RNA – A Magic Molecule*

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie
der Universität Wien

38th Winter Seminar

Biophysical Chemistry, Molecular Biology, and
Cybernetics of Cell Function

Klosters, 15.– 28.01.2003

* Larry Gold at the conference „Frontiers of Life“, Blois (France), 1991

RNA ***as transmitter of genetic information***

DNA

↓ transcription

...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...

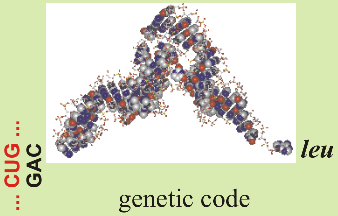
messenger-**RNA**

↓ translation

protein

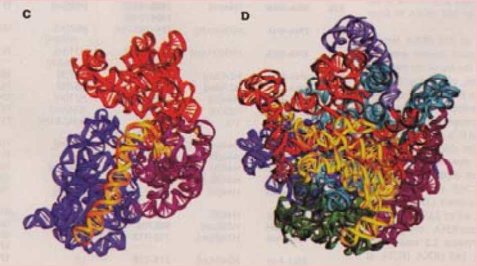
RNA as **working copy** of genetic information

RNA ***as adapter molecule***

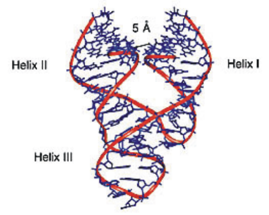


genetic code

RNA ***is the catalytic subunit in supramolecular complexes***



RNA ***as catalyst***



ribozyme

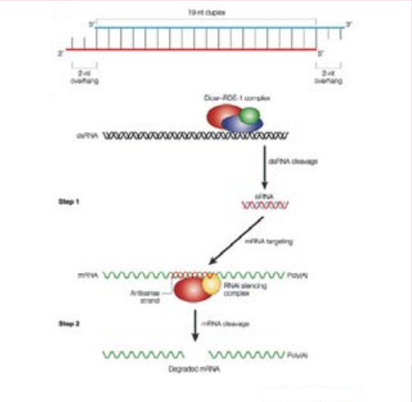
RNA

RNA ***is modified by epigenetic control***

RNA editing

Alternative splicing of messenger **RNA**

RNA ***as regulator of gene expression***



gene silencing by small interfering RNAs

RNA ***as carrier of genetic information***

RNA viruses and retroviruses

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

←

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

Functions of RNA molecules

1. **Introduction**

2. **A few experiments**

3. **Analysing neutral networks**

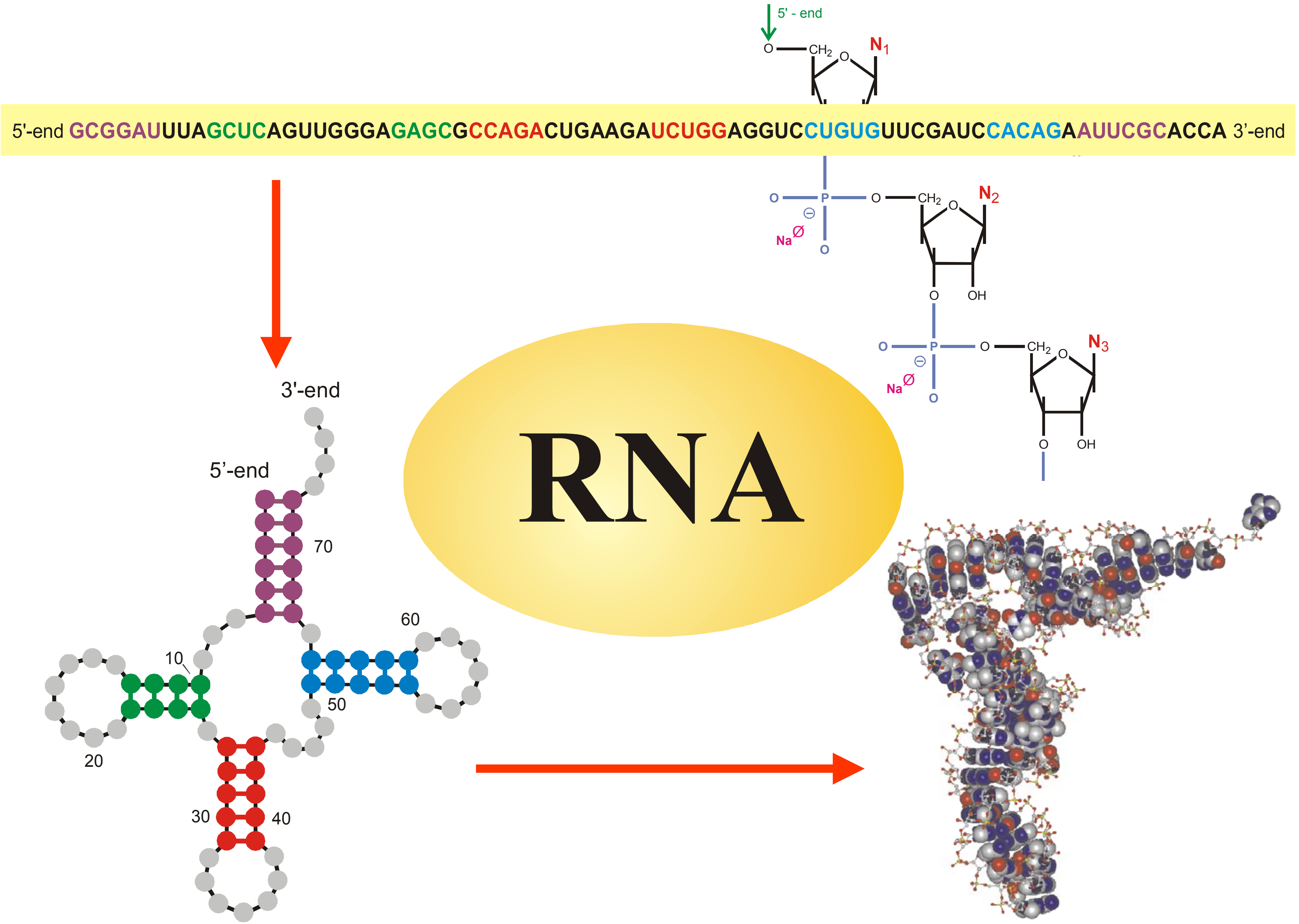
4. **Mechanisms of neutral evolution**

1. **Introduction**

2. A few experiments

3. Analysing neutral networks

4. Mechanisms of neutral evolution



5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end

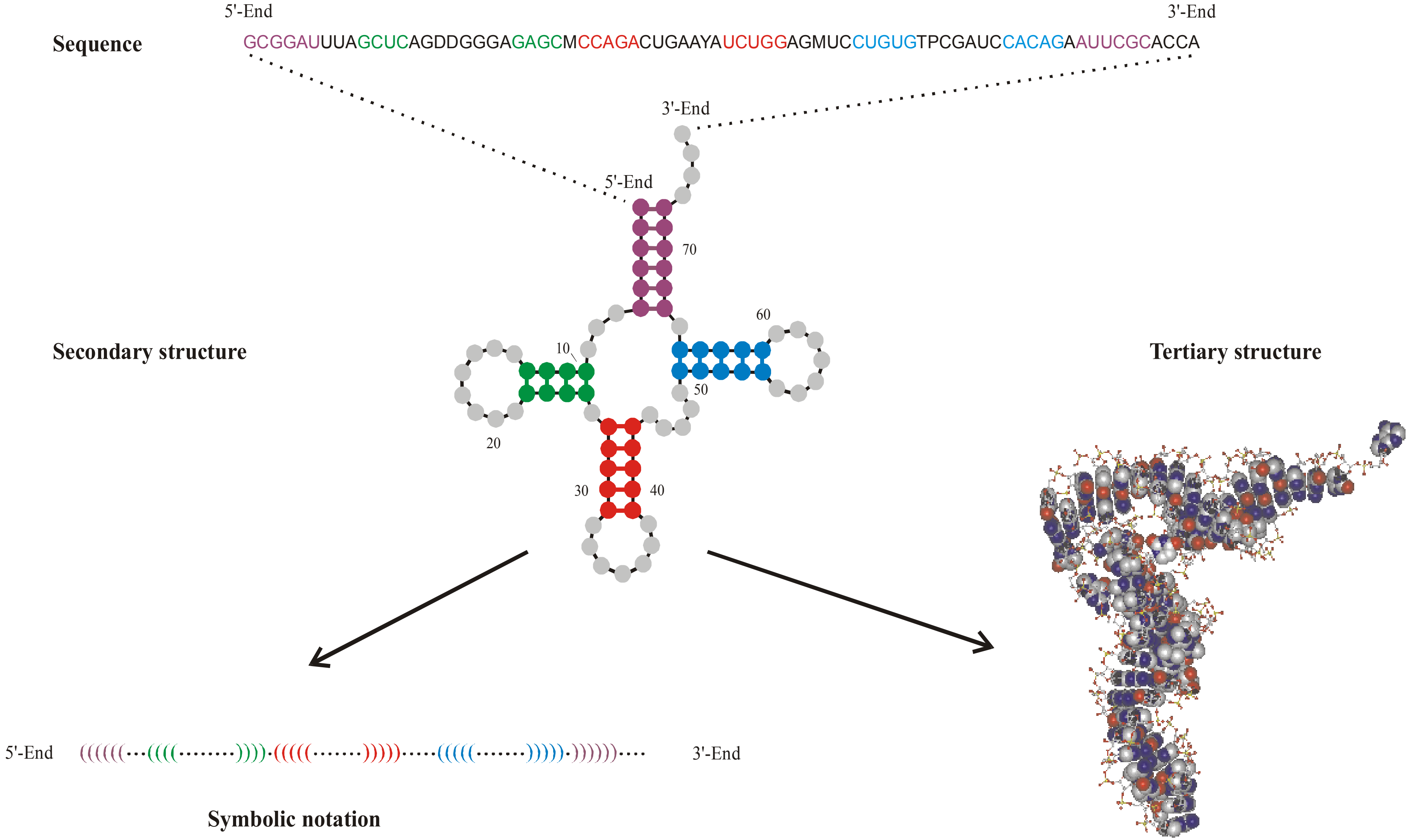
Definition of RNA structure

Definition and physical relevance of RNA secondary structures

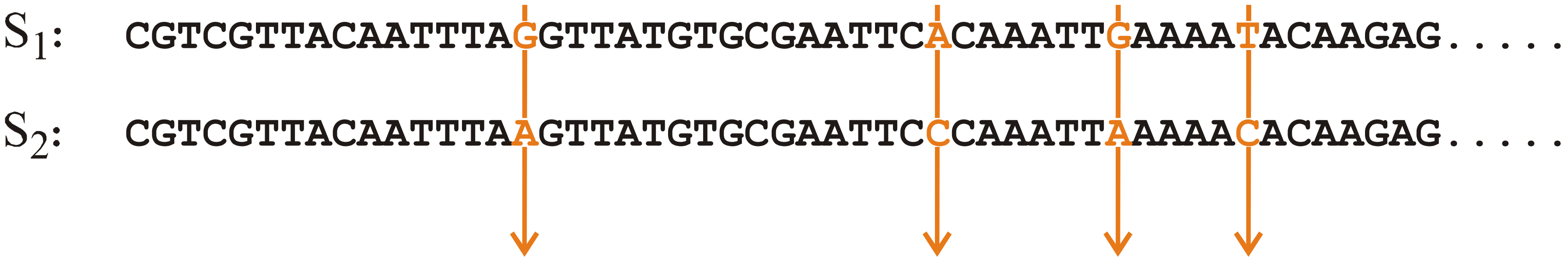
RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots.

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

„Secondary structures are folding intermediates in the formation of full three-dimensional 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 and they are intermediates in RNA folding.



S_1: CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....
S_2: CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAAACACAAGAG.....

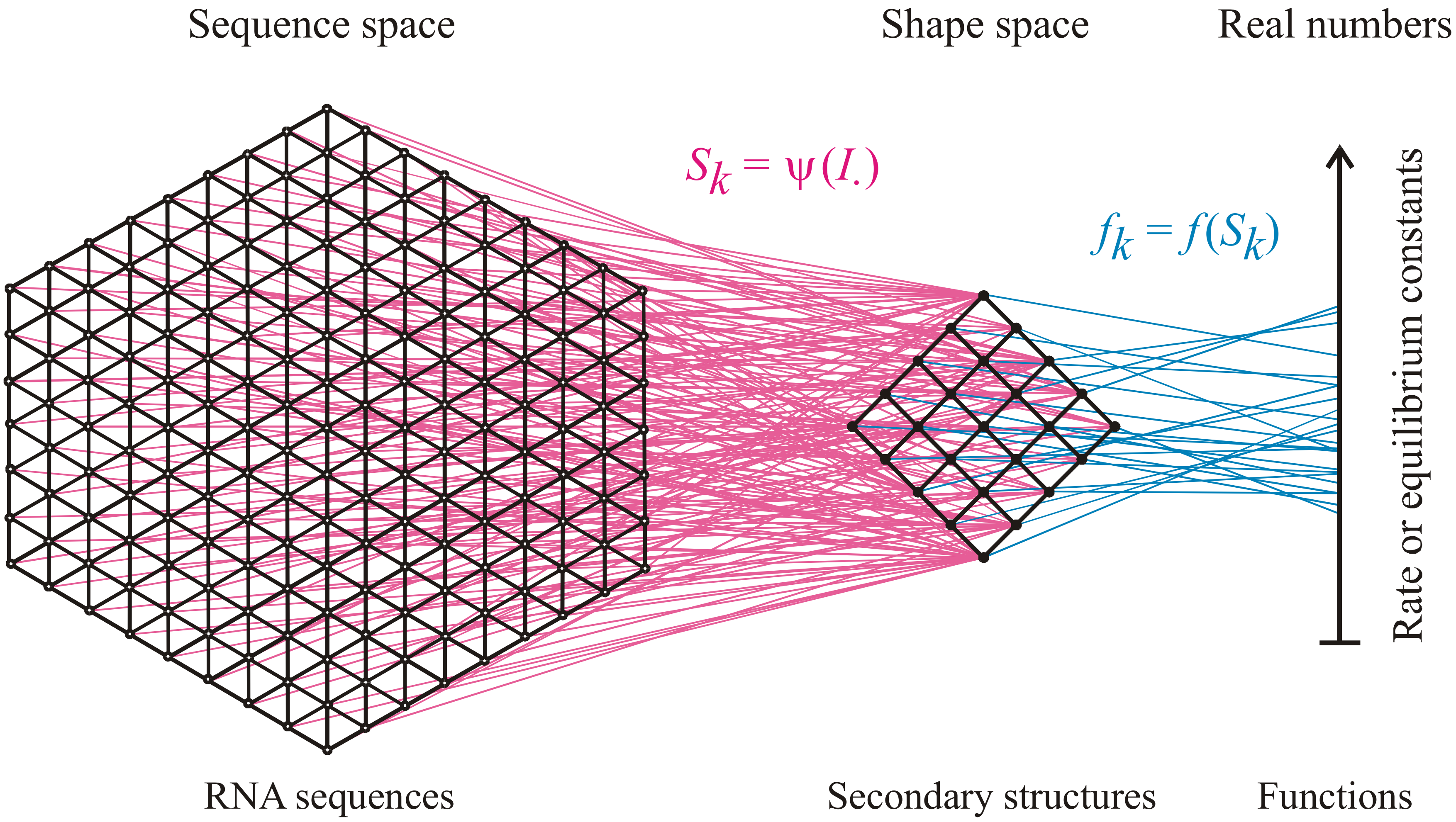
Hamming distance $d_H(S_1,S_2) = 4$

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

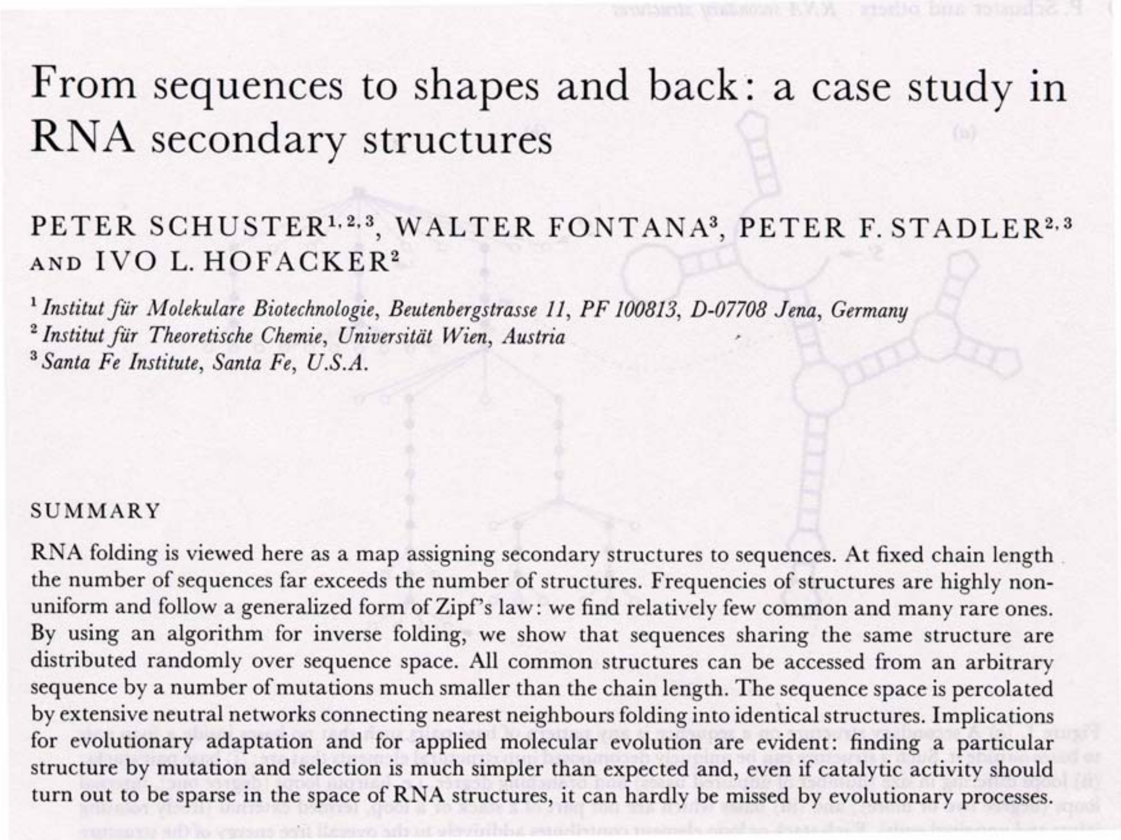
(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

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

The Hamming distance induces a metric in sequence space



Mapping of RNA sequences into structures and structures into functions



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

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

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

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare ones. By using an algorithm for inverse folding, we show that sequences sharing the same structure are distributed randomly over sequence space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected and, even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

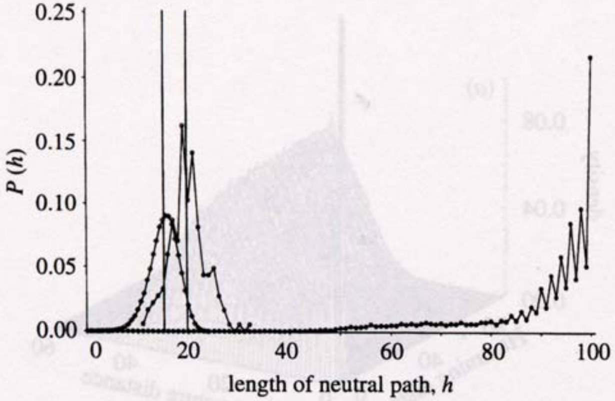
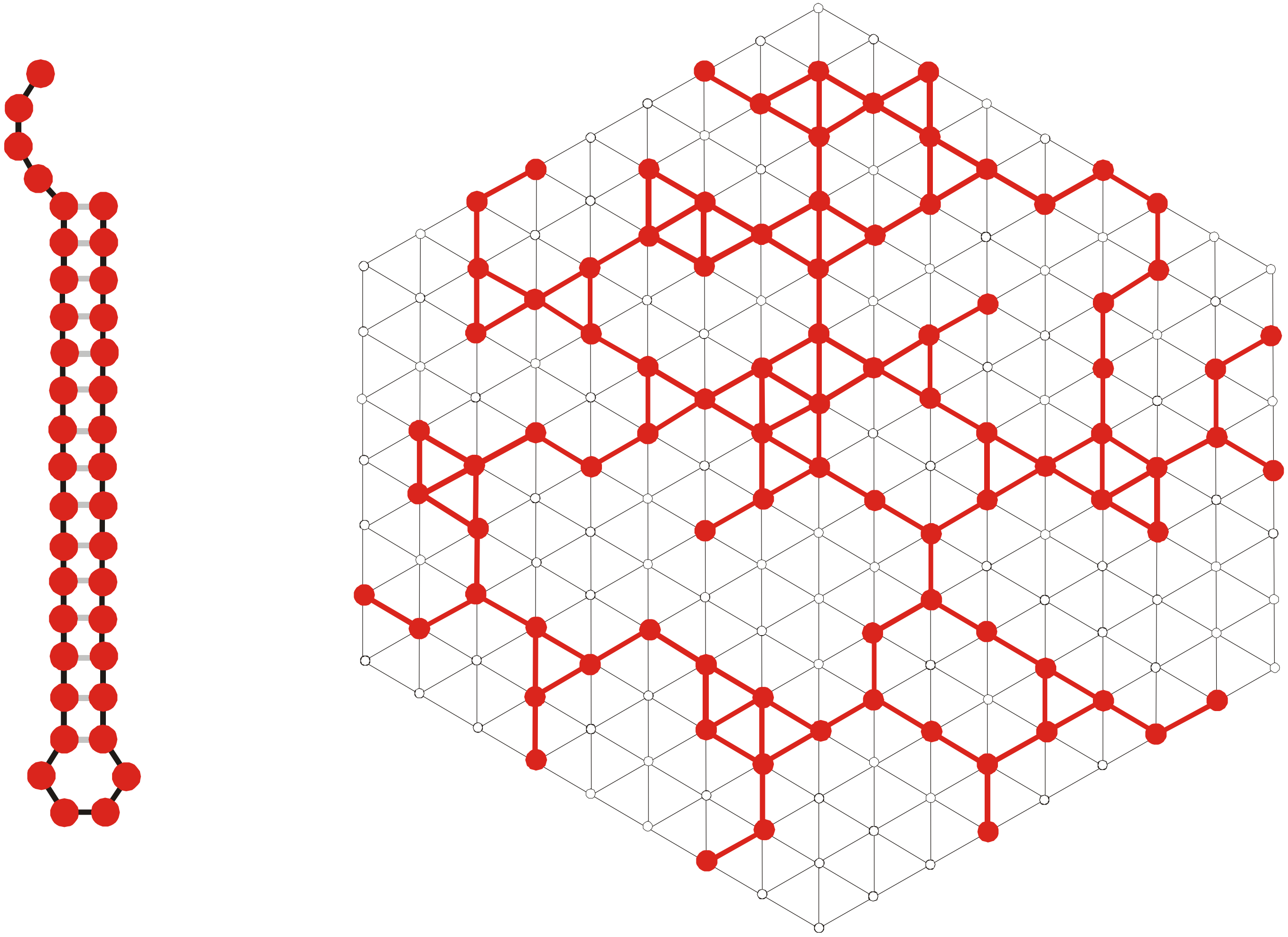


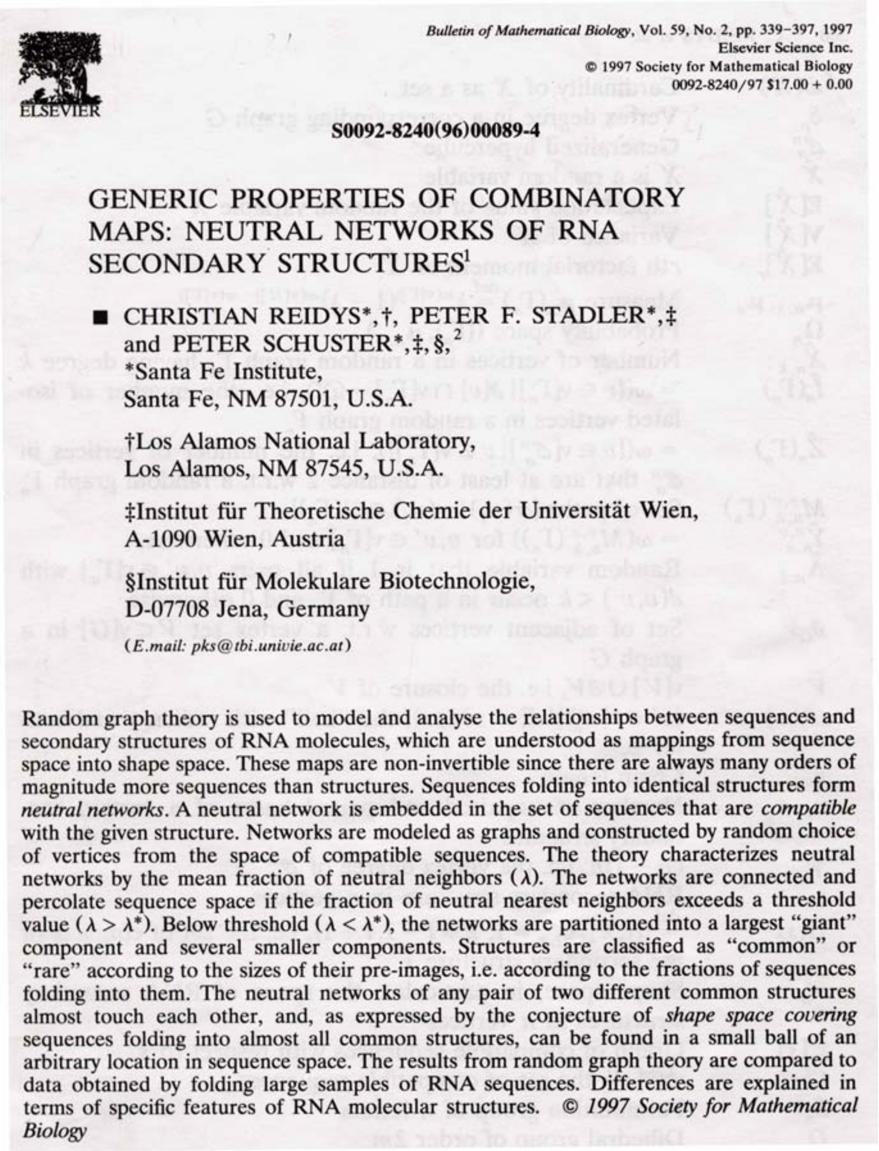
Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and neighbours of Hamming distance 2, resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana *et al.* 1993*a*; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

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

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



A connected neutral network



Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.
© 1997 Society for Mathematical Biology
0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES[1]

■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡ and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

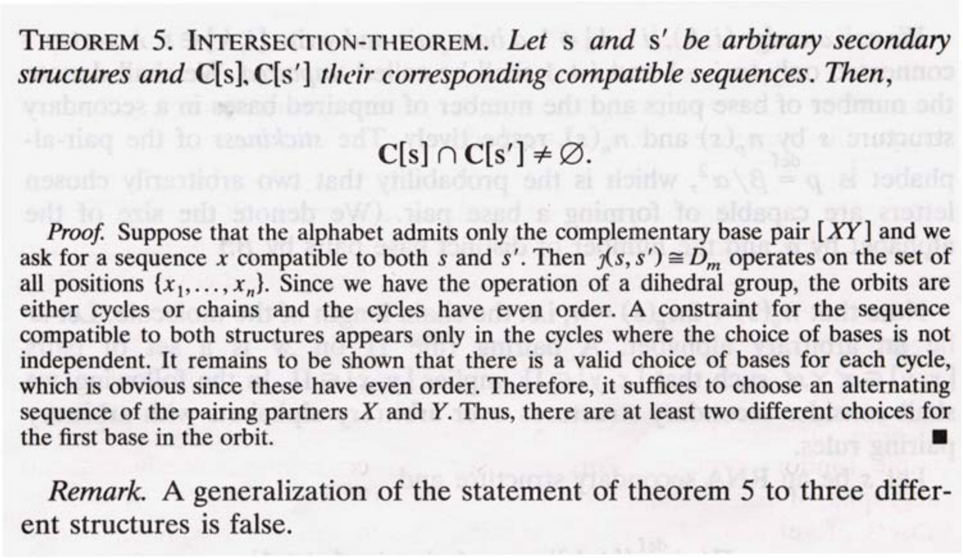
†Los Alamos National Laboratory,
Los Alamos, NM 87545, U.S.A.

‡Institut für Theoretische Chemie der Universität Wien,
A-1090 Wien, Austria

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

(*E.mail: pks@tbi.univie.ac.at*)

Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. Sequences folding into identical structures form *neutral networks*. A neutral network is embedded in the set of sequences that are *compatible* with the given structure. Networks are modeled as graphs and constructed by random choice of vertices from the space of compatible sequences. The theory characterizes neutral networks by the mean fraction of neutral neighbors (λ). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ($\lambda > \lambda^*$). Below threshold ($\lambda < \lambda^*$), the networks are partitioned into a largest "giant" component and several smaller components. Structures are classified as "common" or "rare" according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of *shape space covering* sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. *© 1997 Society for Mathematical Biology*



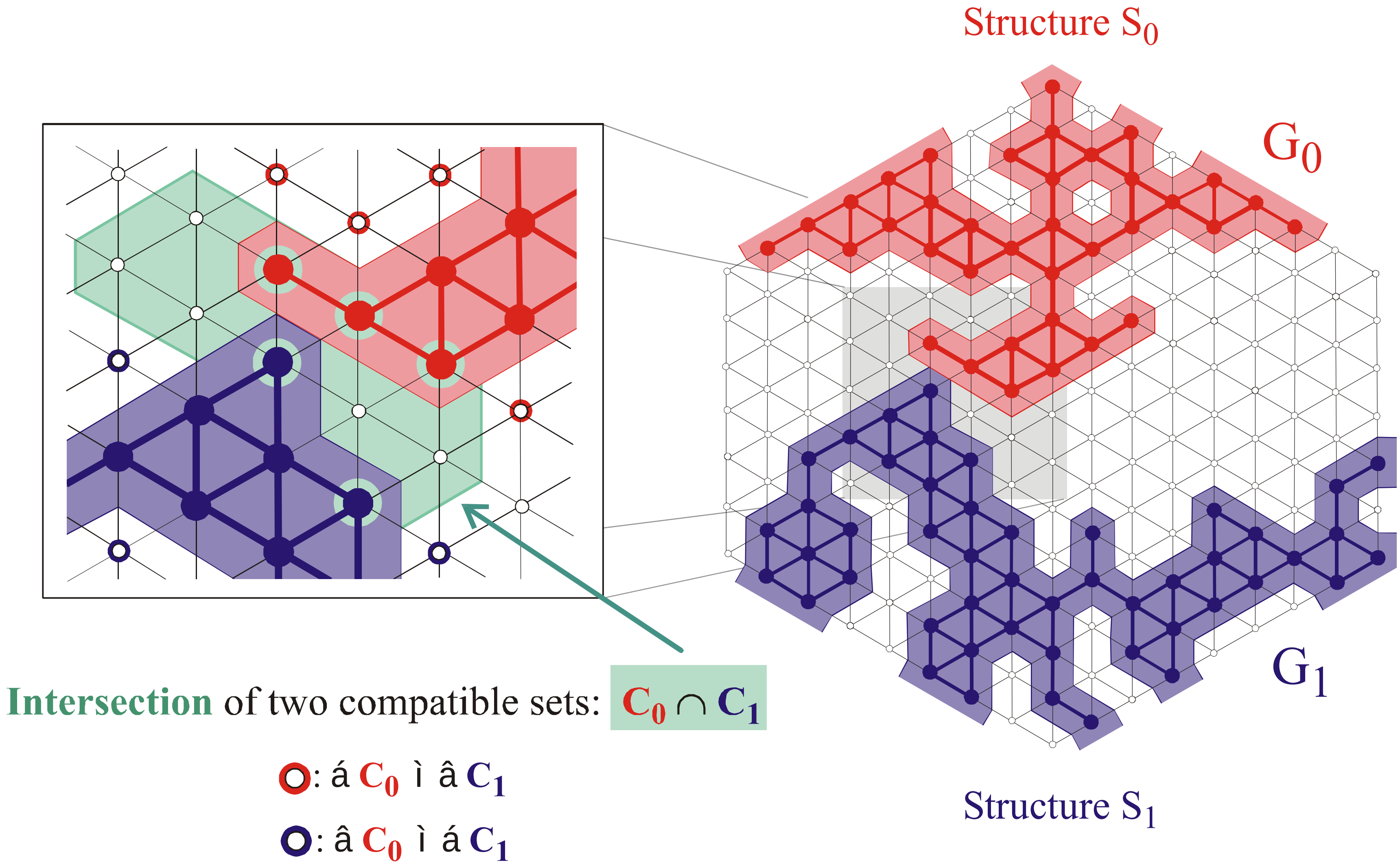
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* C[s], C[s′] *their corresponding compatible sequences. Then,*

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

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 $\mathcal{J}(s,s') \cong D_m$ 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***

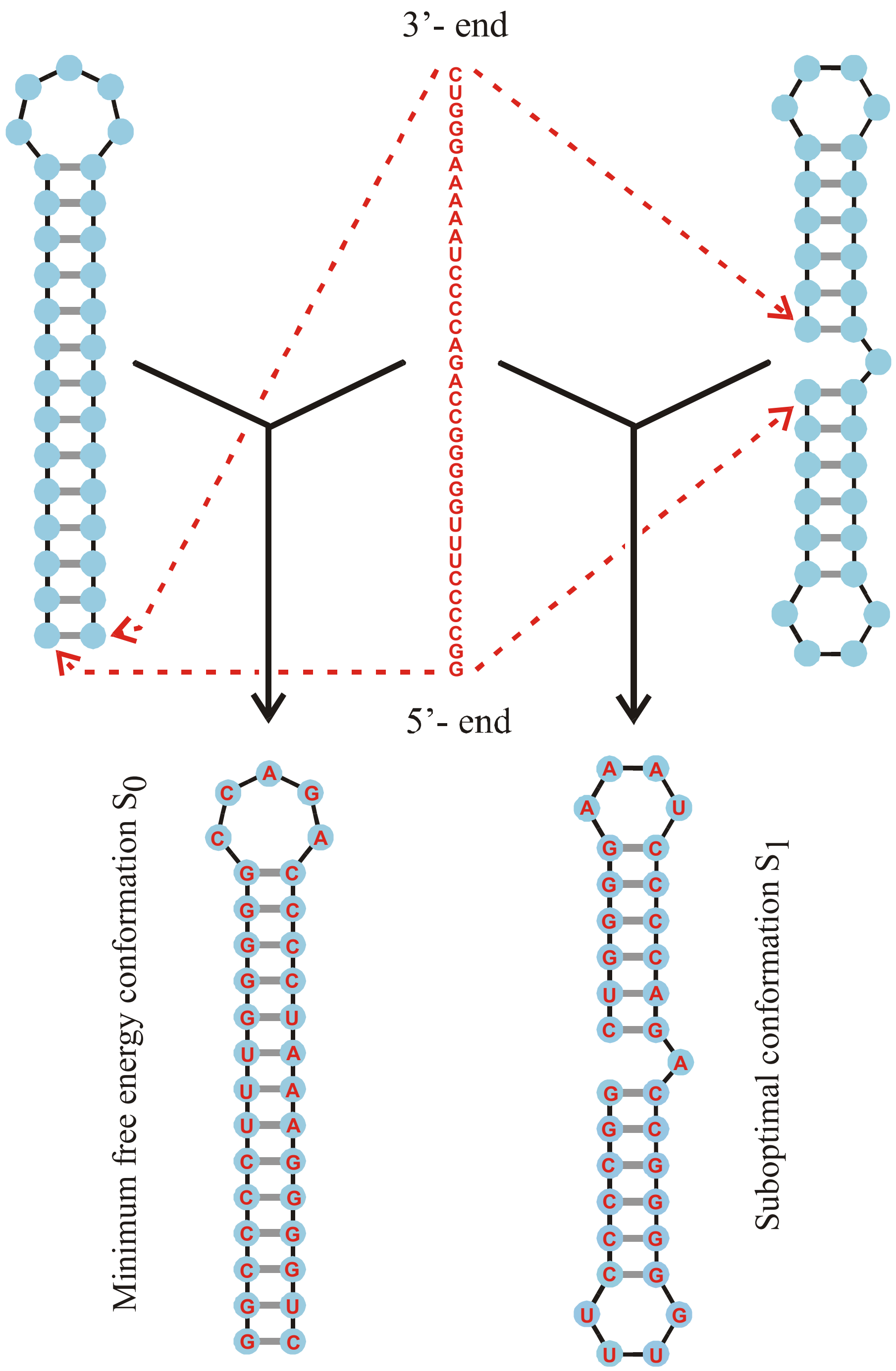


Intersection of two compatible sets: $C_0 \cap C_1$

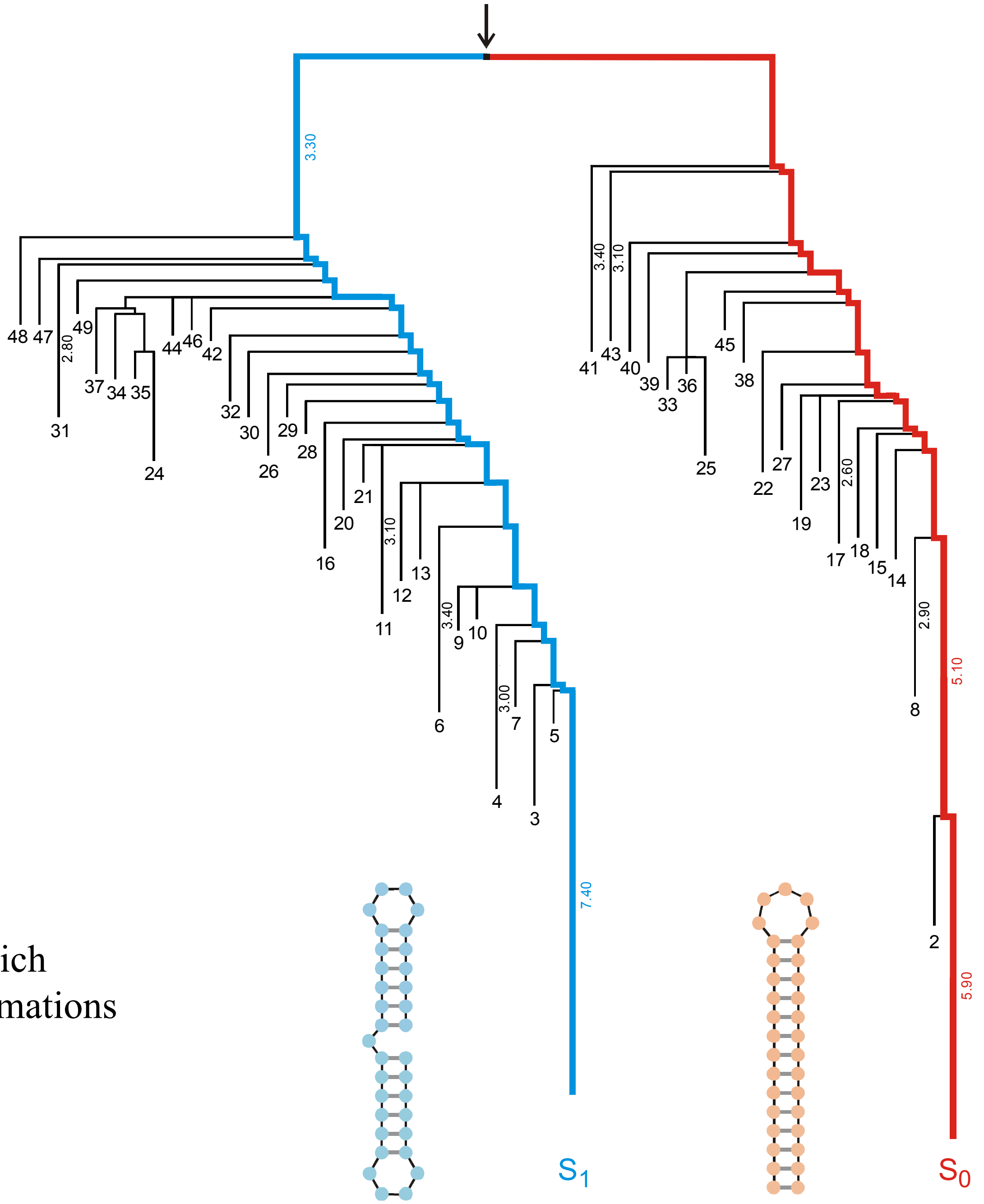
○: á C_0 ì â C_1

○: â C_0 ì á C_1

The intersection of two compatible sets is always non empty: C_0 Ú C_1 â Ù


3’- end
CUGGGAAAAUCCCCAGACCGGGGUUCCCCGG
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



Barrier tree of a sequence which switches between two conformations

1. Introduction

2. **A few experiments**

3. Analysing neutral networks

4. Mechanisms of neutral evolution

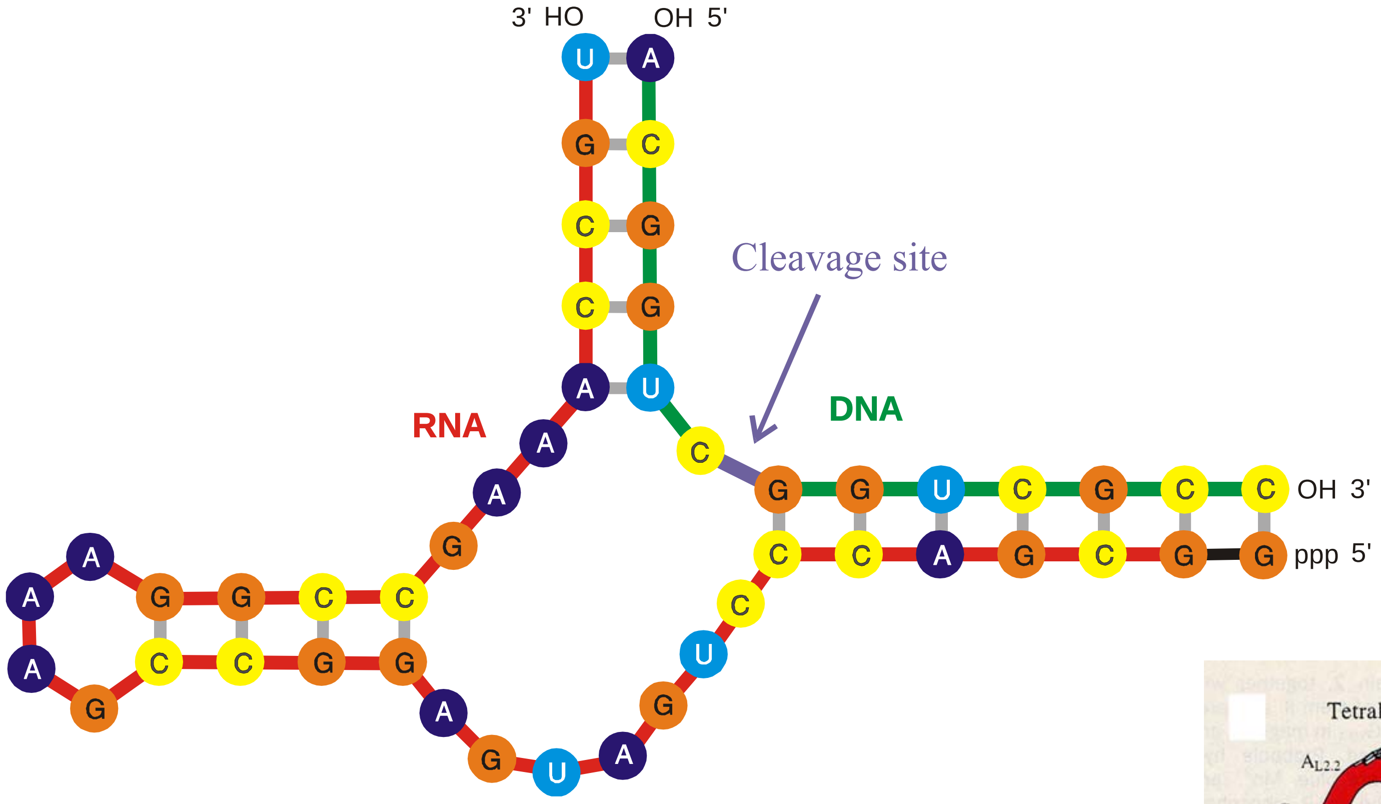
Hammerhead ribozyme – The smallest RNA based catalyst

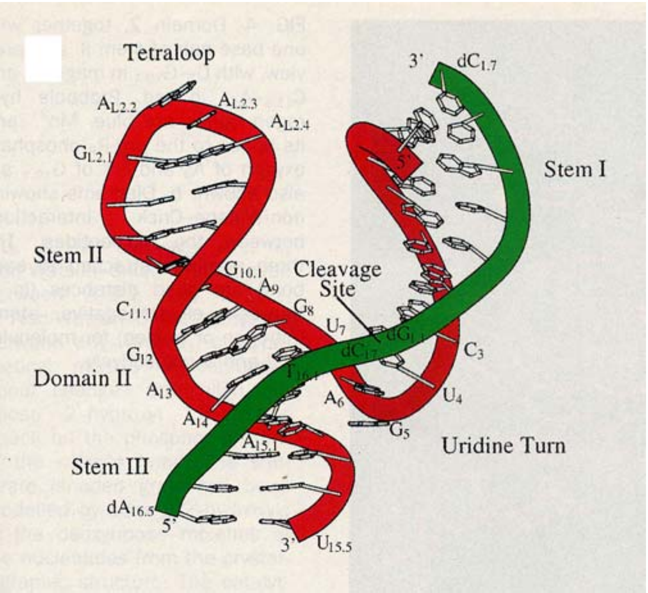
H.W.Pley, K.M.Flaherty, D.B.McKay, ***Three dimensional structure of a hammerhead ribozyme***. Nature **372** (1994), 68-74

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

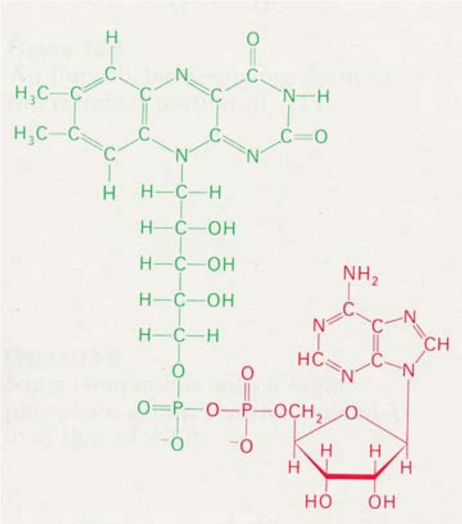




Hammerhead ribozyme: The smallest known catalytically active RNA molecule

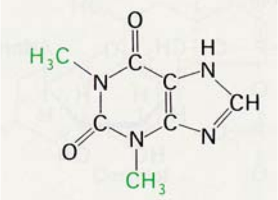
Allosteric effectors:

FMN = flavine mononucleotide



H10 – H12

theophylline

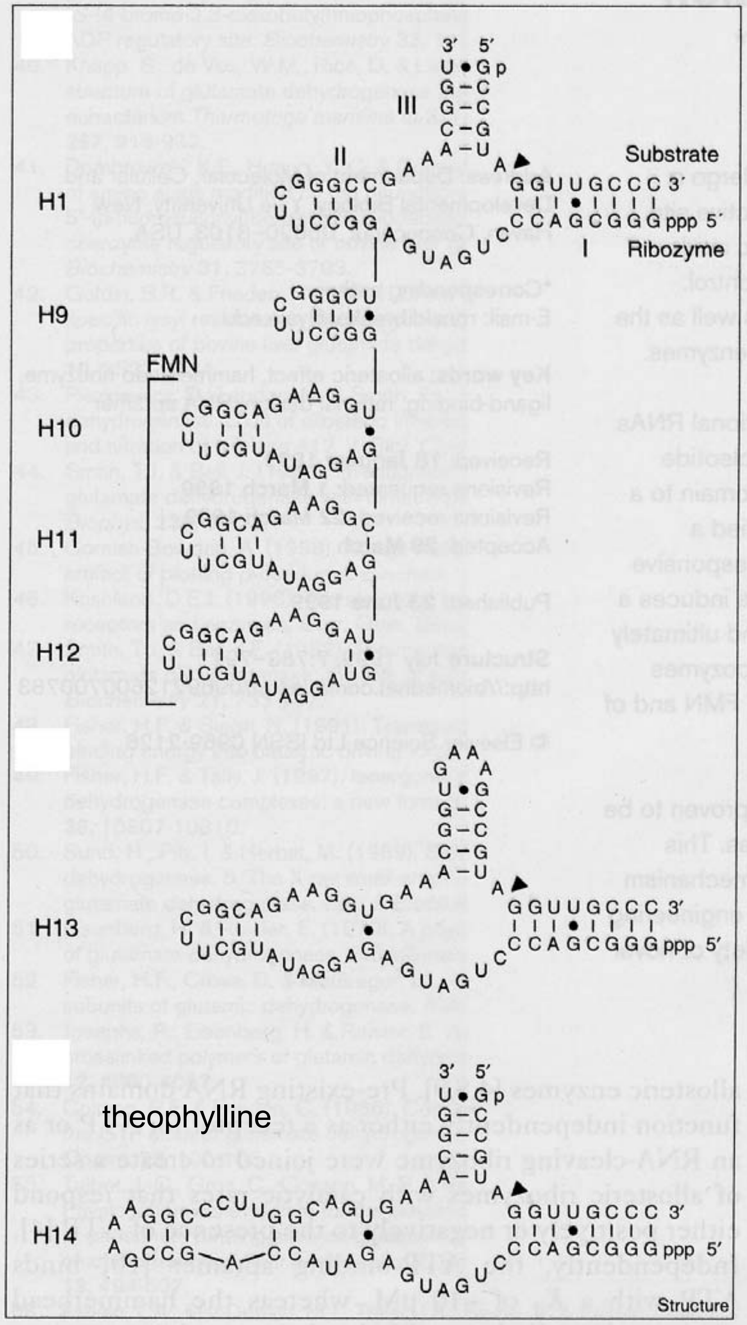


H14

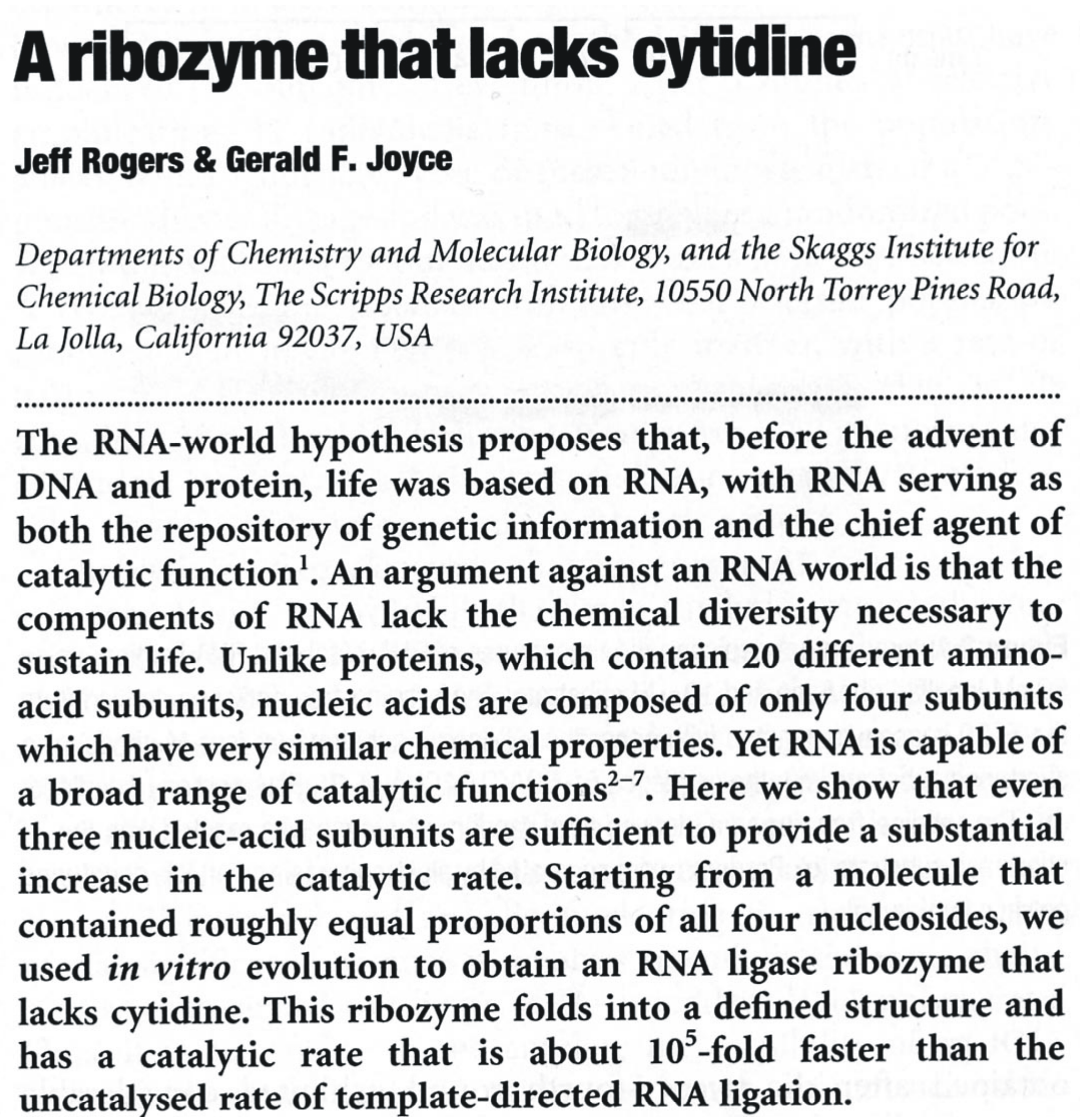
Self-splicing allosteric ribozyme

H13

Hammerhead ribozymes with allosteric effectors



Nature **402**, 323-325, 1999



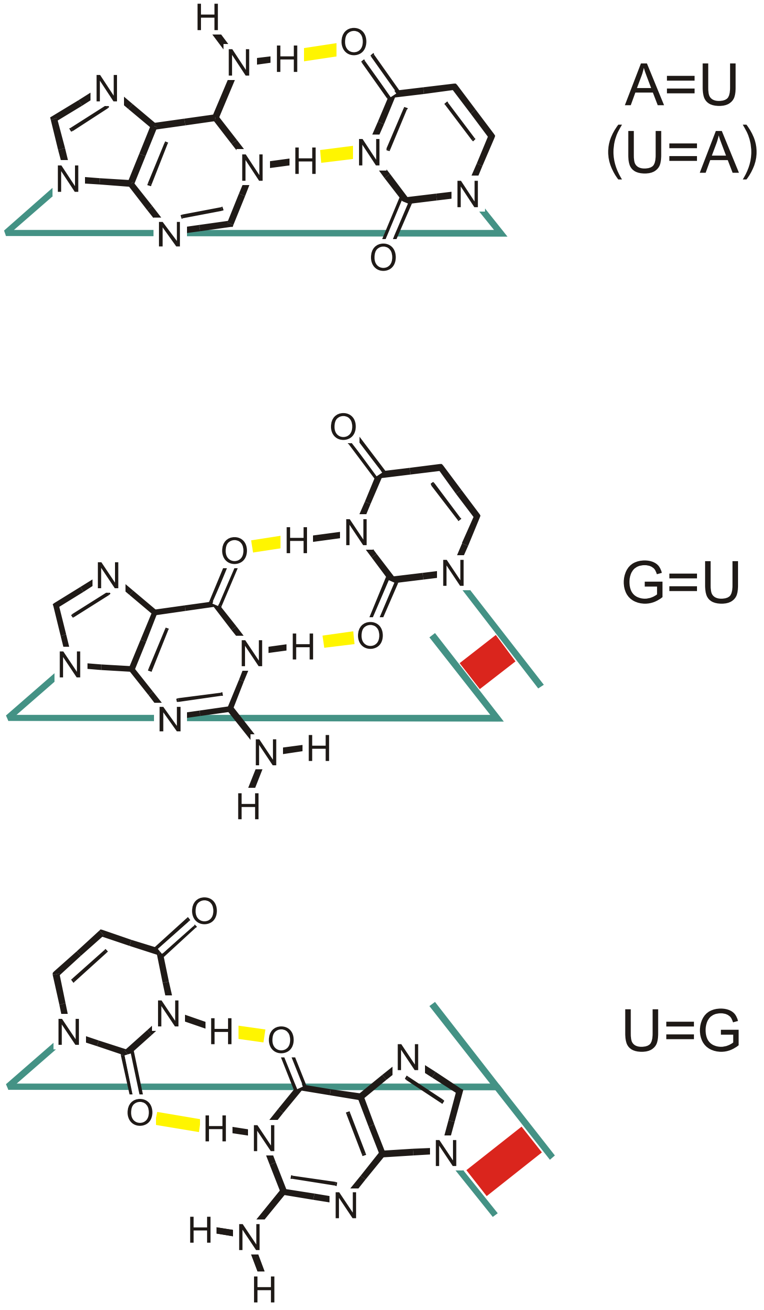
A ribozyme that lacks cytidine

Jeff Rogers & Gerald F. Joyce

Departments of Chemistry and Molecular Biology, and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

The RNA-world hypothesis proposes that, before the advent of DNA and protein, life was based on RNA, with RNA serving as both the repository of genetic information and the chief agent of catalytic function[1]. An argument against an RNA world is that the components of RNA lack the chemical diversity necessary to sustain life. Unlike proteins, which contain 20 different amino-acid subunits, nucleic acids are composed of only four subunits which have very similar chemical properties. Yet RNA is capable of a broad range of catalytic functions[2–7]. Here we show that even three nucleic-acid subunits are sufficient to provide a substantial increase in the catalytic rate. Starting from a molecule that contained roughly equal proportions of all four nucleosides, we used *in vitro* evolution to obtain an RNA ligase ribozyme that lacks cytidine. This ribozyme folds into a defined structure and has a catalytic rate that is about 10^5-fold faster than the uncatalysed rate of template-directed RNA ligation.

Catalytic activity in the **AUG** alphabet



Base pairs in the **AUG** alphabet

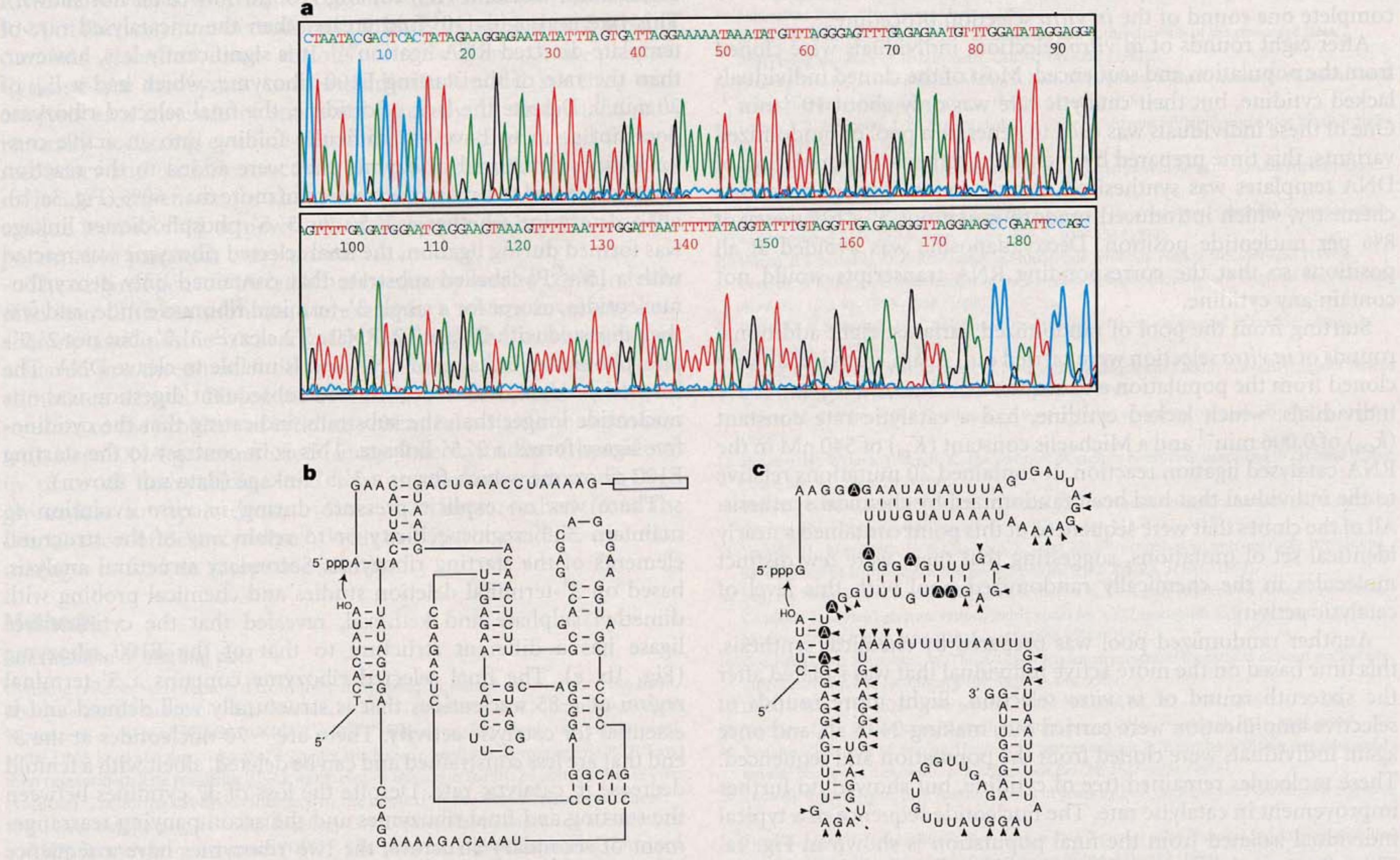
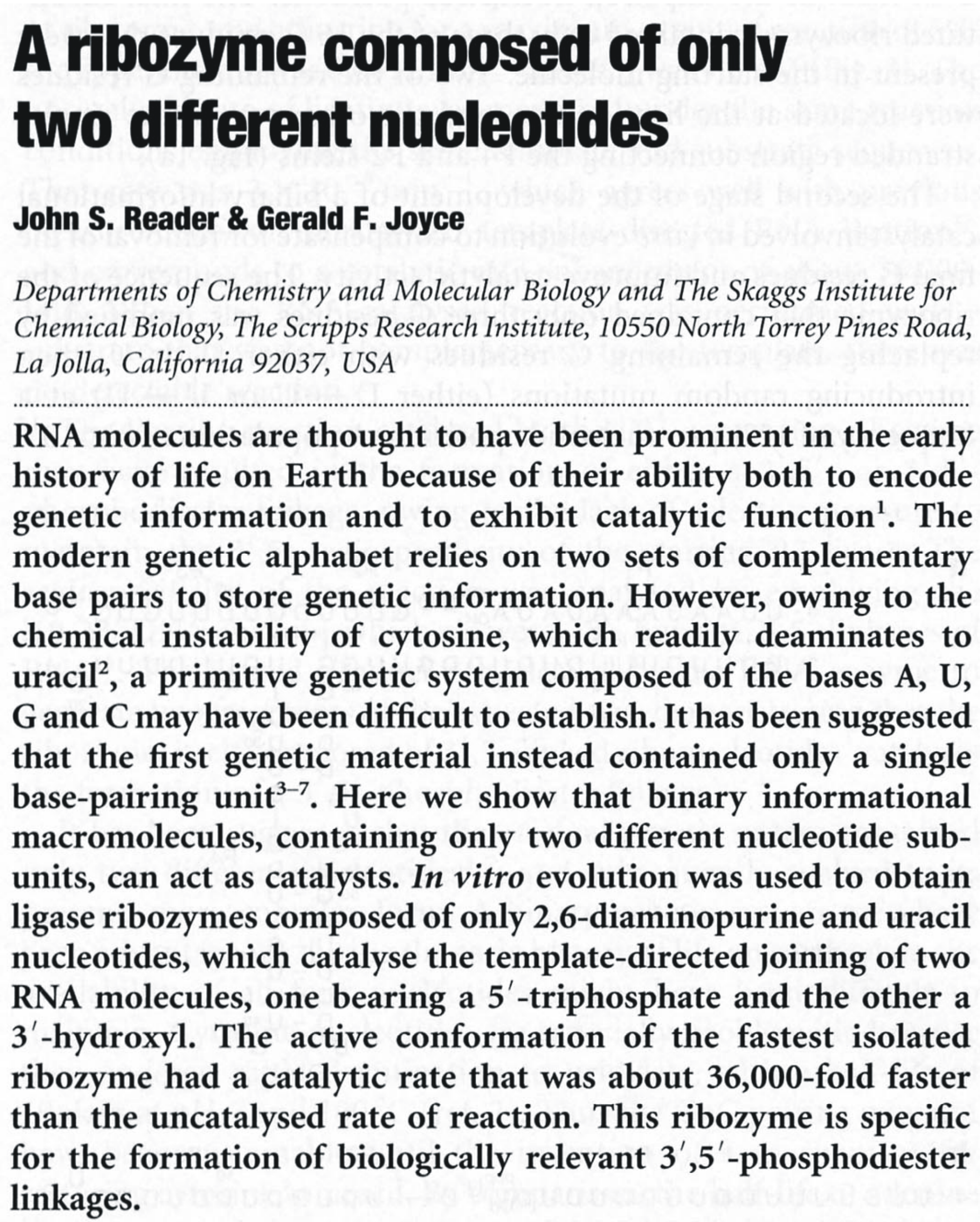

a
CTAATACGACTCACTATAGAAGGAGAATATATTTAGTGATTAGGAAAAATAAATATGTTTAGGGAGTTTGAGAGAATGTTTGGATATAGGAGGA
10 20 30 40 50 60 70 80 90
AGTTTTGAGATGGAATGAGGAAGTAAAGTTTTTAATTTGGATTAATTTTTATAGGTATTTGTAGGTTGAGAAGGGTTAGGAAGCCGAATTCCAGC
100 110 120 130 140 150 160 170 180
b
GAAC-GACUGAACCUAAAAG
5′ pppA
HO
5′
GAAAAGACAAAU
c
AAGGAGAAUAUAUUUA
UUUGUAUAAAU
5′ pppG
HO
AAAGUUUUUAAUUU
3′ GG
5′

Figure 1 Composition of the final selected cytidine-free ribozyme. **a**, Sequence trace showing the lack of cytidines at nucleotide positions 19–173. Positions 2–18 correspond to the T7 promoter sequence and positions 174–188 correspond to the downstream vector sequence (pCR 2.1). Automated sequencing was carried out using an ABI model 373 DNA sequencer and was confirmed by manual sequencing of both strands (data not shown). **b**, Secondary structure of the starting ribozyme (E100) based on that of the class I ligase[10]. Box indicates the primer binding site at the 3′ end of the ribozyme. **c**, Secondary structure of the final selected ribozyme based on chemical modification of unpaired adenosine and guanosine residues (carat marks), carried out in the absence of substrate. Highlighted adenosine residues blocked catalytic activity when methylated at N1. Dashed line indicates the site of the largest 3′-terminal deletion that was compatible with catalytic activity.

Nature **420**, 841-844, 2002



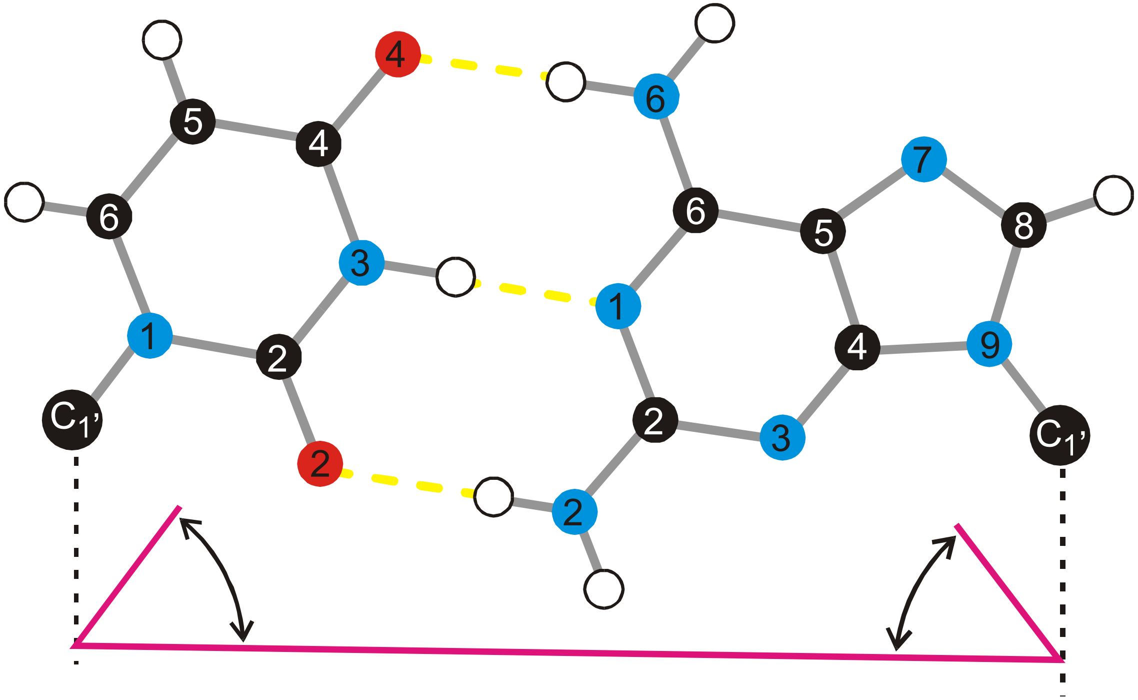
A ribozyme composed of only two different nucleotides

John S. Reader & Gerald F. Joyce

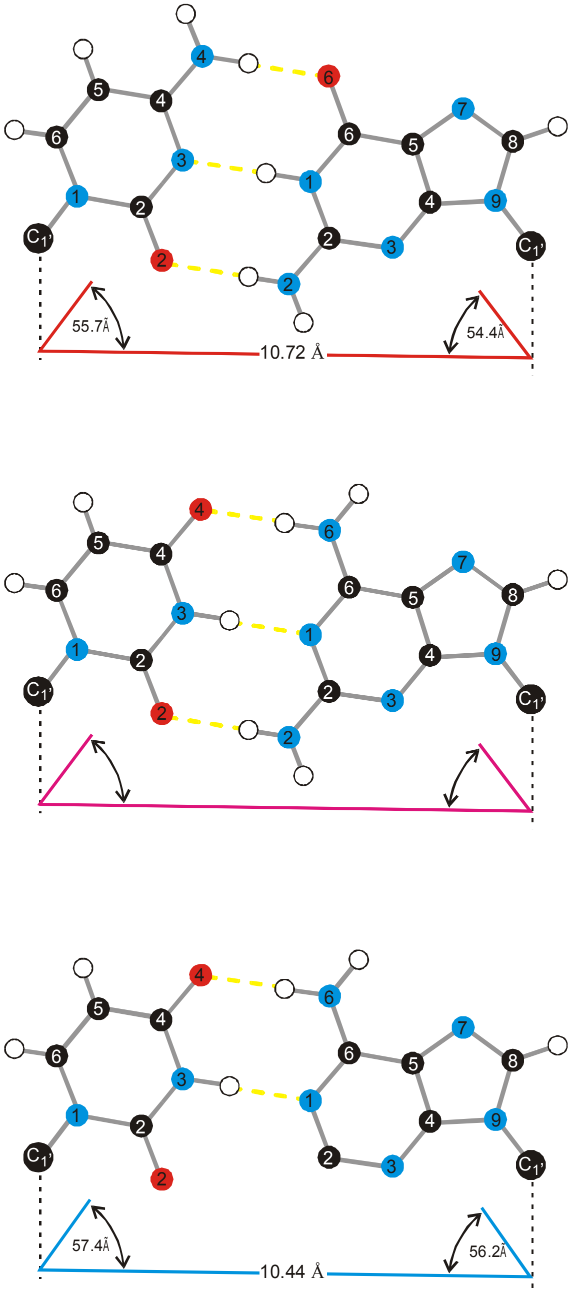
Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

RNA molecules are thought to have been prominent in the early history of life on Earth because of their ability both to encode genetic information and to exhibit catalytic function[1]. The modern genetic alphabet relies on two sets of complementary base pairs to store genetic information. However, owing to the chemical instability of cytosine, which readily deaminates to uracil[2], a primitive genetic system composed of the bases A, U, G and C may have been difficult to establish. It has been suggested that the first genetic material instead contained only a single base-pairing unit[3–7]. Here we show that binary informational macromolecules, containing only two different nucleotide subunits, can act as catalysts. *In vitro* evolution was used to obtain ligase ribozymes composed of only 2,6-diaminopurine and uracil nucleotides, which catalyse the template-directed joining of two RNA molecules, one bearing a 5′-triphosphate and the other a 3′-hydroxyl. The active conformation of the fastest isolated ribozyme had a catalytic rate that was about 36,000-fold faster than the uncatalysed rate of reaction. This ribozyme is specific for the formation of biologically relevant 3′,5′-phosphodiester linkages.

Catalytic activity in the **DU** alphabet



The 2,6-diamino purine – uracil, **DU**, base pair



G Í C

D Í U

A = U

Three Watson-Crick type base pairs

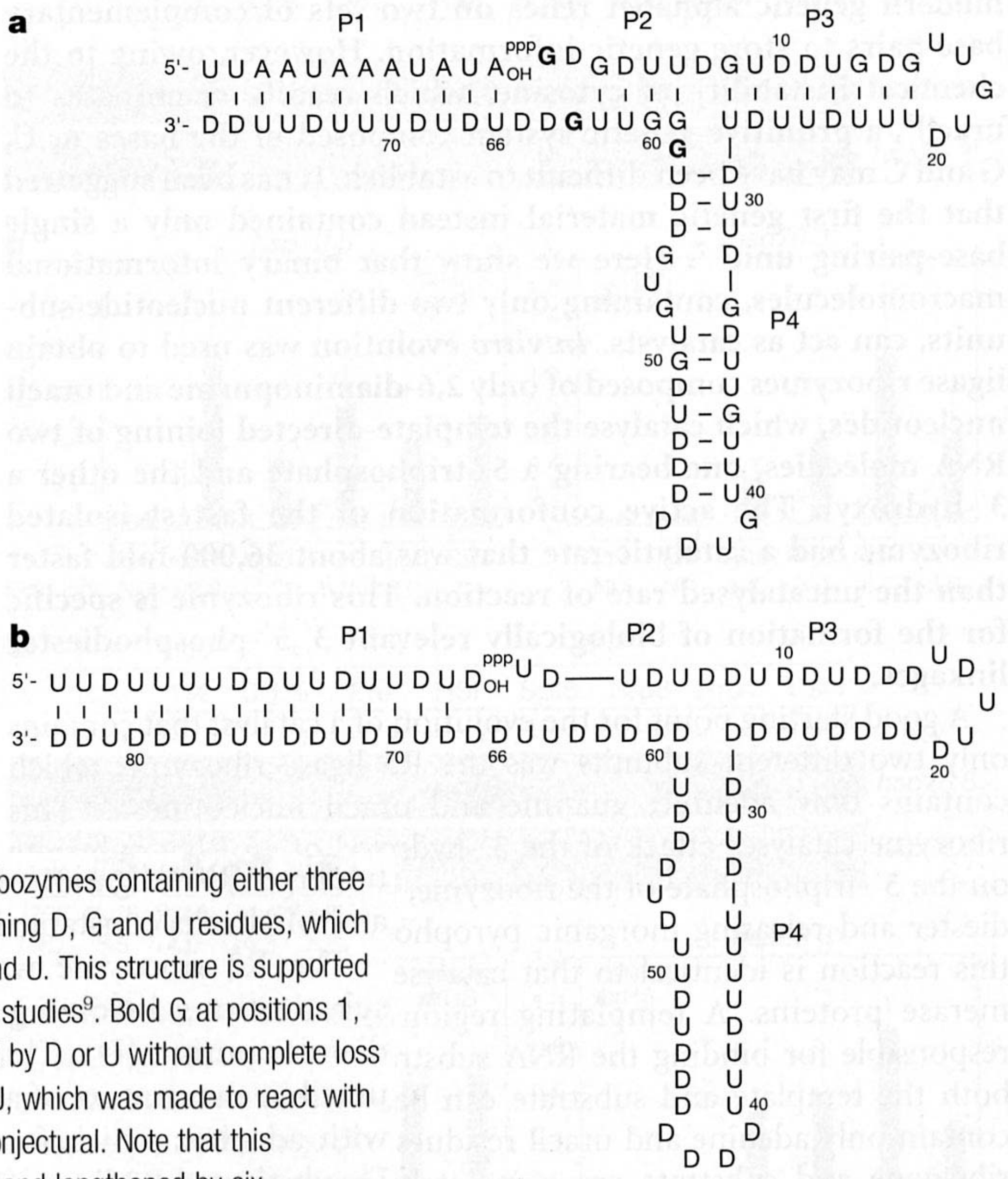
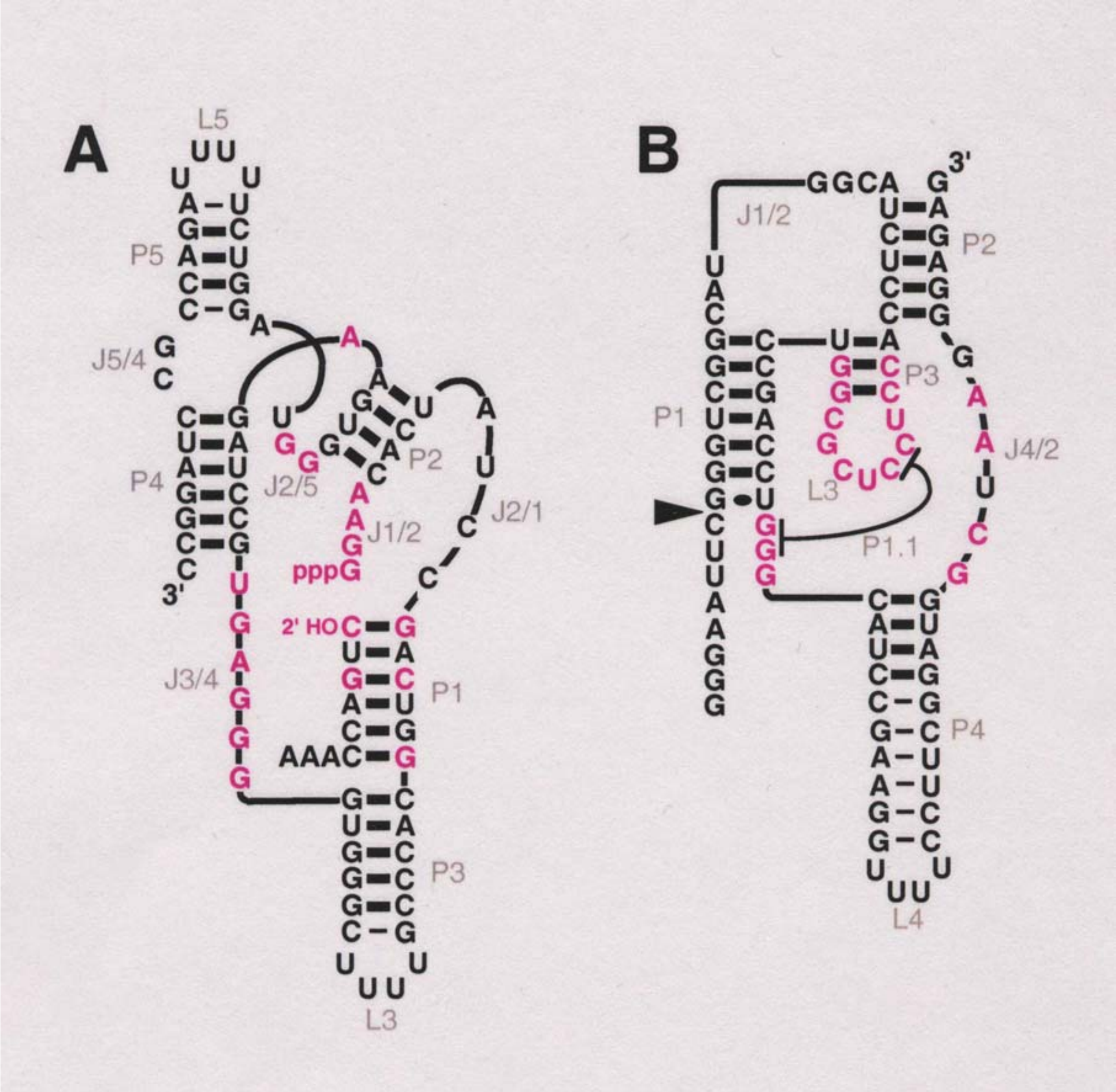

a
P1
P2
P3
P4
5'- U U A A U A A A U A U A
OH
ppp
G D G D U U D G U D D U G D G
3'- D D U U D U U U D U D U D D G U U G G
70
66
60
10
20
30
40
50
b
5'- U U D U U U U D D U U D U U D U D
3'- D D U D D D D U U D D U D D U D D D U U D D D D
80
U D —— U D U D D U D D U D D D

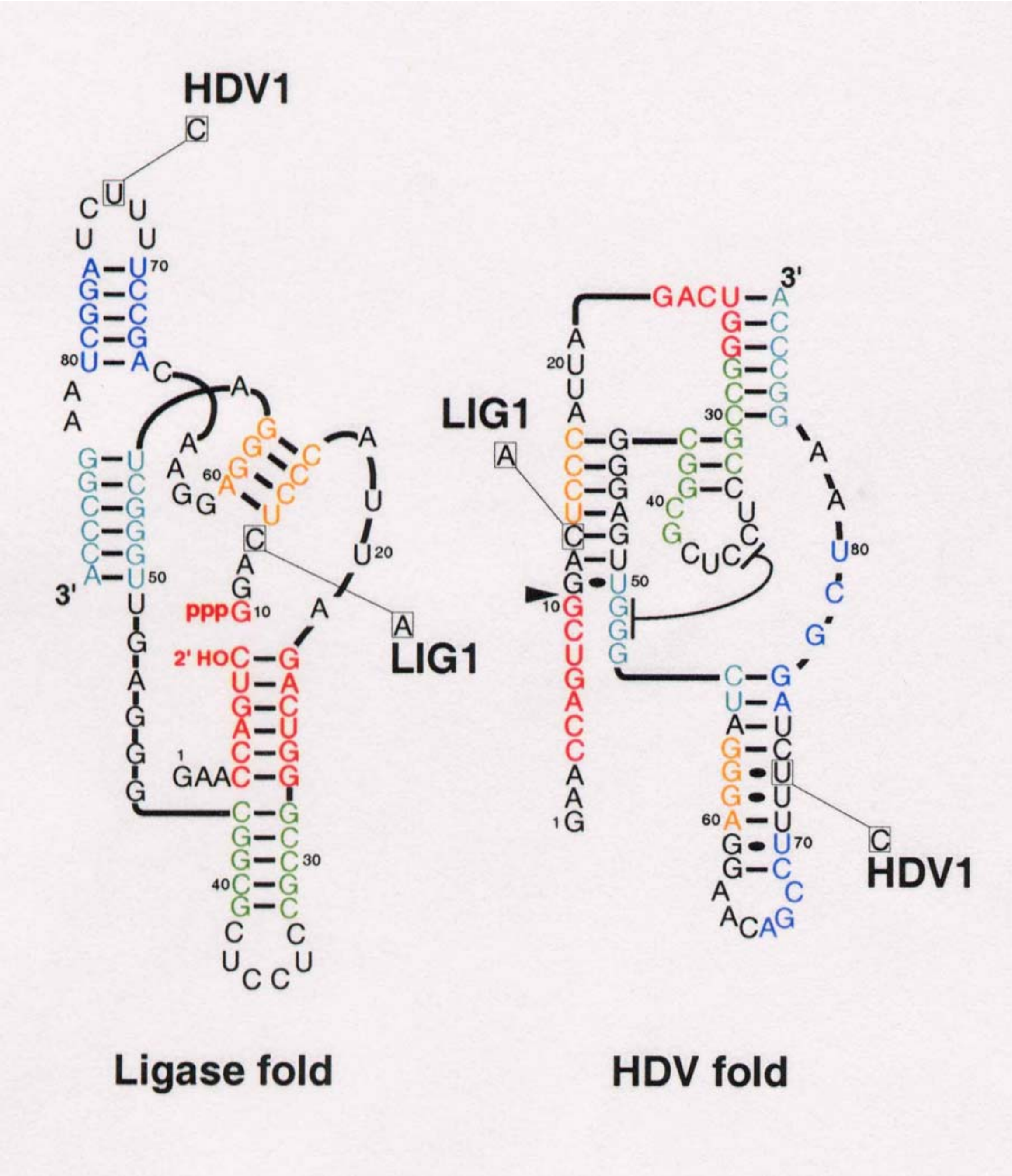
Figure 1 Sequence and secondary structure of ligase ribozymes containing either three or two different nucleotide subunits. **a**, Ribozyme containing D, G and U residues, which was made to react with a substrate containing only A and U. This structure is supported by chemical modification and site-directed mutagenesis studies[9]. Bold G at positions 1, 58 and 63 indicates residues that could not be replaced by D or U without complete loss of catalytic activity. **b**, Ribozyme containing only D and U, which was made to react with a substrate containing only D and U. This structure is conjectural. Note that this molecule is shortened by one nucleotide at the 5′ end and lengthened by six nucleotides at the 3′ end compared with the ribozyme shown in **a**.

A ribozyme switch

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

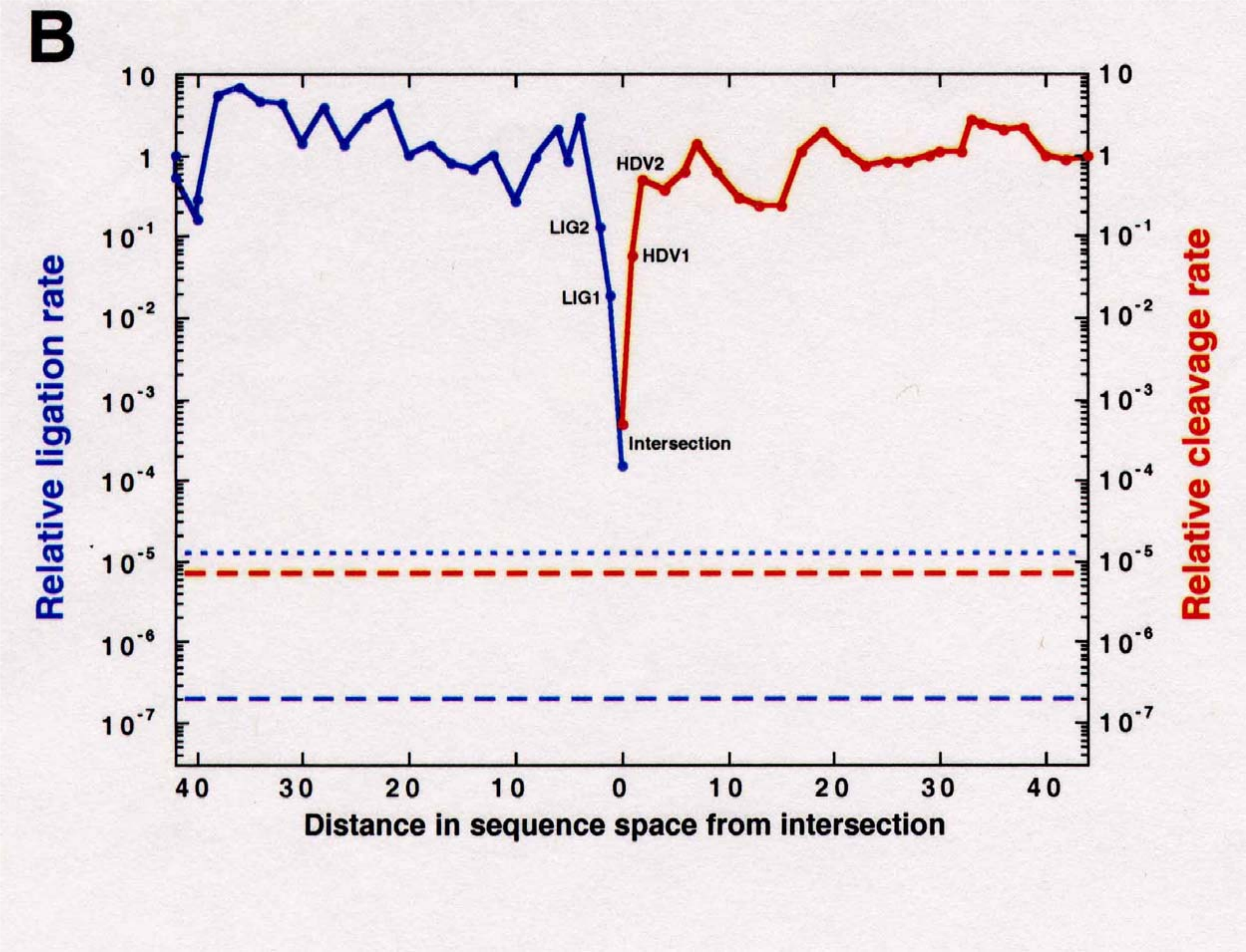


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)



The sequence at the ***intersection***:

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



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

1. Introduction

2. A few experiments

3. **Analysing neutral networks**

4. Mechanisms of neutral 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.)

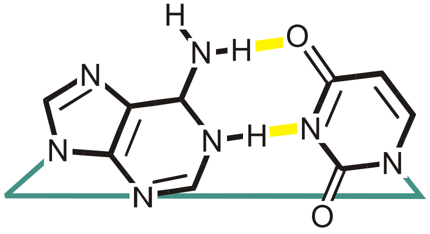
I.L.Hofacker, W. Fontana, P.F.Stadler, L.S.Bonhoeffer, M.Tacker, and P. Schuster. *Mh.Chem*. **125**:167-188 (1994)

Statistics of RNA structures from random sequences over different nucleotide alphabets

Walter Fontana, Danielle A. M. Konings, Peter F. Stadler, Peter Schuster, ***Statistics of RNA secondary structures***. Biopolymers **33** (1993), 1389-1404

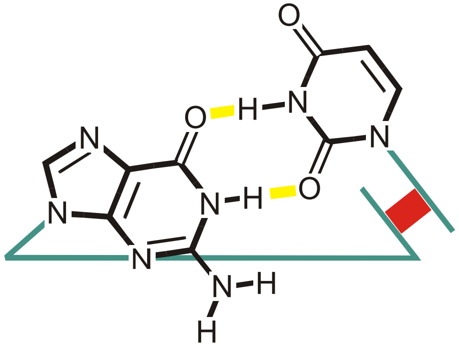
Peter Schuster, Walter Fontana, Peter F. Stadler, Ivo L. Hofacker, ***From sequences to shapes and back: A case study in RNA secondary structures***. Proc.Roy.Soc.London **B 255** (1994), 279-284

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

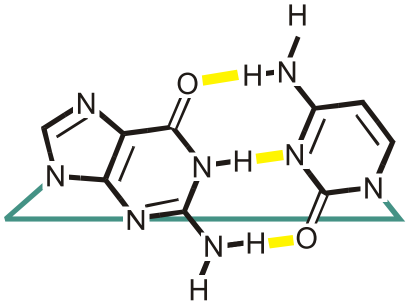


A=U

(U=A)

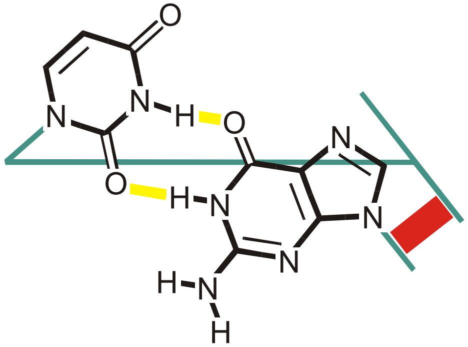


G=U



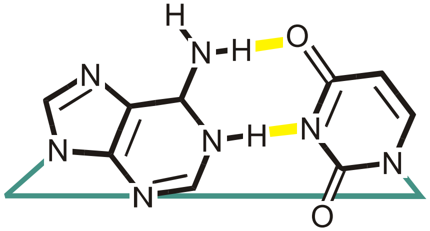
GÍC

(CÍG)



U=G

The six base pairing alphabets built from natural nucleotides **A**, **U**, **G**, and **C**



A=U

(U=A)

The six base pairing alphabets built from natural nucleotides **A**, **U**, **G**, and **C**

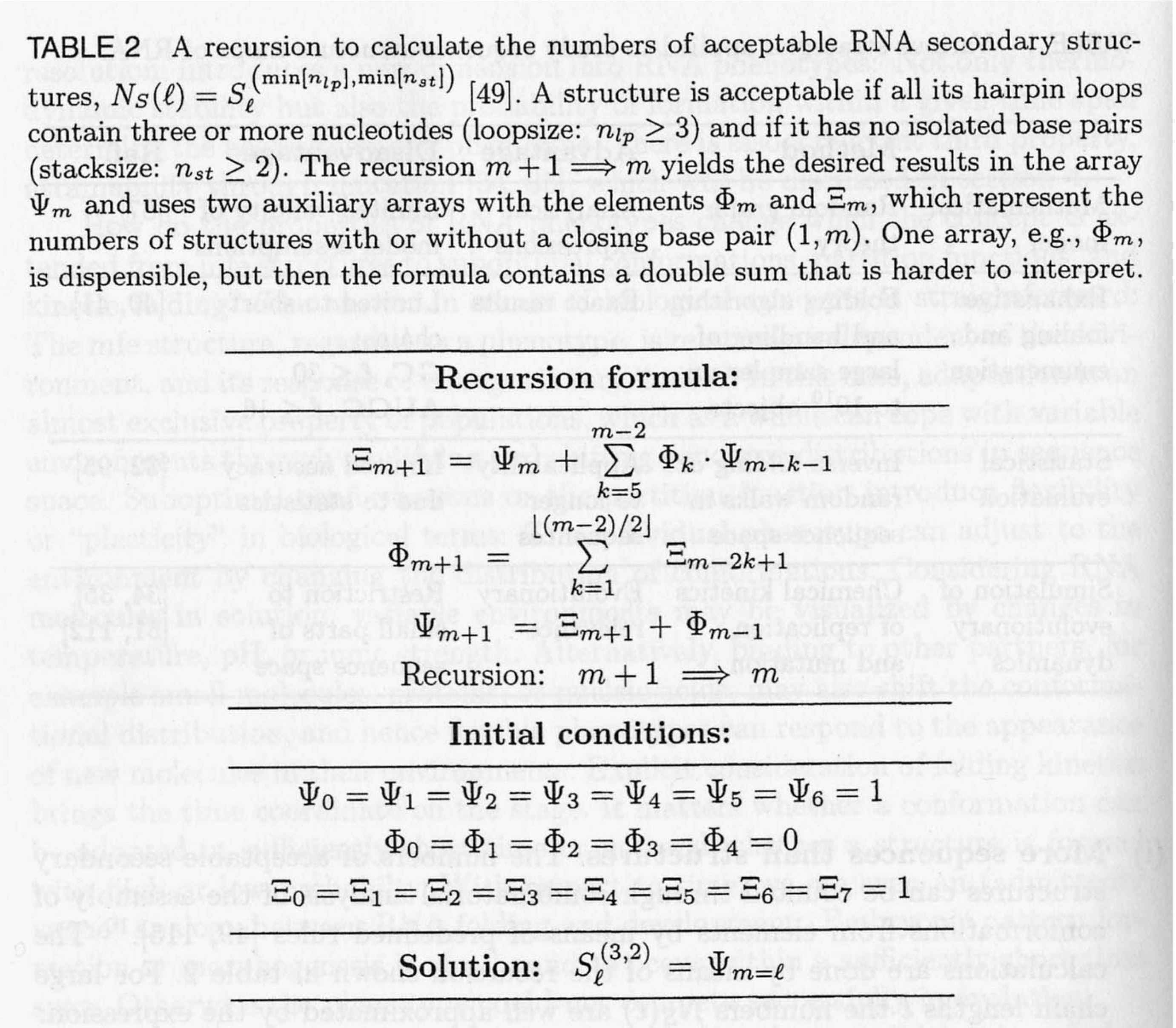


TABLE 2 A recursion to calculate the numbers of acceptable RNA secondary structures, $N_S(\ell) = S_\ell^{(\min[n_{lp}],\min[n_{st}])}$ [49]. A structure is acceptable if all its hairpin loops contain three or more nucleotides (loopsize: $n_{lp} \geq 3$) and if it has no isolated base pairs (stacksize: $n_{st} \geq 2$). The recursion $m+1 \Longrightarrow m$ yields the desired results in the array Ψ_m and uses two auxiliary arrays with the elements Φ_m and Ξ_m, which represent the numbers of structures with or without a closing base pair $(1, m)$. One array, e.g., Φ_m, is dispensible, but then the formula contains a double sum that is harder to interpret.

Recursion formula:

$$\Xi_{m+1} = \Psi_m + \sum_{k=5}^{m-2} \Phi_k \cdot \Psi_{m-k-1}$$

$$\Phi_{m+1} = \sum_{k=1}^{\lfloor (m-2)/2 \rfloor} \Xi_{m-2k+1}$$

$$\Psi_{m+1} = \Xi_{m+1} + \Phi_{m-1}$$

Recursion: $m+1 \Longrightarrow m$

Initial conditions:

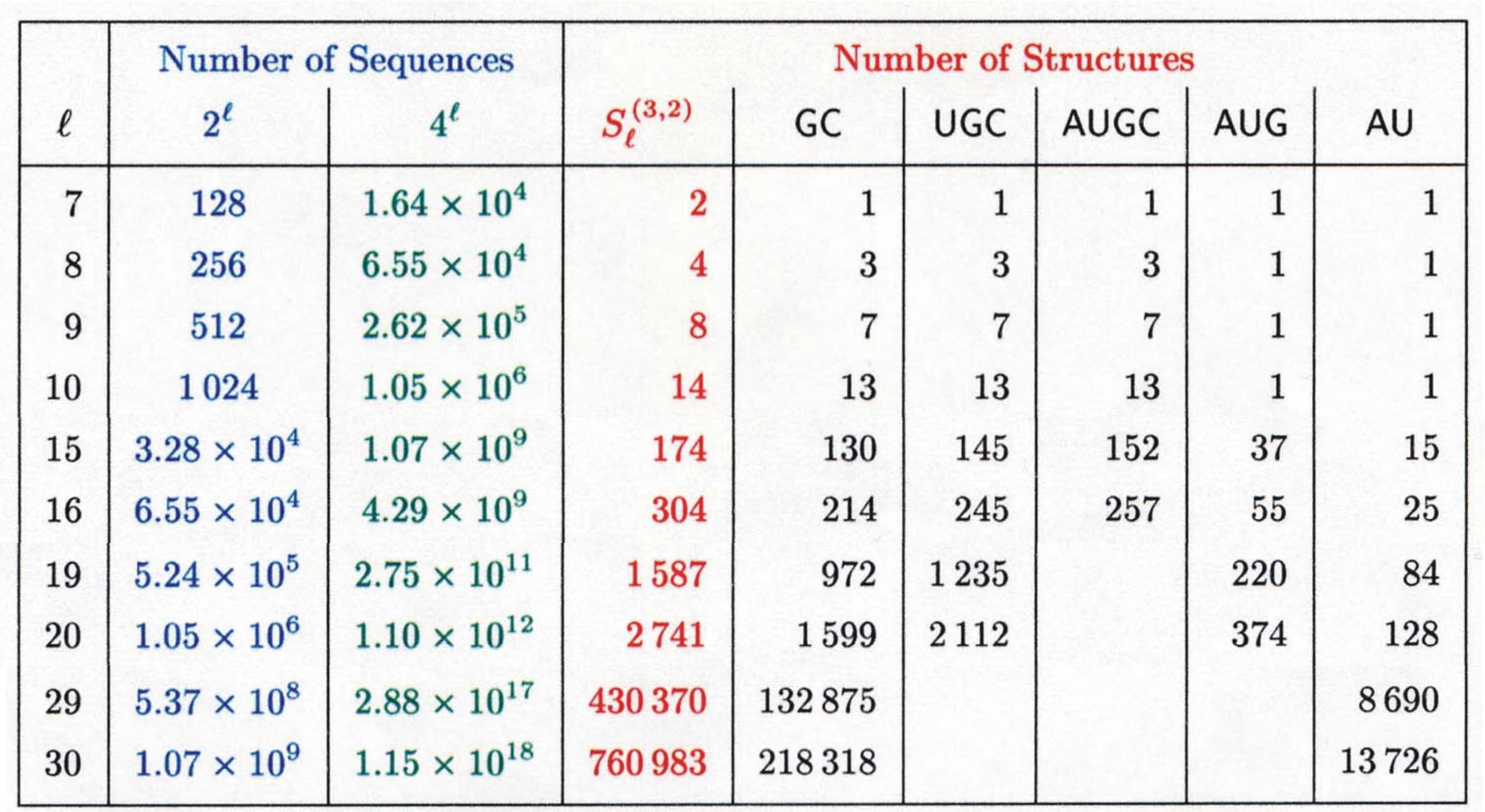
$$\Psi_0 = \Psi_1 = \Psi_2 = \Psi_3 = \Psi_4 = \Psi_5 = \Psi_6 = 1$$

$$\Phi_0 = \Phi_1 = \Phi_2 = \Phi_3 = \Phi_4 = 0$$

$$\Xi_0 = \Xi_1 = \Xi_2 = \Xi_3 = \Xi_4 = \Xi_5 = \Xi_6 = \Xi_7 = 1$$

Solution: $S_\ell^{(3,2)} = \Psi_{m=\ell}$

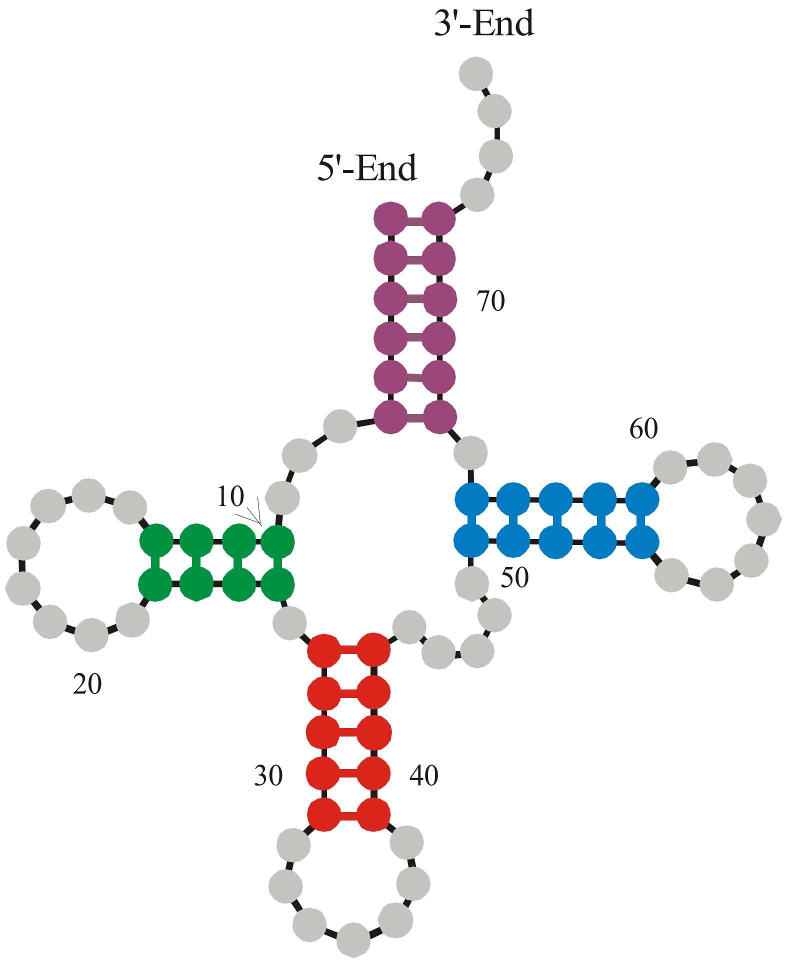
Recursion formula for the number of acceptable RNA secondary structures



	Number of Sequences		Number of Structures					
ℓ	2^ℓ	4^ℓ	$S_\ell^{(3,2)}$	GC	UGC	AUGC	AUG	AU
7	128	1.64×10^4	2	1	1	1	1	1
8	256	6.55×10^4	4	3	3	3	1	1
9	512	2.62×10^5	8	7	7	7	1	1
10	1 024	1.05×10^6	14	13	13	13	1	1
15	3.28×10^4	1.07×10^9	174	130	145	152	37	15
16	6.55×10^4	4.29×10^9	304	214	245	257	55	25
19	5.24×10^5	2.75×10^{11}	1 587	972	1 235		220	84
20	1.05×10^6	1.10×10^{12}	2 741	1 599	2 112		374	128
29	5.37×10^8	2.88×10^{17}	430 370	132 875				8 690
30	1.07×10^9	1.15×10^{18}	760 983	218 318				13 726

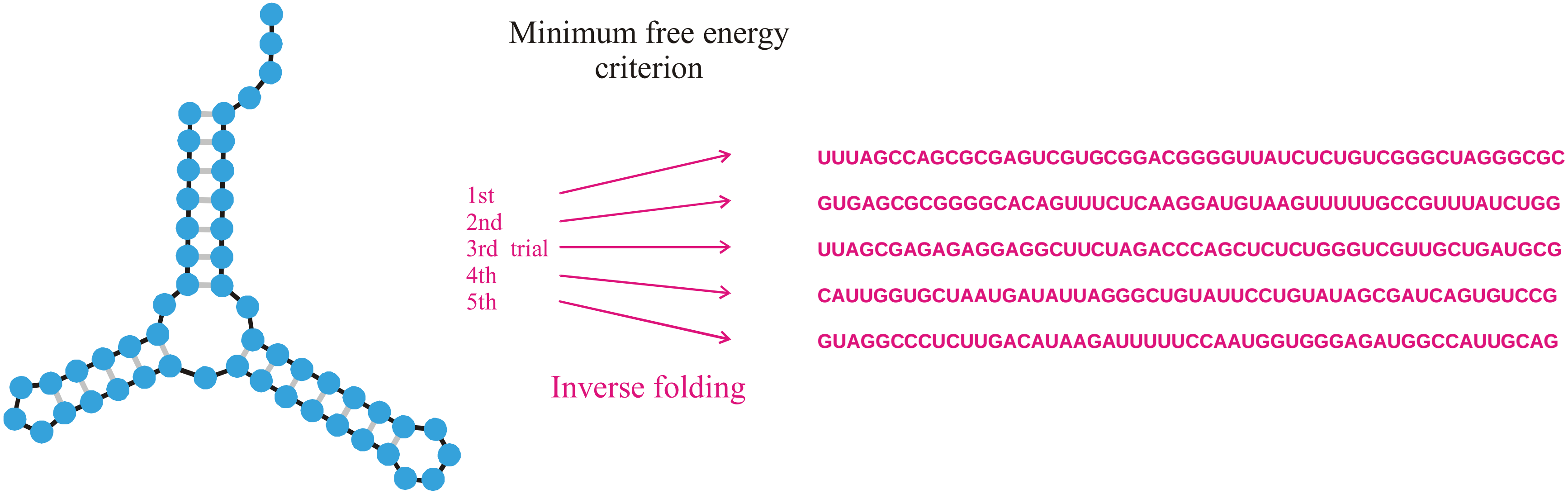
Computed numbers of minimum free energy structures over different alphabets

P. Schuster, *Molecular insights into evolution of phenotypes*. In: J. Crutchfield & P.Schuster, Evolutionary Dynamics. Oxford University Press, New York 2003, pp.163-215.



tRNA^{phe}

RNA clover-leaf secondary structures of sequences with chain length n=76



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

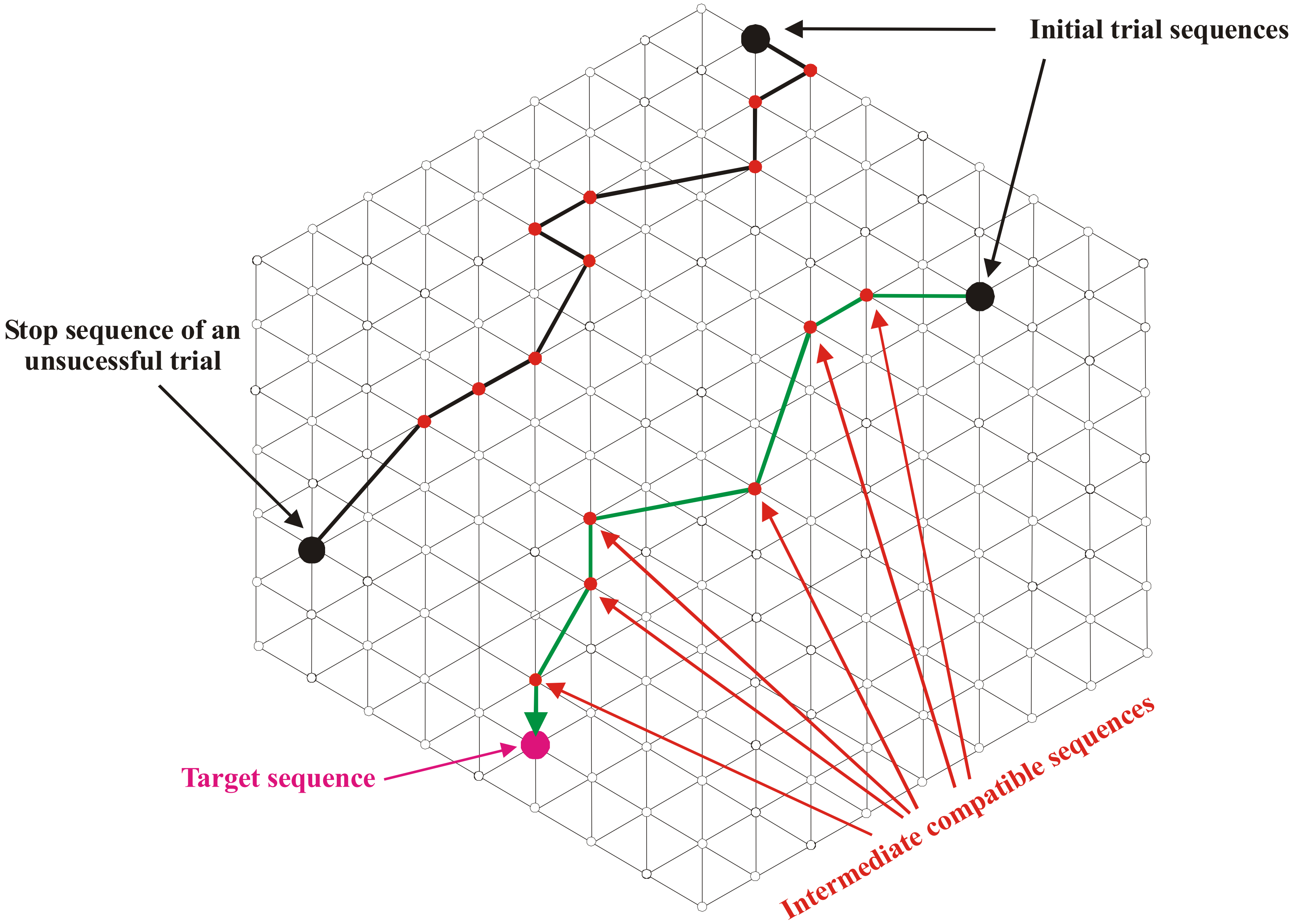
UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

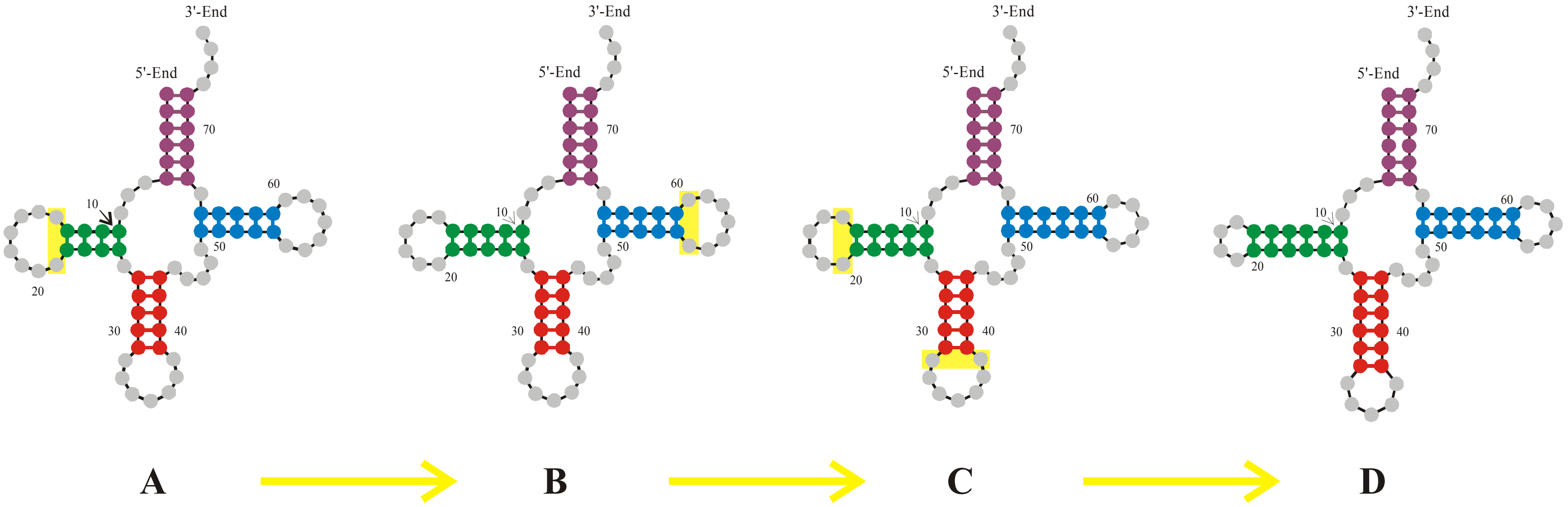
GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG

Inverse folding

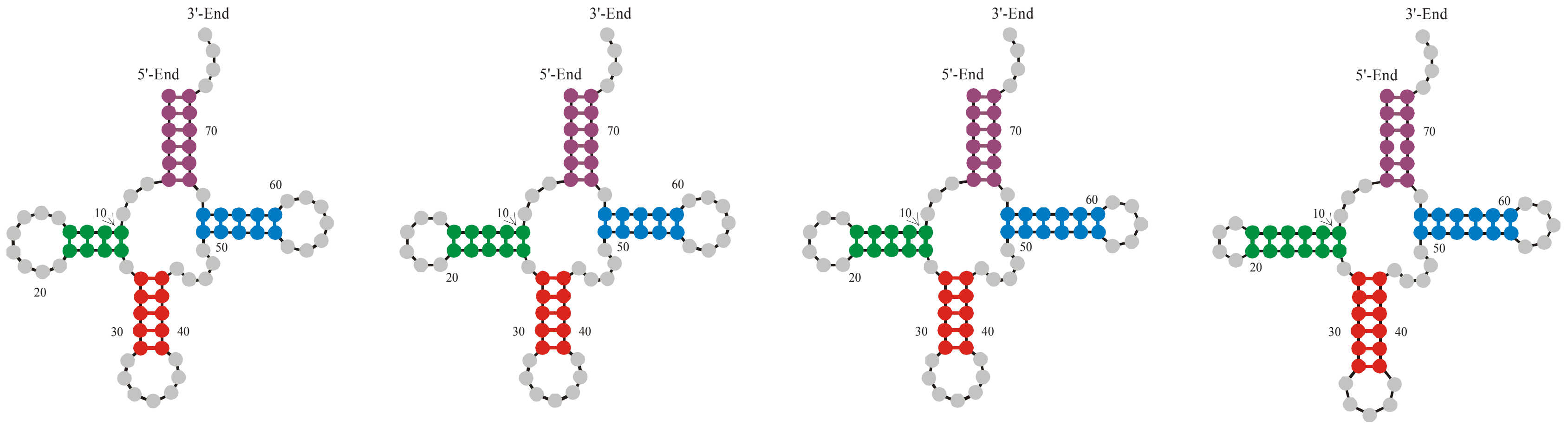
The **inverse folding algorithm** searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.


Initial trial sequences
Stop sequence of an unsucessful trial
Target sequence
Intermediate compatible sequences

Approach to the target structure in the inverse folding algorithm



RNA clover-leaf secondary structures of sequences with chain length n=76



Alphabet	Number of successful inverse foldings out of 1000 trials			
AU	- - -	- - -	- - -	- - -
AUG	- - -	4 Ä 2	24 Ä 8	30 Ä 6
AUGC	790	900	940	960
UGC	570	630	710	740
GC	64 Ä 6	89 Ä 15	84 Ä 10	77 Ä 5

Search for clover-leef structures by means of the inverse folding algorithm

Theory of sequence – structure mappings

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

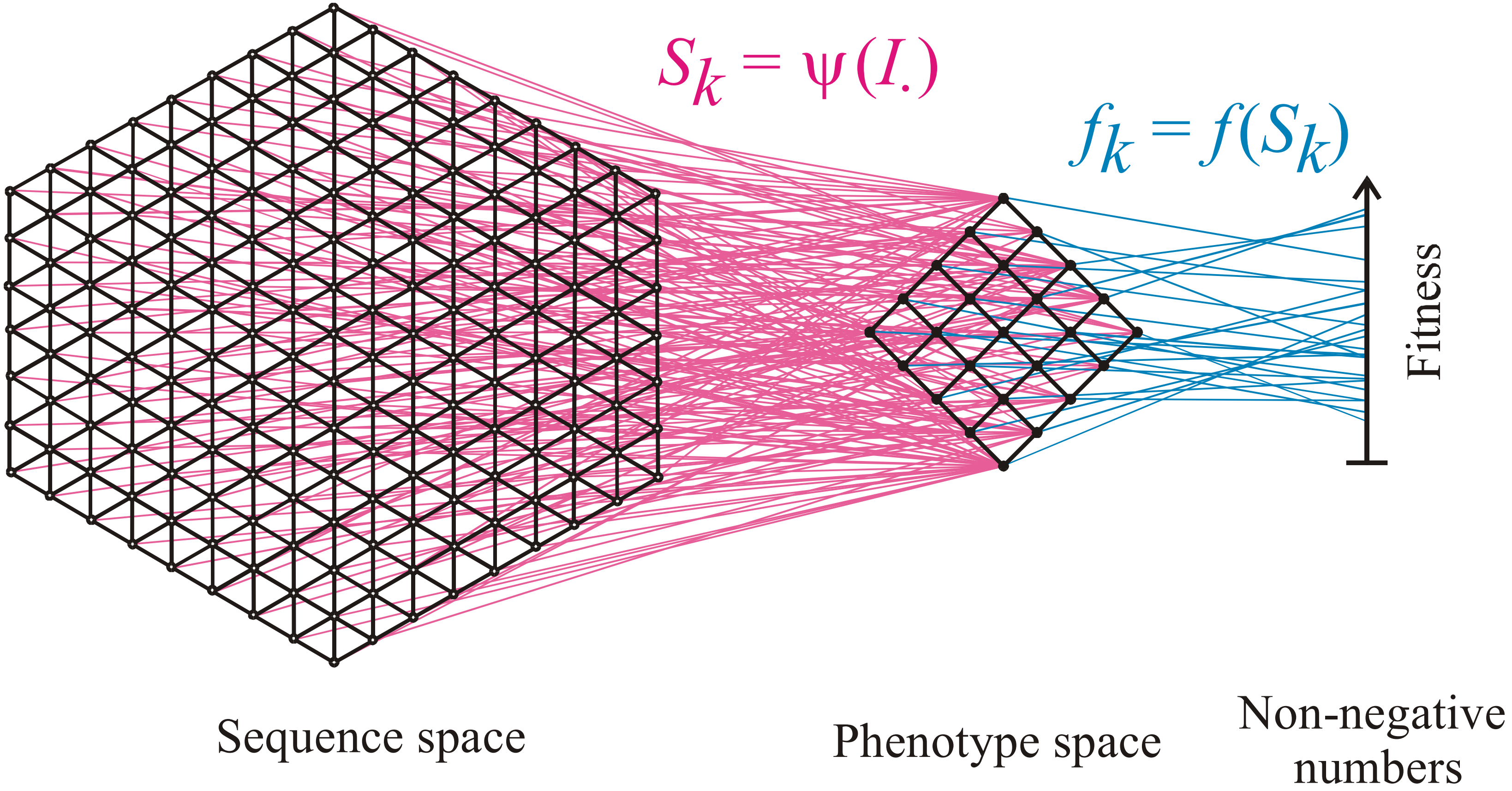
C.M.Reidys, P.F.Stadler, P.Schuster, ***Generic properties of combinatory maps***. Bull.Math.Biol. **59** (1997), 339-397

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

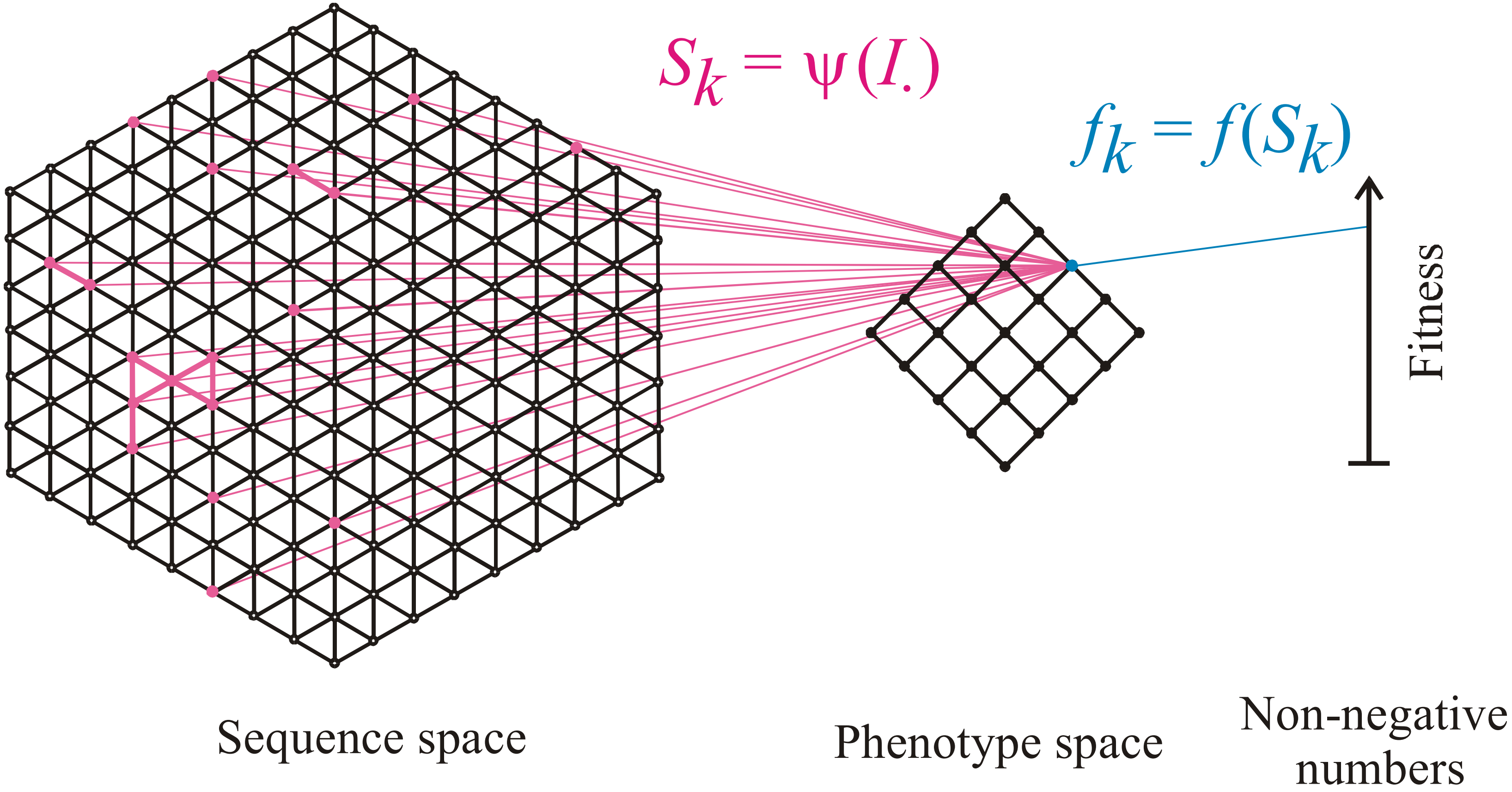
C.M.Reidys, P.F.Stadler, ***Combinatory landscapes***. SIAM Review **44** (2002), 3-54

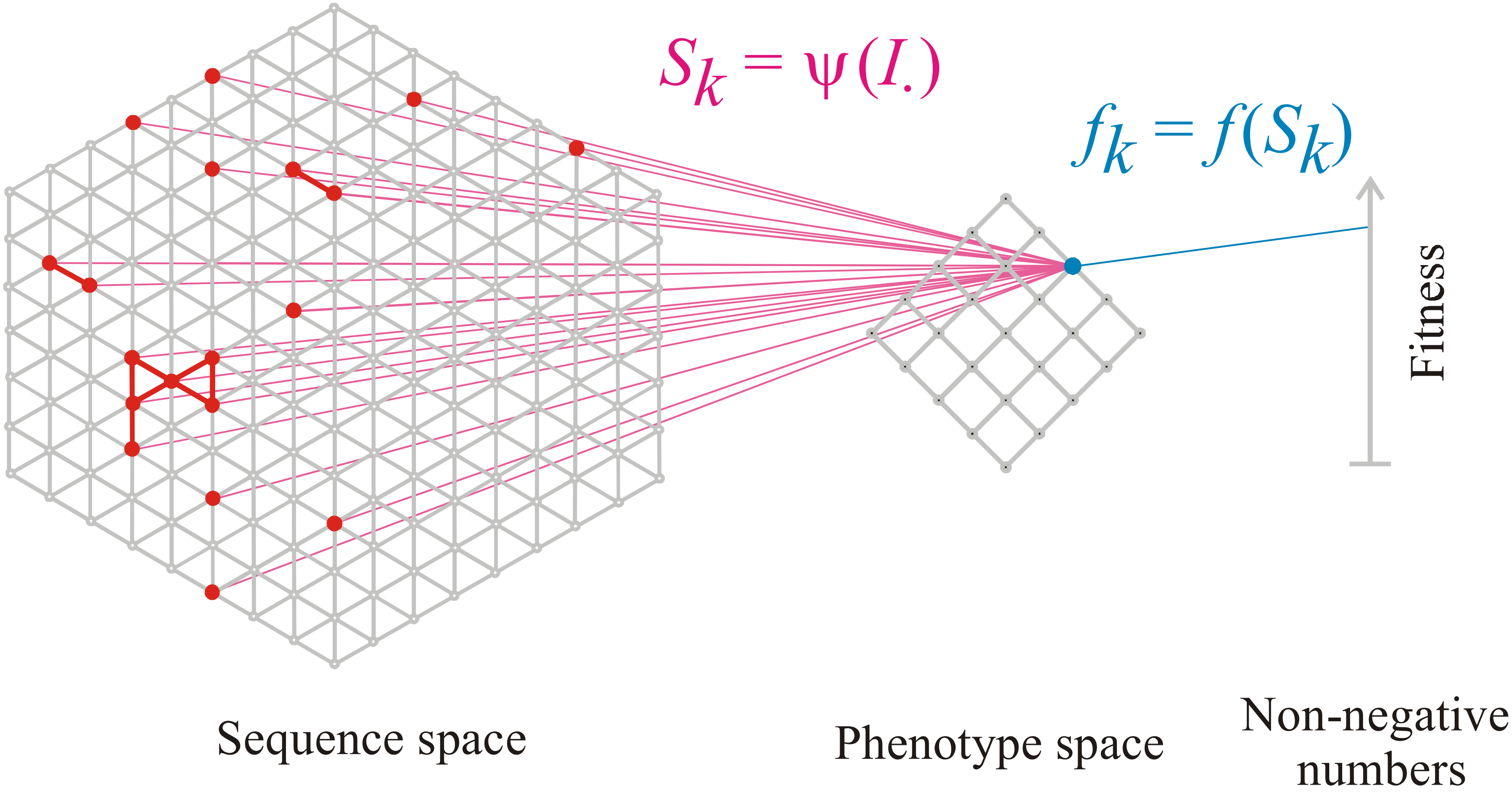
Sequence-structure relations are highly complex and only the simplest case can be studied. An example is the folding of RNA sequences into RNA structures represented in course-grained form as secondary structures.

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



Mapping from sequence space into phenotype space and into function


$S_k = \psi(I.)$
$f_k = f(S_k)$
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:

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

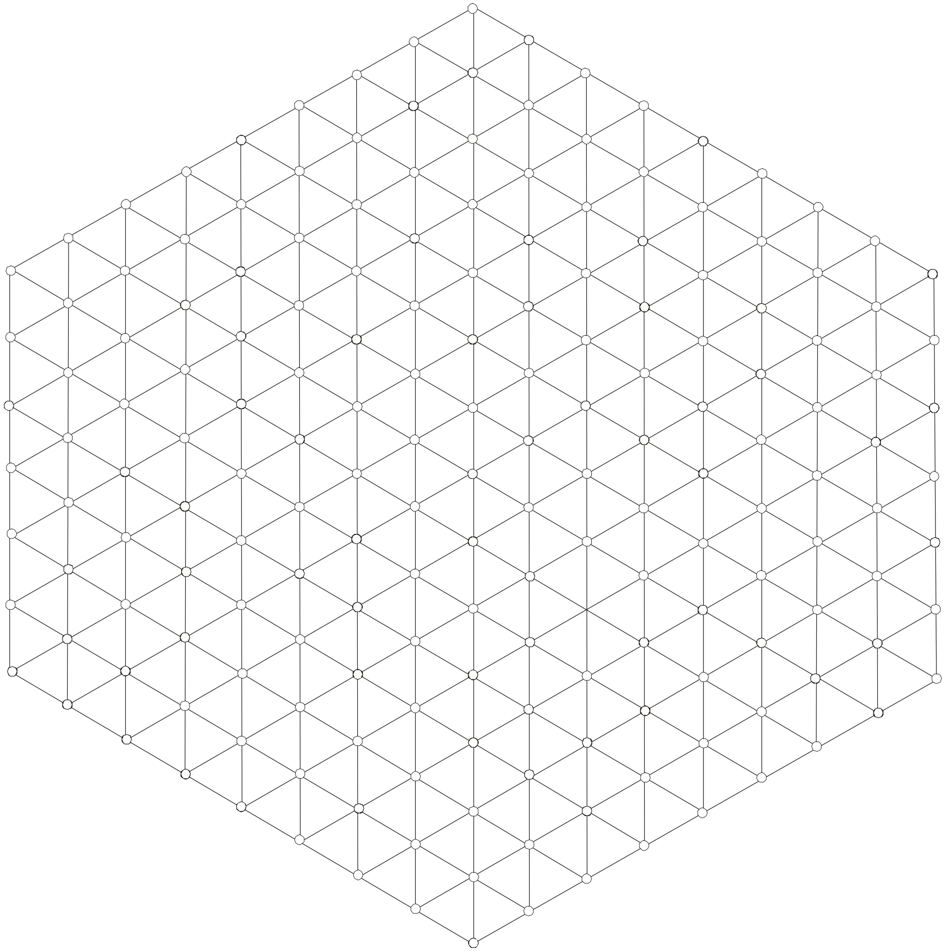
The set is converted into a graph by connecting all sequences of Hamming distance one.

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.

Step 00

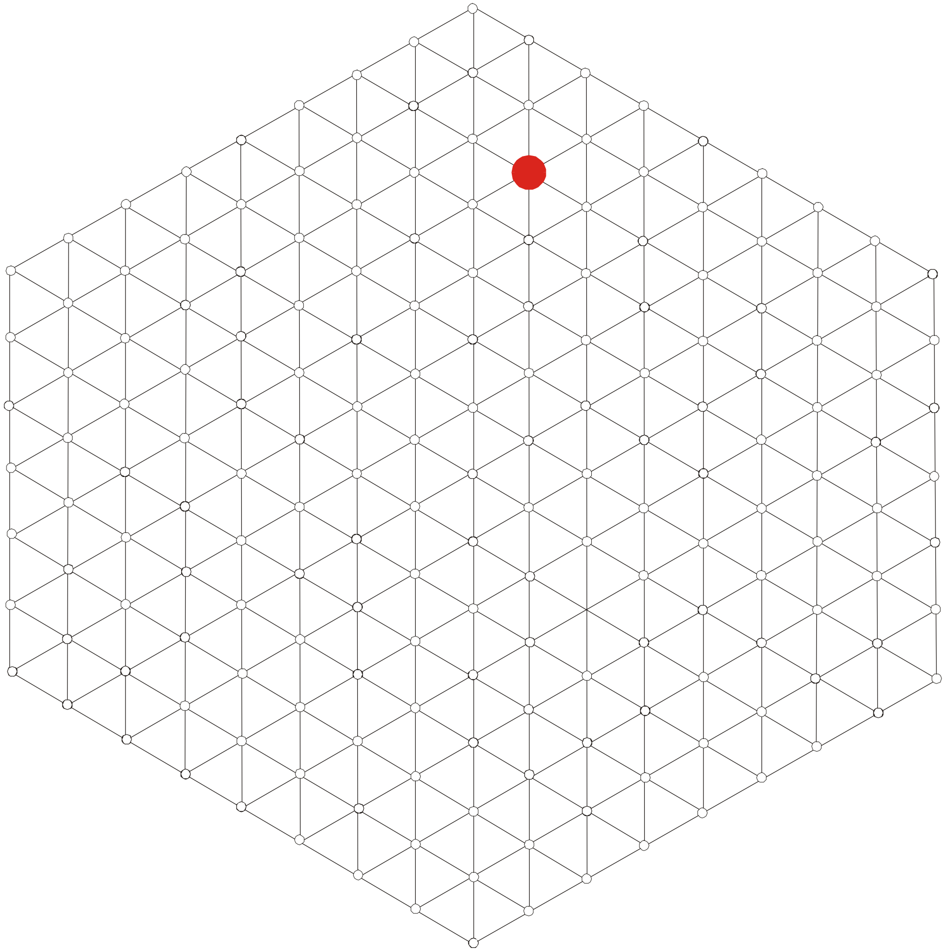
Sketch of sequence space



Random graph approach to neutral networks

Step 01

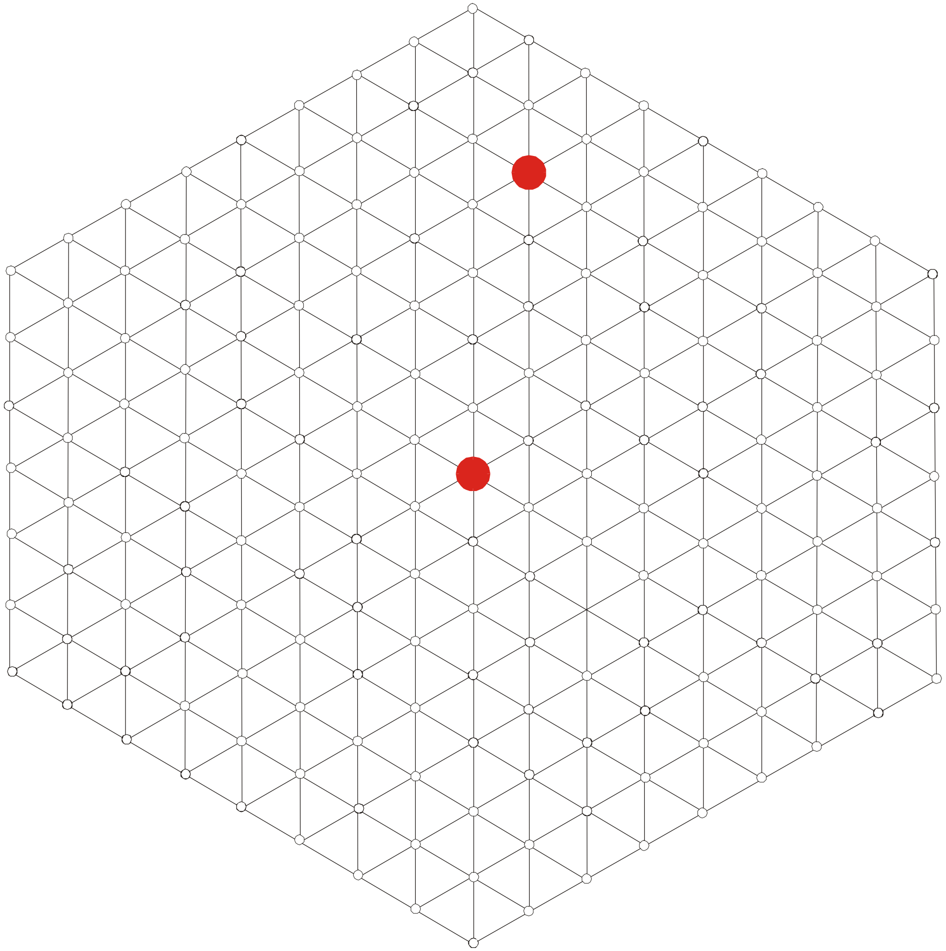
Sketch of sequence space



Random graph approach to neutral networks

Step 02

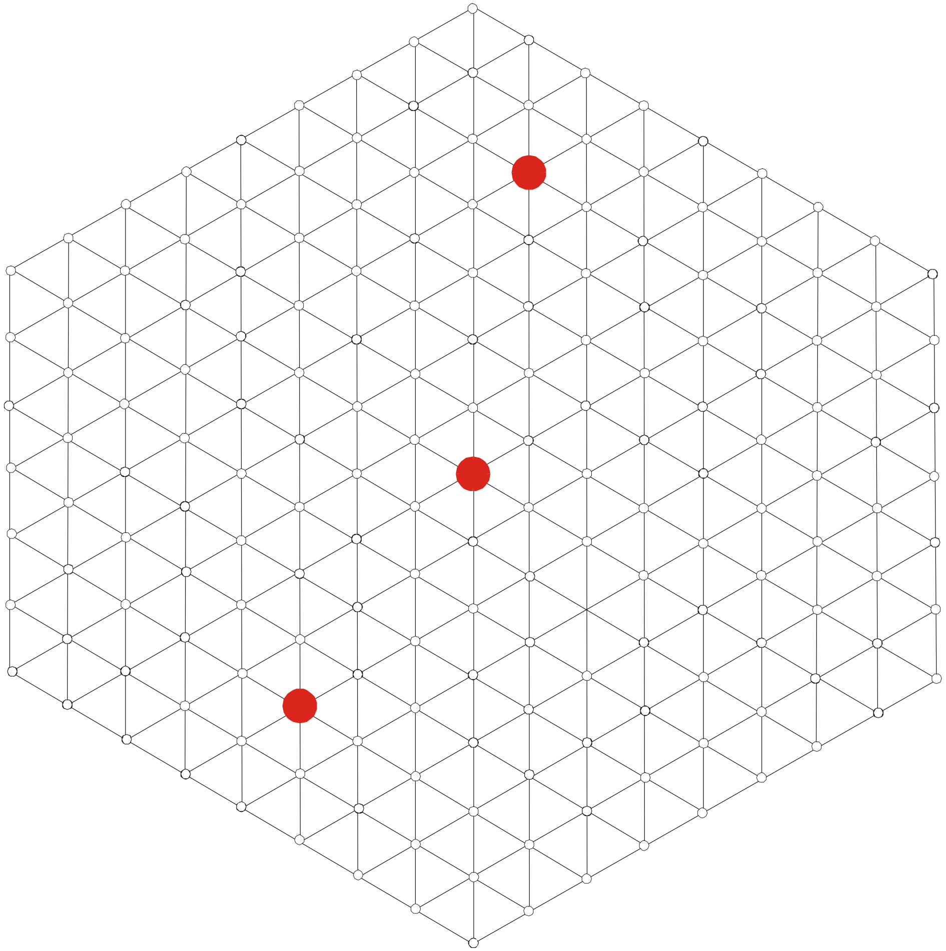
Sketch of sequence space



Random graph approach to neutral networks

Step 03

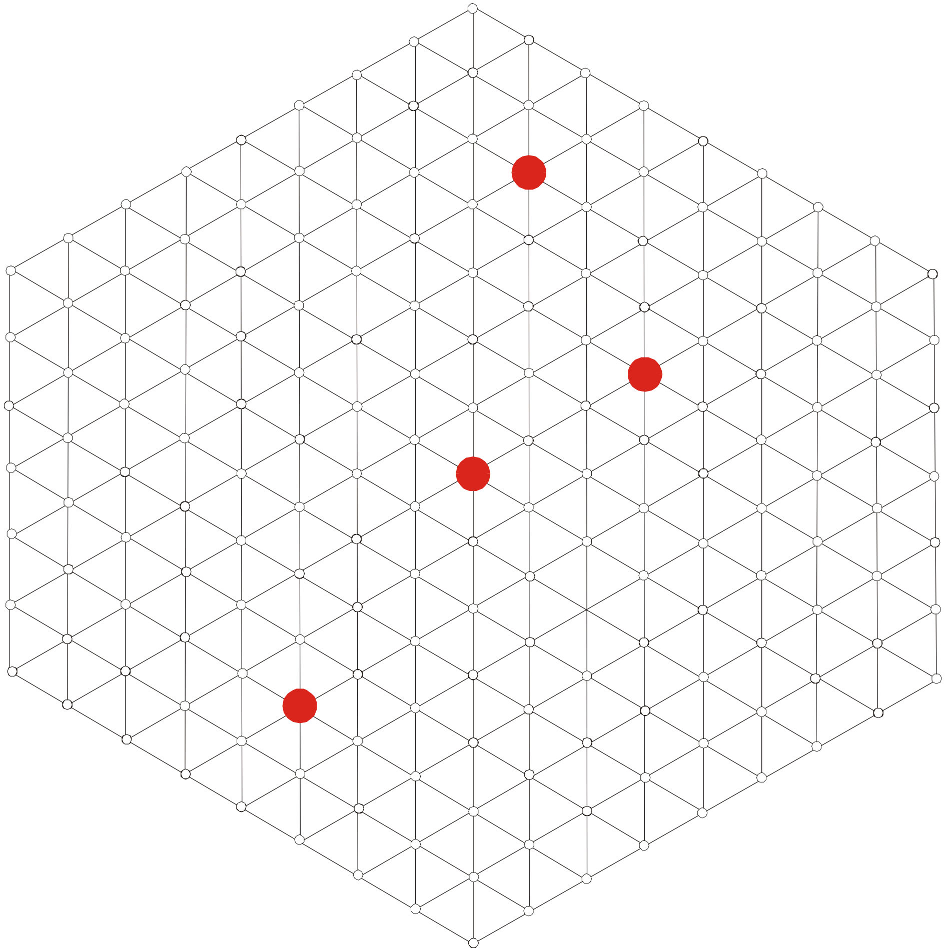
Sketch of sequence space



Random graph approach to neutral networks

Step 04

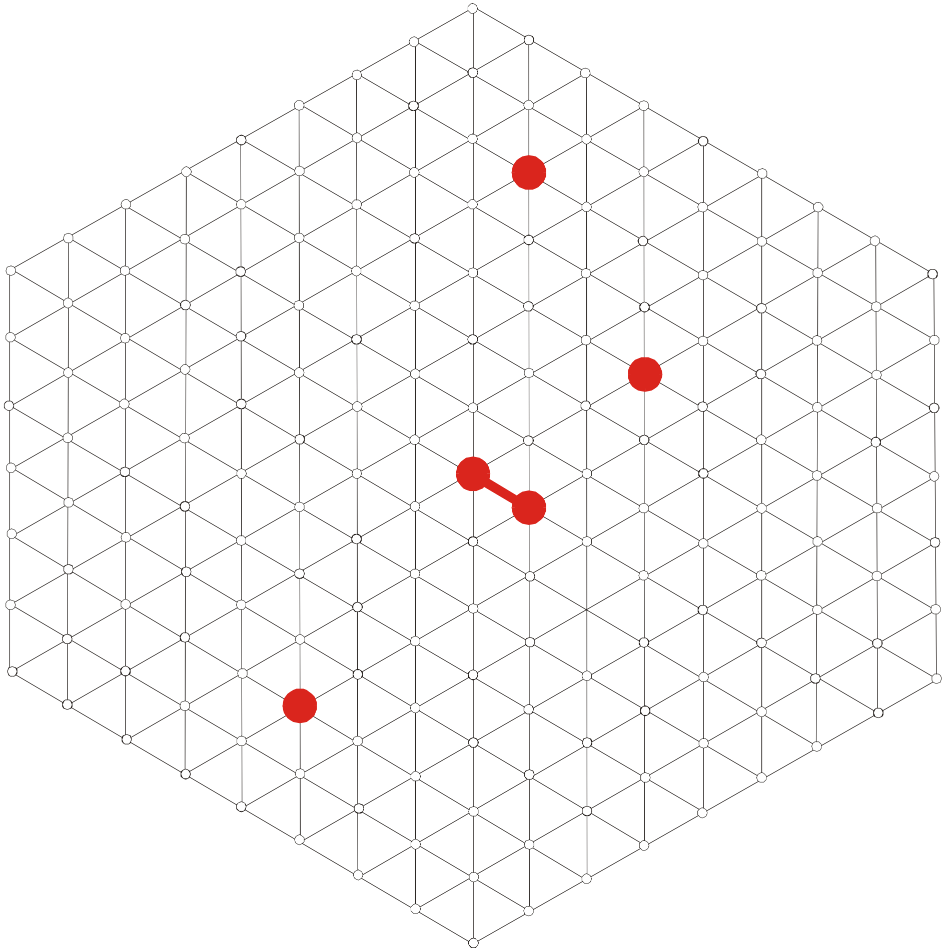
Sketch of sequence space



Random graph approach to neutral networks

Step 05

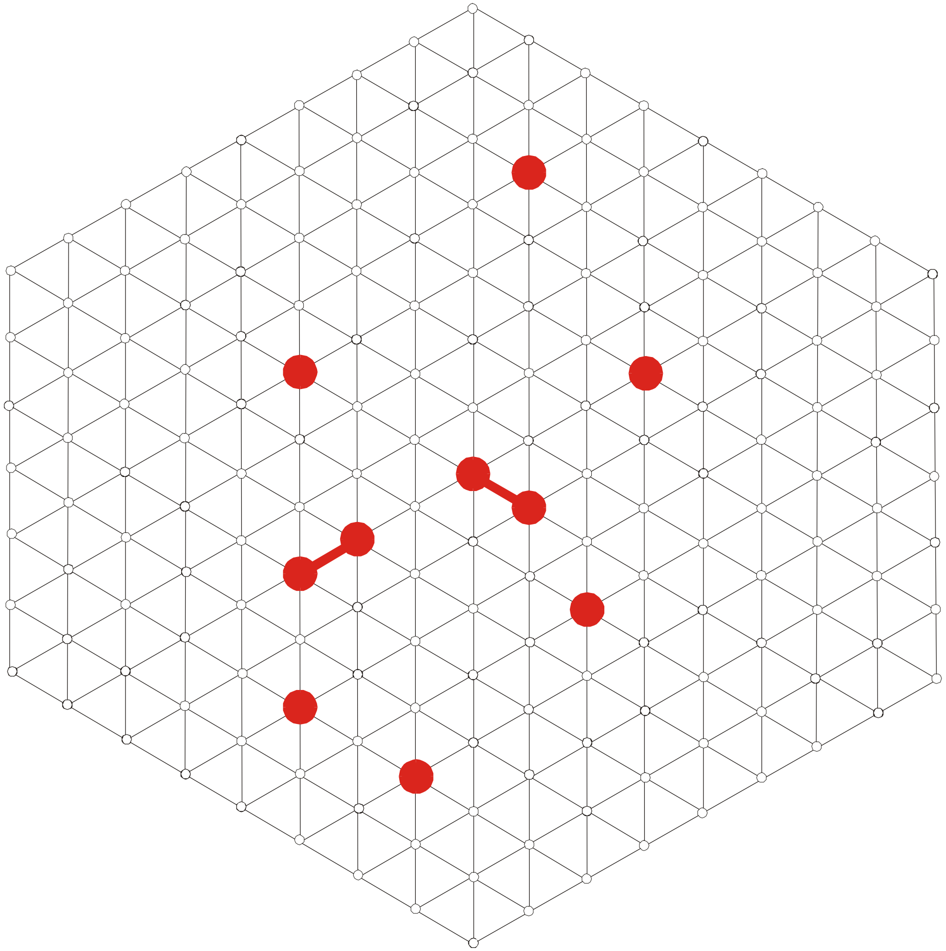
Sketch of sequence space



Random graph approach to neutral networks

Step 10

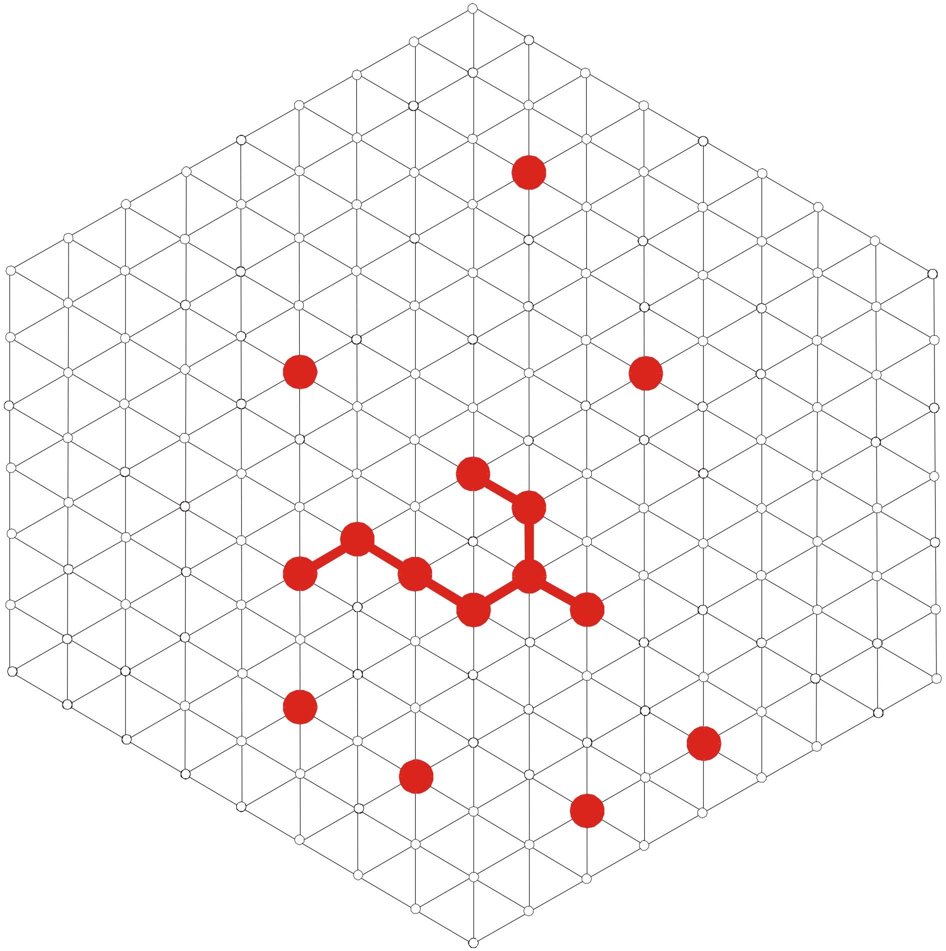
Sketch of sequence space



Random graph approach to neutral networks

Step 15

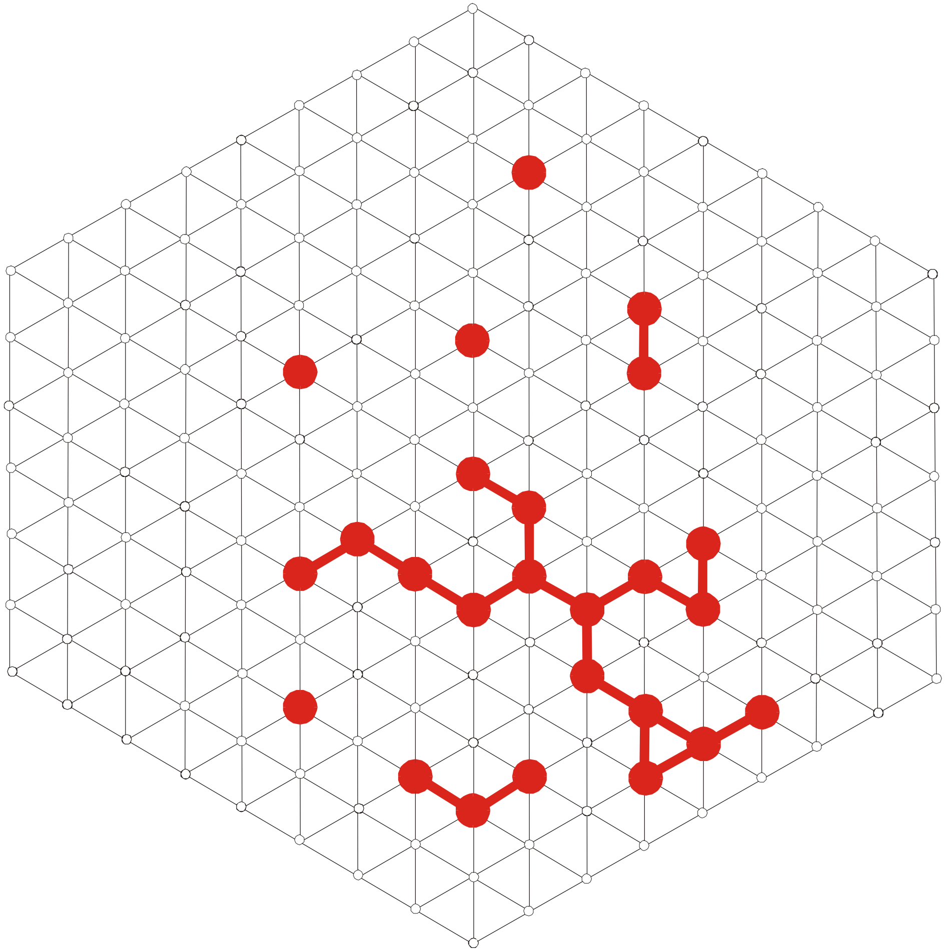
Sketch of sequence space



Random graph approach to neutral networks

Step 25

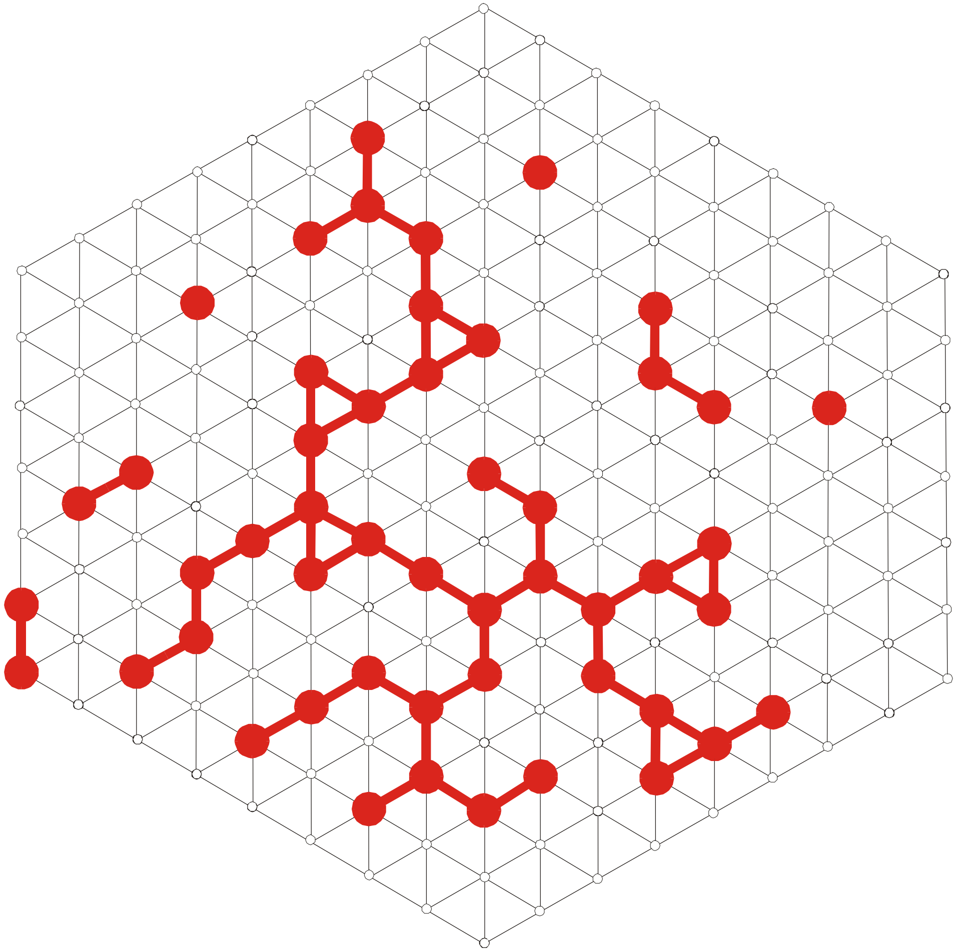
Sketch of sequence space



Random graph approach to neutral networks

Step 50

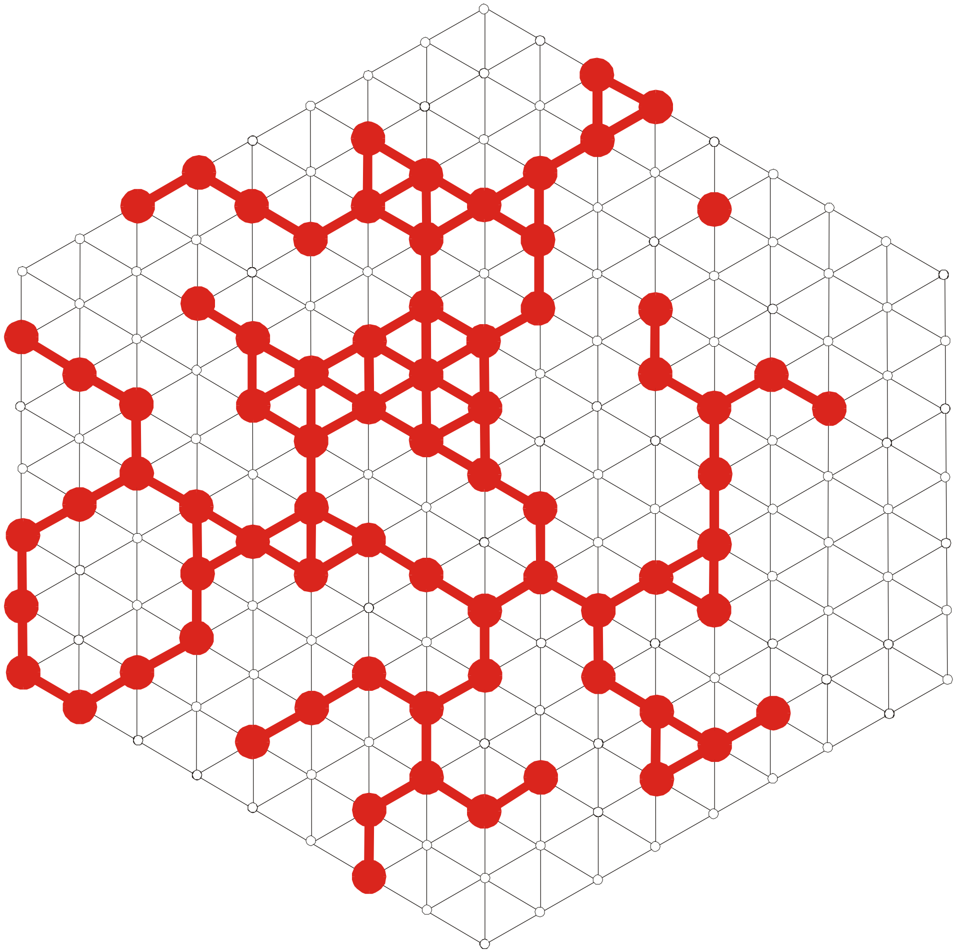
Sketch of sequence space



Random graph approach to neutral networks

Step 75

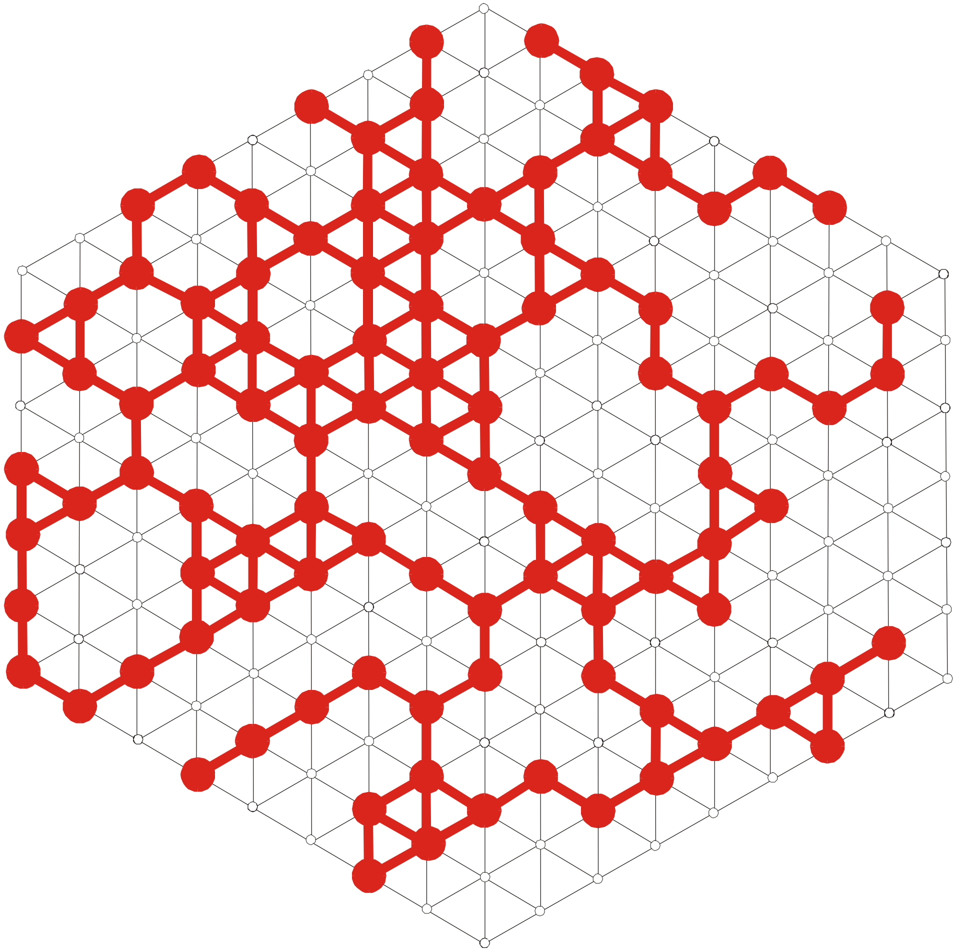
Sketch of sequence space



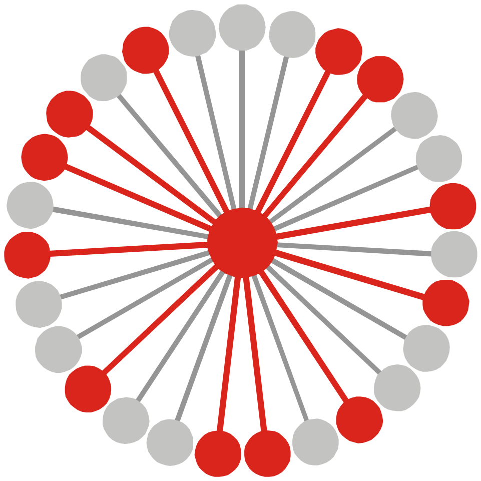
Random graph approach to neutral networks

Step 100

Sketch of sequence space



Random graph approach to neutral networks



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

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

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

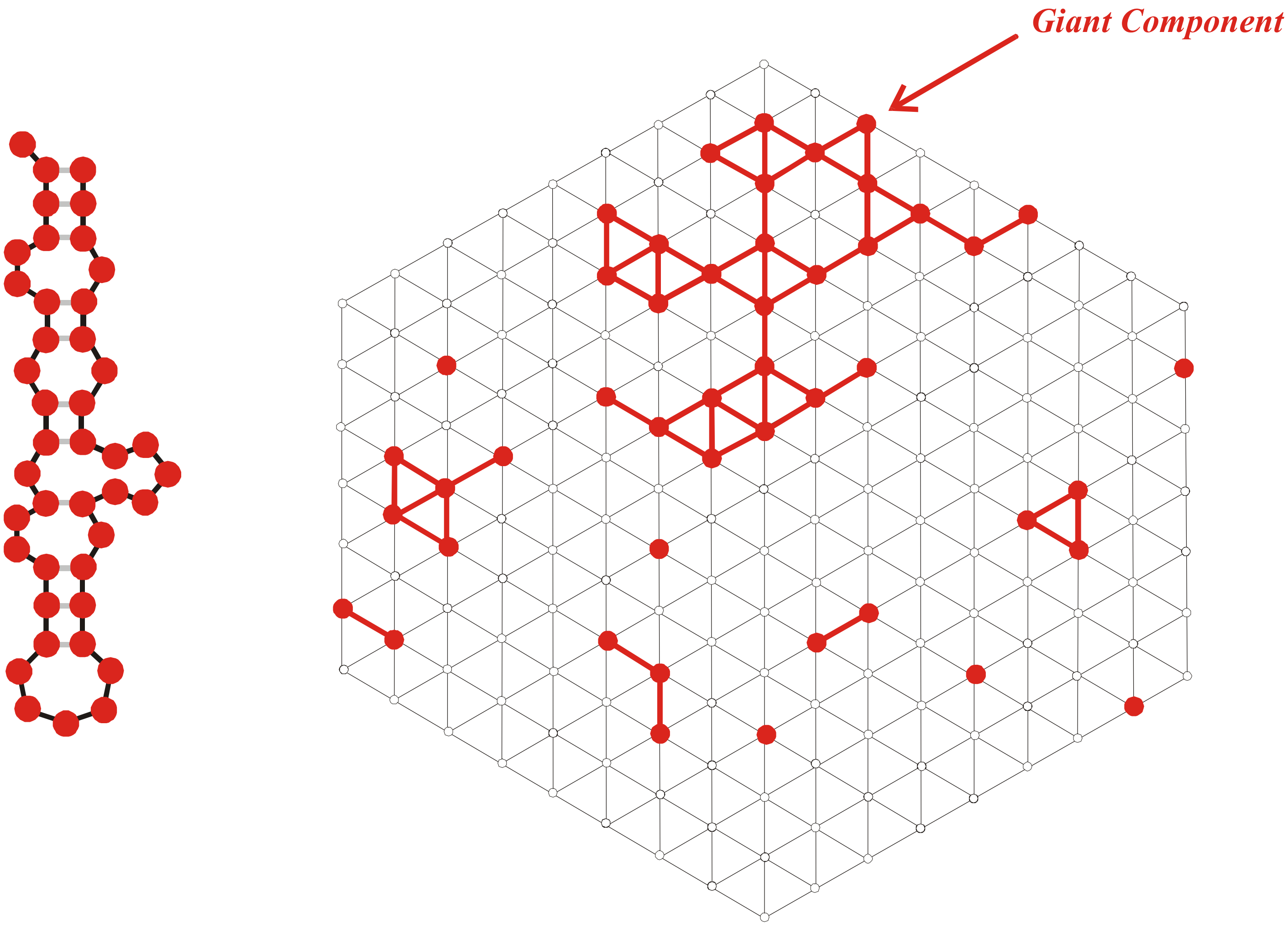
Alphabet size k : **AUGC** ñ $k = 4$

k	l_{cr}
2	0.5
3	0.4226
4	0.3700

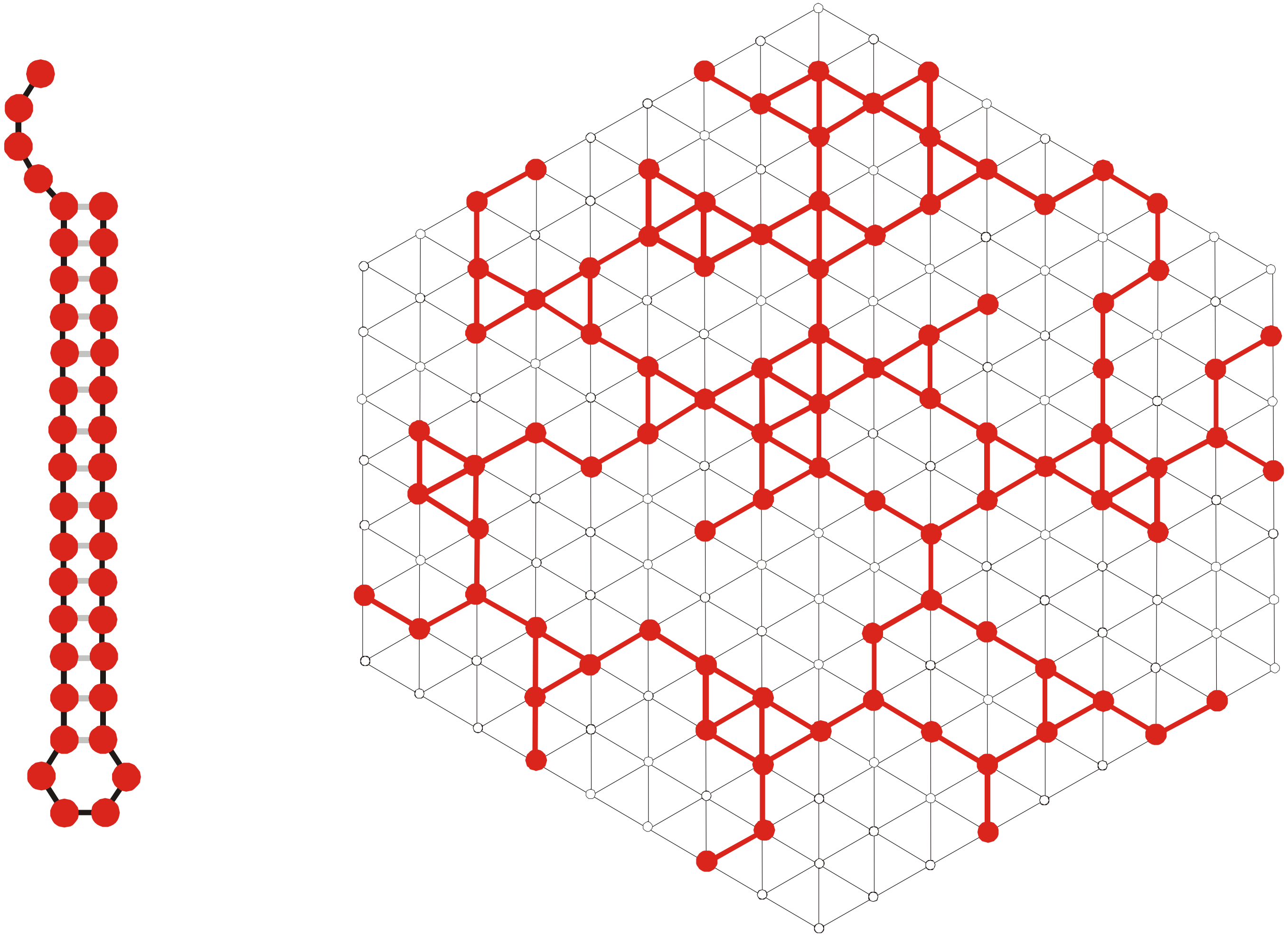
$\bar{\lambda}_k > \lambda_{cr}$ . . . . network **G_k** is connected

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

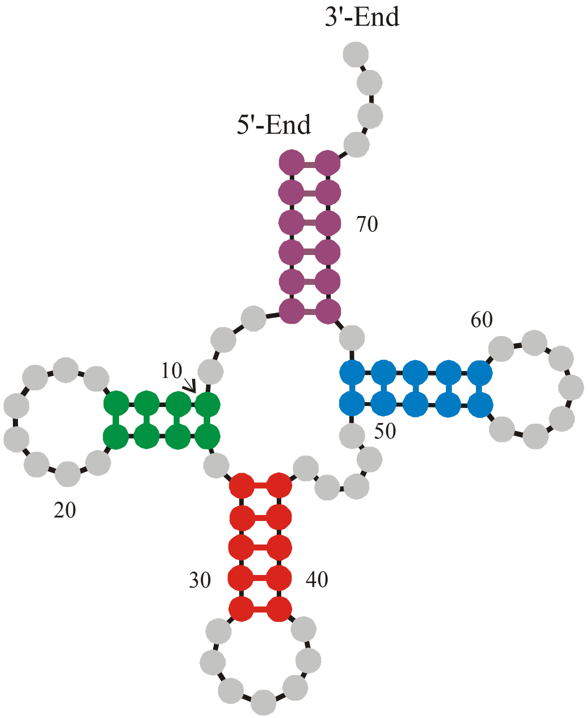
Mean degree of neutrality and connectivity of **neutral networks**



A multi-component neutral network

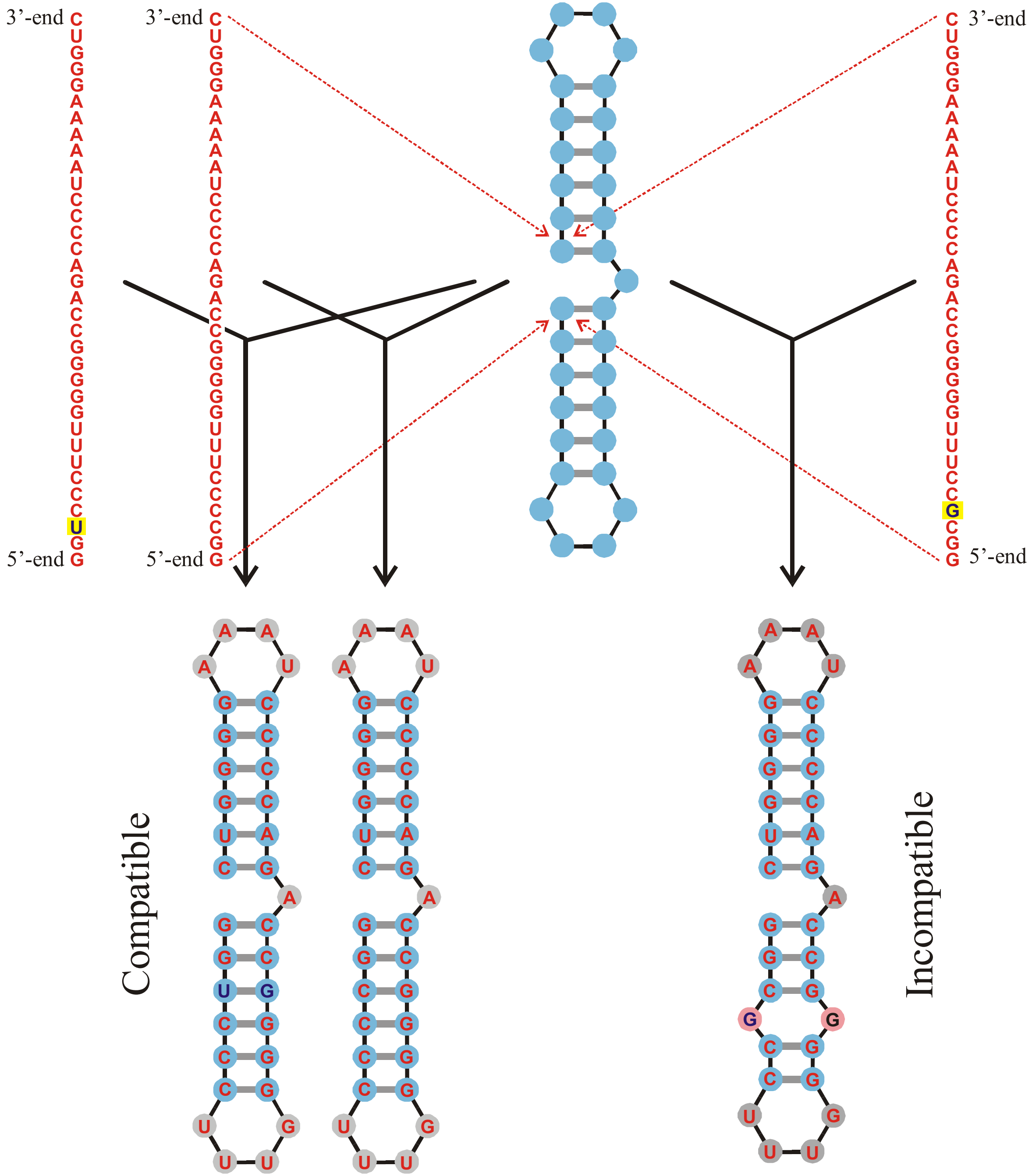


A connected neutral network

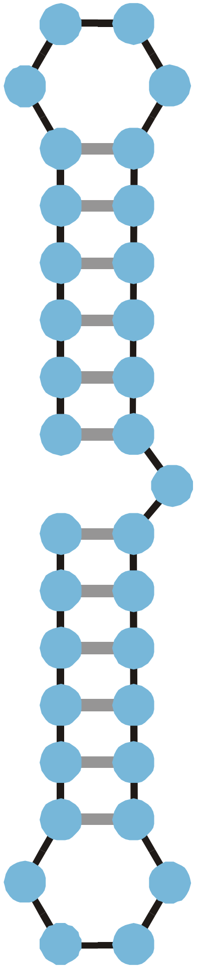


Alphabet	Degree of neutrality Ī
AUGC	0.27 Ä 0.07
UGC	0.26 Ä 0.07
GC	0.06 Ä 0.03

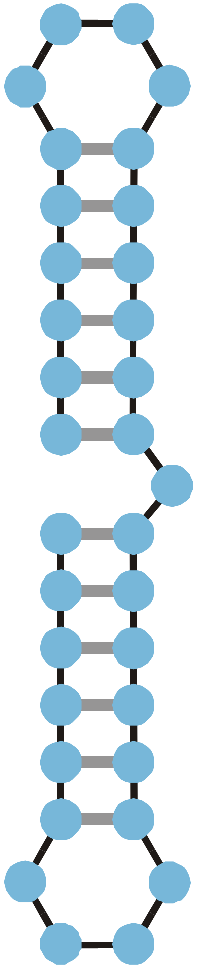
Computated degree of neutrality for the tRNA neutral network

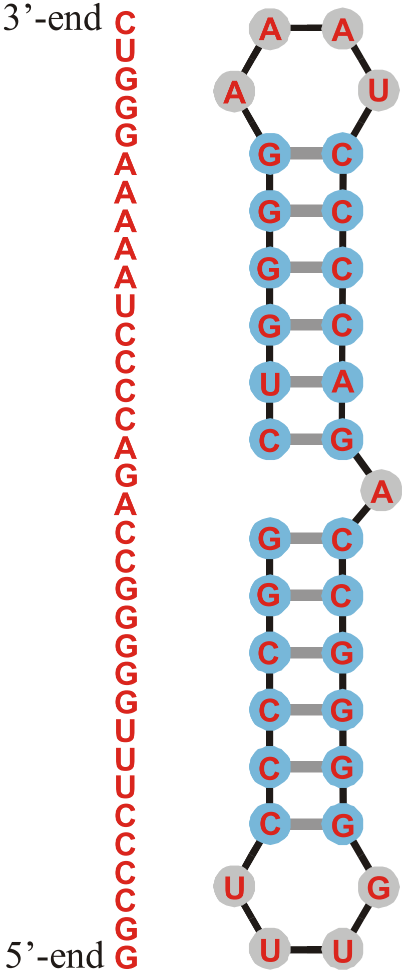

3’-end
CUGGGAAAAAUCCCCAGACCGGGGGUUUCCCUGG
5’-end
3’-end
CUGGGAAAAAUCCCCAGACCGGGGGUUUCCCCGG
5’-end
3’-end
CUGGGAAAAAUCCCCAGACCGGGGGUUUCCGCGG
5’-end
Compatible
Incompatible

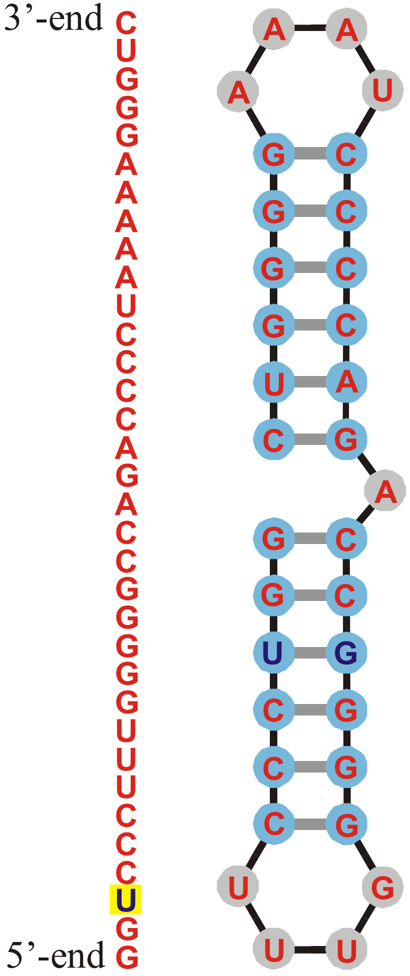
Definition of **compatibility** of sequences and structures



Structure

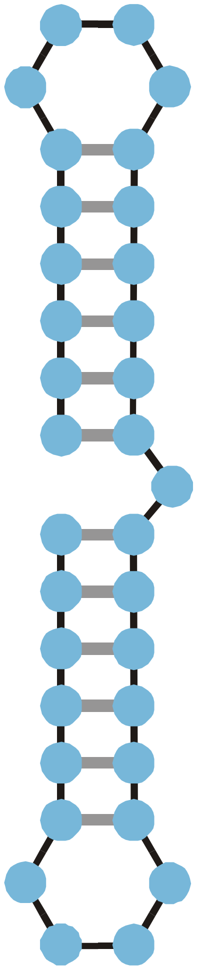




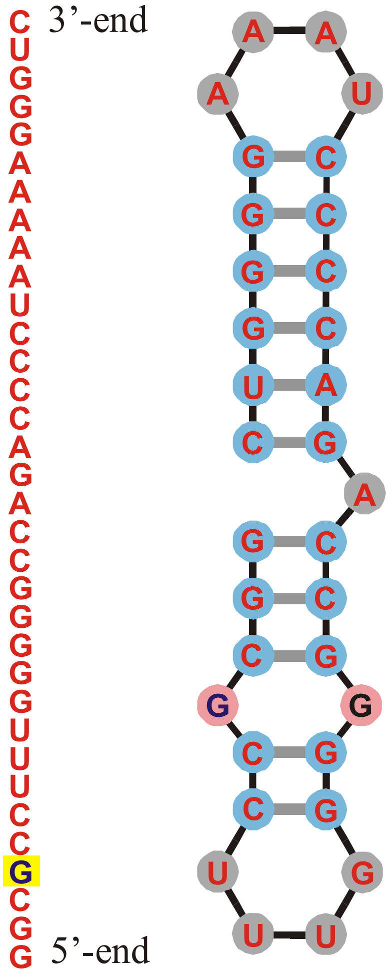


Structure

Compatible sequences

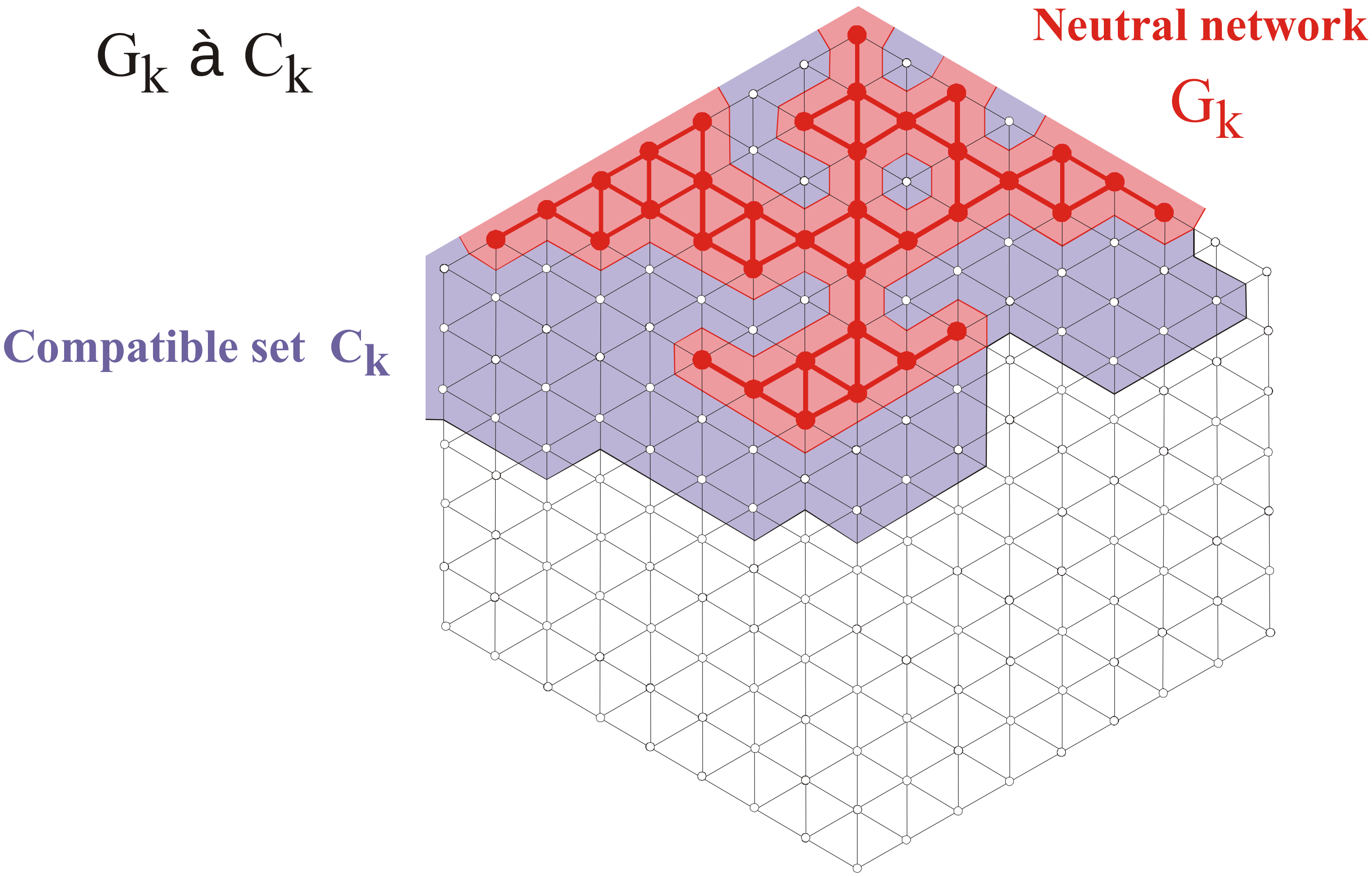


Structure

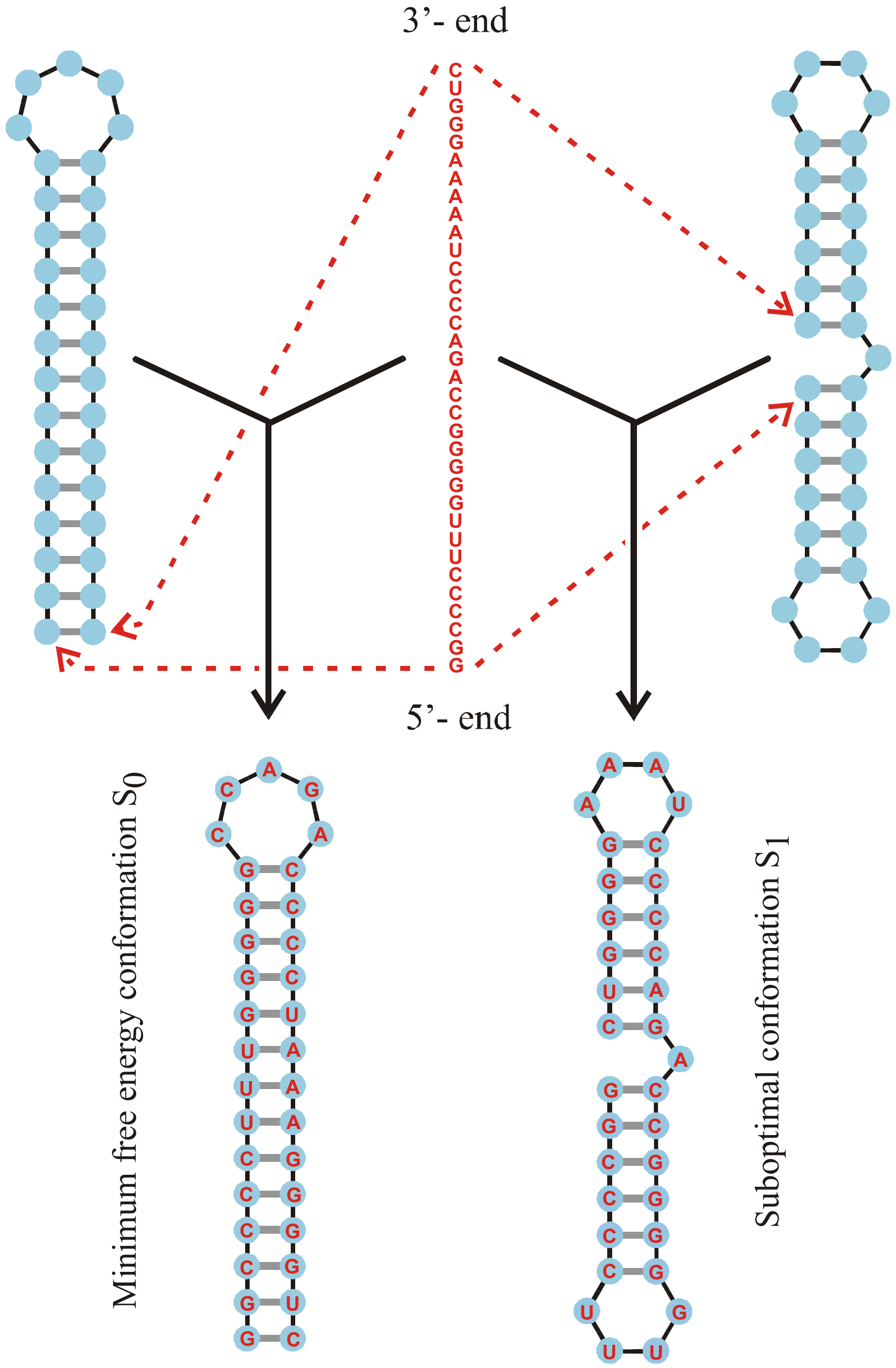


Incompatible sequence

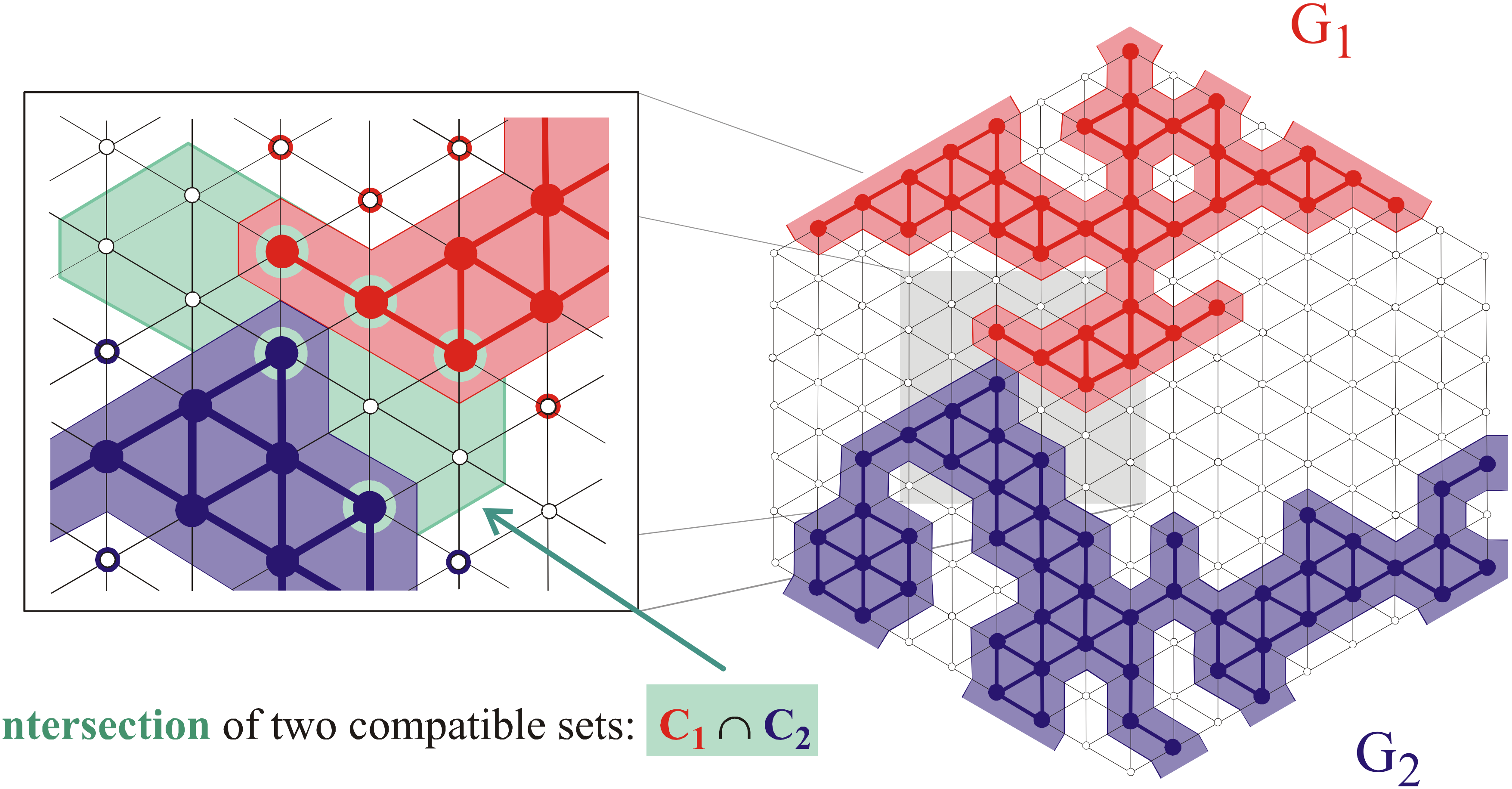
G_k à C_k



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
CUGGGAAAAUCCCCAGACCGGGGUUUCCCCGG
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: $C_1 \cap C_2$

○: á C_1 ì â C_2

○: â C_1 ì á C_2

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

1. Introduction

2. A few experiments

3. Analysing neutral networks

4. **Mechanisms of neutral evolution**

Optimization of RNA molecules *in silico*

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

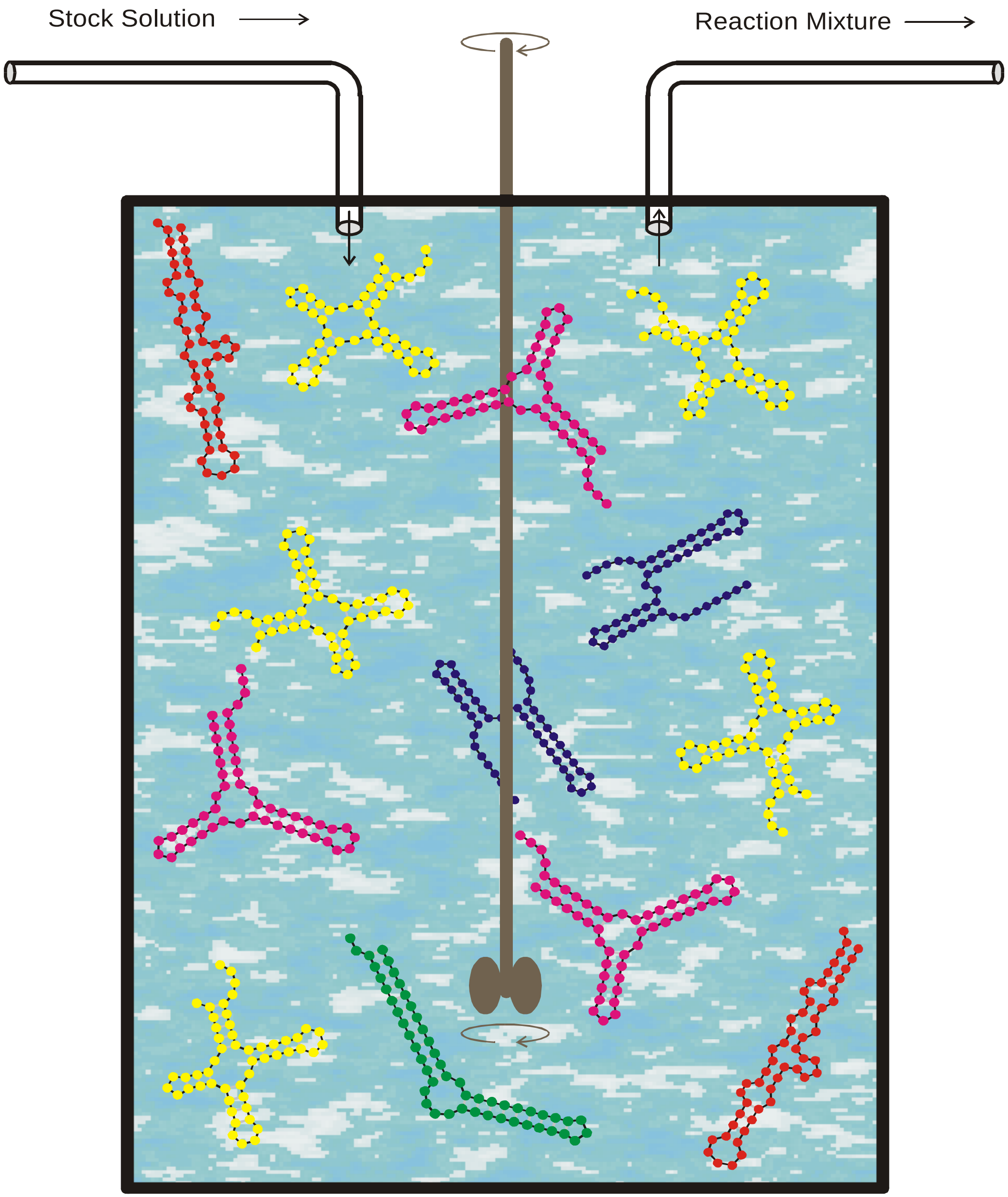
W.Fontana, W.Schnabl, P.Schuster, ***Physical aspects of evolutionary optimization and adaptation***. Phys.Rev.A **40** (1989), 3301-3321

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

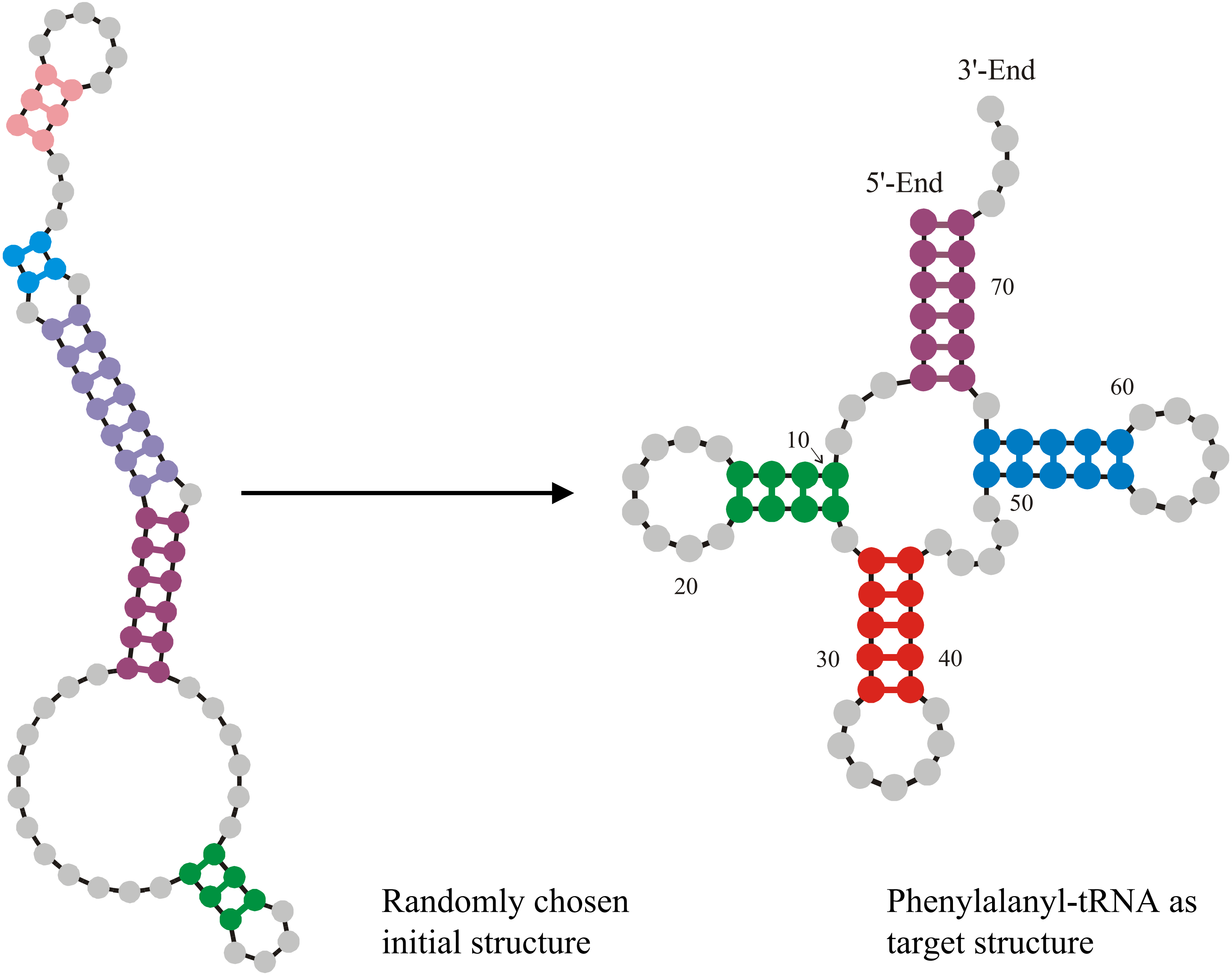


Fitness function:

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

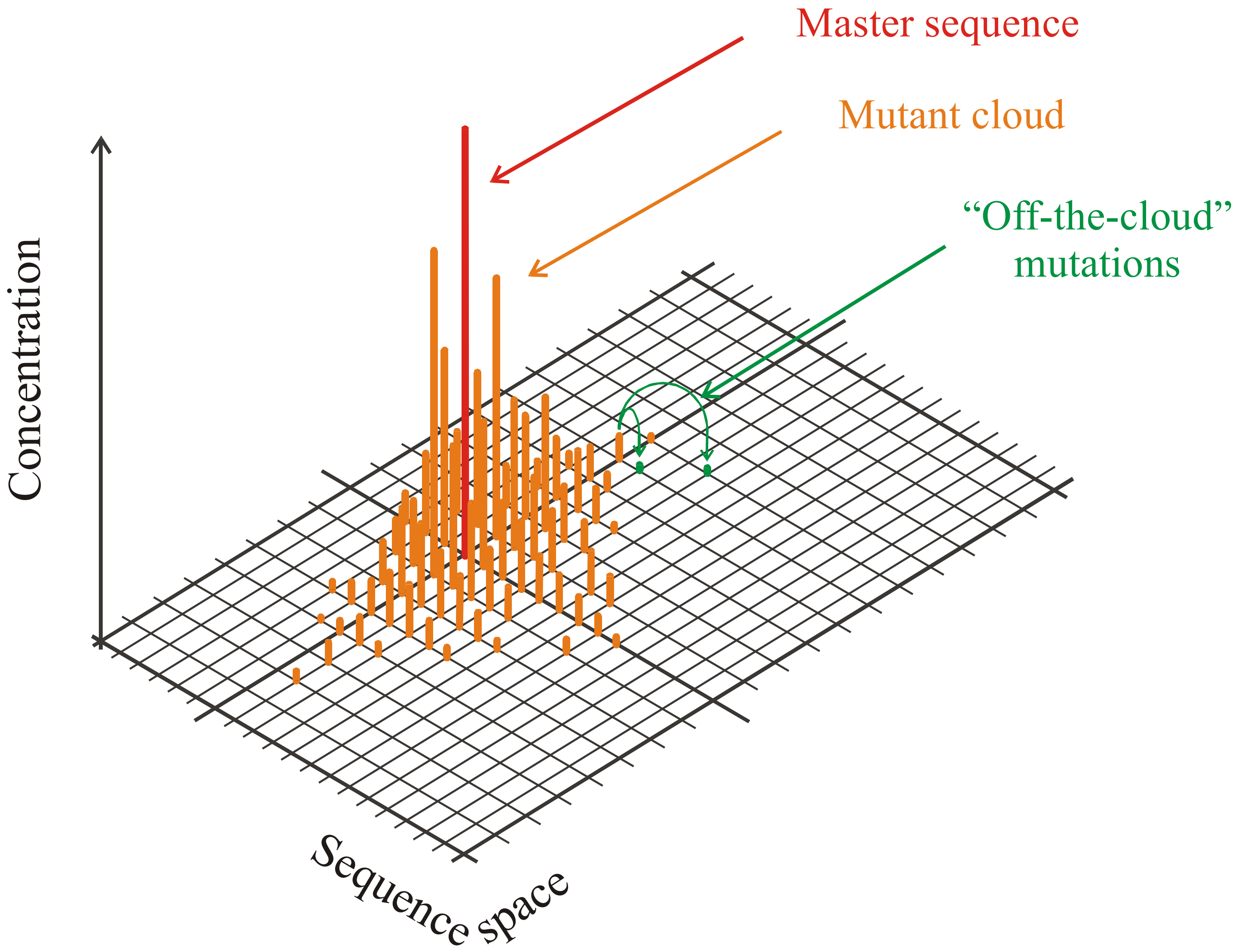
$$Dd_S^{\,(k)} = d^s(I_k, I_t)$$

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

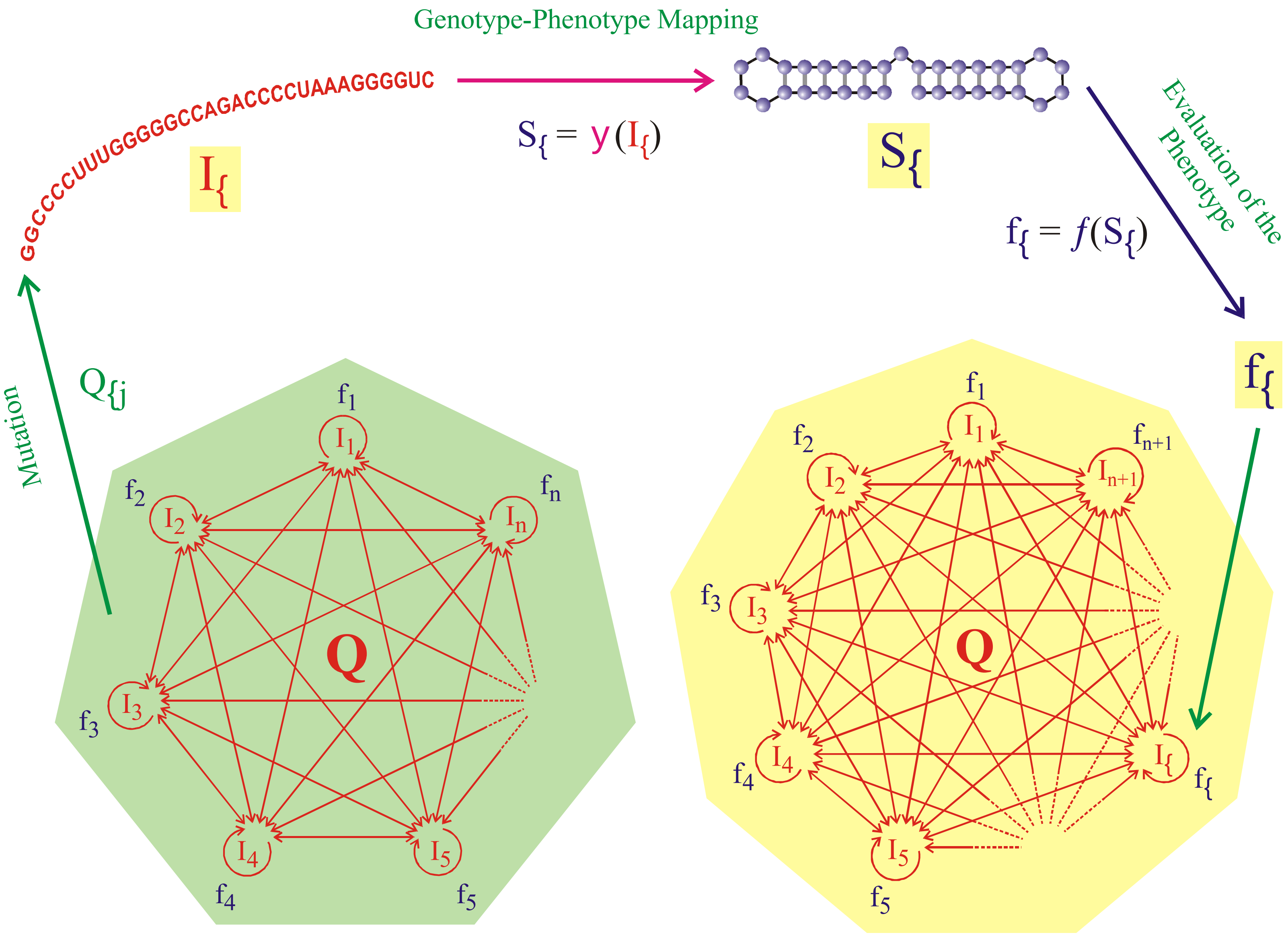


Randomly chosen initial structure

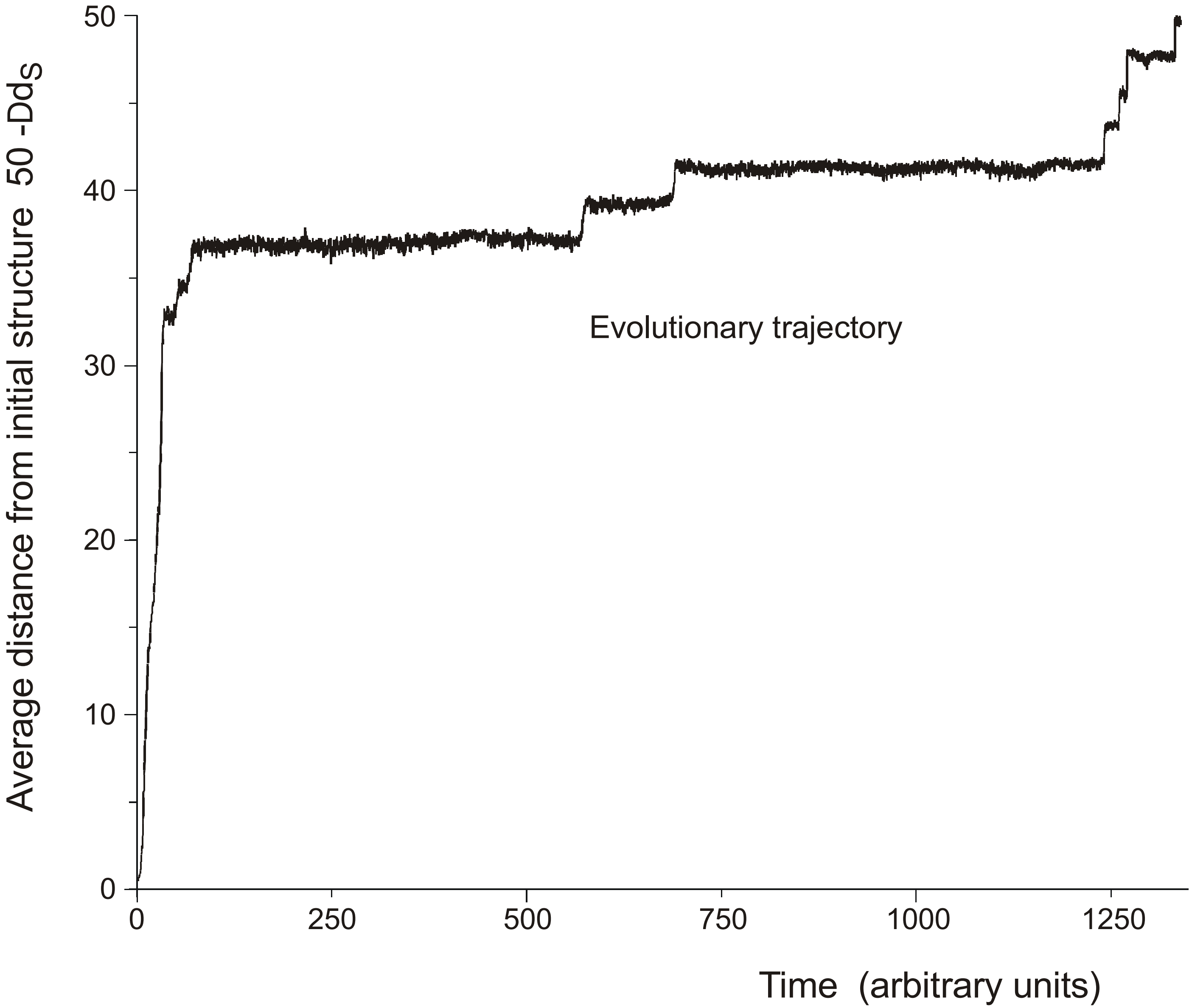
Phenylalanyl-tRNA as target structure



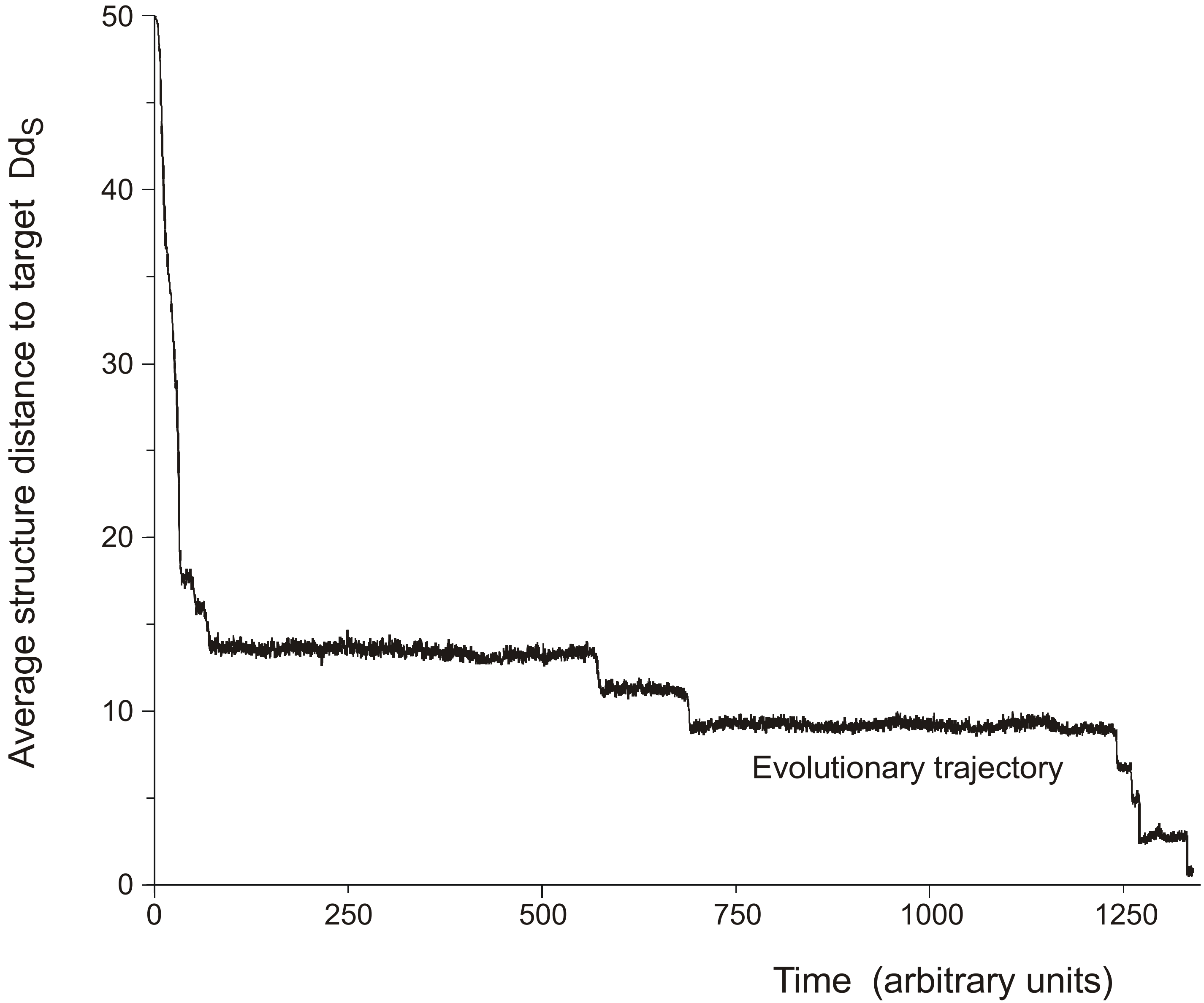
The molecular quasispecies
in sequence space



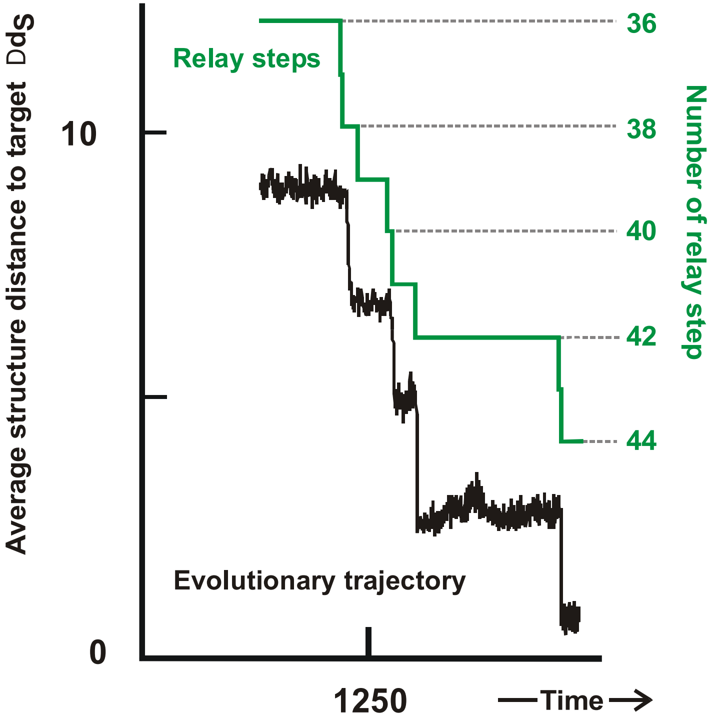
Evolutionary dynamics
including molecular phenotypes

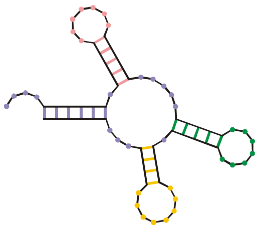


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)



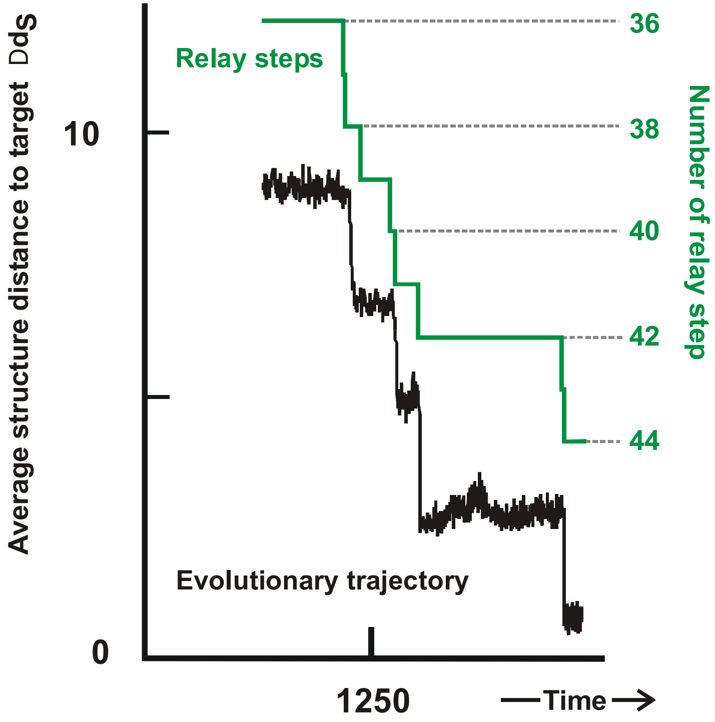
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)

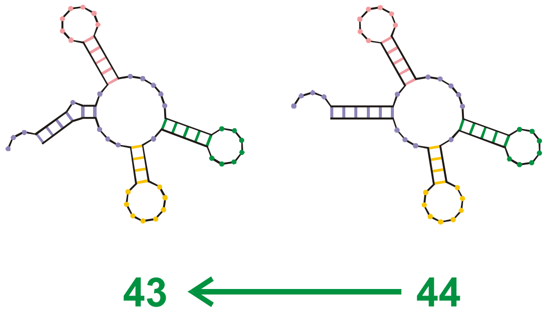




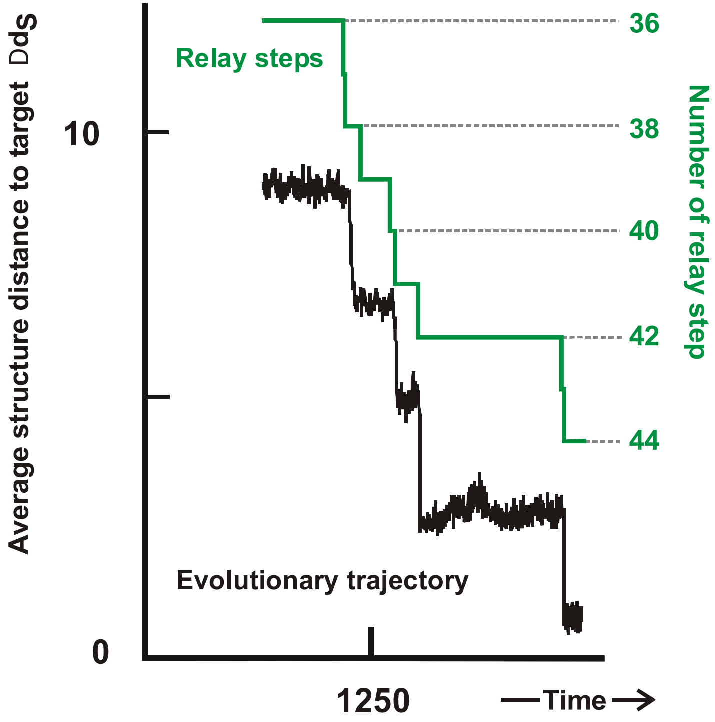
44

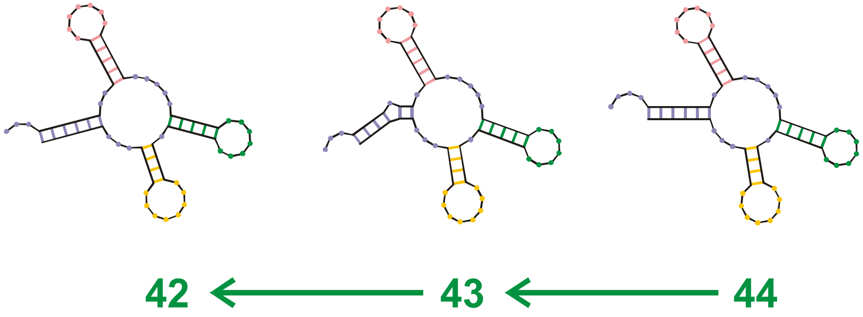
Endconformation of optimization



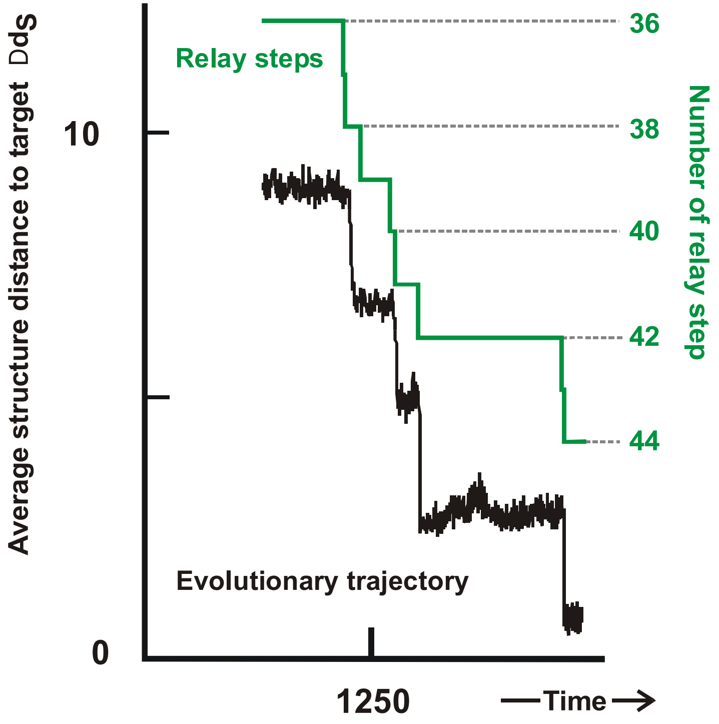


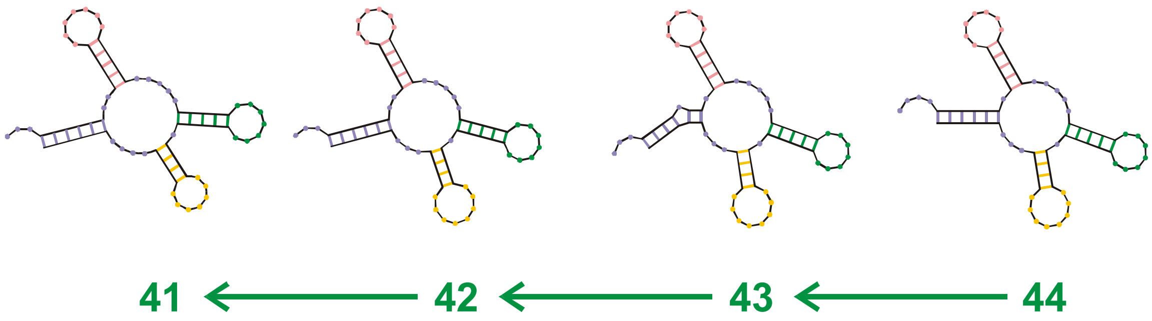
Reconstruction of the last step 43 Á 44



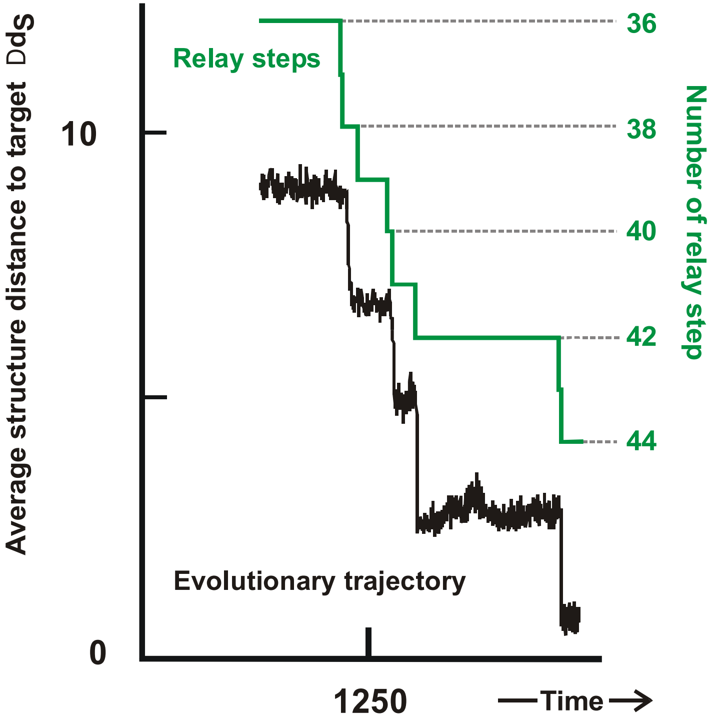


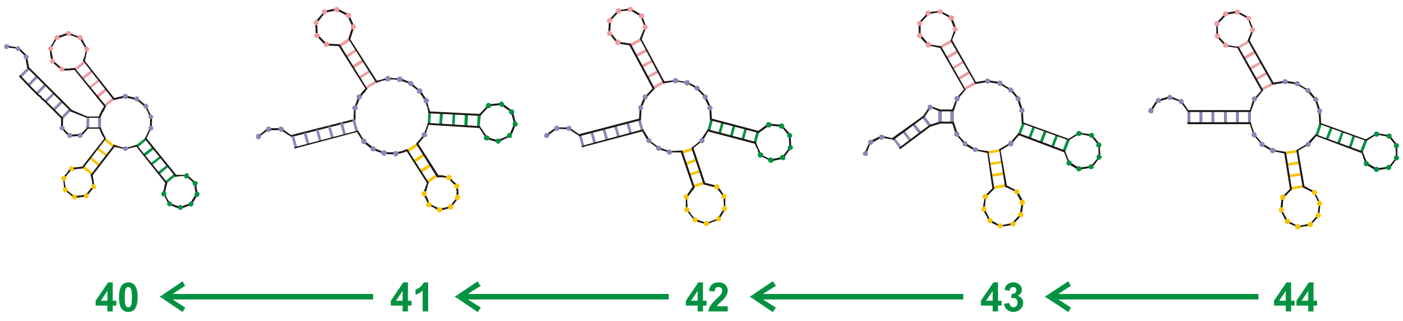
Reconstruction of last-but-one step 42 Á 43 (Á 44)



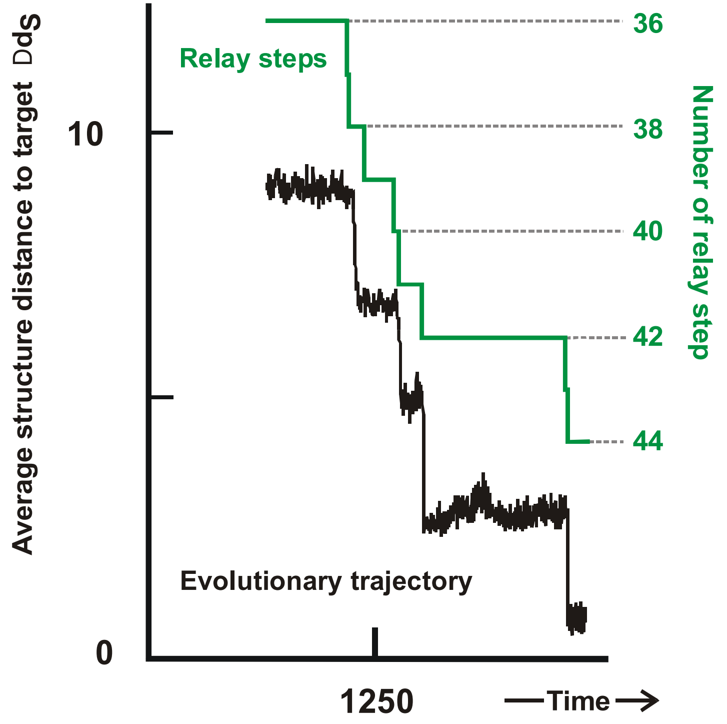


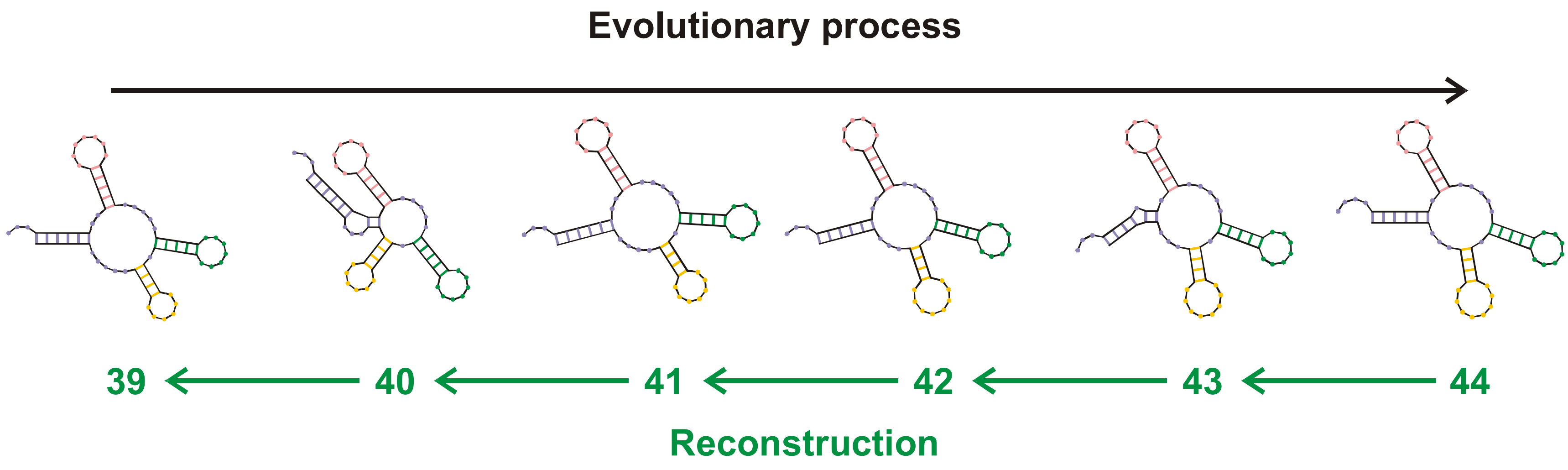
Reconstruction of step 41 Á 42 (Á 43 Á 44)



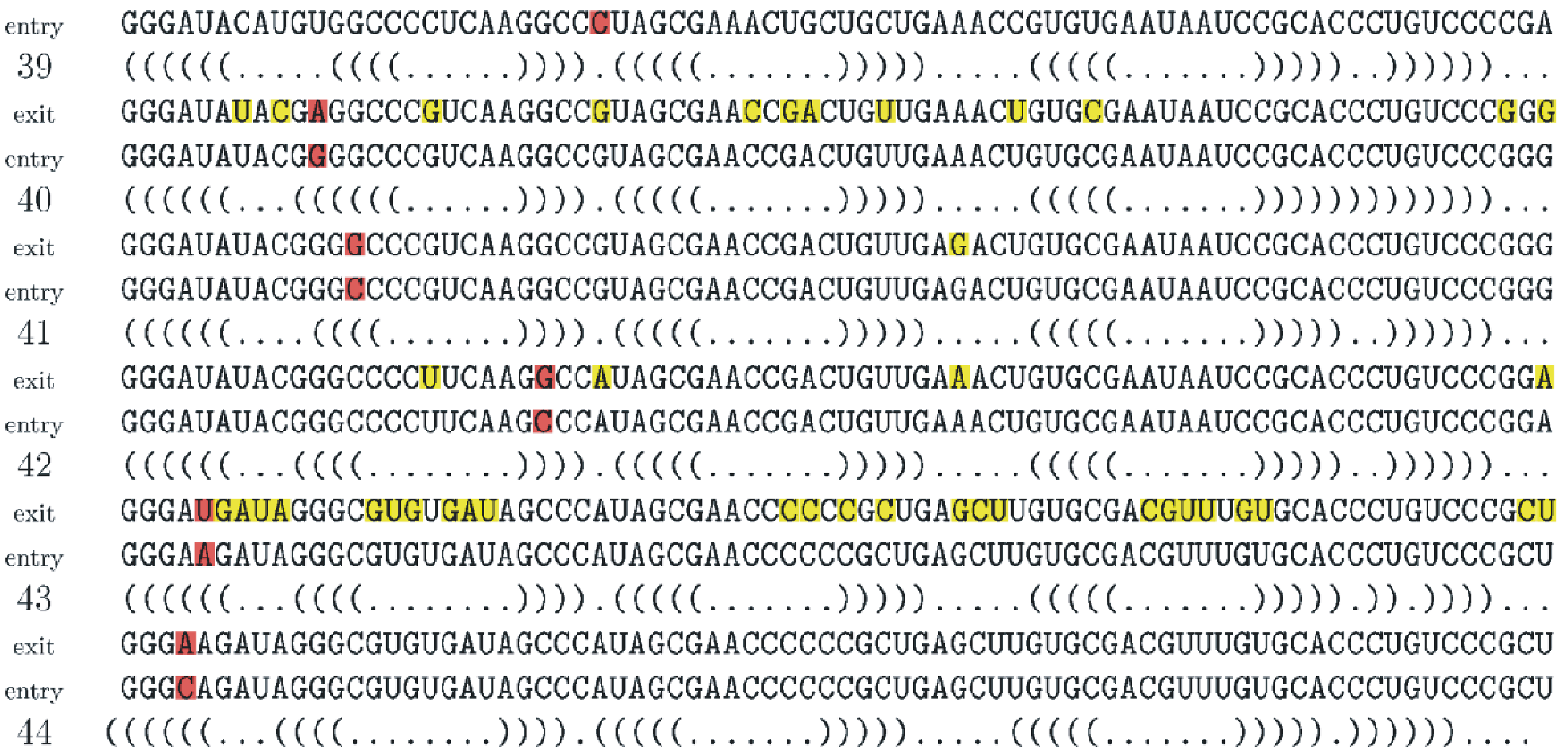


Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)





Reconstruction of the relay series

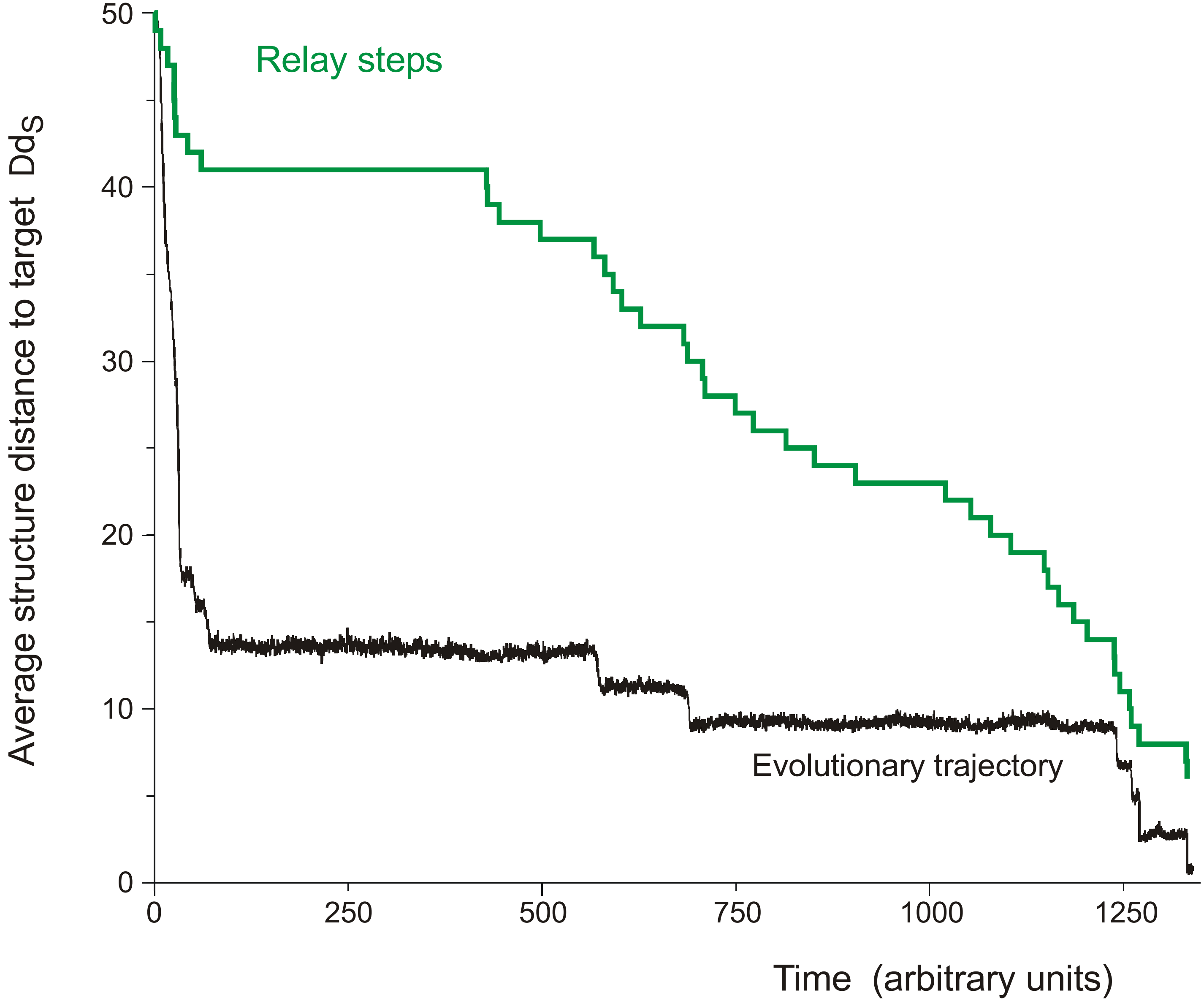


```
entry  GGGAUACAUGUGGCCCCUCAAGGCCCUAGCGAAACUGCUGCUGAAACCGUGUGAAUAAUCCGCACCCUGUCCCCGA
39     ((((((.....((((......)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUAUACGAGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
cntry  GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
40     ((((((...(((((......)))).(((((.......))))).....(((((.......))))))))))))...
exit   GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
entry  GGGAUAUACGGGCCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG
41     ((((((....((((.......)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUAUACGGGCCCCUUCAAGGCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA
entry  GGGAUAUACGGGCCCCUUCAAGCCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA
42     ((((((...((((........)))).(((((.......))))).....(((((.......)))))..))))))...
exit   GGGAUGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
entry  GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
43     ((((((...((((........)))).(((((.......))))).....(((((.......))))).)).))))...
exit   GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
entry  GGGCAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU
44    ((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....
```

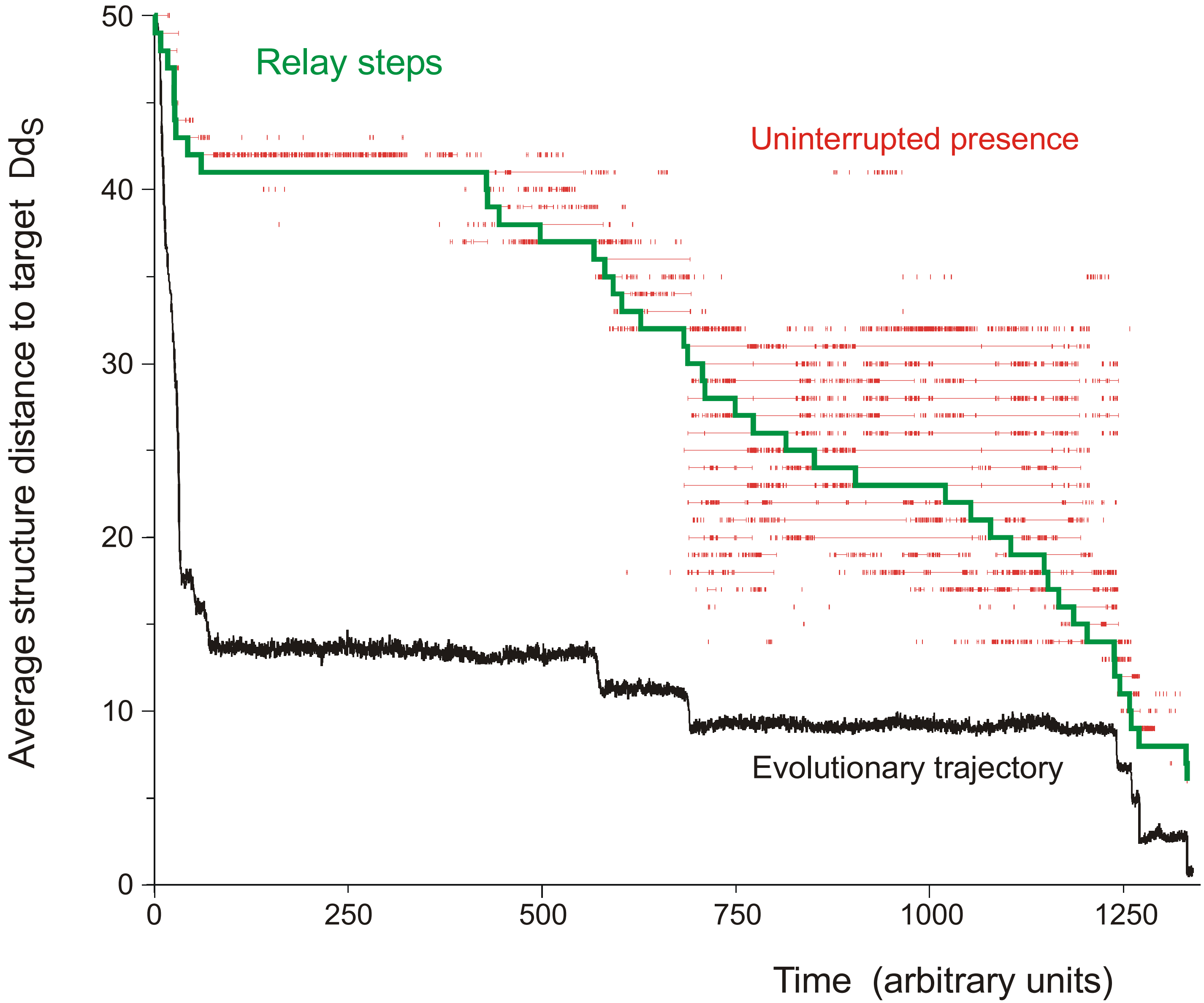
Transition inducing point mutations

Neutral point mutations

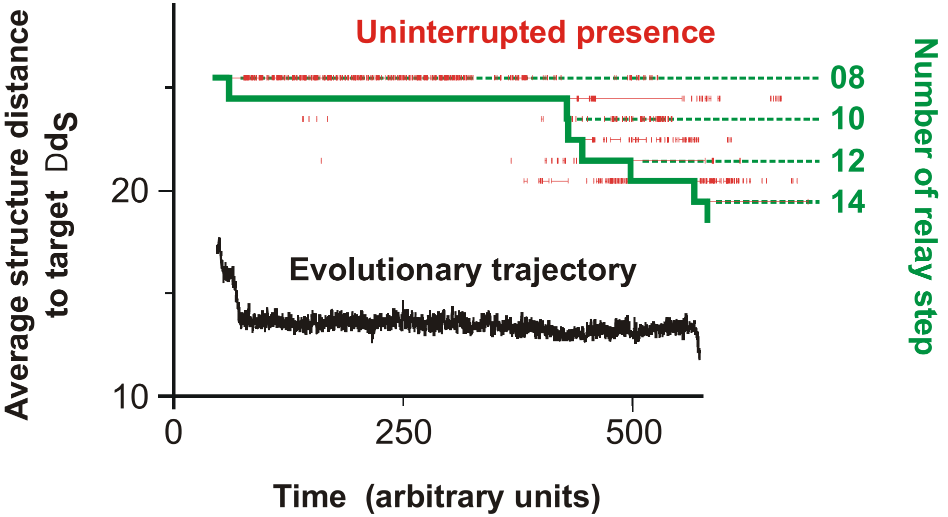
Change in RNA sequences during the final five relay steps 39 Á 44



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



In silico optimization in the flow reactor: Uninterrupted presence

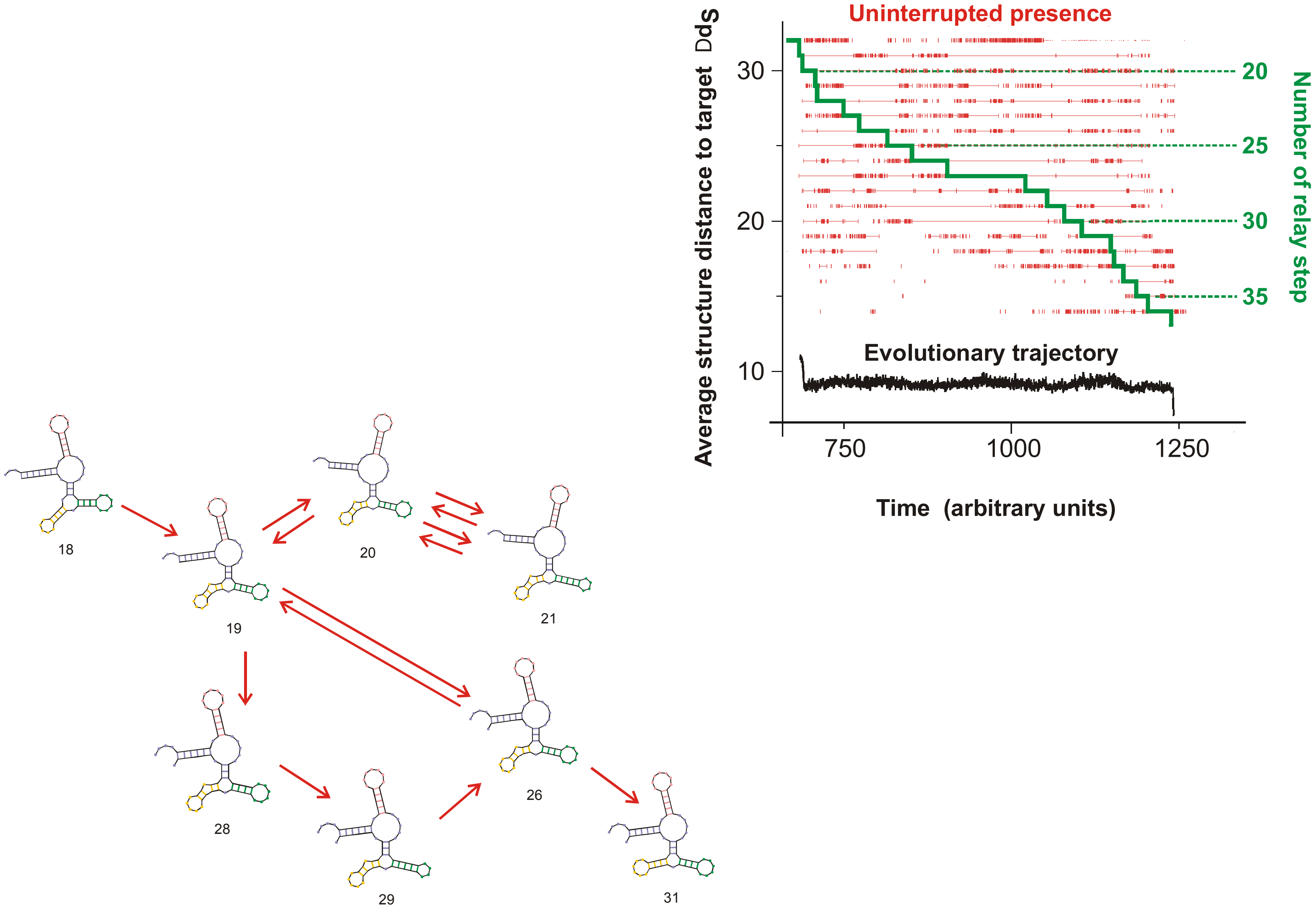


```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
  8    .((((((((((((........(((....))).......))))).....(((((.......))))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

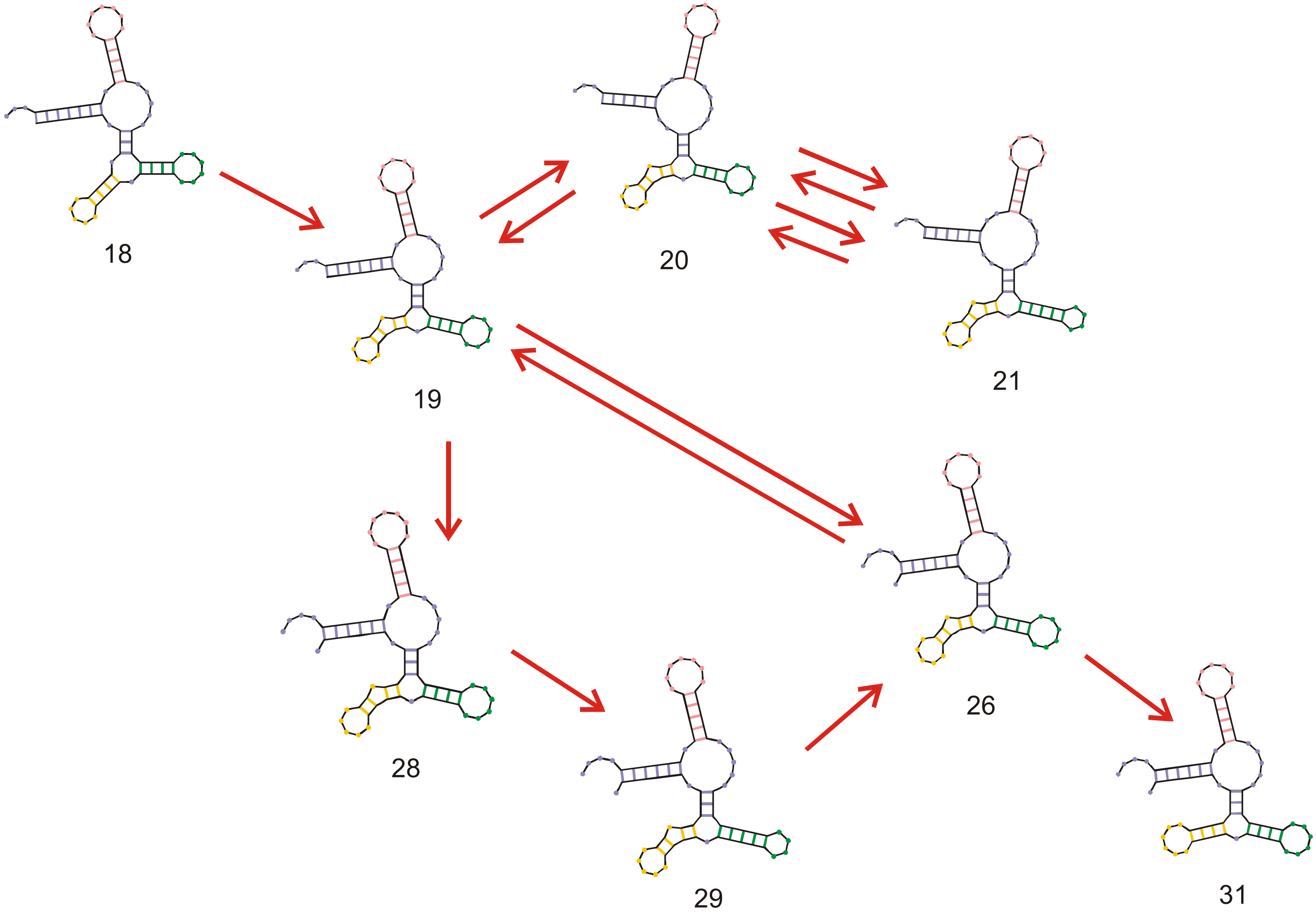
Transition inducing point mutations

Neutral point mutations

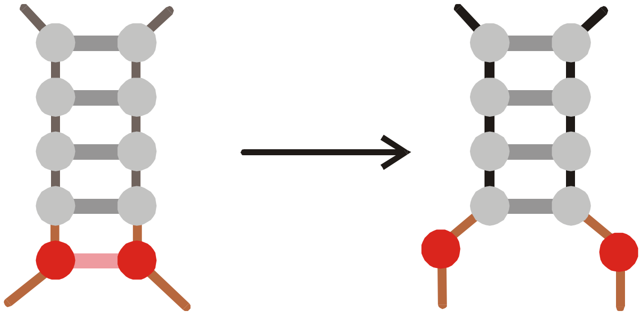
Neutral genotype evolution during phenotypic stasis



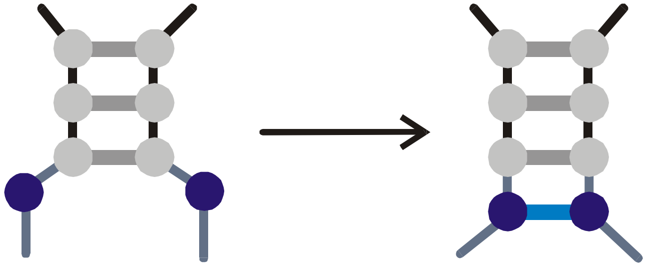
A random sequence of **minor** or continuous **transitions** in the relay series



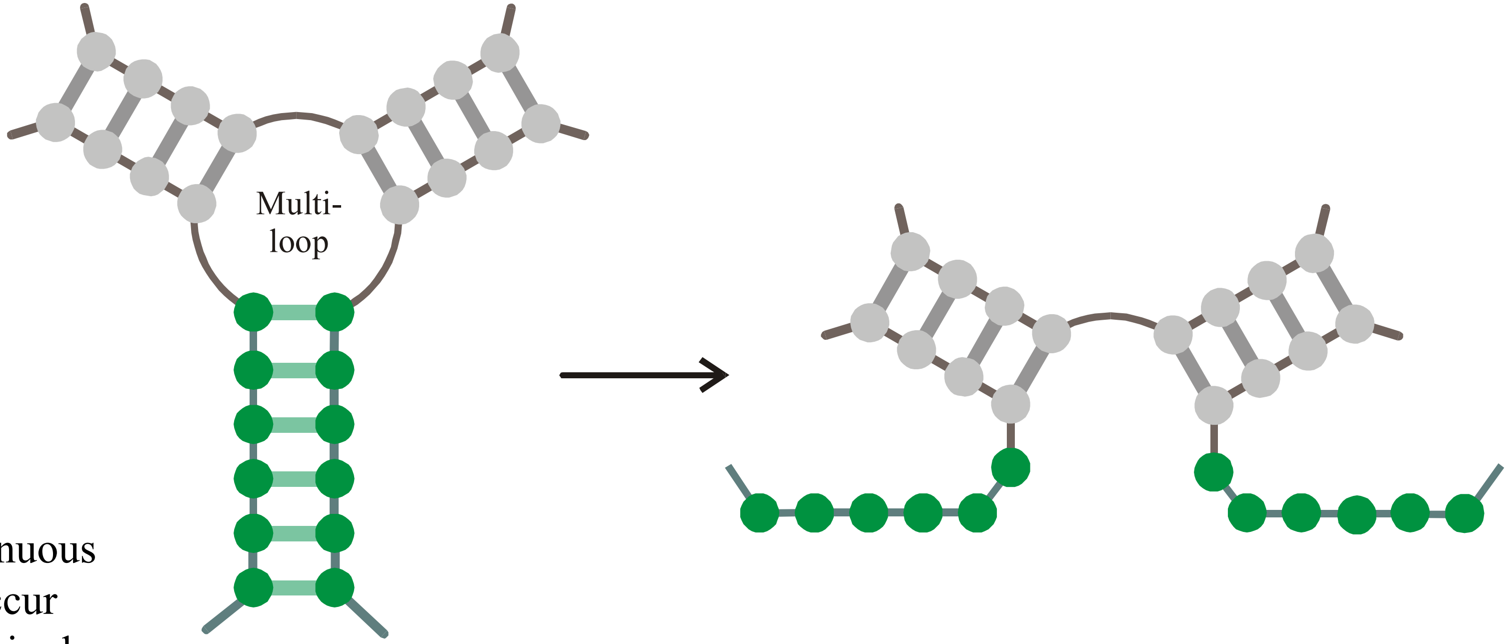
A random sequence of **minor** or continuous **transitions** in the relay series



Shortening of Stacks

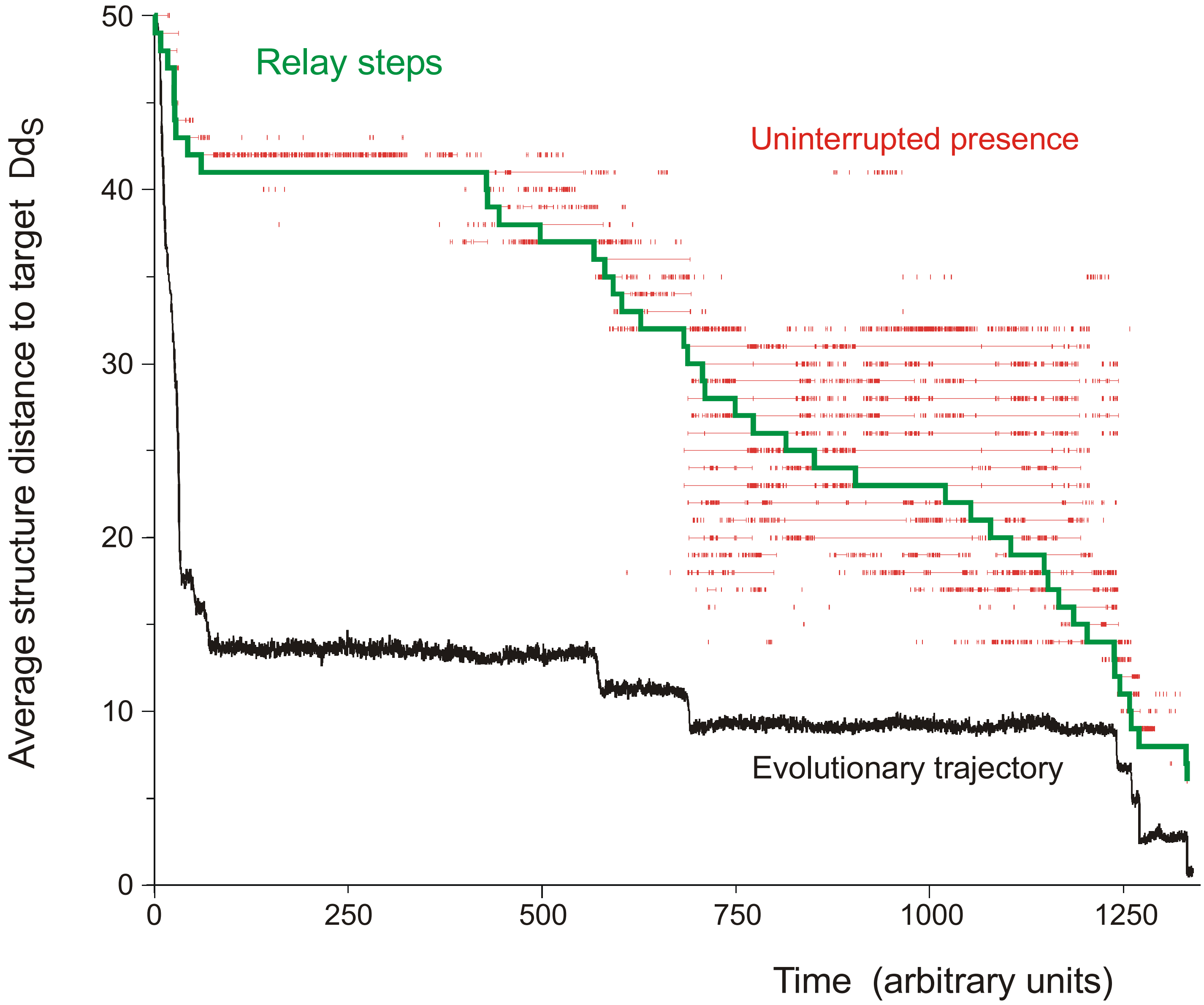


Elongation of Stacks



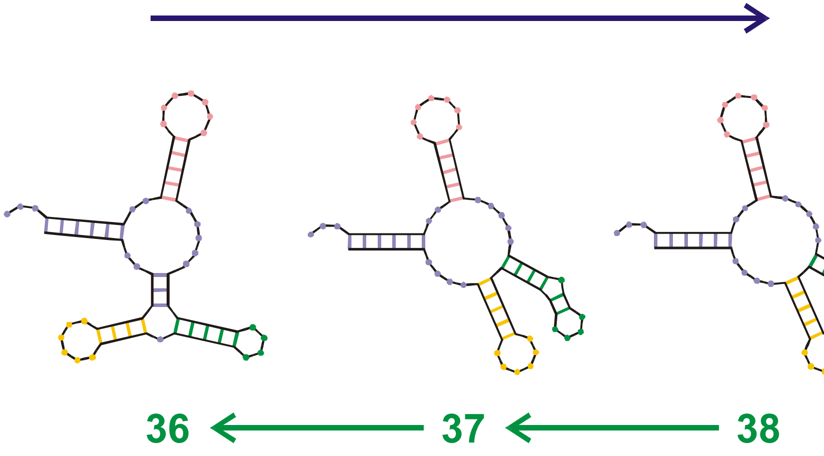
Opening of Constrained Stacks

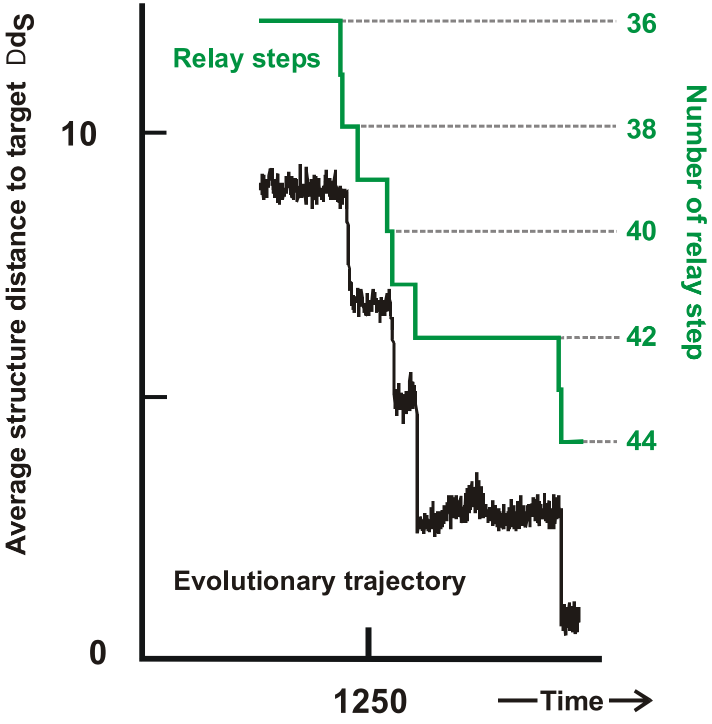
Minor or continuous **transitions**: Occur **frequently** on single point mutations



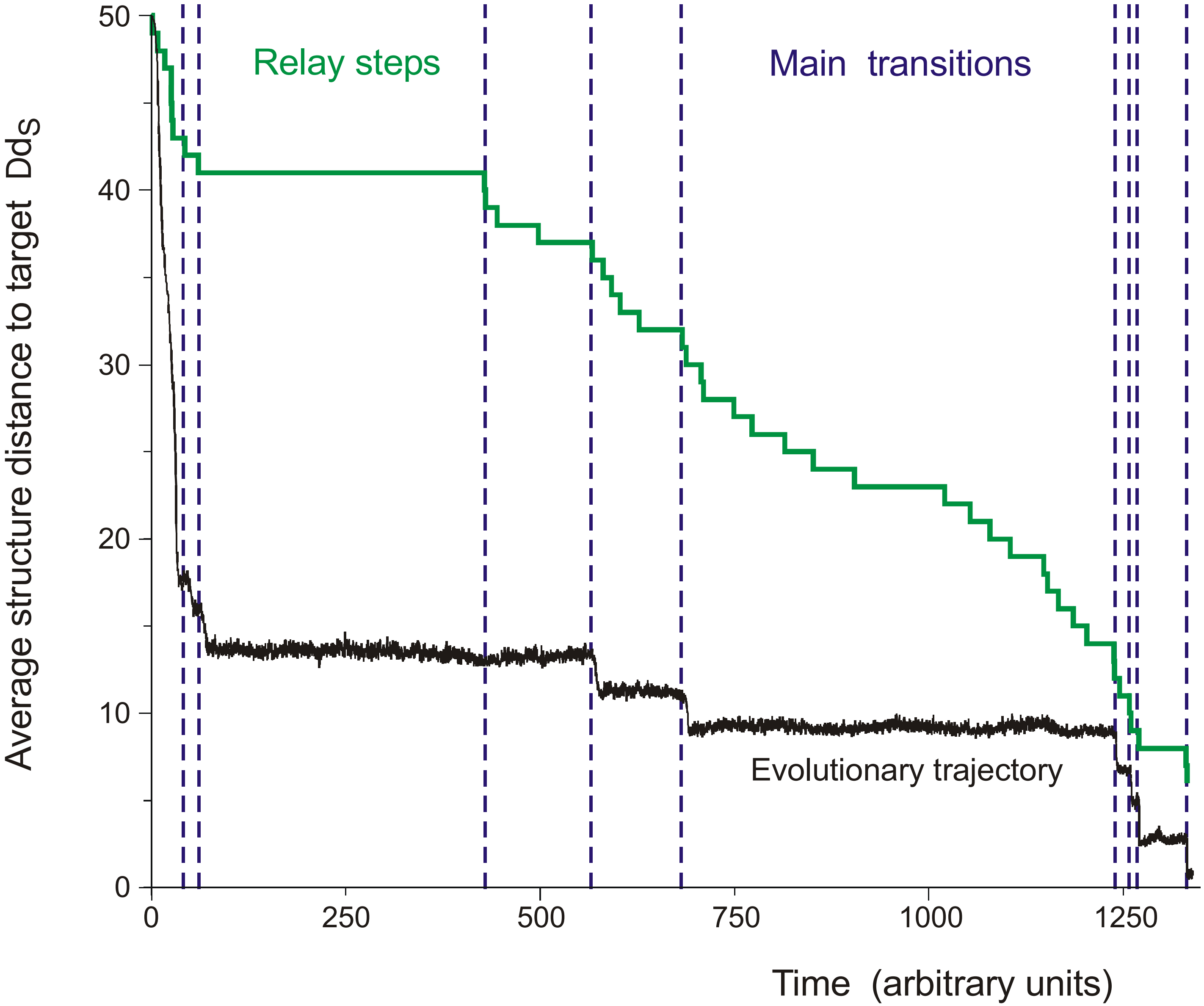
In silico optimization in the flow reactor: **Uninterrupted presence**

Main transition leading to clover leaf

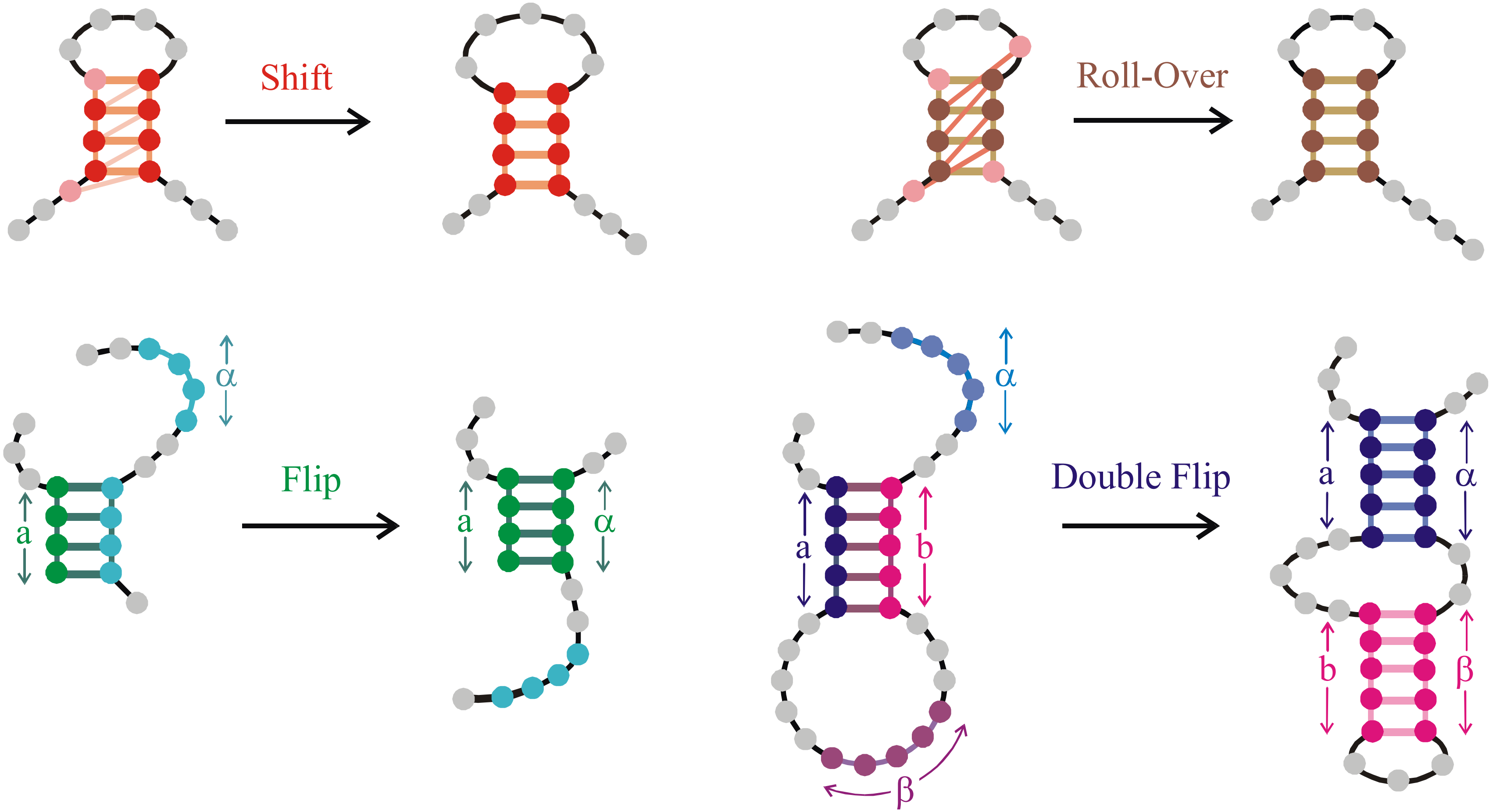




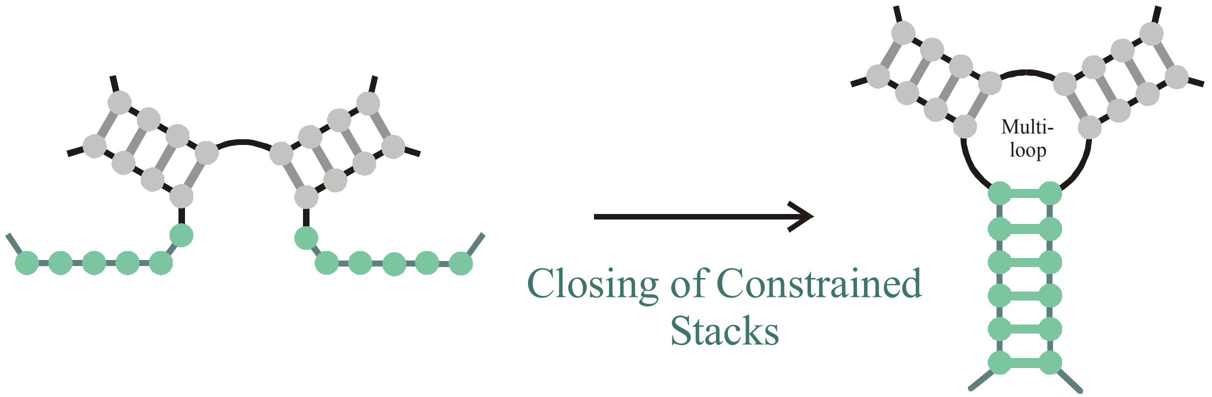
Reconstruction of a main transitions 36 Á 37 (Á 38)

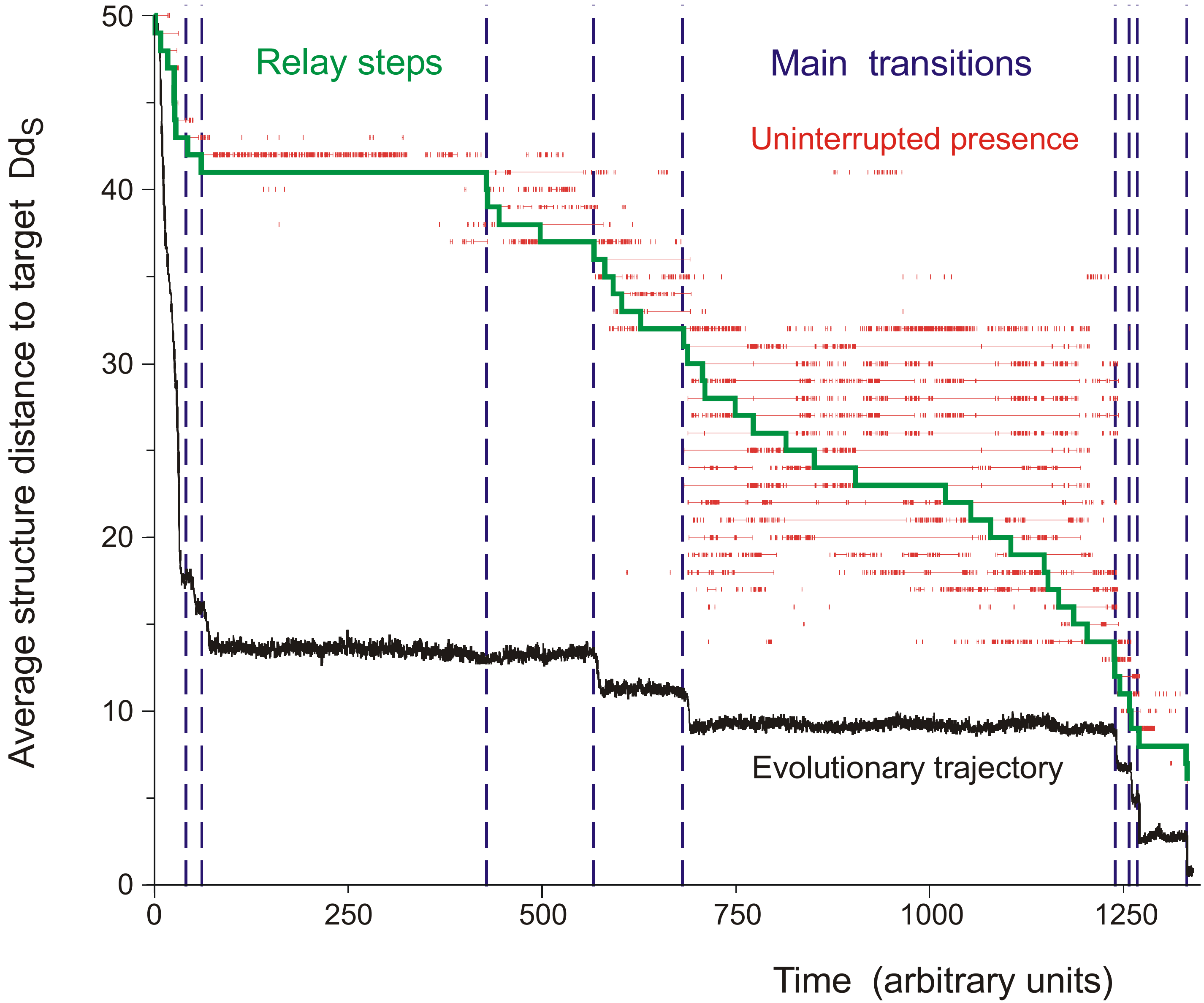


In silico optimization in the flow reactor: Main transitions



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



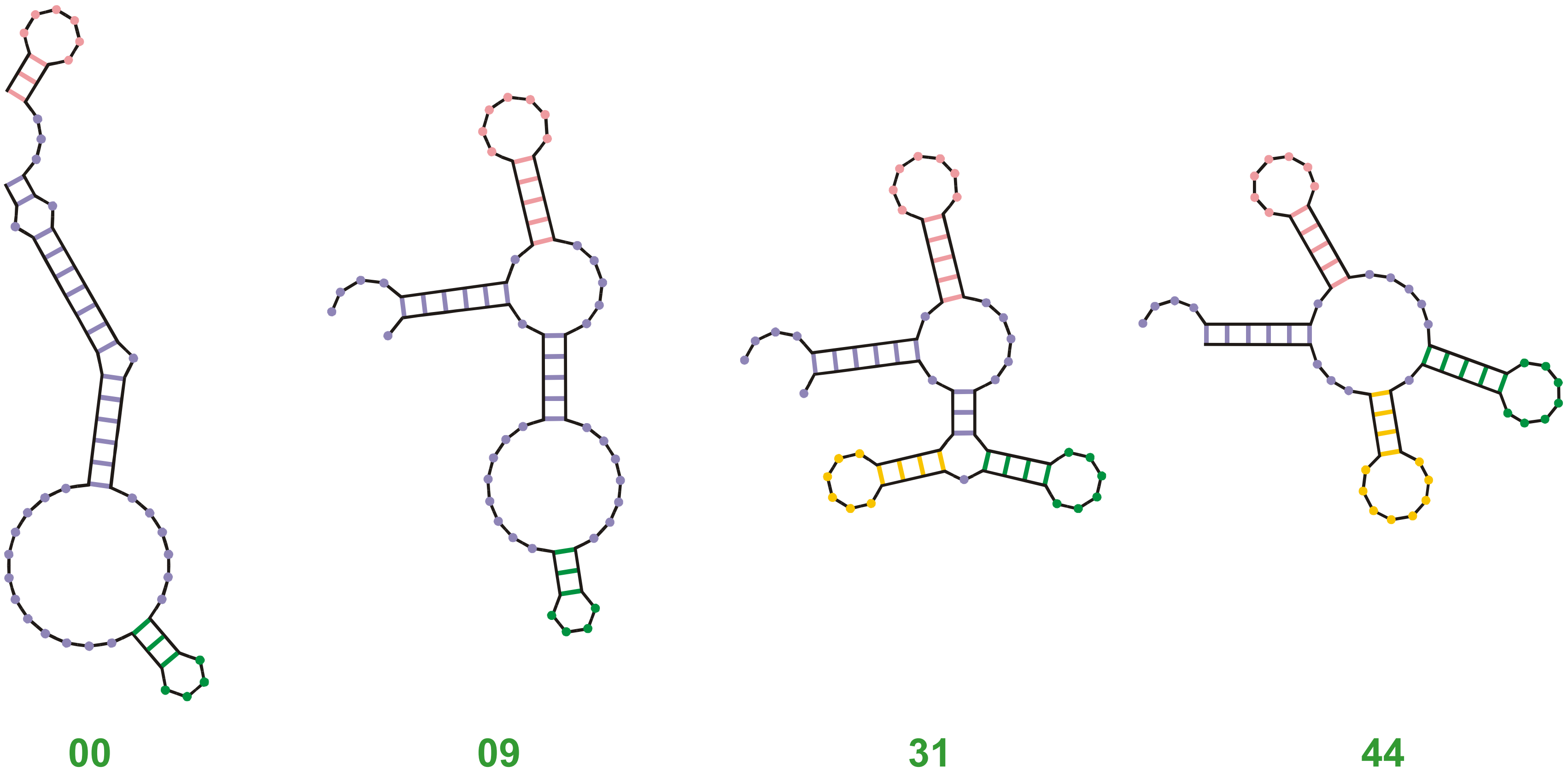


In silico optimization in the flow reactor

Statistics of evolutionary trajectories

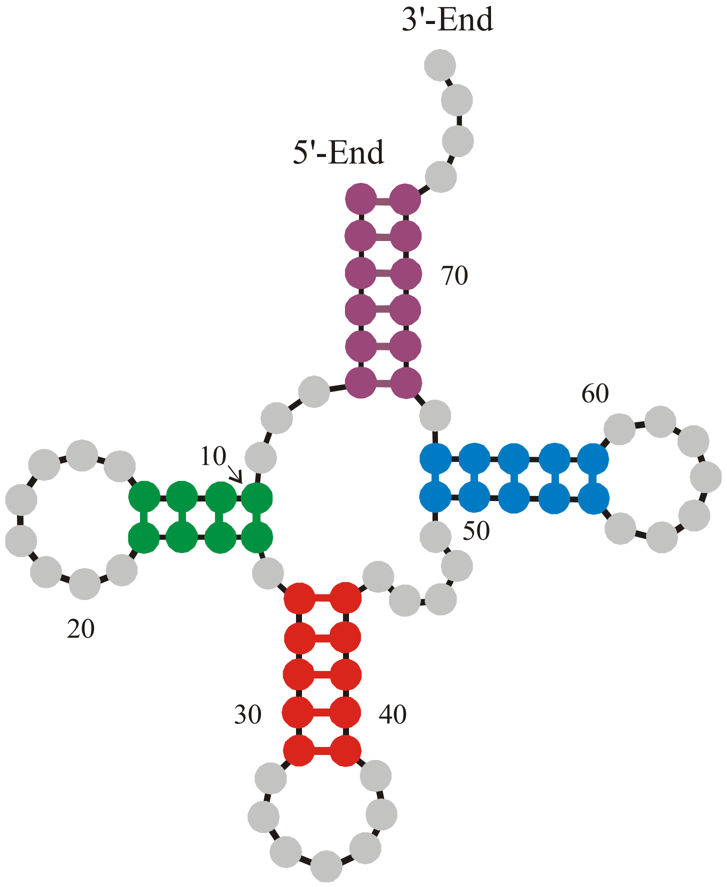
Population size N	Number of replications $< n_{rep} >$	Number of transitions $< n_{tr} >$	Number of main transitions $< n_{dtr} >$
1 000	$(5.5 \pm [6.9,3.1]) \times 10^7$	$92.7 \pm [80.3,43.0]$	$8.8 \pm [2.4,1.9]$
2 000	$(6.0 \pm [11.1,3.9]) \times 10^7$	$55.7 \pm [30.7,19.8]$	$8.9 \pm [2.8,2.1]$
3 000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

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



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

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



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

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

Main results of computer simulations of molecular evolution

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

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

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

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

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

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

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