

Evolutionary Optimization at the Molecular Level

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria

and

The Santa Fe Institute, Santa Fe, New Mexico, USA



Physikalisches Kolloquium

TU Wien, 28.11.2005

Web-Page for further information:

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

Genotype, Genome

Collection of genes

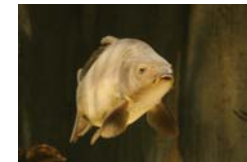
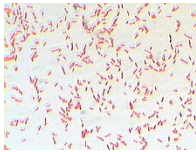
Developmental program

Unfolding of the genotype

Highly specific environmental conditions

Phenotype

Evolution explains the origin of species and their interactions

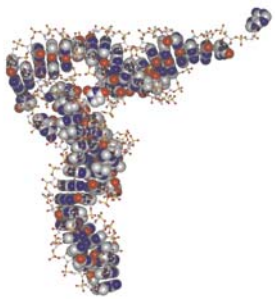


Genotype, Genome

CGGGATTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCCTGTGTTTCGATCCACAGAATTTCGCACCA

Omics

'the new biology is the chemistry of living matter'



evolution of RNA molecules, ribozymes and splicing, the idea of an RNA world, selection of RNA molecules, RNA editing, the ribosome is a ribozyme, small RNAs and RNA switches.

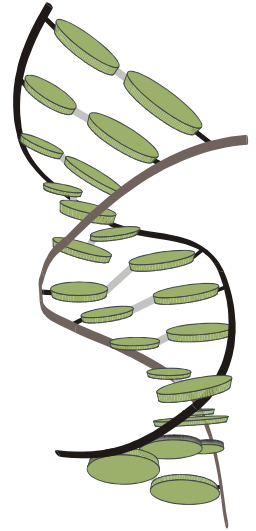
The exciting RNA story

Biochemistry
molecular biology
structural biology
molecular evolution
molecular genetics
systems biology
bioinformatics

Unfolding of the genotype

Phenotype

Highly specific environmental conditions



John Kendrew



Manfred Eigen



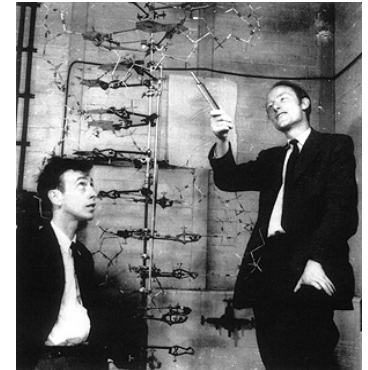
Molecular evolution
Linus Pauling and
Emile Zuckerkandl



Hemoglobin sequence
Gerhard Braunitzer



Max Perutz



James D. Watson und
Francis H.C. Crick



Three necessary conditions for Darwinian evolution are:

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

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

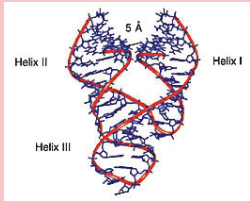
One important property of the Darwinian scenario is that **variations** in the form of mutations or recombination events occur **uncorrelated** with their **effects on the selection process**.

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

	Generation time	Selection and adaptation 10 000 generations	Genetic drift in small populations 10^6 generations	Genetic drift in large populations 10^7 generations
RNA molecules	10 sec 1 min	27.8 h = 1.16 d 6.94 d	115.7 d 1.90 a	3.17 a 19.01 a
Bacteria	20 min 10 h	138.9 d 11.40 a	38.03 a 1 140 a	380 a 11 408 a
Multicellular organisms	10 d 20 a	274 a 200 000 a	27 380 a 2×10^7 a	273 800 a 2×10^8 a

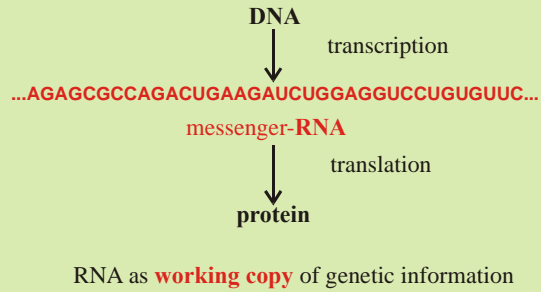
Time scales of evolutionary change

RNA as catalyst

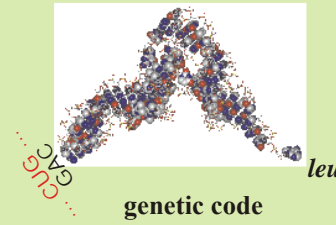


Ribozyme

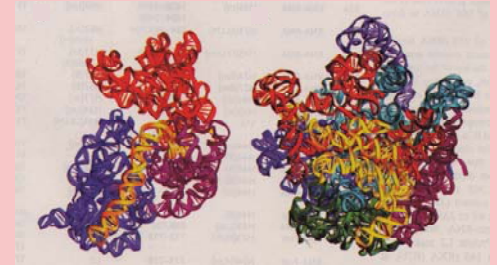
RNA as transmitter of genetic information



RNA as adapter molecule



RNA is the catalytic subunit in supramolecular complexes



The **ribosome** is a **ribozyme** !

RNA

RNA is modified by epigenetic control

RNA editing

Alternative splicing of messenger RNA

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

RNA as carrier of genetic information

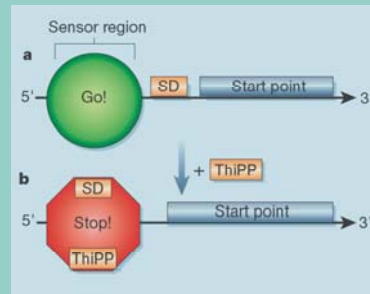
RNA viruses and retroviruses

RNA evolution *in vitro*

Evolutionary biotechnology

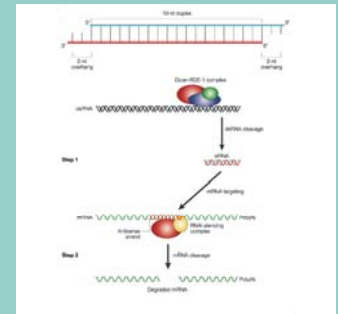
RNA aptamers, artificial ribozymes, allosteric ribozymes

Allosteric control of transcribed RNA



Riboswitches controlling transcription and translation through **metabolites**

RNA as regulator of gene expression

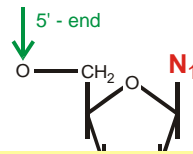


Gene silencing by small interfering RNAs

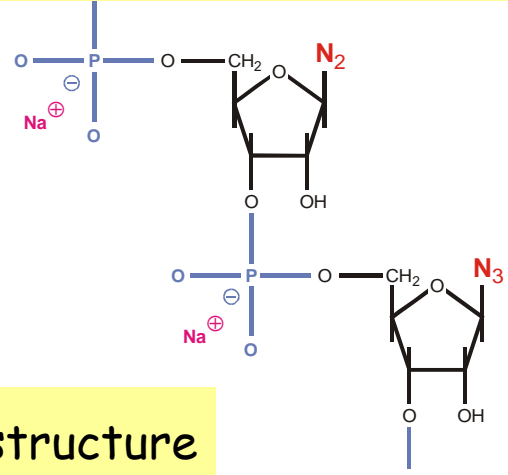
Functions of RNA molecules

1. RNA sequences and structures
2. Neutral networks
3. Evolutionary optimization of structure
4. Suboptimal structures and kinetic folding
5. Comparison of kinetic folding and evolution

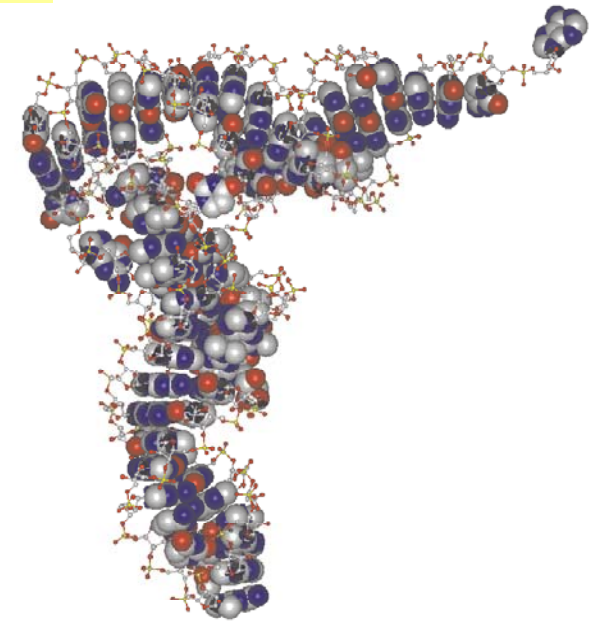
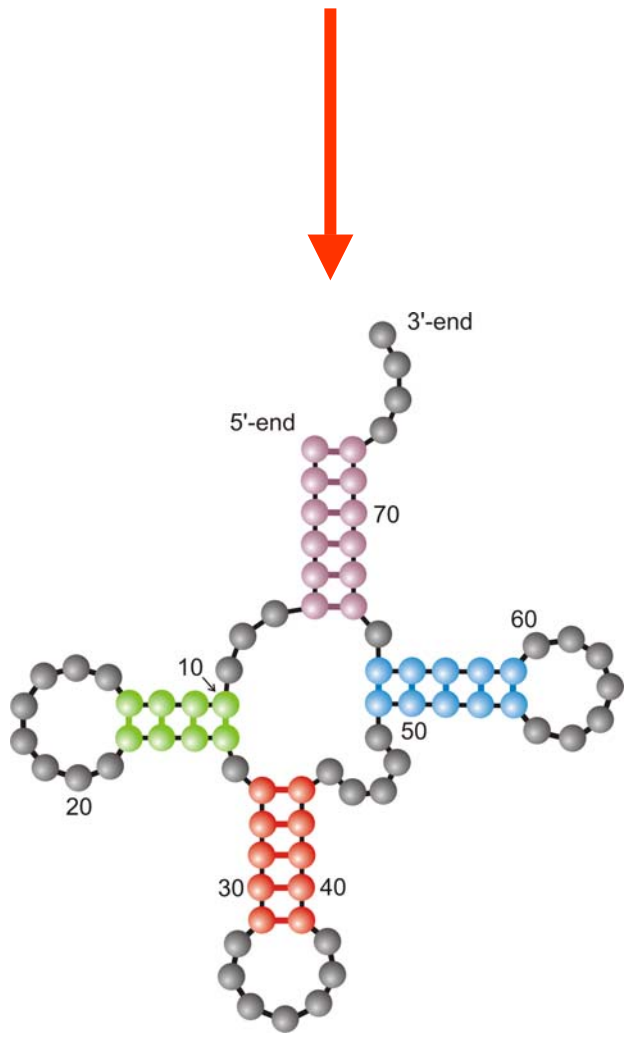
- 1. RNA sequences and structures**
2. Neutral networks
3. Evolutionary optimization of structure
4. Suboptimal structures and kinetic folding
5. Comparison of kinetic folding and evolution

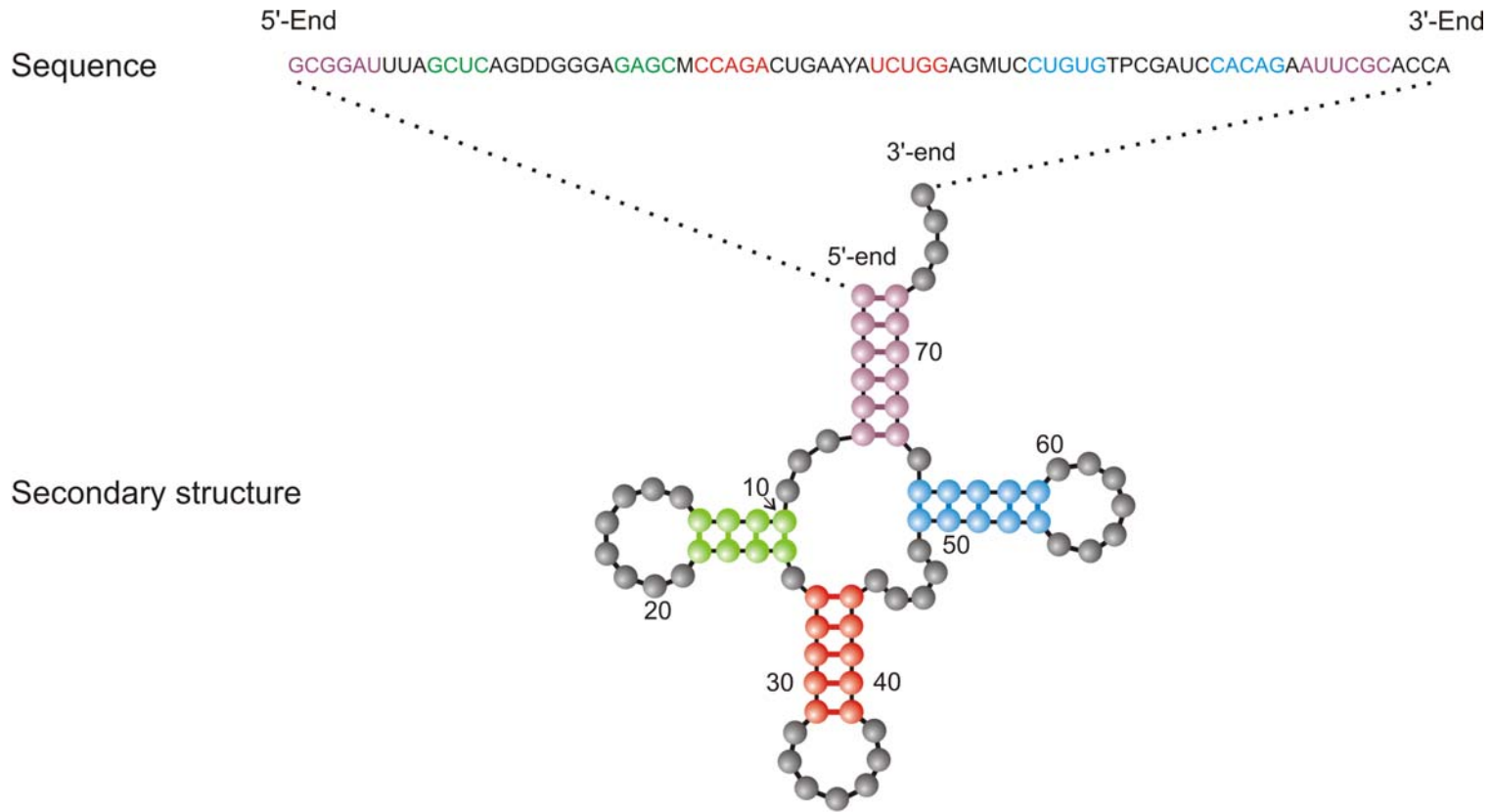


5'-end **GCGGAUUUAGCUC**AGUUGGGAGAG**CGCCAGACUGAAGAUCUGG**AGGUC**CUGUGUUCGAUCCACAGAAUUCGCACCA** 3'-end

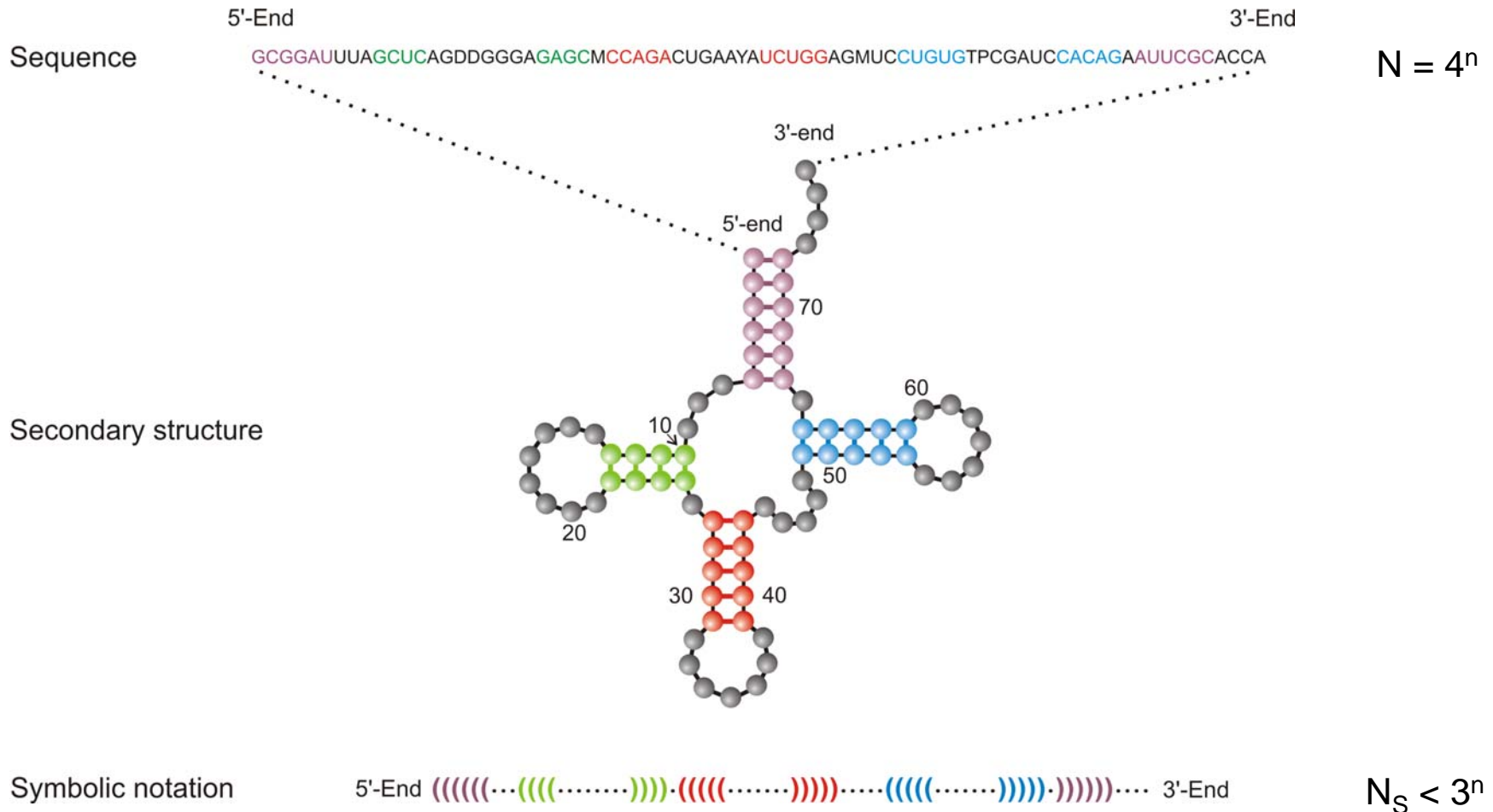


Definition of RNA structure





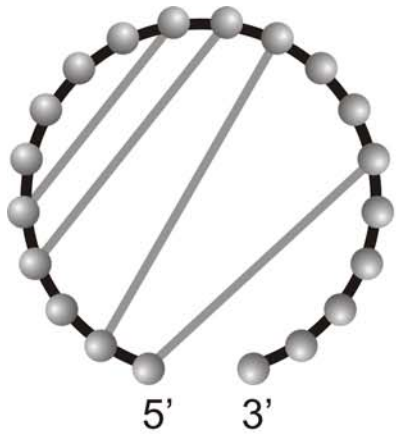
A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs



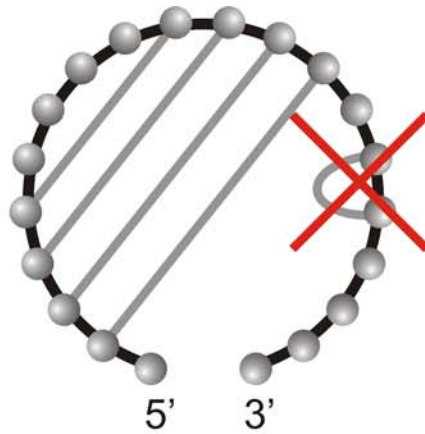
Criterion: Minimum free energy (mfe)

Rules: $_ (_) _ \in \{\text{AU, CG, GC, GU, UA, UG}\}$

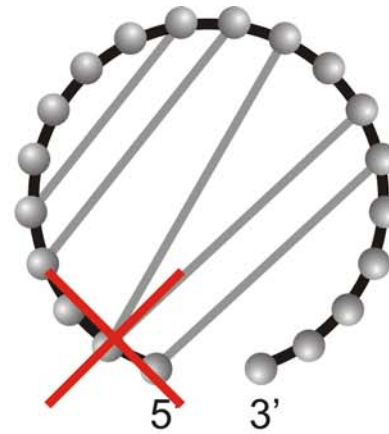
A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs



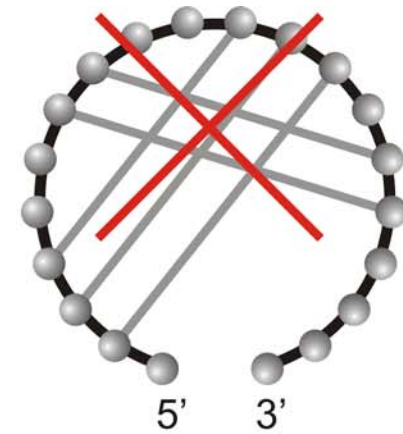
Base pairing



No nearest neighbor pair rule



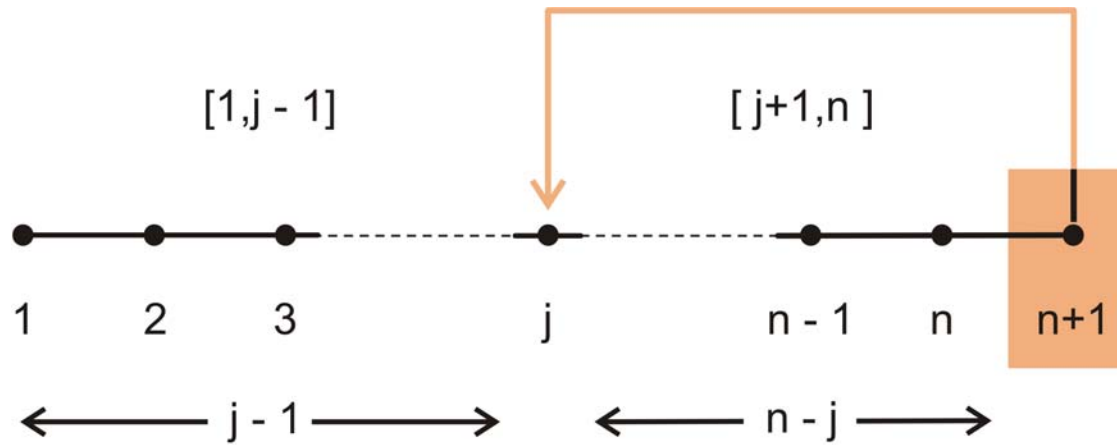
No base triplet rule



No pseudoknot rule

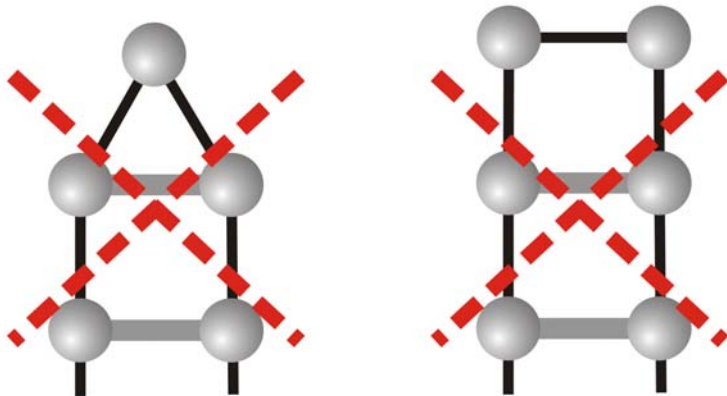
Base pairs $\in \{\mathbf{AU}, \mathbf{CG}, \mathbf{GC}, \mathbf{GU}, \mathbf{UA}, \mathbf{UG}\}$

Conventional definition of RNA secondary structures

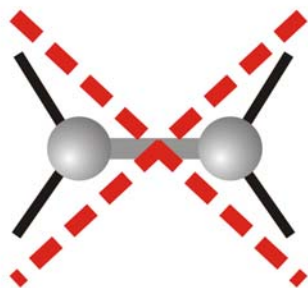
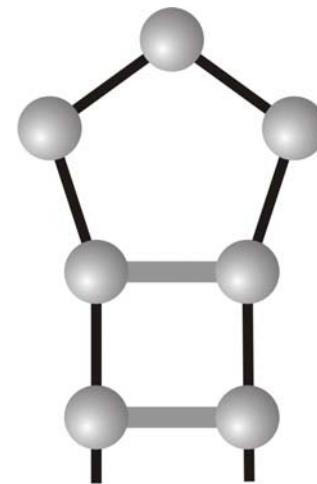


$$S_{n+1} = S_n + \sum_{j=1}^{n-1} S_{j-1} \cdot S_{n-j}$$

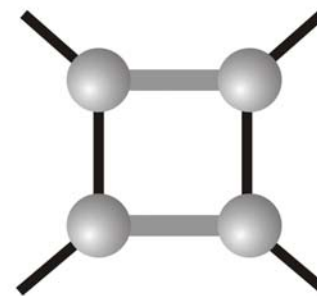
Counting the numbers of structures of chain length $n \Rightarrow n+1$



Impossible (extremely high free energies)
for steric reasons



High free energies because of lack of stacking and
very rare in minimum free energy structures



Restrictions on physically acceptable mfe-structures: $\lambda \geq 3$ and $\sigma \geq 2$

Size restriction of elements: (i) hairpin loop $n_{\text{loop}} \geq \lambda$
(ii) stack $n_{\text{stack}} \geq \sigma$

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

$$\Xi_{m+1} = S_m + \sum_{k=\lambda+2\sigma-2}^{m-2} \Phi_k \cdot S_{m-k+1}$$

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

$S_n \approx \#$ structures of a sequence with chain length n

Recursion formula for the number of physically acceptable stable structures

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

RNA sequence

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

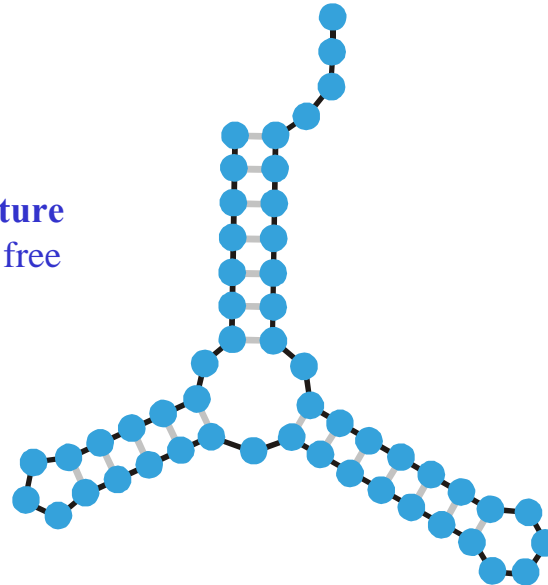
RNA folding:
Structural biology,
spectroscopy of
biomolecules,
understanding
molecular function

Biophysical chemistry:
thermodynamics and
kinetics



Empirical parameters

RNA structure
of minimal free
energy

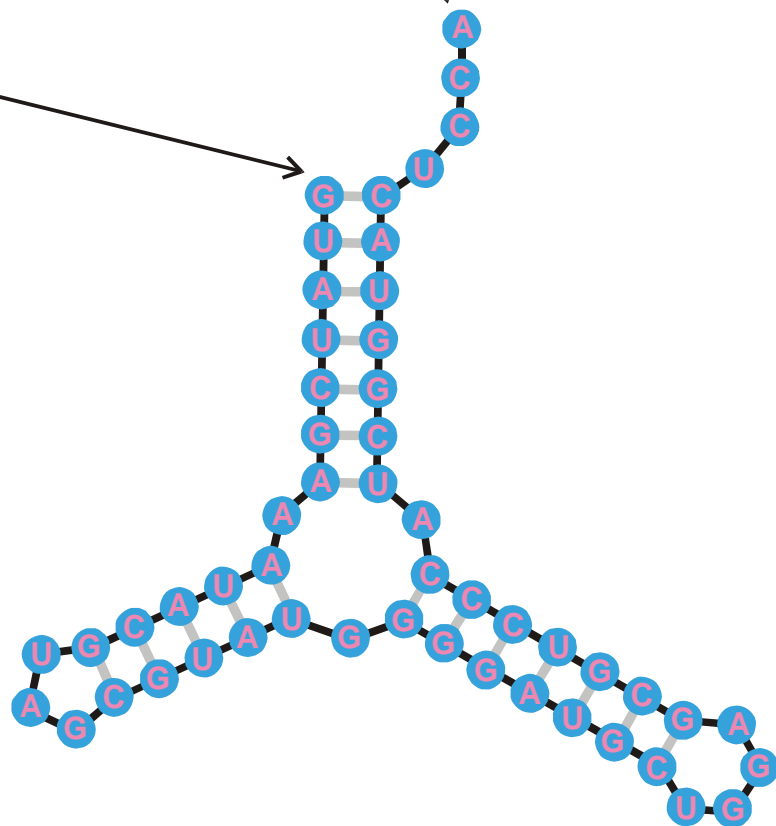
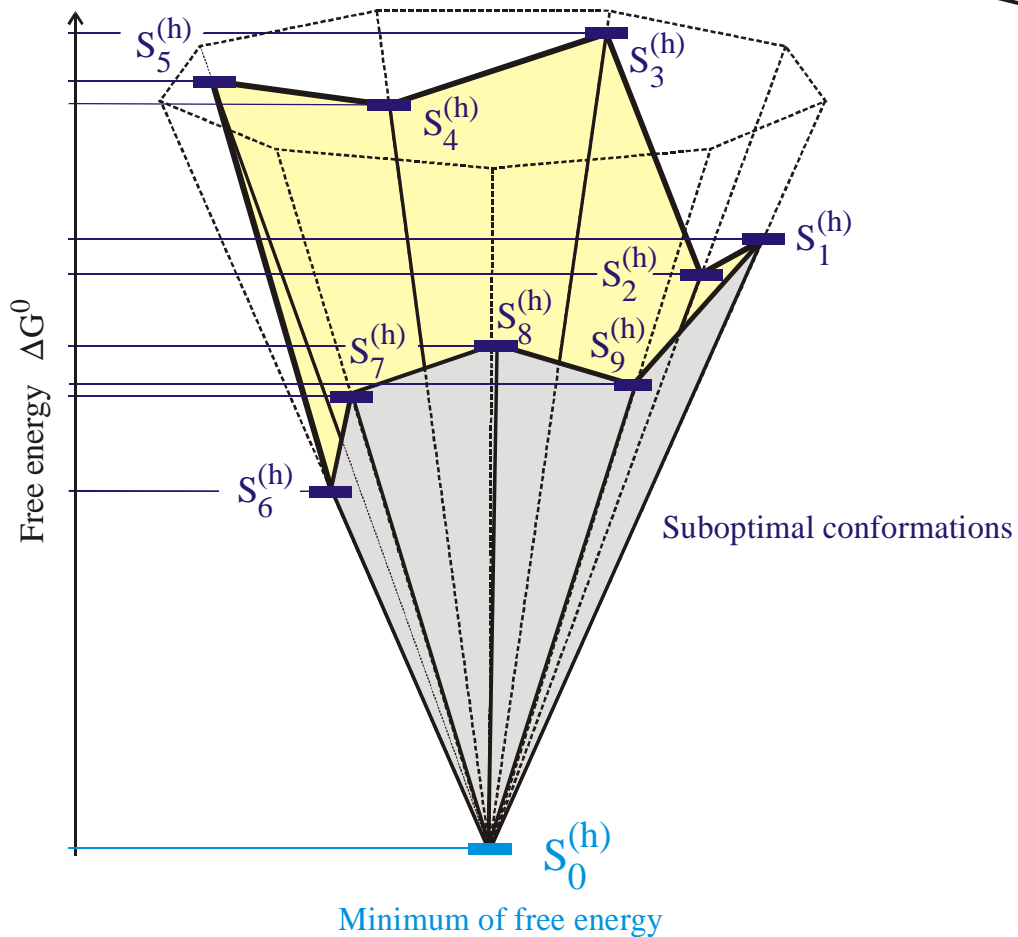


Sequence, structure, and design

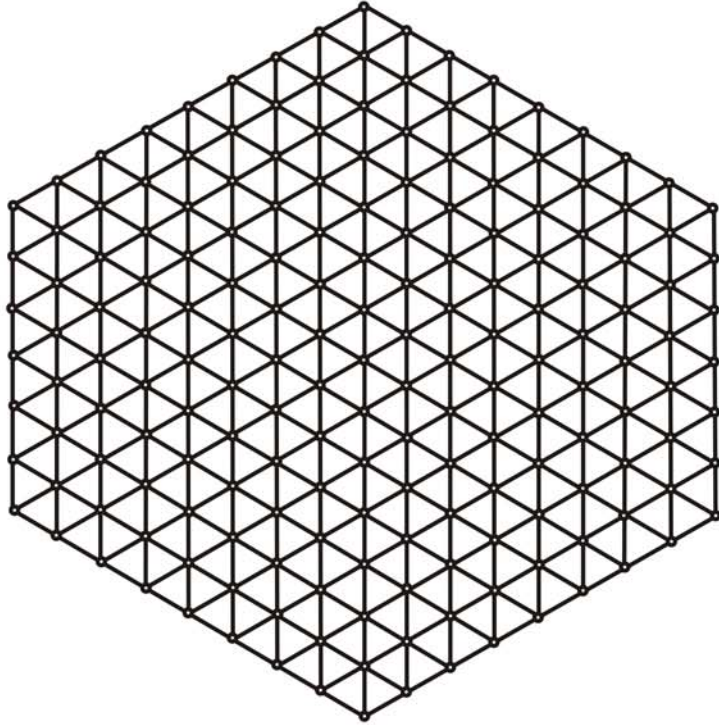
5'-end

3'-end

GUAUCGAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



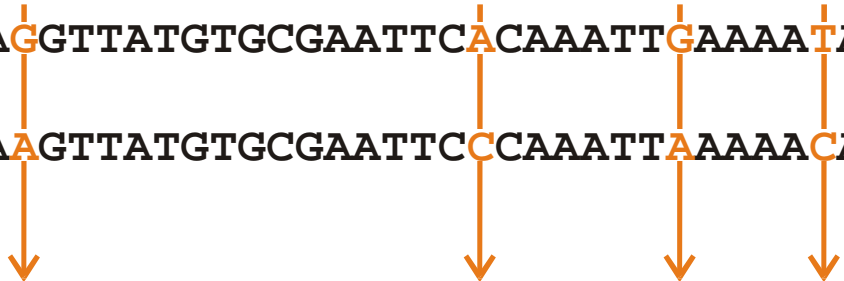
The minimum free energy structures on a discrete space of conformations



Sequence space

Sequence space

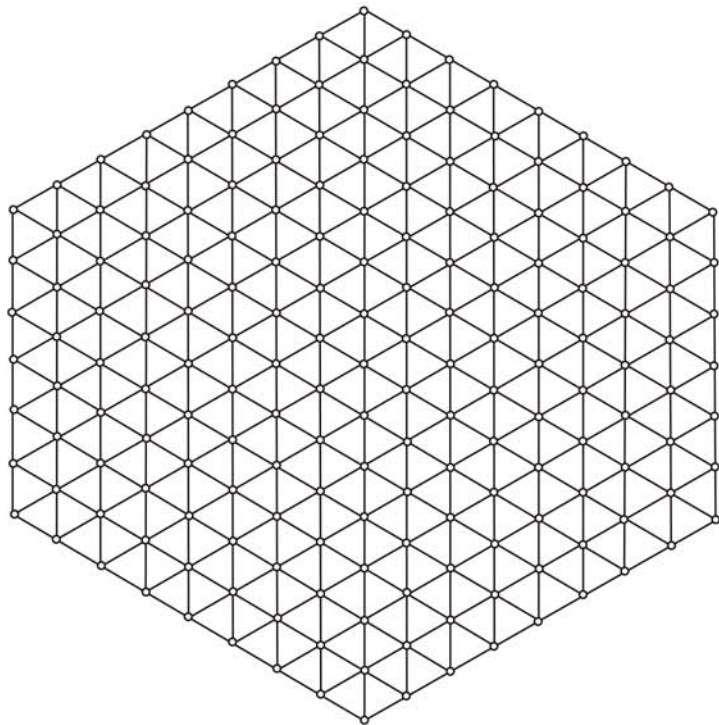
I_1 : CGTCGTTACAATTTA**G**GTTATGTGCGAATTC**A**CAAATT**G**AAAA**T**ACAAGAG.....
 I_2 : CGTCGTTACAATTTA**A**GTTATGTGCGAATTC**C**CAAATT**A**AAAA**C**ACAAGAG.....



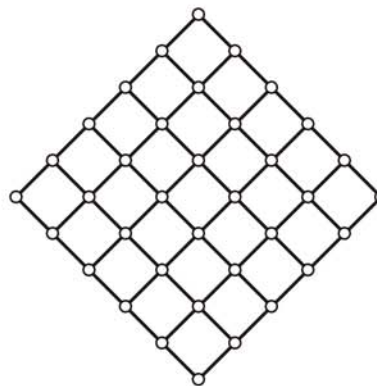
Hamming distance $d_H(I_1, I_2) = 4$

- (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

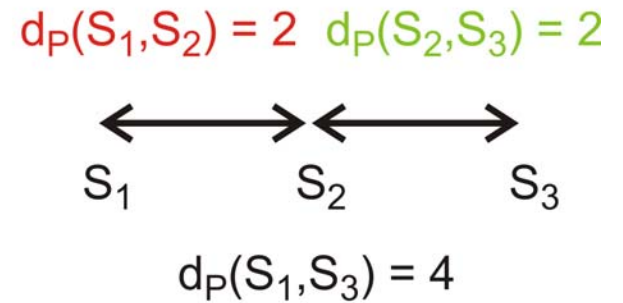
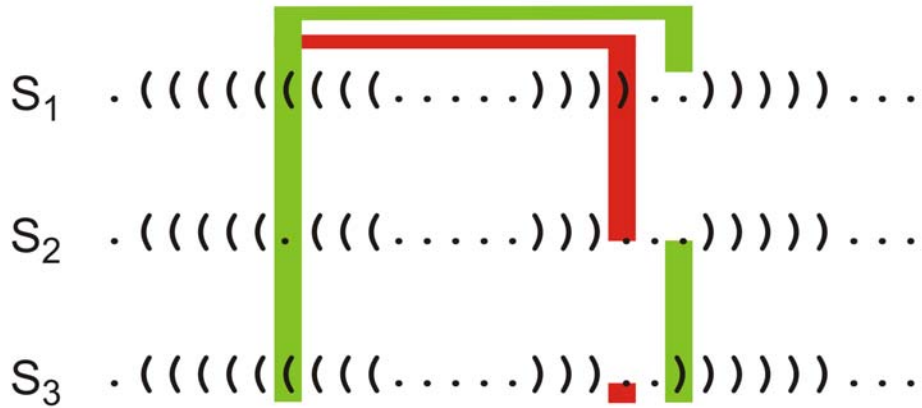
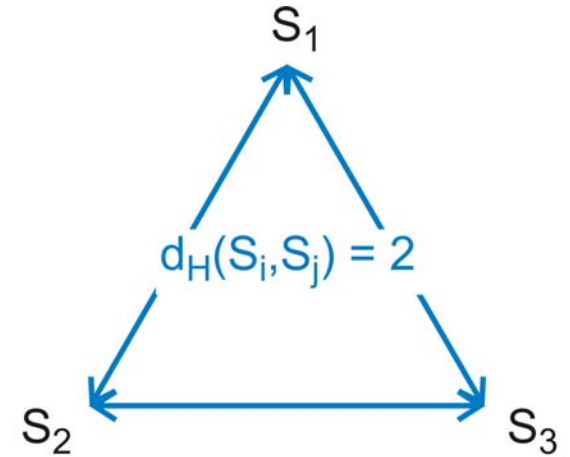
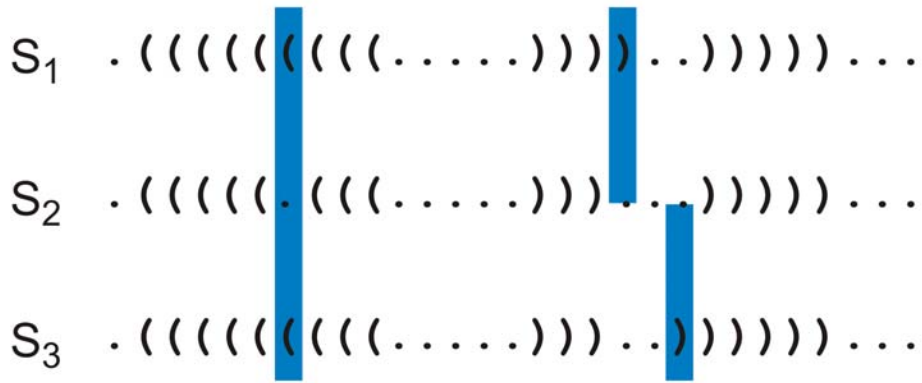


Sequence space



Structure space

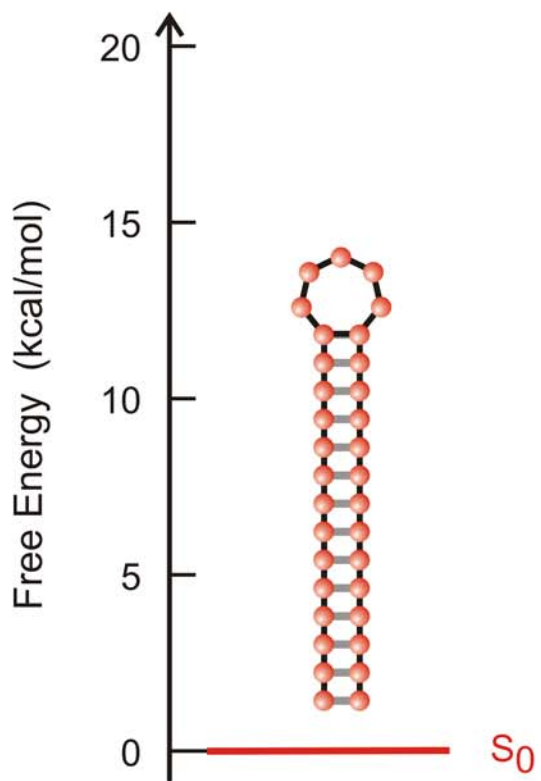
Sequence space and structure space



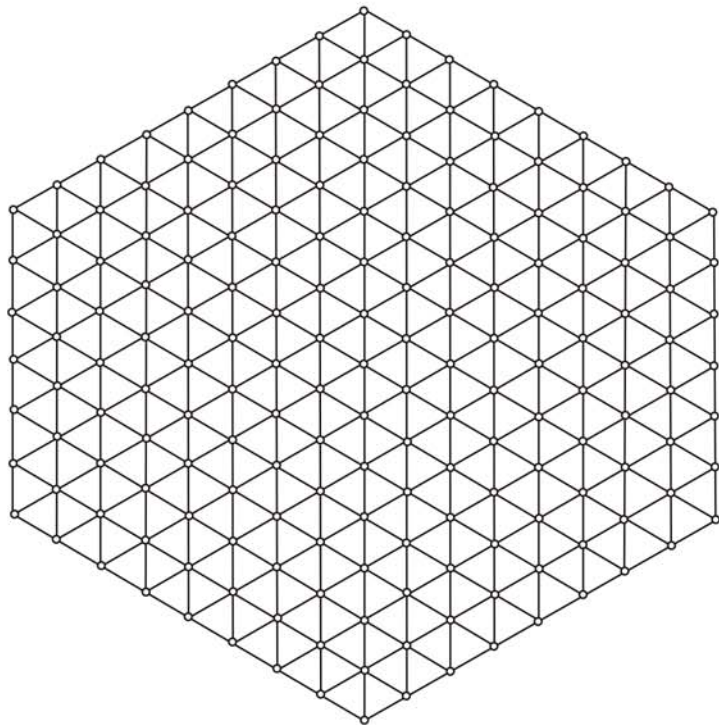
Two measures of distance in shape space:

Hamming distance between structures, $d_H(S_i, S_j)$ and base pair distance, $d_P(S_i, S_j)$

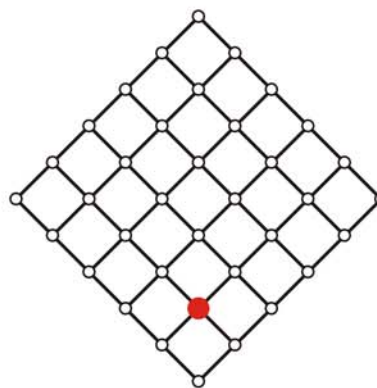
One sequence - one structure



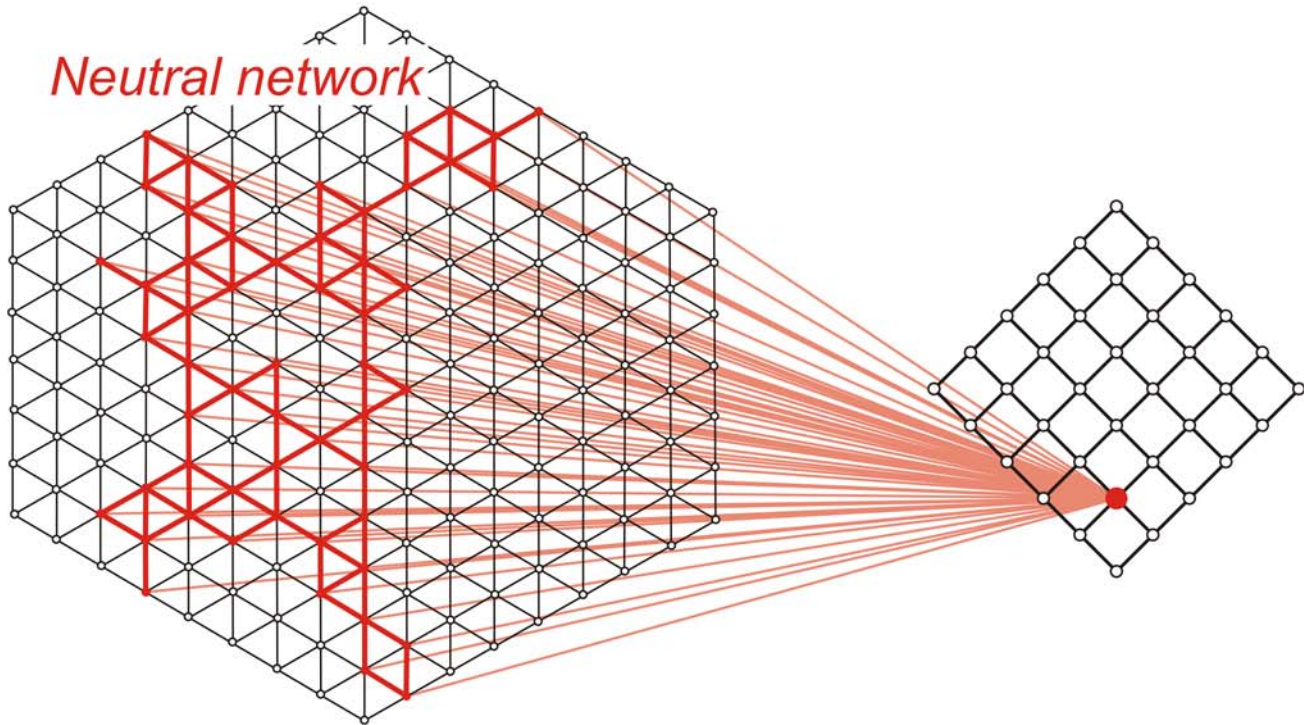
Minimum free energy structure



Sequence space



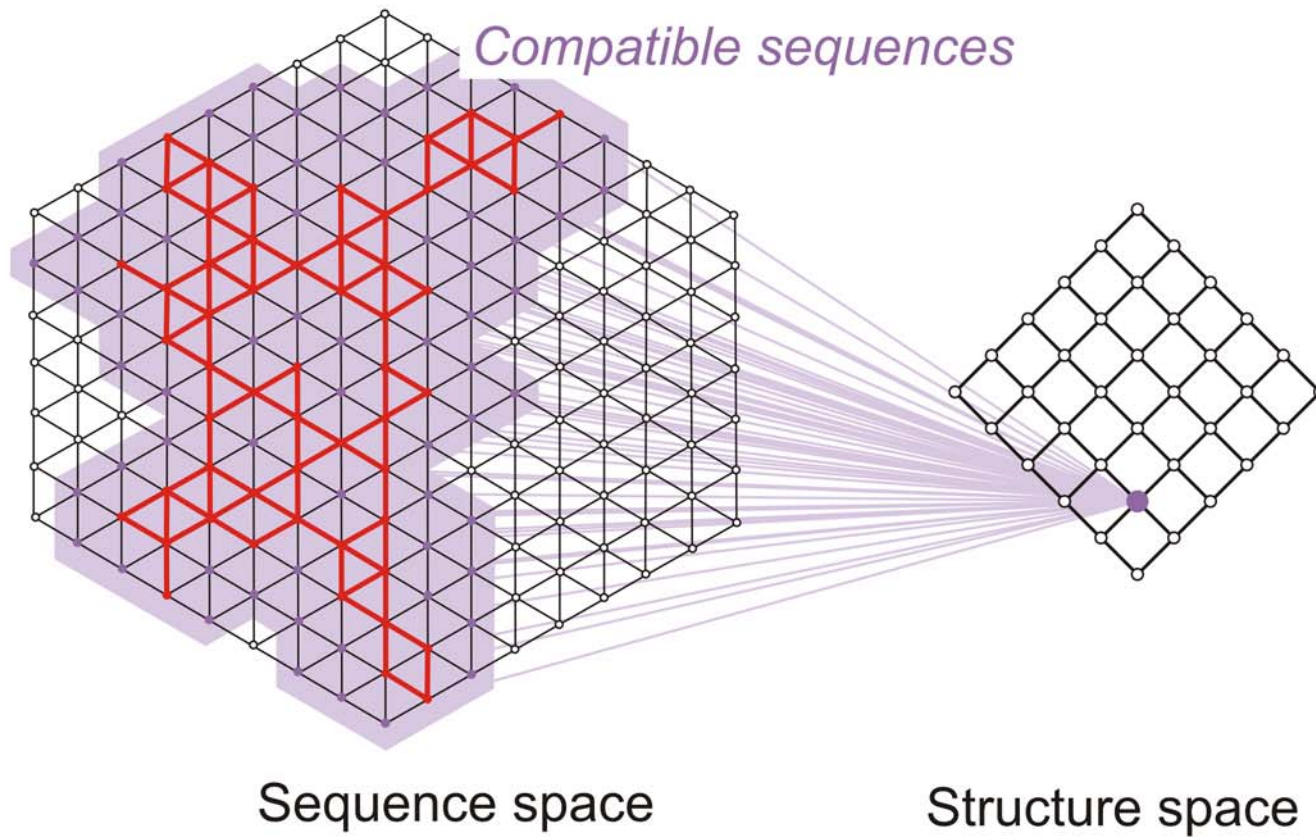
Structure space



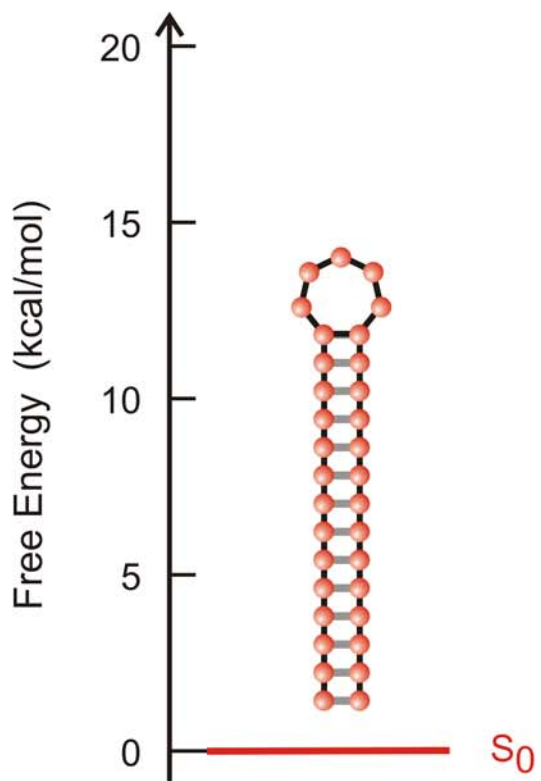
Neutral network

Sequence space

Structure space

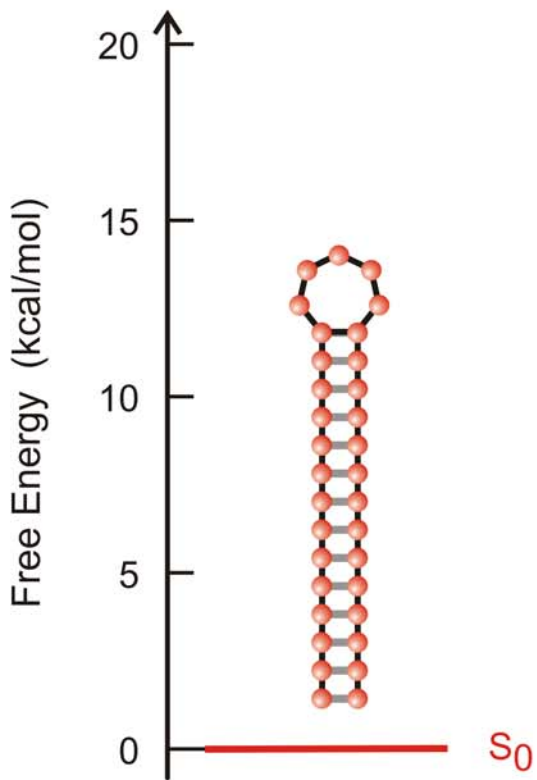


One sequence - one structure



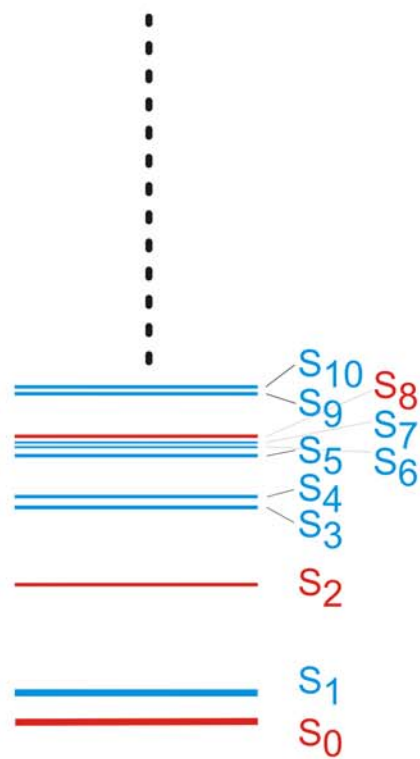
Minimum free energy structure

One sequence - one structure

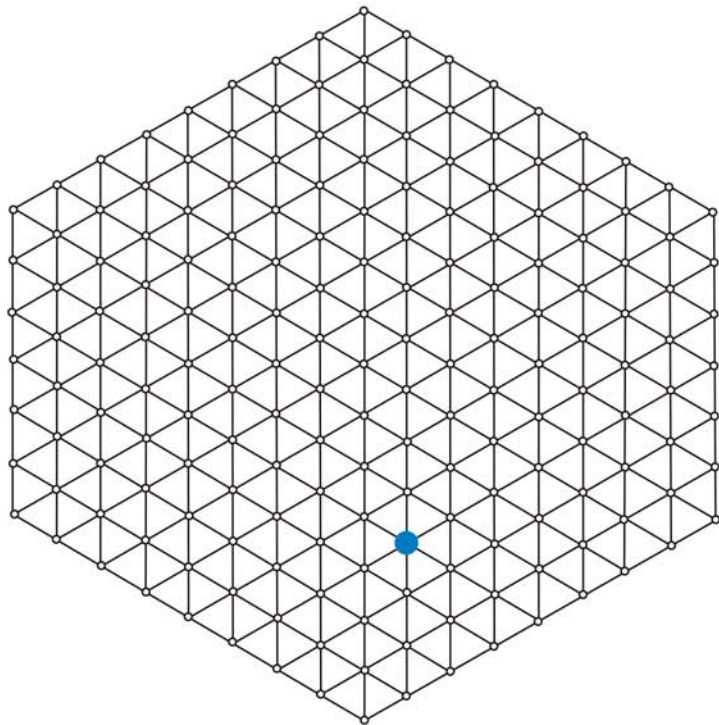


Minimum free energy structure

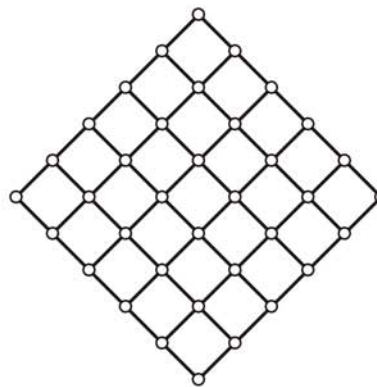
Many suboptimal structures
Partition function



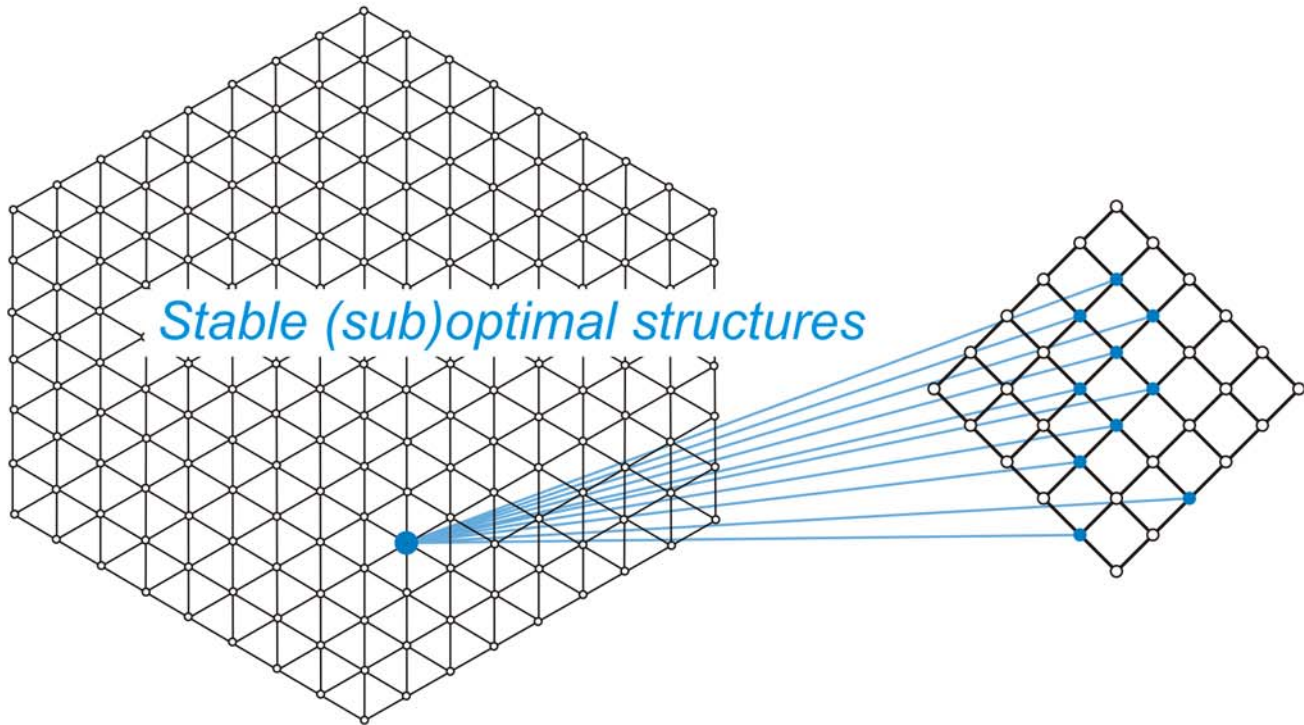
Suboptimal structures



Sequence space

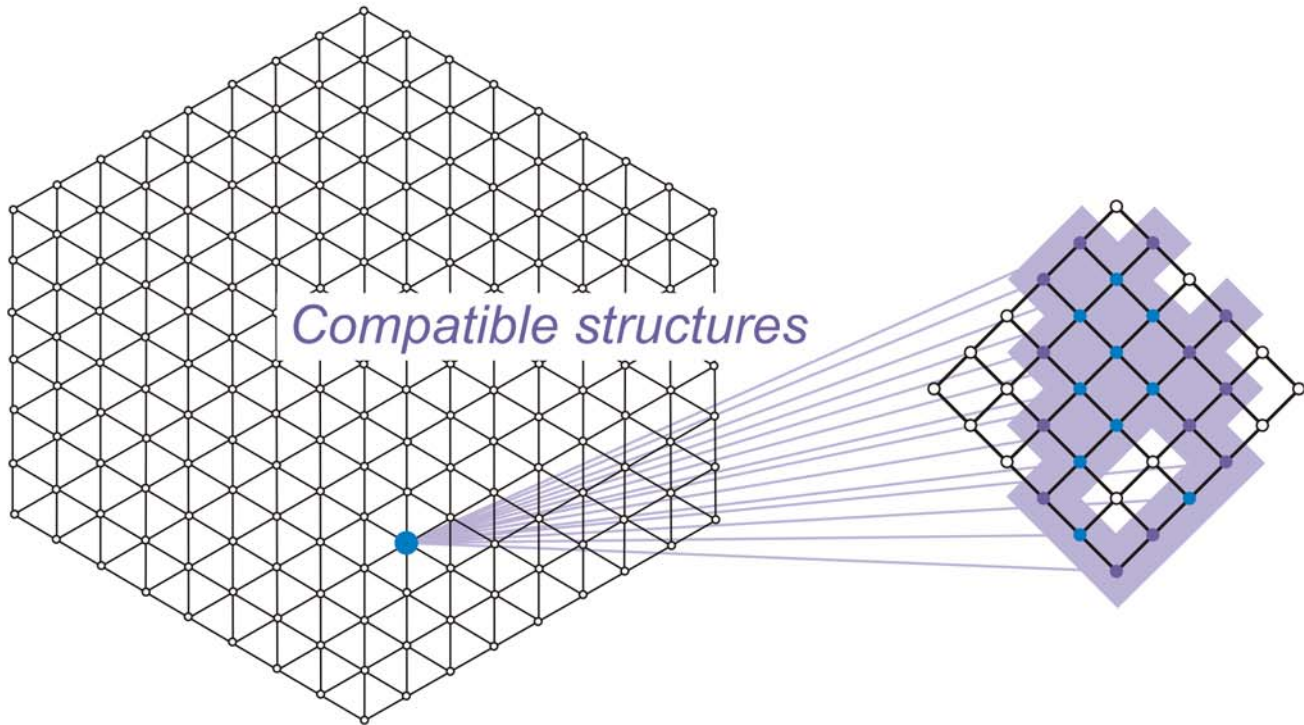


Structure space



Sequence space

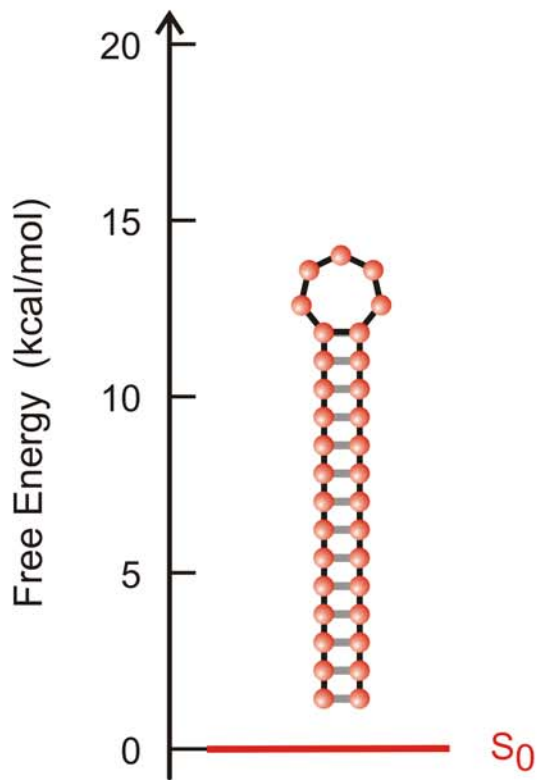
Structure space



Sequence space

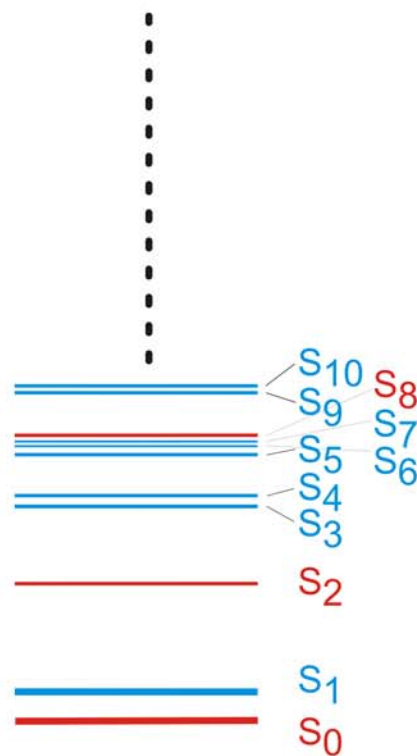
Structure space

One sequence - one structure



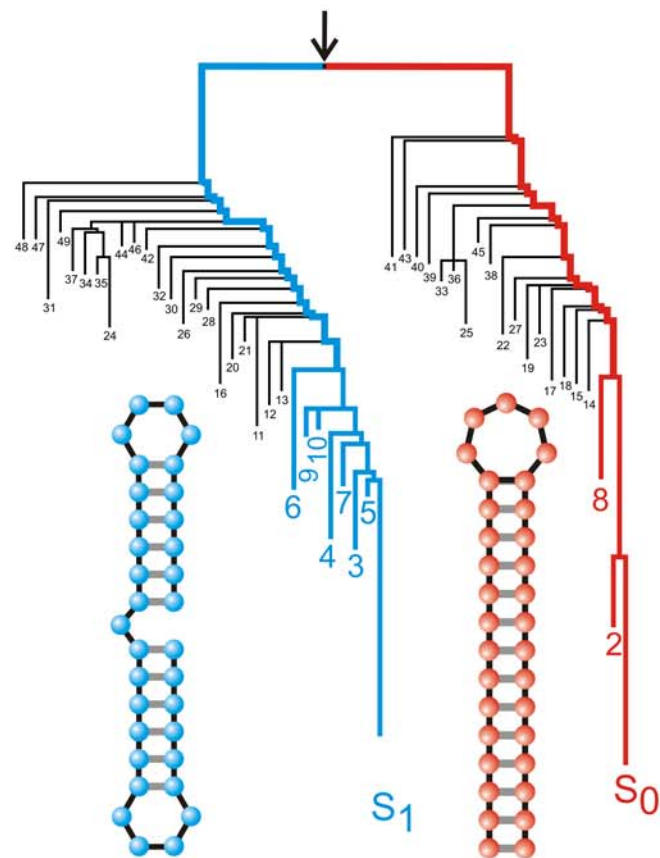
Minimum free energy structure

Many suboptimal structures
Partition function



Suboptimal structures

Metastable structures
Conformational switches



Kinetic structures

1. RNA sequences and structures
- 2. Neutral networks**
3. Evolutionary optimization of structure
4. Suboptimal structures and kinetic folding
5. Comparison of kinetic folding and evolution
6. How to model evolution of kinetic folding?

RNA sequence

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

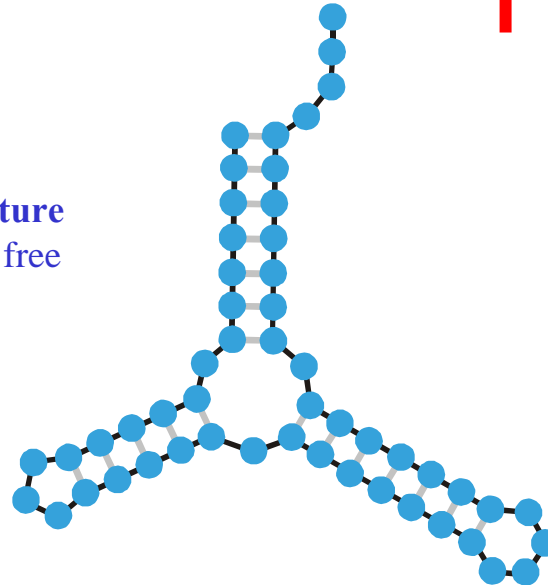
RNA folding:
Structural biology,
spectroscopy of
biomolecules,
understanding
molecular function

Iterative determination
of a sequence for the
given secondary
structure

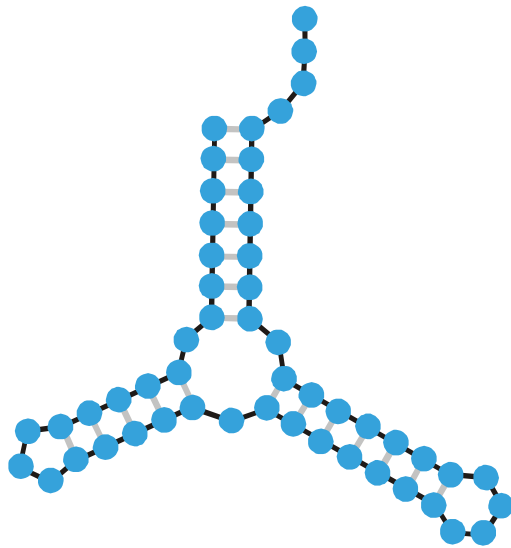
**Inverse Folding
Algorithm**

Inverse folding of RNA:
Biotechnology,
design of biomolecules
with predefined
structures and functions

RNA structure
of minimal free
energy

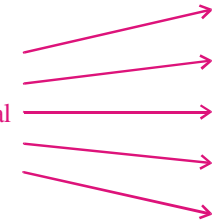


Sequence, structure, and design



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th



UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC
 GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUAUCUGG
 UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG
 CAUUGGUGCUAAUGAUUUAGGGCUGUAUUCUGUAUAGCGAUCAGUGUCCG
 GUAGGCCCUUGACAUAAGAUUUUCCAUGGUGGGAGAUGGCCAUUGCAG

Inverse folding

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

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

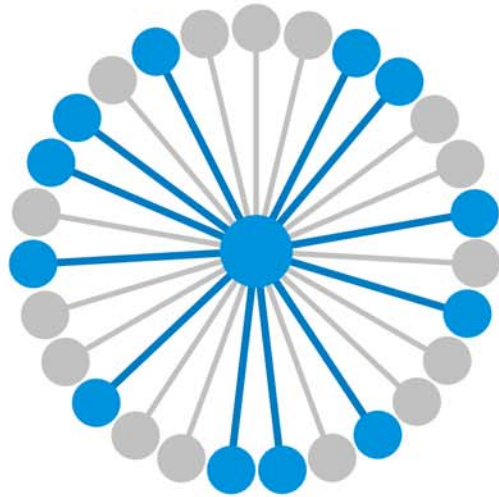
Space of phenotypes: $\mathcal{S} = \{S_1, S_2, S_3, S_4, \dots, S_M\}$; metric (not required)

$$N \gg M$$

$$\psi(I_j) = S_k$$

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

A mapping ψ and its inversion



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

$$\mathbf{G}_k = \psi^{-1}(\mathbf{S}_k) \doteq \{ I_j \mid \psi(I_j) = \mathbf{S}_k \}$$

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

Alphabet size κ :

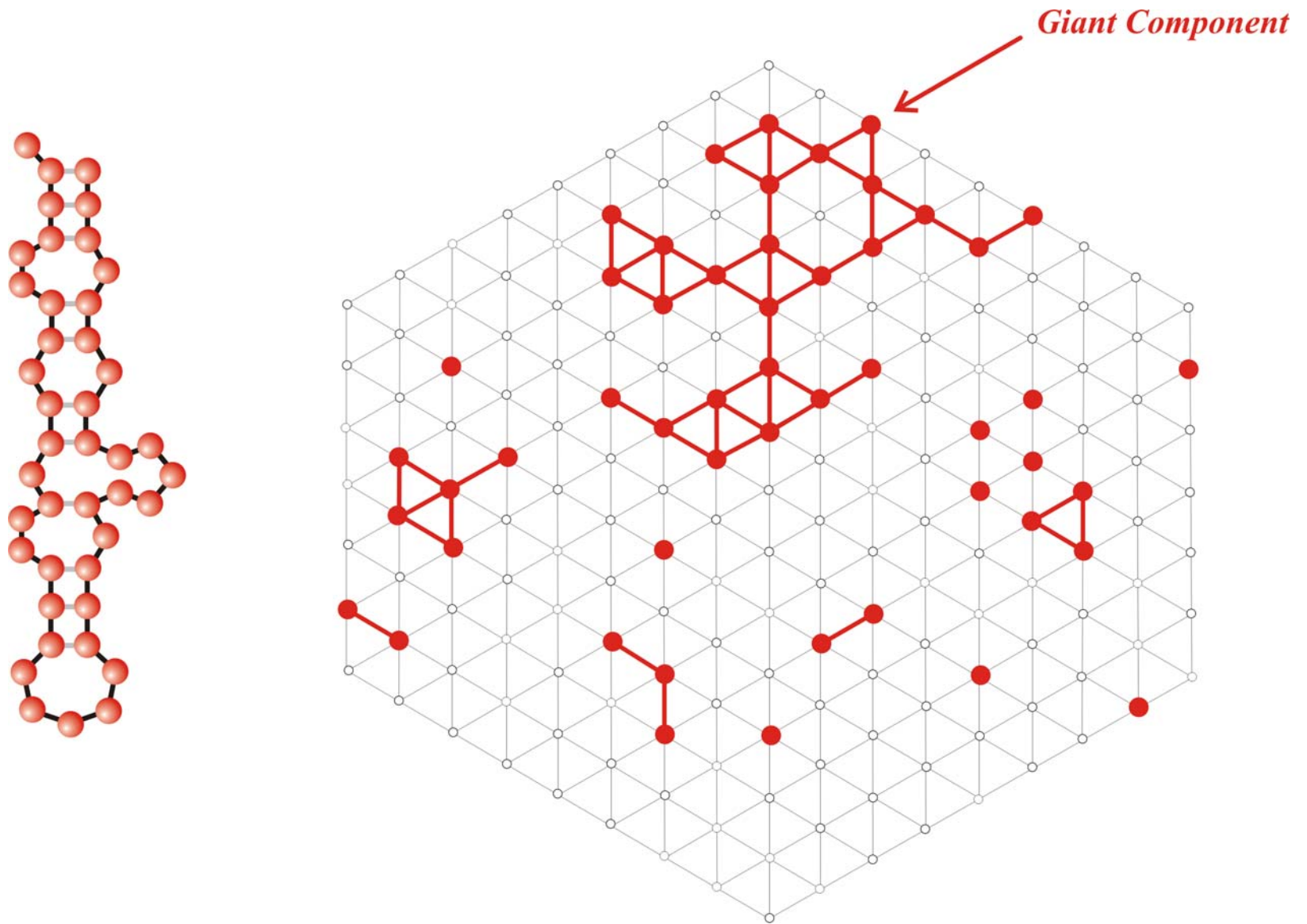
κ	λ_{cr}	
2	0.5	AU,GC,DU
3	0.423	AUG , UGC
4	0.370	AUGC

$\bar{\lambda}_k > \lambda_{cr}$ network \mathbf{G}_k is connected

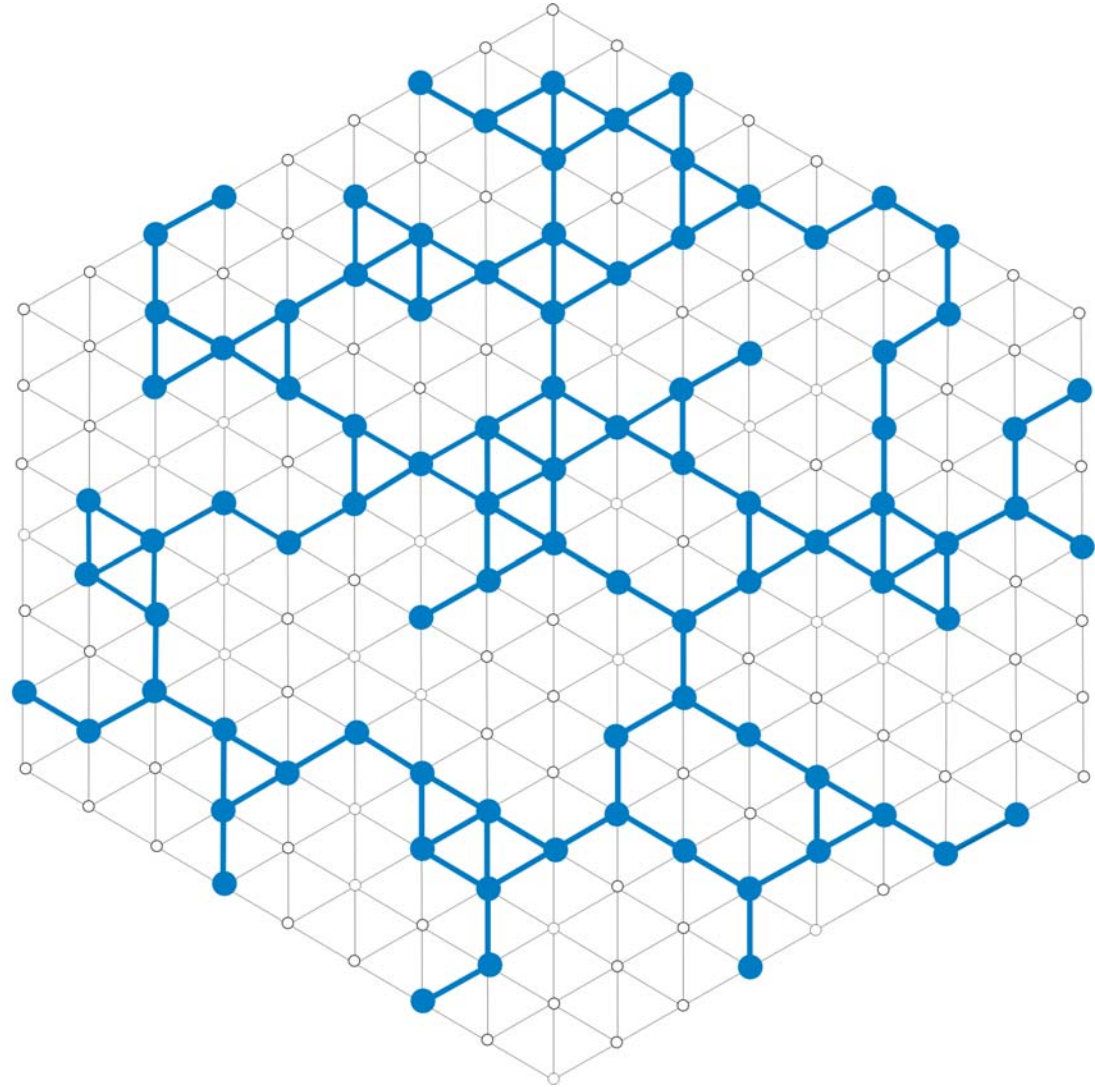
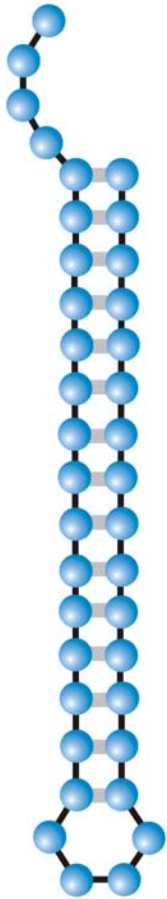
$\bar{\lambda}_k < \lambda_{cr}$ network \mathbf{G}_k is **not** connected

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

Degree of neutrality of neutral networks and the connectivity threshold



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



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

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

PETER SCHUSTER^{1,2,3}, WALTER FONTANA³, PETER F. STADLER^{2,3}
AND IVO L. HOFACKER²

¹ Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany

² Institut für Theoretische Chemie, Universität Wien, Austria

³ Santa Fe Institute, Santa Fe, U.S.A.

SUMMARY

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

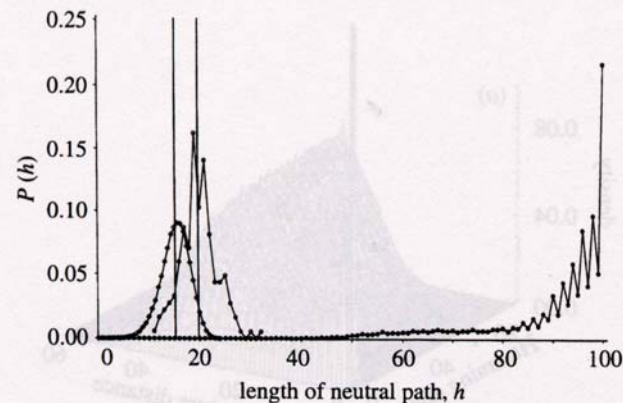


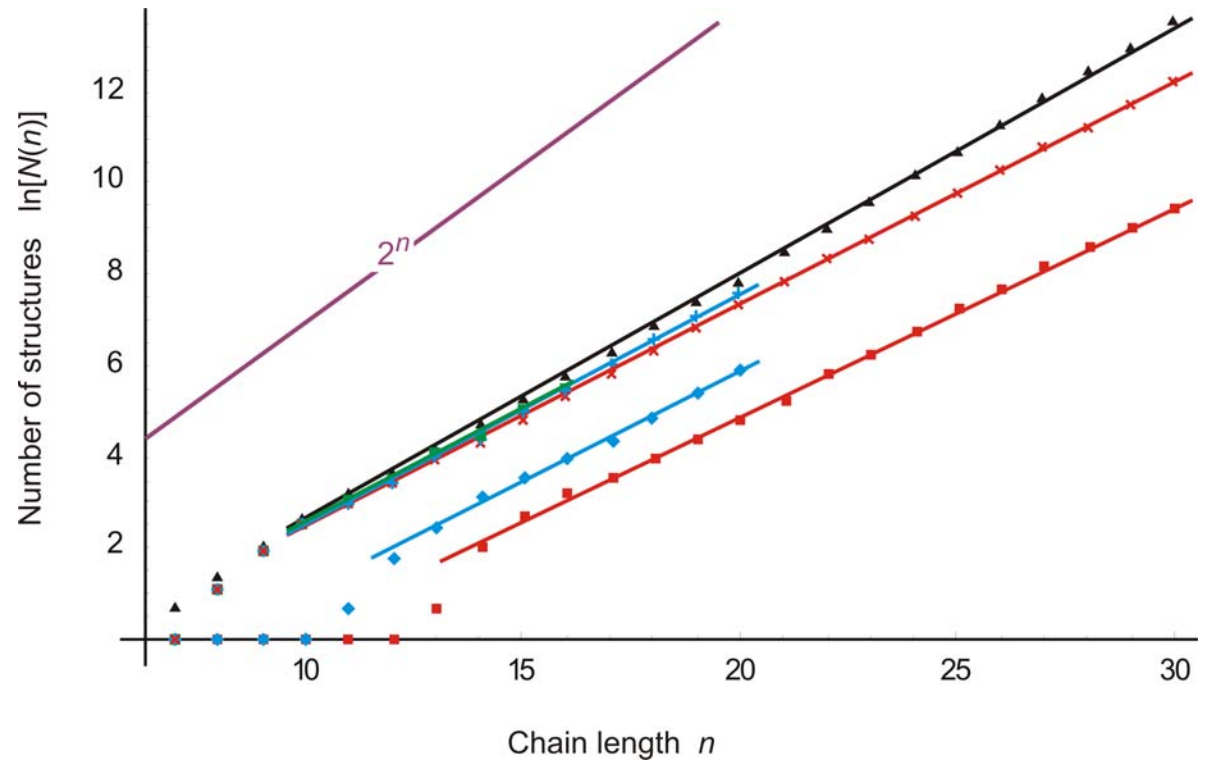
Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

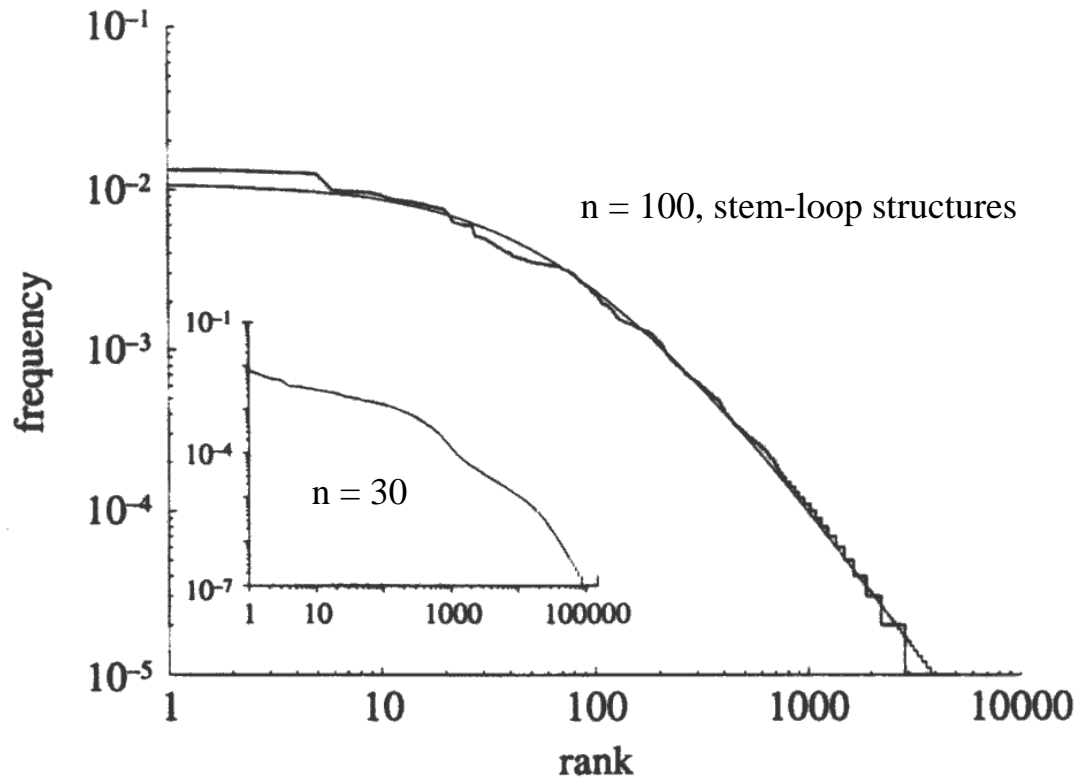


Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures



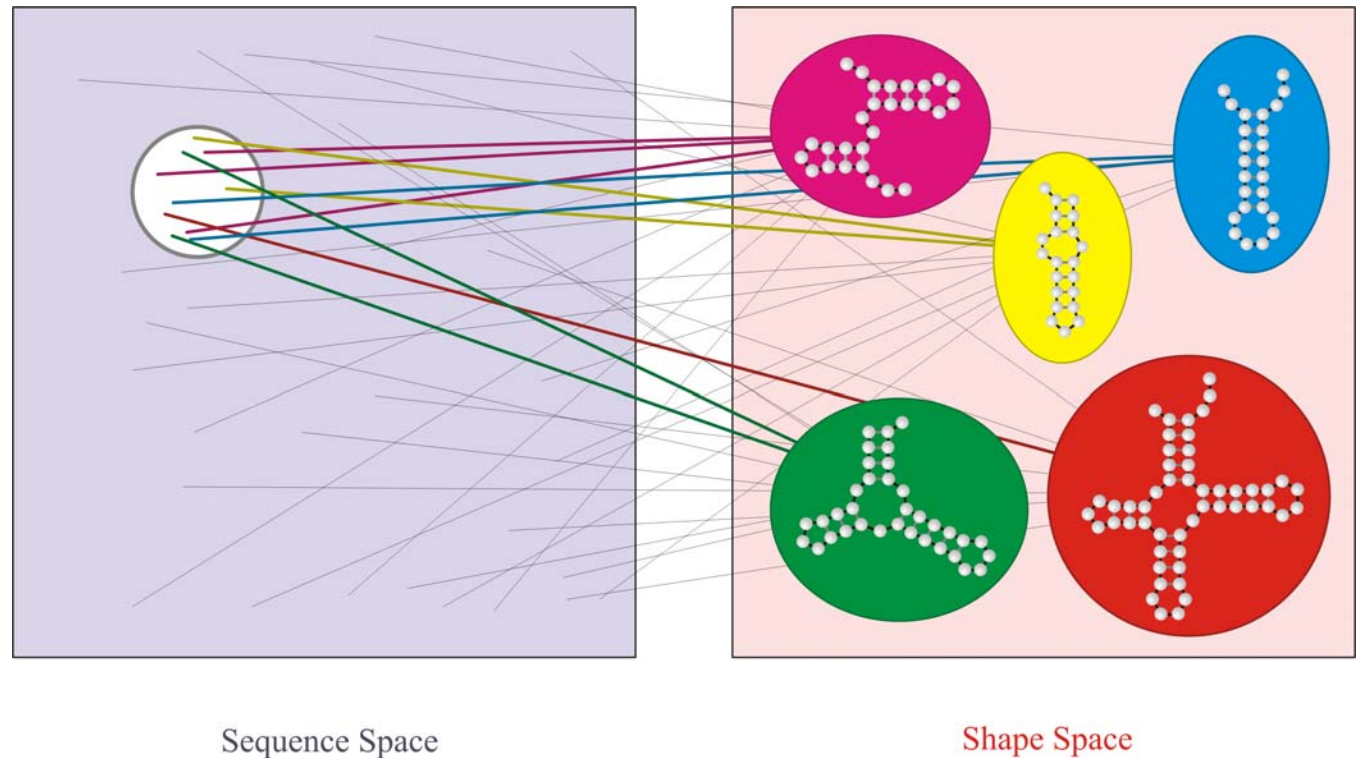
RNA secondary structures and Zipf's law

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures



Properties of RNA sequence to secondary structure mapping

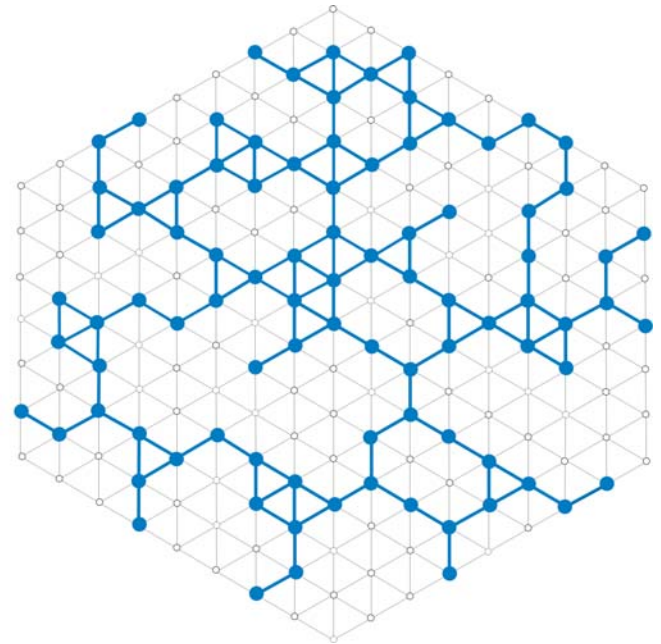
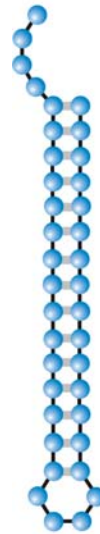
1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected

Alphabet size κ :

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



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

ZHEN HUANG¹ and JACK W. SZOSTAK²

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

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

ABSTRACT

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

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

RNA 9:1456-1463, 2003

Evidence for neutral networks and shape space covering

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

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

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

Received: 15 November 2002 / Accepted: 8 April 2003

Abstract. The evolution of ligand specificity underlies many important problems in biology, from the appearance of drug resistant pathogens to the re-engineering of substrate specificity in enzymes. In studying biomolecules, however, the contributions of macromolecular sequence to binding specificity can be obscured by other selection pressures critical to bioactivity. Evolution of ligand specificity *in vitro*—unconstrained by confounding biological factors—is addressed here using variants of three flavin-binding RNA aptamers. Mutagenized pools based on the three aptamers were combined and allowed to compete during *in vitro* selection for GMP-binding activity. The sequences of the resulting selection isolates were diverse, even though most were derived from the same flavin-binding parent. Individual GMP aptamers differed from the parental flavin aptamers by 7 to 26 mutations (20 to 57% overall change). Acquisition of GMP recognition coincided with the loss of FAD (flavin-adenine dinucleotide) recognition in all isolates, despite the absence of a counter-selection to remove FAD-binding RNAs. To examine more precisely the proximity of these two activities within a defined sequence space, the complete set of all intermediate sequences between an FAD-binding aptamer and a GMP-binding aptamer were synthesized and assayed for activity. For this set of sequences, we observe a portion of a neutral network for FAD-binding function separated from GMP-binding function by a distance of three muta-

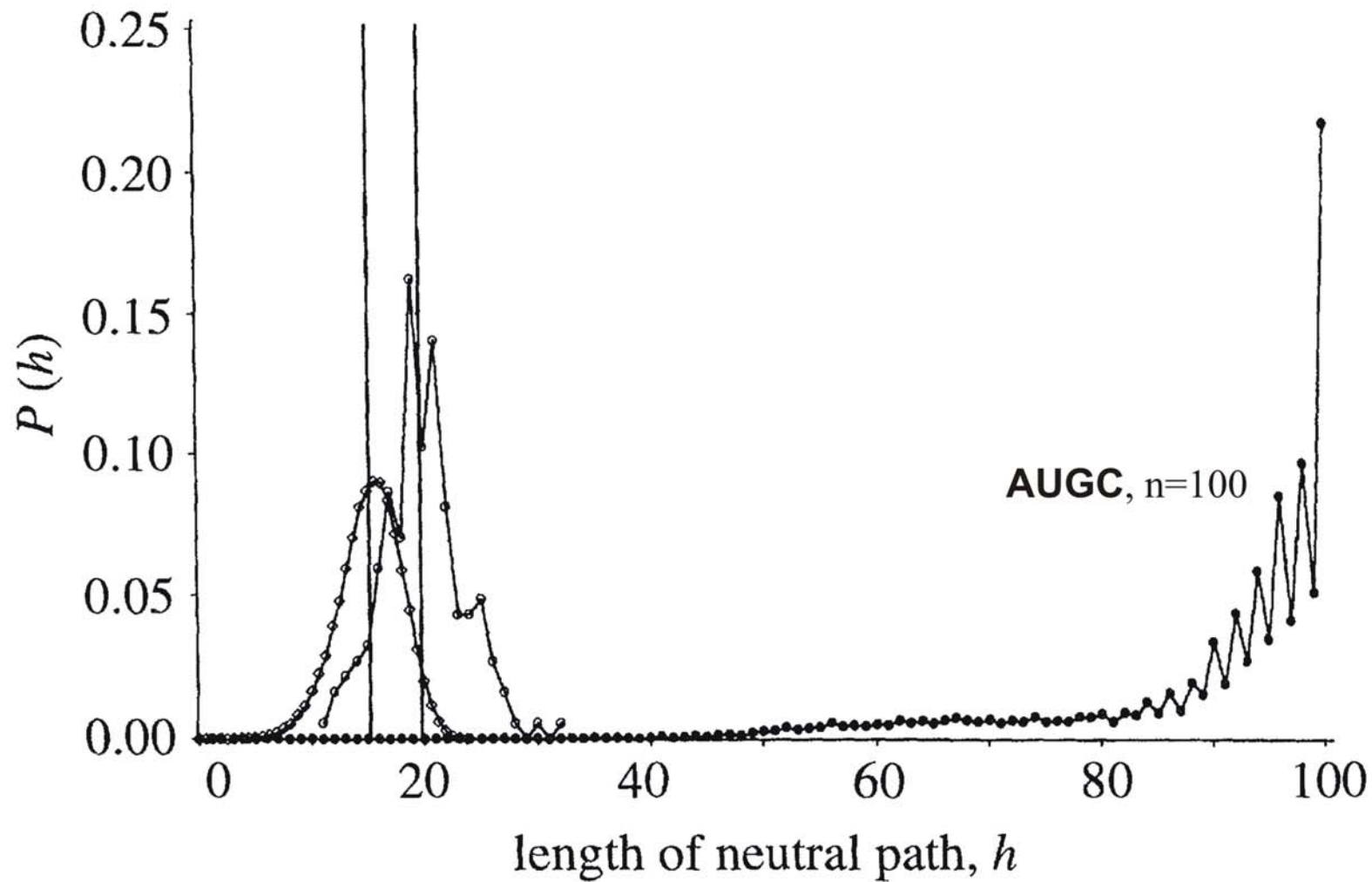
tions. Furthermore, enzymatic probing of these aptamers revealed gross structural remodeling of the RNA coincident with the switch in ligand recognition. The capacity for neutral drift along an FAD-binding network in such close approach to RNAs with GMP-binding activity illustrates the degree of phenotypic 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.

Evidence for neutral networks and intersection of aptamer functions



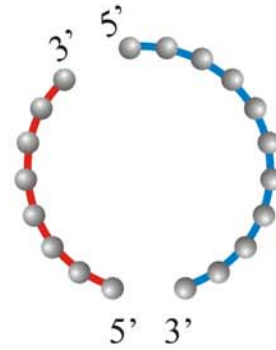
P. Schuster, W. Fontana, P.F. Stadler, I.L. Hofacker. 1994
Proc.Roy.Soc. London **B 255**:279-284.



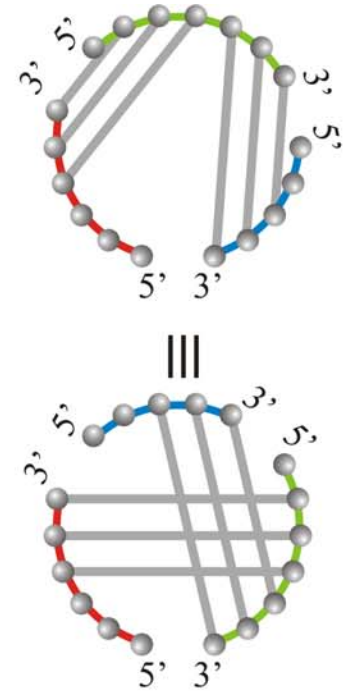
Circular RNA



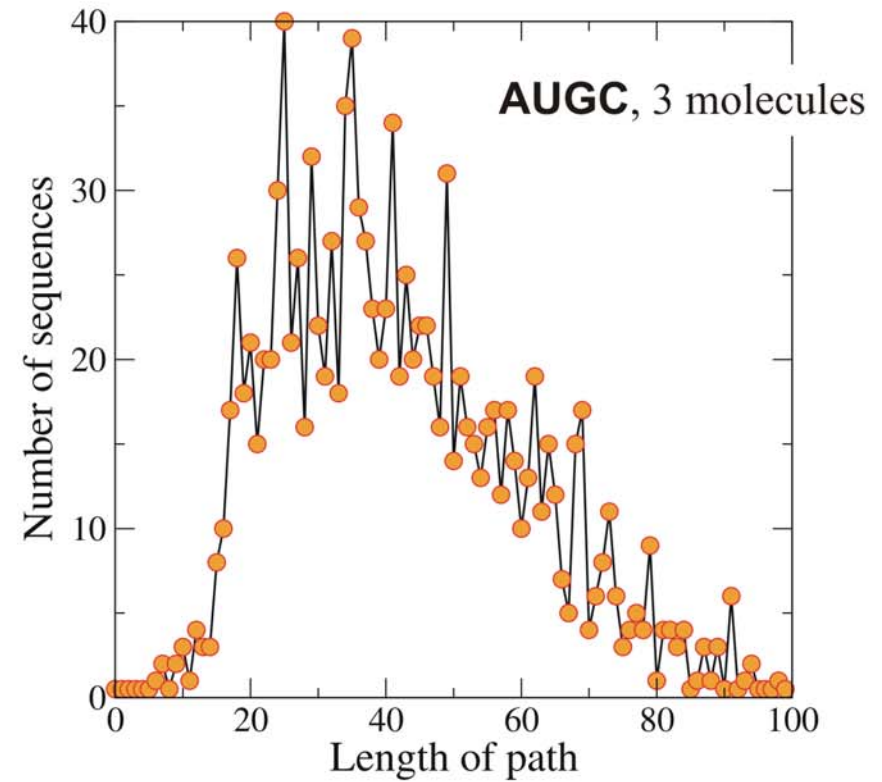
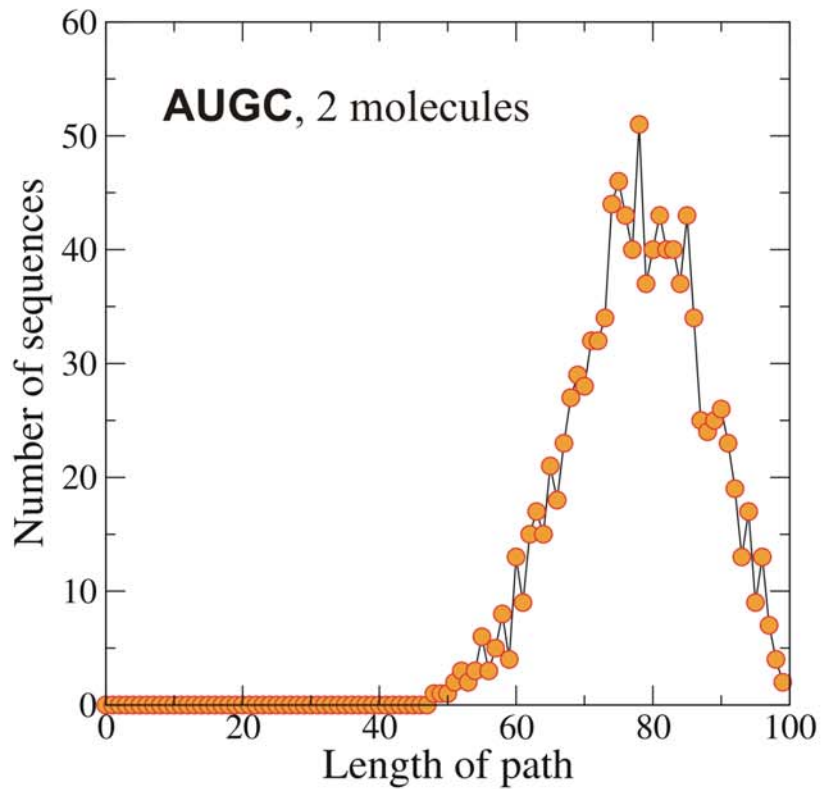
One RNA molecule



Two RNA molecules



Three RNA molecules



Total chain length: $n=100$; cofolding with one or two fixed sequences

C. Stephan-Otto Attolini, P.F. Stadler, 2005.
 Adv.Complex Syst., in press.

AUGC, $n = 100$

	Degree of neutrality λ	Mean length of path h
Unconstrained fold	0.33	> 95
Cofold with one sequence	0.32	75
Cofold with two sequences	0.18	40

Folding constraints, degree of neutrality and lengths of neutral path

1. RNA sequences and structures
2. Neutral networks
- 3. Evolutionary optimization of structure**
4. Suboptimal structures and kinetic folding
5. Comparison of kinetic folding and evolution

random individuals. The primer pair used for genomic DNA amplification is 5'-TCTCCCTGGATTCT-CATTTA-3' (forward) and 5'-TCTTTGTCTTCTGT-TGCACC-3' (reverse). Reactions were performed in 25 μ l using 1 unit of Taq DNA polymerase with each primer at 0.4 μ M, 200 μ M each dATP, dTTP, dCTP, and dGTP, and PCR buffer [10 mM Tris-HCl (pH 8.3), 50 mM KCl, 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 for 30 s 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, *MROD 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. Fiedel, 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 labyrinth) 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'-GCATGACCTGCGGGTAAT-GCG-3'; reverse, 5'-CTCAAGGCTTCTGGCATGGT-GCTCGCTGCG-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. Ferguson, 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 J. Barber, S. Sullivan, E. Green, D. Drayna, and T. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00335-01 and Z01 DC 00338-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.G.M.), the National Institute of Child Health and Human Development (R01 HD04028 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 empir-

ically 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 (replicable 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.

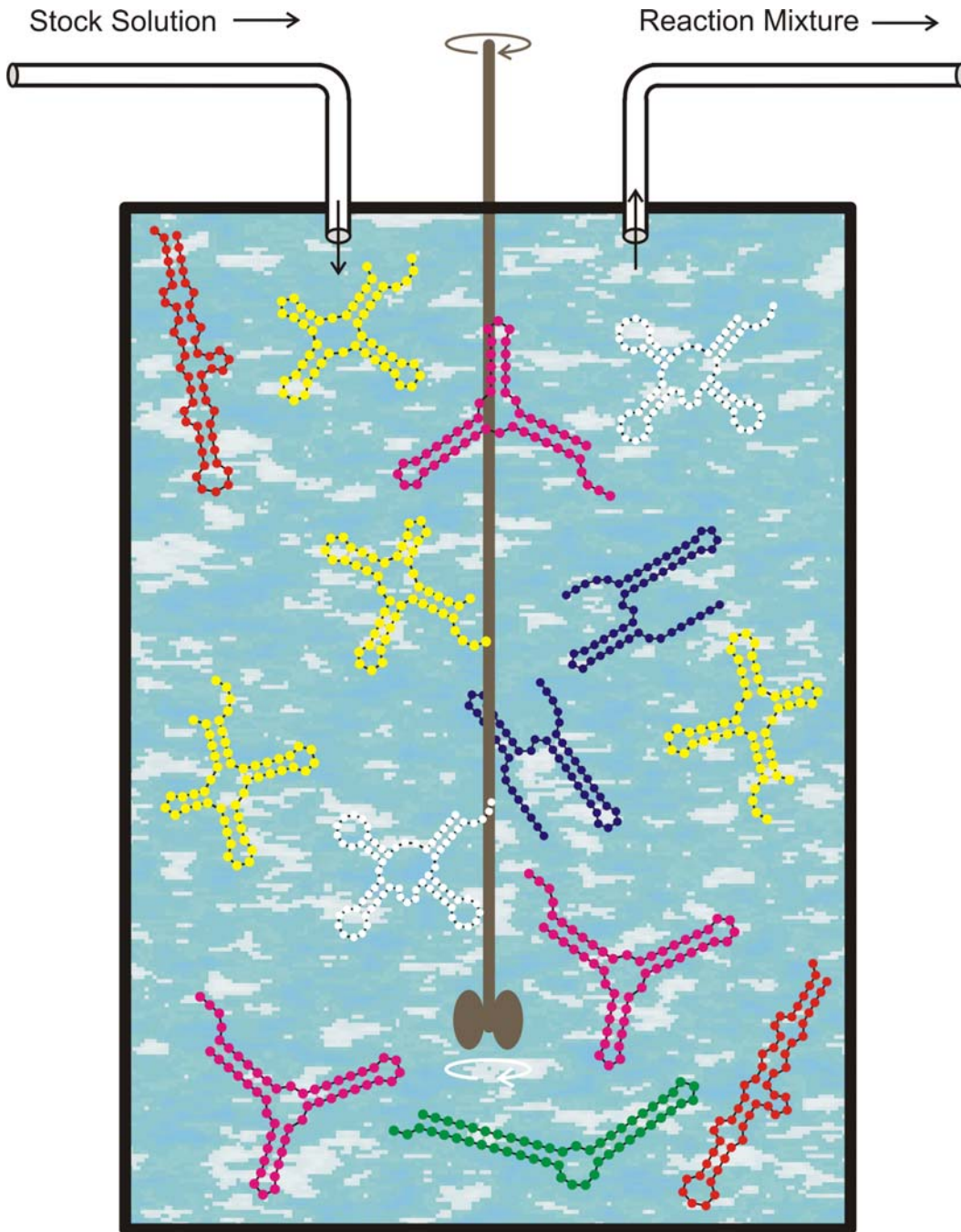
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces 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

Evolution *in silico*

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

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.



Replication rate constant:

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

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

Selection constraint:

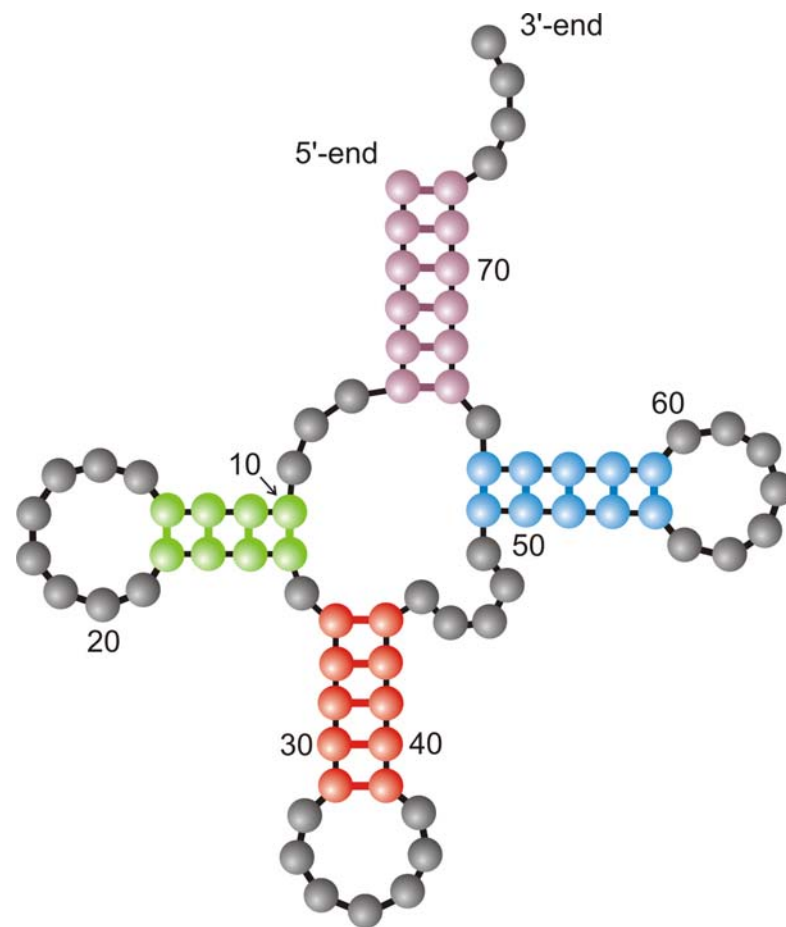
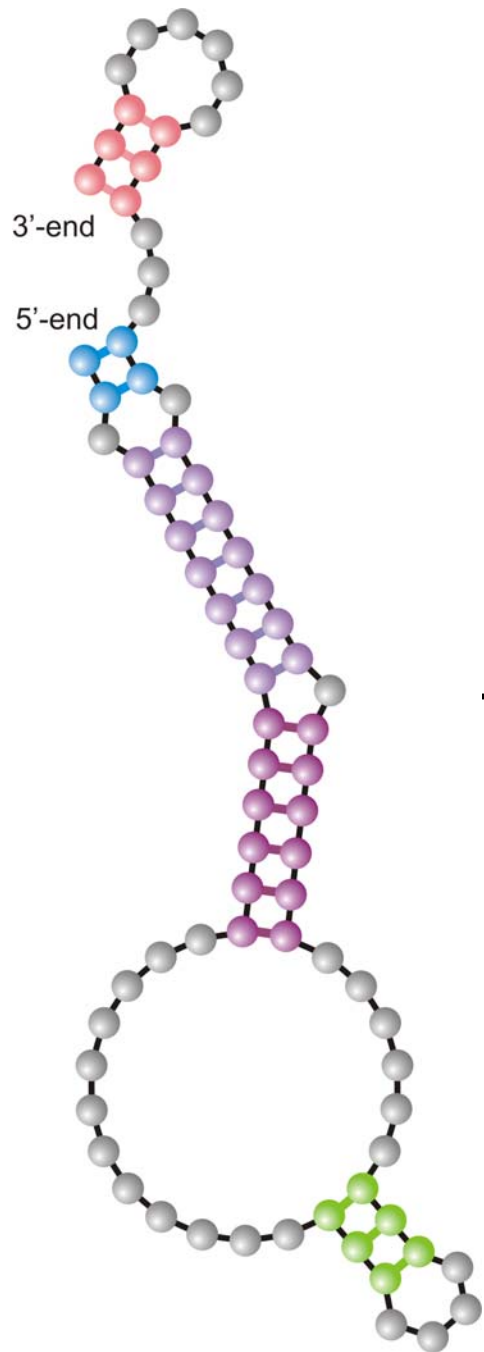
Population size, $N = \#$ RNA molecules, is controlled by the flow

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

Mutation rate:

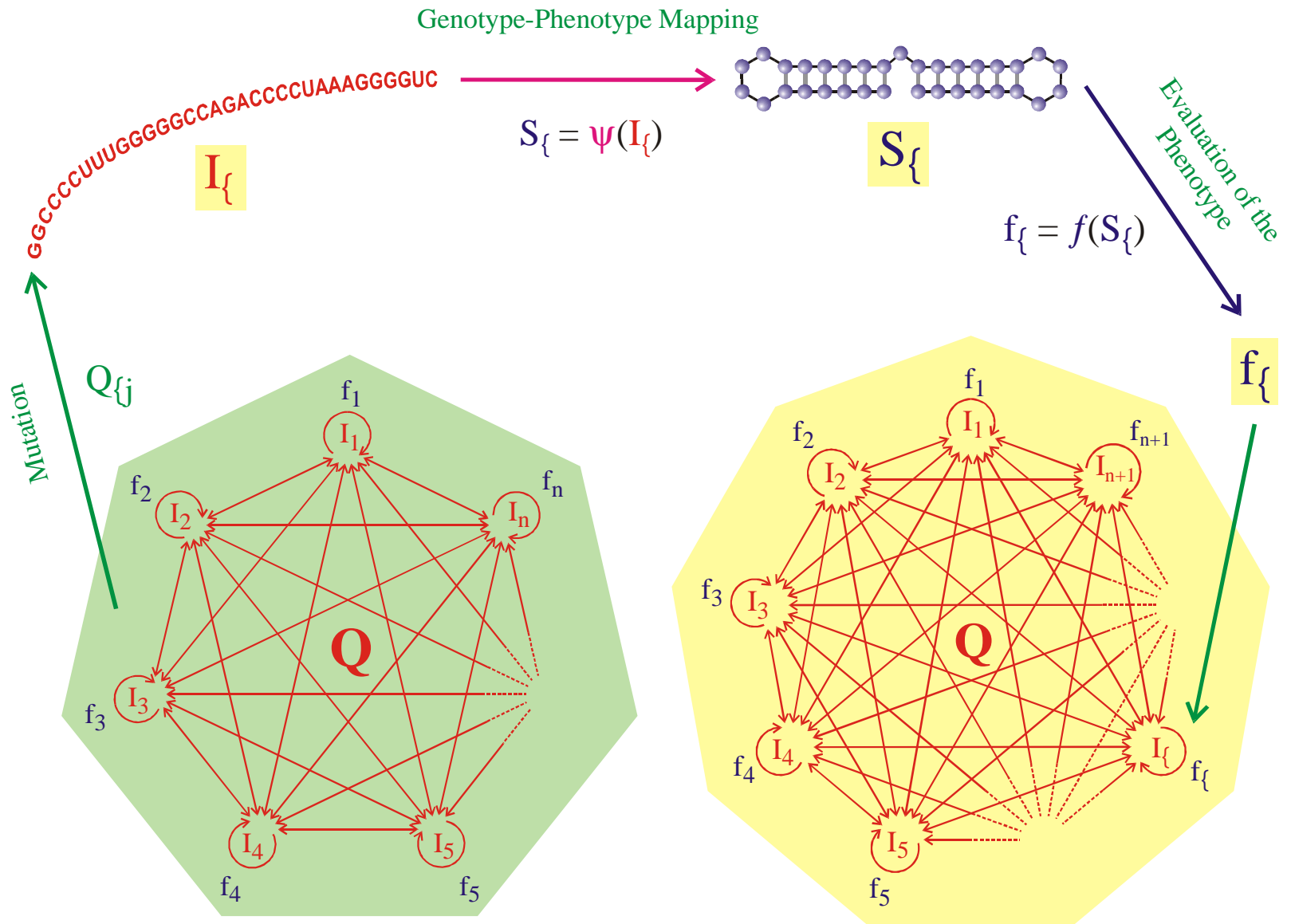
$p = 0.001 / \text{site} \times \text{replication}$

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

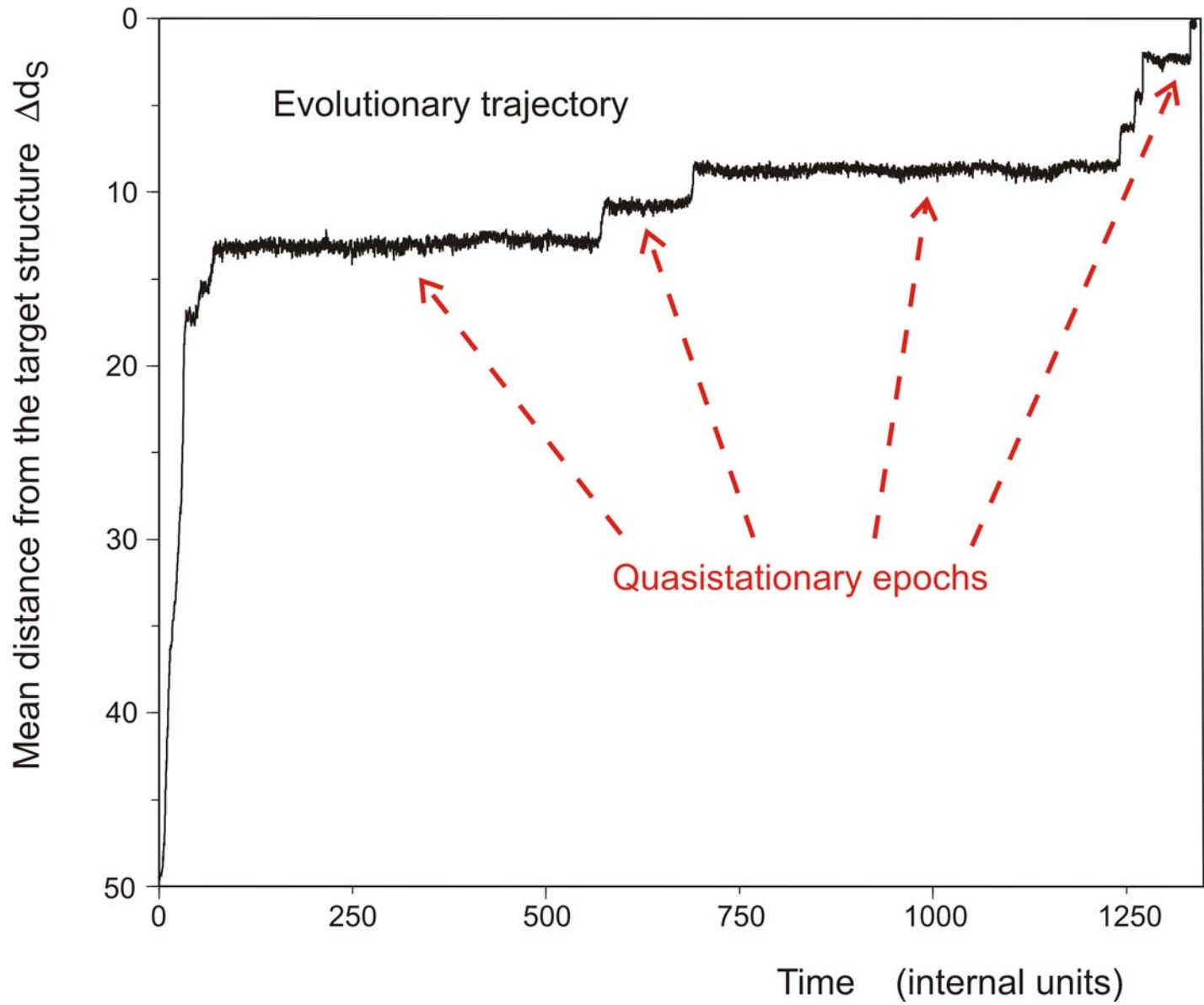


Randomly chosen
initial structure

Phenylalanyl-tRNA as
target structure

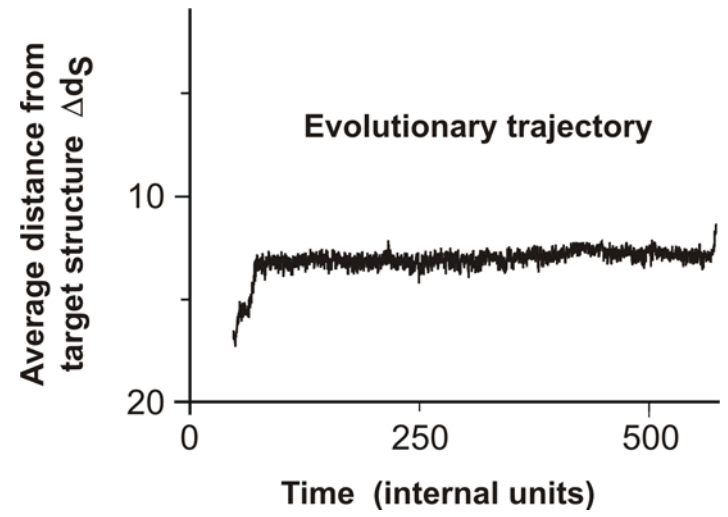


Evolutionary dynamics
including molecular phenotypes



In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch



entry	GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGG	CAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8	.(((((((((((((. (((.))))))))(((((.))))))))	
exit	GGUAUGGGCGUUGAAUA	UAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCAUAACAGAA
entry	GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCAUAACAGAA	
9	.((((((.(. (((.))))))))(((((.))))))))	
exit	UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG	
entry	UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG	
10	.(((((. (((.))))))))(((((.))))))))	
exit	UGGAUGGACGUUGAAUAACAAGGUAUCG	ACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG

Transition inducing point mutations change the molecular structure

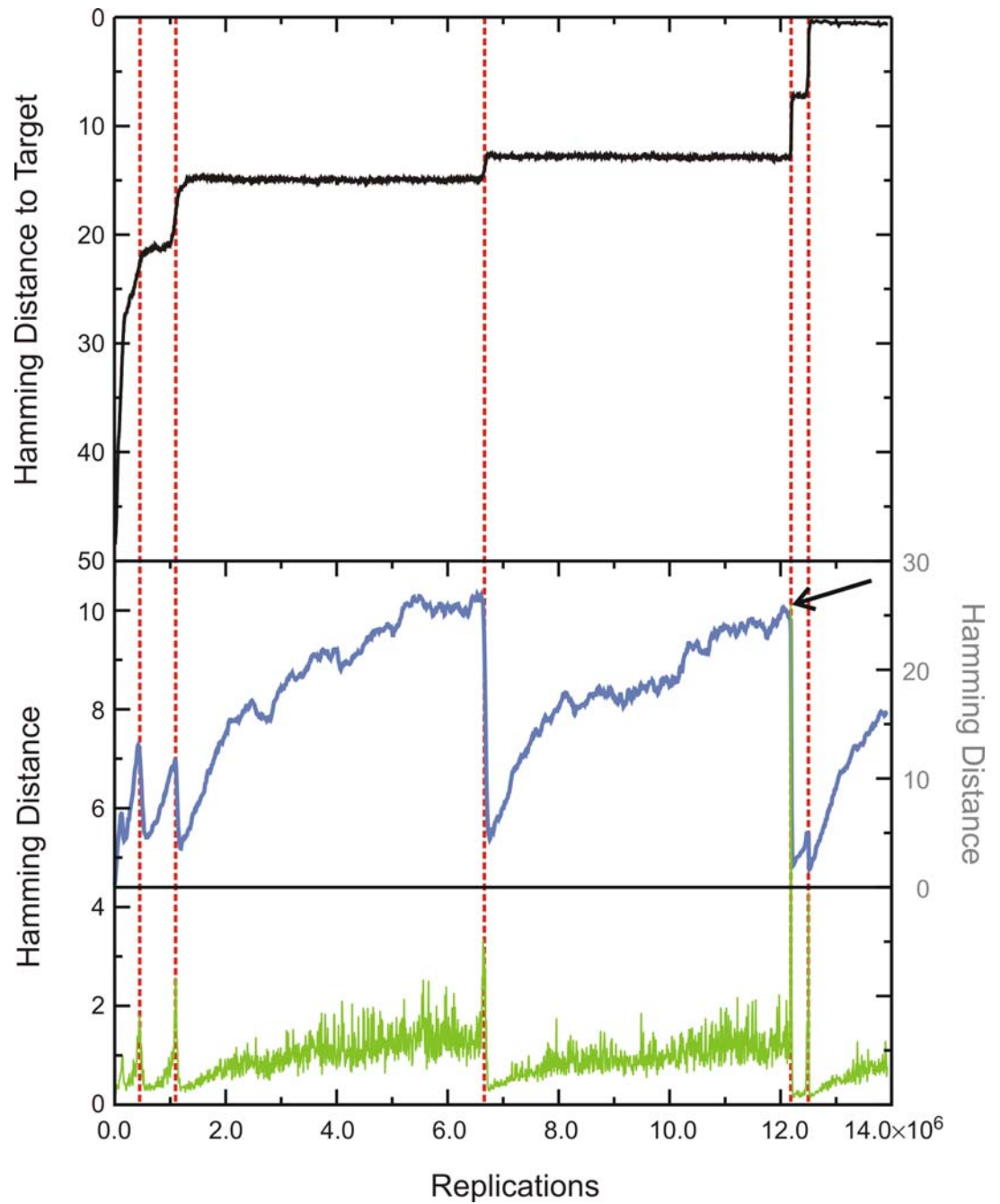
Neutral point mutations leave the molecular structure unchanged

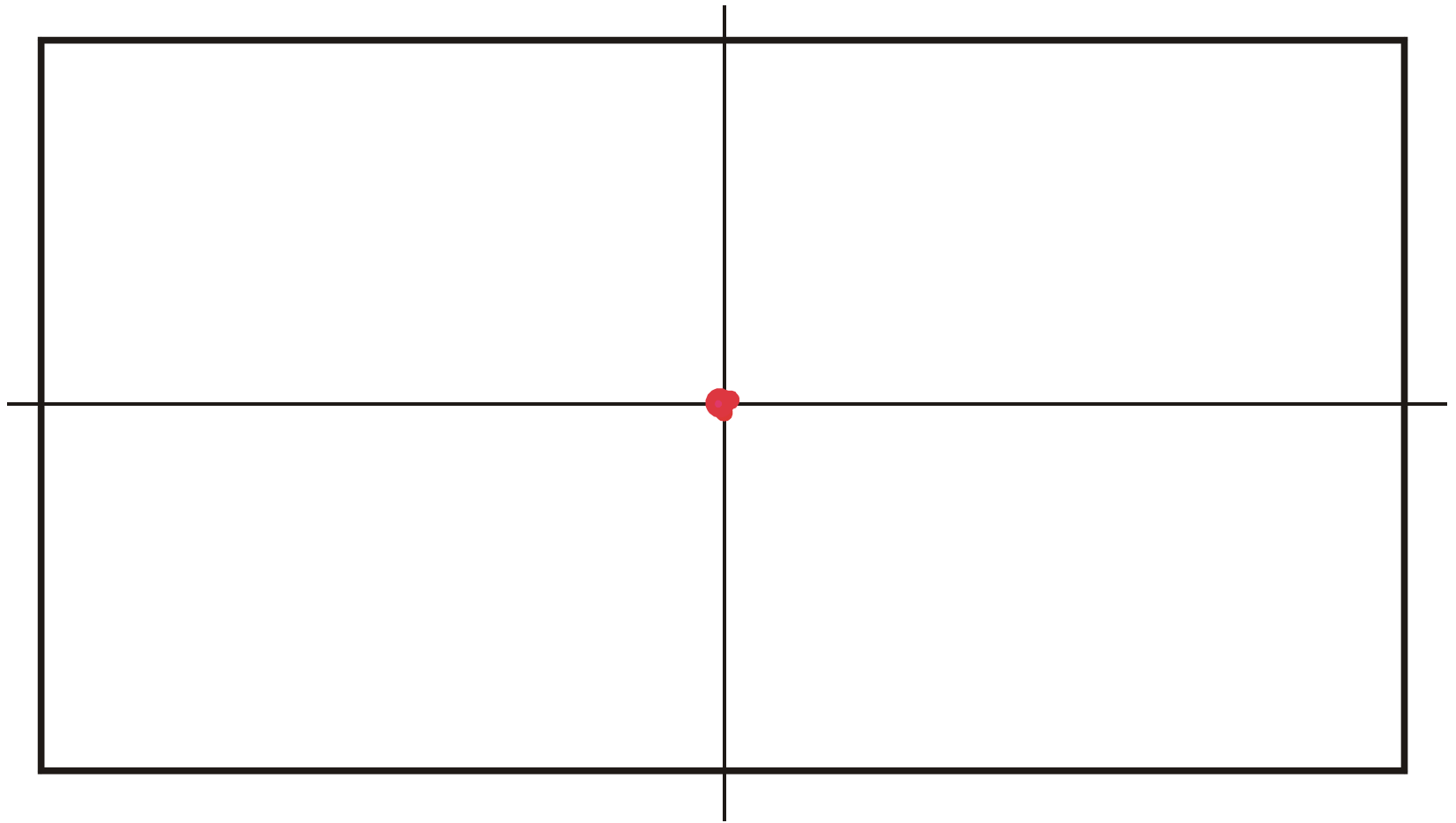
Neutral genotype evolution during phenotypic stasis

Evolutionary trajectory

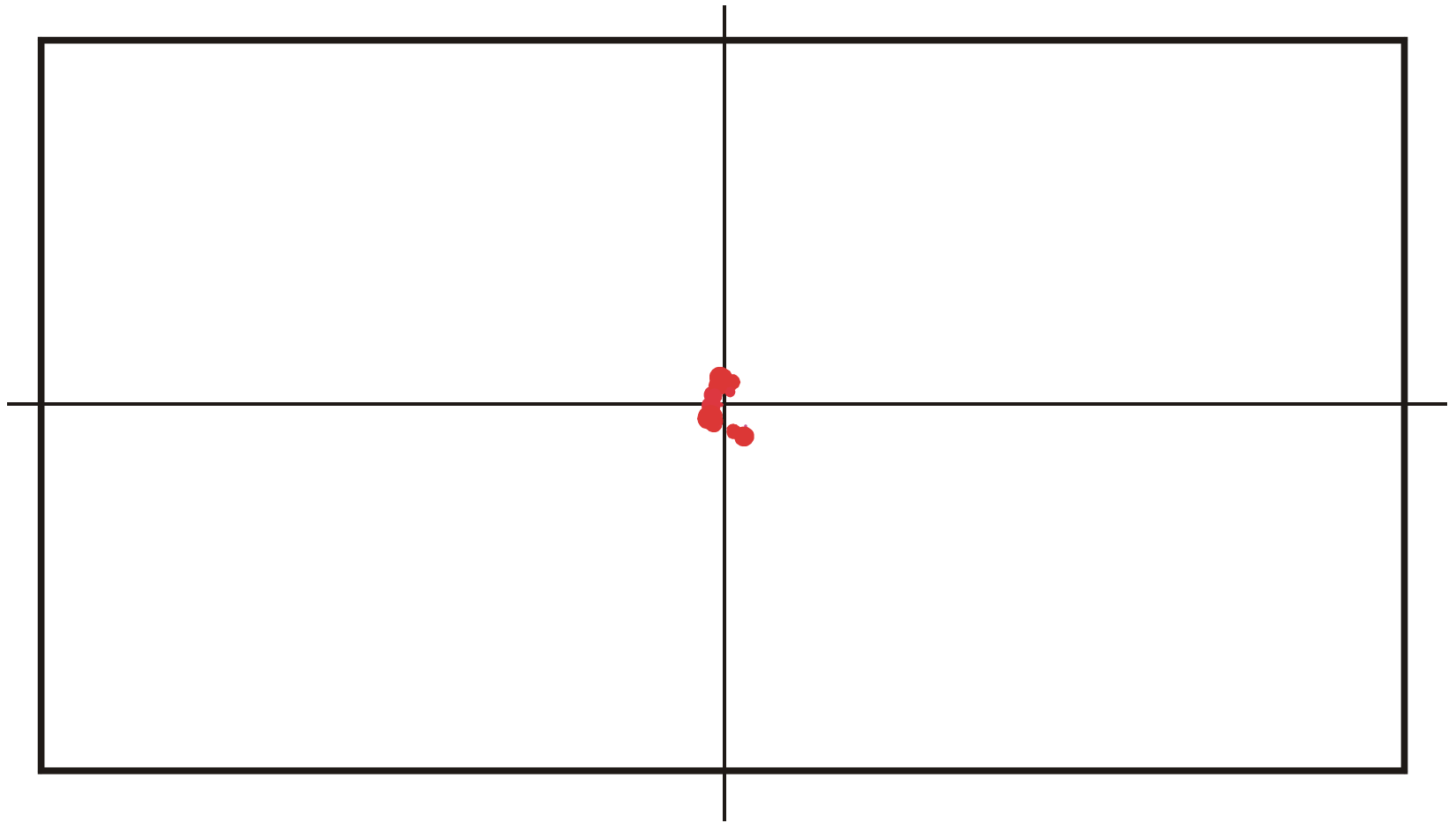
Spreading of the population on neutral networks

Drift of the population center in sequence space

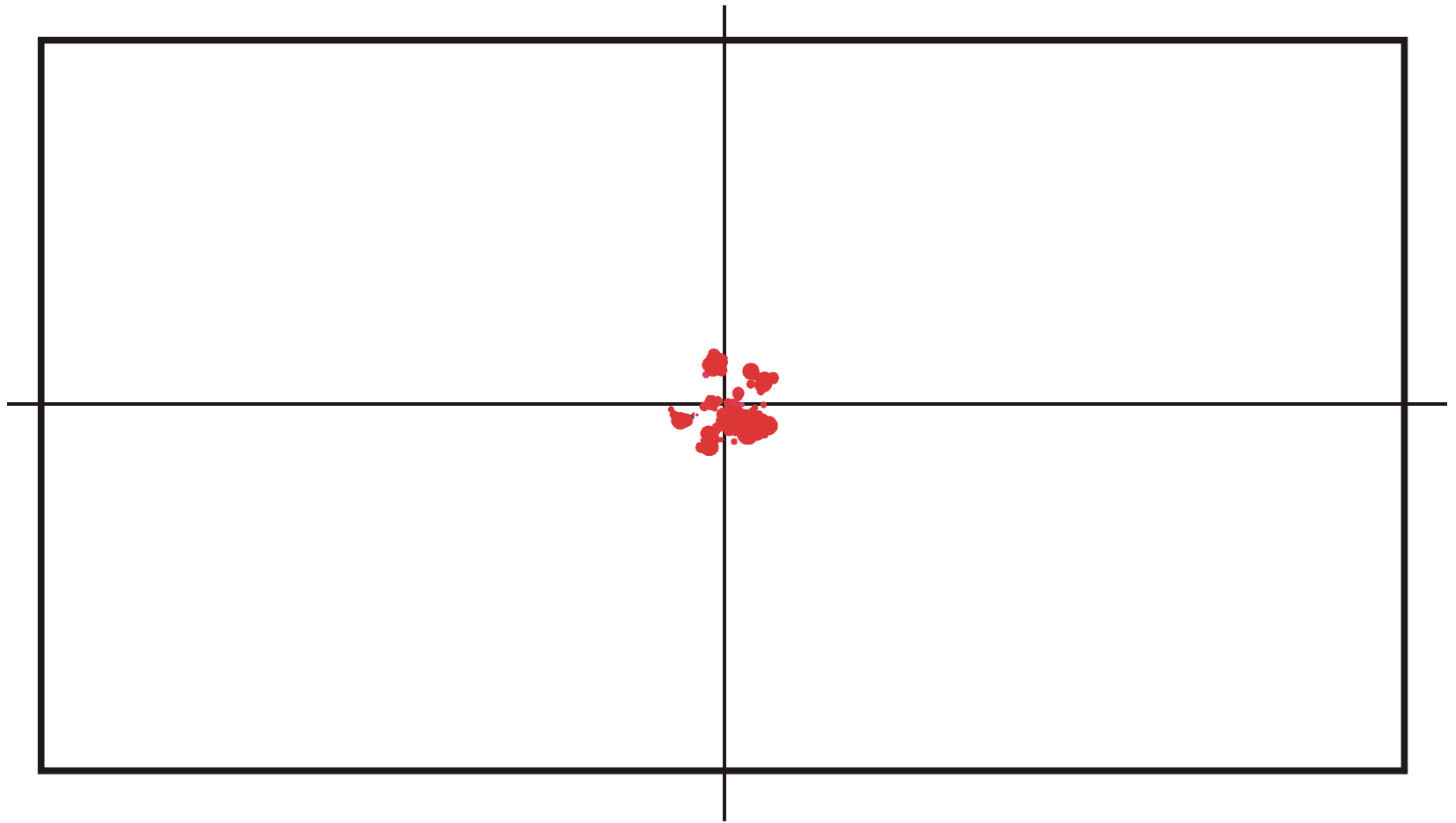




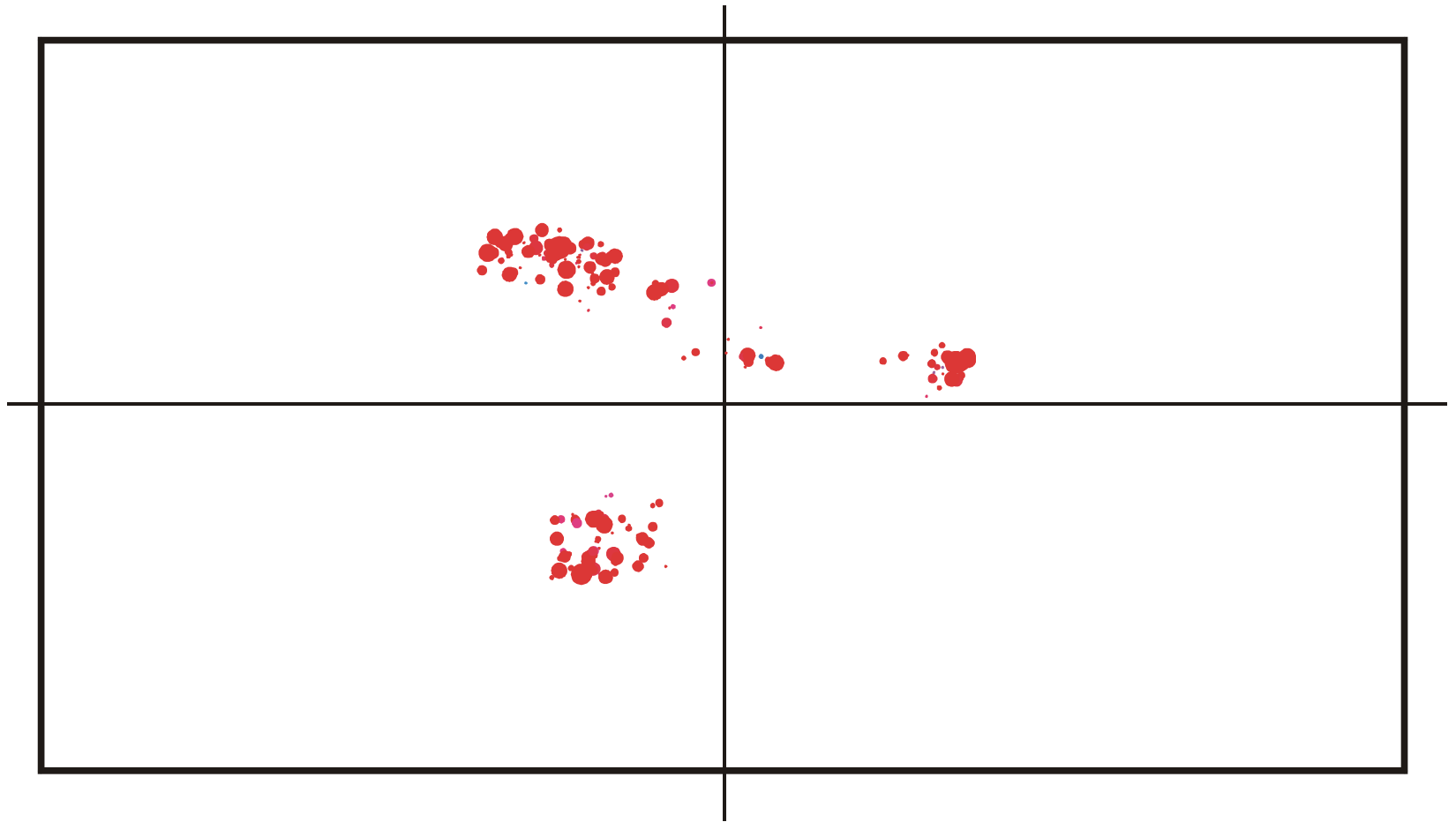
Spreading and evolution of a population on a neutral network: $t = 150$



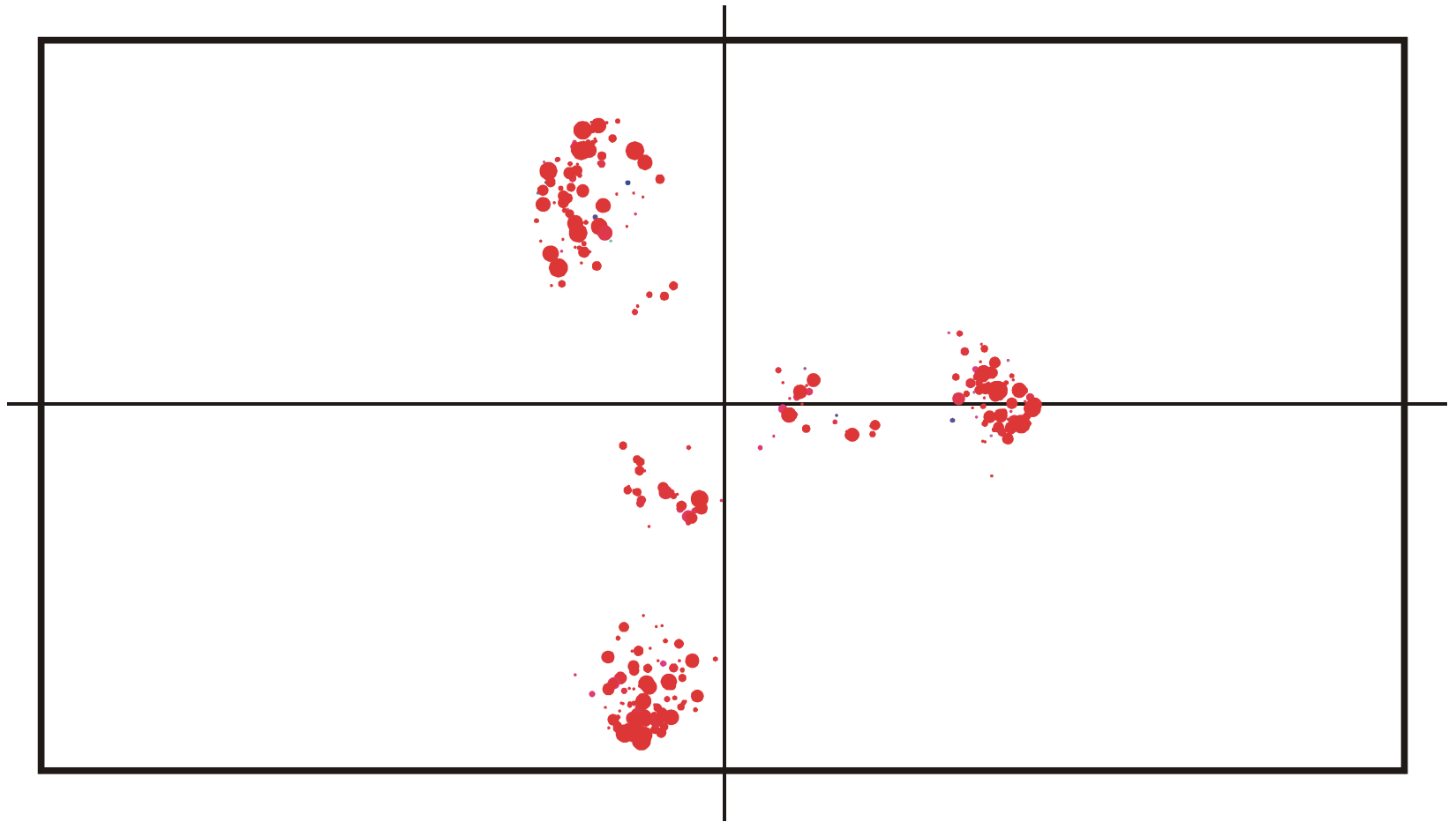
Spreading and evolution of a population on a neutral network : $t = 170$



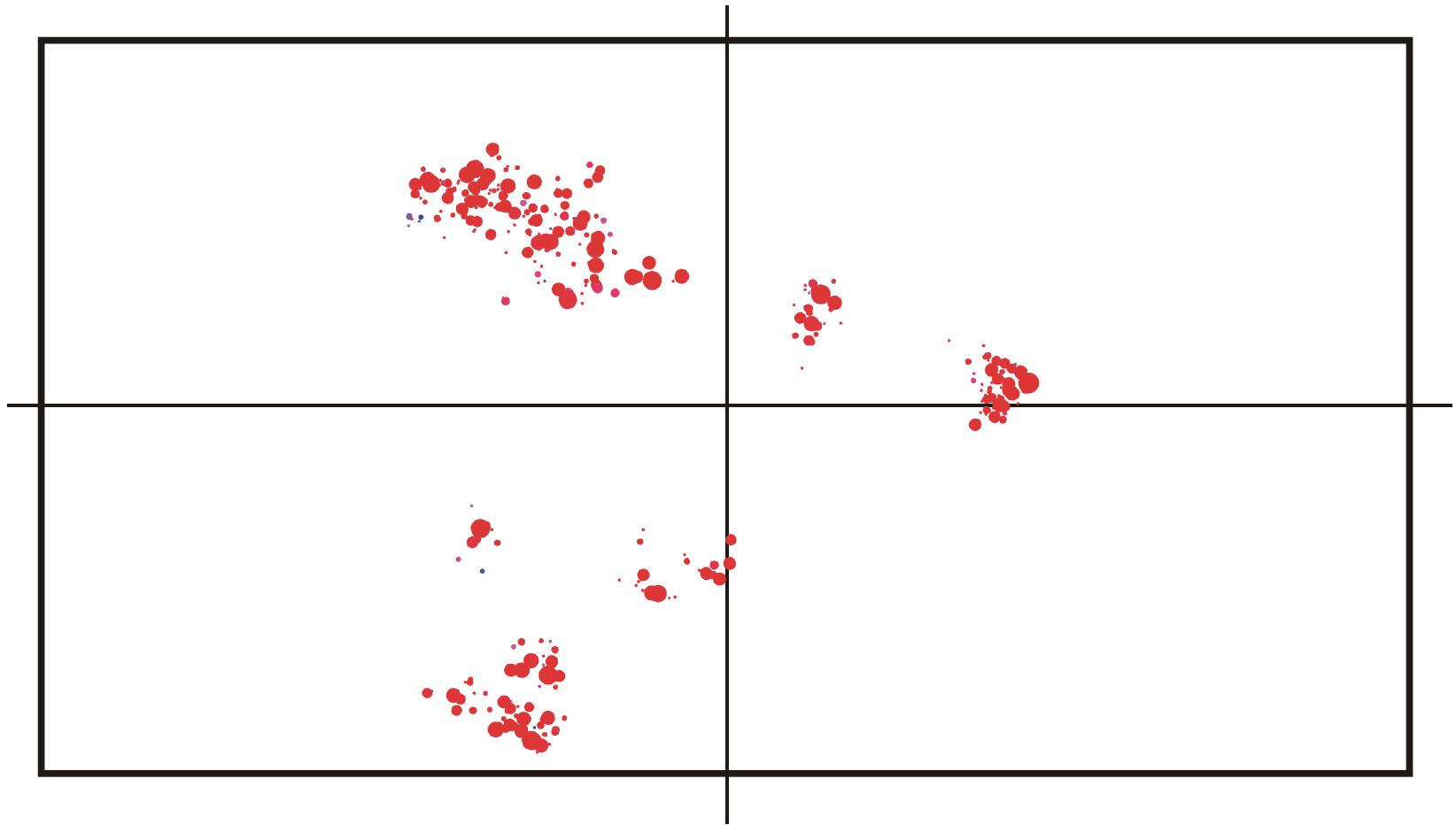
Spreading and evolution of a population on a neutral network : $t = 200$



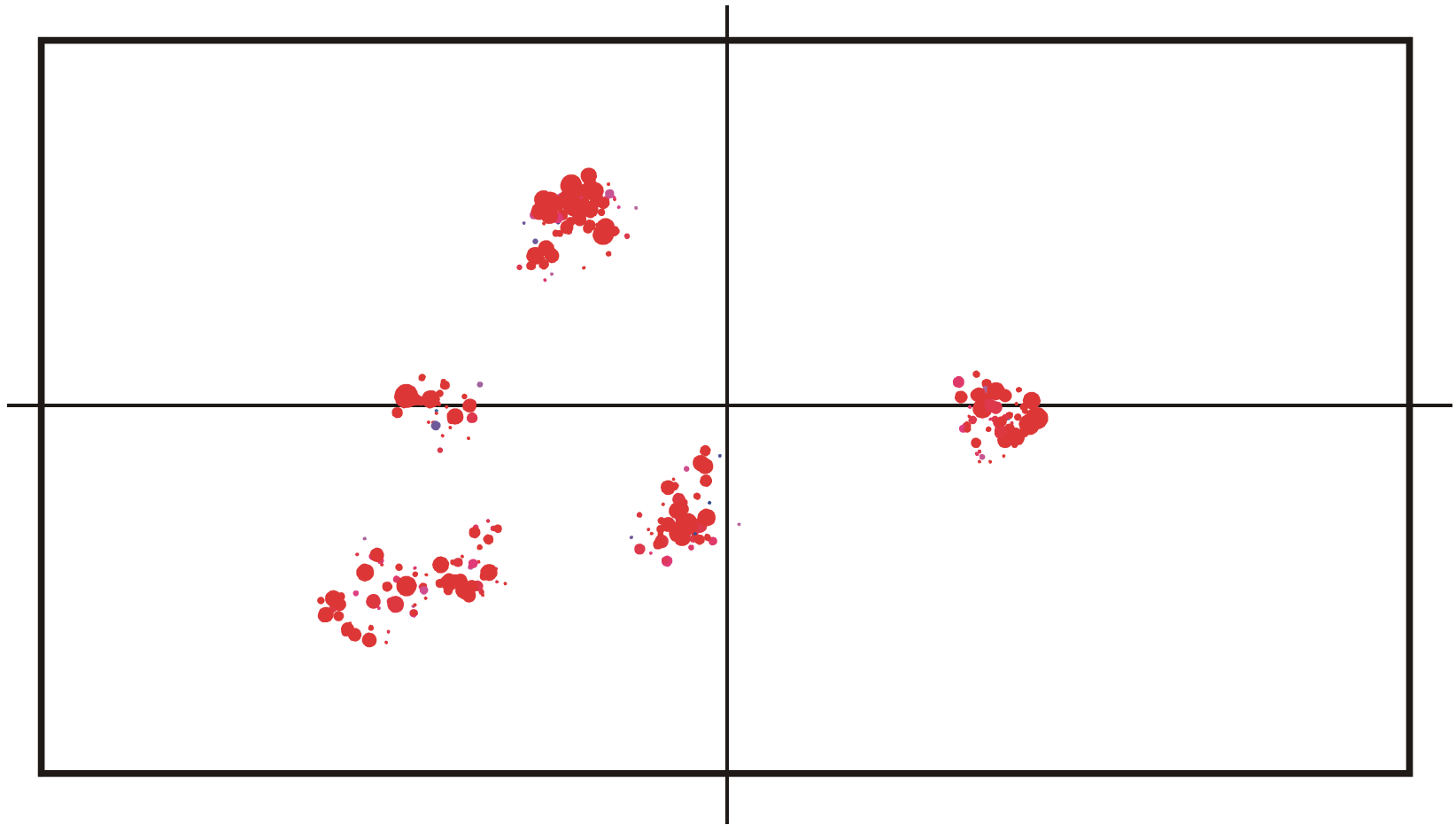
Spreading and evolution of a population on a neutral network : $t = 350$



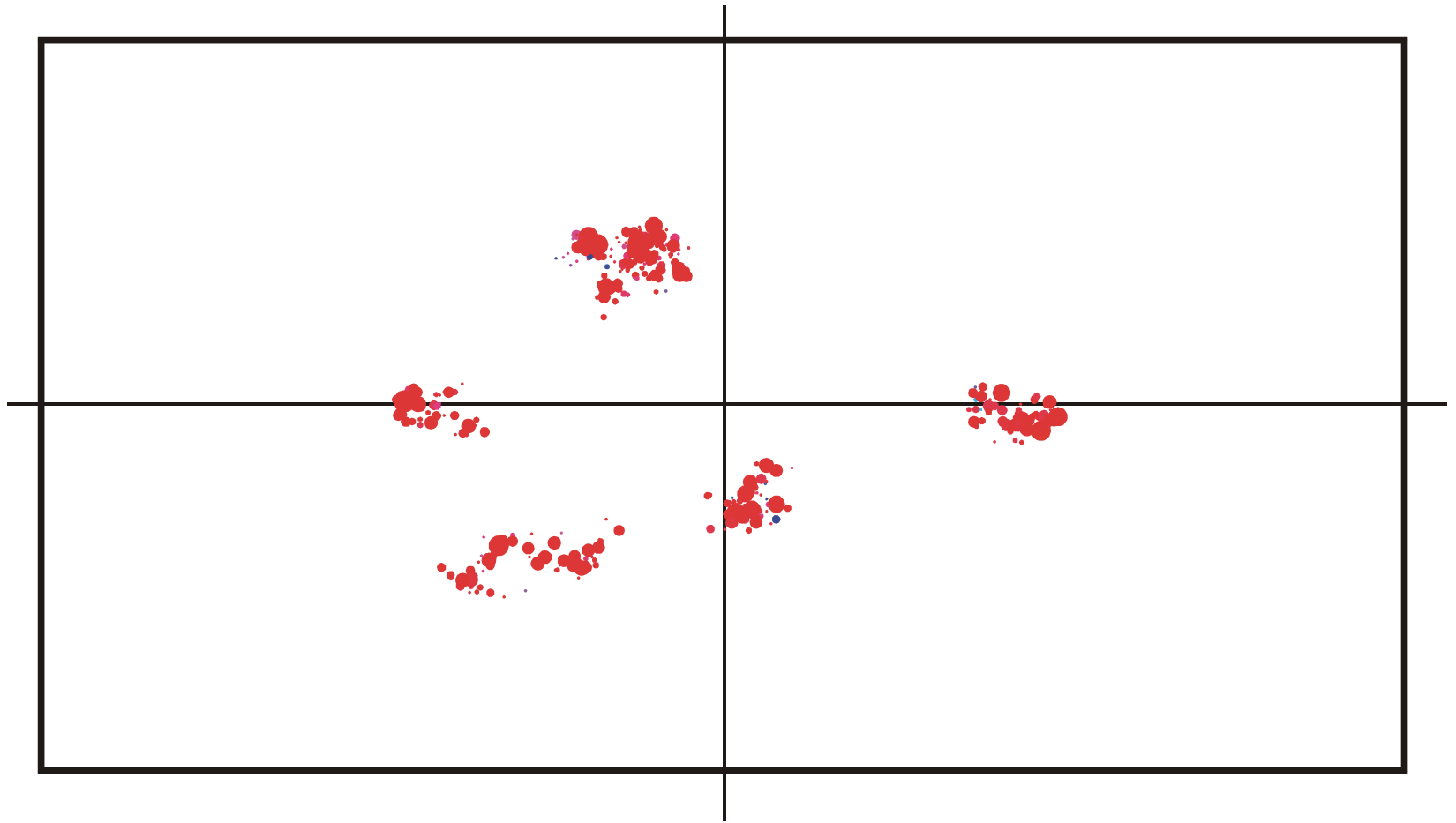
Spreading and evolution of a population on a neutral network : $t = 500$



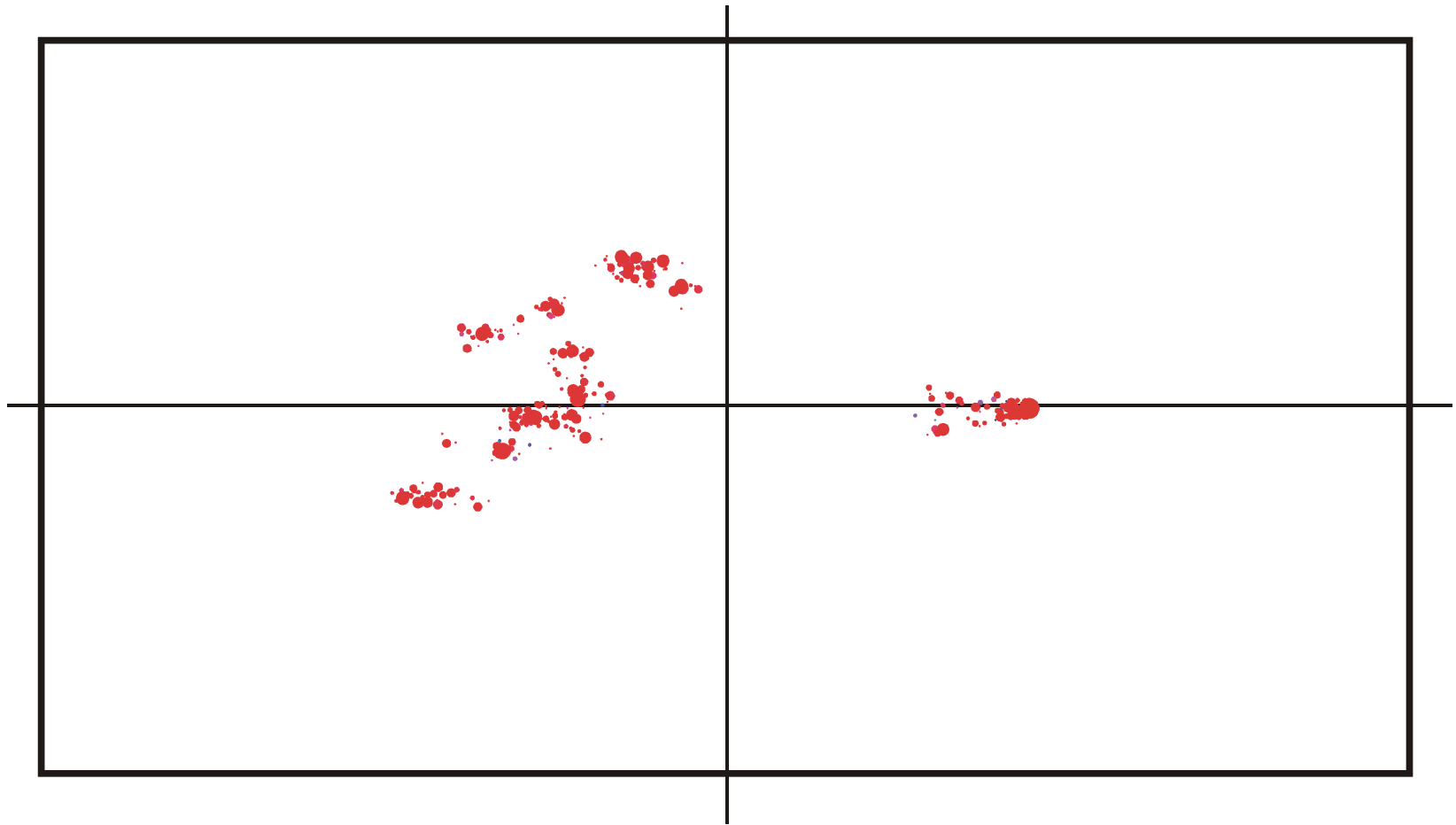
Spreading and evolution of a population on a neutral network : $t = 650$



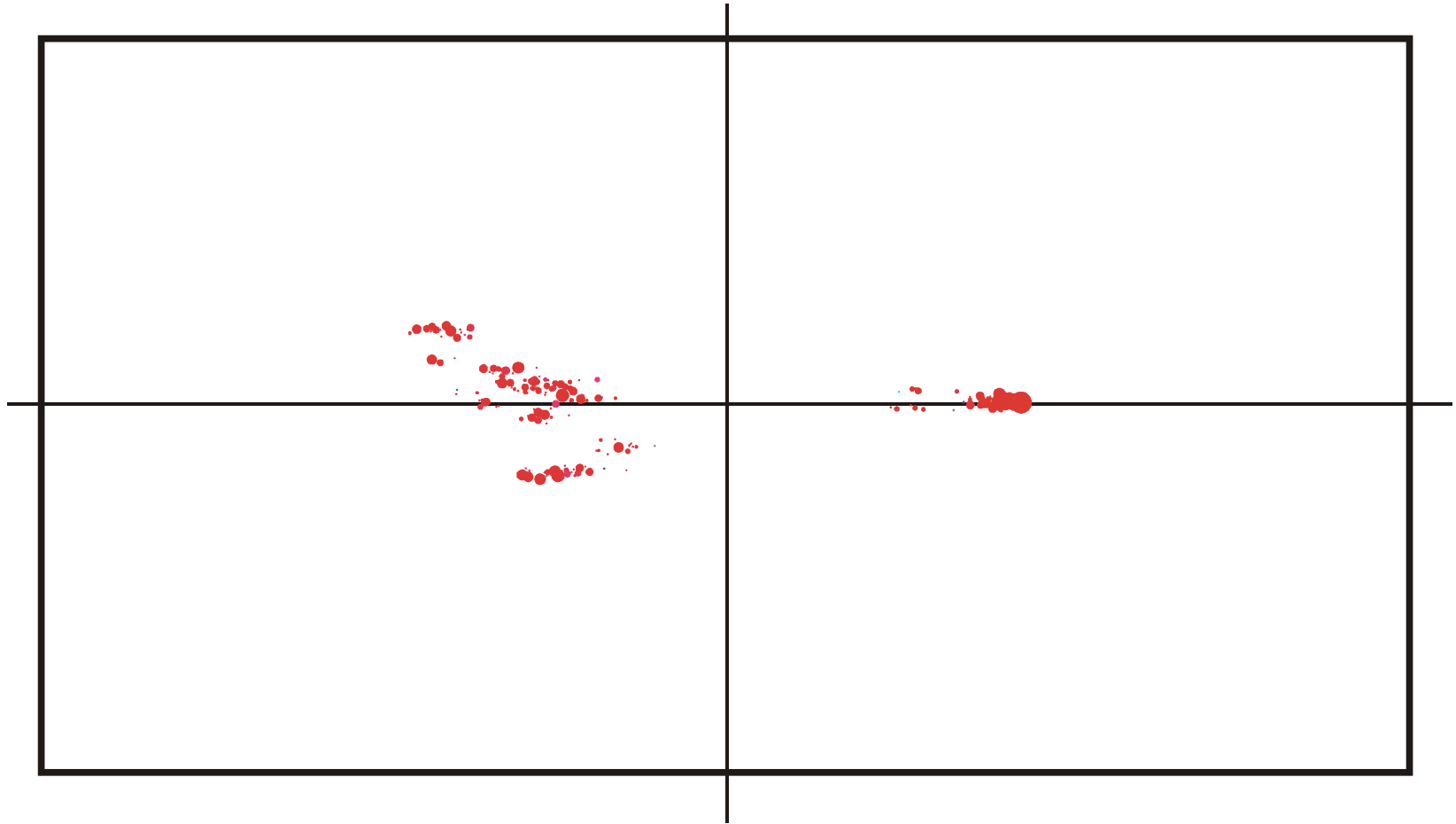
Spreading and evolution of a population on a neutral network : $t = 820$



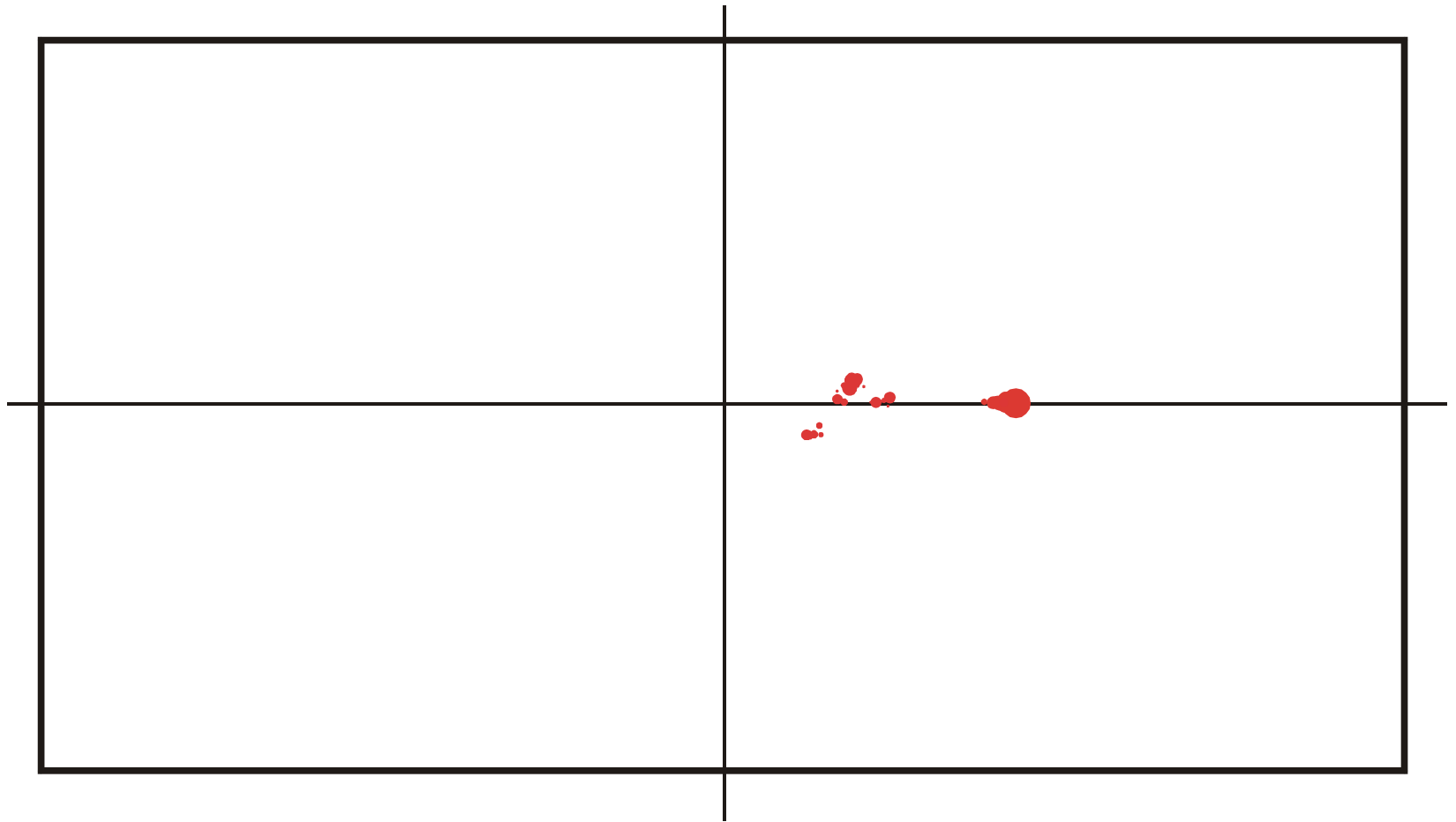
Spreading and evolution of a population on a neutral network : $t = 825$



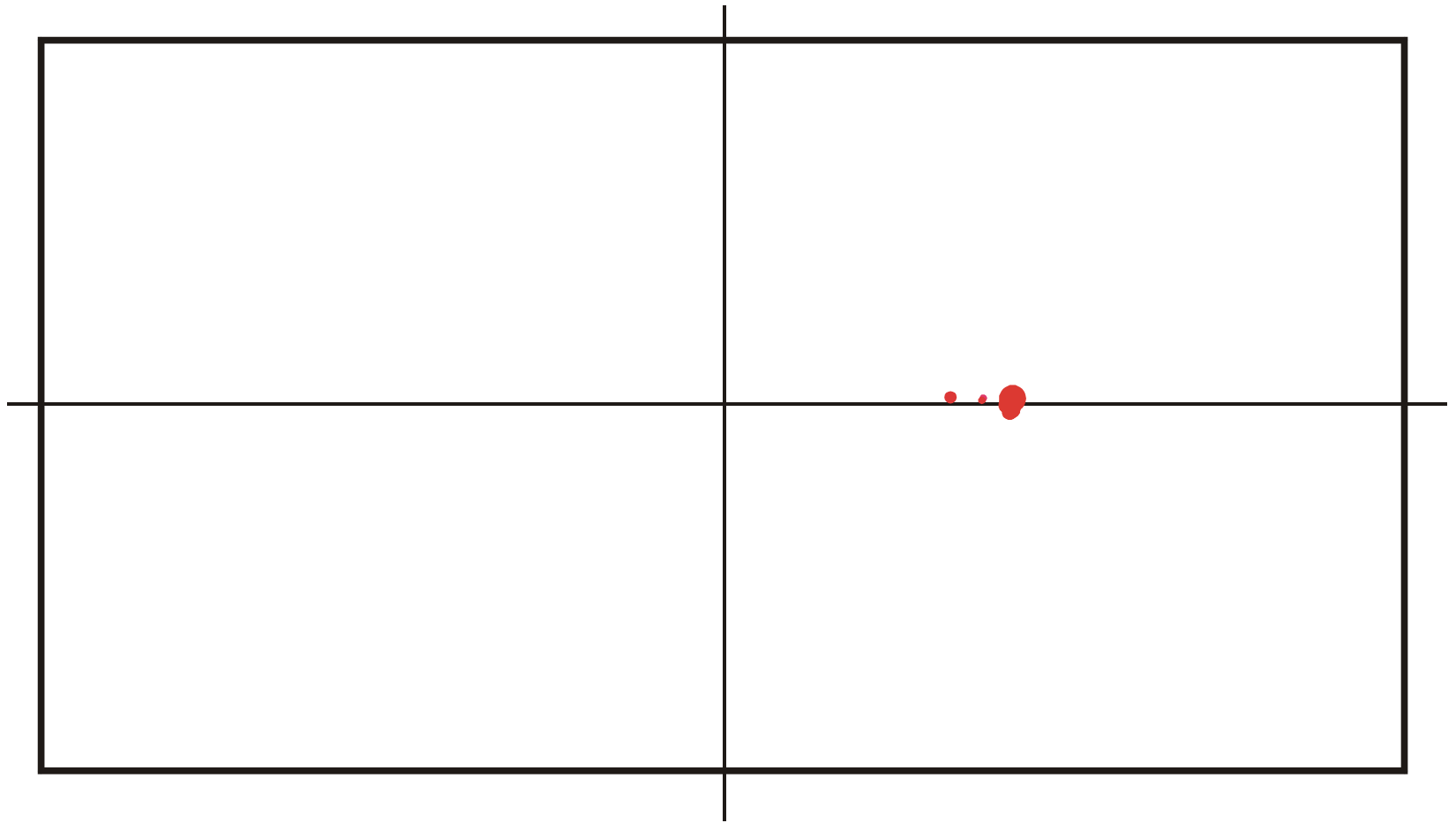
Spreading and evolution of a population on a neutral network : $t = 830$



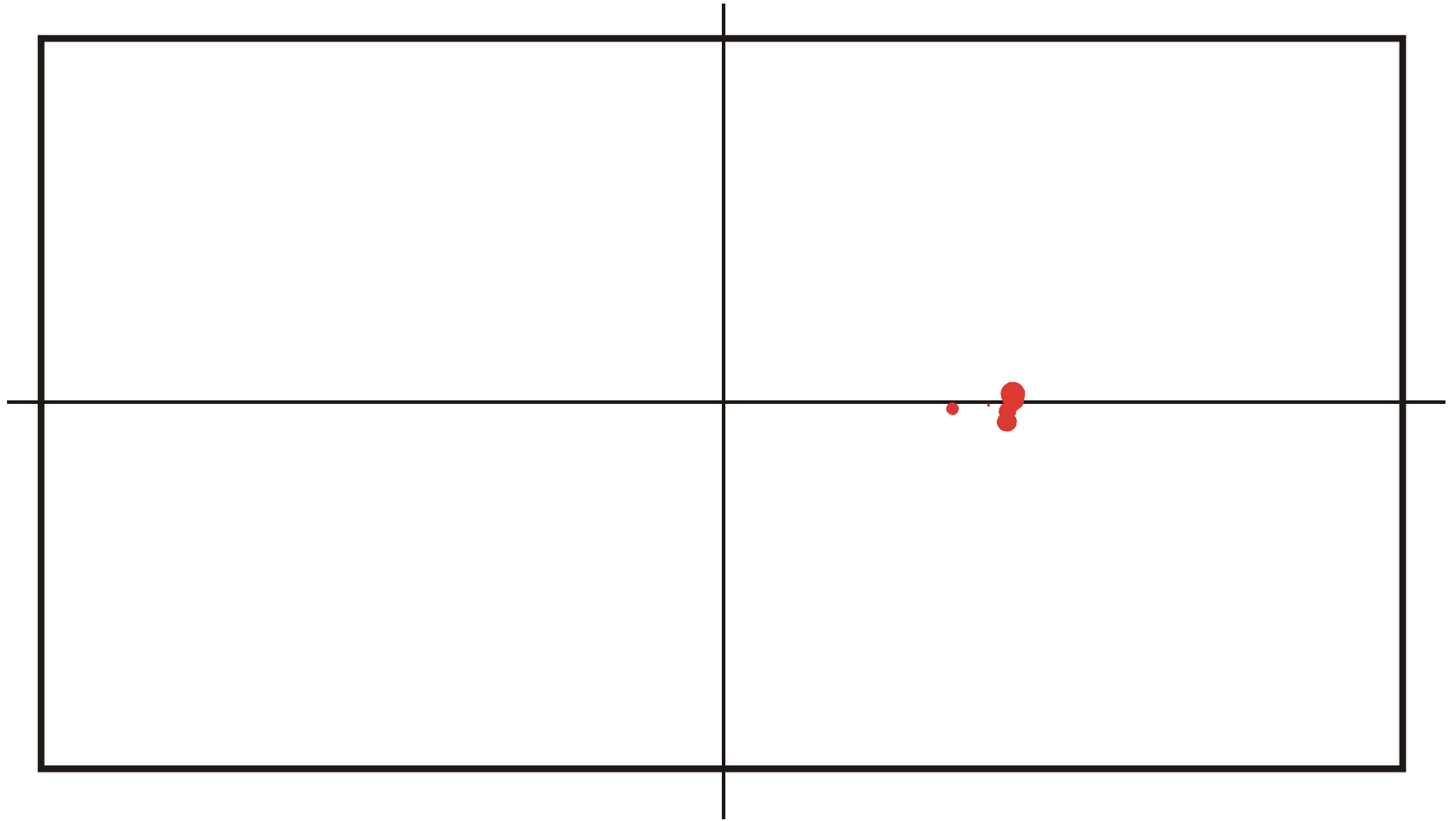
Spreading and evolution of a population on a neutral network : $t = 835$



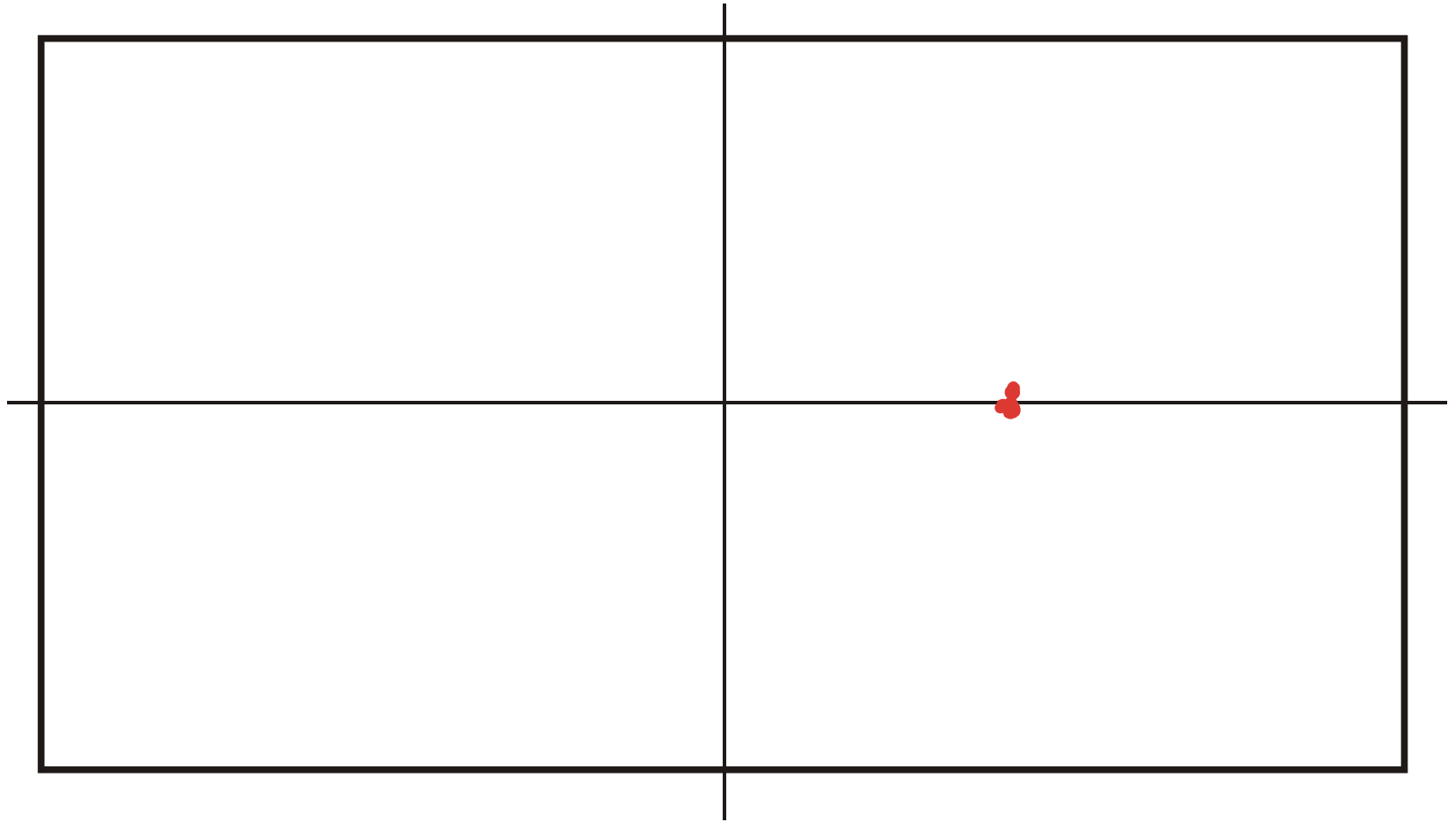
Spreading and evolution of a population on a neutral network : $t = 840$



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



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



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

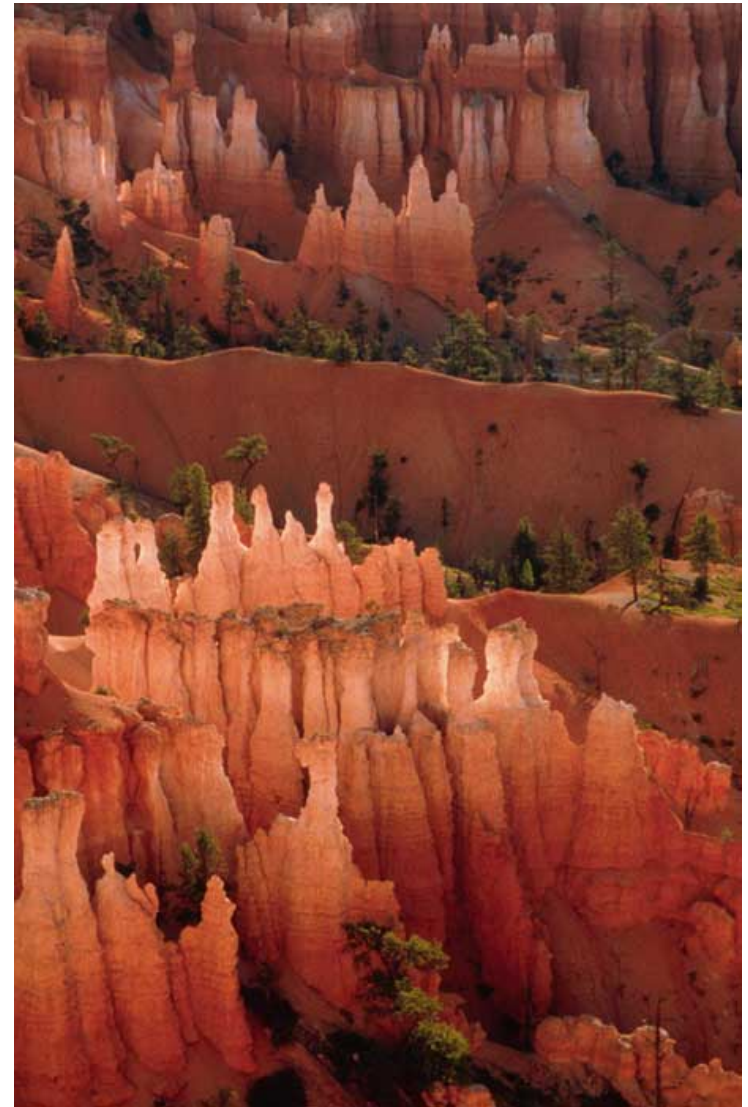


Mount Fuji

Example of a smooth landscape on Earth

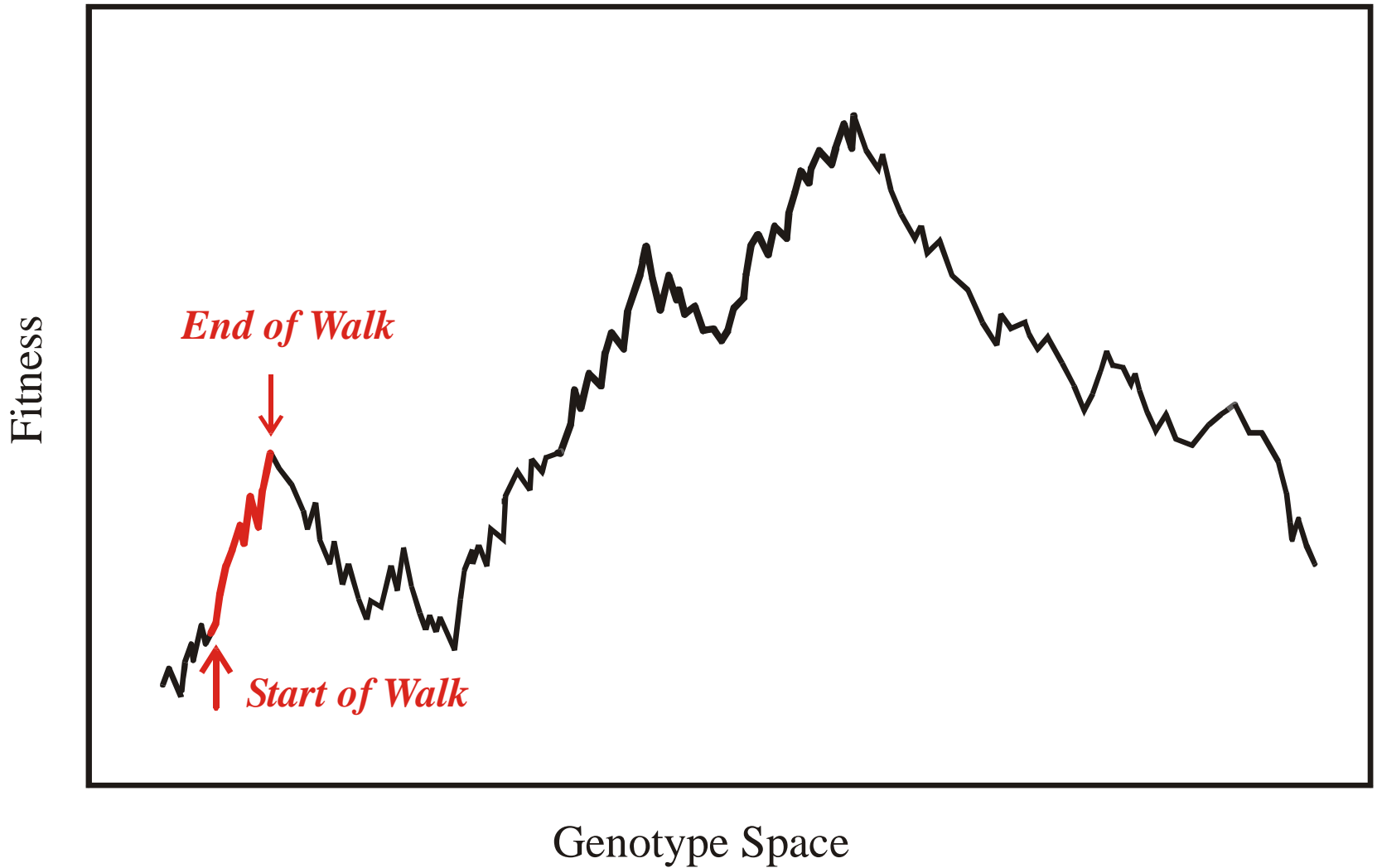


Dolomites

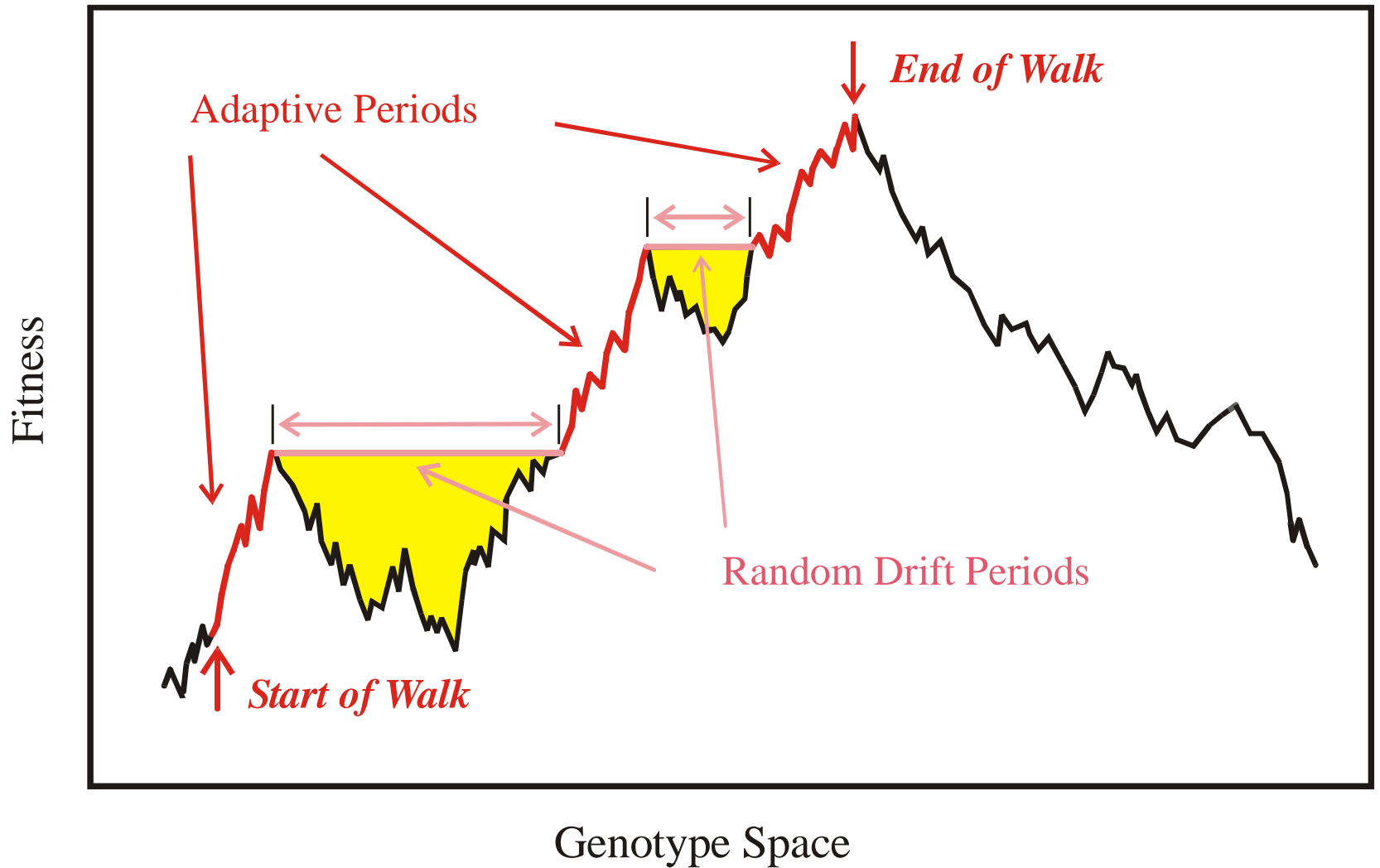


Bryce Canyon

Examples of rugged landscapes on Earth



Evolutionary optimization in absence of neutral paths in sequence space



Evolutionary optimization including neutral paths in sequence space



Grand Canyon

Example of a landscape on Earth with ‘neutral’
ridges and plateaus

1. RNA sequences and structures
2. Neutral networks
3. Evolutionary optimization of structure
- 4. Suboptimal structures and kinetic folding**
5. Comparison of kinetic folding and evolution

The Folding Algorithm

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

$$\mathfrak{S}(\mathbf{I}) = \{\mathbf{S}_0, \mathbf{S}_1, \dots, \mathbf{S}_m, \mathbf{O}\}$$

A trajectory $\mathfrak{Z}_k(\mathbf{I})$ is a time ordered series of structures in $\mathfrak{S}(\mathbf{I})$. A folding trajectory is defined by starting with the open chain \mathbf{O} and ending with the global minimum free energy structure \mathbf{S}_0 or a metastable structure \mathbf{S}_k which represents a local energy minimum:

$$\mathfrak{Z}_0(\mathbf{I}) = \{\mathbf{O}, \mathbf{S}(1), \dots, \mathbf{S}(t-1), \mathbf{S}(t), \mathbf{S}(t+1), \dots, \mathbf{S}_0\}$$

$$\mathfrak{Z}_k(\mathbf{I}) = \{\mathbf{O}, \mathbf{S}(1), \dots, \mathbf{S}(t-1), \mathbf{S}(t), \mathbf{S}(t+1), \dots, \mathbf{S}_k\}$$

Formulation of kinetic RNA folding as a stochastic process

Master equation

$$\frac{dP_k}{dt} = \sum_{i=0}^{m+1} (P_{ik}(t) - P_{ki}(t)) = \sum_{i=0}^{m+1} k_{ik} P_i - P_k \sum_{i=0}^{m+1} k_{ki}$$

$$k = 0, 1, \dots, m+1$$

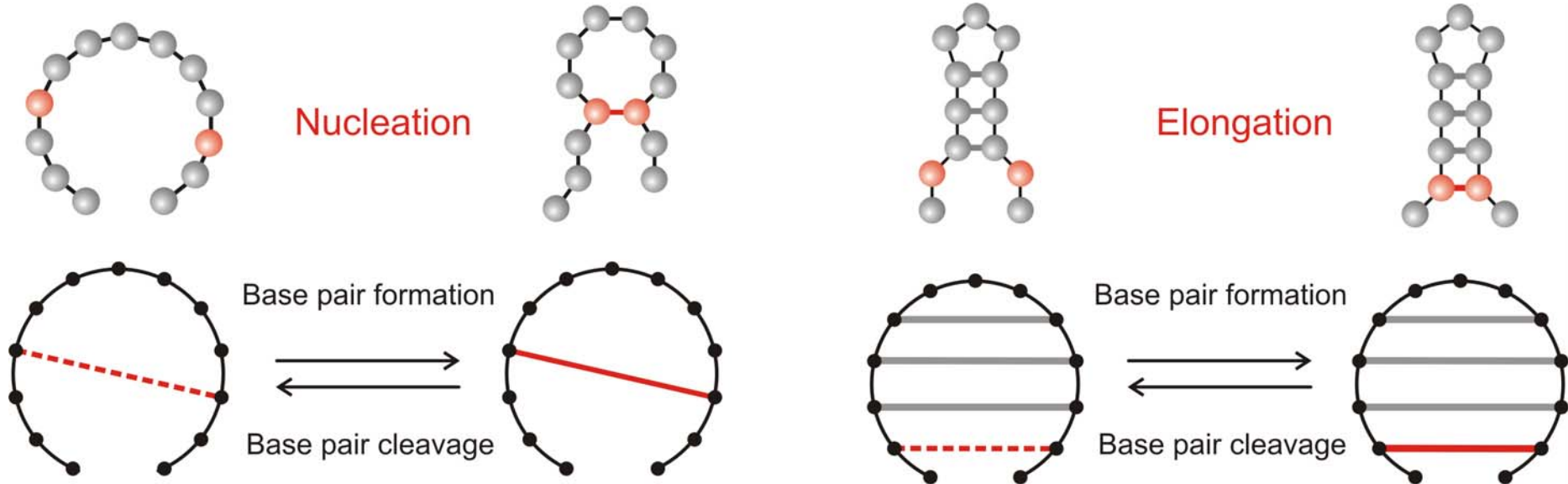
Transition probabilities $P_{ij}(t) = \text{Prob}\{\mathbf{S}_i \rightarrow \mathbf{S}_j\}$ are defined by

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

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

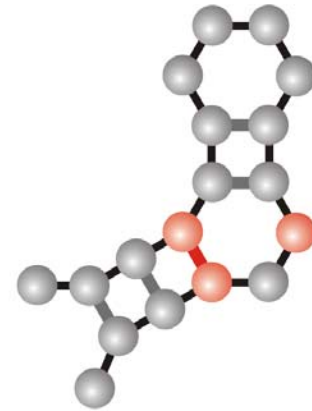
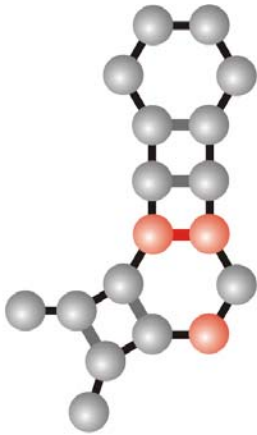
$$\Sigma_k = \sum_{k=1, k \neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

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



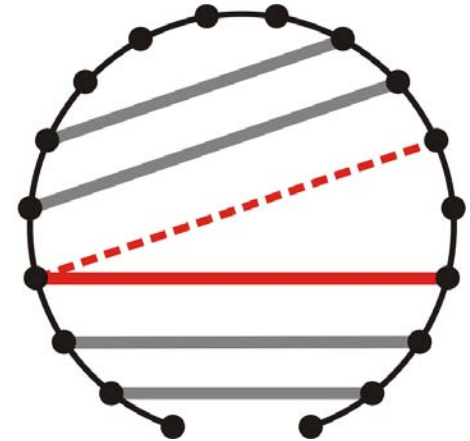
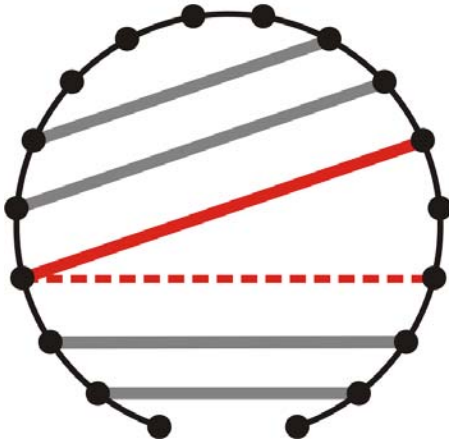
Corresponds to base pair distance: $d_p(S_1, S_2)$

Base pair formation and base pair cleavage moves for nucleation and elongation of stacks



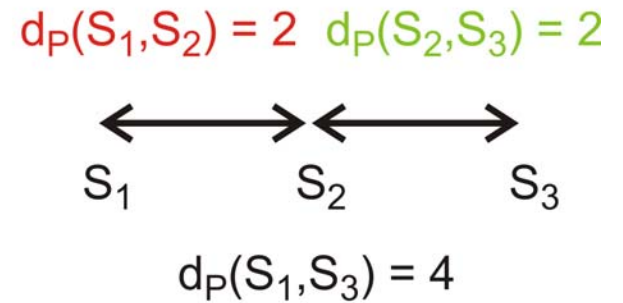
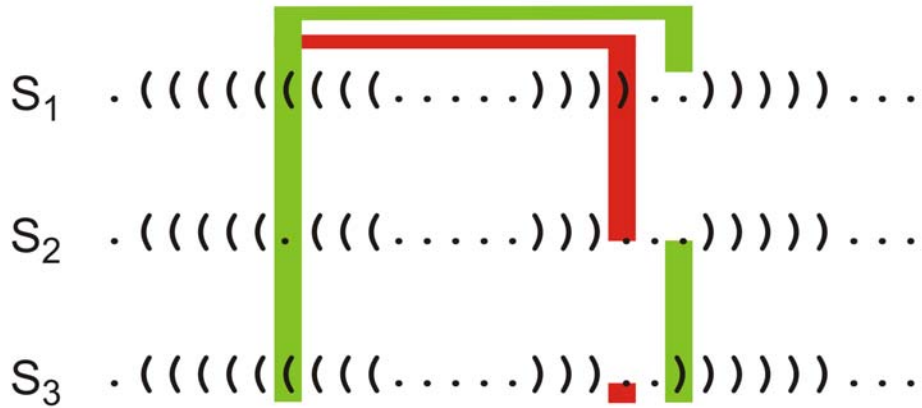
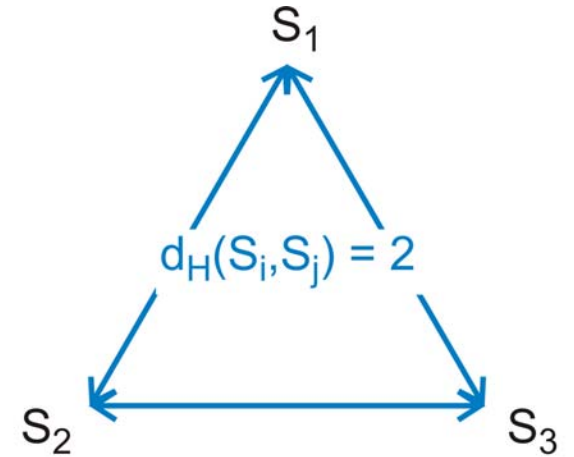
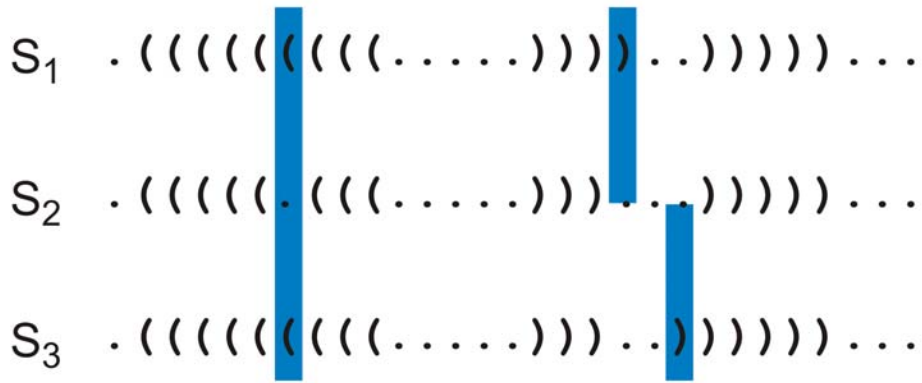
Base pair shift

Class 1



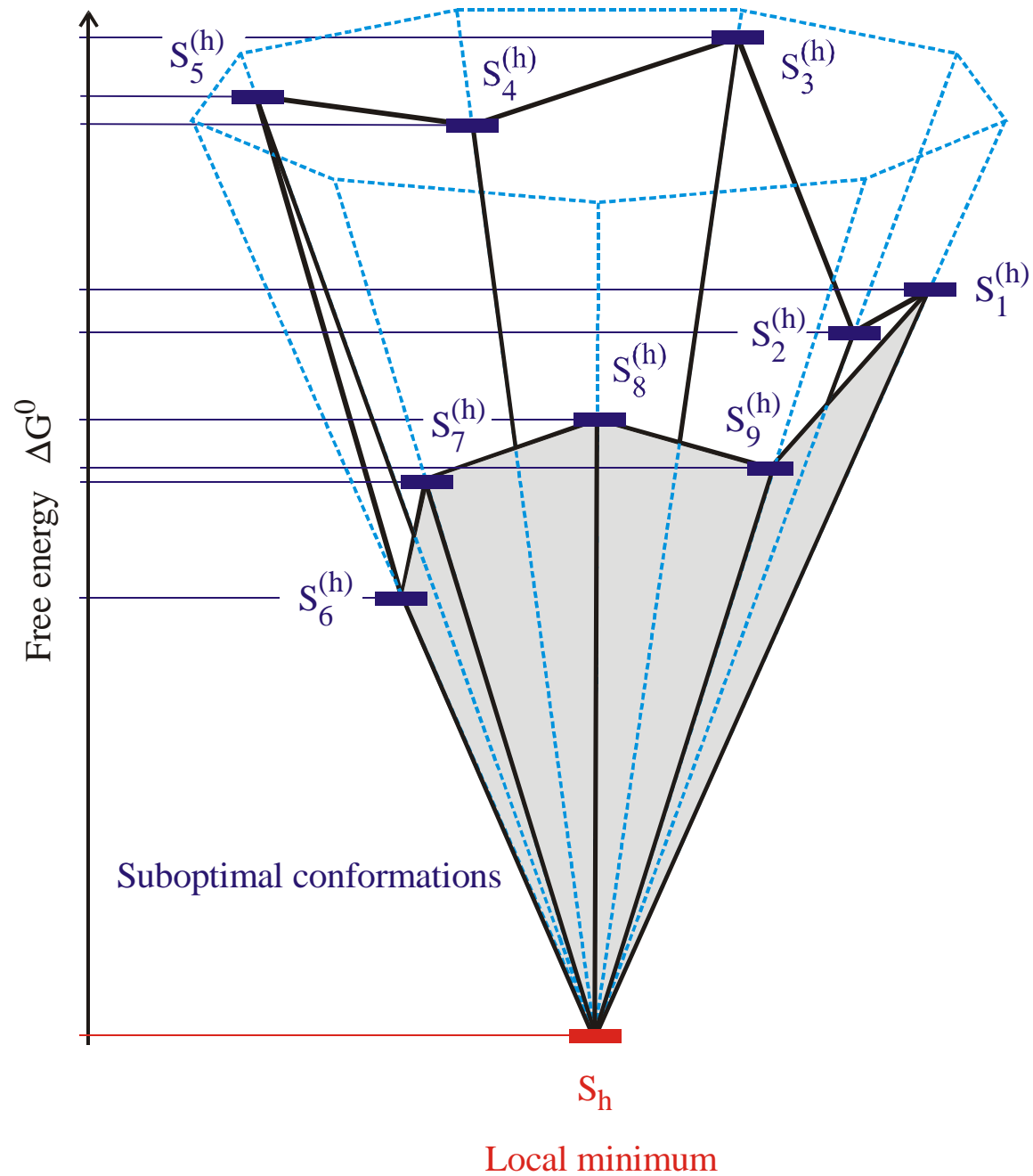
Base pair closure, opening and shift corresponds to Hamming distance: $d_H(S_1, S_2)$

Base pair shift move of class 1: Shift inside internal loops or bulges

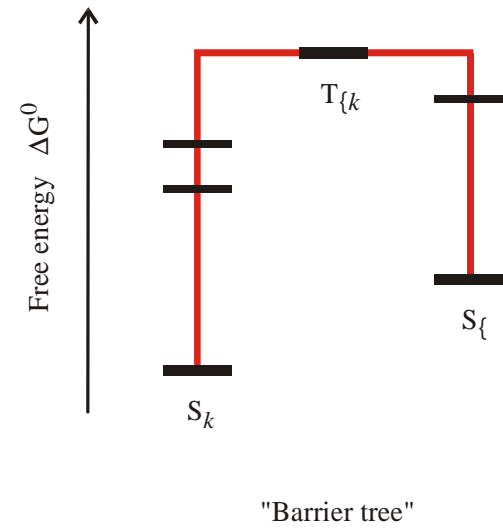
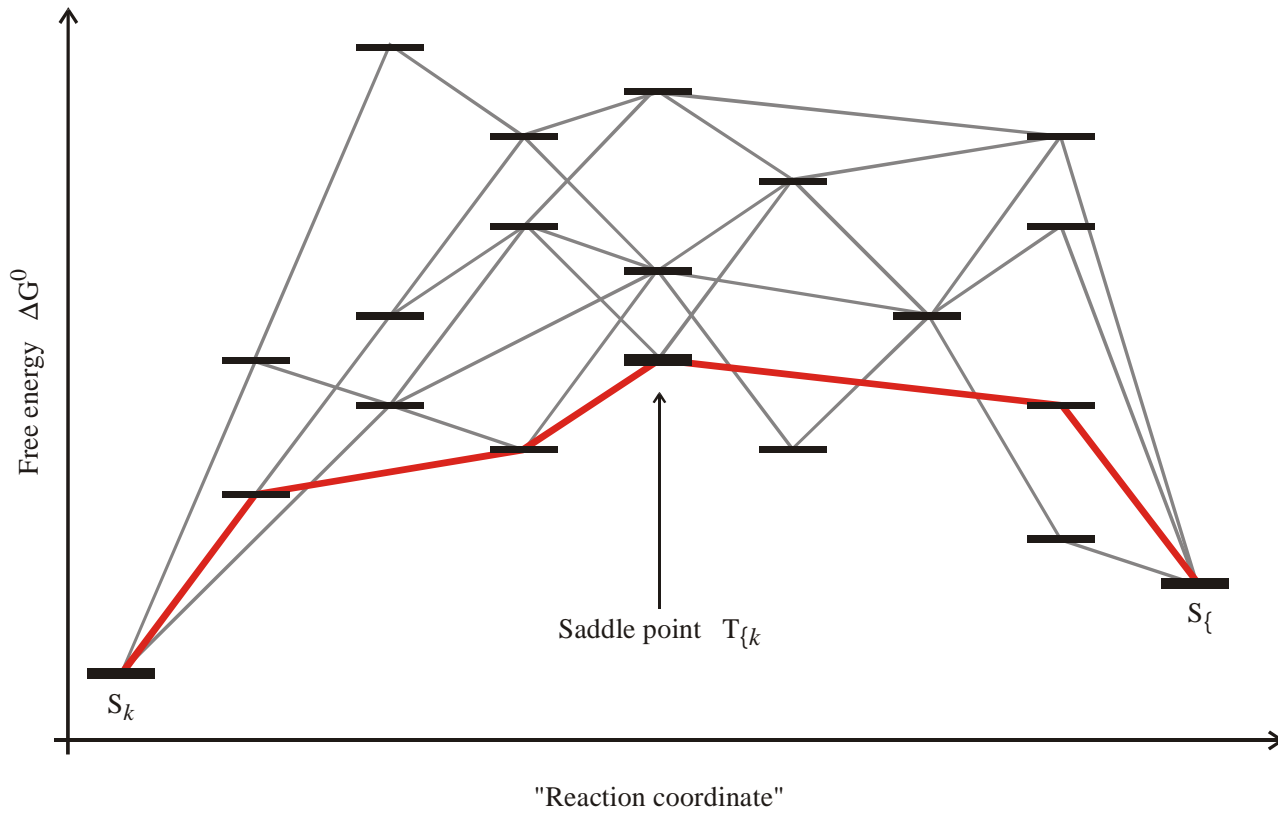


Two measures of distance in shape space:

Hamming distance between structures, $d_H(S_i, S_j)$ and base pair distance, $d_P(S_i, S_j)$



Search for local minima in conformation space



Definition of a ,barrier tree‘

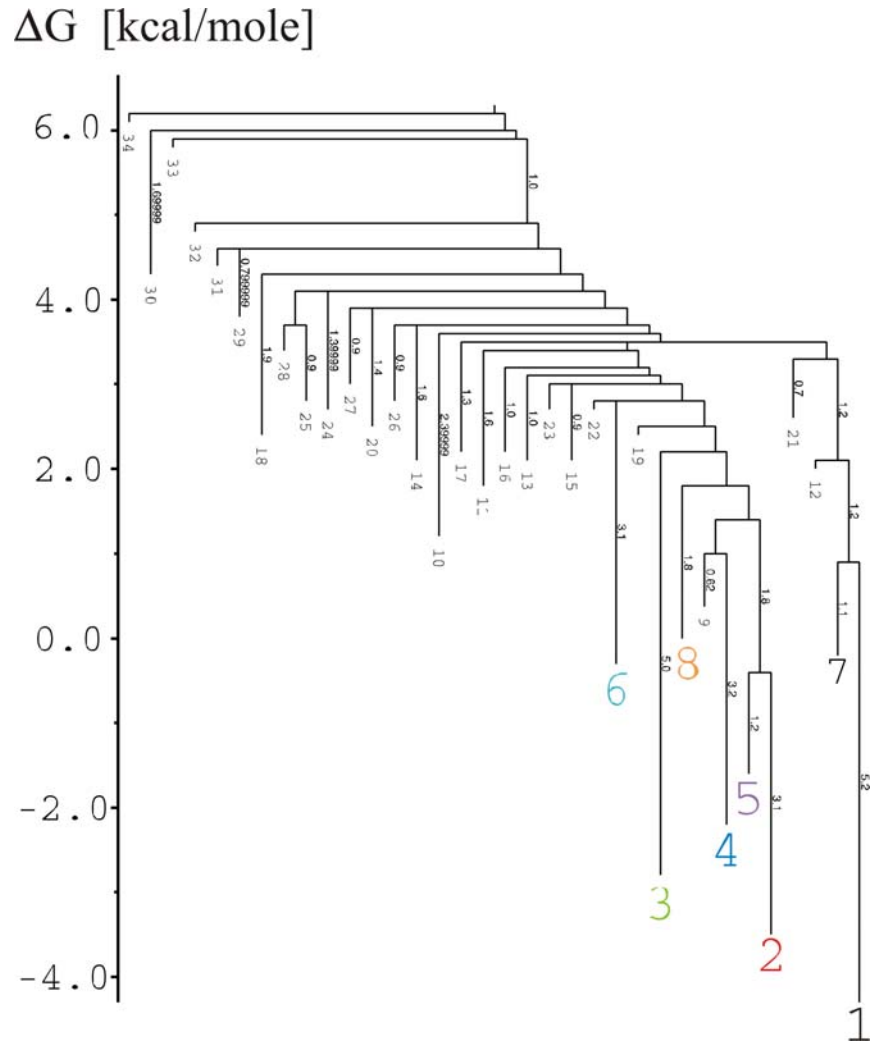
CUGCGGCUUUGGCUCUAGCC

.....((((.....)))))	-4.30
(((.....)).....)	-3.50
((.....)).....	-3.10
.....(((.....)))	-2.80
.....(((.....)))	-2.20
.....(((.....)))	-2.20
((.....)).....	-2.00
.....(((.....)))	-1.60
.....(((.....)))	-1.60
.....(((.....)))	-1.50
.....(((.....)))	-1.40
.....(((.....)))	-1.40
.....(((.....)))	-1.00
.....(((.....)))	-0.90
.....(((.....)))	-0.90
.....(((.....)))	-0.80
.....(((.....)))	-0.80
.....(((.....)))	-0.60
.....(((.....)))	-0.60
.....(((.....)))	-0.50
.....(((.....)))	-0.50
.....(((.....)))	-0.40
.....(((.....)))	-0.30
.....(((.....)))	-0.30
.....(((.....)))	-0.20
.....(((.....)))	-0.20
.....(((.....)))	-0.20
.....(((.....)))	0.00
.....(((.....)))	0.00
.....(((.....)))	0.10



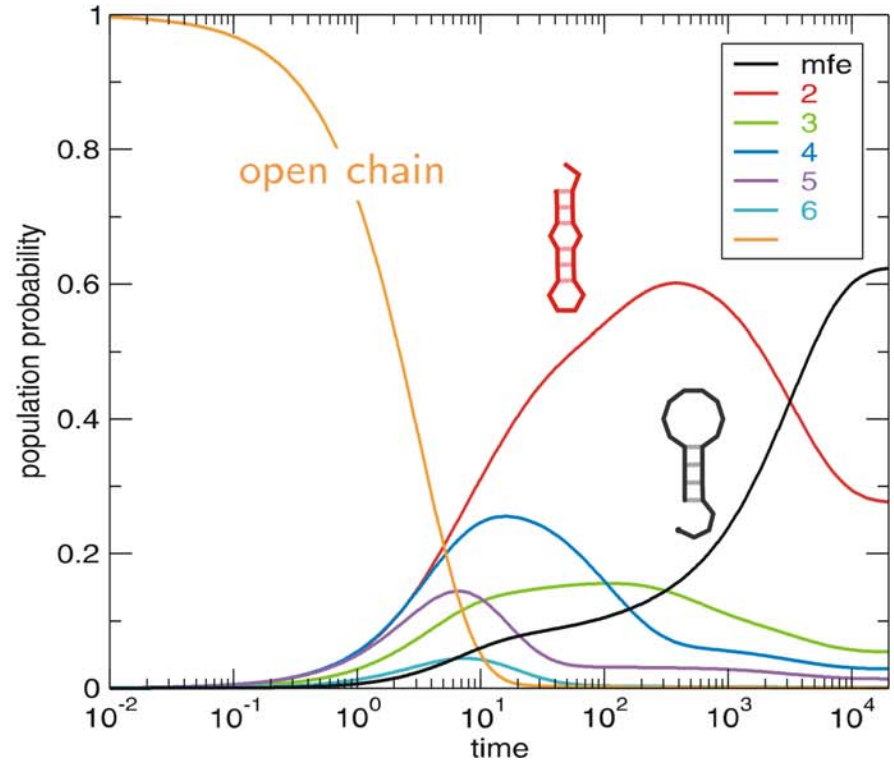
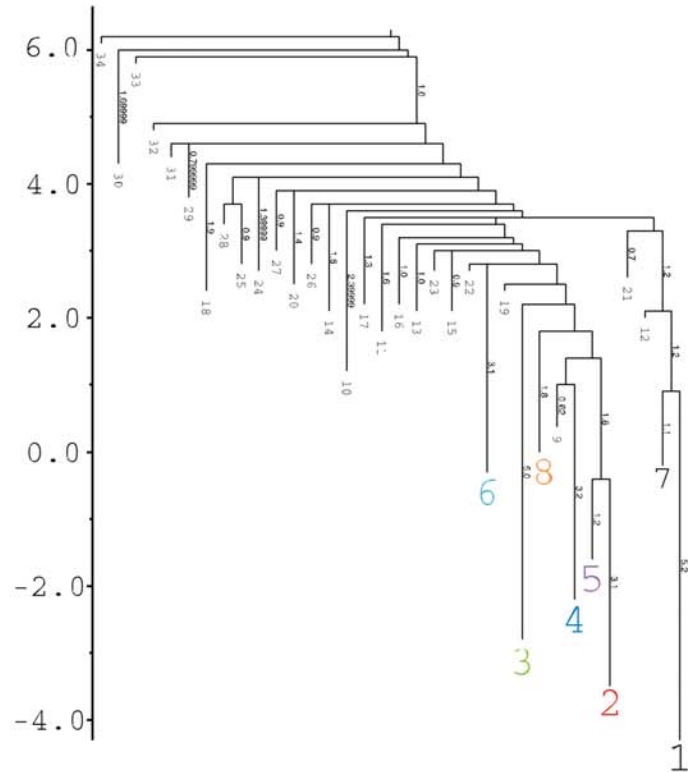
M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:*
Math.Gen. **37**:4731-4741.

CUGCGGCUUUGGCUCUAGCC	
.....((((.....)))))	-4.30
(((.....)).....))..	-3.50
((.....)).....)	-3.10
.....(((.....)))	-2.80
.....((((.....)).....)	-2.20
.....((((.....)).....)	-2.20
((.....)).....)	-2.00
.....((((.....)).....)	-1.60
.....((((.....)).....)	-1.60
.....((((.....)).....)	-1.50
.....((((.....)).....)	-1.40
.....((((.....)).....)	-1.40
.....((((.....)).....)	-1.00
.....((((.....)).....)	-0.90
.....((((.....)).....)	-0.90
.....((((.....)).....)	-0.80
.....((((.....)).....)	-0.80
.....((((.....)).....)	-0.60
.....((((.....)).....)	-0.60
.....((((.....)).....)	-0.50
.....((((.....)).....)	-0.50
.....((((.....)).....)	-0.40
.....((((.....)).....)	-0.30
.....((((.....)).....)	-0.30
.....((((.....)).....)	-0.20
.....((((.....)).....)	-0.20
.....((((.....)).....)	0.00
.....((((.....)).....)	0.00
.....((((.....)).....)	0.10



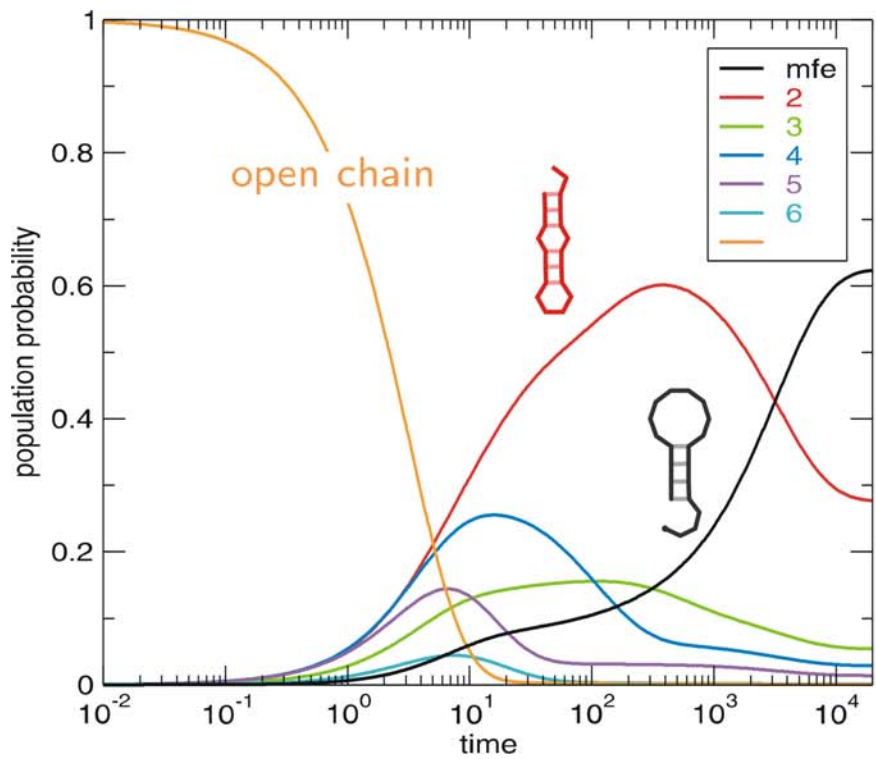
M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A: Math.Gen.* **37**:4731-4741.

ΔG [kcal/mole]

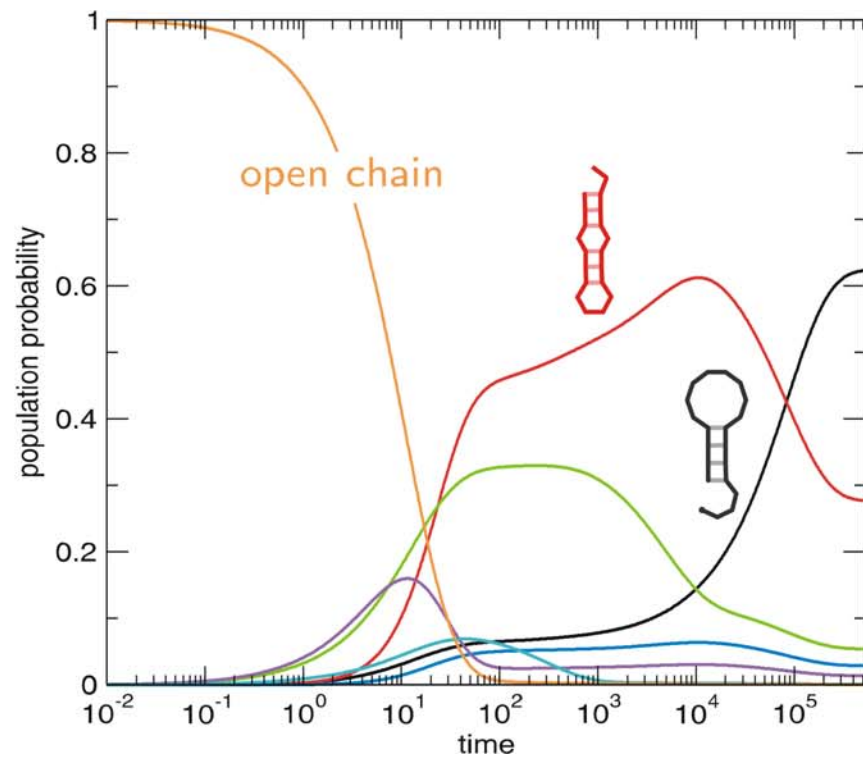


Arrhenius kinetics

M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:*
Math.Gen. **37**:4731-4741.



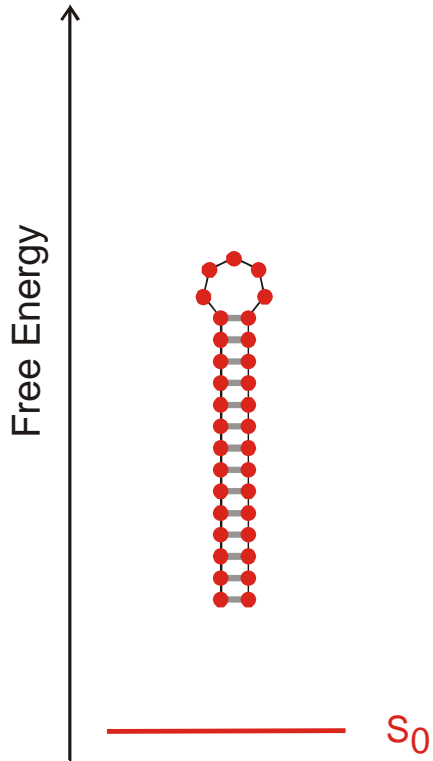
Arrhenius kinetic



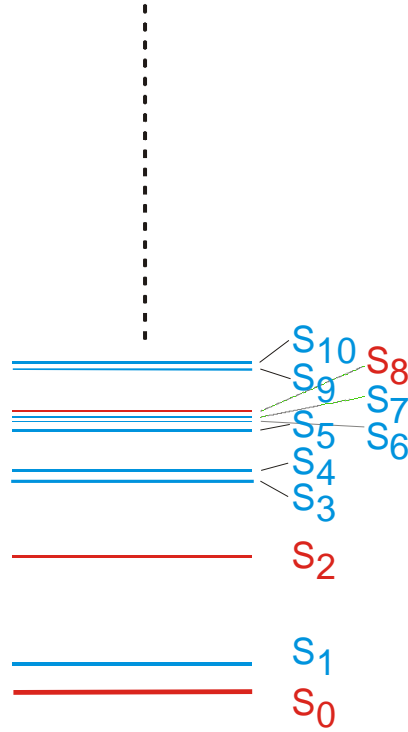
Exact solution of the master equation

M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm,
 I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A:*
Math.Gen. **37**:4731-4741.

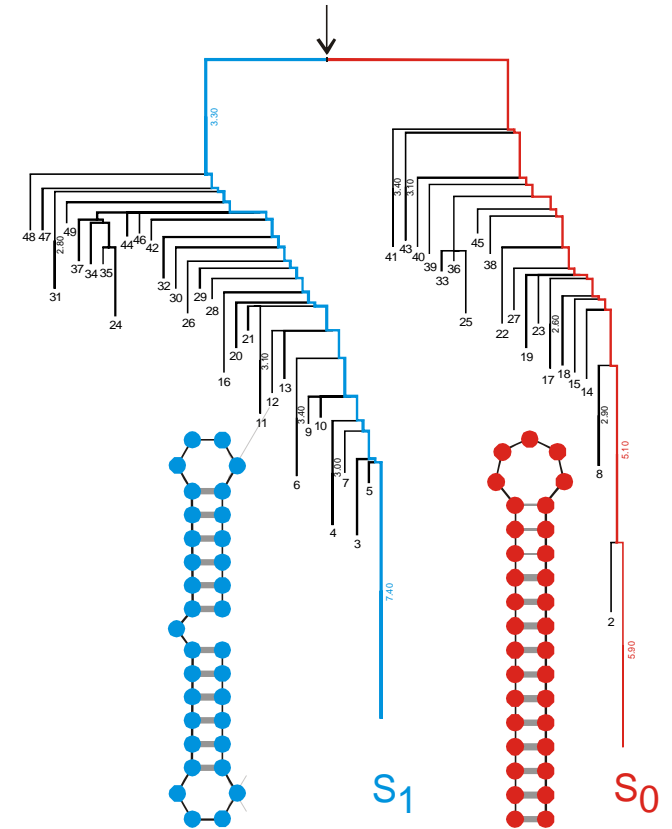
One sequence - one structure



Many suboptimal structures
Partition function



Metastable structures
Conformational switches

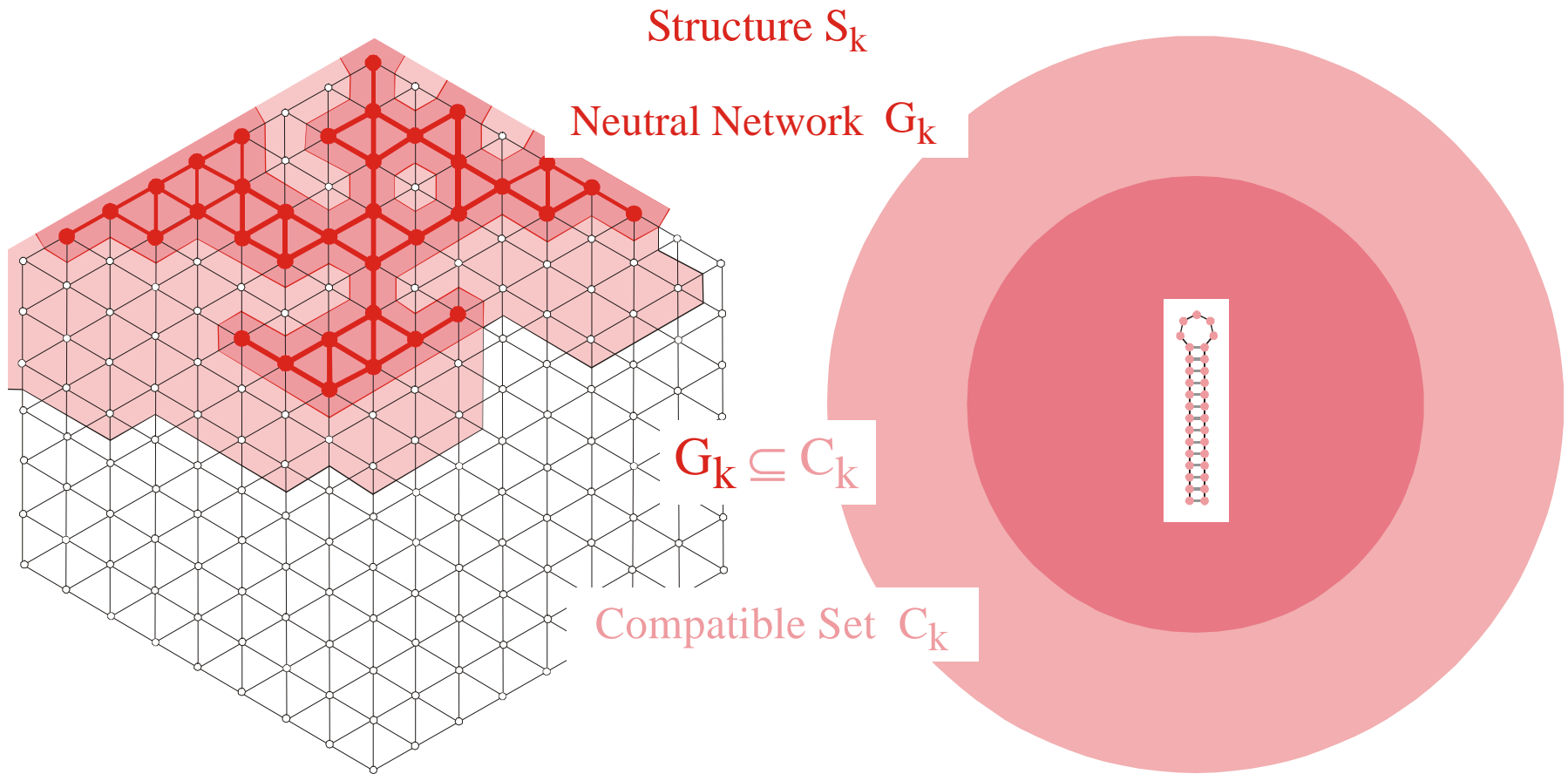


Minimum free energy structure

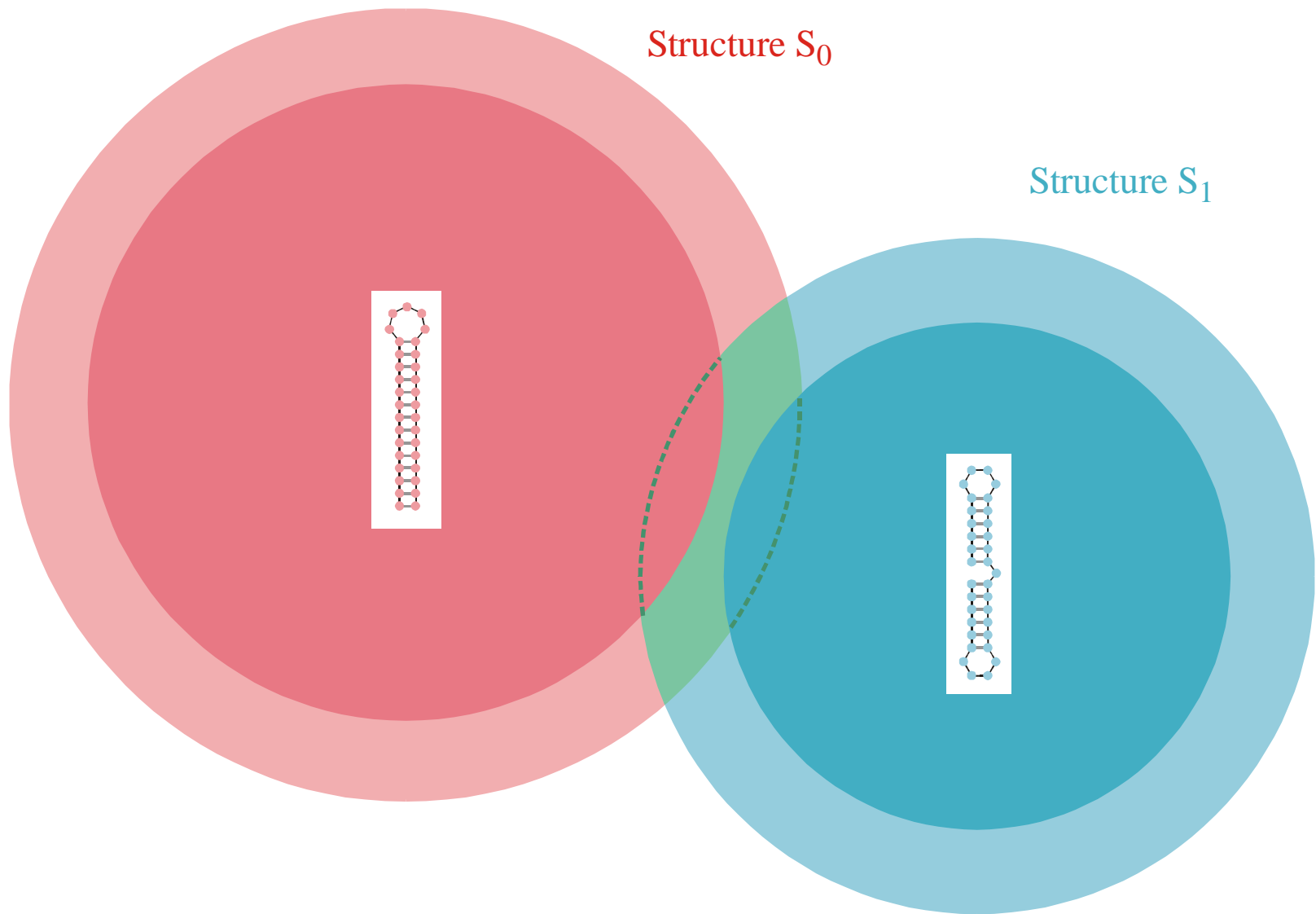
Suboptimal structures

Kinetic structures

RNA secondary structures derived from a single sequence



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



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

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \neq \emptyset$



S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES¹

■ CHRISTIAN REIDYS*, †, PETER F. STADLER*, ‡
 and PETER SCHUSTER*, ‡, §, ¶²

*Santa Fe Institute,
 Santa Fe, NM 87501, U.S.A.

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

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

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

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

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

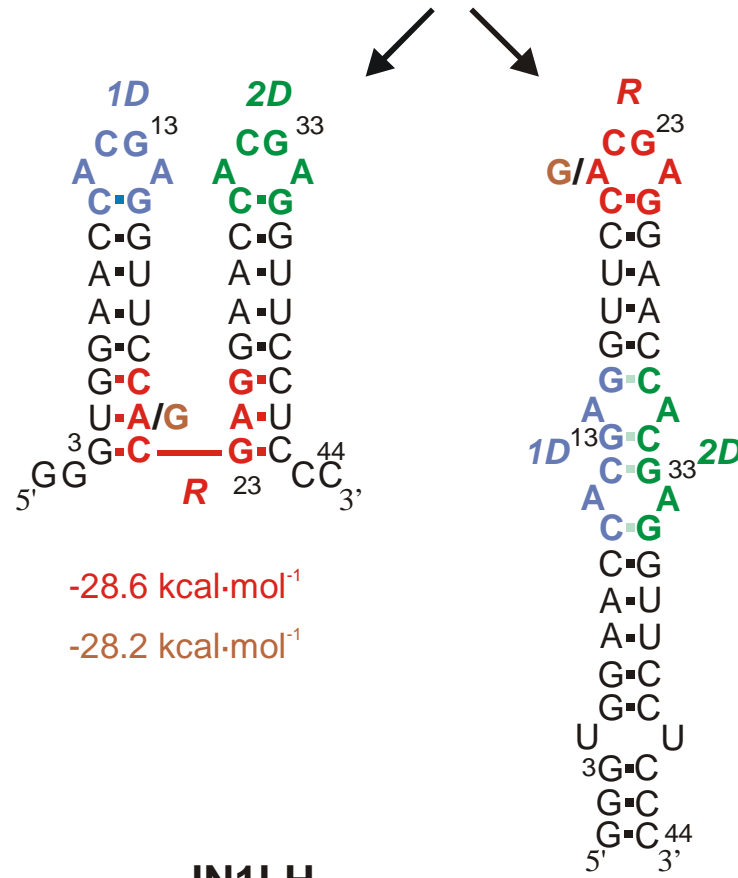
THEOREM 5. INTERSECTION-THEOREM. *Let s and s' be arbitrary secondary structures and $C[s], C[s']$ their corresponding compatible sequences. Then,*

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

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

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

Reference for the definition of the intersection and the proof of the [intersection theorem](#)



-28.6 kcal·mol⁻¹

-28.2 kcal·mol⁻¹

-28.6 kcal·mol⁻¹

-31.8 kcal·mol⁻¹

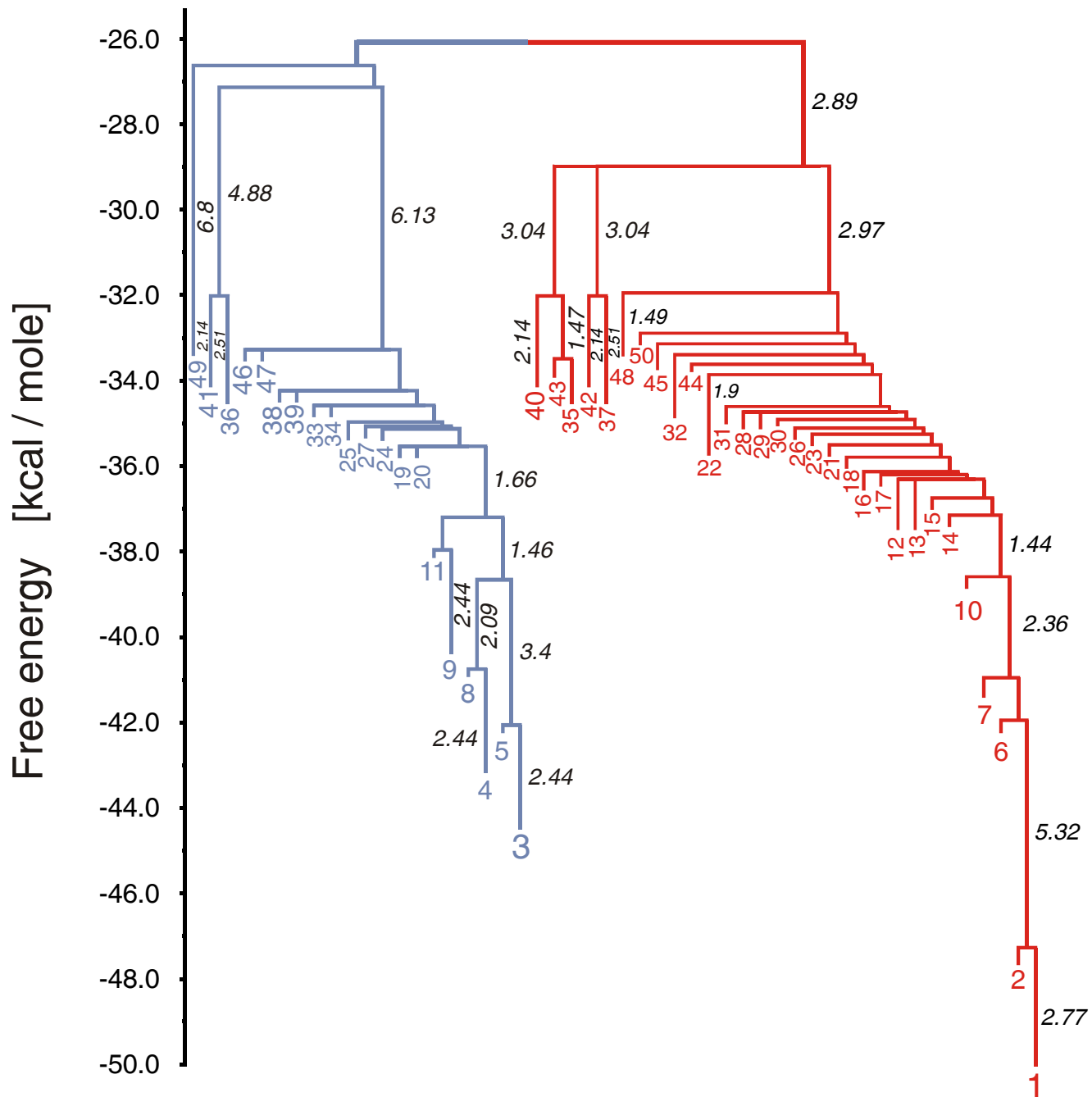
An RNA switch

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, Nucleic Acids Res., in press 2005.*

J1LH barrier tree



- 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 μ M) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μ M) 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 μ l 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/CH₂Cl₂ 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. Horadzovsky, *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 rbt1, 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 dis-

parate 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 non-functional 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).

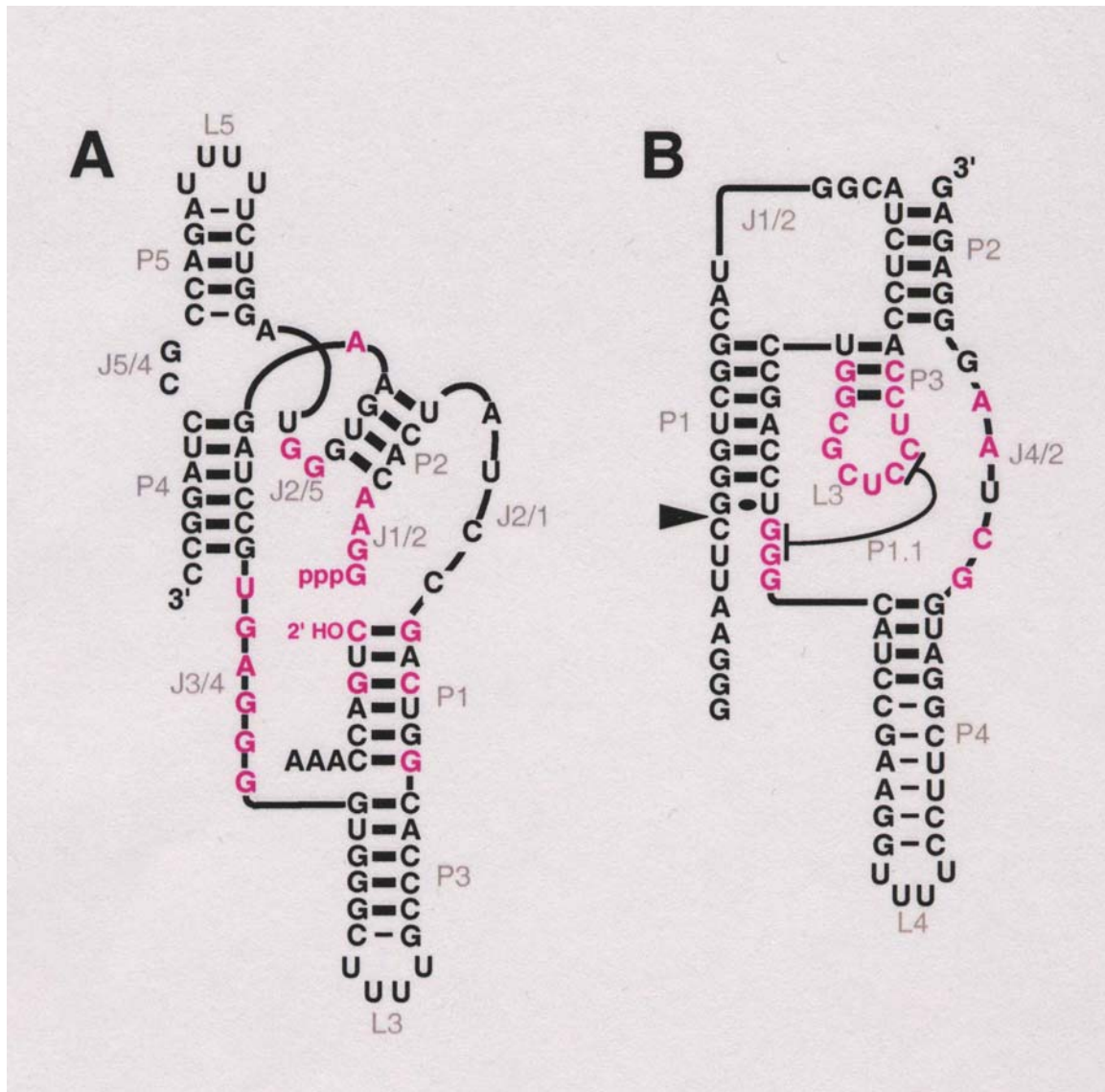
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

A ribozyme switch

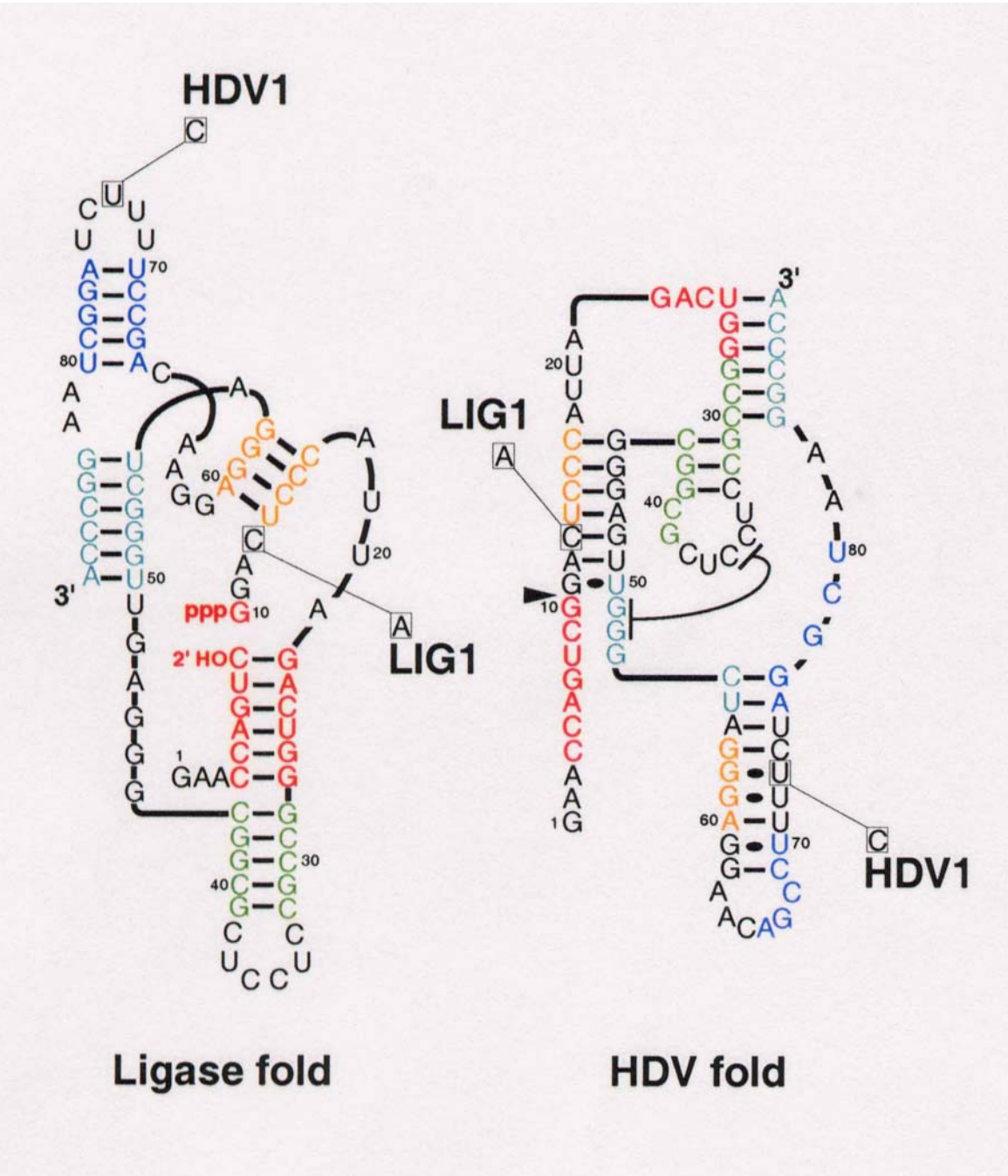
E.A.Schultes, D.B.Bartel, *Science*
289 (2000), 448-452

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

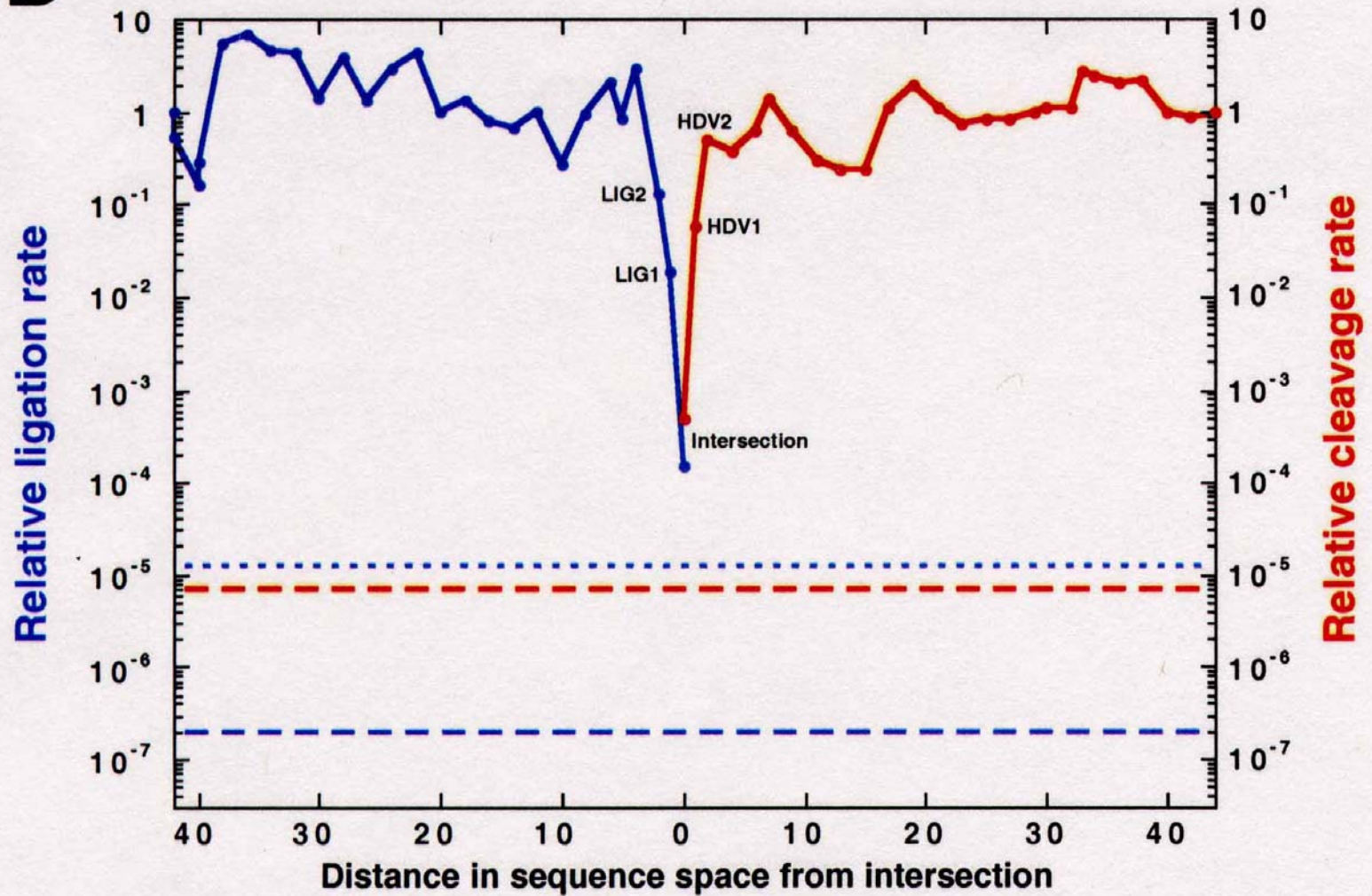


Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (A) and a natural cleavage ribozyme of hepatitis- δ -virus (B)



The sequence at the *intersection*:

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

B

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

1. RNA sequences and structures
2. Neutral networks
3. Evolutionary optimization of structure
4. Suboptimal structures and kinetic folding
5. **Comparison of kinetic folding and evolution**

Kinetic Folding

Compatible structures:

Set of **structures compatible with**
a given **sequence**

stability restriction



Conformation space

Folding trajectory in conformation space:

Time ordered series of structures

Folding process:

Average of trajectories on the
ensemble level

Criterion: minimizing free energy

Evolutionary optimization

Compatible sequences:

Set of **sequences compatible with**
a given **structure**

mfe restriction



Neutral network

Genealogy on a neutral network:

Time ordered series of sequences

Optimization process:

Average over genealogies on the
population level

Criterion: maximizing fitness

Acknowledgement of support

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

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

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

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

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

Österreichische Akademie der Wissenschaften

Berlin-Brandenburgische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute



Universität Wien

Coworkers

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

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

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

Jord Nagel, Kees Pleij, Universiteit Leiden, NL

Walter Fontana, Harvard Medical School, MA

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

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

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

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

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



Universität Wien

Web-Page for further information:

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

