RNA Evolution, Kinetic Folding, and Molecular Switches

Peter Schuster

Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien

The Origin of Life and the Early Evolution of Cells and Organelles

Les Treilles, 10.– 16.09.2004

Web-Page for further information:

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

1. RNA structures

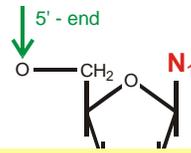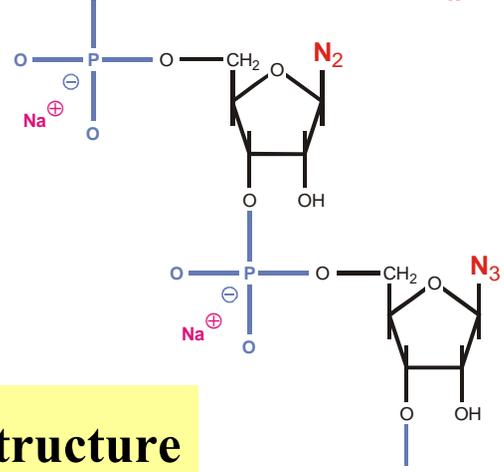2. Neutrality and RNA optimization

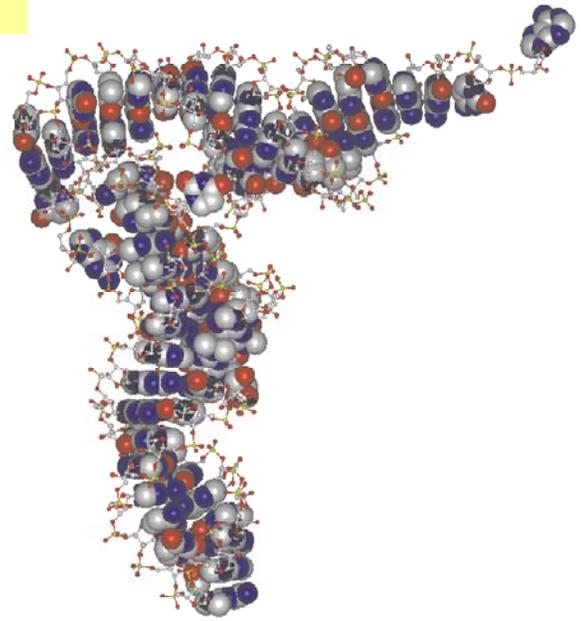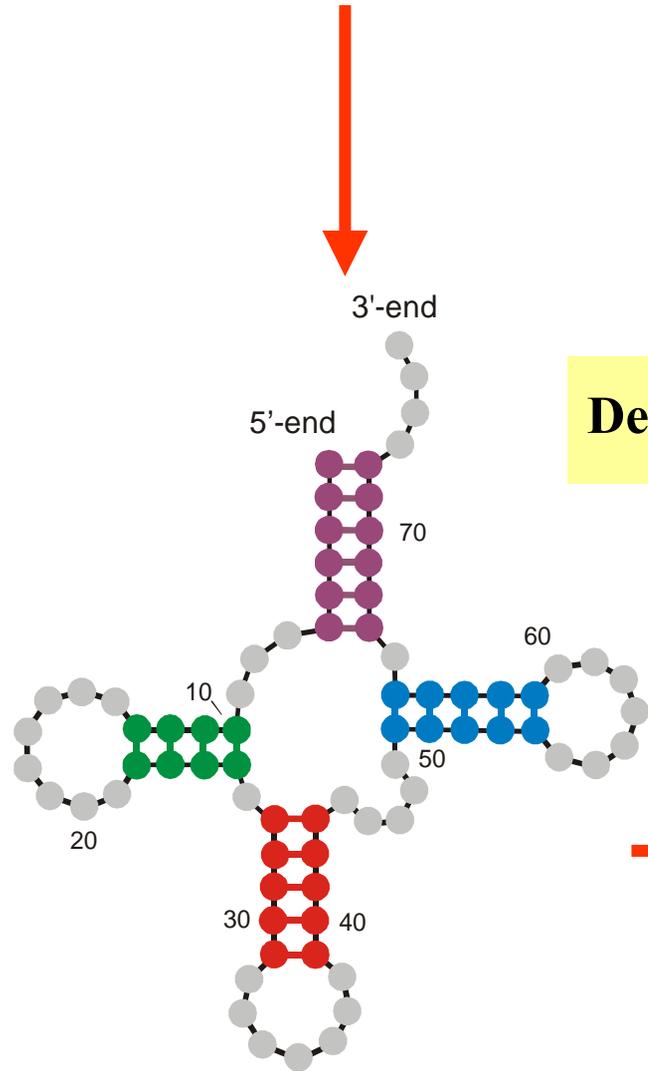3. Kinetic folding of RNA

4. RNA switches and their design

1. **RNA structures**

2. Neutrality and RNA optimization

3. Kinetic folding of RNA

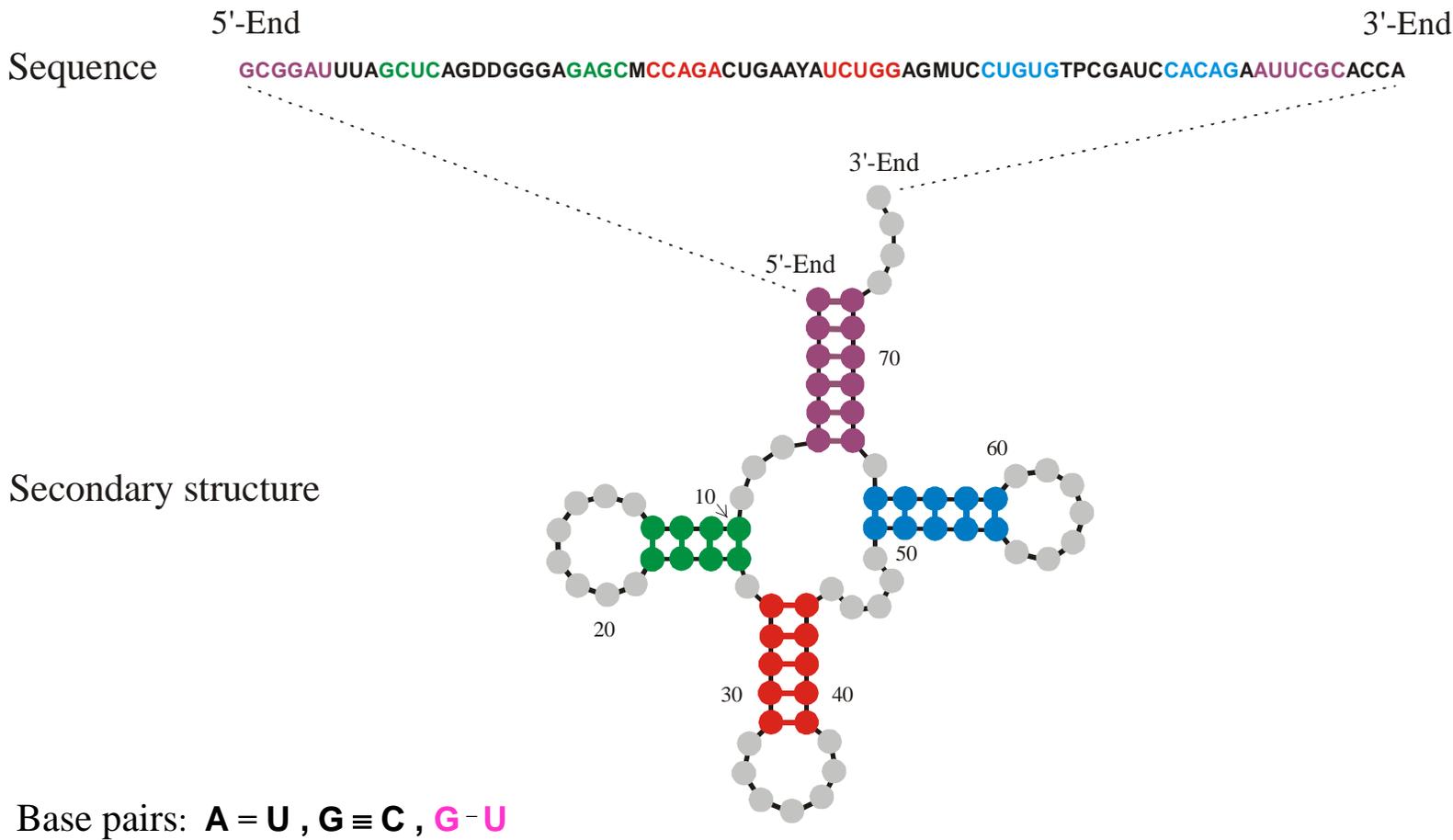4. RNA switches and their design

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

Definition of RNA structure

5' - end

N₁

N₂

N₃

3'-end

5'-end

70

60

10

50

20

30   40

Sequence

5'-End
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA
3'-End

Secondary structure

Base pairs: $A = U$, $G \equiv C$, $G^- U$

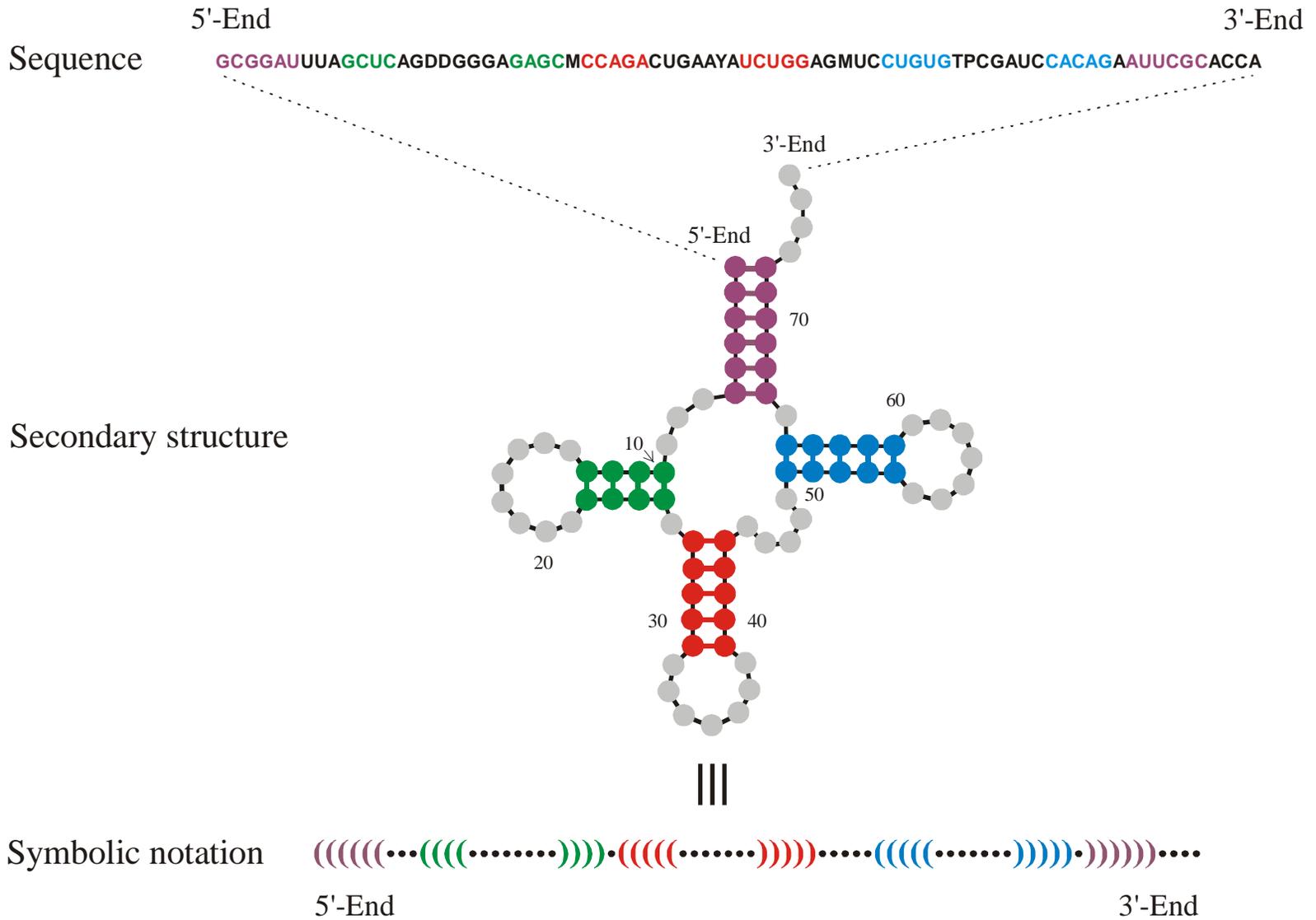A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs

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.

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

How to compute RNA secondary structures

Efficient algorithms based on **dynamic programming** are available for computation of minimum free energy and **many** suboptimal secondary structures for given sequences.

M.Zuker and P.Stiegler. *Nucleic Acids Res*. **9**:133-148 (1981)

M.Zuker, *Science* **244**: 48-52 (1989)

Equilibrium partition function and base pairing probabilities in Boltzmann ensembles of suboptimal structures.

J.S.McCaskill. *Biopolymers* **29**:1105-1190 (1990)

The **Vienna RNA Package** provides in addition: **inverse folding** (computing sequences for given secondary structures), computation of melting profiles from partition functions, **all suboptimal structures** within a given energy interval, barrier tress of suboptimal structures, **kinetic folding** of RNA sequences, RNA-hybridization and RNA/DNA-hybridization through **cofolding** and **kinetic cofolding** of two and more sequences, simultaneous alignment of sequences and structures, etc..

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

S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. *Biopolymers* **49**:145-165 (1999)

C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. *RNA* **6**:325-338 (1999)

Vienna RNA Package: http://www.tbi.univie.ac.at
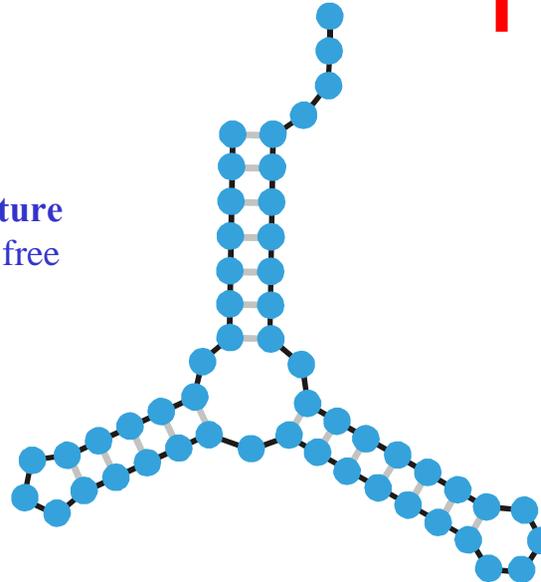
Sequence, structure, and design

5'-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

3'-end

Free energy ΔG^0

$S_5^{(h)}$

$S_3^{(h)}$

$S_4^{(h)}$

$S_1^{(h)}$

$S_2^{(h)}$

$S_8^{(h)}$

$S_7^{(h)}$

$S_9^{(h)}$

$S_6^{(h)}$

Suboptimal conformations

$S_0^{(h)}$

Minimum of free energy
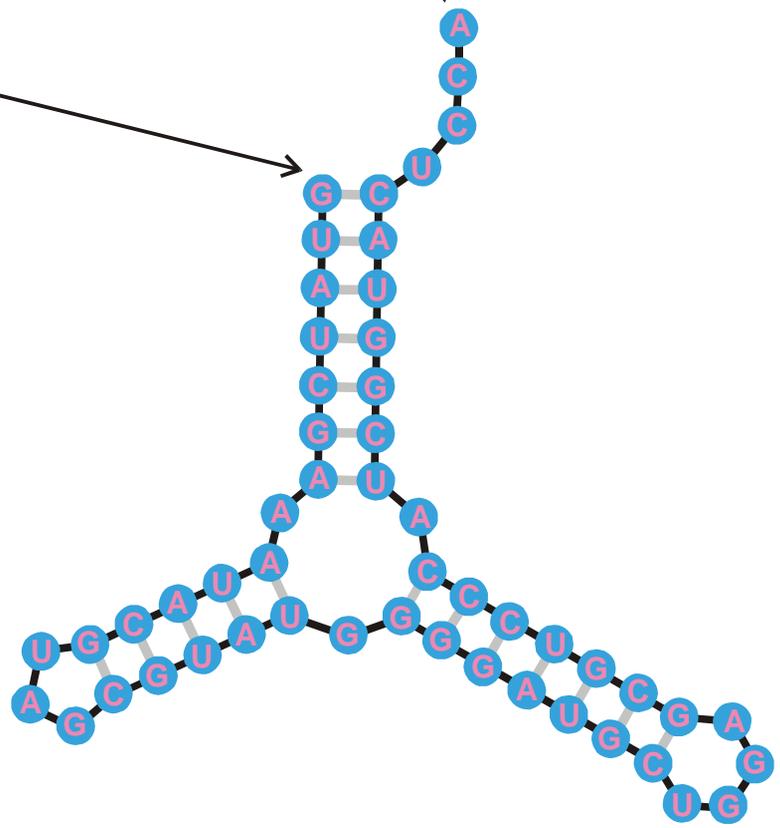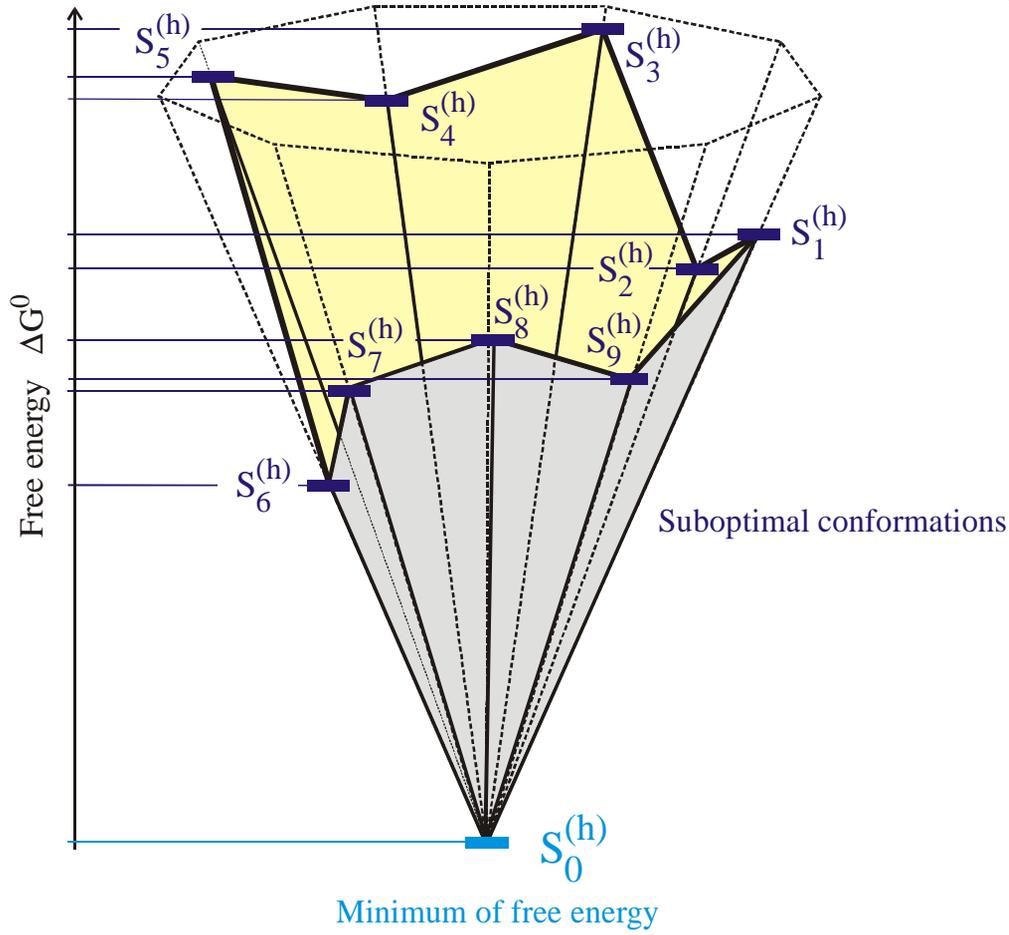
The minimum free energy structures on a discrete space of conformations

S_{10}
S_9 S_8
S_7
S_5 S_6
S_4
S_3

S_2

S_1
S_0
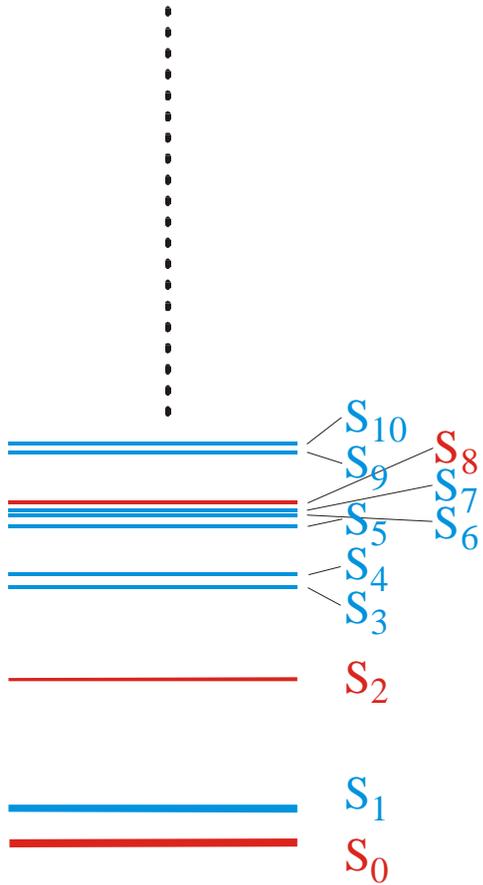
Suboptimal structures

Suboptimal secondary structures of an RNA sequence
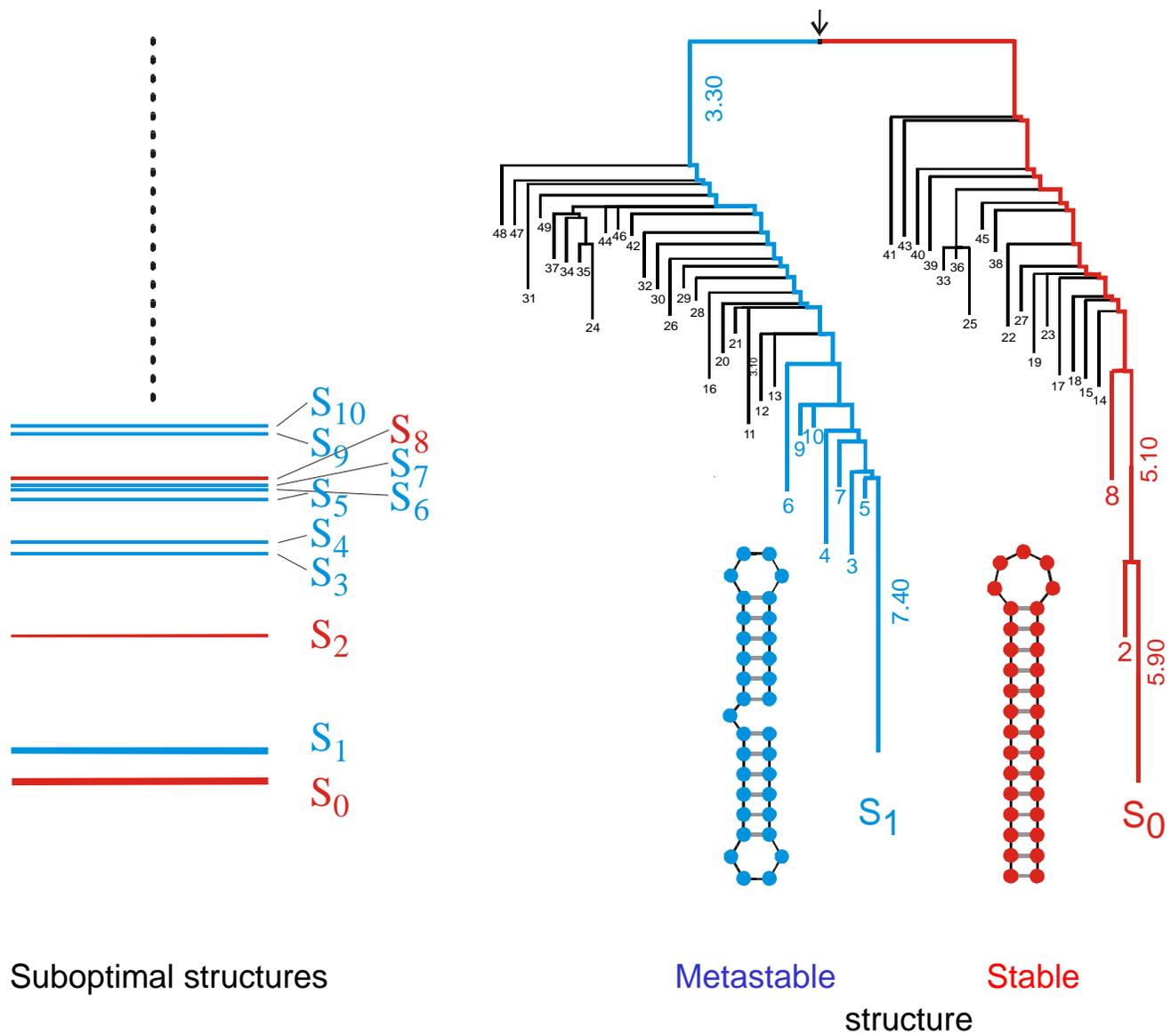
An RNA molecule with two (**meta**)**stable** conformations
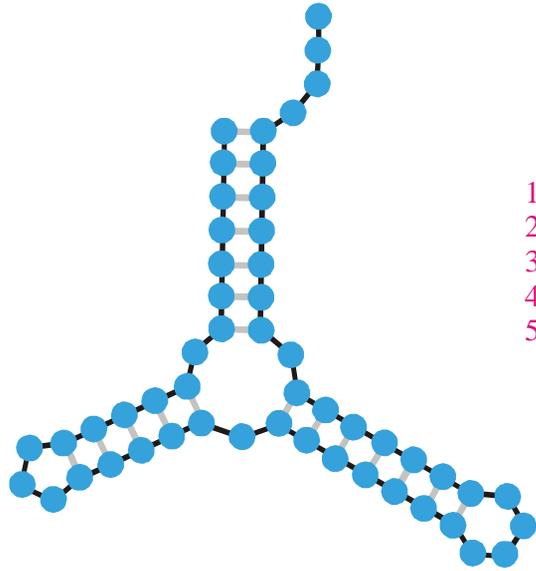
Minimum free energy criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

Inverse folding of RNA secondary structures

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

One-error neighborhood

The surrounding of **GUCAAUCAG** in sequence space

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG   GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**C**UGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space
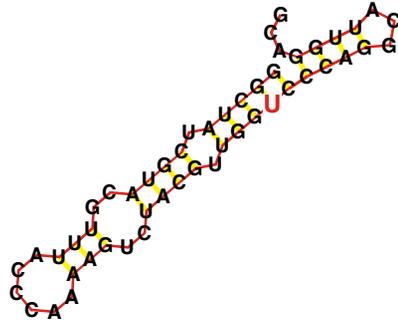
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG
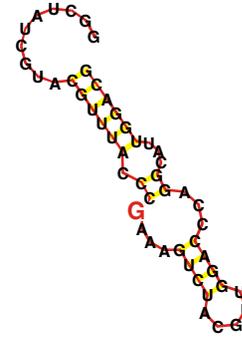
GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
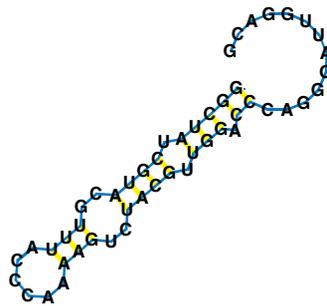
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**C**UGGACG
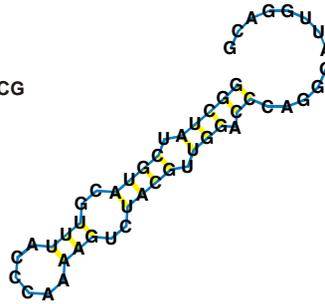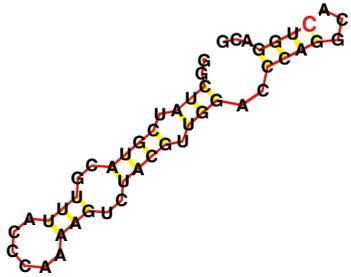
GGCUAUCGUACGU**G**UACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule in sequence and shape space

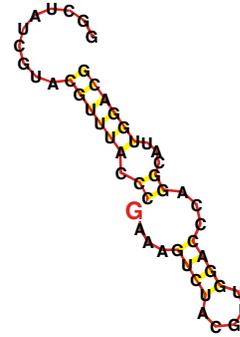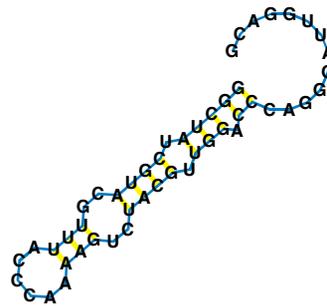GGCUAUCGUA**U**GUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUU**A**GACG

GGCUAUCGUACGUUUAC**U**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACG**C**UUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGC**C**AUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGU**G**UACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUA**A**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCC**U**GGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**C**UGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG

GGCUA**G**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAG**C**CUACGUUGGACCCAGGCAUUGGACG
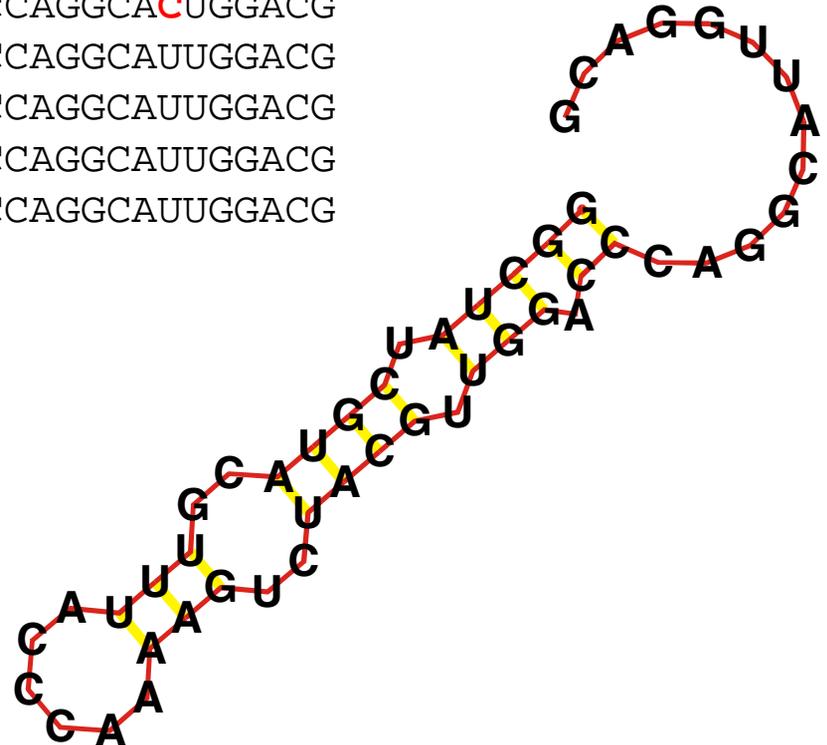
One error neighborhood – Surrounding of an RNA molecule
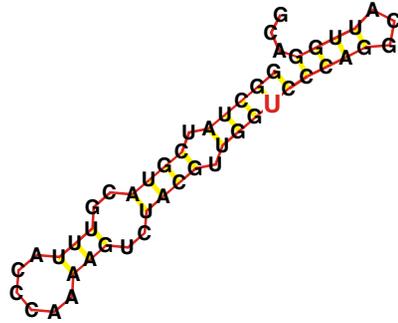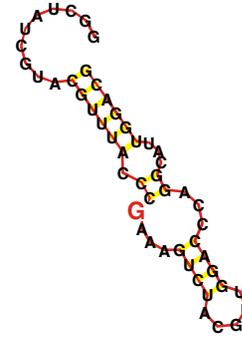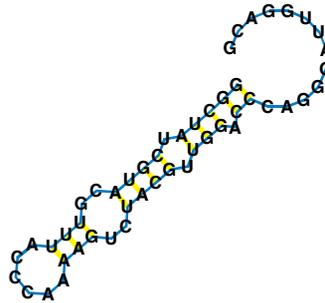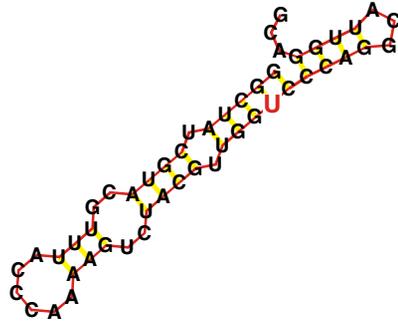in sequence and shape space

Mount Fuji

The sourroundings of RNA sequences in sequence space **do not** resemble segments of **smooth landscapes**. They are rather indicating a very high degree of **ruggedness** ...

Example of a smooth landscape on Earth

Dolomites
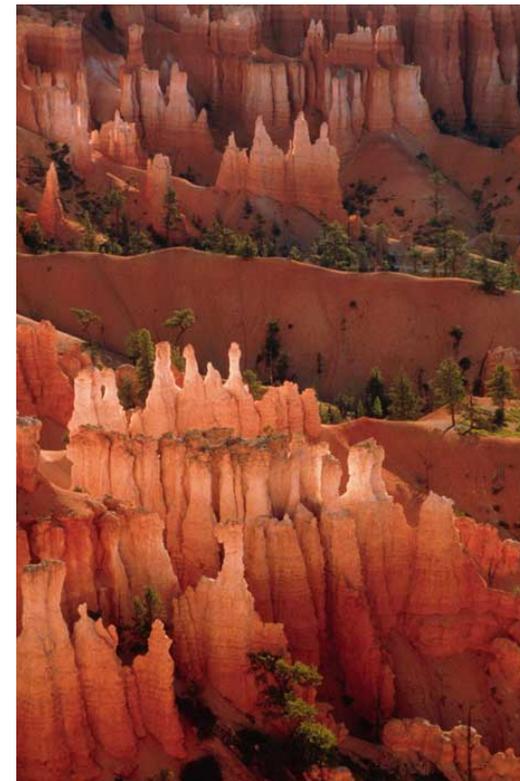
Bryce Canyon

Examples of rugged landscapes on Earth

	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:	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

	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

Computer simulation of RNA optimization

Walter Fontana and Peter Schuster,
Biophysical Chemistry 26:123-147, 1987

Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

A computer model of evolutionary optimization

Walter Fontana and Peter Schuster

Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

Molecular evolution is viewed as a typical combinatorial optimization problem. We analyse a chemical reaction model which considers RNA replication including correct copying and point mutations together with hydrolytic degradation and the dilution flux of a flow reactor. The corresponding stochastic reaction network is implemented on a computer in order to investigate some basic features of evolutionary optimization dynamics. Characteristic features of real molecular systems are mimicked by folding binary sequences into unknotted two-dimensional structures. Selective values are derived from these molecular 'phenotypes' by an evaluation procedure which assigns numerical values to different elements of the secondary structure. The fitness function obtained thereby contains nontrivial long-range interactions which are typical for real systems. The fitness landscape also reveals quite involved and bizarre local topologies which we consider also representative of polynucleotide replication in actually occurring systems. Optimization operates on an ensemble of sequences via mutation and natural selection. The strategy observed in the simulation experiments is fairly general and resembles closely a heuristic widely applied in operations research areas. Despite the relative smallness of the system – we study 2000 molecules of chain length $\nu = 70$ in a typical simulation experiment – features typical for the evolution of real populations are observed as there are error thresholds for replication, evolutionary steps and quasistationary sequence distributions. The relative importance of selectively neutral or almost neutral variants is discussed quantitatively. Four characteristic ensemble properties, entropy of the distribution, ensemble correlation, mean Hamming distance and diversity of the population, are computed and checked for their sensitivity in recording major optimization events during the simulation.

1. Molecular evolution and optimization

Conventional population genetics treats mutation as an external stochastic source. Moreover, mutations are considered as very rare events. In the absence of genetic recombination populations of haploid organisms are expected to be usually homogeneous. Experimental evidence on viral and bacterial populations is available now and it contradicts these expectations. Mutations appear much more frequently than was originally assumed.

The molecular approach considers error-free replication and mutation as parallel reactions within the same mechanism. Detailed information on the molecular mechanisms of polynucleotide replication provides direct insight into the nature of mutations and their role in evolution. Several classes of mutations are properly distinguished: point mutations, deletions and insertions. Point mutations are of special importance: they represent the most frequent mutations and are easily incorporated into theoretical models of molecular evolution. This does not mean, however, that the other classes of mutations are not important in evolution. To give an example: there is a general belief that insertions leading to gene duplication played a major role in the development of present day enzyme families.
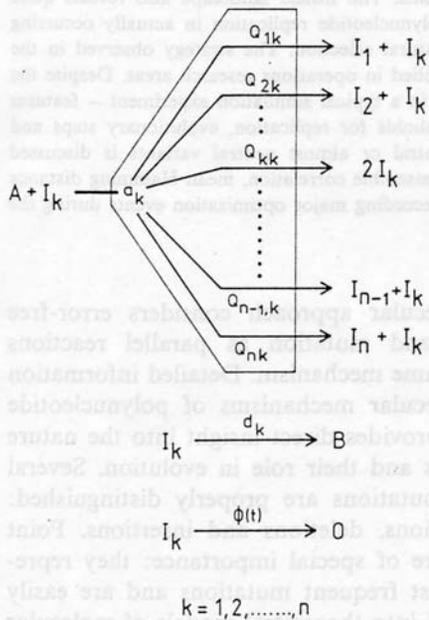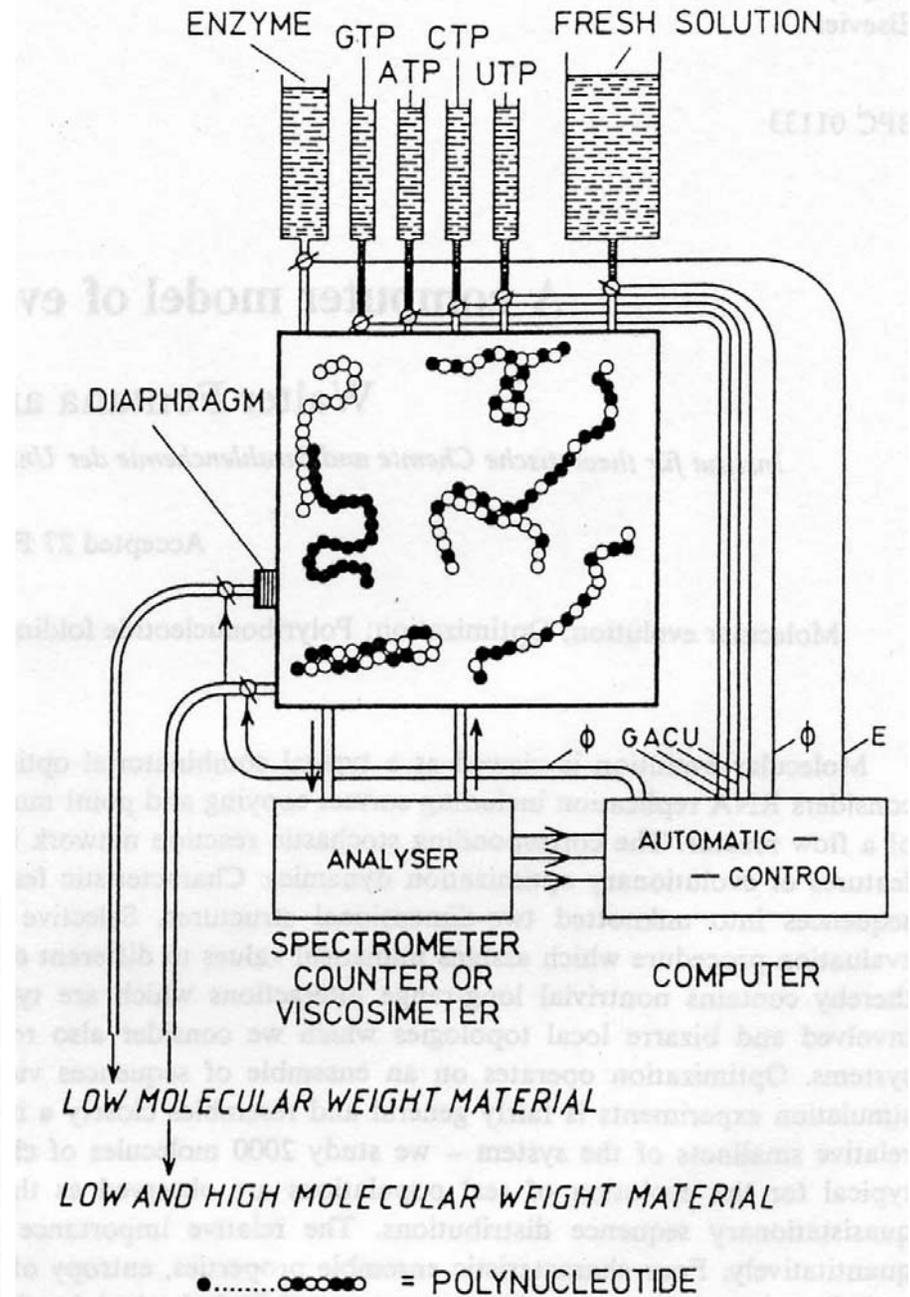
The first theoretical model of molecular evolu-

Dedicated to Professor Manfred Eigen on the occasion of his 60th birthday.
Correspondence address: P. Schuster, Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria.

Physical aspects of evolutionary optimization and adaptation

Walter Fontana, Wolfgang Schnabl, and Peter Schuster*
Institut für Theoretische Chemie der Universität Wien, Währingerstrasse 17, A 1090 Wien, Austria
(Received 2 February 1989; revised manuscript received 5 May 1989)

A model of an objective function based on polynucleotide folding is used to investigate the dynamics of evolutionary adaptation in finite populations. Binary sequences are optimized with respect to their kinetic properties through a stochastic process involving mutation and selection. The objective function consists in a mapping from the set of all binary strings with given length into a set of two-dimensional structures. These structures then encode the kinetic properties, expressed in terms of parameters of reaction probability distributions. The objective function obtained thereby represents a realistic example of a highly "rugged landscape." Ensembles of molecular strings adapting to this landscape are studied by tracing their escape path from local optima and by applying multivariate analysis. Effects of small population numbers in the tail of the sequence distribution are discussed quantitatively. Close upper bounds to the number of distinct values produced by our objective function are given. The distribution of values is explored by means of simulated annealing and reveals a random scatter in the locations of optima in the space of all sequences. The genetic optimization protocol is applied to the "traveling salesman" problem.

Fig. 1. The reaction network. Synthesis on template I_k proceeds with the rate constant a_k and leads with frequency Q_{ik} to a new template I_i preserving the old copy. Materials A needed for polymerization are assumed to be buffered. Degradation to waste products B occurs with rate d_k and a controlled unspecific flux $\Phi(t)$ removes templates from the system.

Fig. 2. The evolution reactor. This kind of flow reactor consists of a reaction vessel which allows for temperature and pressure control. Its walls are impermeable to polynucleotides. Energy-rich material is poured from the environment into the reactor. The degradation products are removed steadily. Material transport is adjusted in such a way that the concentration of monomers is constant in the reactor. A dilution flux Φ is installed in order to remove excess of polynucleotides produced by replication. Thus, the sum of the numbers of individual particles $\sum_i X_i(t) = N(t)$ may be controlled by the flux Φ. Under 'constant organization' Φ is adjusted such that $N(t) = \Theta$ is essentially constant. By this we indicate that fluctuations with standard deviation $\sigma = \sqrt{N}$ occur regularly. The regulation of Φ requires internal control, which can be achieved by logistic coupling.

$...\ 0000000000000000000000000000000\ ...$ (d)

Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

Evolution *in silico*

W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455

random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCT-CATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGT-TCCACC-3′ (reverse). Reactions were performed in 25 µl using 1 unit of Taq DNA polymerase with each primer at 0.4 µM; 200 µM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl₂,1.5 mM MgCl₂] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C at 72°C followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].

33. Data not shown; a dot blot with poly (A)⁺ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (*13*).

34. Smith–Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [(*6*); K-S Chen, L. Potocki, J. R. Lupski, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.

35. R. A. Fridell, data not shown.

36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.

37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)⁺ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGACCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengkala, Bali, and the two families from India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

To distinguish continuous from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. This nearness relation is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and provides a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much-debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (*1*). Our goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicatable sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (*3*, *4*).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (*5*, *6*). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (*4*). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection criterion. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces the average distance to the target (inversely related to fitness) in the population against time. Aside from a short initial phase, the entire history is dominated by steps, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target structure (*7*). However, the dominant shapes in the population not only change at these marked events but undergo several fitness-neutral transformations during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive phenotypes itself can be called continuous or discontinuous.

A set of entities is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The issue arises because, in contrast to sequences, there are

Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.

Stock Solution ⟶   Reaction Mixture ⟶
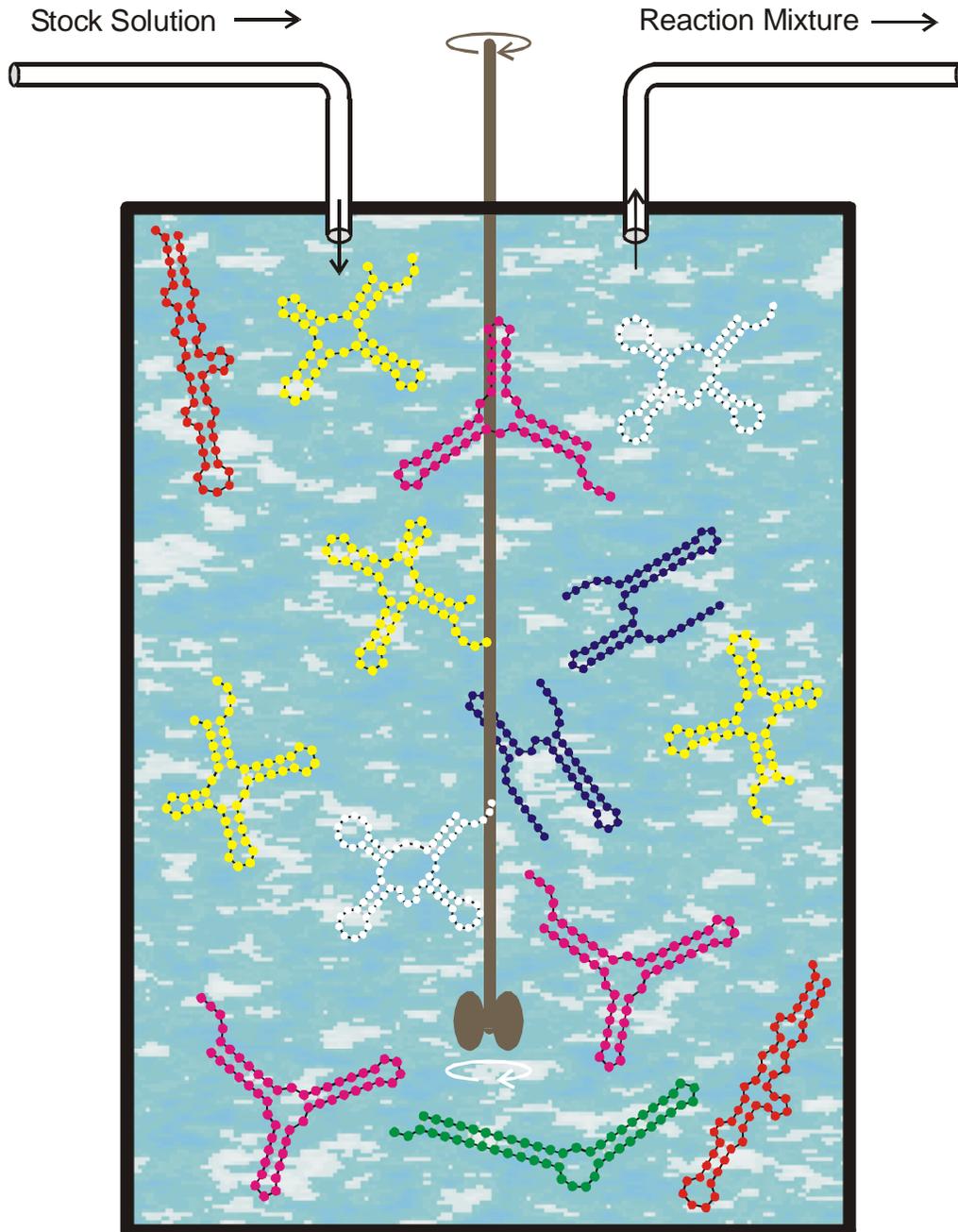
Replication rate constant:

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

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

Selection constraint:

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

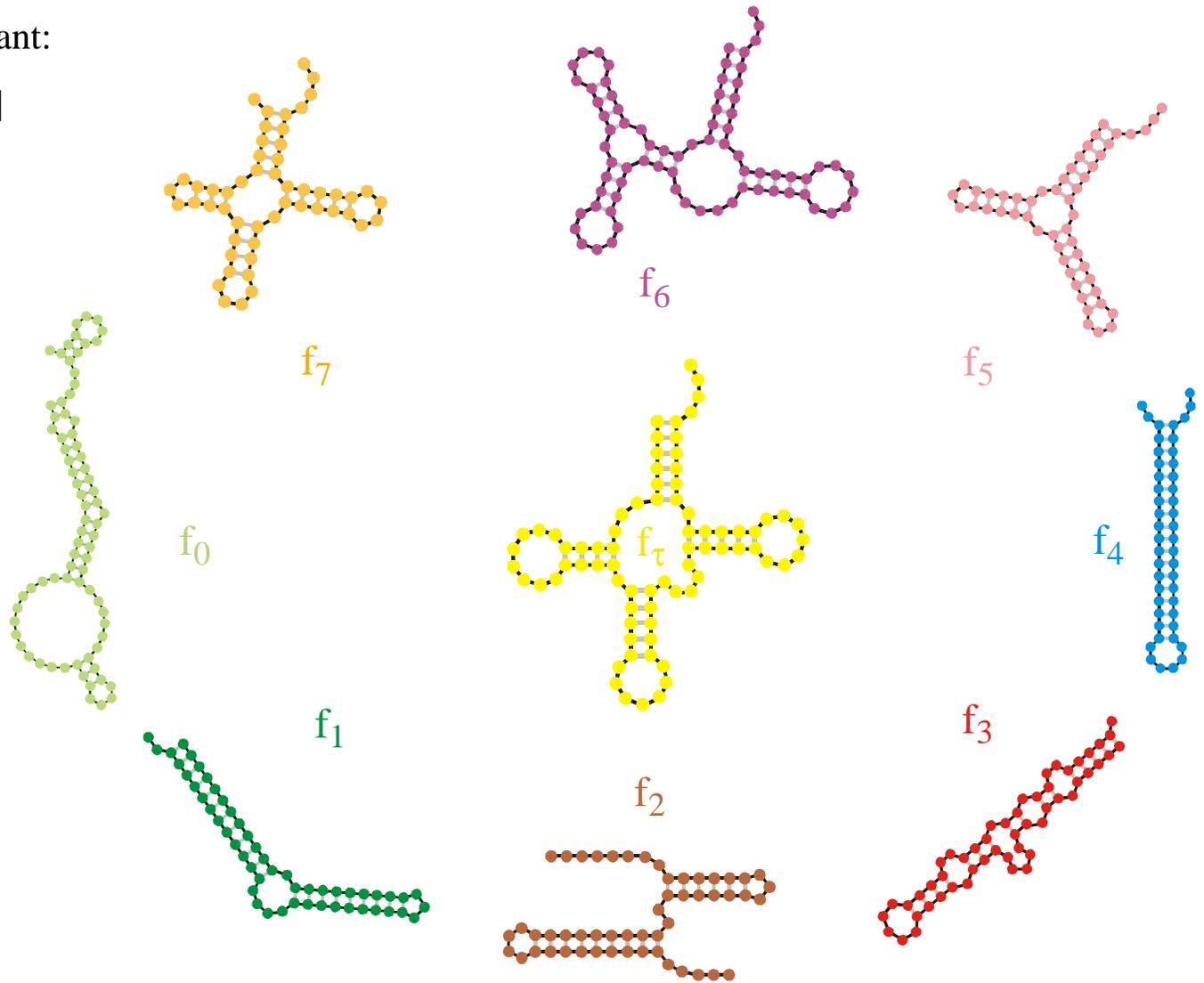$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

Mutation rate:

$p = 0.001 \,/\, \text{site} \times \text{replication}$

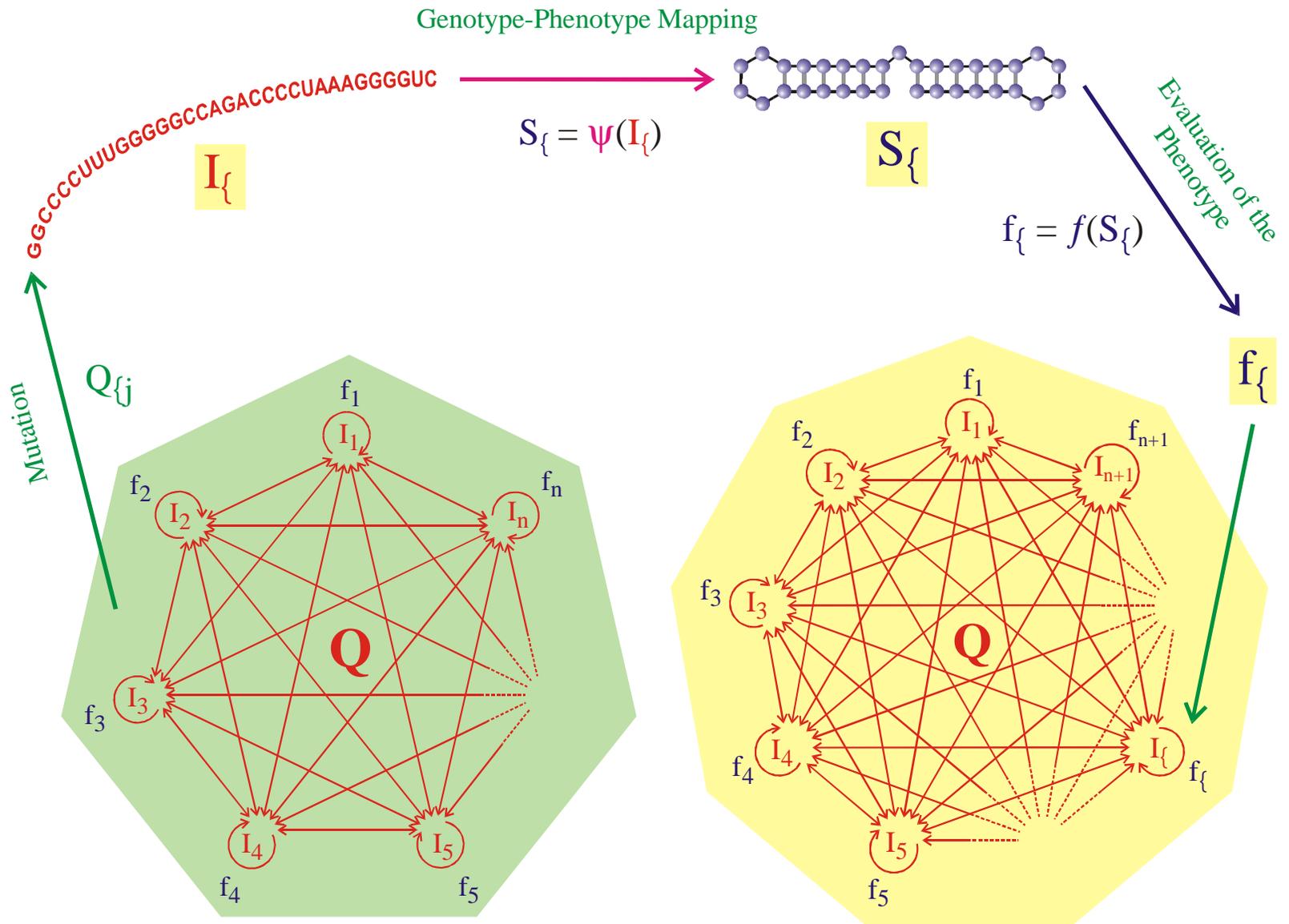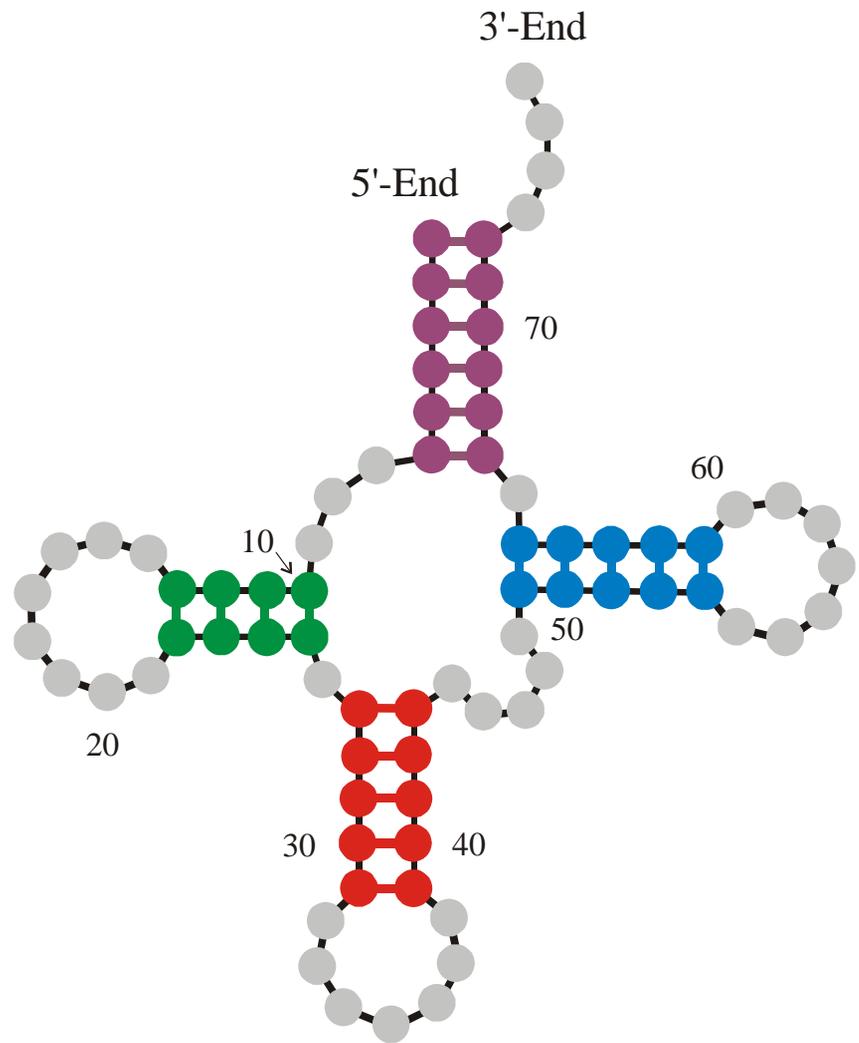The flowreactor as a device for **studies** of evolution ***in vitro*** and ***in silico***

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

Genotype-Phenotype Mapping

$S_\{ = \psi(I_\{)$

GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC

$I_\{$

$S_\{$

Evaluation of the Phenotype

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

$f_\{$

$Q_{\{j}$

Mutation

f$_1$

I_1

f$_2$ I_2 f$_n$ I_n

Q

f$_3$ I_3

f$_4$ I_4 f$_5$ I_5

f$_1$ I_1

f$_2$ I_2 f$_{n+1}$ I_{n+1}

f$_3$ I_3

Q

f$_4$ I_4 $I_\{$ f$_\{$

f$_5$ I_5

f$_5$

Evolutionary dynamics
including molecular phenotypes
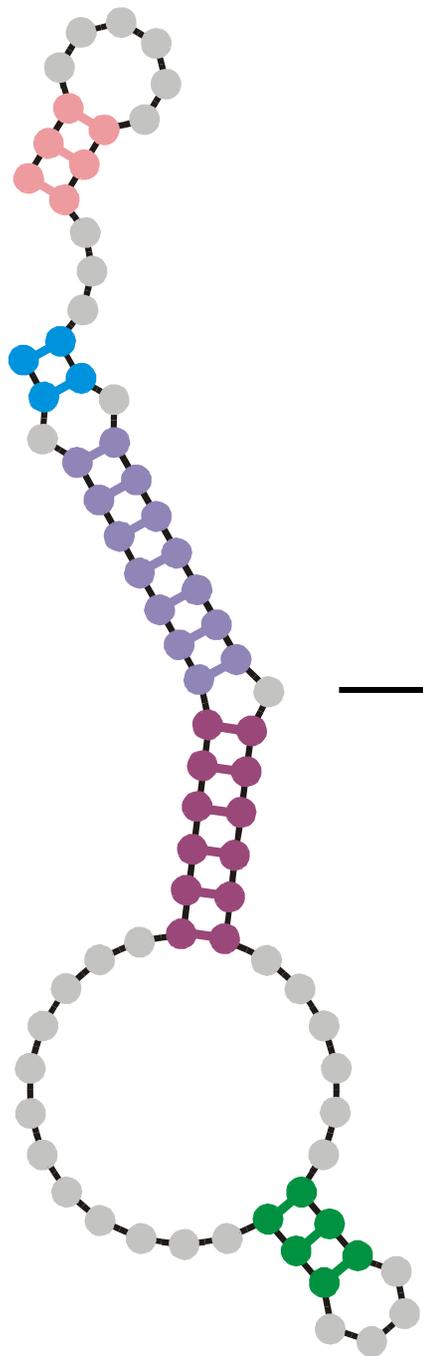
Randomly chosen
initial structure

Phenylalanyl-tRNA as
target structure

Axis labels: Average distance from initial structure $50 - \Delta d_S$ (vertical), Time (arbitrary units) (horizontal)

Annotation: Evolutionary trajectory
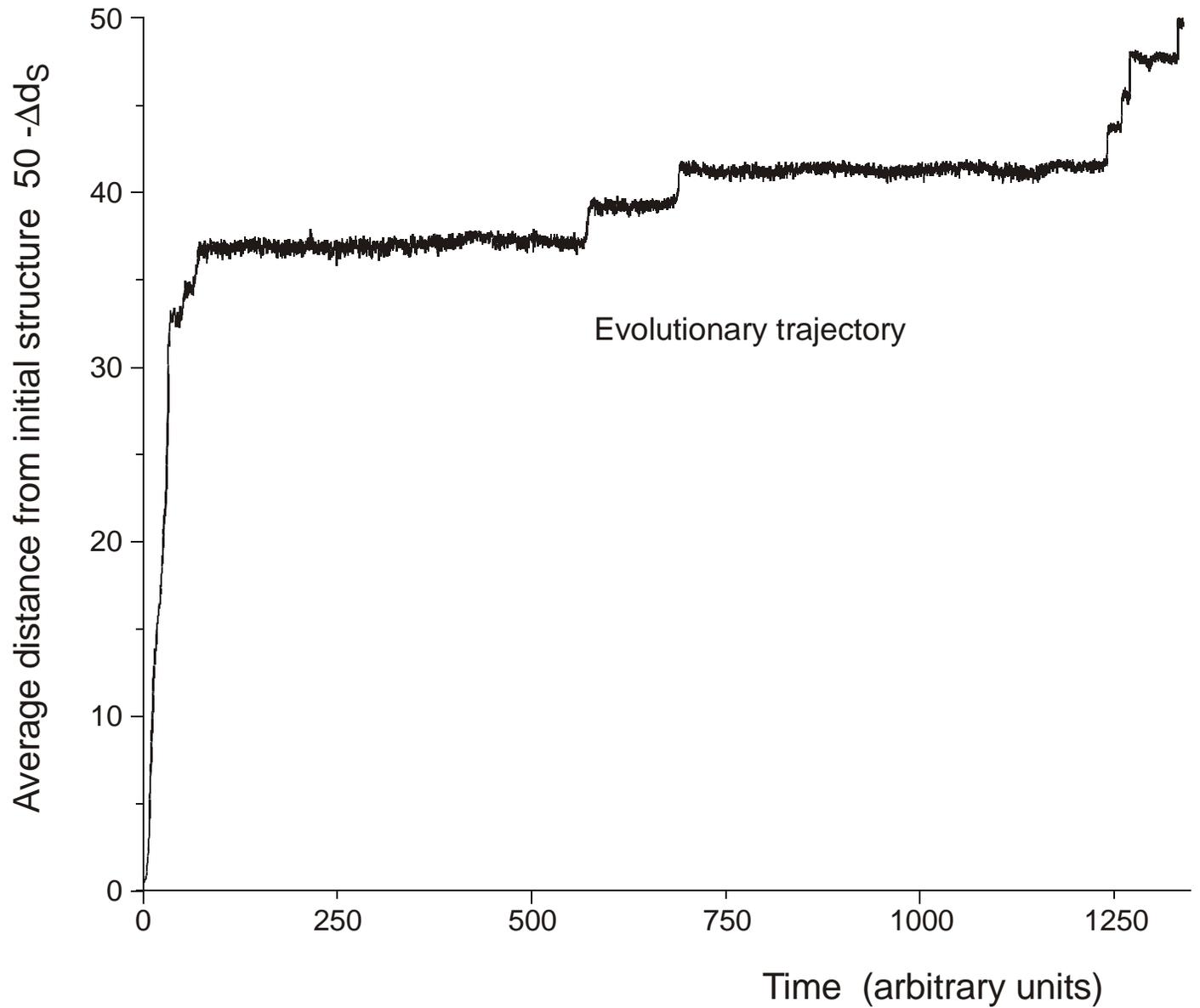
In silico optimization in the flow reactor: Trajectory (**biologists' view**)
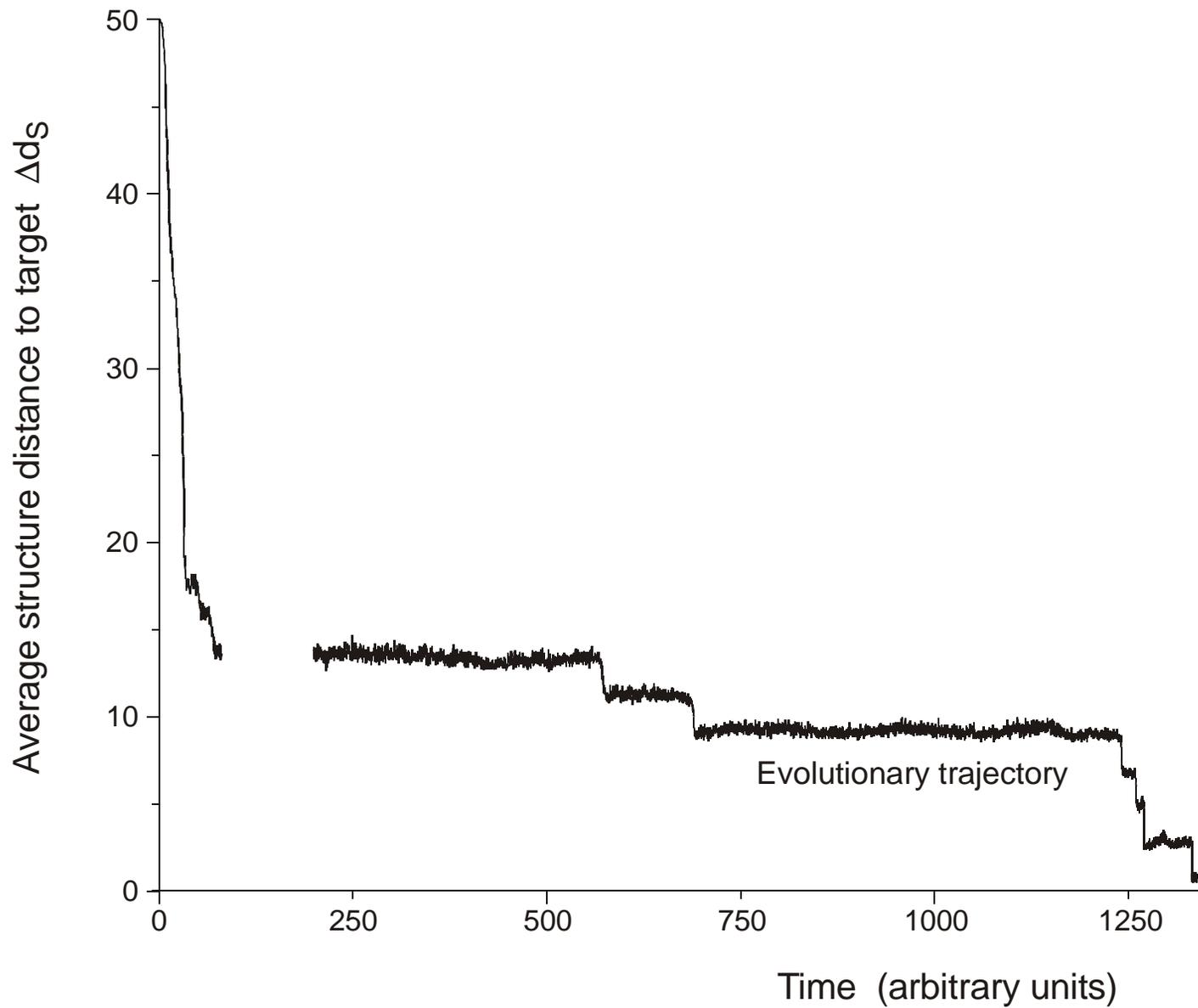
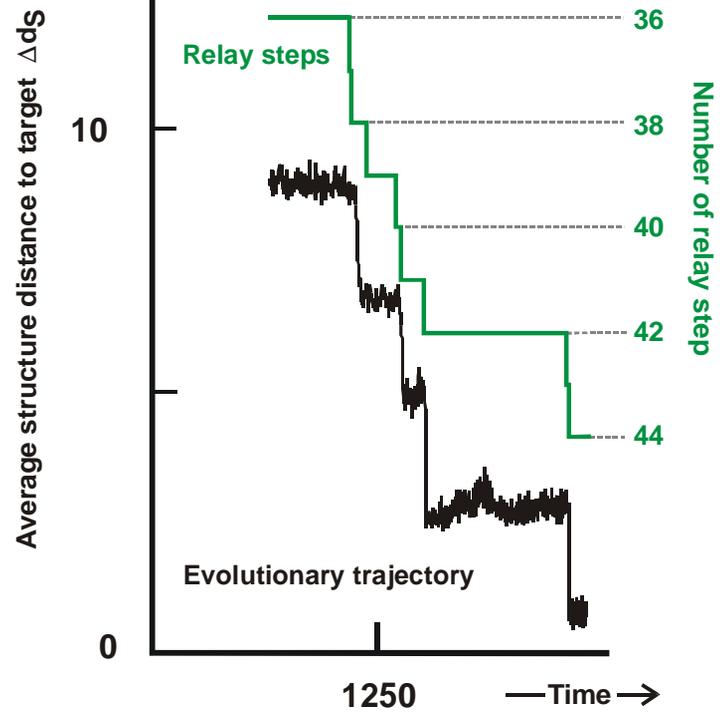*In silico* optimization in the flow reactor: Trajectory (**physicists' view**)

Average structure distance to target ΔdS

Relay steps

36
38
40
42
44

Number of relay step

10

0

Evolutionary trajectory

1250 — Time →
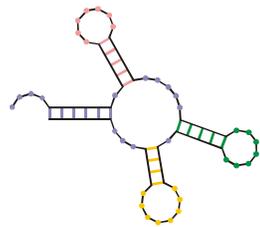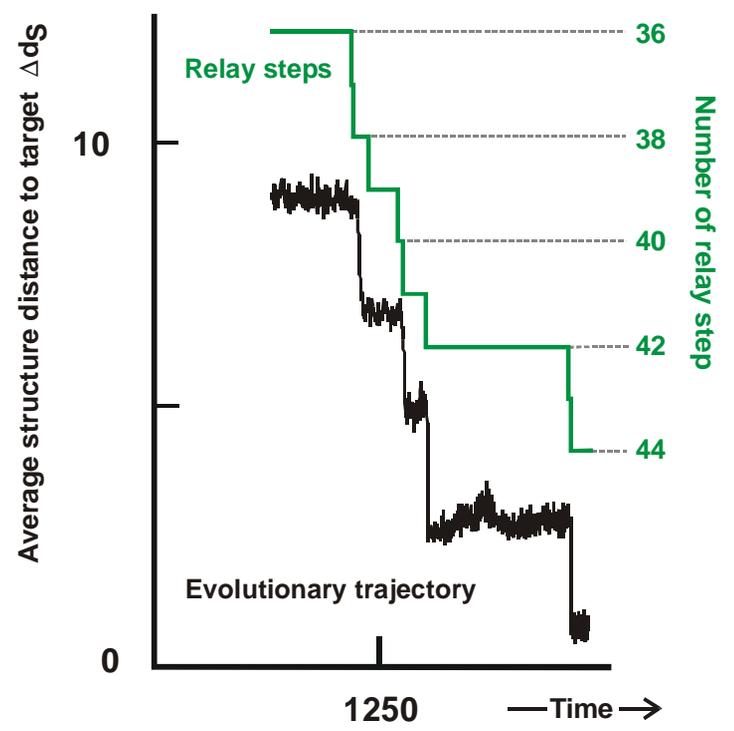
44

Endconformation of optimization
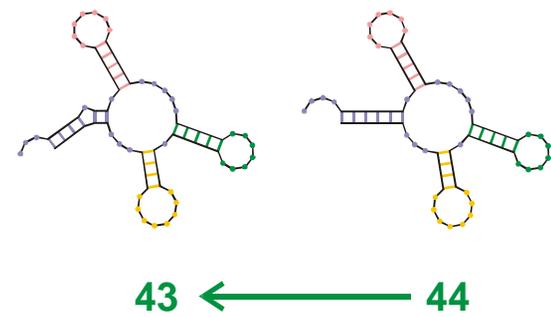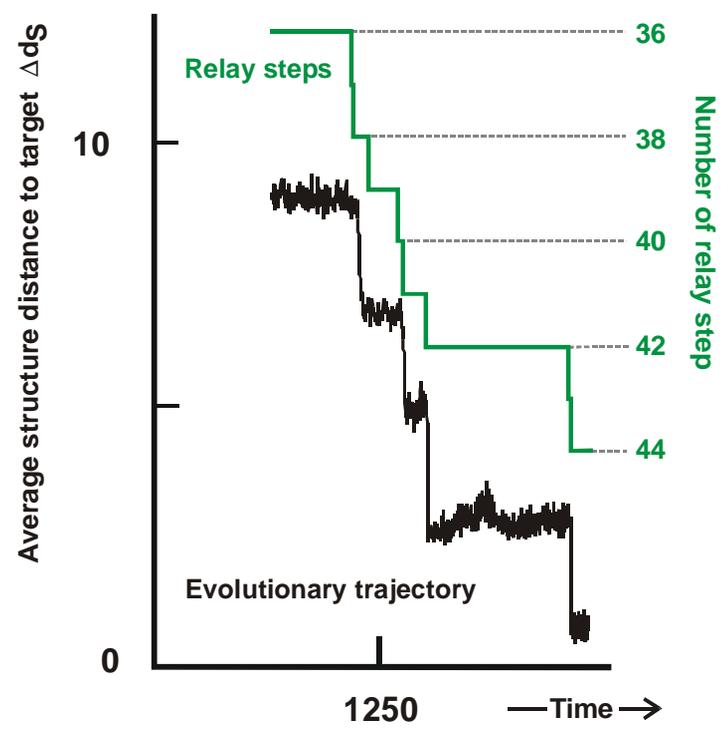
Reconstruction of the last step 43 → 44

Average structure distance to target Δd_S

Number of relay step

36

Relay steps

10

38

40

42

44

Evolutionary trajectory

0

1250 ——Time →
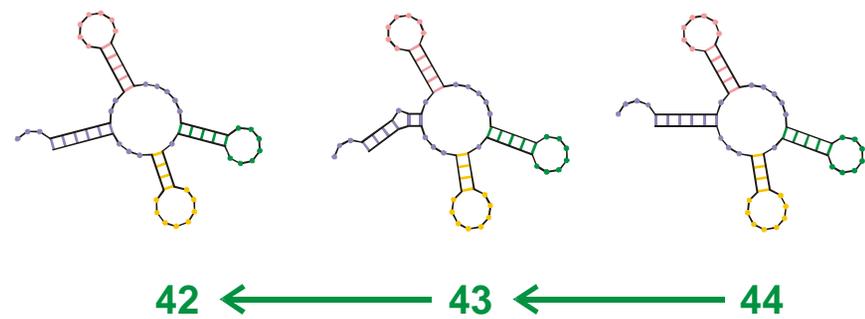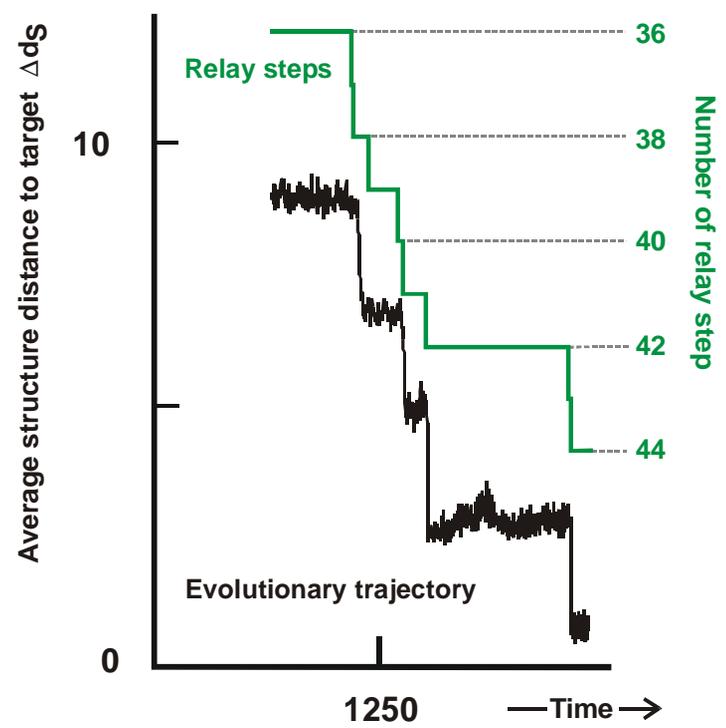
42 ⟵ 43 ⟵ 44

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

Average structure distance to target ΔdS

Number of relay step

Relay steps

36
38
40
42
44

10

Evolutionary trajectory

0

1250 — Time →

Evolutionary process

39 ← 40 ← 41 ← 42 ← 43 ← 44
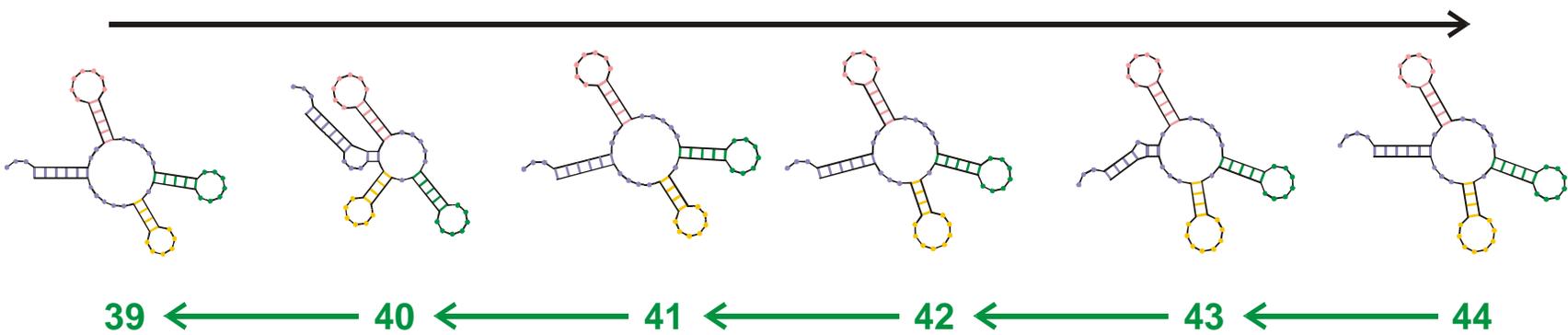
Reconstruction

Reconstruction of the relay series

entry

GGGAUACAUGUGGCCCCUCAAGGCCCUAGCGAAACUGCUGCUGAAACCGUGUGAAUAAUCCGCACCCUGUCCCCGA

39

((((((.....((((......))))).(((((......))))).....((((((......))))).)))))...

exit

GGGAUAUACGAGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG

entry

GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGG

40

((((((...(((((.....)))).(((((......))))).....((((((......))))))))))))...

exit

GGGAUAUACGGGGCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG

entry

GGGAUAUACGGGGCCCCGUCAAGGCCGUAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCGGG

41

((((((....((((......))))).(((((......))))).....((((((......)))))).)))))...

exit

GGGAUAUACGGGCCCCCUUCAAGGCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA

entry

GGGAUAUACGGGCCCCCUUCAAGCCCAUAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCGGA

42

((((((...((((........)))).(((((......))))).....((((((......))))).)))))...

exit

GGGAUGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU

entry

GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU

43

((((((...((((........)))).(((((......))))).....((((((......))))).)).)))))...

exit

GGGAAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU

entry

GGGCAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAGCUUGUGCGACGUUUGUGCACCCUGUCCCGCU

44

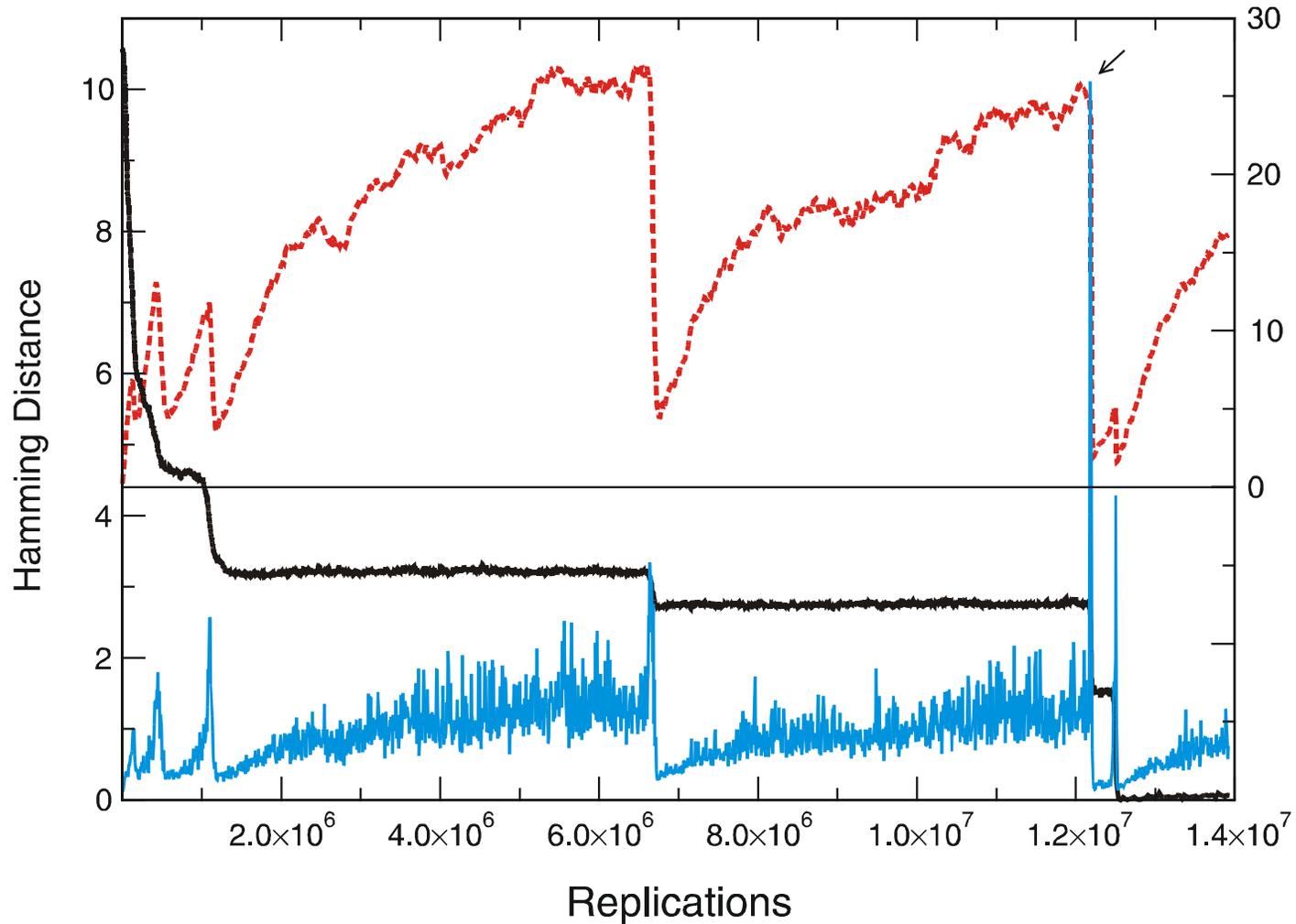((((((...((((........)))).(((((......))))).....((((((......))))).)))))....

Transition inducing point mutations   **Neutral point mutations**

Change in RNA sequences during the final five relay steps 39 → 44
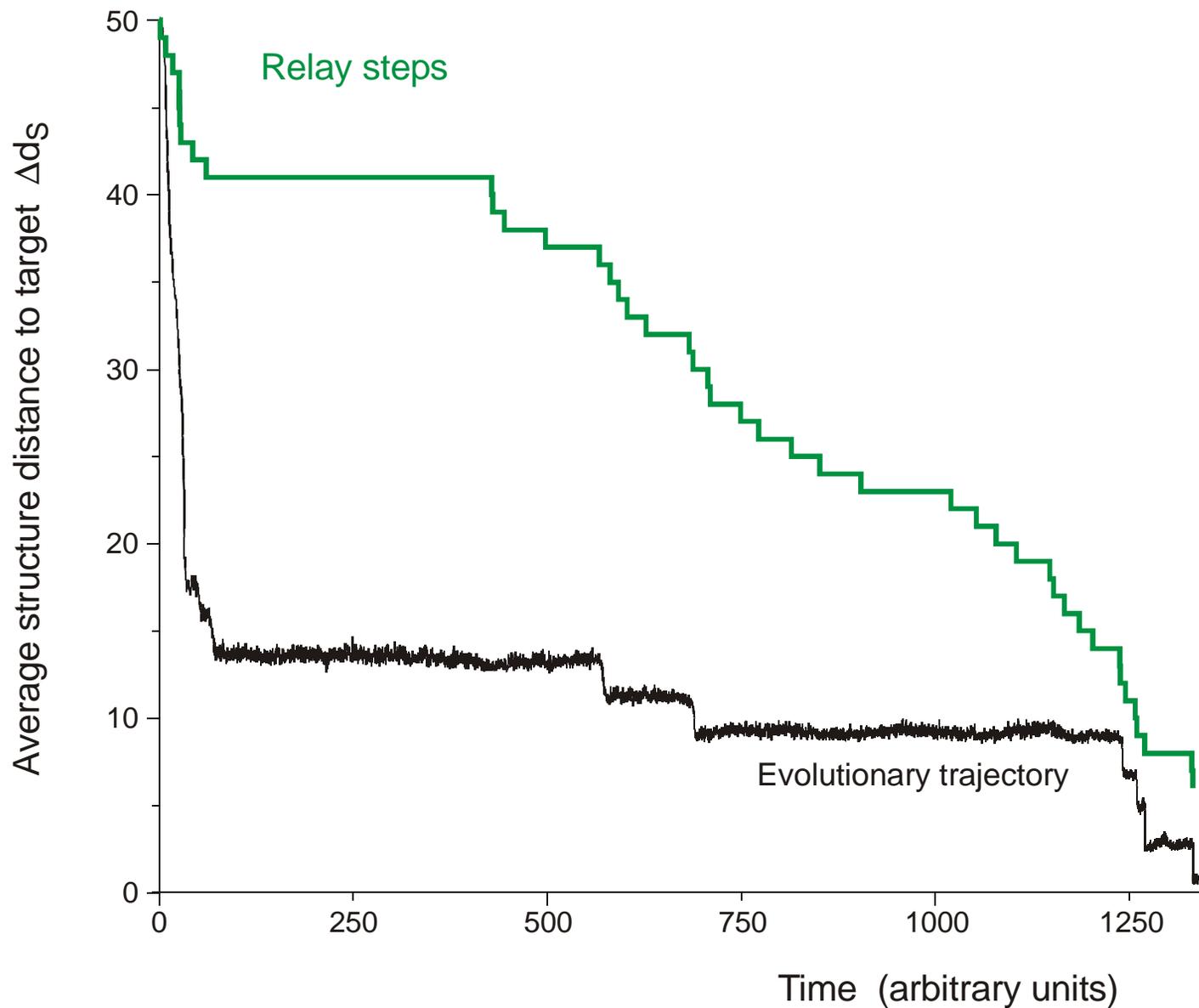
*In silico* optimization in the flow reactor: Trajectory and relay steps

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   **Neutral point mutations**

Neutral genotype evolution during phenotypic stasis

Variation in genotype space during optimization of phenotypes

Mean Hamming distance within the population and **drift velocity of the population center** in sequence space.
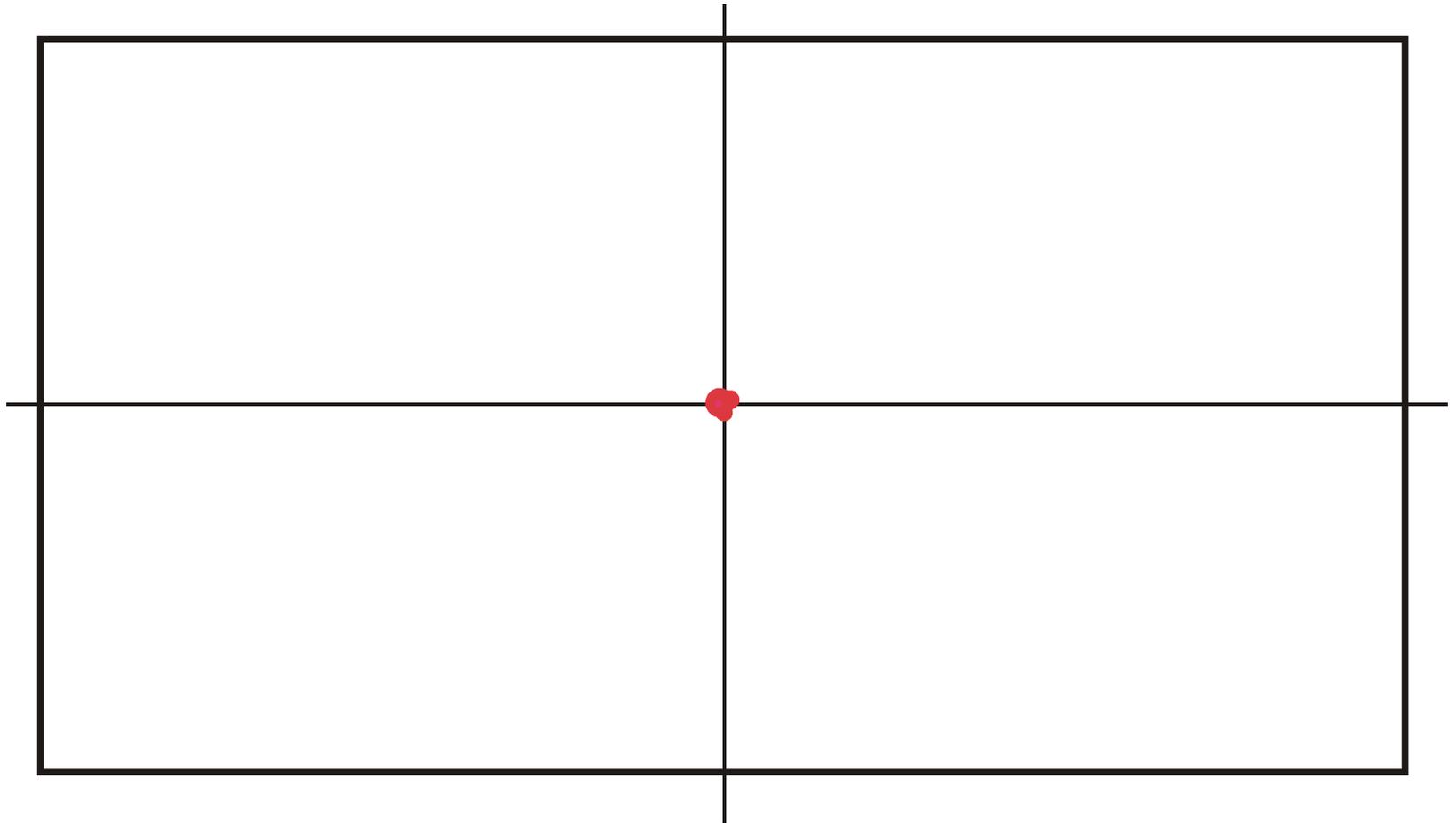
Spread of population in sequence space during a quasistationary epoch: t = 150
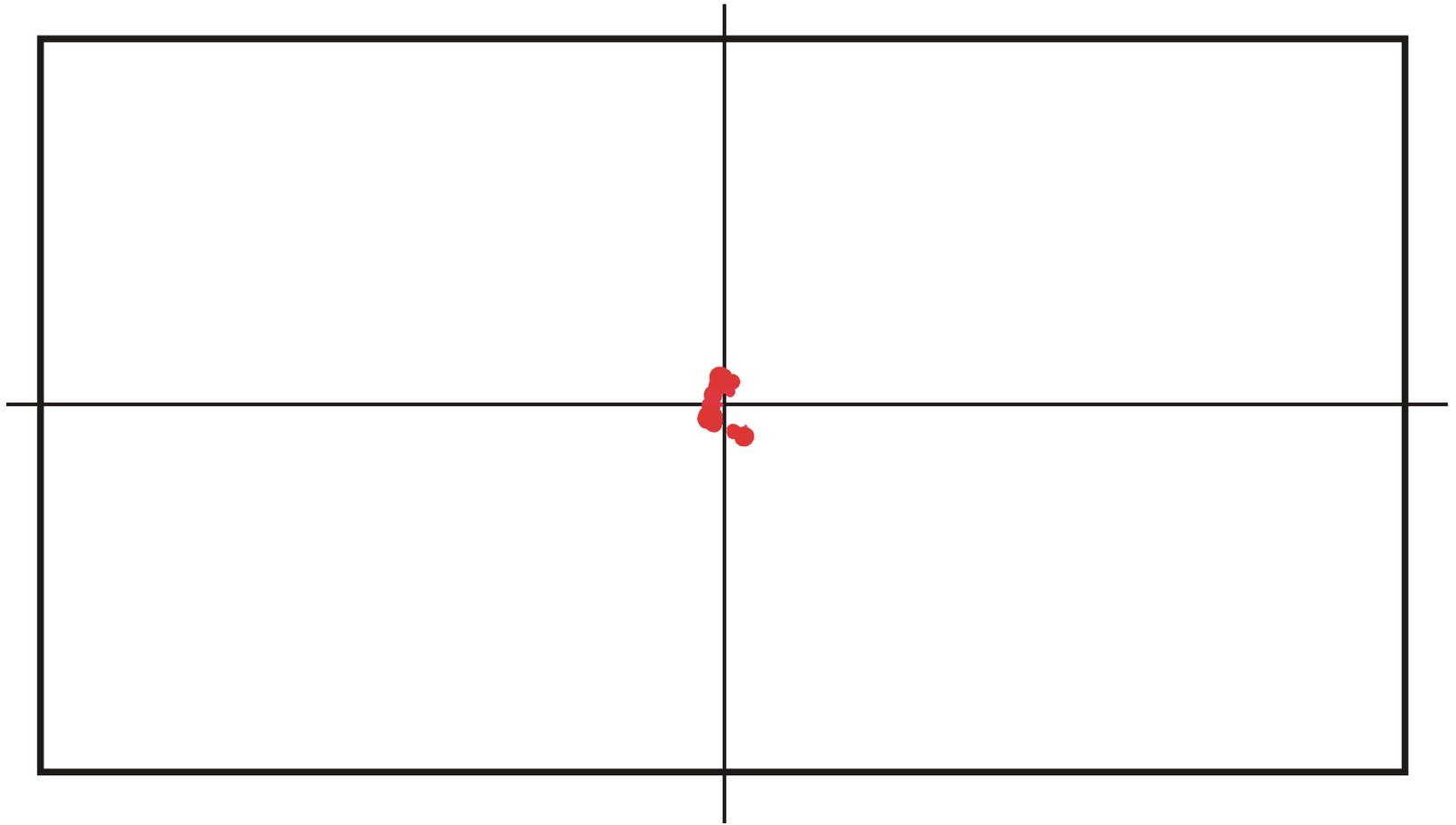
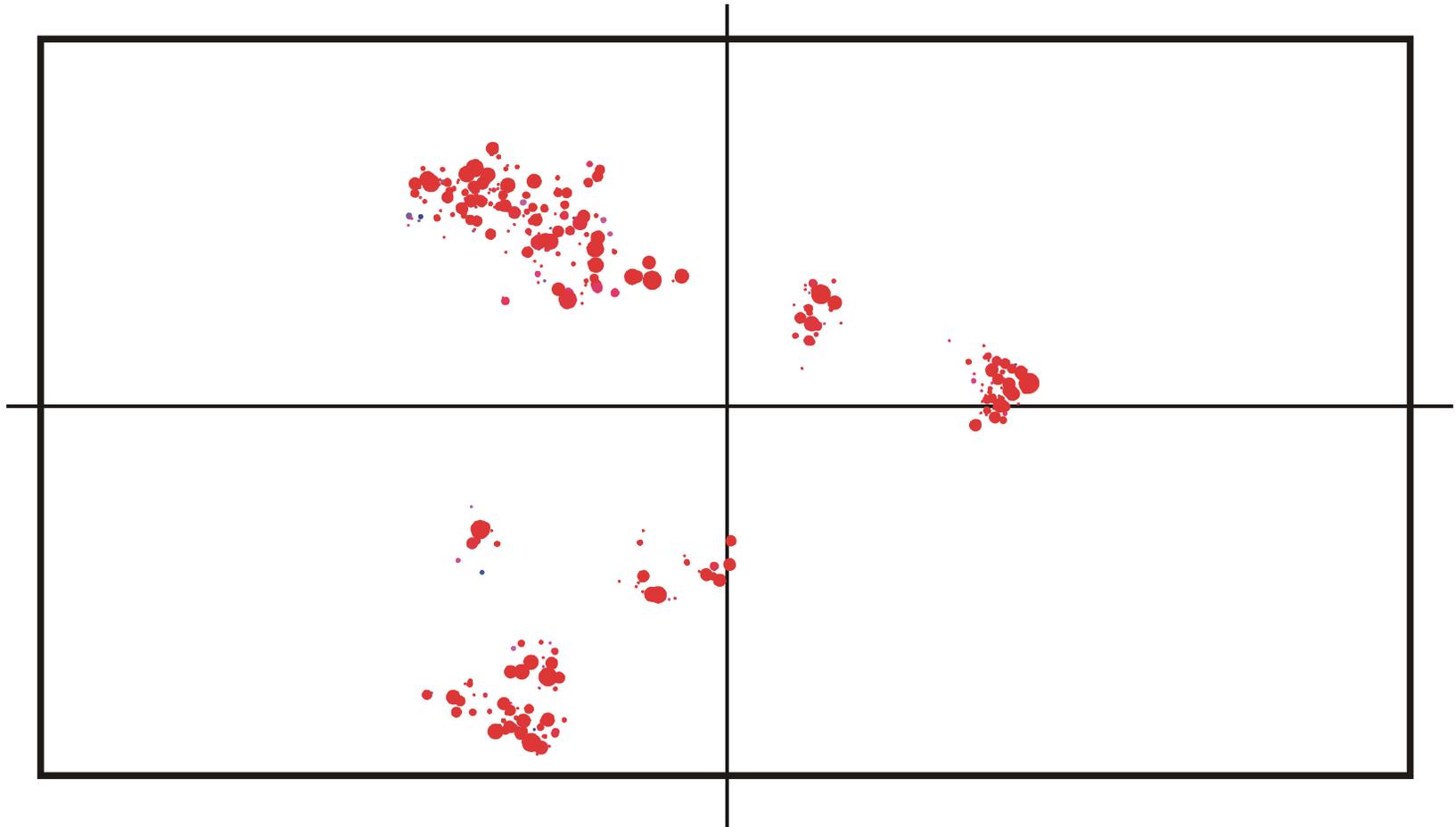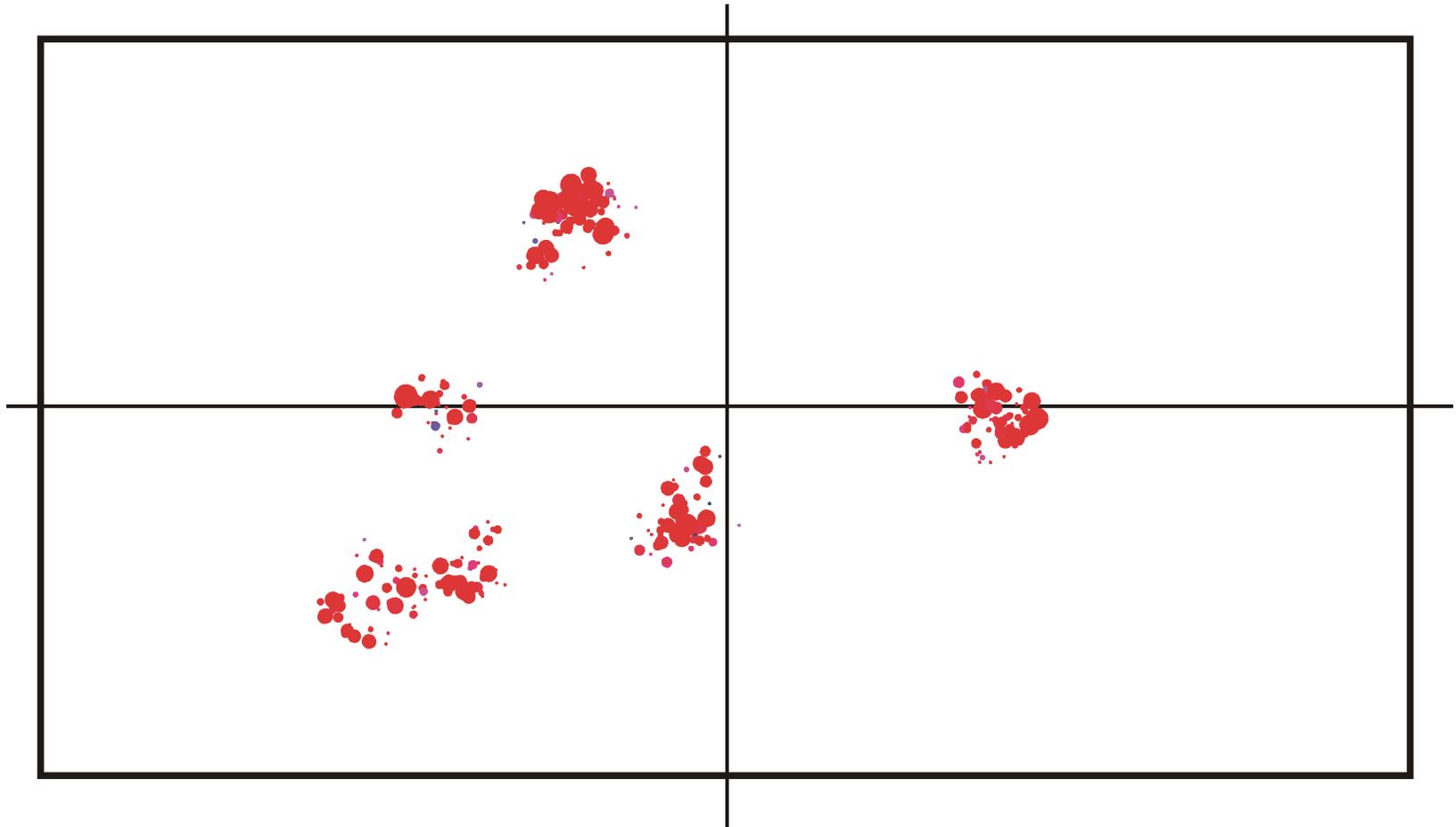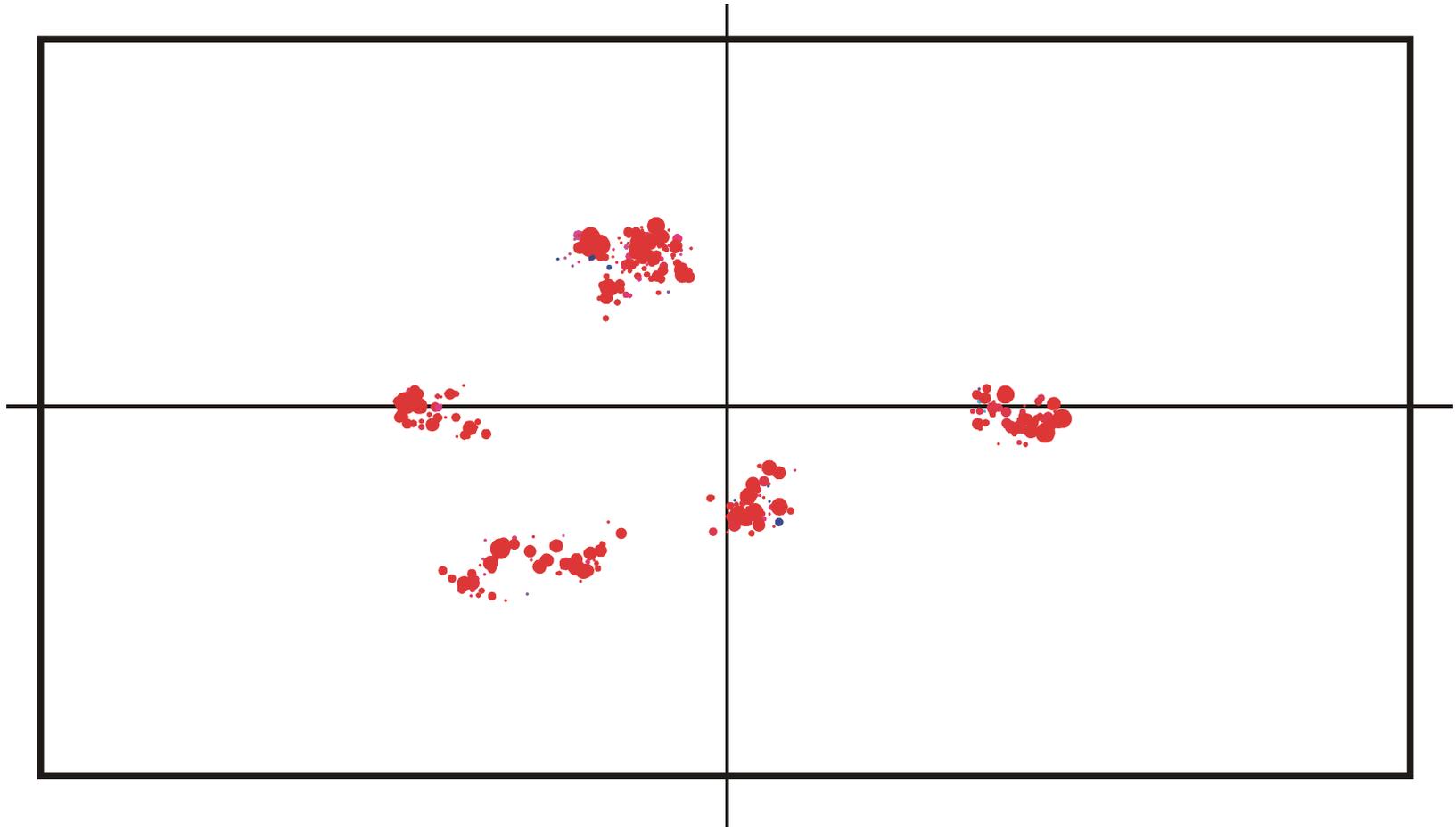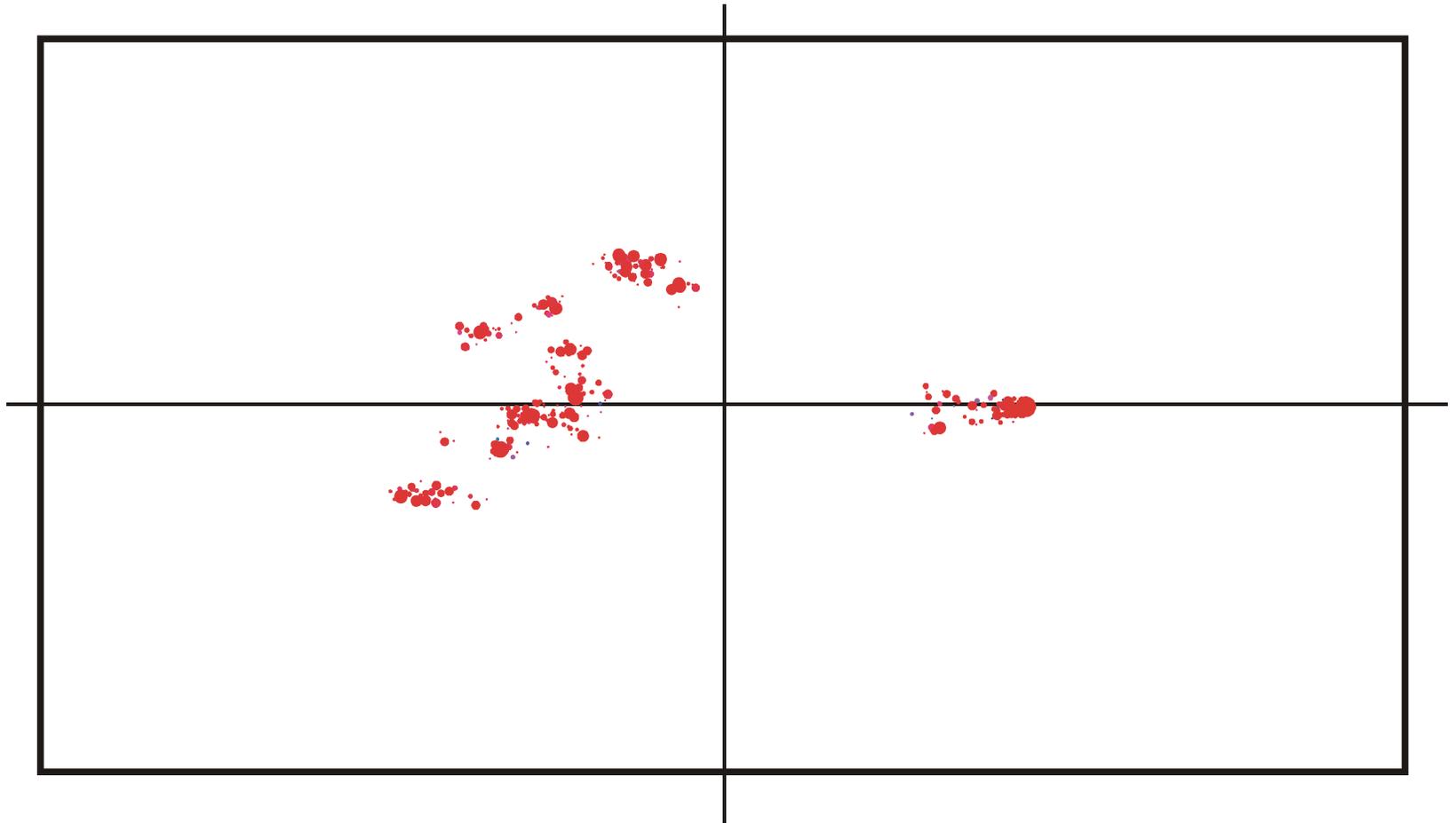Spread of population in sequence space during a quasistationary epoch: t = 170
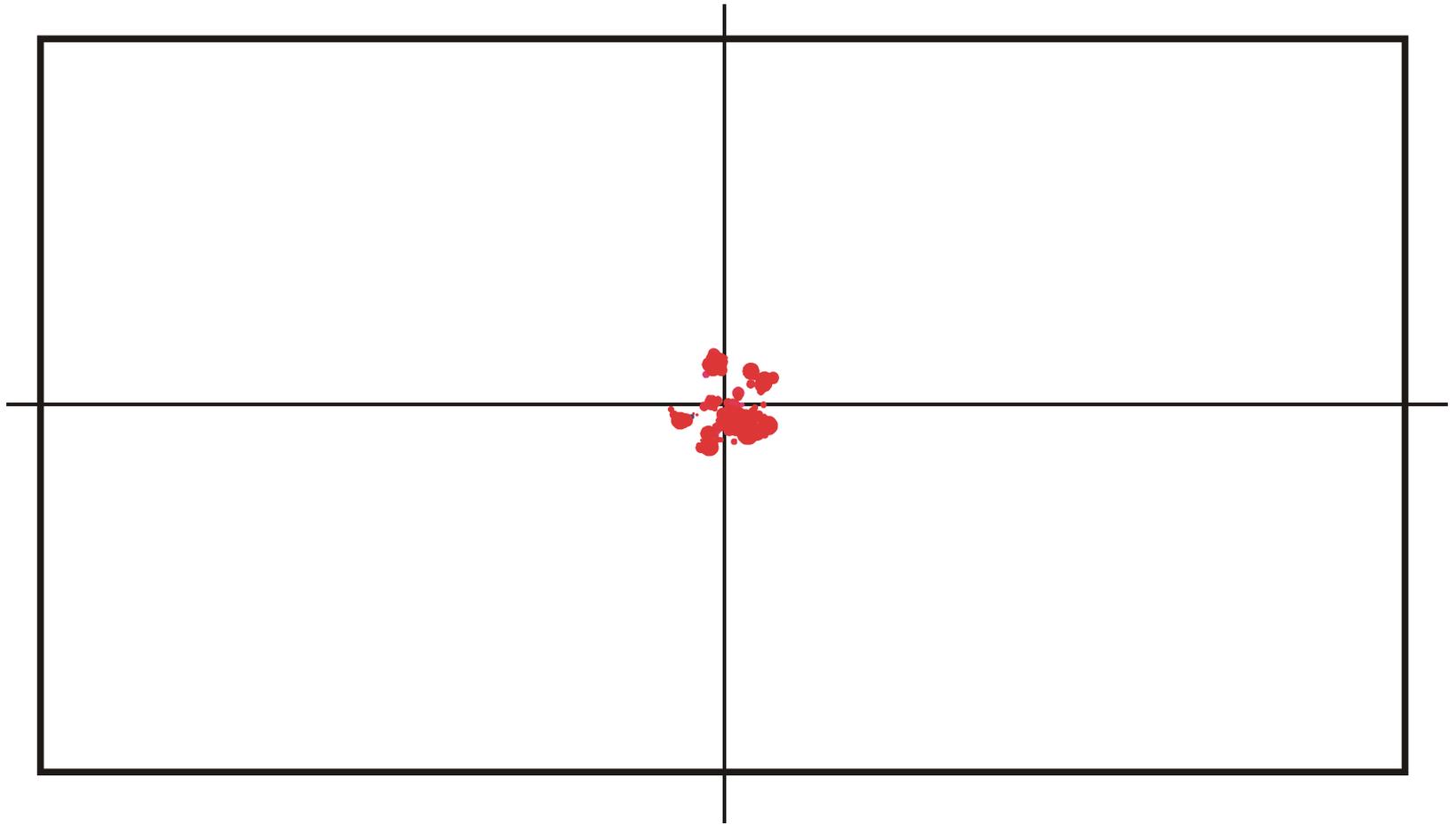
Spread of population in sequence space during a quasistationary epoch: t = 200
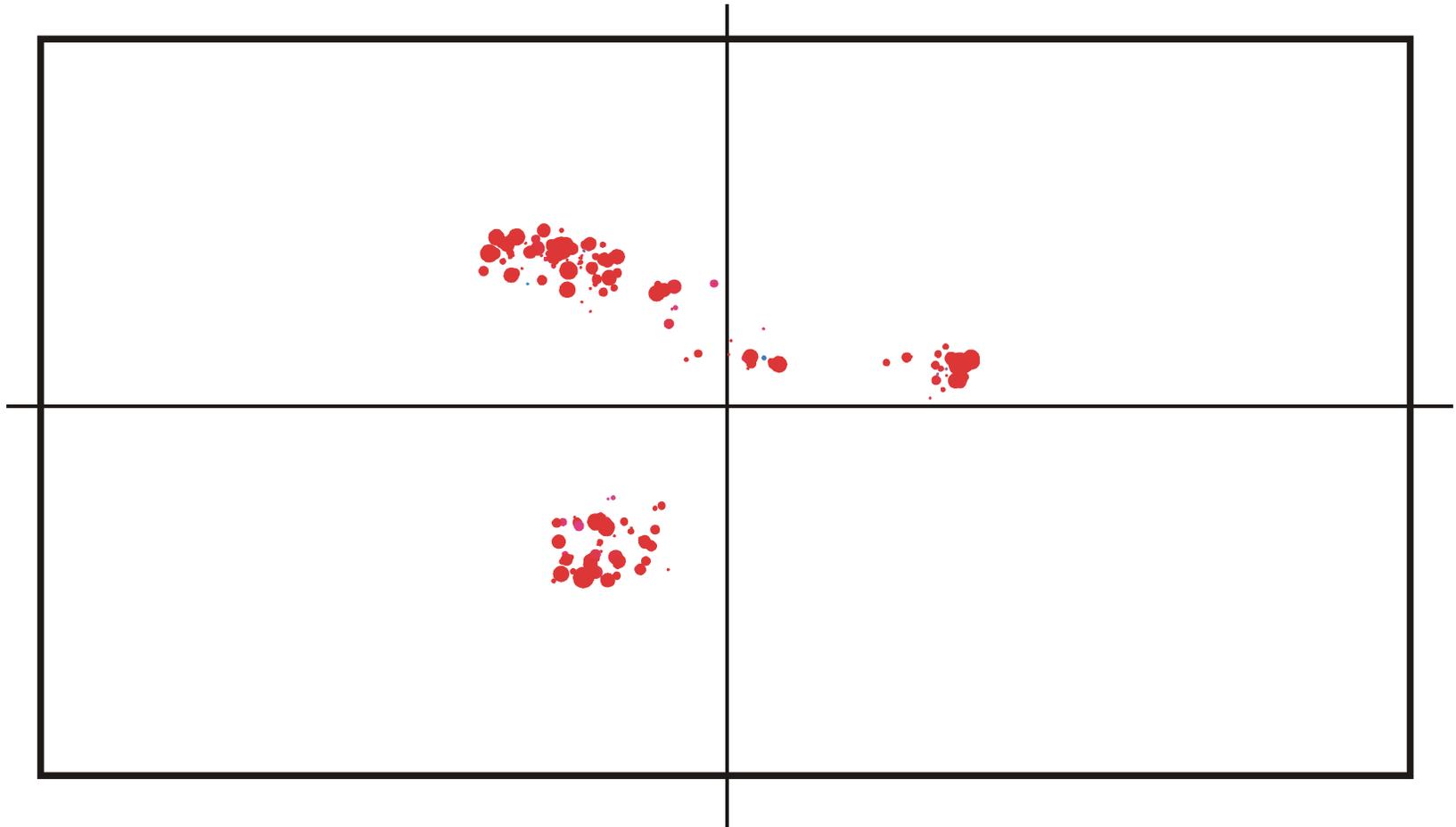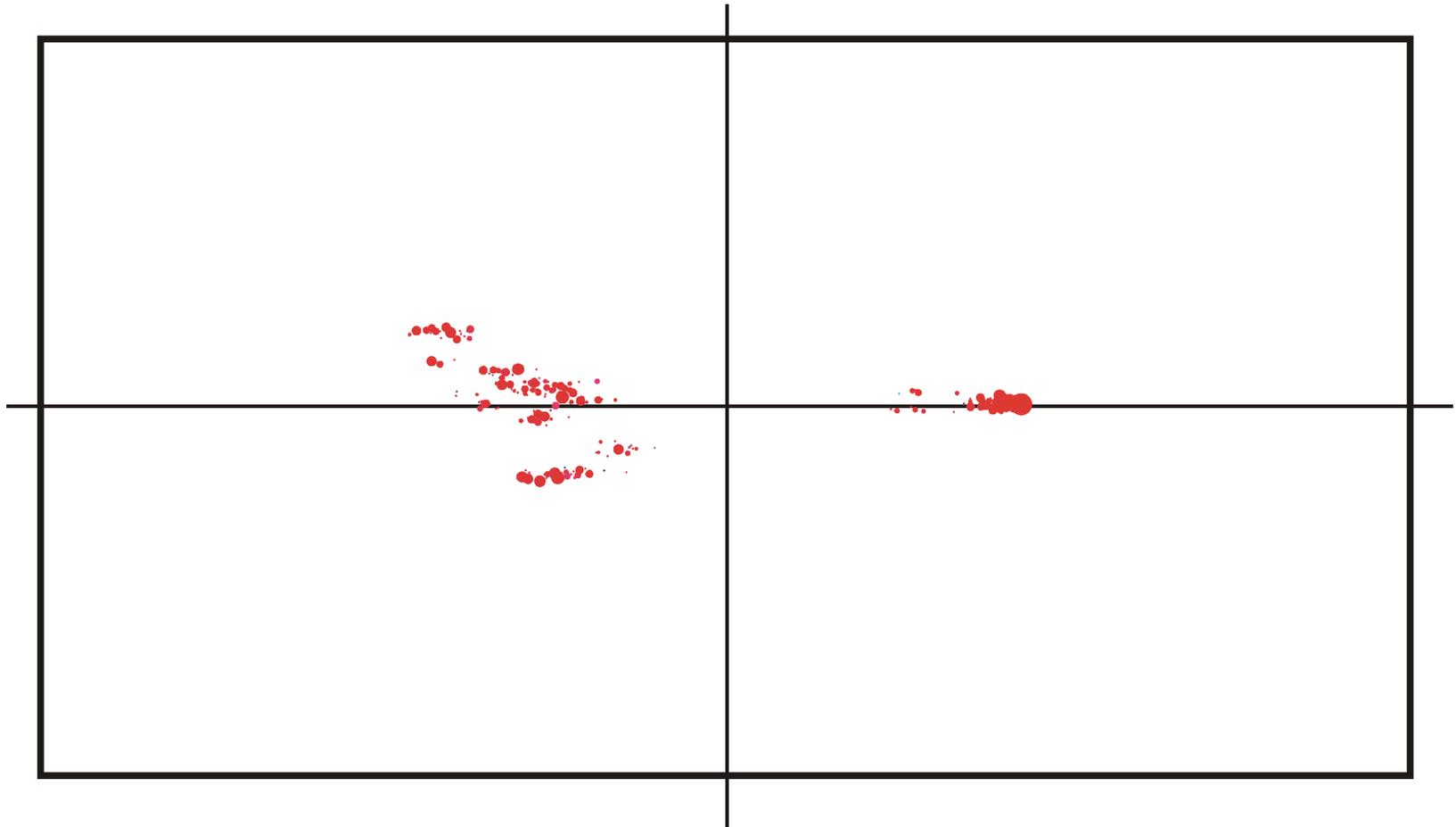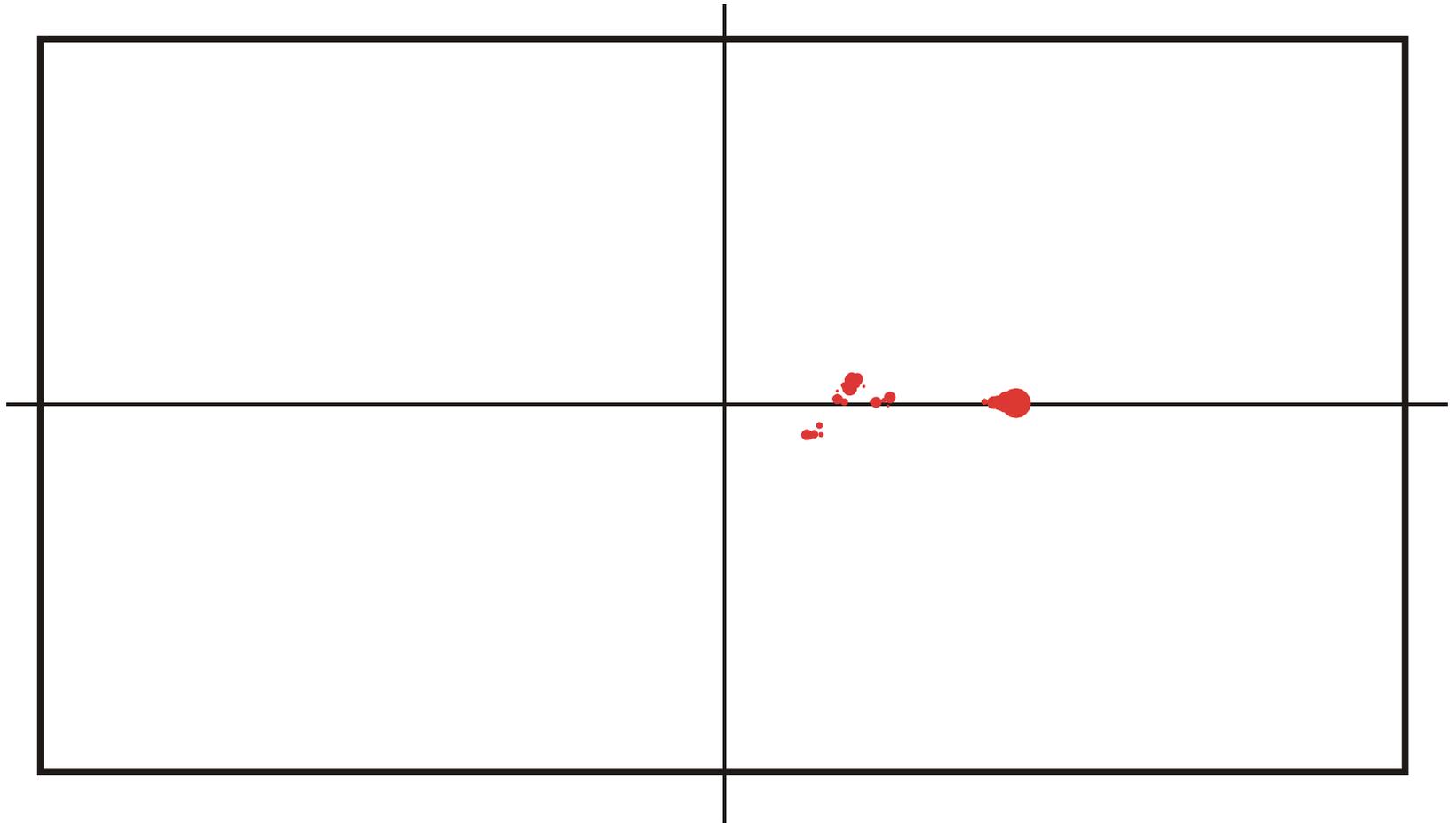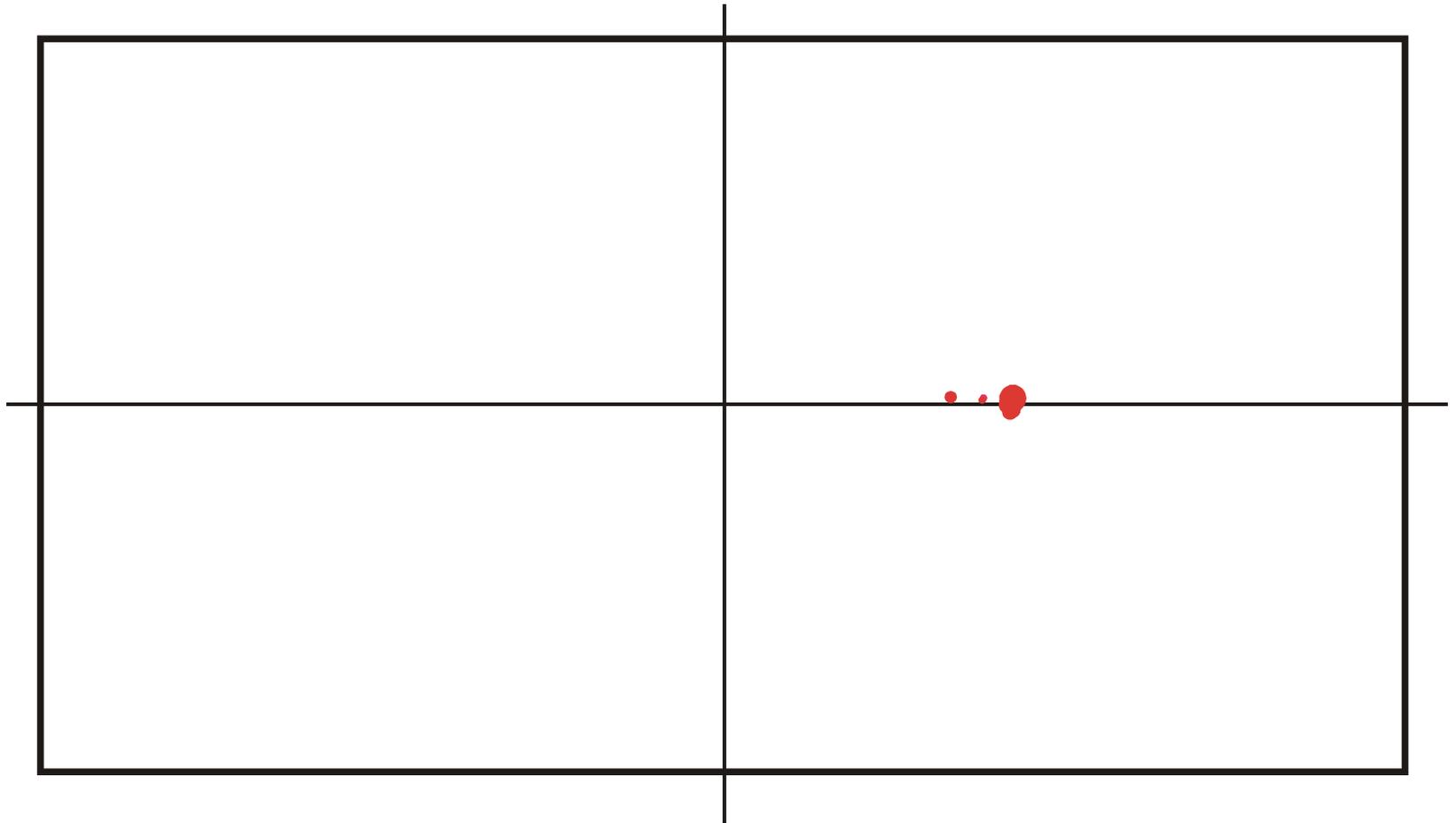
Spread of population in sequence space during a quasistationary epoch: $t = 350$
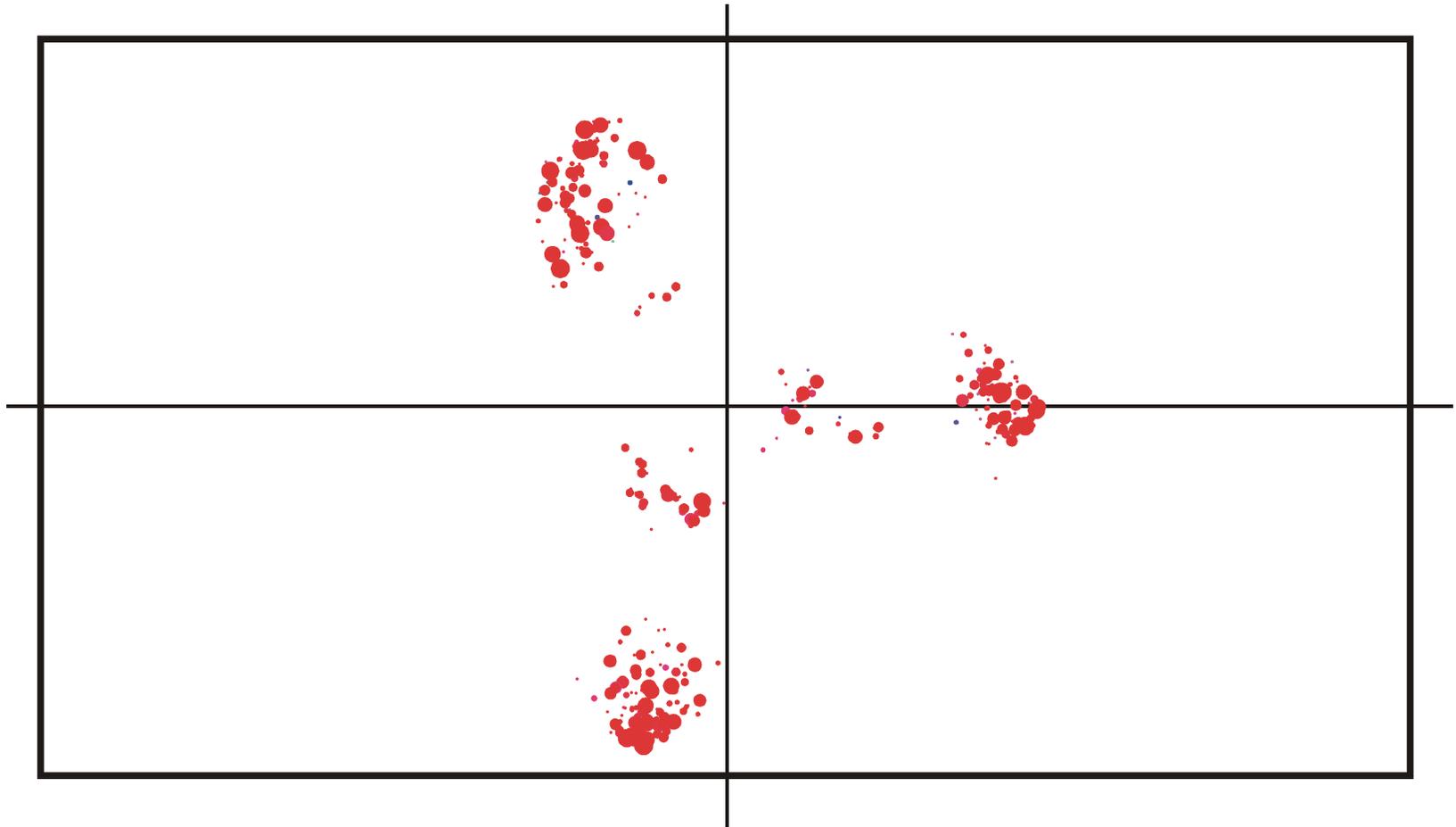
Spread of population in sequence space during a quasistationary epoch: t = 500

Spread of population in sequence space during a quasistationary epoch: t = 650

Spread of population in sequence space during a quasistationary epoch: t = 820

Spread of population in sequence space during a quasistationary epoch: t = 825

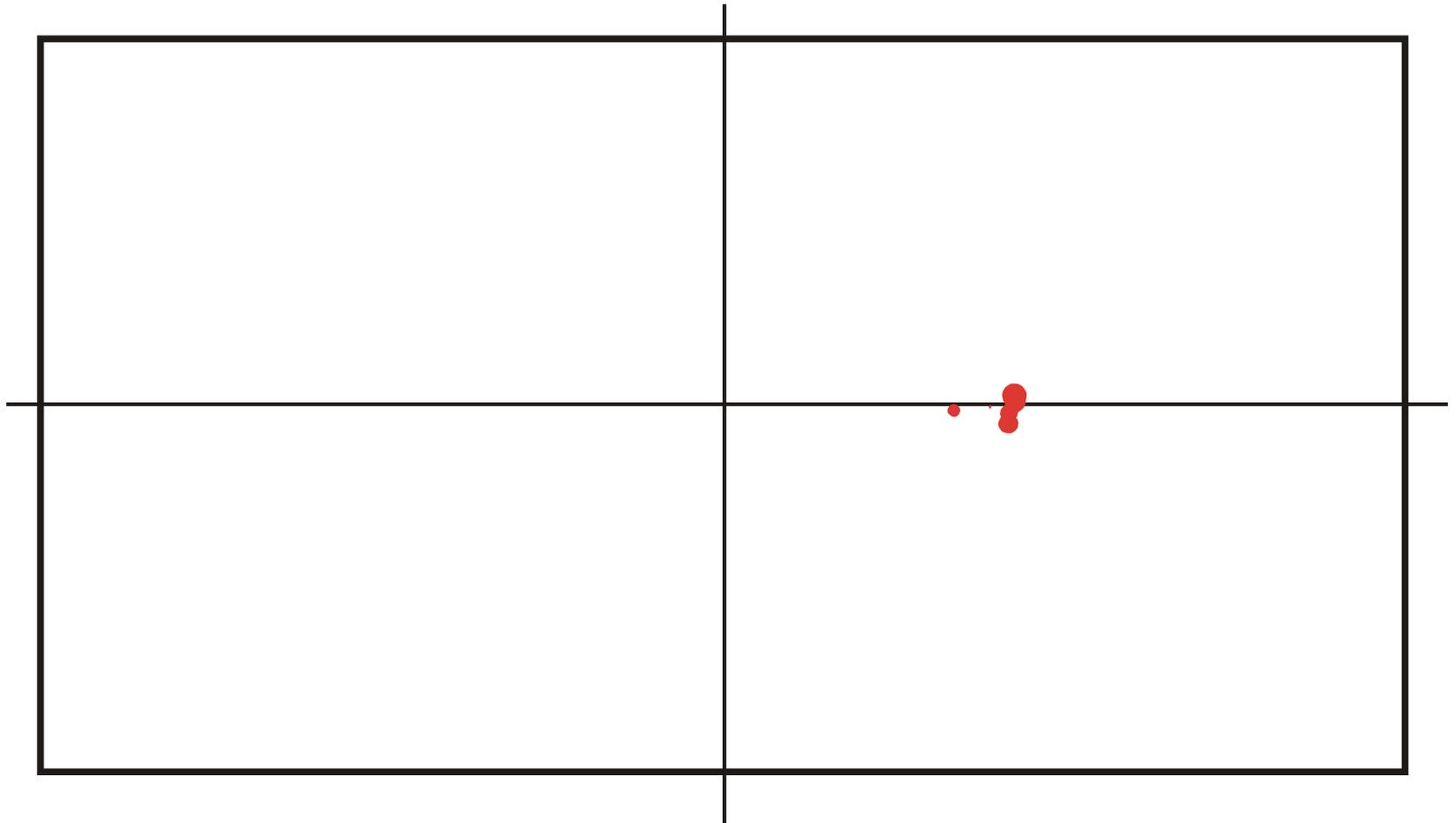Spread of population in sequence space during a quasistationary epoch: t = 830

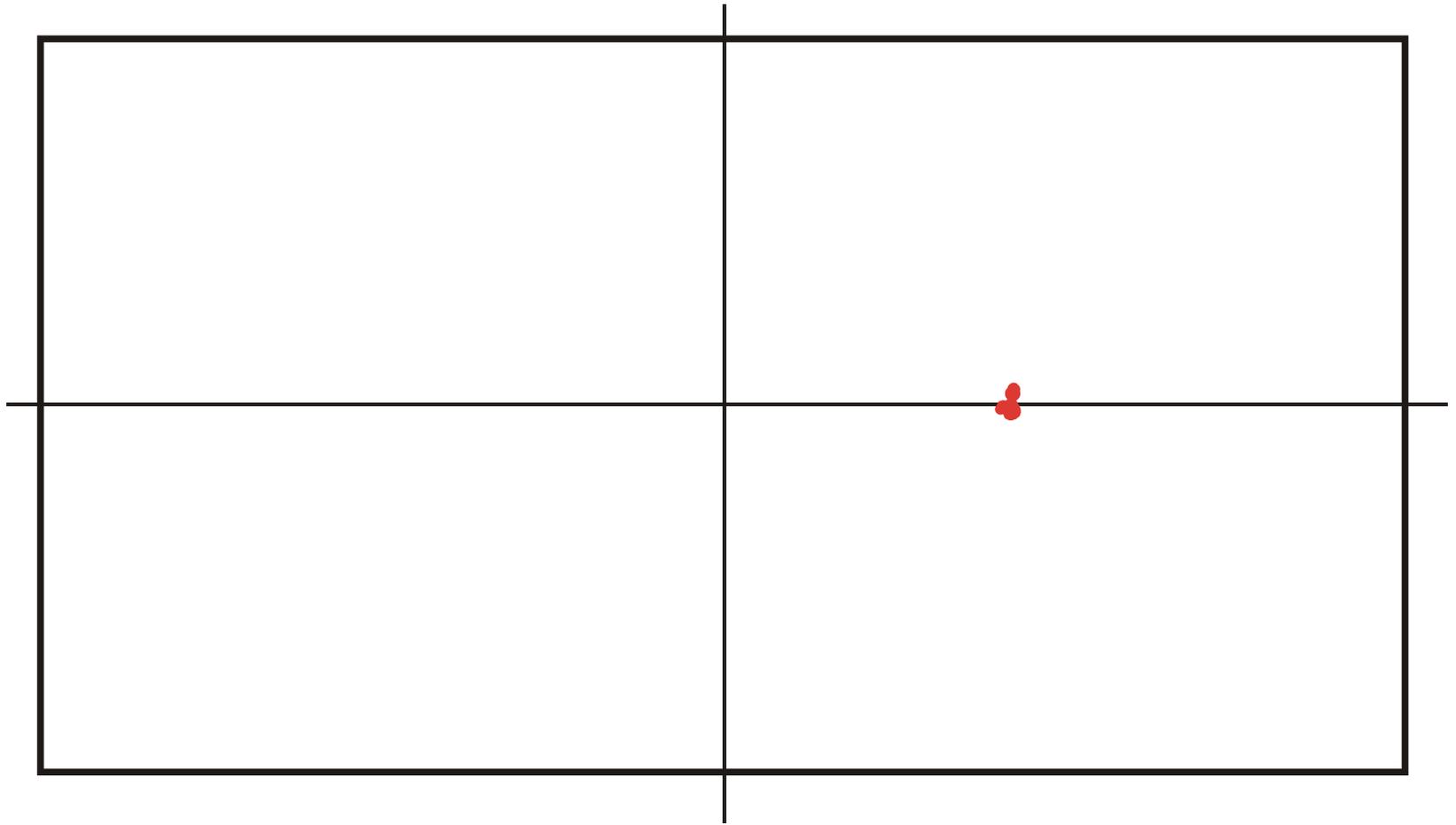Spread of population in sequence space during a quasistationary epoch: t = 835

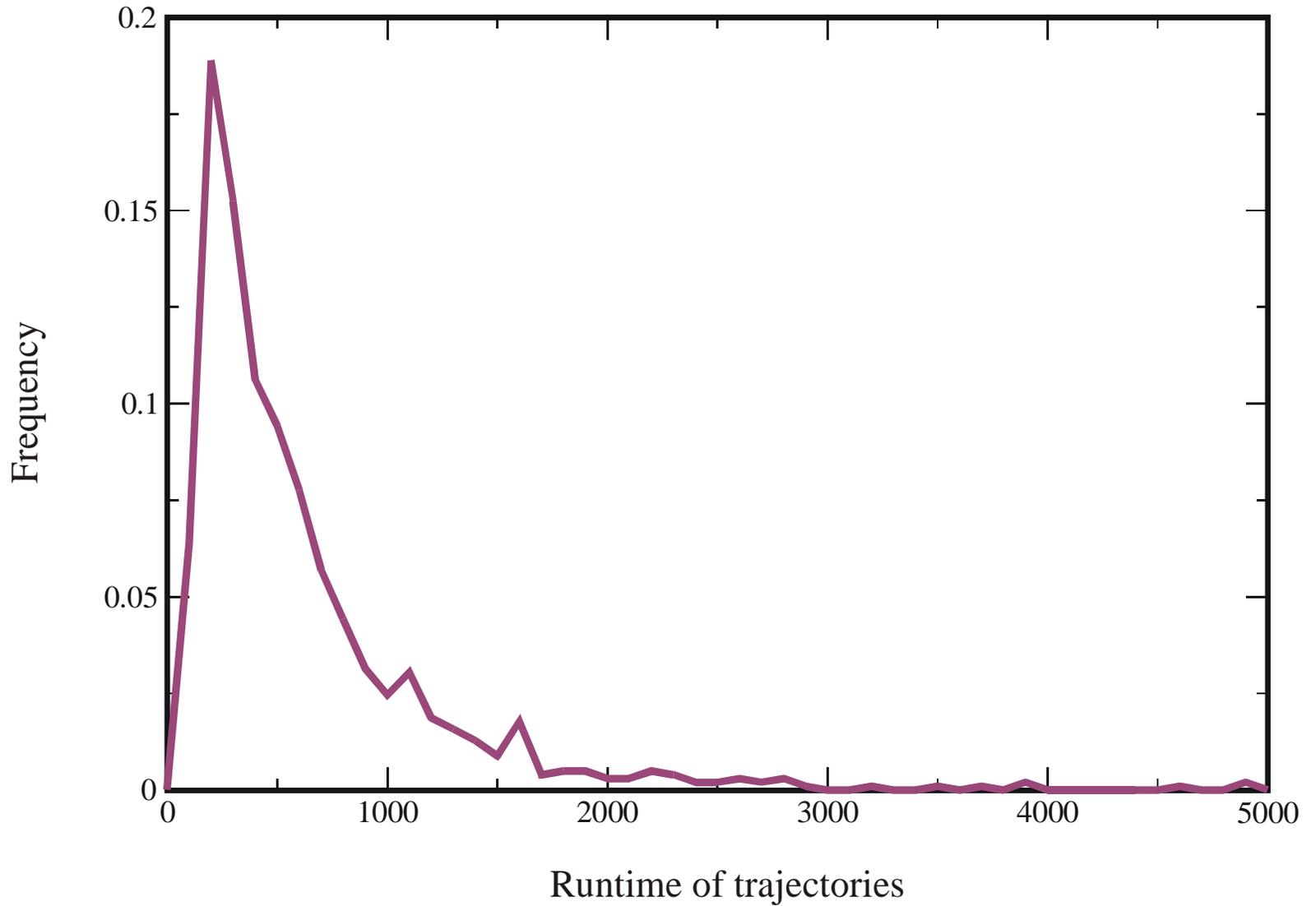Spread of population in sequence space during a quasistationary epoch: t = 840

Spread of population in sequence space during a quasistationary epoch: t = 845

Spread of population in sequence space during a quasistationary epoch: t = 850

Spread of population in sequence space during a quasistationary epoch: t = 855

Statistics of the lengths of trajectories from initial structure to target (**AUGC**-sequences)

Alphabet	Real time	Transitions	Major transitions	Sample size
AUGC	**398.3**	22.8	12.7	1199
GUC	448.9	30.5	16.5	611
GC	**1908.7**	38.7	20.1	278

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

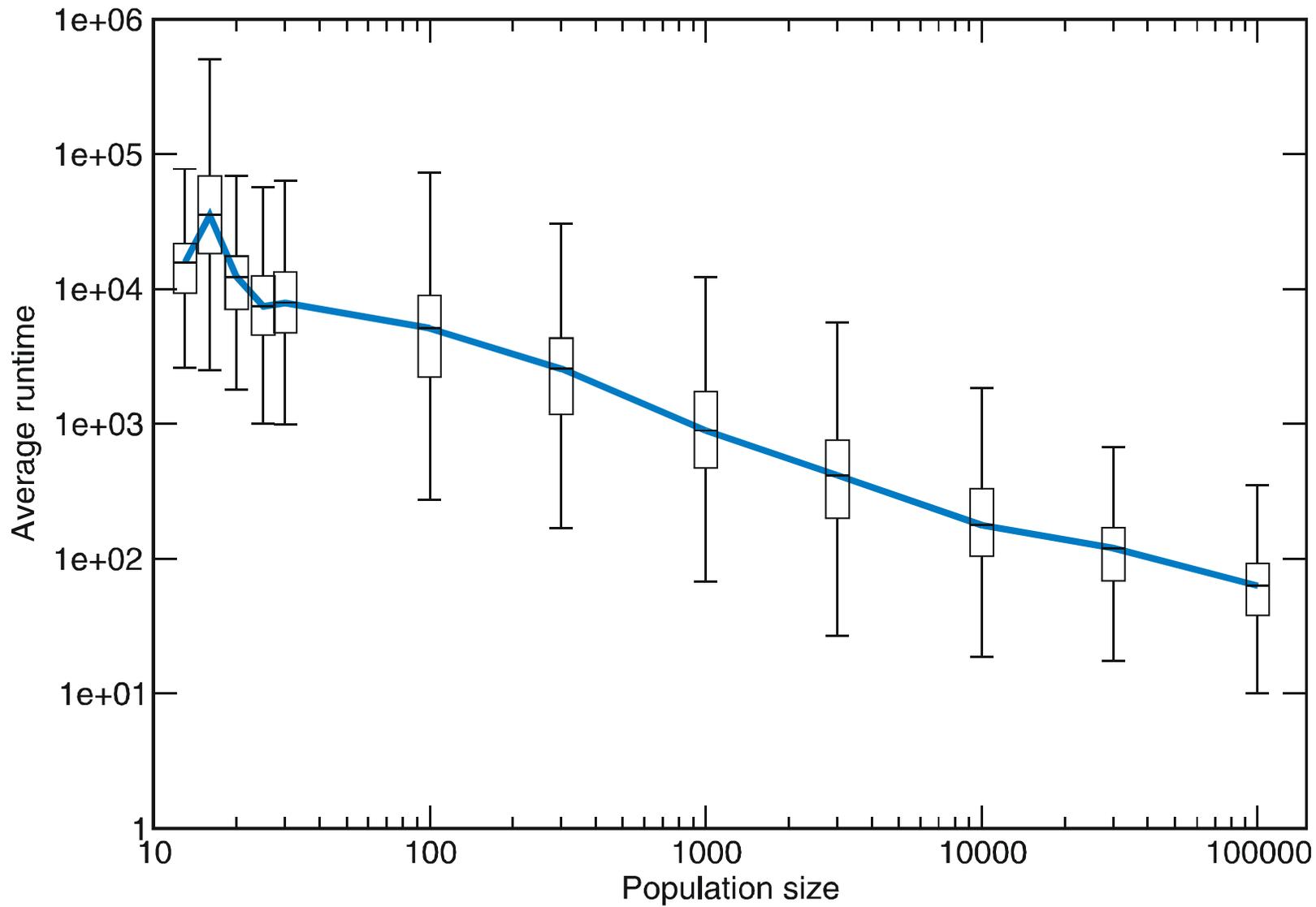Statistics of trajectories and relay series (mean values of log-normal distributions).

AUGC neutral networks of tRNAs are near the connectivity threshold, **GC** neutral networks are way below.
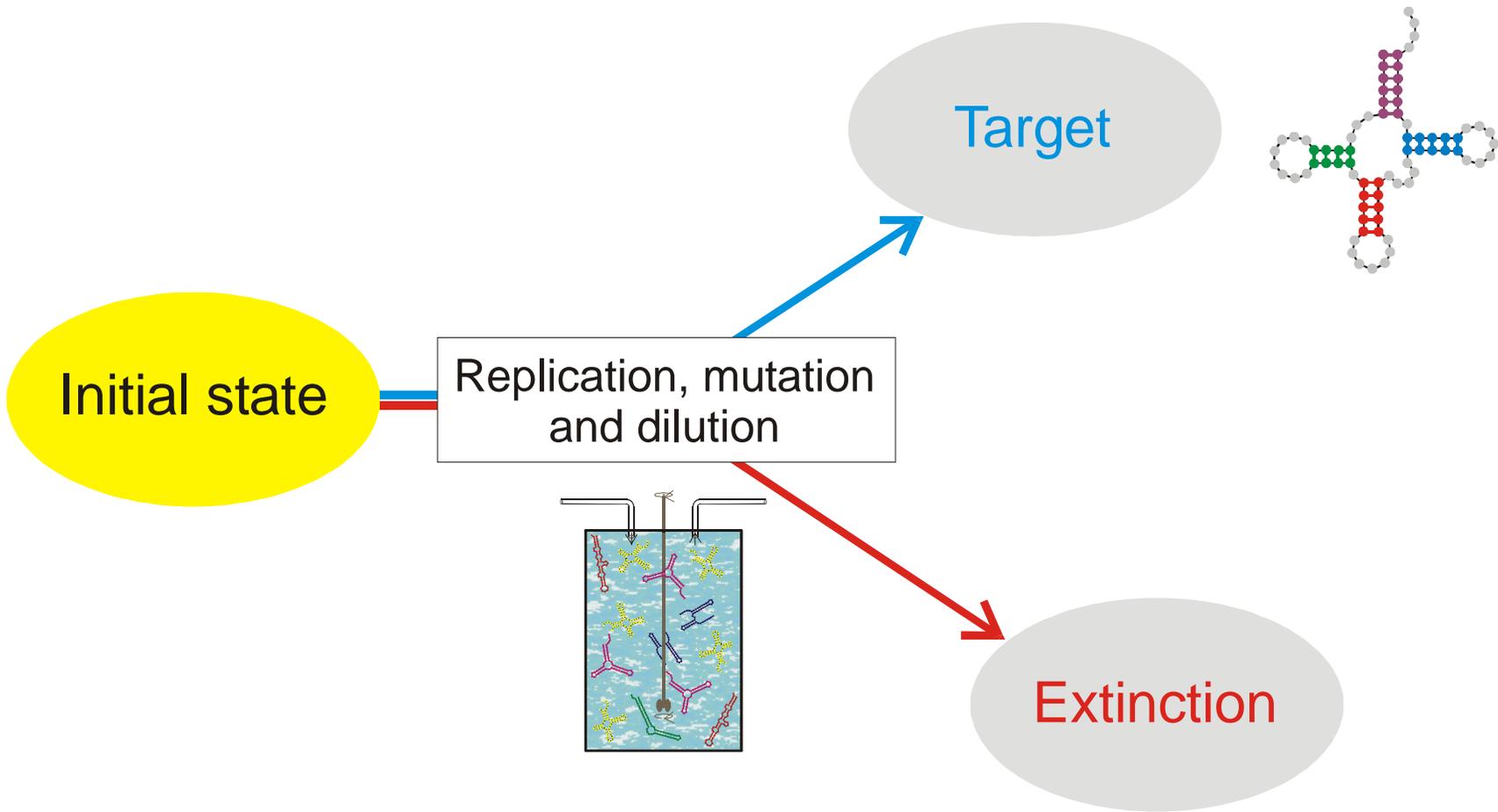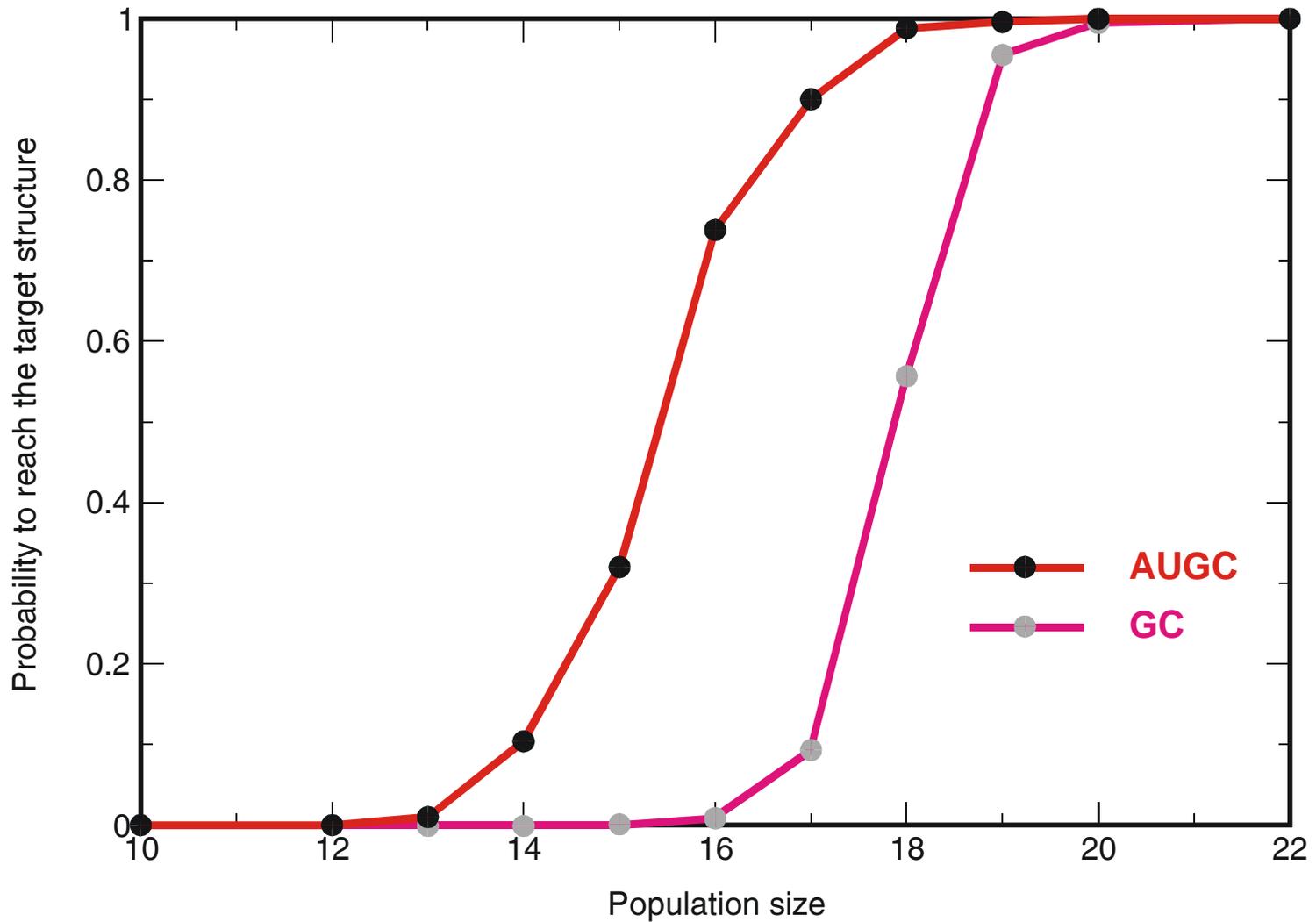
Probability of a single trajectory to reach the target structure
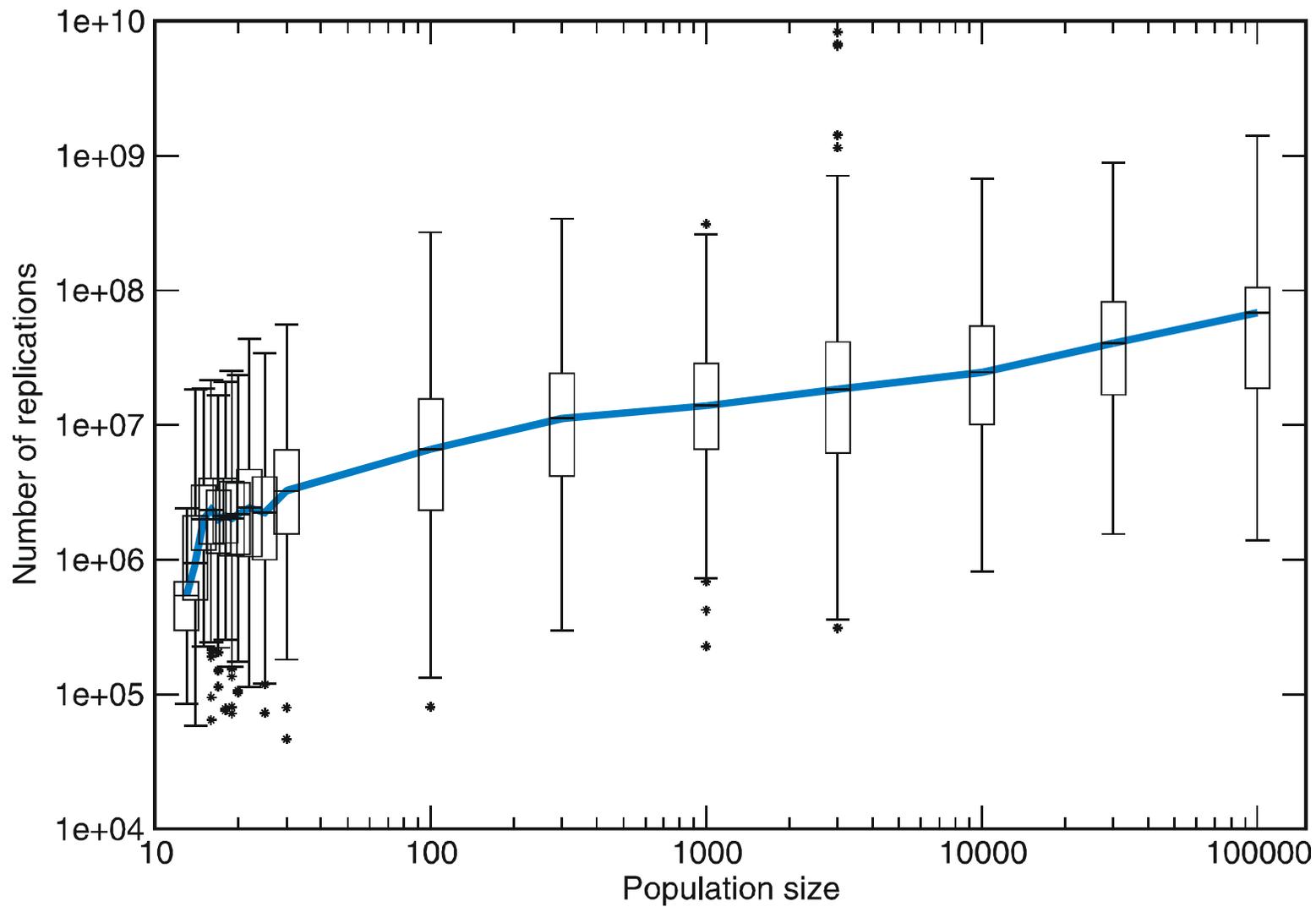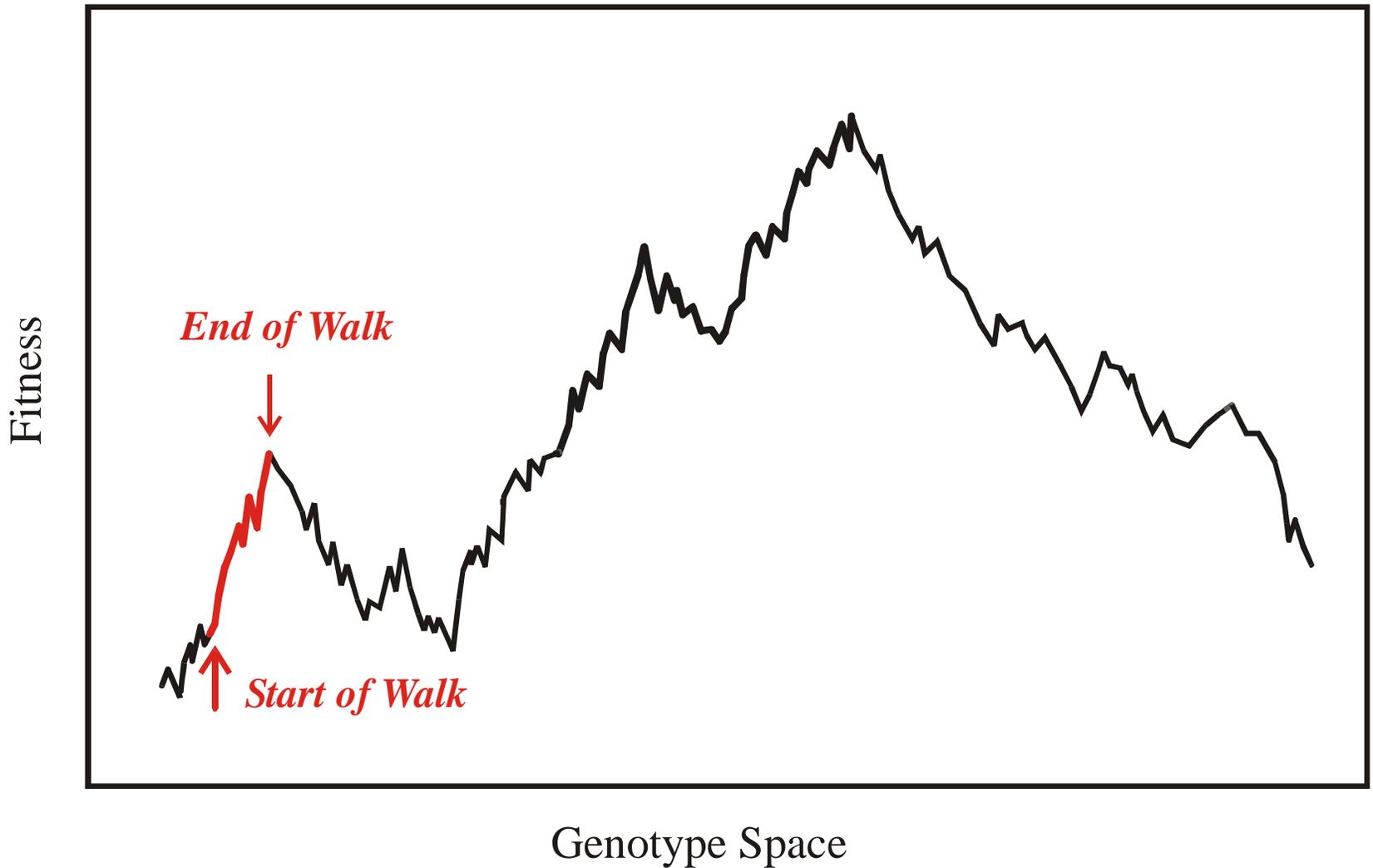
Evolutionary optimization in absence of neutral paths in sequence space
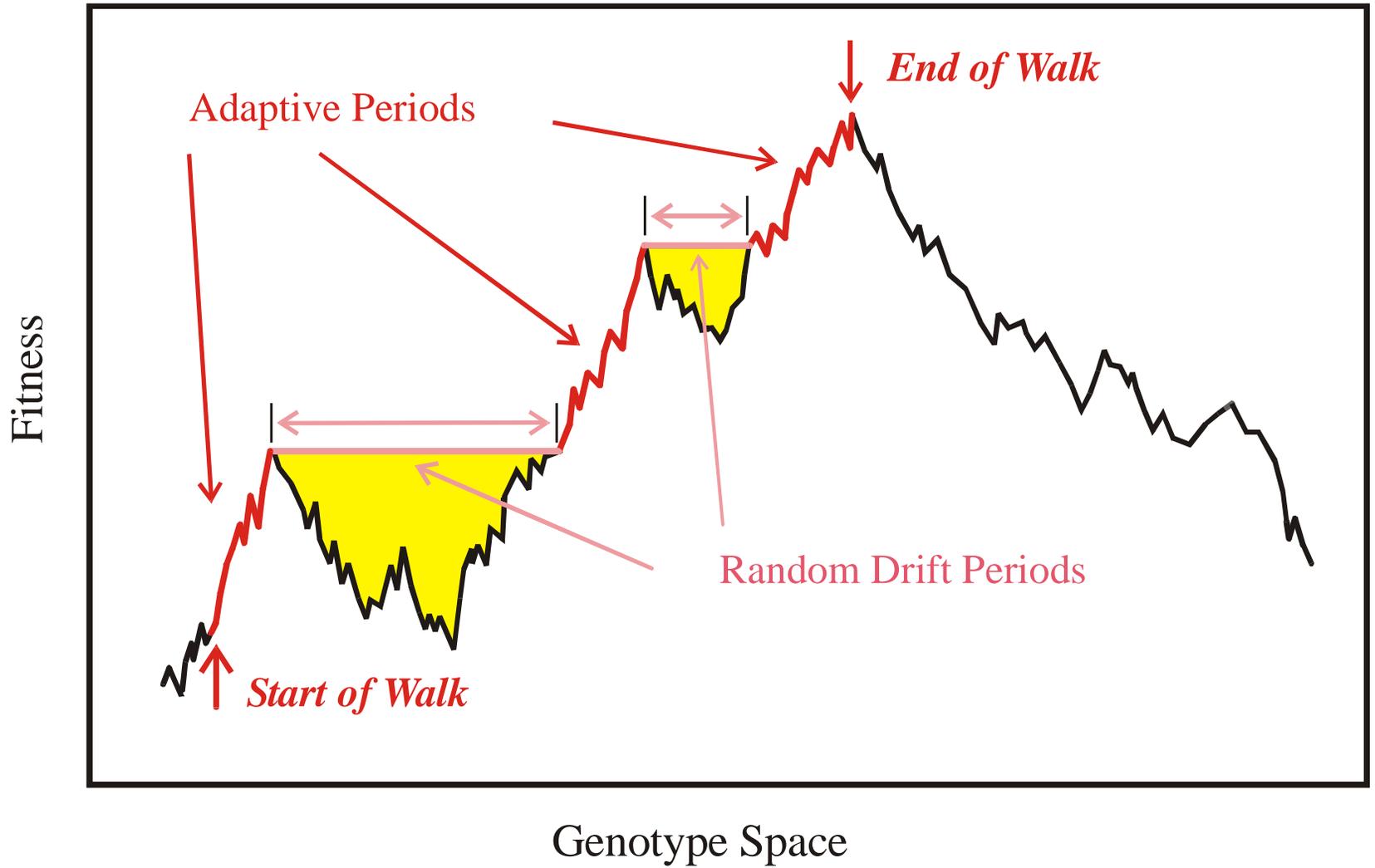
Evolutionary optimization including neutral paths in sequence space

Dolomites
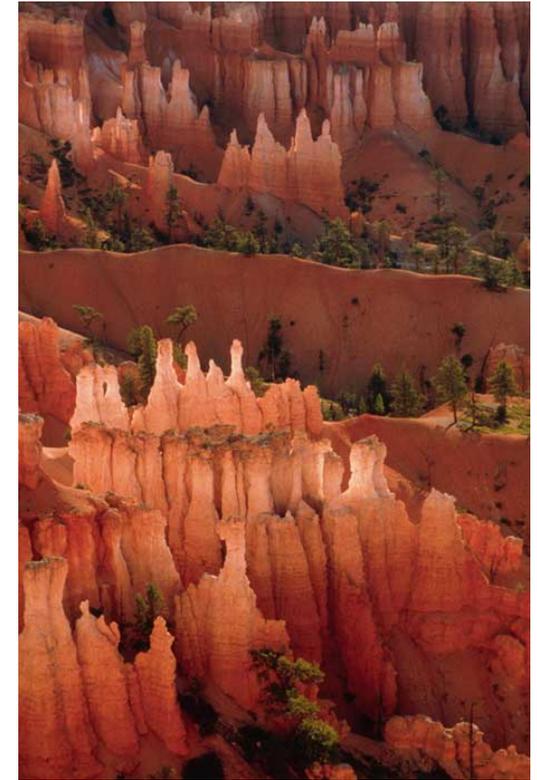
Bryce Canyon

Examples of rugged landscapes on Earth

Grand Canyon

Example of a landscape on Earth with 'neutral' ridges and plateaus

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

Sir Peter Medawar, 1985

Space of genotypes: $\quad \mathcal{I} = \{I_1, I_2, I_3, I_4, \dots, I_N\}$; Hamming metric

Space of phenotypes: $\quad \mathcal{S} = \{S_1, S_2, S_3, S_4, \dots, S_M\}$; metric (not required)

$$N \gg M$$

$$\psi(I_j) = S_k$$

$$\mathbf{G}_k = \psi^{-1}(S_k) \cup \{ \, I_j \mid \psi(I_j) = S_k \, \}$$

A mapping ψ and its inversion

.... GCCAUC ....

$d_H=1$

.... GCGAUC ....

$d_H=2$

.... GCCUUC ....

$d_H=1$

.... GCGUUC ....
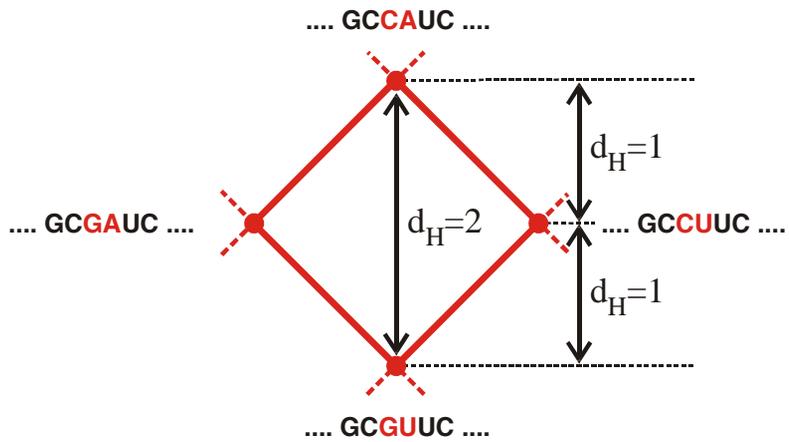
City-block distance in sequence space
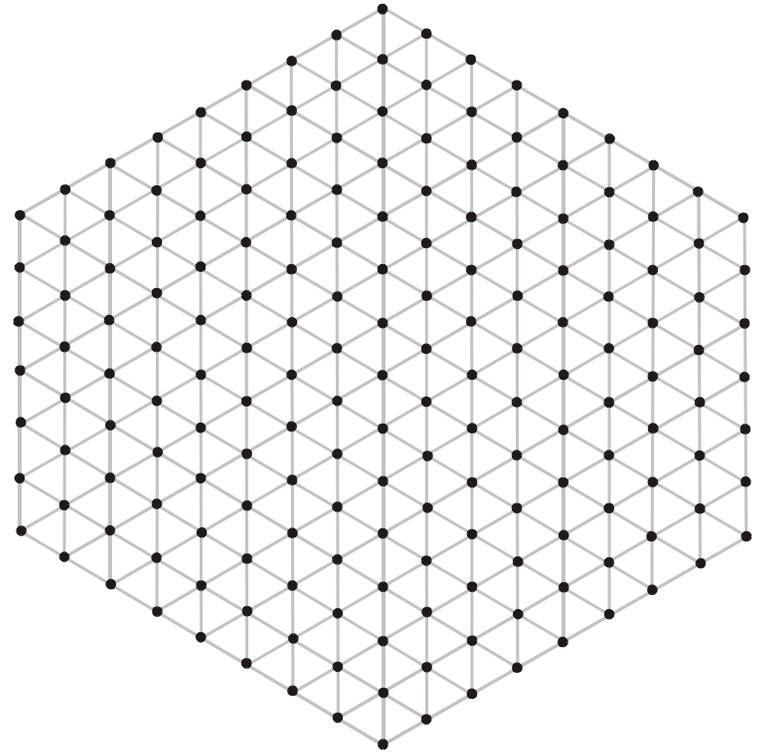
2D Sketch of sequence space

Single point mutations as moves in sequence space

I$_1$: CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....

I$_2$: CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAACACAAGAG.....

Hamming distance $d_H(I_1,I_2) = 4$

(i) $d_H(I_1,I_1) = 0$

(ii) $d_H(I_1,I_2) = d_H(I_2,I_1)$

(iii) $d_H(I_1,I_3) \leq d_H(I_1,I_2) + d_H(I_2,I_3)$

The Hamming distance between sequences induces a metric in sequence space

S1: . . . . . . ((((((. . (((((. . . . . .))))))) . . . ((((. . . . .))))) . .))))))

S2: . . . . . . ((((((. . ((. (((. . . . . .))) .)) . . (((((. . . . .)))))) .))))))

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

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

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

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

The Hamming distance between structures in parentheses notation forms a metric in structure space

$S_k = \psi(I.)$

$f_k = f(S_k)$

Function

Sequence space   Structure space   Real numbers
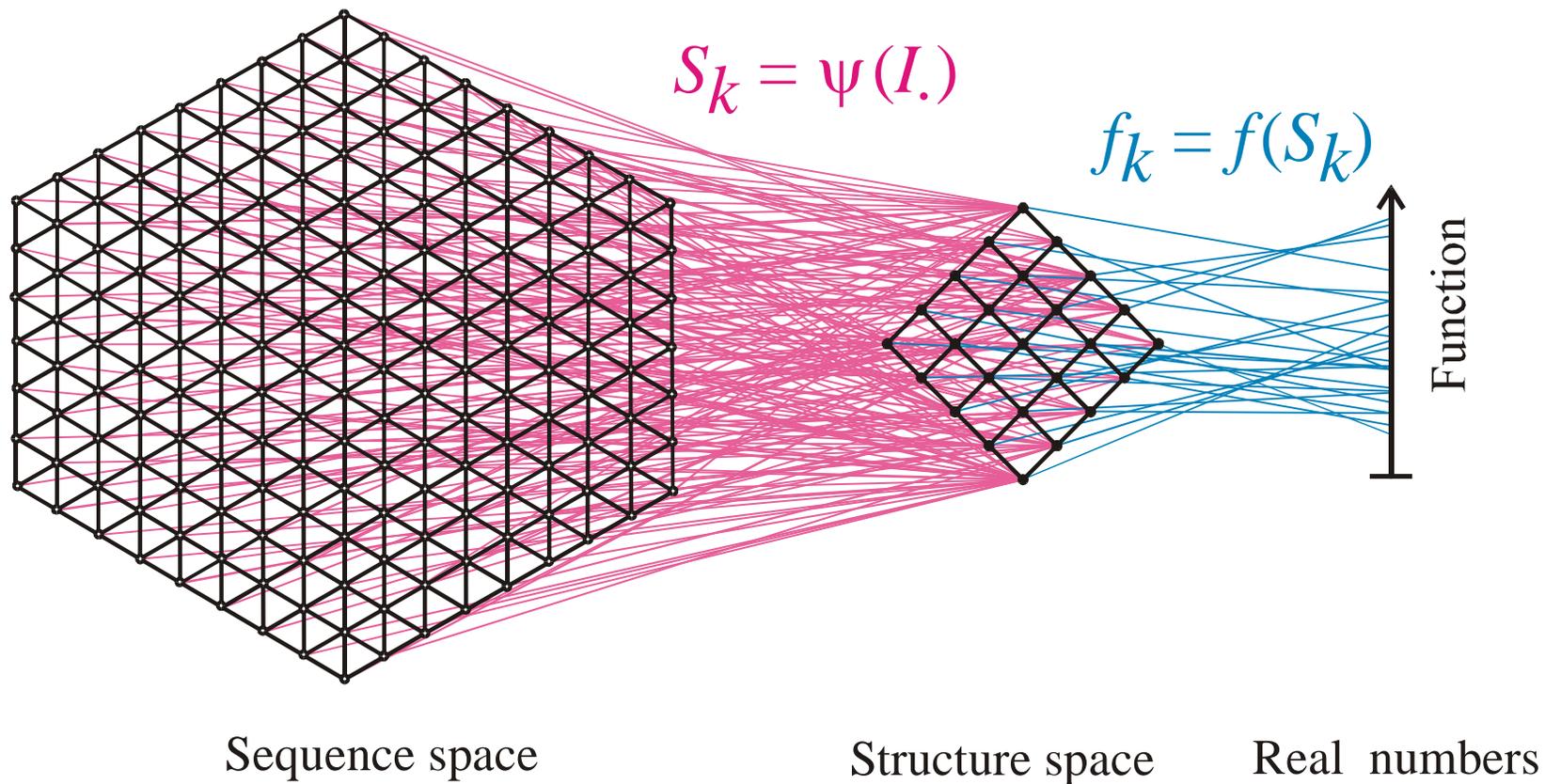
Mapping from sequence space into structure space and into function

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

$$f_k = f(S_k)$$

Function

Sequence space   Structure space   Real numbers

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

Sequence space   Structure space

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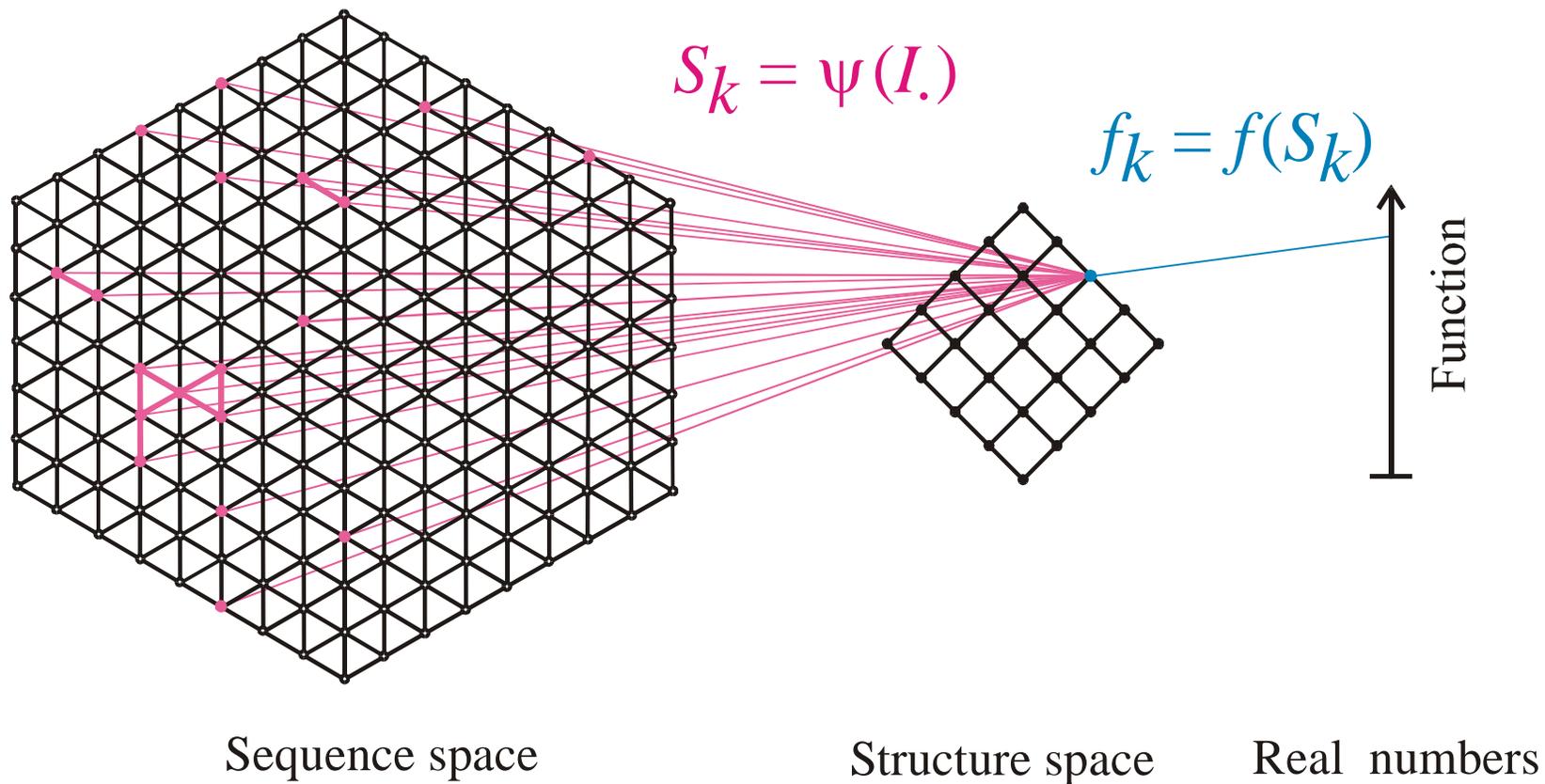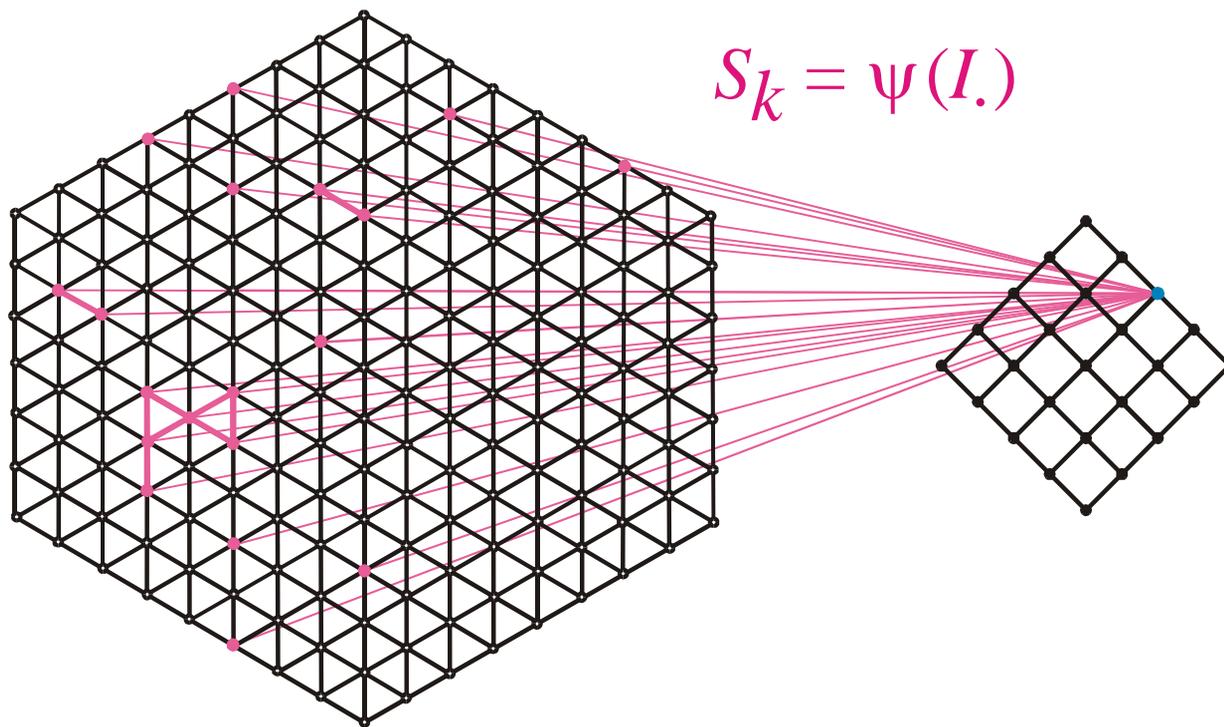
Sequence space
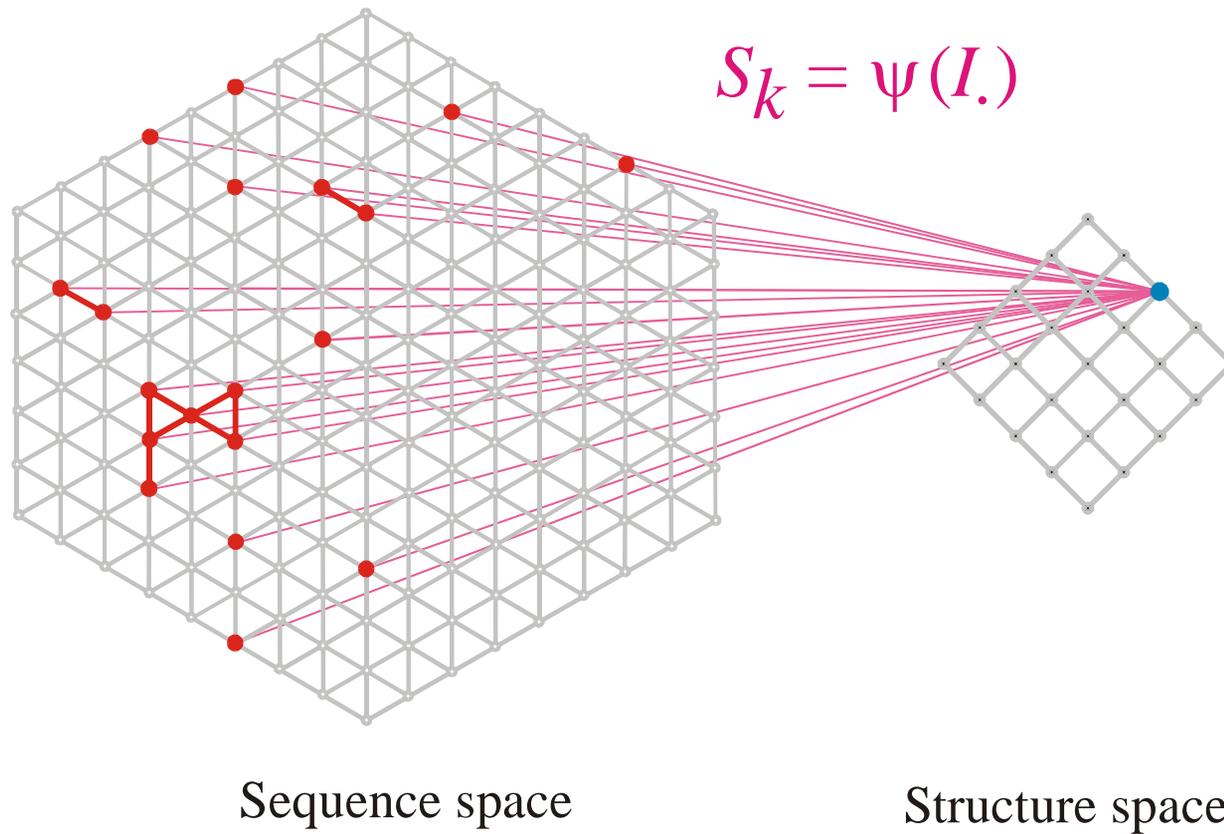
Structure space

The pre-image of the structure S_k in sequence space is the **neutral network G_k**
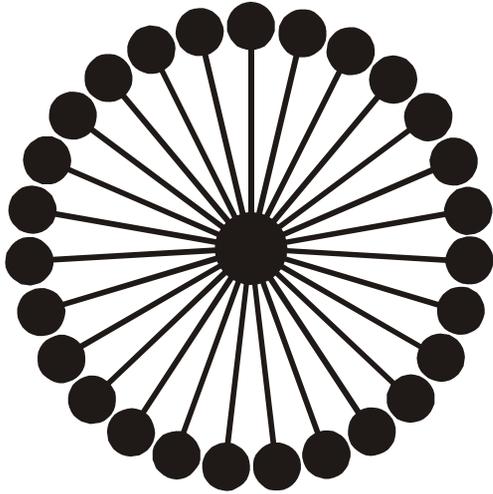
n = 9 ; 3n = 27

Degree of neutrality of neutral networks and the connectivity threshold
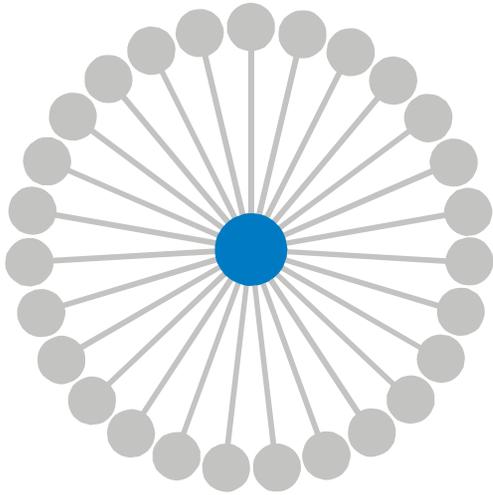
Degree of neutrality of neutral networks and the connectivity threshold
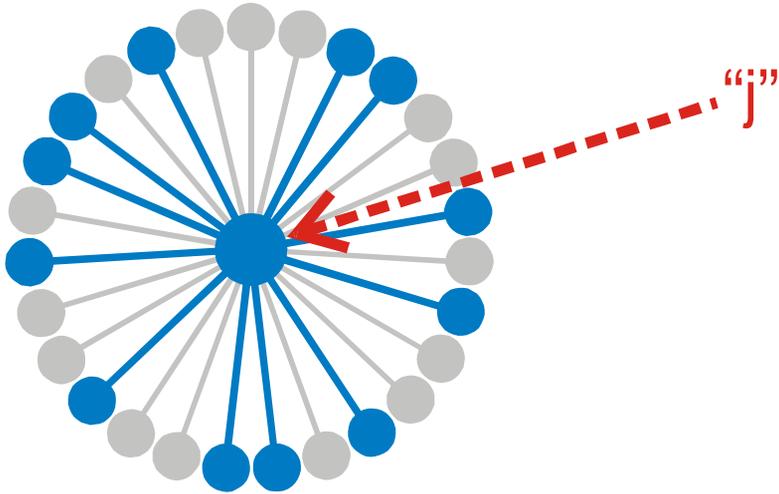
$$\lambda_j = \mathbf{12} / 27 = 0.444$$

Degree of neutrality of neutral networks and the connectivity threshold
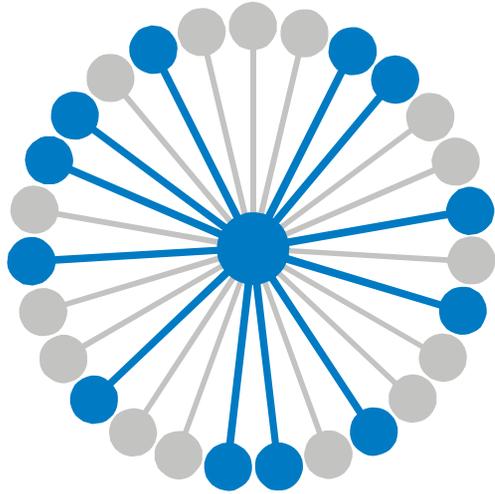
$$G_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

$$\overline{\lambda}_k = \frac{\sum\limits_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

$$\lambda_j = 12 / 27 = 0.444$$

Degree of neutrality of neutral networks and the connectivity threshold
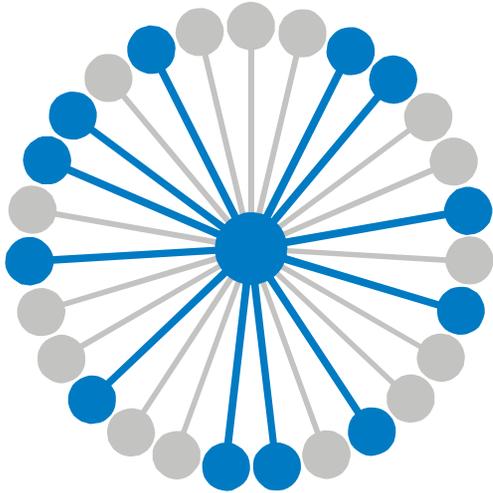
$$G_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

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

$$\lambda_j = 12 / 27 = 0.444$$

$\bar{\lambda}_k > \lambda_{cr}$ .... 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
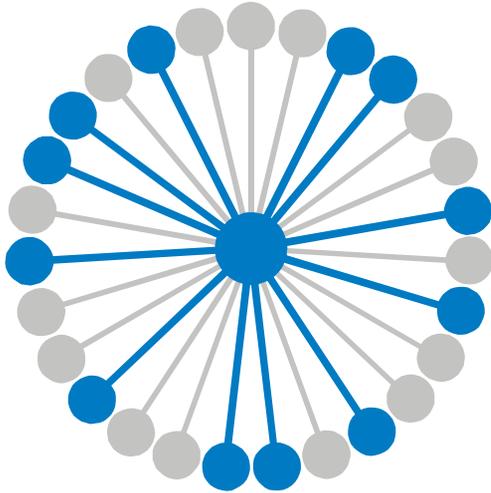
$$G_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

$$\overline{\lambda}_k = \frac{\sum\limits_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

Alphabet size κ :

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

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

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

$\overline{\lambda}_k < \lambda_{cr}$ .... 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

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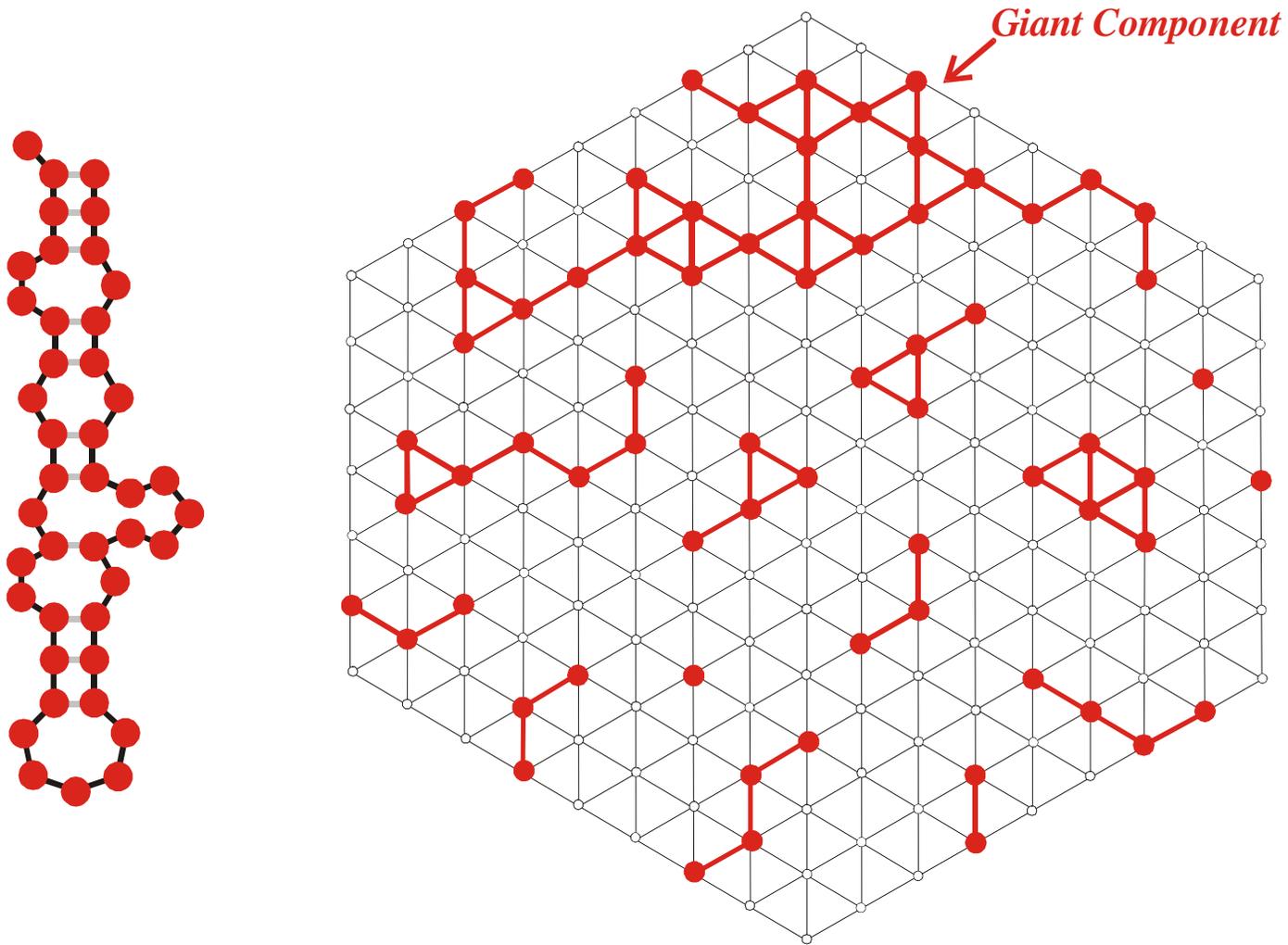
A multi-component neutral network formed by a rare structure: $\lambda < \lambda_{cr}$
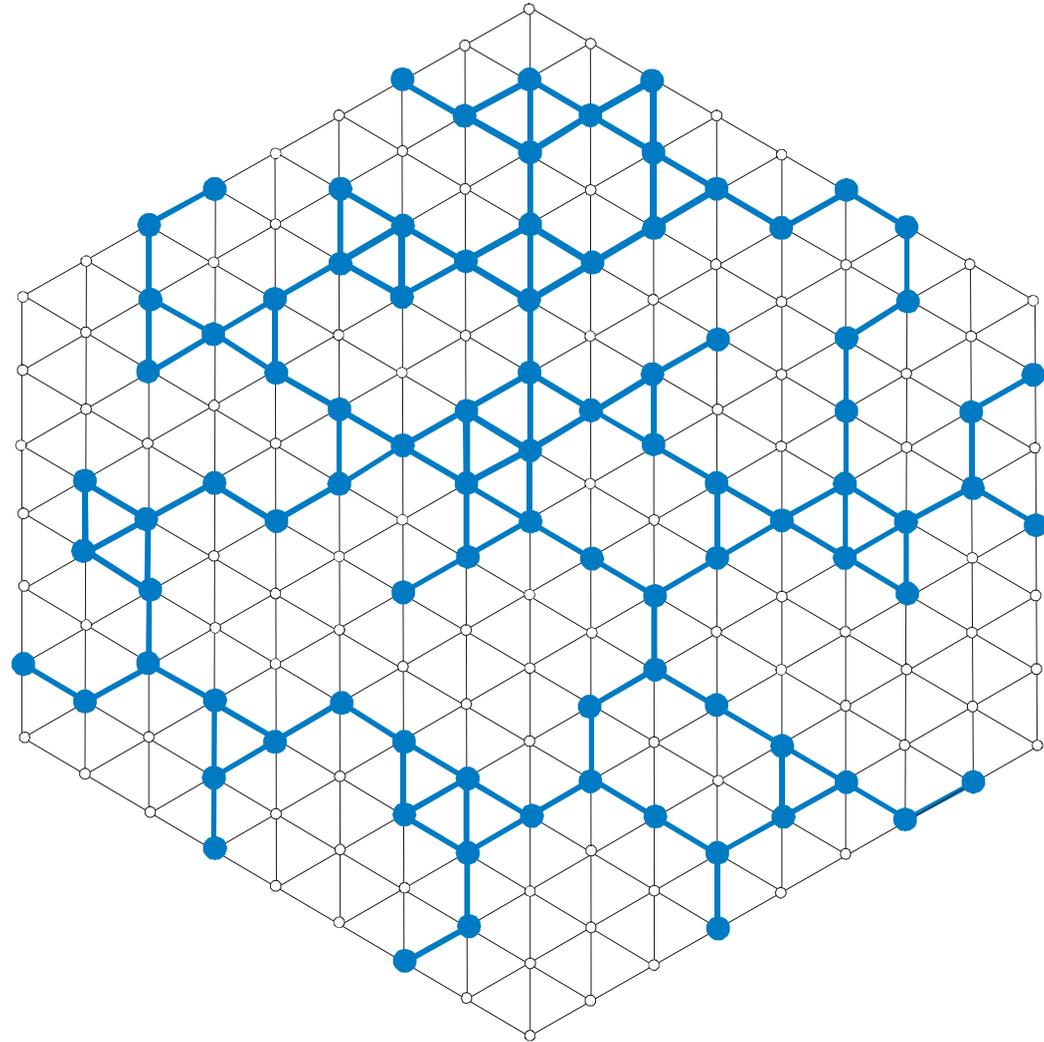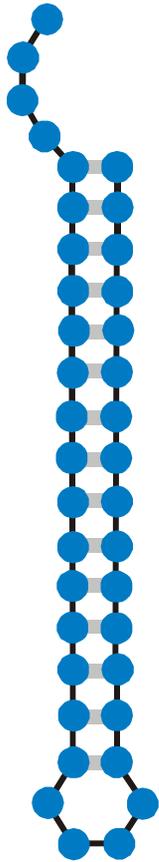
A connected neutral network formed by a common structure: $\lambda > \lambda_{cr}$

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.

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

SUMMARY

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

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

279

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

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

Alphabet	Real time	Transitions	Major transitions	Sample size
AUGC	**398.3**	22.8	12.7	1199
GUC	448.9	30.5	16.5	611
GC	**1908.7**	38.7	20.1	278

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.

An RNA molecule with two (**meta**)**stable** conformations

Kinetic Folding of RNA Secondary Structures

Christoph Flamm, Walter Fontana, Ivo L. Hofacker, Peter Schuster. *RNA folding kinetics at elementary step resolution.* RNA **6**:325-338, 2000

Christoph Flamm, Ivo L. Hofacker, Sebastian Maurer-Stroh, Peter F. Stadler, Martin Zehl. *Design of multistable RNA molecules.* RNA **7**:325-338, 2001

The Folding Algorithm

A sequence I specifies an energy ordered set of compatible structures $\mathcal{S}(I)$:

$$\mathcal{S}(I) = \{S_0, S_1, \ldots, S_m, O\}$$

A trajectory $\mathfrak{T}_k(I)$ is a time ordered series of structures in $\mathcal{S}(I)$. A folding trajectory is defined by starting with the open chain O and ending with the global minimum free energy structure S_0 or a metastable structure S_k which represents a local energy minimum:

$$\mathfrak{T}_0(I) = \{O, S(1), \ldots, S(t-1), S(t), S(t+1), \ldots, S_0\}$$
$$\mathfrak{T}_k(I) = \{O, S(1), \ldots, S(t-1), S(t), S(t+1), \ldots, S_k\}$$

Transition probabilities $P_{ij}(t) = \mathcal{P}rob\{S_i \rightarrow S_j\}$ are defined by

$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t)\, \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

$$P_{ji}(t) = P_j(t)\, k_{ji} = P_j(t)\, \exp(-\Delta G_{ji}/2RT) / \Sigma_j$$

$$\Sigma_k = \sum_{k=1, k \neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

The symmetric rule for transition rate parameters is due to Kawasaki (K. Kawasaki, *Diffusion constants near the critical point for time depen-dent Ising models.* Phys.Rev. **145**:224-230, 1966).

Formulation of kinetic RNA folding as a stochastic process
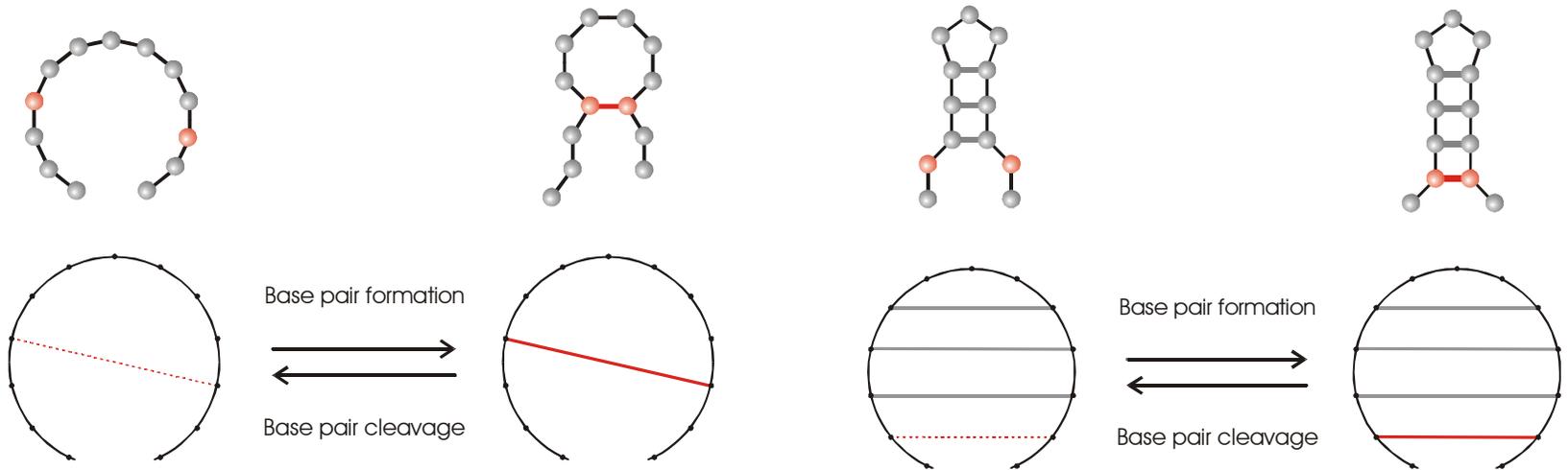
Base pair formation and base pair cleavage moves for nucleation and elongation of stacks
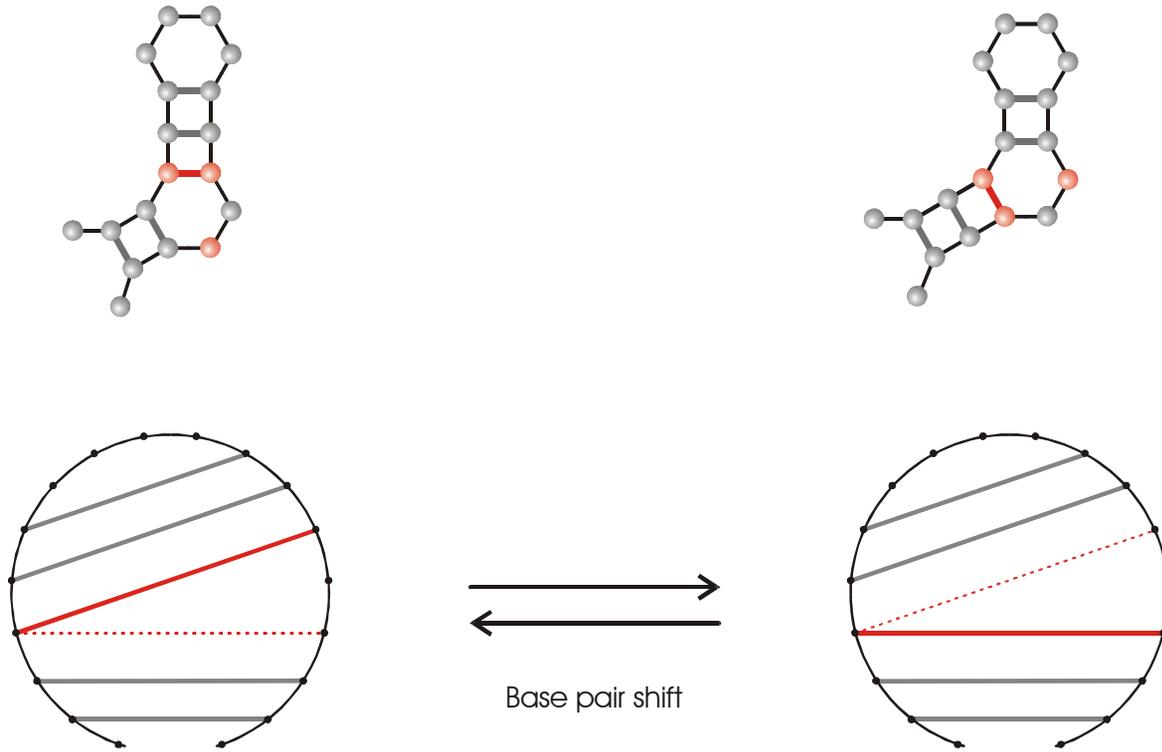
Base pair shift move of class 1: Shift inside internal loops or bulges
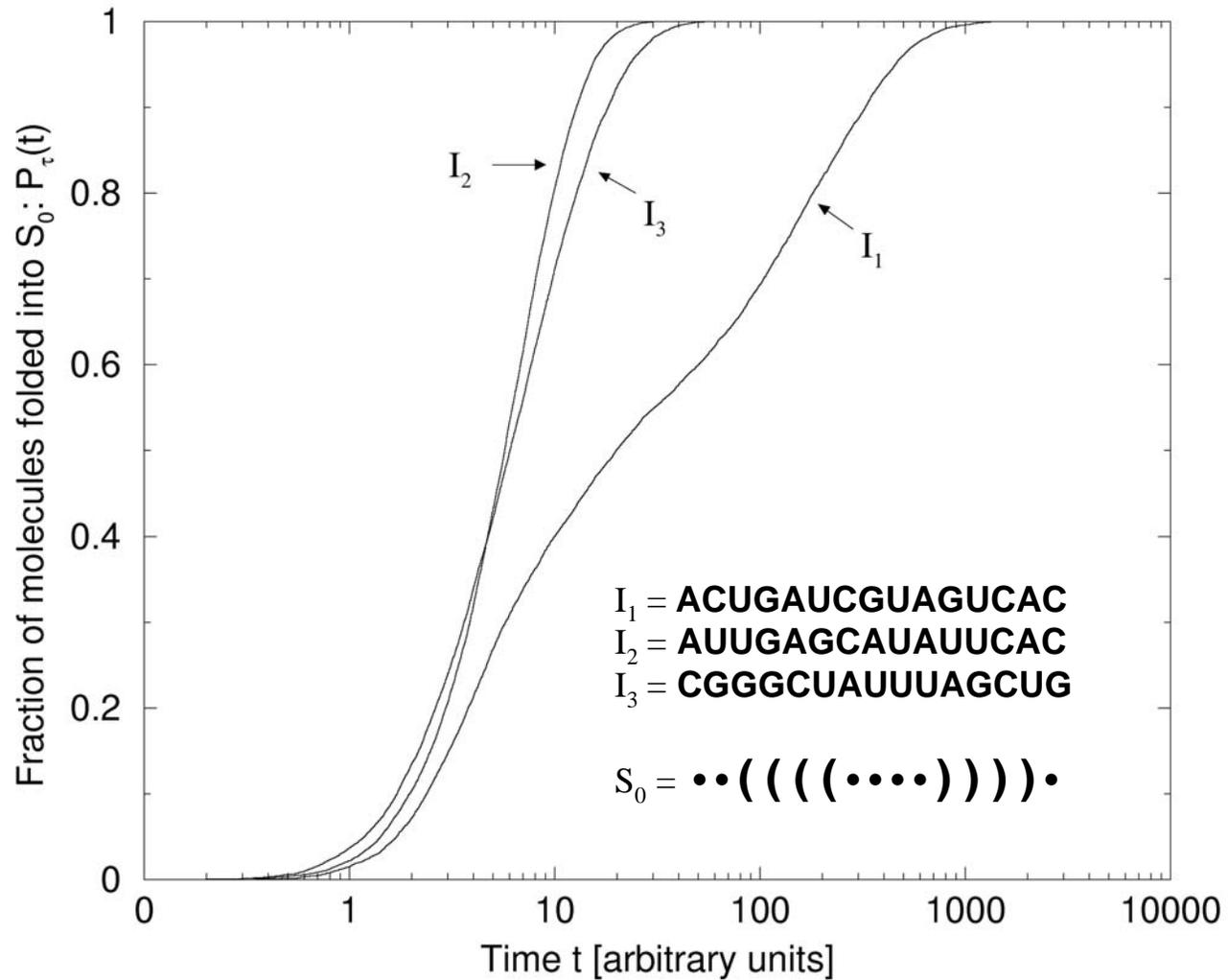
Mean folding curves for three small RNA molecules with different folding behavior

Free energy ΔG^0

$S_5^{(h)}$

$S_4^{(h)}$

$S_3^{(h)}$

$S_1^{(h)}$

$S_2^{(h)}$

$S_8^{(h)}$

$S_7^{(h)}$

$S_9^{(h)}$

$S_6^{(h)}$

Suboptimal conformations
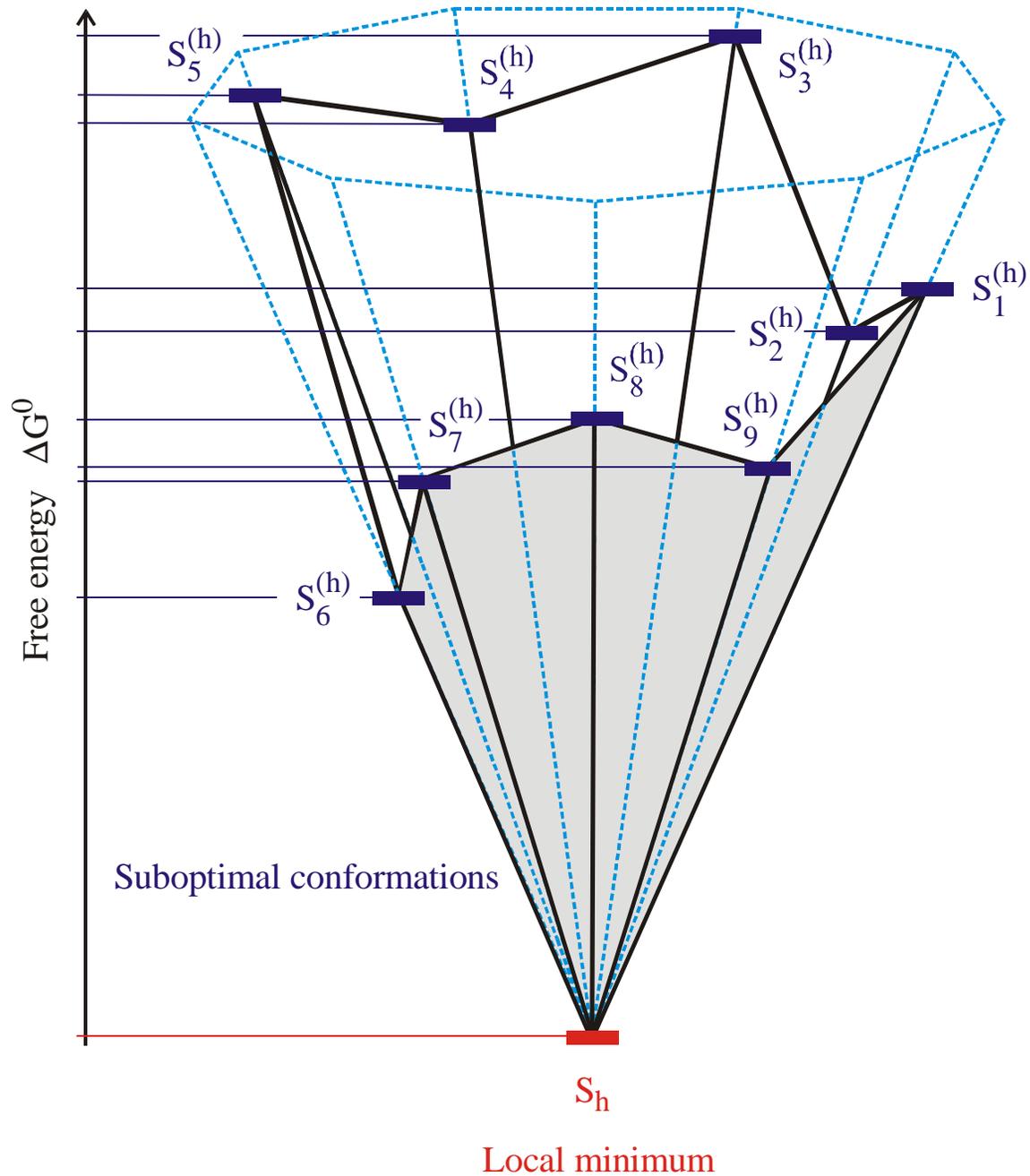
Search for local minima in conformation space

S_h

Local minimum

Free energy ΔG^0

S_k

Saddle point $T_{\{k}$

"Reaction coordinate"

Free energy ΔG^0

$T_{\{k}$

$S_{\{}$

S_k
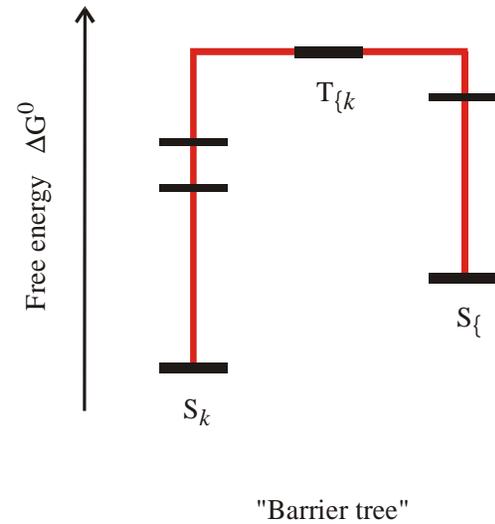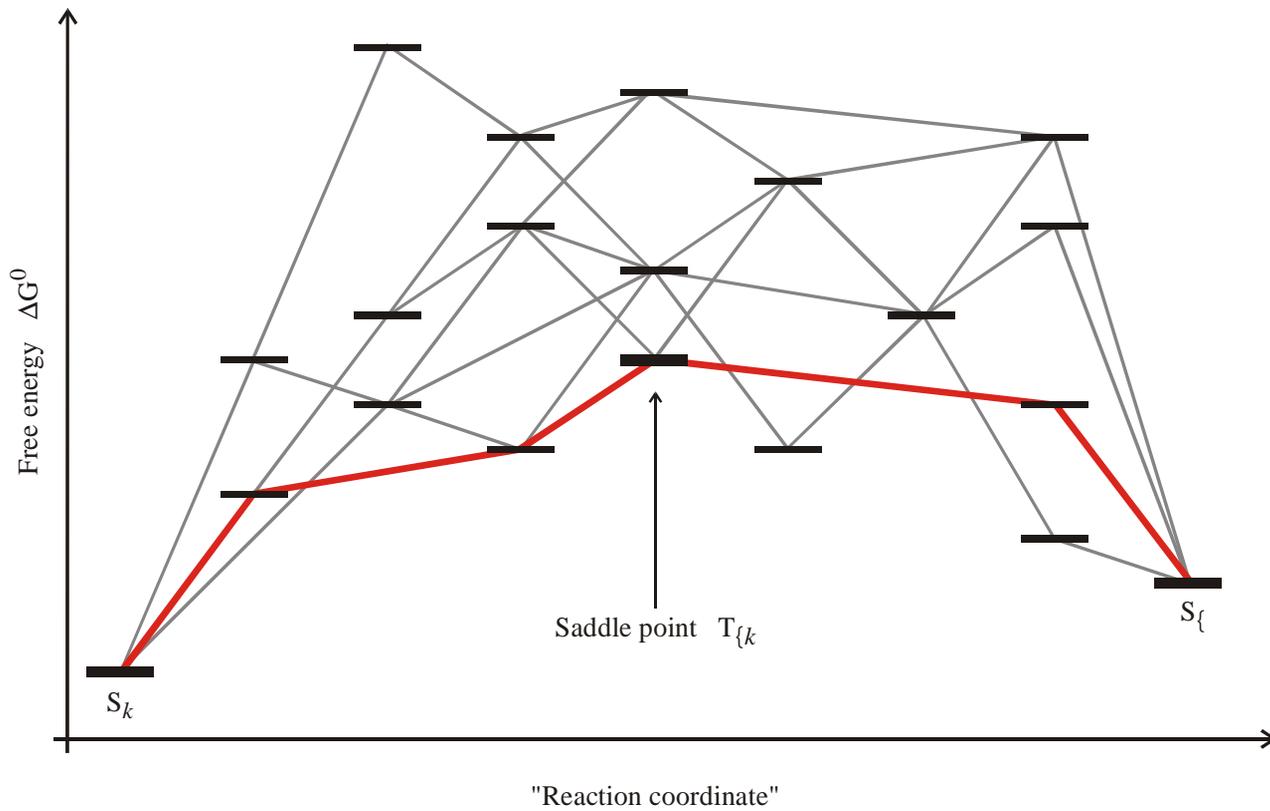
"Barrier tree"

Definition of a ‚barrier tree'

1.70

.......((....))
S_3

1.30

1.70

1.80

3.30

3.80

((.....))......
S_2

....
O

I_1 = **ACUGAUCGUAGUCAC**

((((....))))...
S_1

..((((....)))).
S_0
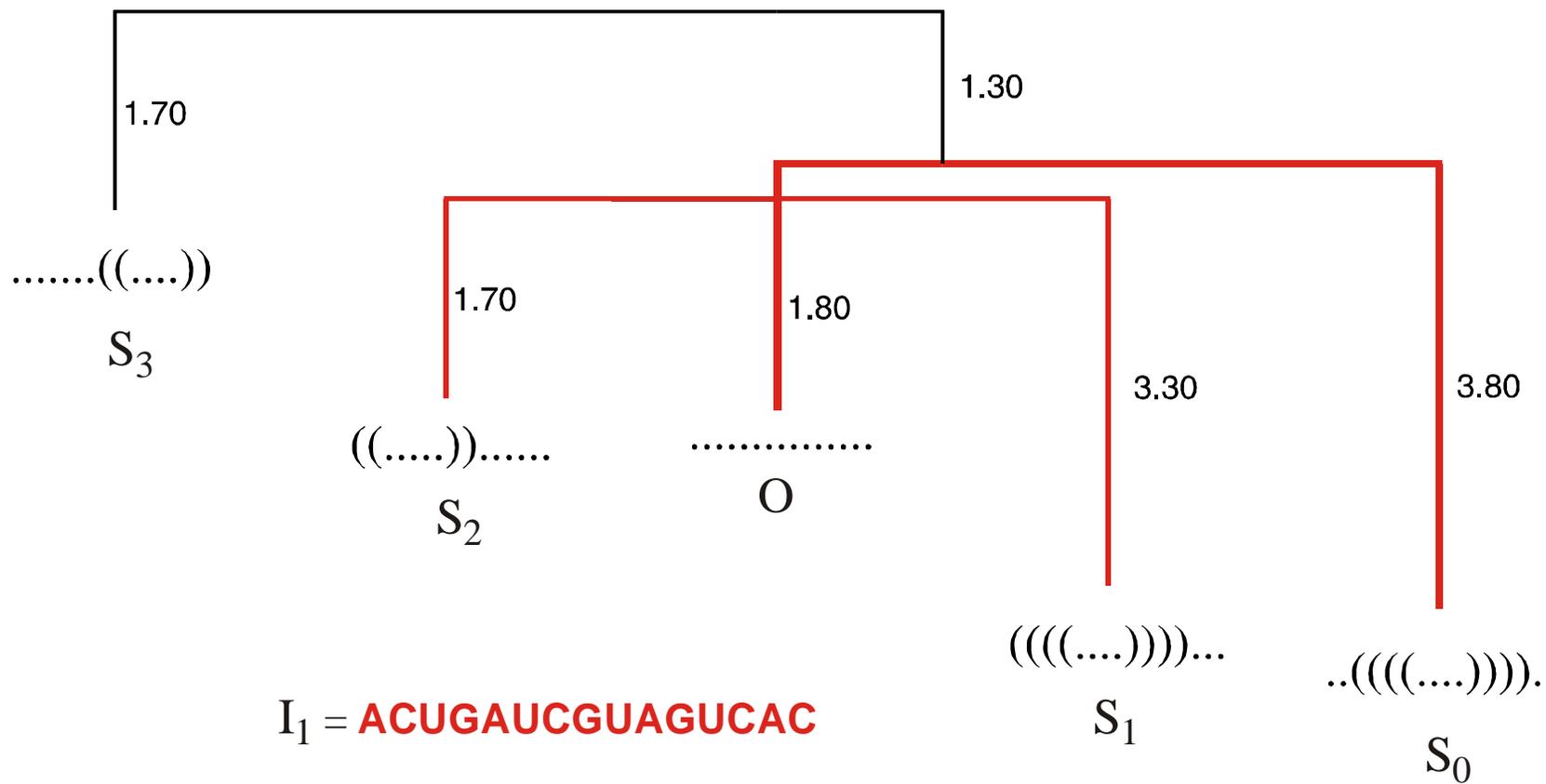
Example of an unefficiently folding small RNA molecule with n = 15

S_4

0.70

((........))....

0.70

((.....))......

S_2

..((....)).....

S_3

1.30

...((.....))...

S_1

2.10

....

O

3.20

..((((....)))).

S_0

$I_2 = $ **AUUGAGCAUAUUCAC**
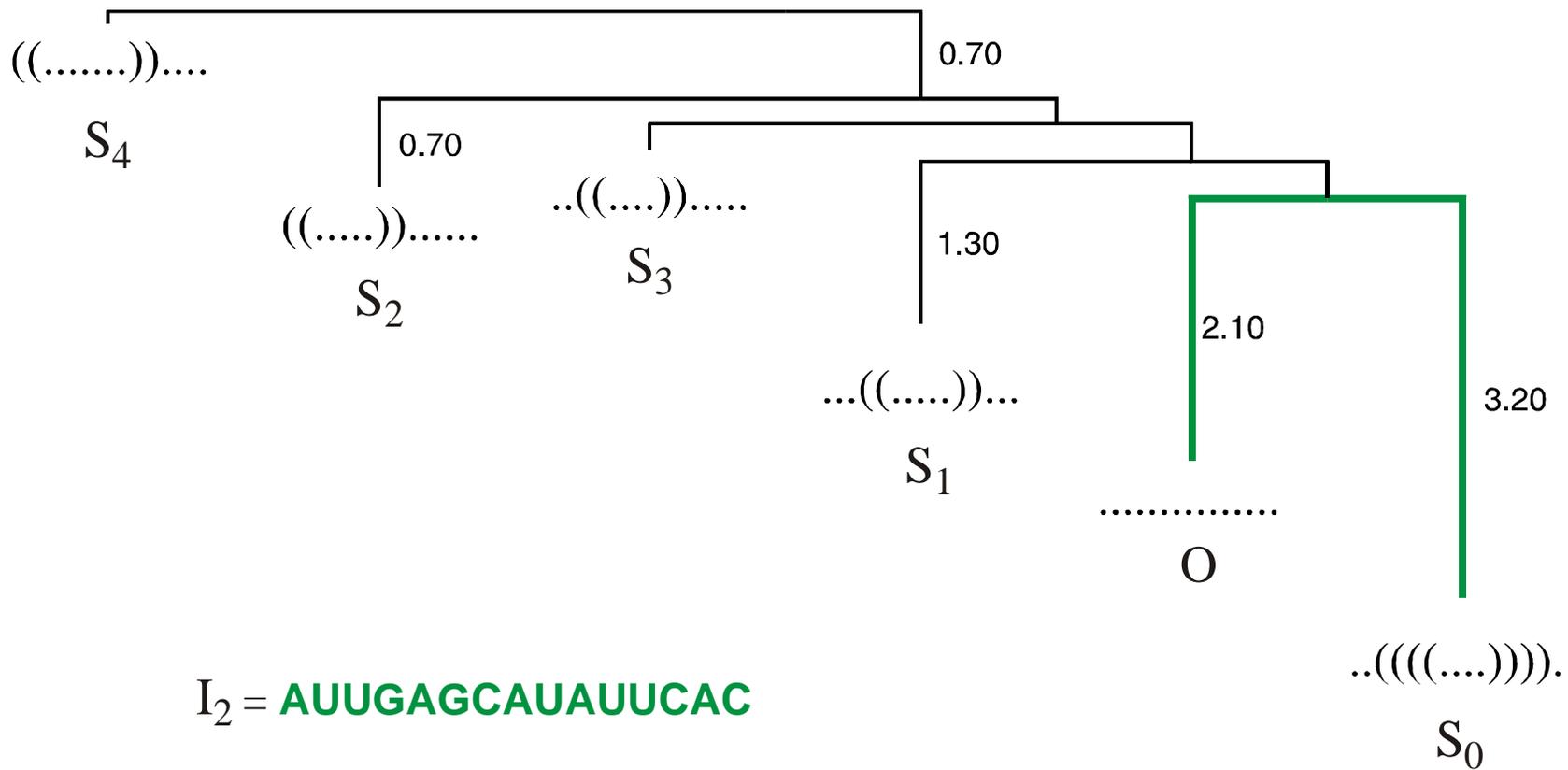
Example of an easily folding small RNA molecule with n = 15

.....((......))
S_3

0.50

0.40

0.90

.((....))......
S_2

1.40

.((.....)).....
S_1

1.10

....
O

7.50

I_3 = **CGGGCUAUUUAGCUG**
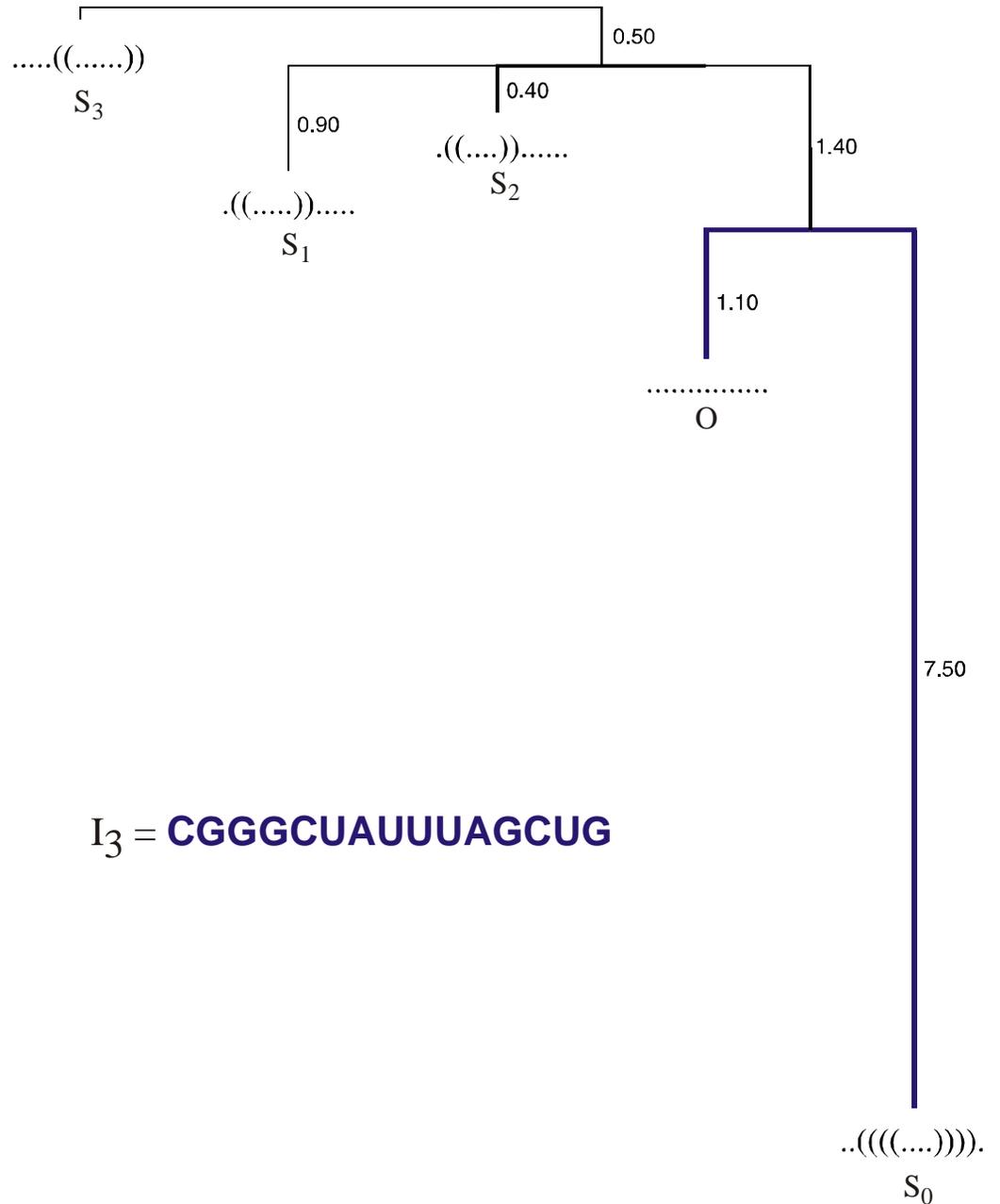
Example of an easily folding
and especially stable small
RNA molecule with n = 15

..((((....)))).
S_0

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA
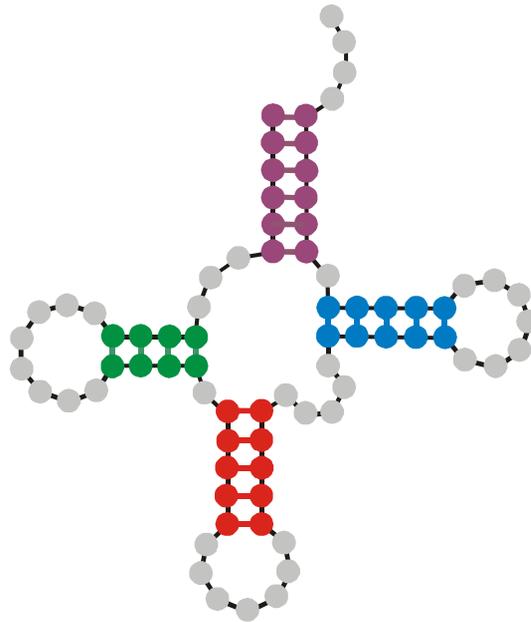
Kinetic folding of phenylalanyl-tRNA
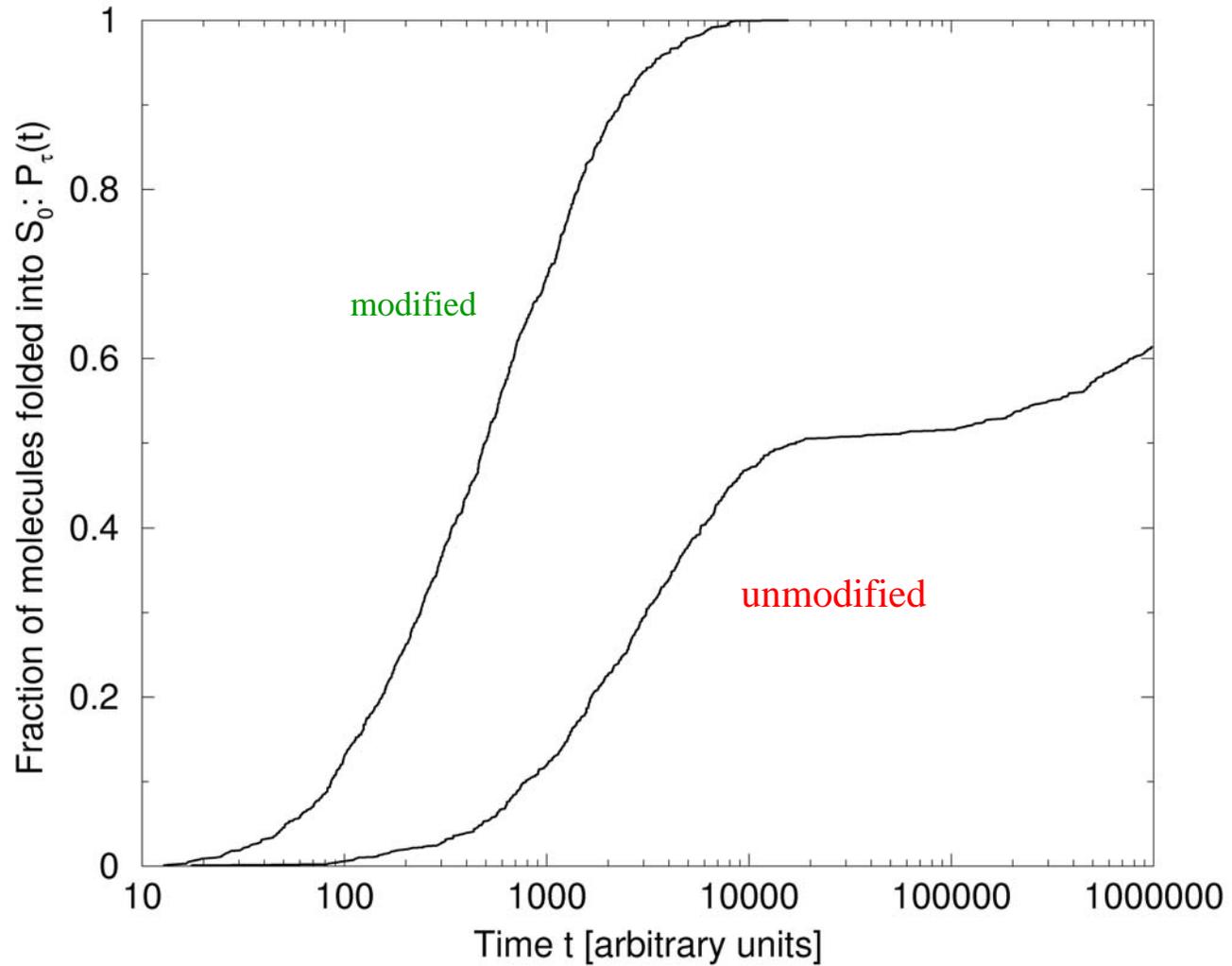
Folding dynamics of tRNA**phe** with and without modified nucelotides
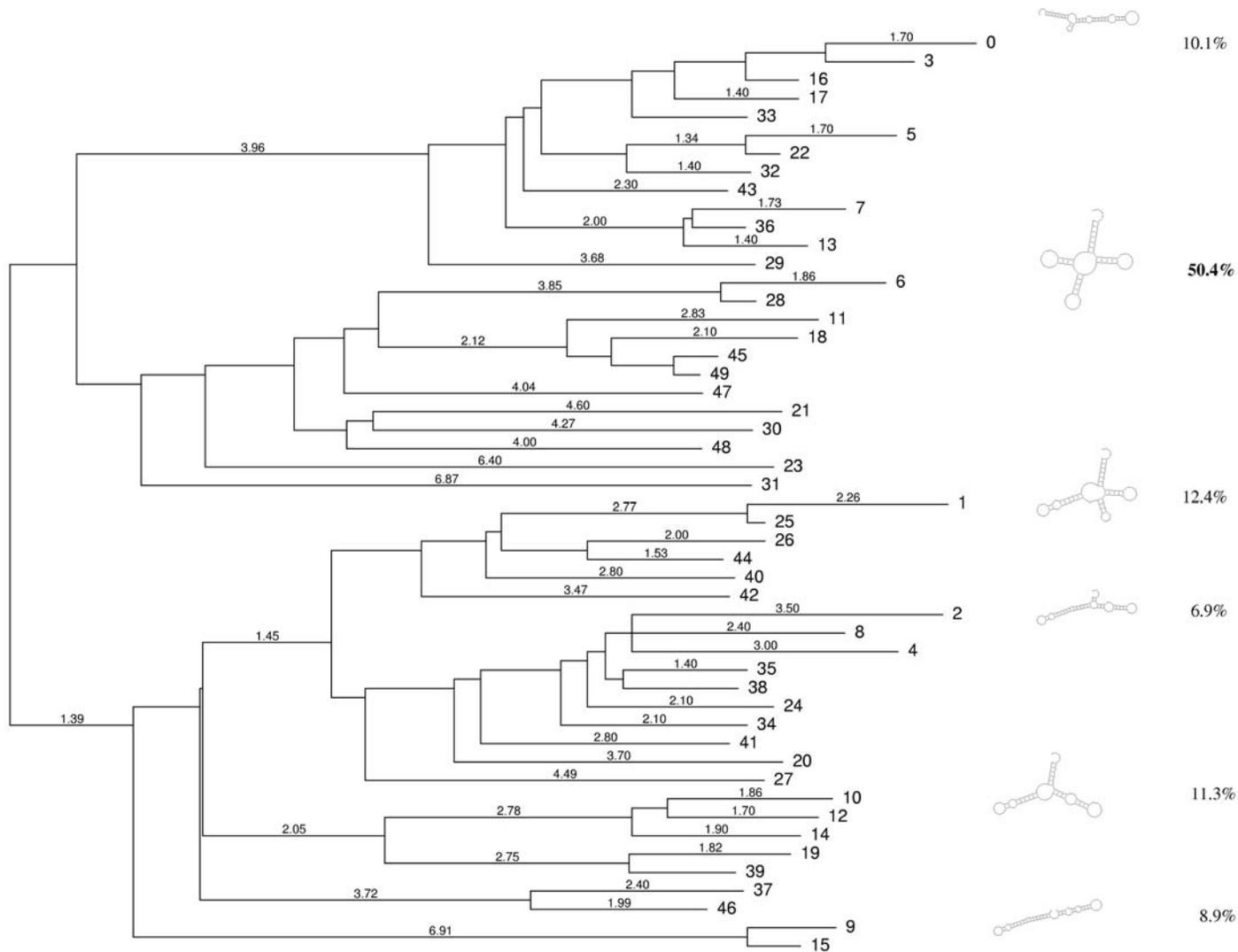
Barrier tree of tRNA**phe** without modified nucelotides
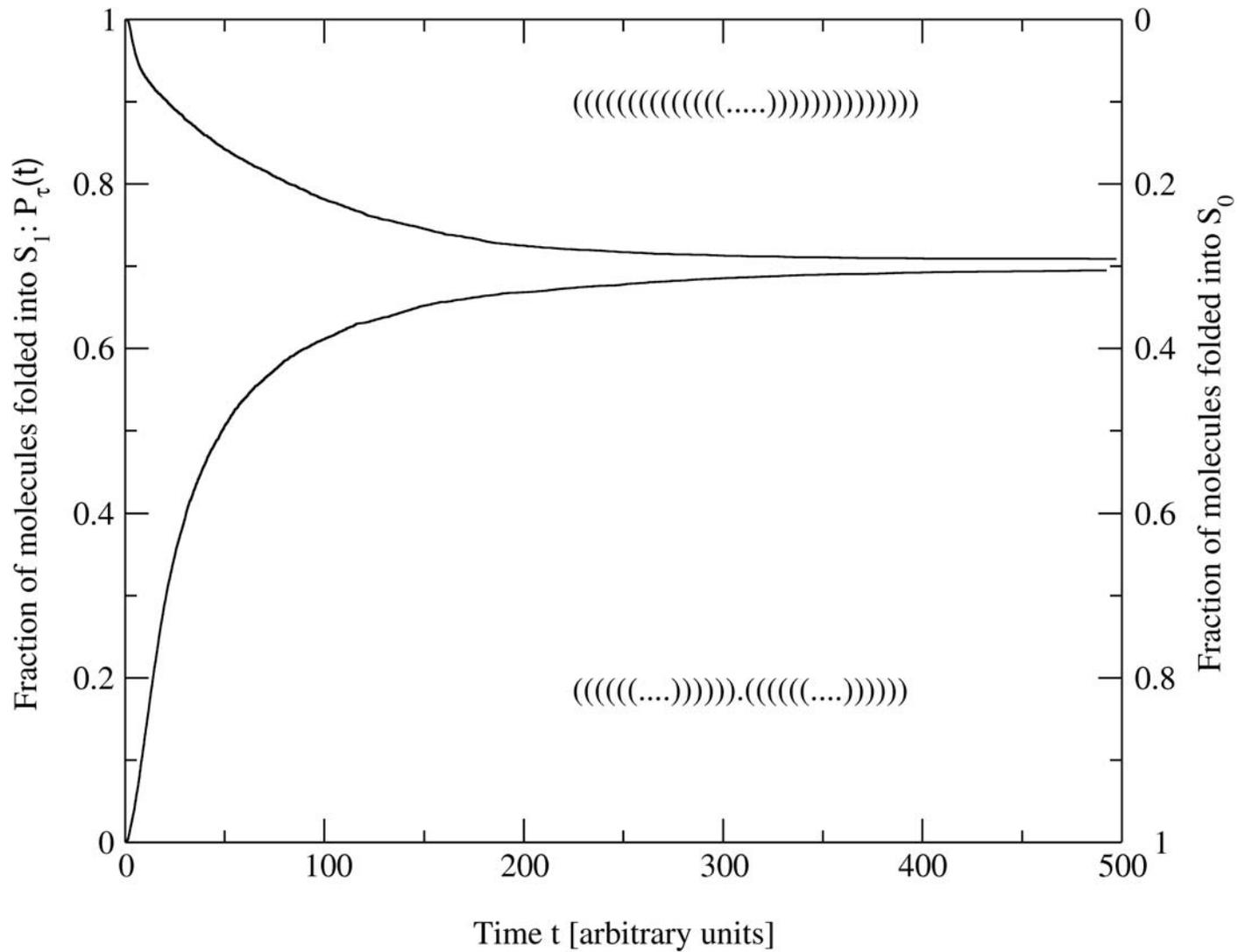
Folding dynamics of the sequence **GGCCCCUUUGGGGGCCAGACCCCUAAAAAGGGUC**

3'-end

CUGGGAAAAAUCCCCAGAGACCGGGGGGUUUCCCCGG

5'-end

Minimum free energy conformation S_0

Suboptimal conformation S_1

One sequence is compatible with two structures

Barrier tree of a sequence with two conformations

Structure S_0
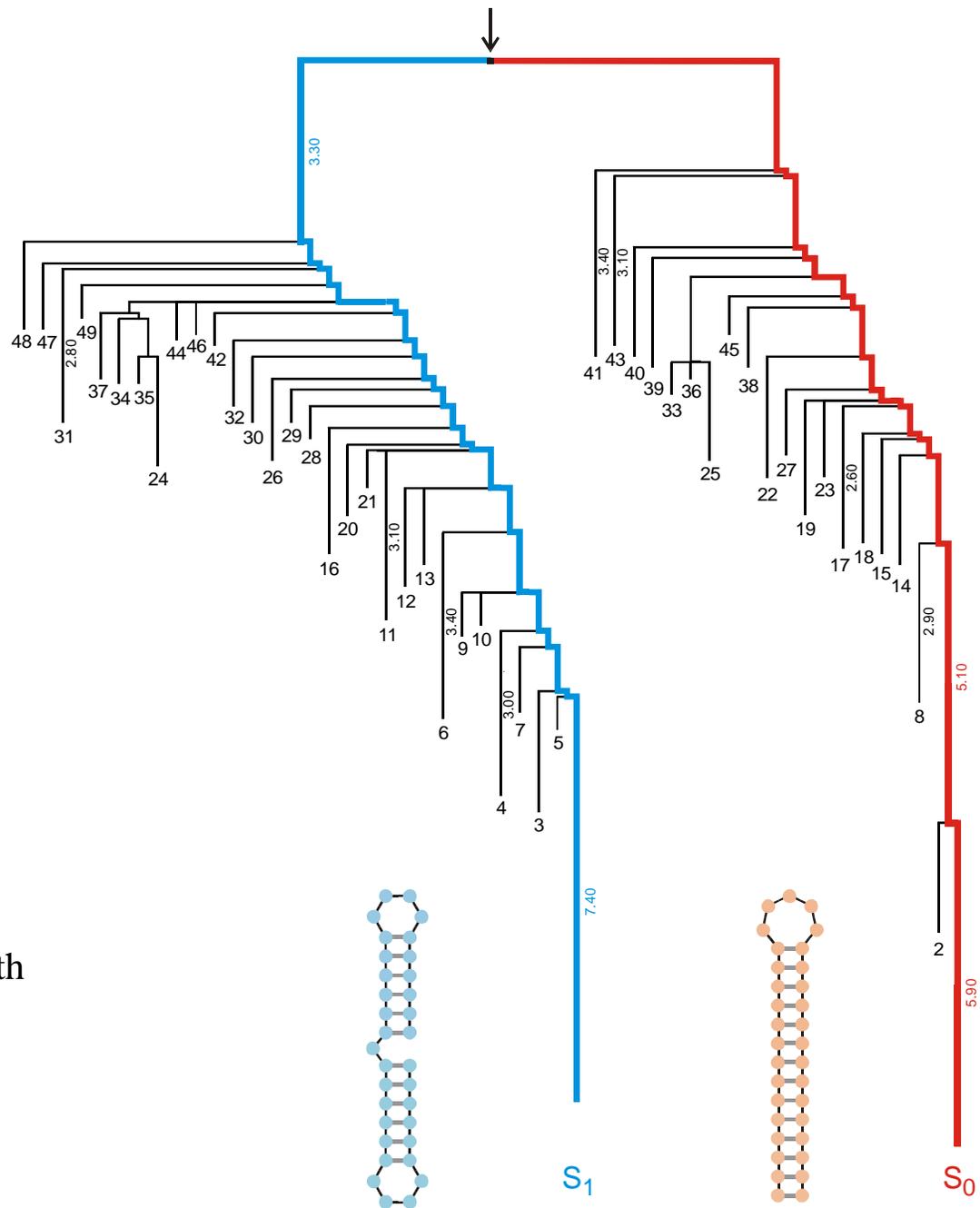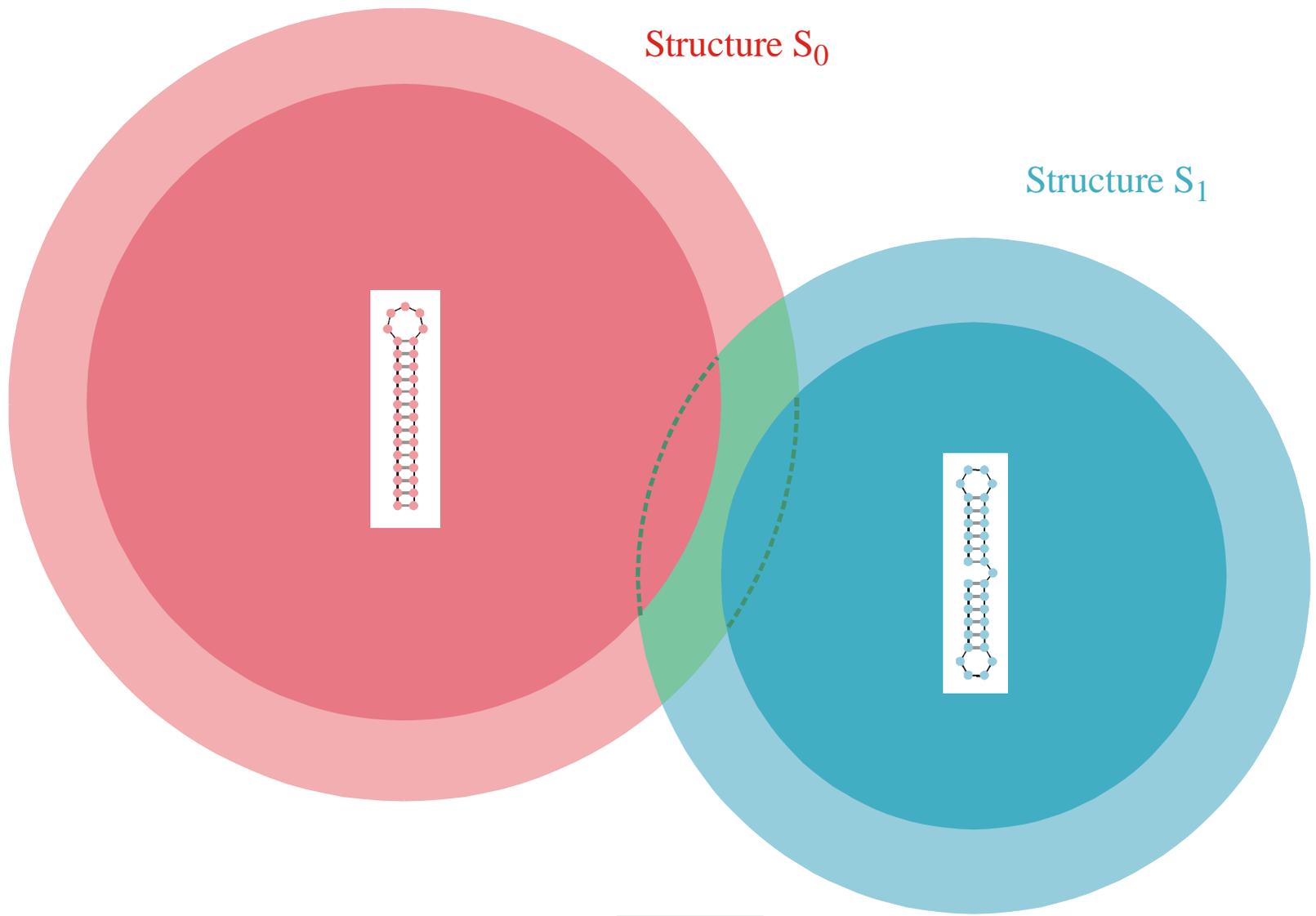
Structure S_1

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

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

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

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

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

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

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

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/CH₃Cl 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 dispar-

ate 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**)
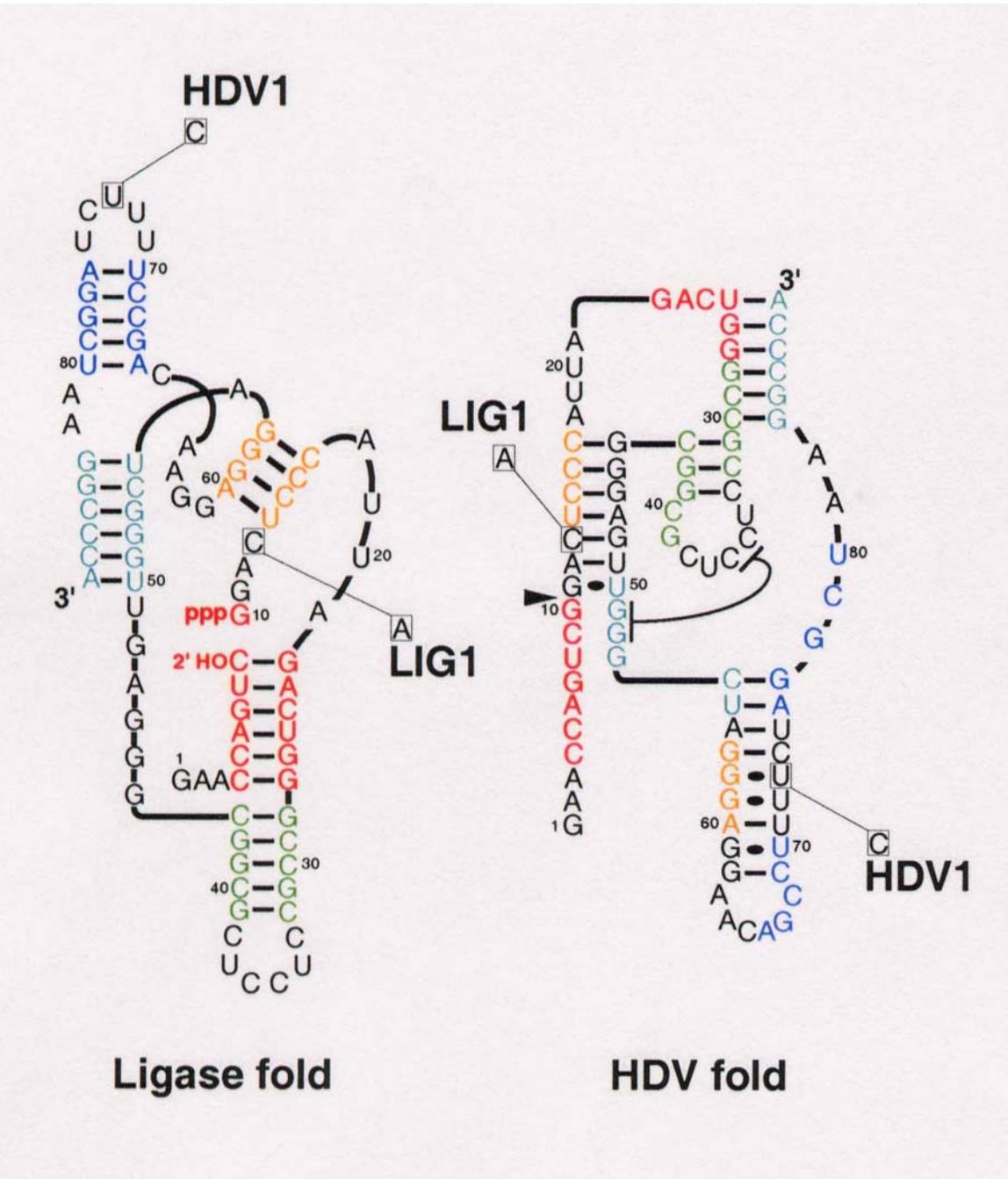
The sequence at the *intersection*:

An RNA molecules which is 88 nucleotides long and can form both structures
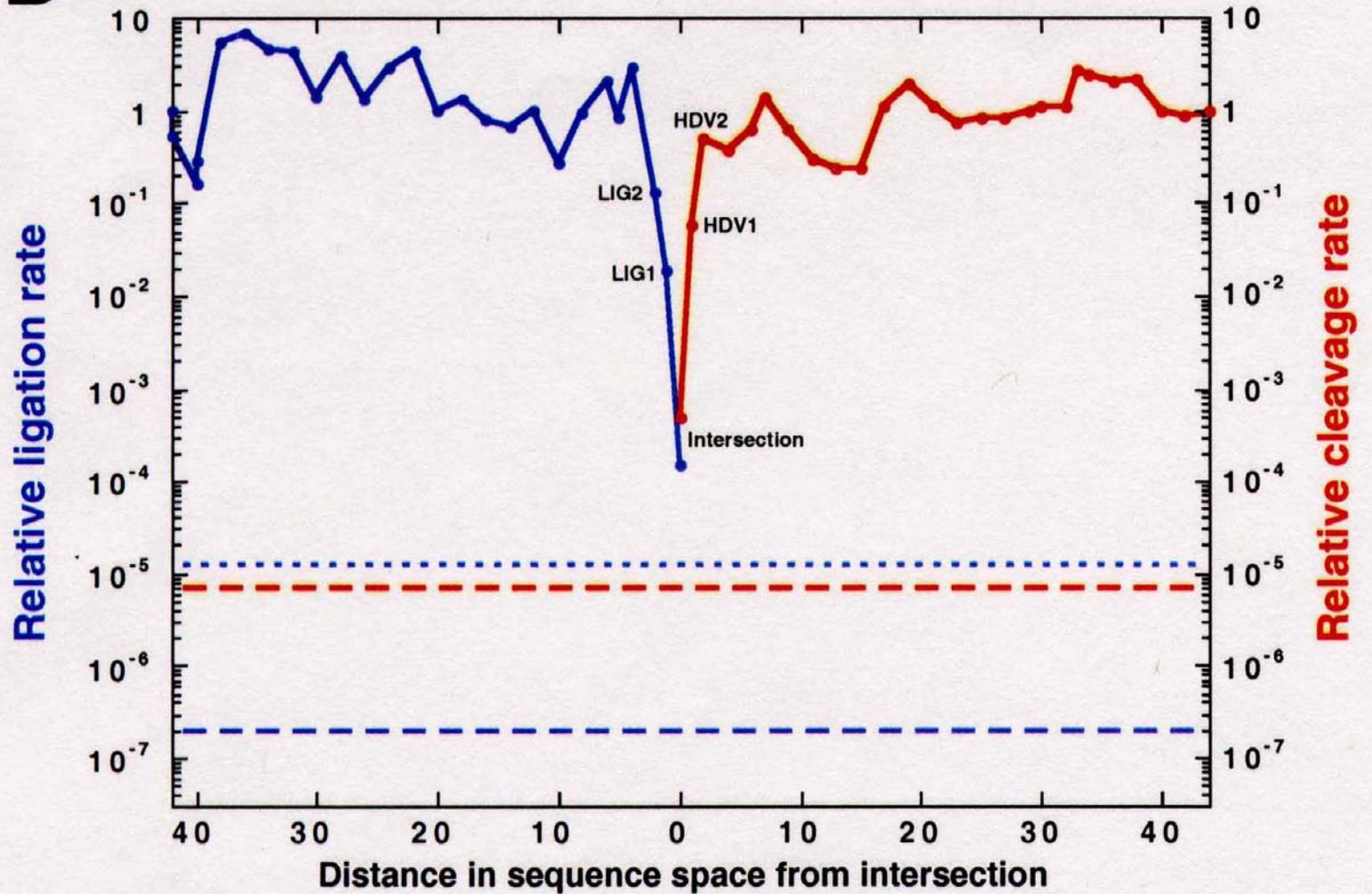
Two neutral walks through sequence space with conservation of structure and catalytic activity
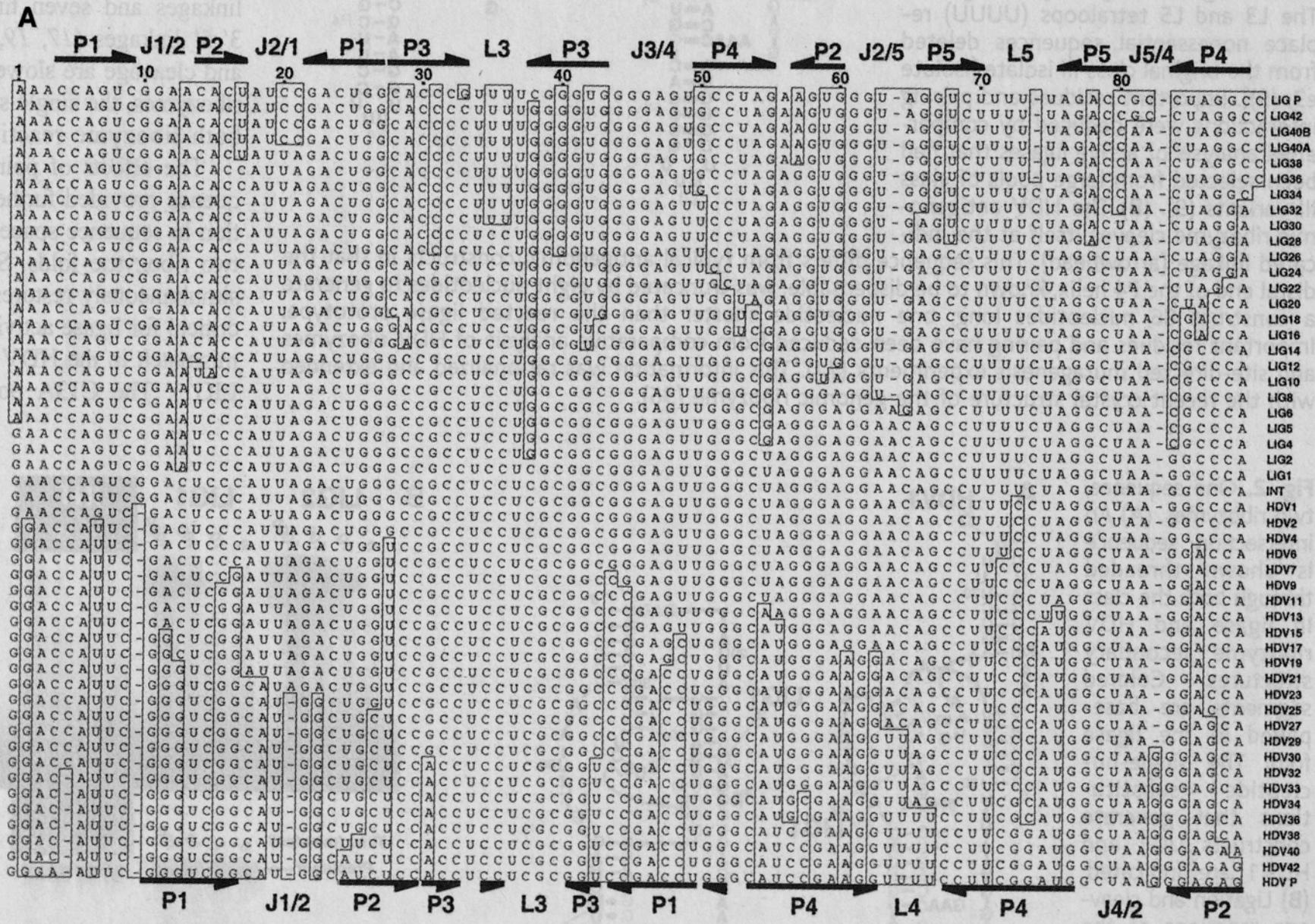
Sequence of mutants from the intersection to both reference ribozymes

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.

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 2004.

A   **B**   **C**

CUGUUUUUGCA**GCAGAAGC**U**GCAGAAGC**AGCUUCUGUUG

3   15   24   36

15
  G A
 A   A
  C-G
  G-C
  A-U
A C-G **B**
  G-C
  U-A
  U-G 24
  U-A
  U-A
  U-G
3 G-C
  U-A   **C**
5' C-G C U U C U G U U G 3'
               36

24
  G A
 A   A
  C-G
  G-C
  U-A
B C-G **C**
  G-C
  A-U
  A-U
15 G-C
  A-U
  C-G 36
  G-U
  A-U   **A**
5' CUGUUUUUG C-G 3'
            3

-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.

1D   **R**   **2D**

GGGUGGAAC**CACGAG**GUUC**CACGAG**GAAC**CACGAG**GUUCCUCCC

3   13   **G** 23   33   44

1D   **2D**

C G¹³   **C G**³³
A A   A A
C•G   **C•G**
C•G   C•G
A•U   A•U
A•U   A•U
G•C   G•C
G•**C**   **G**•C
U•**A/G**   **A**•U
³G•**C** —— **G**•C⁴⁴
GG   **R**   23   CC
5'   3'
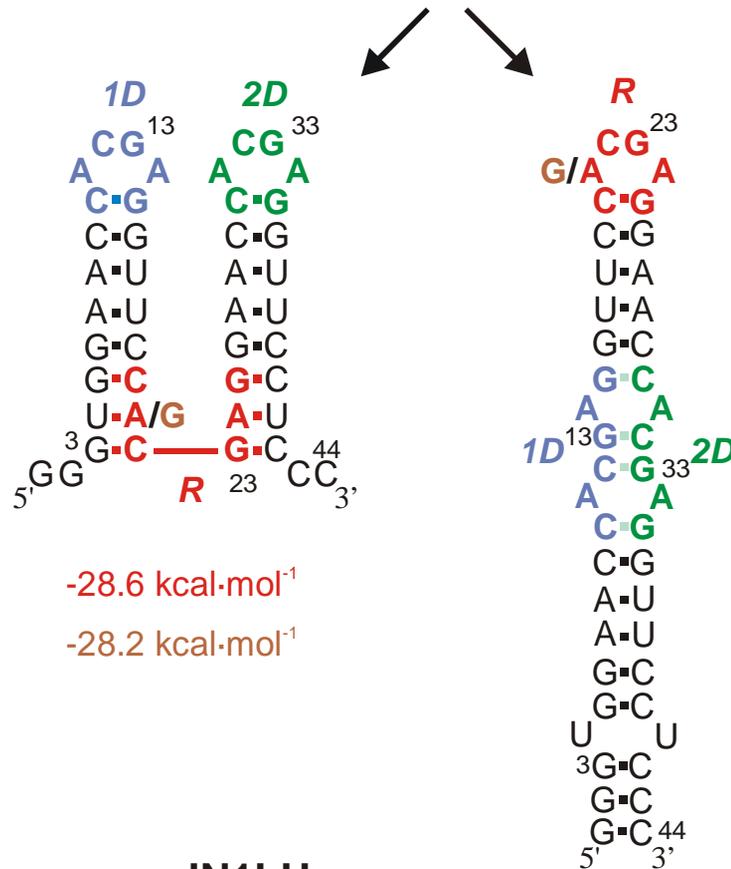
-28.6 kcal·mol⁻¹

-28.2 kcal·mol⁻¹

R

C G²³
G/A   A
C•G
C•G
U•A
U•A
G•C
G•C
A   A
1D ¹³**G•C** **2D**
C•G³³
A   A
C•G
C•G
A•U
A•U
G•C
G•C
U   U
³G•C
G•C
G•C⁴⁴
5'   3'

-28.6 kcal·mol⁻¹

-31.8 kcal·mol⁻¹

JN1LH

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.

J1LH barrier tree

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 2004.

metastable

I

II

tac

ucb

predicted saddle point

I

tac *ucb*

stable

tac *ucb*
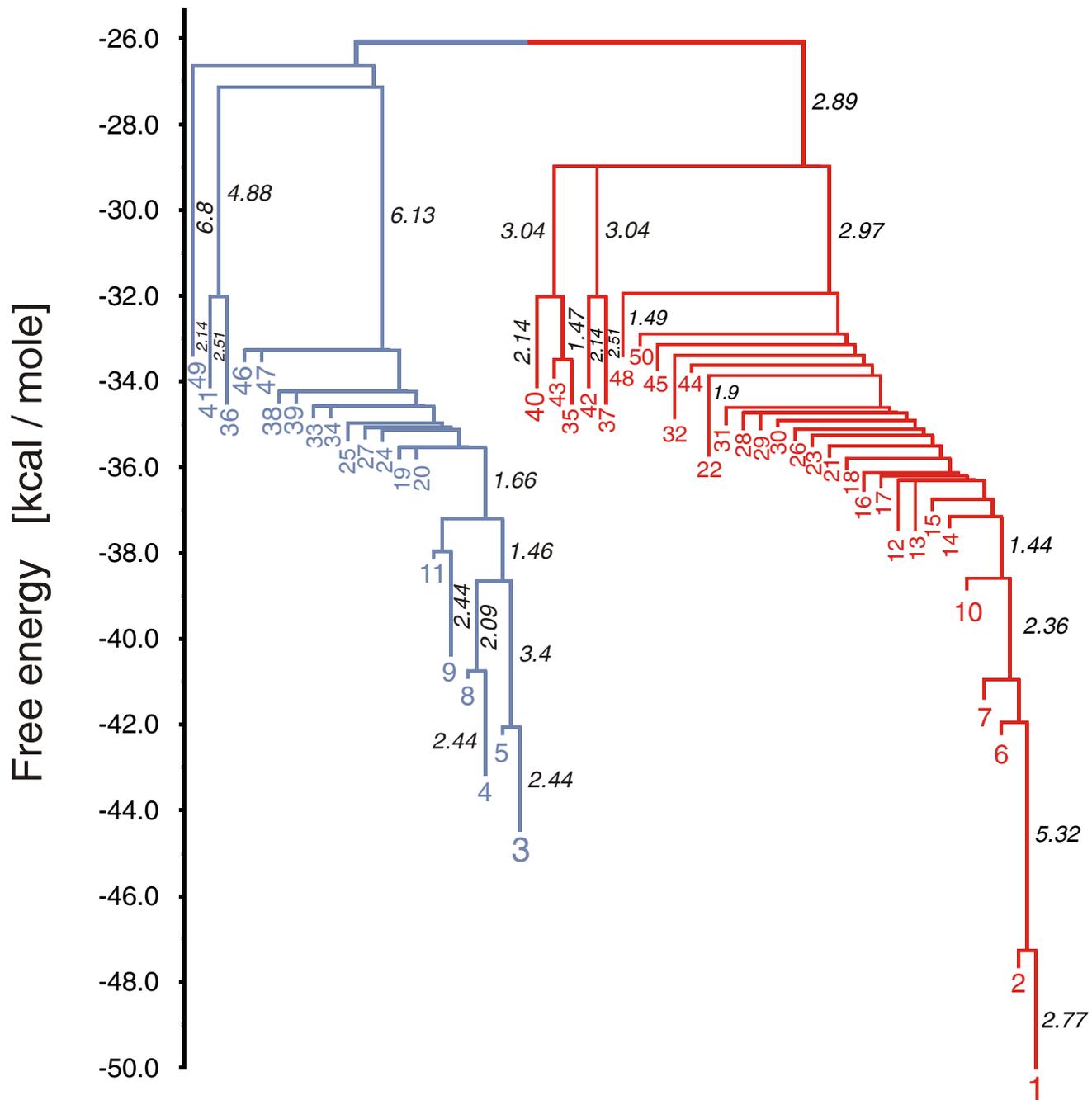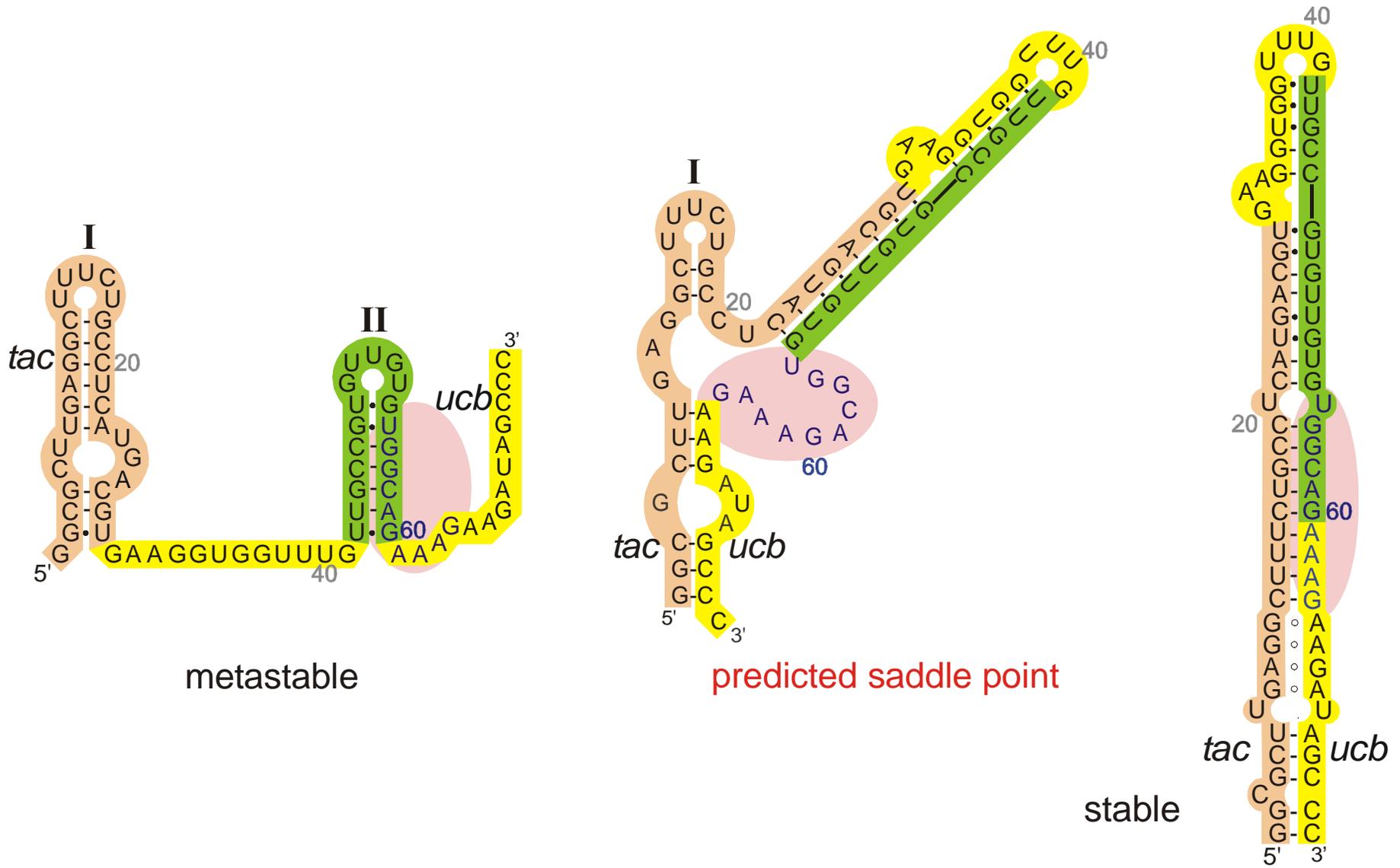
Transition from the metastable to the stable comformation
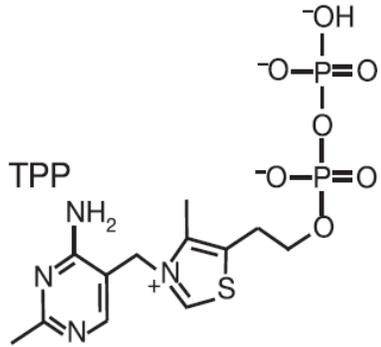
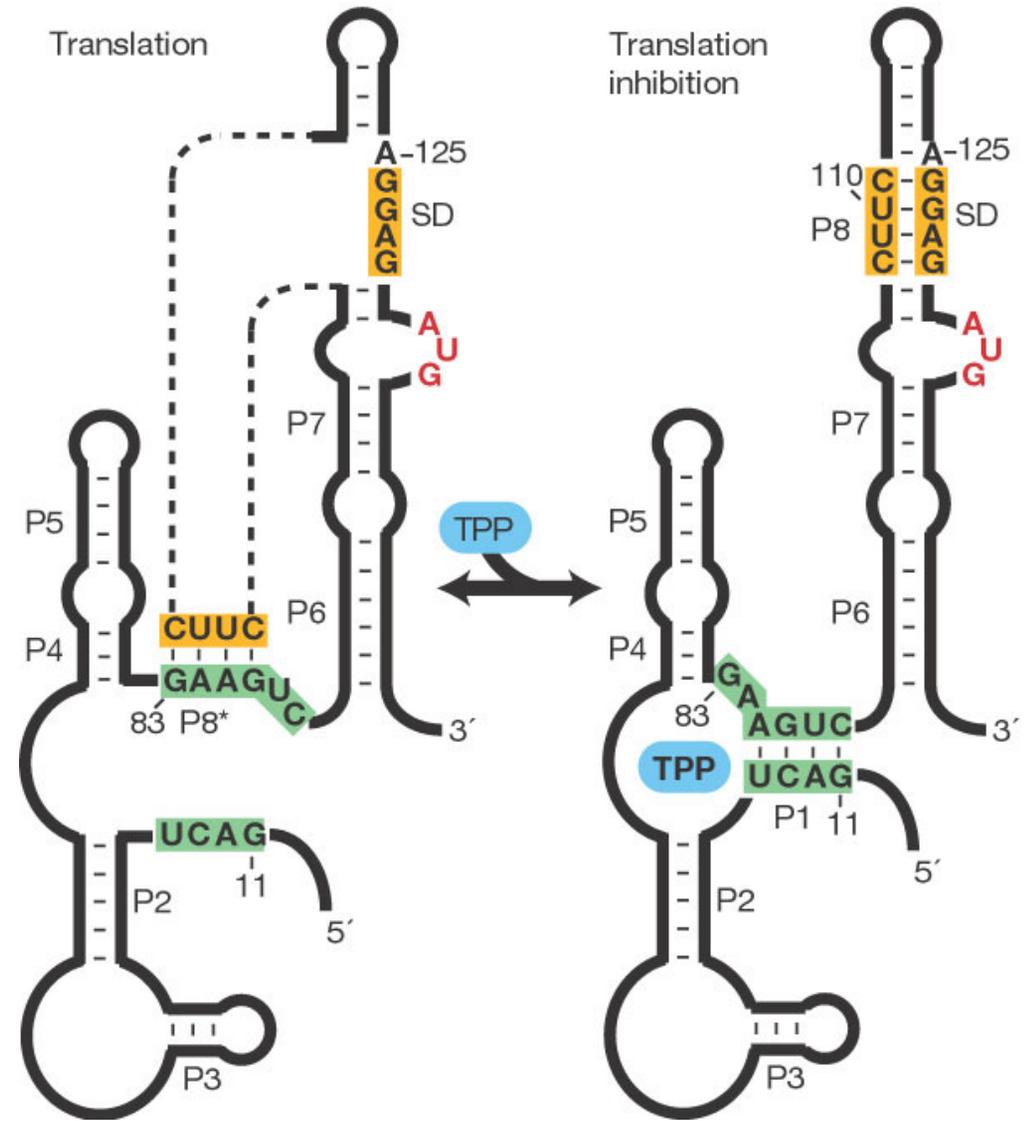Wade Winkler, Ali Nahvi, and Ronald R. Breaker, *Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression*. Nature **419**, 952-956, 2002.

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

Evidence for **neutral networks** and **shape space covering**

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

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 muta-

tions. 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

Acknowledgement of support

Universität Wien

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, MA

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

Peter Stadler, **Bärbel Stadler**, Universität Leipzig, GE

Jord Nagel, **Kees Pleij**, Universiteit Leiden, NL

Ivo L.Hofacker, **Christoph Flamm**, Universität Wien, AT

Andreas Wernitznig, **Michael Kospach**, Universität Wien, AT
Ulrike Langhammer, **Ulrike Mückstein**, **Stefanie Widder**
Jan Cupal, **Kurt Grünberger**, **Andreas Svrček-Seiler**, **Stefan Wuchty**
Stefan Bernhart, **Lukas Endler**

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

Web-Page for further information:

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