



**Evolution with RNA Molecules:  
From experiment to theory and back**

Peter Schuster

Institut für Theoretische Chemie und Molekulare  
Strukturbiologie der Universität Wien

Workshop Biotechnologie

Heidelberg, 30.09.2002

	Generation time	10 000 generations	10 <sup>6</sup> generations	10 <sup>7</sup> 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 multicellular organisms	10 d 20 a	274 a 20 000 a	27 380 a 2 × 10 <sup>7</sup> a	273 800 a 2 × 10 <sup>8</sup> a

Generation times and evolutionary timescales

## Evolution of RNA molecules based on Q $\beta$ phage

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

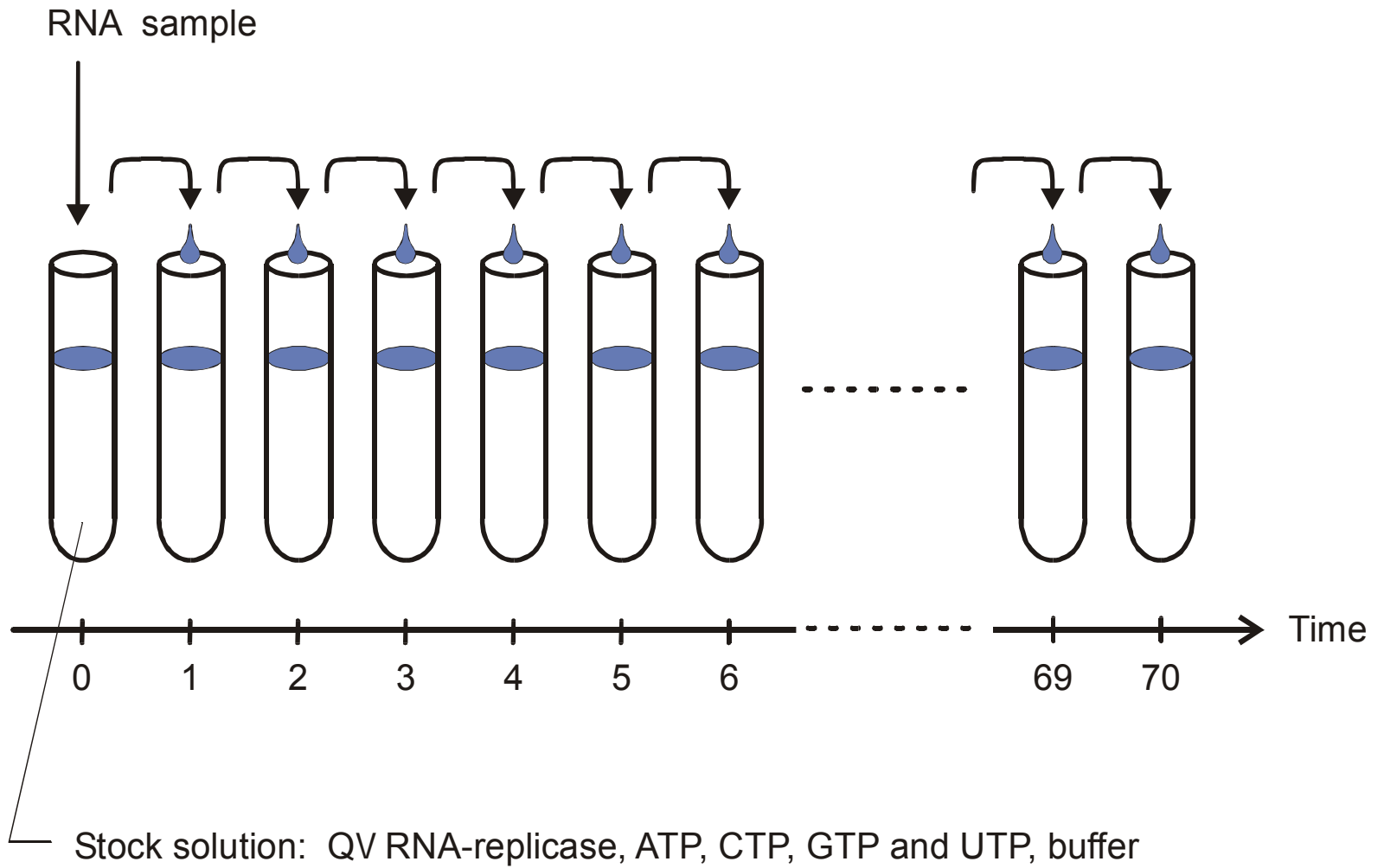
S.Spiegelman, *An approach to the experimental analysis of precellular evolution*. Quart.Rev.Biophys. **4** (1971), 213-253

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

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

C.K.Biebricher, W.C. Gardiner, *Molecular evolution of RNA in vitro*. Biophysical Chemistry **66** (1997), 179-192

G.Strunk, T. Ederhof, *Machines for automated evolution experiments in vitro based on the serial transfer concept*. Biophysical Chemistry **66** (1997), 193-202



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

Reproduction of the original figure of the serial transfer experiment with Q $\beta$  RNA

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

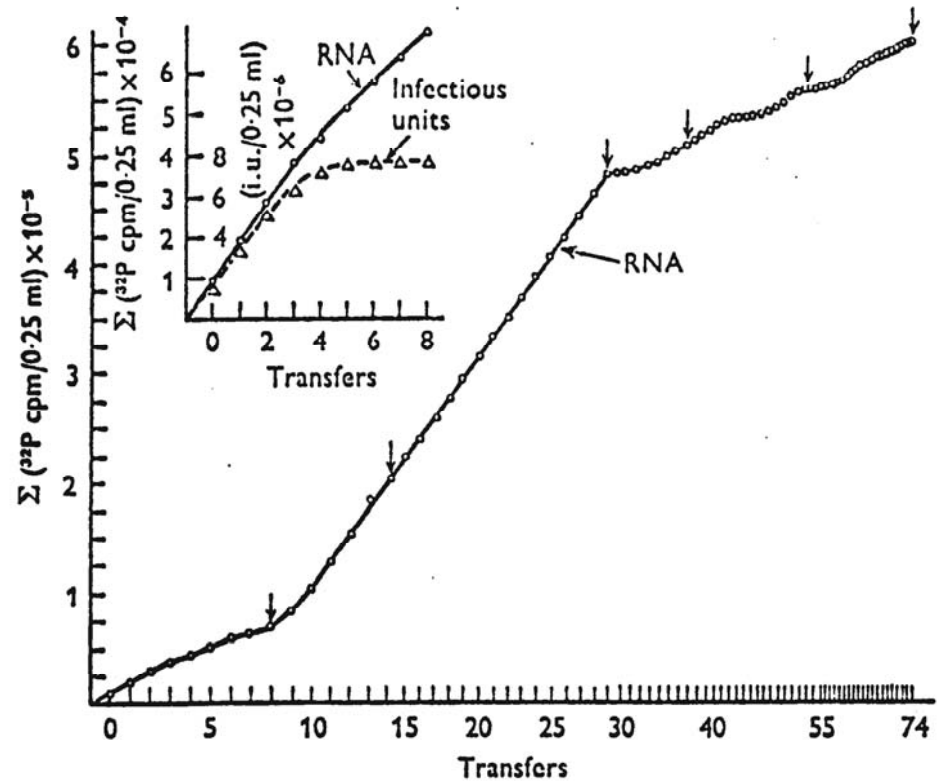
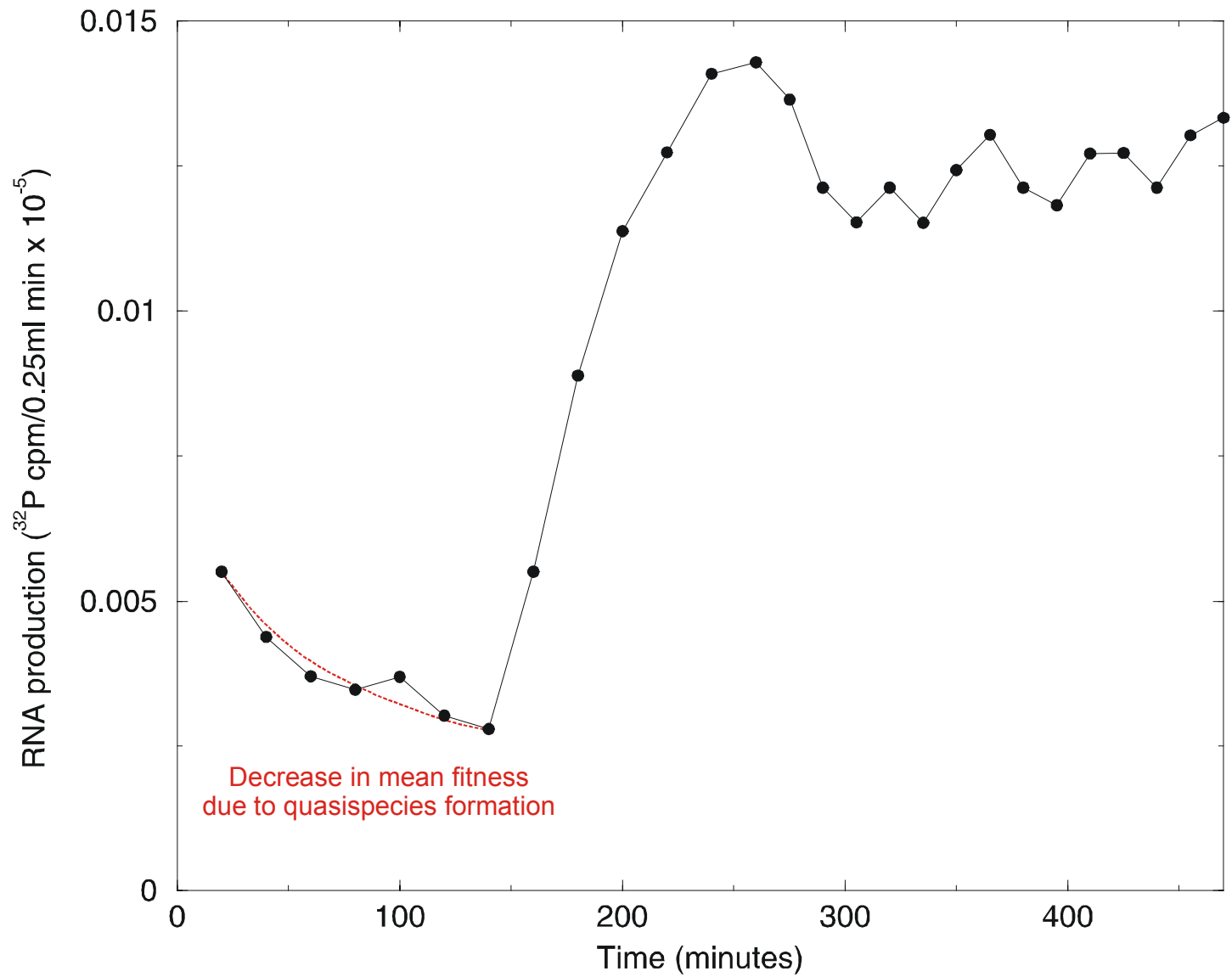


Fig. 9. Serial transfer experiment. Each 0.25 ml standard reaction mixture contained 40  $\mu\text{g}$  of Q $\beta$  replicase and  $^{32}\text{P}$ -UTP. The first reaction (0 transfer) was initiated by the addition of 0.2  $\mu\text{g}$  ts-1 (temperature-sensitive RNA) and incubated at 35  $^{\circ}\text{C}$  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).



The increase in RNA production rate during a serial transfer experiment

Ronald Fisher's conjecture of **optimization of mean fitness in populations** does not hold in general for **replication-mutation systems**: In general evolutionary dynamics the mean fitness of populations may also decrease monotonously or even go through a maximum or minimum. It does also not hold in general for **recombination of many alleles** and general multi-locus systems in population genetics.

**Optimization of fitness** is, nevertheless, fulfilled in most cases, and can be understood as a useful heuristic.



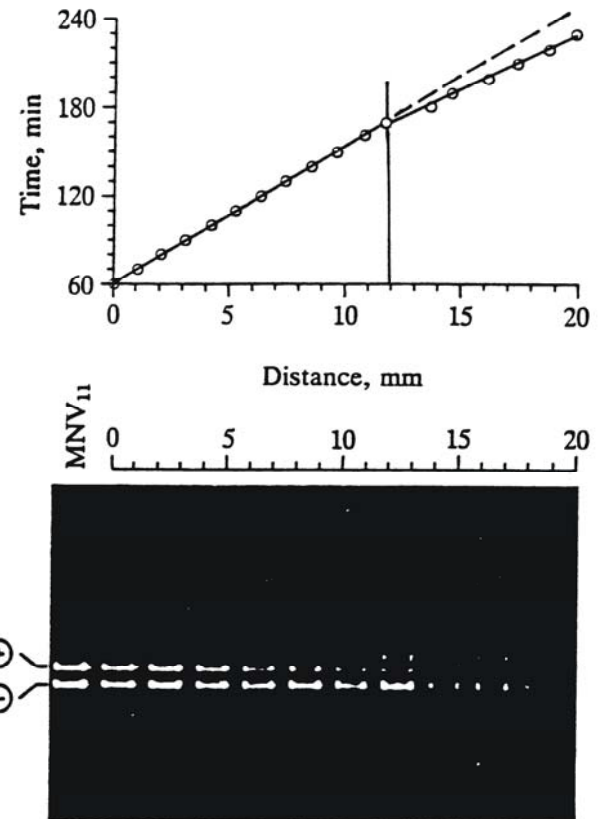
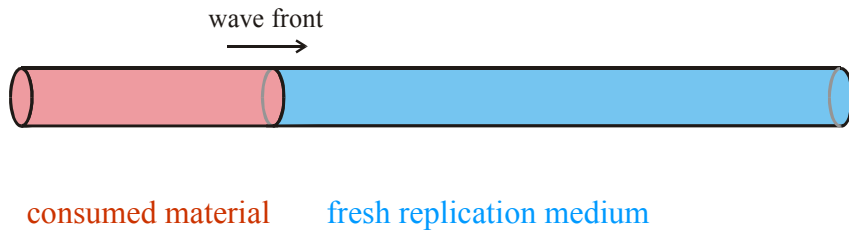


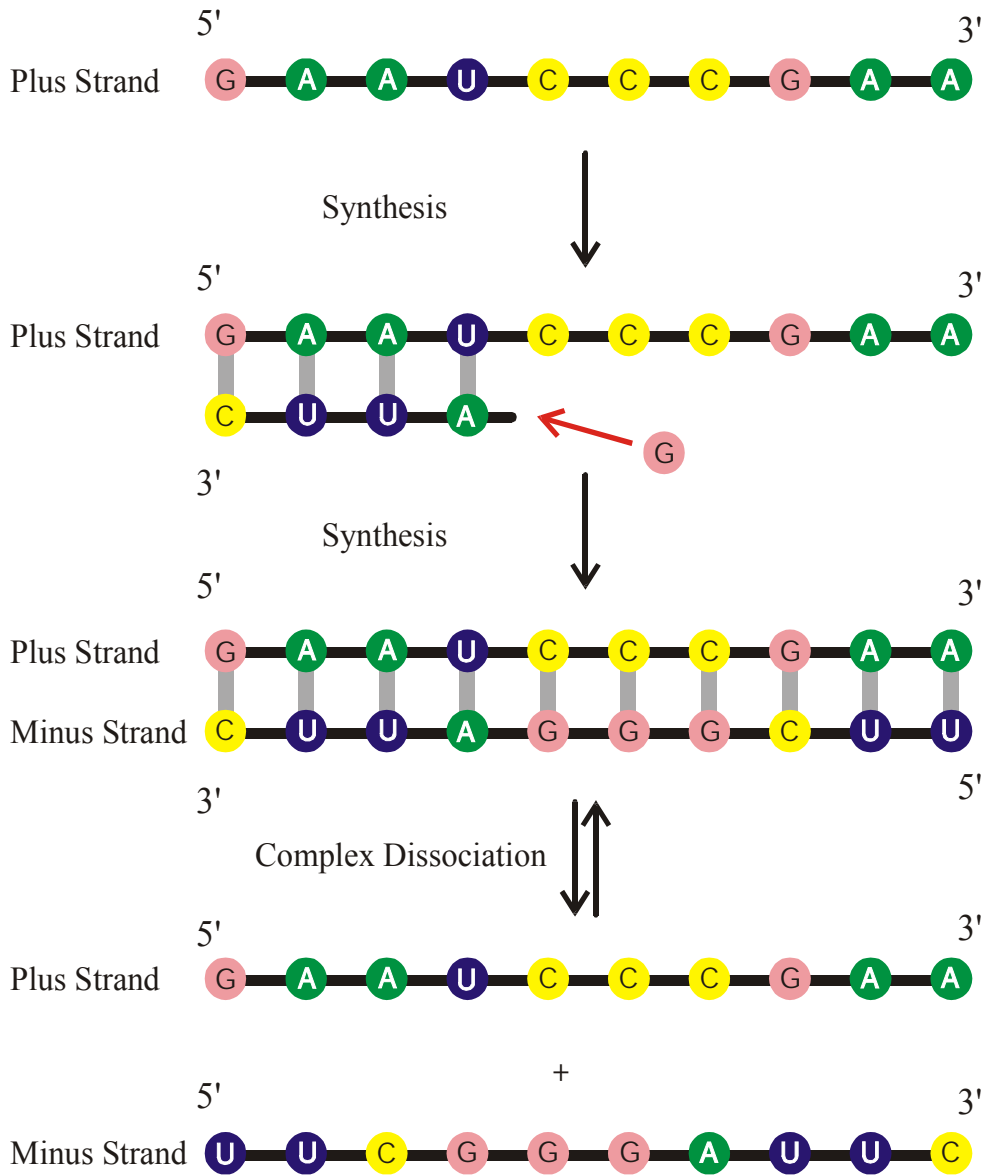
FIG. 3. Evolution of a new quasi-species along the capillary. (Upper) Front position measured using setup B. (Lower) Gel containing the fractions at 2.5-mm intervals. Regression lines are shown for the periods before and after 170 min. Aliquots (2  $\mu$ l) of the fractions were withdrawn after 240 min, mixed with 2  $\mu$ l of loading buffer, boiled for 3 min to melt the double strands, immediately chilled on dry ice, and loaded into the gel slots. The polyacrylamide gel contained 13% (wt/vol) acrylamide and 0.26% *N,N'*-methylene-bisacrylamide in running buffer (100 mM Tris borate, pH 8.3). Electrophoresis was for 6 hr at 5 V/cm at 4°C (16). Lane MNV<sub>11</sub> contains MNV<sub>11</sub> single strands (plus and minus strands) as reference. The concentration shift to new bands is centered at 12 mm where the velocity changes.

## Selection of QV-RNA through replication in a capillary

G.Bauer, H.Otten, J.S. McCaskill,  
*Proc.Natl.Acad.Sci.USA* **90**:4191, 1989

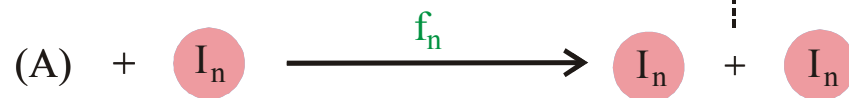
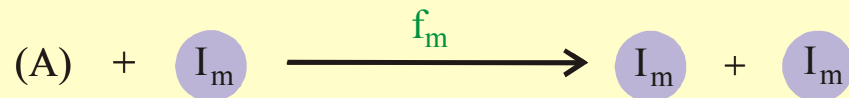
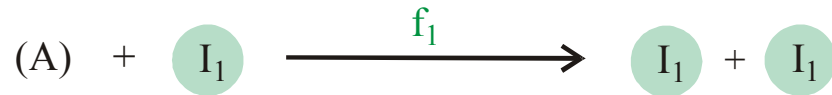
No new principle will declare  
itself from below a heap of  
facts.

Sir Peter Medawar, 1985



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





$$\frac{dx_i}{dt} = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

$$\Phi = \sum_j f_j x_j ; \quad \sum_j x_j = 1 ; \quad i, j = 1, 2, \dots, n$$

$$[I_i] = x_i \geq 0 ; \quad i = 1, 2, \dots, n ;$$

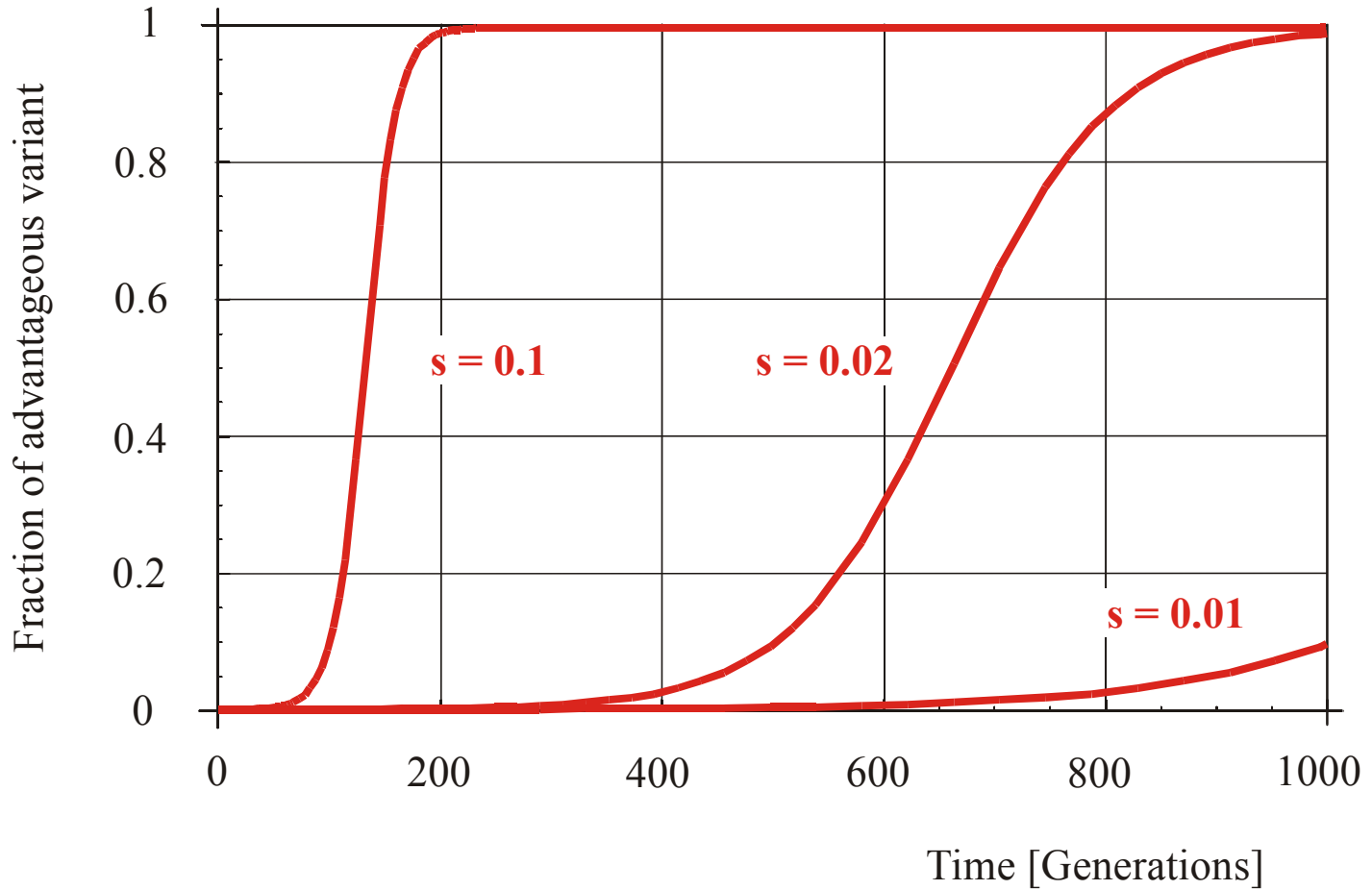
$$[A] = a = \text{constant}$$

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

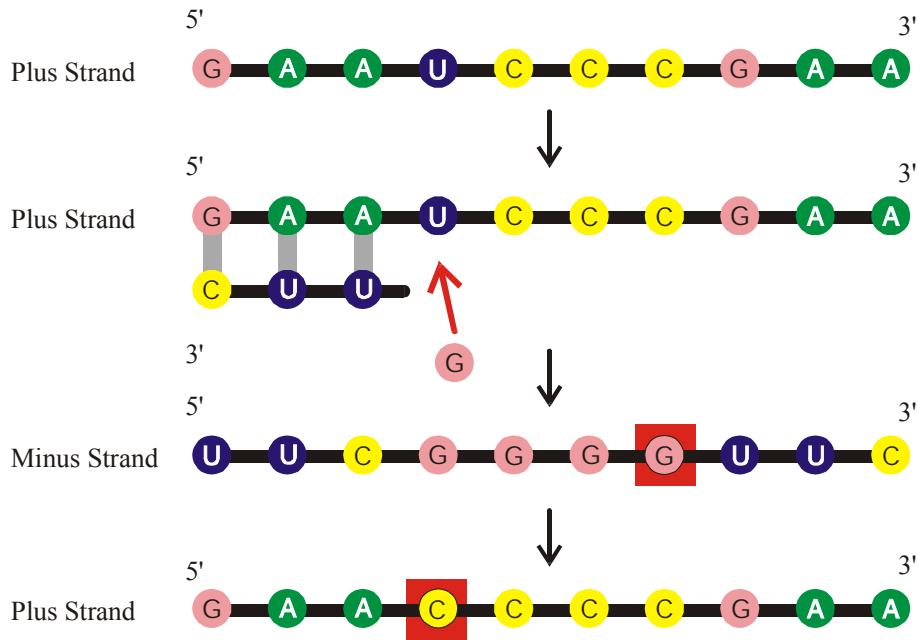
$$x_m(t) \rightarrow 1 \text{ for } t \rightarrow \infty$$

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

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



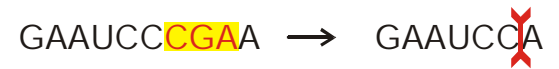
Selection of advantageous mutants in populations of  $N = 10\,000$  individuals



**Point Mutation**



**Insertion**



**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

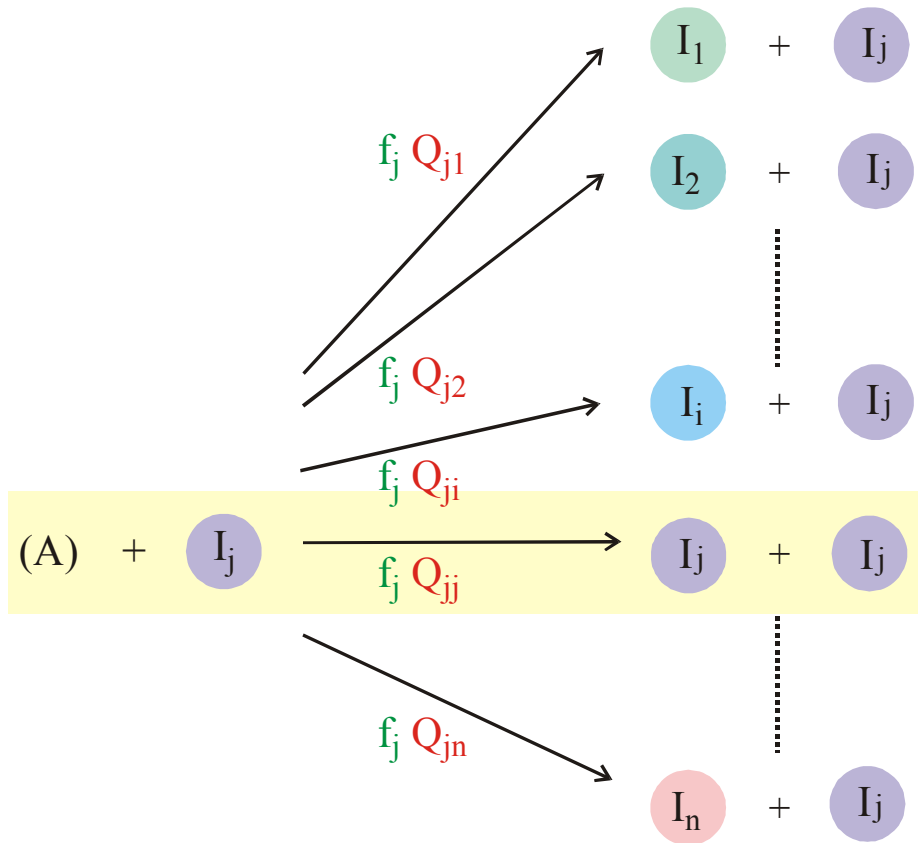
J. Swetina, P. Schuster, *Self-replication with errors - A model for polynucleotide replication*.

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



$$\frac{dx_i}{dt} = \sum_j f_j Q_{ji} x_j - x_i \Phi$$

$$\Phi = \sum_j f_j x_j ; \quad \sum_j x_j = 1 ; \quad \sum_i Q_{ij} = 1$$

$$[I_i] = x_i \ll 1 ; \quad i = 1, 2, \dots, n ;$$

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

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

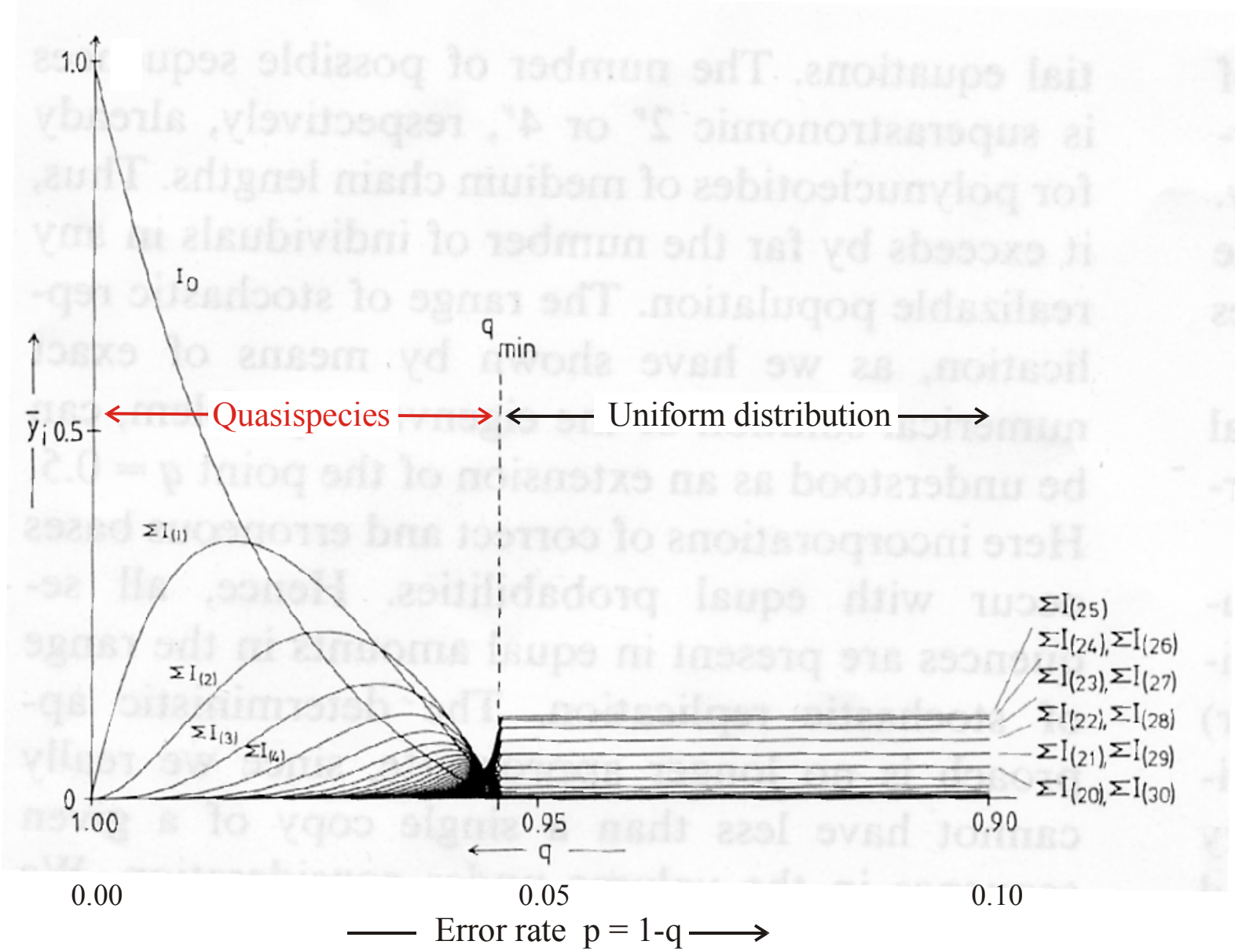
$p$  ..... Error rate per digit

$l$  ..... Chain length of the polynucleotide

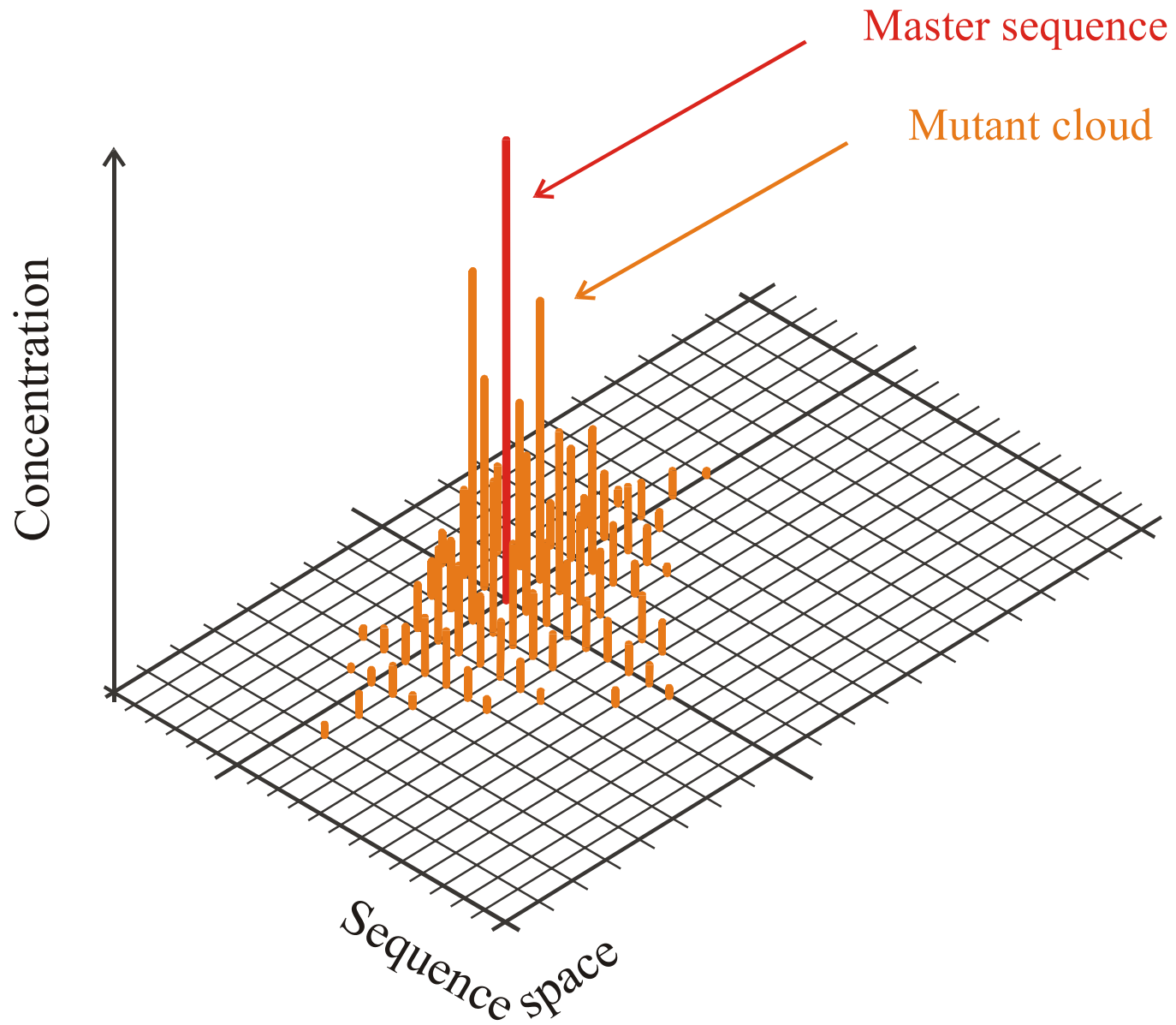
$d(i,j)$  .... Hamming distance between  $I_i$  and  $I_j$

Chemical kinetics of replication and mutation as parallel reactions





**Quasispecies** as a function of the replication accuracy  $q$



The molecular quasispecies in sequence space

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 where the mean fitness decreases monotonously or where 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 increasing 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.

## 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

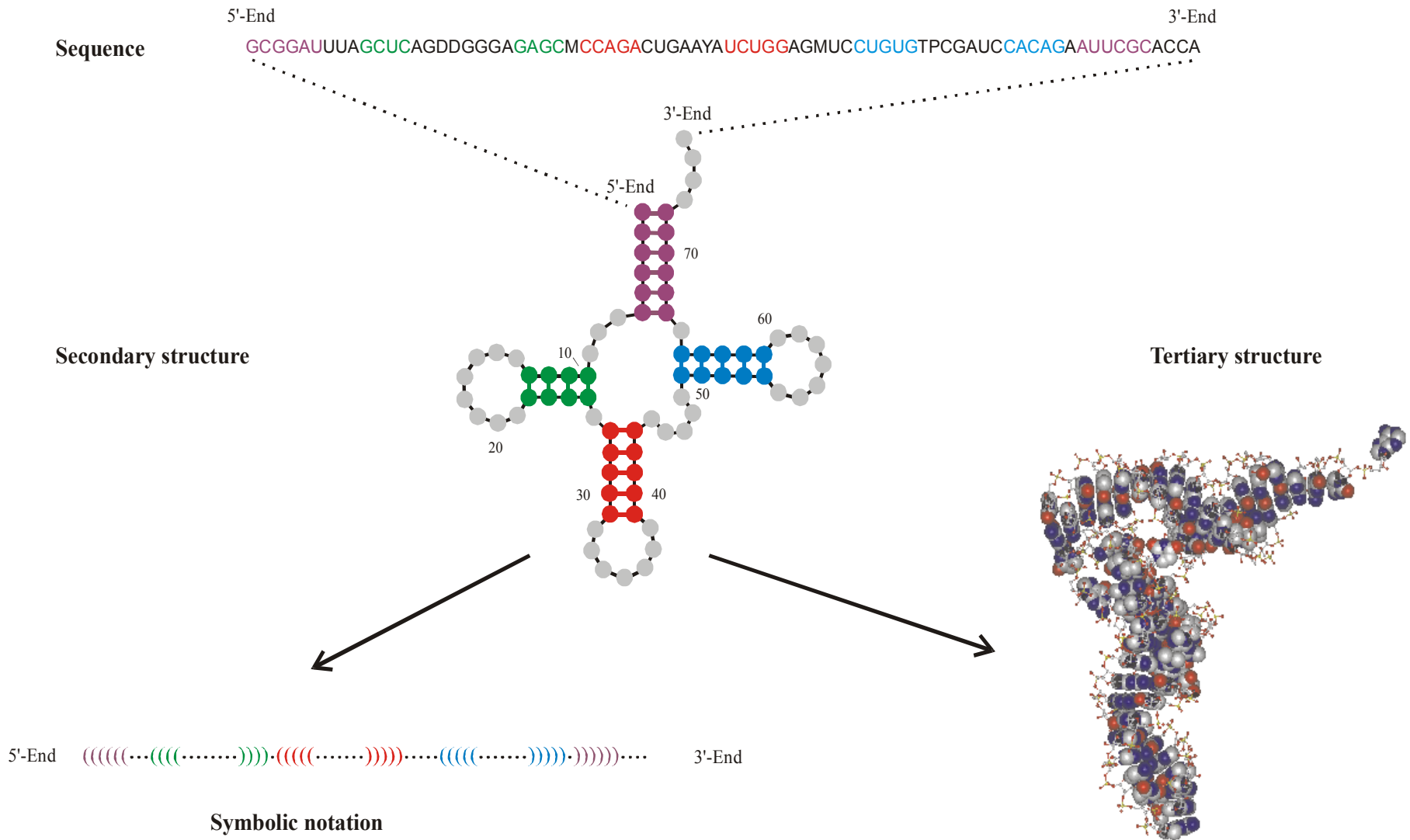
C.M.Reidys, P.F.Stadler, P.Schuster, *Generic properties of combinatory maps*. Bull.Math.Biol. **59** (1997), 339-397

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 coarse-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.

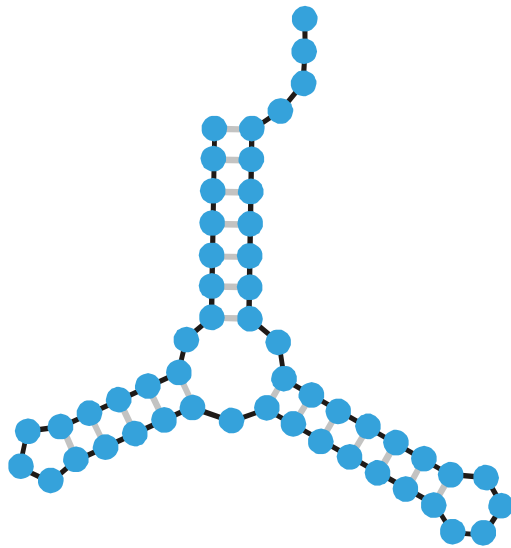
# RNA Minimum Free Energy Structures

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)

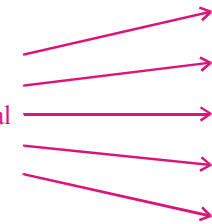
**Vienna RNA Package:** <http://www.tbi.univie.ac.at> (includes inverse folding, suboptimal structures, kinetic folding, etc.)

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



Minimum free energy  
criterion

1st  
2nd  
3rd trial  
4th  
5th



UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUUAUCUCUGUCGGGCUAGGGCGC  
 GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUUAUCUGG  
 UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG  
 CAUJGGUGCUAAUGAUUUAGGGCUGUAUUCUGUAUAGCGAUCAGUGUCCG  
 GUAGGCCCUUUGACAUAAGAUUUUUCCAUGGUGGGAGAUGGCCAUUGCAG

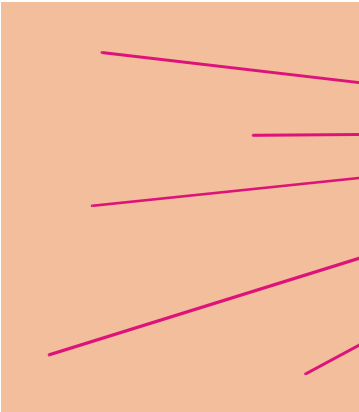
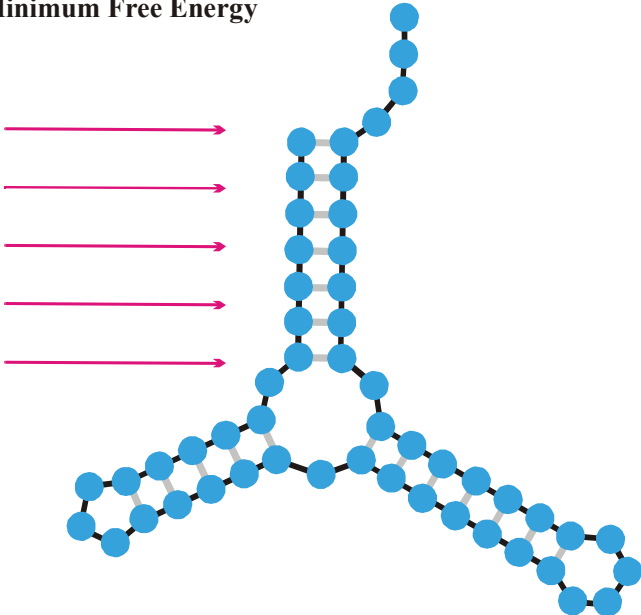
Inverse folding

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

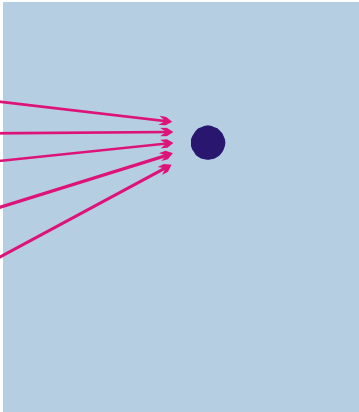


**Criterion of  
Minimum Free Energy**

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC  
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUUUCUGG  
UUAGCGAGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG  
CAUUGGUGCUAAUGAUUUAGGGCUGUAUJCCUGUAUAGCGAUCAGUGUCCG  
GUAGGCCCUUUGACAUAAGAUUUUUCCAUGGGUGGAGAUGGCCAUUGCAG



Sequence Space



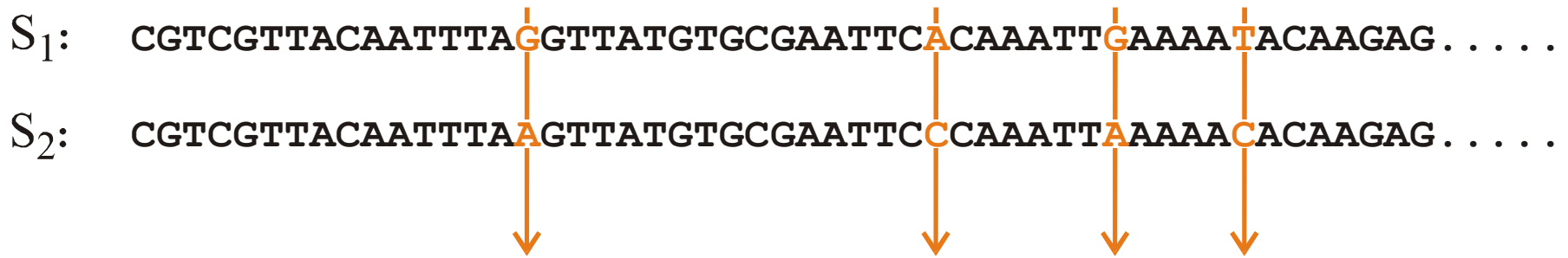
Shape Space

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.

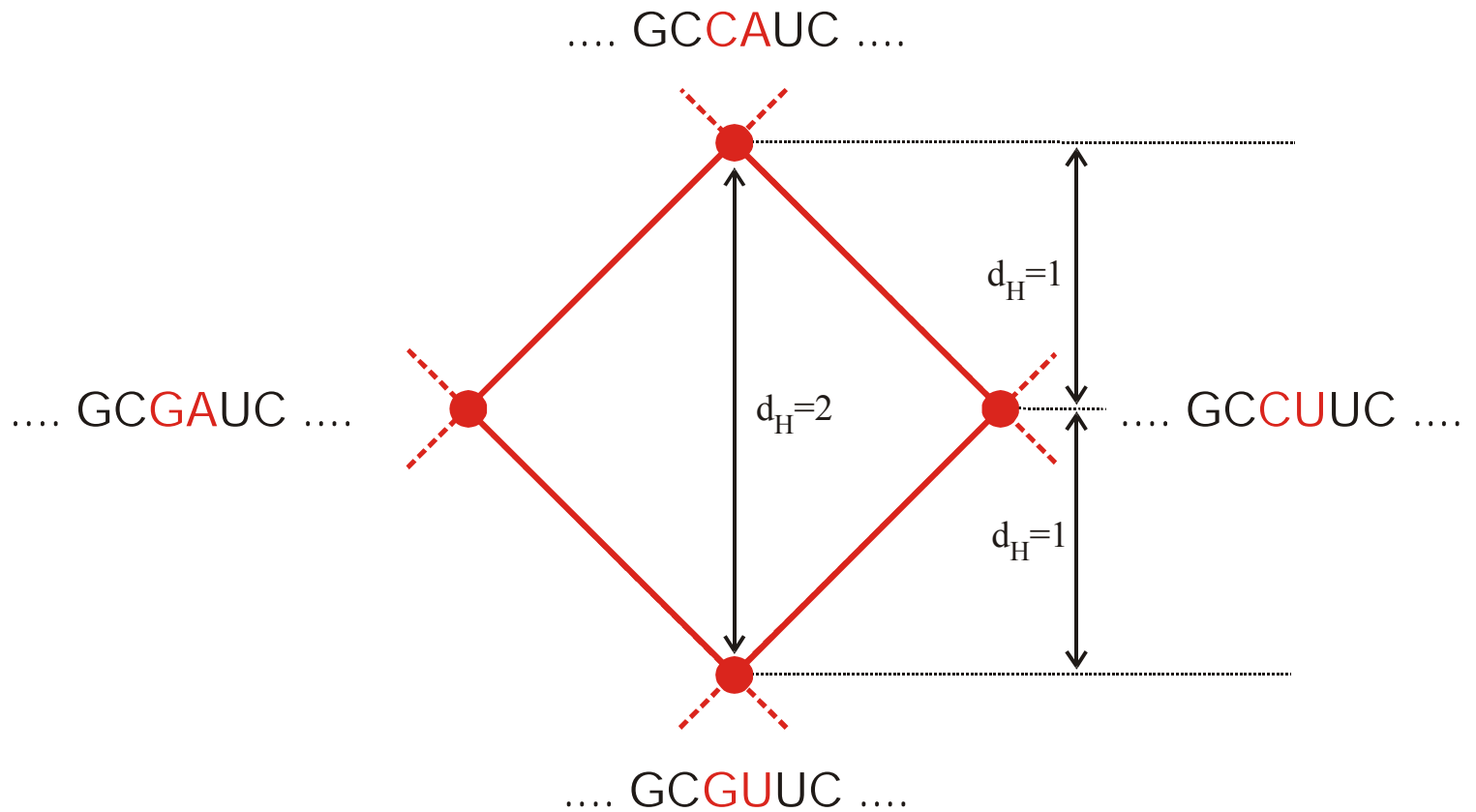
$S_1$ : CGTCGTTACAATTTA **G**GTTATGTGCGAATTC **A**CAAATT **G**AAAA **T**ACAAGAG . . . . .  
 $S_2$ : CGTCGTTACAATTTA **A**GTTATGTGCGAATTC **C**CAAATT **A**AAAA **C**ACAAGAG . . . . .



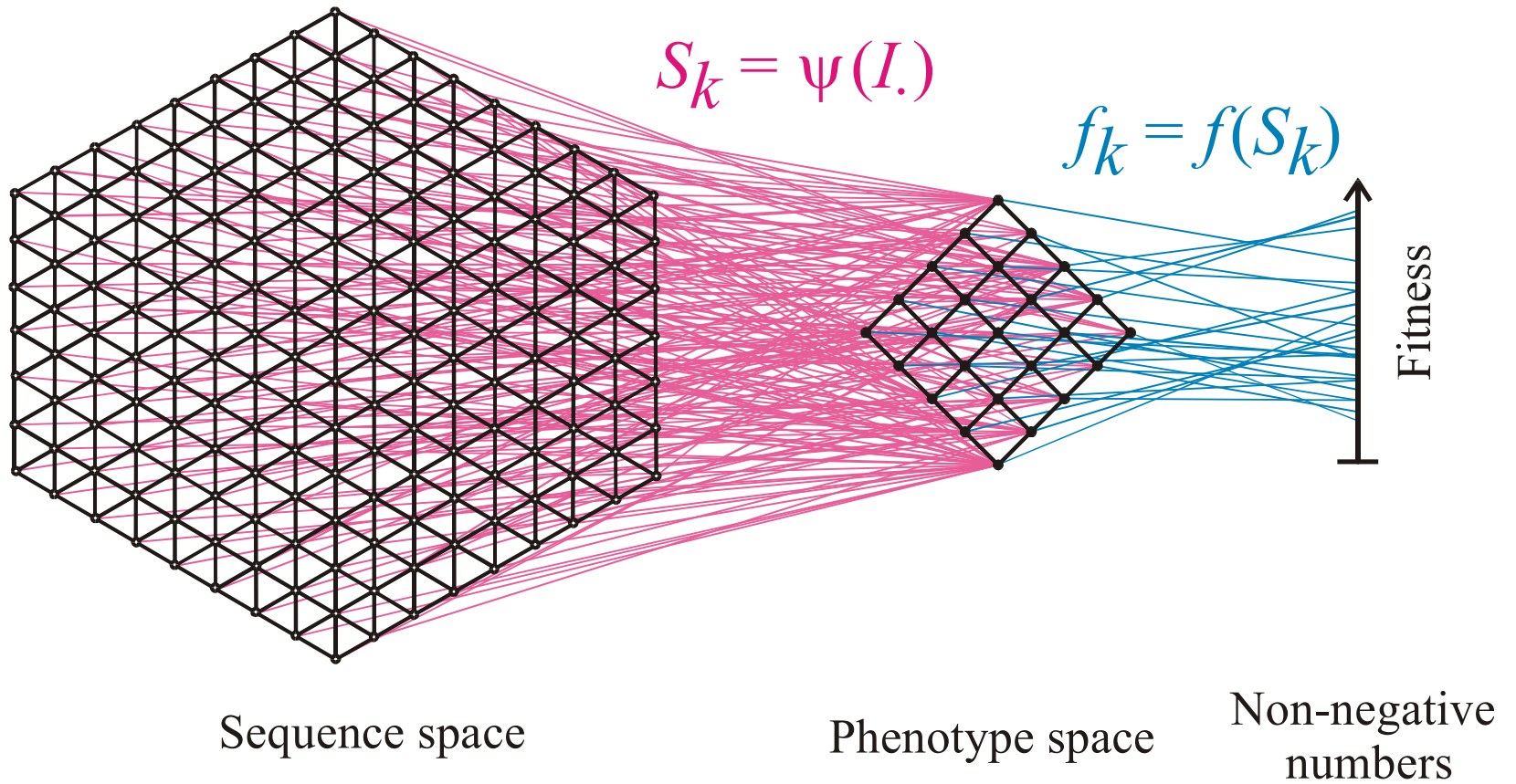
Hamming distance  $d_H(S_1, S_2) = 4$

- (i)  $d_H(S_1, S_1) = 0$
- (ii)  $d_H(S_1, S_2) = d_H(S_2, S_1)$
- (iii)  $d_H(S_1, S_3) < d_H(S_1, S_2) + d_H(S_2, S_3)$

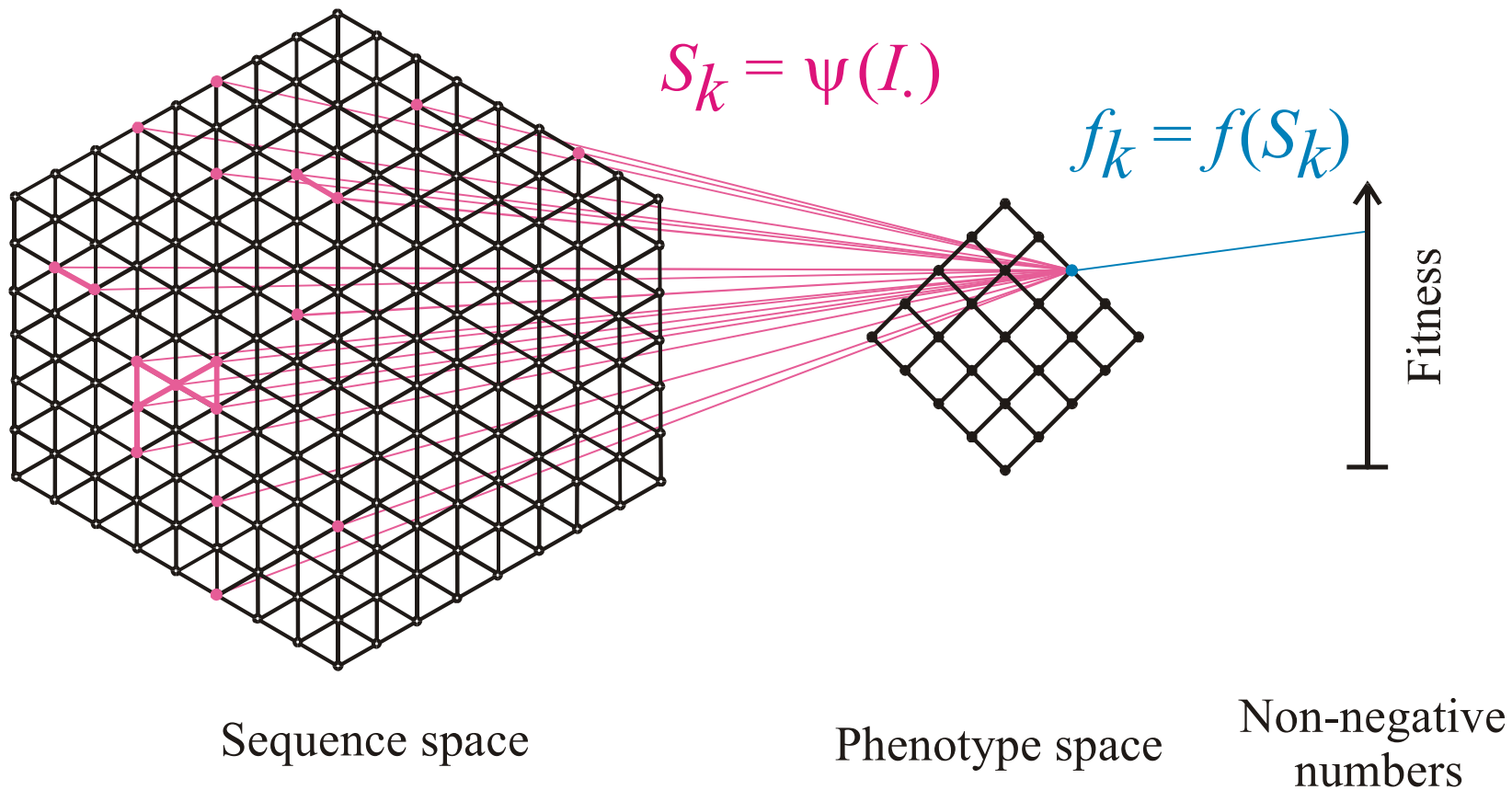
The Hamming distance induces a metric in sequence space

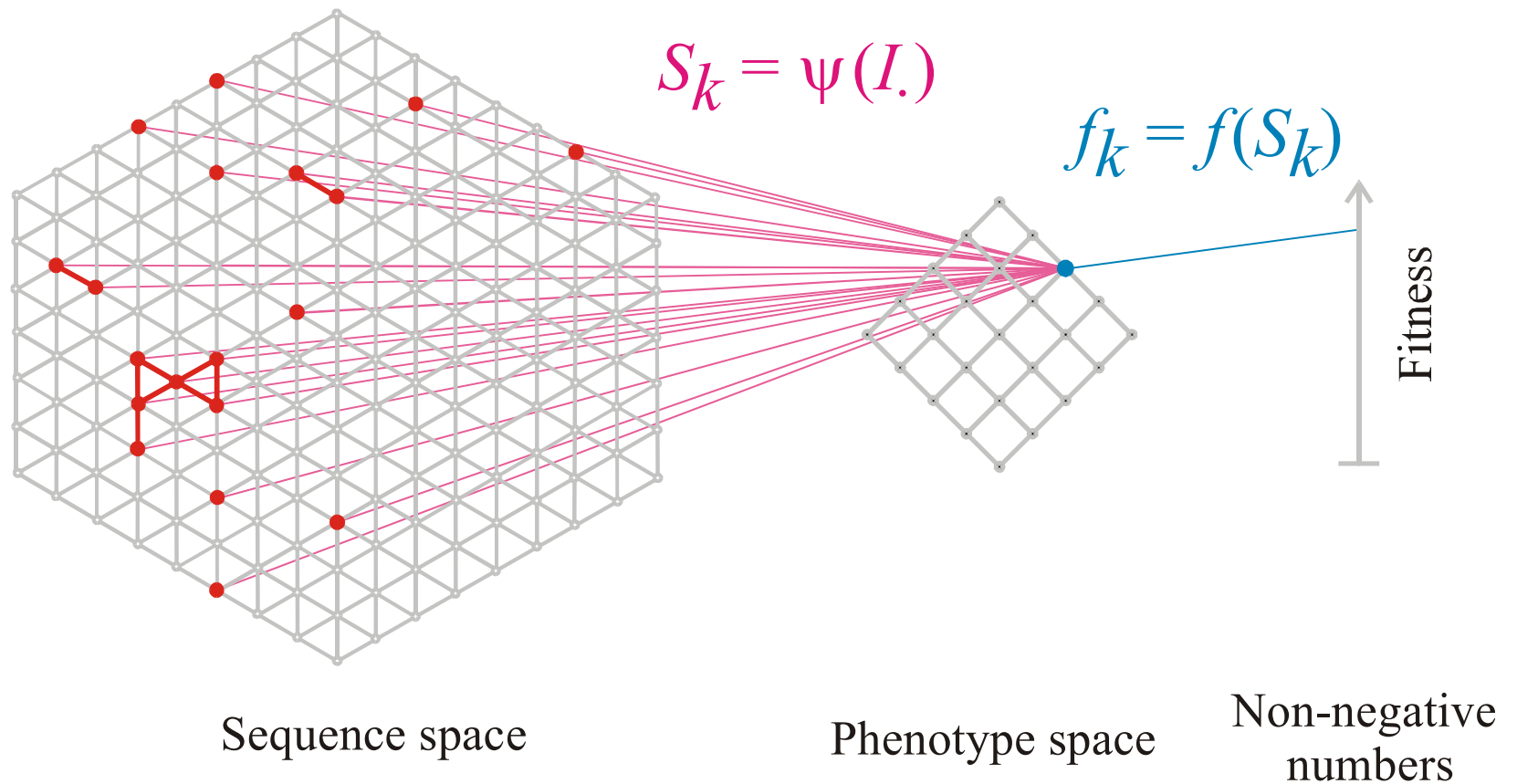


Single point mutations as moves in sequence space



Mapping from sequence space into phenotype space and into fitness values





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

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

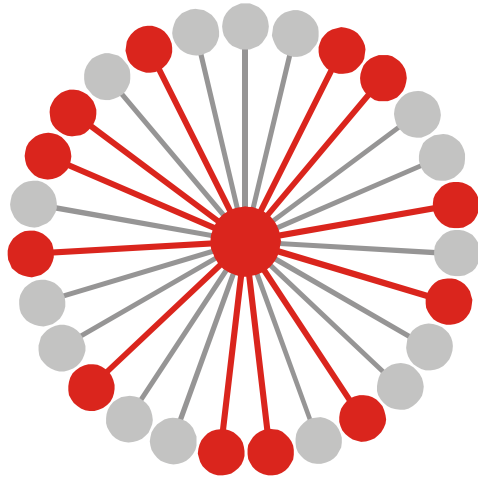
$$G_k = m^{-1}(S_k) \circledast \{m_j \mid m(I_j) = S_k\}$$

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,  $N=4^n$ , becomes very large with increasing length, and is prohibitive for numerical computations.

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





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

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

Connectivity threshold:

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

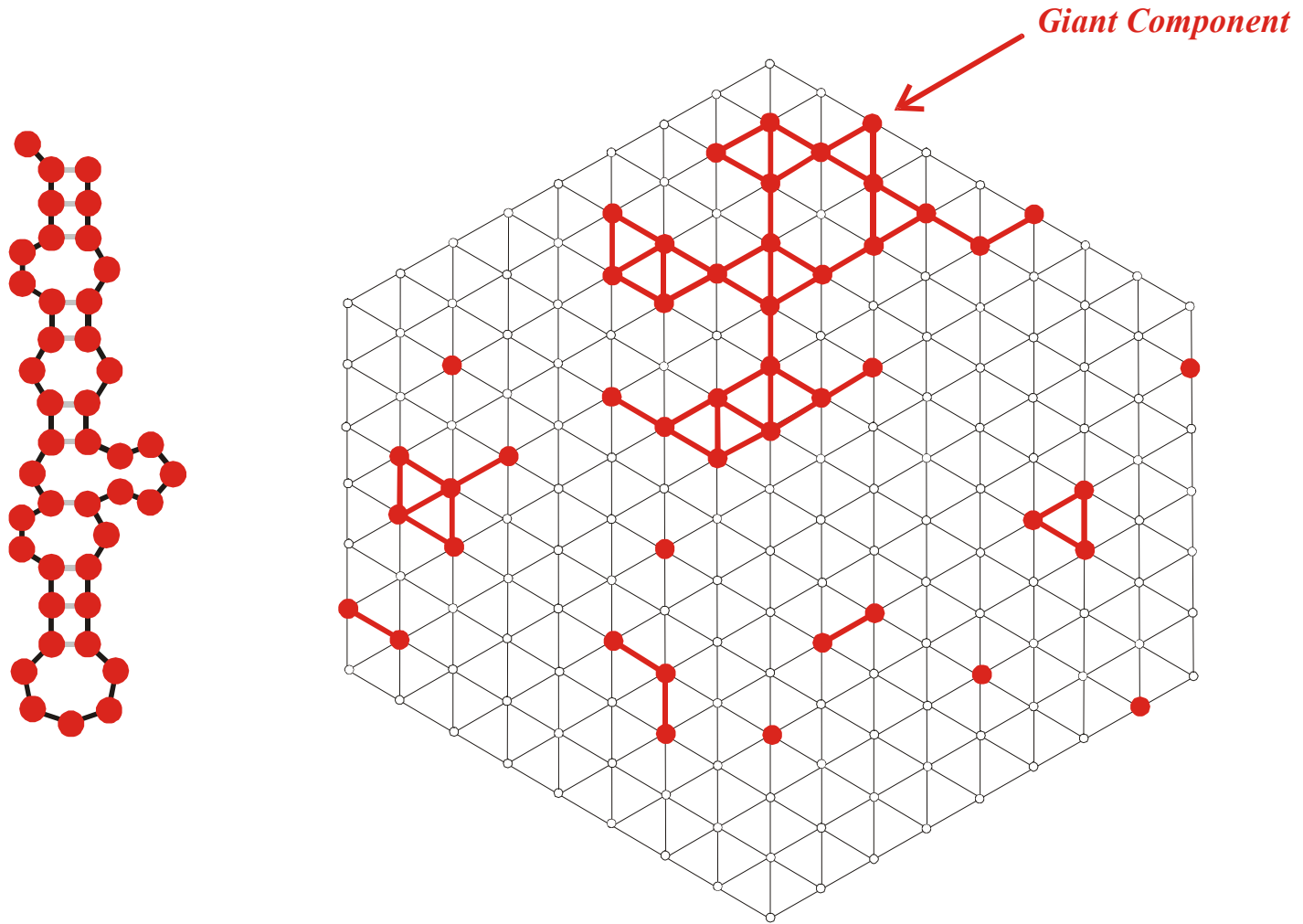
Alphabet size  $\kappa$ : **AUGC**  $\kappa = 4$

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

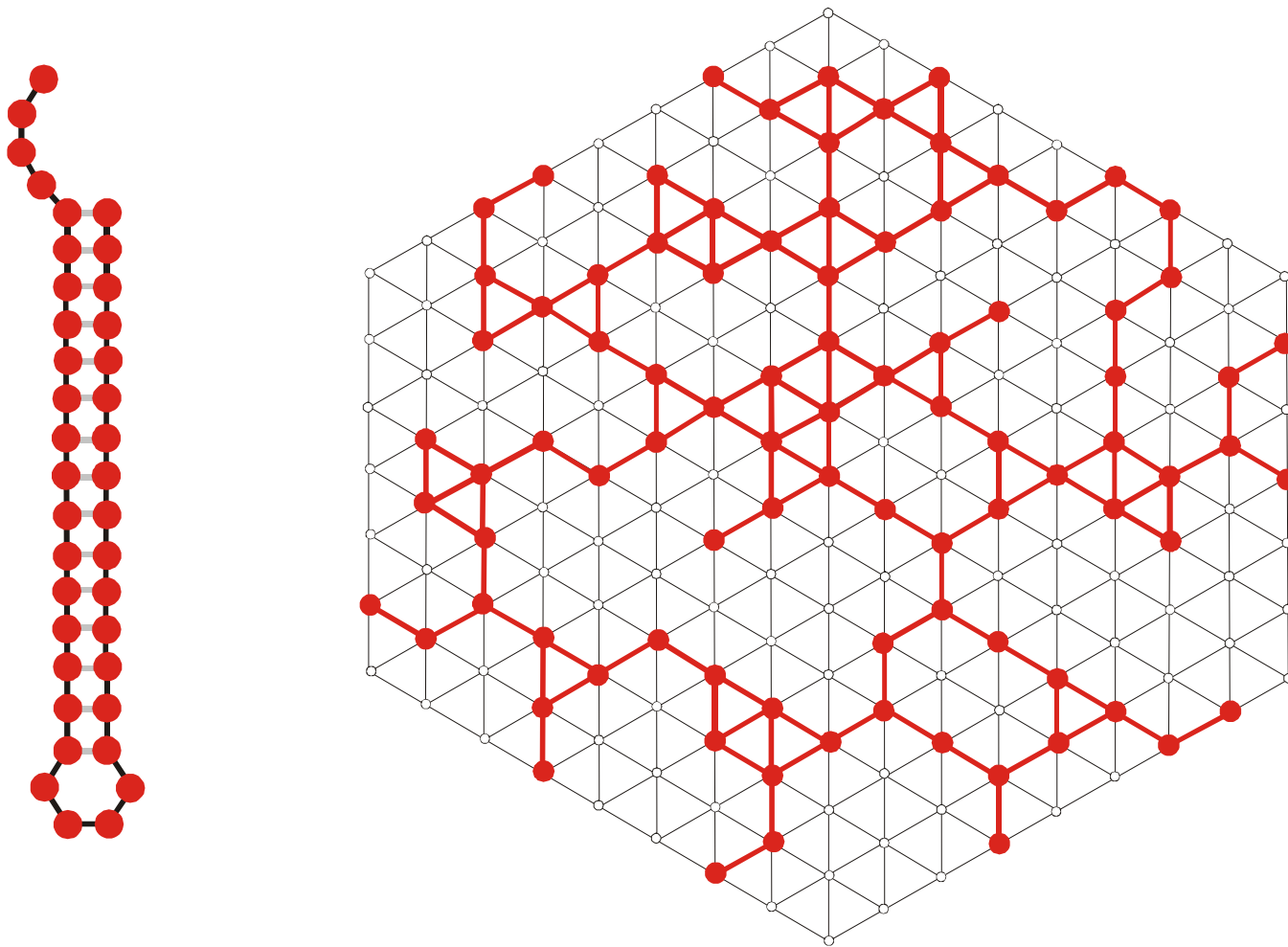
$\bar{\lambda}_k < \lambda_{cr}$  .... network  $G_k$  is **not** connected

$\kappa$	$\lambda_{cr}$
2	0.5
3	0.4226
4	0.3700

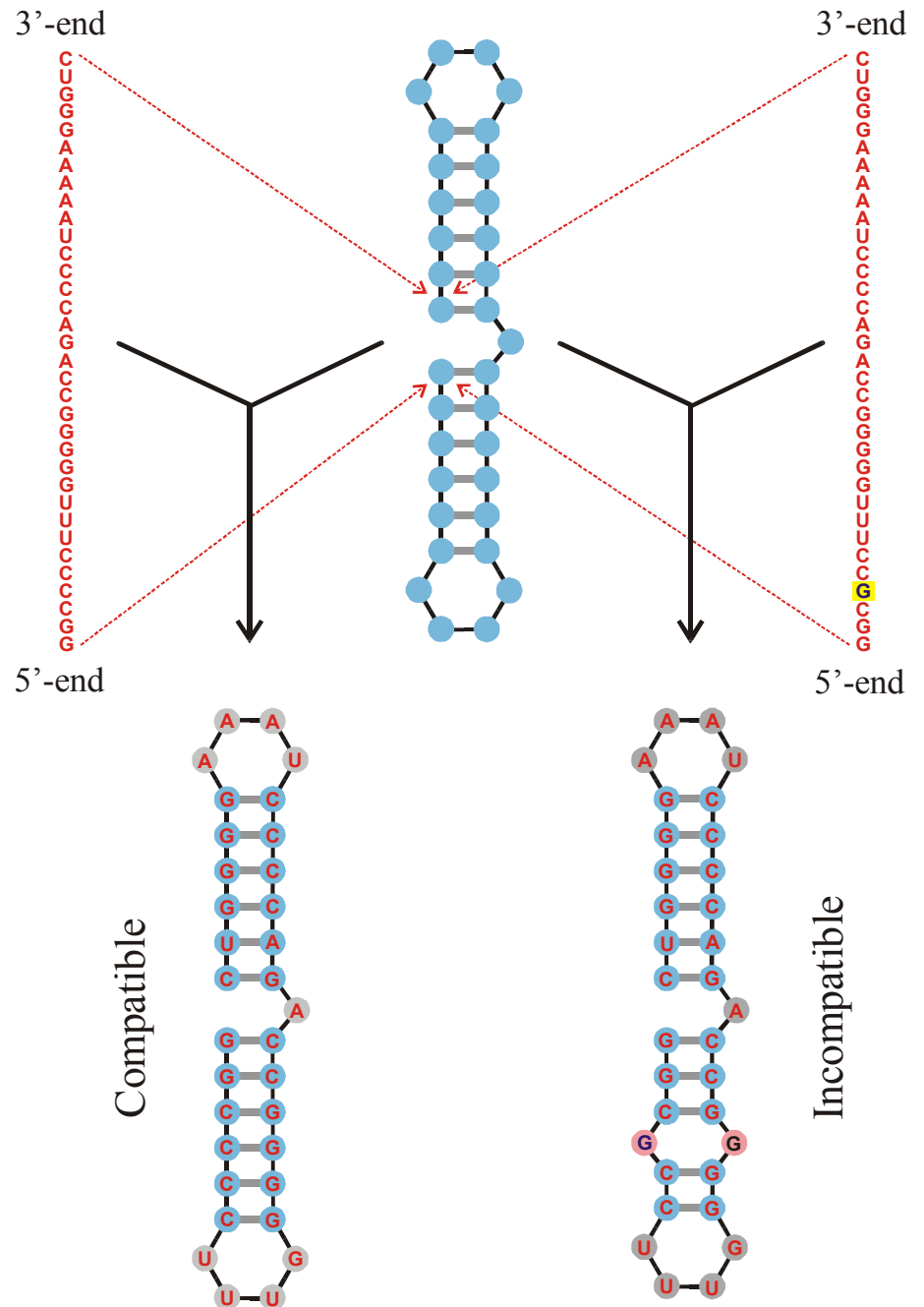
Mean degree of neutrality and connectivity of **neutral networks**



A multi-component neutral network

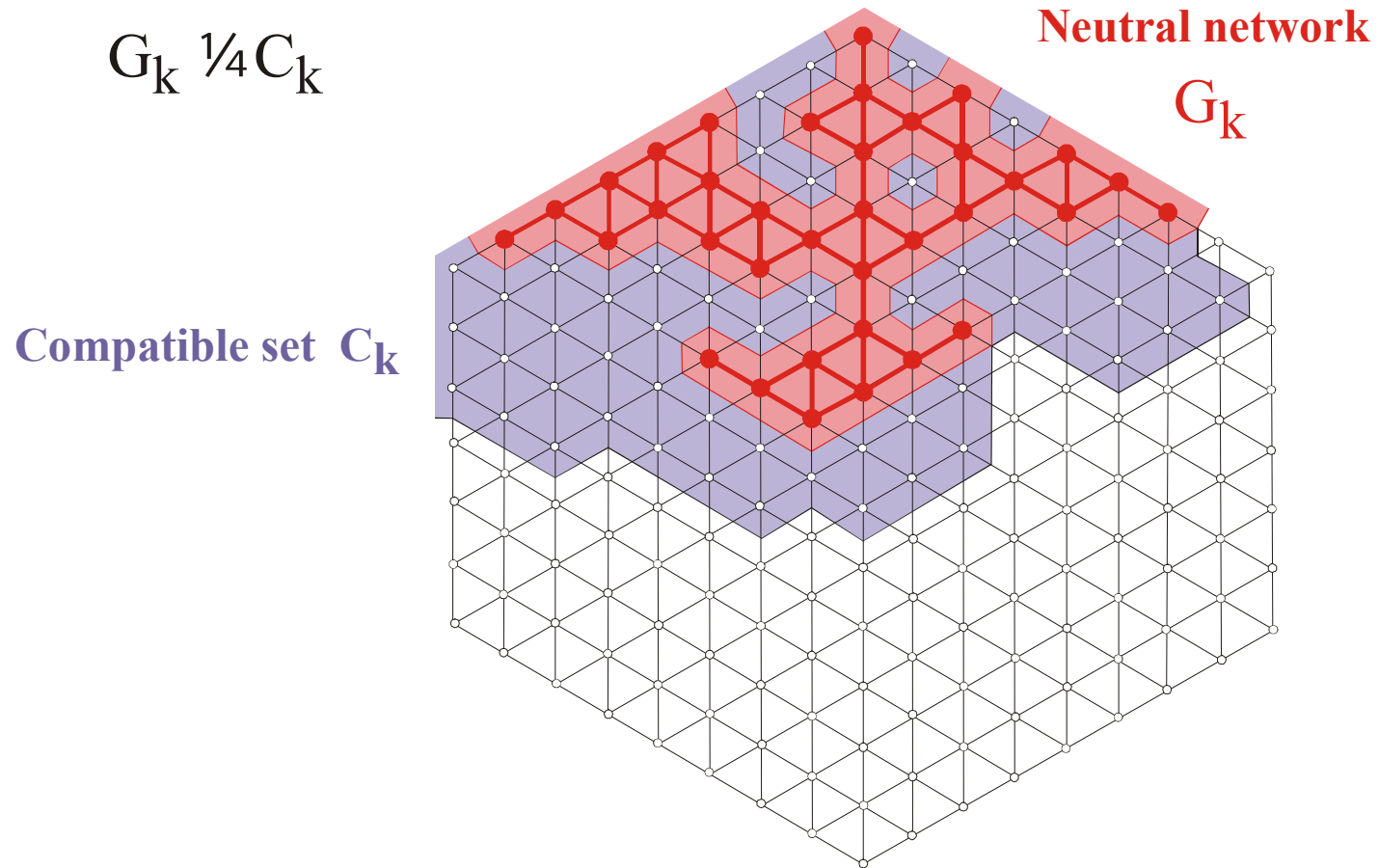


A connected neutral network

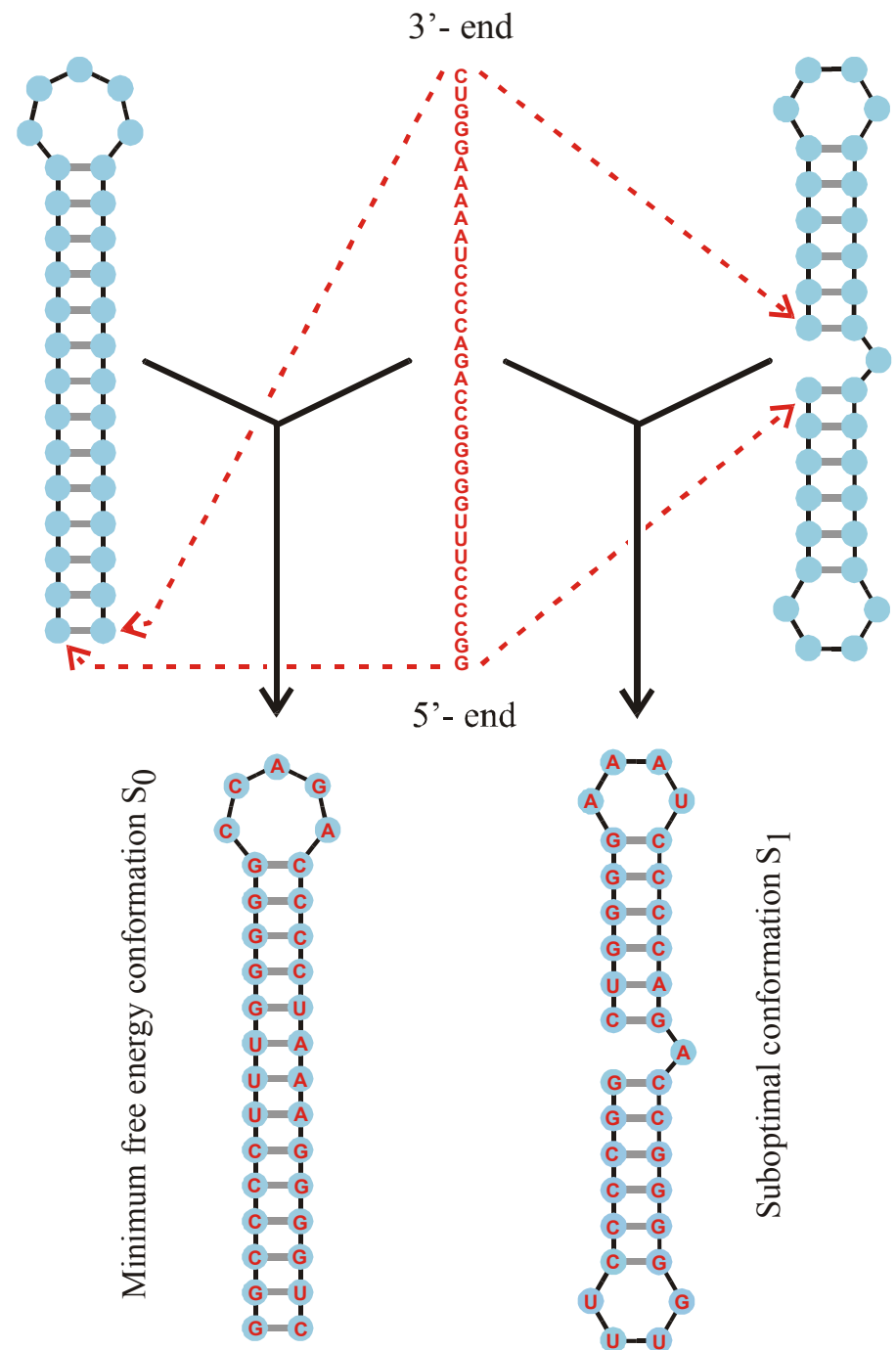


## Compatibility of sequences with structures

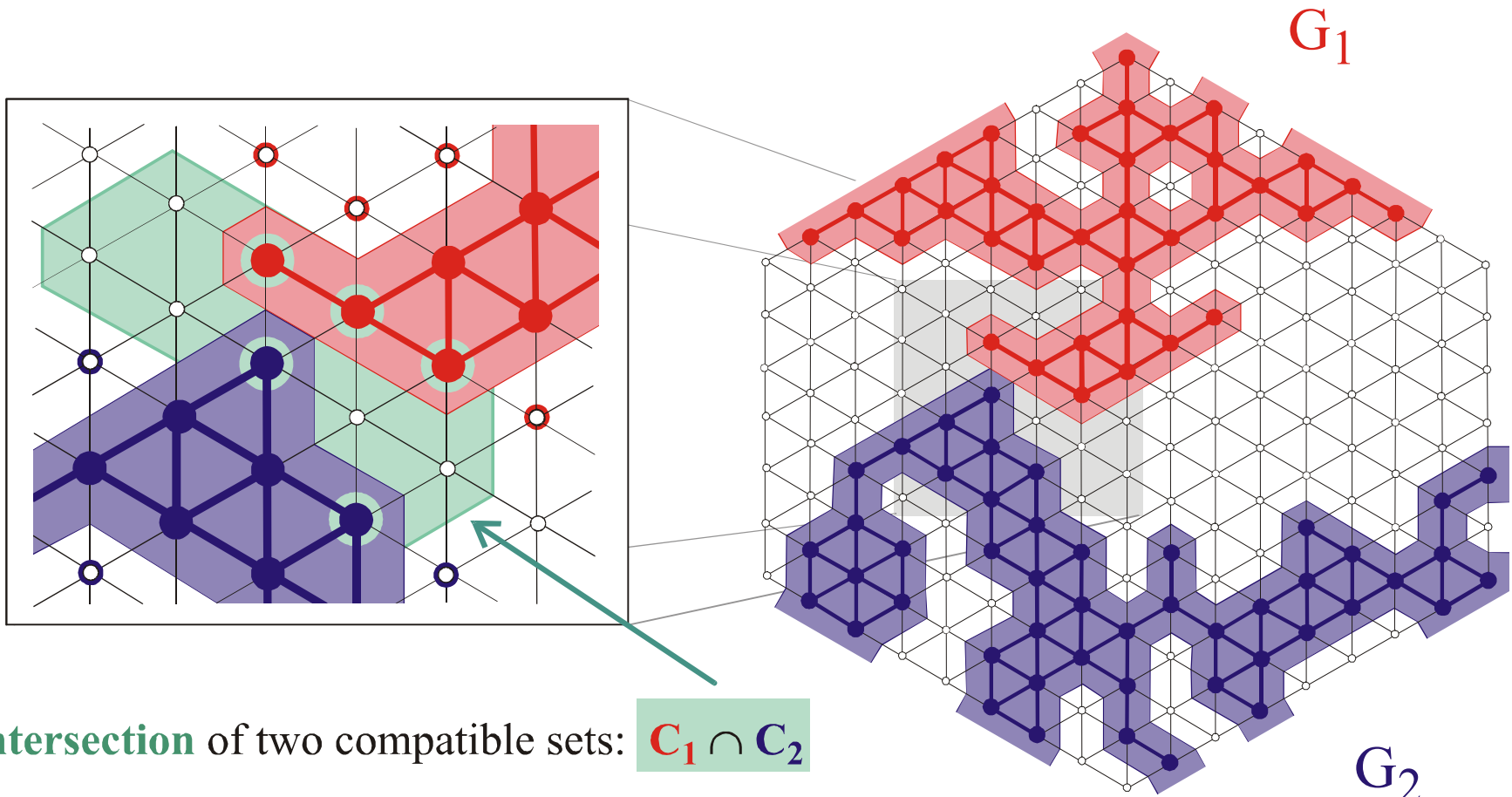
A sequence is compatible with its minimum free energy structure and all its suboptimal structures.



The **compatible set**  $C_k$  of a structure  $S_k$  consists of all sequences which form  $S_k$  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



The intersection of two compatible sets is always non empty:  $C_1 \cap C_2 \neq \emptyset$

## **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

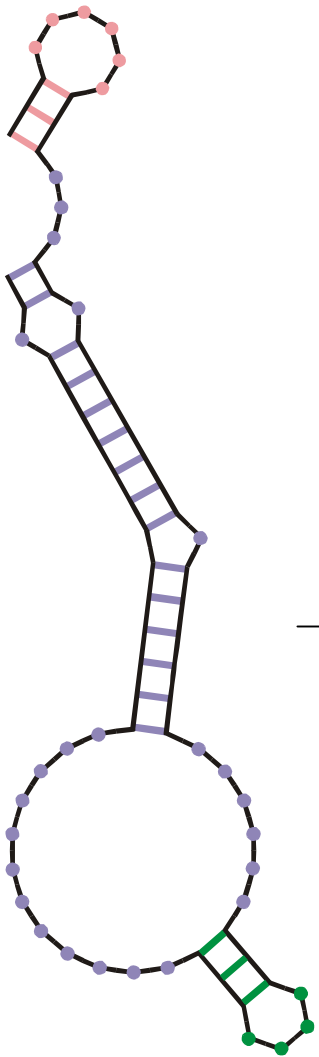
M.A.Huynen, W.Fontana, P.F.Stadler, *Smoothness within ruggedness. The role of neutrality in adaptation*. Proc.Natl.Acad.Sci.USA **93** (1996), 397-401

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

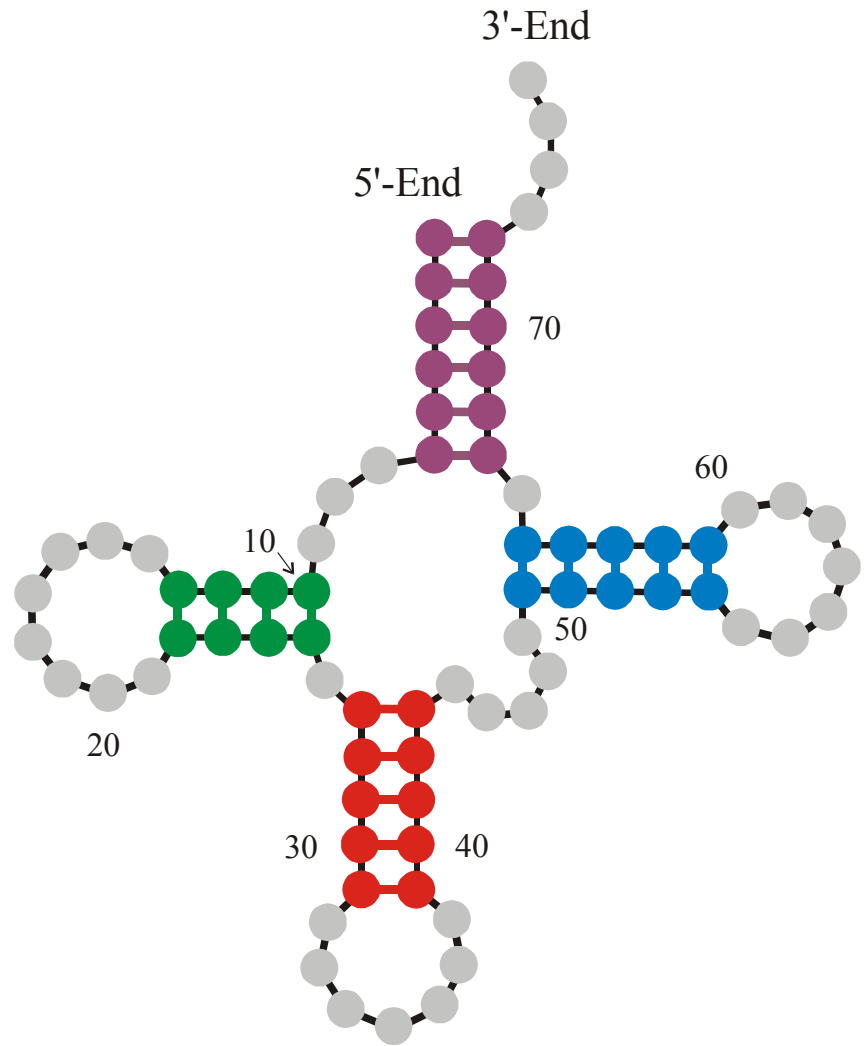
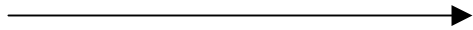
W.Fontana, P.Schuster, *Shaping space. The possible and the attainable in RNA genotype-phenotype mapping*. J.Theor.Biol. **194** (1998), 491-515

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

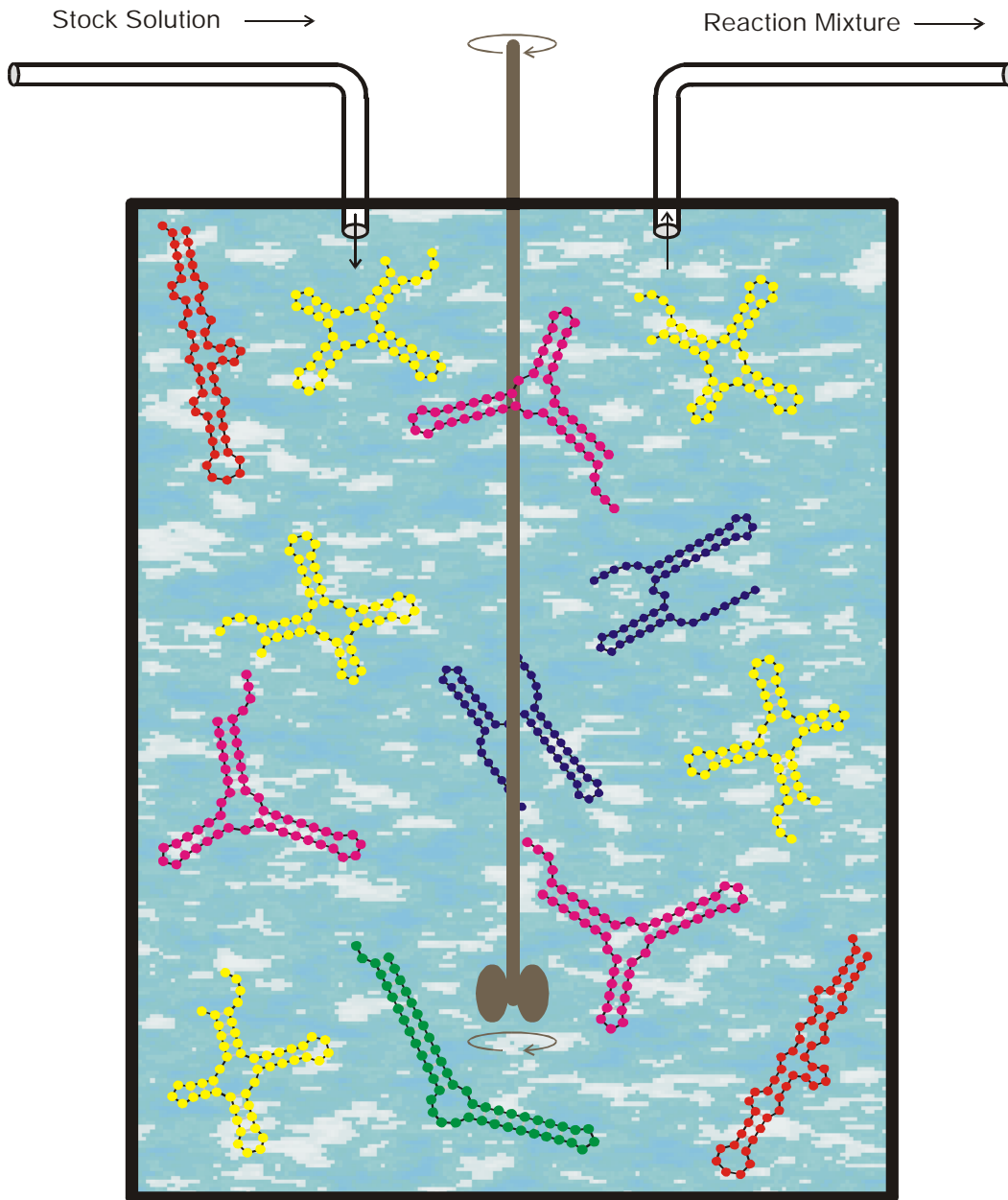




Randomly chosen  
initial structure



Phenylalanyl-tRNA as  
target structure

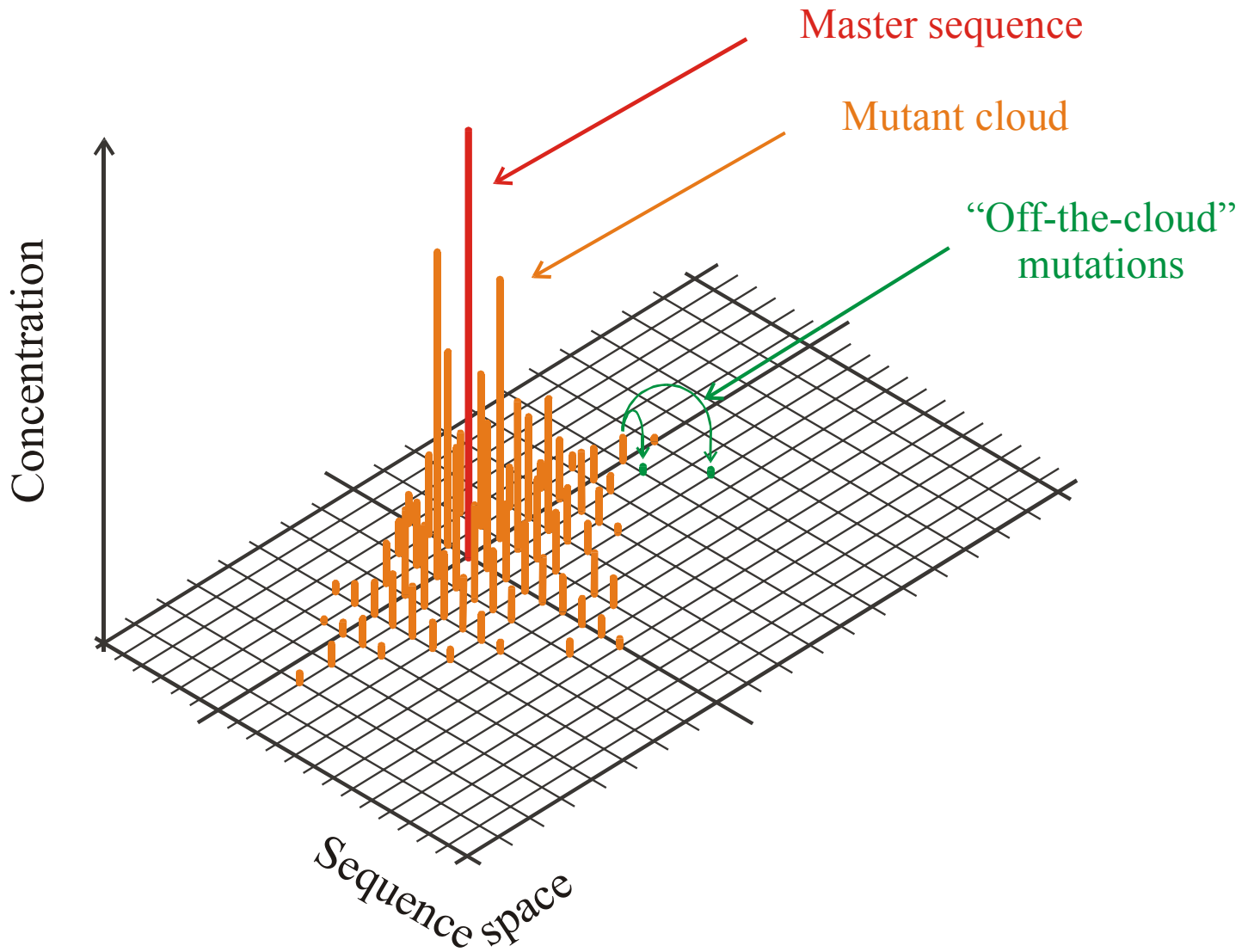


Fitness function:

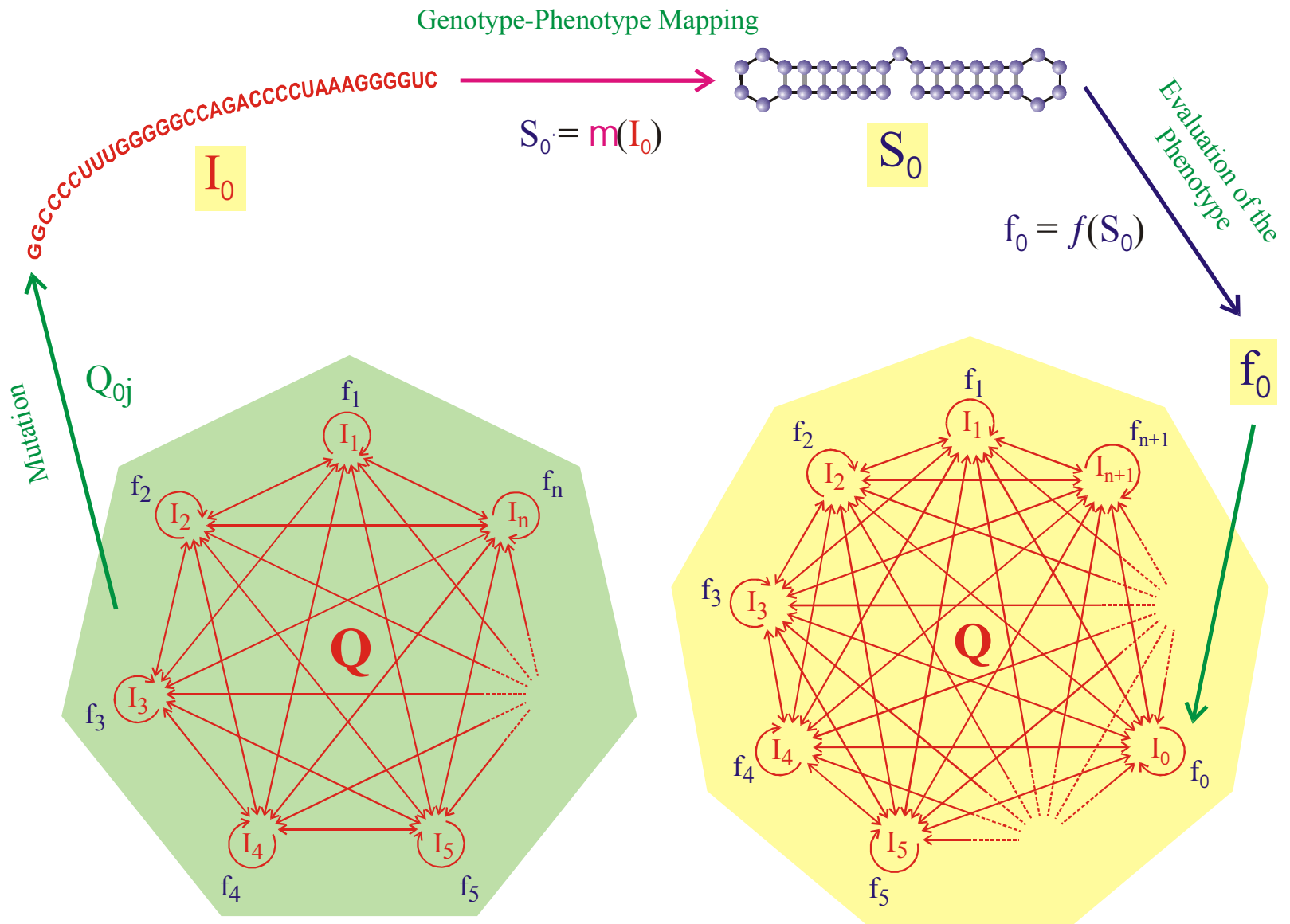
$$f_k = [ / [U + 8d_S^{(k)}]$$

$$8d_S^{(k)} = d^s(I_k, I_h)$$

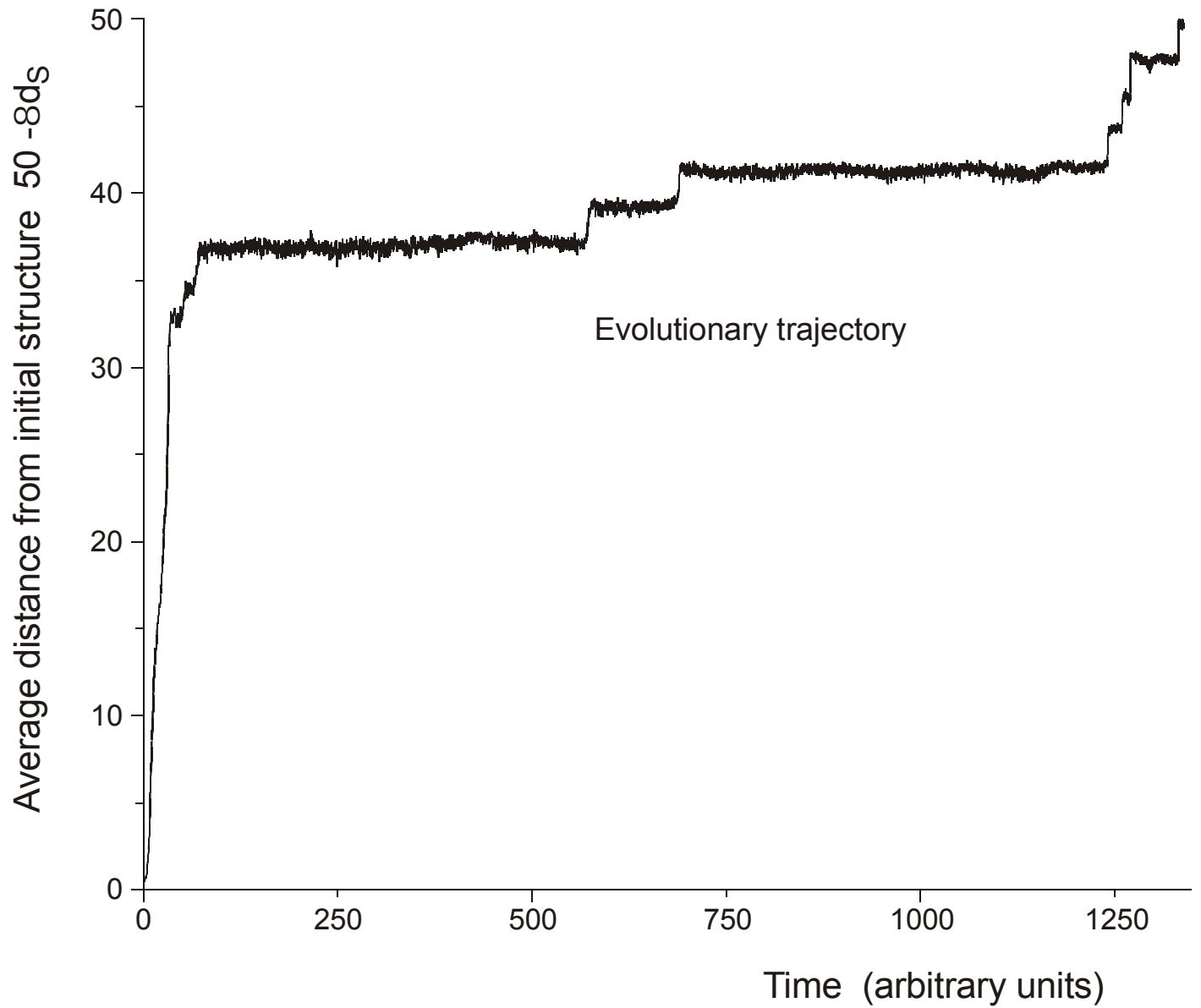
The flowreactor as a device for studies of evolution *in vitro* and *in silico*



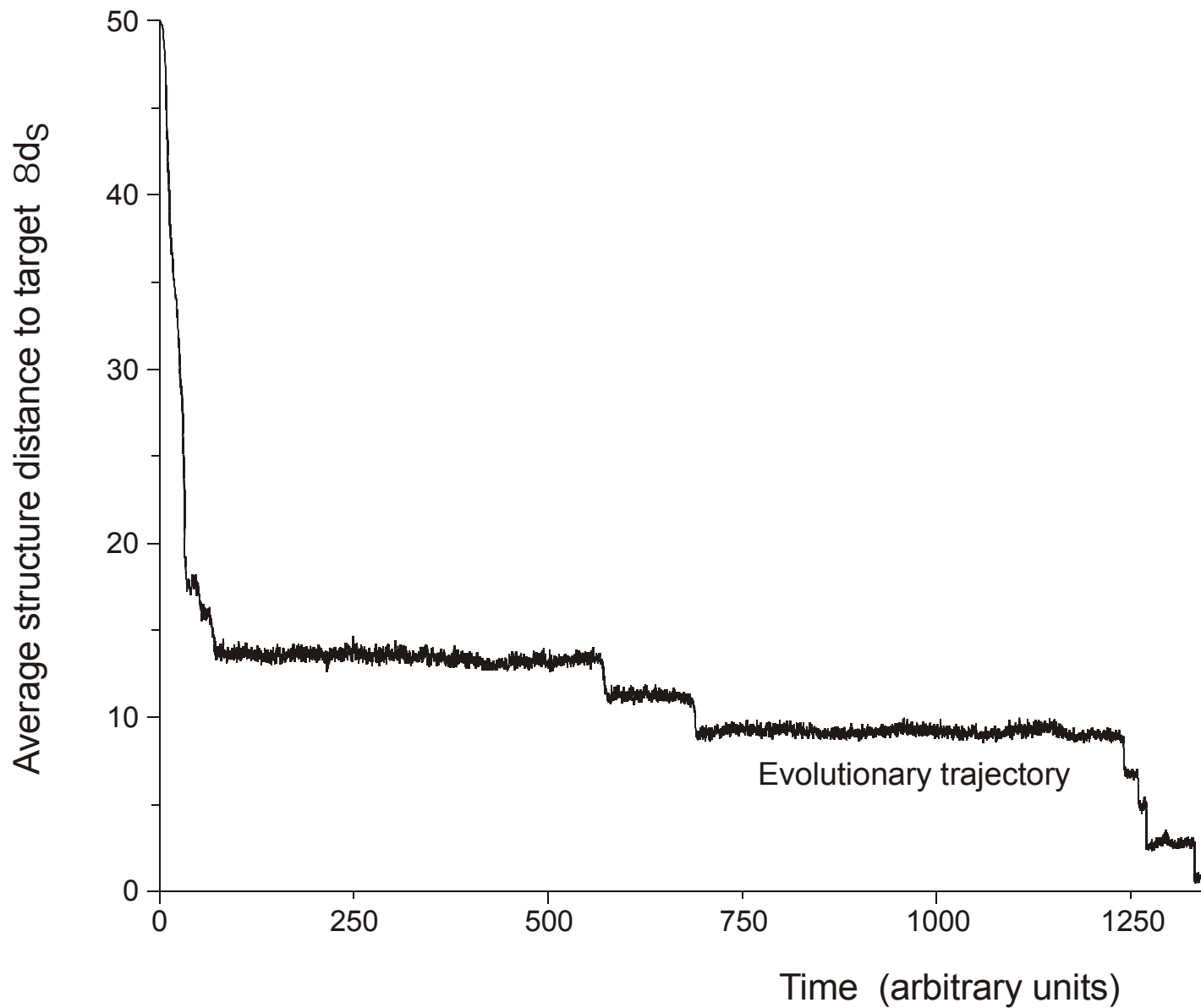
The molecular quasispecies  
in sequence space



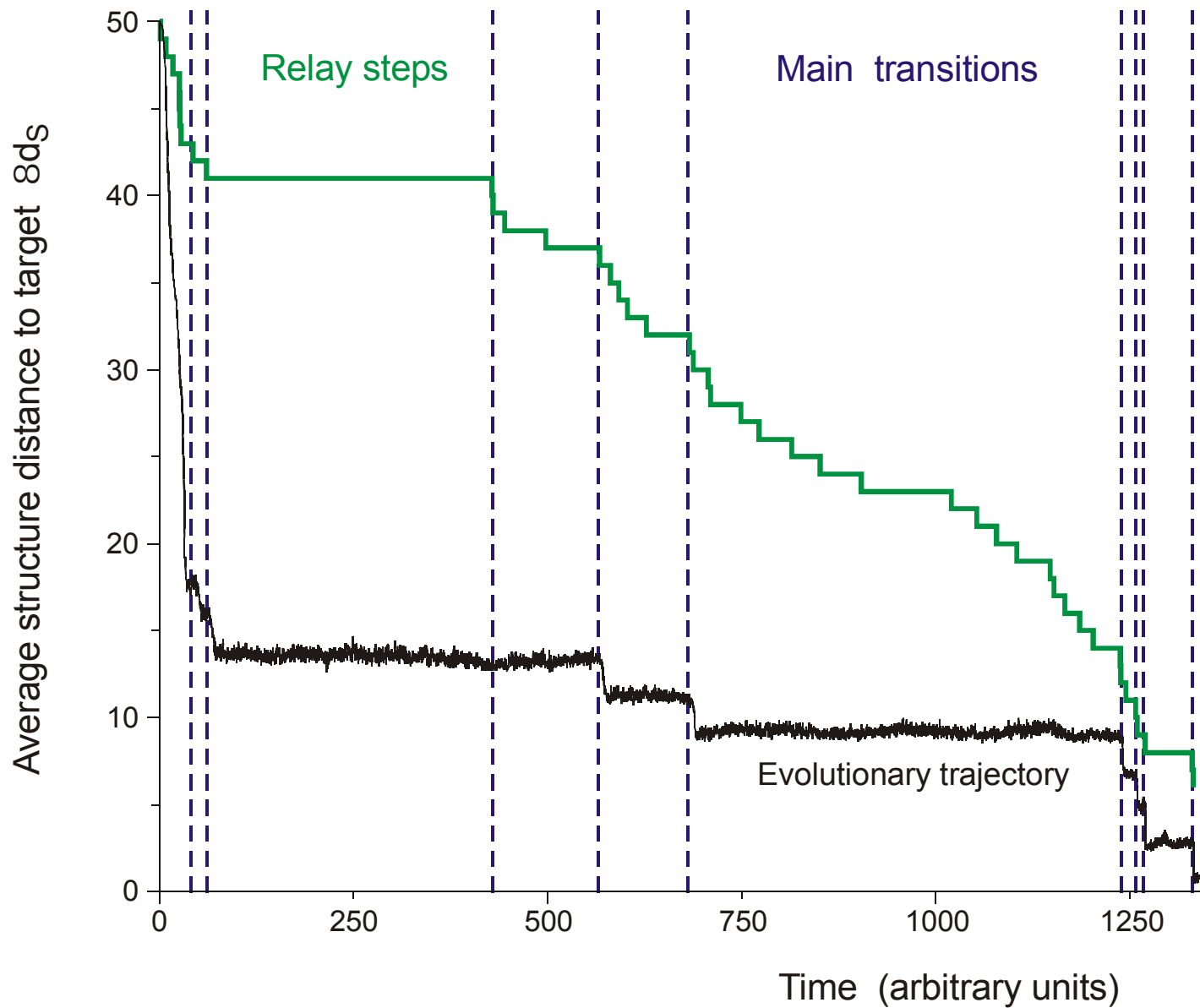
Evolutionary dynamics  
including molecular phenotypes



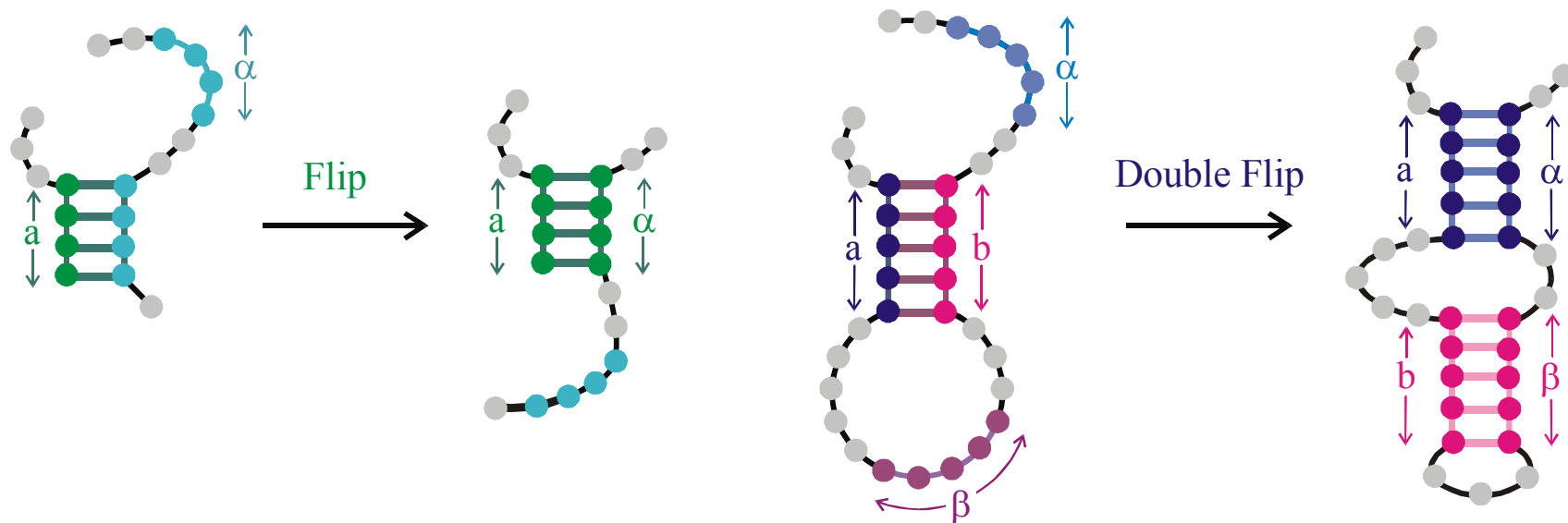
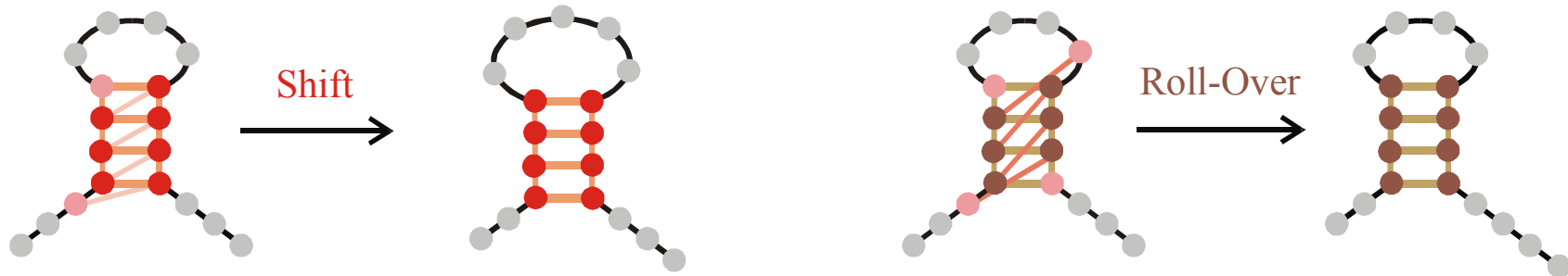
*In silico* optimization in the flow reactor: Trajectory (**biologists' view**)



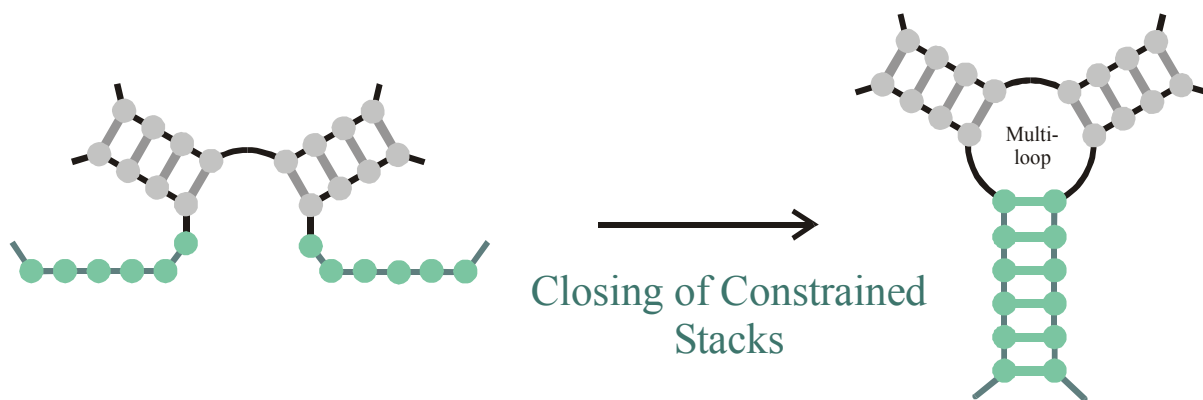
*In silico* optimization in the flow reactor: Trajectory (**physicists' view**)



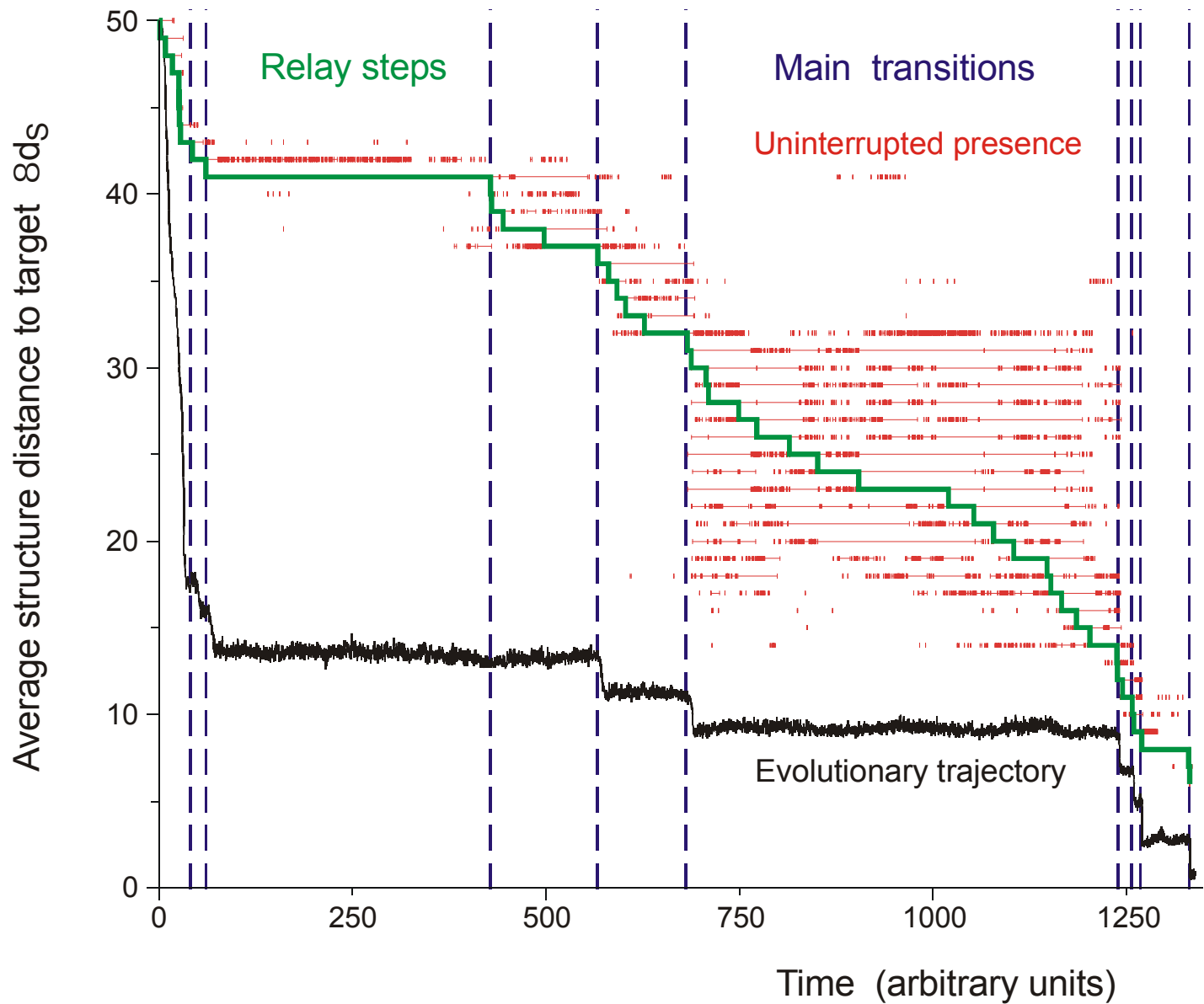
*In silico* optimization in the flow reactor: Main transitions



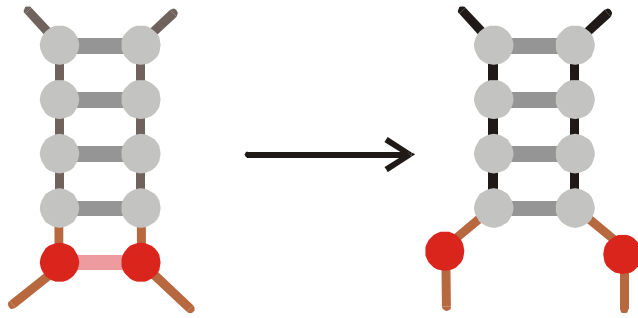
**Main** or discontinuous transitions: **Structural innovations**, occur rarely on single point mutations



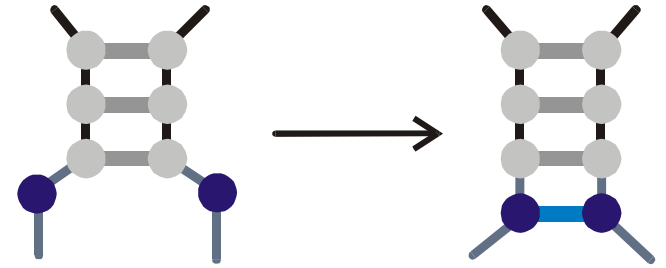




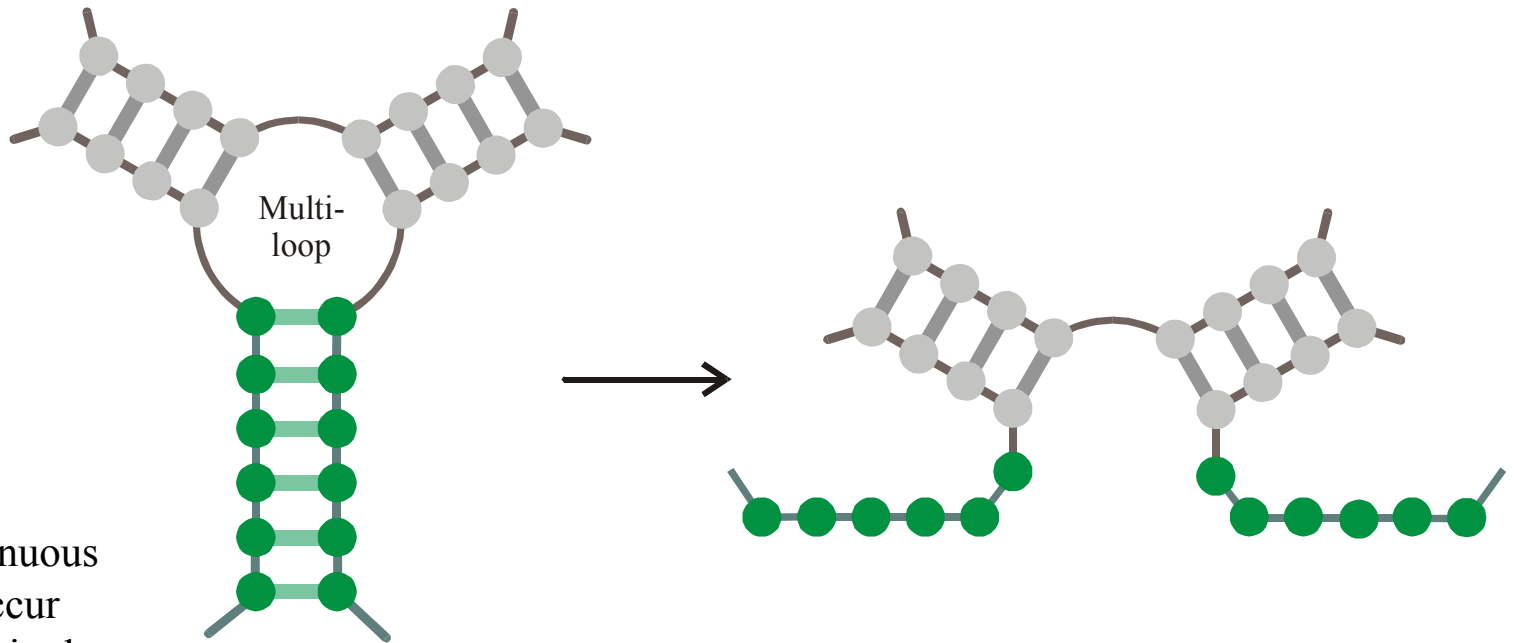
*In silico* optimization in the flow reactor



Shortening of Stacks



Elongation of Stacks



Opening of Constrained Stacks

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

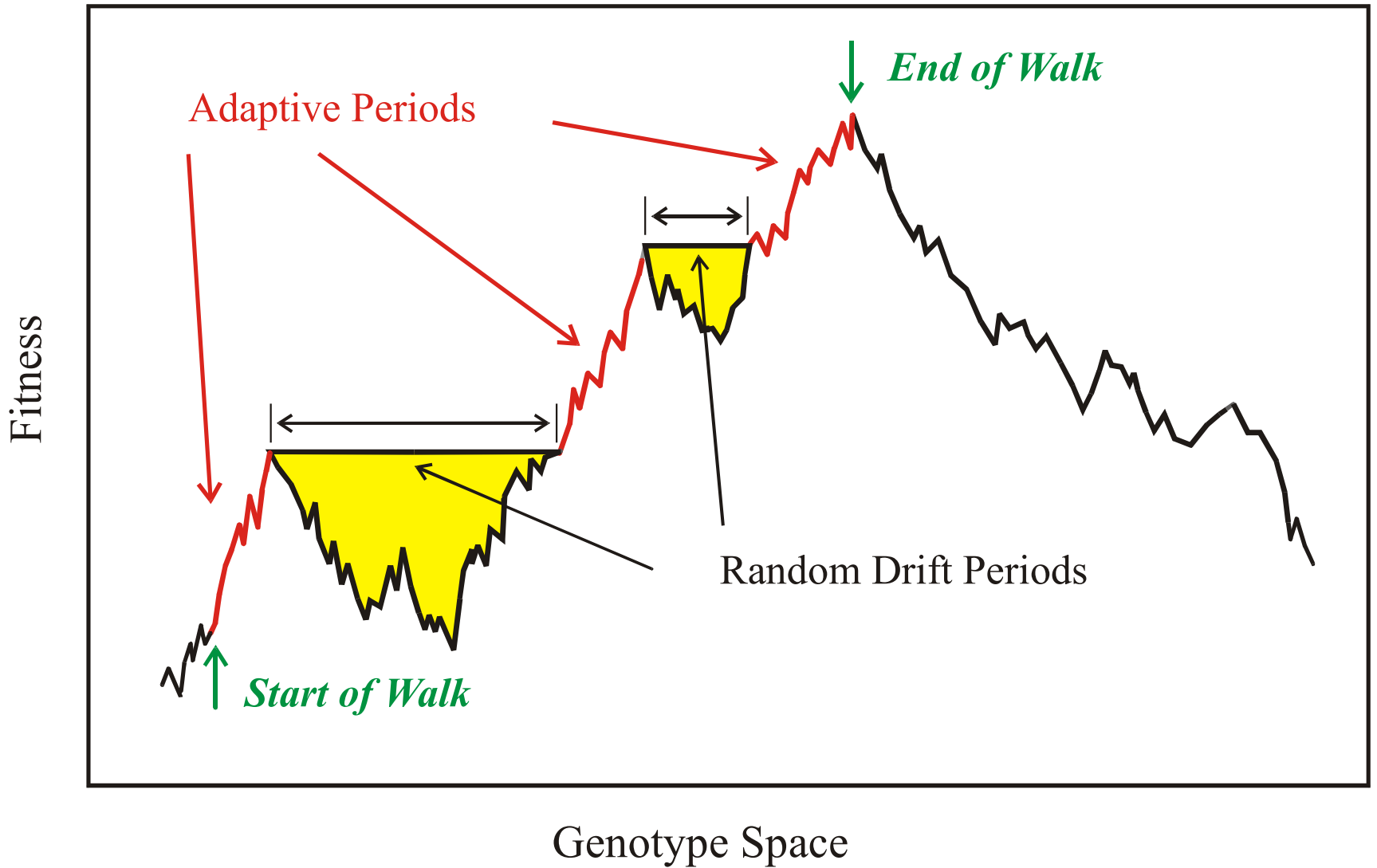
## Statistics of evolutionary trajectories

Population size N	Number of replications < n <sub>rep</sub> >	Number of transitions < n <sub>tr</sub> >	Number of main transitions < n <sub>dtr</sub> >
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	$(3 \pm [2, 1]) \times 10^8$	$24 \pm [6, 5]$	$9 \pm 2$

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

„...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)

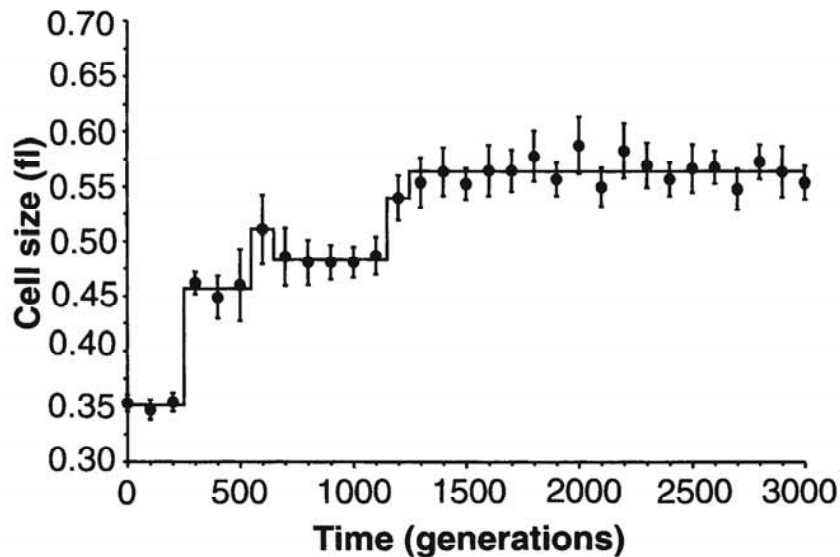


Evolution in genotype space sketched as a non-descending walk in a fitness landscape

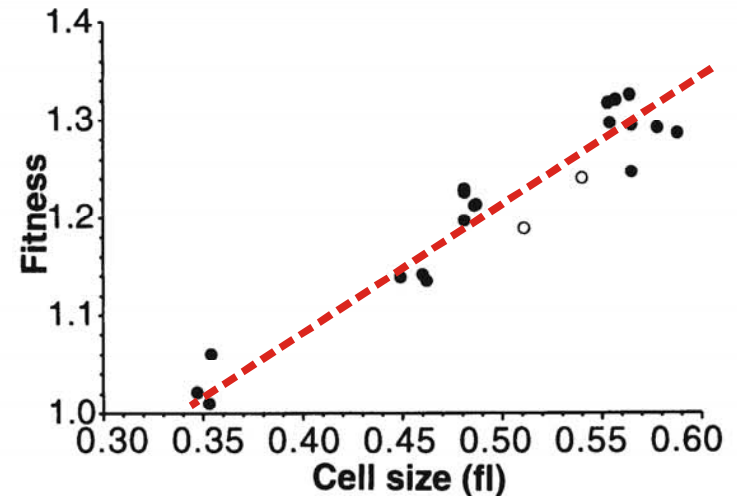
## **Bacterial Evolution**

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants*. Science **272** (1996), 1802-1804

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812



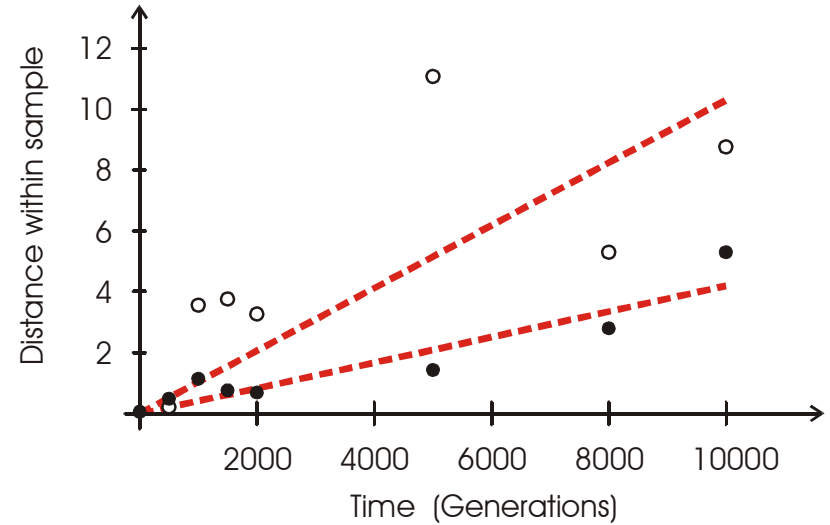
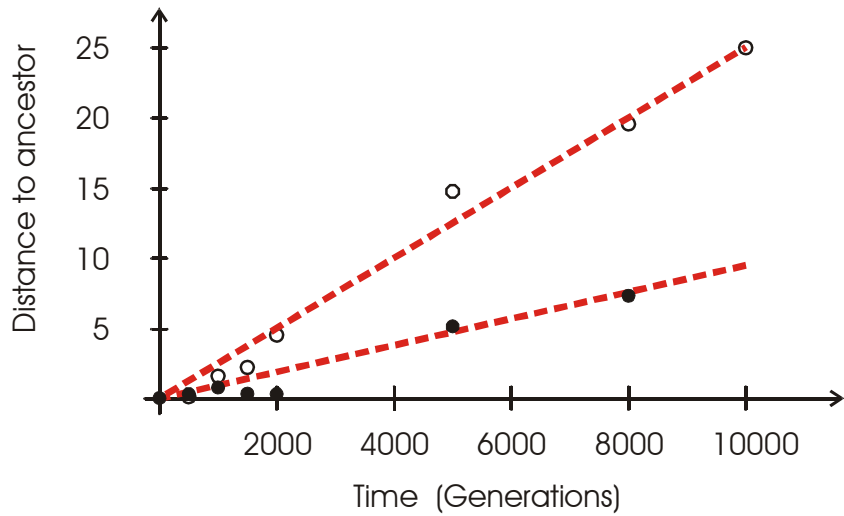
**Fig. 1.** Change in average cell size (1 fl =  $10^{-15}$  L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (22). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these data (Table 1).



**Fig. 2.** Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (6, 7). The open symbols indicate the only two samples assigned to different steps by the cell size and fitness data.

## Epochal evolution of bacteria in serial transfer experiments under constant conditions

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants.* Science **272** (1996), 1802-1804



## Variation of genotypes in a bacterial serial transfer experiment

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812



## 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

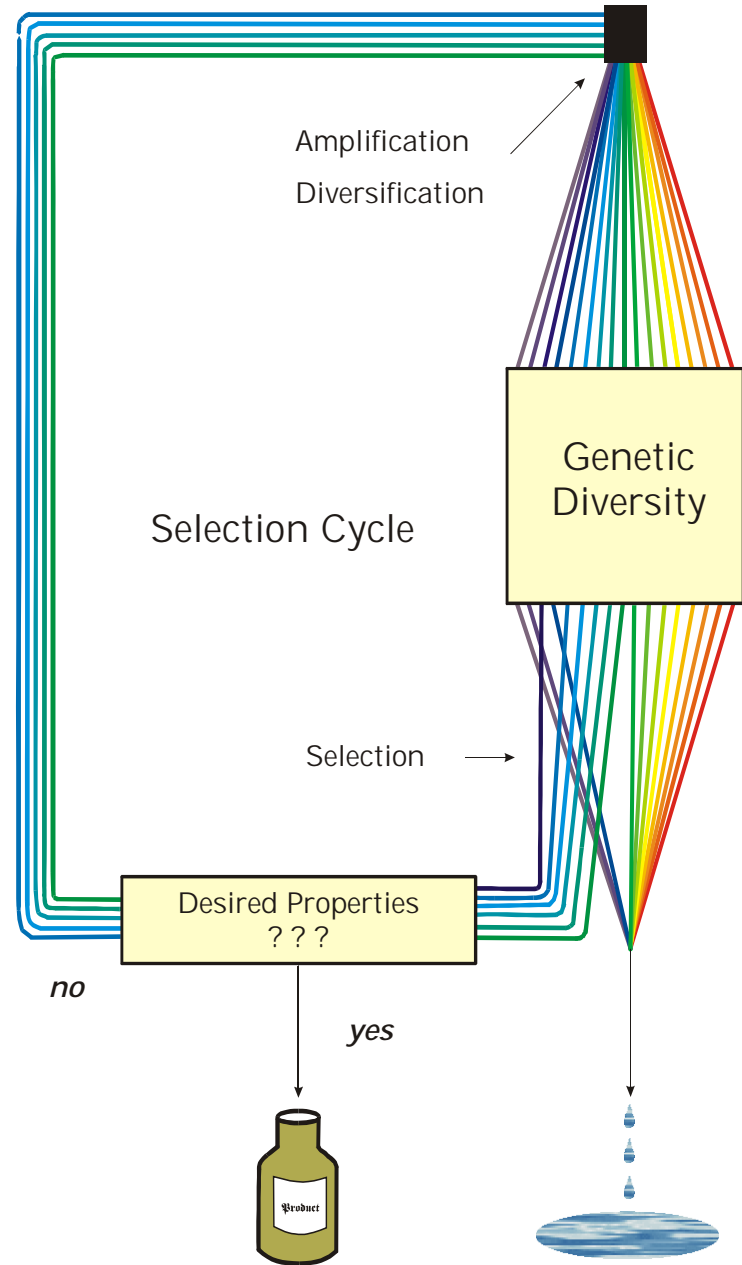
C.Tuerk, L.Gold, *SELEX - Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science **249** (1990), 505-510

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

Y. Wang, R.R.Rando, *Specific binding of aminoglycoside antibiotics to RNA.* Chemistry & Biology **2** (1995), 281-290

Jiang, A. K. Suri, R. Fiala, D. J. Patel, *Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex.* Chemistry & Biology **4** (1997), 35-50

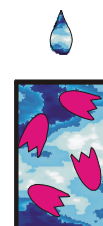
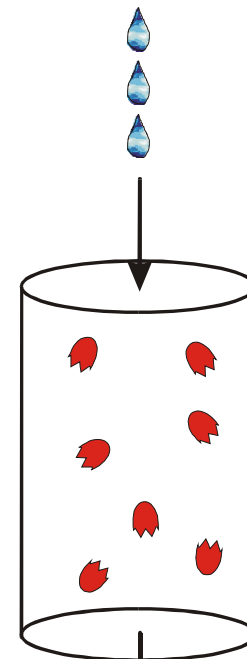
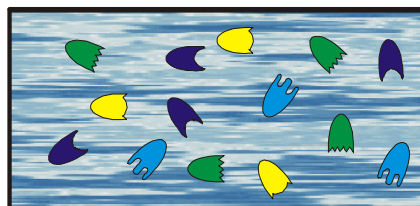
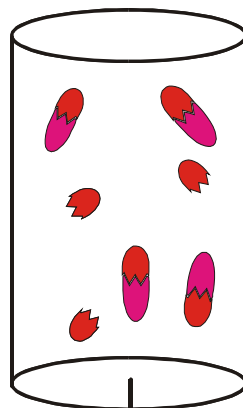
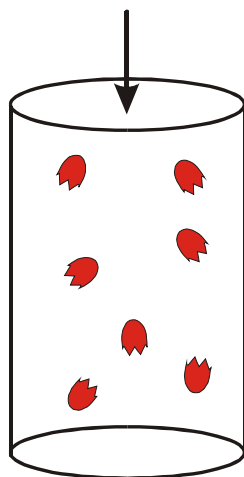
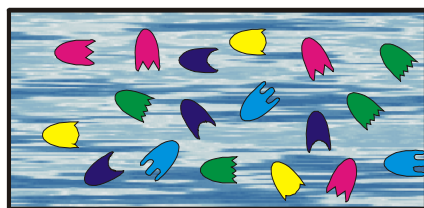


Selection cycle used in applied molecular evolution to design molecules with predefined properties

## Retention of binders

## Elution of binders

Chromatographic column

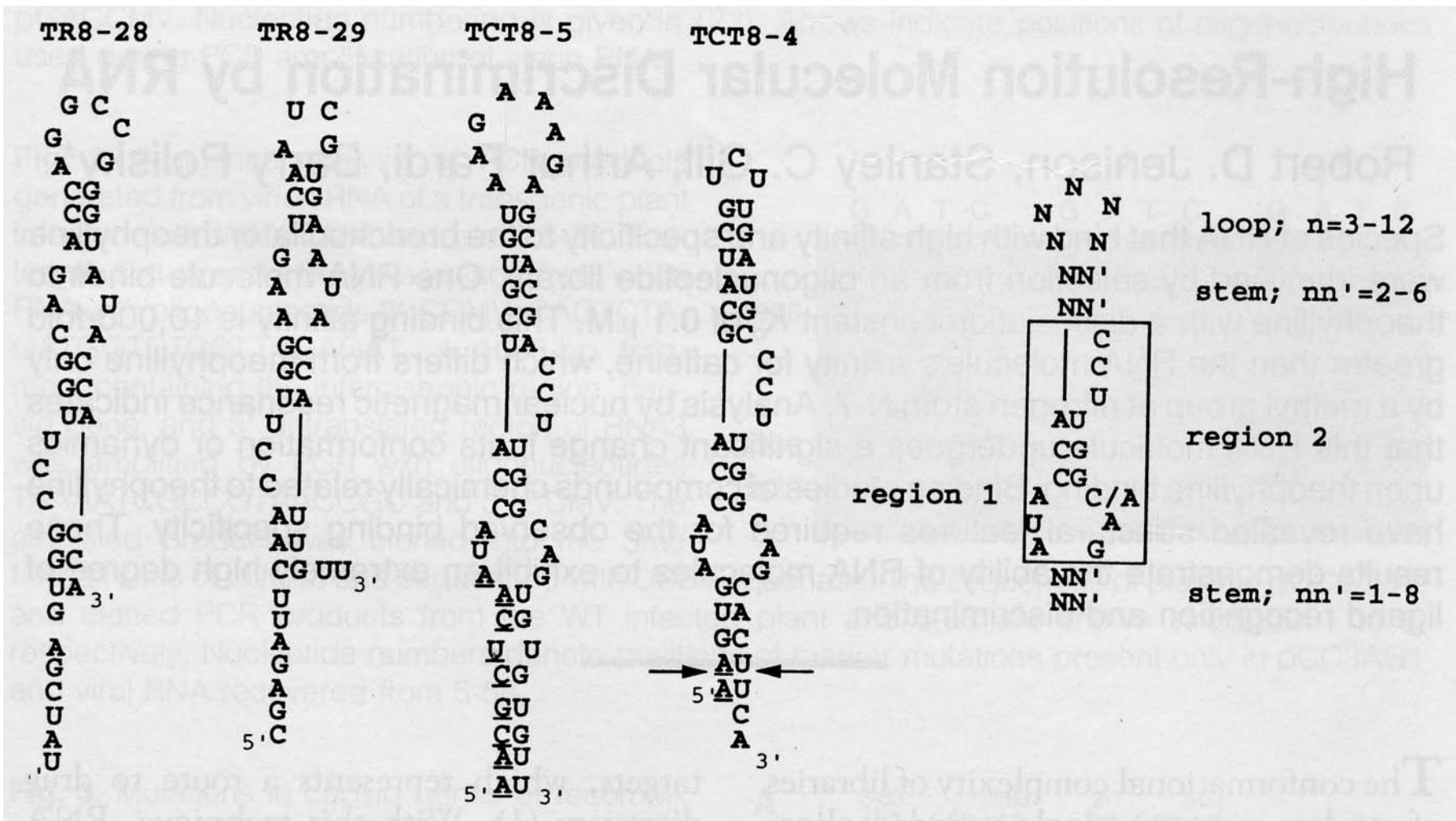


The SELEX technique for the evolutionary design of *aptamers*



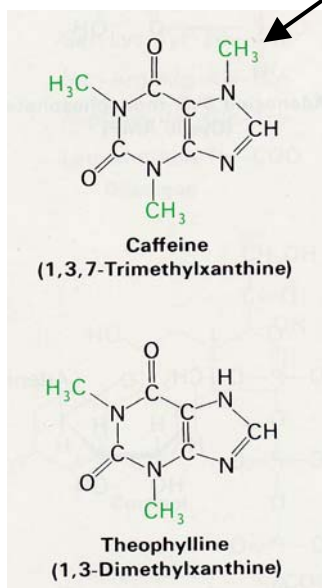
Sequences of aptamers binding theophyllin, caffeine, and related compounds

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



Secondary structures of aptamers binding theophyllin, caffeine, and related compounds

additional methyl group

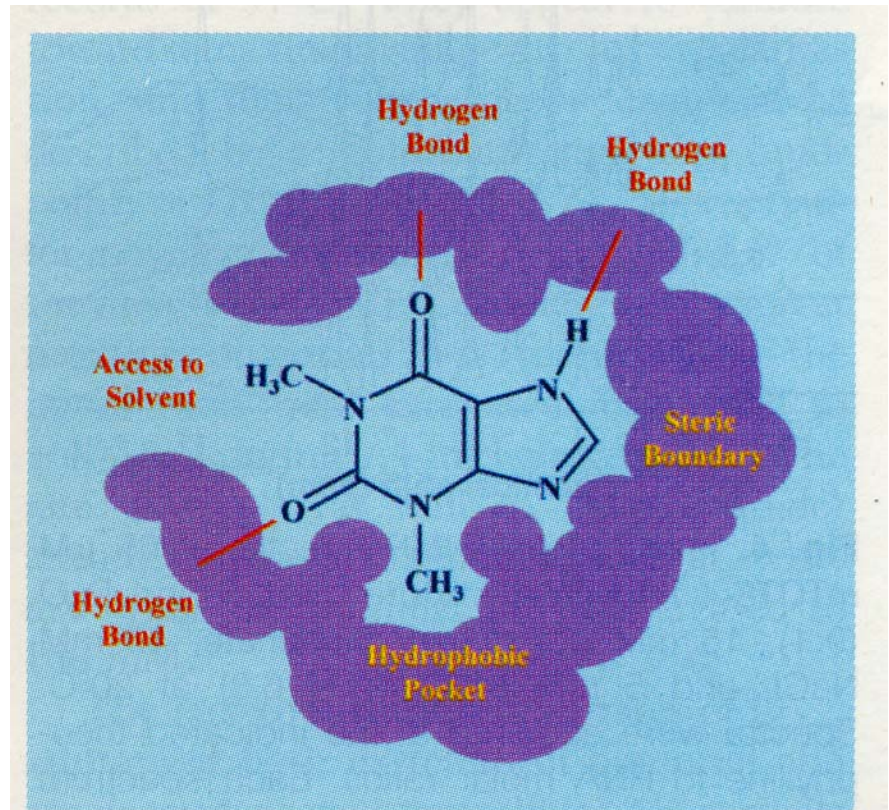


Dissociation constants and specificity of theophylline, caffeine, and related derivatives of uric acid for binding to a discriminating aptamer TCT8-4

**Table 1.** Competition binding analysis with TCT8-4 RNA. The chemical structures are shown for a series of derivatives used in competitive binding experiments with TCT8-4 RNA (Fig. 2) (20). The right column represents the affinity of the competitor relative to theophylline,  $K_d(c)/K_d(t)$ , where  $K_d(c)$  is the individual competitor dissociation constant and  $K_d(t)$  is the competitive dissociation constant of theophylline. Certain data (denoted by >) are minimum values that were limited by the solubility of the competitor. Each experiment was carried out in duplicate. The average error is shown.

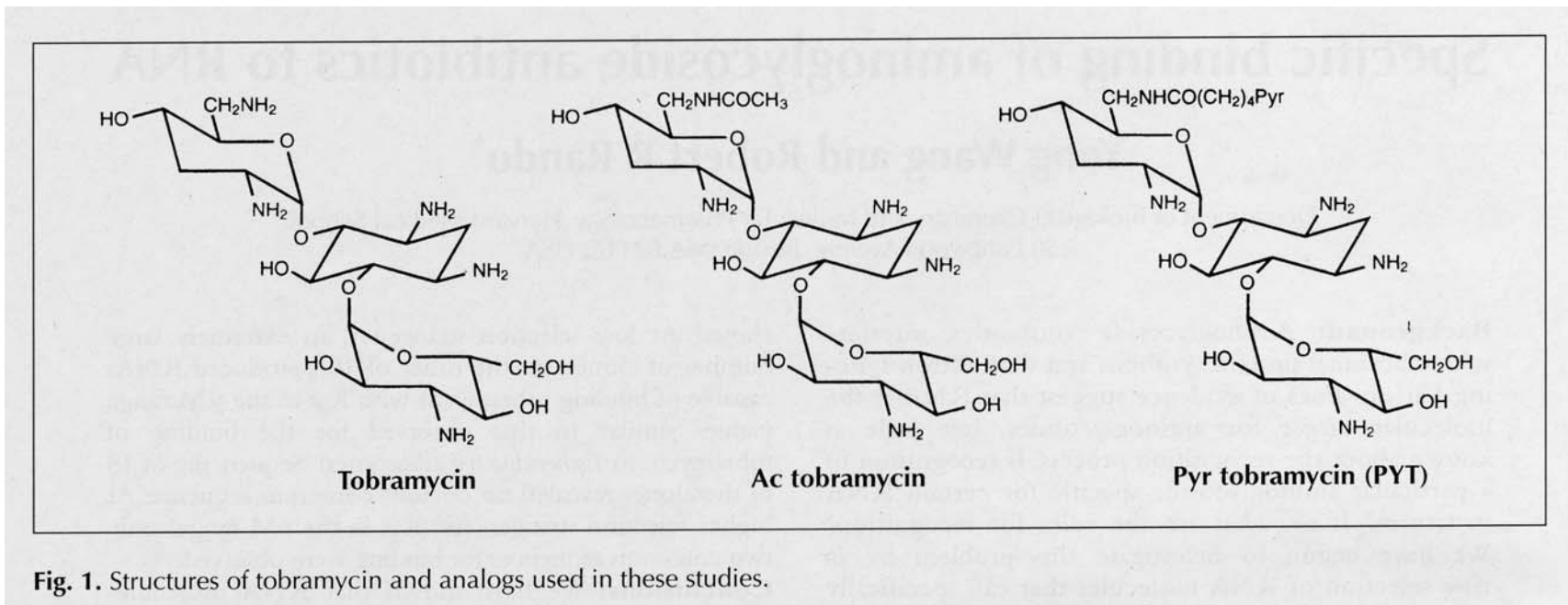
Compound	Structure	$K_d(c)$ ( $\mu\text{M}$ )	$K_d(c)/K_d(t)$
Theophylline		$0.32 \pm 0.13$	1
CP-theophylline		$0.93 \pm 0.20$	2.9
Xanthine		$8.5 \pm 0.40$	27
1-Methylxanthine		$9.0 \pm 0.30$	28
3-Methylxanthine		$2.0 \pm 0.7$	6.3
7-Methylxanthine		> 500	>1500
3,7-Dimethylxanthine		> 500	> 1500
1,3-Dimethyluric acid		> 1000	>3100
Hypoxanthine		$49 \pm 10$	153
Caffeine		$3500 \pm 1500$	10,900





**Fig. 3.** Schematic representation of the RNA (purple) binding site for theophylline (blue).

Schematic drawing of the aptamer binding site for the theophylline molecule



## Aptamer binding to aminoglycosid antibiotics: Structure of ligands

Y. Wang, R.R.Rando, *Specific binding of aminoglycoside antibiotics to RNA*. Chemistry & Biology 2 (1995), 281-290



```

1. 5' ... AAUUGGAUCAUACUGUUAUAAGUAGGUGGCUCAGCGGUUGCCUUCGGCAGGUUGAUCCGU...3'
2. 5' ... GUGGUUAUACCUGGUUGUGCGUAUGUACUCCGGCUUCUAUAGUUCUGGUUGGGUGCGGAC...3'
3. 5' ... UGUGAUAGGACUGUGAUUCGGAUCAUGUUGUGAACAUUGGUGUGCAUCAUGGGGAGCCCGU...3'
4. 5' ... CUCAUUCACGGGAUGGCUCUUAACACAGUCUGUGUAGGGCGUCGCGUGUGGUUAUGUUGGC...3'
5. 5' ... UUCACGGCAUAAUUCUGUUUGAGUUGGCCUUUUGGUAAUGUAGCGUGGGUUUGCGGUGCG...3'
6. 5' ... CCUGUGGGGGCGUACGUUGGUACAAGAGGUUAGGUGGGCCUAAUGUAGCCUACGUGCGGGCGU...3'
7. 5' ... CUGCUGCAGUUGUUAUAGGGCCGUUCUAACUGGGGUGGUACCCGACCGAGUCAGCAGGUG...3'
8. 5' ... CGUUUGGGGUCCCACAACACAGGUCUUUGCUGGUCAUAUAUGCGUGUCCUCUAGGAAGUG...3'
9. 5' ... CGUAAAUGGCGCAACGUCGCUCAUGUCGGGUUGGCCAAGUUUCAUUUGCGUUGUCUCAGU...3'
10. 5' ... GCGUUCAGCGGUUCUUCGGGUAAGGCCGUGAGUGUGCGUACGUGAGGGUUAACGCCCGU...3'
11. 5' ... UAUGUUGCACCCACUGCGGUUCAGUUAUGGUUGUUUCGGCCAUUUGGUUCGCAGCUGUGA...3'
12. 5' ... GGUCAGGCGUCUCAUGUAGUUUGUUCUGUGUAUAUUUGCUCUCUUCGGGGACACCCUGUC...3'
13. 5' ... AGUGAUUAUCACAACAGUAUGUCUGGUUGCUCUAGGAGGUUGCAUUGAGGCCUAAUAGUAG...3'
14. 5' ... UCAAGUCCUGUCUGGUGCGCCUACAGGAUUGGGAUCAUGGGUGUCAAGCAUUUGUGUC...3'
15. 5' ... CACGGUAAAGUAGCAUUGUUCGUUUCUUGUUUUACUACUCGUAGUGACAGCAGAUGCUG...3'
16. 5' ... UGUGUUGCUCACGAGGUUGGCCAAUAUGUAGCGUUGCUGCAUAGUGGUUGUCUUAUAUCUG...3'
17. 5' ... UGCAGGCUUUUGGCUGUGGGUUUAUGUCCAUGUGUUAGAUCAGCGUUUACGGUUAGCGC...3'
18. 5' ... CCACUAGGCUACAUAAGGCGUACAUCUCCCGGAGGUCUUAGCGCCGCUAUGGGGUUGGA...3'

```

**Fig. 3.** Nucleotide sequences of low-affinity aptamers. The sequence of the random region of 18 clones is shown. Total sequence: 5'-GGGAGAAUCCGACCAGAAGCUU-random region-CAUAUGUGCGUCUACAUGGAUCCUCA-3'.

## Aptamer binding to aminoglycosid antibiotics: Sequence of low affinity RNA aptamers

Y. Wang, R.R.Rando, *Specific binding of aminoglycoside antibiotics to RNA*. Chemistry & Biology 2 (1995), 281-290

X1(4): 5' ... CUGGUUAGUUUUGCACAGUGGTCGAUGCUAGACUUGGUUUAGGUA AUGAGUCCA AUAGUC ... 3'  
 X8(2): 5' ... GCGGUUGAGCGCUC CCGGUACA UUGACAGGCAGGCAUGGUUU AACUAAUGUGCUUGGACUU ... 3'  
 K13(1): 5' ... CUUGGUUUAGGUA AUGAGCAUCAUCGGCCGAUUUCAUCUCUGUGGGUAGUACGGAAGCAC ... 3'  
 -----  
 J6(13): 5' ... AGUAUAGCGAGGUUUAGCUACACUCGUGCUGAUCGUUUGGUACGGGACCUGCGUGUAGCC ... 3'  
 J14(2): 5' ... CCAAAAAGUAAGGACGAGAGCCGACUAGGUUUAGCUACACUAGAGCUCUC CGCCAACGG ... 3'  
 X3(8): 5' ... UGGUUGUCUUGUACGUUCACUGUUACGAUUGUGUUAGGUUU AACUACACUUUGCAAUCG ... 3'

**Fig. 8.** Nucleotide sequences of high-affinity aptamers. Six independent sequences are shown with the number of times that each sequence was obtained indicated in parentheses. The sequences fall into two groups. Consensus sequences are highlighted.

**Table 2.** Dissociation constants of RNA aptamers.

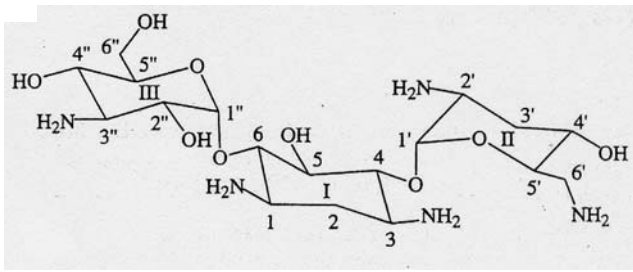
	J6	X1	J6sl	X1sl
$K_1$ (PYT) (nM)	5.4±0.5 <sup>a</sup>	7.6±1.1 <sup>a</sup>	286±10 <sup>a</sup>	65.1±7.4 <sup>b</sup>
$K_1$ (TOB) (nM)	2±1	3±1	9±3	12±5
$K_2$ (TOB) (μM)	6.0±0.4	15.9±0.7	2.7±0.3	13.2±0.5

Binding measurements were made at 20 °C as described in Materials and methods. The high affinity component for <sup>a</sup>PYT, or <sup>b</sup>6'-N-4-pyreneacetyl tobramycin, is measured along with both high- and low-affinity components ( $K_1$  and  $K_2$ , respectively) for the binding of tobramycin (TOB). Values for binding to J6, X1 and their stem-loop (sl) counterparts are given.

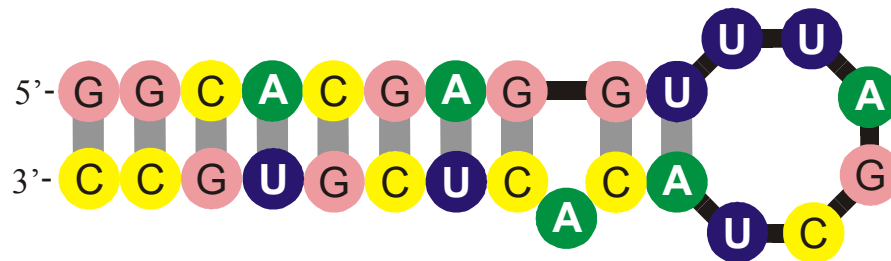
Aptamer binding to aminoglycosid antibiotics: Sequence of high affinity RNA aptamers and dissociation constants of RNA-antibiotic complexes

Y. Wang, R.R.Rando, *Specific binding of aminoglycoside antibiotics to RNA*. Chemistry & Biology 2 (1995), 281-290





tobramycin

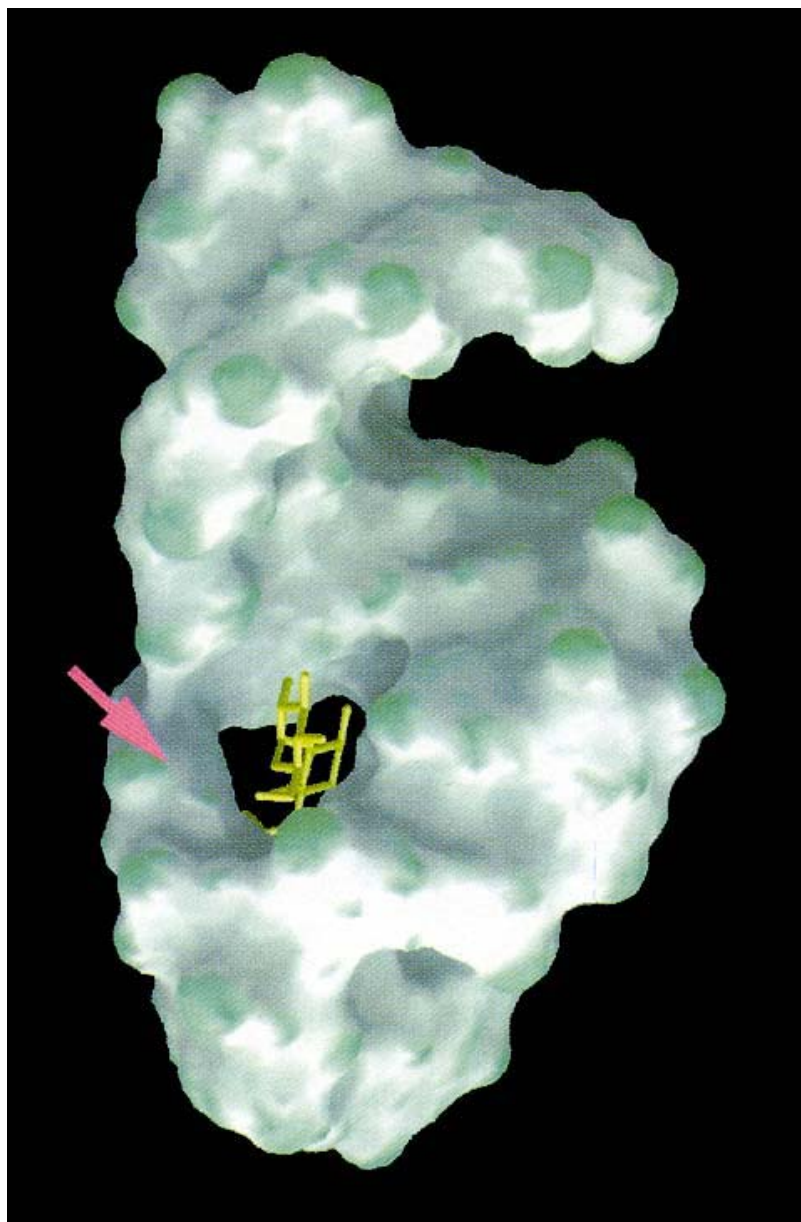


RNA aptamer

Formation of secondary structure of the tobramycin binding RNA aptamer

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, *Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex*. *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)

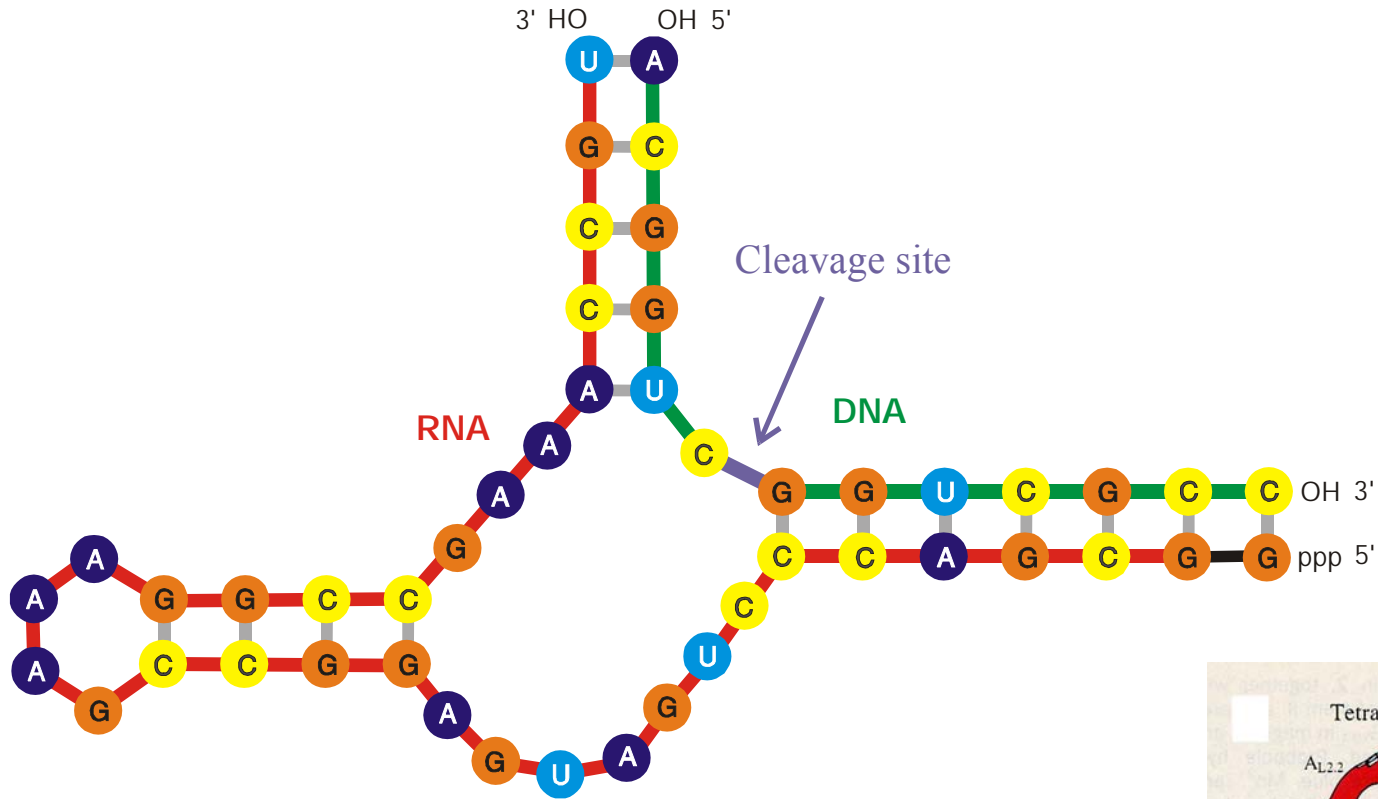
## **Hammerhead ribozyme – The smallest based catalyst**

H.W. Pley, K.M. Flaherty, D.B. McKay, *Three dimensional structure of a hammerhead ribozyme*. Nature **372** (1994), 68-74

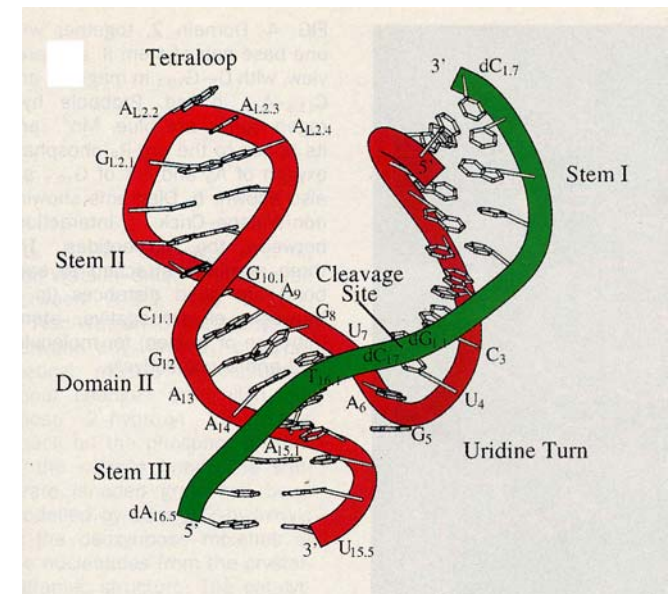
W.G. Scott, J.T. Finch, A. Klug, *The crystal structures of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage*. Cell **81** (1995), 991-1002

J.E. Wedekind, D.B. McKay, *Crystallographic structures of the hammerhead ribozyme: Relationship to ribozyme folding and catalysis*. Annu.Rev.Biophys.Biomol.Struct. **27** (1998), 475-502

G.E. Soukup, R.R. Breaker, *Design of allosteric hammerhead ribozymes activated by ligand-induced structure stabilization*. Structure **7** (1999), 783-791



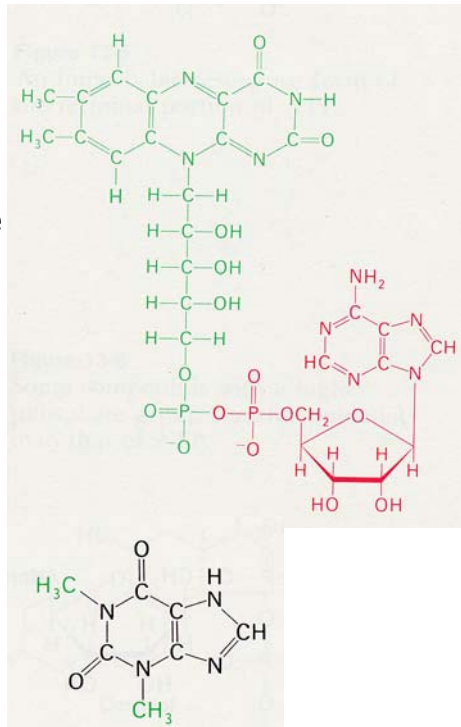
**Hammerhead ribozyme:** The smallest known catalytically active RNA molecule



**Allosteric effectors:**

FMN = flavine mononucleotide

H10 – H12



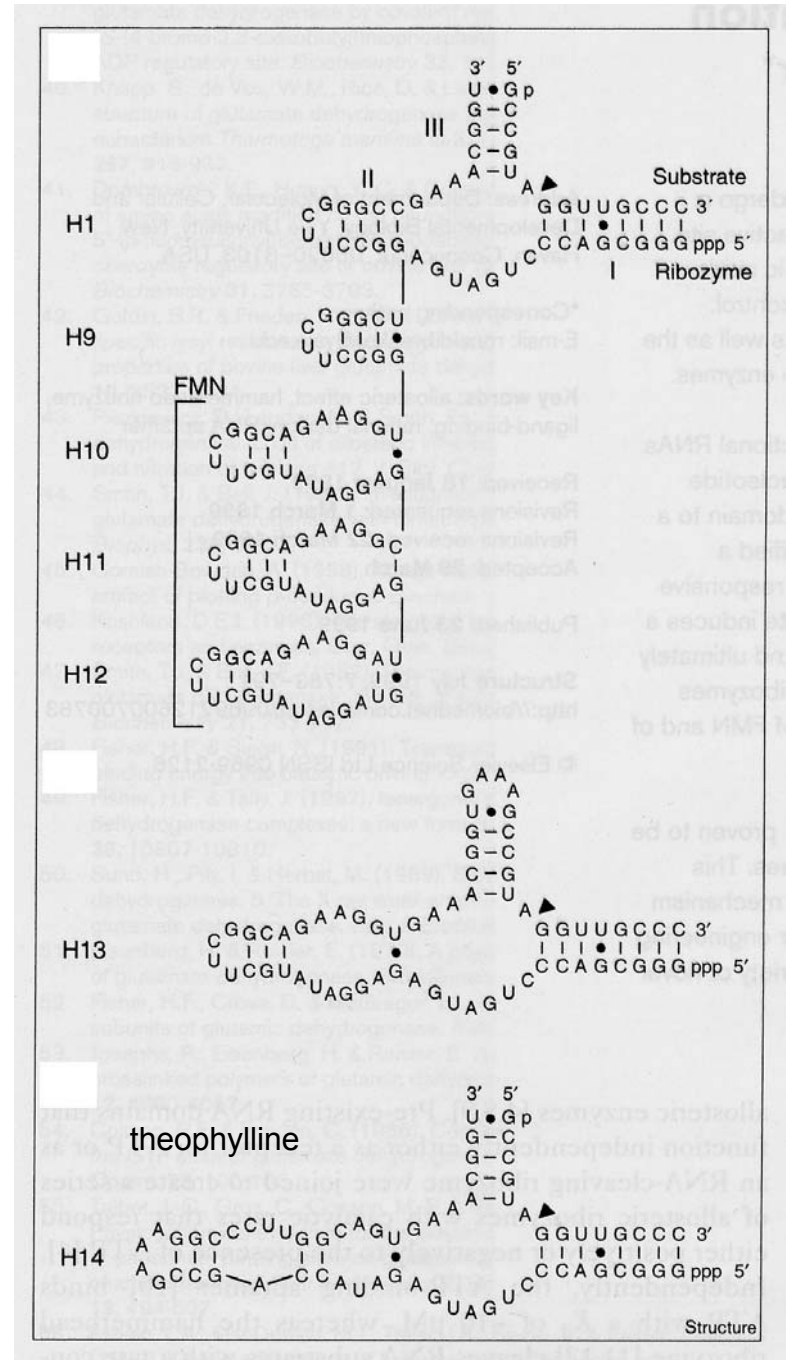
theophylline

H14

Self-splicing allosteric ribozyme

H13

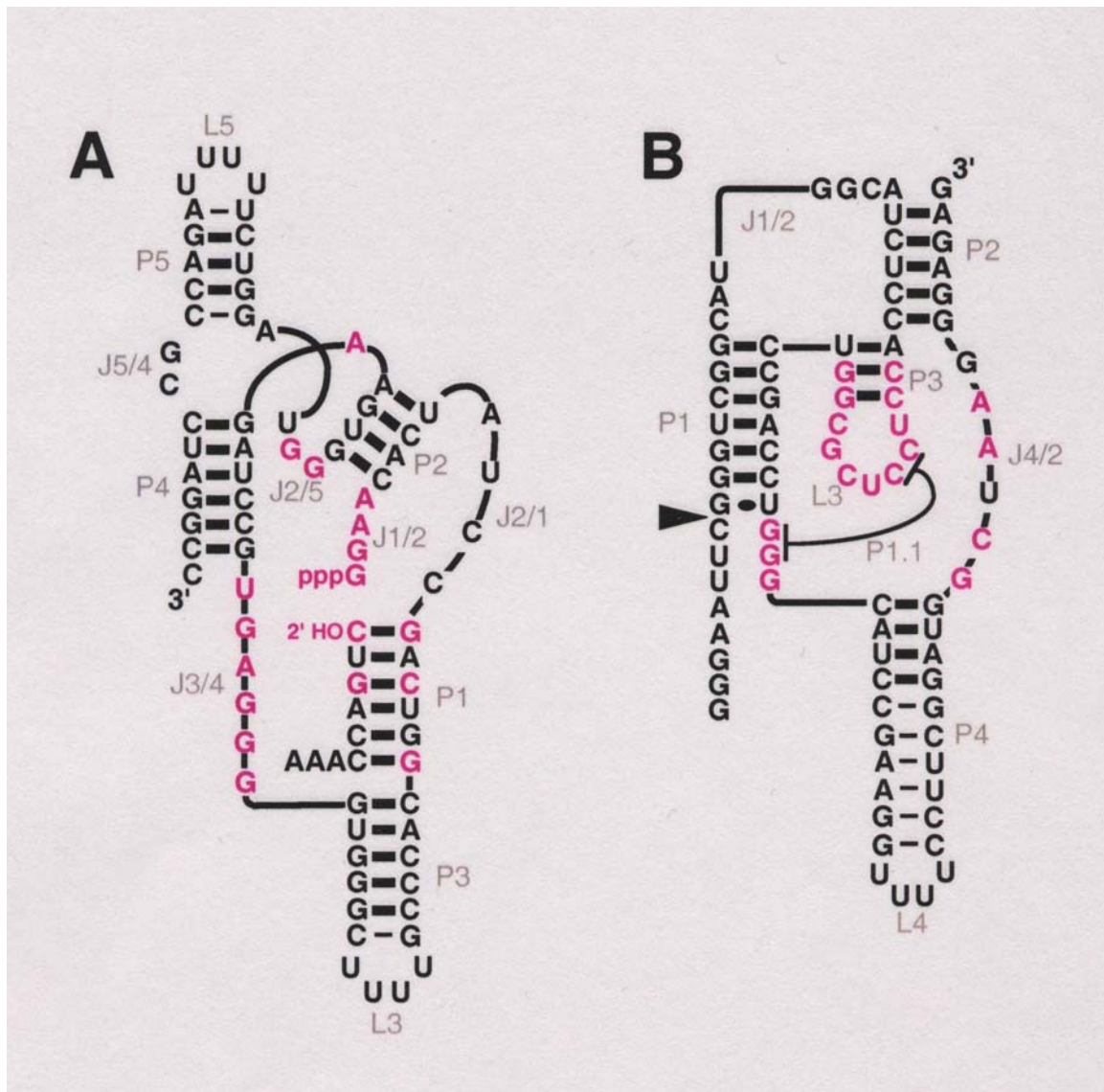
Hammerhead ribozymes with allosteric effectors



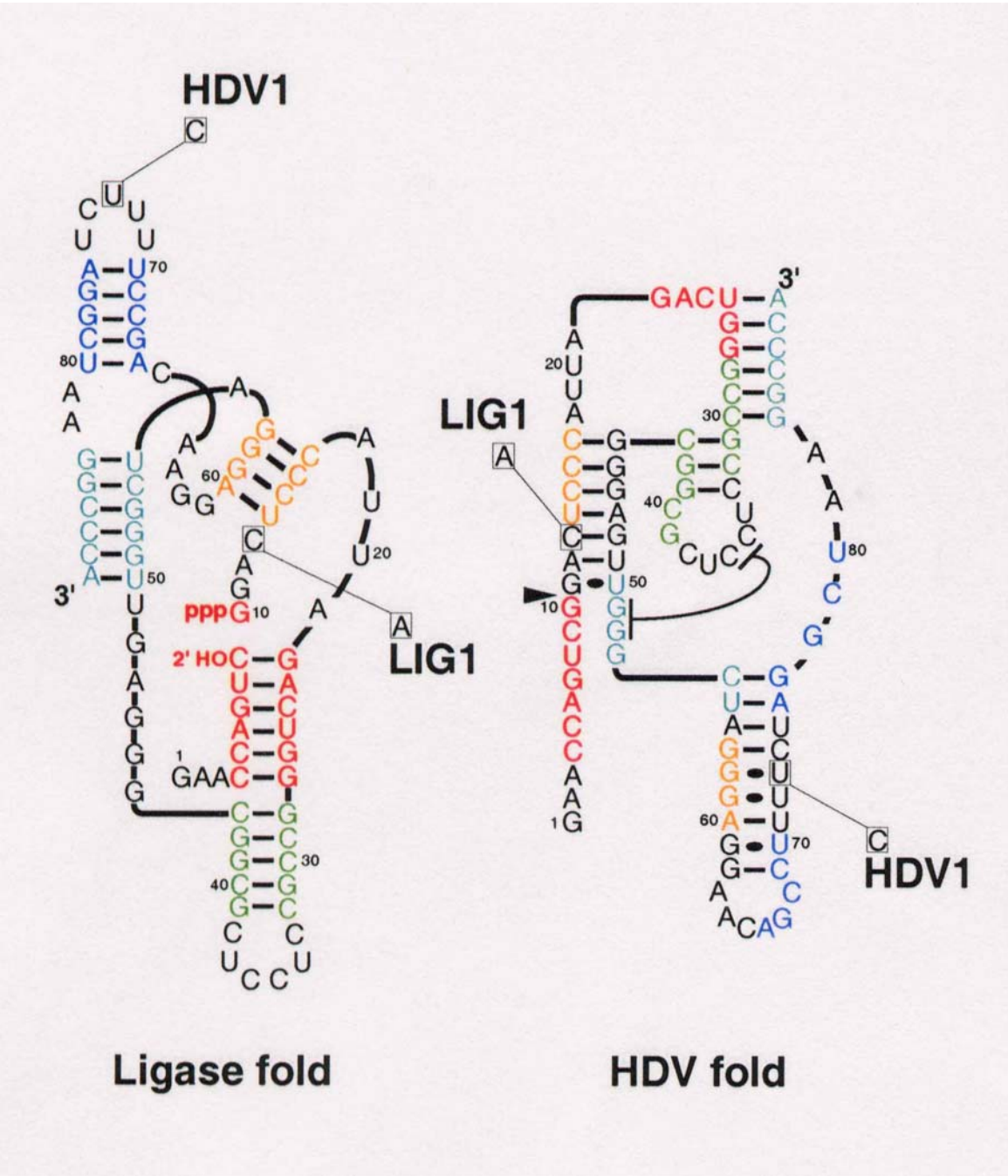


## **A ribozyme switch**

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-X-virus (B)



The sequence at the *intersection*:

An RNA molecules which is 88 nucleotides long and can form both structures



S0092-8240(96)00089-4

## GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES<sup>1</sup>

■ CHRISTIAN REIDYS\*, †, PETER F. STADLER\*, ‡  
 and PETER SCHUSTER\*, ‡, §, ¶

\*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 ( $\lambda$ ). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ( $\lambda > \lambda^*$ ). Below threshold ( $\lambda < \lambda^*$ ), the networks are partitioned into a largest “giant” component and several smaller components. Structures are classified as “common” or “rare” according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of *shape space covering* sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. © 1997 Society for Mathematical Biology

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

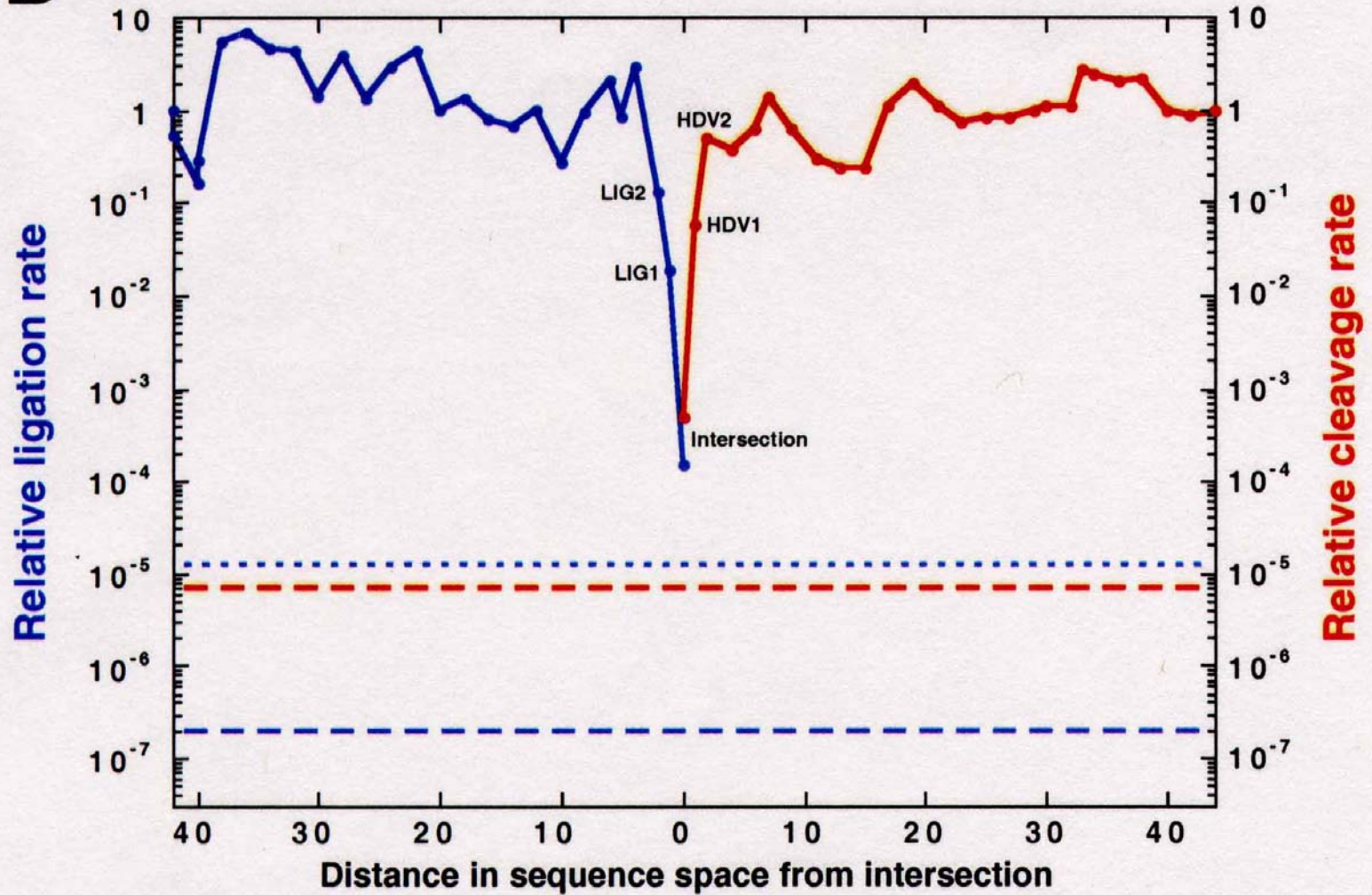
$$C[s] \cap C[s'] \neq \emptyset.$$

*Proof.* Suppose that the alphabet admits only the complementary base pair  $[XY]$  and we ask for a sequence  $x$  compatible to both  $s$  and  $s'$ . Then  $f(s, s') \cong D_m$  operates on the set of all positions  $\{x_1, \dots, x_n\}$ . Since we have the operation of a dihedral group, the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore, it suffices to choose an alternating sequence of the pairing partners  $X$  and  $Y$ . Thus, there are at least two different choices for the first base in the orbit. ■

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

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



**B**

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<sup>1,2,3</sup>, WALTER FONTANA<sup>3</sup>, PETER F. STADLER<sup>2,3</sup>  
AND IVO L. HOFACKER<sup>2</sup>

<sup>1</sup> Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany

<sup>2</sup> Institut für Theoretische Chemie, Universität Wien, Austria

<sup>3</sup> 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.

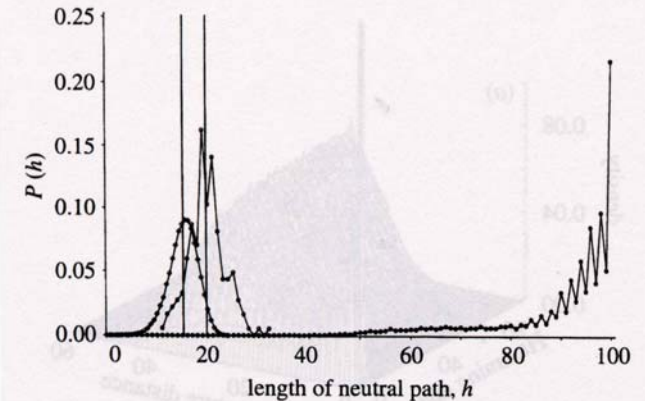


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and 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.* 1993a; 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).

# **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**