Some Implications of the RNA Model for a Prebiotic World

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria
and
The Santa Fe Institute, Santa Fe, New Mexico, USA

Chemibiogenesis 2005, COST Action D 27

Venice, 28.09.– 01.10.2005

Web-Page for further information:

http://www.tbi.univie.ac.at/~pks

Three necessary conditions for Darwinian evolution are:

1. **Multiplication,**
2. **Variation**, and
3. **Selection**.

Variation through mutation and recombination operates on the **genotype** whereas the **phenotype** is the target of **selection**.

One important property of the Darwinian scenario is that variations in the form of mutations or recombination events occur uncorrelated with their effects on the selection process.

All conditions can be fulfilled not only by cellular organisms but also by nucleic acid molecules in suitable cell-free experimental assays.

Population genetics and chemical reaction kinetics count only numbers of molecules and model their changes in time.

Molecular properties and functions enter as parameters and mechanisms and are inputs and no intrinsic part of the model.

Understanding evolution requires a concept of the phenotype as an intergal part of the model.

Genotype = Genome

Mutation → GGCTATCGTACGTTTACCCAAAAAGTCTACGTTGGACCCAGGCATTGGAC.......G

Unfolding of the genotype:

Production and assembly of all parts of a bacterial cell, and cell division

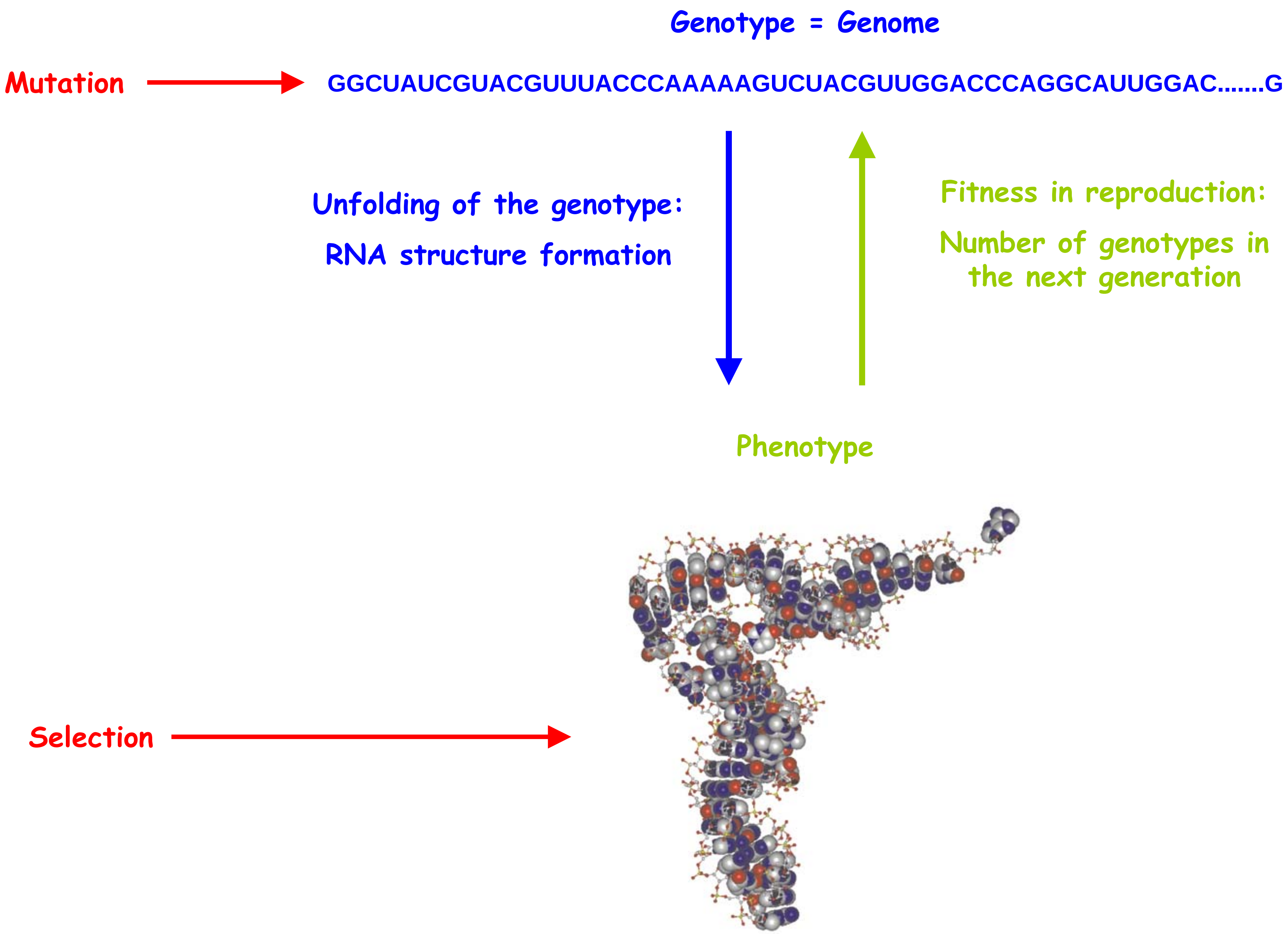
Fitness in reproduction:

Number of bacterial cells in the next generation

Phenotype

Selection →

Evolution of phenotypes: Bacterial cells



Genotype = Genome

Mutation

GGCUAUCGUACGUUUACCCAAAAAGUCUACGUUGGACCCAGGCAUUGGAC.......G

Unfolding of the genotype:

RNA structure formation

Fitness in reproduction:

Number of genotypes in the next generation

Phenotype

Selection

Evolution of phenotypes

Genotype = Genome

Mutation → GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGAC.......G

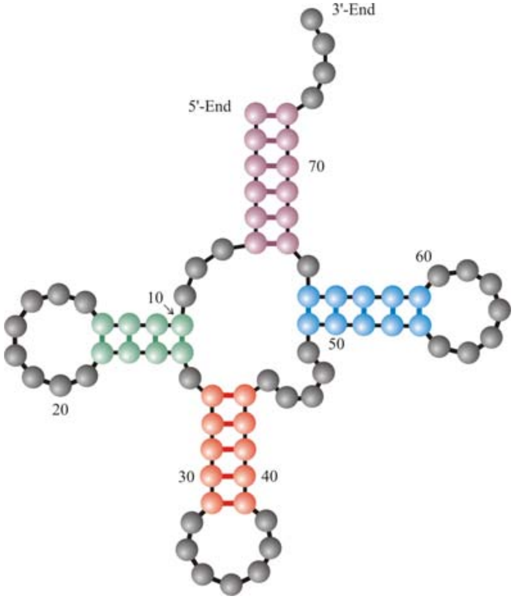
Unfolding of the genotype:

RNA structure formation

Fitness in reproduction:

Number of genotypes in the next generation

Phenotype



Selection →

Evolution of phenotypes

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. This definition allows for rigorous mathematical analysis by means of combinatorics.

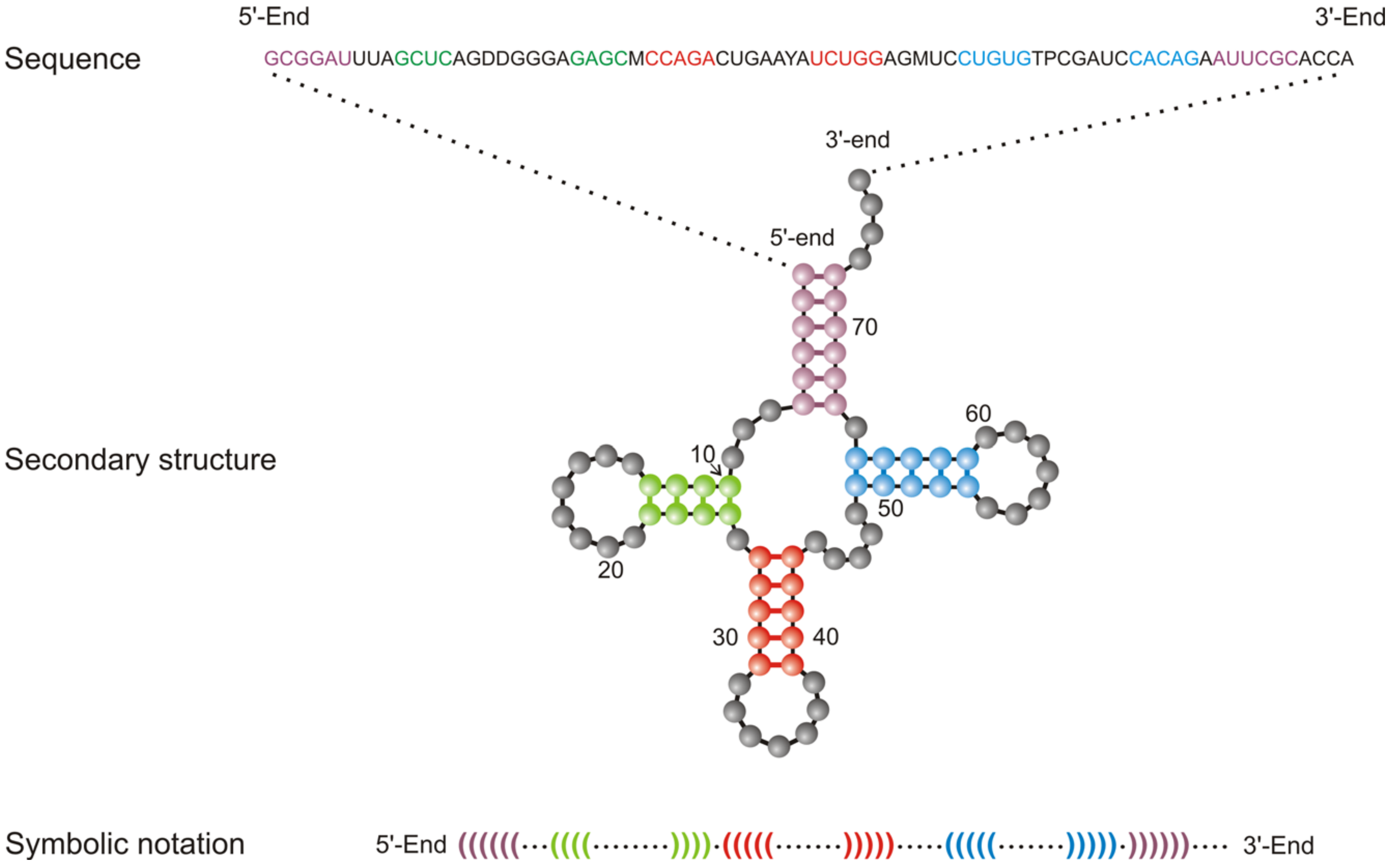
D.Thirumalai, N.Lee, S.A.Woodson, and D.K.Klimov.
Annu.Rev.Phys.Chem. **52**:751-762 (2001):

„Secondary structures are folding intermediates in the formation of full three-dimensional structures.“

Secondary structures have been and still are frequently used to predict and discuss RNA function.

1. The RNA model
2. How many stable structures can be formed?
3. Why not binary (GC or AU) sequences?
4. Evolution on neutral networks
5. Multiconformational RNA molecules

1. **The RNA model**
2. How many stable structures can be formed?
3. Why not binary (GC or AU) sequences?
4. Evolution on neutral networks
5. Multiconformational RNA molecules


Sequence
5'-End
3'-End
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA
3'-end
5'-end
70
60
Secondary structure
10
50
20
30
40
Symbolic notation
5'-End ((((((···((((········))))·(((((·······)))))·····(((((·······)))))·))))))···· 3'-End

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

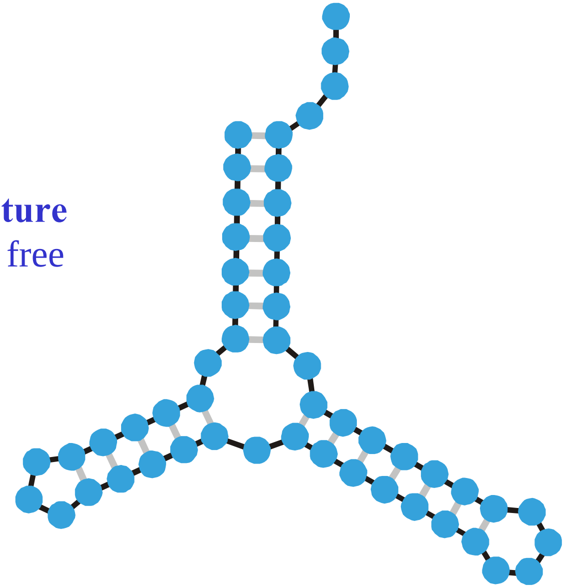
RNA folding:

Structural biology,
spectroscopy of
biomolecules,
understanding
molecular function

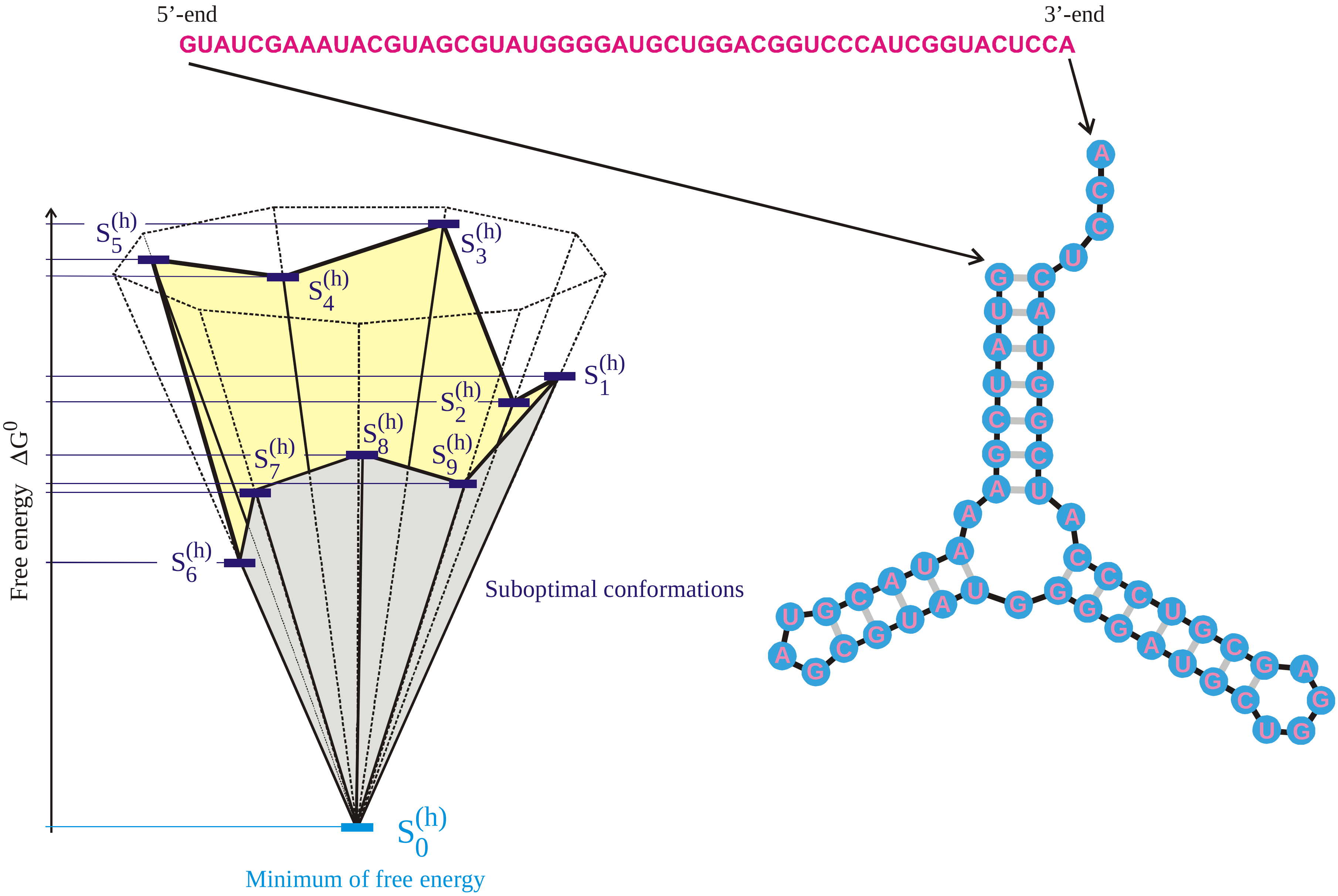
Biophysical chemistry:
thermodynamics and
kinetics

Empirical parameters

RNA structure
of minimal free
energy



Sequence, structure, and design



The minimum free energy structures on a discrete space of conformations

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

Structural biology, spectroscopy of biomolecules, understanding **molecular function**

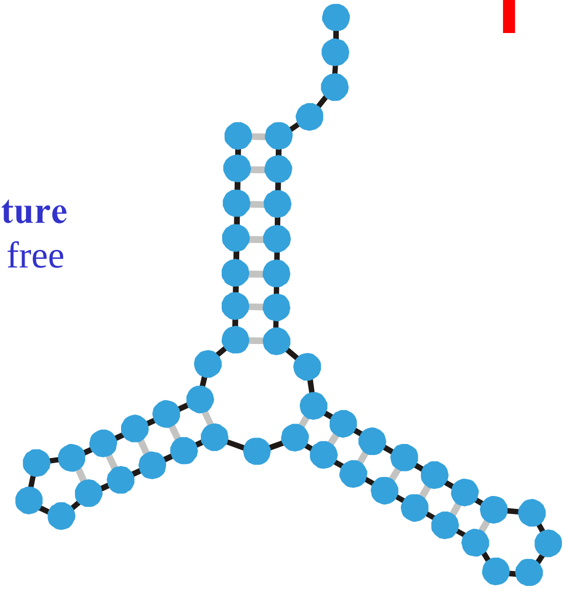
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

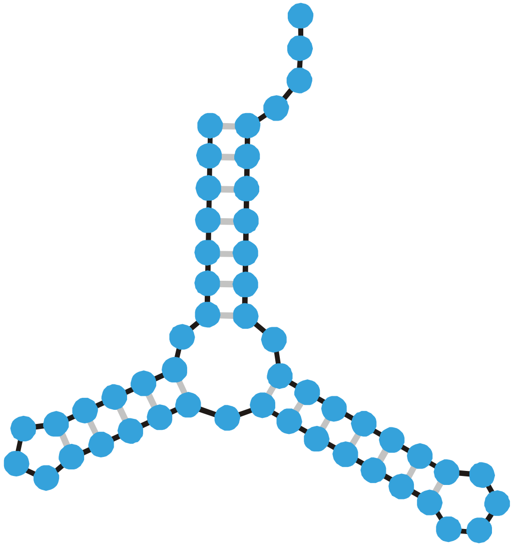
Inverse folding of RNA:

Biotechnology, **design of biomolecules** with predefined structures and functions

RNA structure of minimal free energy



Sequence, structure, and design



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

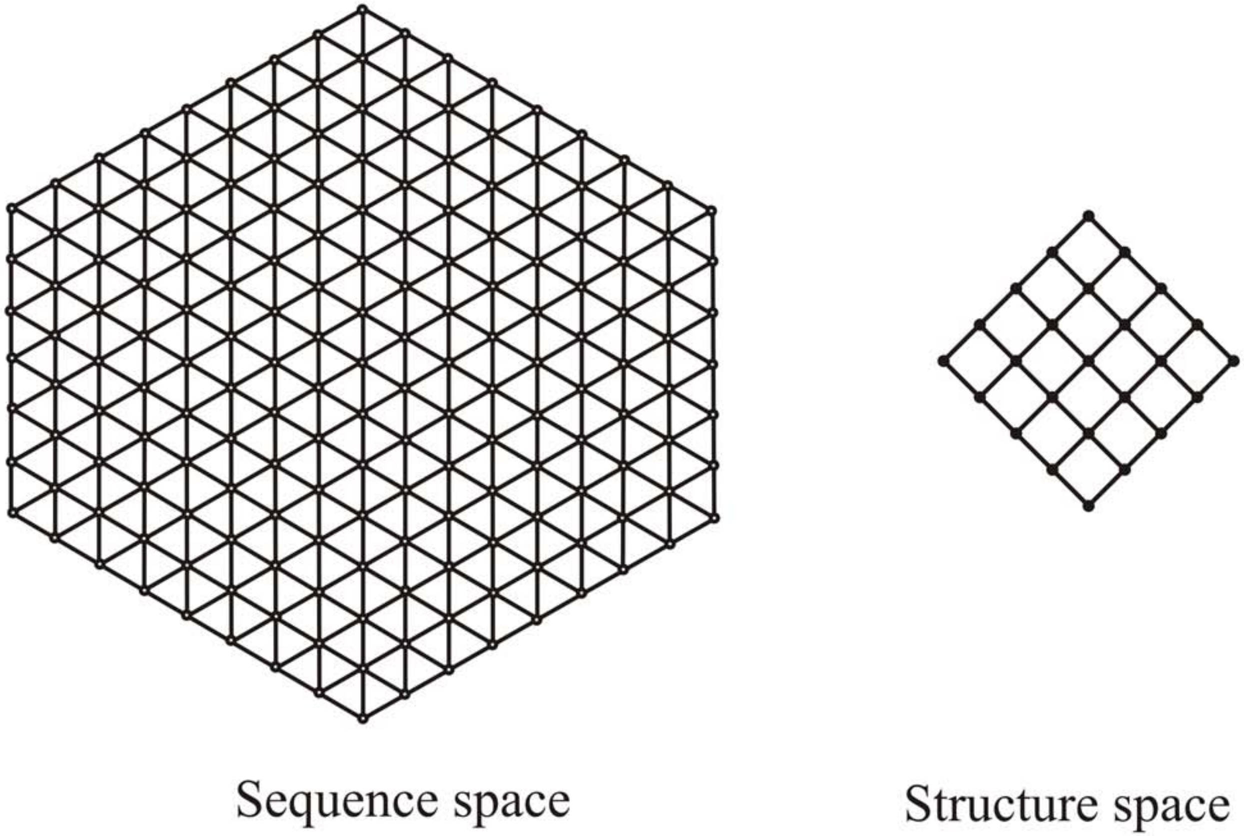
UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG

Inverse folding

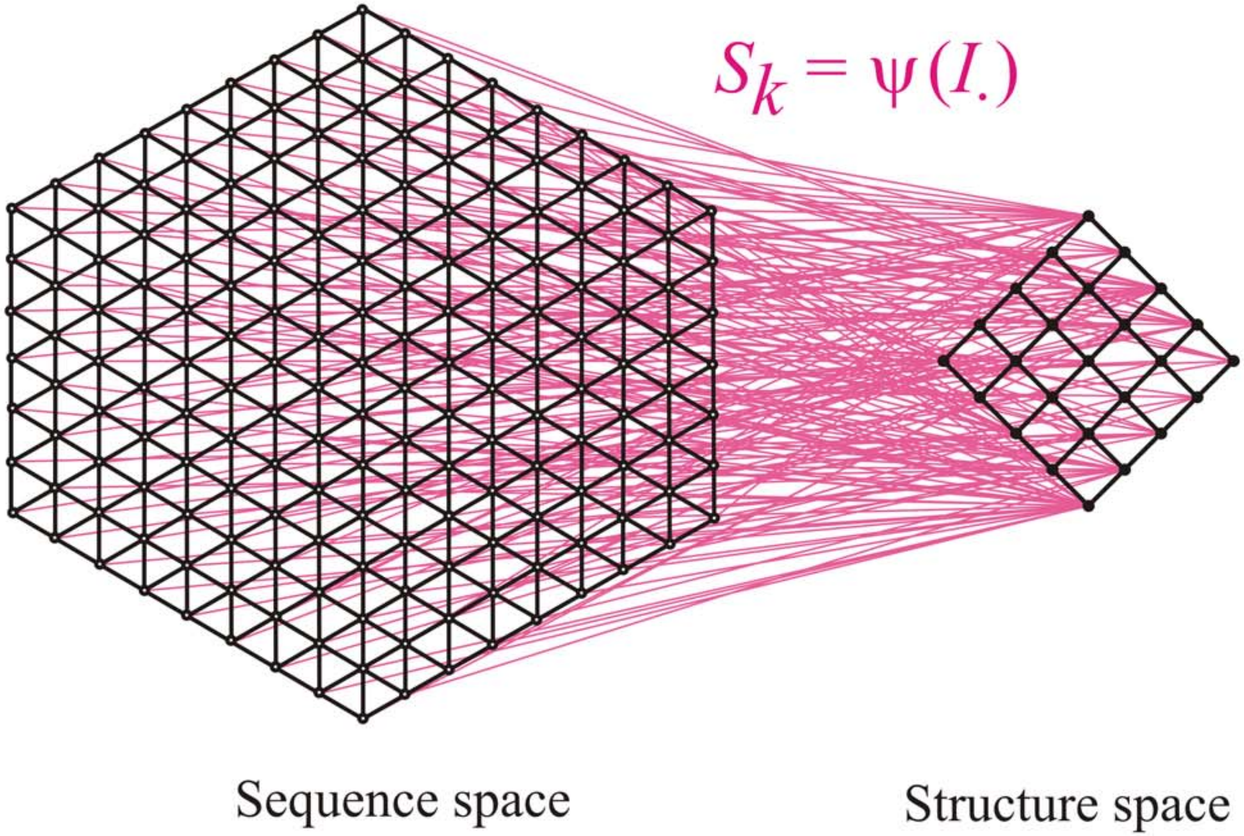
The **inverse folding algorithm** searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.



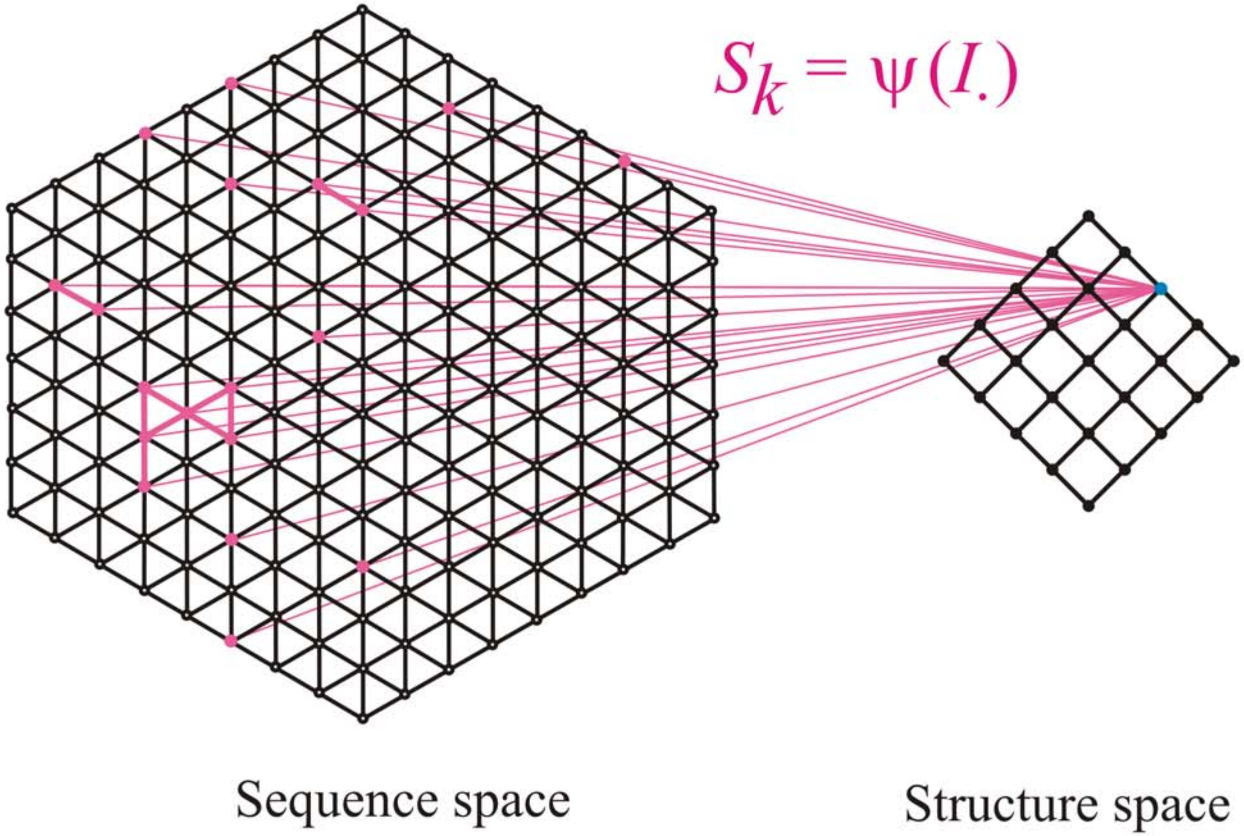
Sequence space

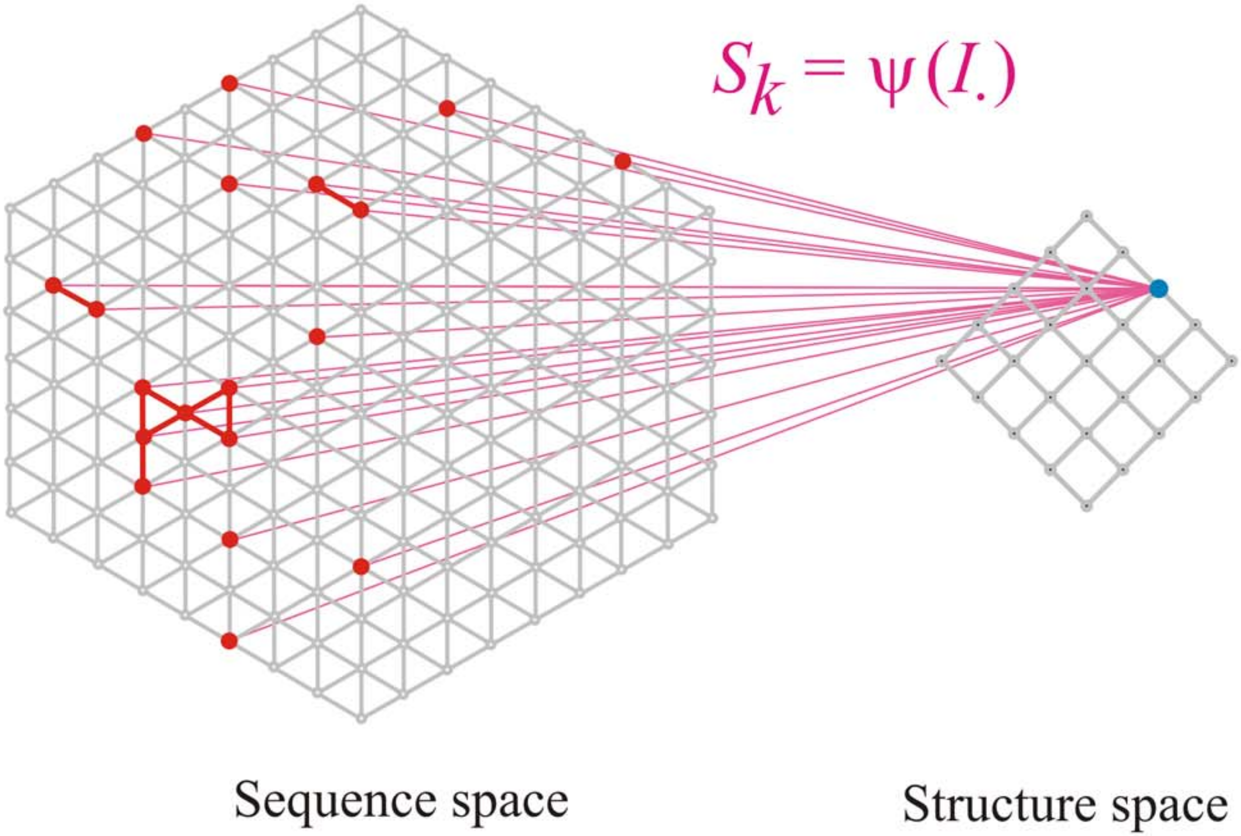
Structure space

Sequence space and structure space



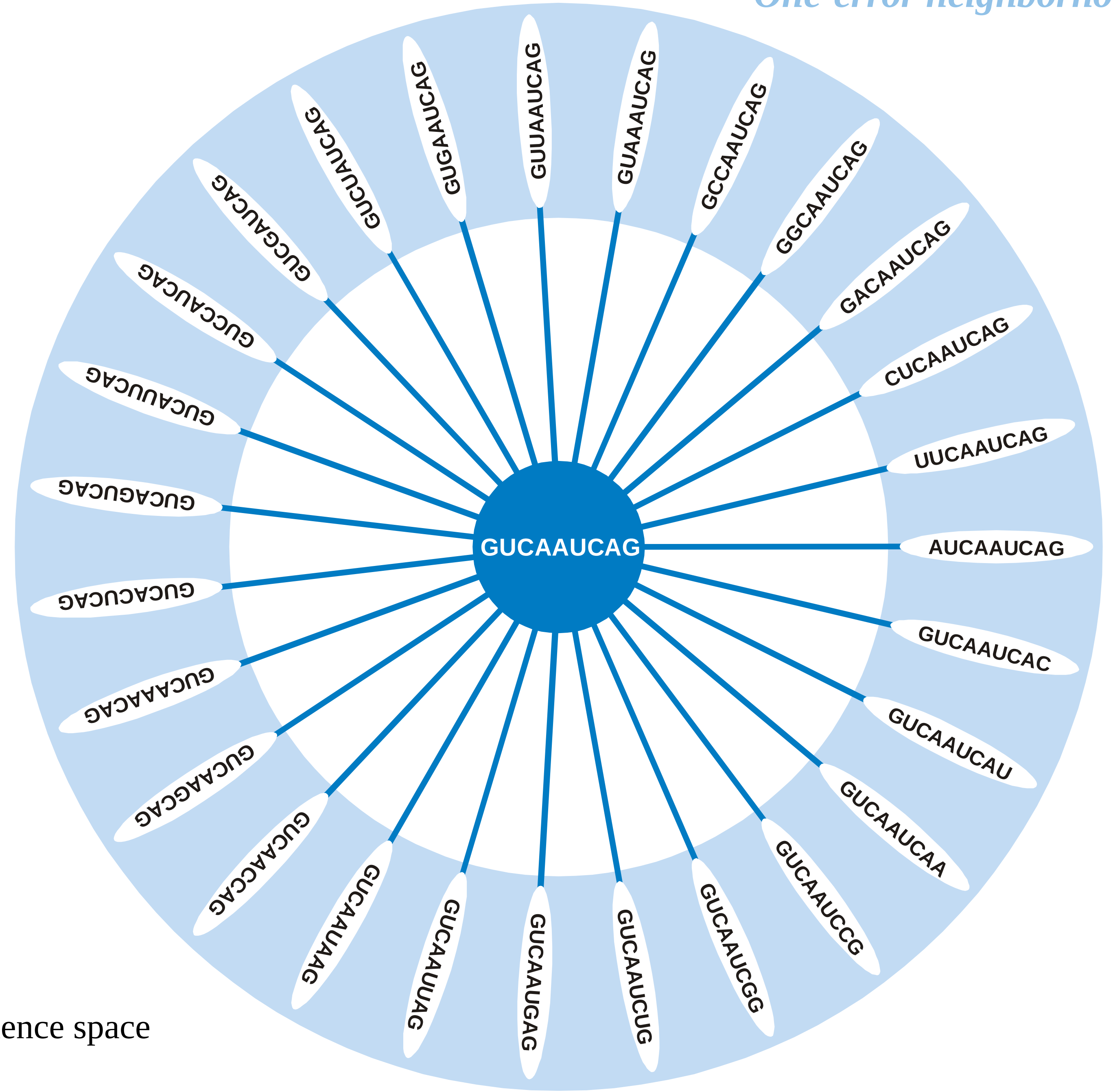
Mapping from sequence space into structure space


$S_k = \psi(I.)$
Sequence space
Structure space



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

One-error neighborhood



The surrounding of
GUCAAUCAG in sequence space

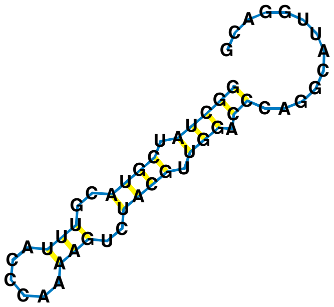
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

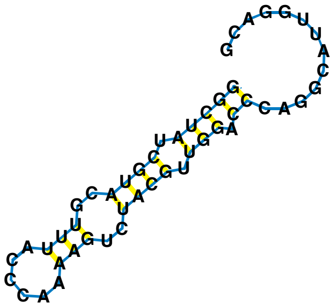
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



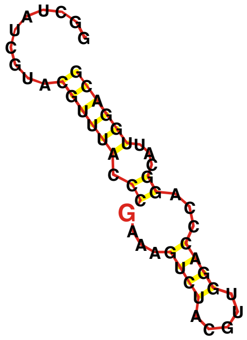
One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

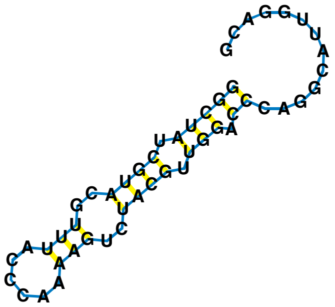


One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

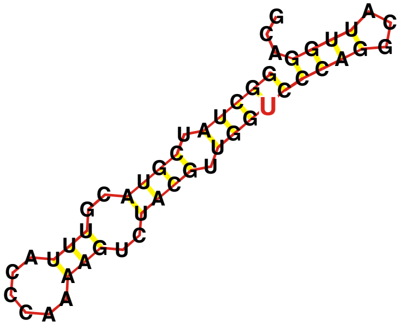


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

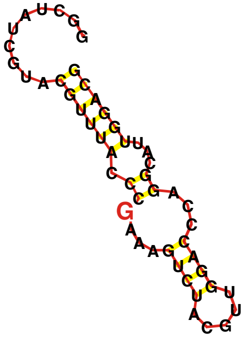
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

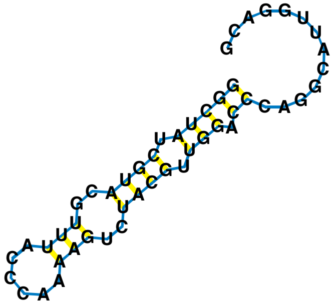


GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

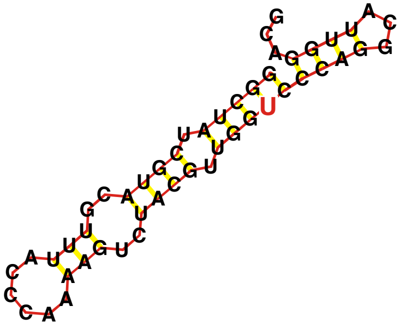


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

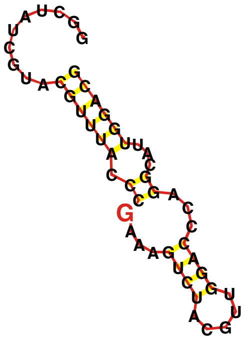
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space



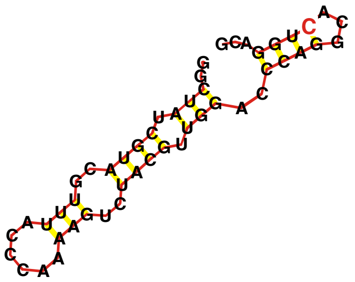
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

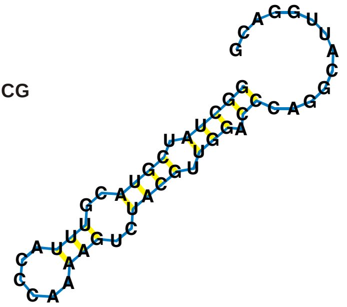


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

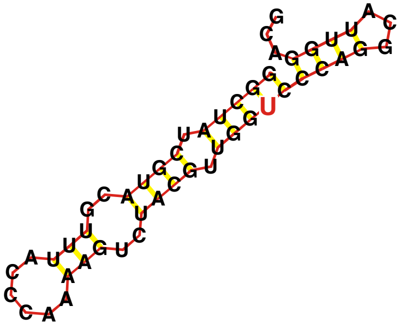
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG

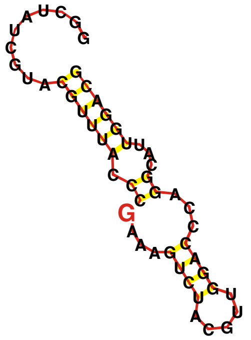




One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space



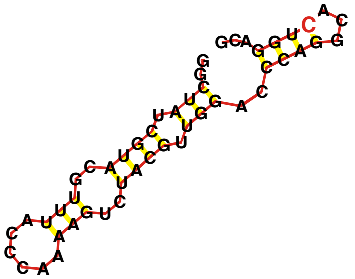
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

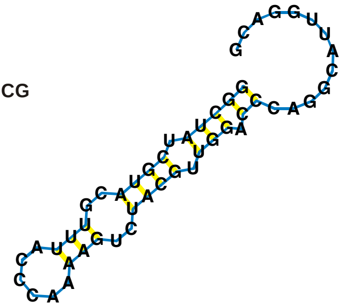


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

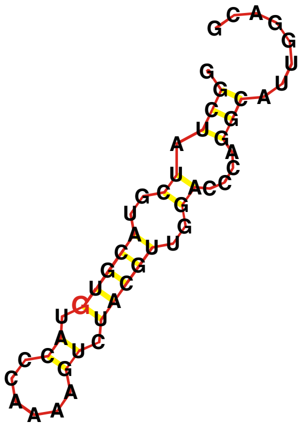
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG



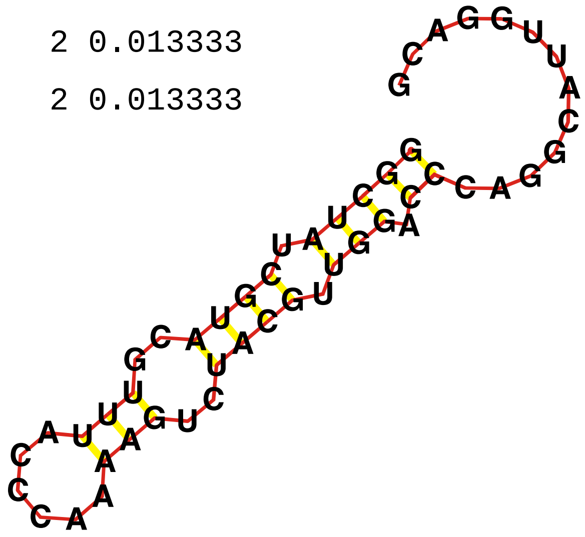


GGCUAUCGUACGUGUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



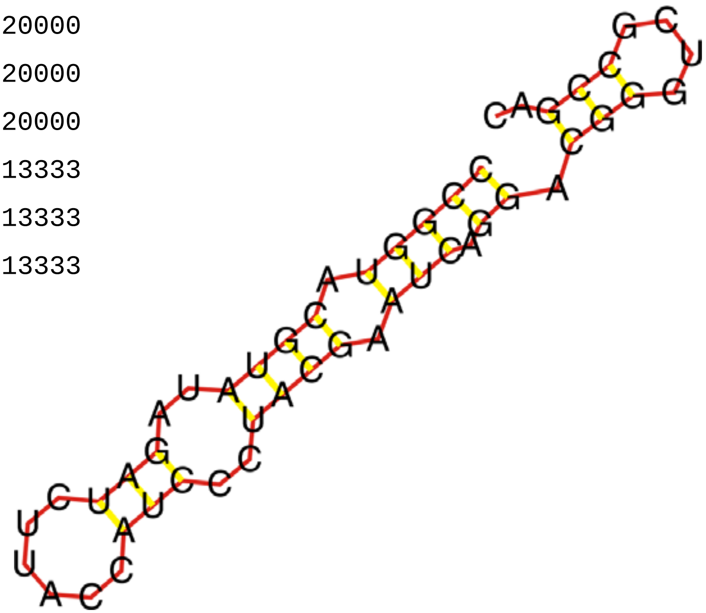
One error neighborhood – Surrounding of an RNA molecule (n=50) in sequence and shape space

```
GCAGCUUGCCCAAUGCAACCCCAUGUGGCGCGCUAGCUAACACCAUCCCC
 1 (((((.((((..(((......)))..)))).))).)).............      65 0.433333
 2 ..(((((((((((((......))).))).)))..))))............       9 0.060000
 3 (((((.((((....(((......))))))).))).)).............       5 0.033333
 4 ..(((.((((..(((......)))..)))).)))................       5 0.033333
 5 ..(((((((((((((......))).)))...)))))))............       4 0.026667
 6 (((((.((((((.((.....)).)).)))).))).)).............       3 0.020000
 7 (((((.((((.((((......)))).)))).))).)).............       3 0.020000
 8 (((((.(((((.(((......))).))))).))).)).............       3 0.020000
 9 ((((((((((..(((......)))..))))))))).)).............       3 0.020000
10 (((((.((((((...........)).)))).))).)).............       3 0.020000
11 (((((..(((..(((......)))..)))..))).)).............       2 0.013333
12 (((((.((((..(((......)))..)))).)).))).............       2 0.013333
13 ..((((.((.(..((((......))))..).)).))))............       2 0.013333
14 (((((.((.((((((......))).))))).))).)).............       2 0.013333
15 .((((((((((((((......))).))).)))..)))))...........       2 0.013333
```

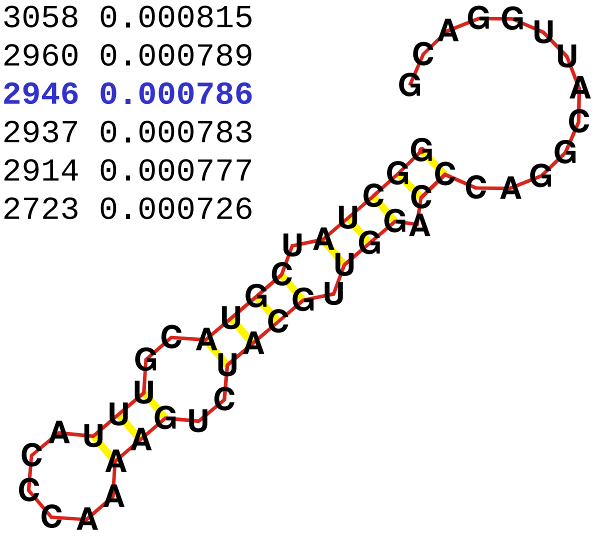


GGAGCUUGCCGAAUGCAACCCCAUGAGGCGCGCUGCCUGGCACCAGCCCC

```
 1 (((((.((((..(((......)))..)))).))).)).(((....)))..       49 0.326667
 2 (((((.((((..(((......)))..)))).))).)).............        7 0.046667
 3 ..(((.((((..(((......)))..)))).)))....(((....)))..        6 0.040000
 4 (((((.((((..((........))..)))).))).)).(((....)))..        5 0.033333
 5 ((.((((((((...(((.((((....)).).).))).)))))..))))).        5 0.033333
 6 (((((.((((...((......))...)))).))).)).(((....)))..        5 0.033333
 7 (((((.((((..(((......)))..)))).))).))..((....))...        4 0.026667
 8 (((((.((((..(((......)))..)))).)))))..(((....)))..        4 0.026667
 9 (((((.(((...(((......)))...))).))).)).(((....)))..        3 0.020000
10 ((((((((((..(((......)))..)))))))))).)).(((....)))..      3 0.020000
11 ((.(((.((((..(((..(.....)..)))..))))..))).))......        3 0.020000
12 (((((...((..(((......)))..))...))).)).(((....)))..        3 0.020000
13 (.(((.((((..(((......)))..)))).))).)..(((....)))..        3 0.020000
14 ((..(.((((..(((......)))..)))).)...)).(((....)))..        3 0.020000
15 (((((.(((((.(((......))).))))).))).)).(((....)))..        3 0.020000
16 (((((.((((.((((......)))).)))).))).)).(((....)))..        3 0.020000
17 (((((..(((..(((......)))..)))..))).)).(((....)))..        3 0.020000
18 ((.((((((((...(((.(.(........).).))).)))))..))))).        2 0.013333
19 (((((.((((..(((......)))..)))).)).))).(((....)))..        2 0.013333
20 ((.((((((((...((((((((....)).).))))).)))))..))))).        2 0.013333
```

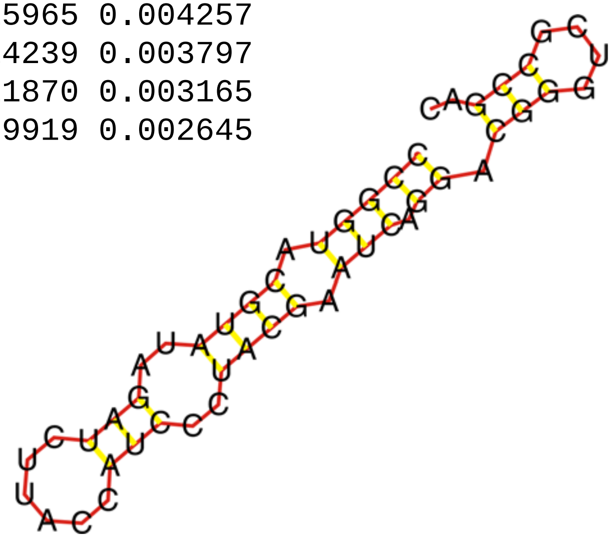


```
                                  Number     Mean Value      Variance      Std.Dev.
Total Hamming Distance:          3750000      11.608372     22.628558      4.756948
Nonzero Hamming Distance:        2493088      16.921998     30.500616      5.522736
Degree of Neutrality:            1256912       0.335177      0.006850      0.082764
Number of Structures:               25000          52.15         84.61          9.20

  1 (((((.((((..(((......)))..)))).))).)).............  1256912 0.335177
  2 ((((((((((..(((......)))..)))))))).)).............    69647 0.018573
  3 ..(((.((((..(((......)))..)))).)))................    69194 0.018452
  4 (((((.((((..((((....))))..)))).))).)).............    61825 0.016487
  5 (((((.((((.((((......)))).)))).))).)).............    56398 0.015039
  6 (((((.(((((.(((......))).))))).))).)).............    55423 0.014779
  7 (((((..(((..(((......)))..)))..))).)).............    34871 0.009299
  8 (((((.((((..((........))..)))).))).)).............    29201 0.007787
  9 ((((..((((..(((......)))..))))..)).)).............    25844 0.006892
 10 (((((.((((..(((......)))..)))).)))))..............    25459 0.006789

 28 (((((.((((..(((......)))..)))).))).))..(((....))).     3629 0.000968
 29 (((((...((..(((......)))..))...))).)).............     3519 0.000938
 30 ...((.((((..(((......)))..)))).)).................     3138 0.000837
 31 (((((.((....(((......)))....)).))).)).............     3067 0.000818
 32 ......((((..(((......)))..))))....................     3058 0.000815
 33 (((((.((((..(((.....)))...)))).))).)).............     2960 0.000789
 34 (((((.((((..(((......)))..)))).))).)).(((....)))..     2946 0.000786
 35 (((((.((((..(((......)))..)))).))).))...(((....)))     2937 0.000783
 36 (((...((((..(((......)))..))))....))).............     2914 0.000777
 37 ..(((.((((..(((......)))..)))).))).(((....))).....     2723 0.000726
```

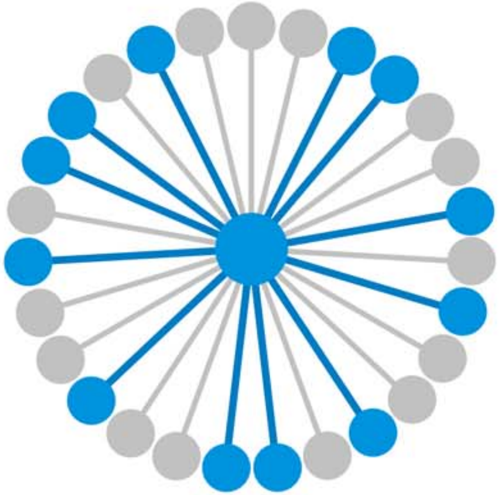


Shadow – Surrounding of RNA structure I in shape space – **AUGC** alphabet

```
                             Number    Mean Value    Variance     Std.Dev.
Total Hamming Distance:      3750000    12.498761   23.352188     4.832410
Nonzero Hamming Distance:    2807992    16.350987   29.476615     5.429237
Degree of Neutrality:         942008     0.251202    0.003690     0.060747
Number of Structures:          25000        54.16       73.46         8.57
   1 (((((.((((..(((......)))..)))).))).)).(((....))).. 942008 0.251202
   2 (((((.((((..(((......)))..)))).))).)).............  166946 0.044519
   3 ..(((.((((..(((......)))..)))).)))....(((....)))..  103673 0.027646
   4 ((((((((((..(((......)))..)))))))).)).(((....)))..   69658 0.018575
   5 (((((.((((..((((....))))..)))).))).)).(((....)))..   62183 0.016582
   6 (((((.((((.((((......)))).)))).))).)).(((....)))..   56510 0.015069
   7 (((((.(((((.(((......))).))))).))).)).(((....)))..   55902 0.014907
   8 (((((..(((..(((......)))..)))..))).)).(((....)))..   35249 0.009400
   9 .((((.((((..(((......)))..)))).))))...(((....)))..   32042 0.008545
  10 (((((.((((..((........))..)))).))).)).(((....)))..   29725 0.007927
  11 (((((.((((..(((......)))..)))).)))))..(((....)))..   27114 0.007230
  12 ((((..((((..(((......)))..))))..)).)).(((....)))..   25820 0.006885
  13 (((((.((((..(((......)))..)))).)).))).(((....)))..   22513 0.006003
  14 (((((.(((...(((......)))...))).))).)).(((....)))..   21640 0.005771
  15 ..(((.((((..(((......)))..)))).)))...((((....)))).   20394 0.005438
  16 ..(((.((((..(((......)))..)))).)))..(((((....)))))   16983 0.004529
  17 (((((.((((...((......))...)))).))).)).(((....)))..   15965 0.004257
  18 (((((.((((..(((......)))..)))).))).))..((....))...   14239 0.003797
  19 (((((.((((..(((......)))..)))).))).)).((......))..   11870 0.003165
  20 (((((.((((..(((......)))..)))).))).))((((....)))).    9919 0.002645
```



Shadow – Surrounding of RNA structure II in shape space – **AUGC** alphabet



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

$$\bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

Alphabet size κ :

κ	λ_{cr}	
2	0.5	**AU,GC,DU**
3	0.423	**AUG , UGC**
4	0.370	**AUGC**

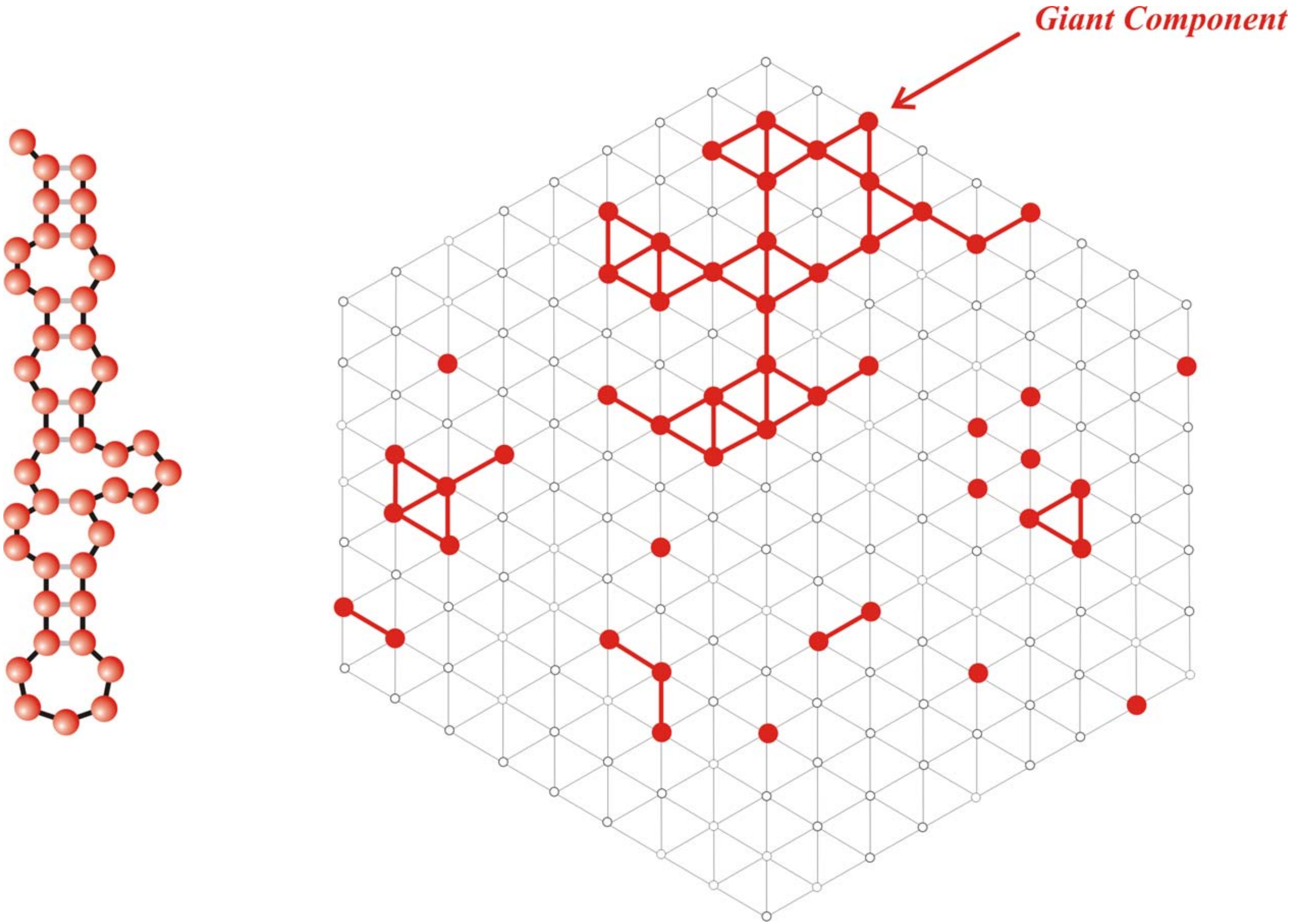
$\lambda_j = \mathbf{12} / 27 = 0.444$

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

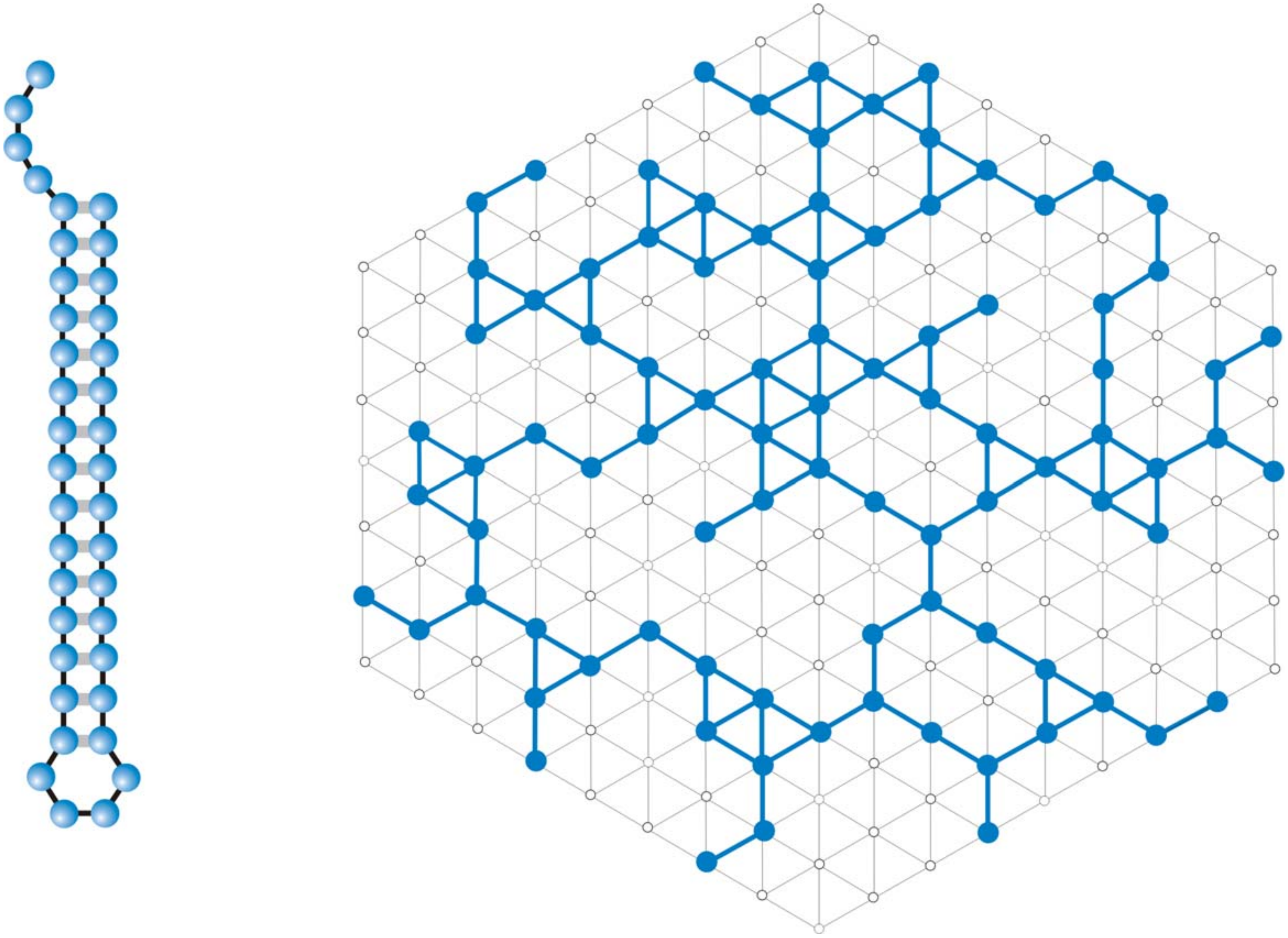
$\bar{\lambda}_k < \lambda_{cr}$ . . . . network G_k is **not** connected

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

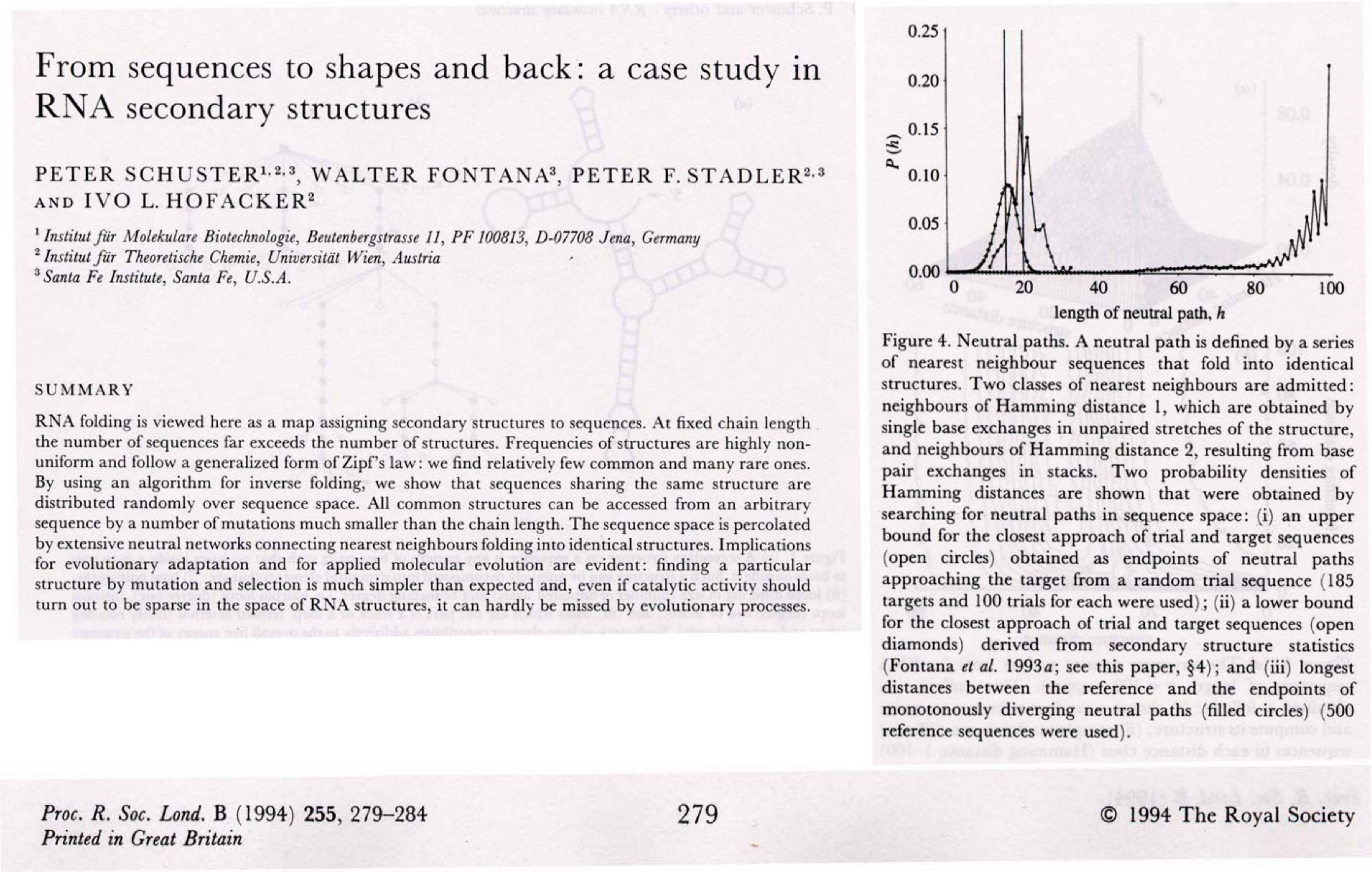
Degree of neutrality of neutral networks and the connectivity threshold



A multi-component neutral network formed by a rare structure: $\lambda < \lambda_{cr}$



A connected neutral network formed by a common structure: $\lambda > \lambda_{cr}$



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

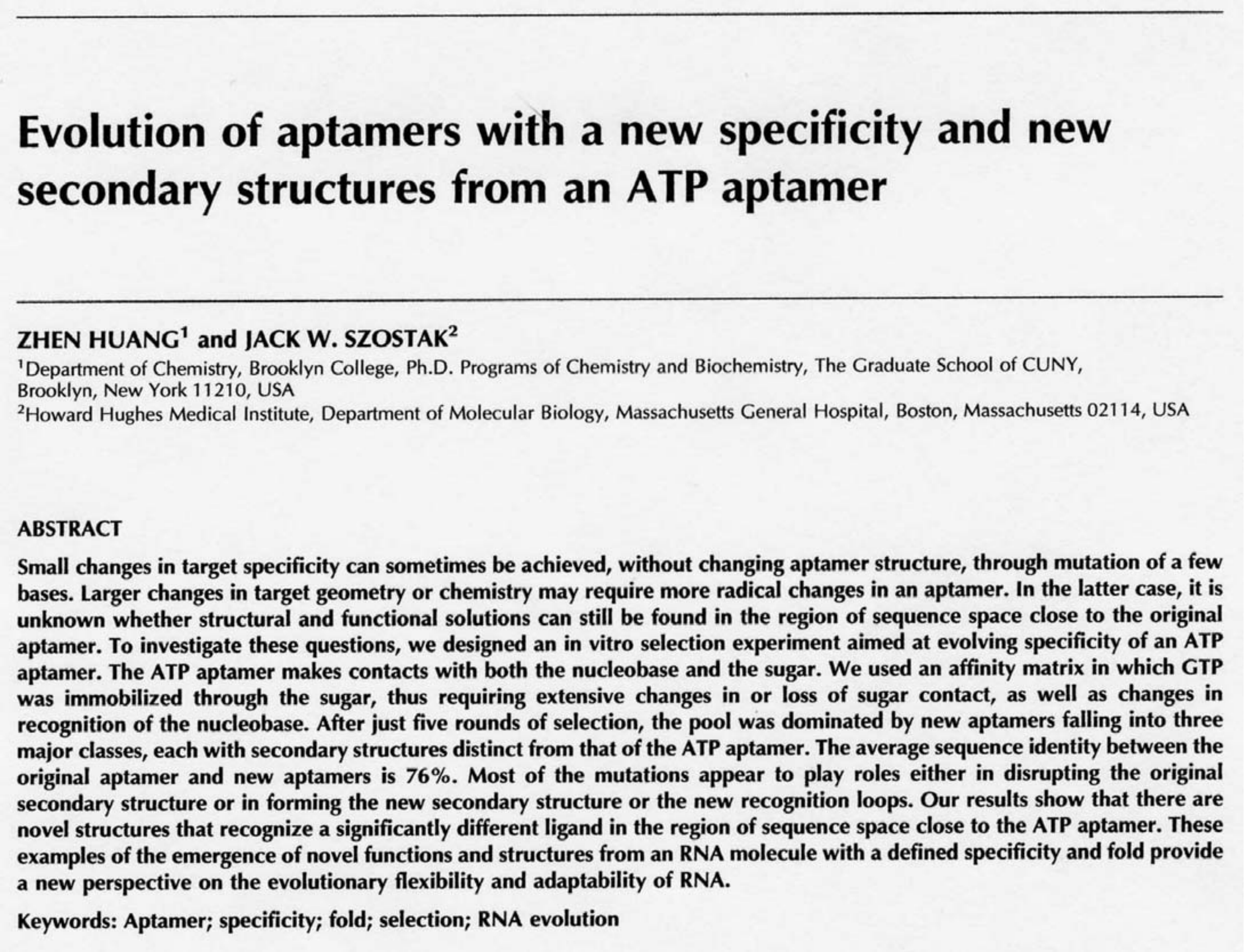
RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare 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

279

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



Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer

ZHEN HUANG[1] **and JACK W. SZOSTAK**[2]

[1]Department of Chemistry, Brooklyn College, Ph.D. Programs of Chemistry and Biochemistry, The Graduate School of CUNY, Brooklyn, New York 11210, USA

[2]Howard Hughes Medical Institute, Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

ABSTRACT

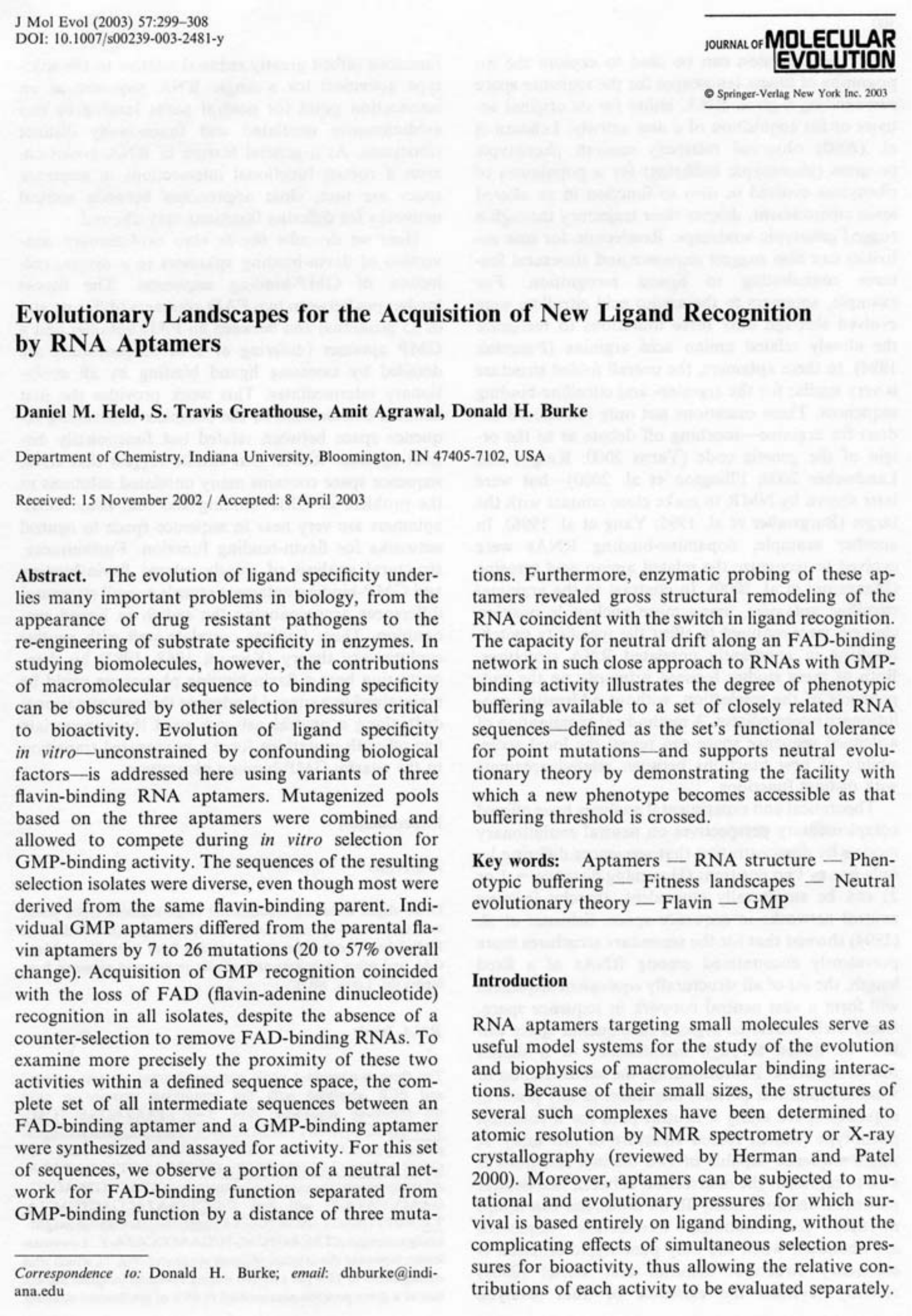
Small changes in target specificity can sometimes be achieved, without changing aptamer structure, through mutation of a few bases. Larger changes in target geometry or chemistry may require more radical changes in an aptamer. In the latter case, it is unknown whether structural and functional solutions can still be found in the region of sequence space close to the original aptamer. To investigate these questions, we designed an in vitro selection experiment aimed at evolving specificity of an ATP aptamer. The ATP aptamer makes contacts with both the nucleobase and the sugar. We used an affinity matrix in which GTP was immobilized through the sugar, thus requiring extensive changes in or loss of sugar contact, as well as changes in recognition of the nucleobase. After just five rounds of selection, the pool was dominated by new aptamers falling into three major classes, each with secondary structures distinct from that of the ATP aptamer. The average sequence identity between the original aptamer and new aptamers is 76%. Most of the mutations appear to play roles either in disrupting the original secondary structure or in forming the new secondary structure or the new recognition loops. Our results show that there are novel structures that recognize a significantly different ligand in the region of sequence space close to the ATP aptamer. These examples of the emergence of novel functions and structures from an RNA molecule with a defined specificity and fold provide a new perspective on the evolutionary flexibility and adaptability of RNA.

Keywords: Aptamer; specificity; fold; selection; RNA evolution

RNA **9**:1456-1463, 2003

Evidence for neutral networks and shape space covering

Evidence for neutral networks and intersection of apatamer functions



J Mol Evol (2003) 57:299–308
DOI: 10.1007/s00239-003-2481-y

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Landscapes for the Acquisition of New Ligand Recognition by RNA Aptamers

Daniel M. Held, S. Travis Greathouse, Amit Agrawal, Donald H. Burke

Department of Chemistry, Indiana University, Bloomington, IN 47405-7102, USA

Received: 15 November 2002 / Accepted: 8 April 2003

Abstract. The evolution of ligand specificity underlies many important problems in biology, from the appearance of drug resistant pathogens to the re-engineering of substrate specificity in enzymes. In studying biomolecules, however, the contributions of macromolecular sequence to binding specificity can be obscured by other selection pressures critical to bioactivity. Evolution of ligand specificity *in vitro*—unconstrained by confounding biological factors—is addressed here using variants of three flavin-binding RNA aptamers. Mutagenized pools based on the three aptamers were combined and allowed to compete during *in vitro* selection for GMP-binding activity. The sequences of the resulting selection isolates were diverse, even though most were derived from the same flavin-binding parent. Individual GMP aptamers differed from the parental flavin aptamers by 7 to 26 mutations (20 to 57% overall change). Acquisition of GMP recognition coincided with the loss of FAD (flavin-adenine dinucleotide) recognition in all isolates, despite the absence of a counter-selection to remove FAD-binding RNAs. To examine more precisely the proximity of these two activities within a defined sequence space, the complete set of all intermediate sequences between an FAD-binding aptamer and a GMP-binding aptamer were synthesized and assayed for activity. For this set of sequences, we observe a portion of a neutral network for FAD-binding function separated from GMP-binding function by a distance of three mutations. Furthermore, enzymatic probing of these aptamers revealed gross structural remodeling of the RNA coincident with the switch in ligand recognition. The capacity for neutral drift along an FAD-binding network in such close approach to RNAs with GMP-binding activity illustrates the degree of phenotypic buffering available to a set of closely related RNA sequences—defined as the set's functional tolerance for point mutations—and supports neutral evolutionary theory by demonstrating the facility with which a new phenotype becomes accessible as that buffering threshold is crossed.

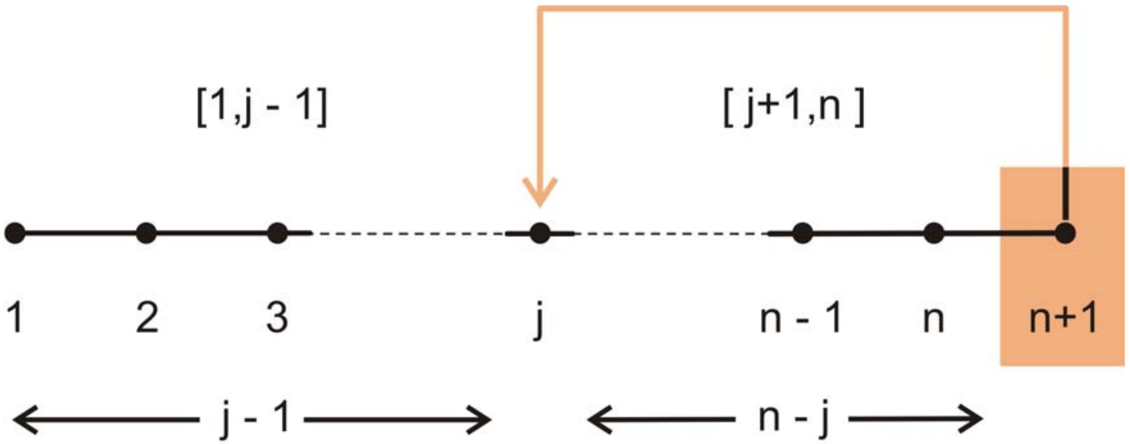
Key words: Aptamers — RNA structure — Phenotypic buffering — Fitness landscapes — Neutral evolutionary theory — Flavin — GMP

Introduction

RNA aptamers targeting small molecules serve as useful model systems for the study of the evolution and biophysics of macromolecular binding interactions. Because of their small sizes, the structures of several such complexes have been determined to atomic resolution by NMR spectrometry or X-ray crystallography (reviewed by Herman and Patel 2000). Moreover, aptamers can be subjected to mutational and evolutionary pressures for which survival is based entirely on ligand binding, without the complicating effects of simultaneous selection pressures for bioactivity, thus allowing the relative contributions of each activity to be evaluated separately.

Correspondence to: Donald H. Burke; *email:* dhburke@indiana.edu

1. The RNA model
2. **How many stable structures can be formed?**
3. Why not binary (GC or AU) sequences?
4. Evolution on neutral networks
5. Multiconformational RNA molecules



$$S_{n+1} = S_n + \sum_{j=1}^{n-1} S_{j-1} \cdot S_{n-j}$$

Counting the numbers of structures of chain length $n \Rightarrow n+1$

M.S. Waterman, T.F. Smith (1978) *Math.Bioscience* **42**:257-266

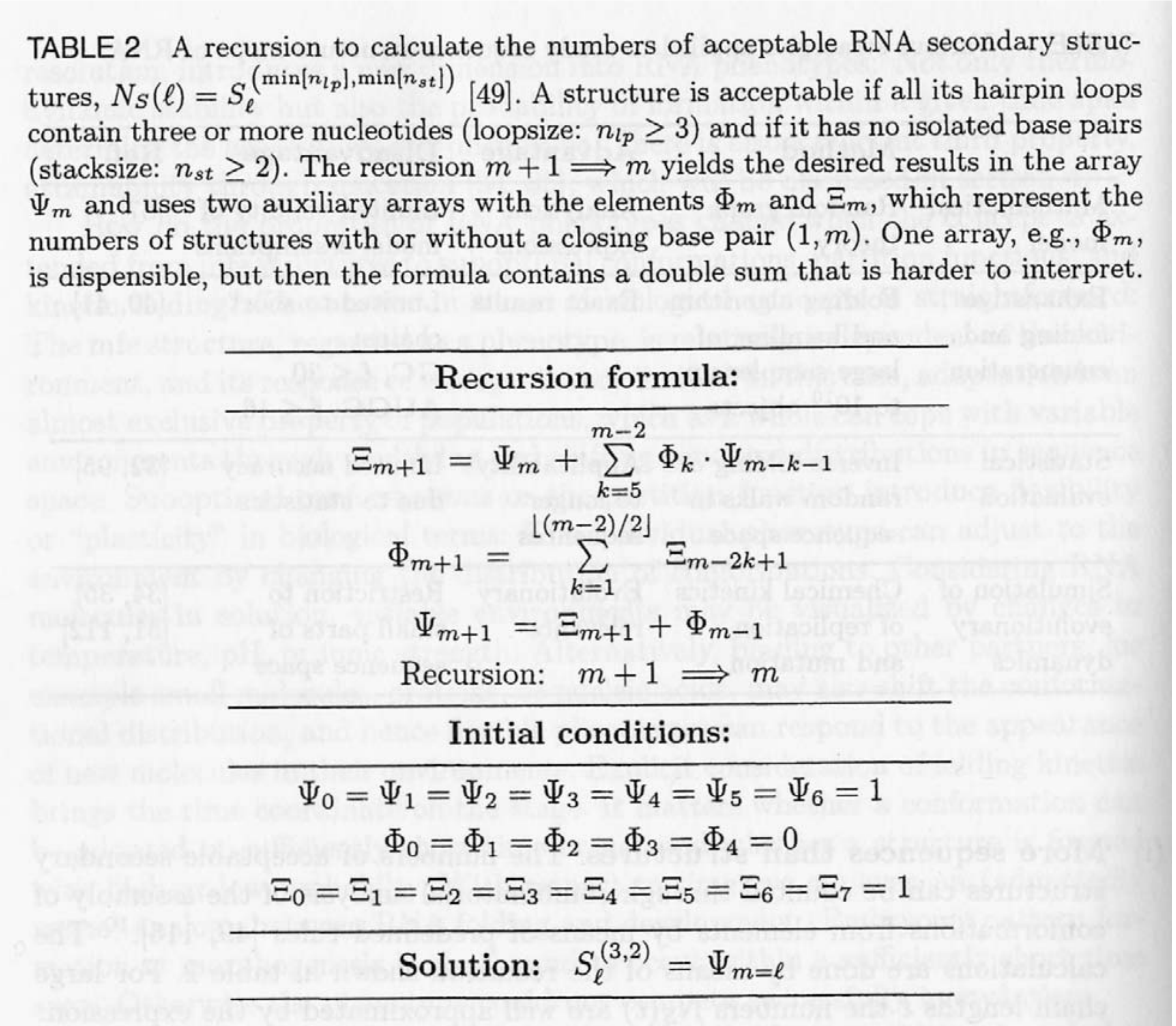


TABLE 2 A recursion to calculate the numbers of acceptable RNA secondary structures, $N_S(\ell) = S_\ell^{(\min[n_{lp}],\min[n_{st}])}$ [49]. A structure is acceptable if all its hairpin loops contain three or more nucleotides (loopsize: $n_{lp} \geq 3$) and if it has no isolated base pairs (stacksize: $n_{st} \geq 2$). 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:

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

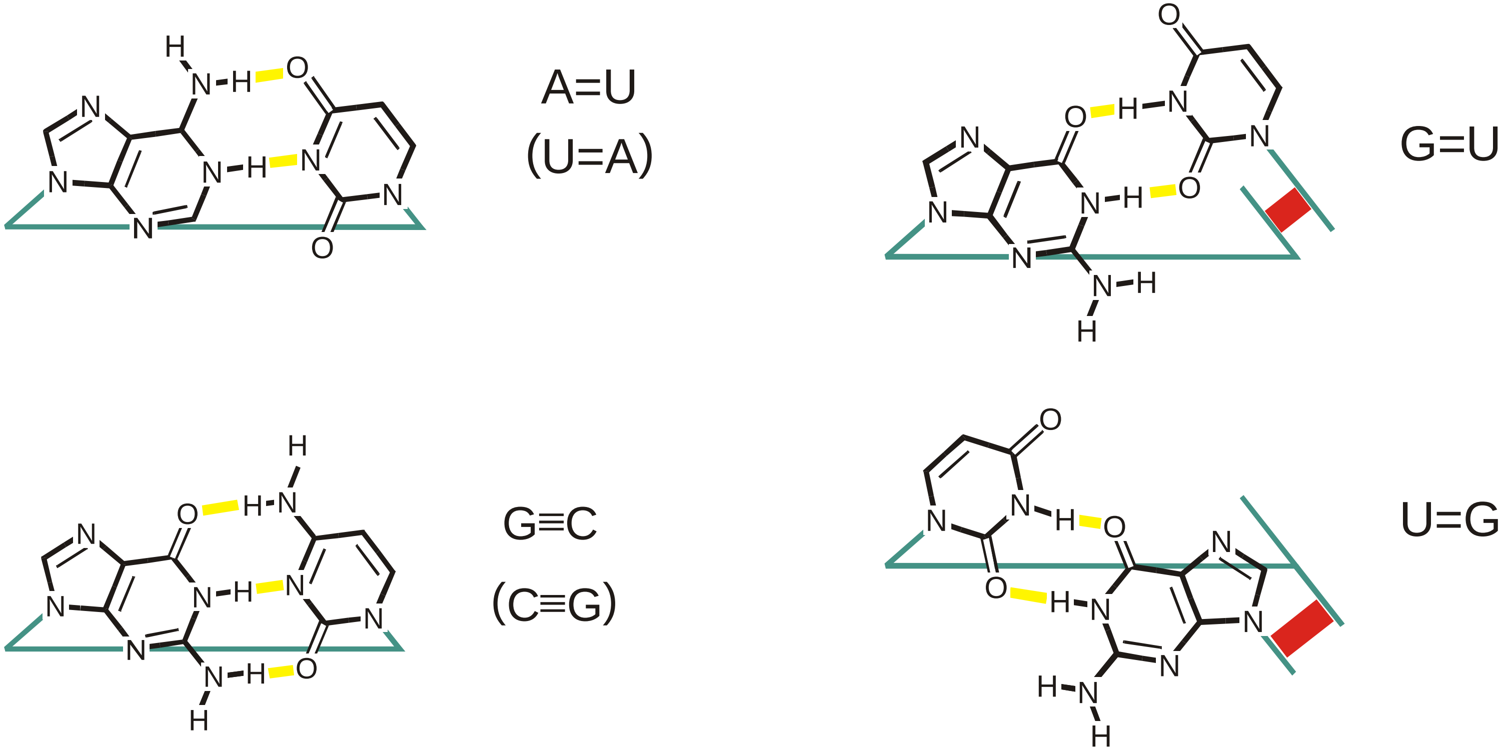
$$\Phi_0 = \Phi_1 = \Phi_2 = \Phi_3 = \Phi_4 = 0$$

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

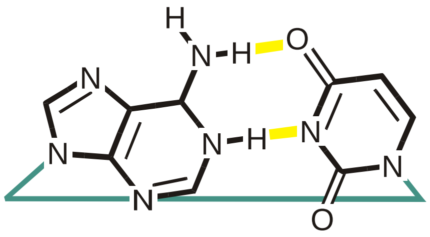
Solution: $S_\ell^{(3,2)} = \Psi_{m=\ell}$

Recursion formula for the number of physically acceptable stable structures

I.L.Hofacker, P.Schuster, P.F. Stadler (1998) *Discr.Appl.Math.* **89**:177-207

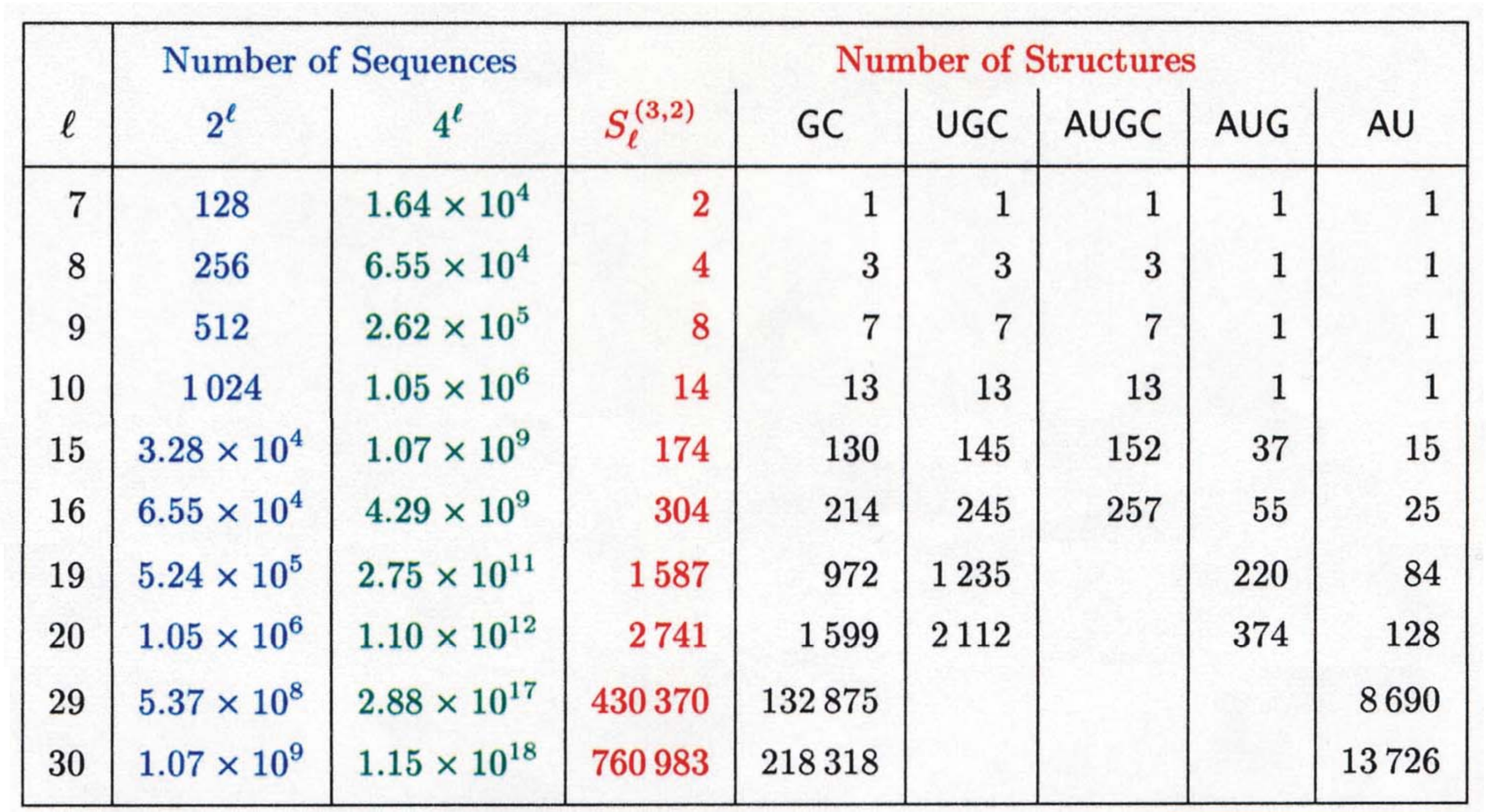


The six base pairing alphabets built from natural nucleotides **A**, **U**, **G**, and **C**



A=U
(U=A)

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

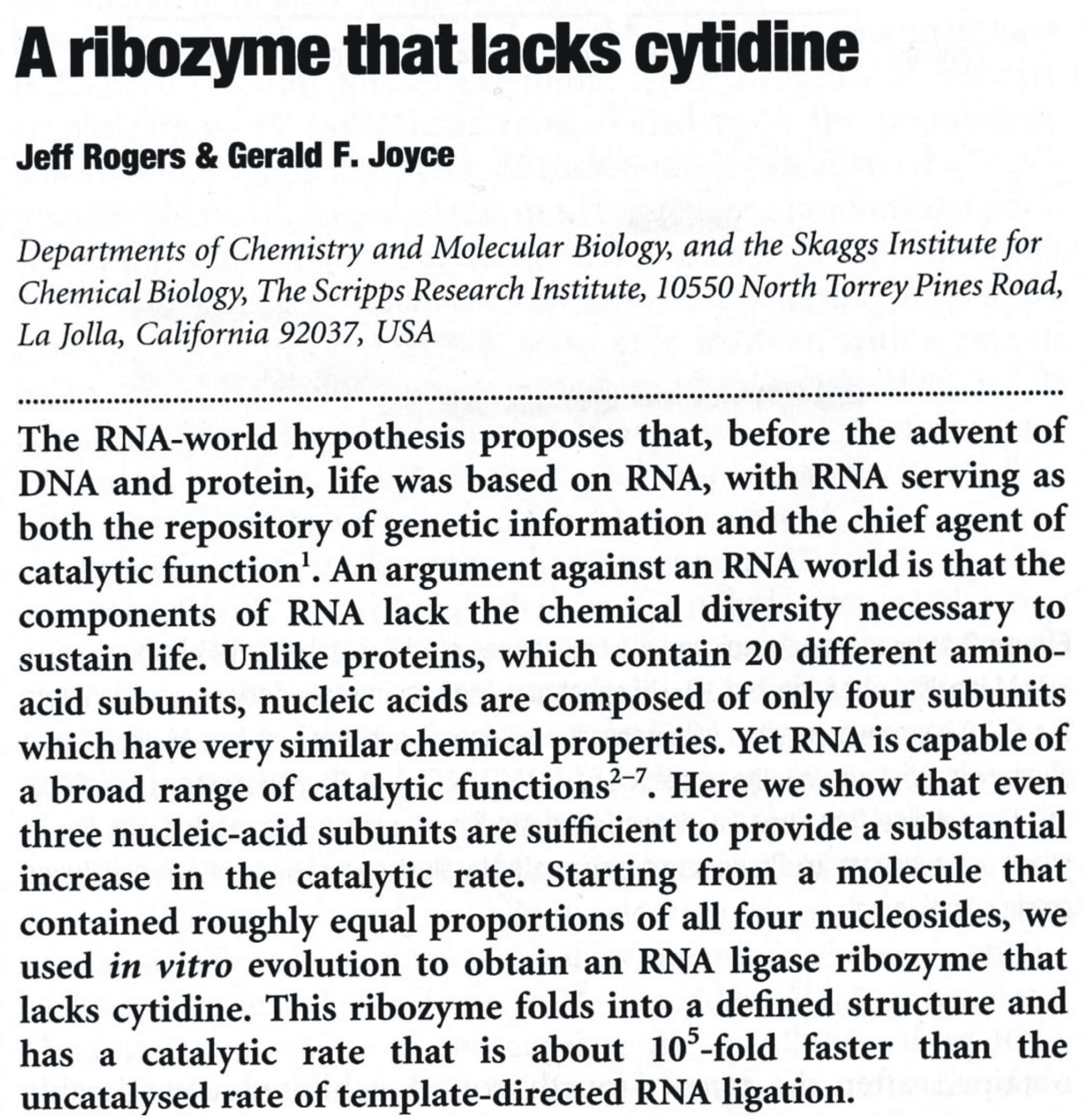


	Number of Sequences		Number of Structures					
ℓ	2^ℓ	4^ℓ	$S_\ell^{(3,2)}$	GC	UGC	AUGC	AUG	AU
7	128	1.64×10^4	2	1	1	1	1	1
8	256	6.55×10^4	4	3	3	3	1	1
9	512	2.62×10^5	8	7	7	7	1	1
10	1 024	1.05×10^6	14	13	13	13	1	1
15	3.28×10^4	1.07×10^9	174	130	145	152	37	15
16	6.55×10^4	4.29×10^9	304	214	245	257	55	25
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.

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



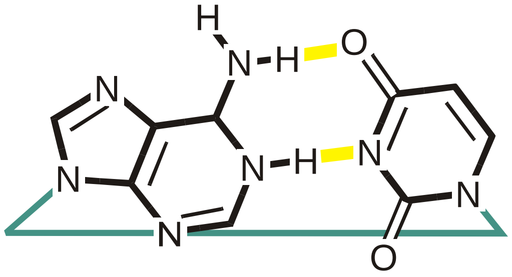
A ribozyme that lacks cytidine

Jeff Rogers & Gerald F. Joyce

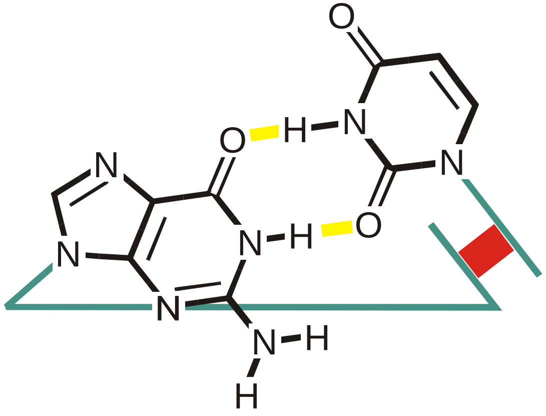
Departments of Chemistry and Molecular Biology, and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

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.

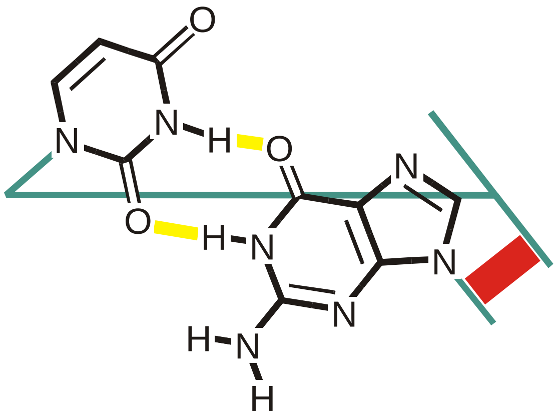
Catalytic activity in the **AUG** alphabet



A=U
(U=A)



G=U



U=G

Base pairs in the **AUG** alphabet

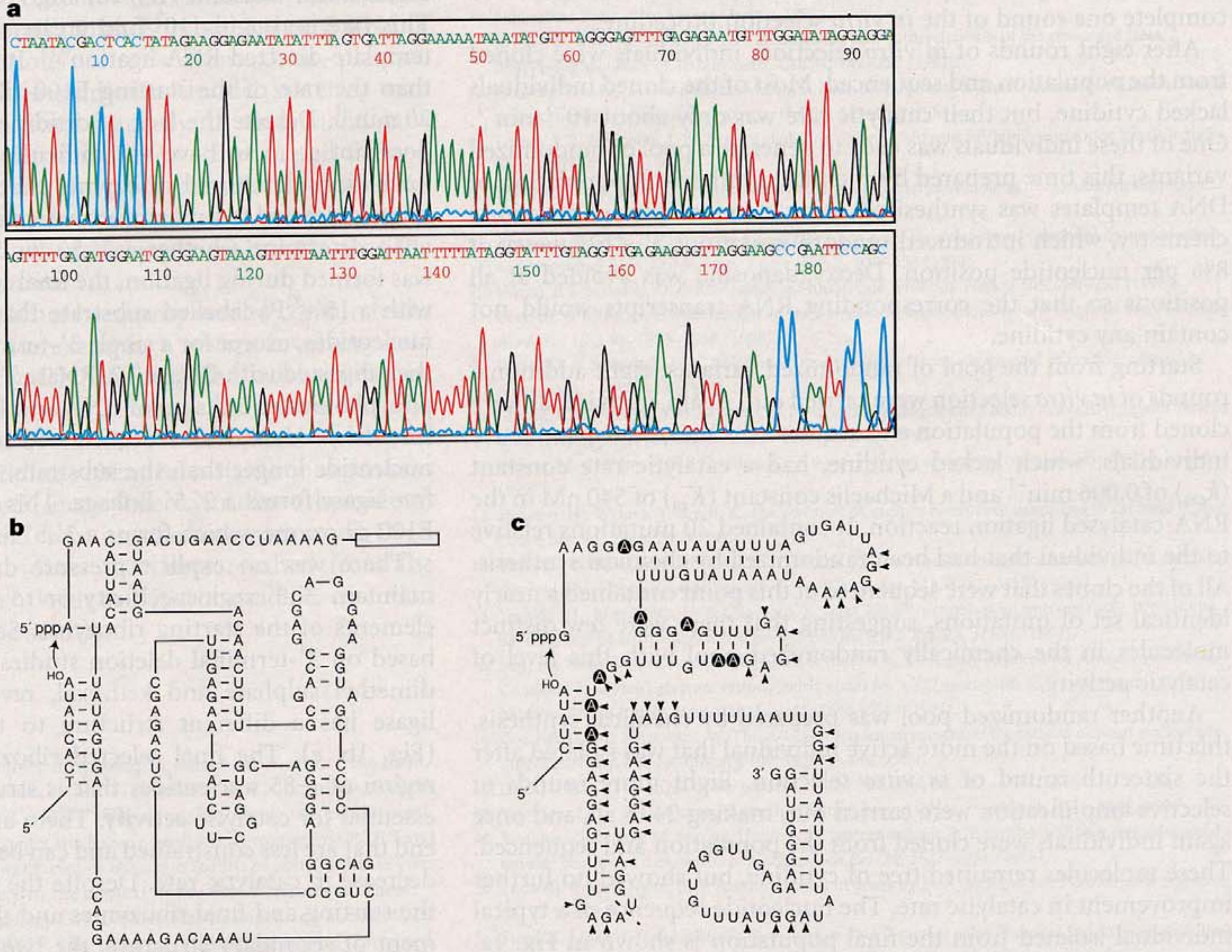
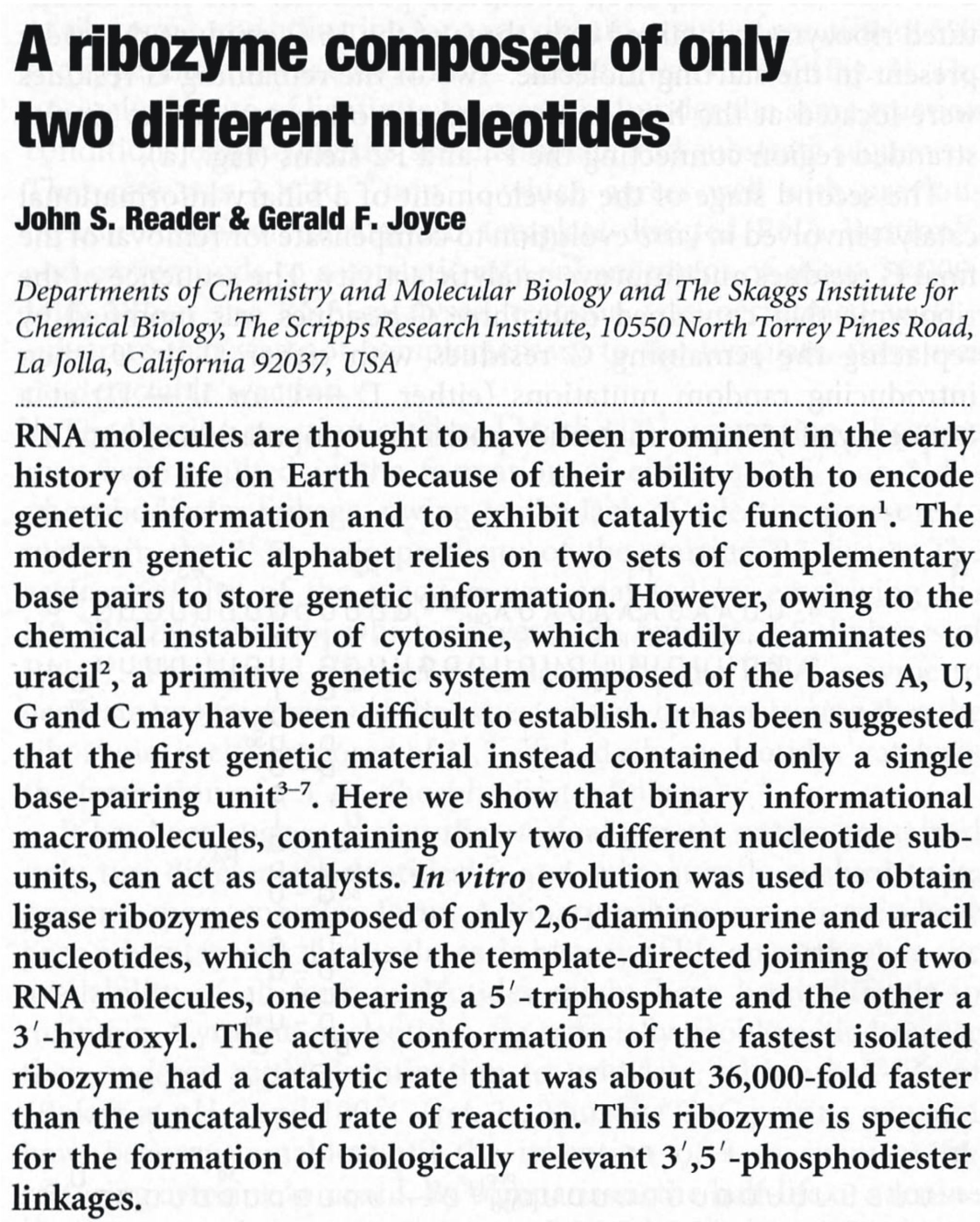

a
CTAATACGACTCACTATAGAAGGAGAATATATTTAGTGATTAGGAAAAATAAATATGTTTAGGGAGTTTGAGAGAATGTTTGGATATAGGAGGA
10 20 30 40 50 60 70 80 90
AGTTTTGAGATGGAATGAGGAAGTAAAGTTTTTAATTTGGATTAATTTTTATAGGTATTTGTAGGTTGAGAAGGGTTAGGAAGCCGAATTCCAGC
100 110 120 130 140 150 160 170 180
b
GAAC–GACUGAACCUAAAAG
5′ pppA
HO
5′
GAAAAGACAAAU
GGCAG
CCGUC
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



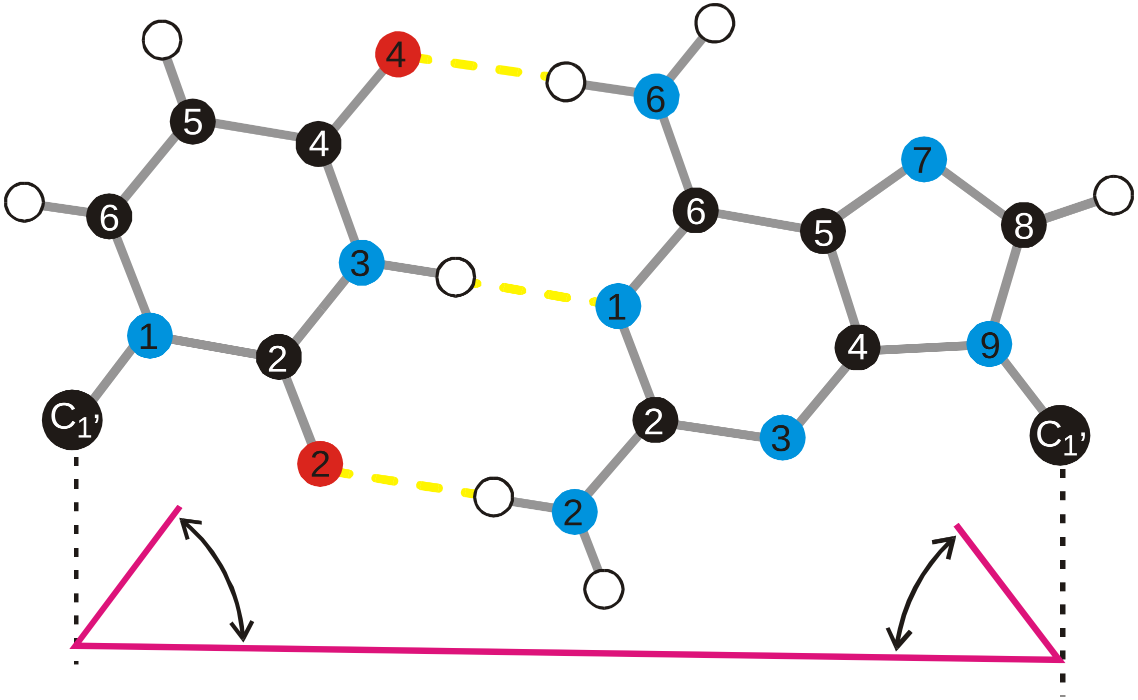
A ribozyme composed of only two different nucleotides

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

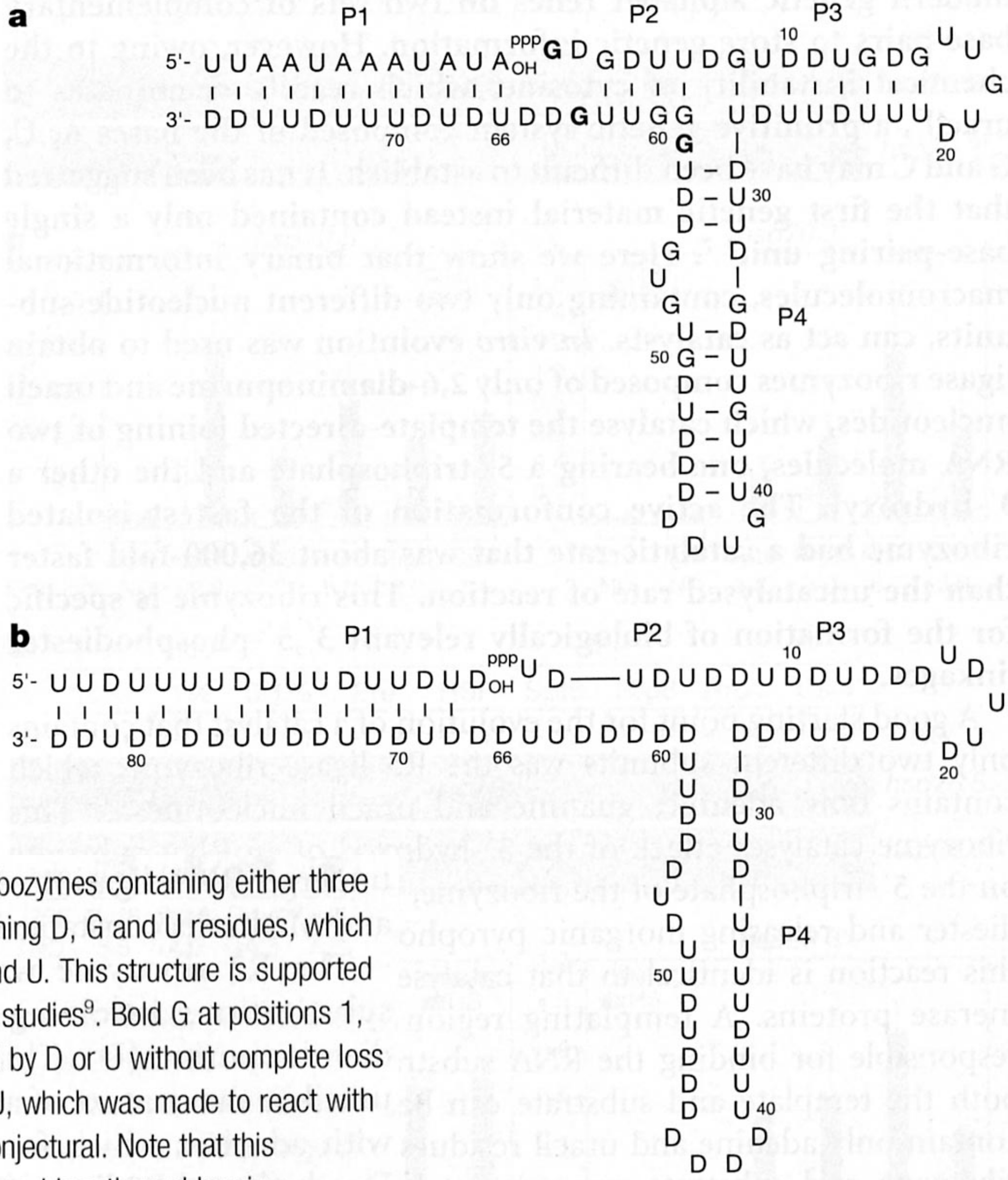

a
P1
P2
P3
P4
5'- U U A A U A A A U A U A
OH
ppp G D G D U U D G U D D U G D G U U G
3'- D D U U D U U U D U D U D D G U U G G
70
66
60
10
20
30
40
50
b
P1
P2
P3
P4
5'- U U D U U U U D D U U D U U D U D
OH
ppp U D U D U D D D U D D U D D D U D U
3'- D D U D D D D U U D D U D D U D D D U U D D D D
80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

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**.

1. The RNA model

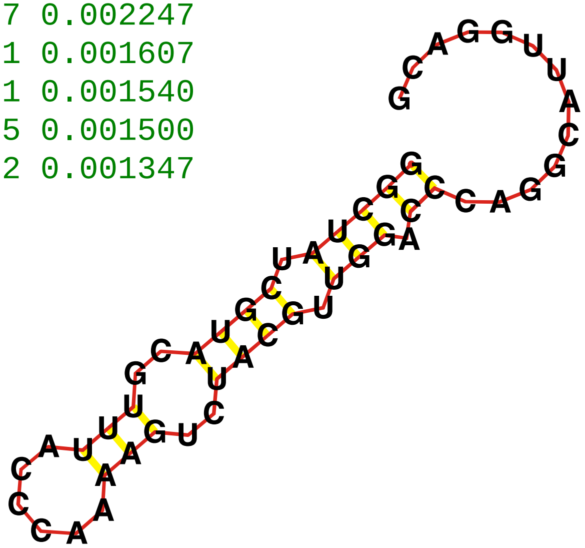
2. How many stable structures can be formed?

3. **Why not binary (GC or AU) sequences?**

4. Evolution on neutral networks

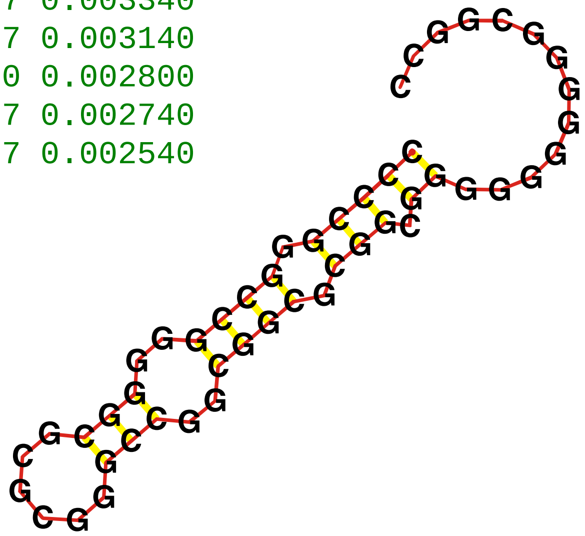
5. Multiconformational RNA molecules

```
                              Number    Mean Value     Variance     Std.Dev.
Total Hamming Distance:       150000     11.647973    23.140715     4.810480
Nonzero Hamming Distance:      99875     16.949991    30.757651     5.545958
Degree of Neutrality:          50125      0.334167     0.006961     0.083434
Number of Structures:            1000         52.31        85.30         9.24

   1 (((((.((((..(((......)))..)))).))).))..............     50125 0.334167
   2 ..(((.((((..(((......)))..)))).))).................      2856 0.019040
   3 ((((((((((..(((......)))..)))))))).))..............      2799 0.018660
   4 (((((.((((..((((....))))..)))).))).))..............      2417 0.016113
   5 (((((.((((.((((......)))).)))).))).))..............      2265 0.015100
   6 (((((.(((((.(((......))).))))).))).))..............      2233 0.014887
   7 (((((..(((..(((......)))..)))..))).))..............      1442 0.009613
   8 (((((.((((..((........))..)))).))).))..............      1081 0.007207
   9 ((((..((((..(((......)))..))))..)).))..............      1025 0.006833
  10 (((((.((((..(((......)))..)))).)))))...............      1003 0.006687
  11 .((((.((((..(((......)))..)))).))))................       963 0.006420
  12 (((((.(((...(((......)))...))).))).))..............       860 0.005733
  13 (((((.((((..(((......)))..)))).)).)))..............       800 0.005333
  14 (((((.((((...((......))...)))).))).))..............       548 0.003653
  15 (((((.((((................)))).))).))..............       362 0.002413
  16 ((.((.((((..(((......)))..)))).))..))..............       337 0.002247
  17 (.(((.((((..(((......)))..)))).))).)...............       241 0.001607
  18 (((((.(((((((((......))))))))).))).))..............       231 0.001540
  19 ((((..((((..(((......)))..))))...))))..............       225 0.001500
  20 ((....((((..(((......)))..)))).....))..............       202 0.001347
```

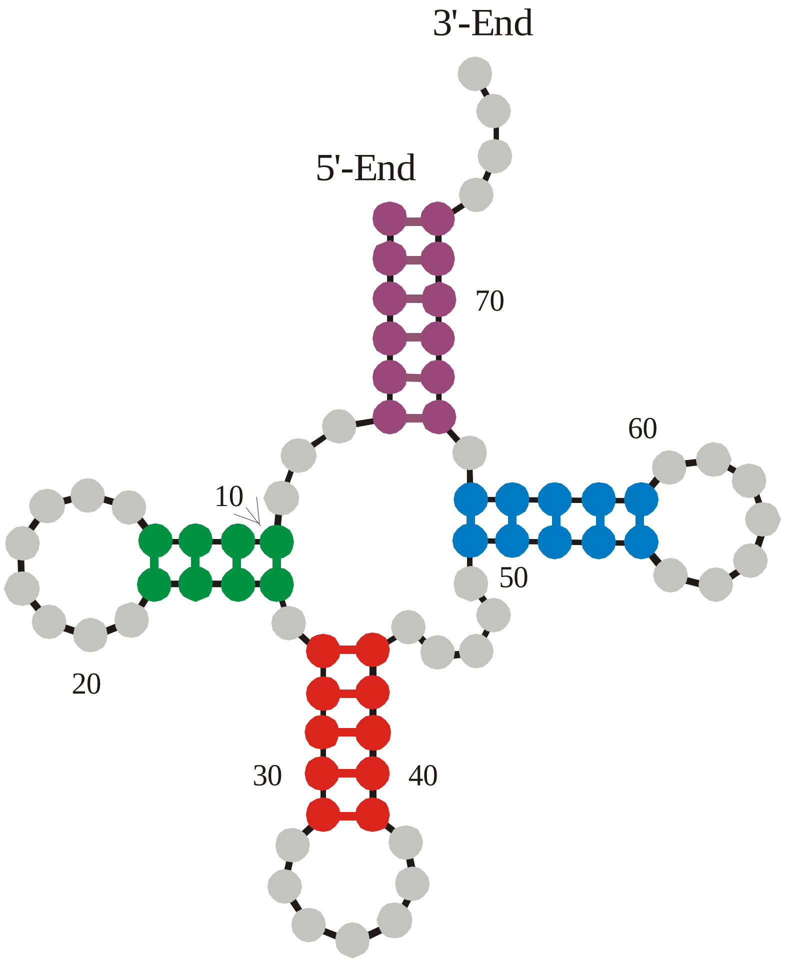


Shadow – Surrounding of an RNA structure in shape space – **AUGC** alphabet

```
                             Number    Mean Value     Variance      Std.Dev.
Total Hamming Distance:       50000     13.673580    10.795762      3.285691
Nonzero Hamming Distance:     45738     14.872054    10.821236      3.289565
Degree of Neutrality:          4262      0.085240     0.001824      0.042708
Number of Structures:          1000         36.24         6.27          2.50

  1 (((((.((((..(((......)))..)))).))).)).............     4262 0.085240
  2 ((((((((((..(((......)))..)))))))).)).............     1940 0.038800
  3 (((((.(((((.(((......))).))))).))).)).............     1791 0.035820
  4 (((((.((((.((((......)))).)))).))).)).............     1752 0.035040
  5 (((((.((((..((((....))))..)))).))).)).............     1423 0.028460
  6 (.(((.((((..(((......)))..)))).))).)..............      665 0.013300
  7 (((((.((((..((........))..)))).))).)).............      308 0.006160
  8 (((((.((((..(((......)))..)))).)))))..............      280 0.005600
  9 (((((.((((..(((......)))..)))).))).))...(((....)))      278 0.005560
 10 (((((.(((...(((......)))...))).))).)).............      209 0.004180
 11 (((((.((((..(((......)))..)))).))).)).(((......)))      193 0.003860
 12 (((((.((((..(((......)))..)))).))).))..(((.....)))      180 0.003600
 13 (((((.((((..((((.....)))).)))).))).)).............      180 0.003600
 14 ..(((.((((..(((......)))..)))).)))................      176 0.003520
 15 (((((.((((.((((.....))))..)))).))).)).............      175 0.003500
 16 (((((.((((..(((......)))..)))))))))...............      167 0.003340
 17 (((((.((((...((......))...)))).))).)).............      157 0.003140
 18 (((((.(.((..(((......)))..)).).))).)).............      140 0.002800
 19 (((((..(((..(((......)))..)))..))).)).............      137 0.002740
 20 .((((.((((..(((......)))..)))).))))...............      127 0.002540
```

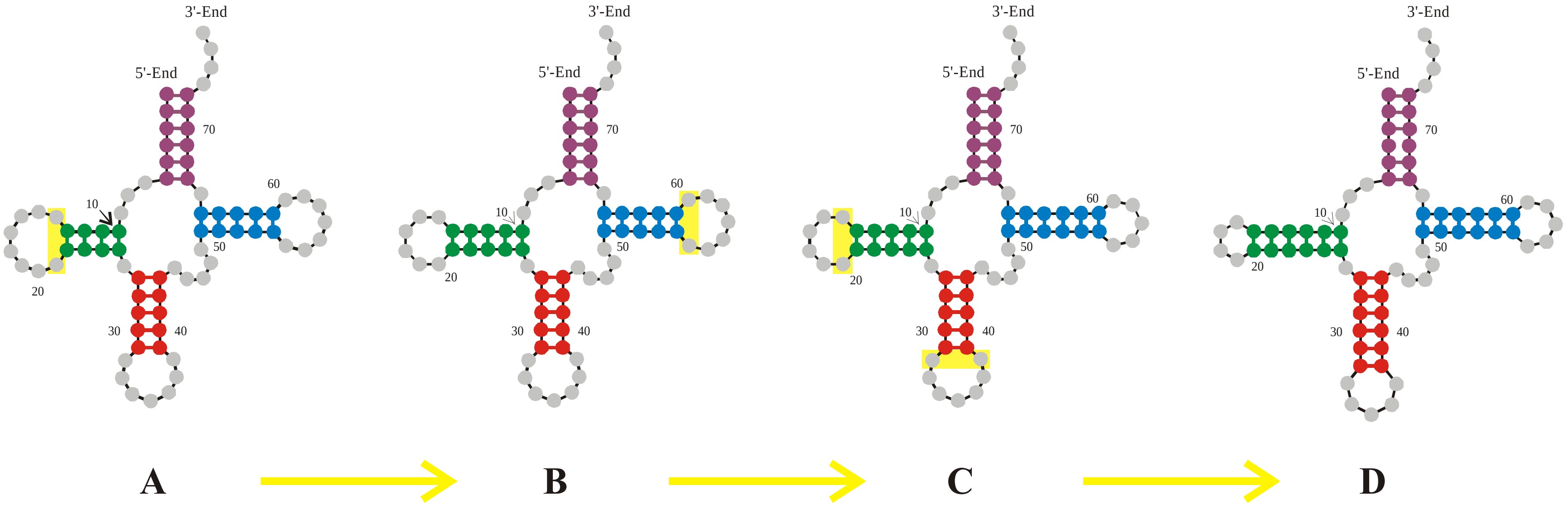


Shadow – Surrounding of an RNA structure in shape space – **GC** alphabet

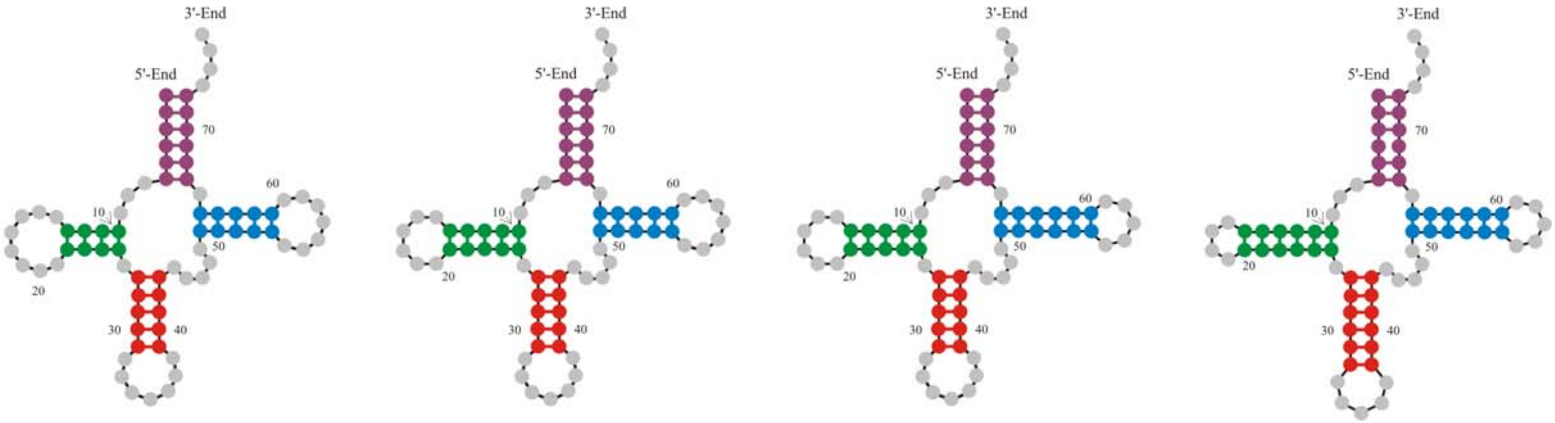


tRNA^{phe}

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

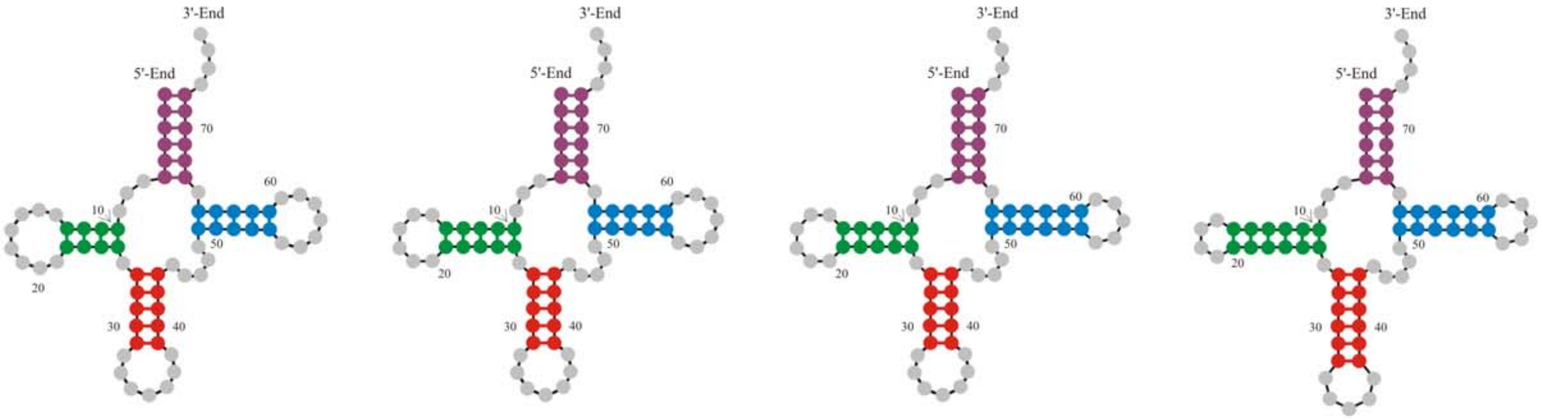


RNA clover-leaf secondary structures of sequences with chain length n=76



Alphabet	Probability of successful trials in inverse folding			
AU	- -	- -	- -	0.051 ± 0.006
AUG	- -	0.003 ± 0.001	0.026 ± 0.006	0.374 ± 0.016
AUGC	0.794 ± 0.007	0.884 ± 0.008	0.934 ± 0.009	0.982 ± 0.004
UGC	0.548 ± 0.011	0.628 ± 0.012	0.697 ± 0.020	0.818 ± 0.012
GC	0.067 ± 0.007	0.086 ± 0.008	0.087 ± 0.008	0.127 ± 0.006

Probability of finding cloverleaf RNA secondary structures from different alphabets



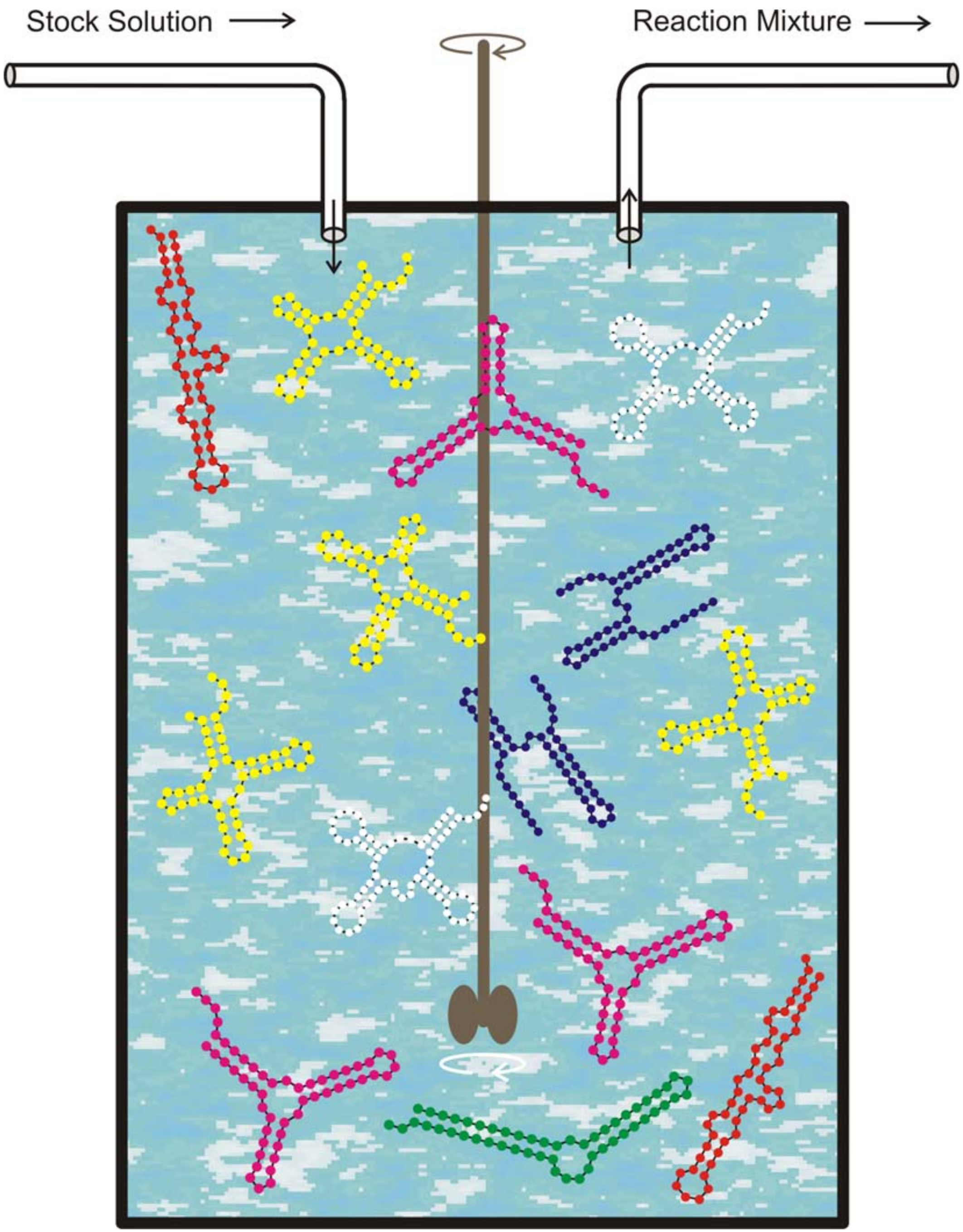
Alphabet	Degree of neutrality $\bar{\lambda}$			
AU	- -	- -	- -	0.073 ± 0.032
AUG	- -	0.217 ± 0.051	0.207 ± 0.055	0.201 ± 0.056
AUGC	0.275 ± 0.064	0.279 ± 0.063	0.289 ± 0.062	0.313 ± 0.058
UGC	0.263 ± 0.071	0.257 ± 0.070	0.251 ± 0.068	0.250 ± 0.064
GC	0.052 ± 0.033	0.057 ± 0.034	0.060 ± 0.033	0.068 ± 0.034

Degree of neutrality of cloverleaf RNA secondary structures over different alphabets

Probability to be able to form a base pair between two arbitrarily chosen nucelotides in a random sequence with uniform base composition

AU,GC	0.5
AUGC	0.375
GCXK	0.25
AUGCXK	0.167

1. The RNA model
2. How many stable structures can be formed?
3. Why not binary (GC or AU) sequences?
4. **Evolution on neutral networks**
5. Multiconformational RNA molecules



Replication rate constant:

$$f_k = \gamma \,/\, [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

Selection constraint:

Population size, N = # RNA molecules, is controlled by the flow

$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

Mutation rate:

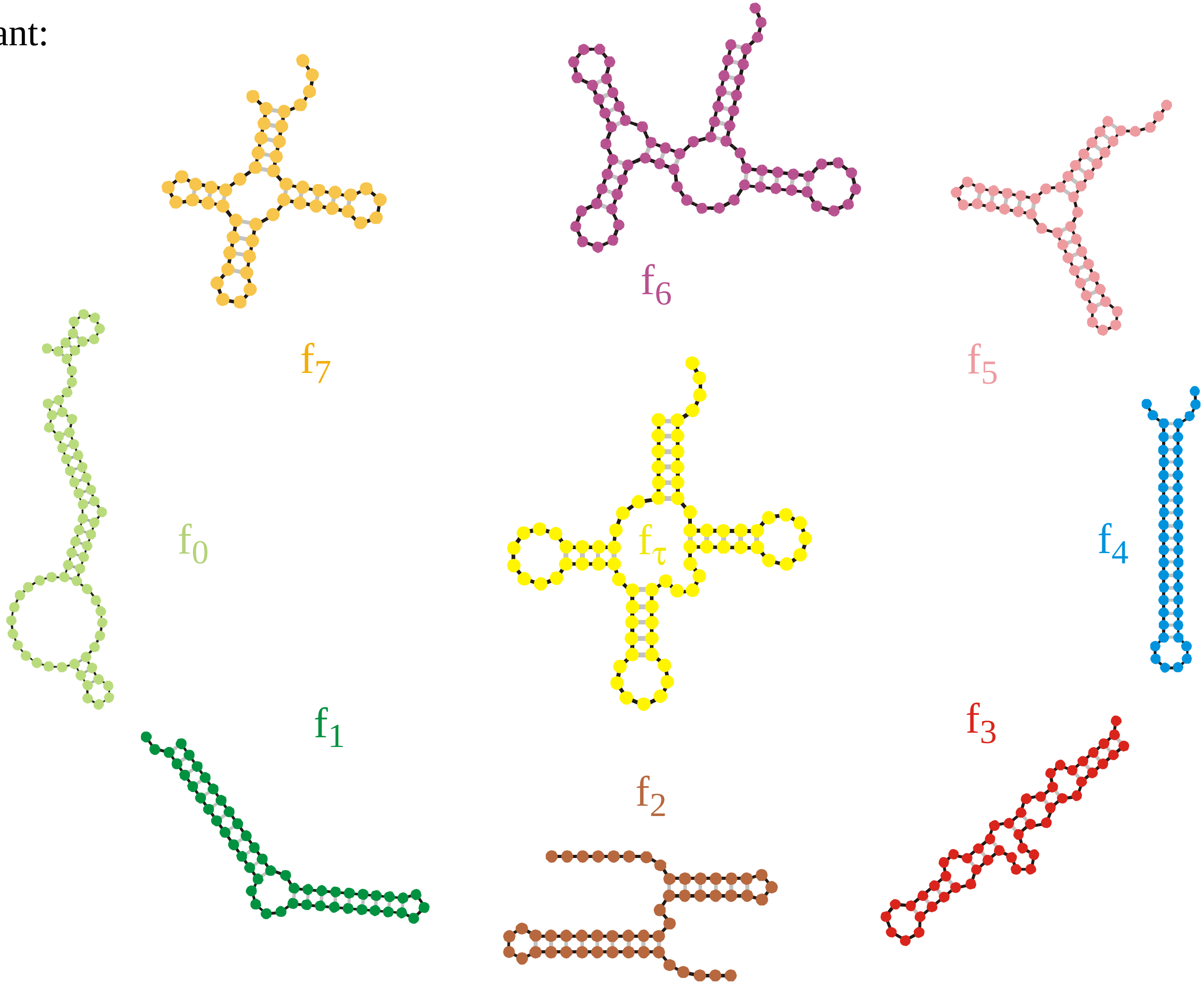
p = 0.001 / site × replication

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

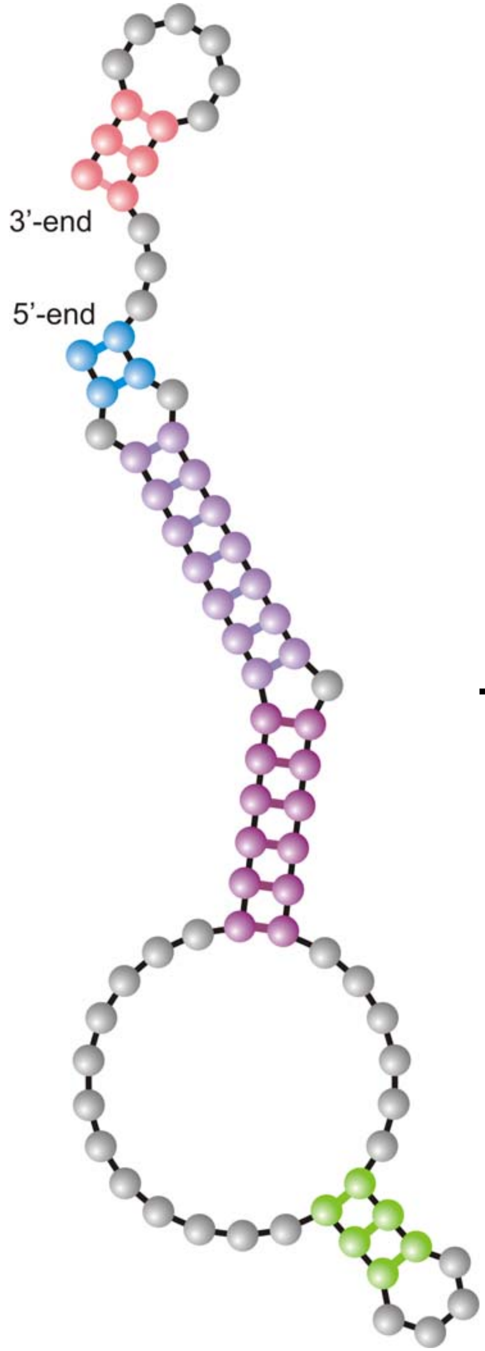
Replication rate constant:

$$f_k = \gamma \, / \, [\alpha + \Delta d_S{}^{(k)}]$$

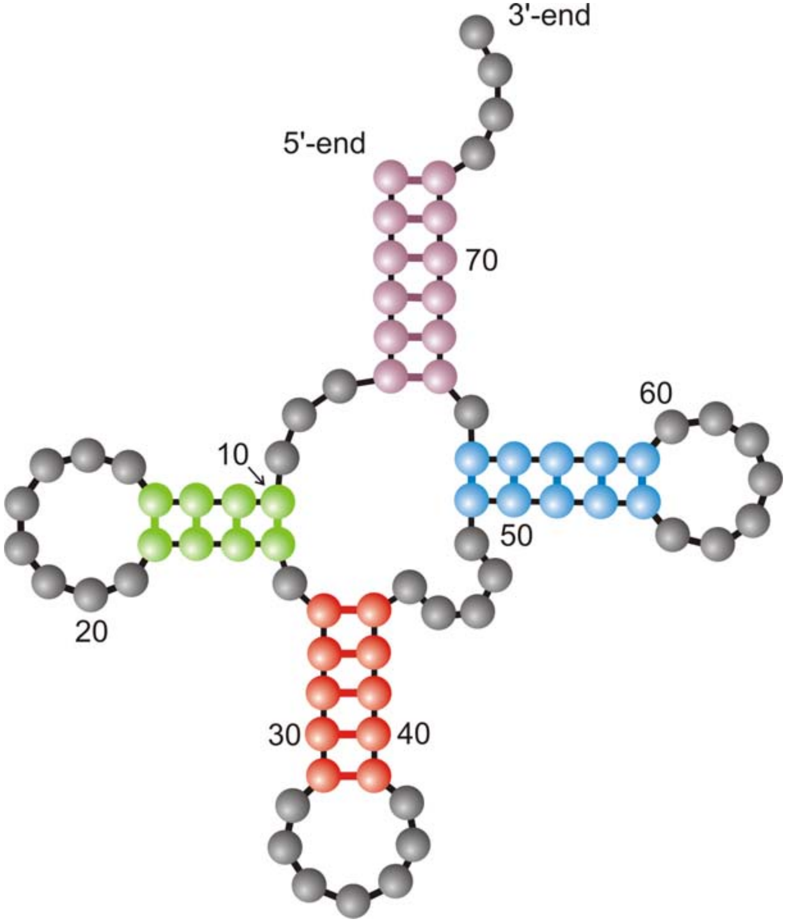
$$\Delta d_S{}^{(k)} = d_H(S_k, S_\tau)$$



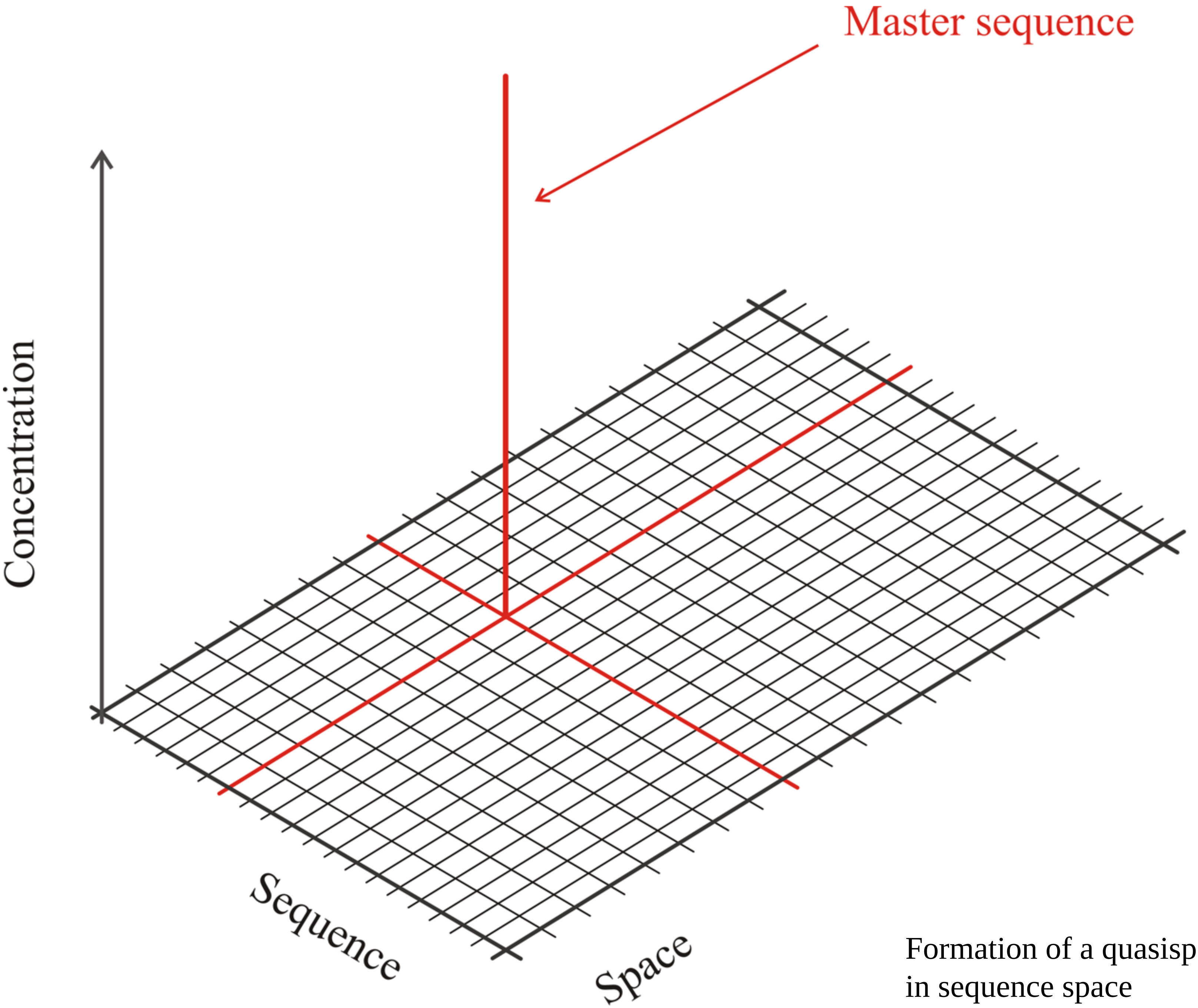
Evaluation of RNA secondary structures yields replication rate constants



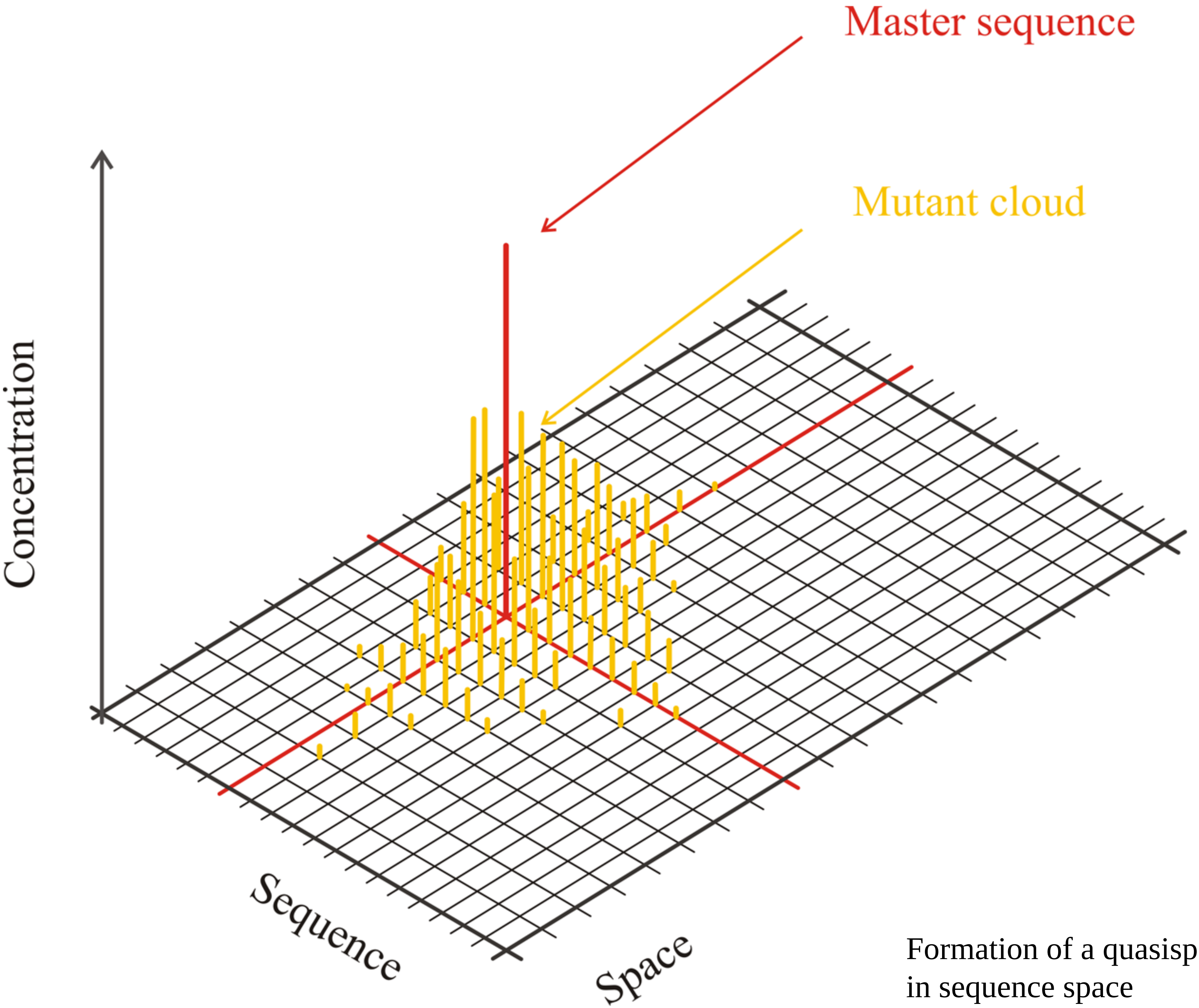
Randomly chosen
initial structure



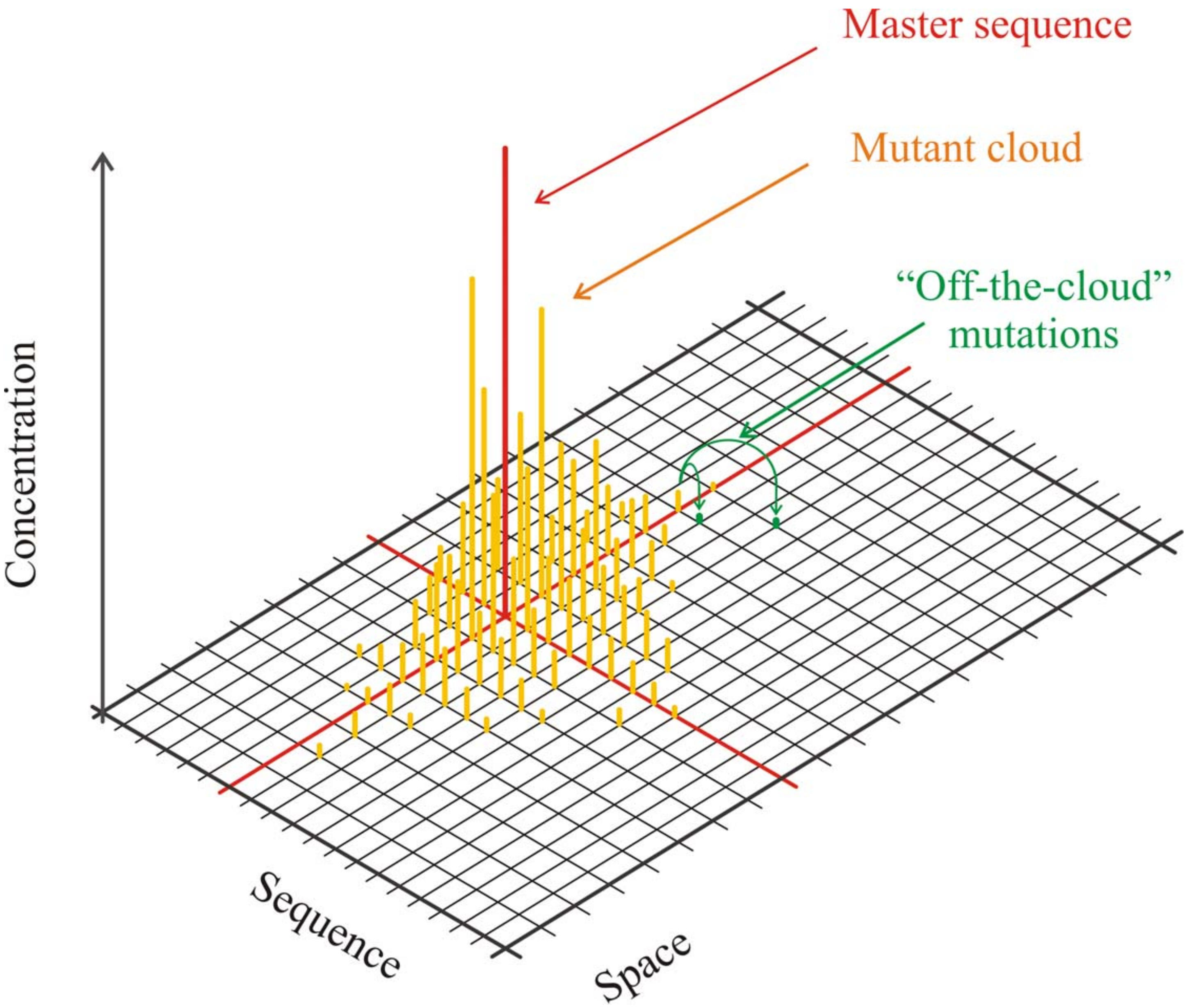
Phenylalanyl-tRNA as
target structure



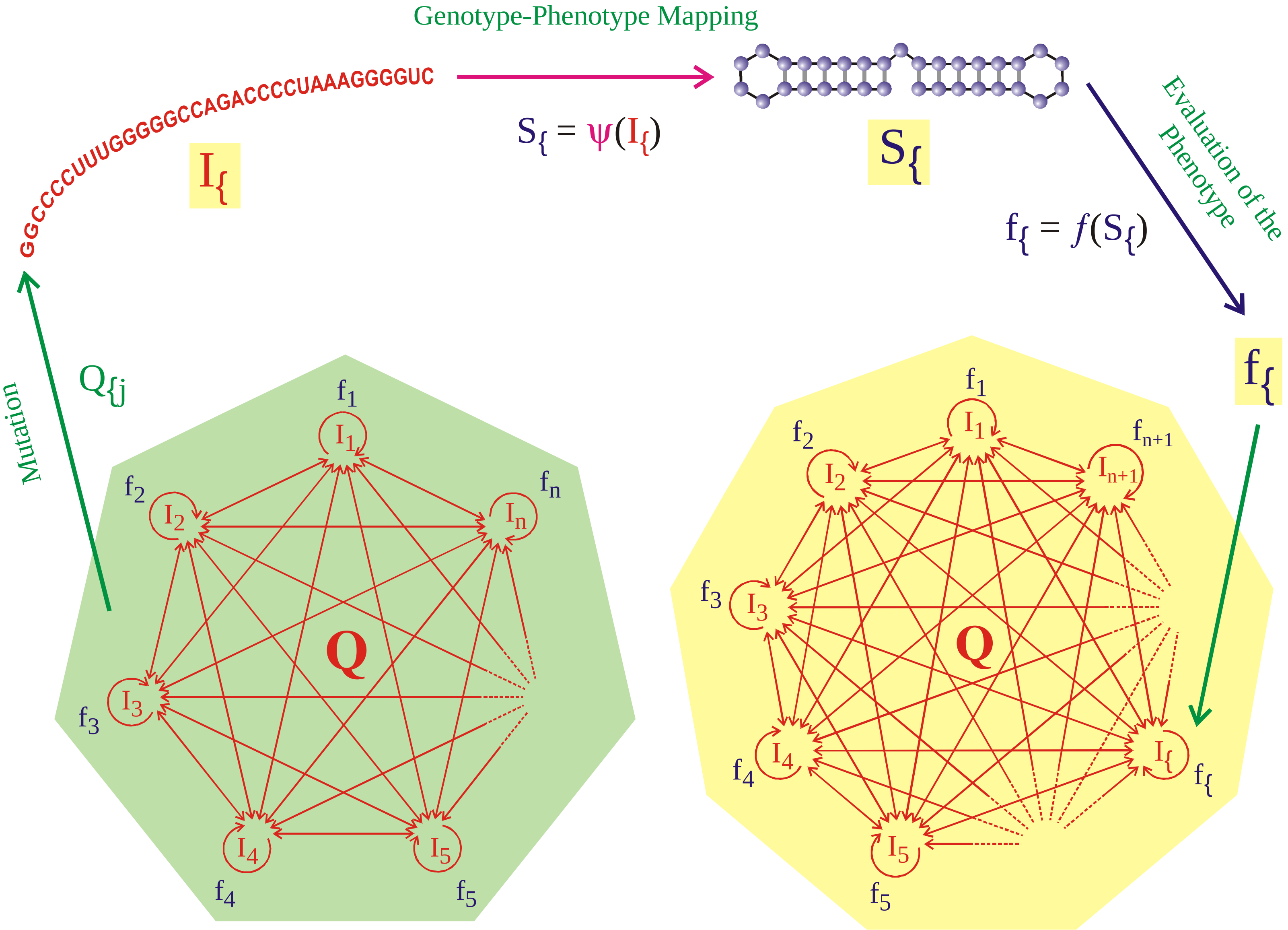
Formation of a quasispecies
in sequence space



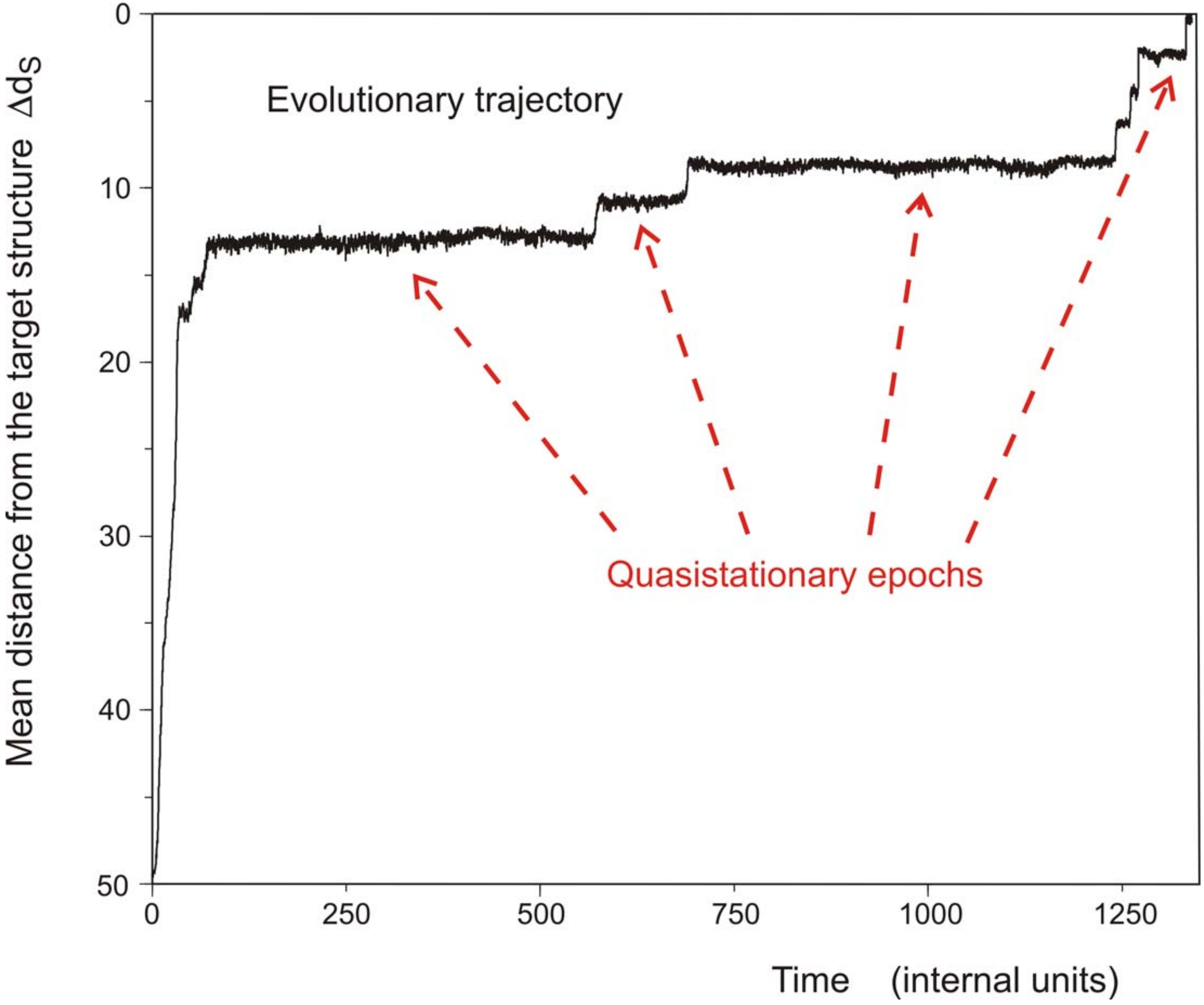
Formation of a quasispecies
in sequence space



Migration of a quasispecies
through sequence space

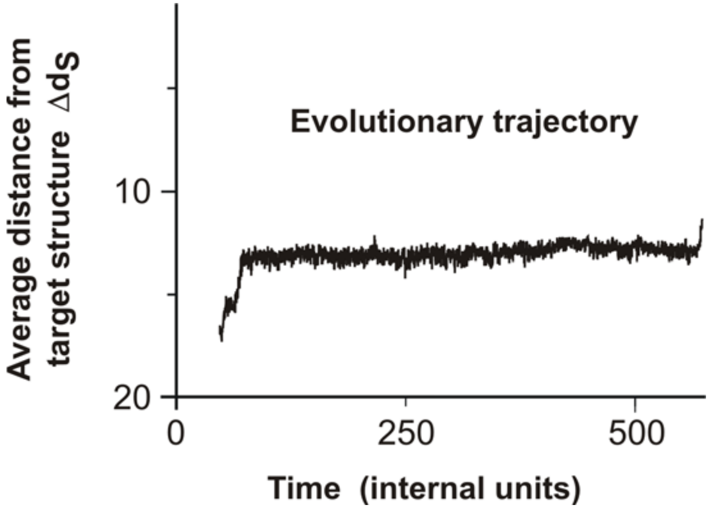


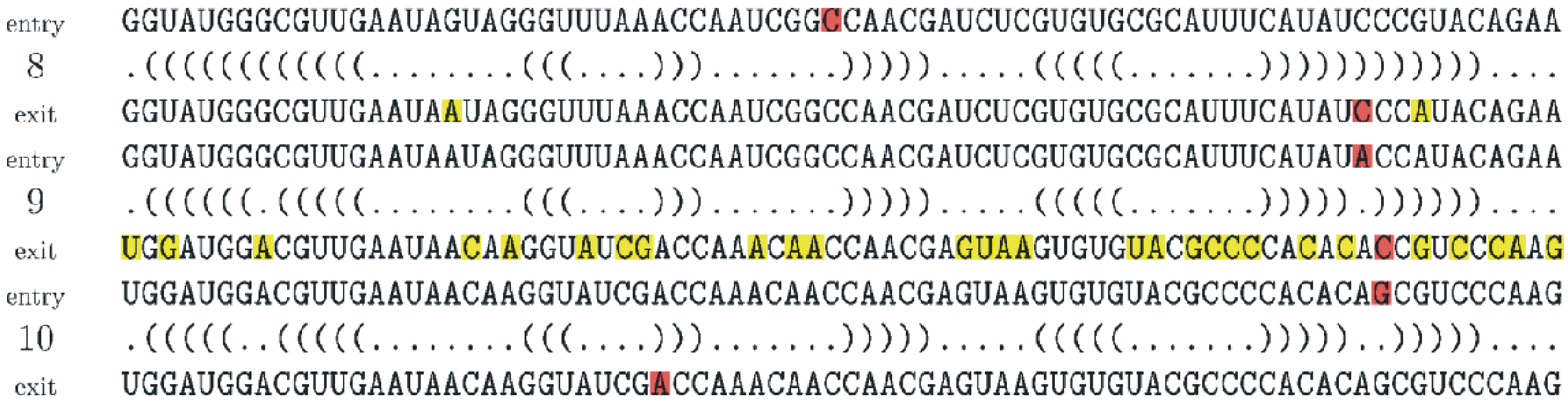
Evolutionary dynamics
including molecular phenotypes



In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch





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

Transition inducing point mutations change the molecular structure

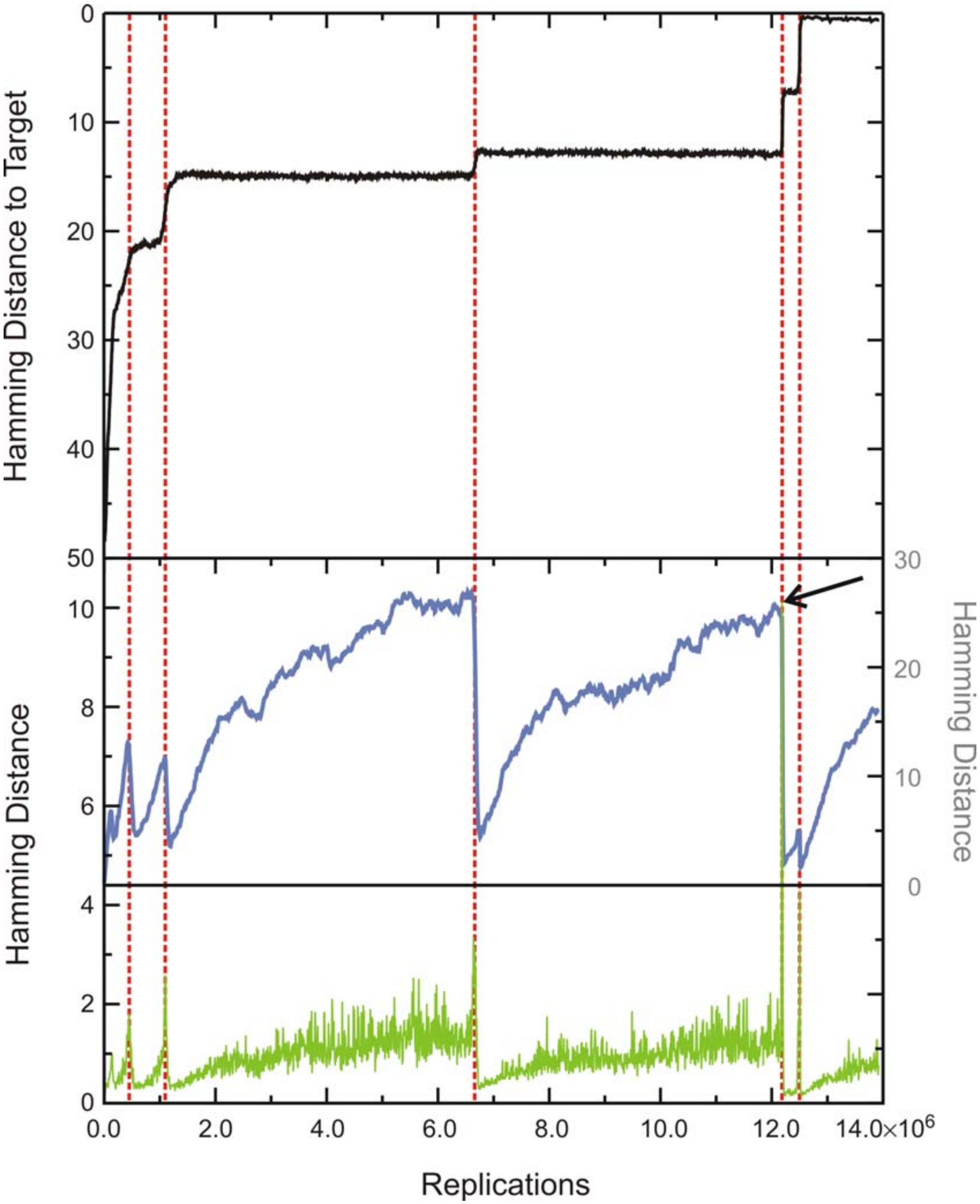
Neutral point mutations leave the molecular structure unchanged

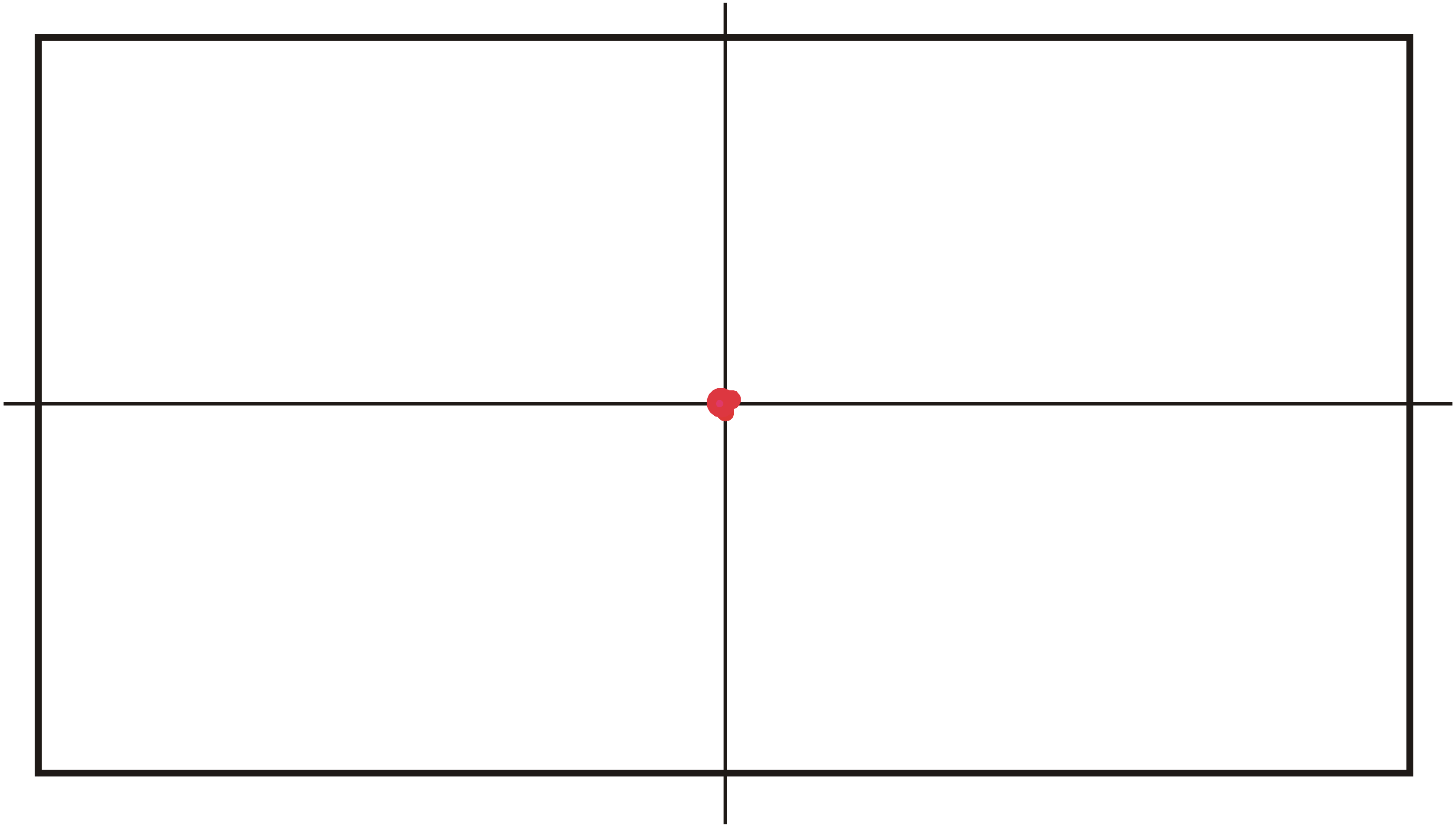
Neutral genotype evolution during phenotypic stasis

Evolutionary trajectory

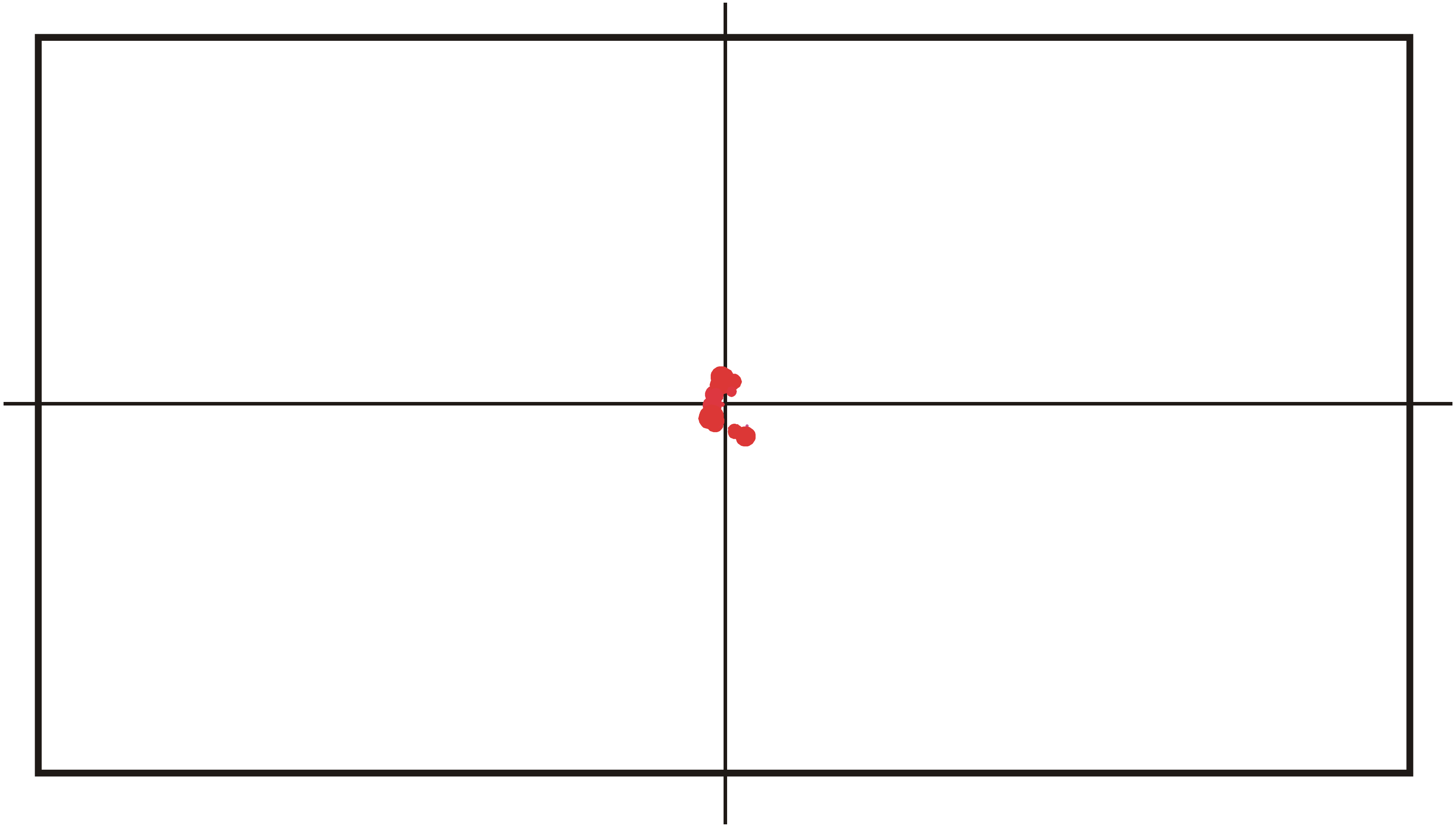
Spreading of the population on neutral networks

Drift of the population center in sequence space

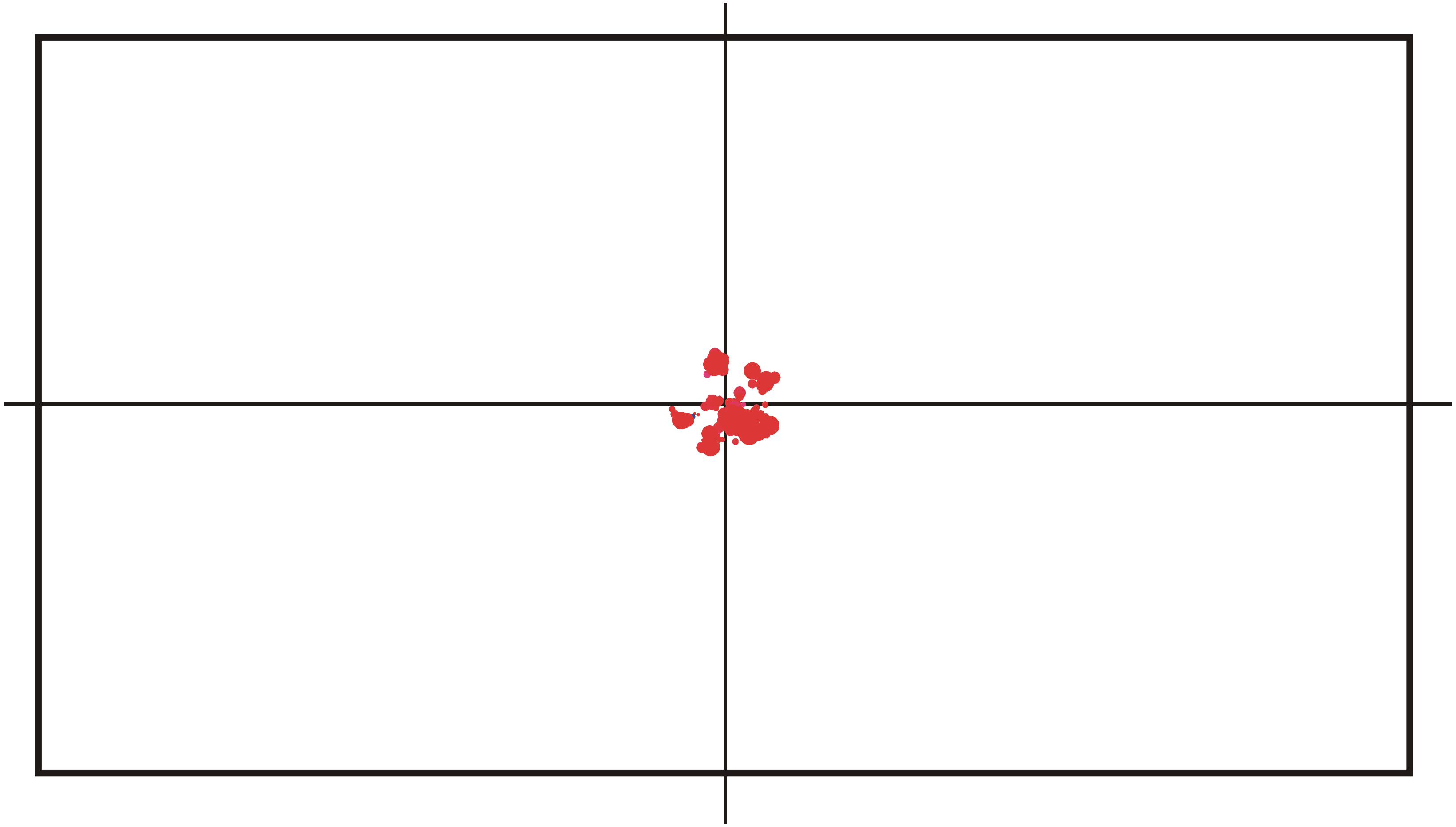




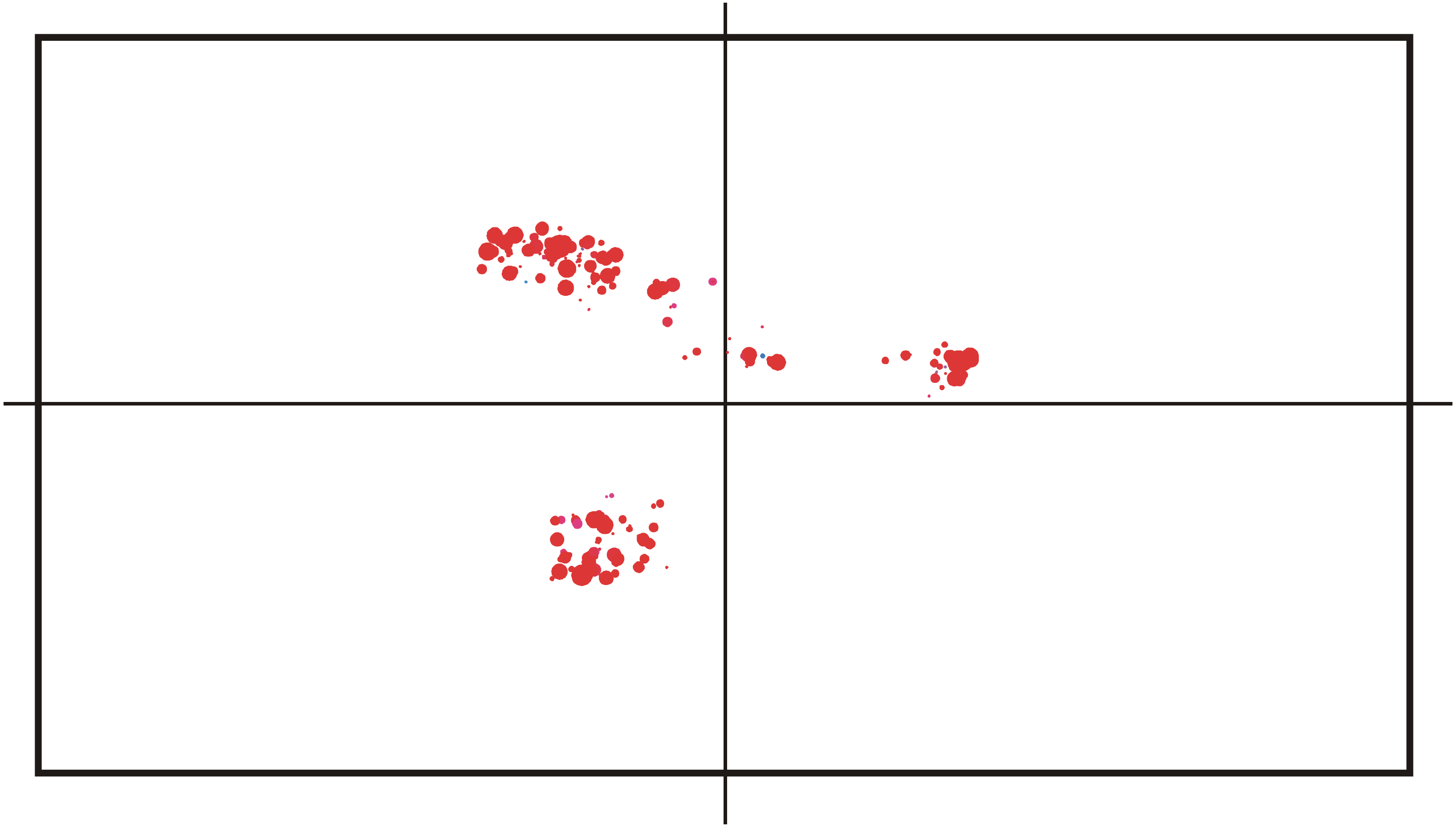
Spreading and evolution of a population on a neutral network: t = 150



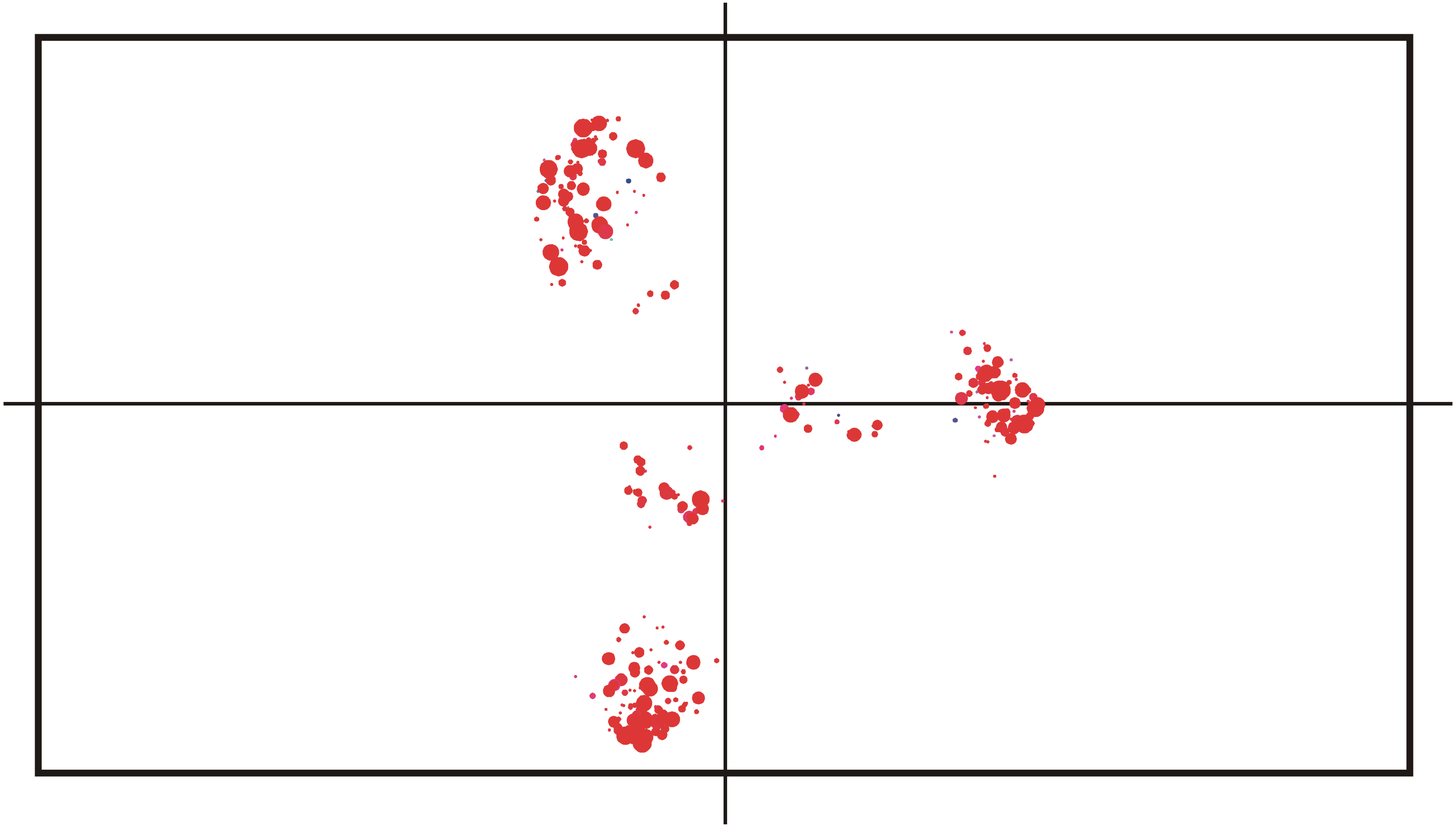
Spreading and evolution of a population on a neutral network : t = 170



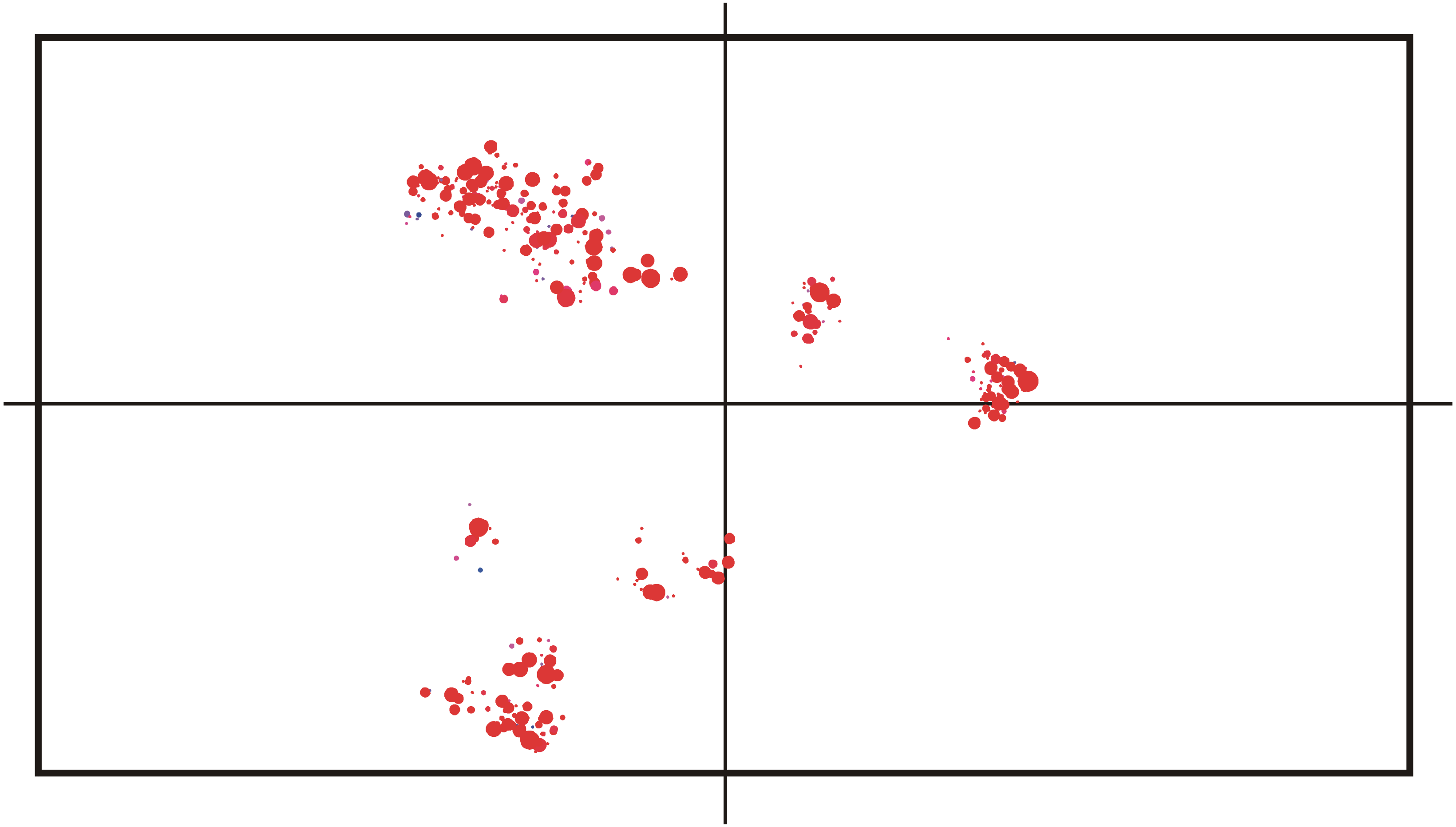
Spreading and evolution of a population on a neutral network : t = 200



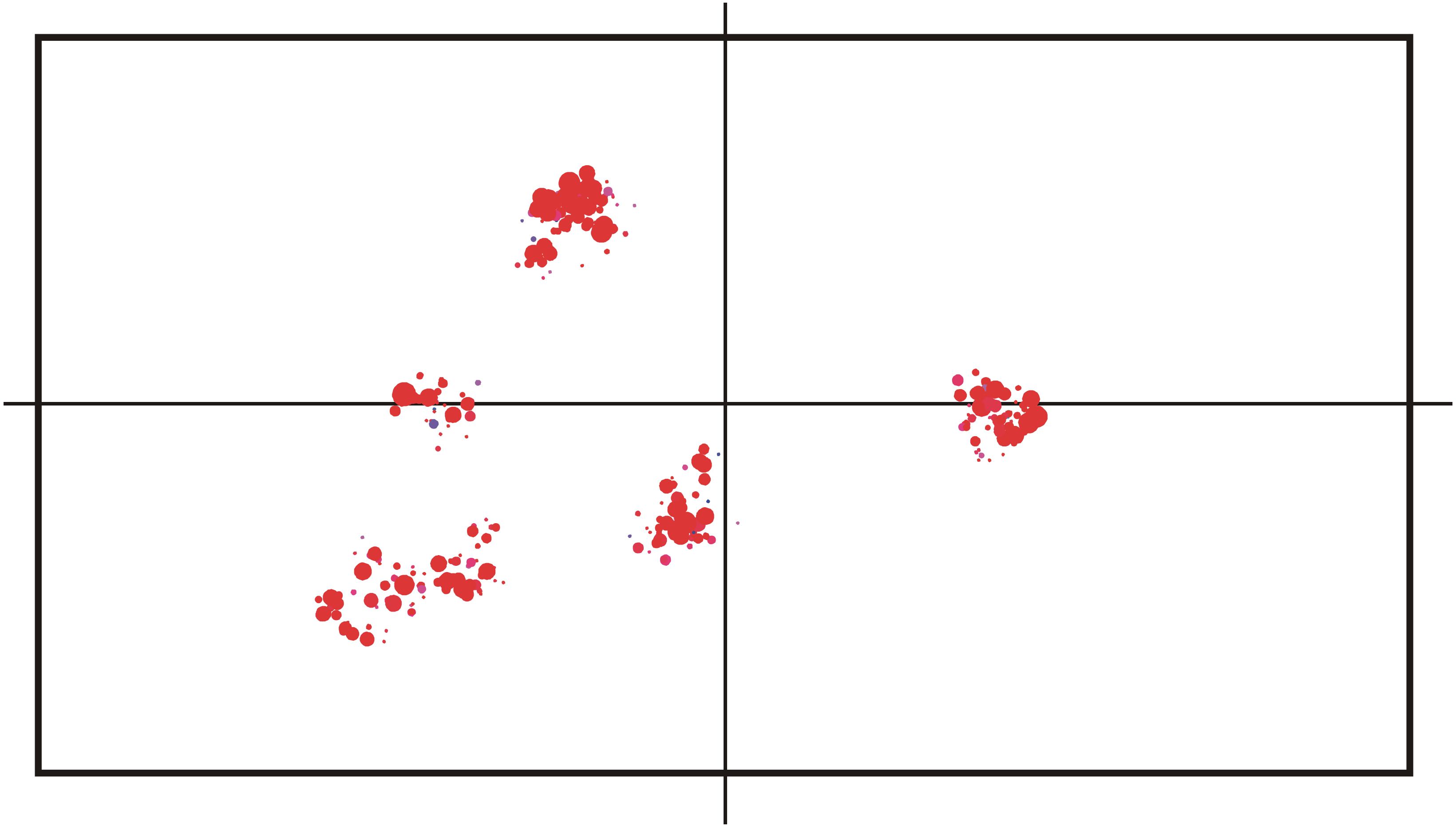
Spreading and evolution of a population on a neutral network : t = 350



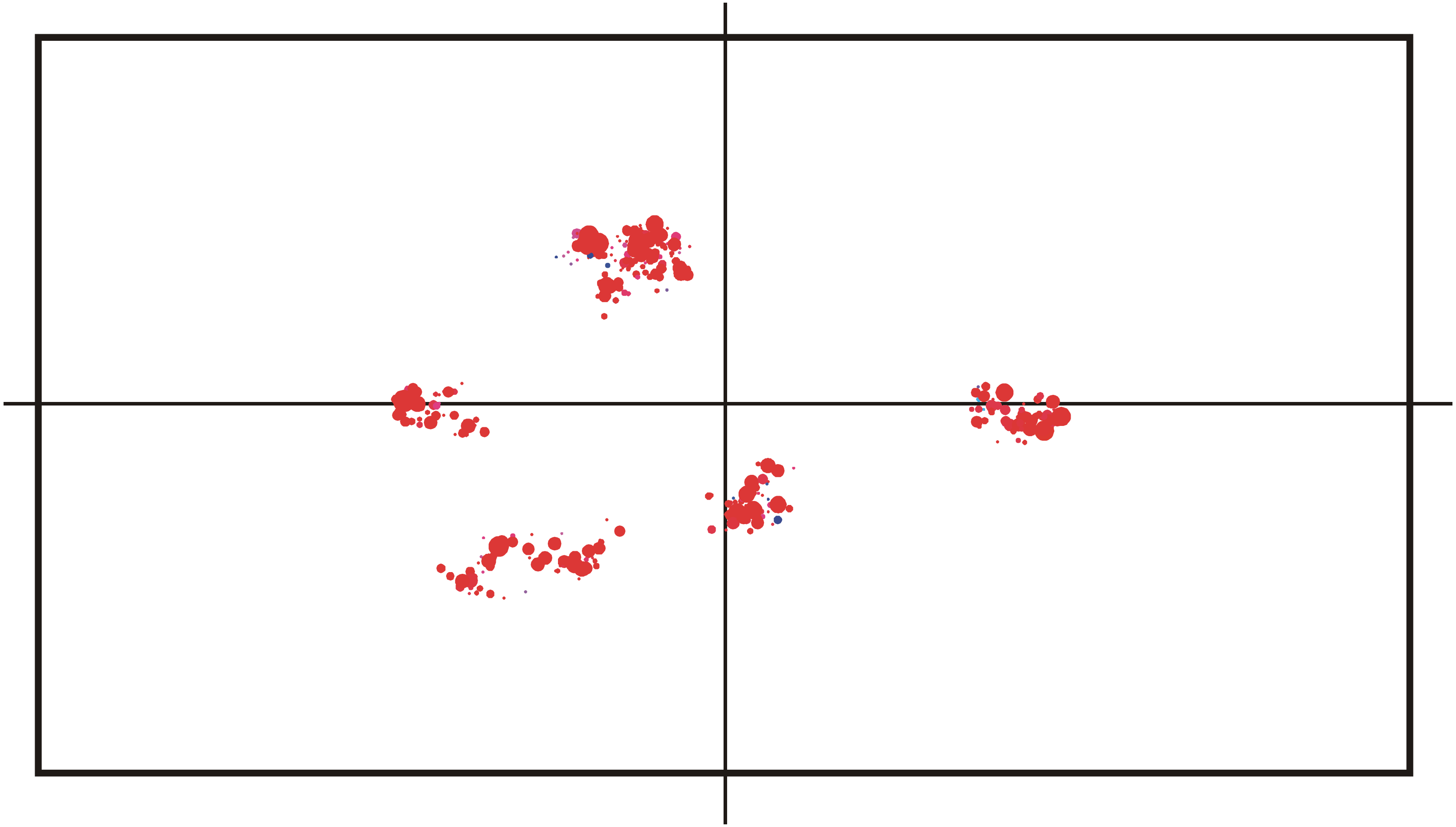
Spreading and evolution of a population on a neutral network : t = 500



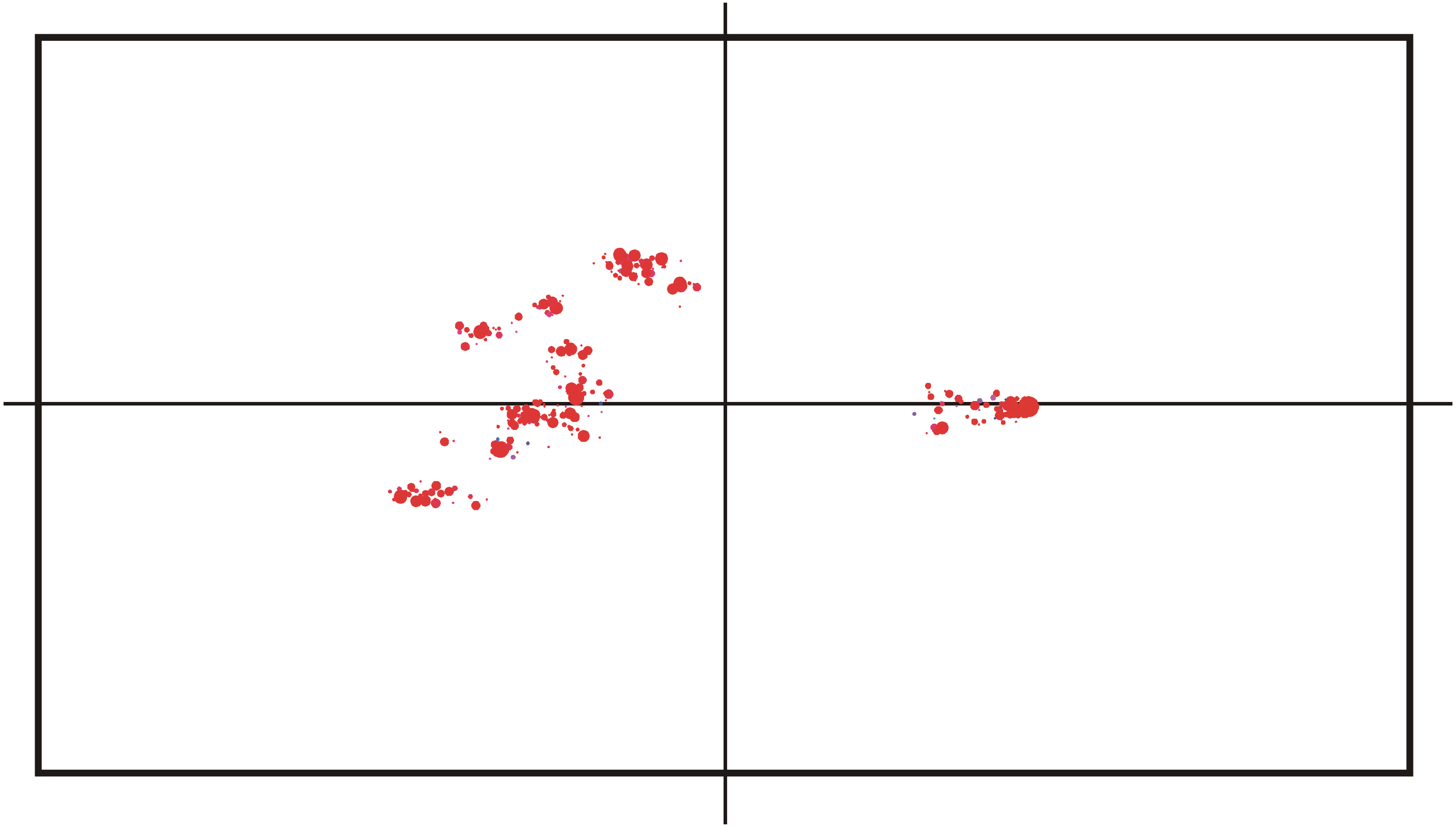
Spreading and evolution of a population on a neutral network : t = 650



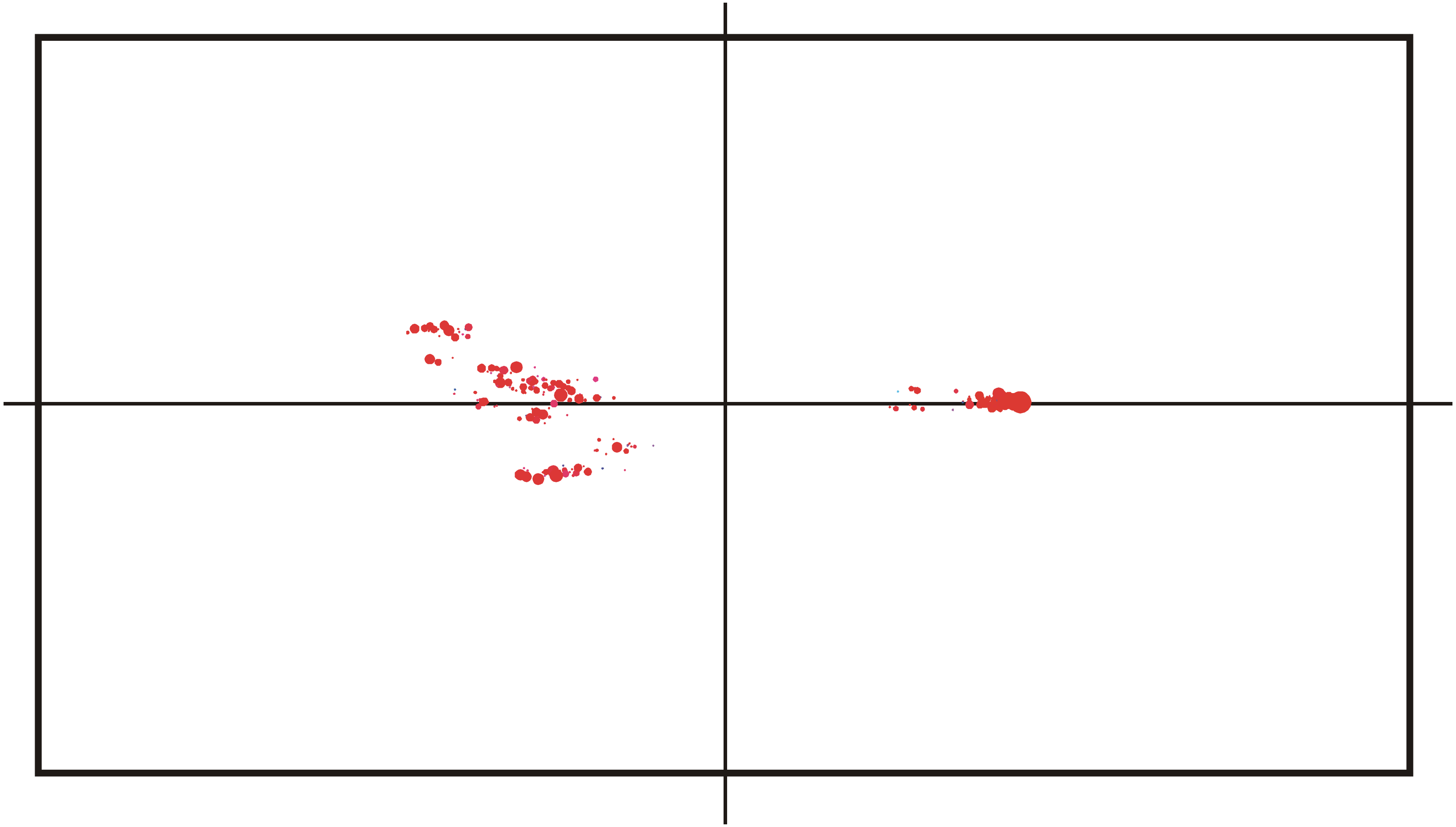
Spreading and evolution of a population on a neutral network : t = 820



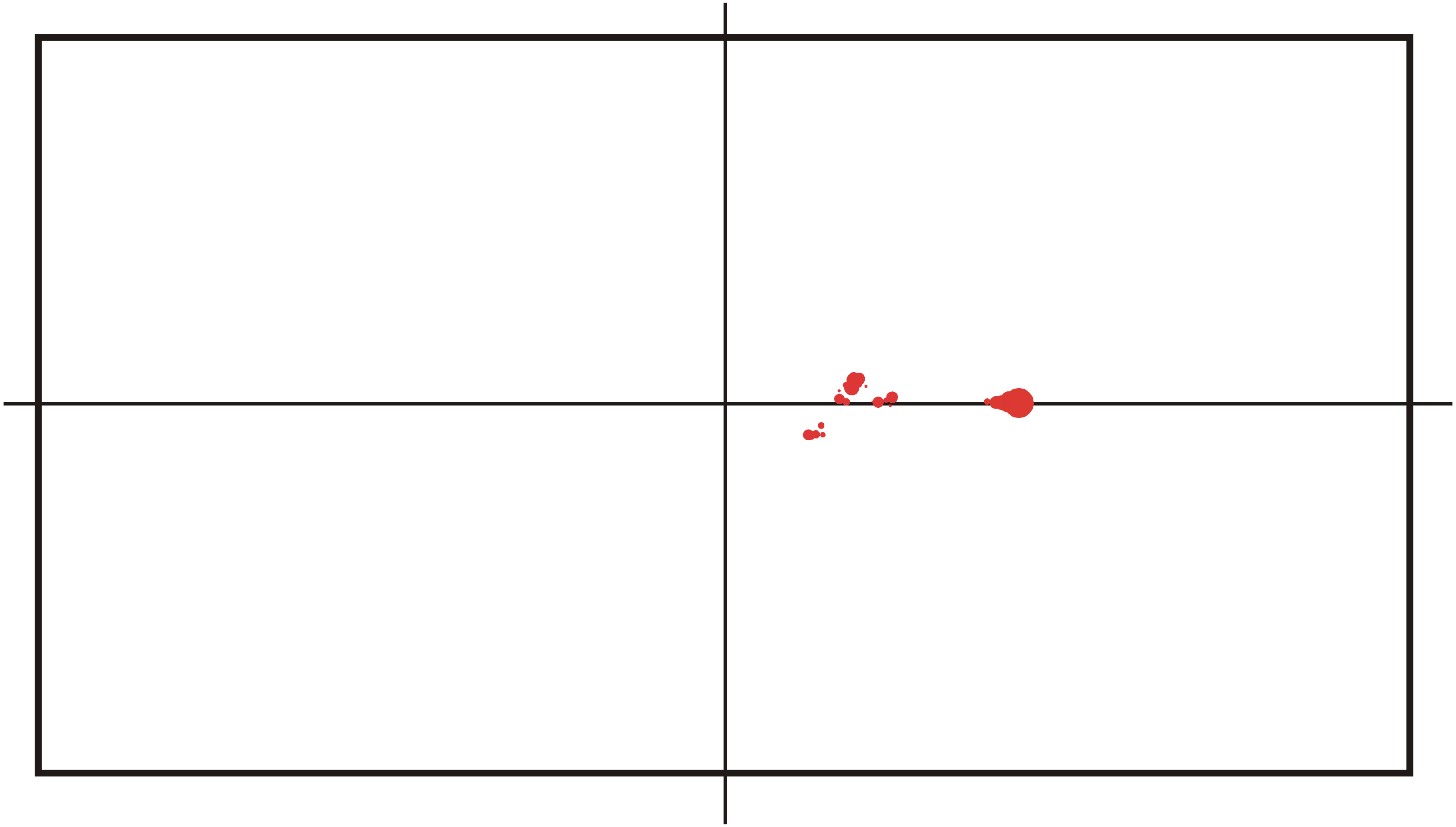
Spreading and evolution of a population on a neutral network : t = 825



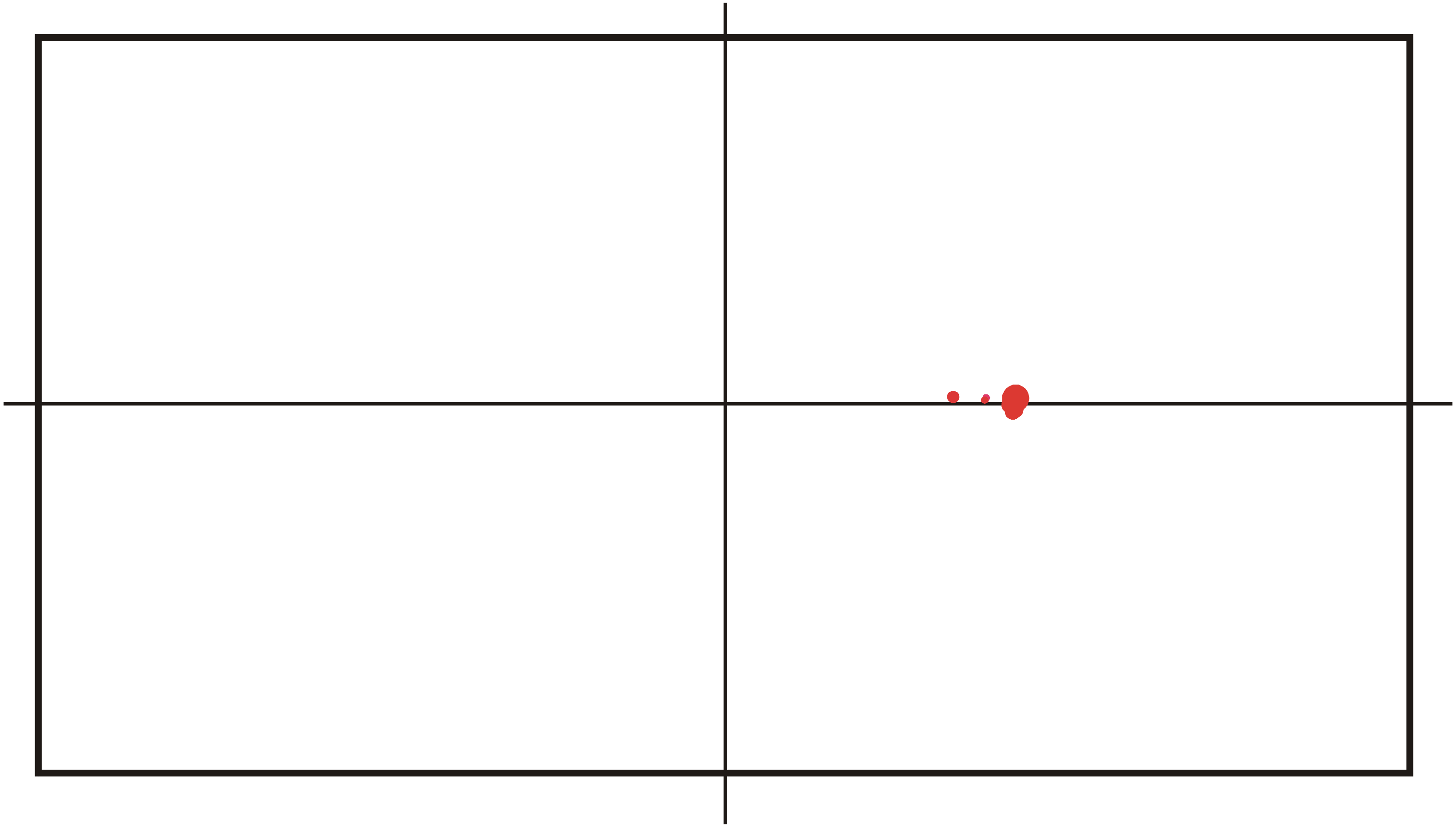
Spreading and evolution of a population on a neutral network : t = 830



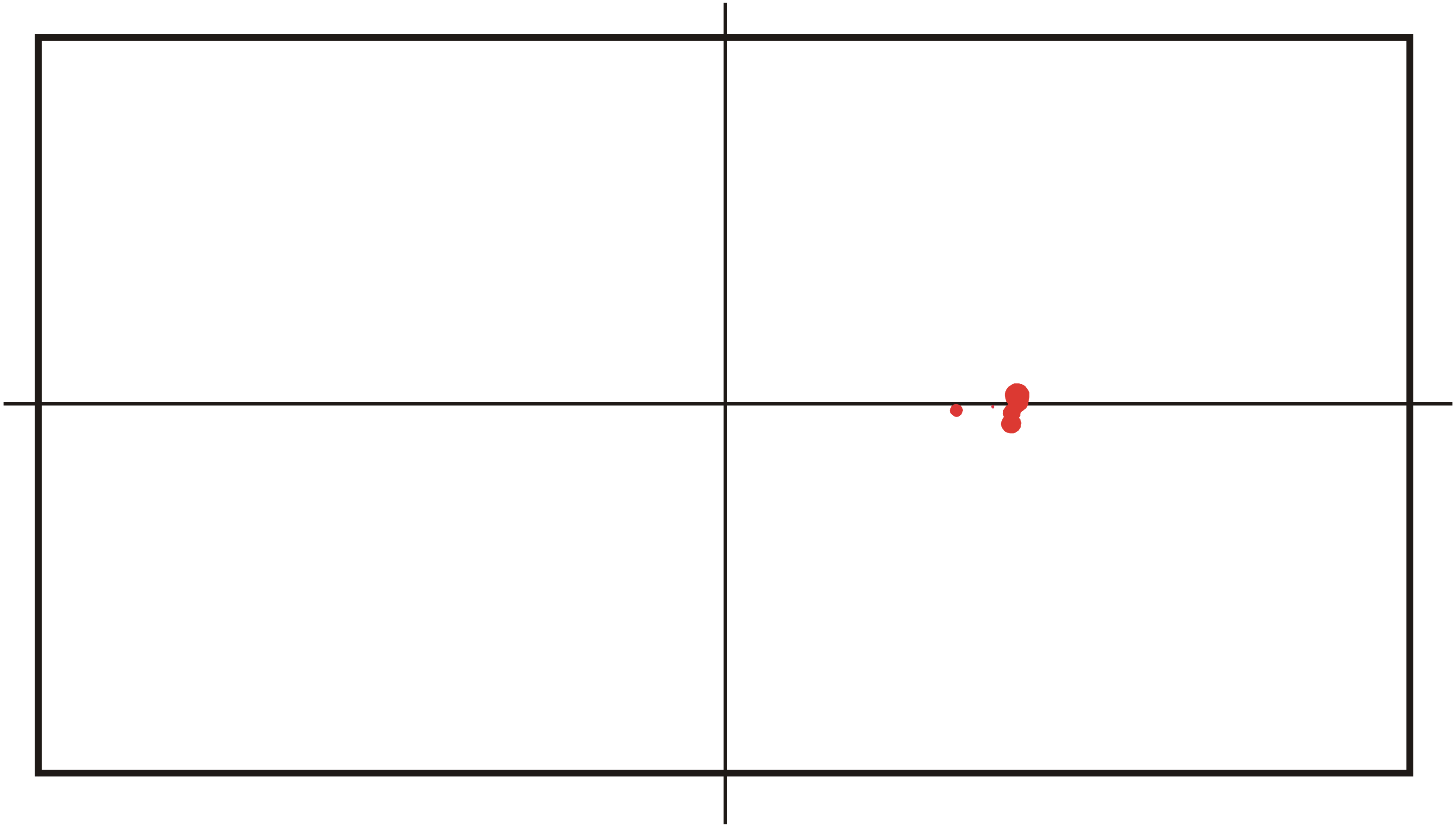
Spreading and evolution of a population on a neutral network : t = 835



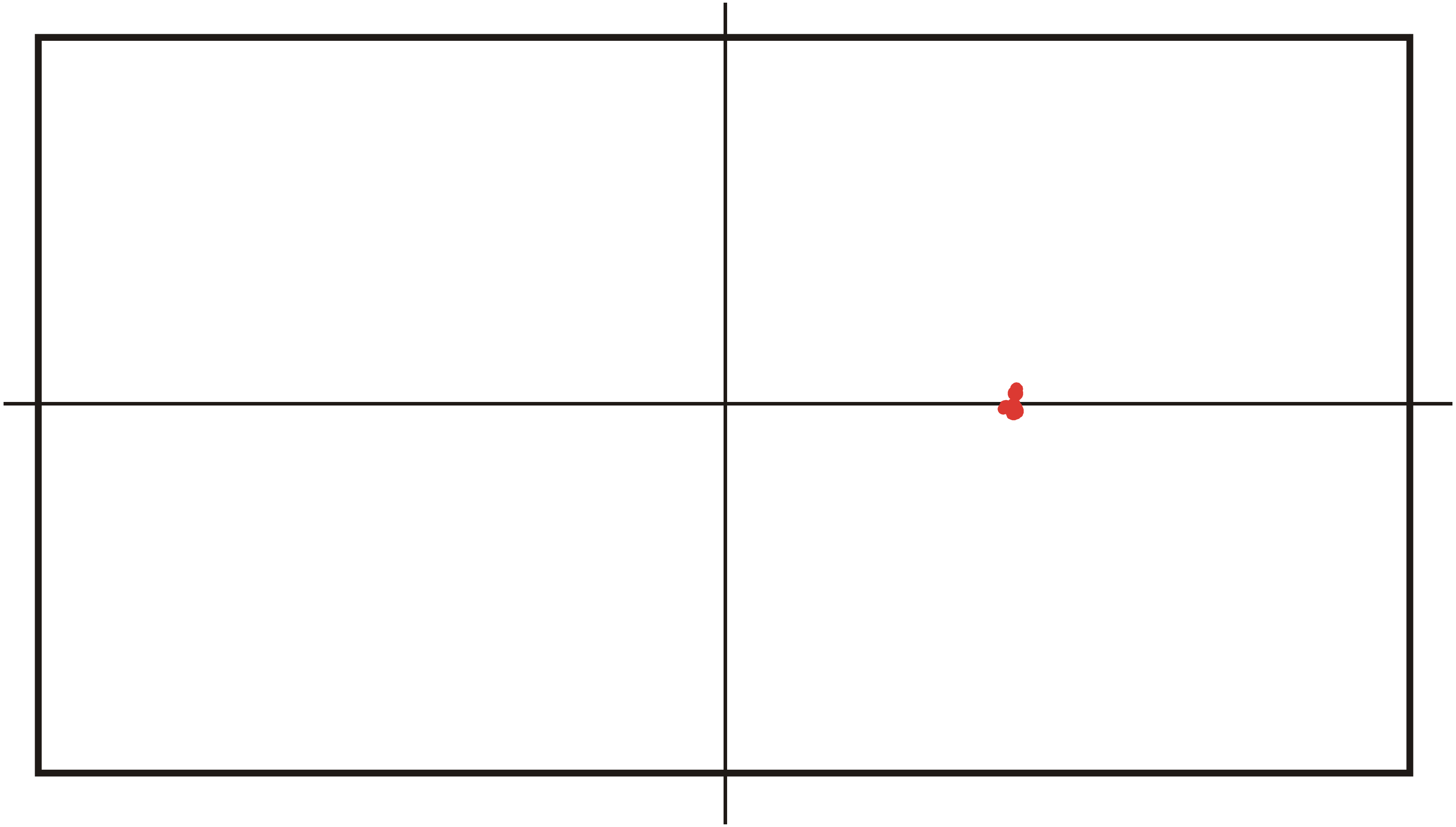
Spreading and evolution of a population on a neutral network : t = 840



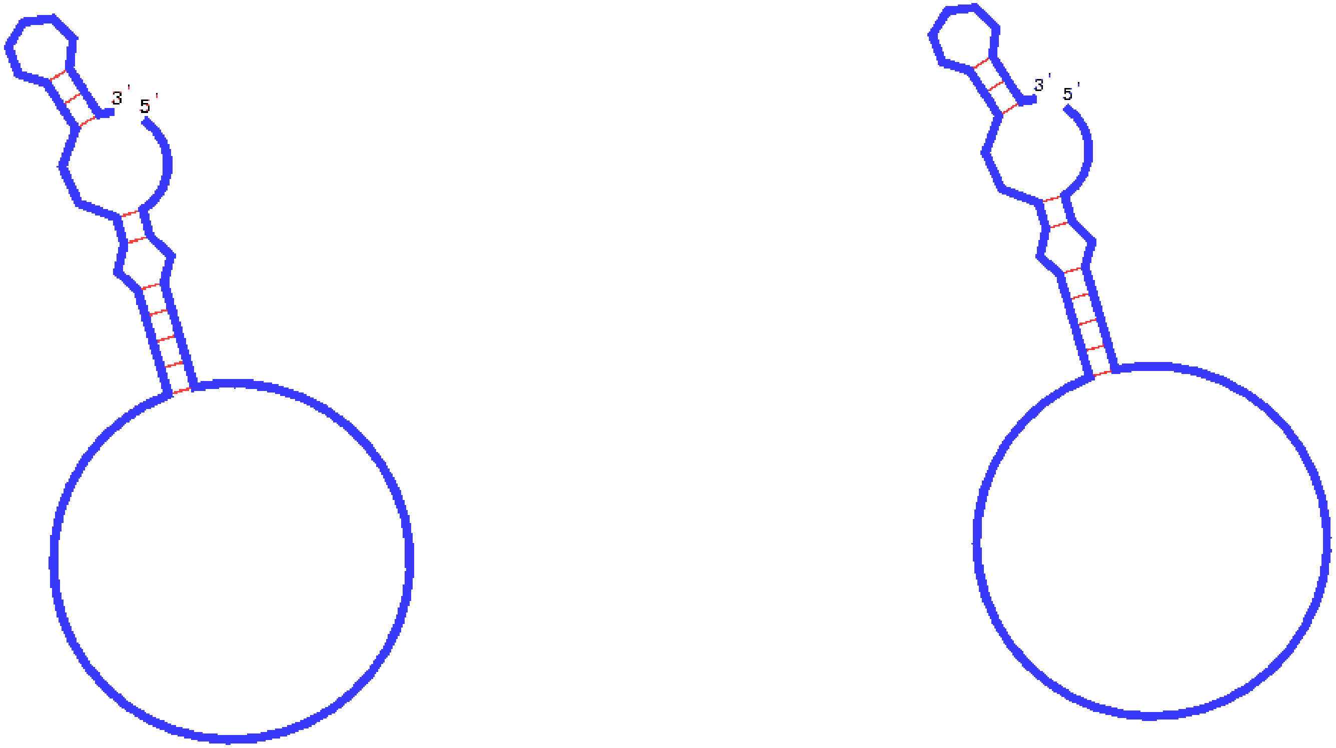
Spreading and evolution of a population on a neutral network : t = 845



Spreading and evolution of a population on a neutral network : t = 850



Spreading and evolution of a population on a neutral network : t = 855



AUGC

GC

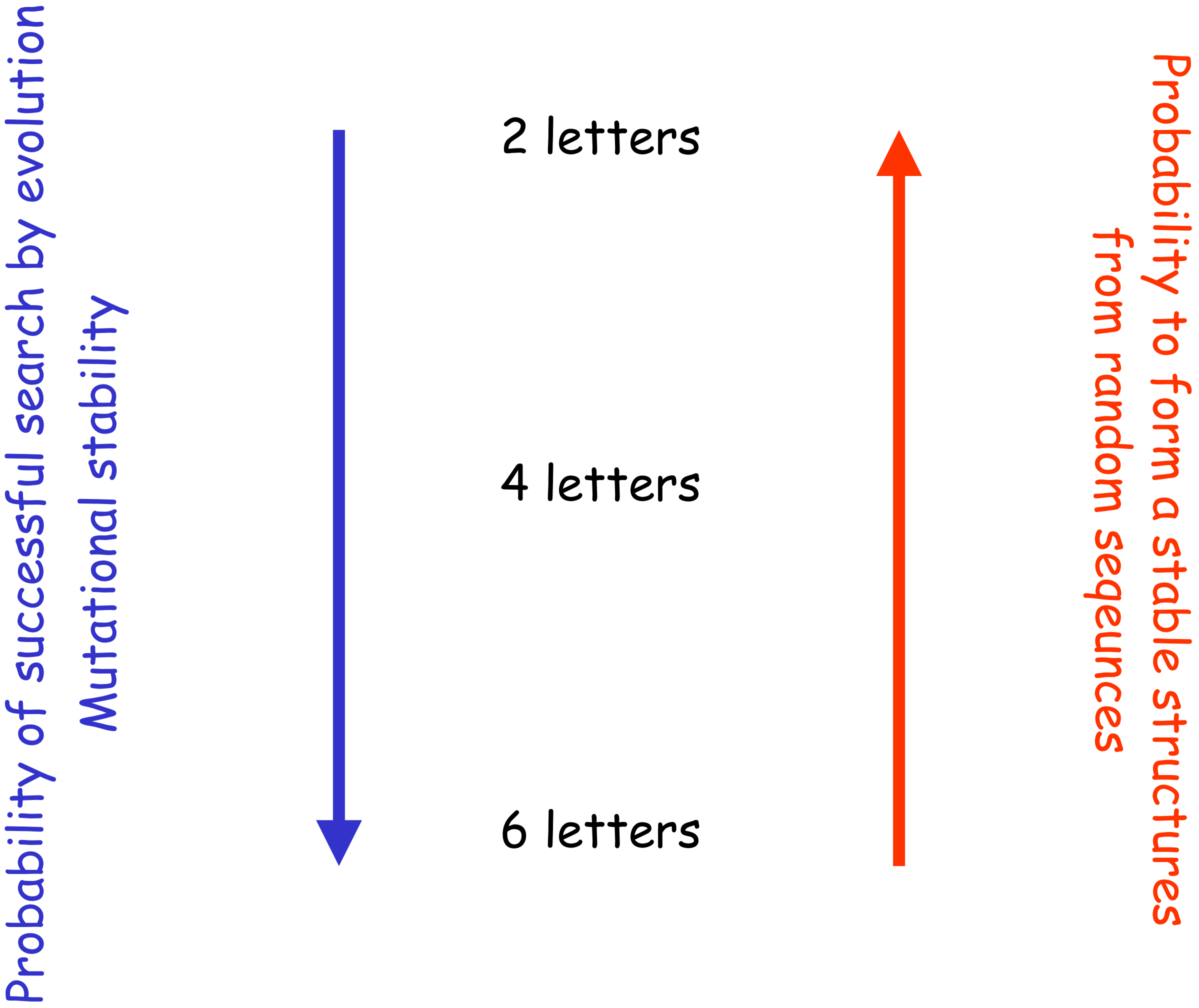
Movies of optimization trajectories over the **AUGC** and the **GC** alphabet

Alphabet	Runtime	Transitions	Main transitions	No. of runs
AUGC	385.6	22.5	12.6	1017
GUC	448.9	30.5	16.5	611
GC	2188.3	40.0	20.6	107

Mean population size: N = 3000 ; mutation rate: p = 0.001

Statistics of trajectories and relay series (mean values of log-normal distributions).

AUGC neutral networks of tRNAs are near the connectivity threshold, **GC** neutral networks are way below.



Probability of successful search by evolution
Mutational stability

2 letters

4 letters

6 letters

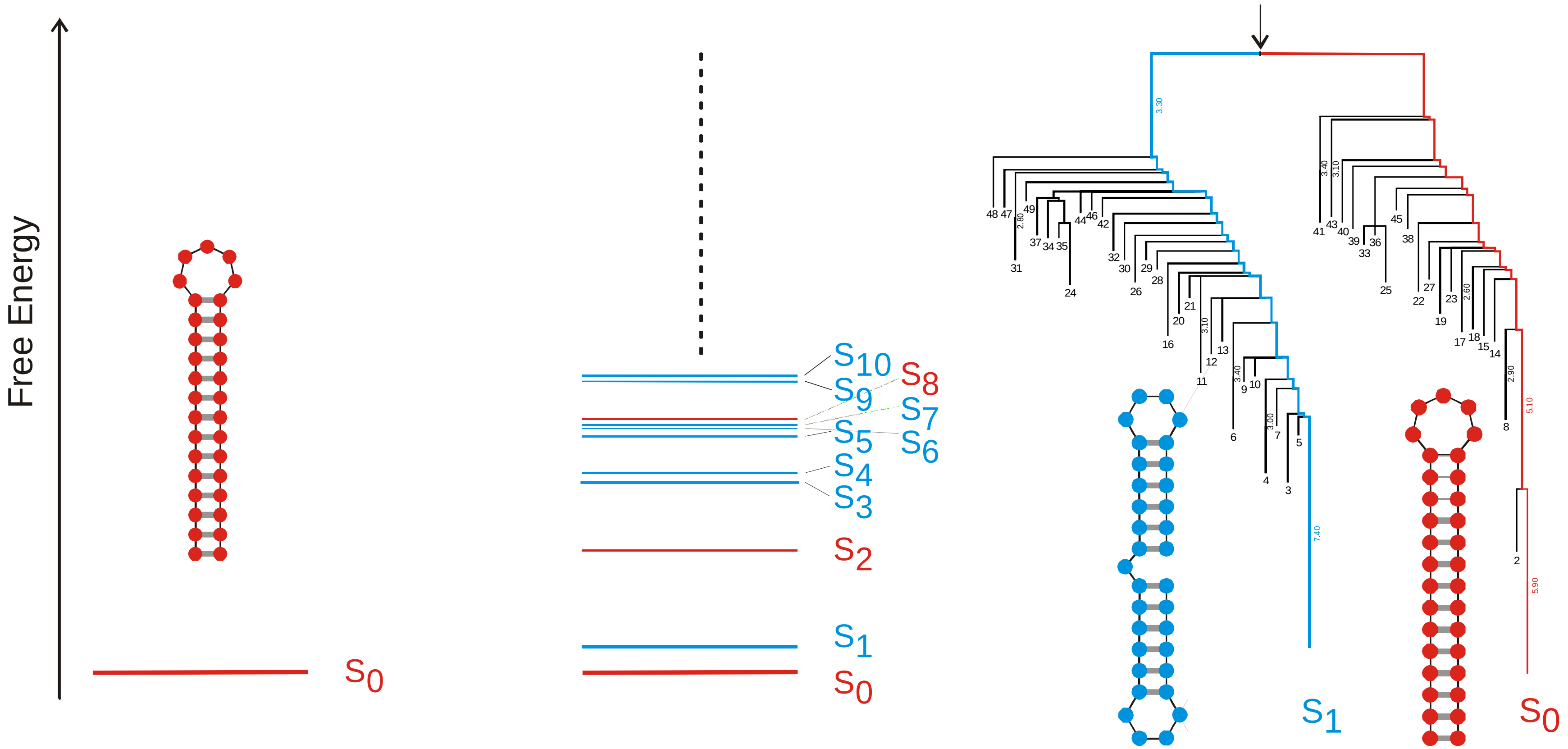
Probability to form a stable structures
from random seqeunces

1. The RNA model
2. How many stable structures can be formed?
3. Why not binary (GC or AU) sequences?
4. Evolution on neutral networks
5. **Multiconformational RNA molecules**

One sequence - one structure

Many suboptimal structures
Partition function

Metastable structures
Conformational switches

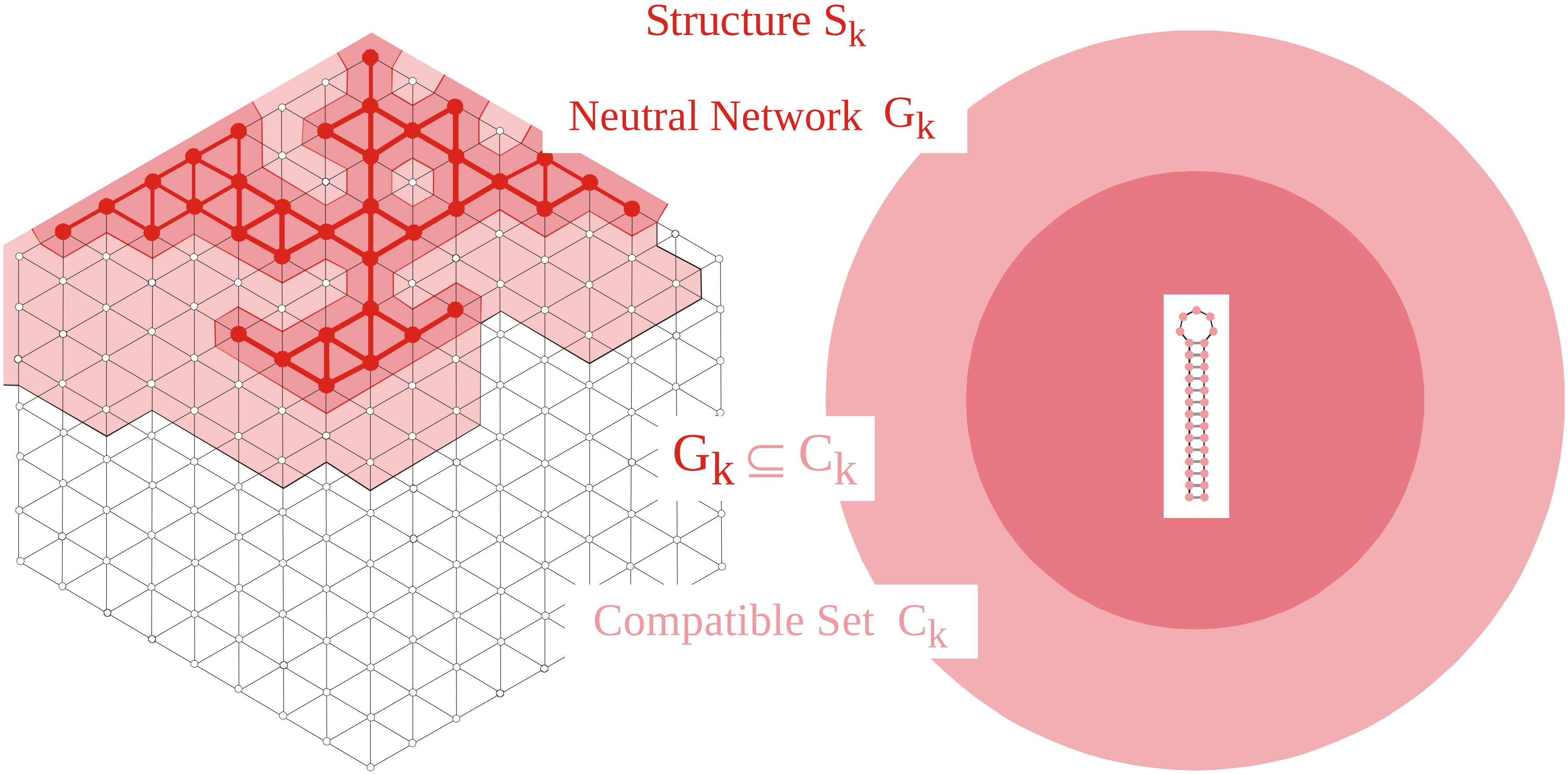


Minimum free energy structure

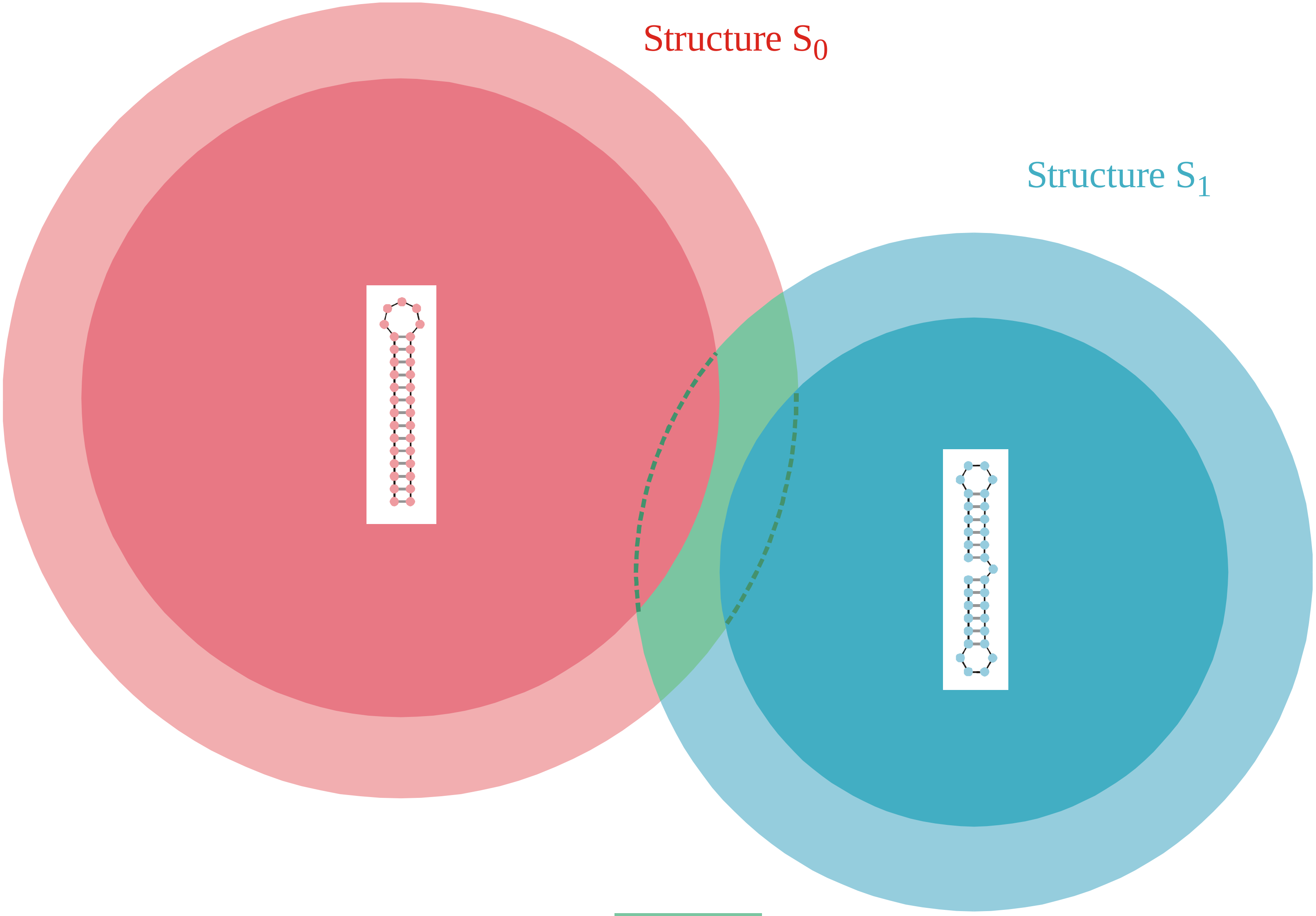
Suboptimal structures

Kinetic structures

RNA secondary structures derived from a single sequence

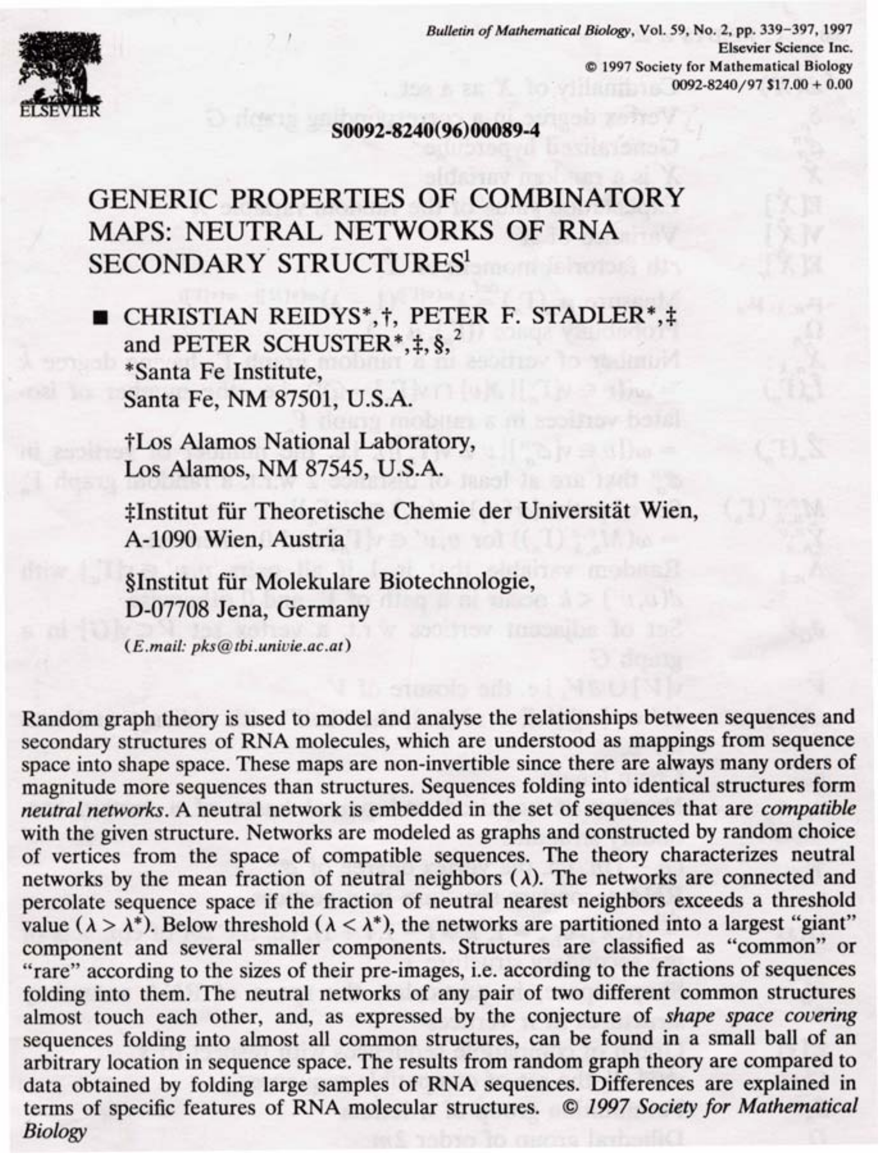


The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



Intersection of two compatible sets: $\boxed{\mathbf{C_0} \cap \mathbf{C_1}}$

The intersection of two compatible sets is always non empty: $\mathbf{C_0} \cap \mathbf{C_1} \notin \emptyset$



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]

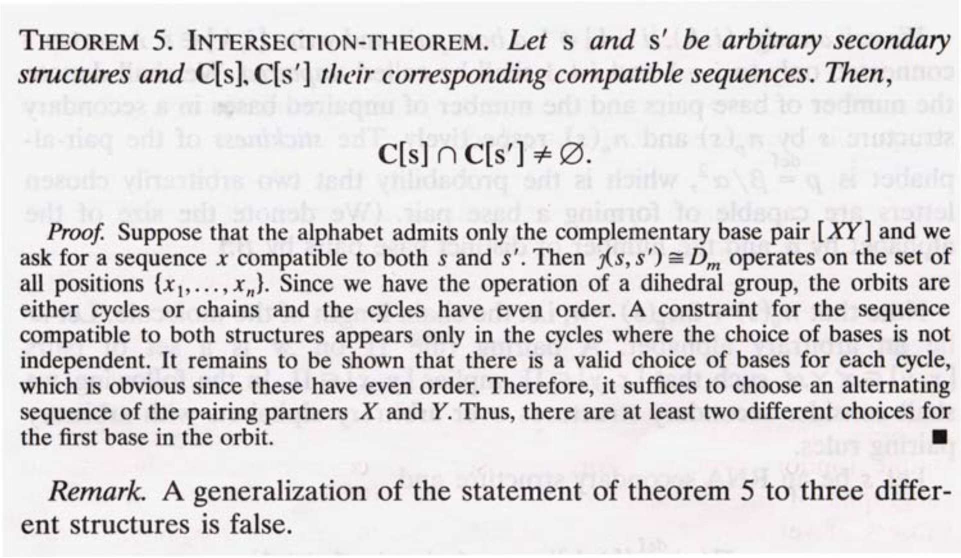
■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡
and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†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.



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

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

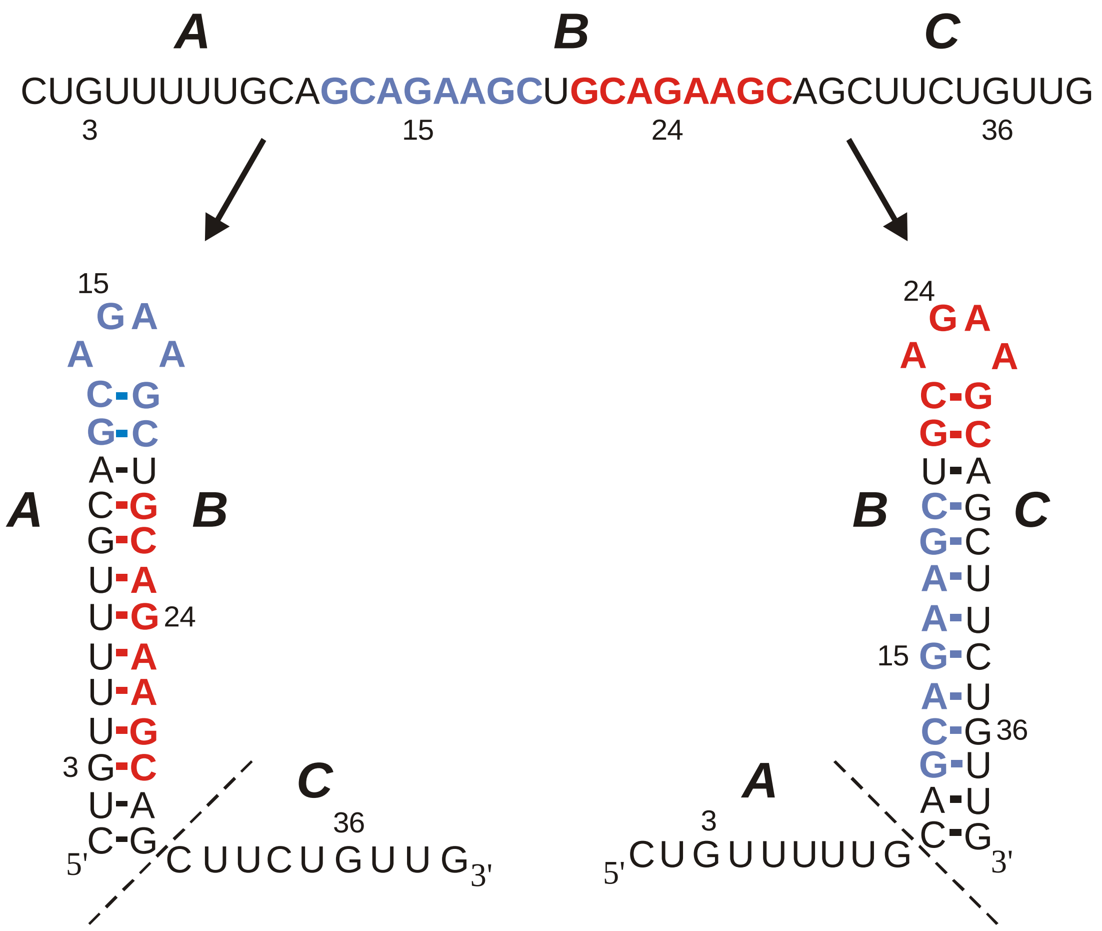
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\mathscr{J}(s,s') \cong D_m$ operates on the set of all positions $\{x_1,\dots,x_n\}$. Since we have the operation of a dihedral group, the orbits are either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

J. H. A. Nagel, C. Flamm, I. L. Hofacker, K. Franke, M. H. de Smit, P. Schuster, and C. W. A. Pleij. ***Structural parameters affecting the kinetic competition of RNA hairpin formation***, in press 2005.

J. H. A. Nagel, J. Møller-Jensen, C. Flamm, K. J. Öistämö, J. Besnard, I. L. Hofacker, A. P. Gultyaev, M. H. de Smit, P. Schuster, K. Gerdes and C. W. A. Pleij. ***The refolding mechanism of the metastable structure in the 5’-end of the*** hok ***mRNA of plasmid*** R1, submitted 2005.



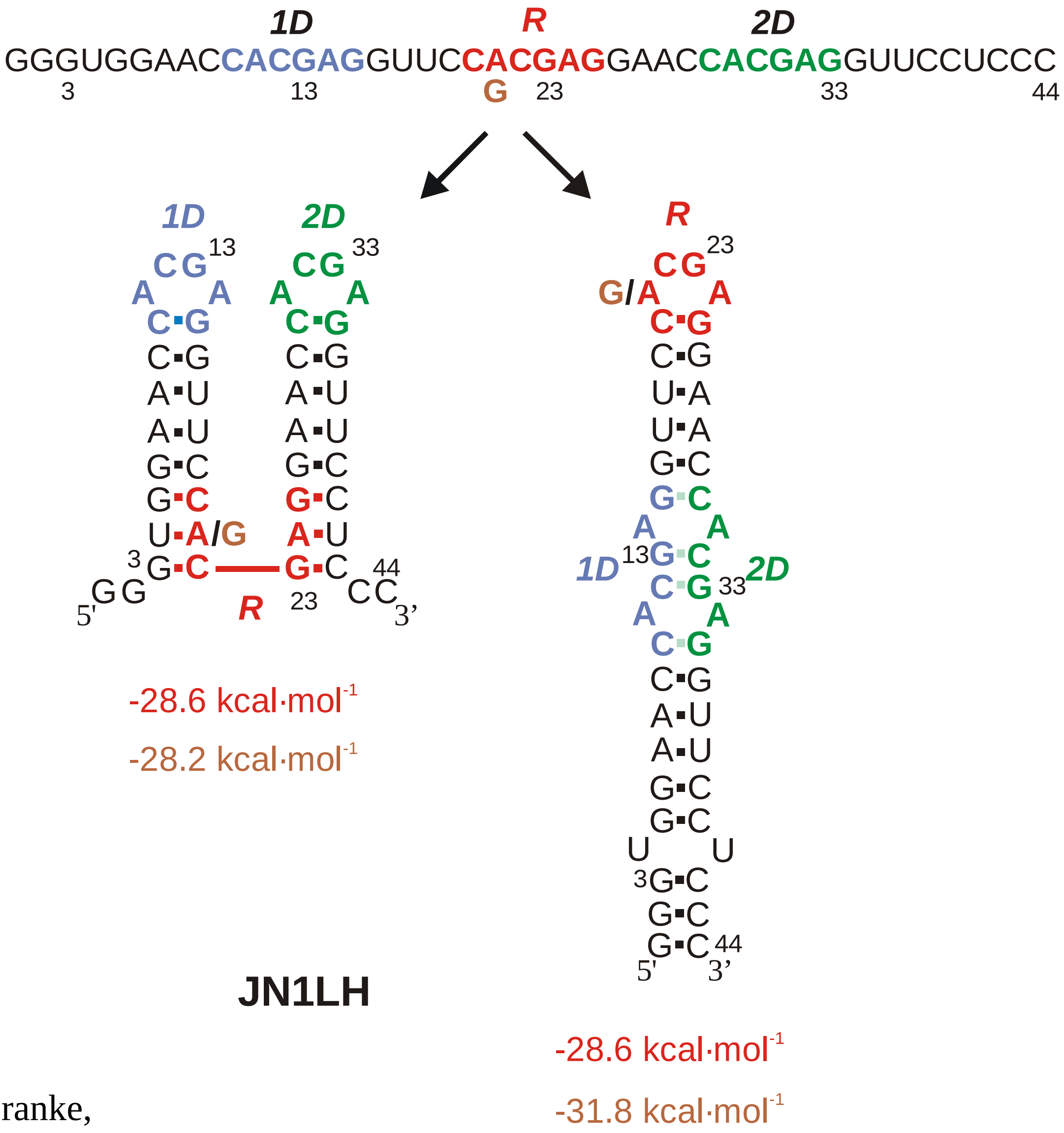
-19.5 kcal·mol^{-1}

-21.9 kcal·mol^{-1}

JN2C

J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke,
M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation, in press 2004.



-28.6 kcal·mol^{-1}

-28.2 kcal·mol^{-1}

JN1LH

-28.6 kcal·mol^{-1}

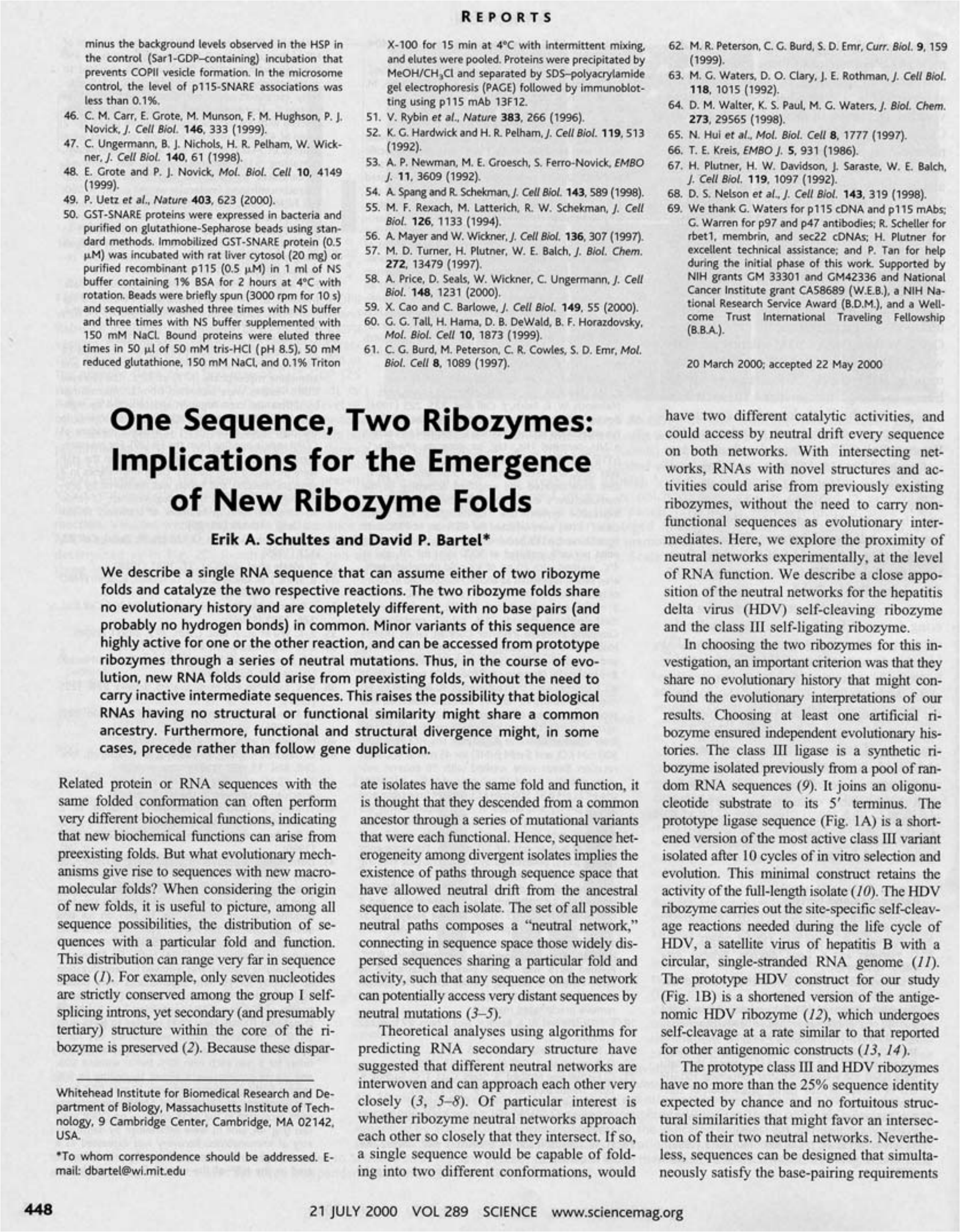
-31.8 kcal·mol^{-1}

J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke,
M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation, in press 2004.

A ribozyme switch

E.A.Schultes, D.B.Bartel, Science **289** (2000), 448-452



minus the background levels observed in the HSP in the control (Sar1-GDP–containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.

46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by $MeOH/CH_3Cl$ and separated by SDS–polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin *et al.*, *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).
62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui *et al.*, *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson *et al.*, *J. Cell Biol.* **143**, 319 (1998).

69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (*1*). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (*2*). Because these disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (*3–5*).

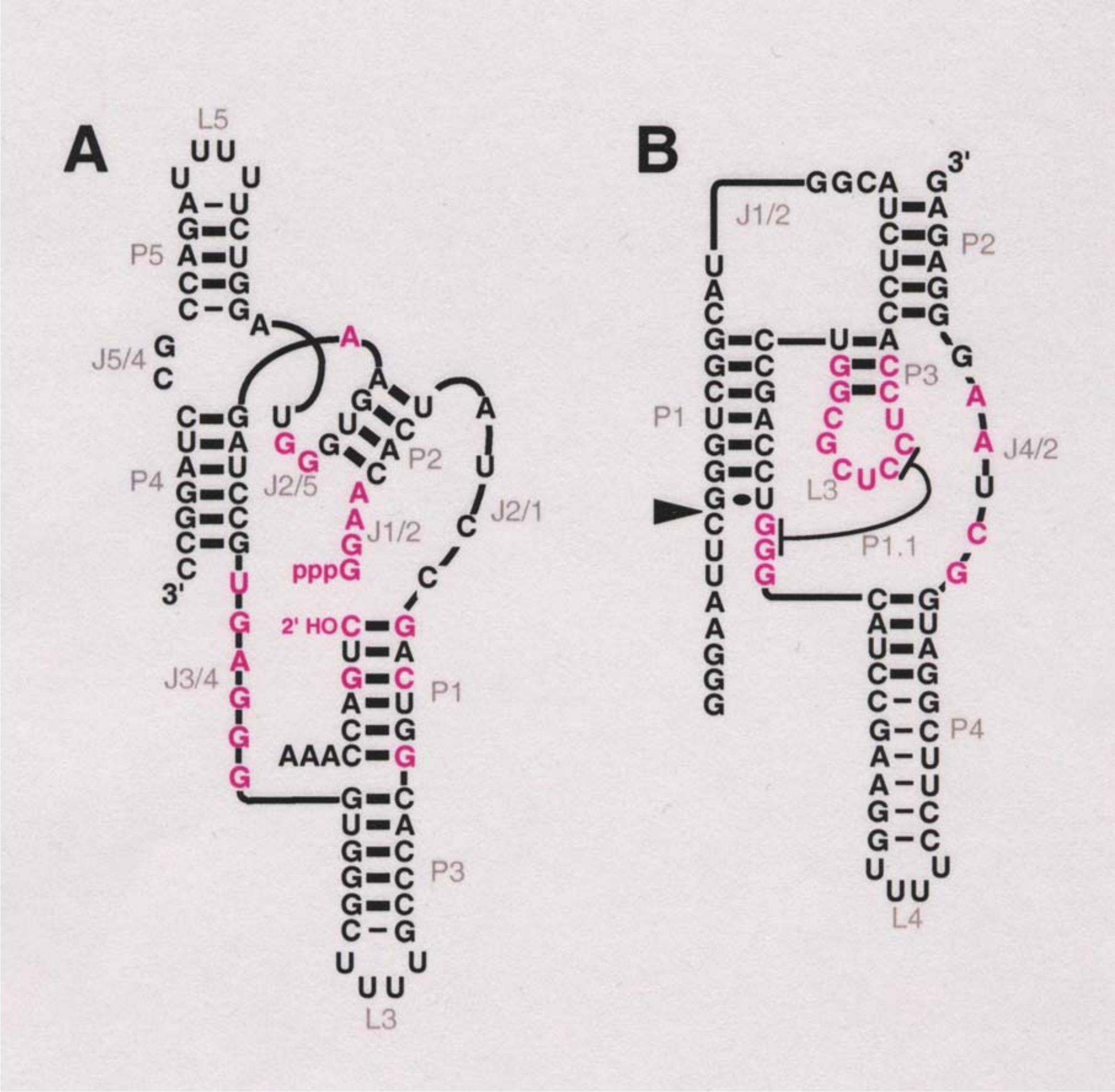
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (*9*). It joins an oligonucleotide substrate to its 5′ terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (*10*). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (*11*). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (*12*), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (*13*, *14*).

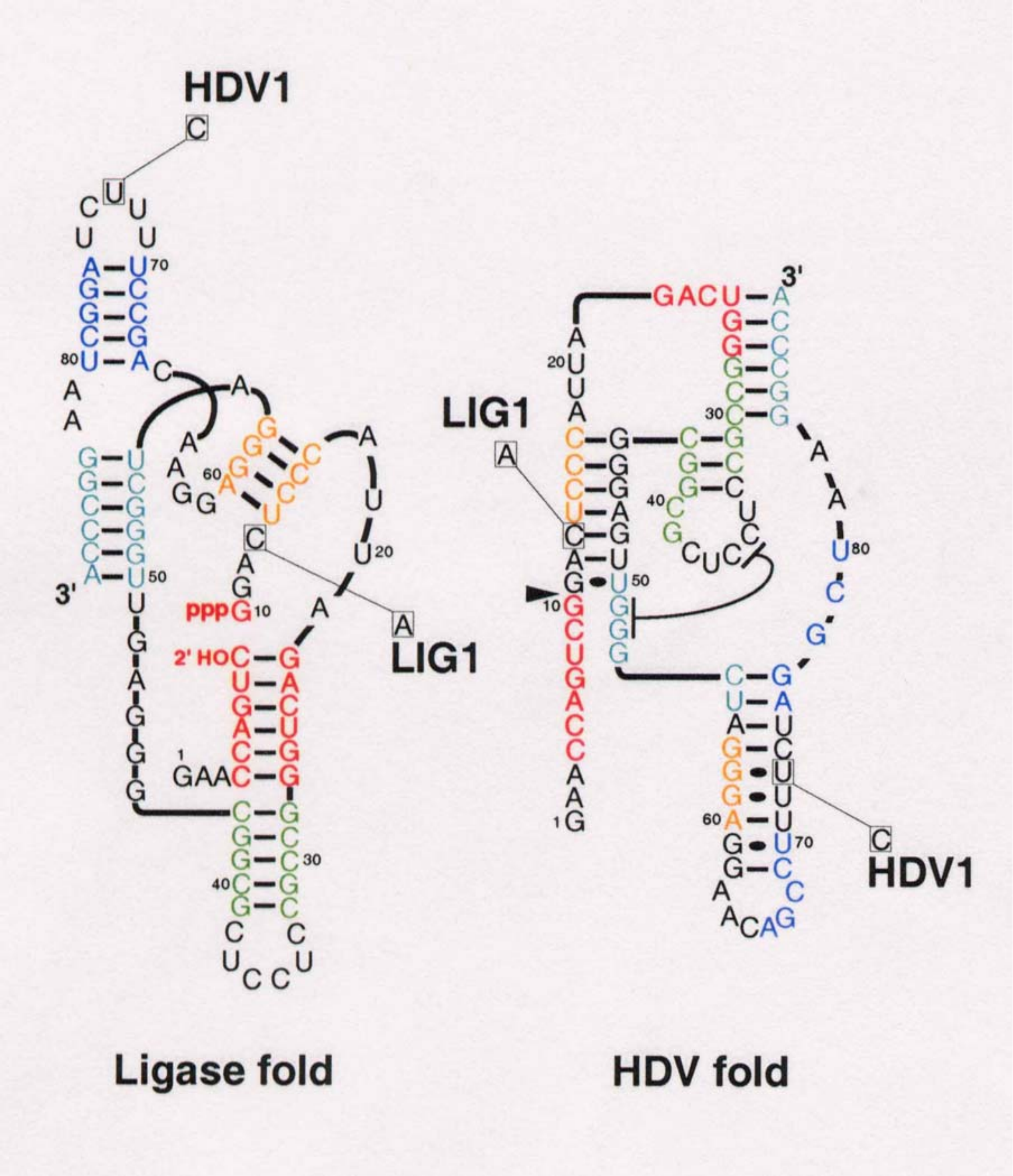
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

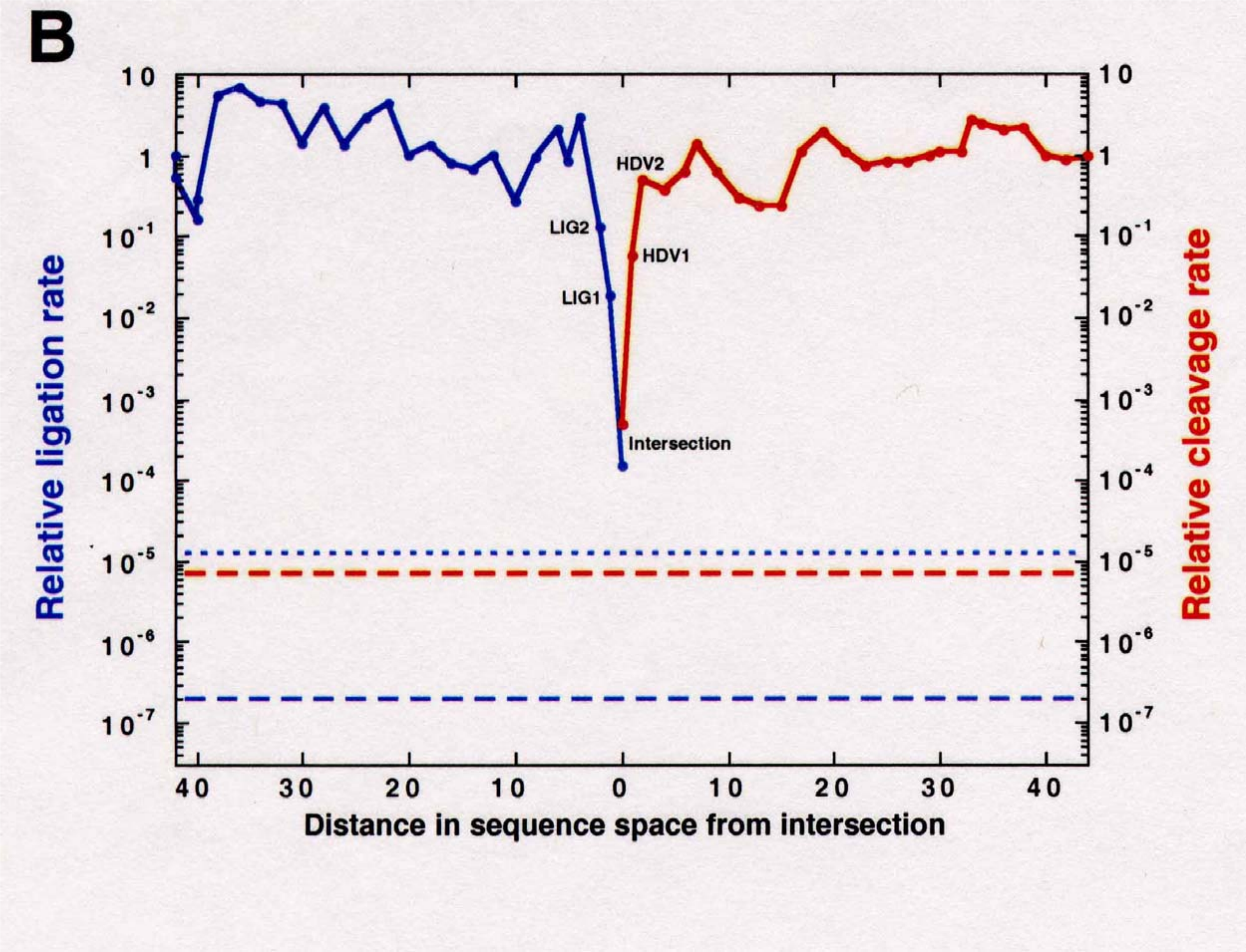


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)



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

Acknowledgement of support

Universität Wien

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)
Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF)
Project No. Mat05

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

European Commission: Contracts No. 98-0189, 12835 (NEST)

Austrian Genome Research Program – GEN-AU: Bioinformatics Network (BIN)

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Coworkers

Universität Wien

Peter Stadler, Bärbel M. Stadler, Universität Leipzig, GE

Paul E. Phillipson, University of Colorado at Boulder, CO

Heinz Engl, Philipp Kügler, James Lu, Stefan Müller, RICAM Linz, AT

Jord Nagel, Kees Pleij, Universiteit Leiden, NL

Walter Fontana, Harvard Medical School, MA

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

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

Ivo L.Hofacker, Christoph Flamm, **Andreas Svrček-Seiler**, Universität Wien, AT

Kurt Grünberger, **Michael Kospach** , **Andreas Wernitznig**, **Stefanie Widder**, **Stefan Wuchty**, Universität Wien, AT

Jan Cupal, **Stefan Bernhart**, **Lukas Endler**, **Ulrike Langhammer**, **Rainer Machne**, **Ulrike Mückstein**, **Hakim Tafer**, **Thomas Taylor**, Universität Wien, AT

Web-Page for further information:

http://www.tbi.univie.ac.at/~pks