

RNA – A Magic Molecule at the Interplay of Structural Biology, Bioinformatics, and Evolution

Peter Schuster

Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien



Genomics, Proteomics, and Bioinformatics for Medicine

Moscow – Ples – Moscow, 14.– 19.07.2004

Web-Page for further information:

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

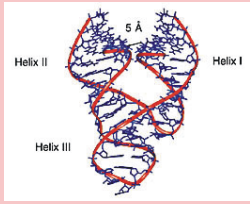
1. From working copy in molecular genetics to riboswitches
2. The never-ending story of RNA structure
3. Neutral networks, intersections, and bistable RNA molecules
4. Evolution of molecules *in silico*
5. A few selected experiments

"Observation and theory get on best when they are mixed together, both helping one another in the pursuit of truth. It is a good rule not to put overmuch confidence in a theory until it has been confirmed by observation. I hope I shall not shock the experimental physicists too much if I add that **it is also a good rule not to put overmuch confidence in the observational results that are put forward until they have been confirmed by theory.**"

Sir Arthur Eddington, *New Pathways in Science*, 1935.

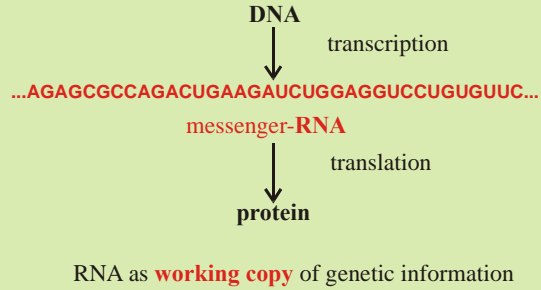
- 1. From working copy in molecular genetics to riboswitches**
2. The never-ending story of RNA structure
3. Neutral networks, intersections, and bistable RNA molecules
4. Evolution of molecules *in silico*
5. A few selected experiments

RNA as catalyst

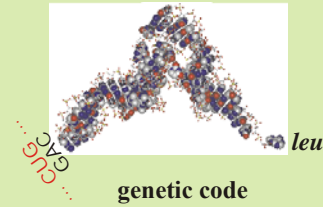


Ribozyme

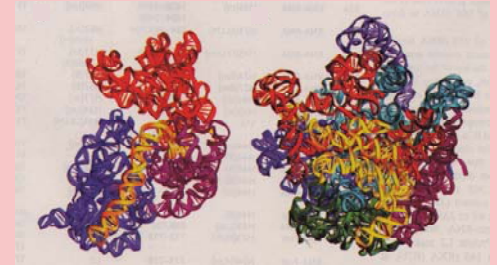
RNA as transmitter of genetic information



RNA as adapter molecule



RNA is the catalytic subunit in supramolecular complexes



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

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

RNA

RNA is modified by epigenetic control

RNA editing

Alternative splicing of messenger RNA

RNA as carrier of genetic information

RNA viruses and retroviruses

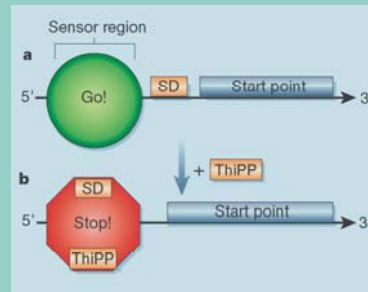
RNA evolution *in vitro*

Evolutionary biotechnology

RNA aptamers, artificial ribozymes, allosteric ribozymes

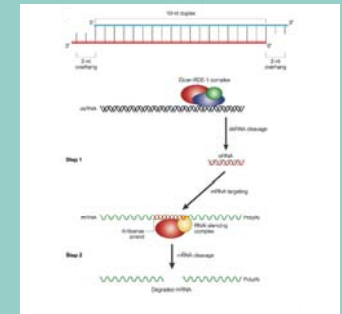
Functions of RNA molecules

Allosteric control of transcribed RNA

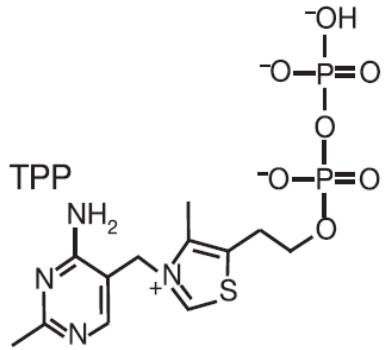


Riboswitches controlling transcription and translation through **metabolites**

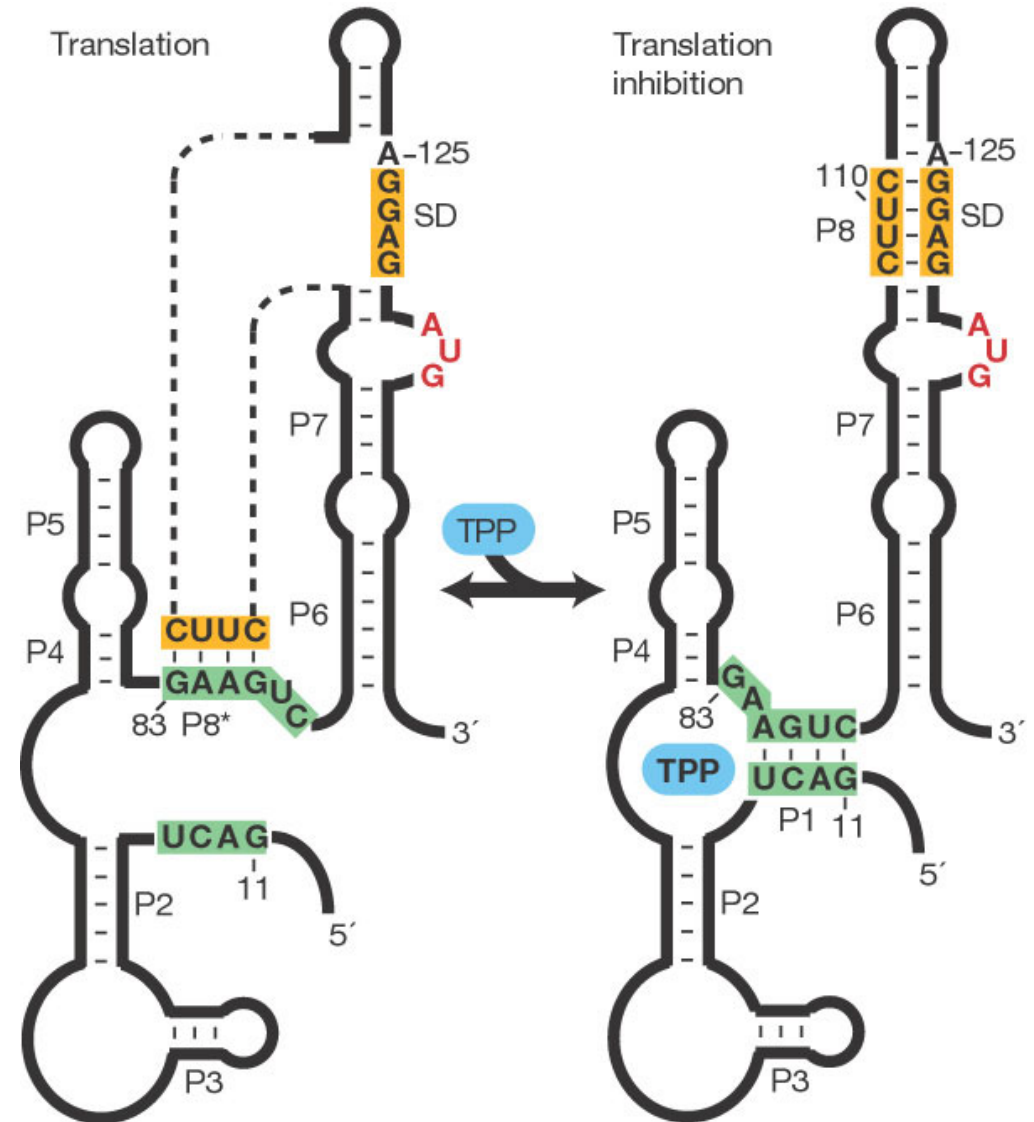
RNA as regulator of gene expression



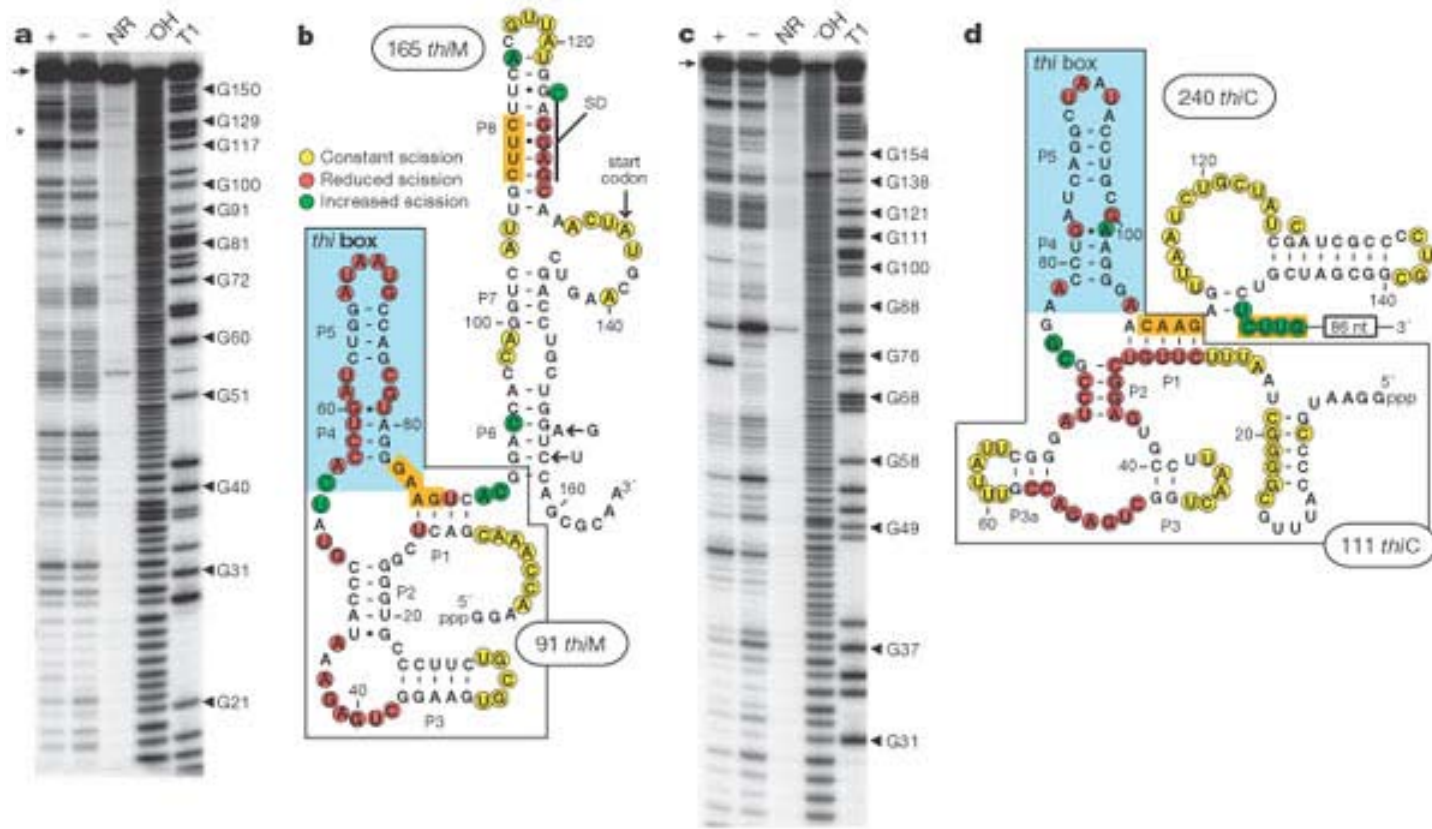
Gene silencing by small interfering RNAs



Thiamine-pyrophosphate



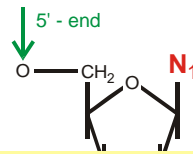
Wade Winkler, Ali Nahvi, and Ronald R. Breaker, *Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression*. Nature **419**, 952-956, 2002.



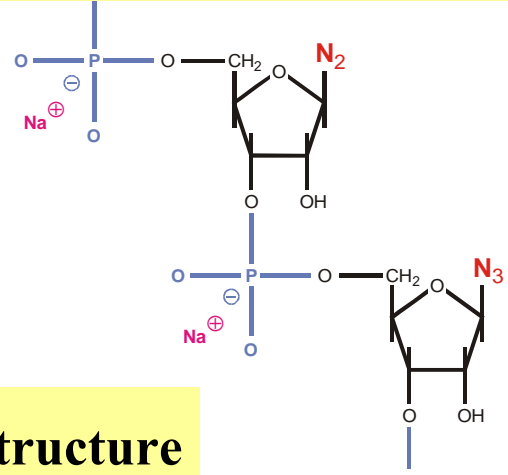
Wade Winkler, Ali Nahvi, and Ronald R. Breaker, *Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression*. Nature **419**, 952-956, 2002.

Even in 2004 RNA is still good for surprises!

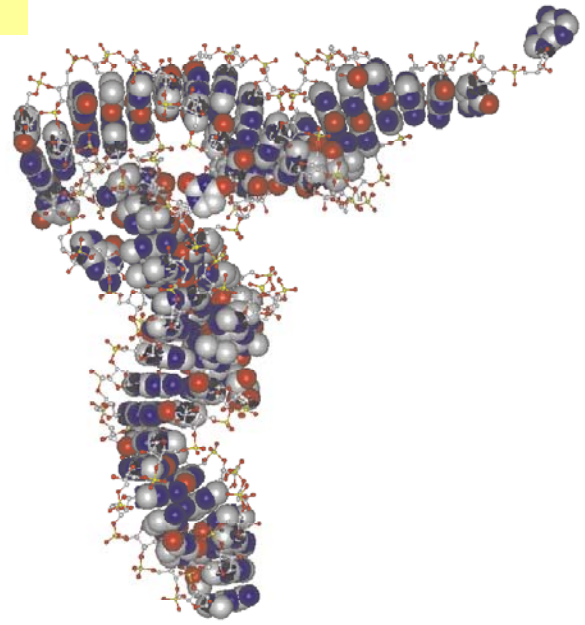
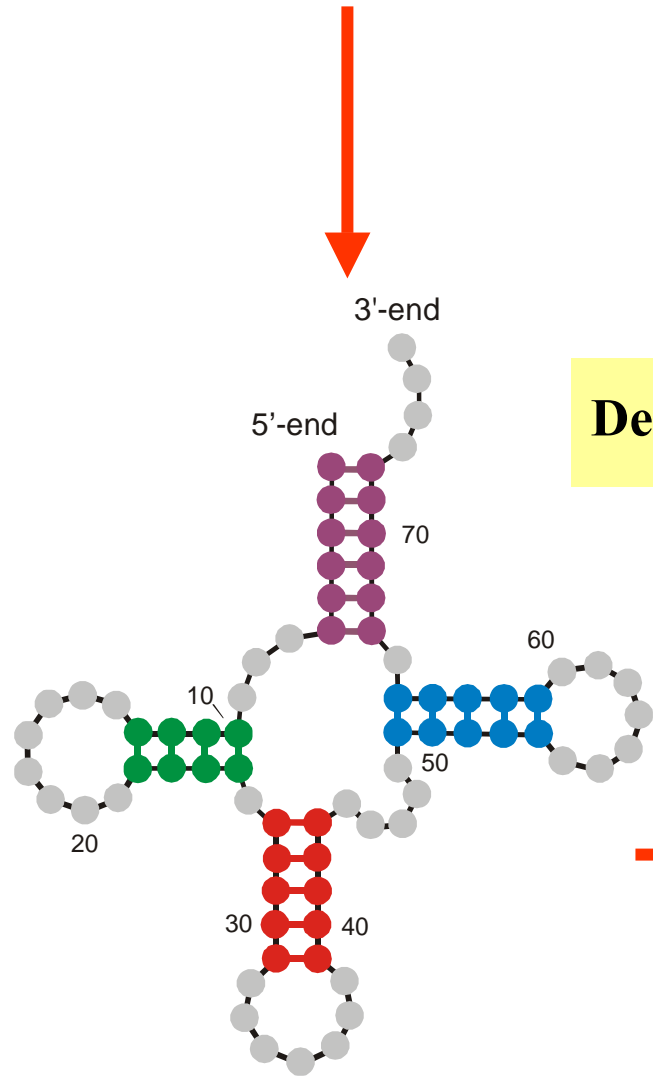
1. From working copy in molecular genetics to riboswitches
- 2. The never-ending story of RNA structure**
3. Neutral networks, intersections, and bistable RNA molecules
4. Evolution of molecules *in silico*
5. A few selected experiments

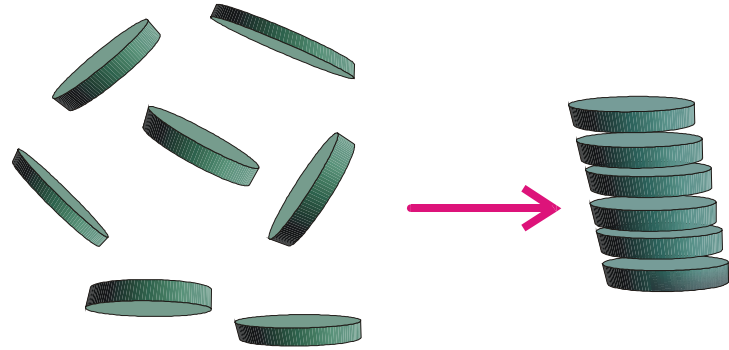


5'-end **GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCUGUGUUCGAUCCACAGAAUUCGCACCA** 3'-end

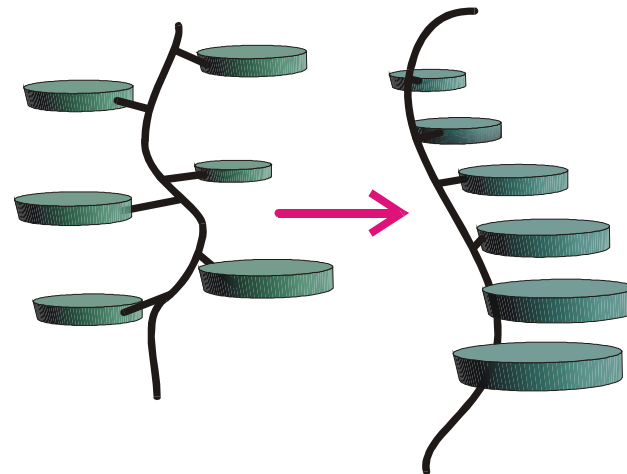


Definition of RNA structure



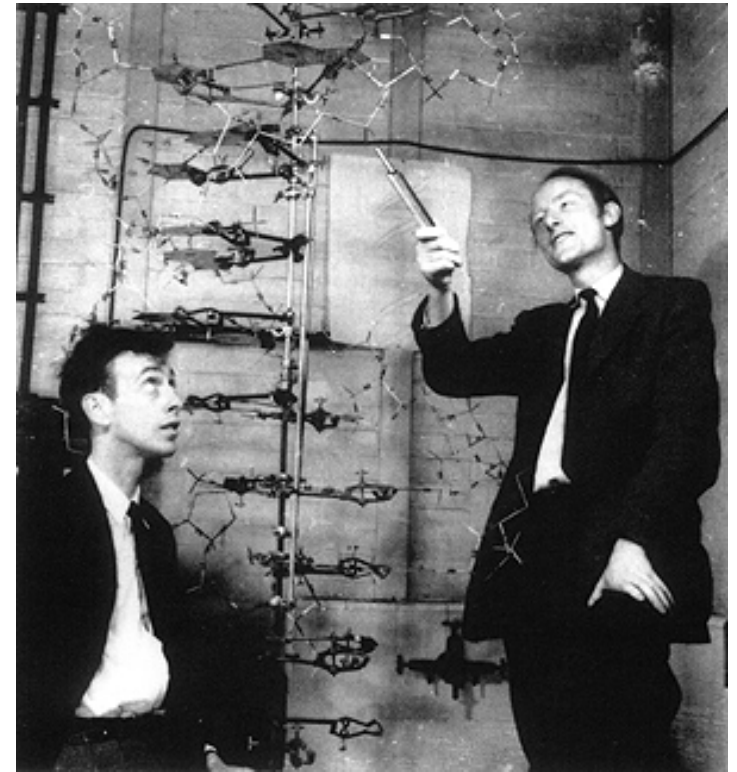
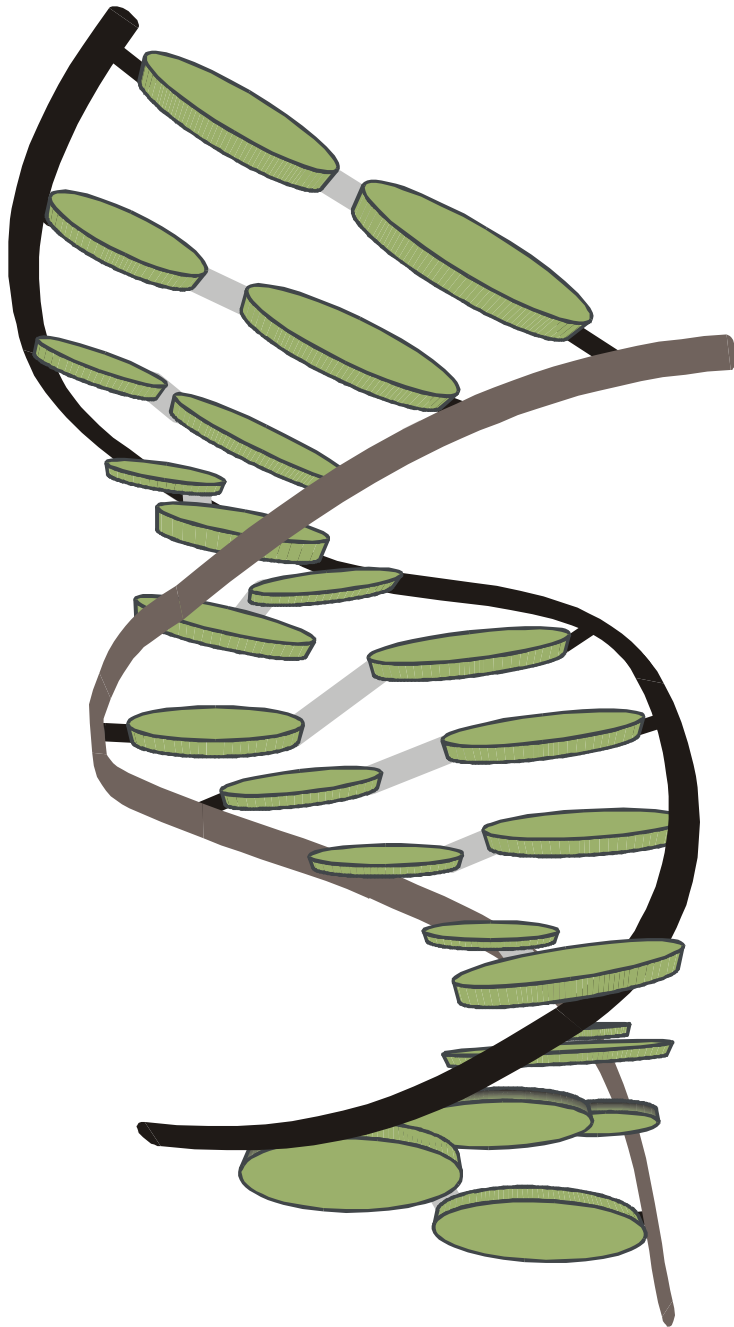


Stacking of free nucleobases or other planar heterocyclic compounds (N6,N9-dimethyl-adenine)



The stacking interaction as driving force of structure formation in nucleic acids

Stacking of nucleic acid single strands (poly-A)



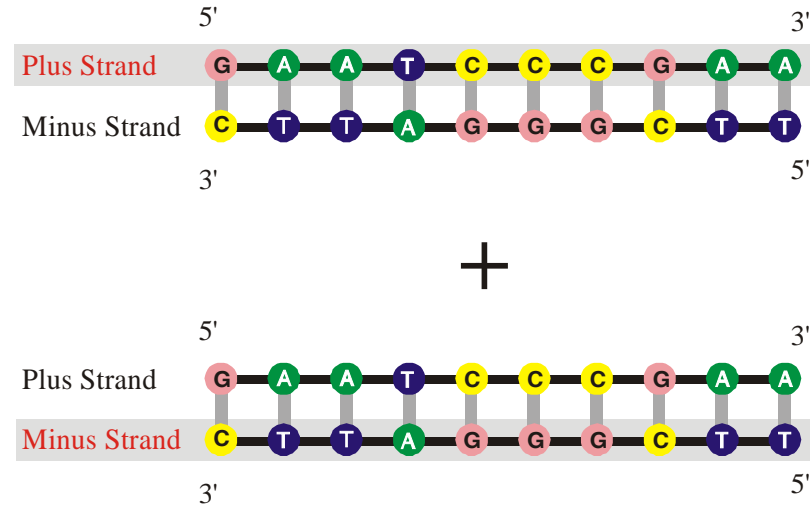
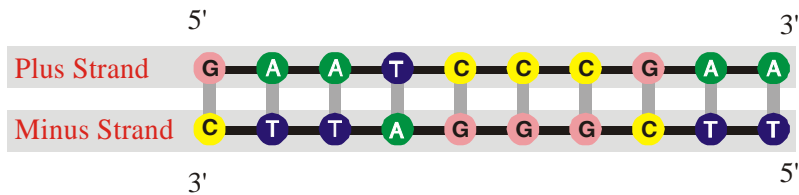
James D. Watson and Francis H.C. Crick

Nobel prize 1962

1953 – 2003 fifty years double helix

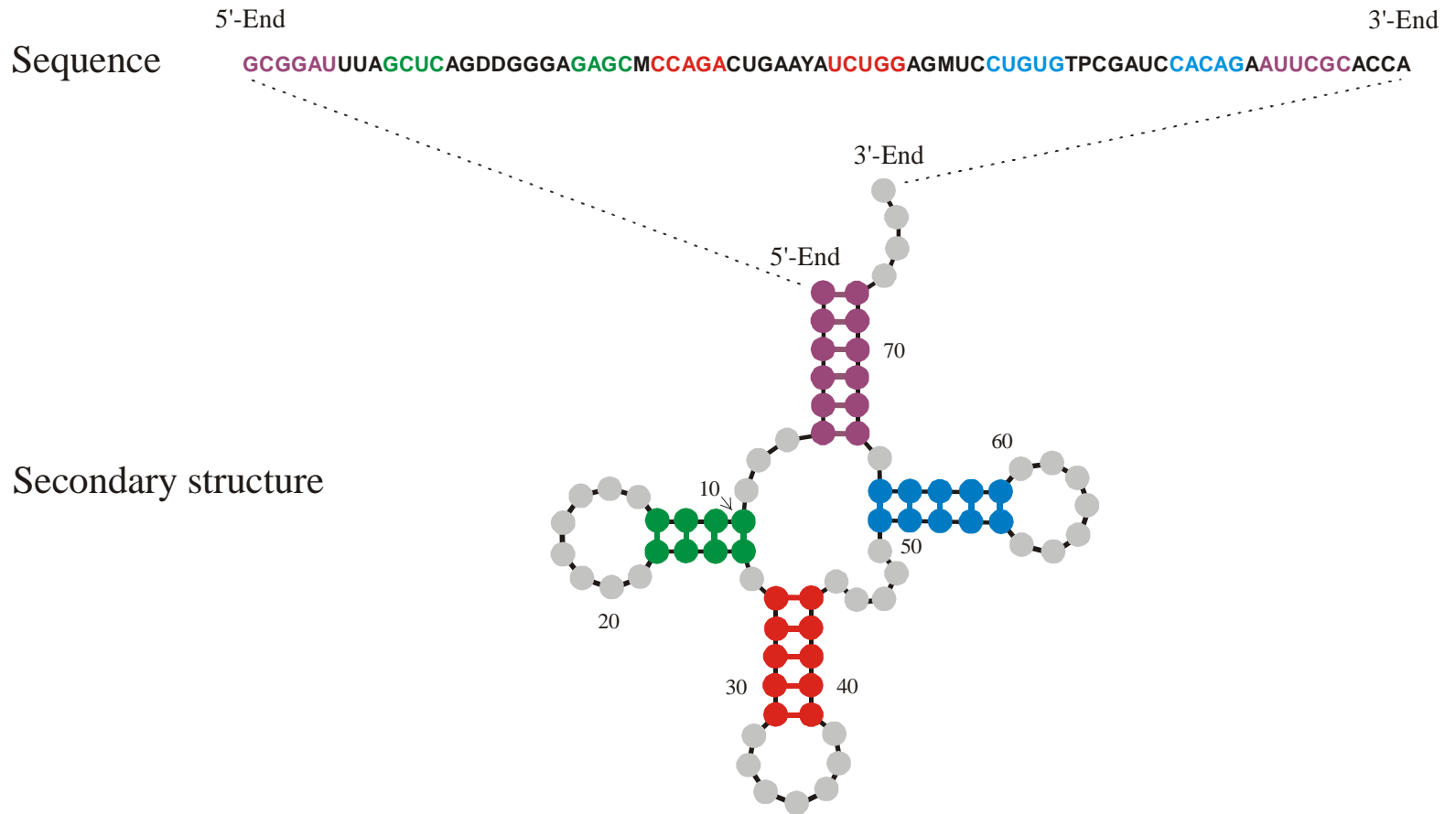
Base pairs: **A** = **T** and **G** ≡ **C**

Stacking of base pairs in nucleic acid double helices (B-DNA)



Direct replication of DNA is a highly complex copying mechanism involving more than ten different protein molecules. Complementarity is determined by Watson-Crick base pairs:





RNA folding is initiated and guided by double helix formation within one single strand

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 pseudoknots.

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.“

RNA sequence

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

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

Biophysical chemistry:
thermodynamics and
kinetics

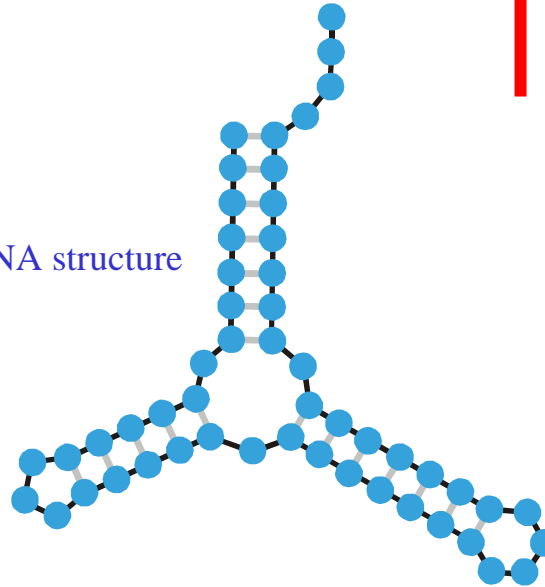


Empirical parameters

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

Sequence, structure, and function

RNA structure



Secondary structure
computation



GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

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 sequences, alignment, 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)

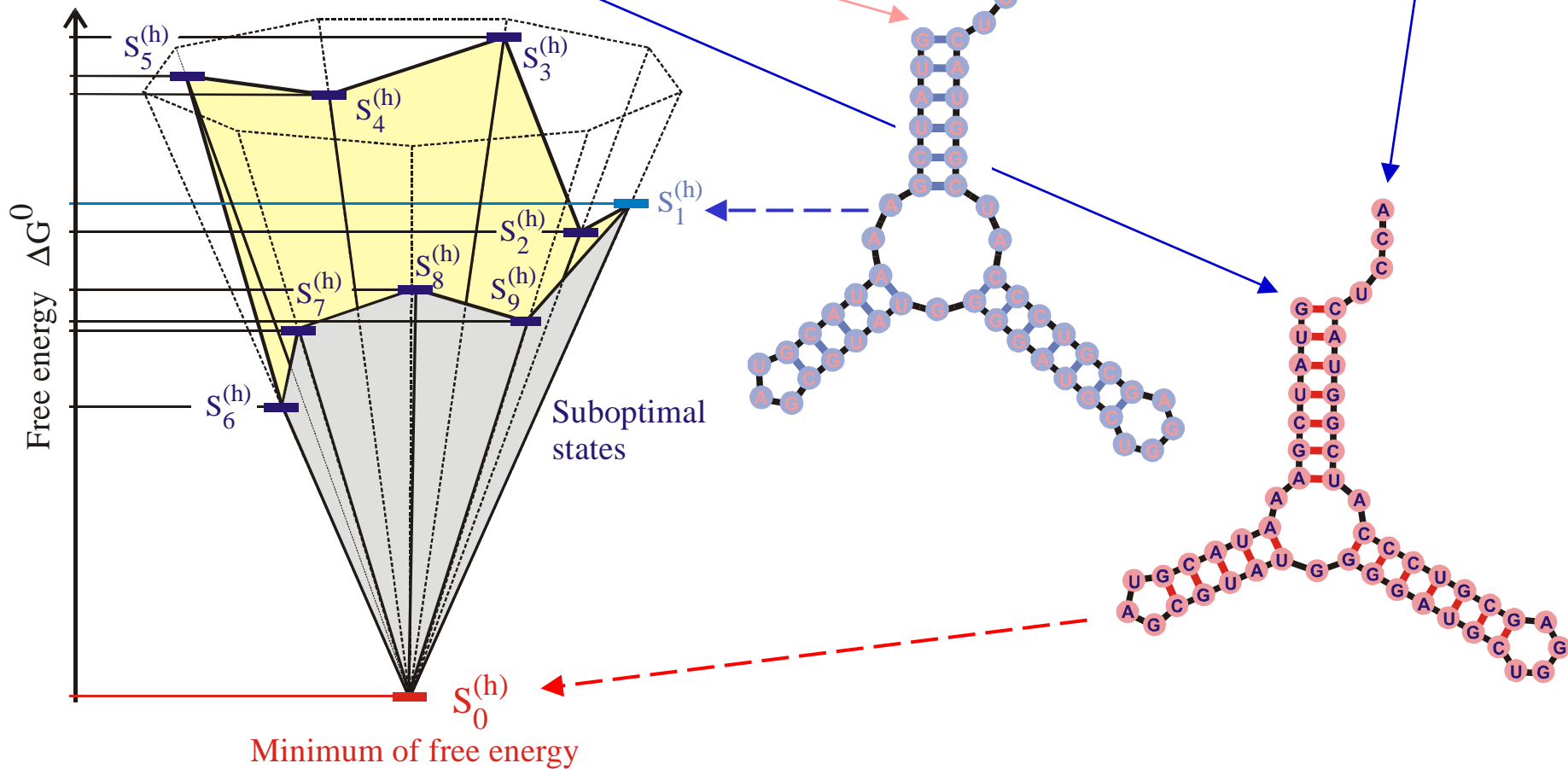
C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. *RNA* **6**:325-338 (1999)

Vienna RNA Package: <http://www.tbi.univie.ac.at>

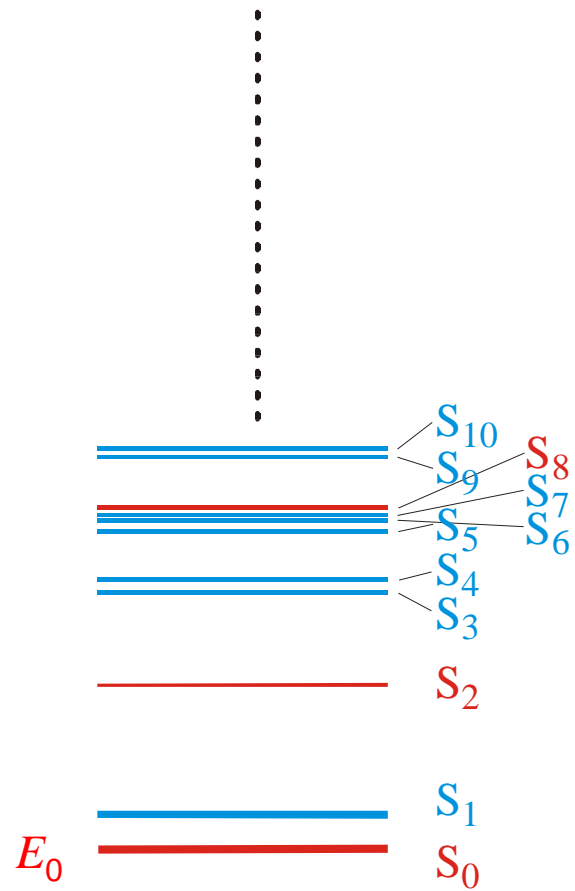
5'-end

3'-end

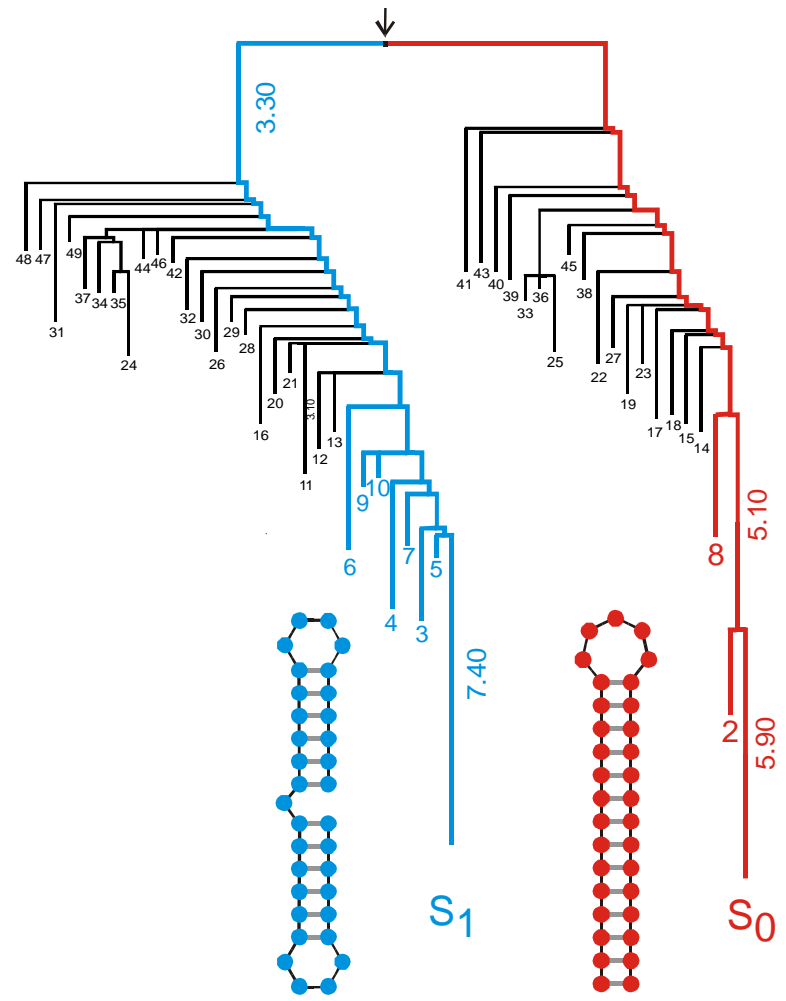
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGAGCGUCCCAUCGGUACUCCA



The minimum free energy and suboptimal structures on a discrete space of conformations



minimum free energy E_0 and energies of suboptimal structures



metastable structure

stable structure

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**:254-265, 2001

Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, Michael T. Wolfinger. *Barrier trees of degenerate landscapes*. Z. Phys. Chem. **216**:155-173, 2002

Michael T. Wolfinger, W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, Martin Zehl. *Exact folding dynamics of RNA secondary structures*. J. Phys. A: Math. Gen. **37**: 4731-4741, 2004

The Folding Algorithm

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

$$\mathfrak{S}(\mathbf{I}) = \{S_0, S_1, \dots, S_m, \mathbf{O}\}$$

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 S_0 or a metastable structure S_k which represents a local energy minimum:

$$\mathfrak{Z}_0(\mathbf{I}) = \{\mathbf{O}, S(1), \dots, S(t-1), S(t), \\ S(t+1), \dots, S_0\}$$

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

Transition probabilities $P_{ij}(t) = \text{Prob}\{S_i \rightarrow S_j\}$ are defined by

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

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

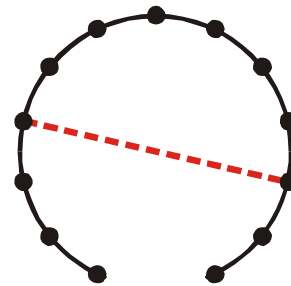
$$\Sigma_k = \sum_{k=1, k \neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

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

Formulation of kinetic RNA folding as a stochastic process



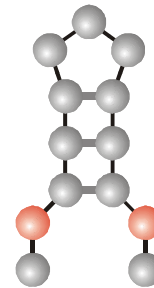
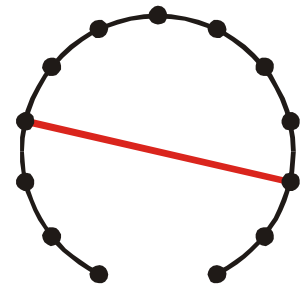
Nucleation



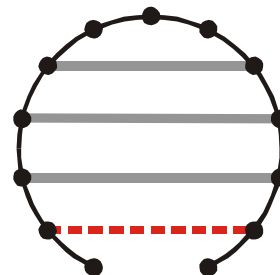
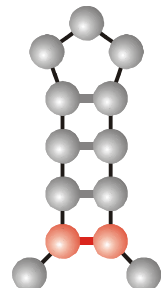
Base pair formation



Base pair cleavage



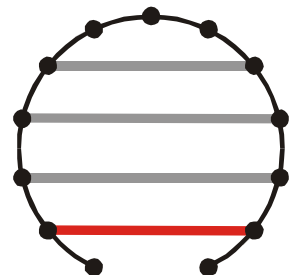
Elongation



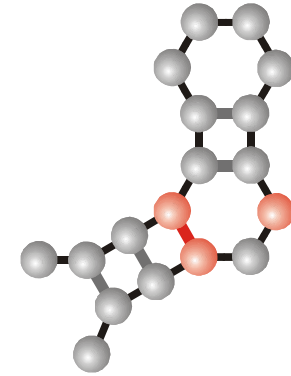
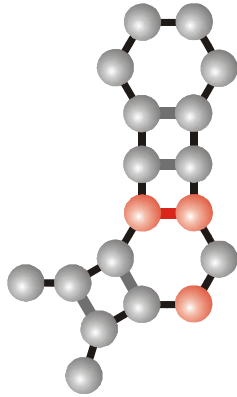
Base pair formation



Base pair cleavage

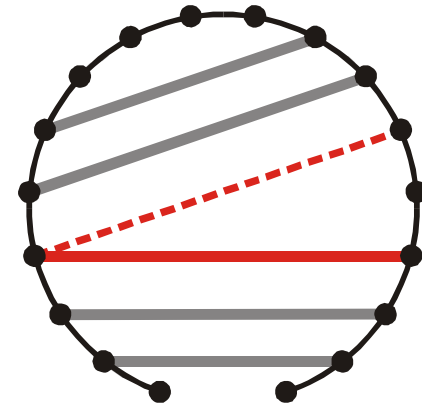
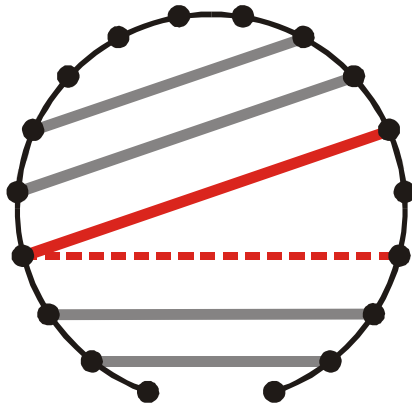


Base pair formation and base pair cleavage moves for **nucleation** and **elongation** of stacks

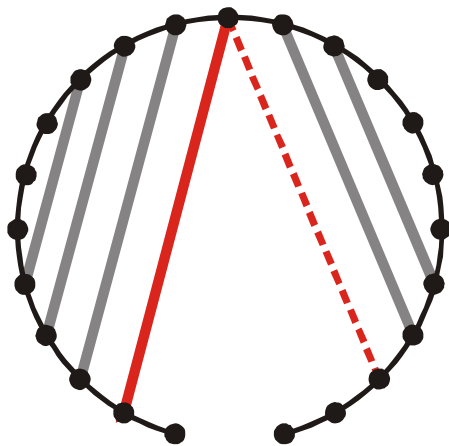
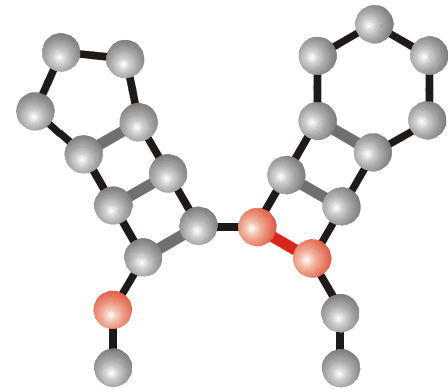
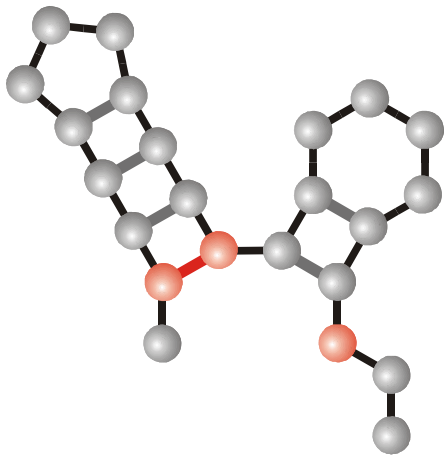


Base pair shift

Class 1

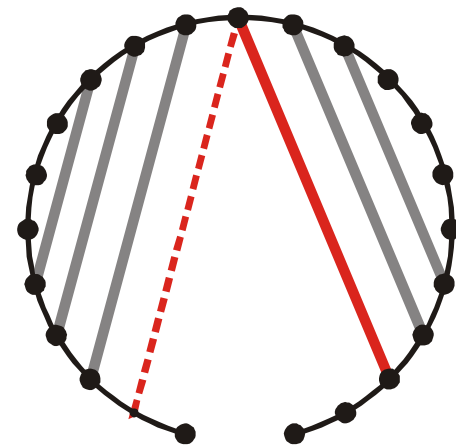


Base pair shift move of class 1: Shift inside internal loops or bulges

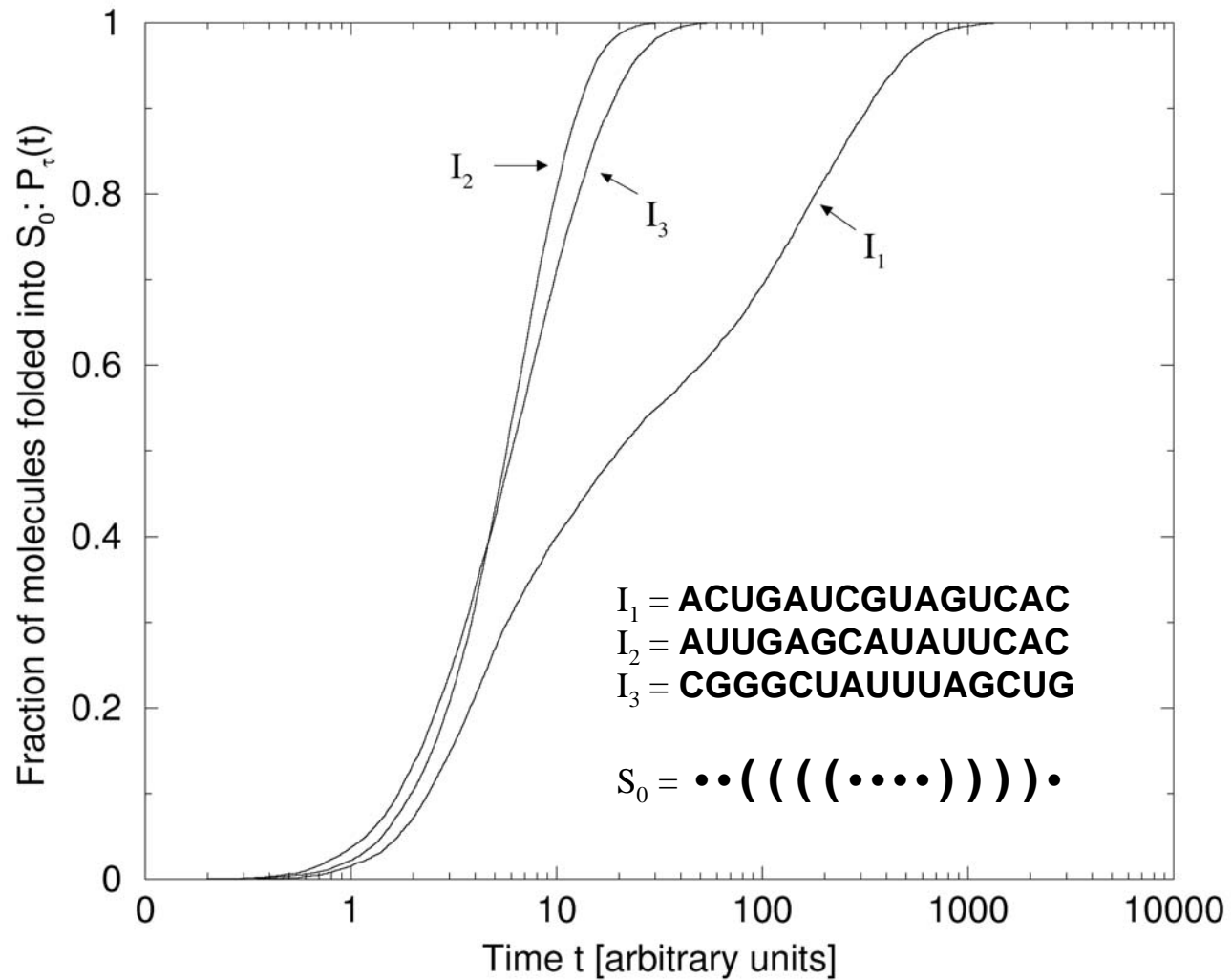


Base pair shift

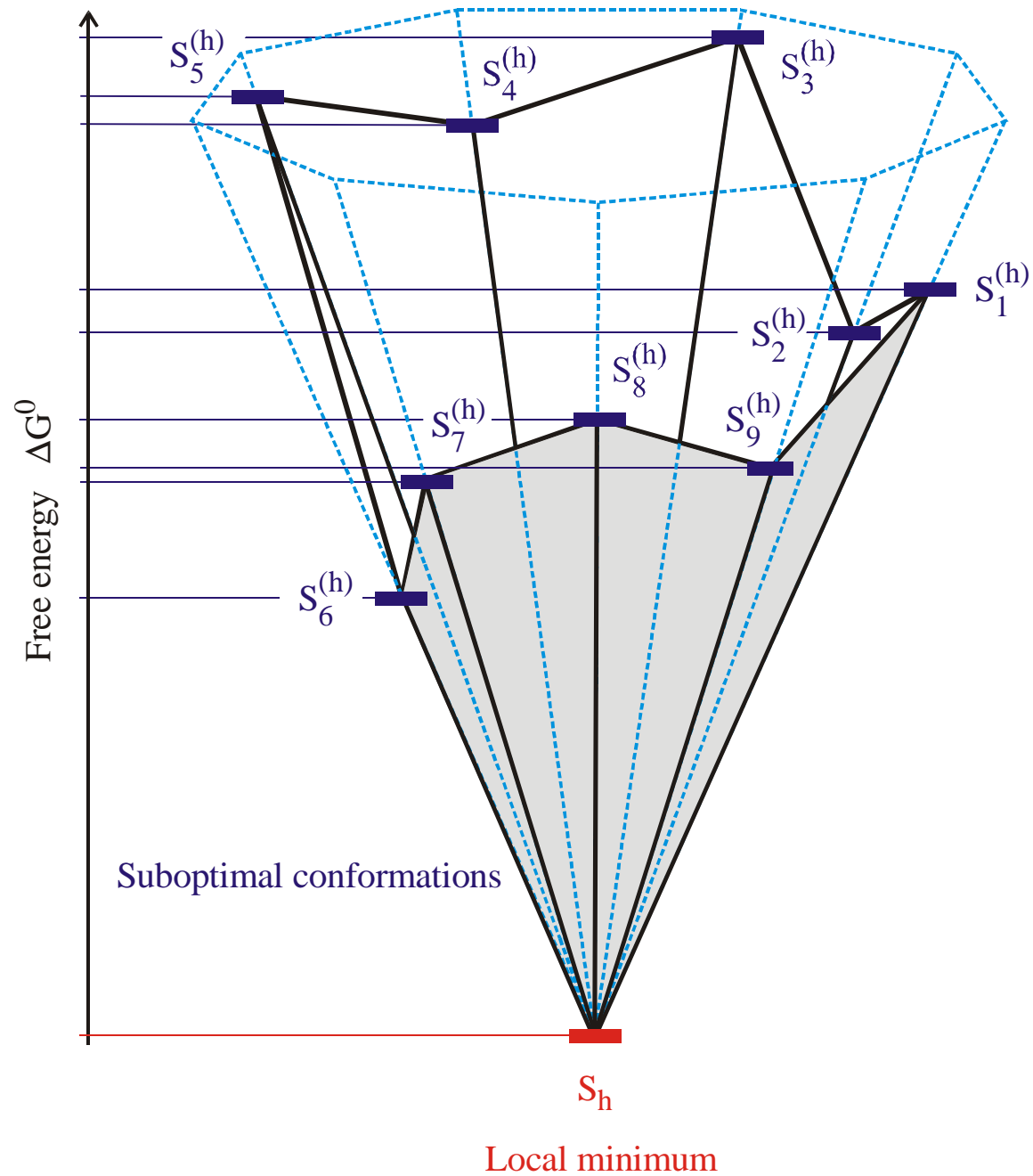
Class 2



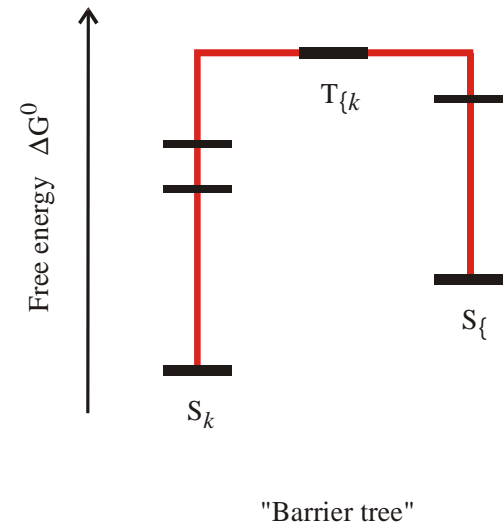
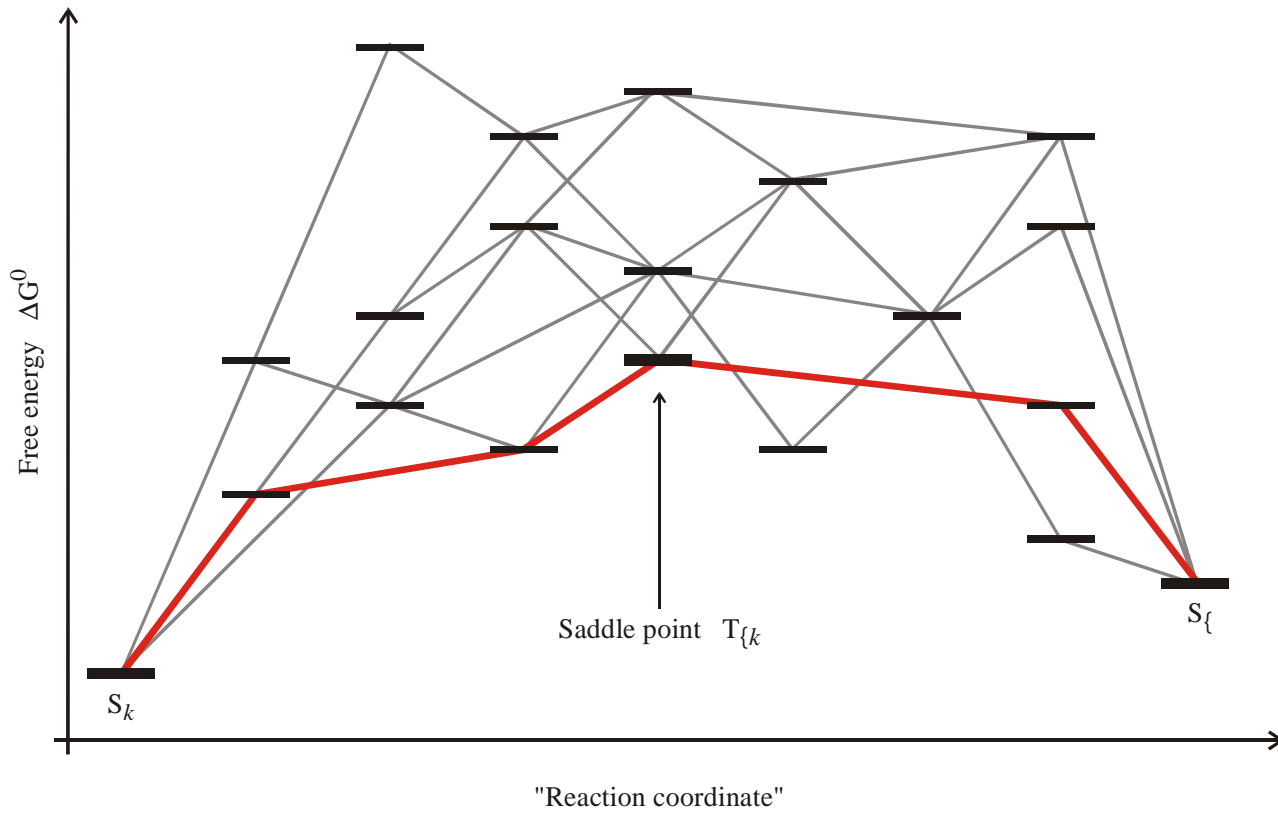
Base pair shift move of class 2: Shift involving free ends



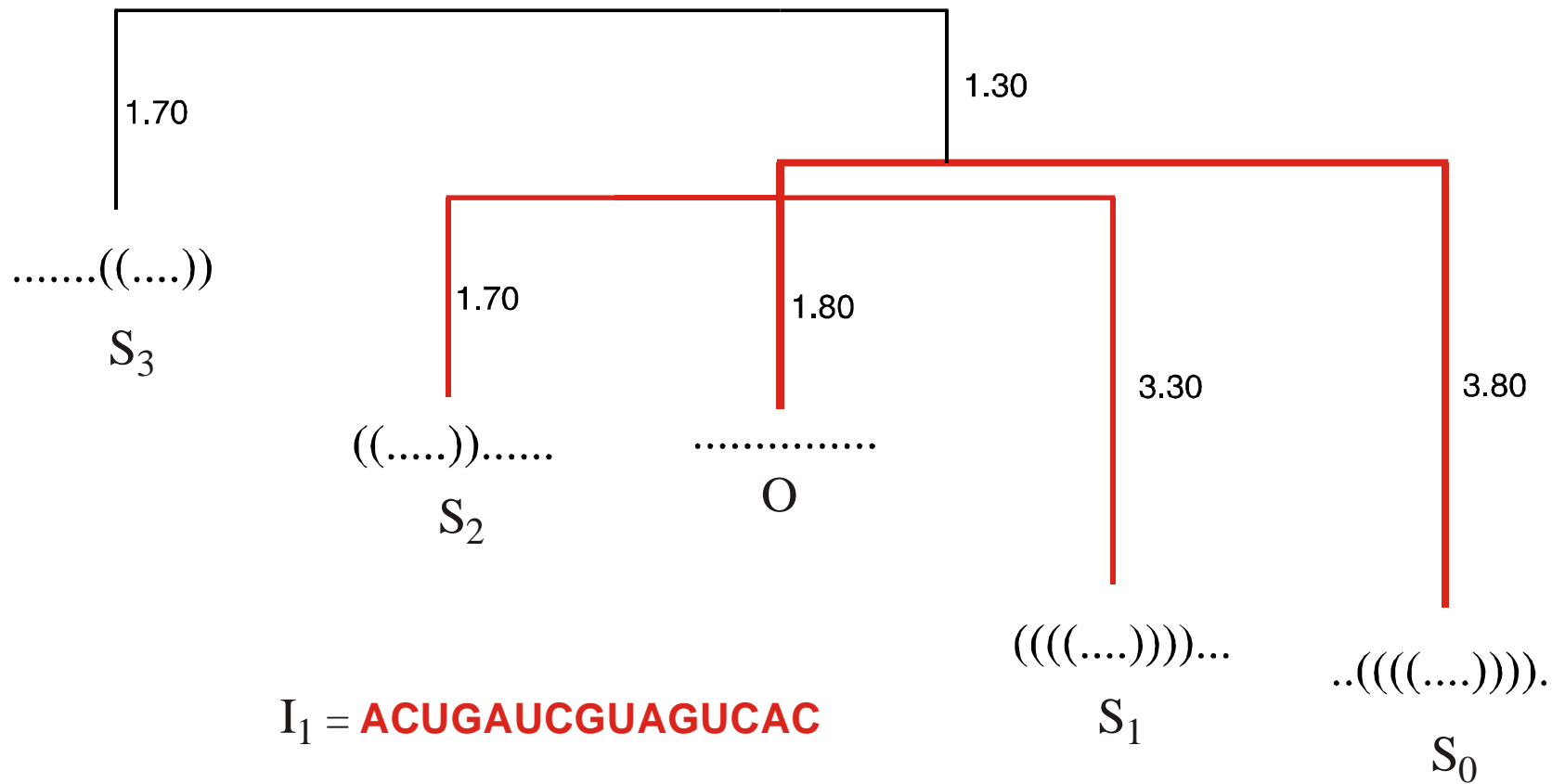
Mean folding curves for three small RNA molecules with different folding behavior



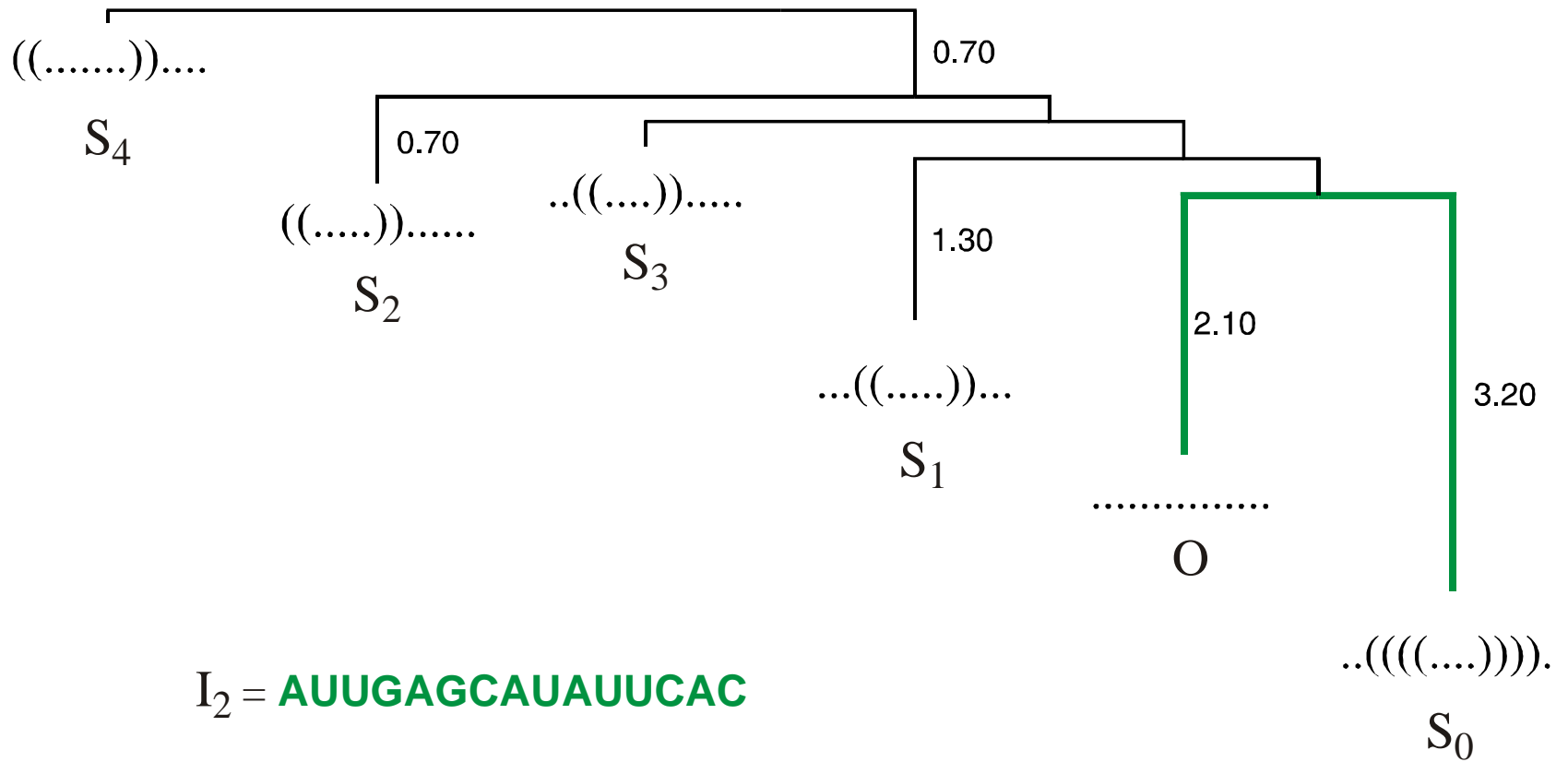
Search for local minima in conformation space



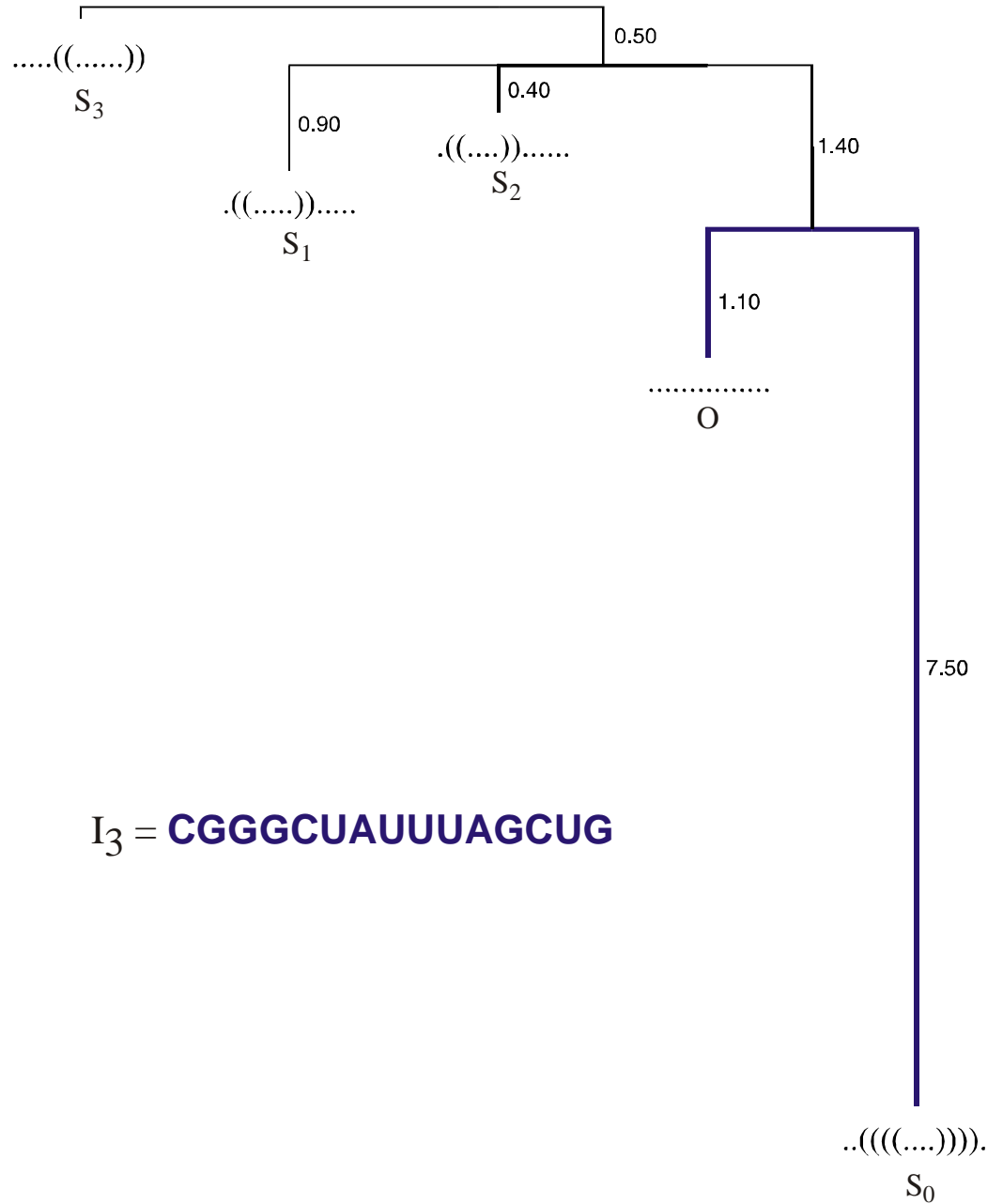
Definition of a ,barrier tree‘



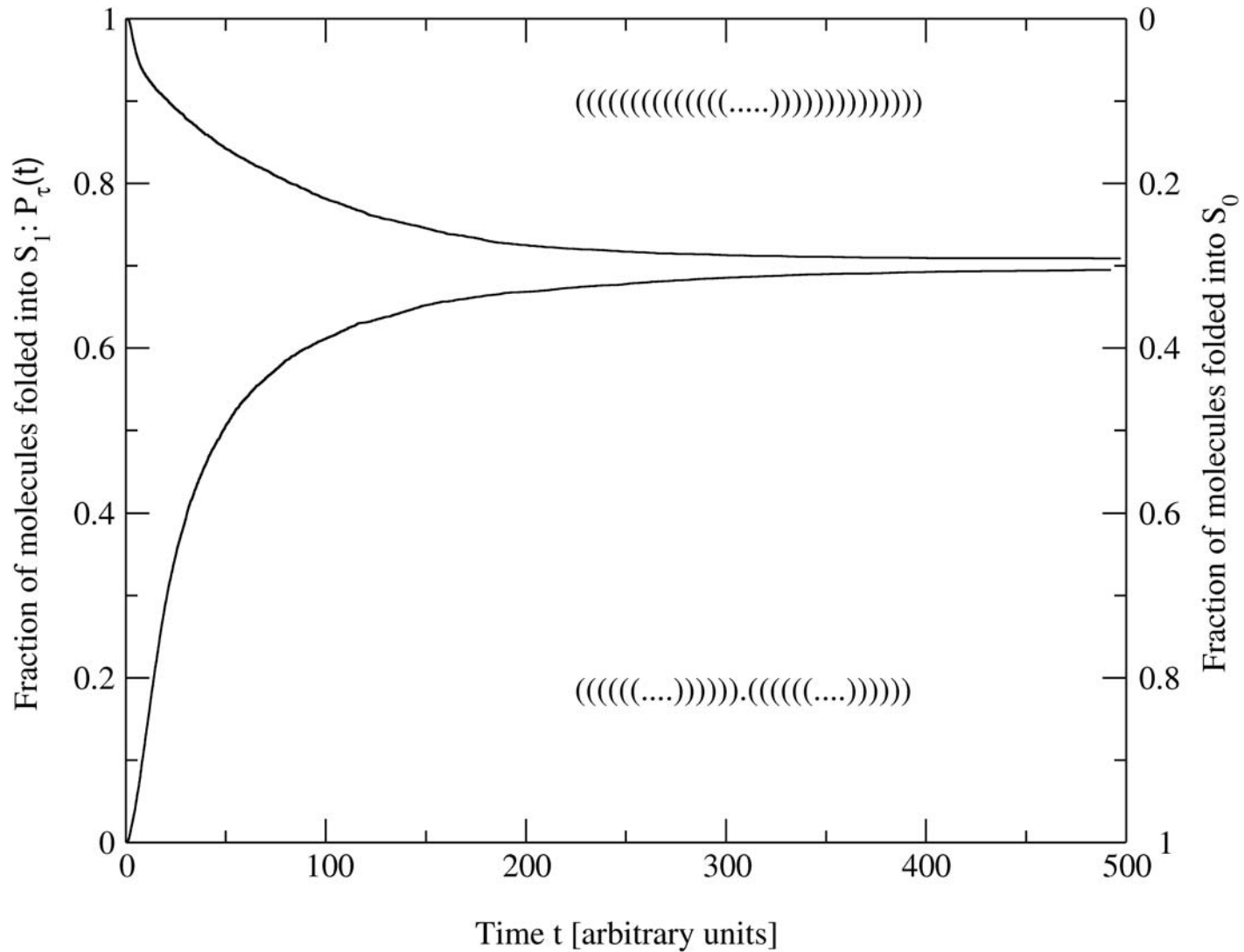
Example of an unefficiently folding small RNA molecule with $n = 15$



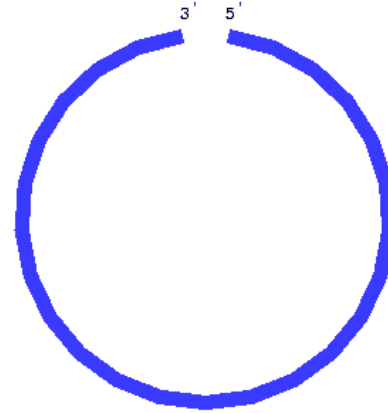
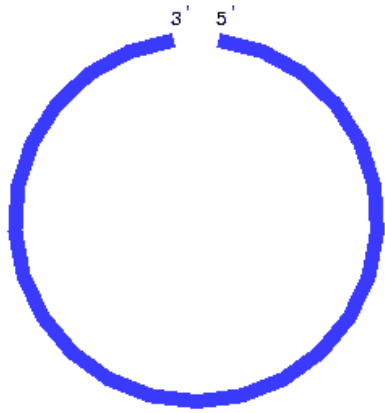
Example of an easily folding small RNA molecule with $n = 15$



Example of an easily folding
and especially stable small
RNA molecule with $n = 15$

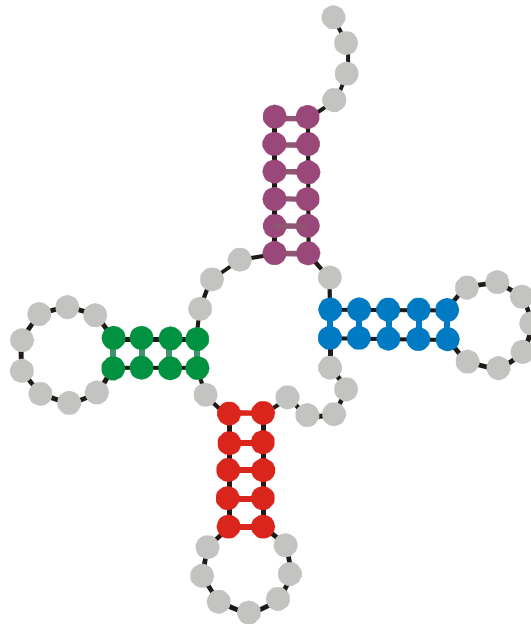


Folding dynamics of the sequence **GGCCCUUUGGGGCCAGACCCUAAAAGGGUC**

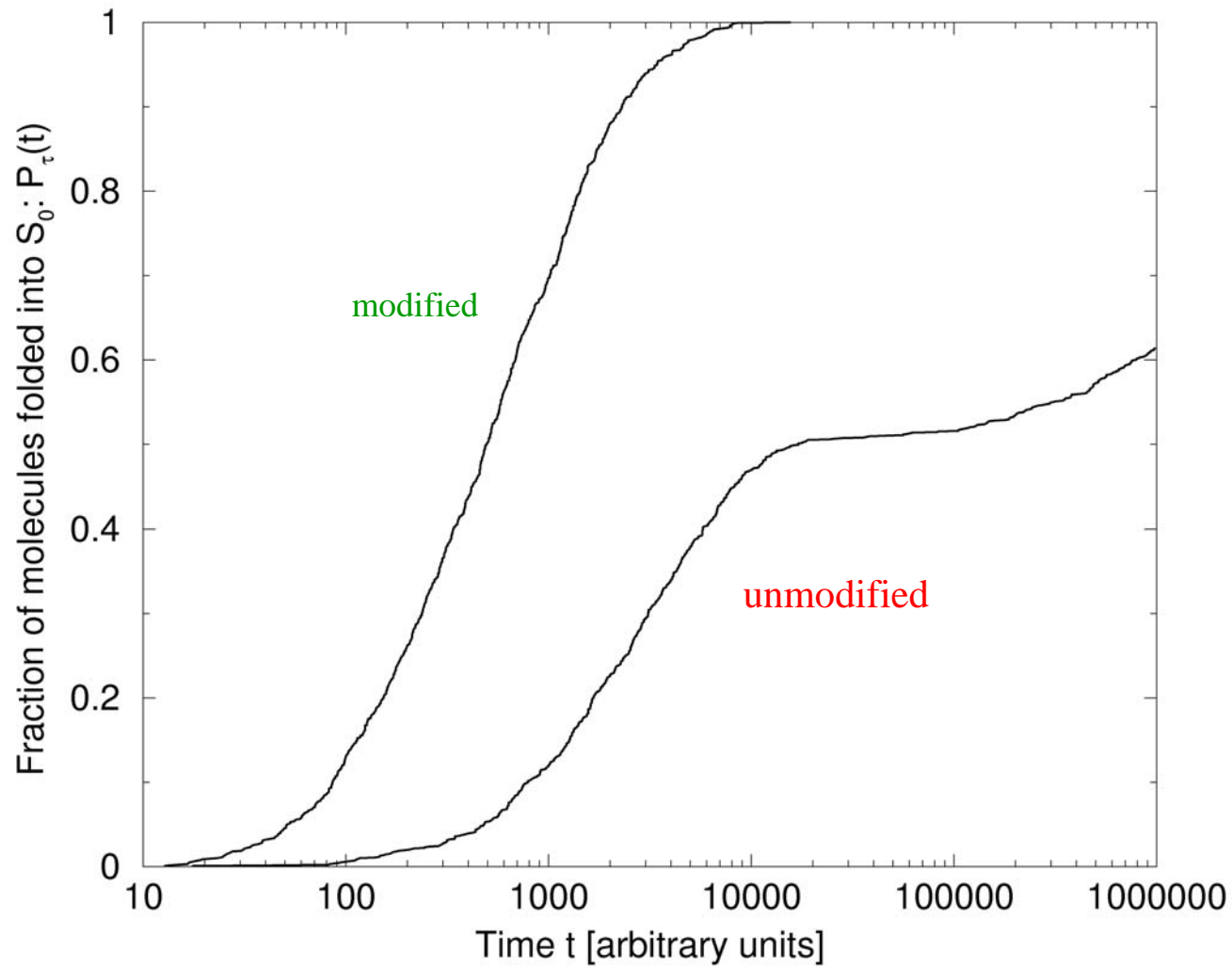


Examples of two folding trajectories leading to different local minima

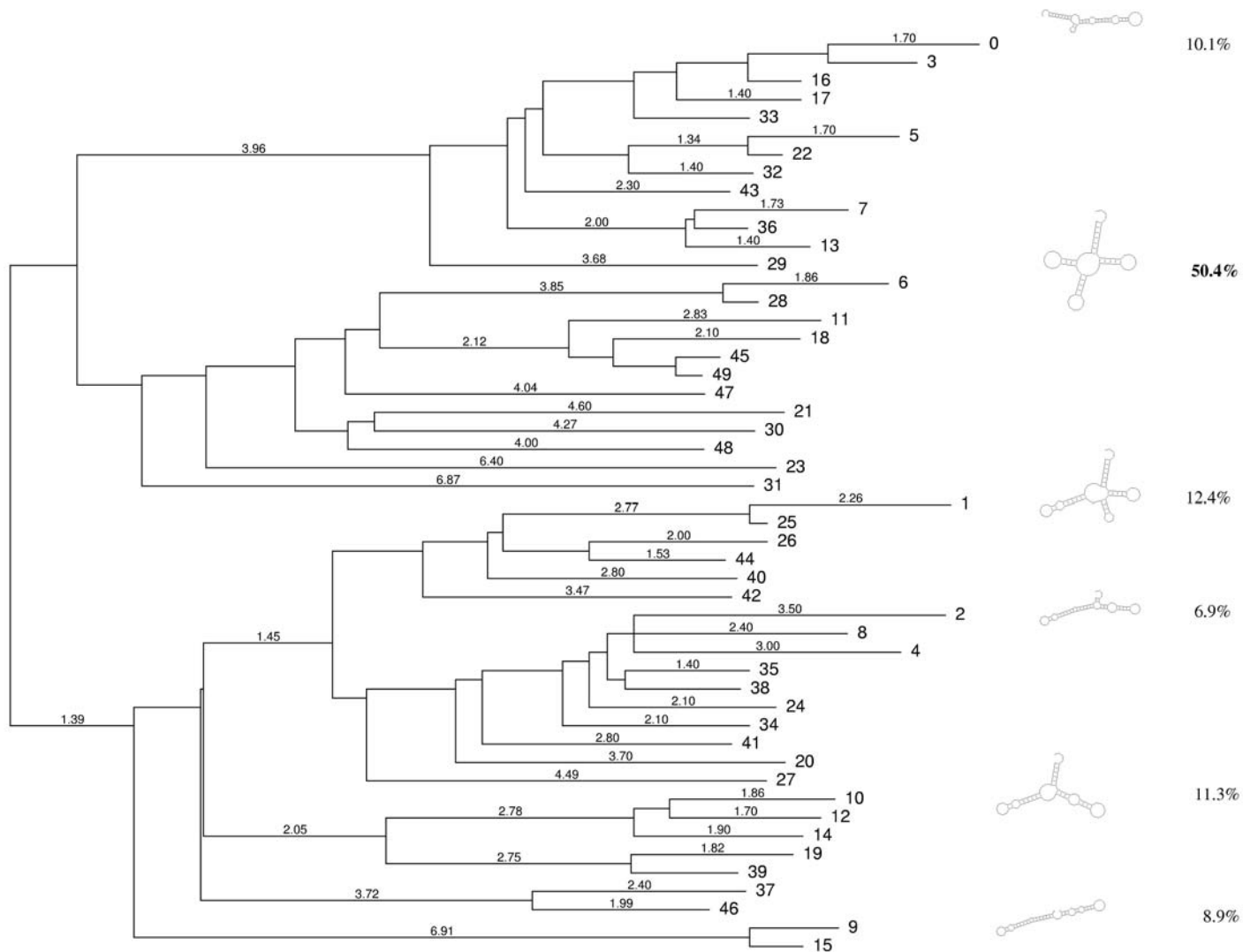
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCUGUGTPCGAUCCACAGAAUUCGCACCA



Kinetic folding of phenylalanyl-tRNA



Folding dynamics of tRNA^{phe} with and without modified nucleotides



Barrier tree of tRNA^{phe} without modified nucleotides

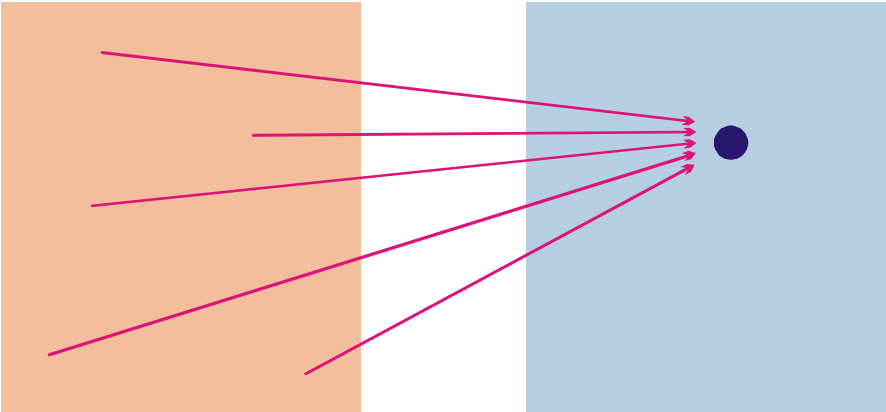
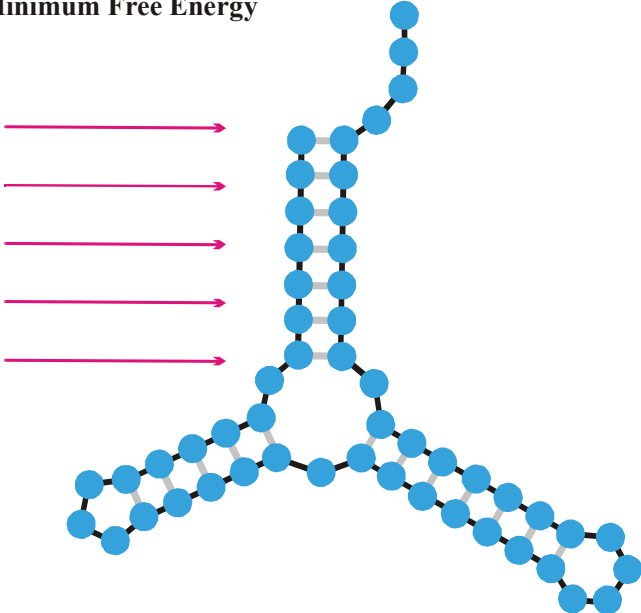
*The one sequence-one structure paradigm fails for
RNA molecules!*

*RNA structures can be predicted and engineered
in silico and in vitro!*

1. From working copy in molecular genetics to riboswitches
2. The never-ending story of RNA structure
- 3. Neutral networks, intersections, and bistable RNA molecules**
4. Evolution of molecules *in silico*
5. A few selected experiments

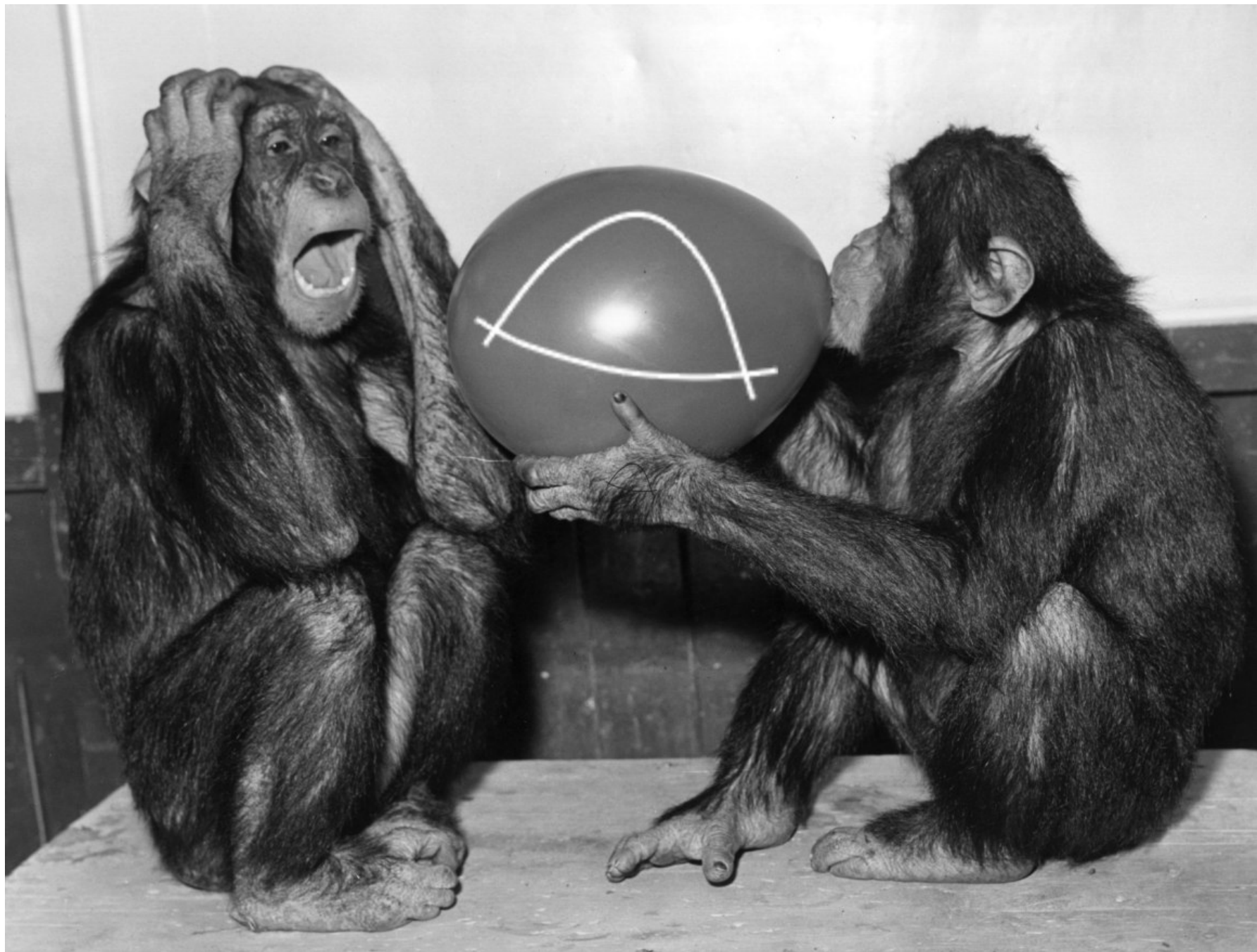
**Criterion of
Minimum Free Energy**

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG
UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG
CAUUGGUGCUAAUGAUUUAGGGCUGUAUUCUGUAUAGCGAUCAGUGUCCG
GUAGGCCUCUUGACAUAAGAUUUUUCCAUGGUGGGAGAUGGCCAUUGCAG



Sequence Space

Shape Space



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.

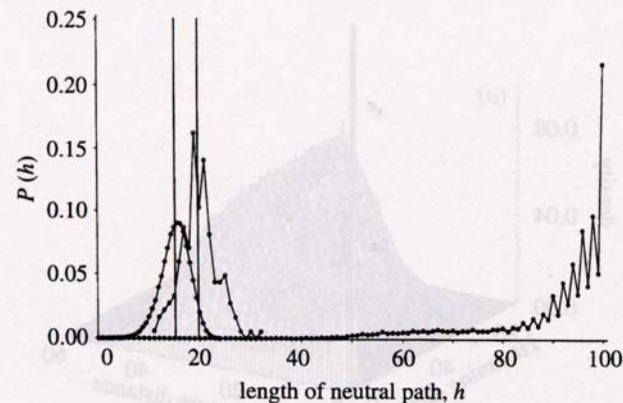
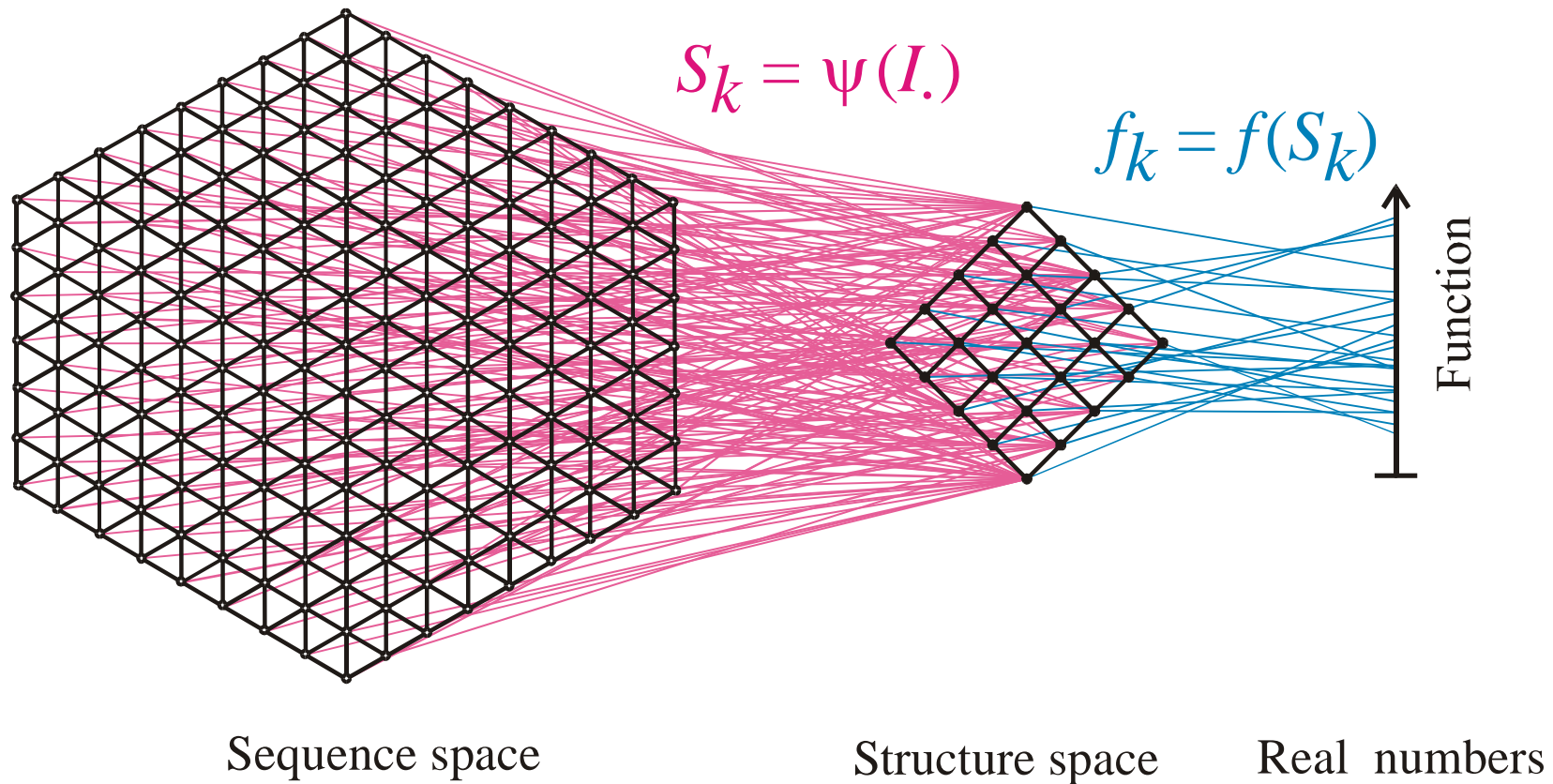
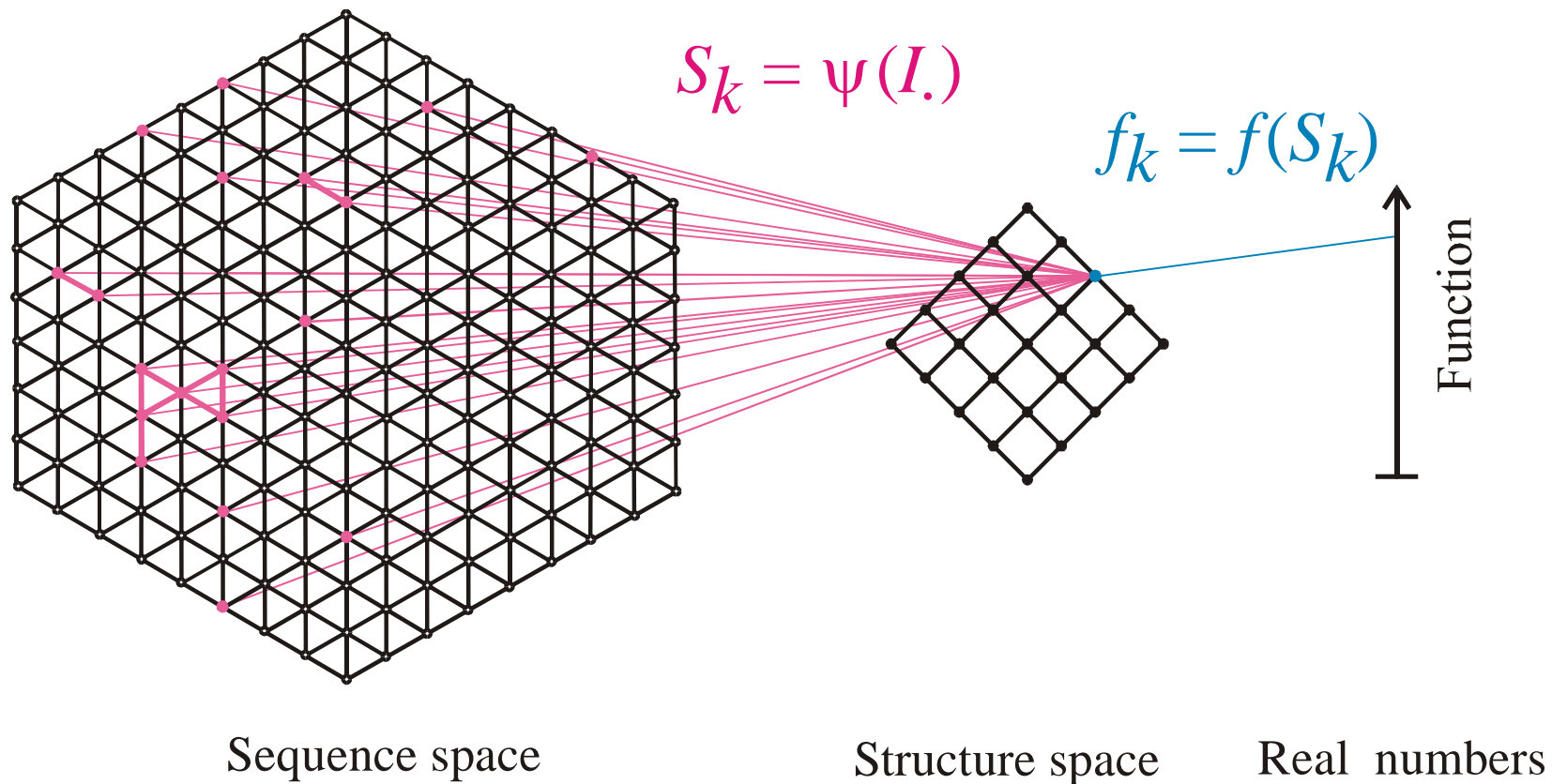
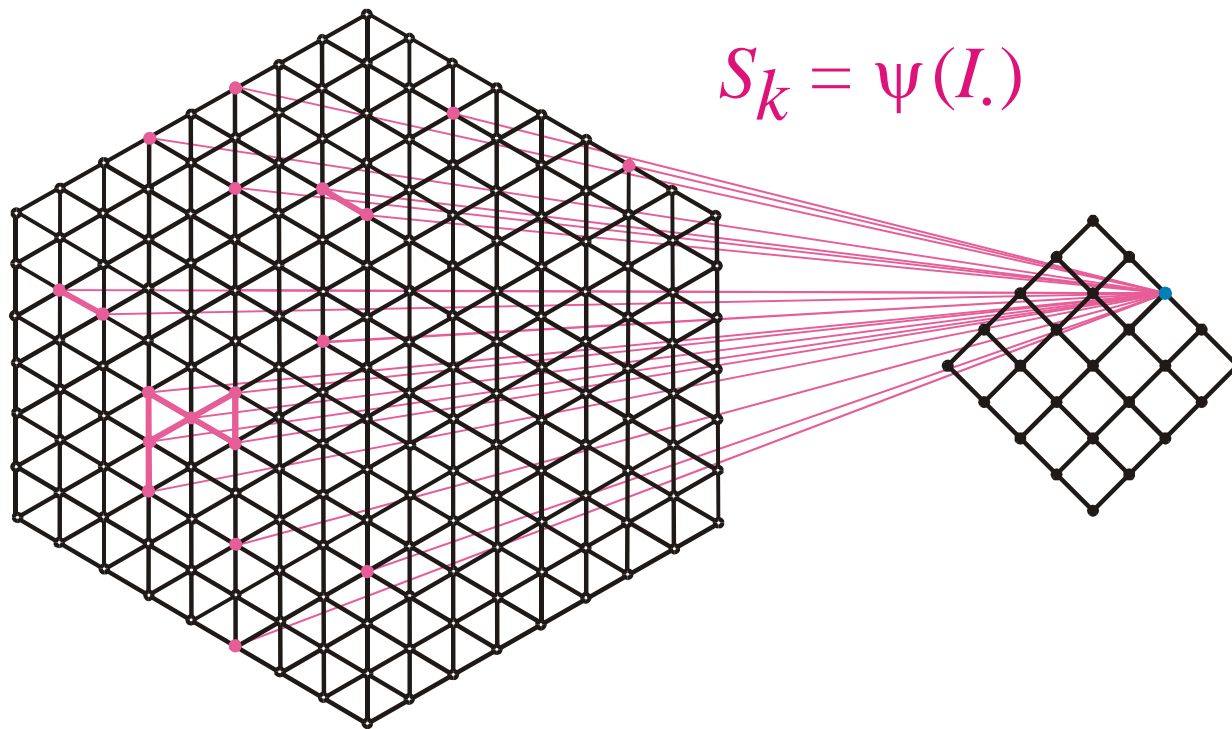


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and 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).



Mapping from sequence space into structure space and into function

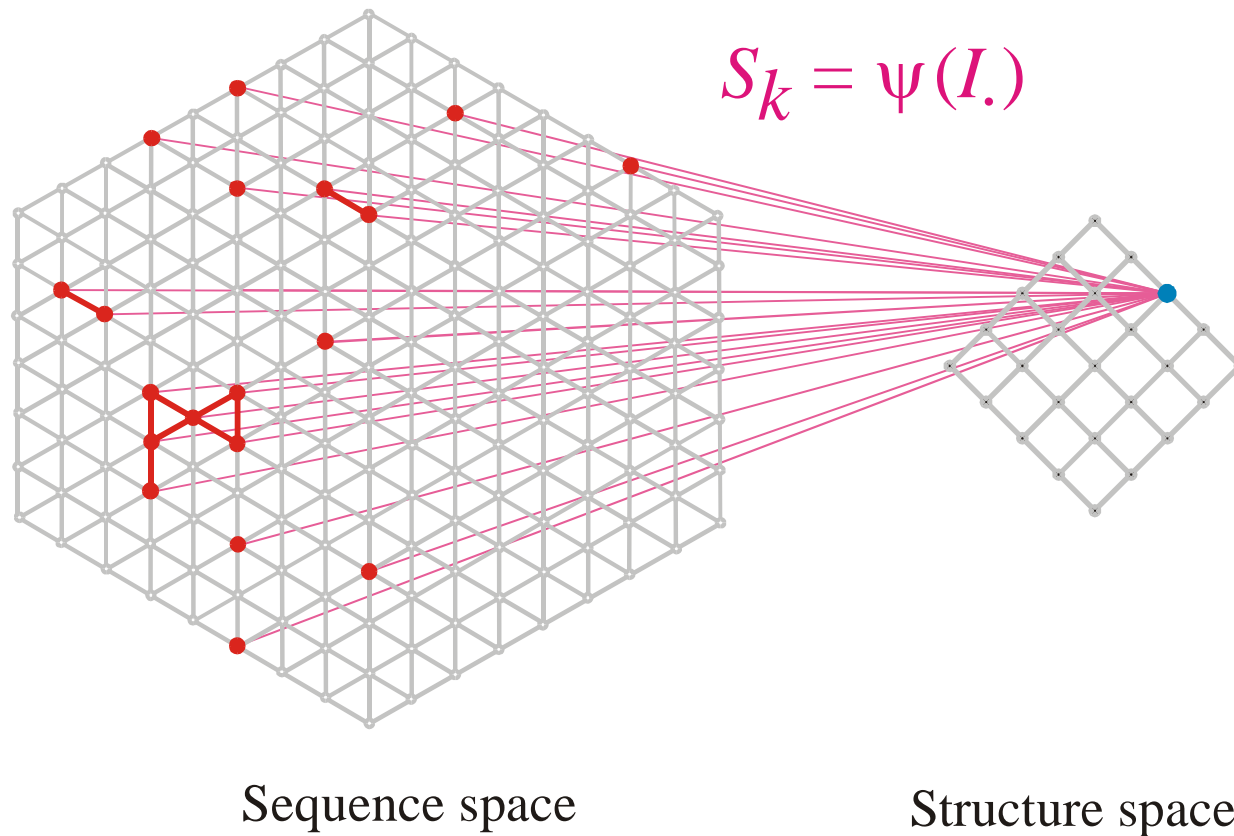




$$S_k = \psi(I.)$$

Sequence space

Structure space



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

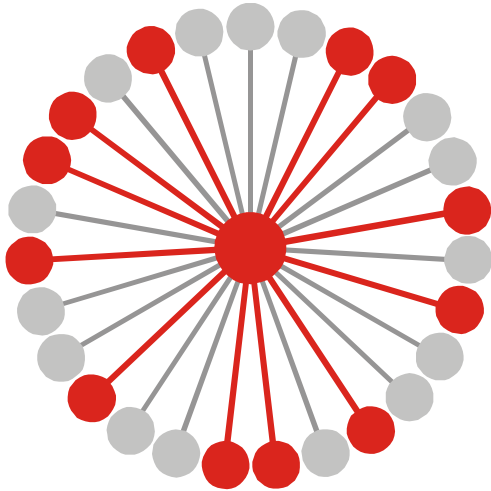
Neutral networks are sets of sequences forming the same object in a phenotype space. The neutral network \mathbf{G}_k is, for example, the pre-image of the structure S_k in sequence space:

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

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

Neutral networks of small biomolecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, $N=4^n$, becomes very large with increasing length, and is prohibitive for numerical computations.

Neutral networks can be modelled by **random graphs** in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.



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

$$\lambda_j = 12 / 27 = 0.444, \quad \bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

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

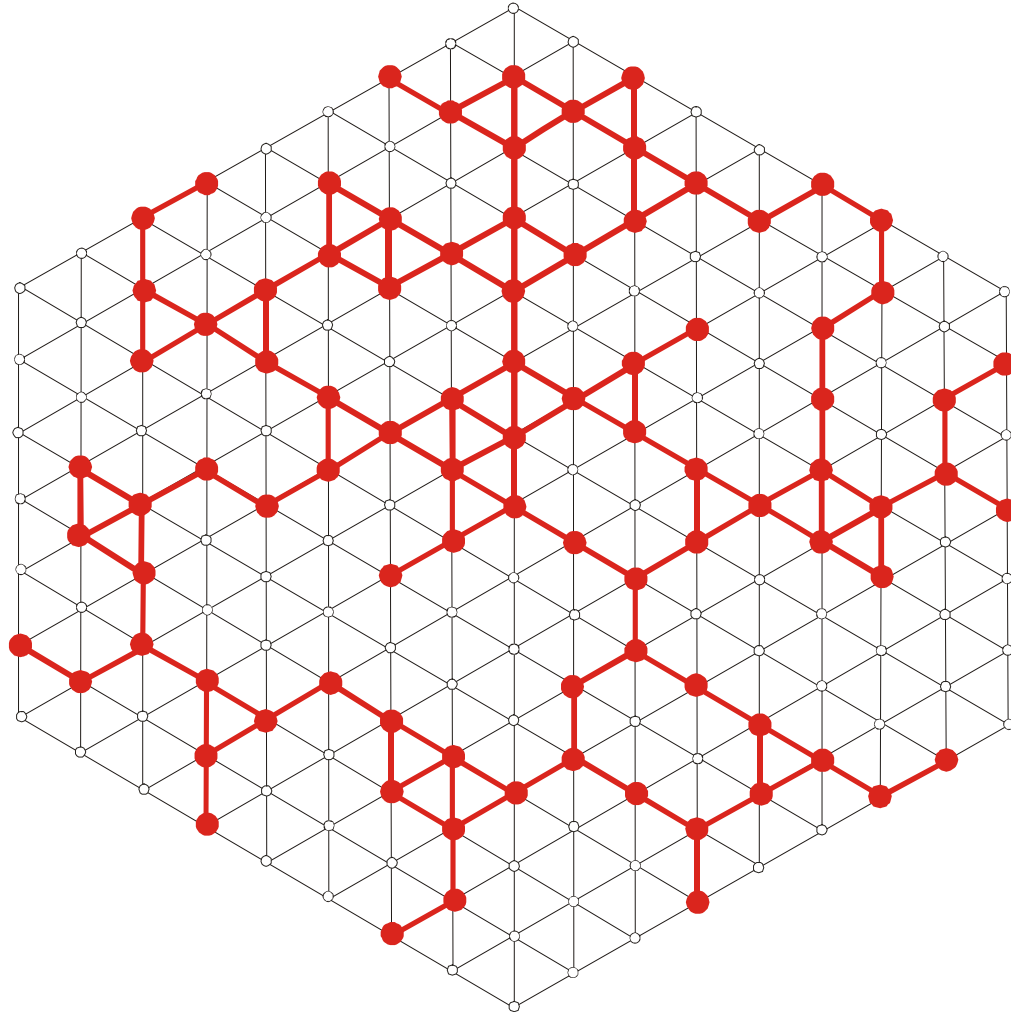
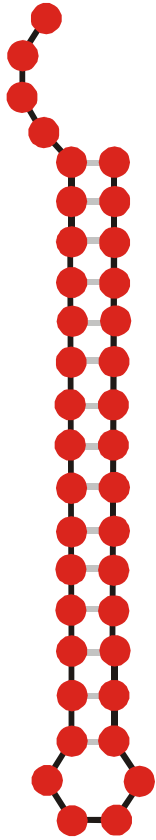
Alphabet size κ : **AUGC** $\Rightarrow \kappa = 4$

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

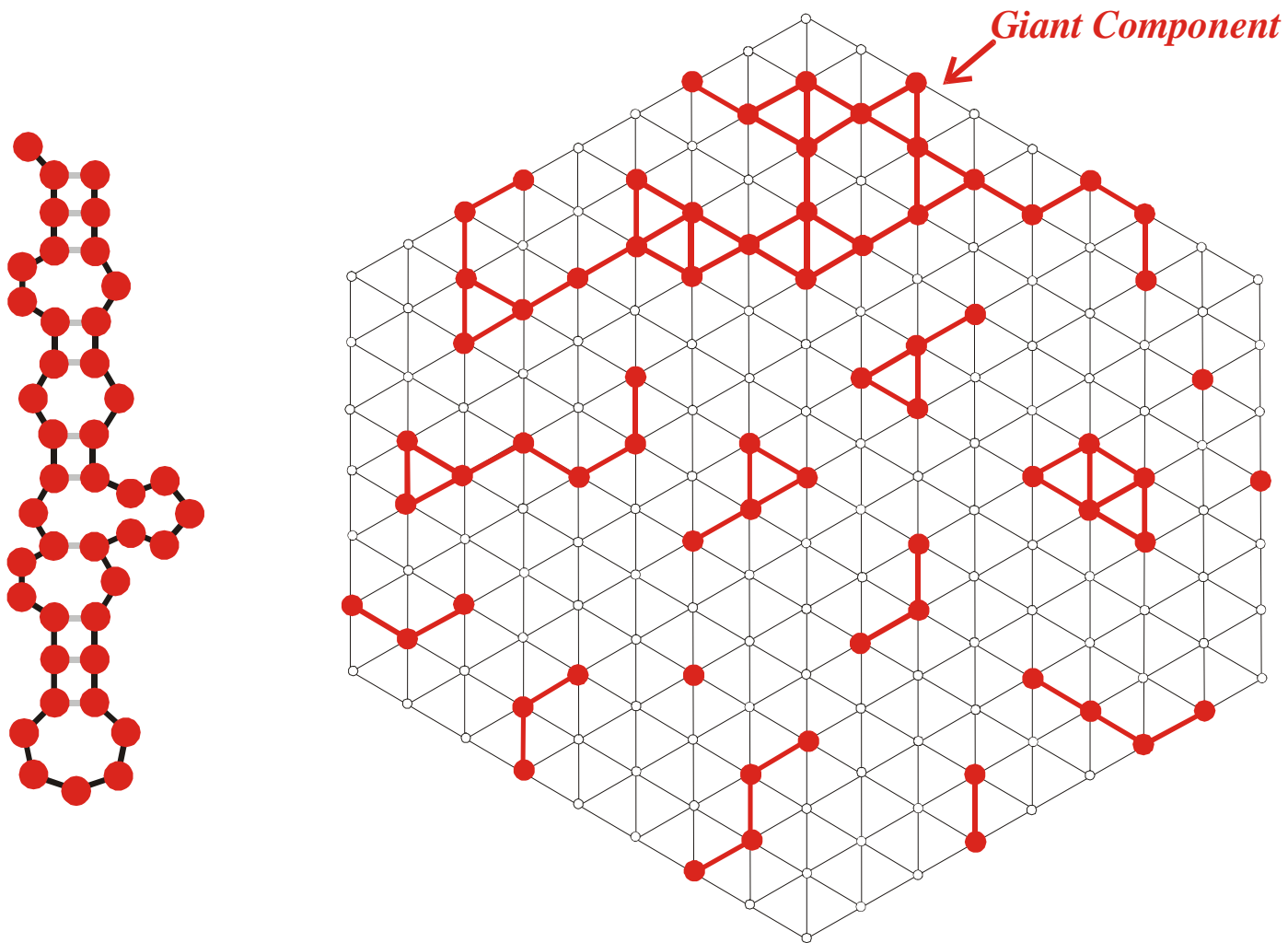
$\bar{\lambda}_k < \lambda_{cr}$ network **G_k** is **not** connected

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

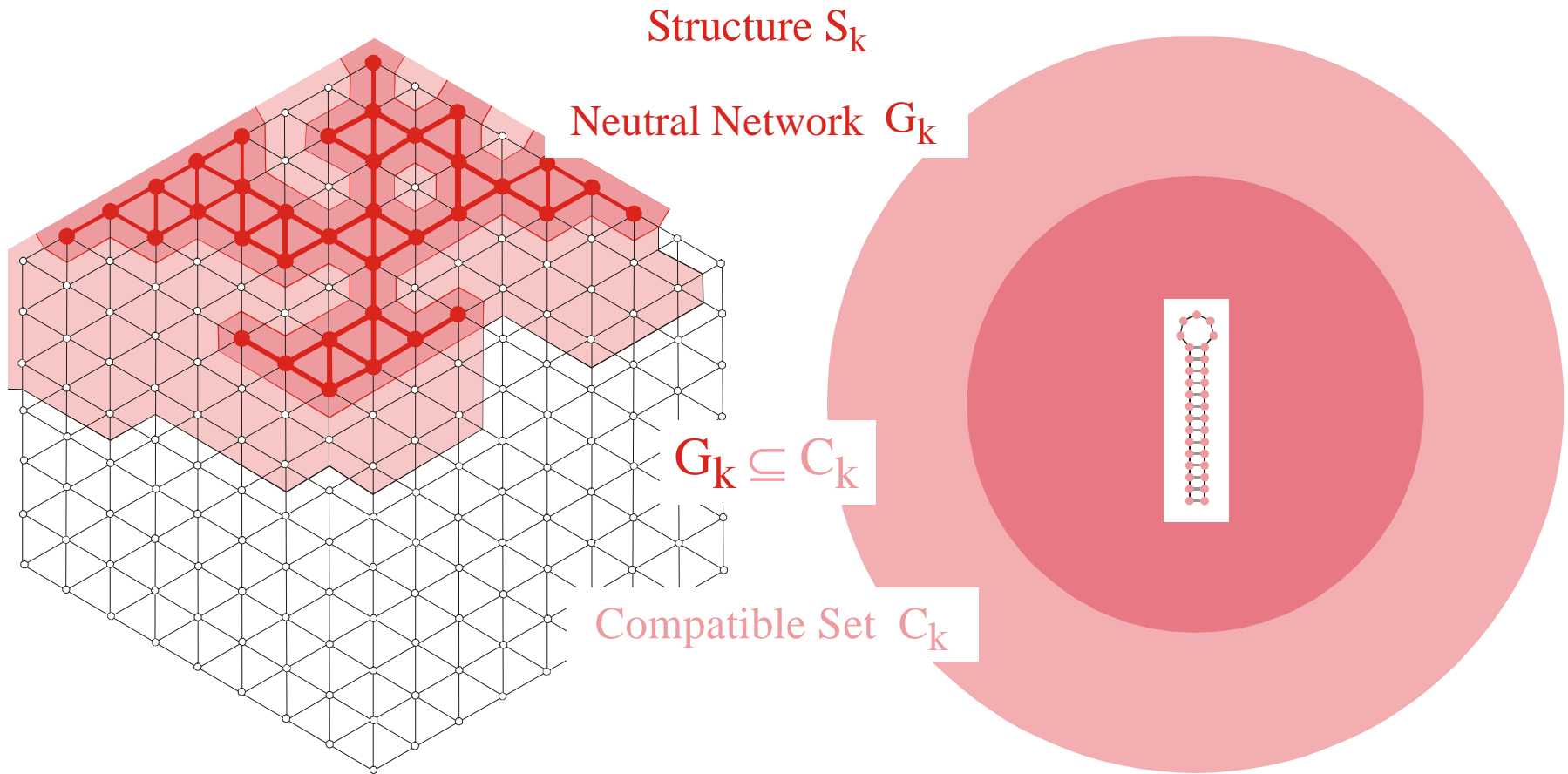
Mean degree of neutrality and connectivity of neutral networks



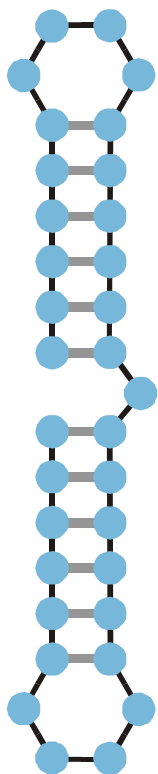
A connected neutral network formed by a common structure



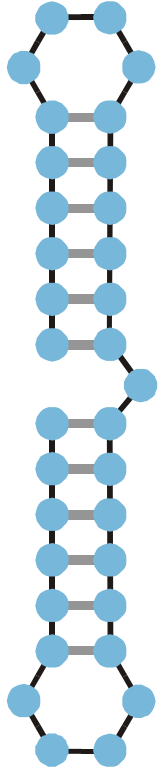
A multi-component neutral network formed by a rare structure



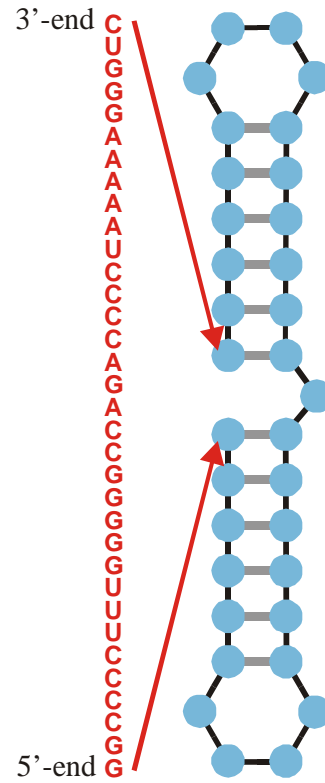
The **compatible set** C_k of a structure S_k consists of all sequences which form S_k as its minimum free energy structure (the **neutral network** G_k) or one of its suboptimal structures.



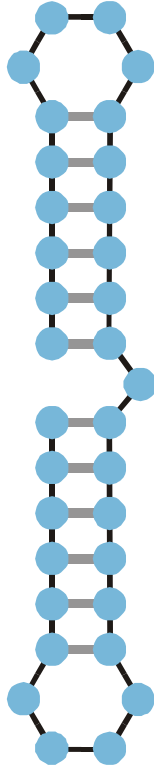
Structure



Structure

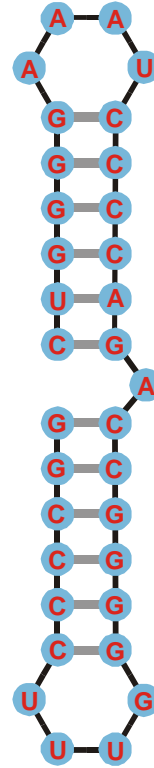


Compatible sequence

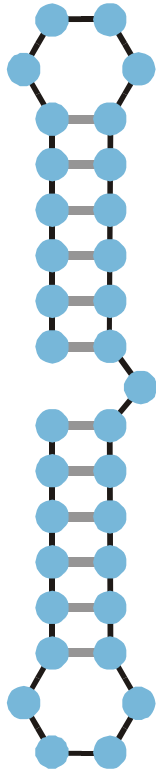


Structure

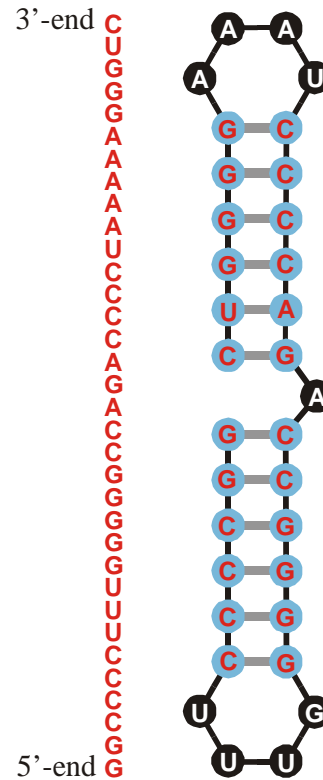
3'-end C
U
G
G
A
A
A
A
A
U
C
C
C
C
A
G
A
C
C
G
G
G
G
G
U
U
U
C
C
C
C
G
5'-end



Compatible sequence



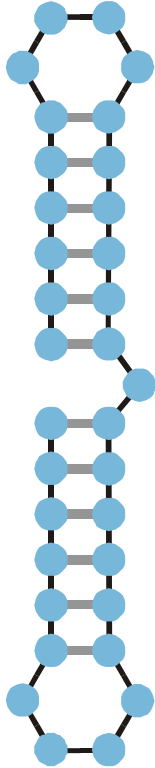
Structure



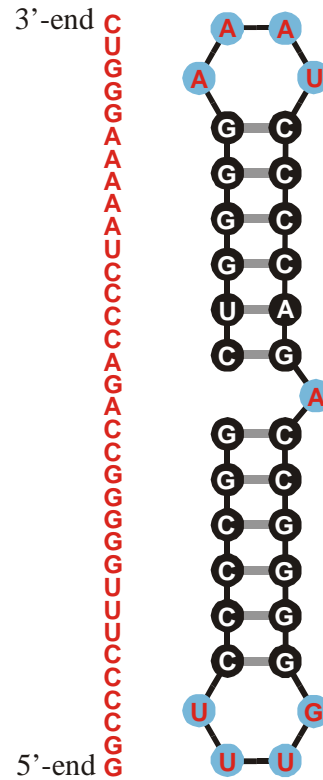
Compatible sequence

Single nucleotides: **A,U,G,C**

Single bases pairs are varied independently



Structure

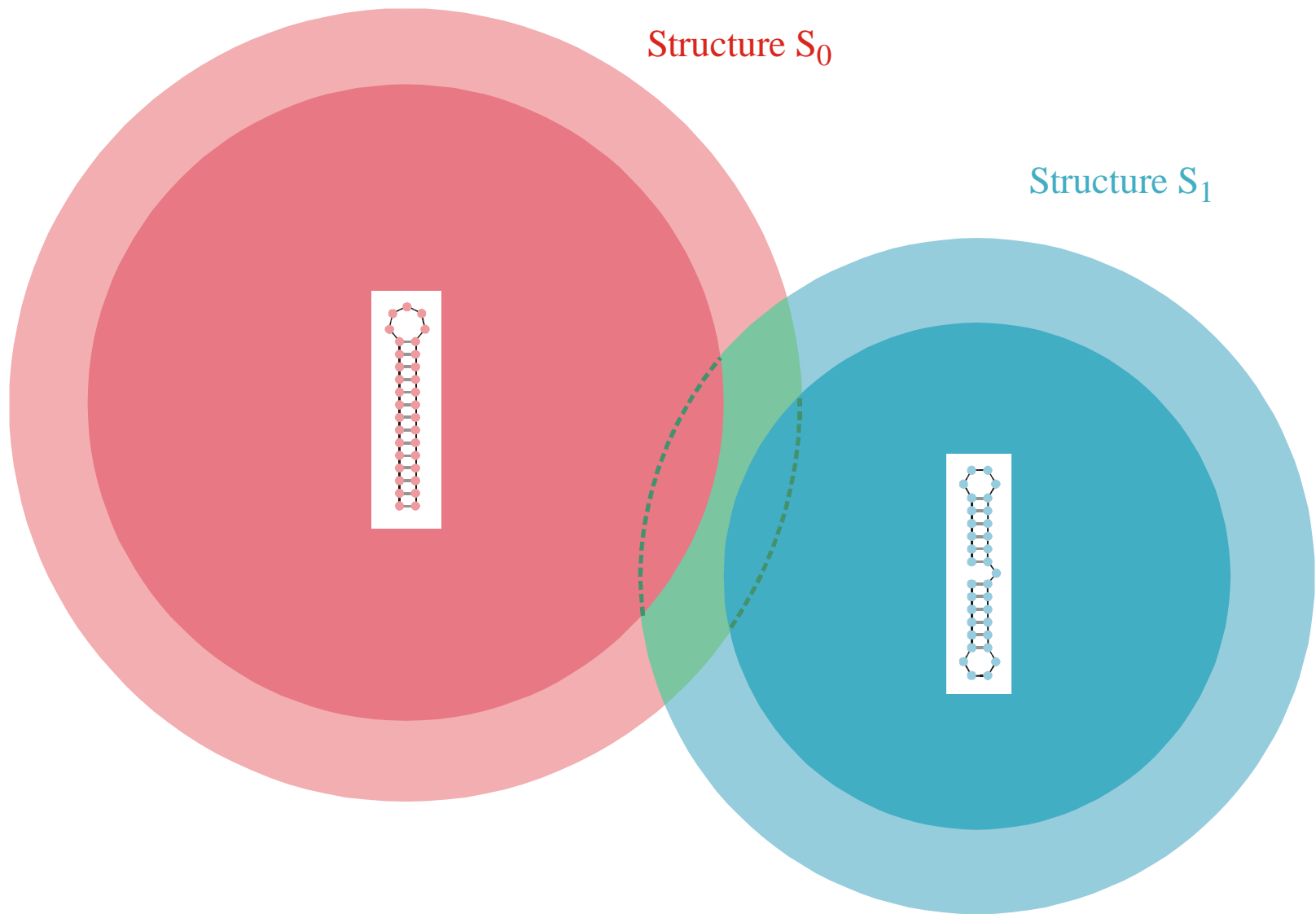


Compatible sequence

Base pairs:

**AU , UA
GC , CG
GU , UG**

Base pairs are varied in strict correlation



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

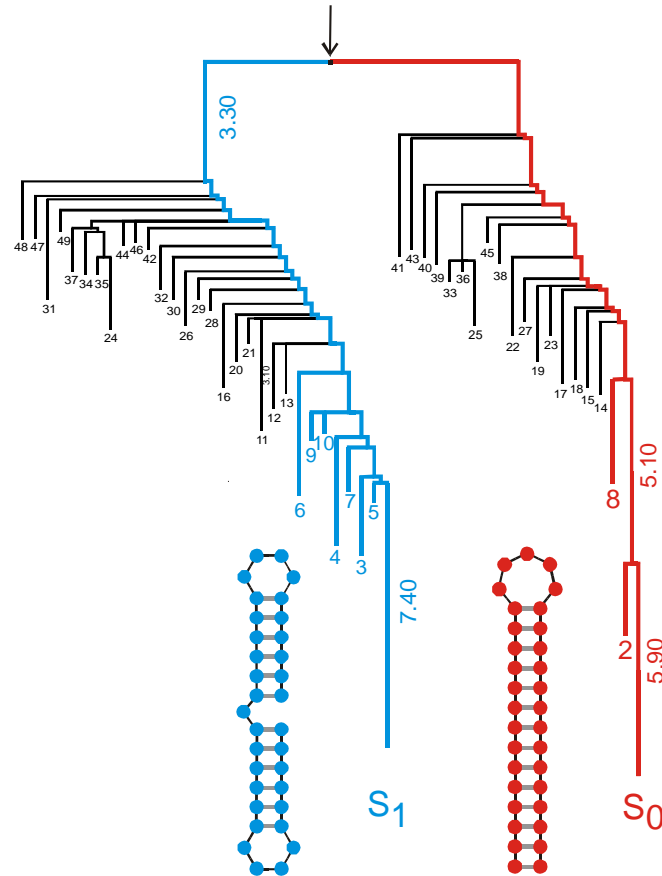
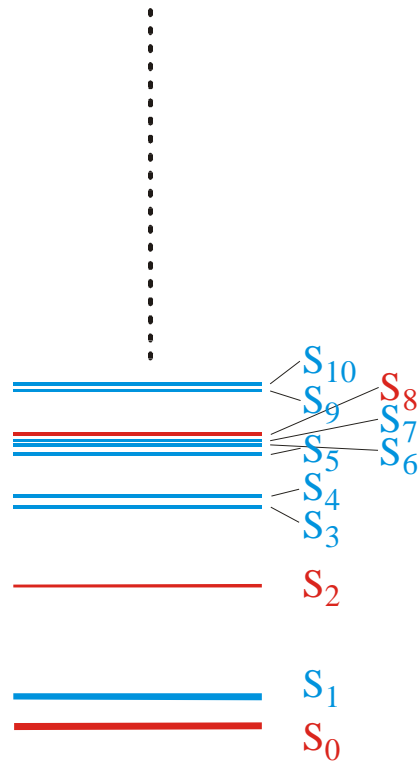
THEOREM 5. INTERSECTION-THEOREM. *Let s and s' be arbitrary secondary structures and $C[s], C[s']$ their corresponding compatible sequences. Then,*

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

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. ■

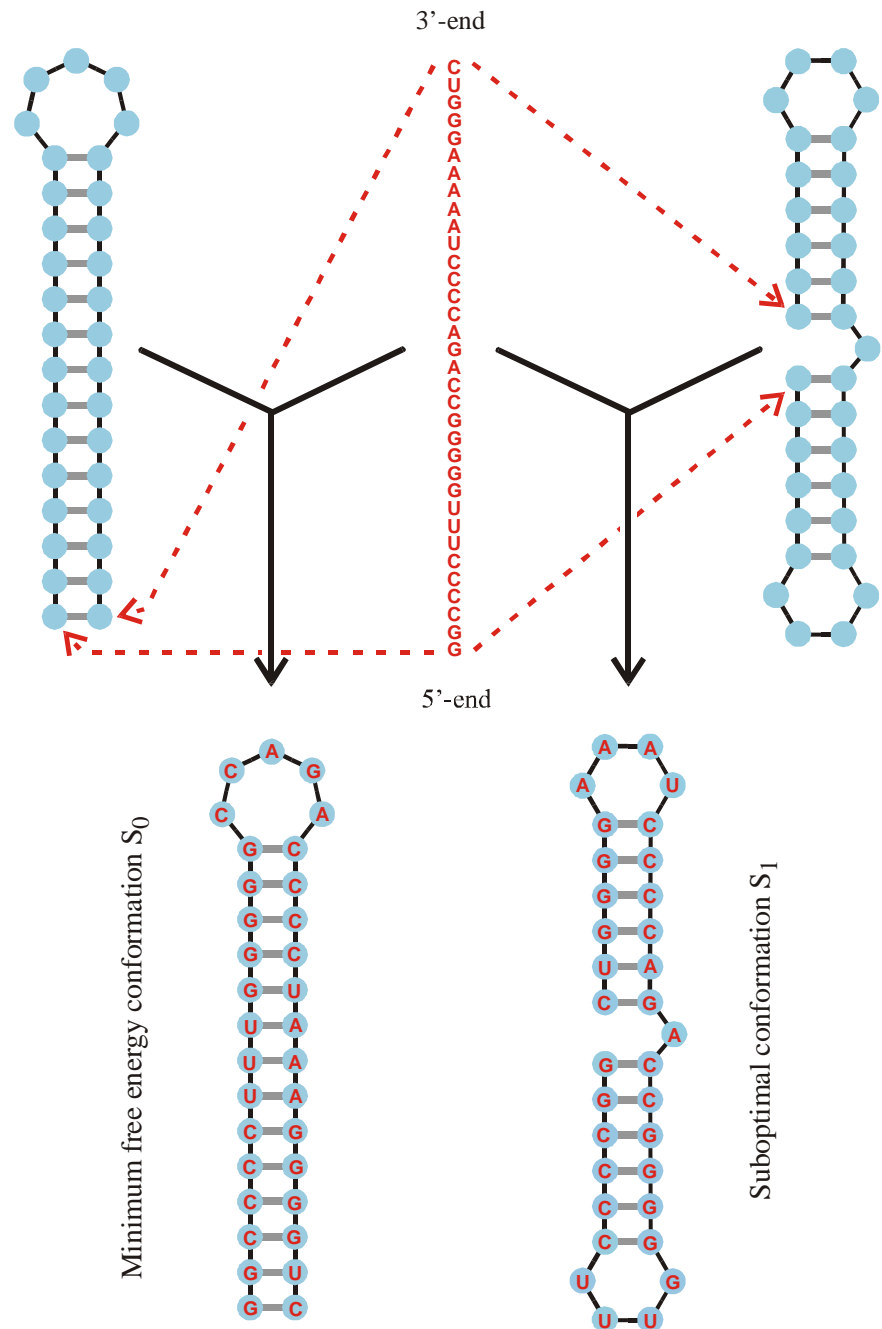
Remark. A generalization of the statement of theorem 5 to three different structures is false.

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

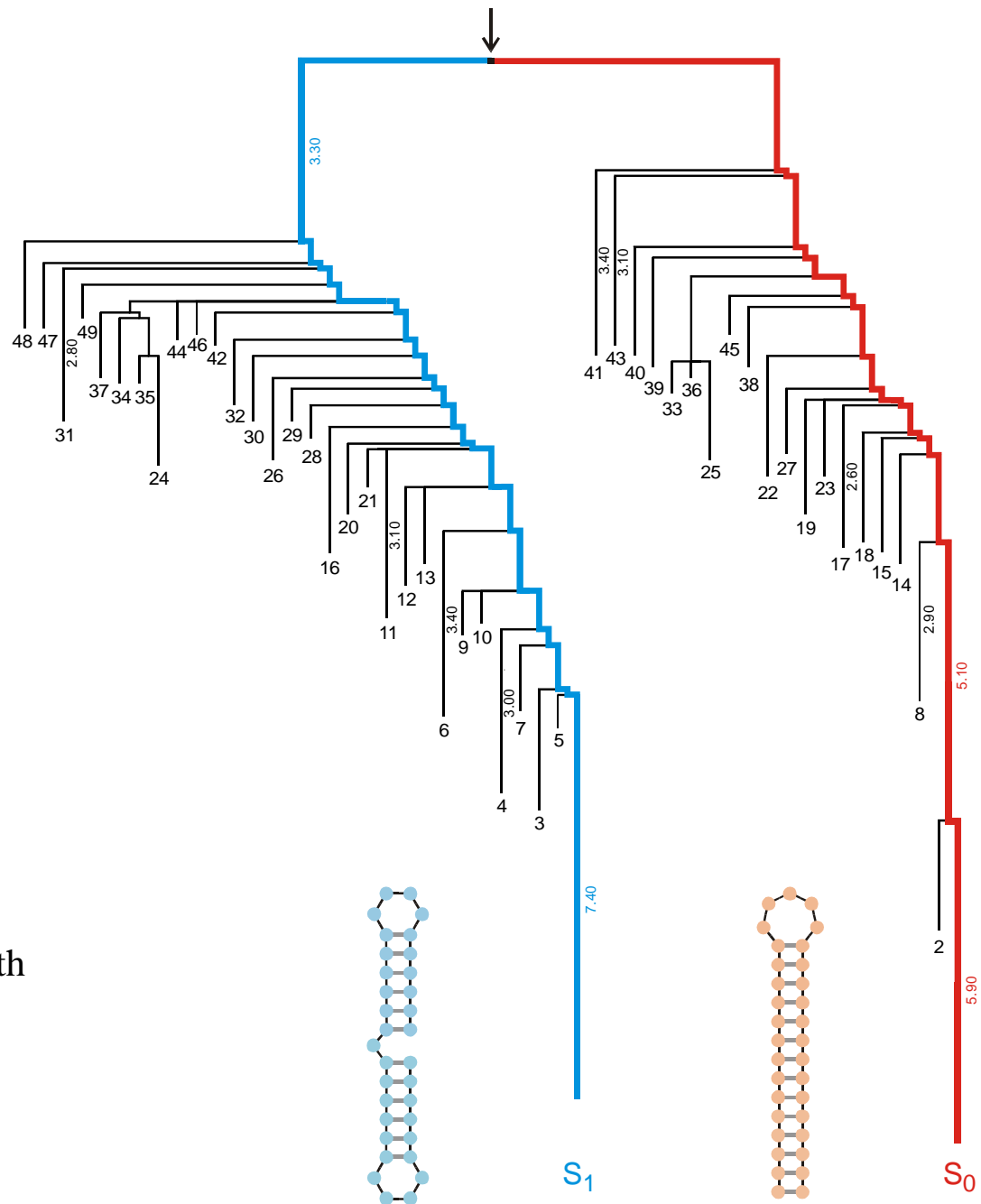


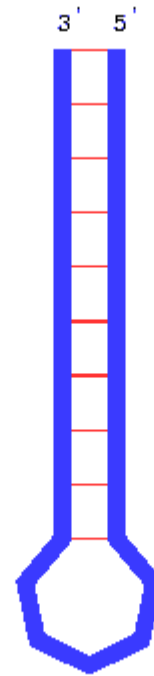
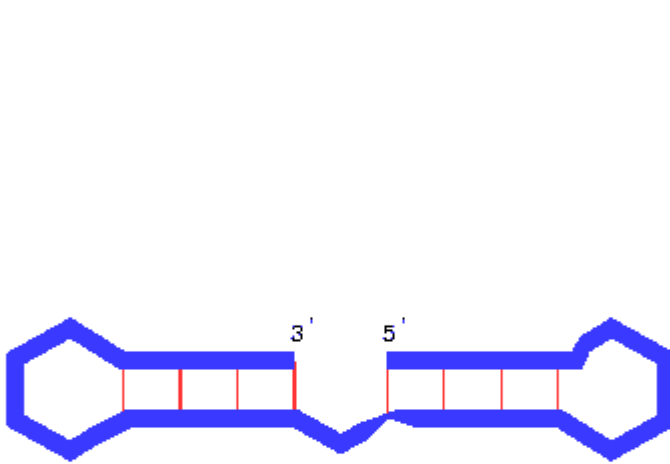
A typical energy landscape of a sequence with two (meta)stable conformations

One sequence is compatible with two structures



Barrier tree of a sequence with two conformations





Kinetics of RNA refolding between a long living metastable conformation and the minimum free energy structure

There are many more sequences than RNA secondary structures!

Common RNA structures form extended neutral networks in sequence space!

RNA molecules can exist in two (or more) long-lived conformations!

1. From working copy in molecular genetics to riboswitches
2. The never-ending story of RNA structure
3. Neutral networks, intersections, and bistable RNA molecules
- 4. Evolution of molecules *in silico***
5. A few selected experiments

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. Torzkadash, 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 00035-01 and Z01 DC 00038-01 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 HD00428 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*

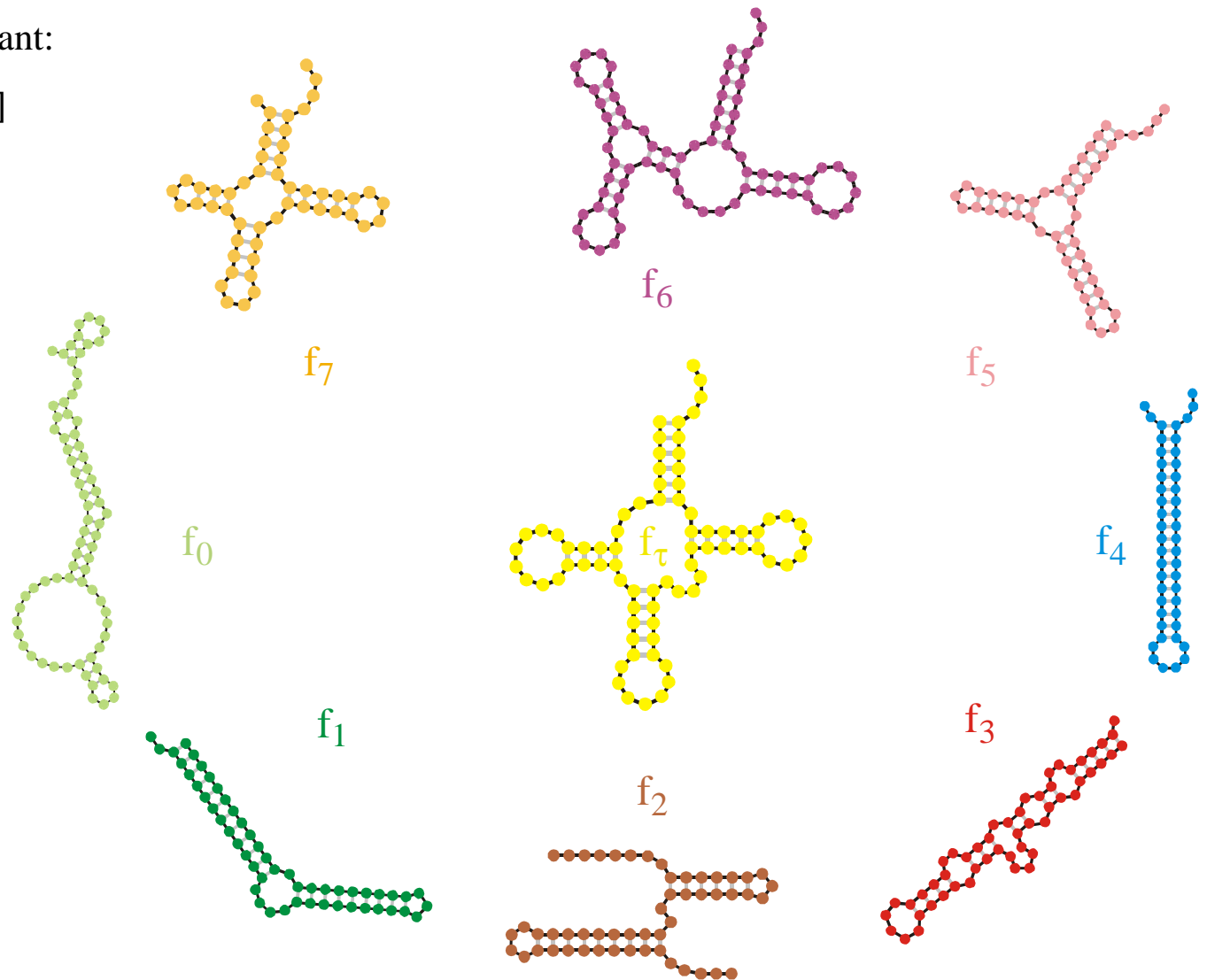
W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455

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.

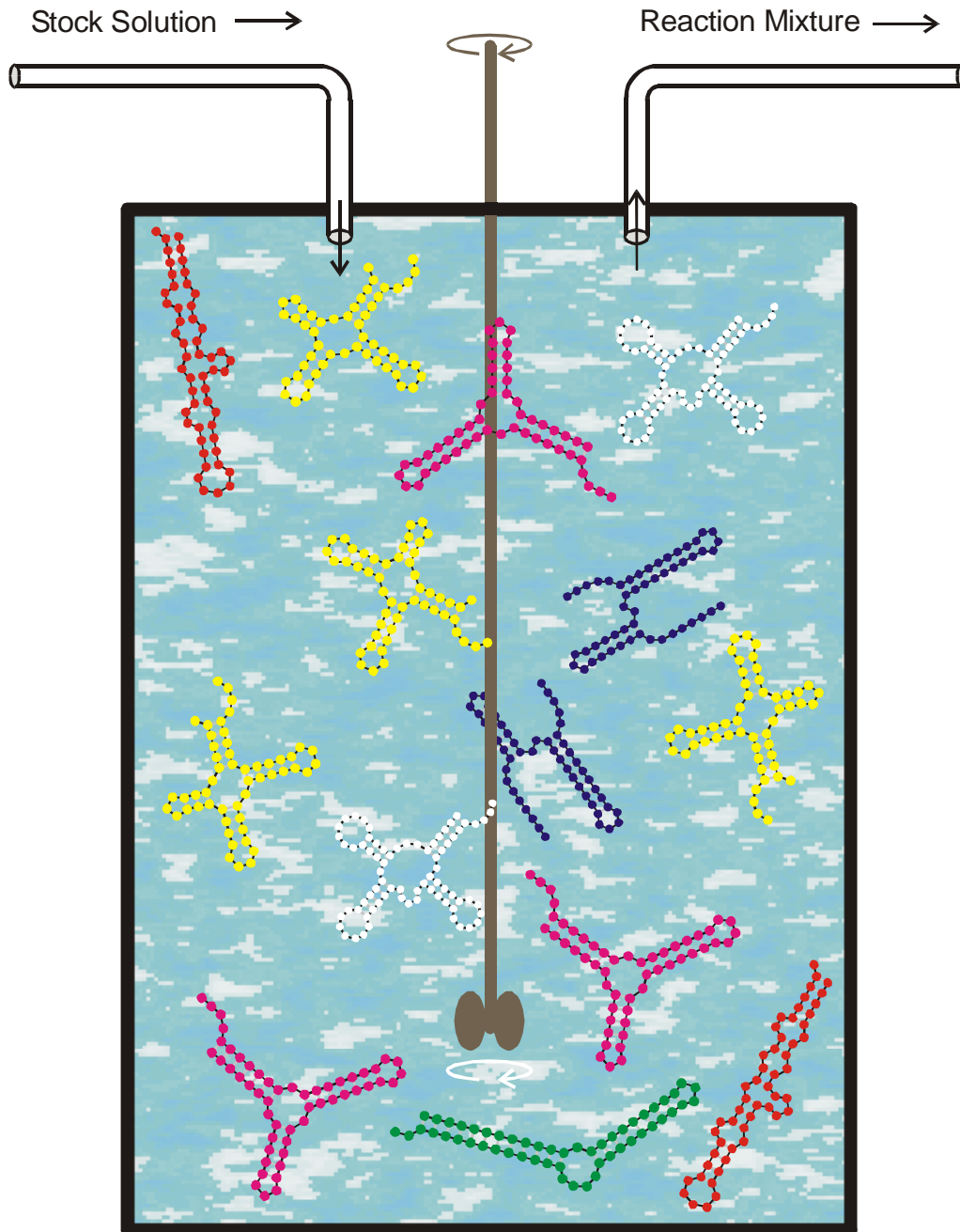
Replication rate constant:

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

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



Evaluation of RNA secondary structures yields replication rate constants



Replication rate constant:

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

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

Selection constraint:

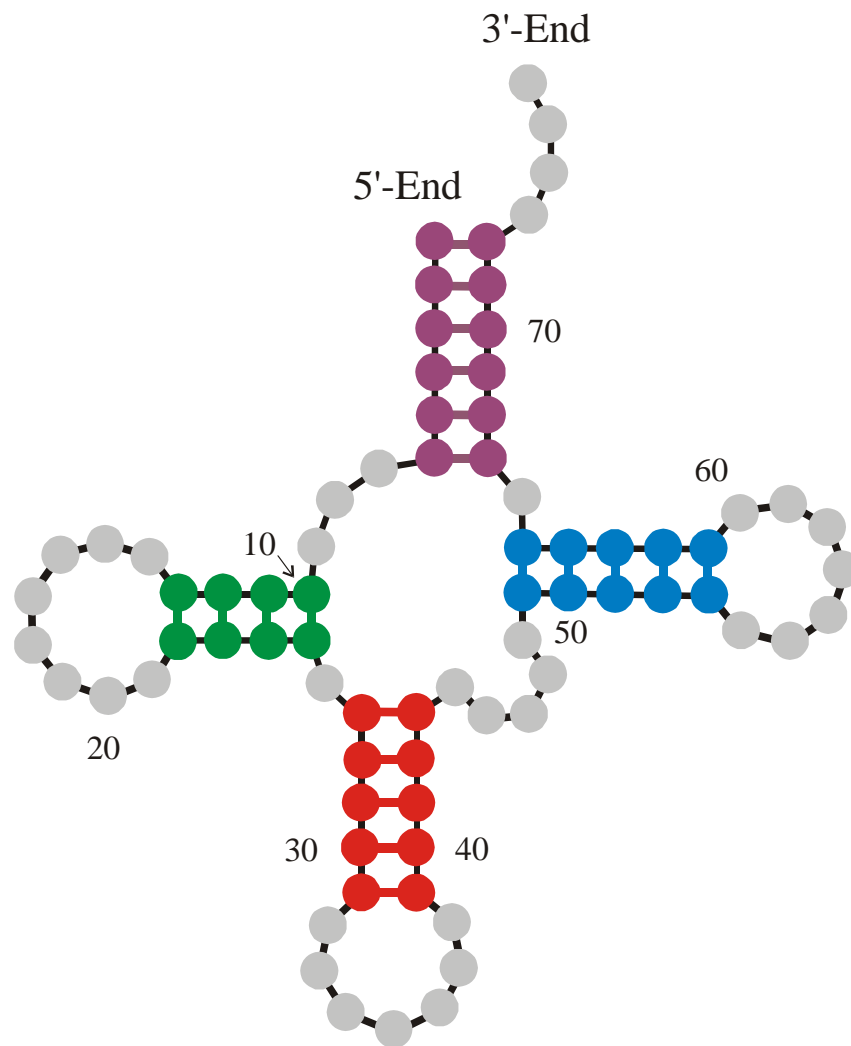
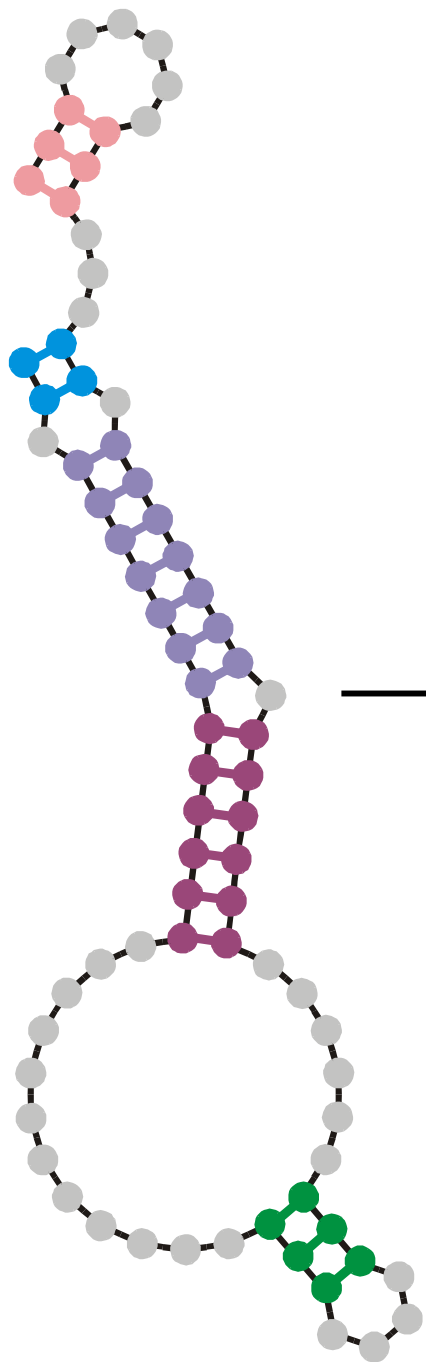
Population size, $N = \#$ RNA molecules, is controlled by the flow

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

Mutation rate:

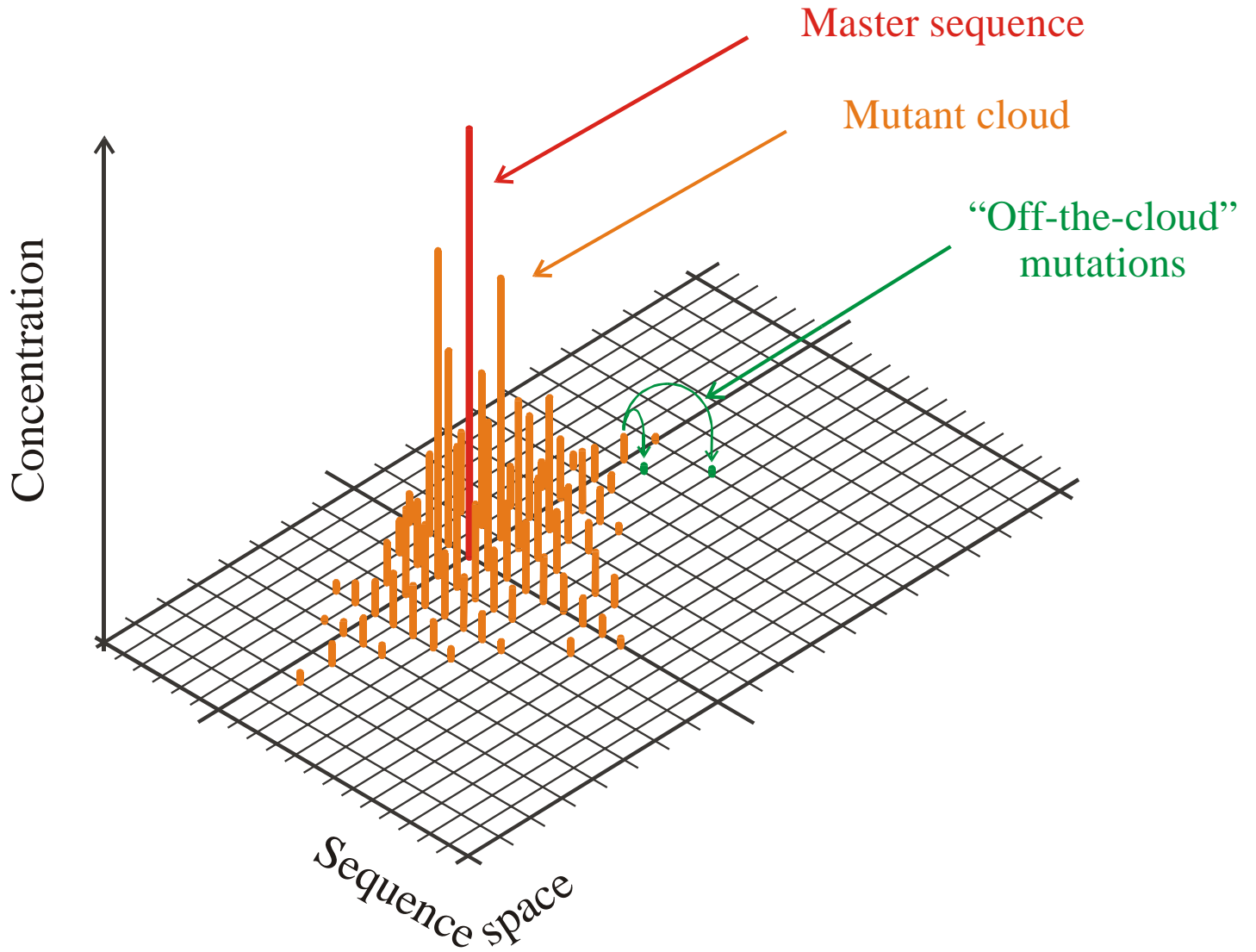
$p = 0.001 / \text{site} \times \text{replication}$

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

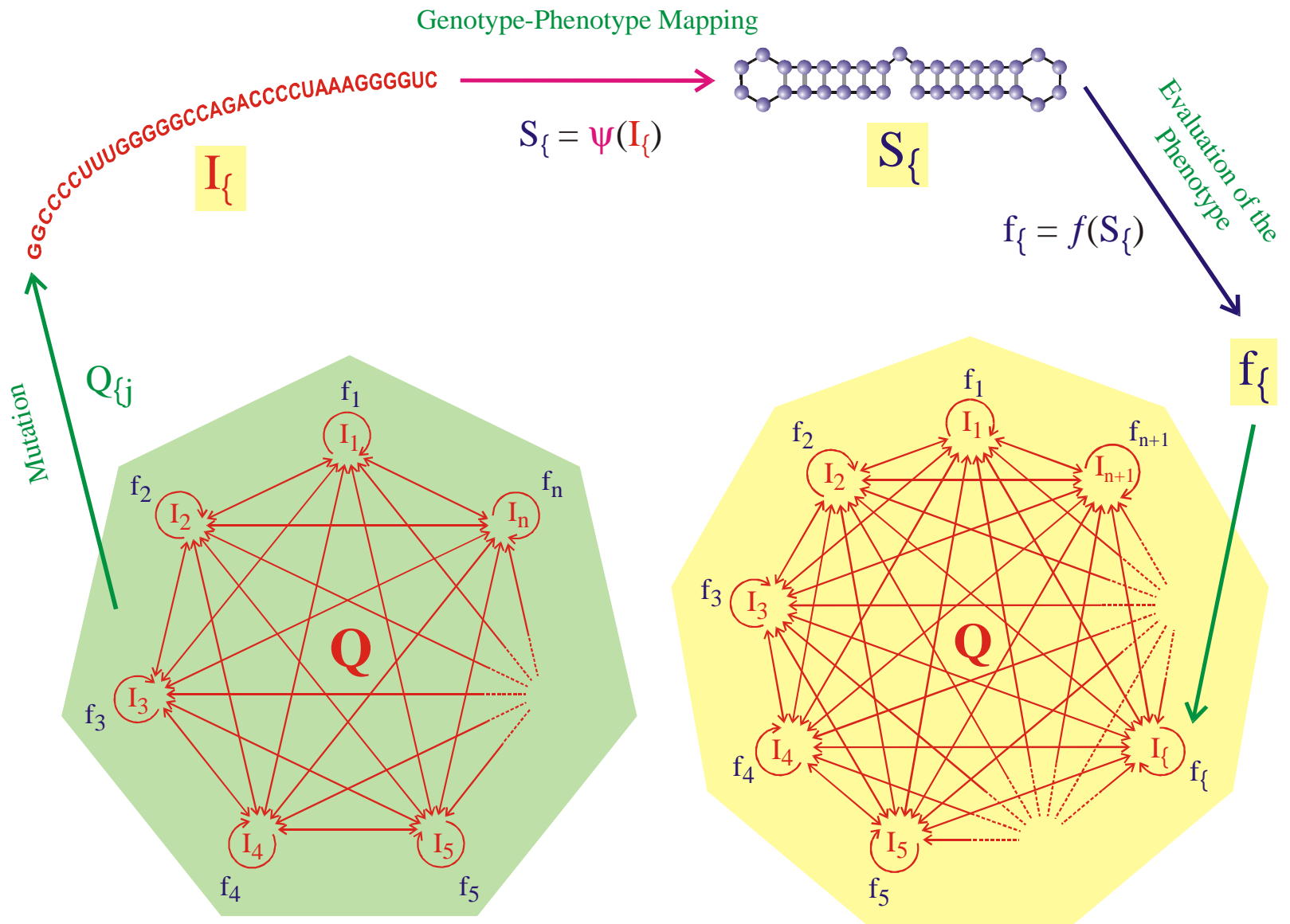


Randomly chosen
initial structure

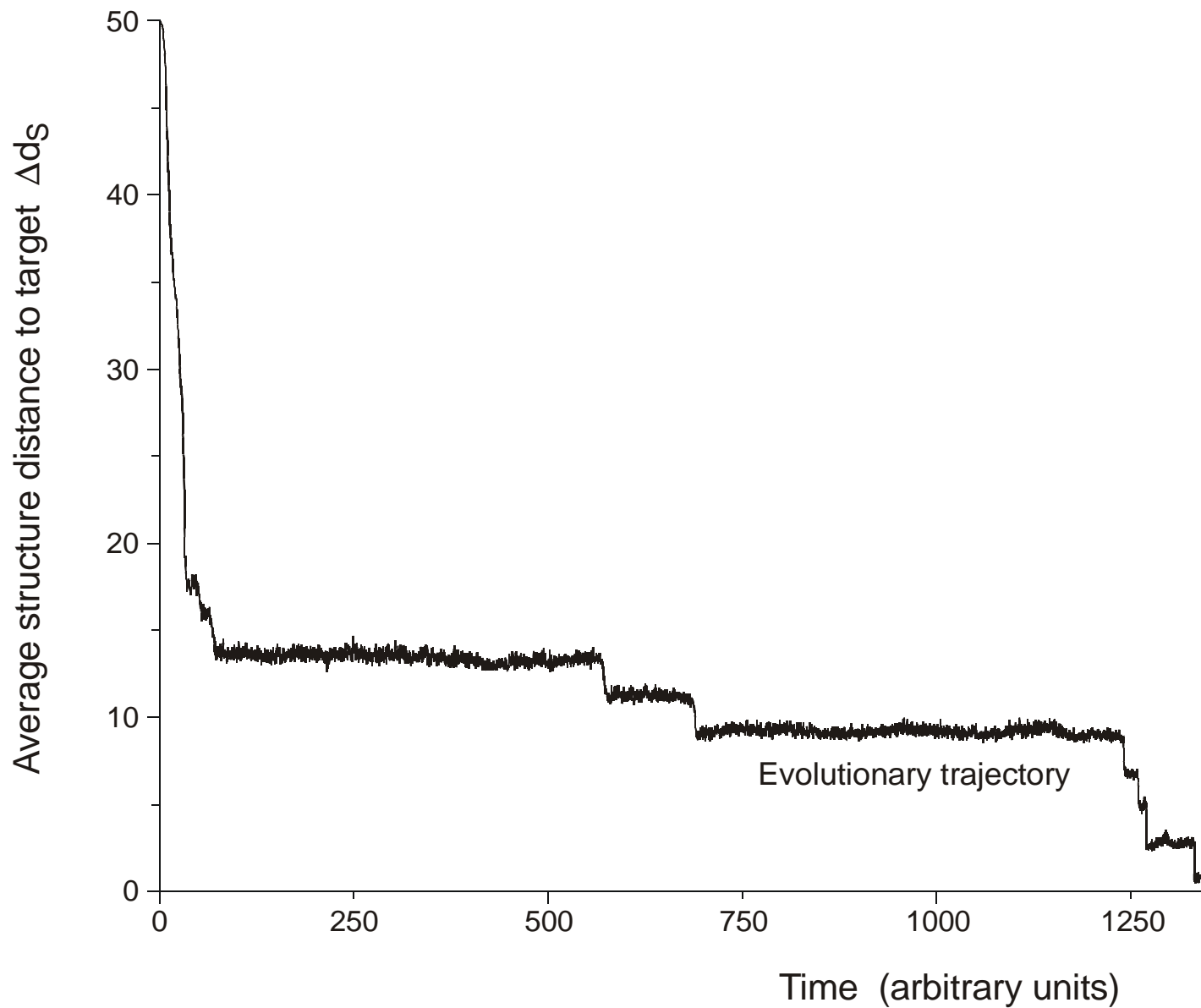
Phenylalanyl-tRNA as
target structure



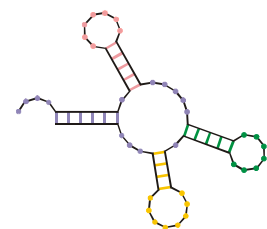
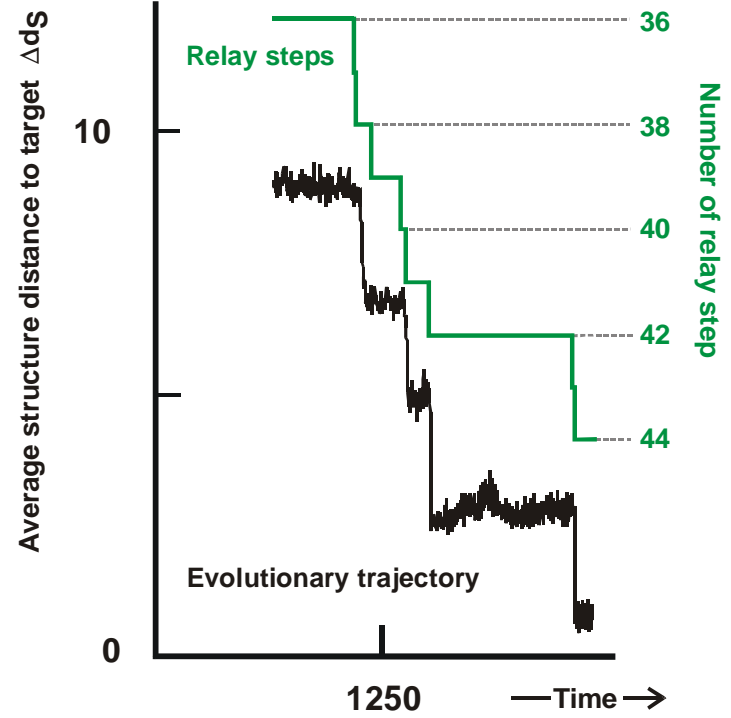
The molecular quasispecies
in sequence space



Evolutionary dynamics including molecular phenotypes

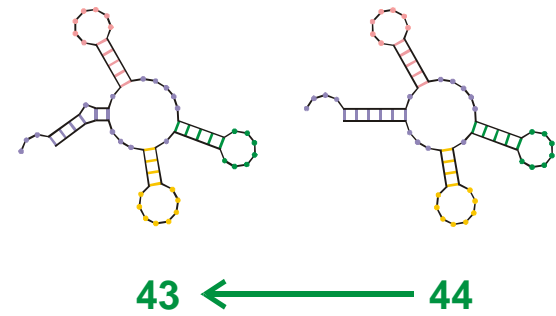
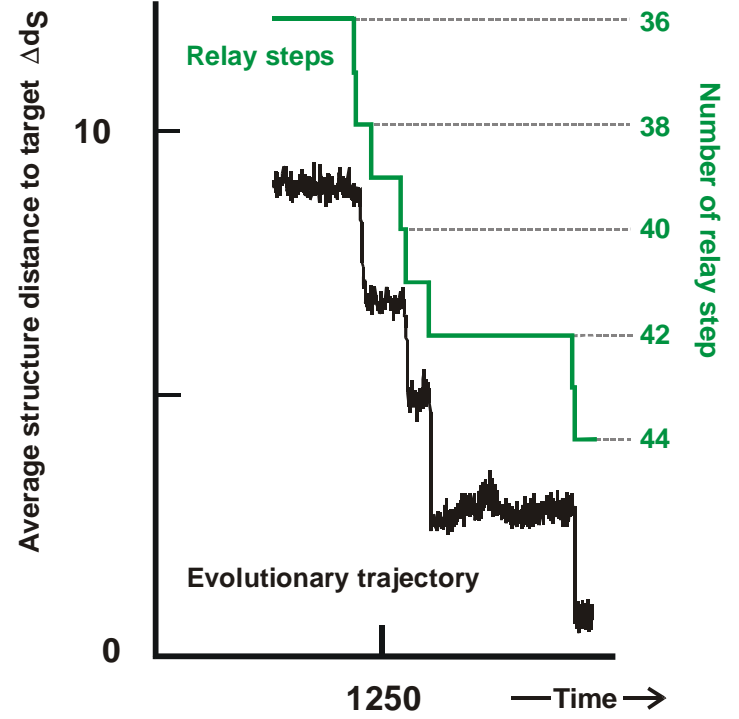


In silico optimization in the flow reactor: Evolutionary trajectory

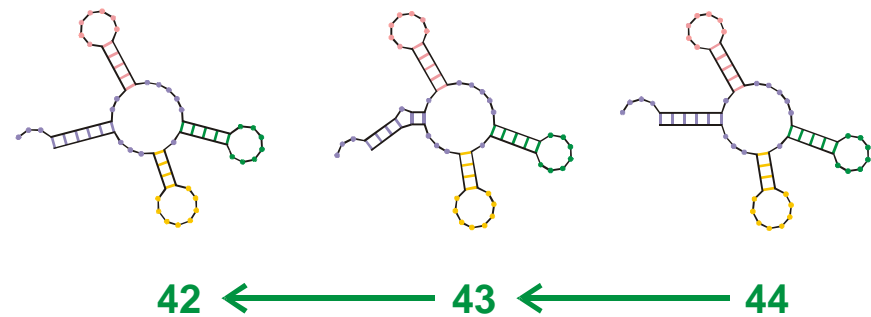
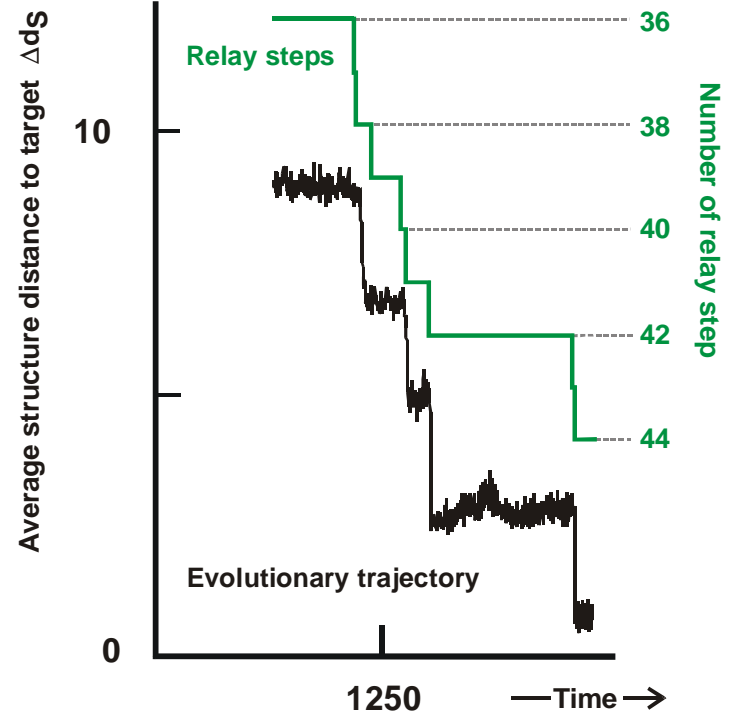


44

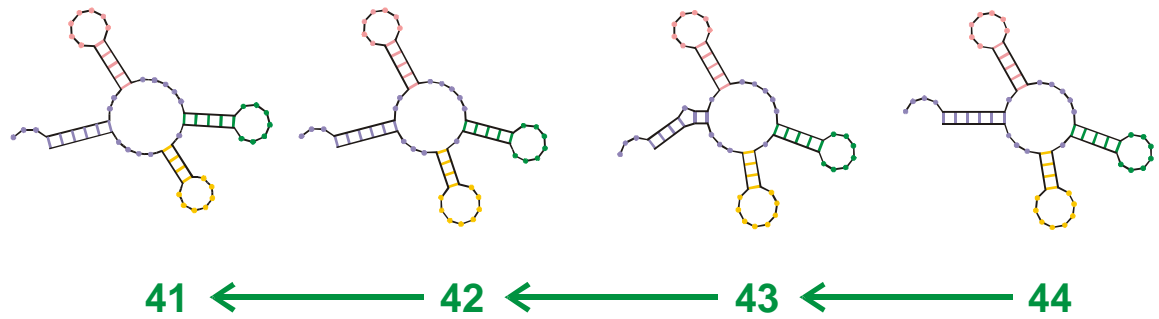
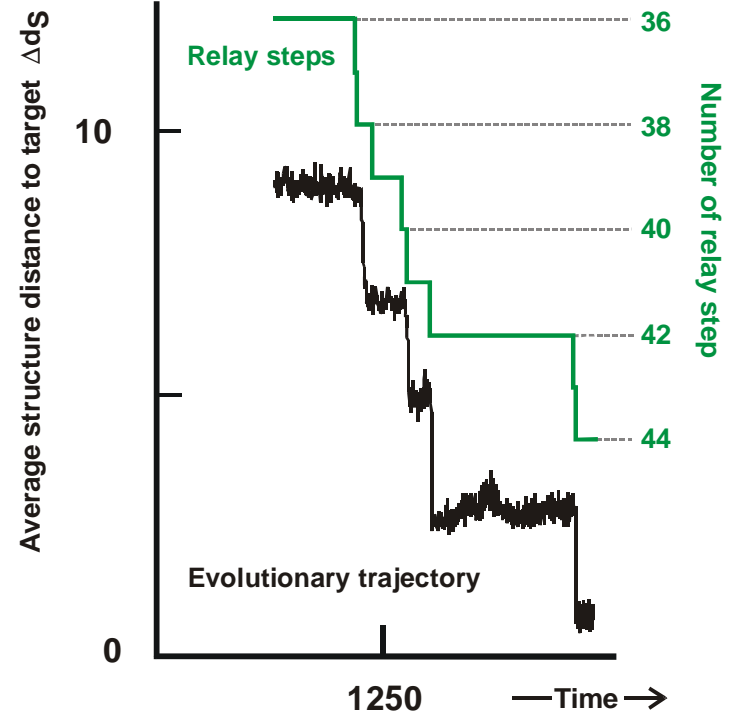
Final conformation of optimization



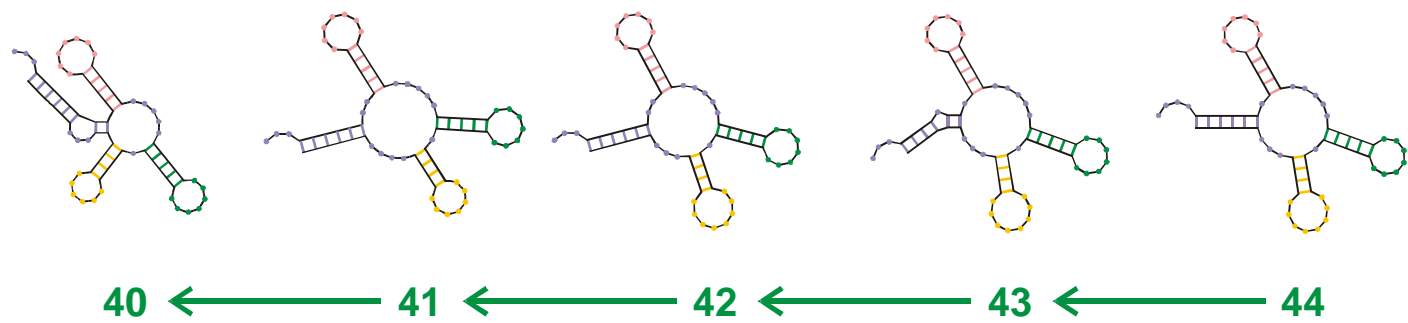
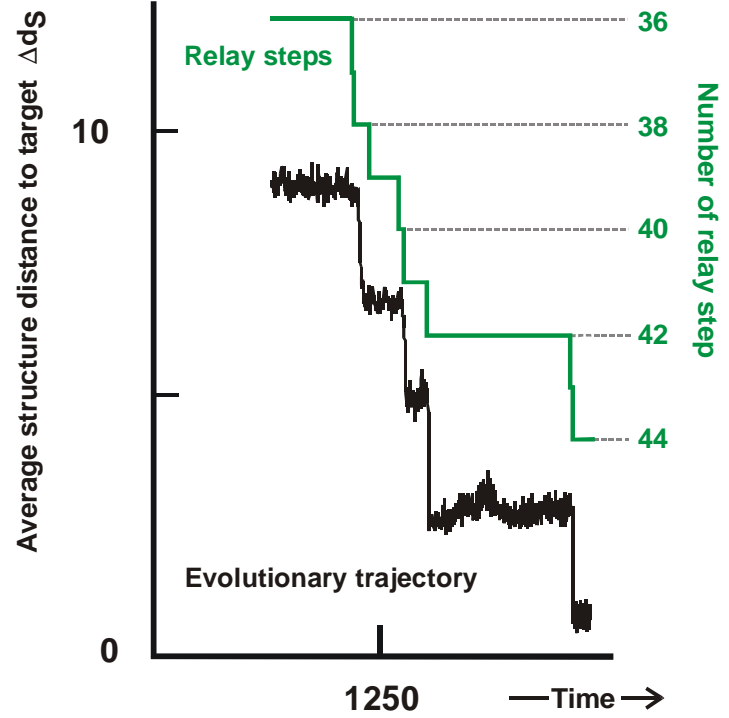
Reconstruction of the last step 43 → 44



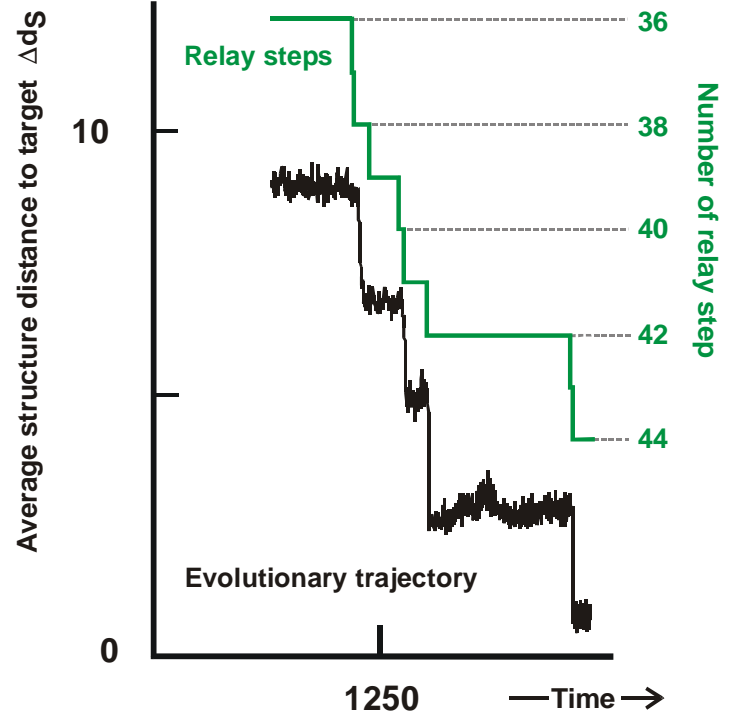
Reconstruction of last-but-one step 42 \rightarrow 43 (\rightarrow 44)



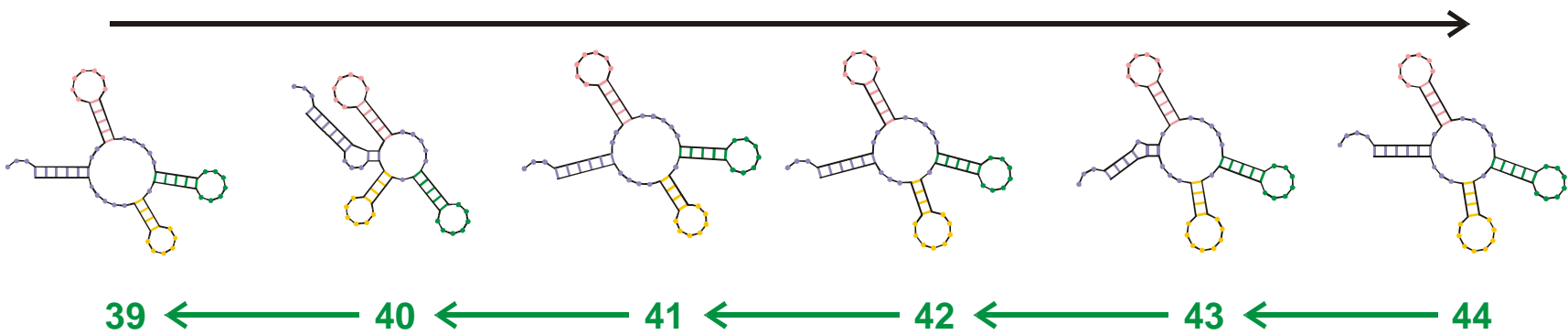
Reconstruction of step 41 \rightarrow 42 (\rightarrow 43 \rightarrow 44)



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



Evolutionary process



Reconstruction

Reconstruction of the relay series

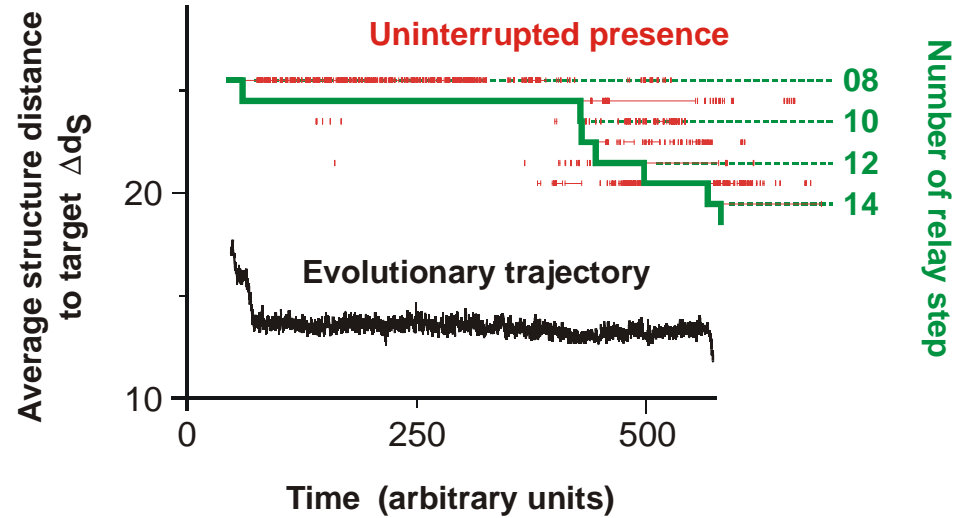
entry 39 GGGAUACAUGUGGCCCCUCAAGGCCCUAGCGAAACUGCUGCUGAAACCGUGUGAAUAAUCCGCACCCUGUCCCCGA
 (((((((.....((((.....))))).((((.....))))).(((.....))))).))))))...
 exit GGGAUAUACGAGGCCCGUCAAGGCCGUAAGCGAACGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCCGGG
 entry 40 GGGAUAUACGGGGGCCCGUCAAGGCCGUAAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCCGGG
 (((((((...((((.....))))).((((.....))))).(((.....))))).))))))...
 exit GGGAUAUACGGGGGCCCGUCAAGGCCGUAAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCCGGG
 entry 41 GGGAUAUACGGGGGCCCGUCAAGGCCGUAAGCGAACCGACUGUUGAGACUGUGCGAAUAAUCCGCACCCUGUCCCCGGG
 (((((((.....((((.....))))).((((.....))))).(((.....))))).))))))...
 exit GGGAUAUACGGGGCCCUUCAAGGCCAUAAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCCGGA
 entry 42 GGGAUAUACGGGGCCCUUCAAGGCCAUAAGCGAACCGACUGUUGAAACUGUGCGAAUAAUCCGCACCCUGUCCCCGGA
 (((((((...((((.....))))).((((.....))))).(((.....))))).))))))...
 exit GGGAUGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCGCUGAGCUUGUGCGACGUUUGUGCACCUGUCCCCGC
 entry 43 GGGAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCGCUGAGCUUGUGCGACGUUUGUGCACCUGUCCCCGC
 (((((((...((((.....))))).((((.....))))).(((.....))))).))))))...
 exit GGGAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCGCUGAGCUUGUGCGACGUUUGUGCACCUGUCCCCGC
 entry 44 GGGAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCGCUGAGCUUGUGCGACGUUUGUGCACCUGUCCCCGC
 (((((((...((((.....))))).((((.....))))).(((.....))))).))))))...)

Transition inducing point mutations

Neutral point mutations

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

28 neutral point mutations during a long quasi-stationary epoch

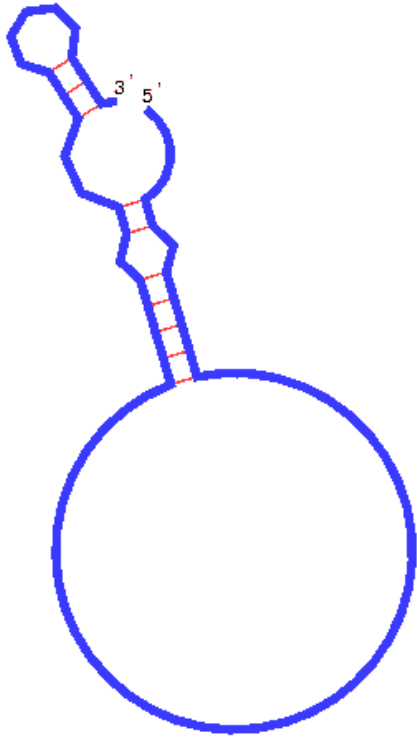


entry	GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGG	CAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8	.(((((((((((((. (((.))))))))(((((.))))))))	
exit	GGUAUGGGCGUUGAAUA	AJAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAU
entry	GGUAUGGGCGUUGAAUA	AUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAU
9	.((((((.(.(((((.))))))))(((((.))))))))	
exit	UGGAUGGACGUUGAAUAACA	AGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAC
entry	UGGAUGGACGUUGAAUAACA	AGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAC
10	.(((((.(((((.))))))))(((((.))))))))	
exit	UGGAUGGACGUUGAAUAACA	AGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG

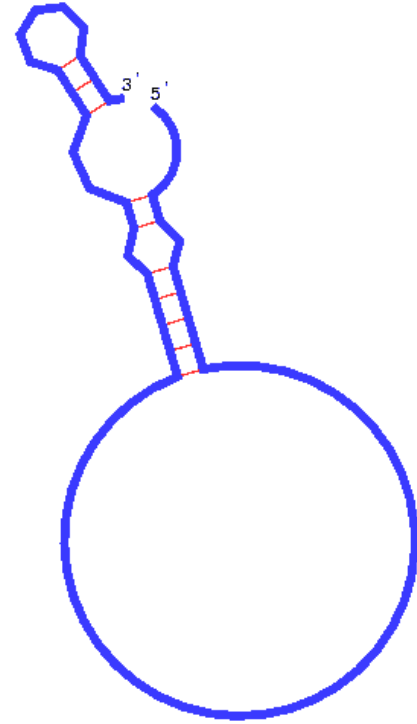
Transition inducing point mutations

Neutral point mutations

Neutral genotype evolution during phenotypic stasis

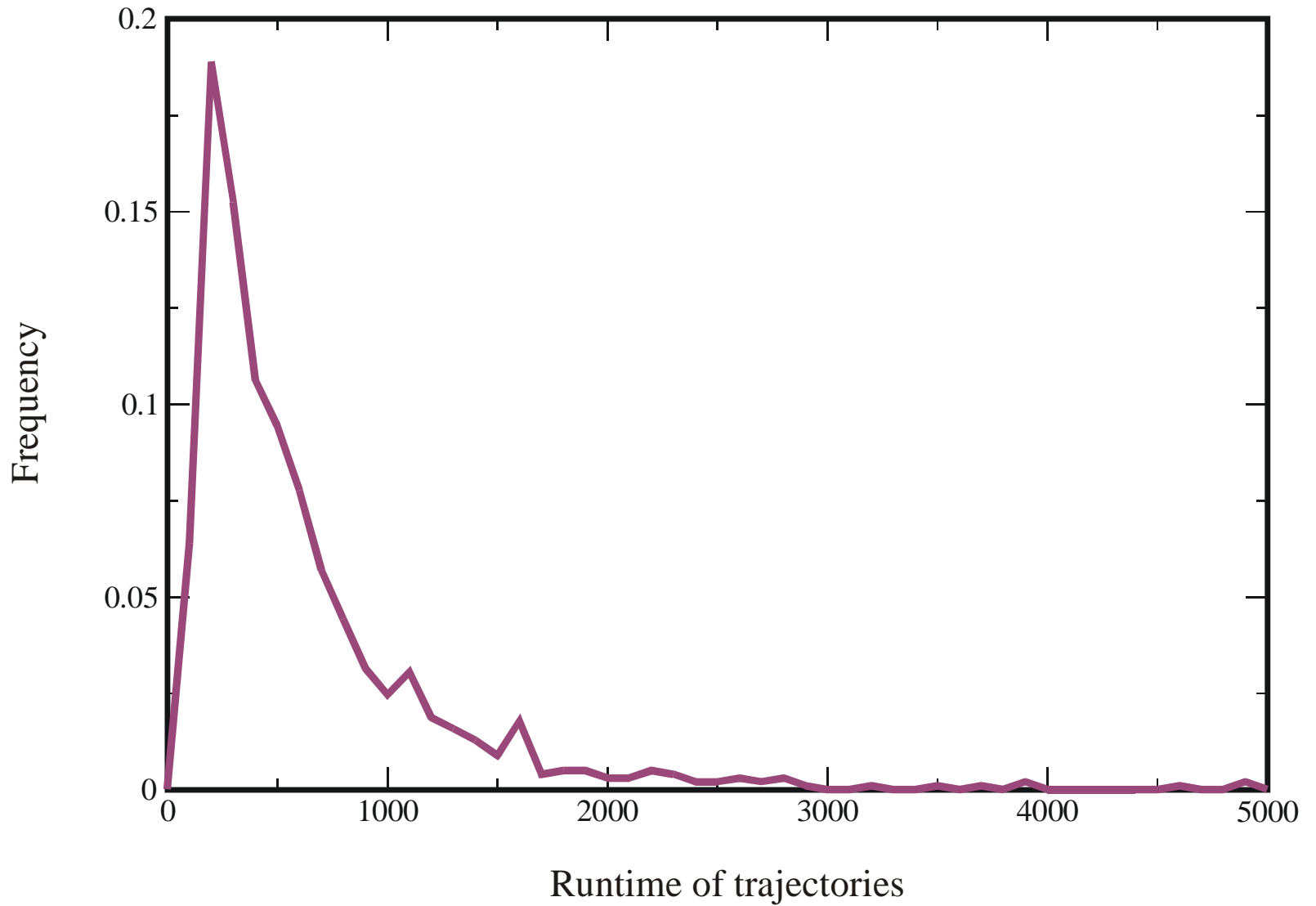


AUGC



GC

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



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

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

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

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

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

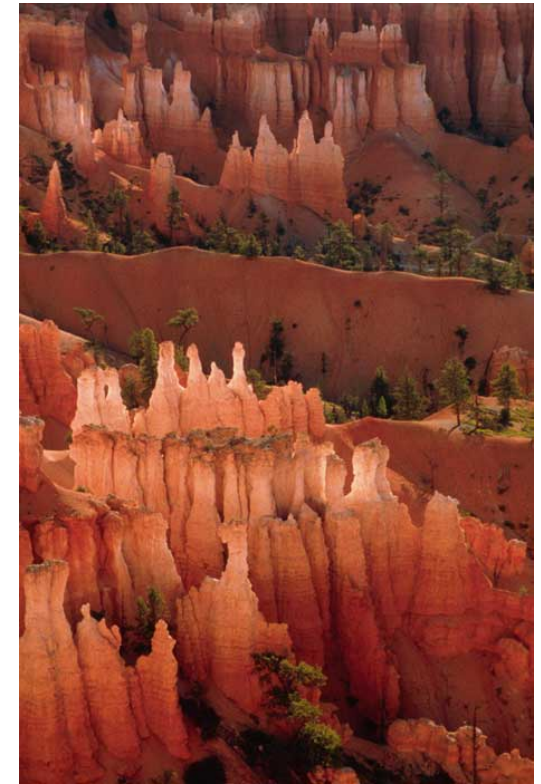


Mount Fuji

Example of a smooth landscape on Earth

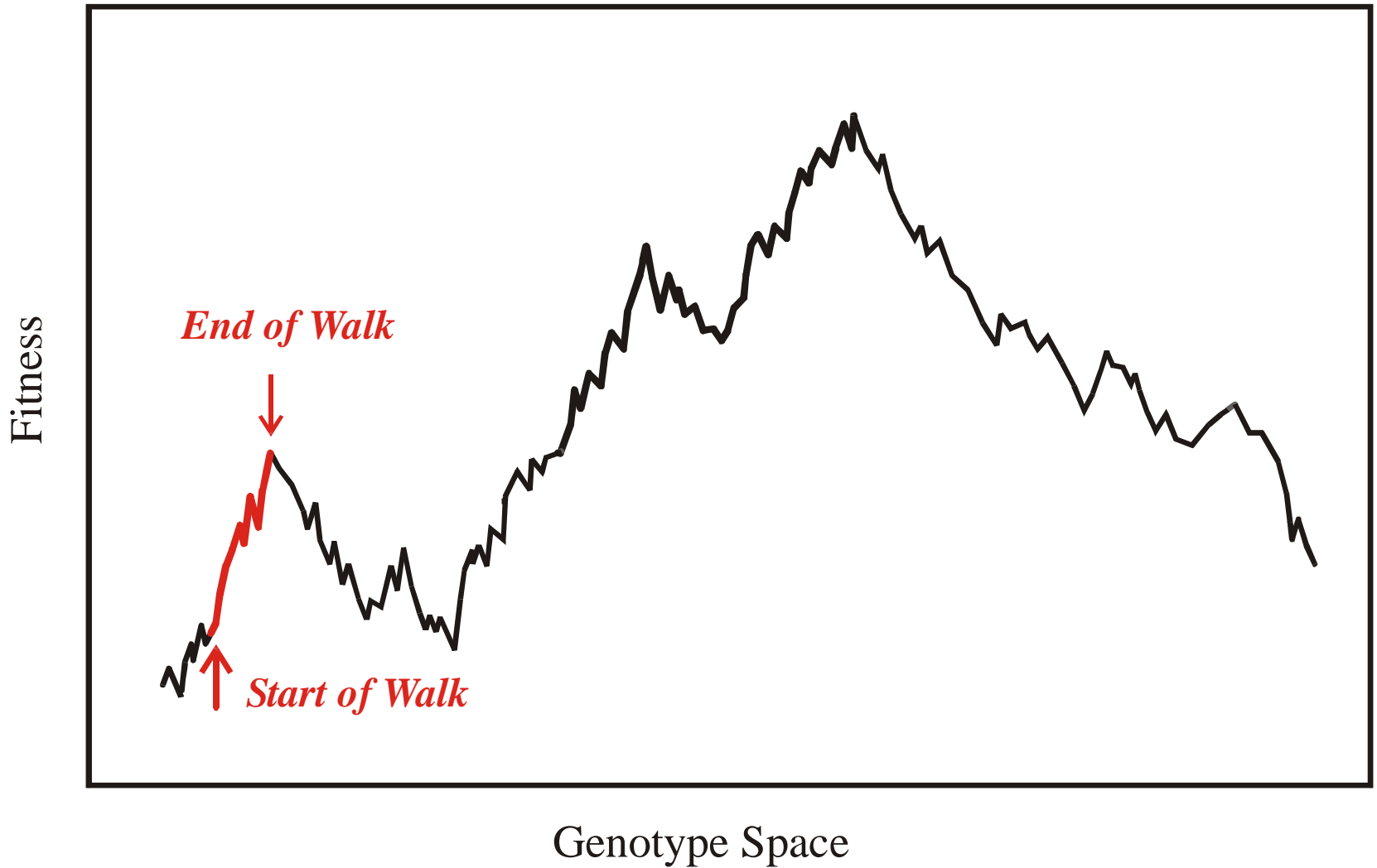


Dolomites

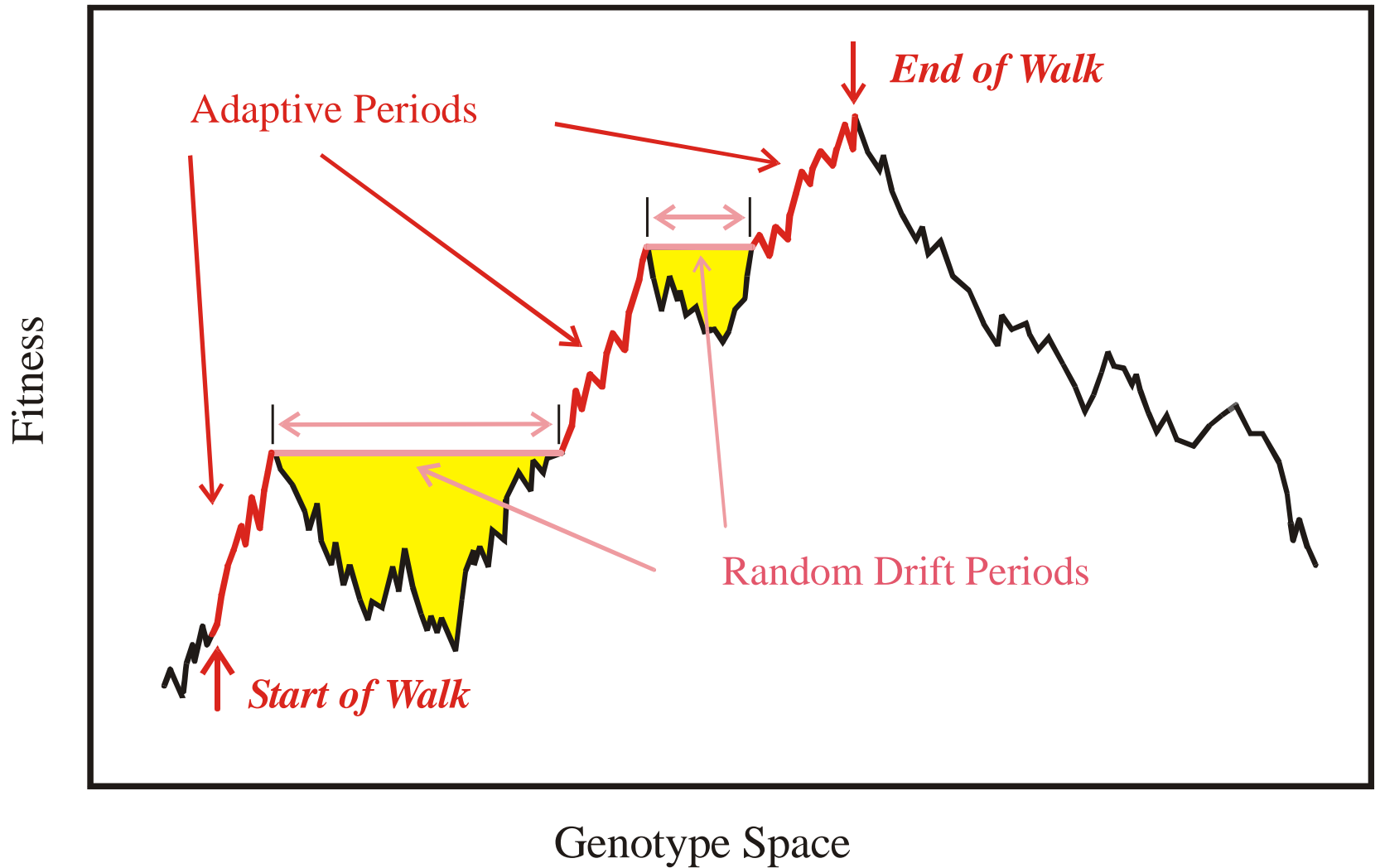


Bryce Canyon

Examples of rugged landscapes on Earth



Evolutionary optimization in absence of neutral paths in sequence space

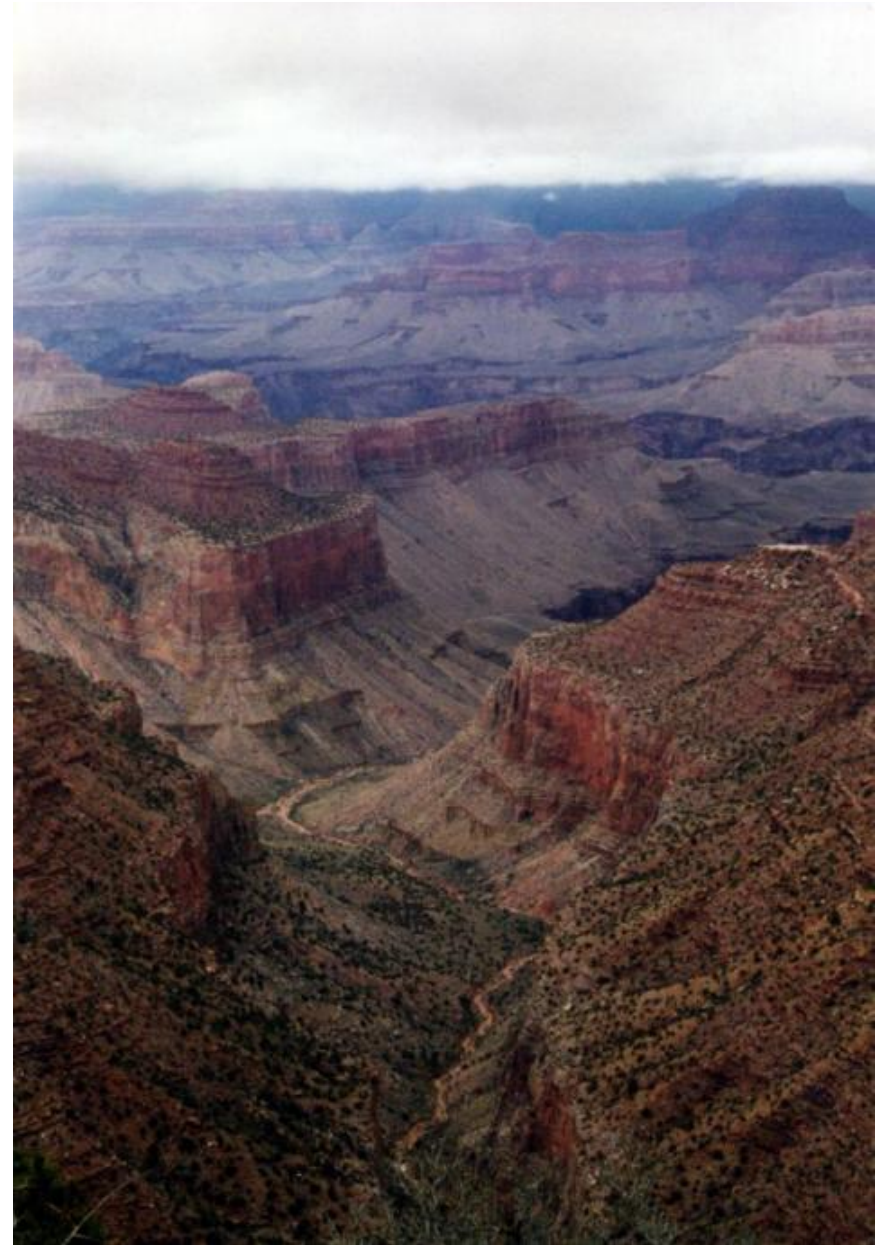


Evolutionary optimization including neutral paths in sequence space

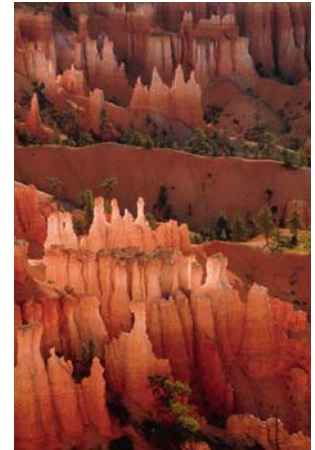


Grand Canyon

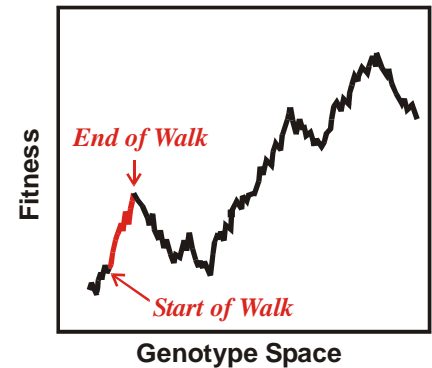
Example of a landscape on Earth with 'neutral' ridges and plateaus



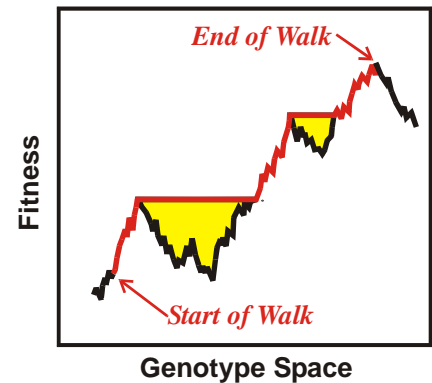
- Conformational and mutational landscapes of biomolecules as well as fitness landscapes of evolutionary biology are **rugged**.



- **Adaptive** or **non-descending walks** on rugged landscapes end commonly at one of the low lying local maxima.



- Selective neutrality in the form of **neutral networks** plays an active role in evolutionary optimization and enables populations to reach high local maxima or even the global optimum.



Evolutionary optimization of RNA structures consists of short adaptive phases and long quasi-stationary epochs!

Neutral networks of RNA structures are essential for success and efficiency of optimization!

1. From working copy in molecular genetics to riboswitches
2. The never-ending story of RNA structure
3. Neutral networks, intersections, and bistable RNA molecules
4. Evolution of molecules *in silico*
- 5. A few selected experiments**





- 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).

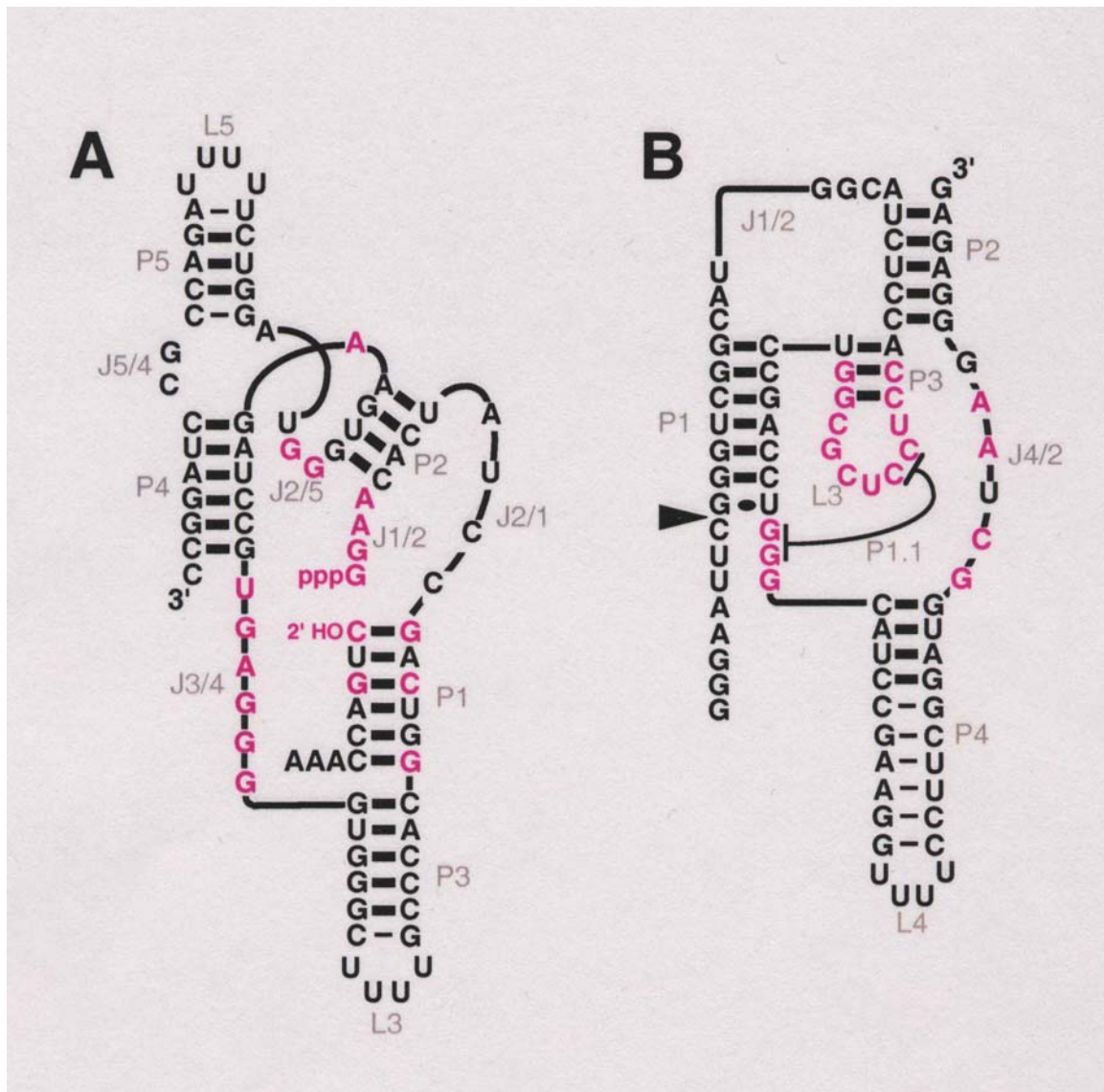
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

A ribozyme switch

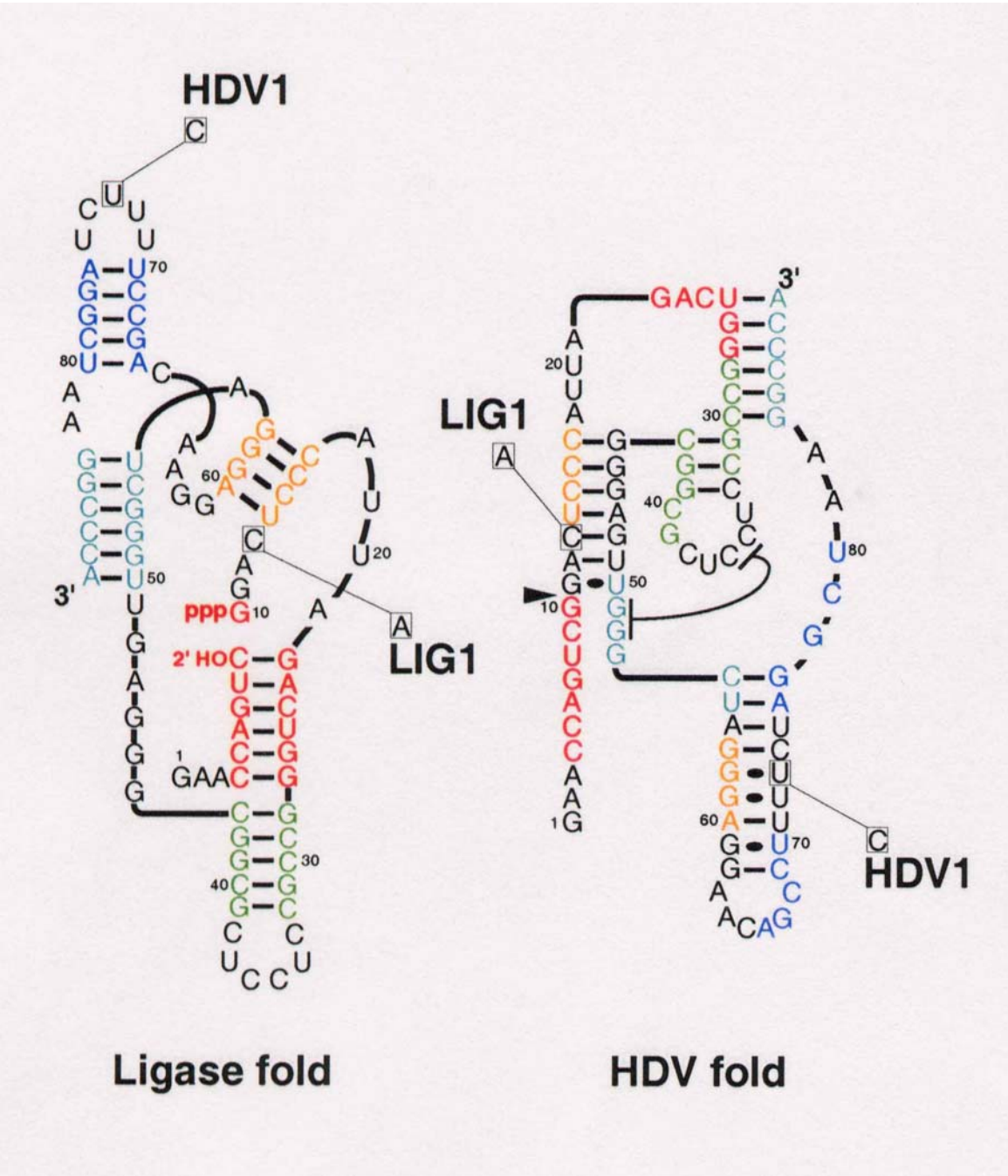
E.A.Schultes, D.B.Bartel, *Science*
289 (2000), 448-452

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

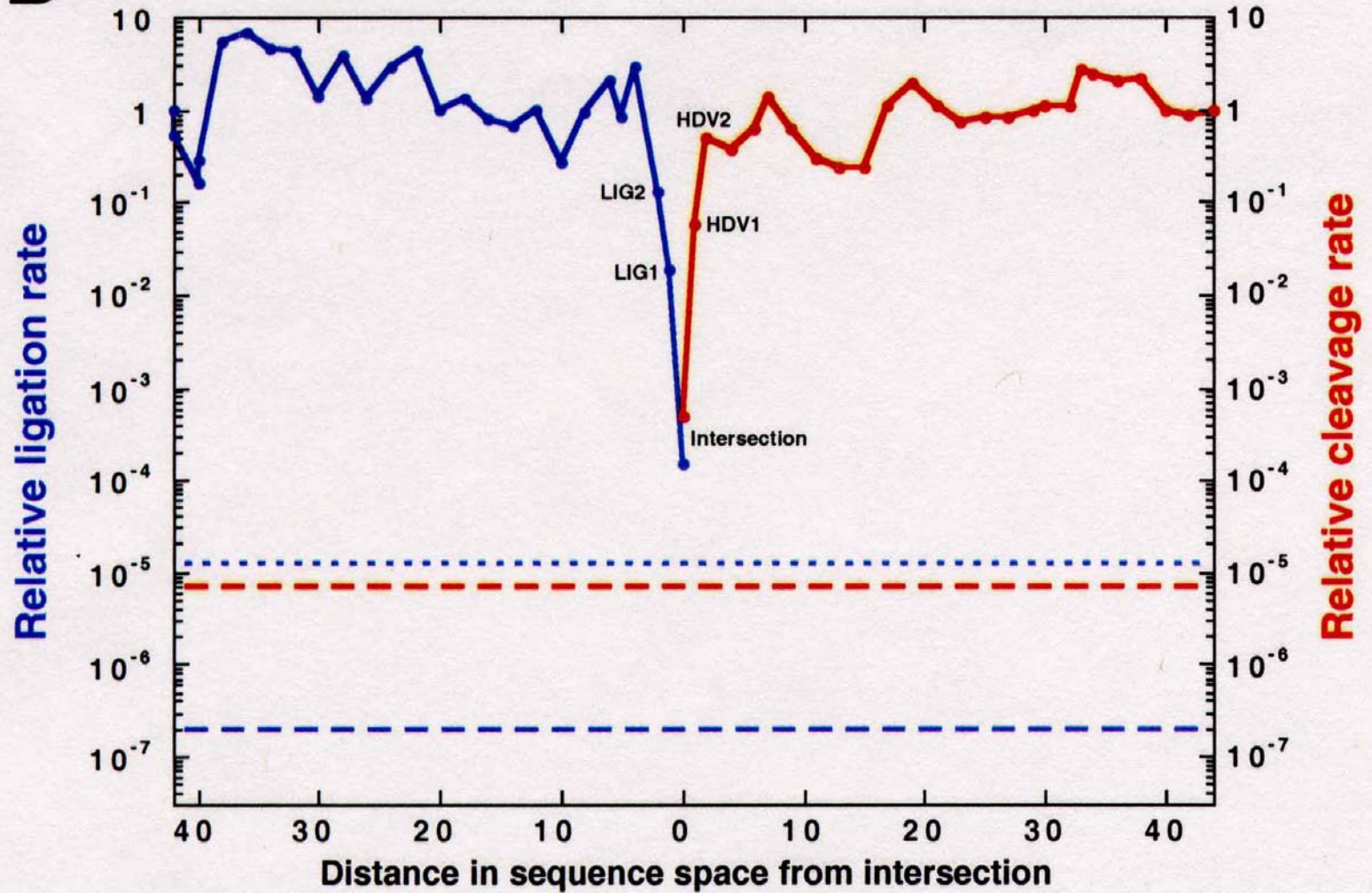


Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis- δ -virus (**B**)

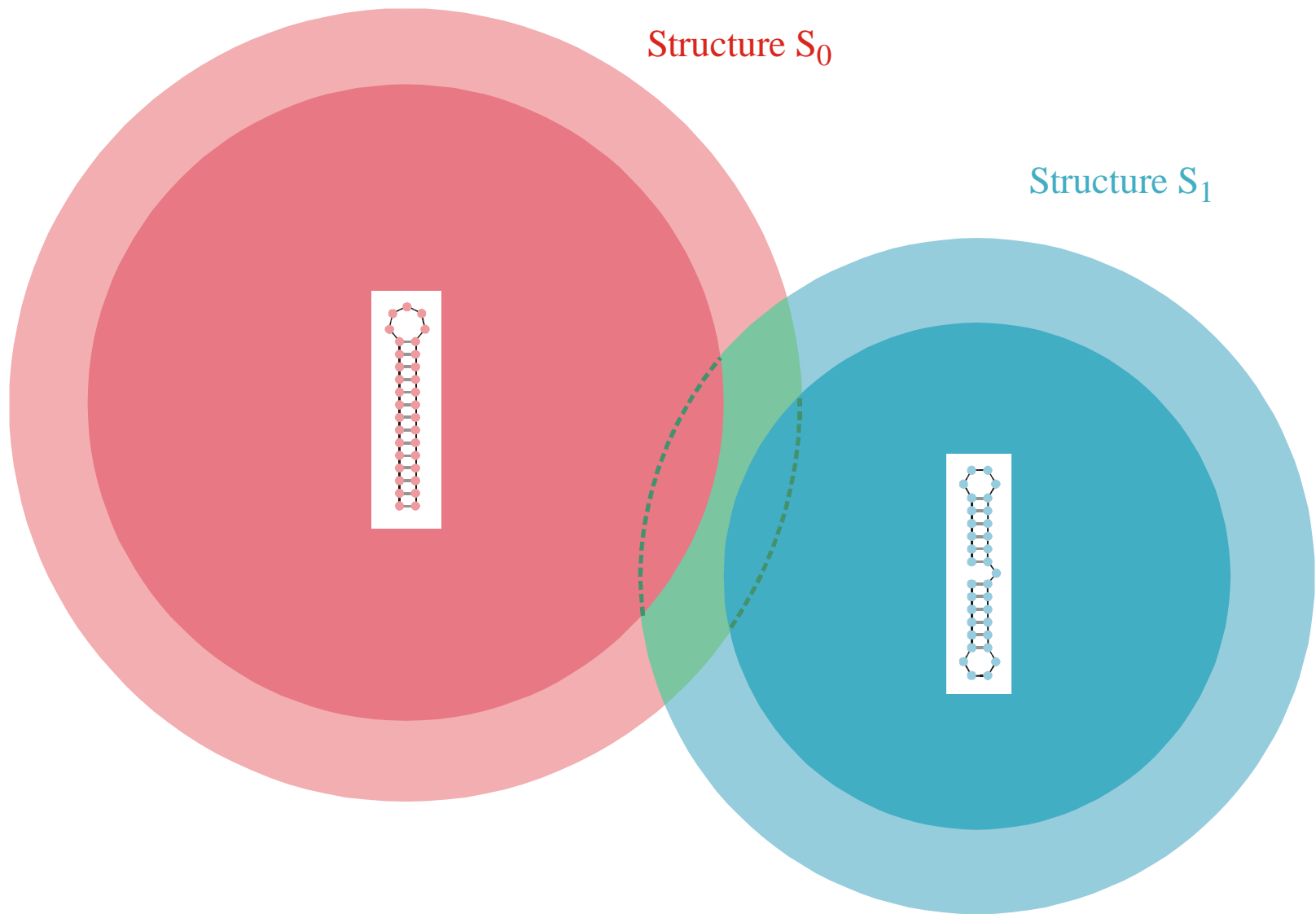


The sequence at the *intersection*:

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



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

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

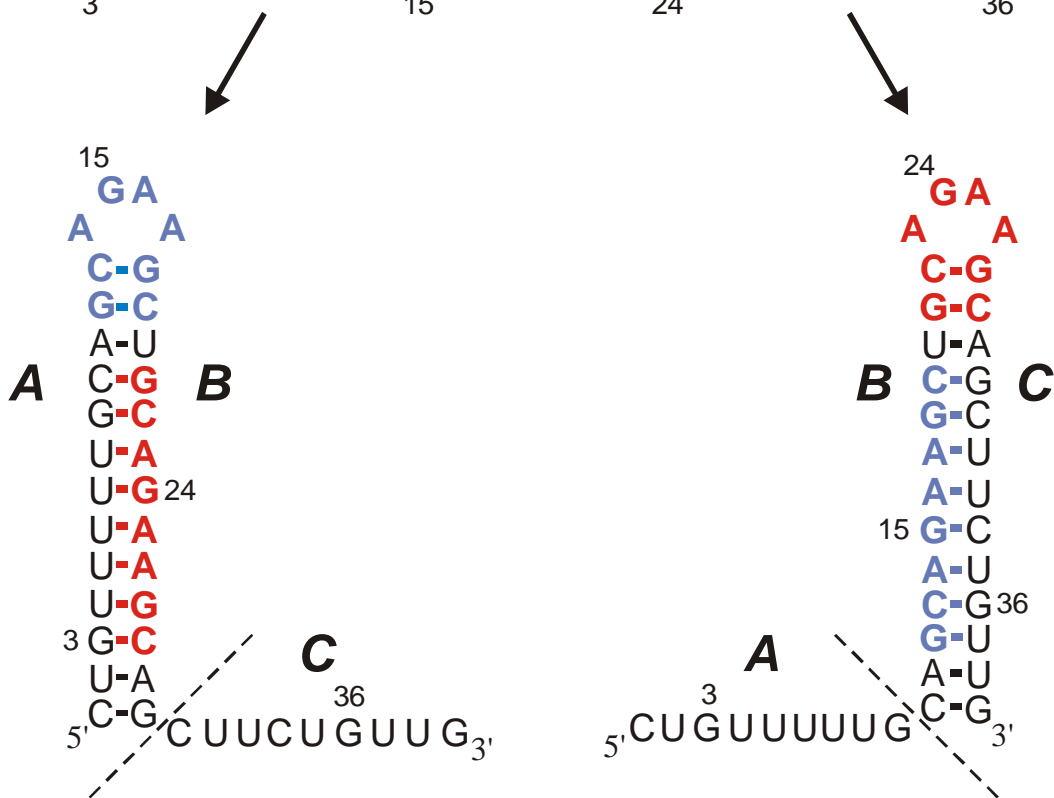
J. H. A. Nagel, C. Flamm, I. L. Hofacker, K. Franke, M. H. de Smit, P. Schuster, and C. W. A. Pleij. *Structural parameters affecting the kinetic competition of RNA hairpin formation*, in press 2004.

J. H. A. Nagel, J. Møller-Jensen, C. Flamm, K. J. Öistämö, J. Besnard, I. L. Hofacker, A. P. Gulyaev, 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

CUGUUUUUGCAG**GCAGAAGCU****GCAGAAGC**AGCUUCUGUUG

3 15 24 36



-19.5 kcal·mol⁻¹

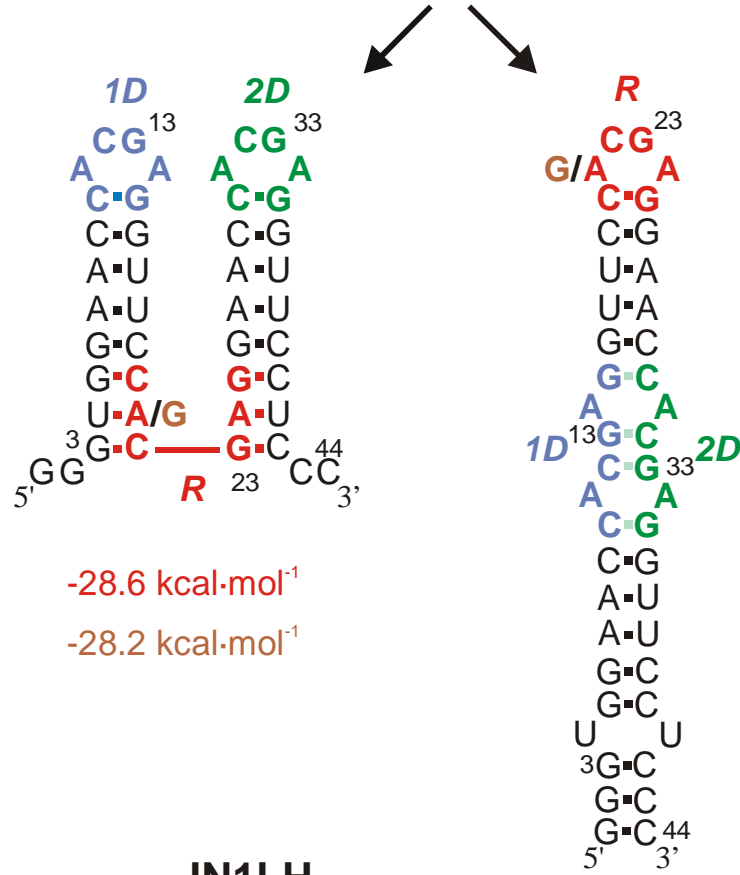
-21.9 kcal·mol⁻¹

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.*

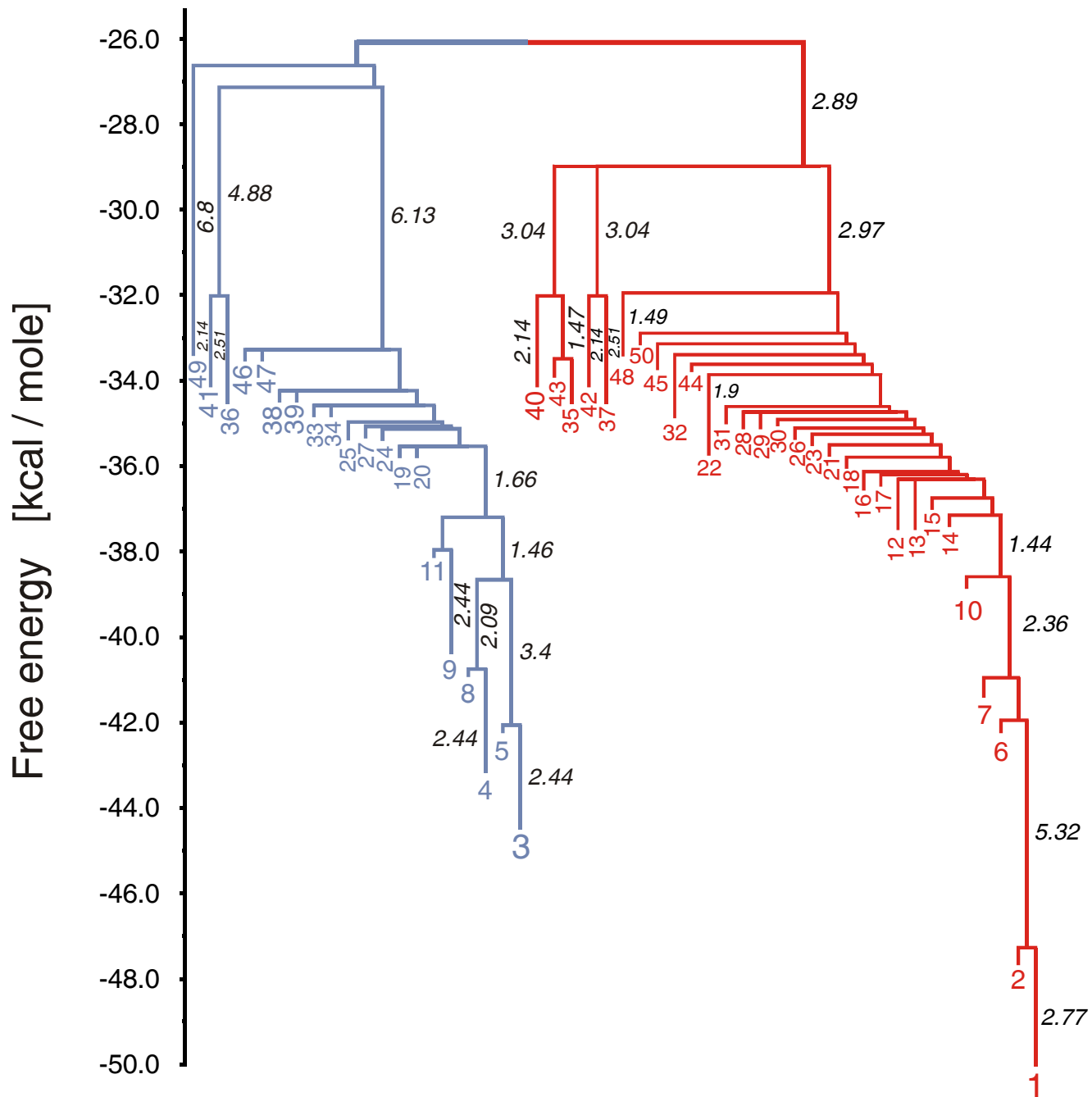
1D **R** 2D
 GGGUGGAAC**CACGAG**GUUC**CACGAG**GAACC**CACGAG**GUUCCUCCC
 3 13 **G** 23 33 44



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



Riboswitches

Jord H. A. Nagel and Cornelius W. A. Pleij. *Self-induced structural switches in RNA*. Biochimie **84** (2002), 913-923

Wade Winkler, Ali Nahvi, and Ronald R. Breaker. *Thiamine derivatives bind messenger RNA directly to regulate bacterial gene expression*. Nature **419** (2002), 952-956

Ronald Micura and Claudia Höbartner. *On Secondary Structure Rearrangements and Equilibria of Small RNAs*. Nature **419** (2002), 952-956

Alexey G. Vitreschak, Dimitry A. Rodionov, Andrey A. Mironov, and Mikhail S. Gefland. *Riboswitches: The oldest mechanism for the regulation of gene expression?* Trends in Genetics **20** (2004), 44-50

Jeffrey E. Barrick, Keith A. Corbino, Wade C. Winkler, Ali Nahvi, Maumita Mandal, Jennifer Collins, Mark Lee, Adam Roth, Narasimhan Sundarasan, Inbal Jona, J. Kenneth Wickiser, and Ronald R. Breaker. *New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control*. Proc.Natl.Acad.Sci.USA **101** (2004), 6421-6426

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

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

A ribozyme that lacks cytidine

Jeff Rogers & Gerald F. Joyce

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

.....
The RNA-world hypothesis proposes that, before the advent of DNA and protein, life was based on RNA, with RNA serving as both the repository of genetic information and the chief agent of catalytic function¹. An argument against an RNA world is that the components of RNA lack the chemical diversity necessary to sustain life. Unlike proteins, which contain 20 different amino-acid subunits, nucleic acids are composed of only four subunits which have very similar chemical properties. Yet RNA is capable of a broad range of catalytic functions²⁻⁷. Here we show that even three nucleic-acid subunits are sufficient to provide a substantial increase in the catalytic rate. Starting from a molecule that contained roughly equal proportions of all four nucleosides, we used *in vitro* evolution to obtain an RNA ligase ribozyme that lacks cytidine. This ribozyme folds into a defined structure and has a catalytic rate that is about 10⁵-fold faster than the uncatalysed rate of template-directed RNA ligation.

Catalytic activity in the
AUG alphabet

A ribozyme composed of only two different nucleotides

John S. Reader & Gerald F. Joyce

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

RNA molecules are thought to have been prominent in the early history of life on Earth because of their ability both to encode genetic information and to exhibit catalytic function¹. The modern genetic alphabet relies on two sets of complementary base pairs to store genetic information. However, owing to the chemical instability of cytosine, which readily deaminates to uracil², a primitive genetic system composed of the bases A, U, G and C may have been difficult to establish. It has been suggested that the first genetic material instead contained only a single base-pairing unit³⁻⁷. Here we show that binary informational macromolecules, containing only two different nucleotide subunits, can act as catalysts. *In vitro* evolution was used to obtain ligase ribozymes composed of only 2,6-diaminopurine and uracil nucleotides, which catalyse the template-directed joining of two RNA molecules, one bearing a 5'-triphosphate and the other a 3'-hydroxyl. The active conformation of the fastest isolated ribozyme had a catalytic rate that was about 36,000-fold faster than the uncatalysed rate of reaction. This ribozyme is specific for the formation of biologically relevant 3',5'-phosphodiester linkages.

Catalytic activity in the
DU alphabet

*Theory and experiment are successful when they go
hand-in-hand!*

*RNA has still more surprises waiting to be discovered
in vivo and in vitro!*

Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank

Project No. Nat-7813

European Commission: Project No. EU-980189

Siemens AG, Austria

Universität Wien

The software for producing RNA movies was developed by
Robert Giegerich and coworkers at the Universität Bielefeld



Universität Wien

Coworkers



Universität Wien

Walter Fontana, Santa Fe Institute, NM

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

Peter Stadler, Bärbel Stadler, Universität Leipzig, GE

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

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>

