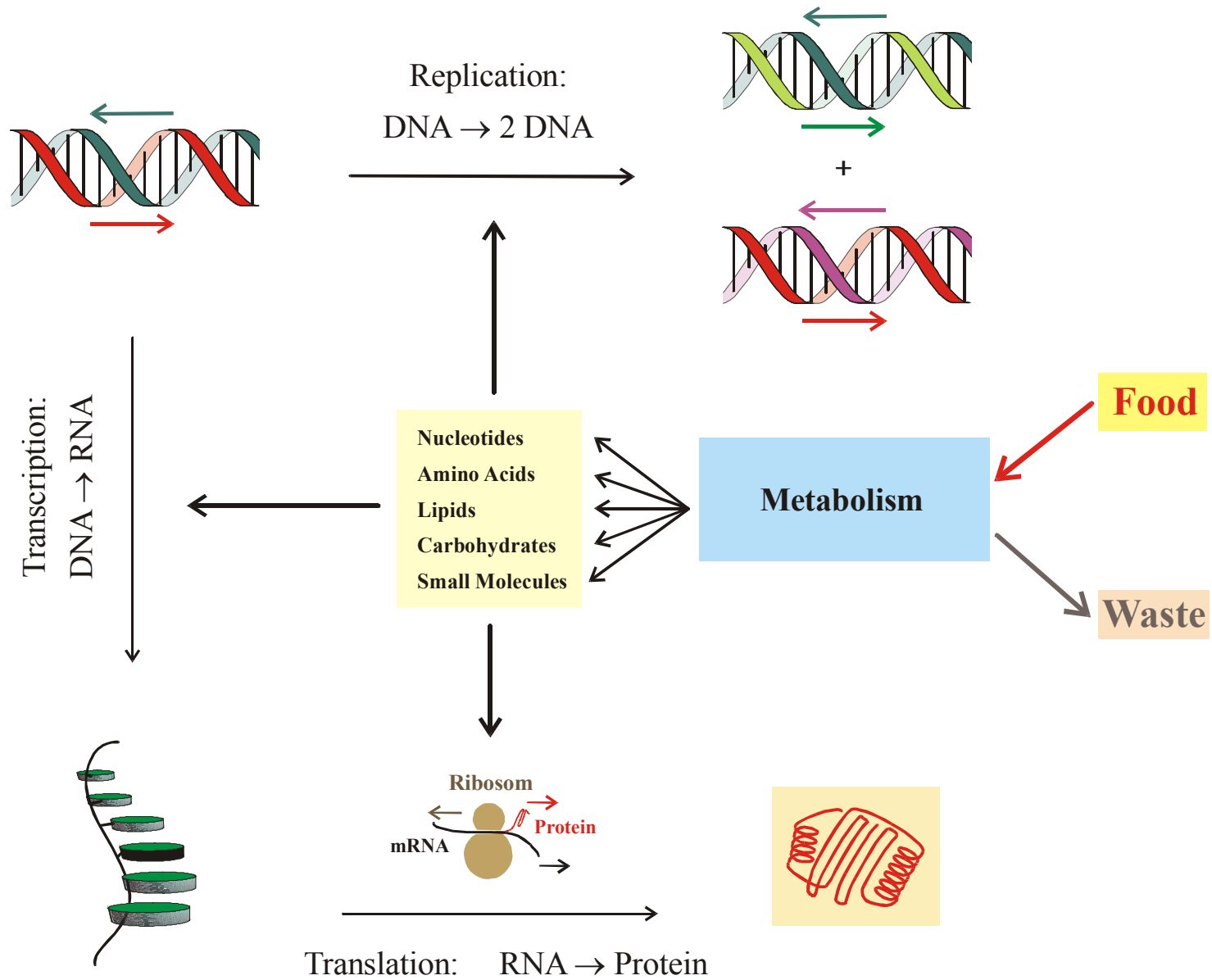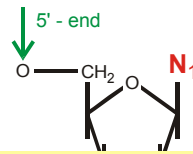# RNA – Das Zaubermolekül

Peter Schuster

Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien

Dies Academicus
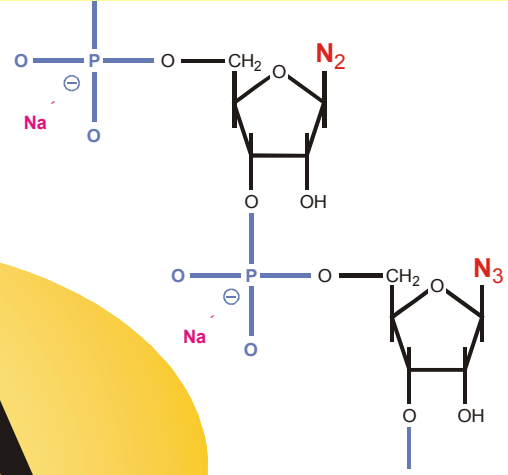
Leipzig, 02.12.2002

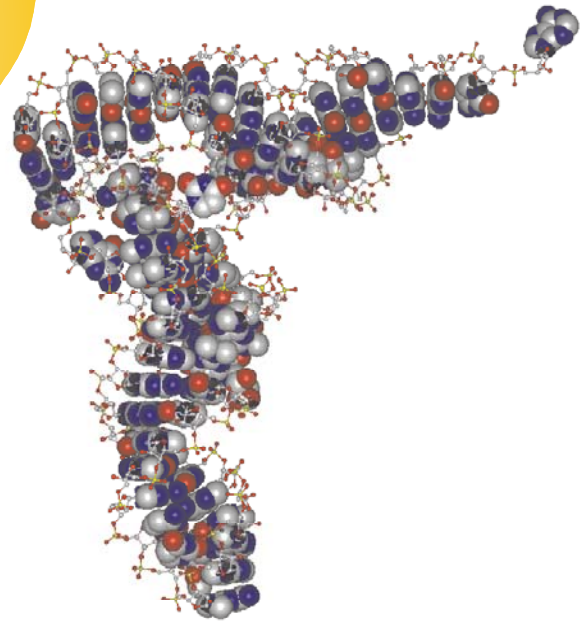Replication:
DNA → 2 DNA

+

Transcription: DNA → RNA

Nucleotides
Amino Acids
Lipids
Carbohydrates
Small Molecules

Metabolism

Food

Waste
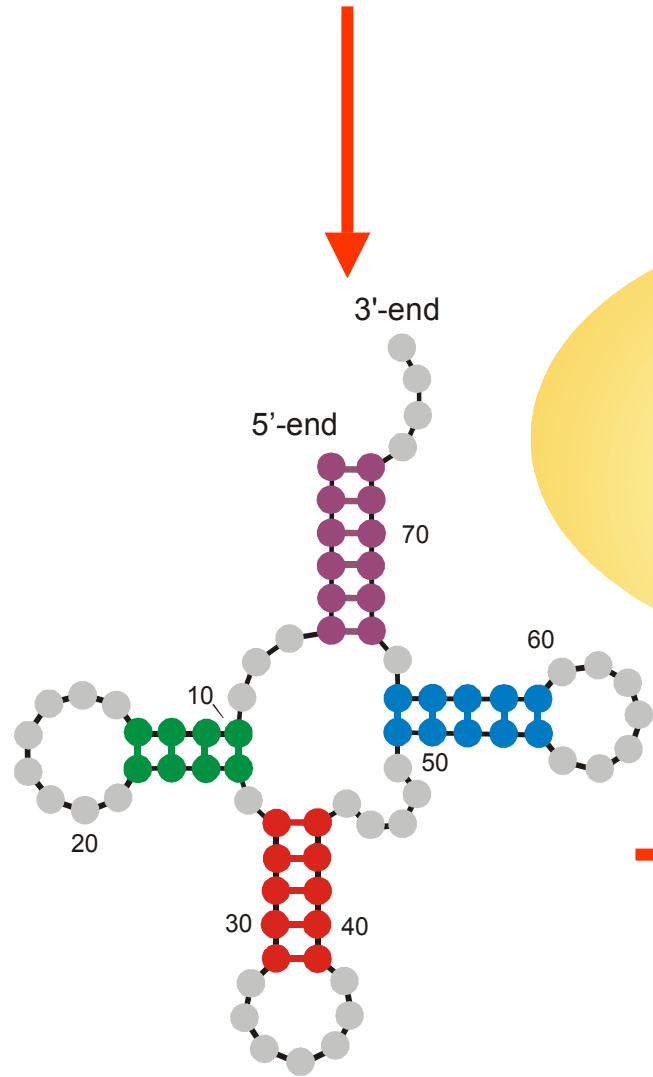
Ribosom

Protein

mRNA

Translation: RNA → Protein

A conventional simplified sketch of cellular metabolism

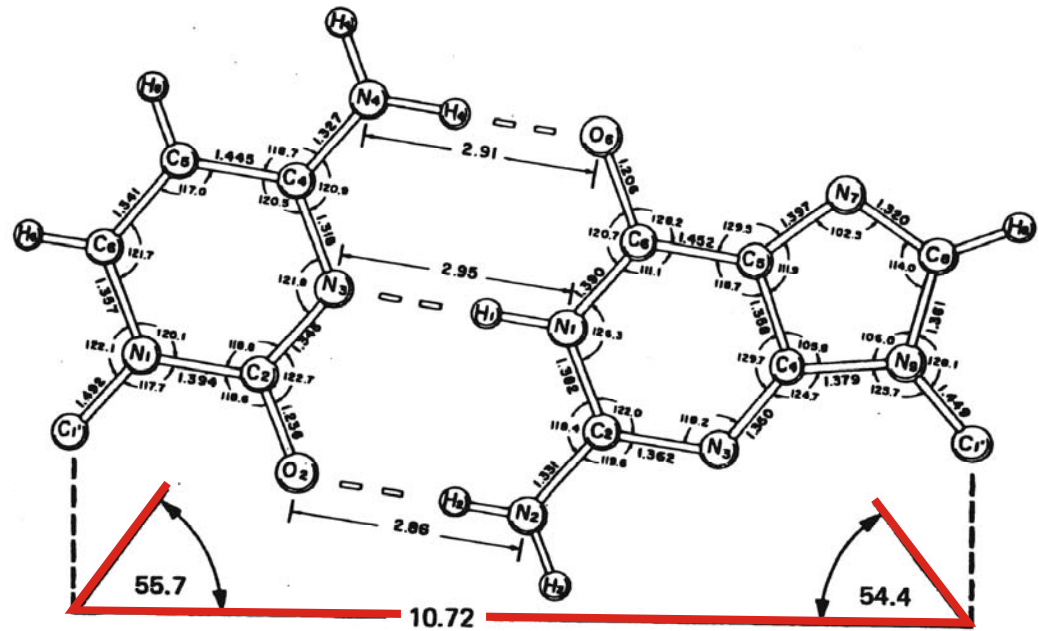5'-end **GCGGAU**UUA**GCUC**AGUUGGGA**GAGC**G**CCAGA**CUGAAGA**UCUGG**AGGUC**CUGUG**UUCGAUC**CACAG**A**AUUCGC**ACCA 3'-end
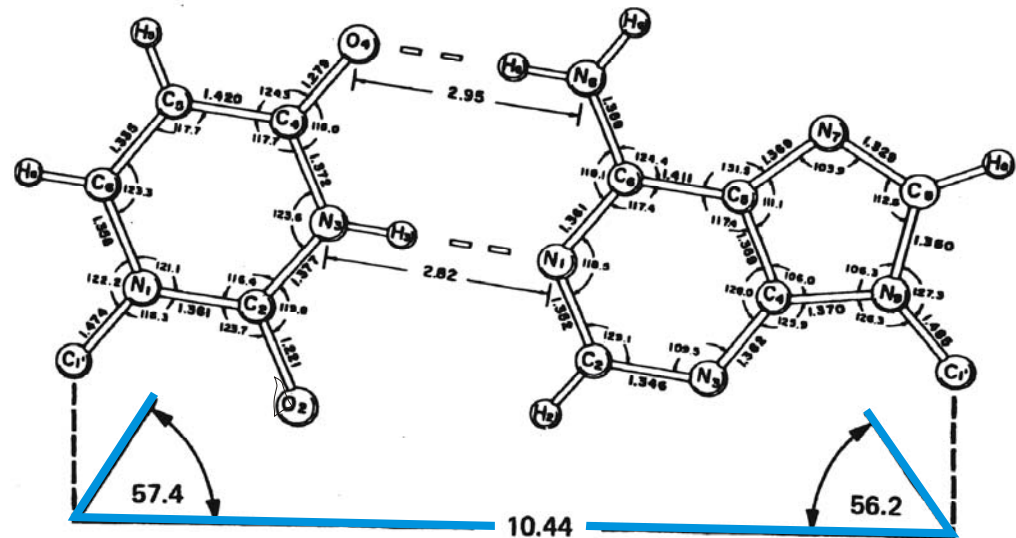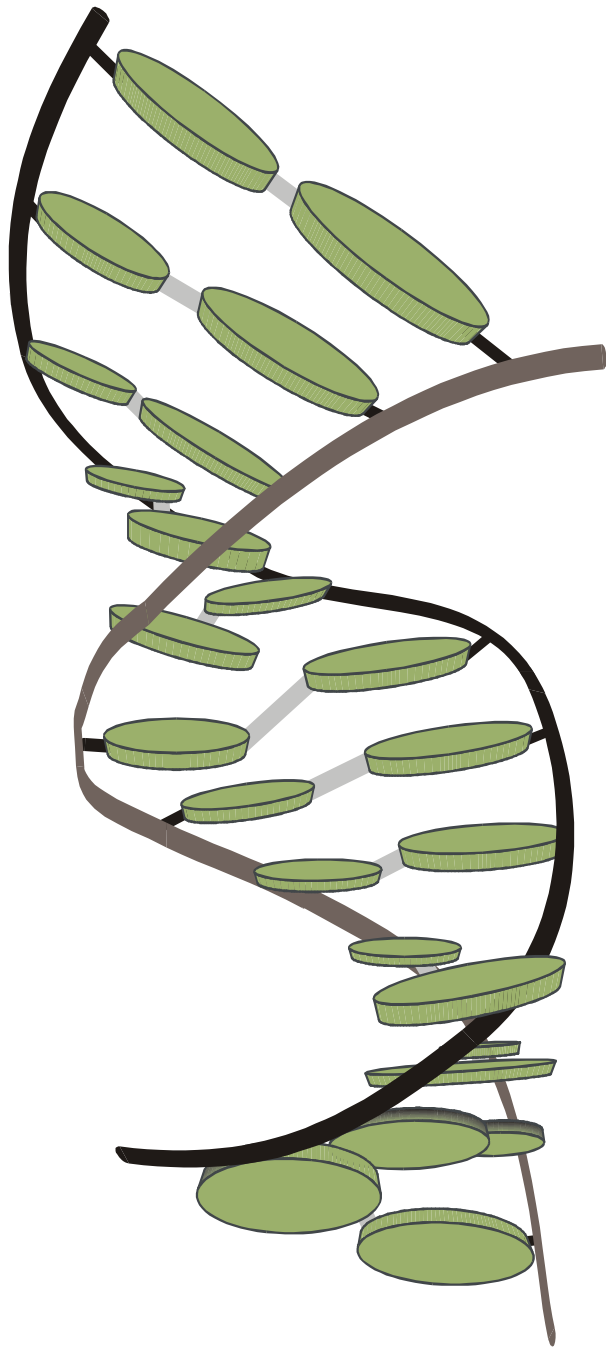
Definition of RNA structure

Canonical or Watson-Crick base pairs:

cytosine – guanine
G©C

uracil – adenine
A=U

W.Saenger, Principles of Nucleic Acid Structure, Springer, Berlin 1984
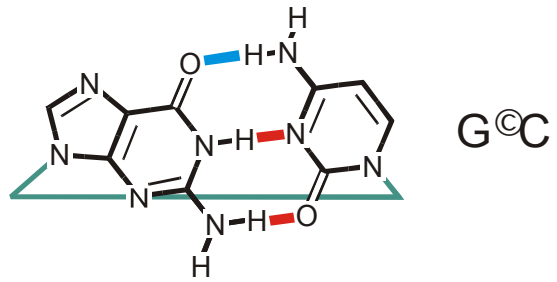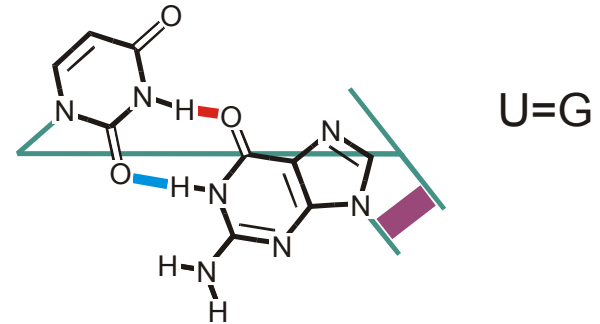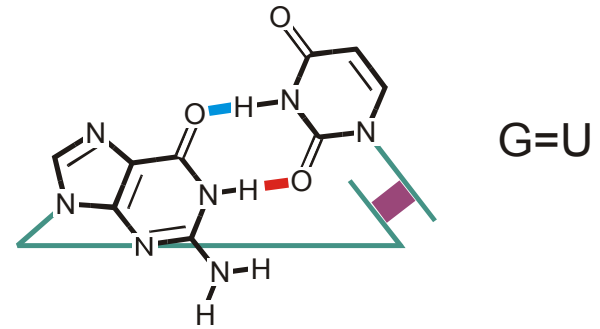
The three-dimensional structure of
a short double helical stack

G©C

Canonical Watson-Crick base-pair

G=U

U=G

Wobble base-pairs

Wobble base pairs in RNA double-helical stacks

RNA *as transmitter of genetic information*

DNA
↓ transcription
...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...
messenger-**RNA**
↓ translation
protein
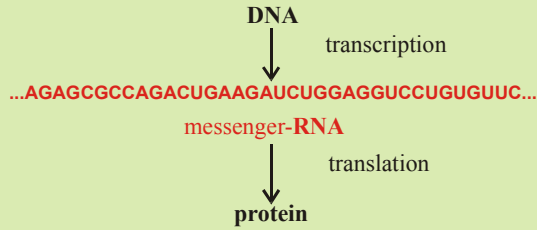
RNA as **working copy** of genetic information

RNA *as adapter molecule*

...CUG... GAC *leu*
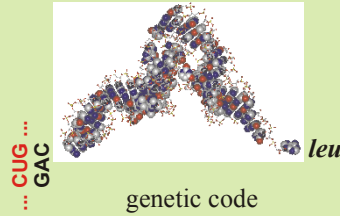
genetic code

RNA *is the catalytic subunit in supramolecular complexes*
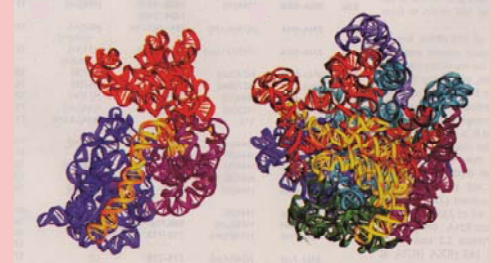
The ribosome is a ribozyme !

RNA *as catalyst*
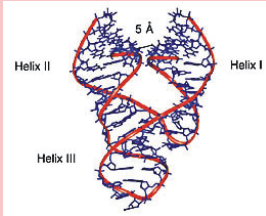
ribozyme

RNA *is modified by epigenetic control*

RNA editing

Alternative splicing of messenger **RNA**

The RNA *world as a precursor of the current* DNA + protein *biology*

RNA *as carrier of genetic information*

RNA viruses and retroviruses

RNA as information carrier in evolution *in vitro* and evolutionary biotechnology

RNA *as regulator of gene expression*
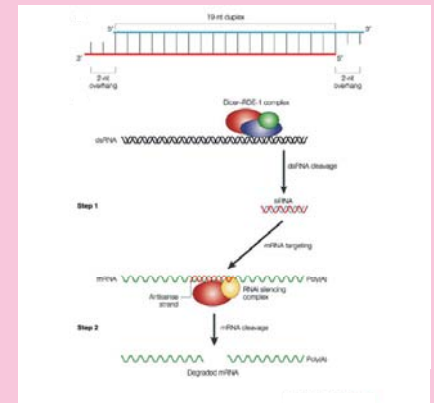
gene silencing by small interfering RNAs

Functions of RNA molecules

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

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

RNA sample

Stock solution: QV RNA-replicase, ATP, CTP, GTP and UTP, buffer

Time

0   1   2   3   4   5   6   69   70

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

Reproduction of the original figure of the serial transfer experiment with Qβ 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
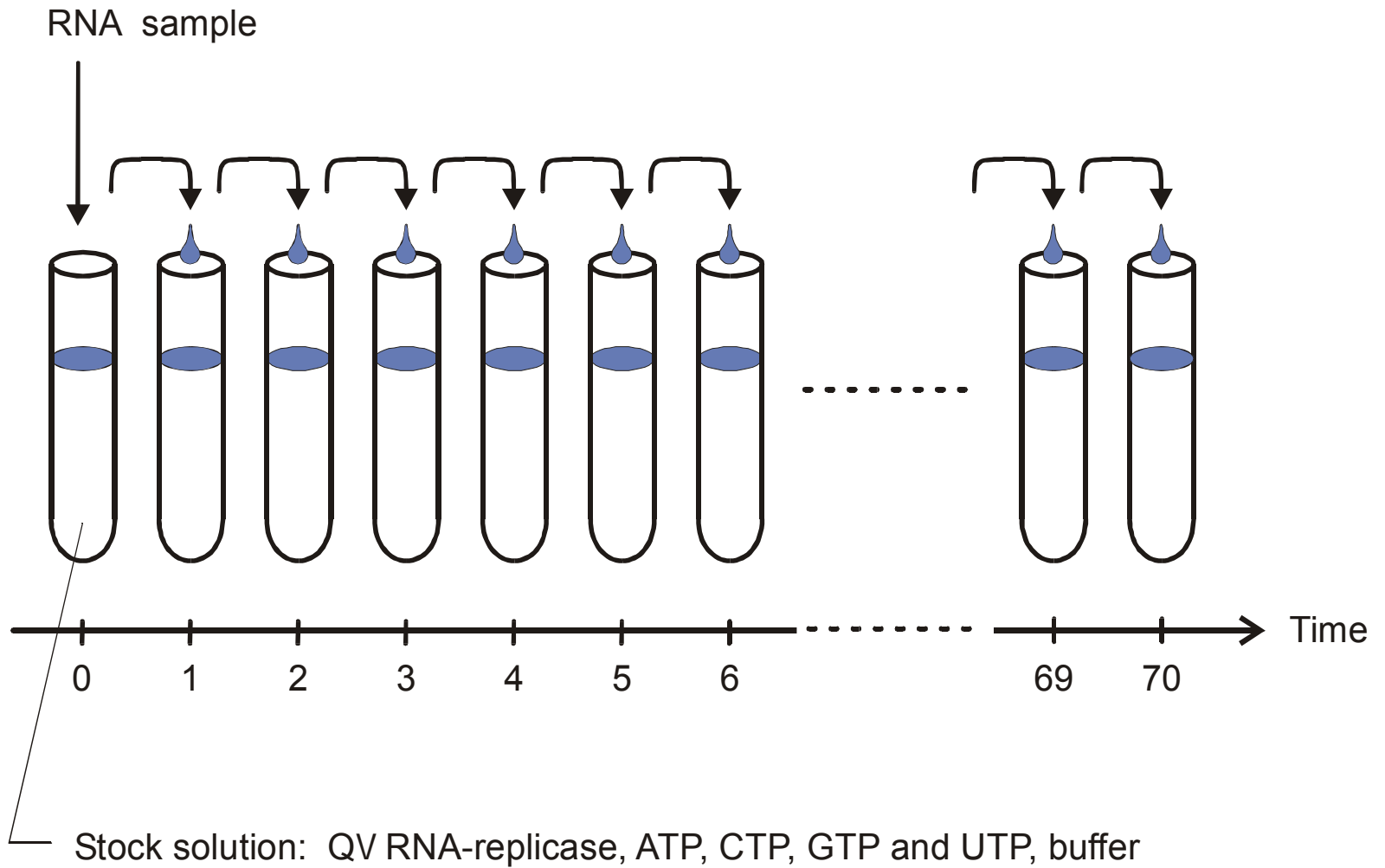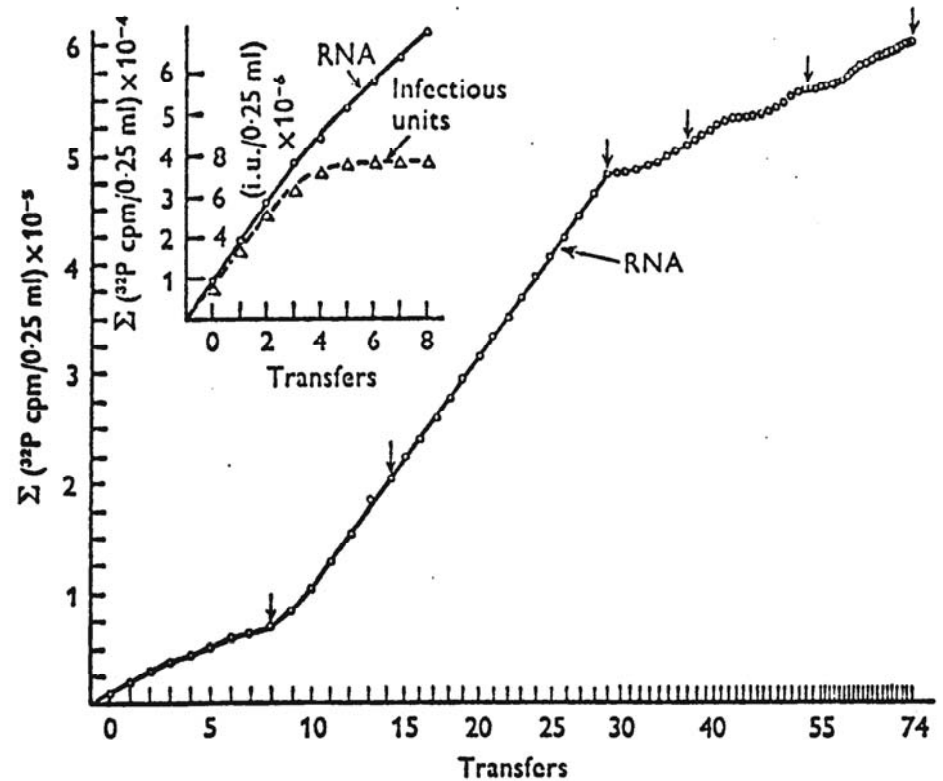
Fig. 9. Serial transfer experiment. Each 0·25 ml standard reaction mixture contained 40 μg of Qβ replicase and ³²P-UTP. The first reaction (0 transfer) was initiated by the addition of 0·2 μg ts-1 (temperature-sensitive RNA) and incubated at 35 °C for 20 min, whereupon 0·02 ml was drawn for counting and 0·02 ml was used to prime the second reaction (first transfer), and so on. After the first 13 reactions, the incubation periods were reduced to 15 min (transfers 14–29). Transfers 30–38 were incubated for 10 min. Transfers 39–52 were incubated for 7 min, and transfers 53–74 were incubated for 5 min. The arrows above certain transfers (0, 8, 14, 29, 37, 53, and 73) indicate where 0·001–0·1 ml of product was removed and used to prime reactions for sedimentation analysis on sucrose. The inset examines both infectious and total RNA. The results show that biologically competent RNA ceases to appear after the 4th transfer (Mills *et al.* 1967).
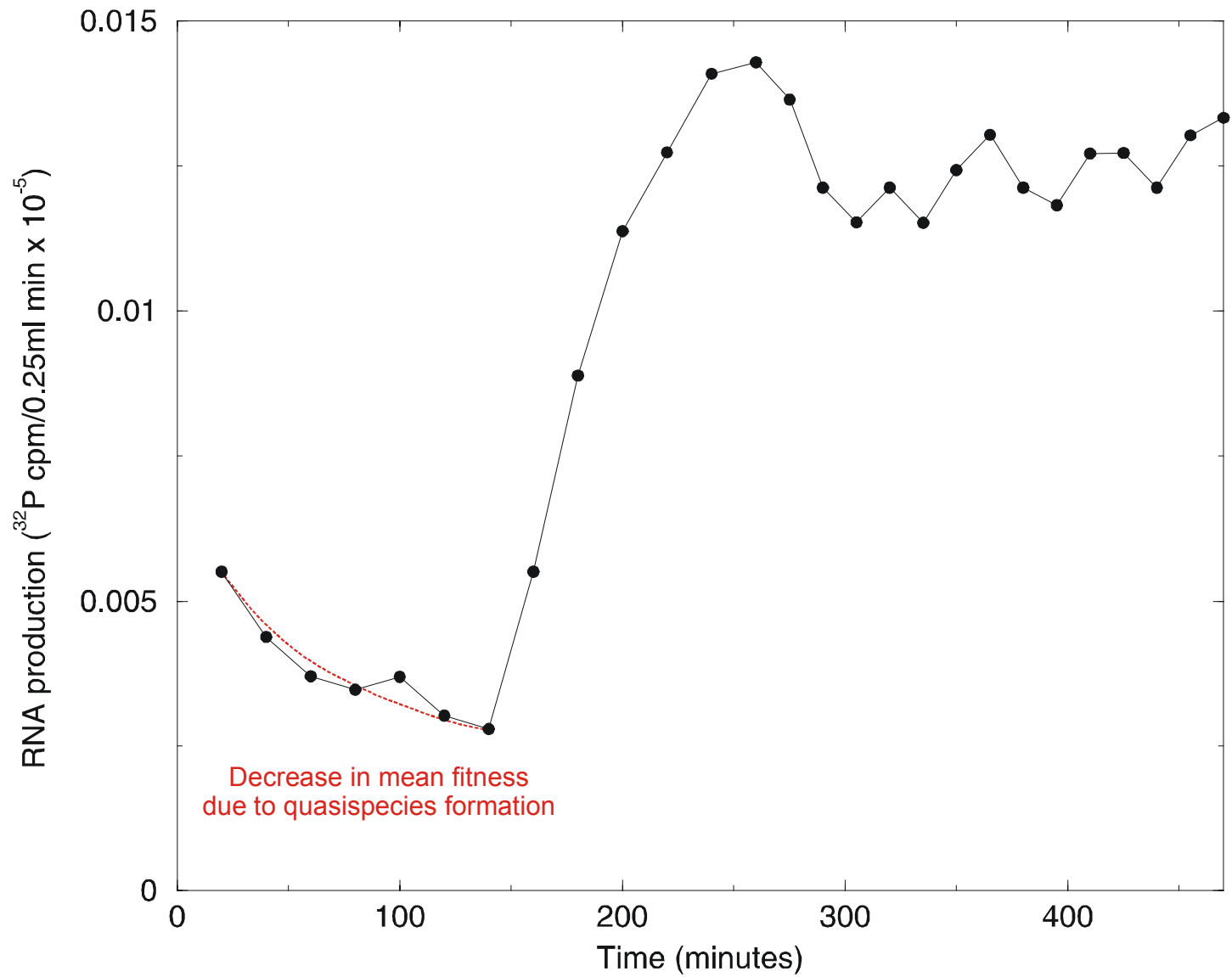
The increase in RNA production rate during a serial transfer experiment

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

Sir Peter Medawar, 1985
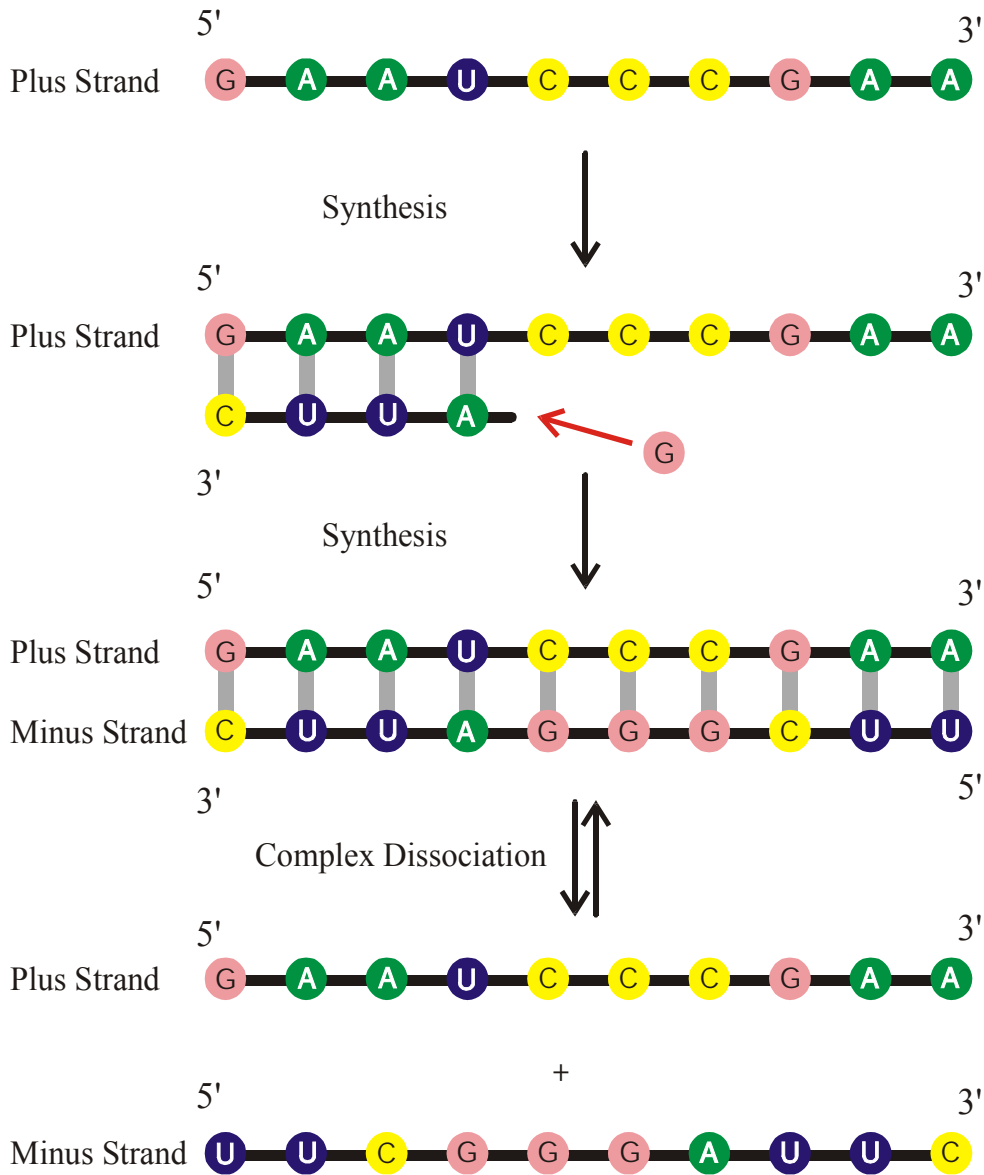
Complementary replication as the simplest copying mechanism of RNA Complementarity is determined by Watson-Crick base pairs:

G©C and A=U
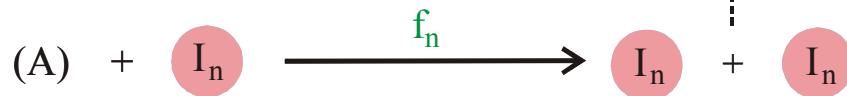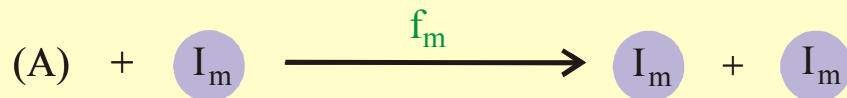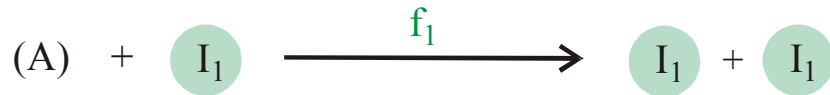
$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

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

$$[I_i] = x_i \not\subset 0 ; \quad i = 1,2,\ldots,n ;$$

$$[A] = a = constant$$

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

$$x_m(t) \, \check{s} \, 1 \; for \; t \, \check{s} \; '$$

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

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

s = 0.1   s = 0.02   s = 0.01

Fraction of advantageous variant

Time [Generations]
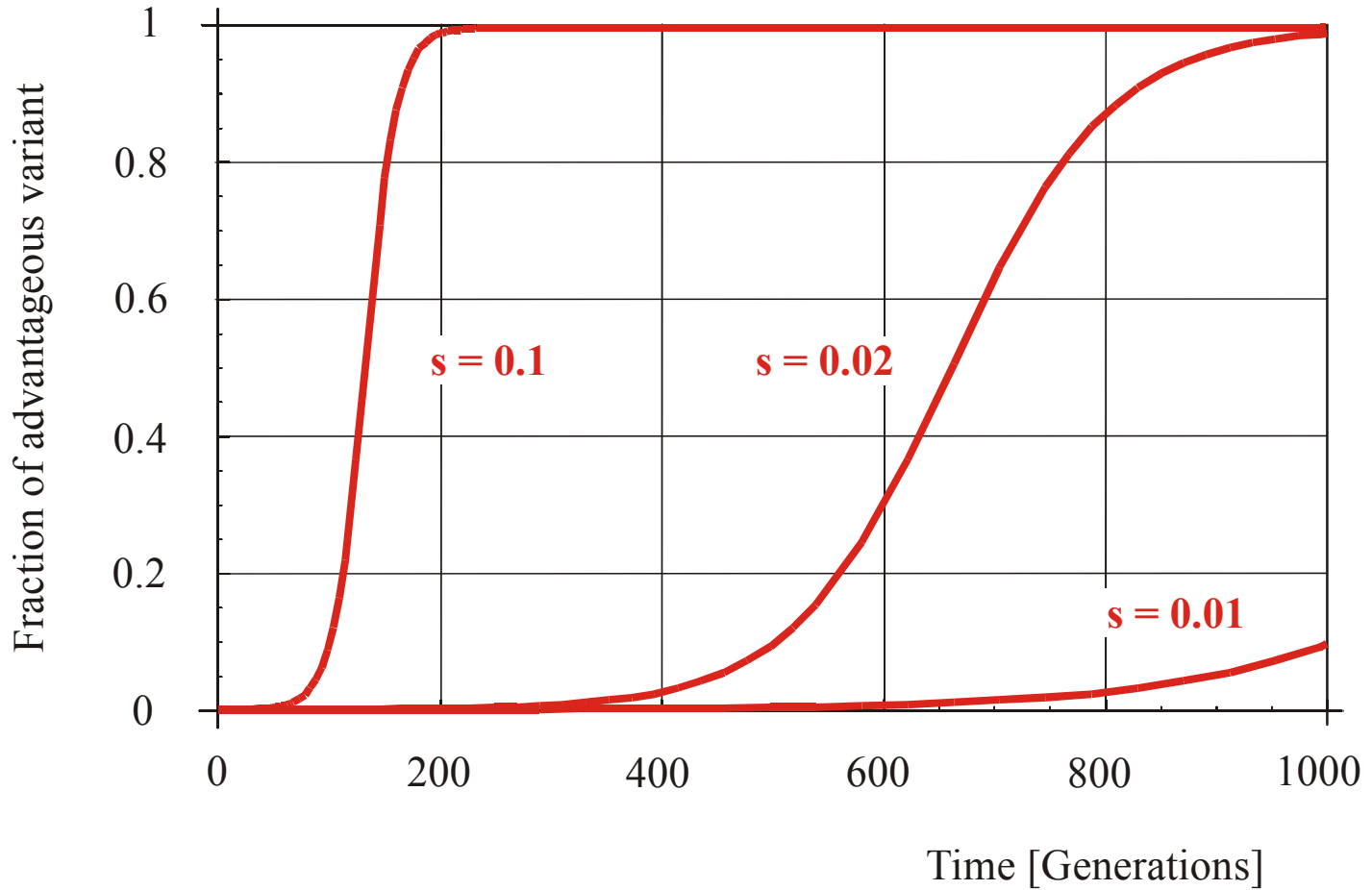
Selection of advantageous mutants in populations of N = 10 000 individuals

Point Mutation

GAA`UCCCG`AA $\longrightarrow$ GAA`UCCCGUCCCG`AA

Insertion

GAAUCC`CGA`A $\longrightarrow$ GAAUCCA
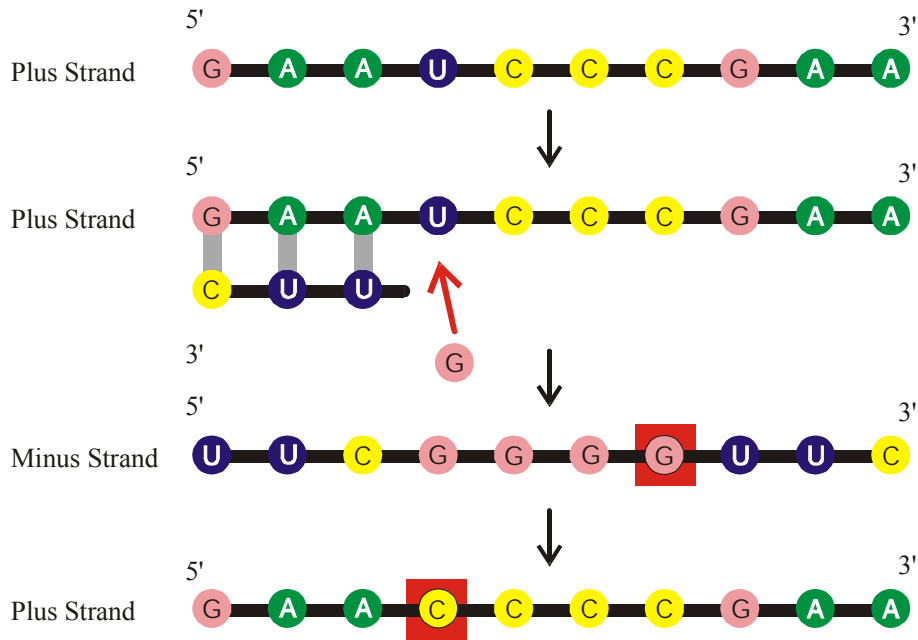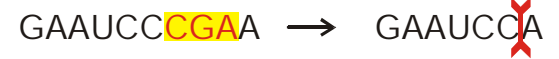
Deletion

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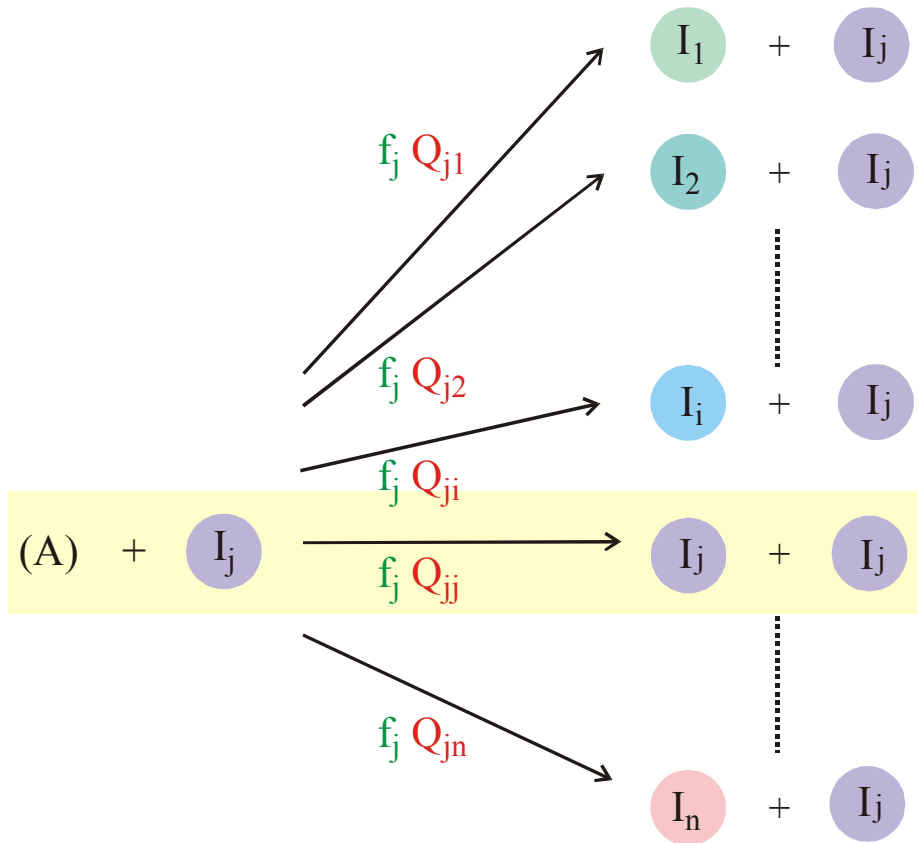
$$\frac{dx_i}{dt} = \sum_j f_j Q_{ji} x_j - x_i \Phi$$

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

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

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

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

p .... Error rate per digit

ℓ .... Chain length of the polynucleotide

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

Chemical kinetics of replication and mutation as parallel reactions

Quasispecies →← Uniform distribution →

$\bar{y}_i$ 0.5

1.0

I_0

q_{min}

$\Sigma I_{(1)}$
$\Sigma I_{(2)}$
$\Sigma I_{(3)}$
$\Sigma I_{(4)}$

$\Sigma I_{(25)}$
$\Sigma I_{(24)}, \Sigma I_{(26)}$
$\Sigma I_{(23)}, \Sigma I_{(27)}$
$\Sigma I_{(22)}, \Sigma I_{(28)}$
$\Sigma I_{(21)}, \Sigma I_{(29)}$
$\Sigma I_{(20)}, \Sigma I_{(30)}$

0

1.00   0.95   0.90

← q

0.00   0.05   0.10

Error rate $p = 1-q$ →
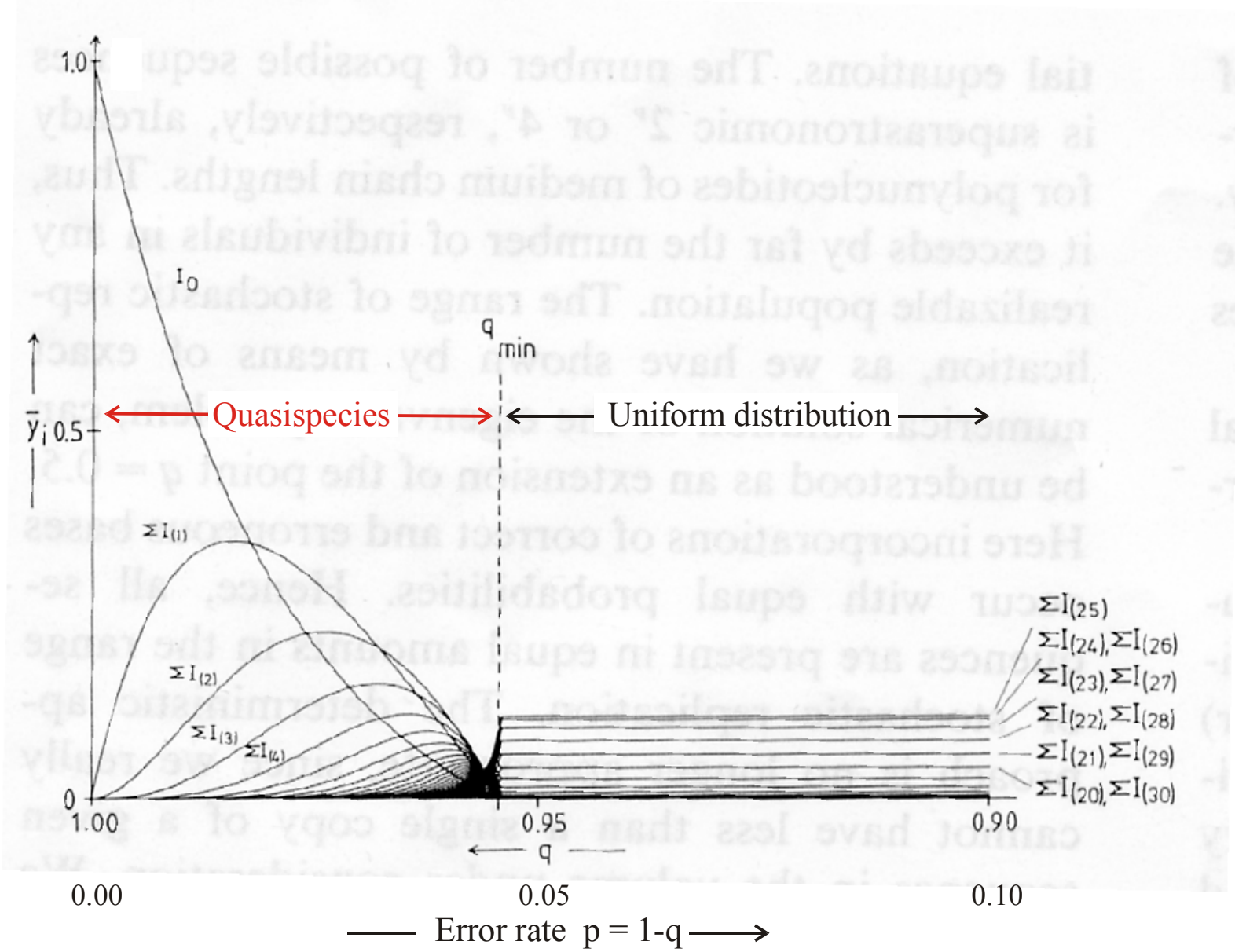
Quasispecies as a function of the replication accuracy q

Master sequence

Mutant cloud

Concentration
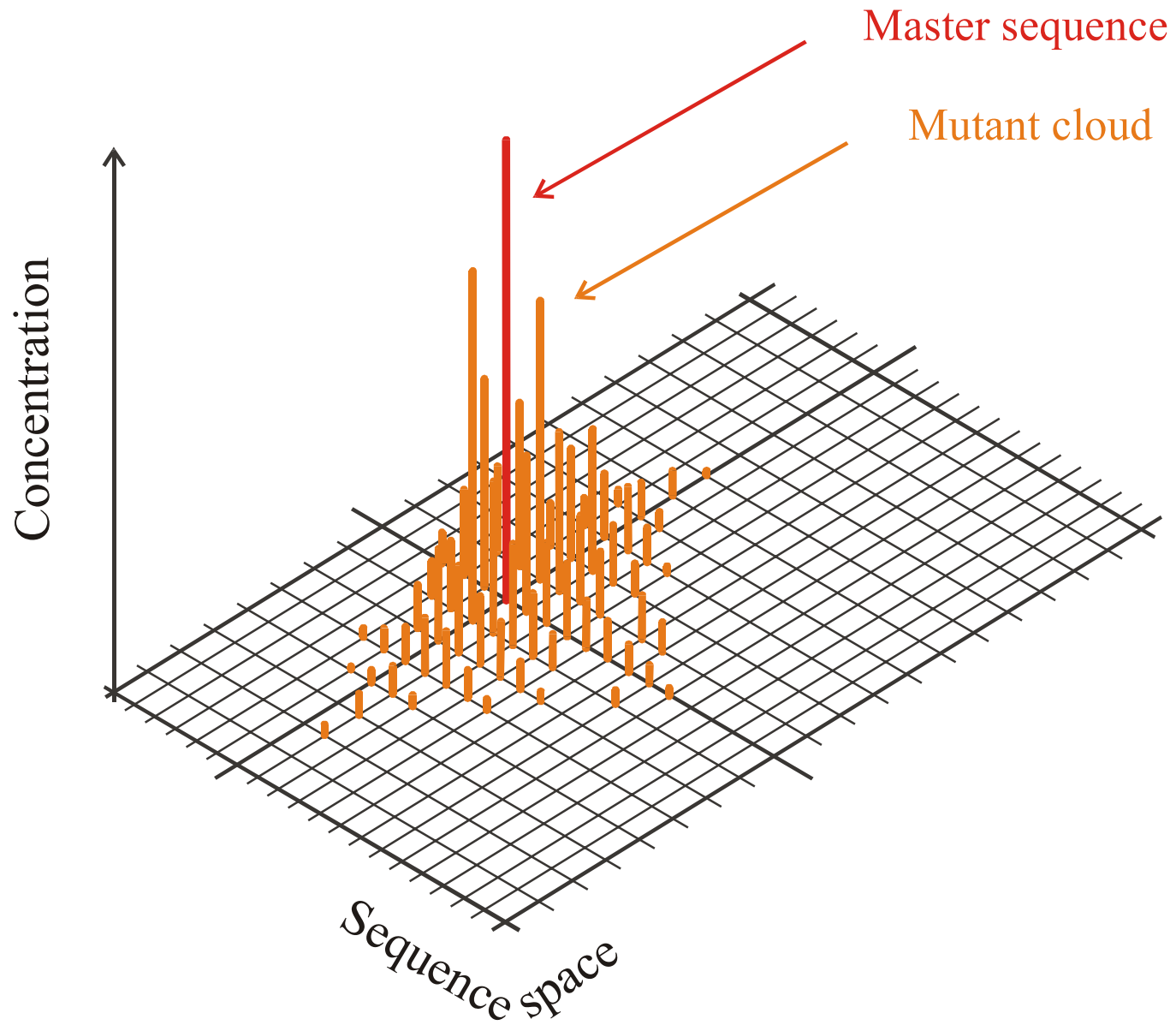
Sequence space

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 were the mean fitness decreases monotonously or were it may show non-monotonous behavior. The volume of the part of the simplex where mean fitness is non-decreasing in the conventional sense decreases with inreasing mutation rate p.

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

Theory of genotype – phenotype mapping

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

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

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, *Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering*. Mh.Chem. **127** (1996), 375-389
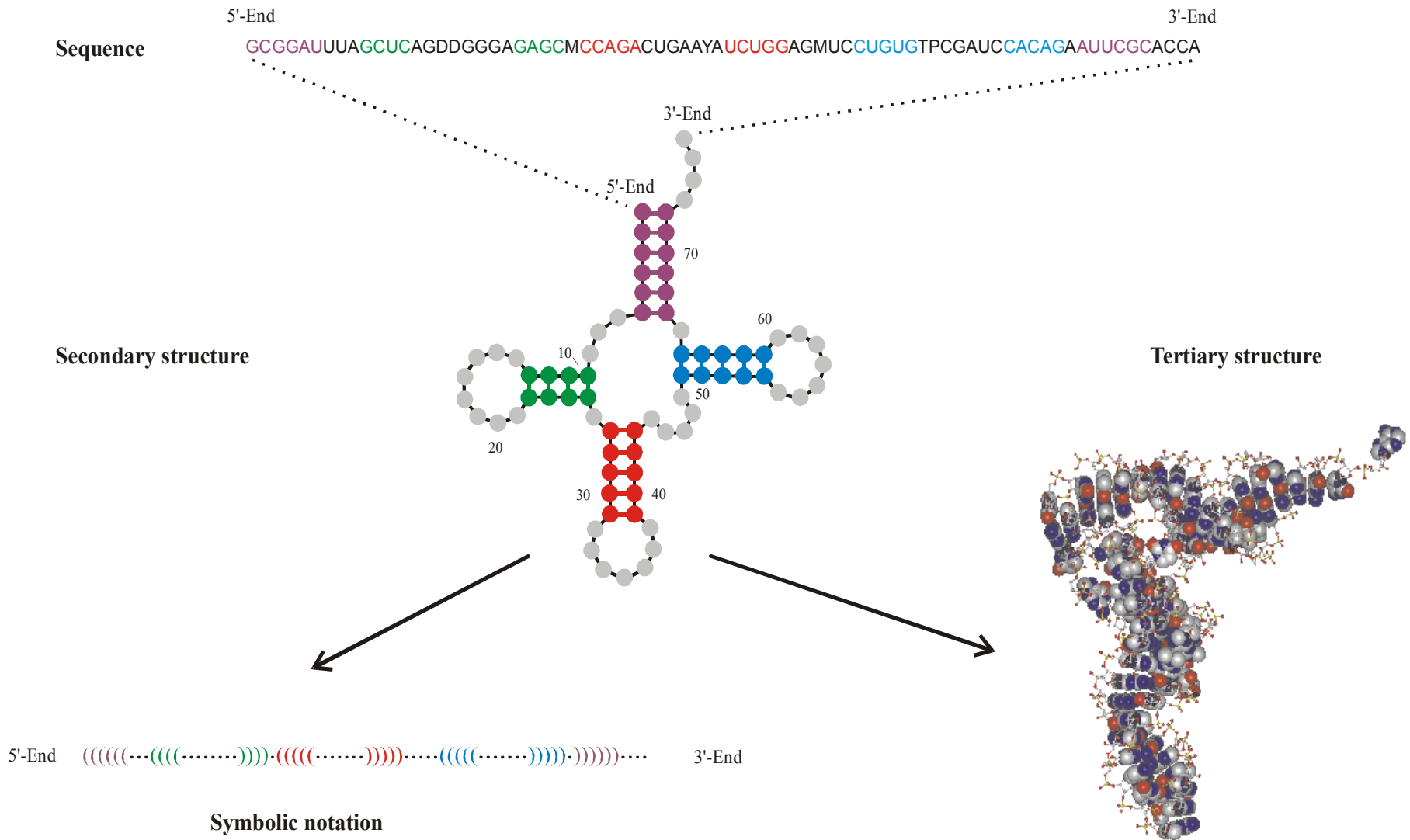
C.M.Reidys, P.F.Stadler, P.Schuster, *Generic properties of combinatory maps*. Bull.Math.Biol. **59** (1997), 339-397

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

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

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

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

Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

Secondary structure

Tertiary structure

Symbolic notation

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

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)
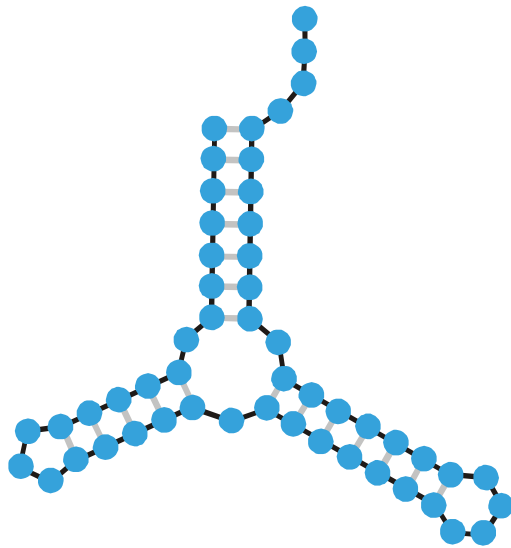
Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG

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

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

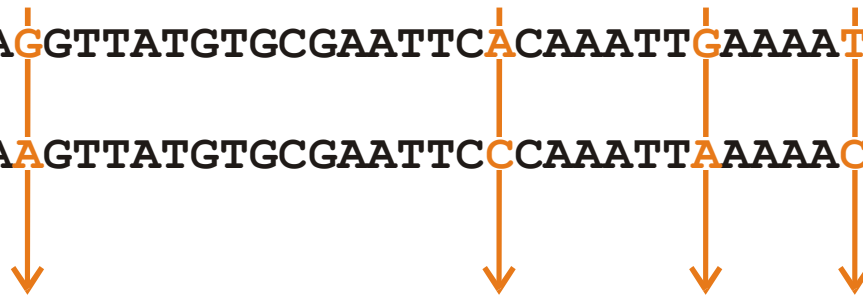GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG
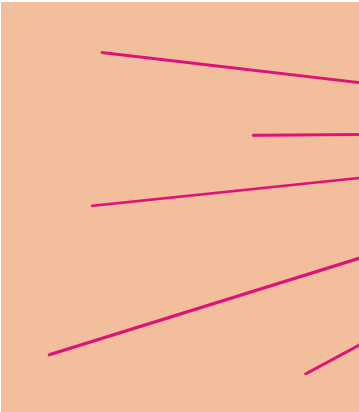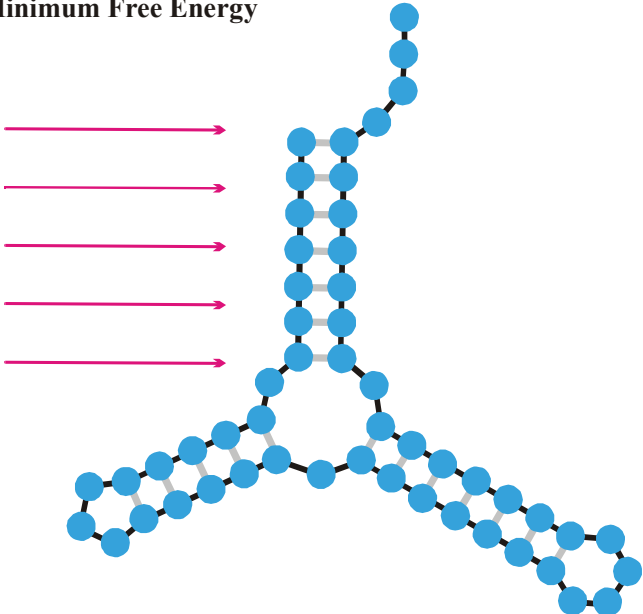
Sequence Space
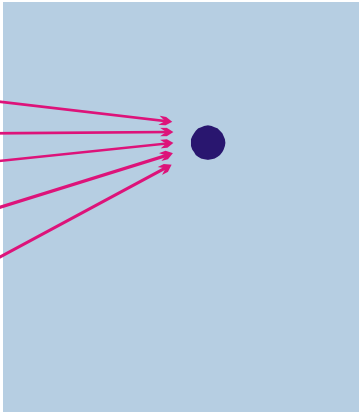
Shape Space

S₁: CGTCGTTACAATTTA**G**GTTATGTGCGAATTC**A**CAAATT**G**AAAA**T**ACAAGAG.....

S₂: 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) $d_H(S_1,S_3) < d_H(S_1,S_2) + d_H(S_2,S_3)$

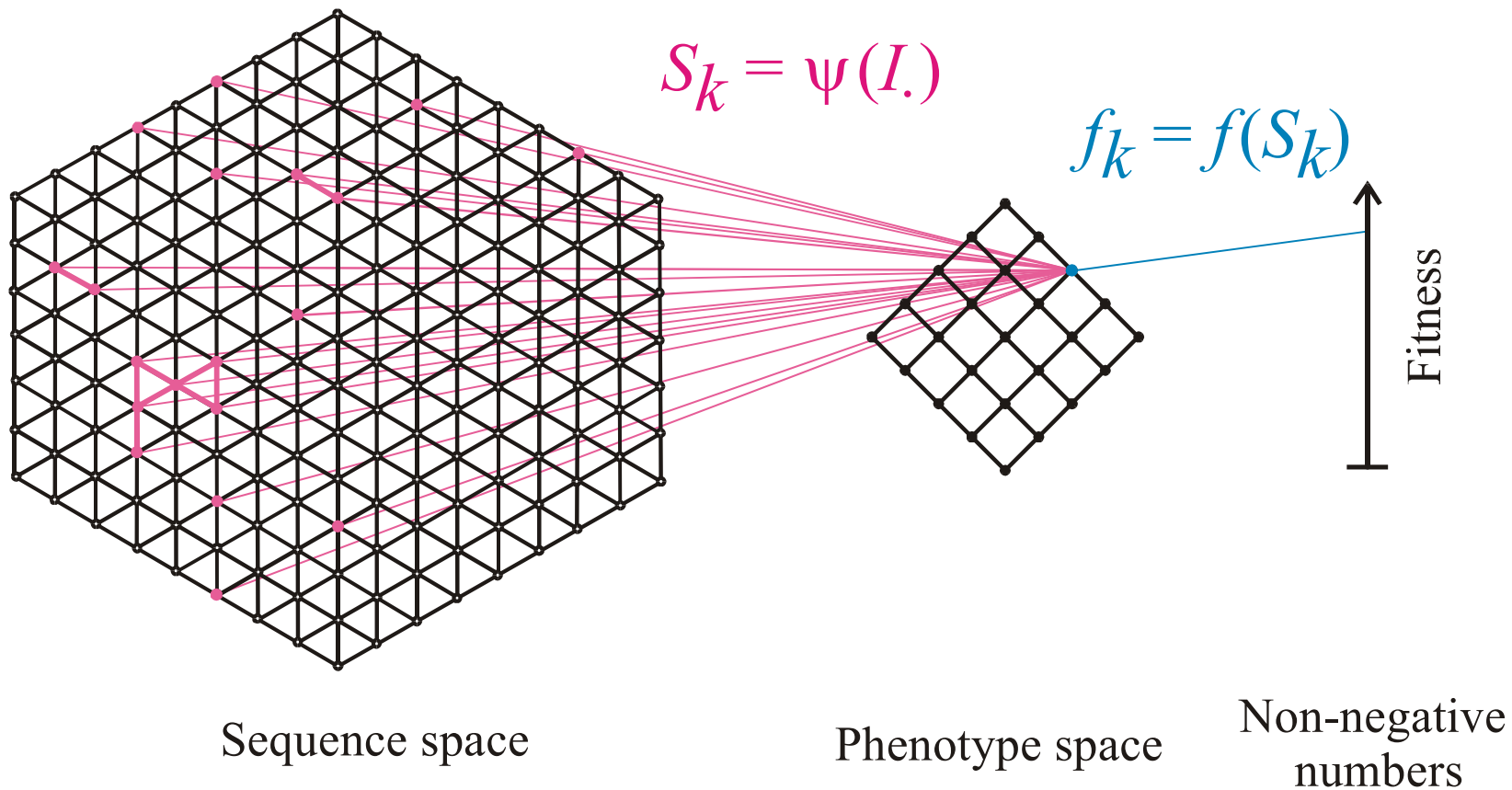The Hamming distance induces a metric in sequence space

The **RNA model** considers RNA sequences as genotypes and simplified RNA structures, called secondary structures, as phenotypes.
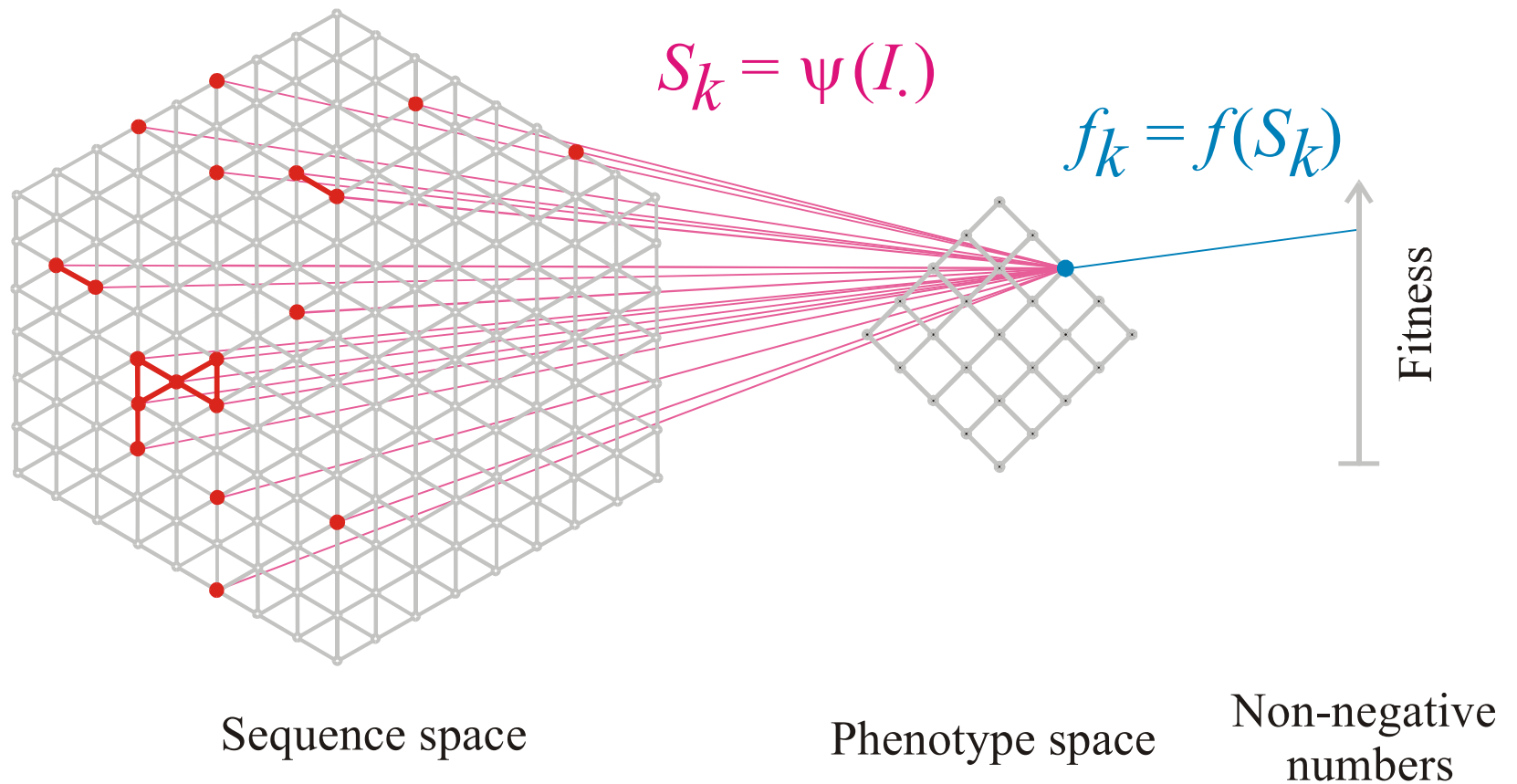
The **mapping** from genotypes into phenotypes is many-to-one. Hence, it is redundant and not invertible.

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.

$$S_k = \psi(I.)$$

$$f_k = f(S_k)$$

Fitness

Sequence space

Phenotype space

Non-negative numbers
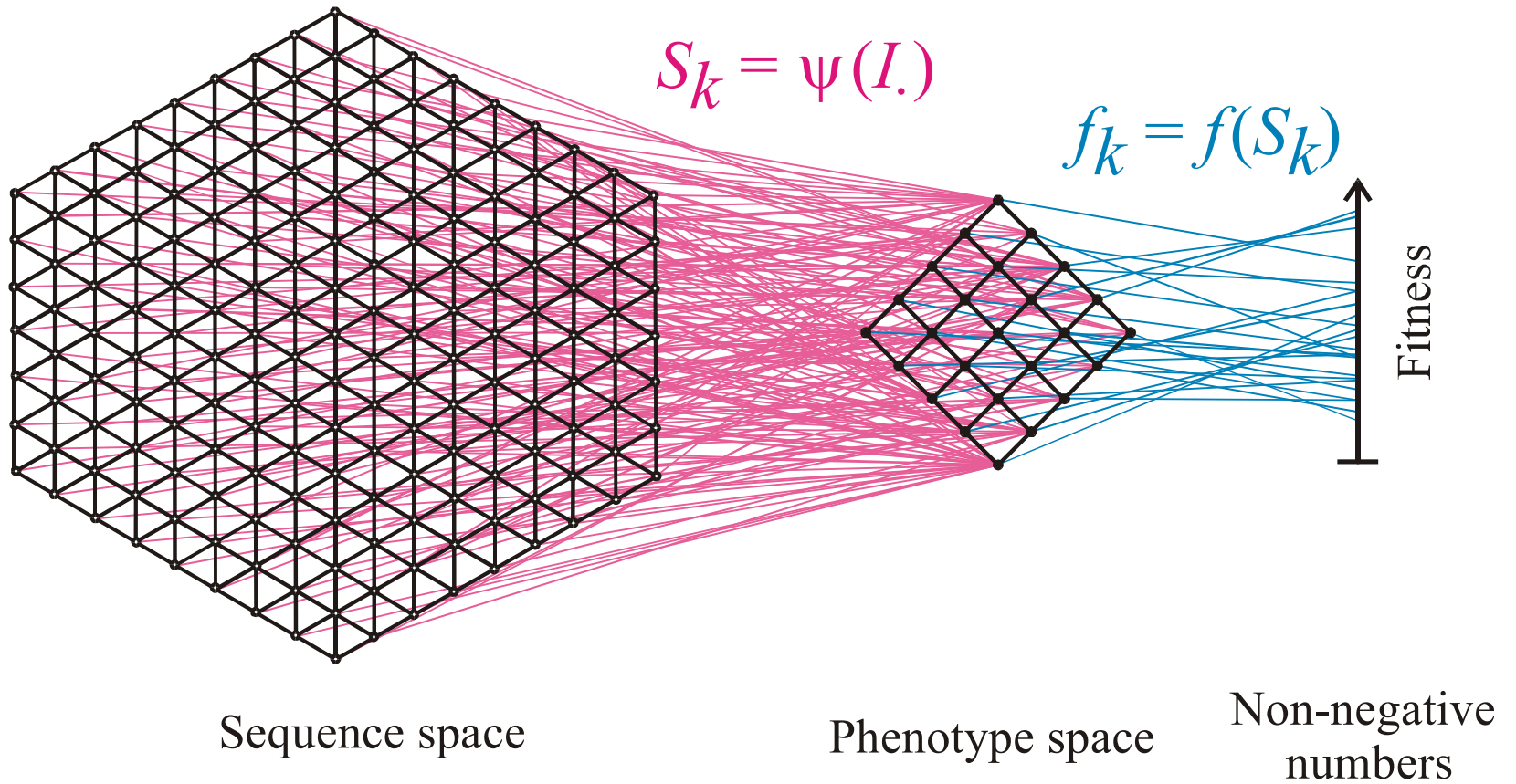
Mapping from sequence space into phenotype space and into fitness values

$S_k = \psi(I.)$

$f_k = f(S_k)$

Fitness

Sequence space

Phenotype space

Non-negative numbers

$S_k = \psi(I.)$

$f_k = f(S_k)$

Fitness

Sequence space   Phenotype space   Non-negative numbers

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

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

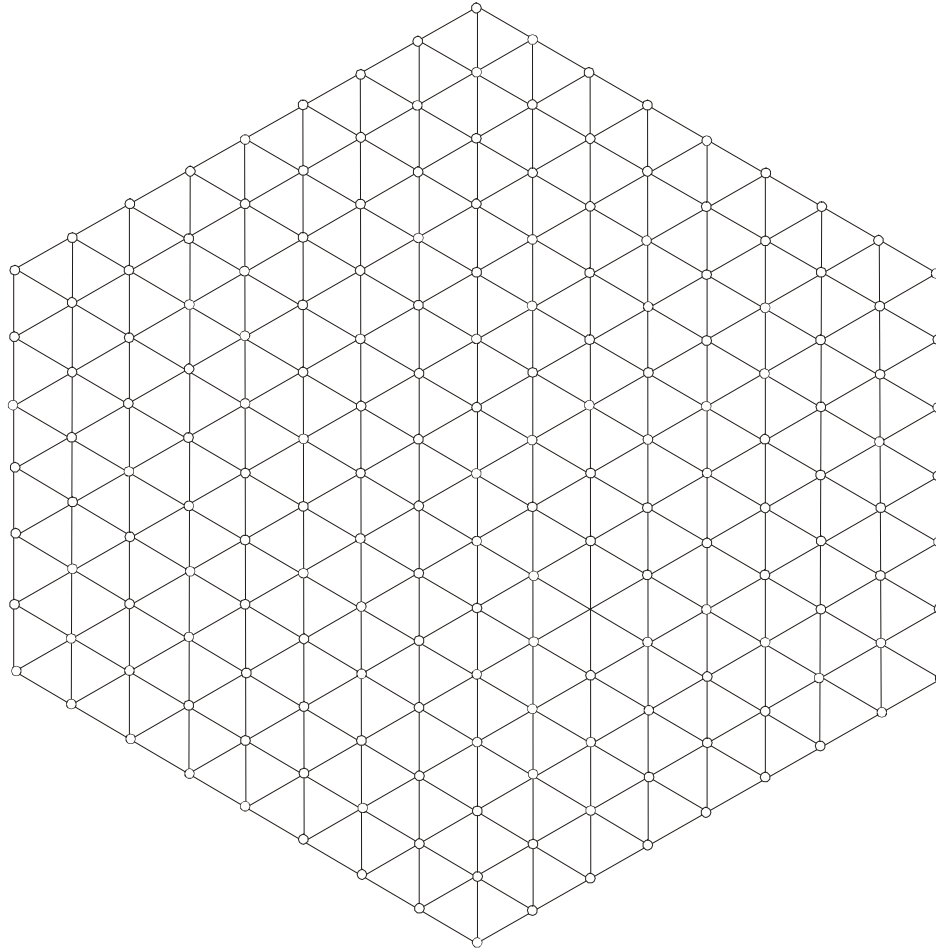$$G_k = m^{-1}(S_k) \quad \{m_j \mid m(I_j) = S_k\}$$

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.
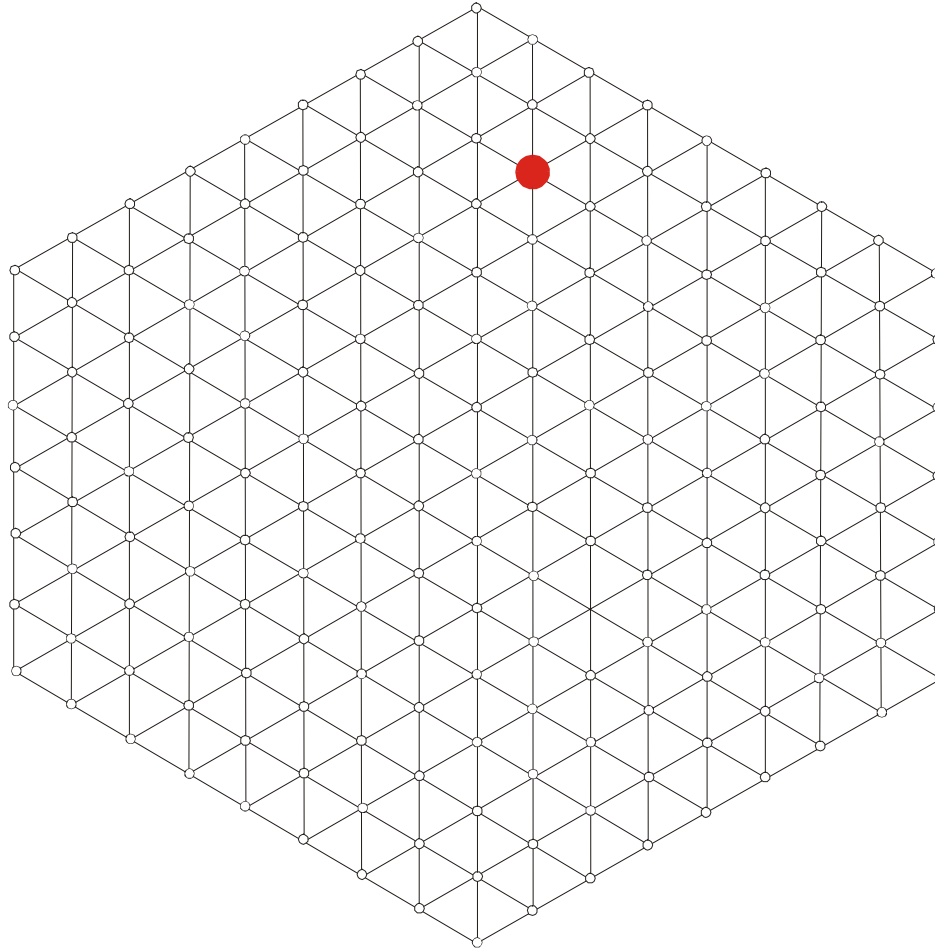
Sketch of sequence space
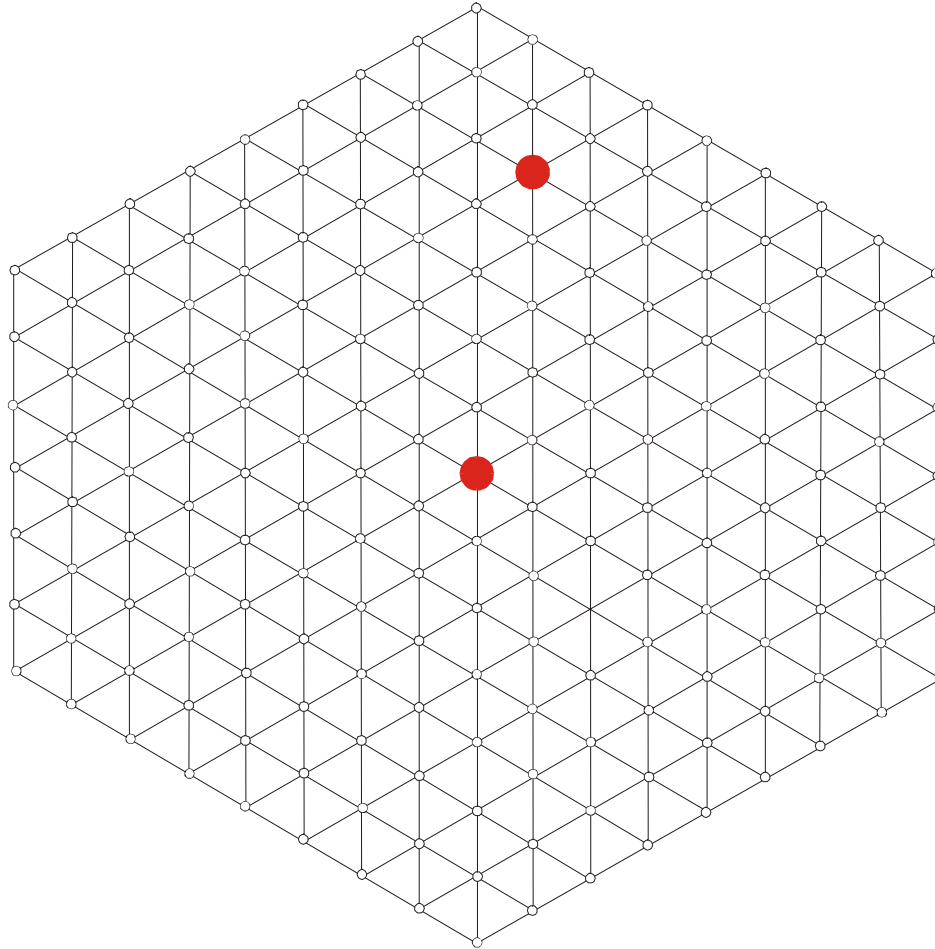

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks

Random graph approach to neutral networks
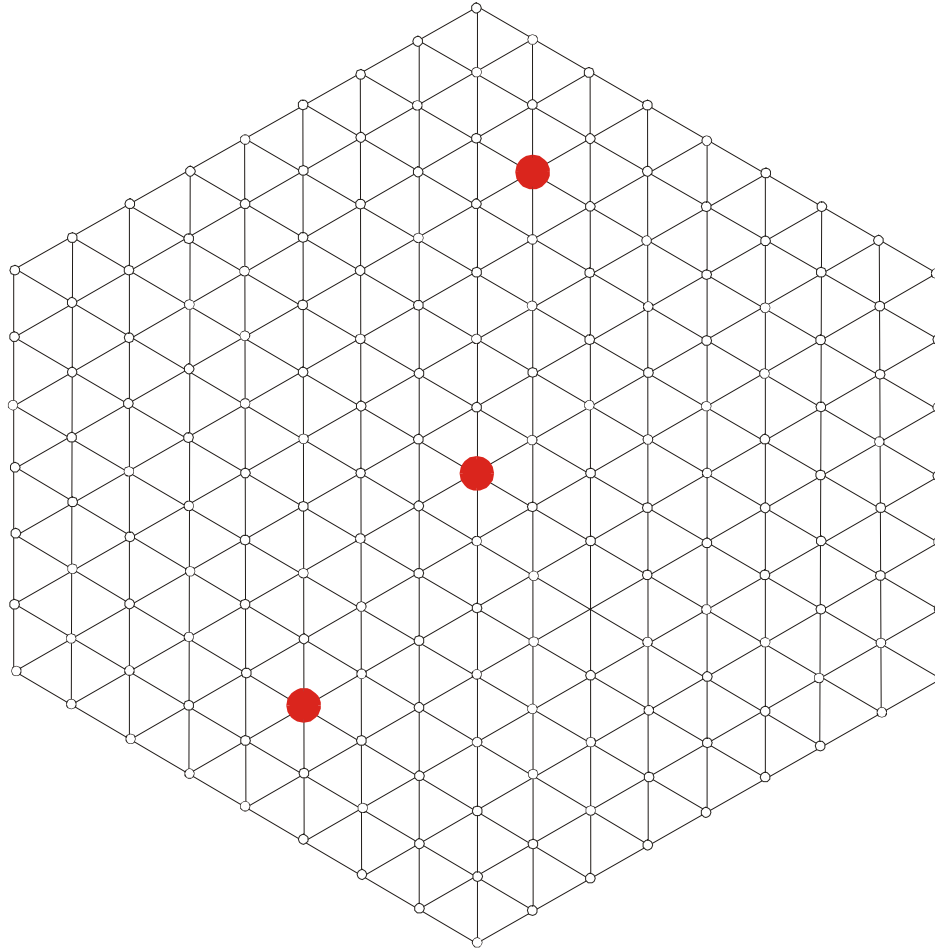
Sketch of sequence space

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks
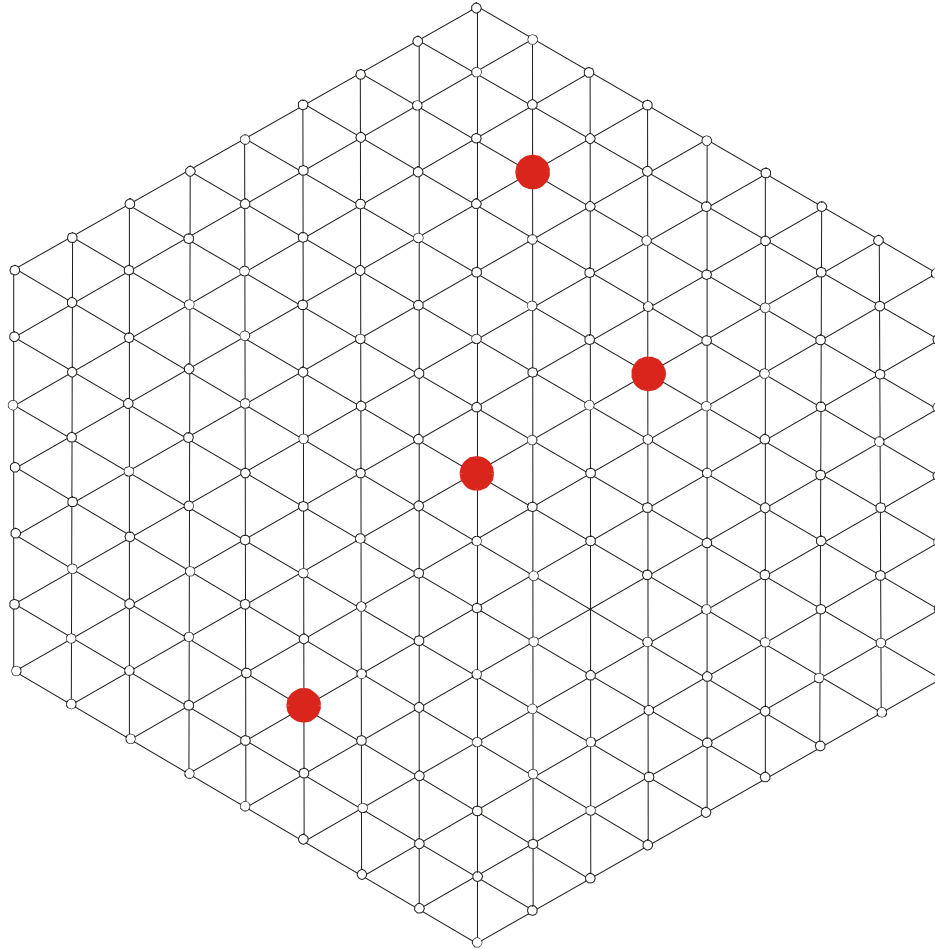
Sketch of sequence space

Random graph approach to neutral networks
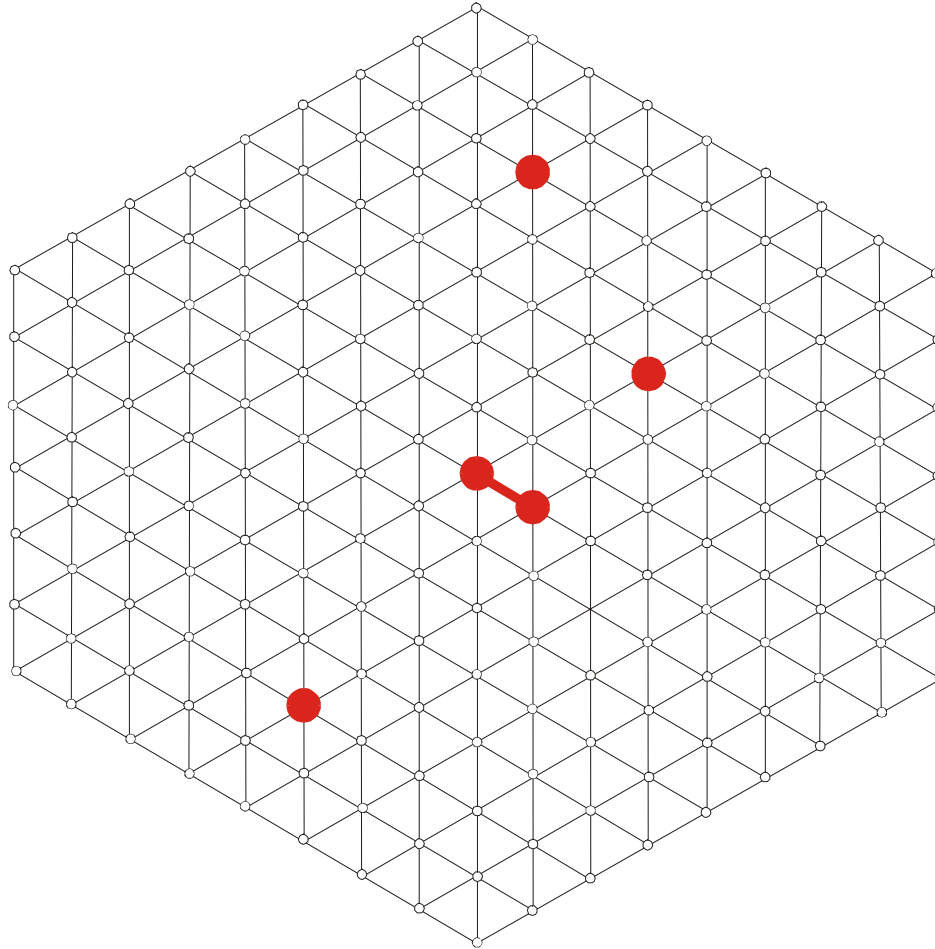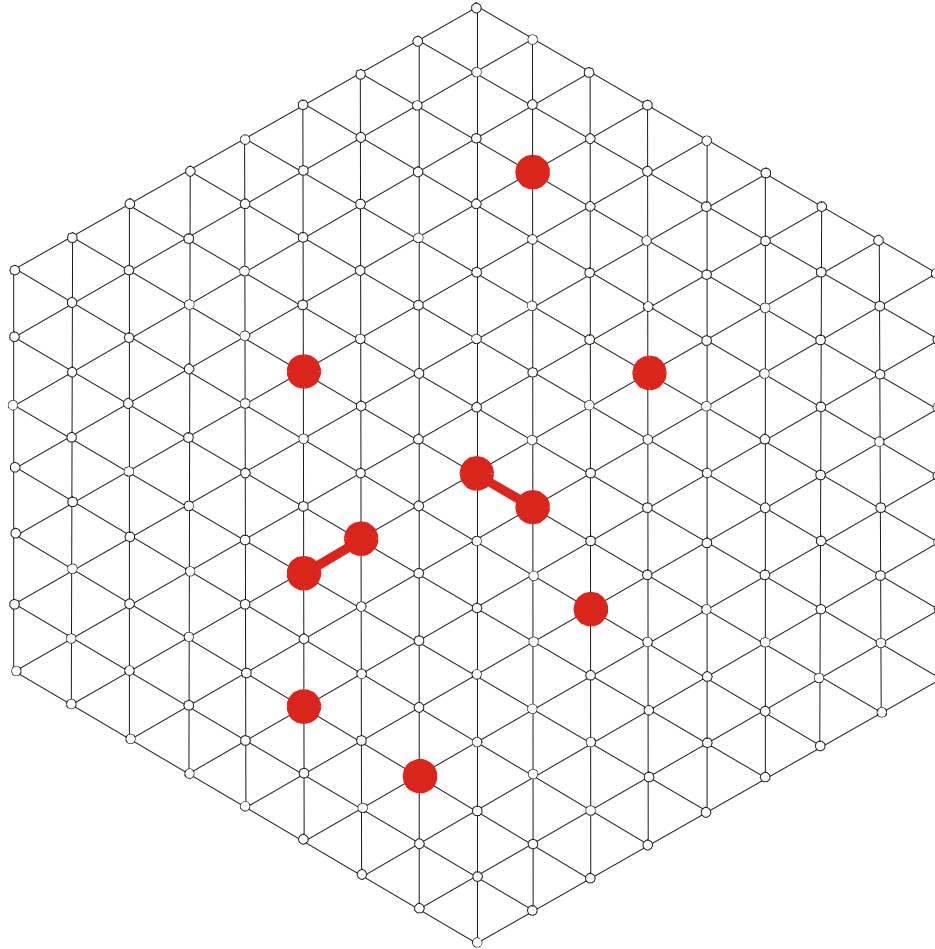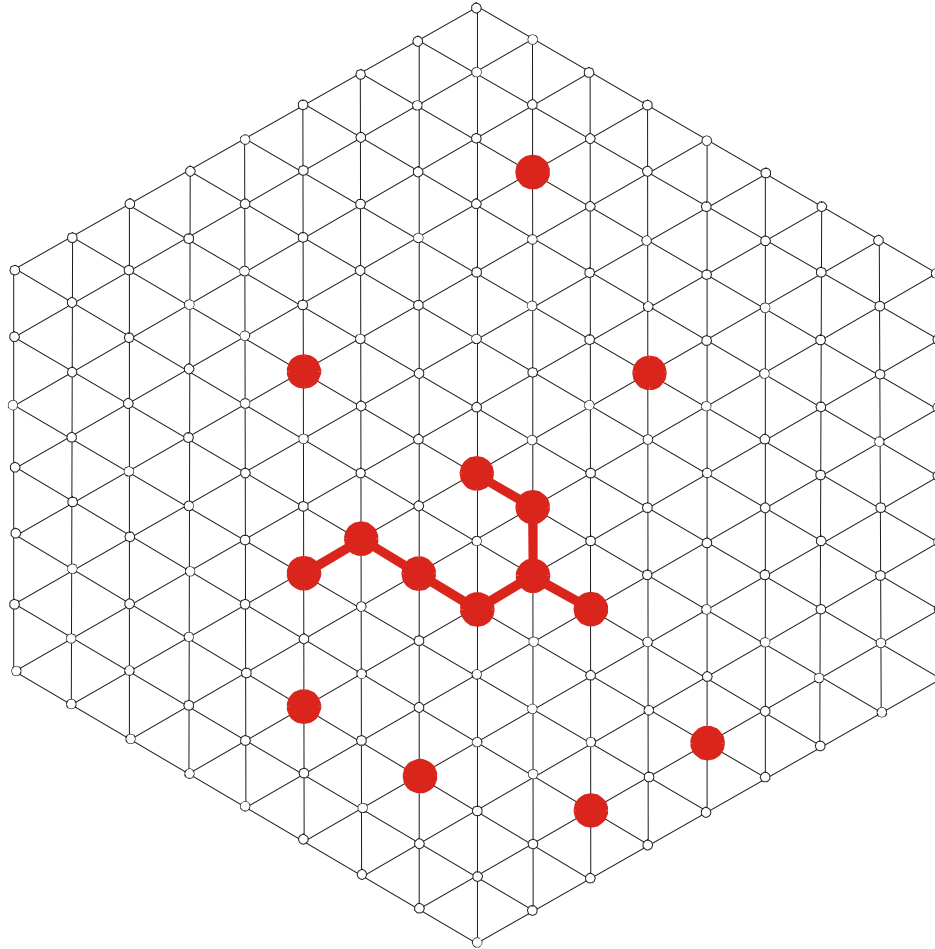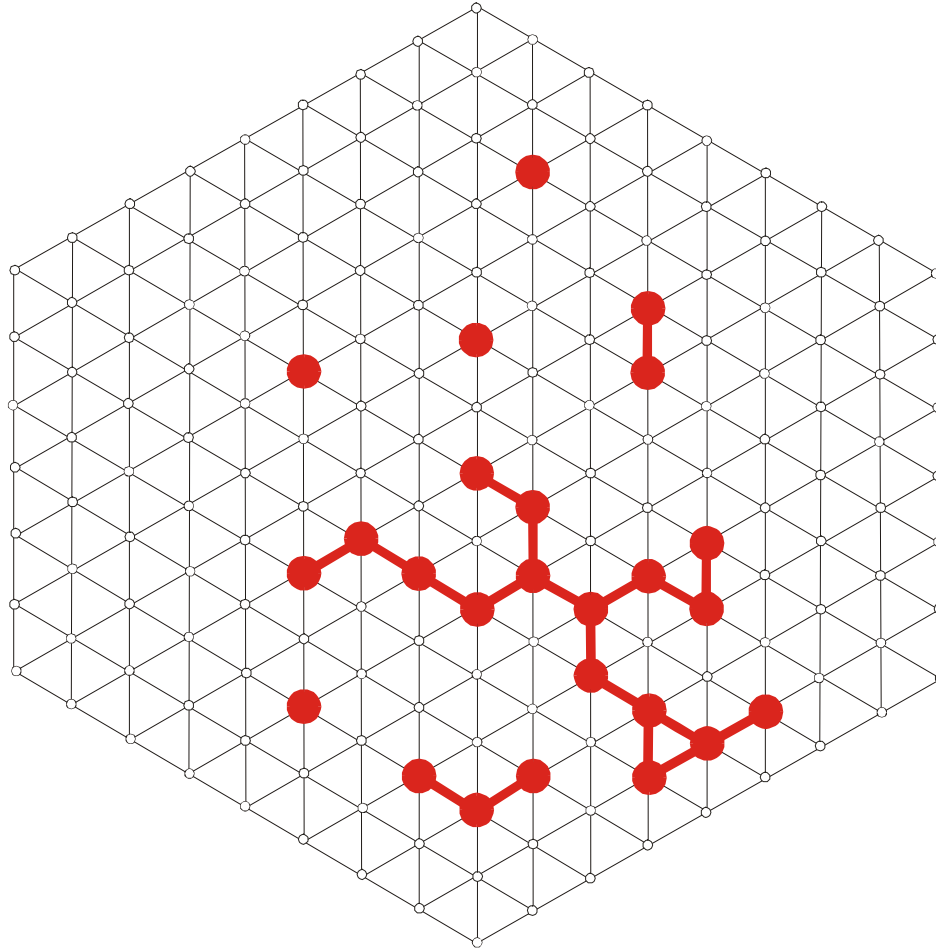
Sketch of sequence space

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks

Random graph approach to neutral networks

Sketch of sequence space

Random graph approach to neutral networks

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

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

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

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

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

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

κ	λ_{cr}
2	0.5
3	0.4226
4	0.3700

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

Giant Component
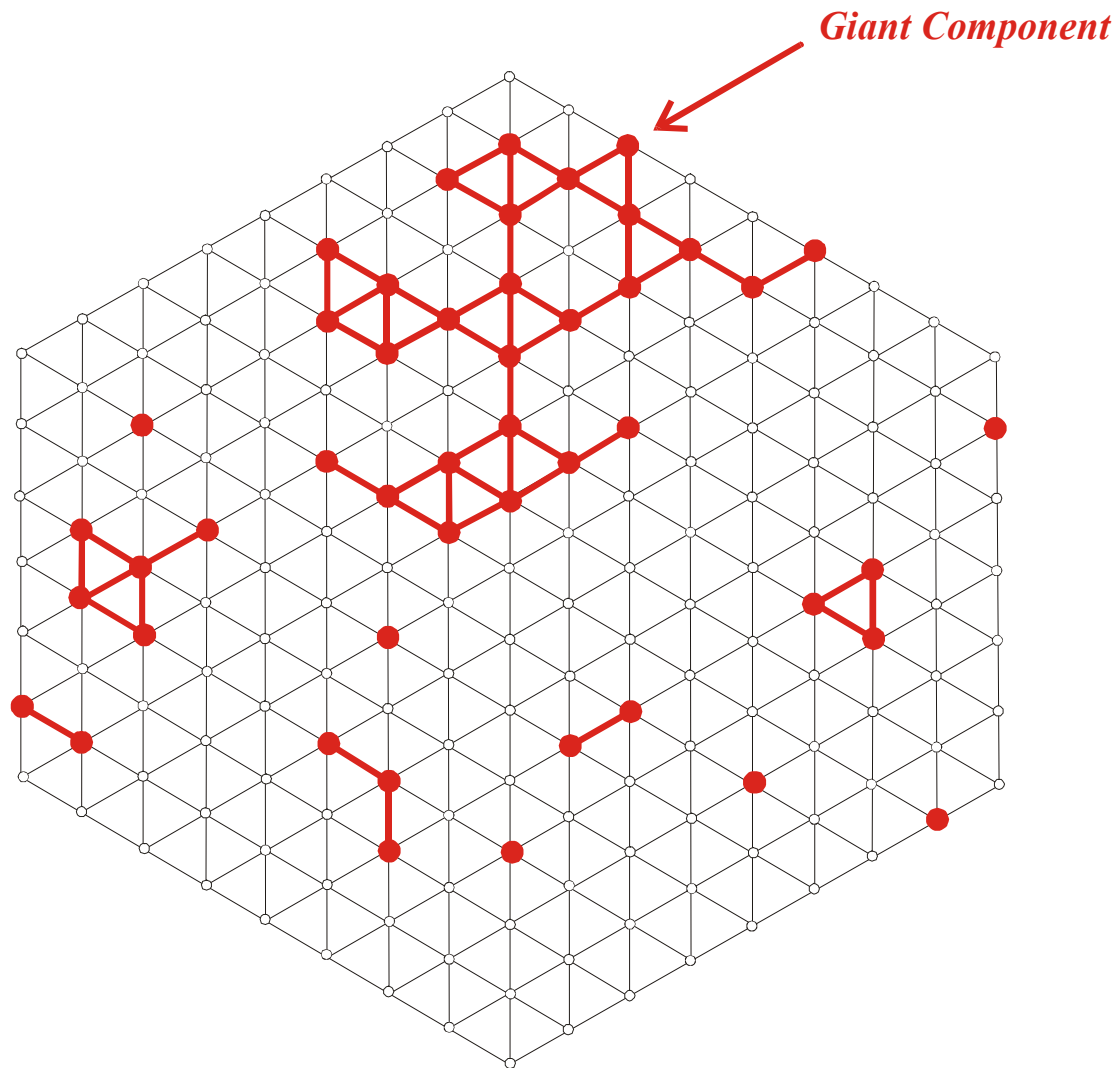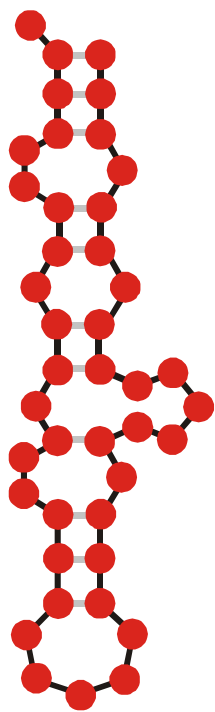
A multi-component neutral network
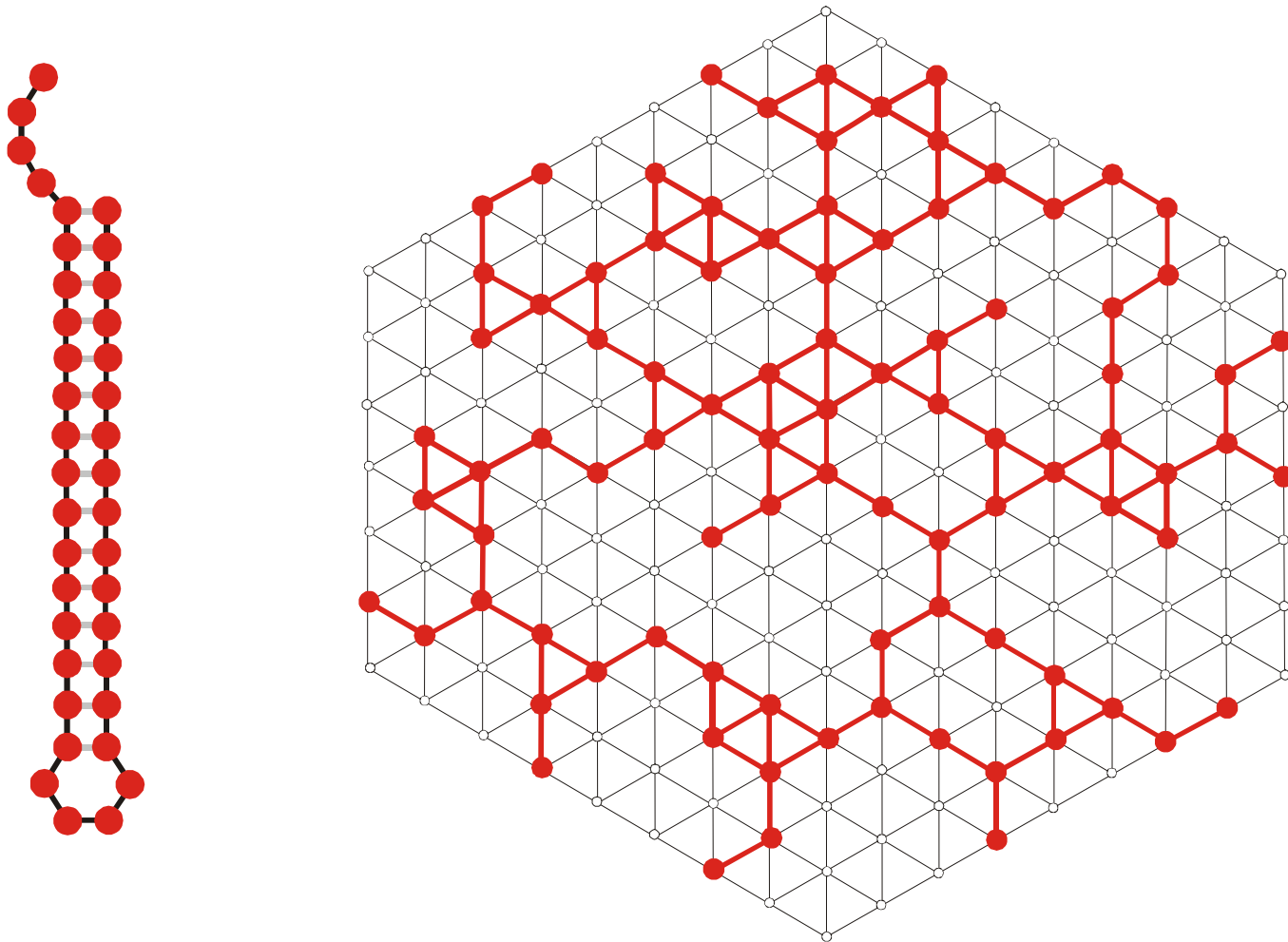
A connected neutral network

3'-end

CUGGGAAAAAUCCCCAGACCGGGGGUUUCCCGG

5'-end

3'-end

CUGGGAAAAAUCCCCAGACCGGGGGUUUCCGCGG

5'-end

Compatible
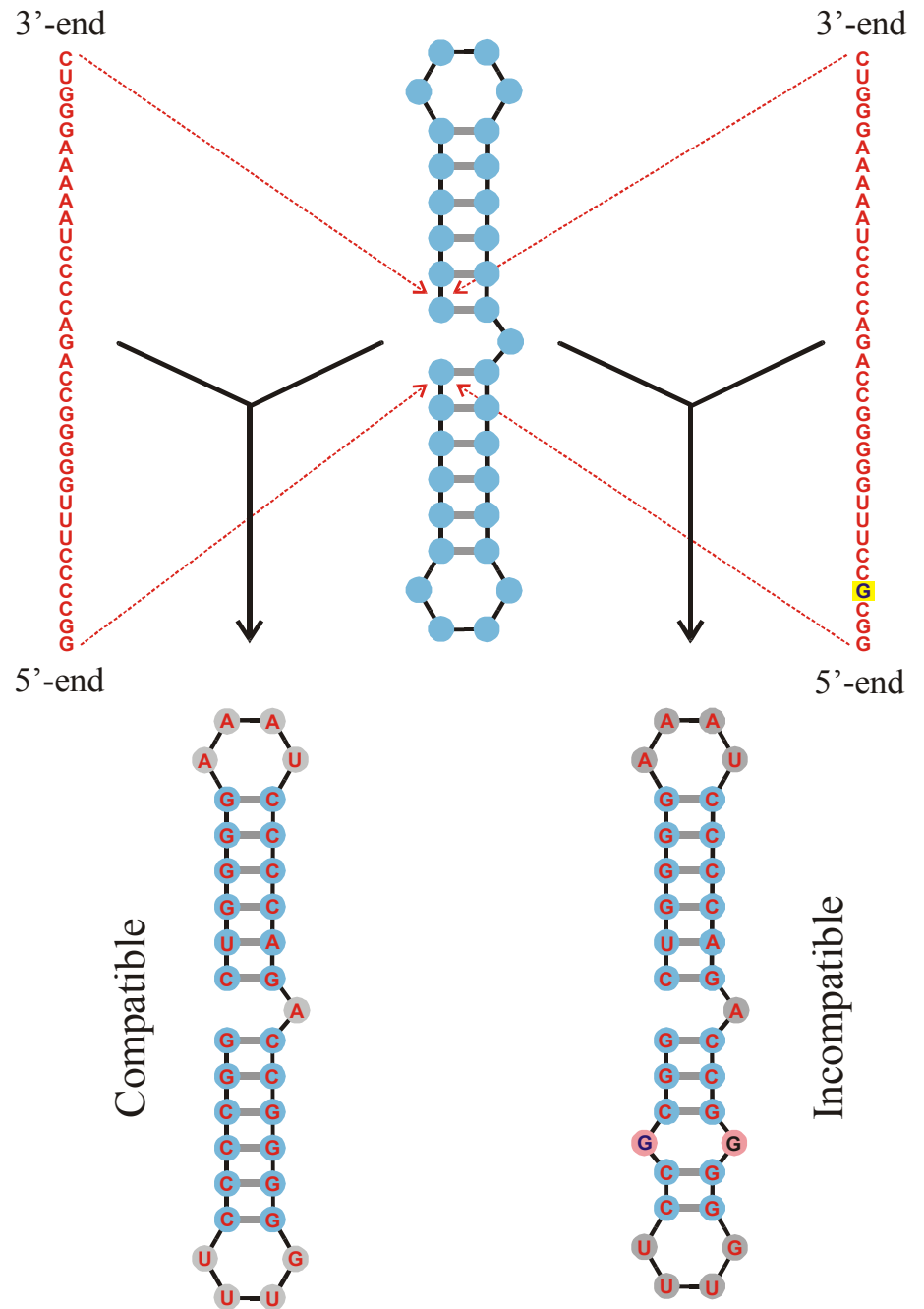
Incompatible

Compatibility of sequences with structures

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

$G_k \subseteq C_k$

Neutral network G_k
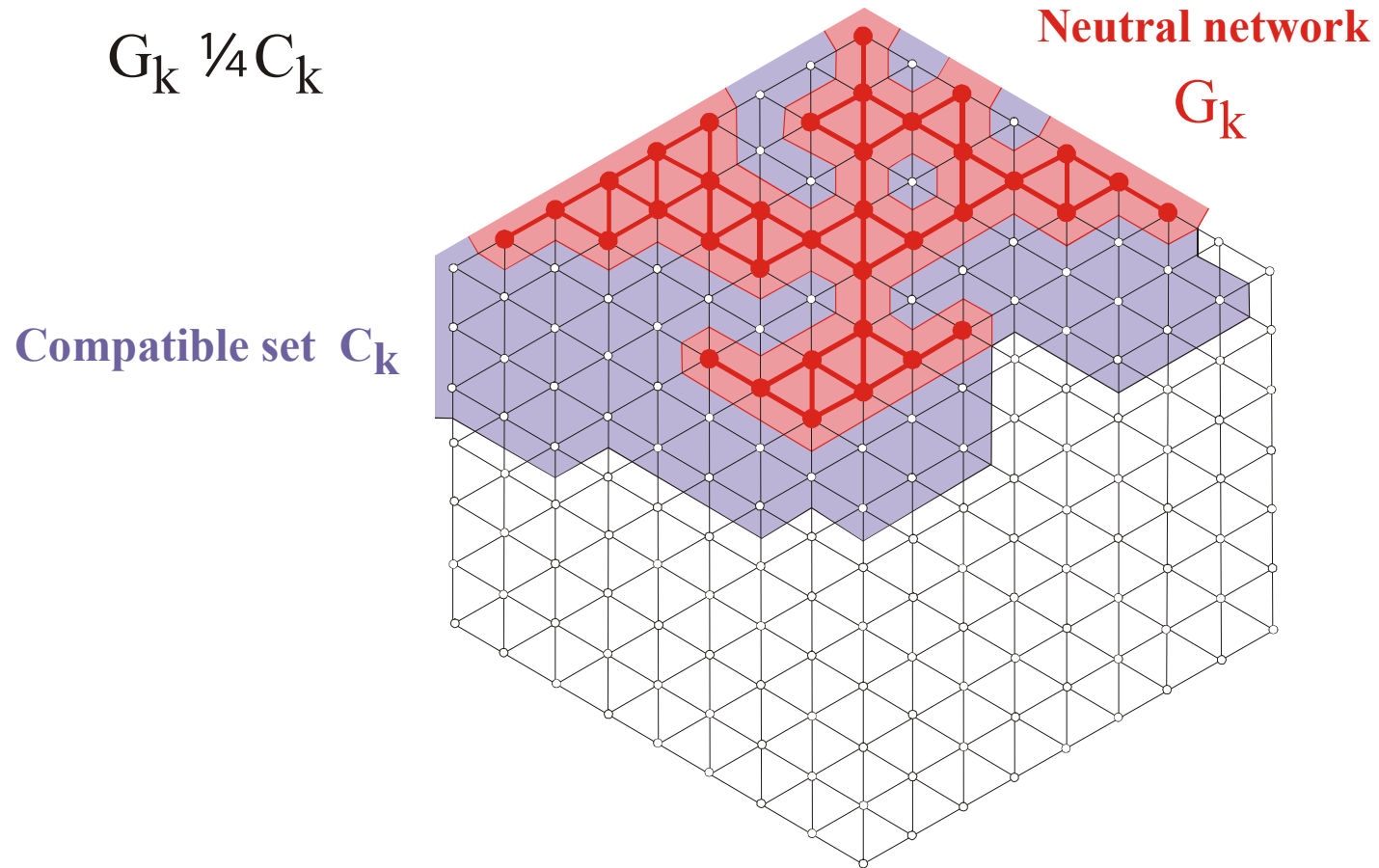
Compatible set C_k

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.

3'- end

CUGGGAAAAAUCCCCAGACCGGGGGUUUCCCGG

5'- end

Minimum free energy conformation S_0

Suboptimal conformation S_1

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

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

- ◯ : $\frac{1}{2}C_1 \; \grave{E} \; \frac{3}{4}C_2$
- ◯ : $\frac{3}{4}C_1 \; \grave{E} \; \frac{1}{2}C_2$
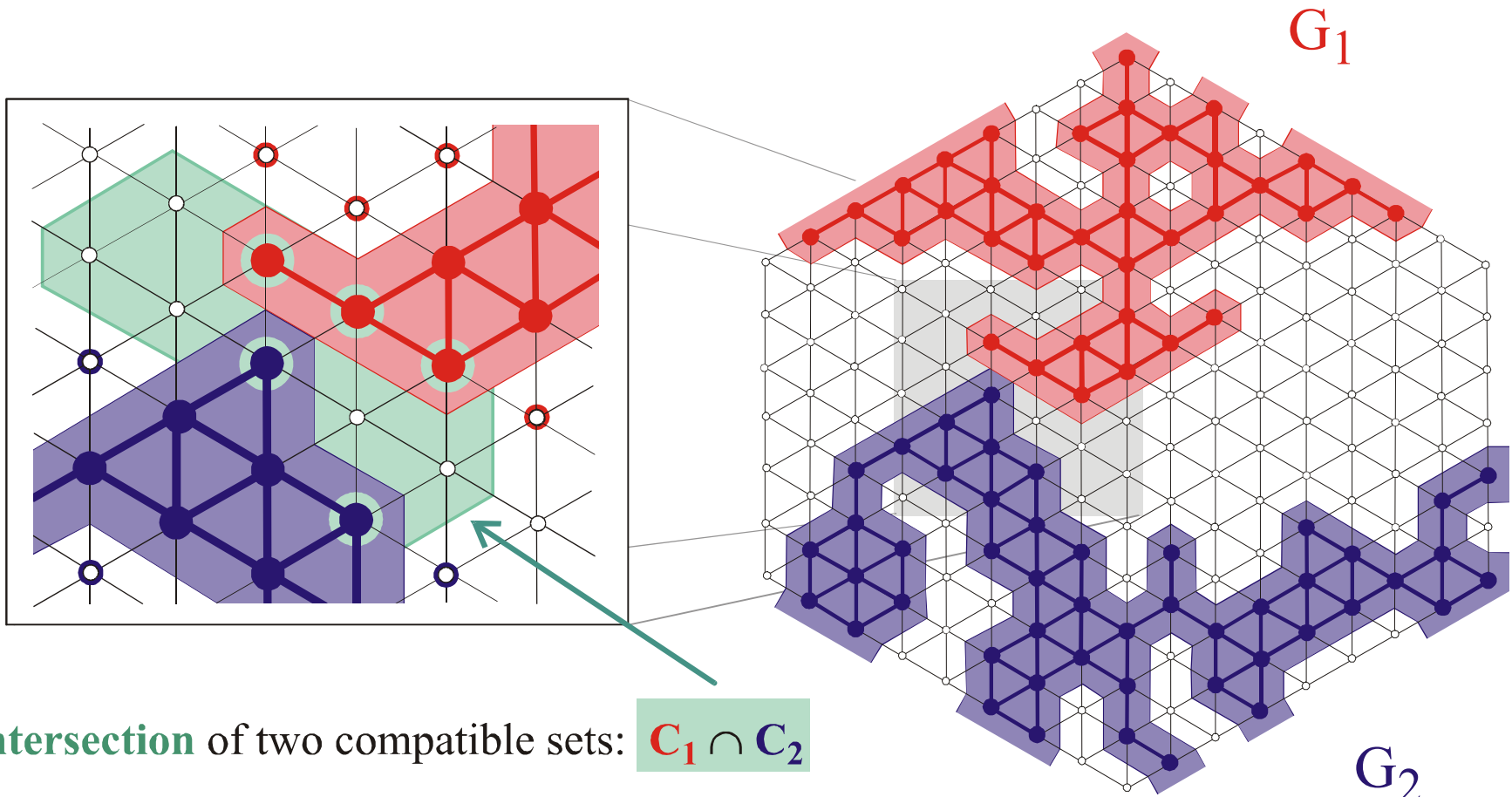
The intersection of two compatible sets is always non empty: $C_1 \; \P \; C_2 \; \frac{3}{4}\mu$

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* $\mathbf{C}[s], \mathbf{C}[s']$ *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 $\jmath(s, s') \cong D_m$ 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***

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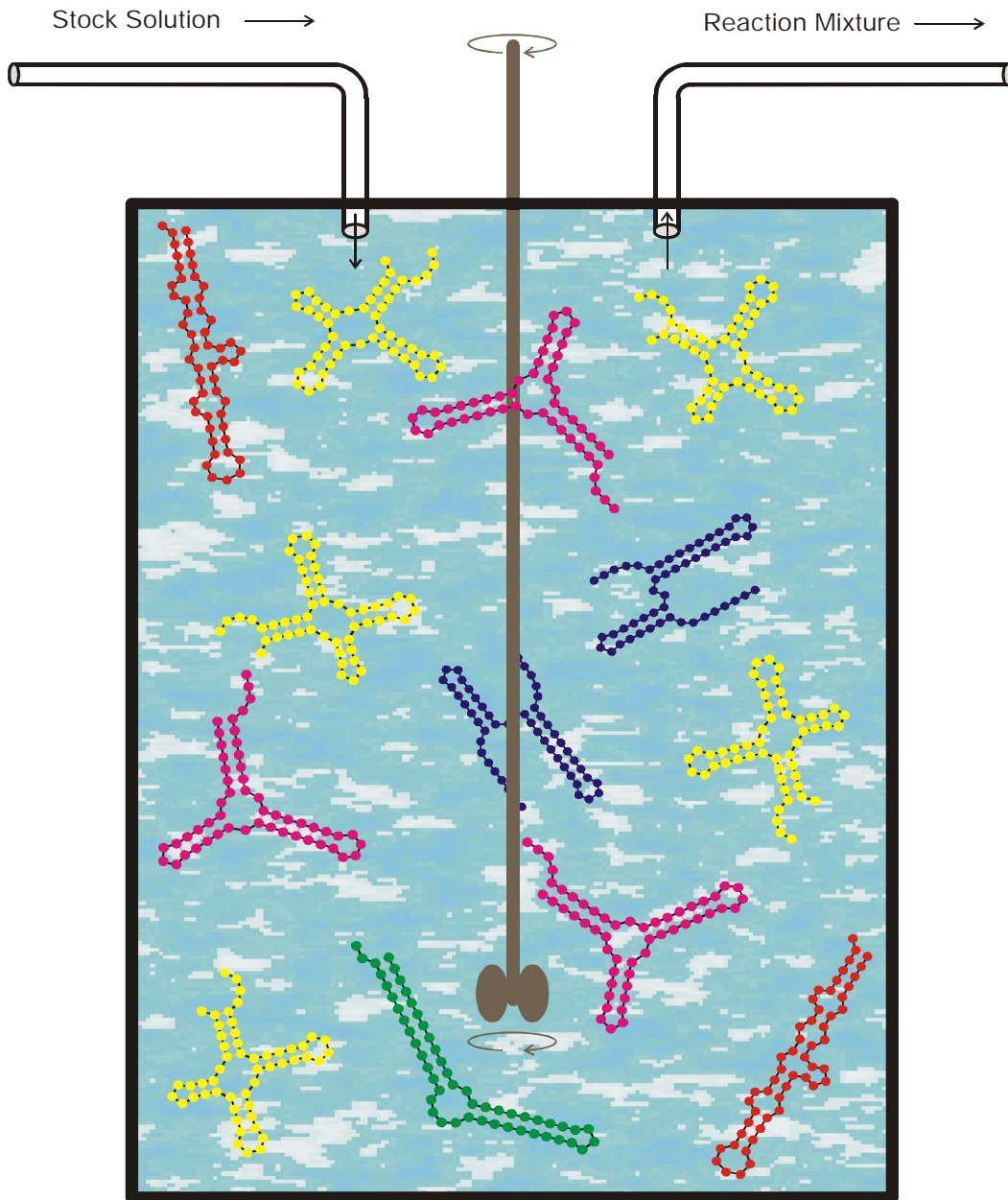
B.M.R.Stadler, P.F.Stadler, G.P.Wagner, W.Fontana, ***The topology of the possible: Formal spaces underlying patterns of evolutionary change.*** J.Theor.Biol. **213** (2001), 241-274

3'-End

5'-End

70

60

10

50

20

30   40
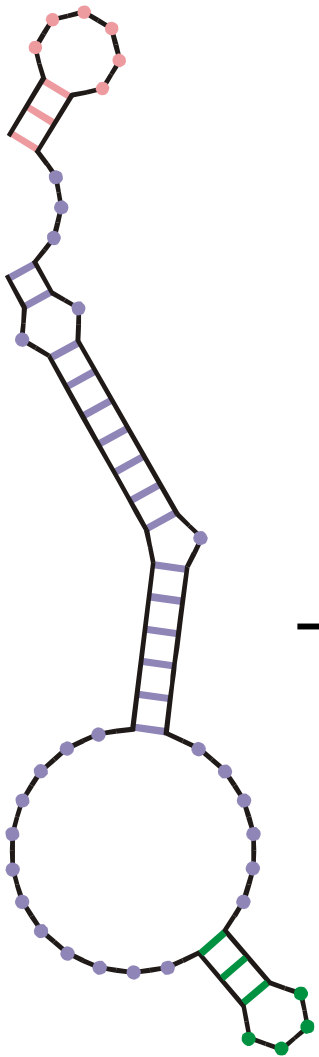
Randomly chosen
initial structure
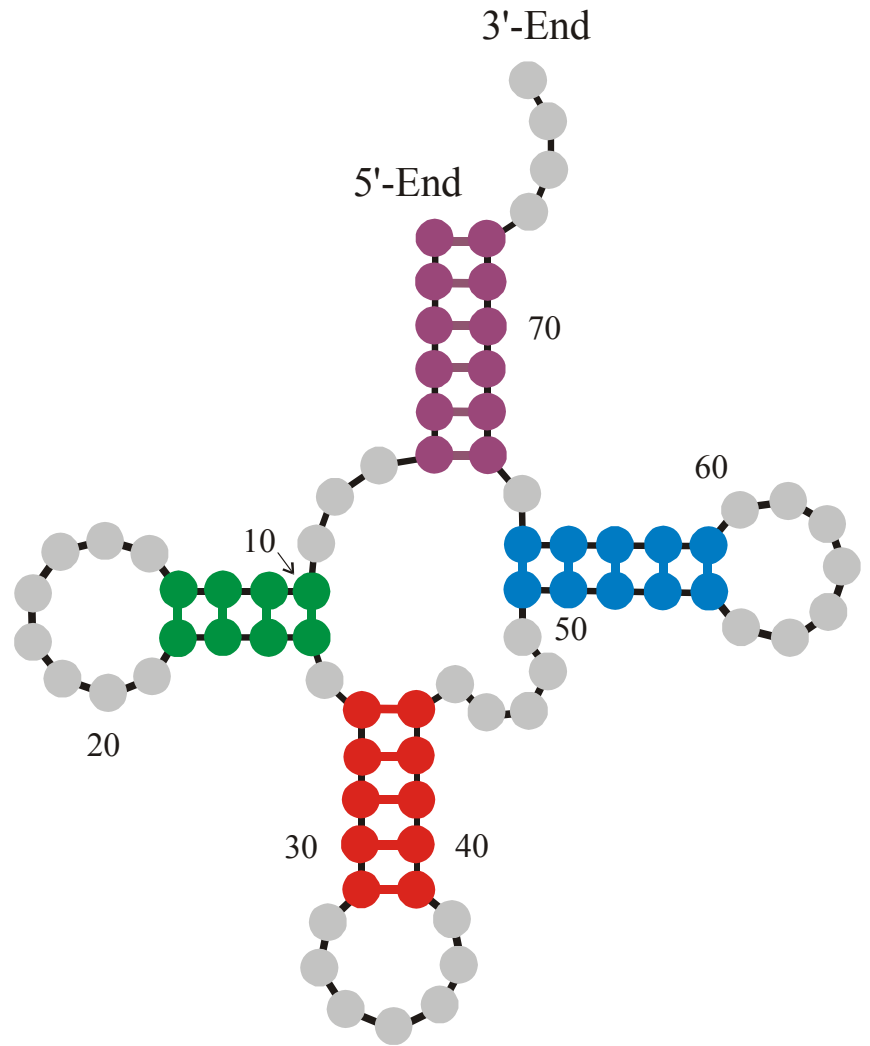
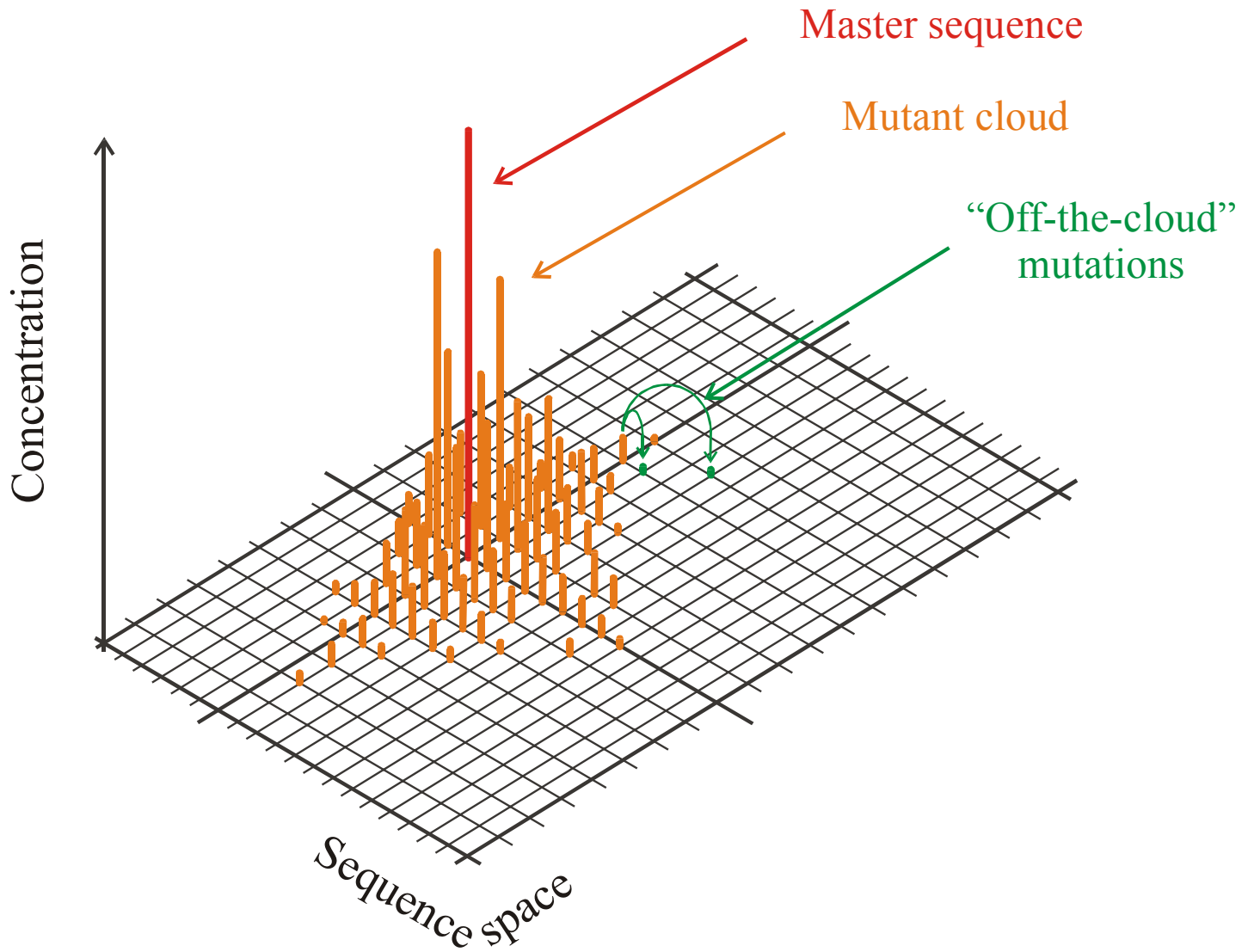Phenylalanyl-tRNA as
target structure

Stock Solution ⟶

Reaction Mixture ⟶

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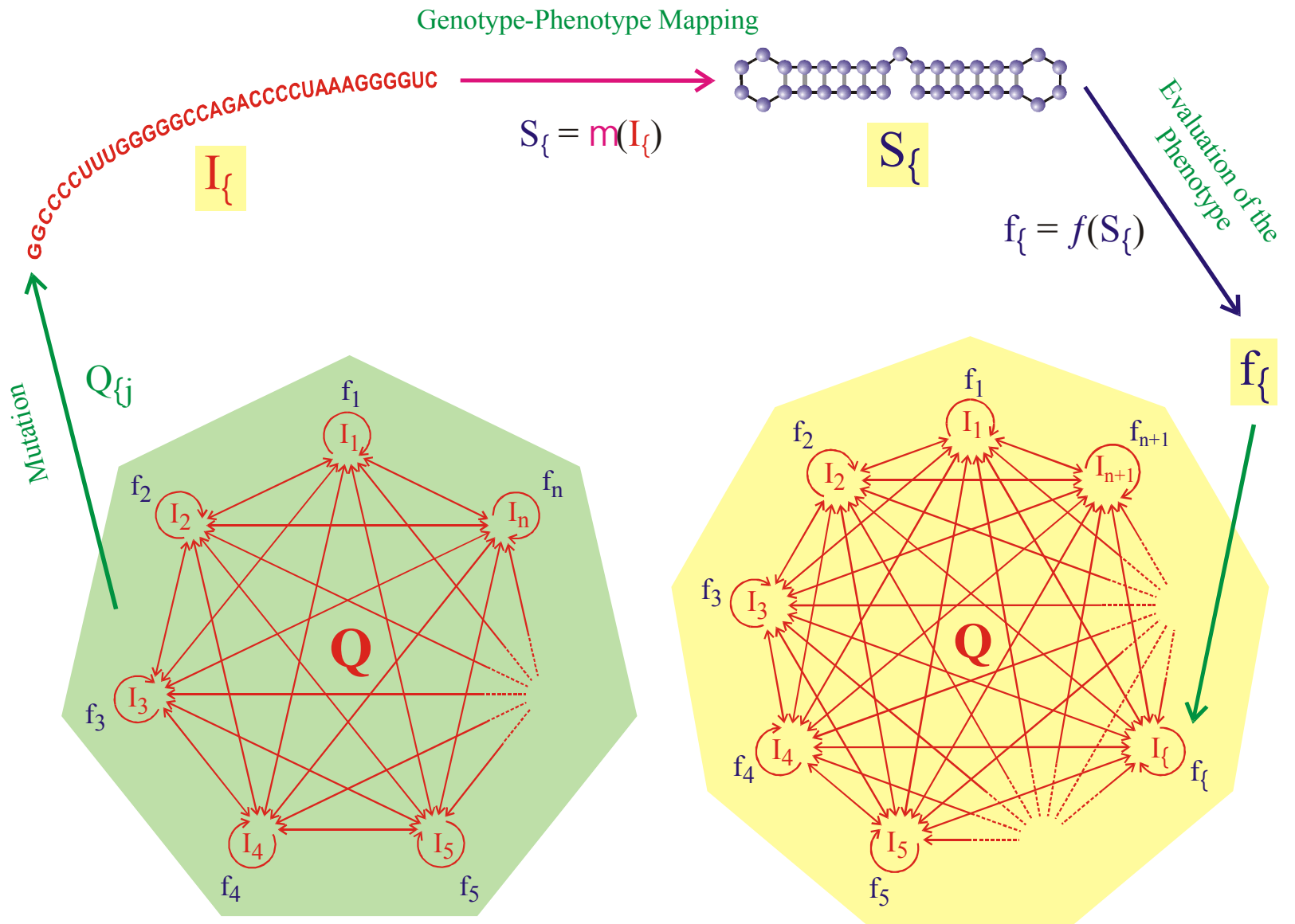
$f_k = [\ / [U + 8d_S^{(k)}]$

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

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

Master sequence

Mutant cloud

"Off-the-cloud" mutations

Concentration

Sequence space

The molecular quasispecies
in sequence space

Genotype-Phenotype Mapping

$I_{\{}$

$S_{\{} = m(I_{\{})$

$S_{\{}$

Evaluation of the Phenotype

$f_{\{} = f(S_{\{})$

$f_{\{}$

$Q_{\{j}$

Mutation

f_1

I_1

f_2

I_2

f_n

I_n

f_3

I_3

Q

f_4

I_4

f_5

I_5

f_1

I_1

f_2

I_2

f_{n+1}

I_{n+1}

f_3

I_3

Q

f_4

I_4

$I_{\{}$

$f_{\{}$

f_5

I_5

Evolutionary dynamics
including molecular phenotypes
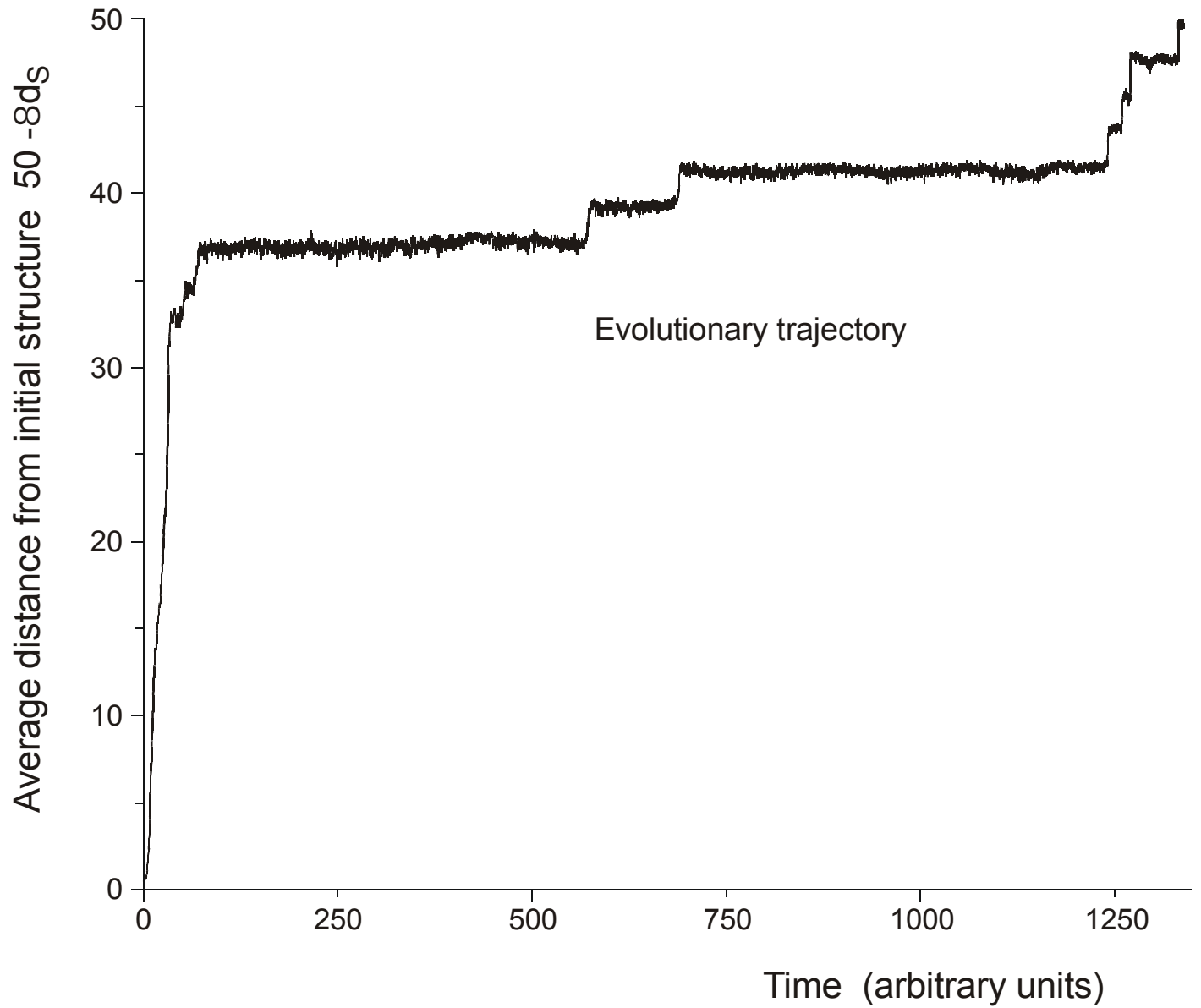
In silico optimization in the flow reactor: Trajectory (**biologists' view**)
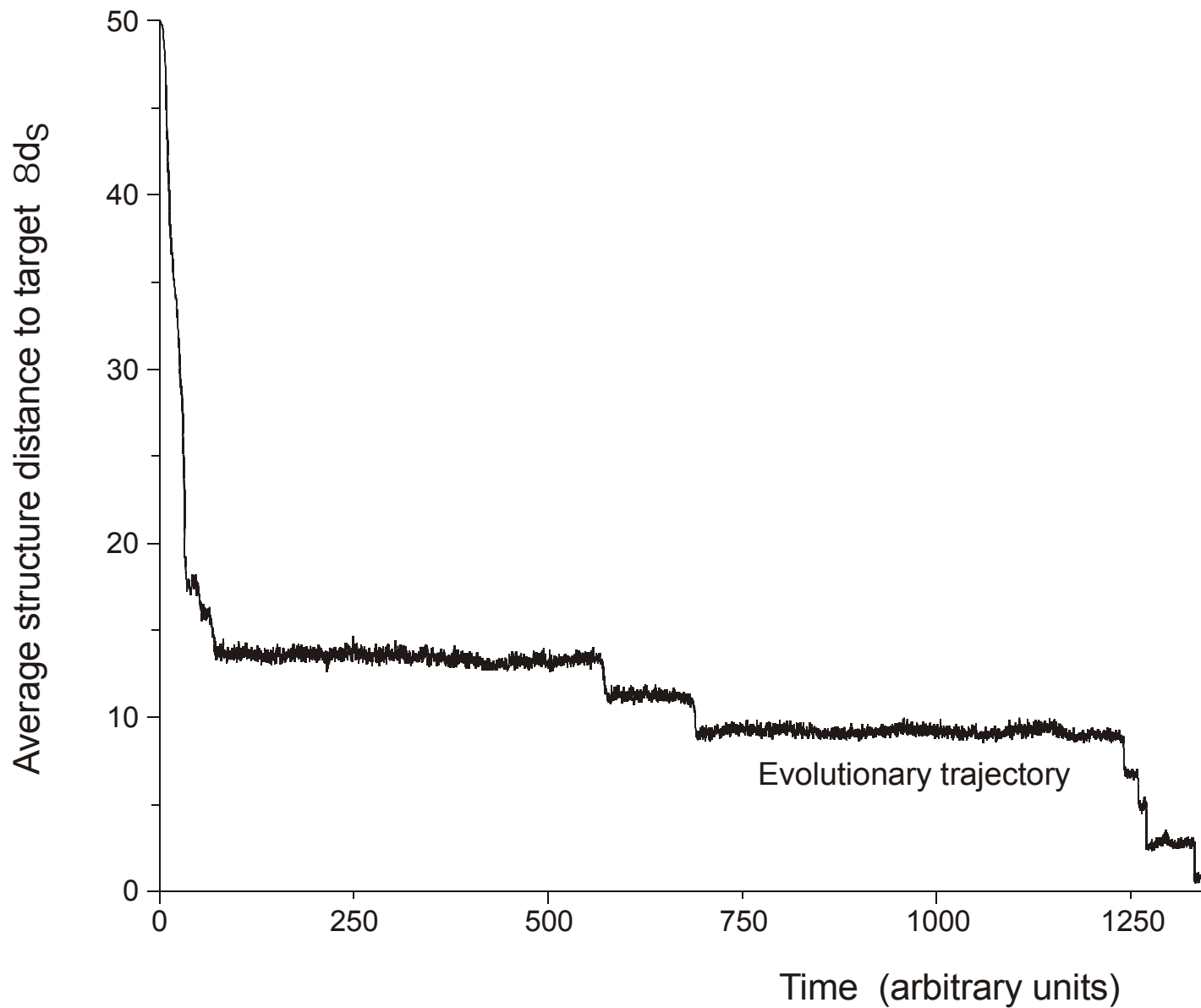
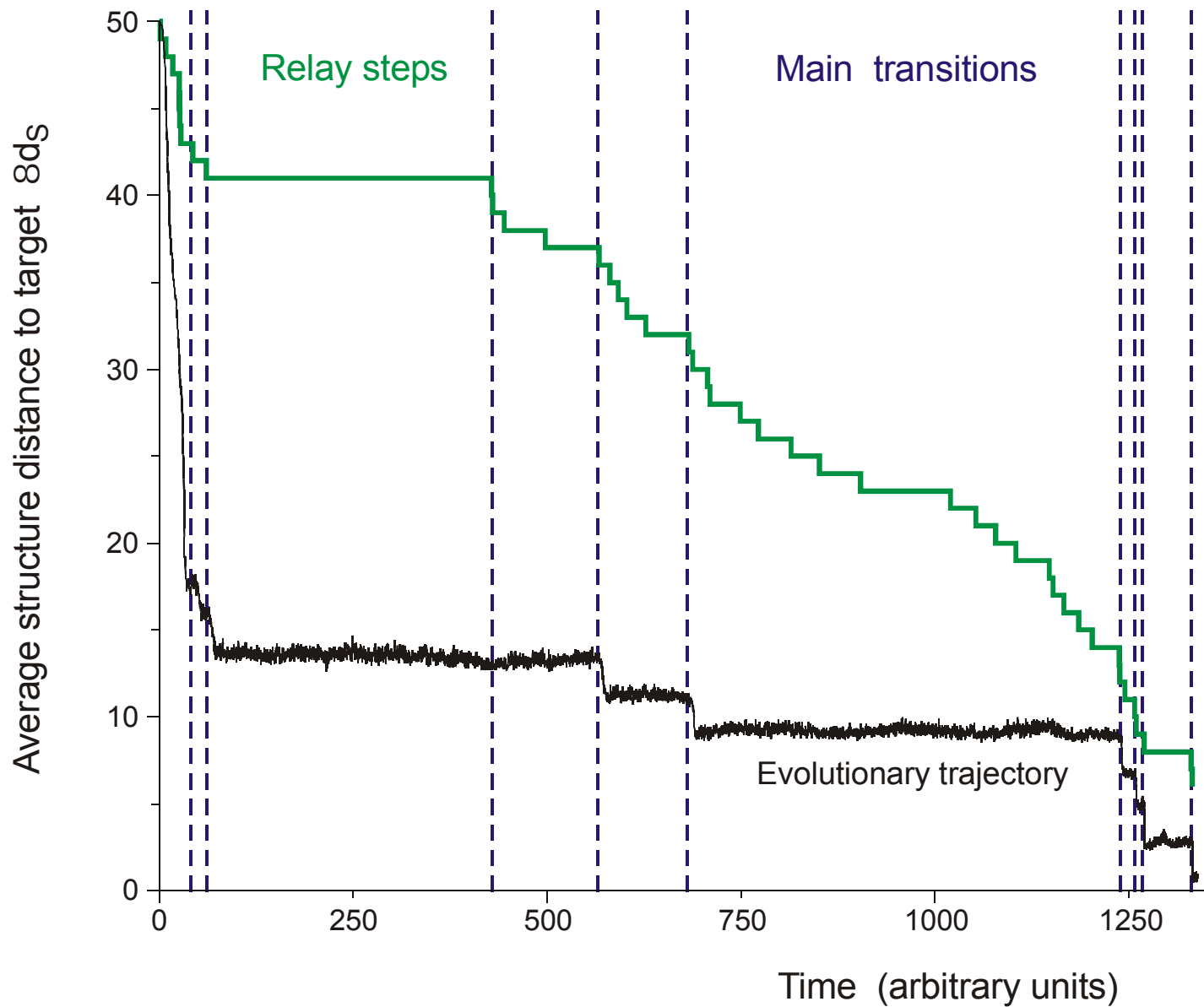*In silico* optimization in the flow reactor: Trajectory (**physicists' view**)

In silico optimization in the flow reactor: Main transitions

Shift

Roll-Over

Flip

Double Flip

α

a

a

α

α

a

b

β

a

α

b

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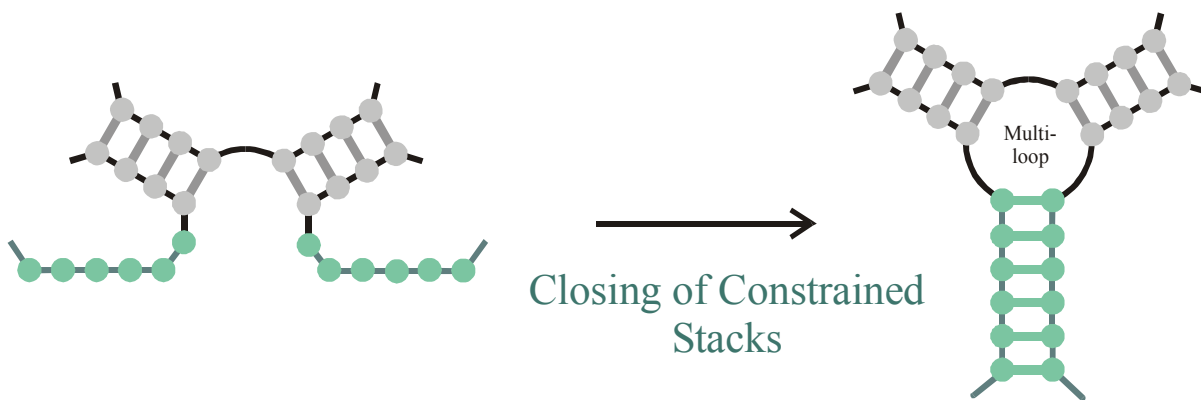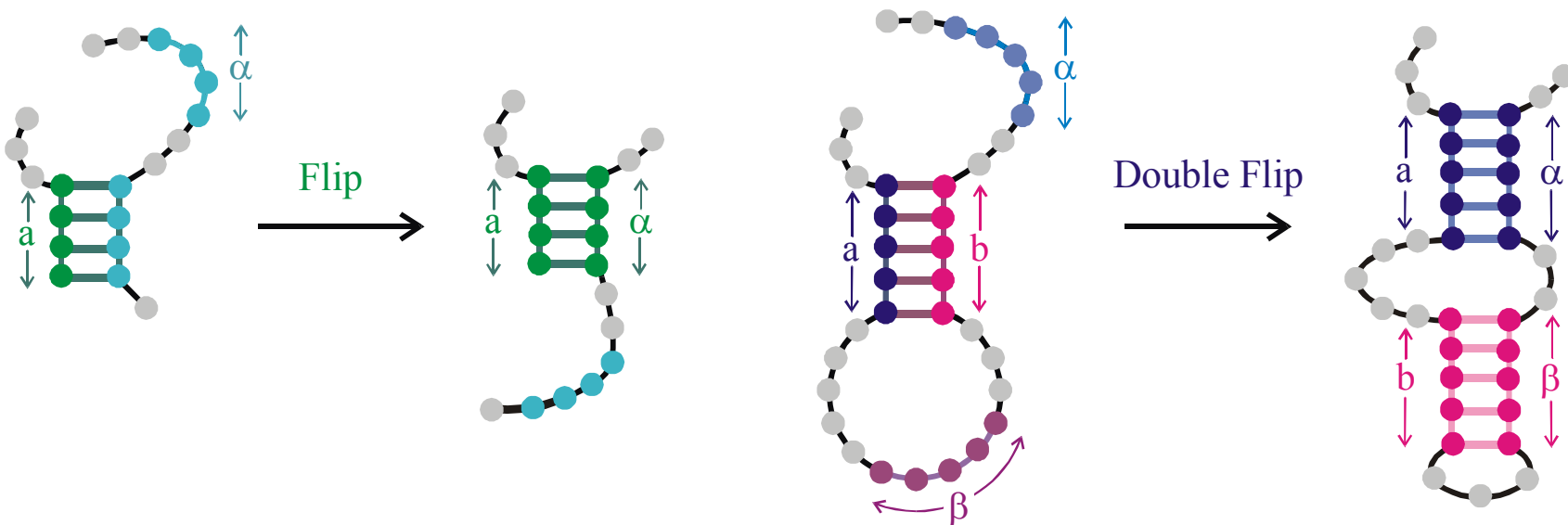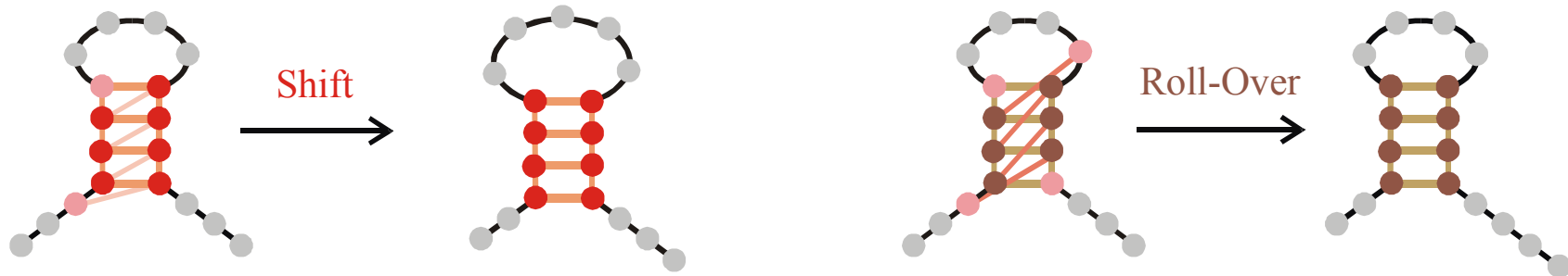
Main or discontinuous **transitions**: *Structural innovations*, occur **rarely** on single point mutations

Closing of Constrained Stacks

Multi-loop

Relay steps

Main transitions

Uninterrupted presence

Evolutionary trajectory

Average structure distance to target $8d_S$

Time (arbitrary units)
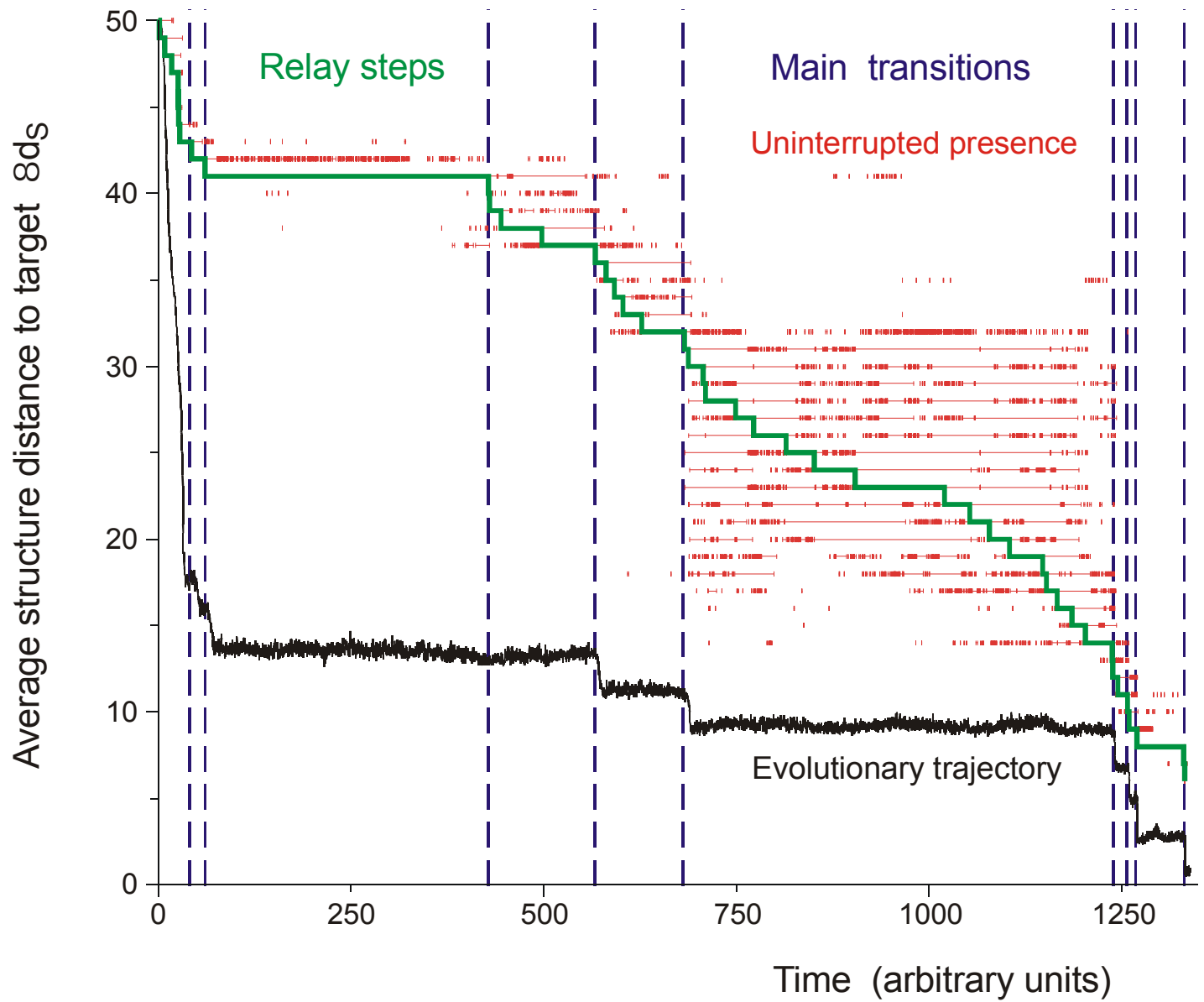
In silico optimization in the flow reactor
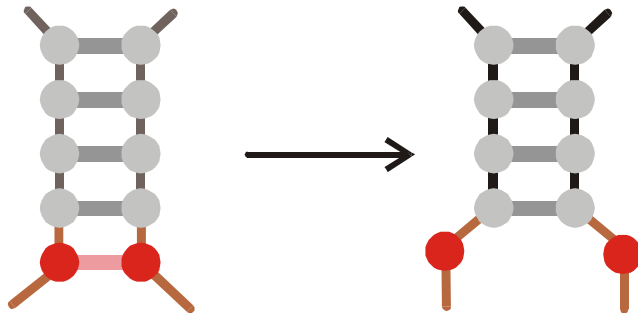
Shortening of Stacks
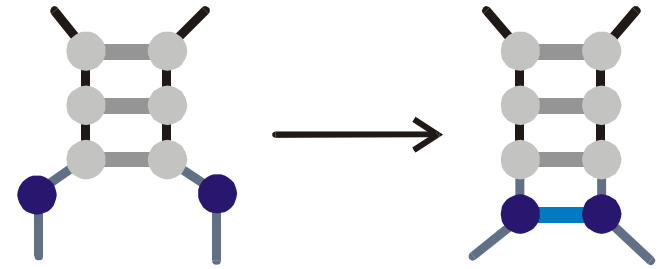
Elongation of Stacks

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

Multi-loop

Opening of Constrained Stacks

Statistics of evolutionary trajectories

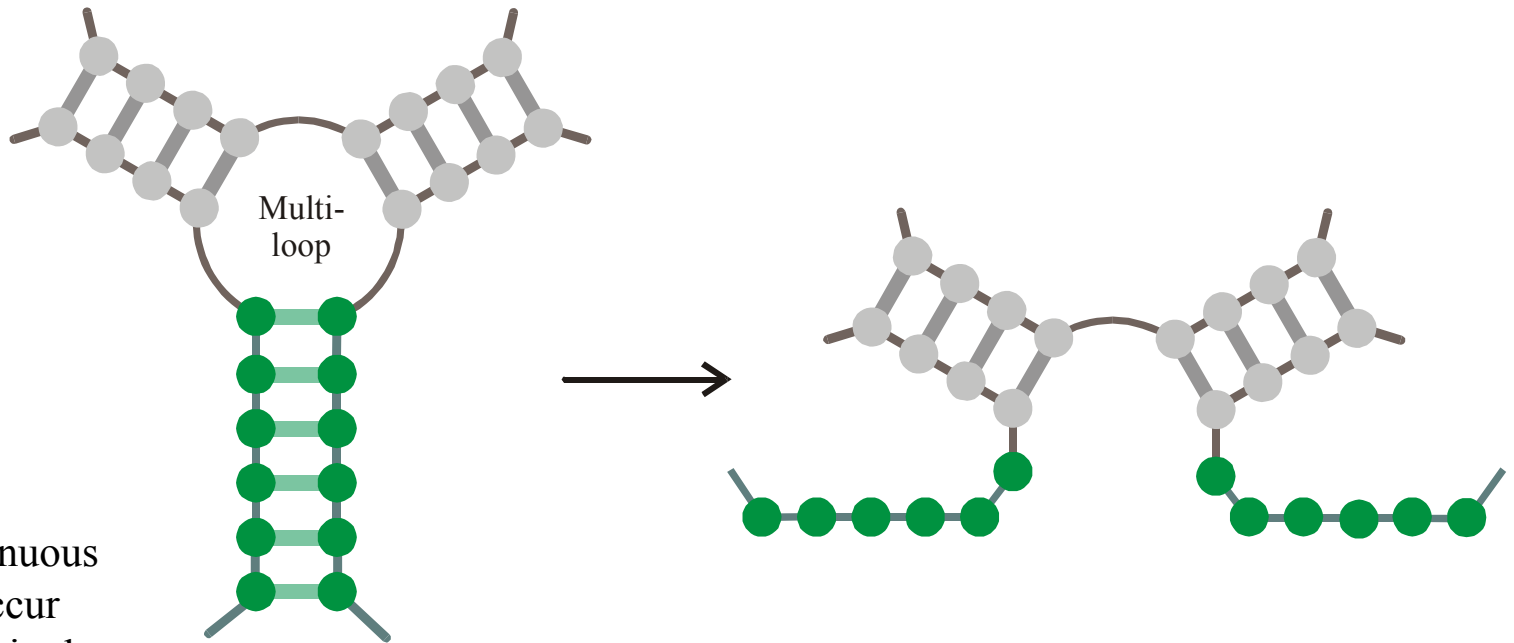| Population size N	Number of replications $< n_{rep} >$	Number of transitions $< n_{tr} >$	Number of main transitions $< n_{dtr} >$
1 000	$(5.5 \pm [6.9, 3.1]) \times 10^7$	$92.7 \pm [80.3, 43.0]$	$8.8 \pm [2.4, 1.9]$
2 000	$(6.0 \pm [11.1, 3.9]) \times 10^7$	$55.7 \pm [30.7, 19.8]$	$8.9 \pm [2.8, 2.1]$
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.

„…*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)
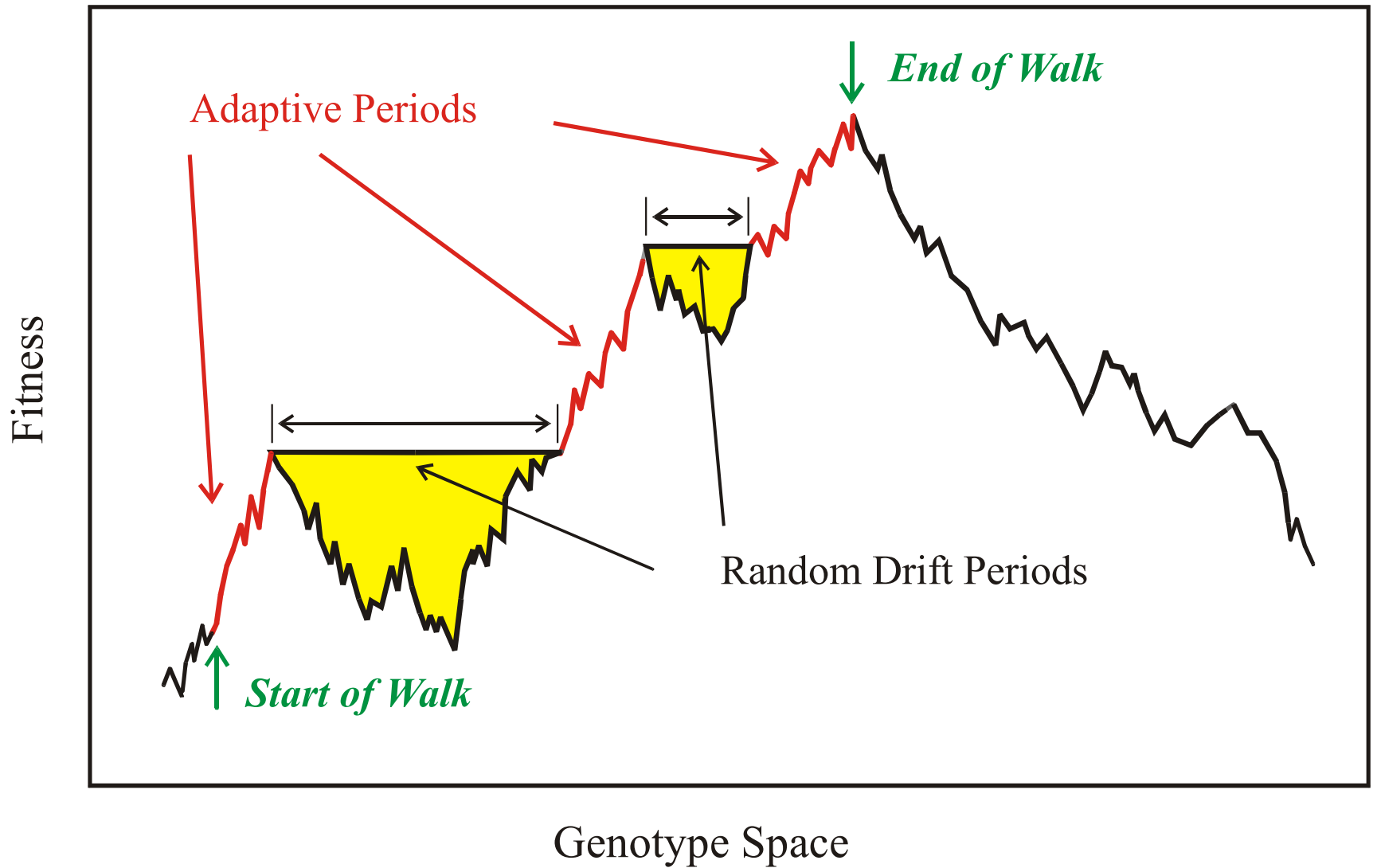
Evolution in genotype space sketched as a non-descending walk in a fitness landscape

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

L.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

Amplification
Diversification

Genetic
Diversity

Selection Cycle

Selection

Desired Properties
? ? ?

no

yes

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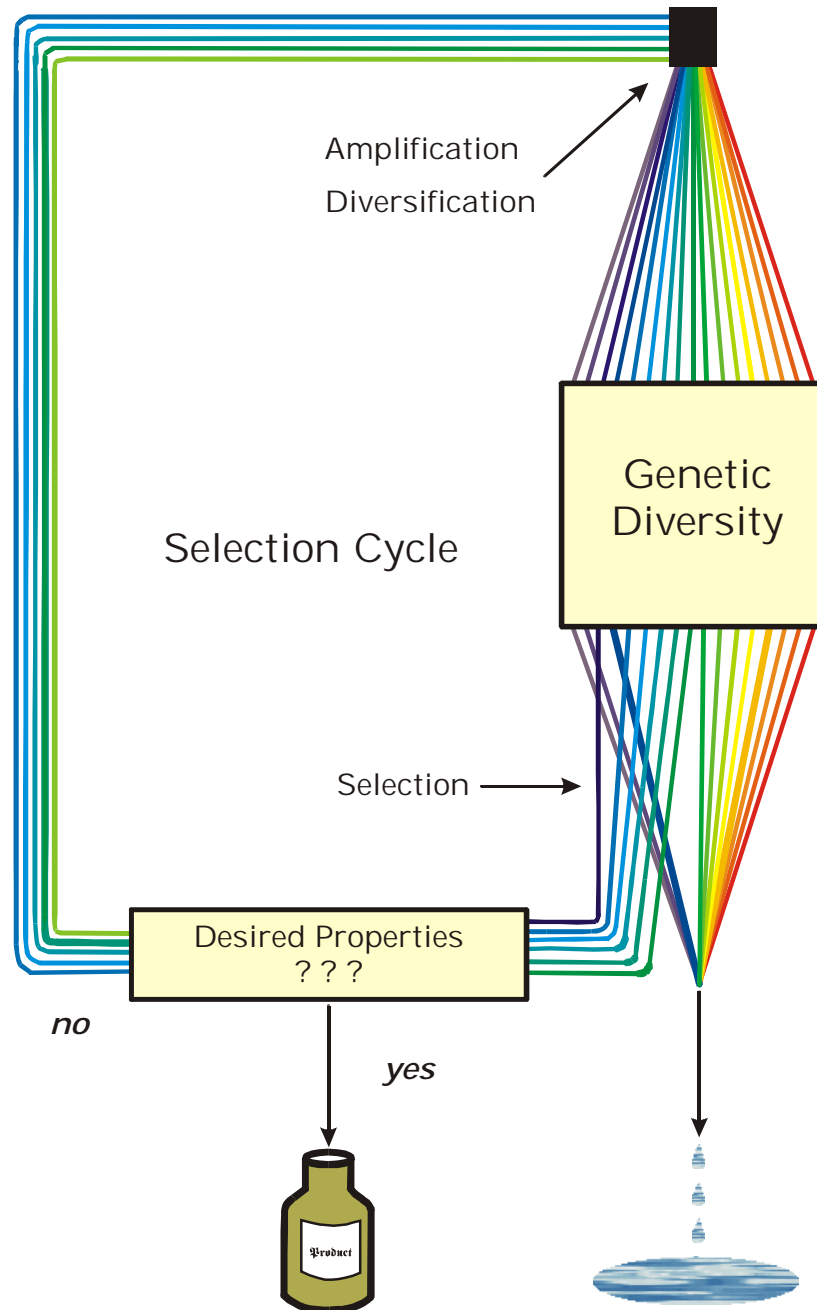
Selection cycle used in
applied molecular evolution
to design molecules with
predefined properties

Retention of binders

Elution of binders

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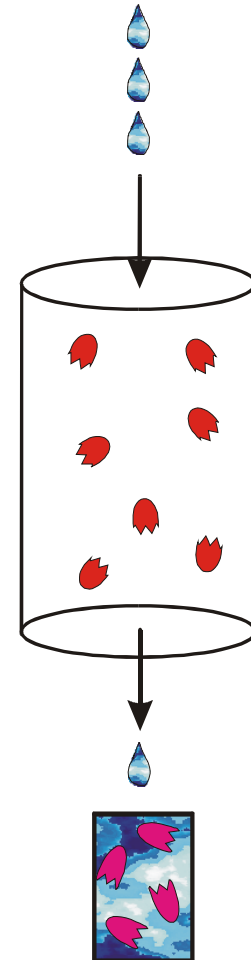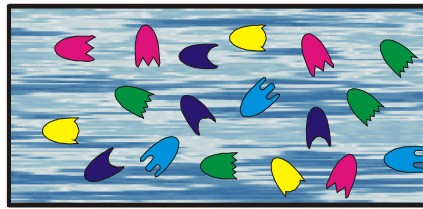
The SELEX technique for the evolutionary design of *aptamers*

TR8-28   TR8-29   TCT8-5   TCT8-4

G C c
G   C
G   G
A   G
CG
CG
AU
G   A
A   U
A   U
C   A
GC
GC
UA
U
C
C
GC
GC
UA 3'
G
A
G
C
U
A
U

U C
A   G
AU
CG
UA
G   A
A   U
A   A
GC
GC
UA
U
C
C
AU
AU
CG UU 3'
U
U
A
G
A
G
5'C

A   A
G   A
A   G
A   A
UG
GC
UA
GC
CG
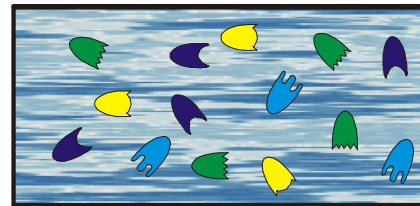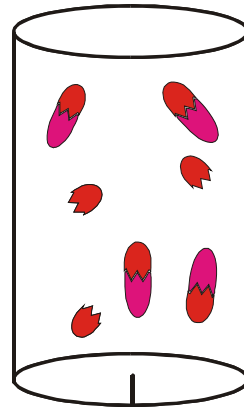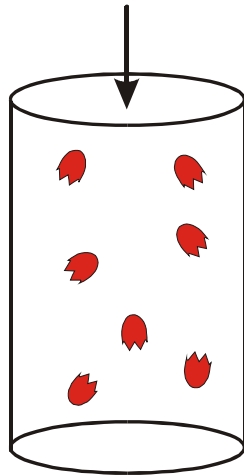UA
C
C
U
AU
CG
CG
A   C
U   A
A   C
A   G
AU
CG
U   U
CG
GU
CG
AU
5' AU 3'

C
U   U
GU
CG
UA
AU
CG
GC
C
C
U
AU
CG
CG
A   C
U   A
A   C
A   G
GC
UA
GC
AU
5' AU
C
A 3'

N
N   N
N   N
NN'
NN'
C
C
U
AU
CG
CG
region 1   A   C/A
A
UA
A   G
NN'
NN'

loop; n=3-12

stem; nn'=2-6

region 2

stem; nn'=1-8
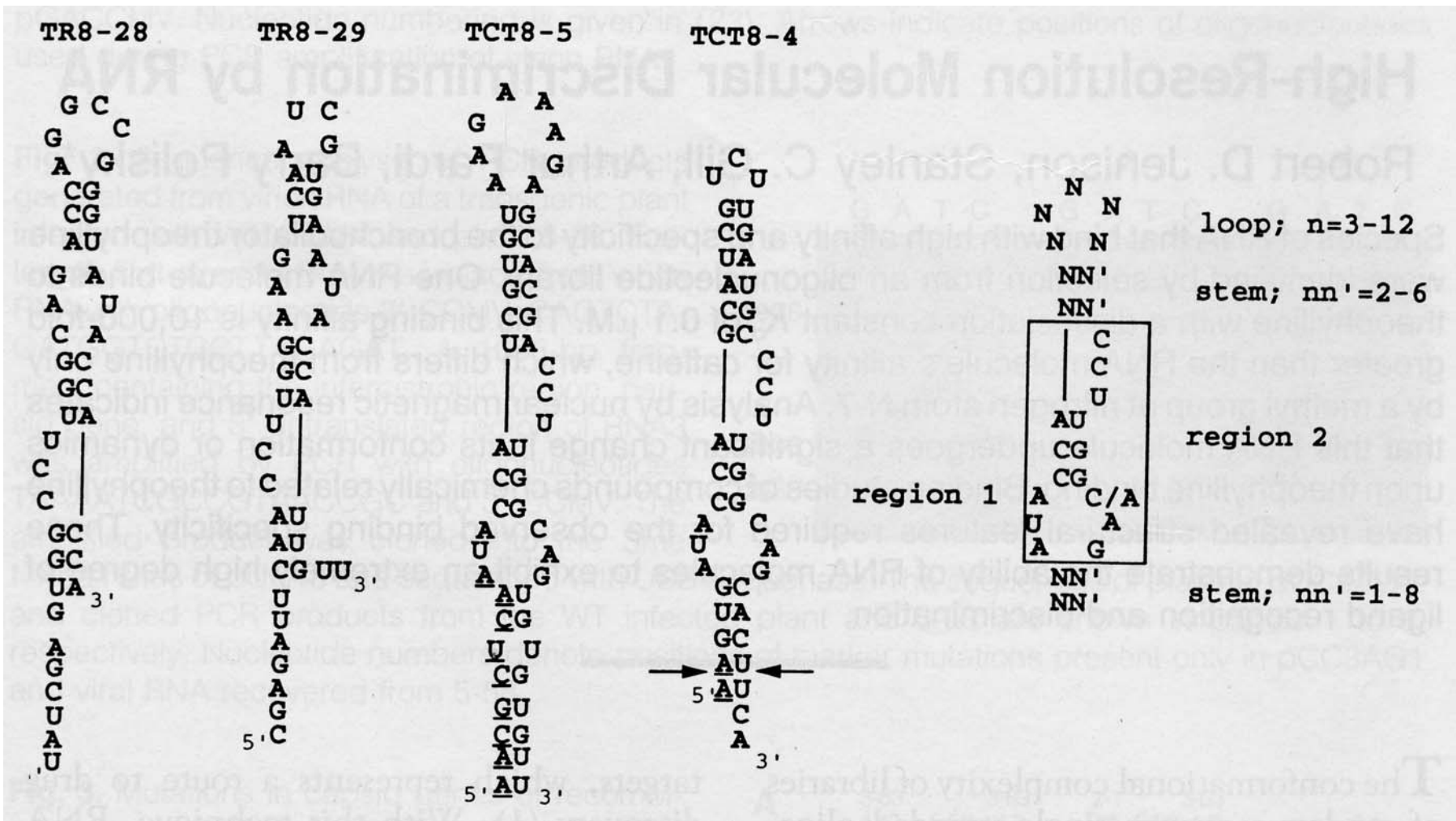
Secondary structures of aptamers binding theophyllin, caffeine, and related compounds

additional methyl group

Caffeine
(1,3,7-Trimethylxanthine)

Theophylline
(1,3-Dimethylxanthine)
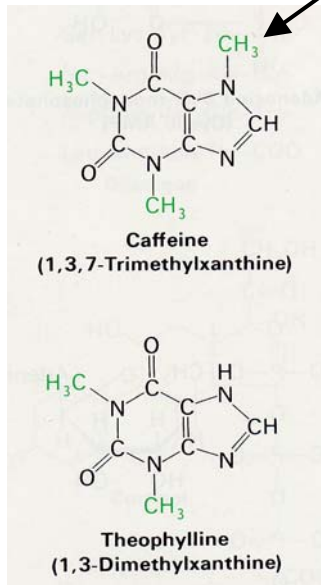
Dissociation constants and specificity of theophylline, caffeine, and related derivatives of uric acid for binding to a discriminating aptamer TCT8-4

**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	$K_d(c)$ (µM)	$K_d(c)/K_d(t)$
Theophylline		0.32 ± 0.13	1
CP-theophylline		0.93 ± 0.20	2.9
Xanthine		8.5 ± 0.40	27
1-Methylxanthine		9.0 ± 0.30	28
3-Methylxanthine		2.0 ± 0.7	6.3
7-Methylxanthine		> 500	>1500
3,7-Dimethylxanthine		> 500	> 1500
1,3-Dimethyluric acid		> 1000	>3100
Hypoxanthine		49 ± 10	153
Caffeine		3500 ± 1500	10,900
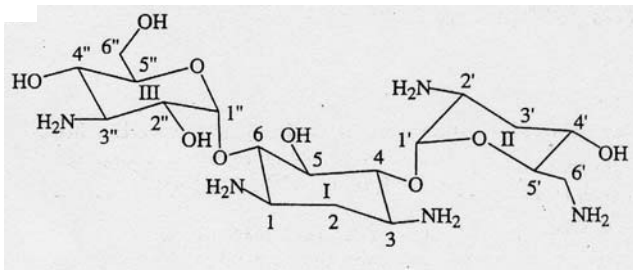
tobramycin
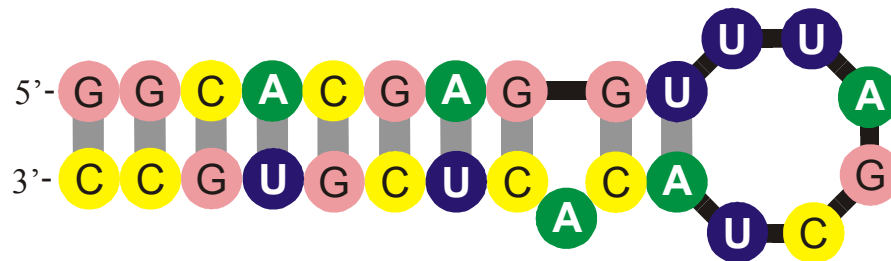
RNA aptamer

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)
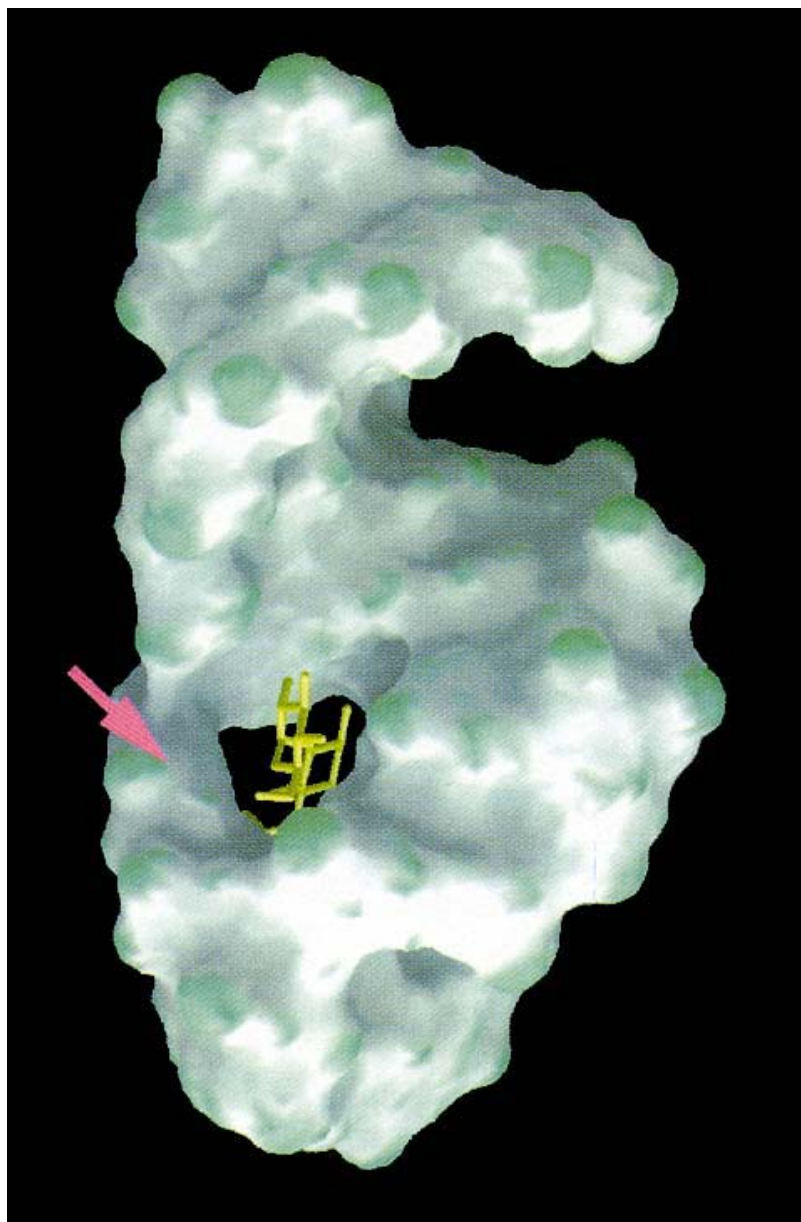
The three-dimensional structure of the tobramycin aptamer complex

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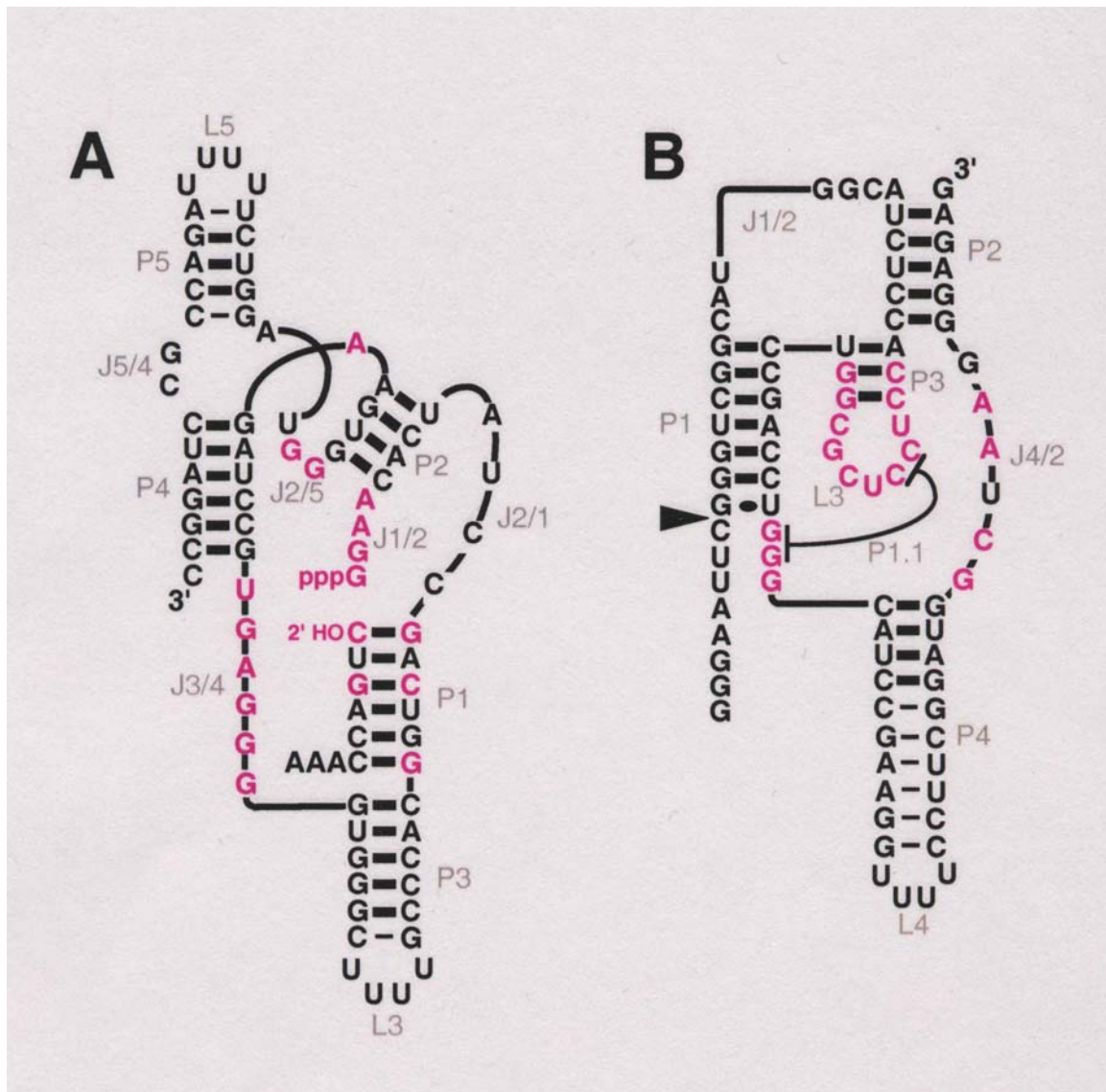
Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-X-virus (**B**)

Ligase fold   **HDV fold**
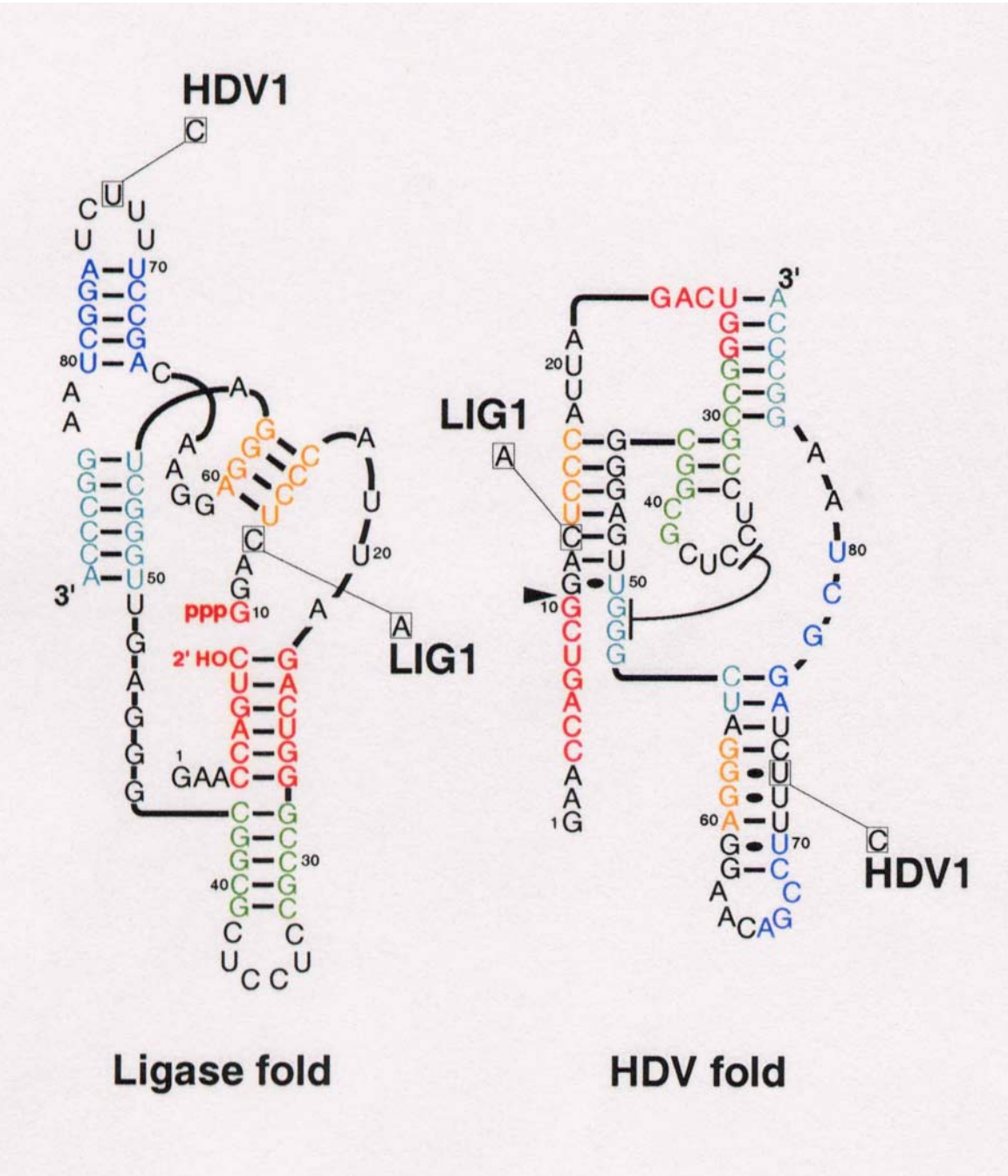
The sequence at the *intersection*:

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

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

From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER[1,2,3], WALTER FONTANA[3], PETER F. STADLER[2,3]
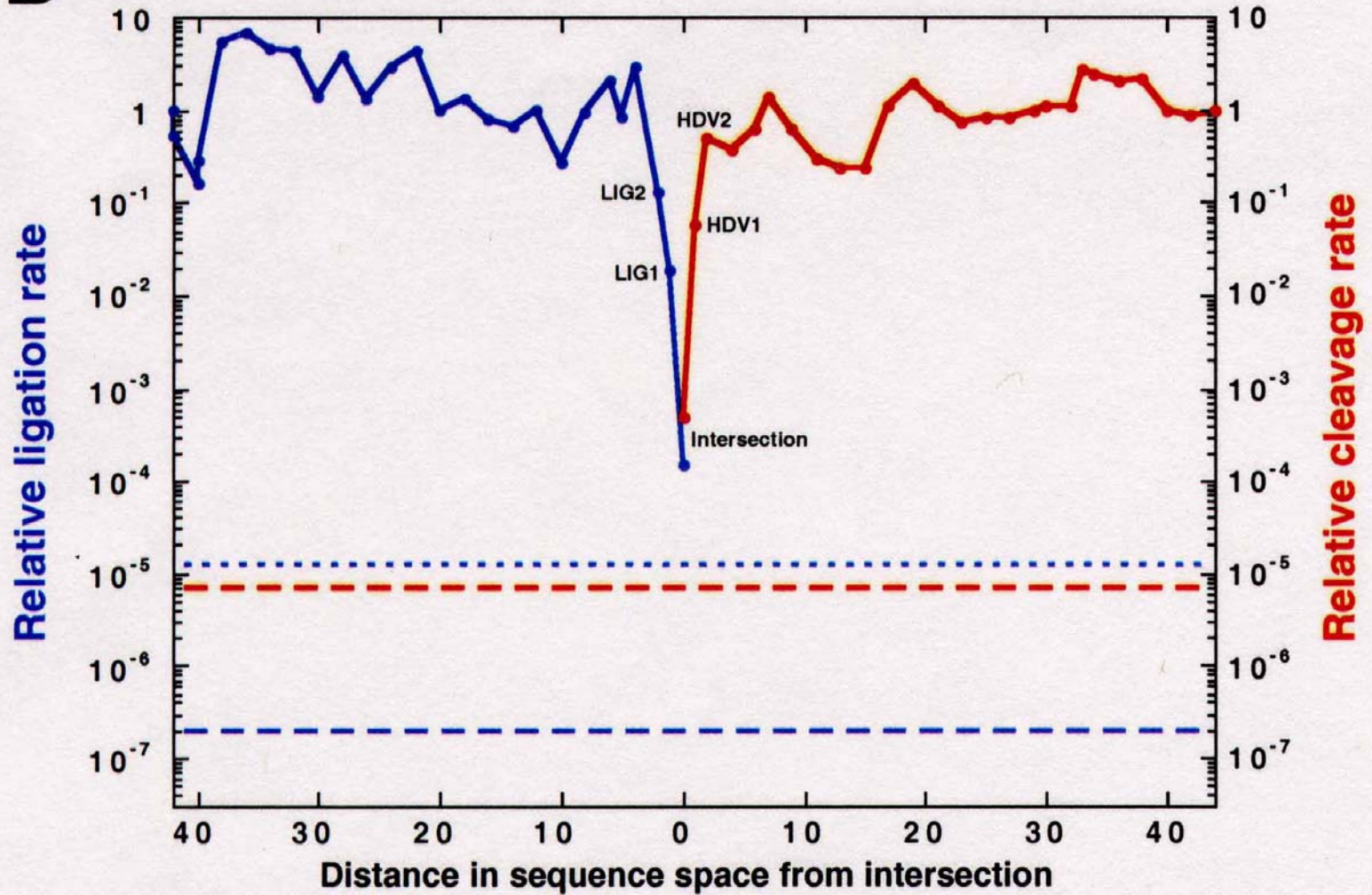AND IVO L. HOFACKER[2]

[1] *Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany*
[2] *Institut für Theoretische Chemie, Universität Wien, Austria*
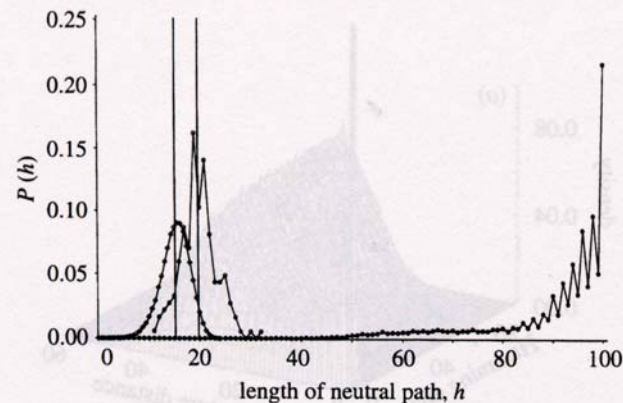[3] *Santa Fe Institute, Santa Fe, U.S.A.*

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).

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.

Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

279

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

Coworkers

Peter Stadler, Universität Leipzig, GE

Walter Fontana, Santa Fe Institute, NM

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

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