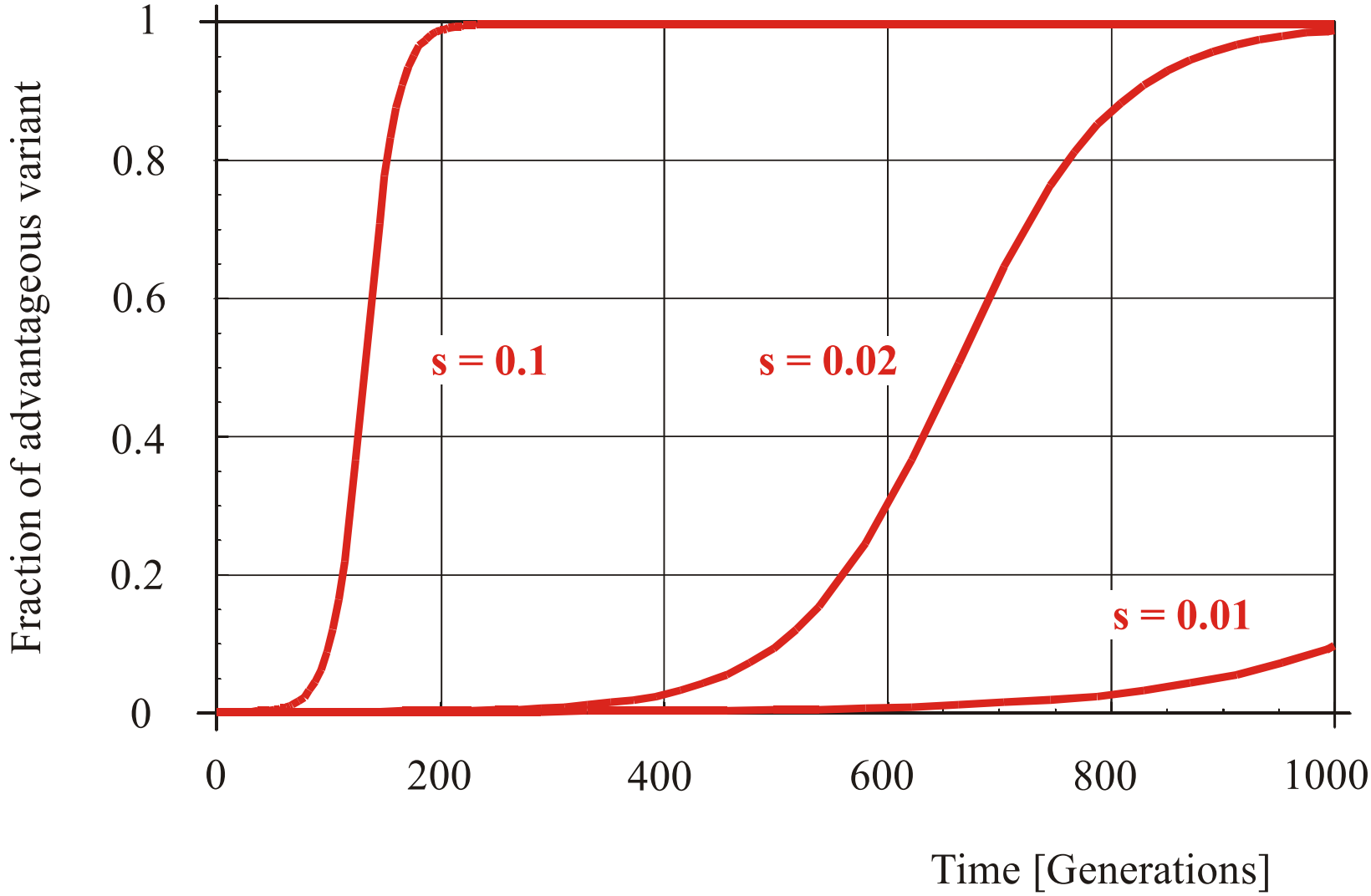
$$A + I_3 \xrightarrow{f_3} 2\,I_3$$

$$A + I_4 \xrightarrow{f_4} 2\,I_4$$

$$A + I_5 \xrightarrow{f_5} 2\,I_5$$

$$f_1 > f_2 > f_3 > f_4 > f_5$$

Selection in the flow reactor: Irreversible replication reactions



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

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

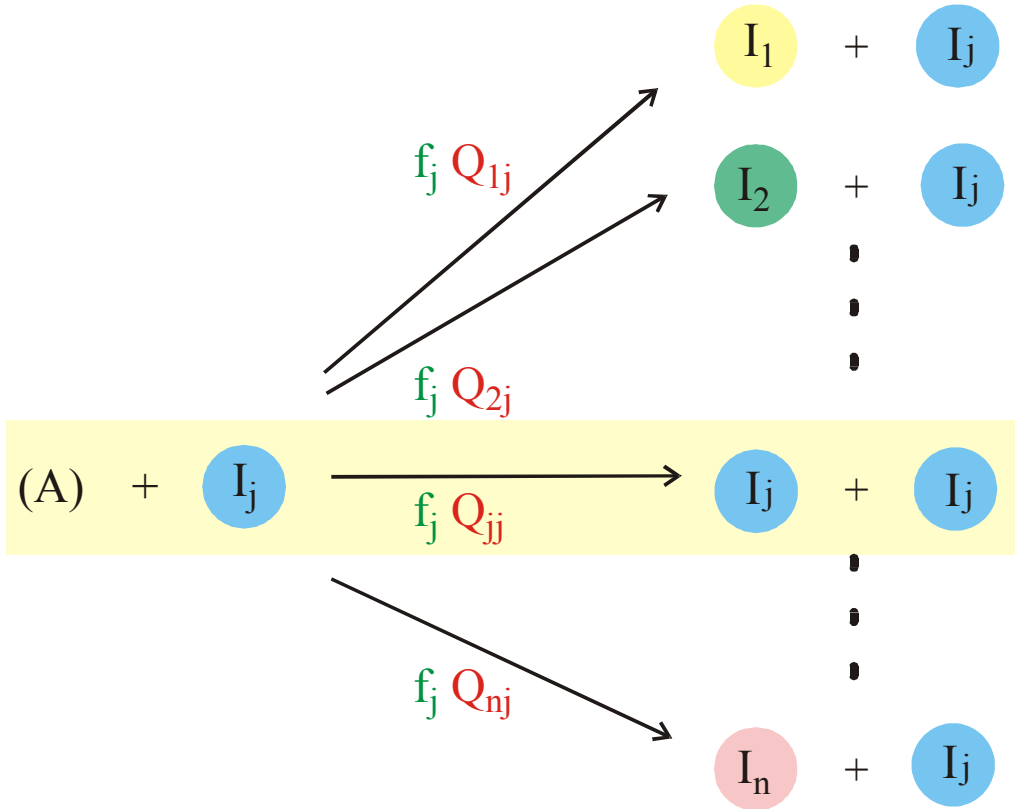
$$\Phi = \Sigma_i f_i x_i ; \quad \Sigma_i x_i = 1 ; \quad \Sigma_i Q_{ij} = 1$$

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

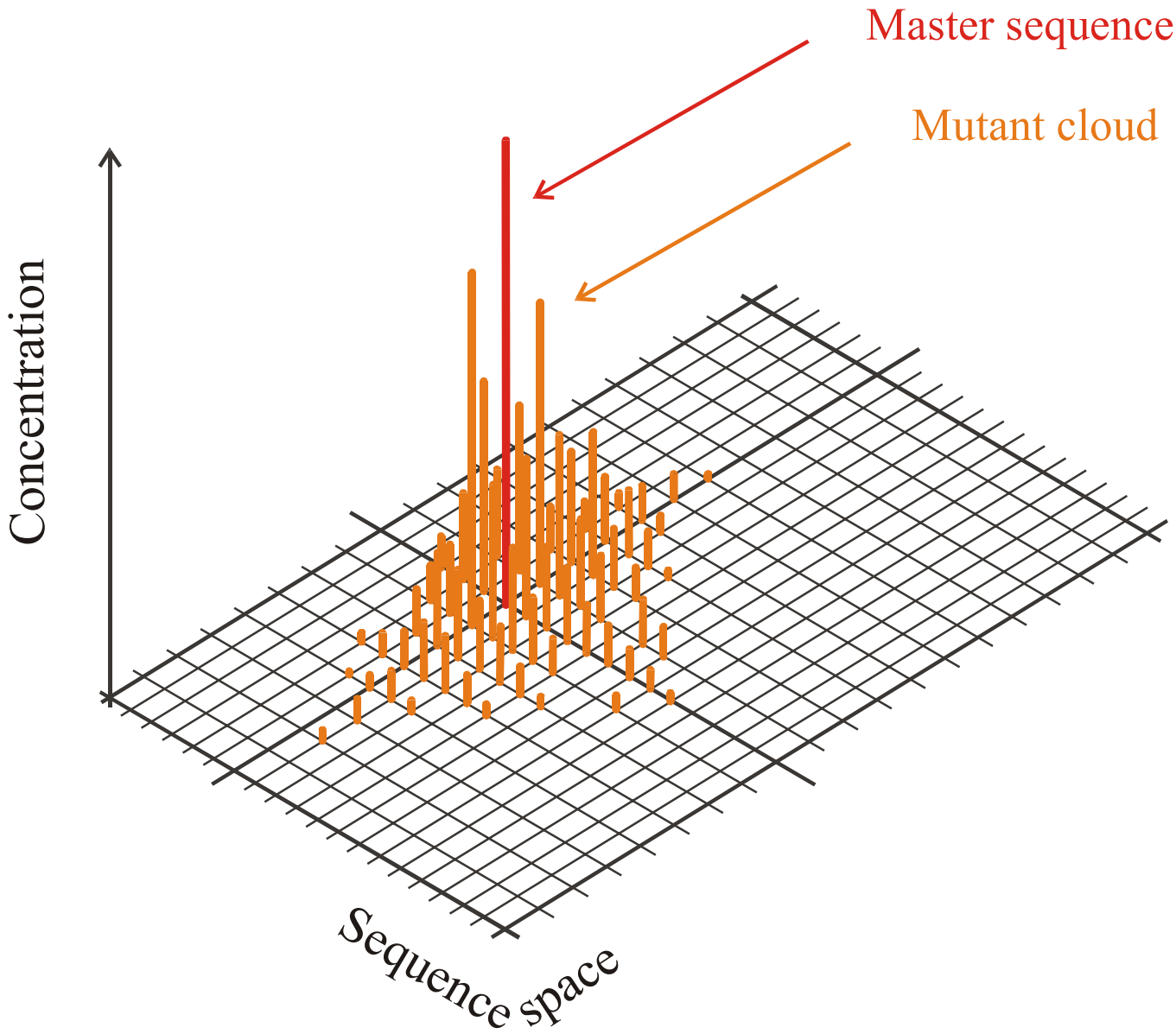
p ..... Error rate per digit

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

[A] = a = constant



Chemical kinetics of replication and mutation

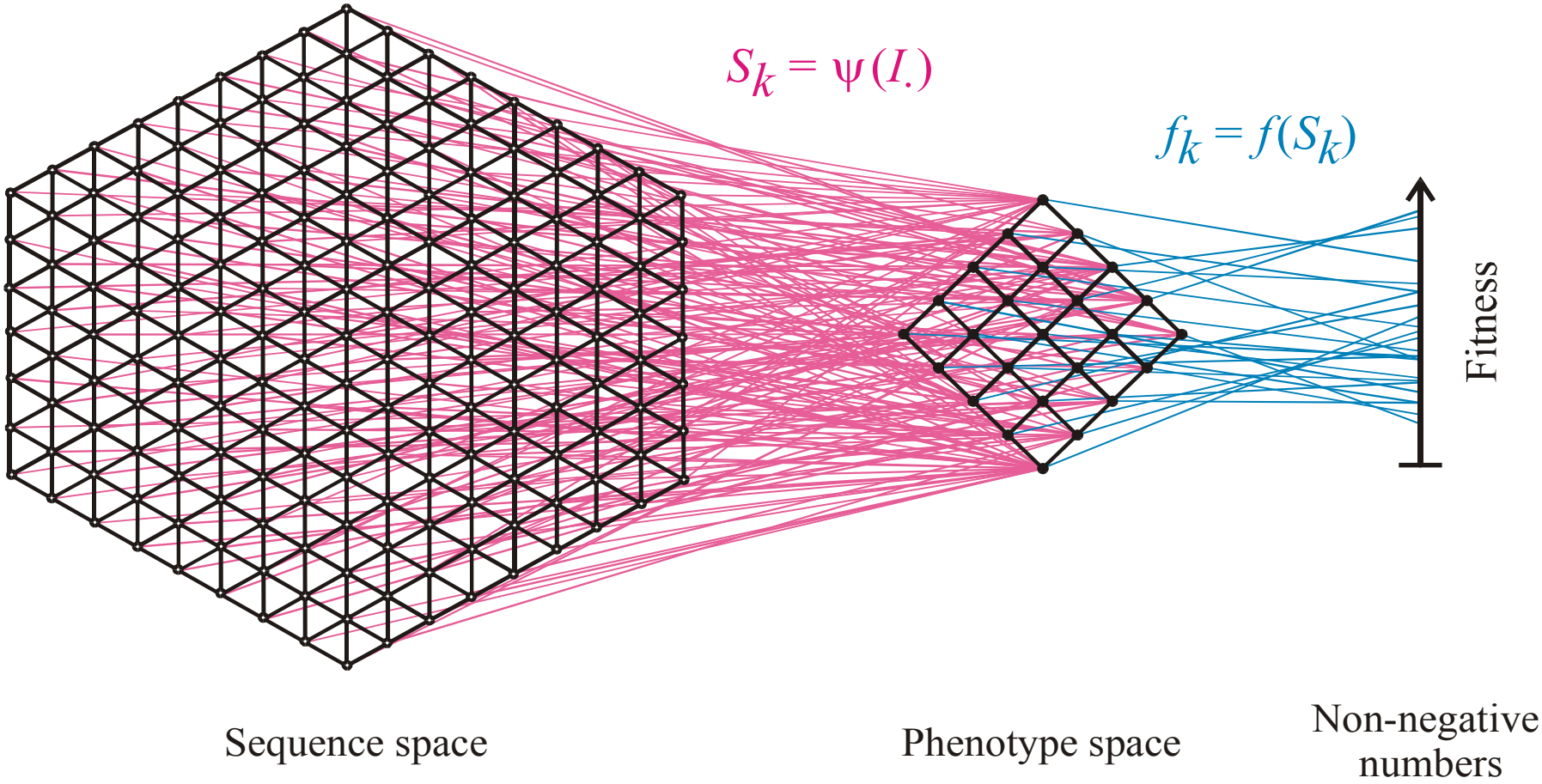


The molecular quasispecies in sequence space

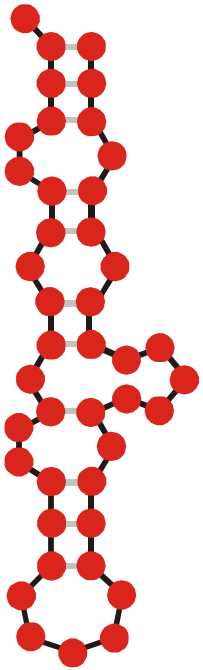
The **RNA model** considers RNA sequences as genotypes and simplified RNA structures, called secondary structures, as phenotypes.

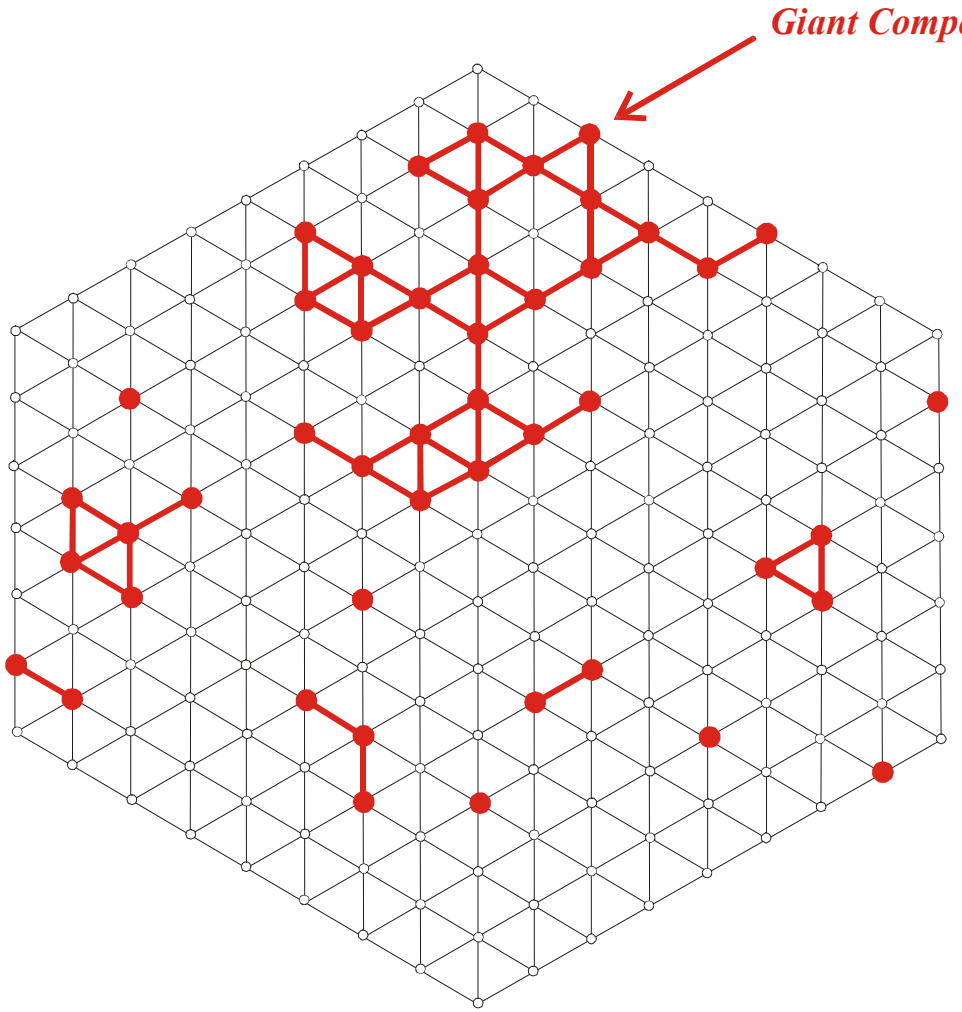
The **mapping** from genotypes into phenotypes is many-to-one. Hence, it is redundant and not invertible.

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.

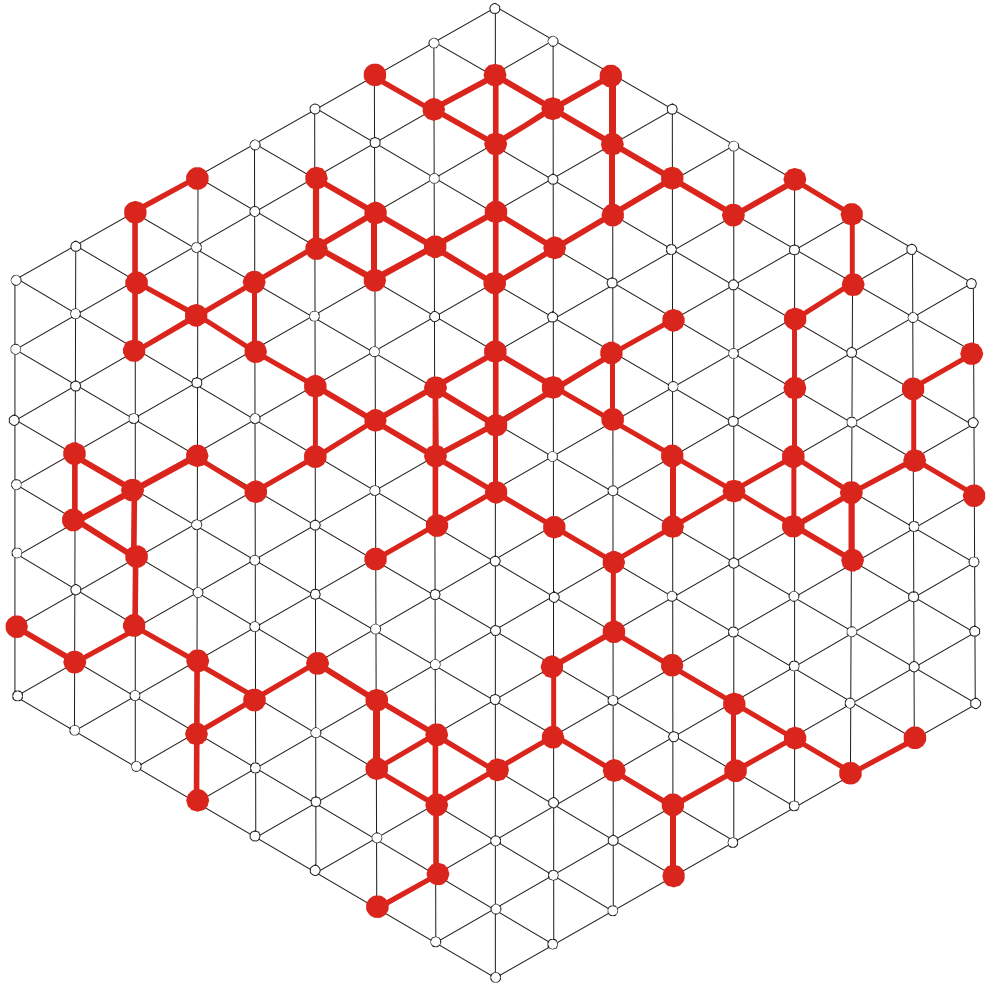


Mapping from sequence space into phenotype space and into fitness values





A multi-component neutral network



A connected neutral network

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

Evolution in the Flow Reactor: The RNA Model

Sequence-structure map $\psi : \left\{I; d_{ij}^{h}\right\} \Rightarrow \left\{S; d_{ij}^{s}\right\}.$

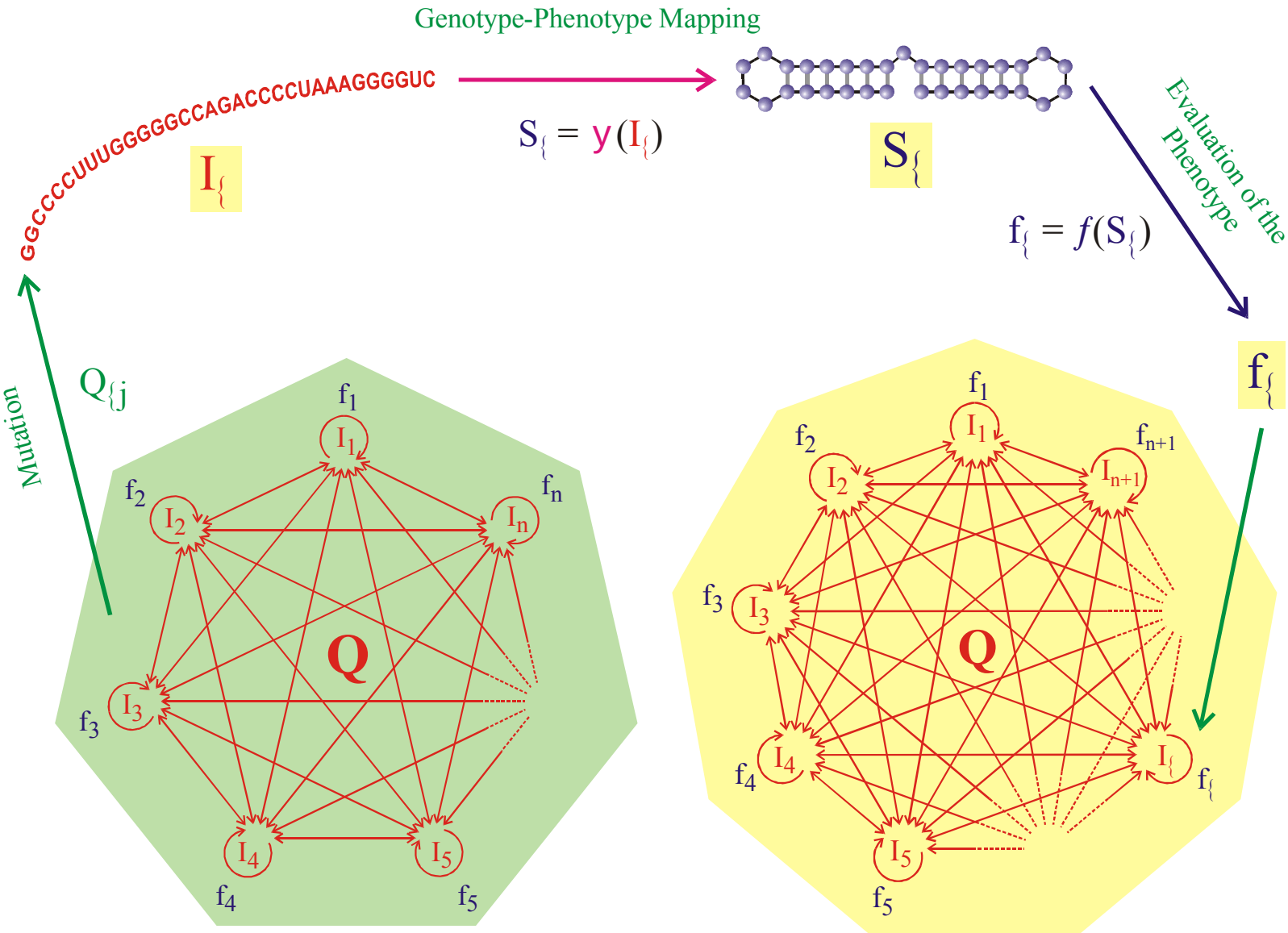
Structure-function map $f : \left\{S; d_{ij}^{s}\right\} \Rightarrow \Re_{+}$.

Environment $\Omega(t)$

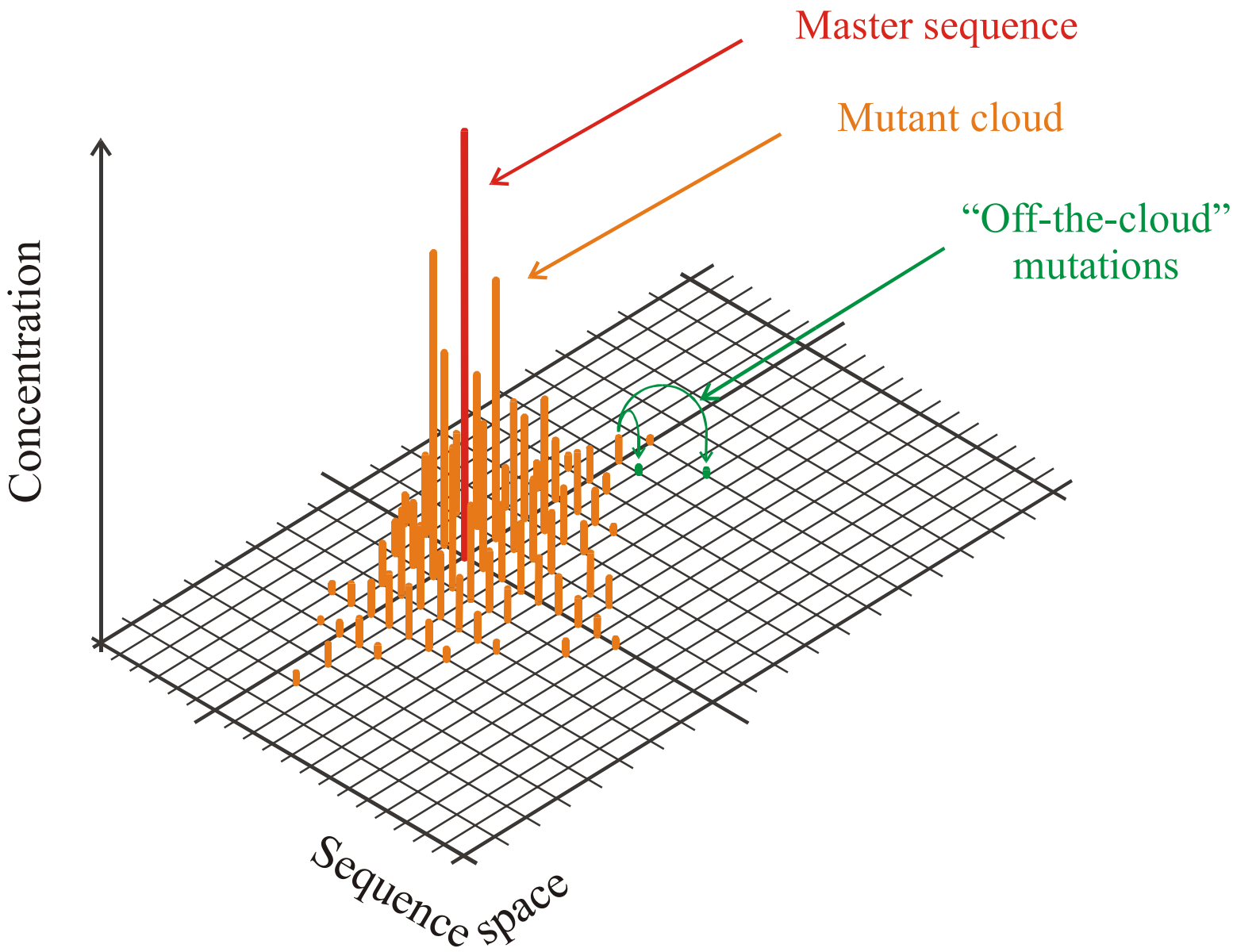
Mapping into fitness values $f_k(t) = f\left(S_k, \Omega(t)\right) = f\left(\psi(I_k, \Omega(t)), \Omega(t)\right)$

$$\frac{dx_k}{dt} = x_k\left(Q_{kk}\, f_k(t) - \Phi(t)\right) + \sum_{j=1,\, j\neq k}^{n} Q_{kj}\, f_j(t)\, x_j + \eta_k(x_k, t)\,\omega(t), \quad k = 1,\ldots,n$$

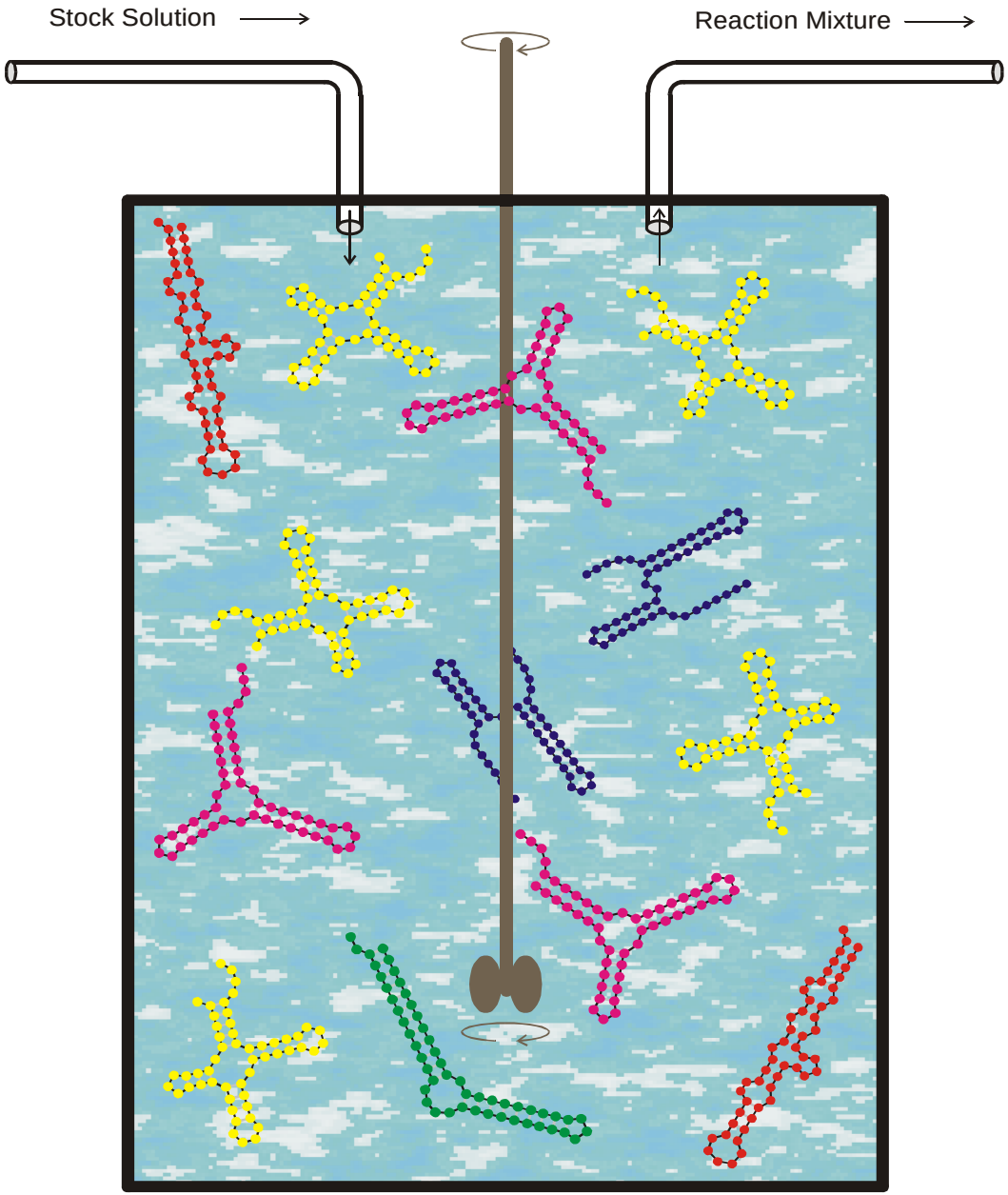
Wiener process $dW(t) = \omega(t)\, dt$



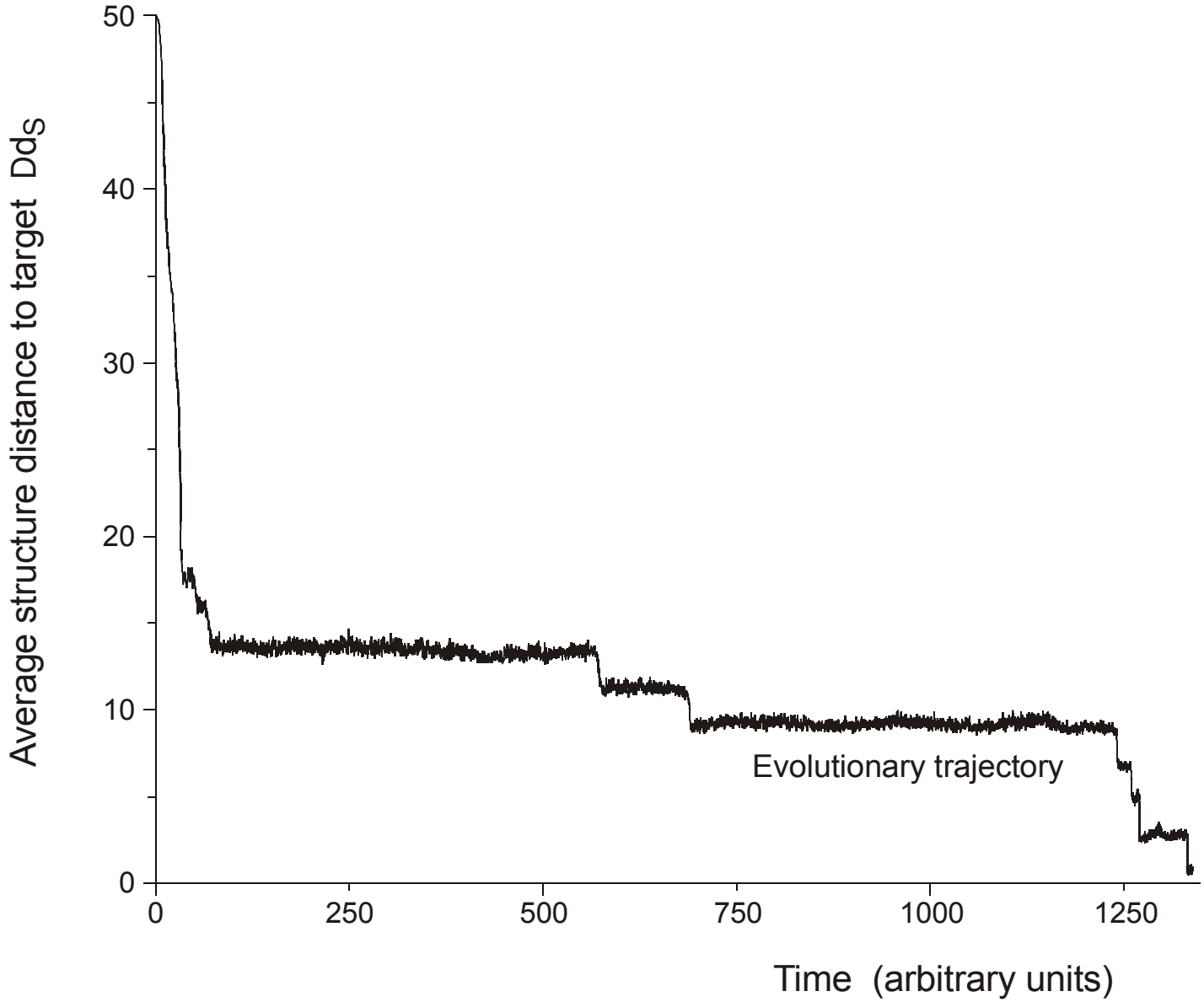
Evolutionary dynamics
including molecular phenotypes



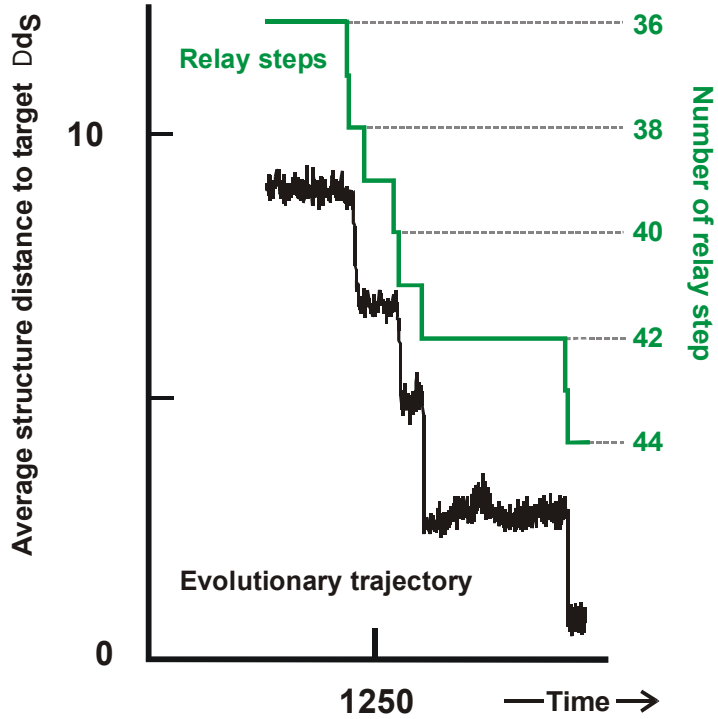
The molecular quasispecies and mutations producing new variants

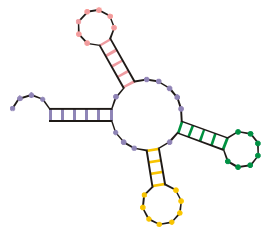


The flowreactor as a device for studies of evolution ***in vitro*** and ***in silico***



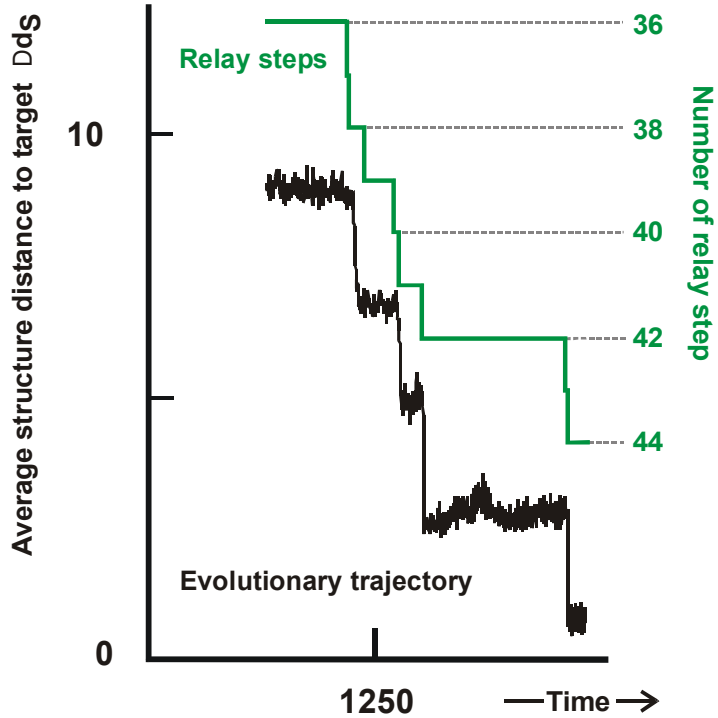
In silico optimization in the flow reactor: Trajectory

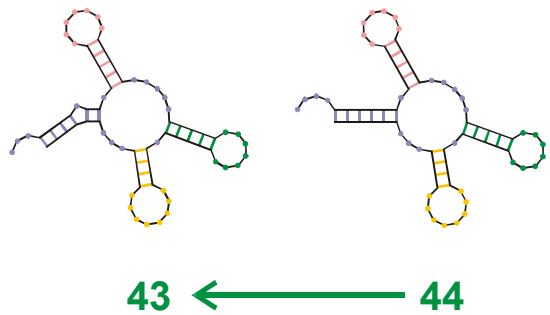




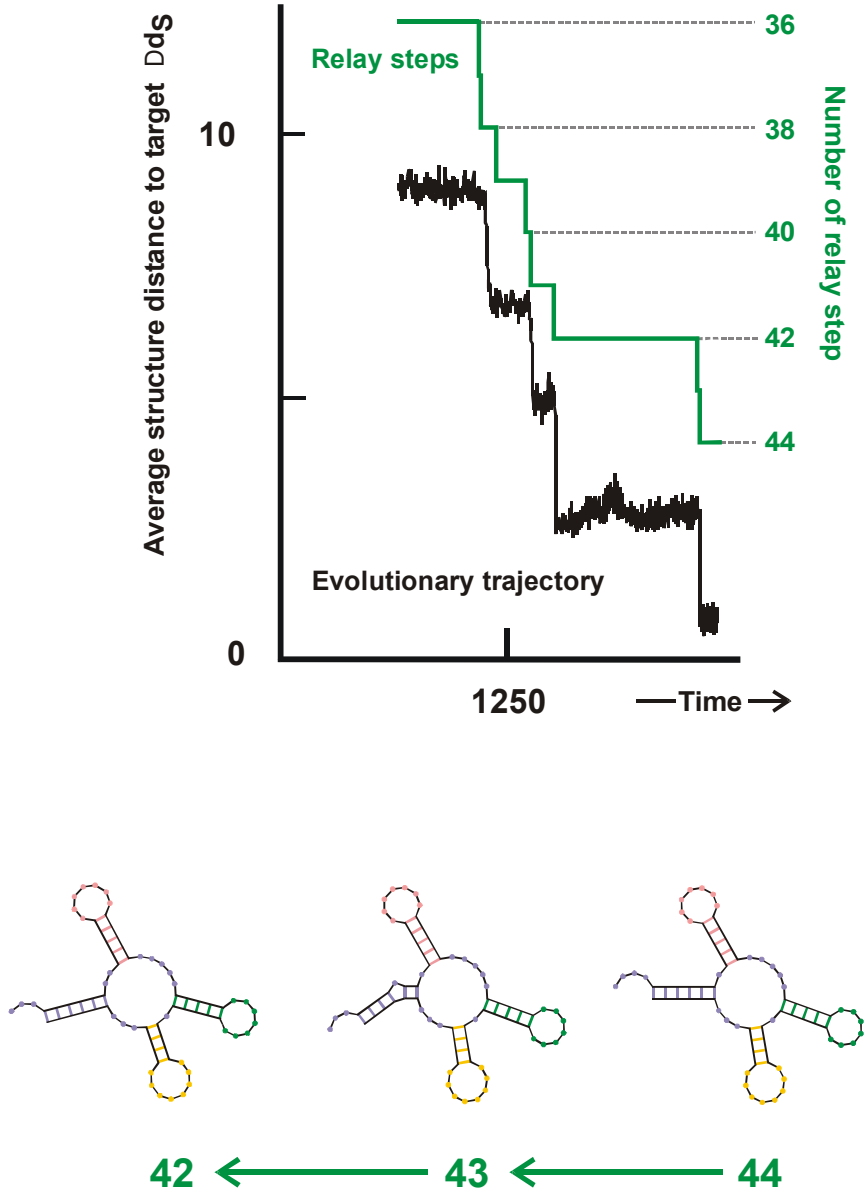
44

Endconformation of optimization

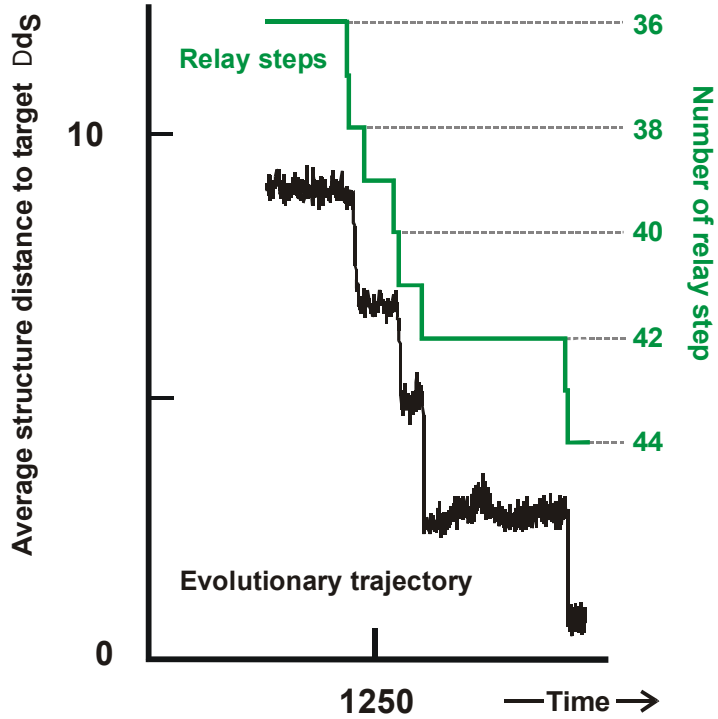


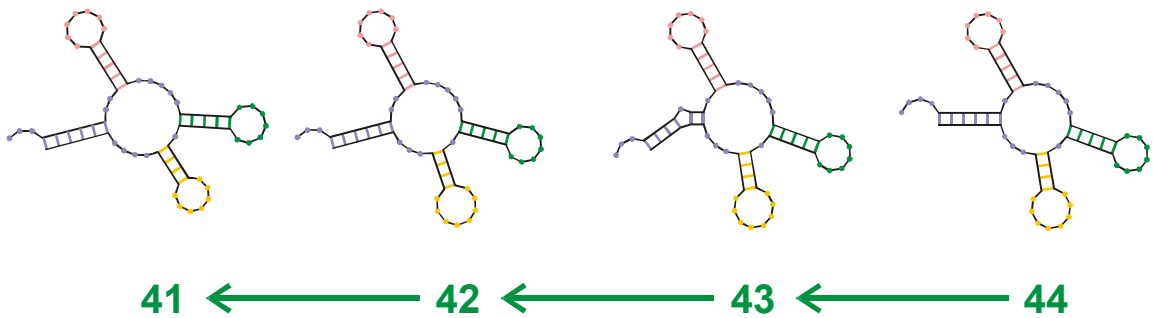


Reconstruction of the last step 43 Á 44

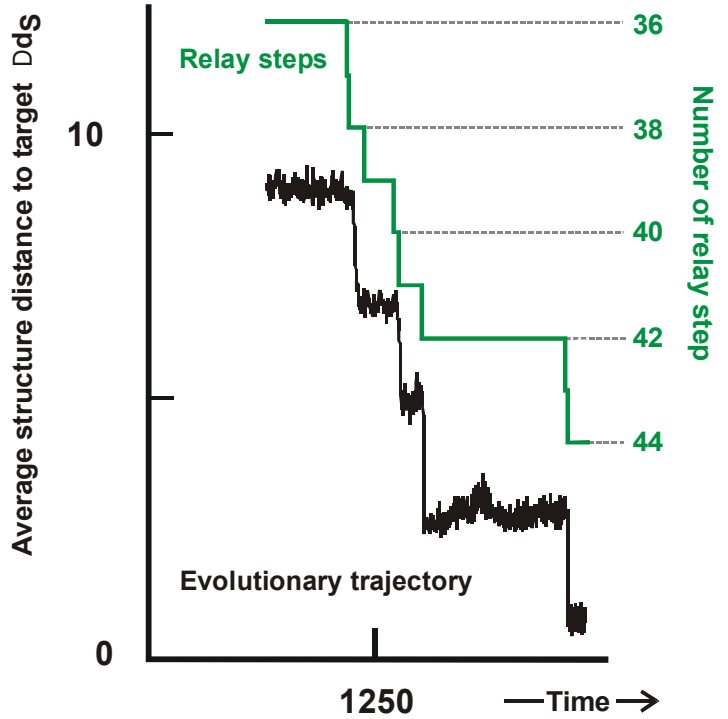


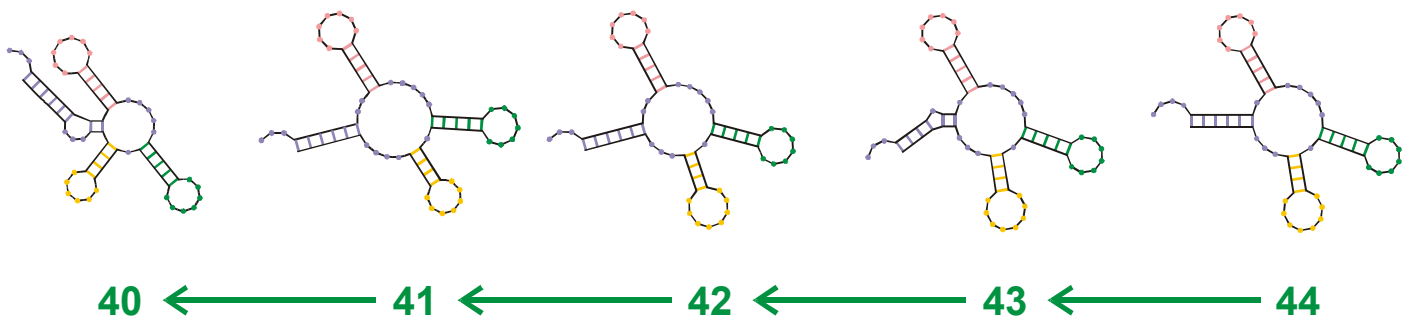
Reconstruction of last-but-one step 42 Á 43 (Á 44)



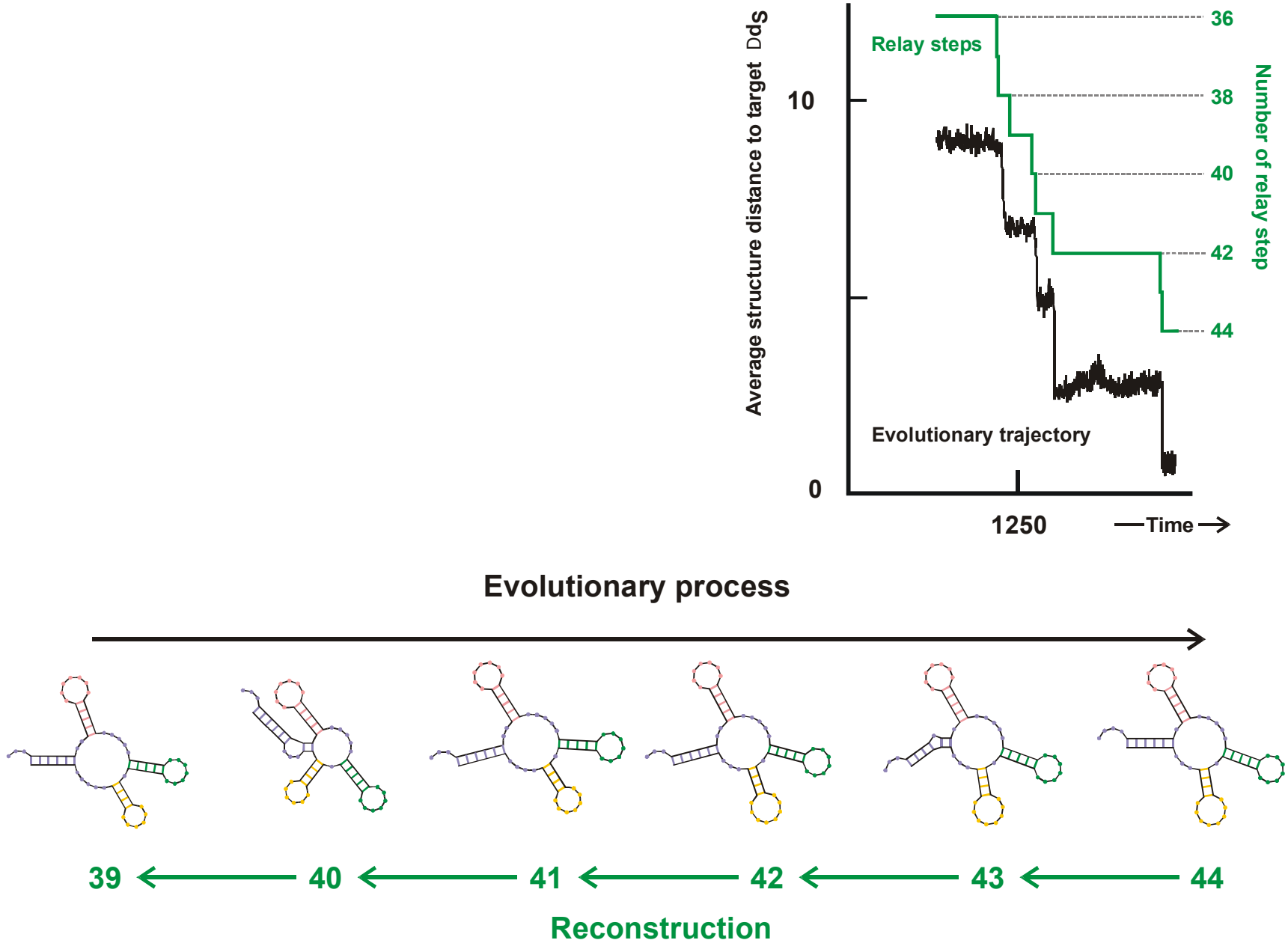


Reconstruction of step 41 Á 42 (Á 43 Á 44)

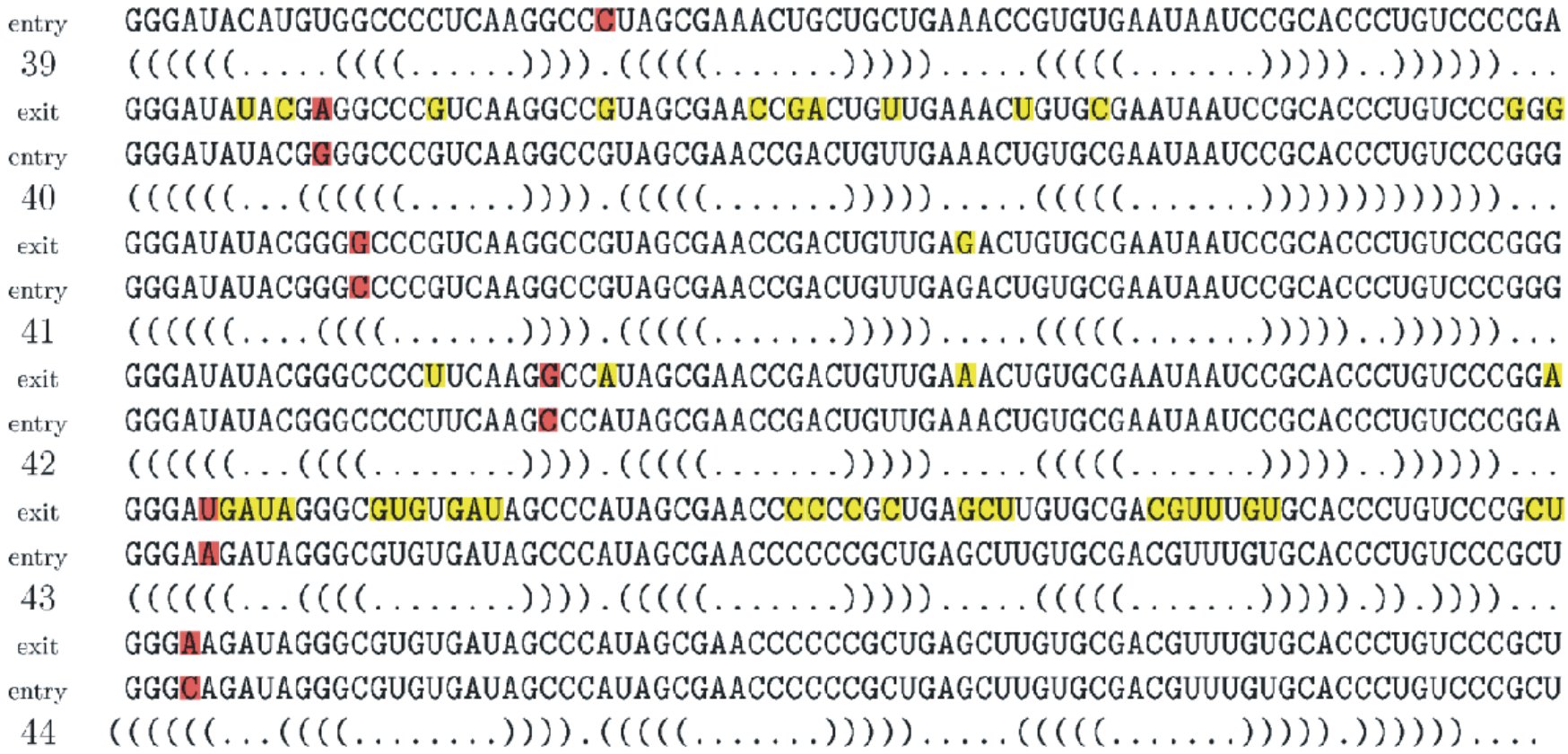




Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)

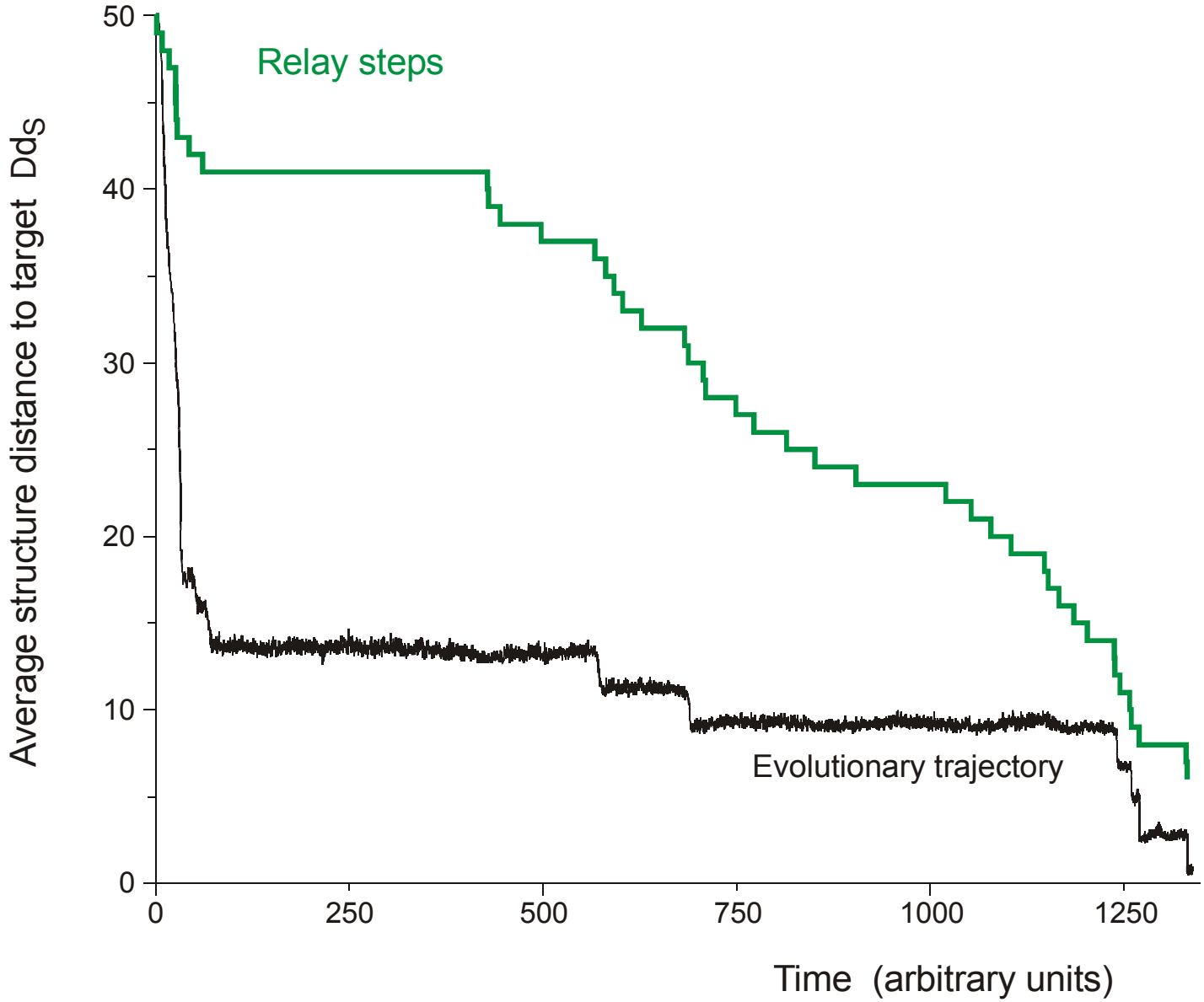


Reconstruction of the relay series

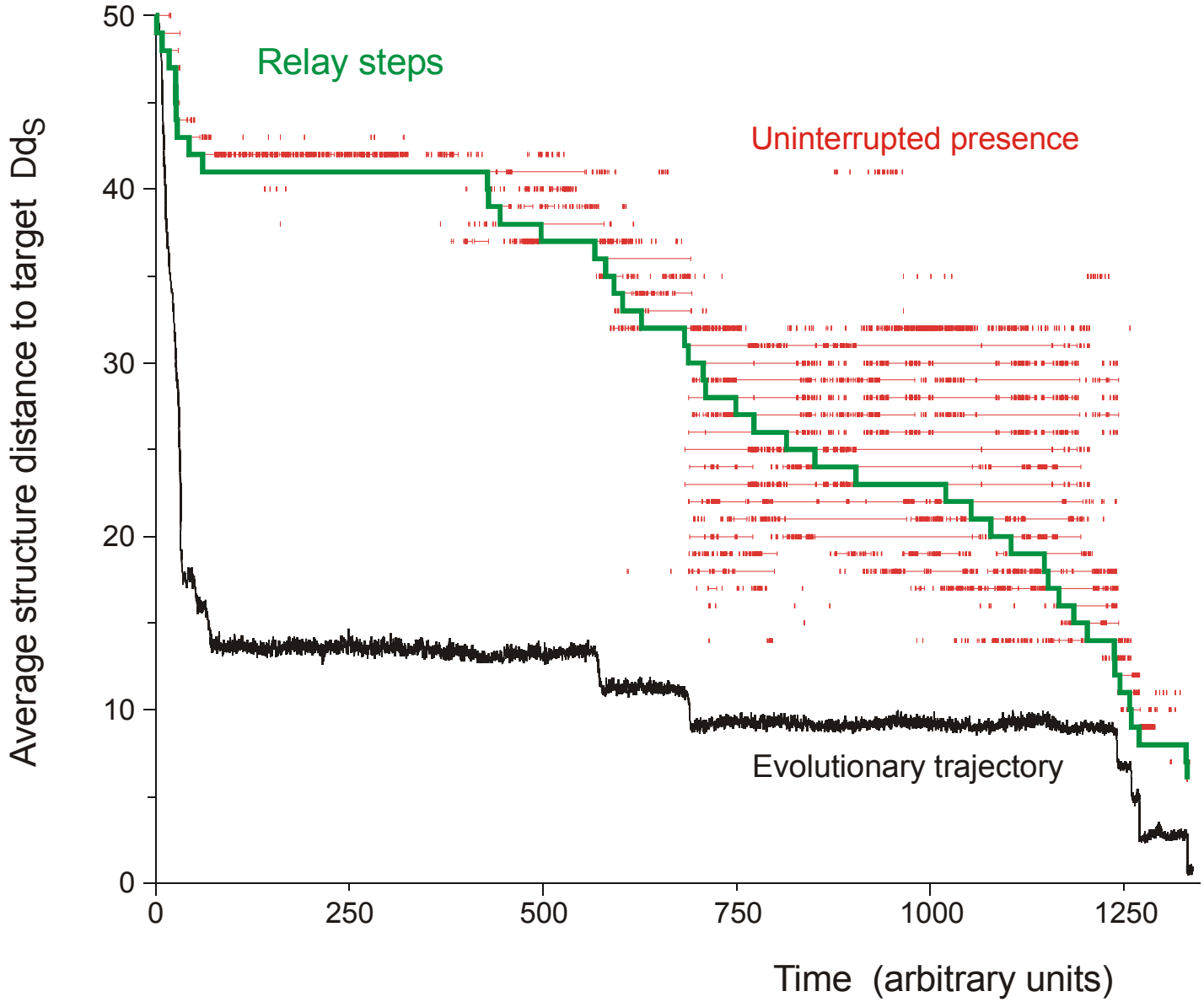


Transition inducing point mutations **Neutral point mutations**

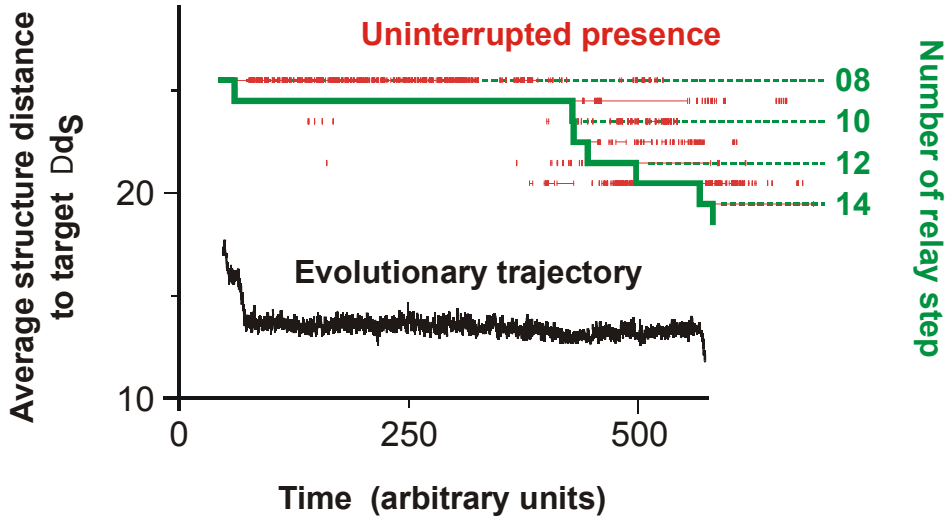
Change in RNA sequences during the final five relay steps 39 Á 44



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



In silico optimization in the flow reactor: Uninterrupted presence

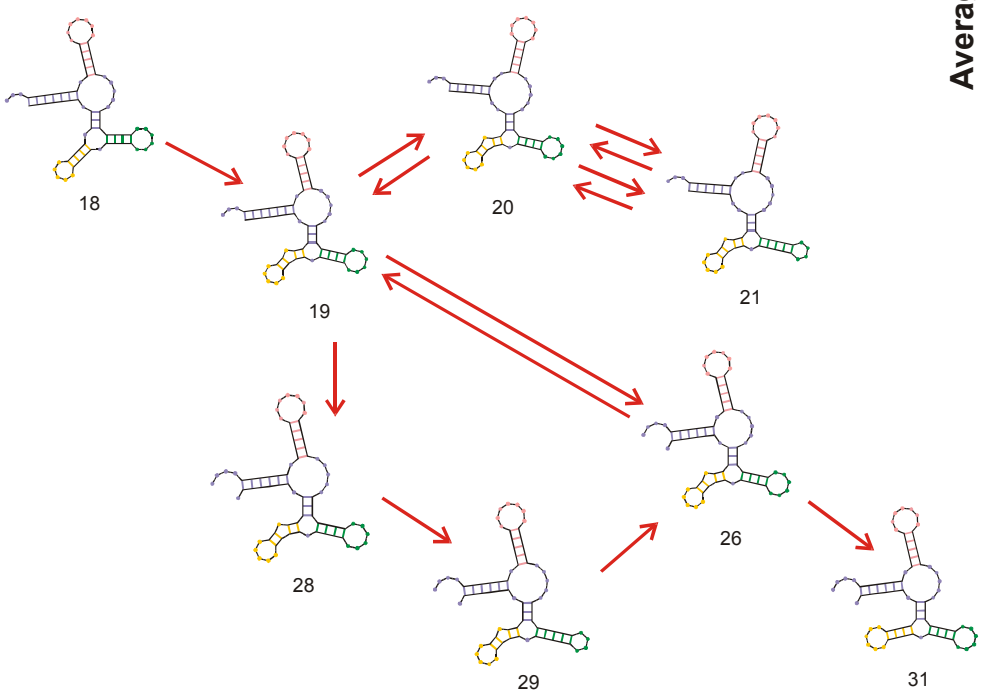


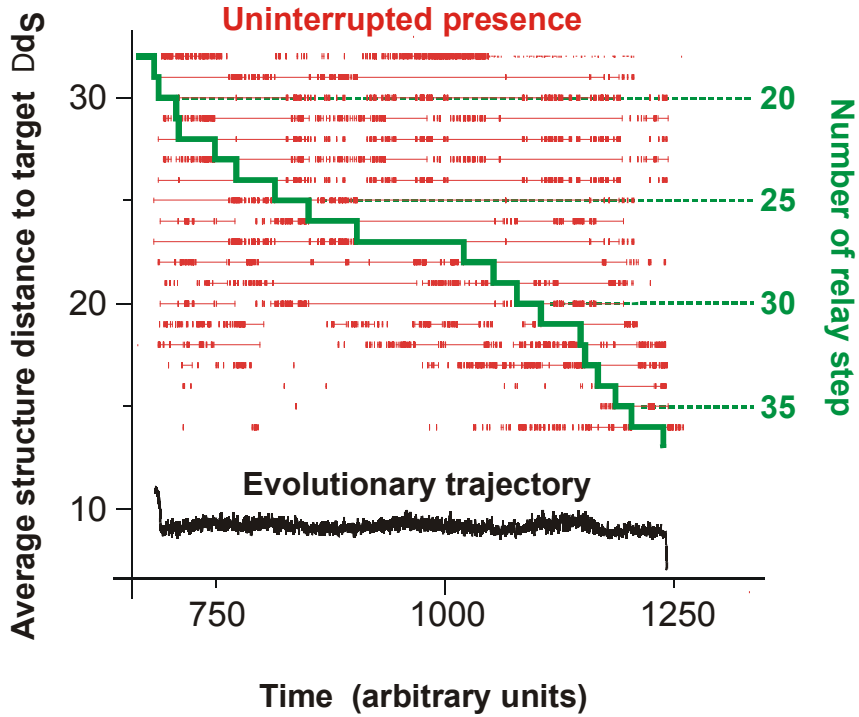

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations

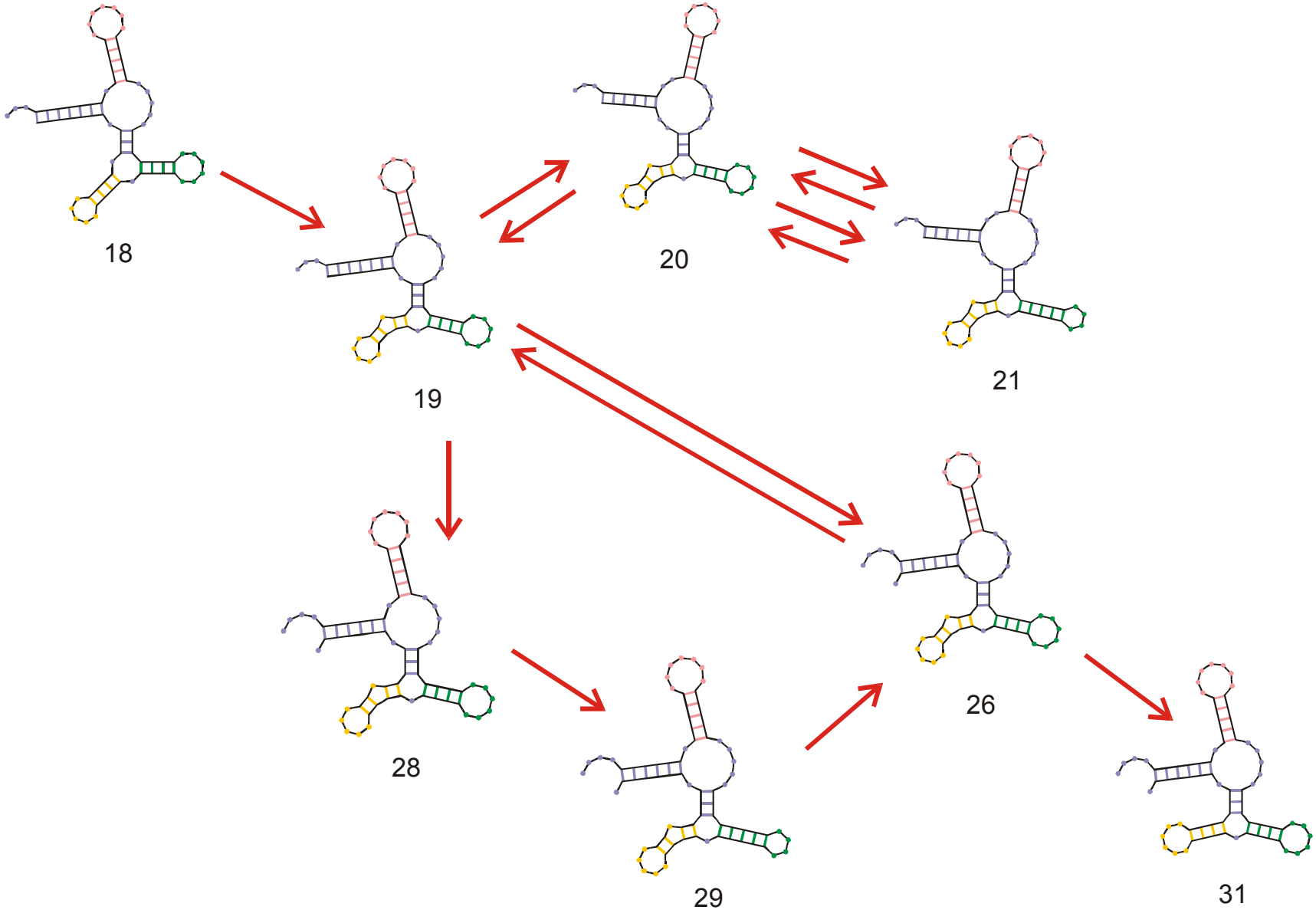
Neutral point mutations

Neutral genotype evolution during phenotypic stasis

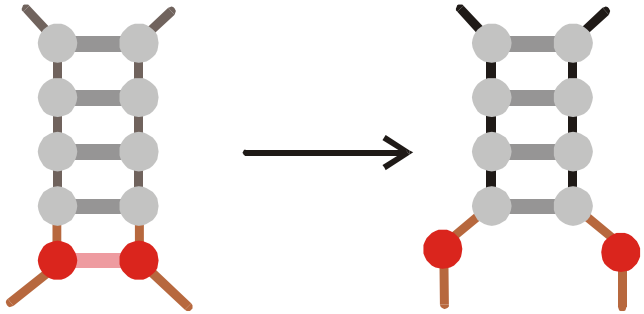




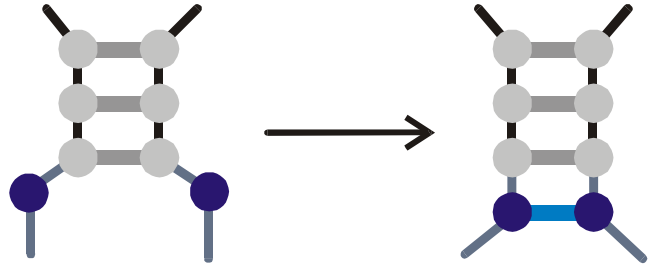
A random sequence of **minor** or continuous **transitions** in the relay series



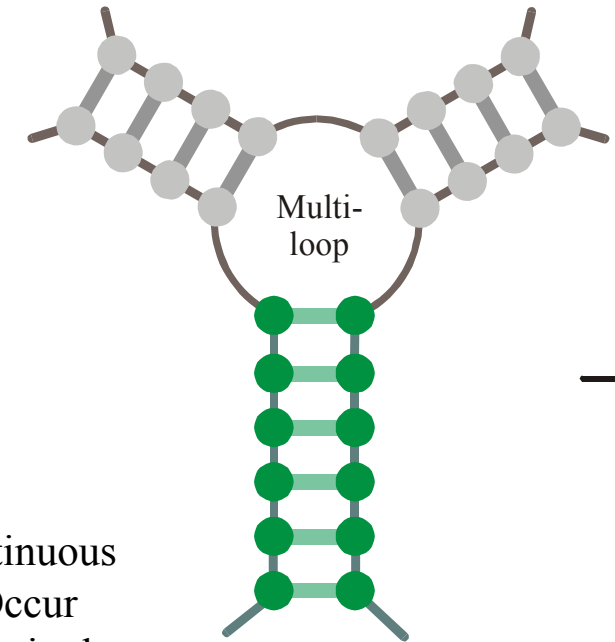
A random sequence of **minor** or continuous **transitions** in the relay series

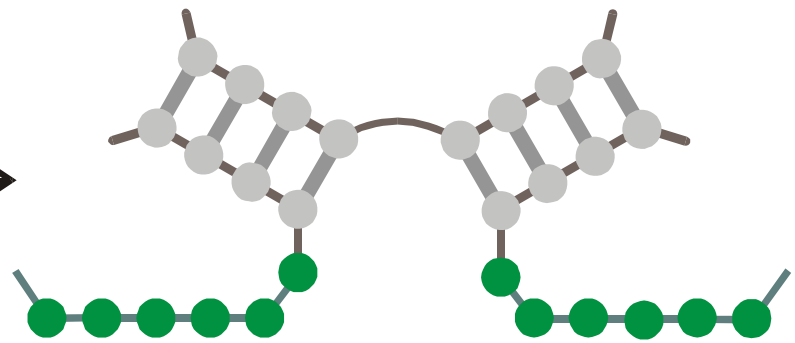


Shortening of Stacks



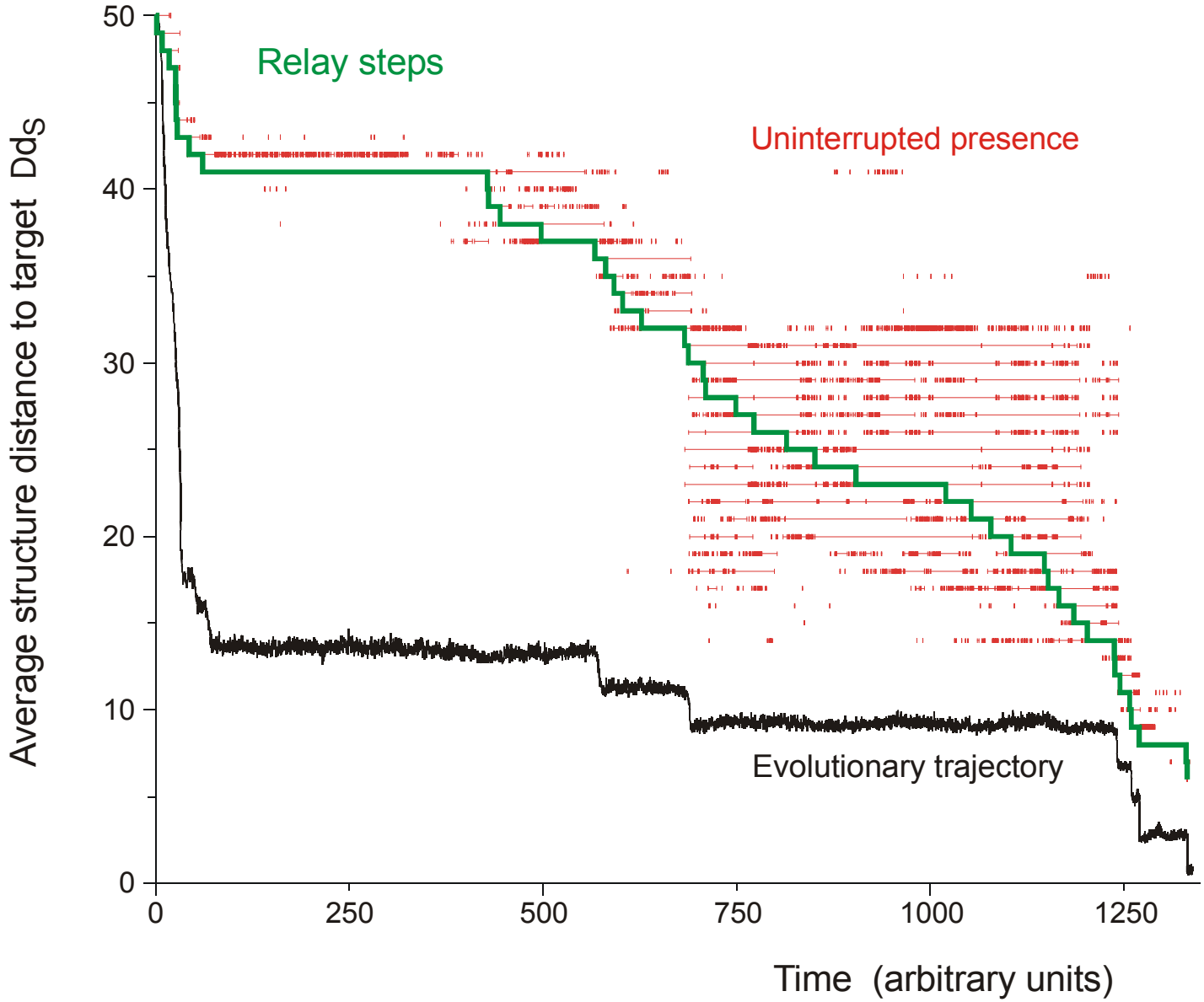
Elongation of Stacks



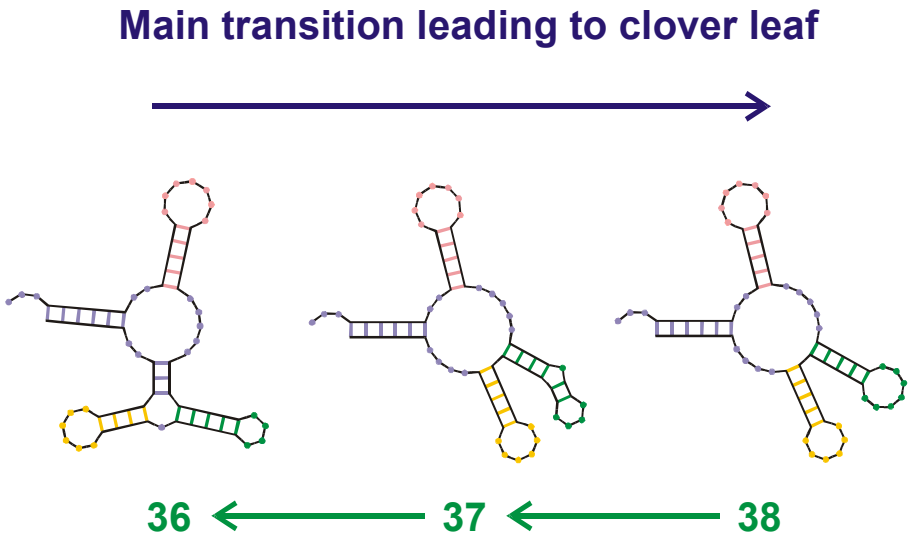


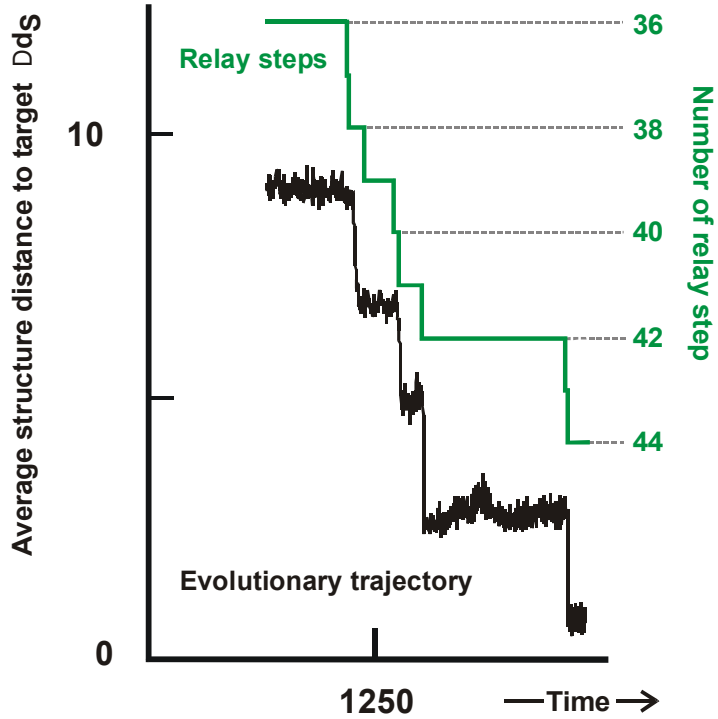
Opening of Constrained Stacks

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

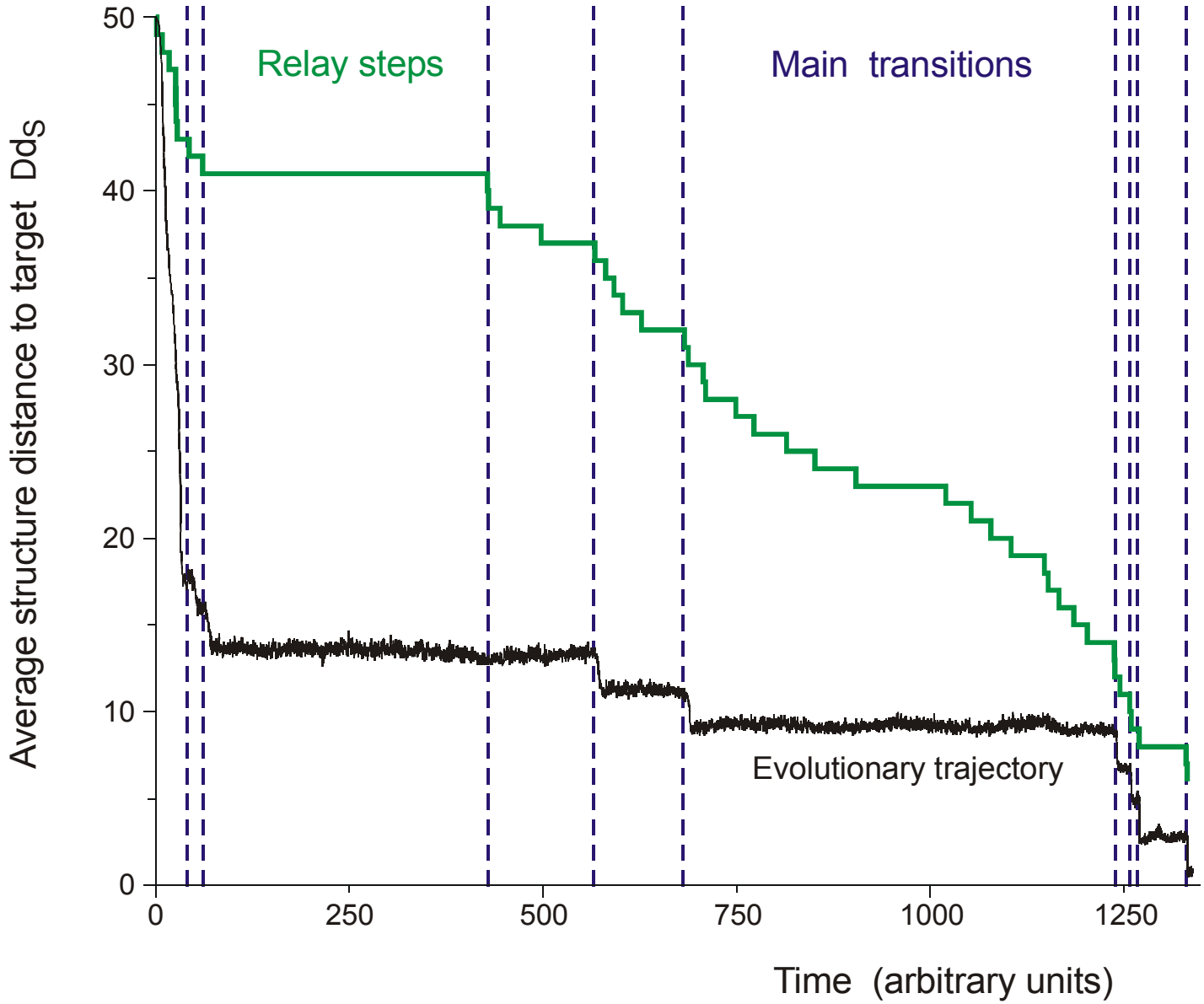


In silico optimization in the flow reactor: **Uninterrupted presence**

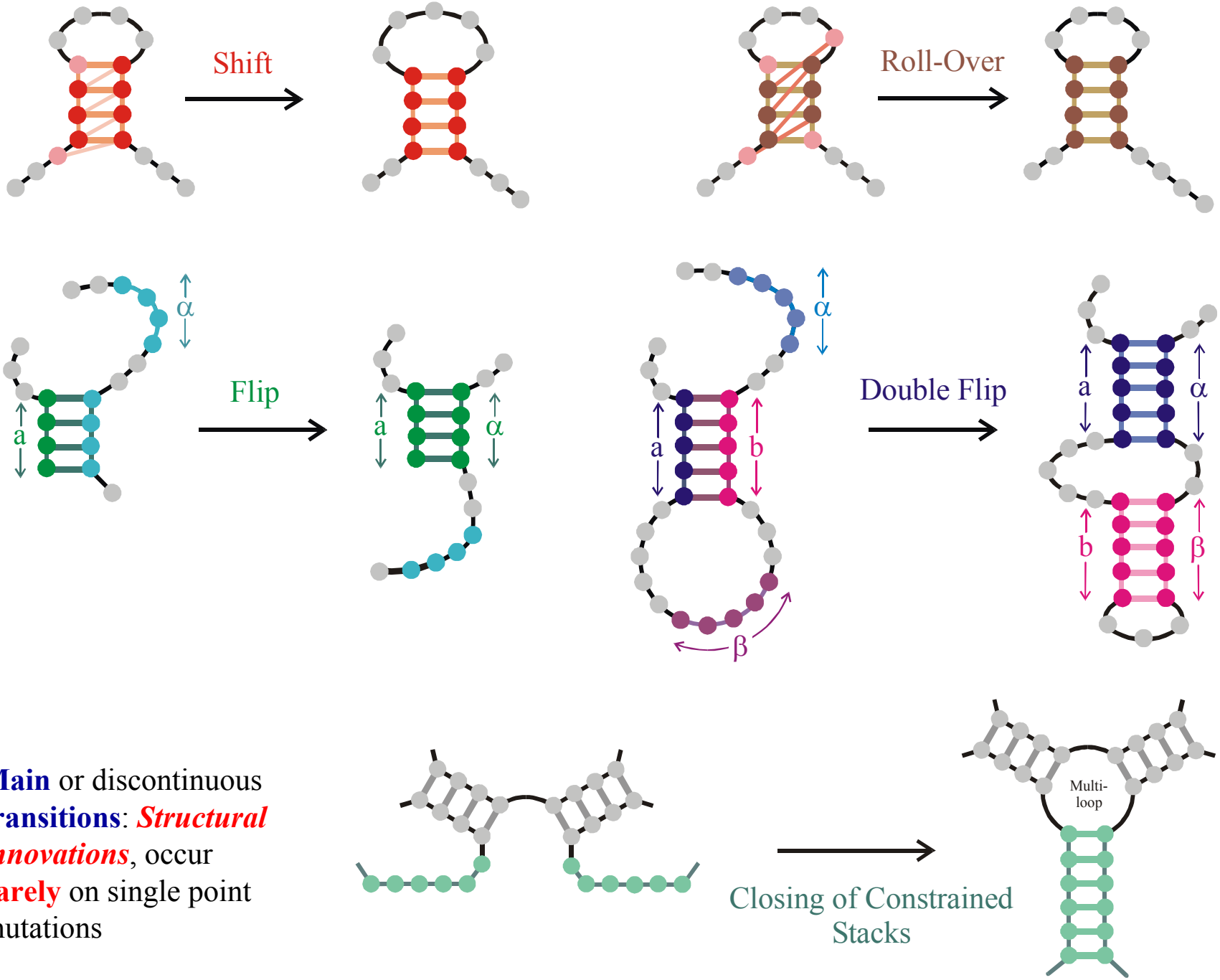




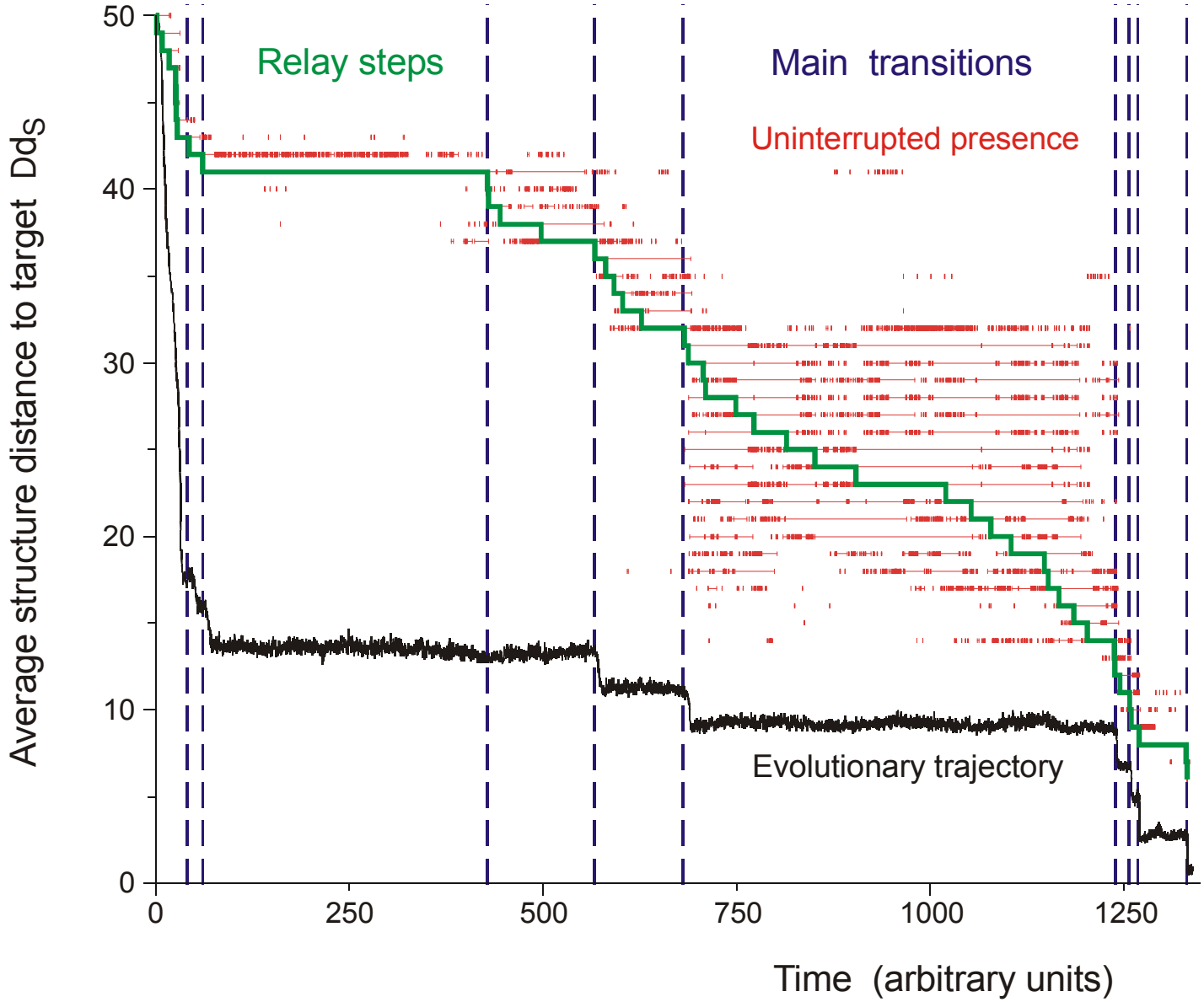
Reconstruction of a main transitions 36 Á 37 (Á 38)



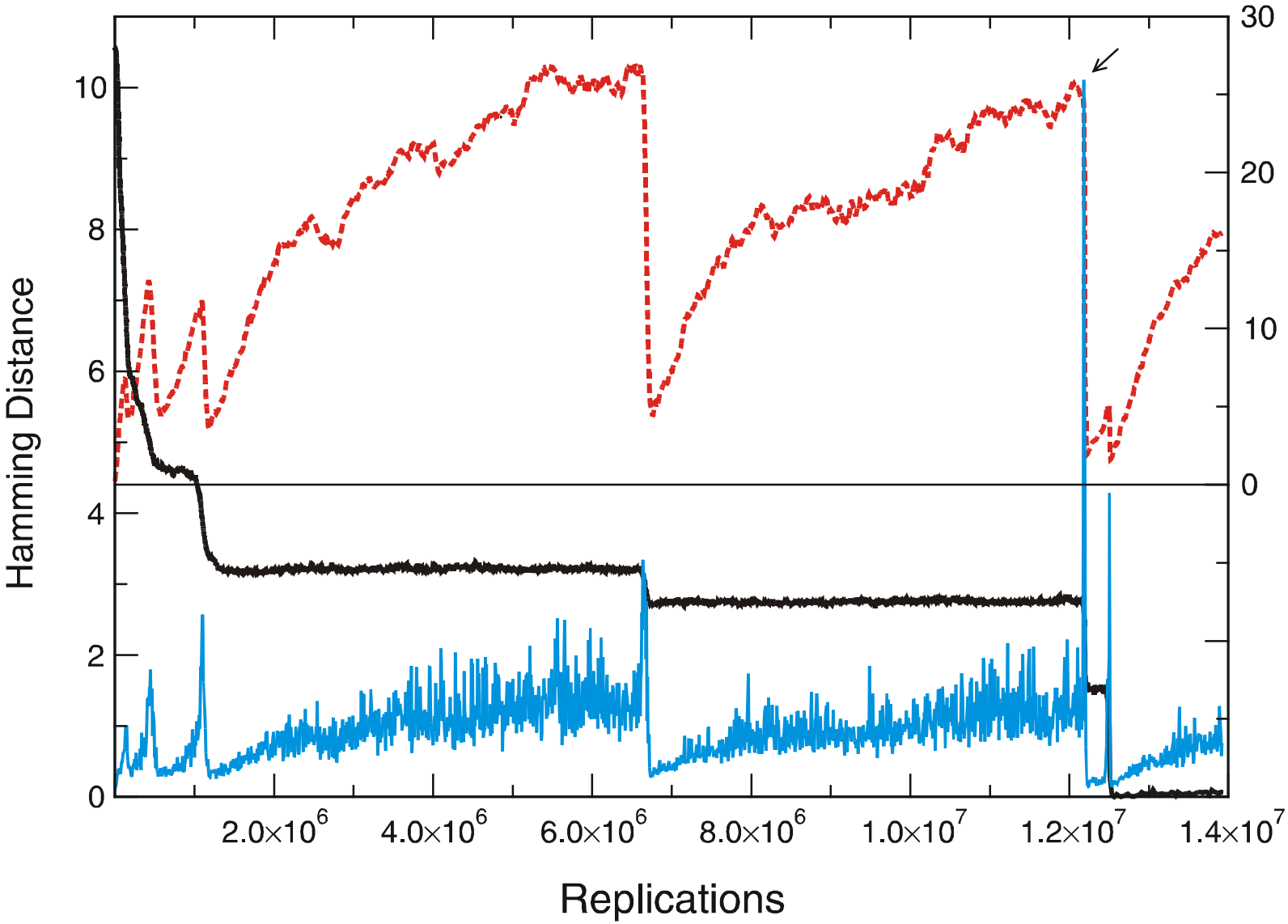
In silico optimization in the flow reactor: Main transitions



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



In silico optimization in the flow reactor

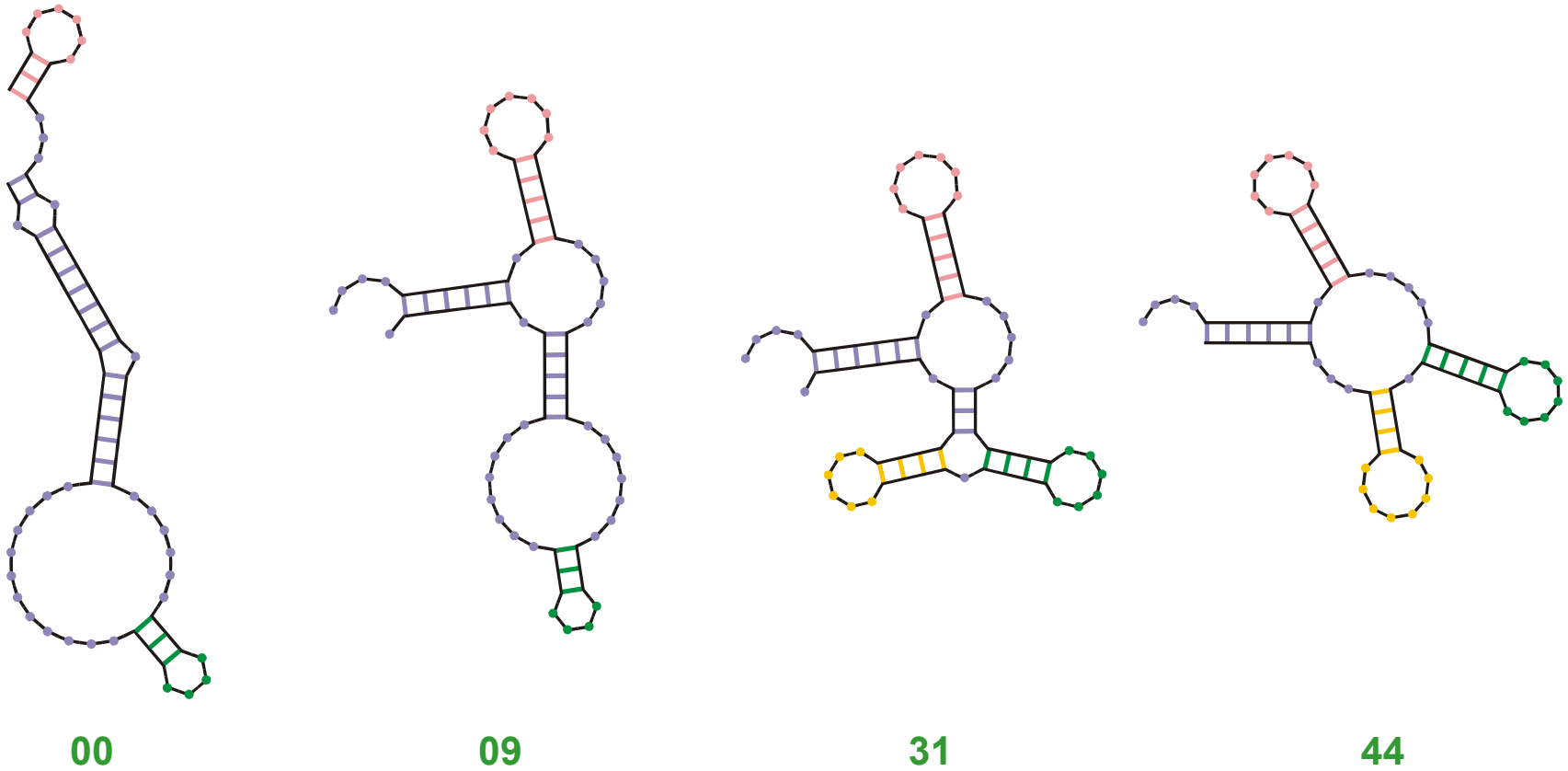


Variation in genotype space during optimization of phenotypes

Statistics of evolutionary trajectories

Population size N	Number of replications $< n_{rep} >$	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.



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

Main results of computer simulations of molecular evolution

• No trajectory was reproducible in detail. Sequences of target structures were different. Nevertheless solutions of comparable or 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.

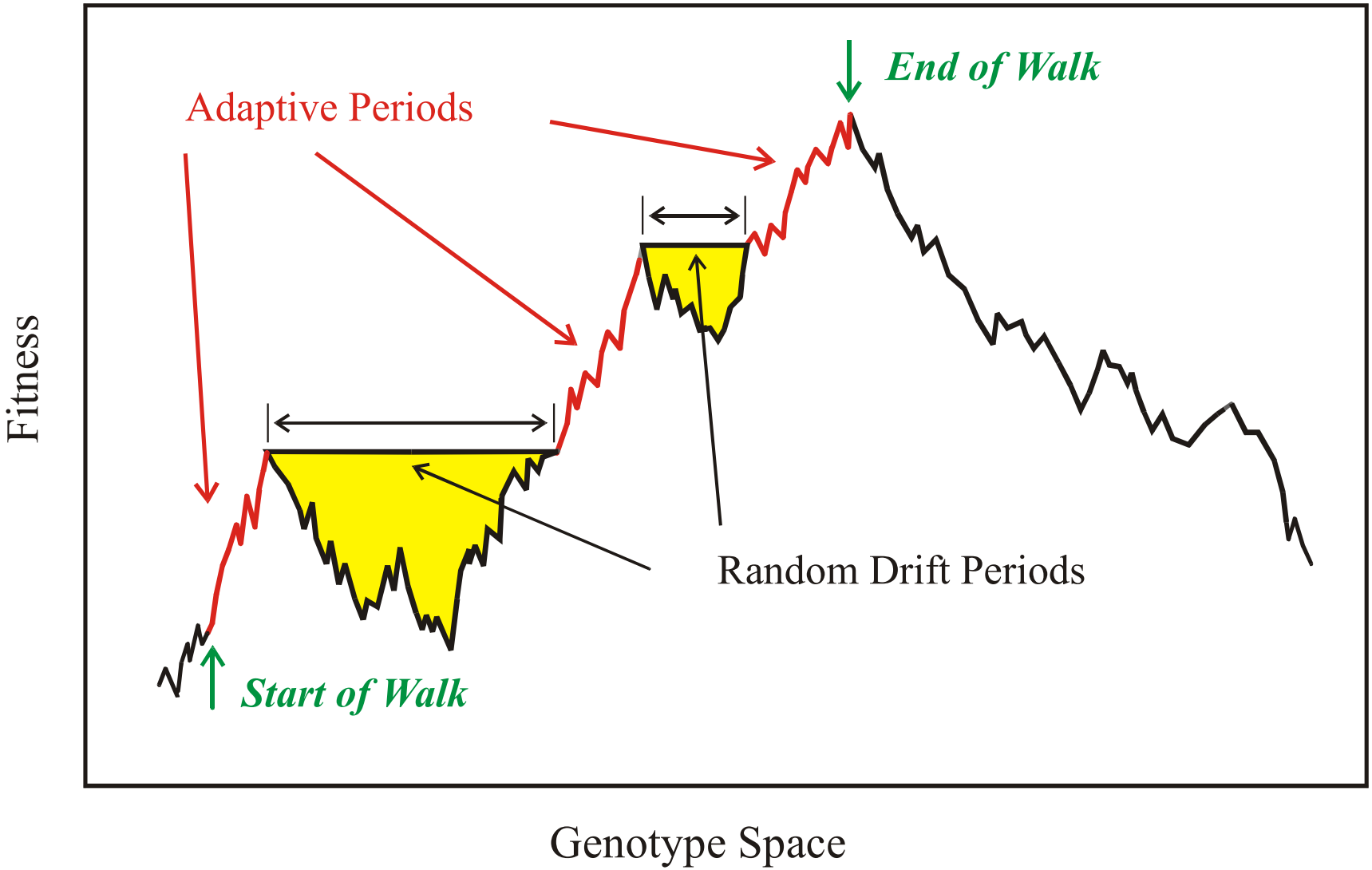
• The number of **minor transitions** decreases with increasing population size.

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

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

„...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 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