Kinetic Folding of RNA and the Design of Molecules with Predefined Secondary Structures

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Gunnar and Gunnel Källén Memorial Lecture

Lund University, 10.– 11.05.2004

Web-Page for further information:

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

RNA *as transmitter of genetic information*

DNA

↓ transcription

...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...

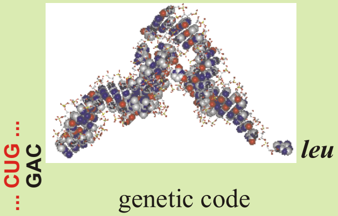
messenger-**RNA**

↓ translation

protein

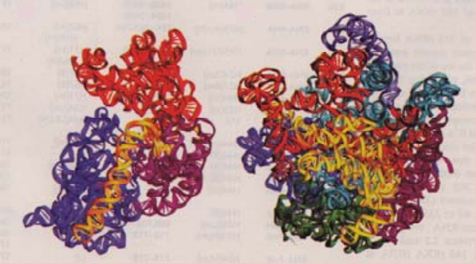
RNA as **working copy** of genetic information

RNA *as adapter molecule*



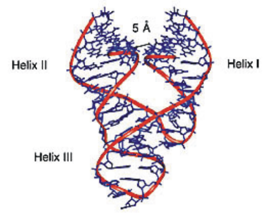
genetic code

RNA *is the catalytic subunit in supramolecular complexes*



The ribosome is a ribozyme !

RNA *as catalyst*



ribozyme

RNA

RNA *is modified by epigenetic control*

RNA editing

Alternative splicing of messenger **RNA**

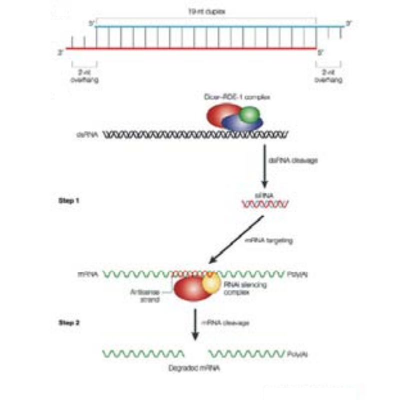
The RNA *world as a precursor of the current* DNA + protein *biology*

RNA *as carrier of genetic information*

RNA viruses and retroviruses

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

RNA *as regulator of gene expression*

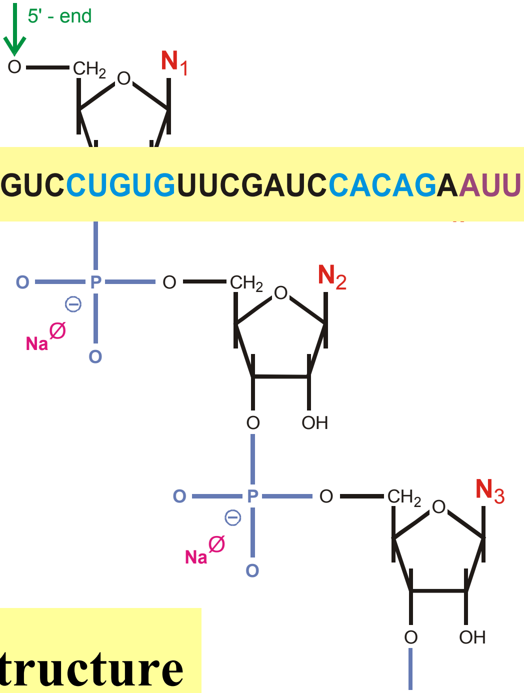


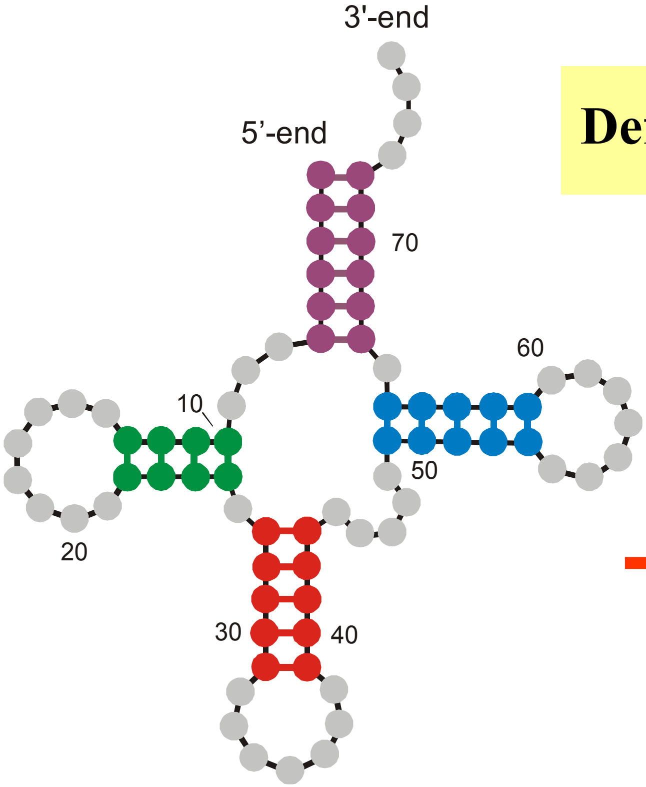
gene silencing by small interfering RNAs

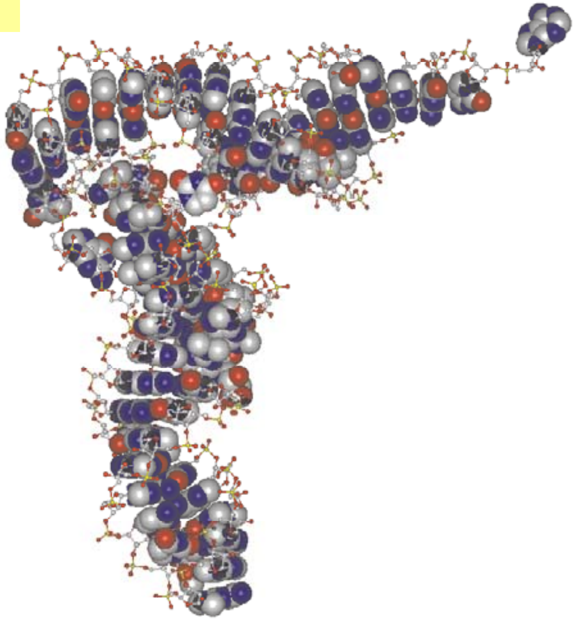
Functions of RNA molecules

Definition of RNA structure

5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





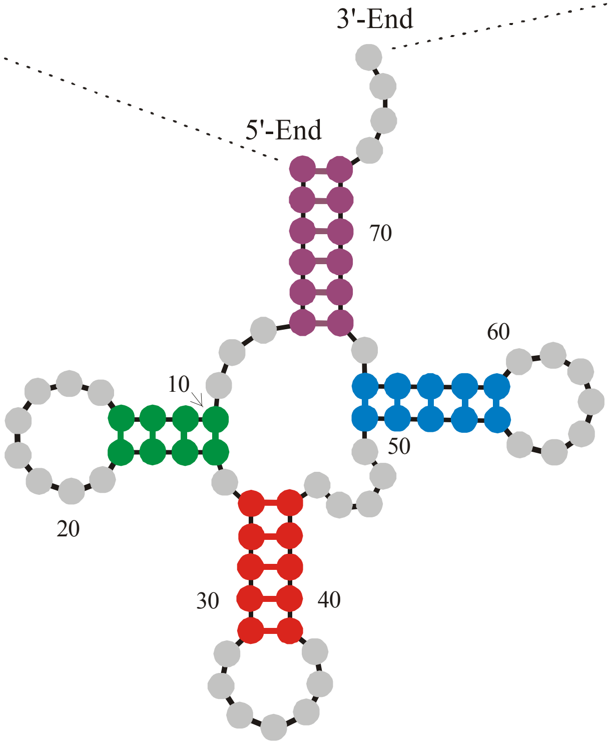


Sequence

5'-End 3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure

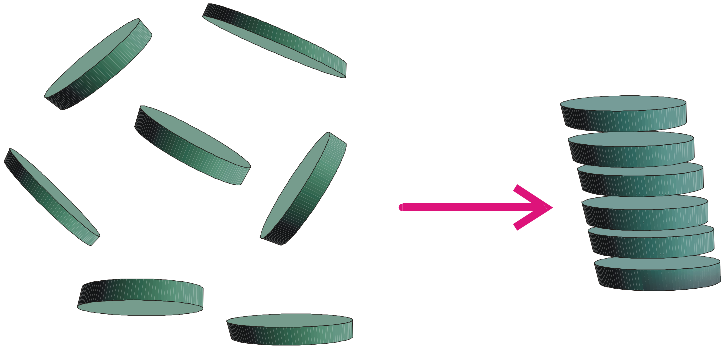


Definition and **physical relevance** of RNA secondary structures

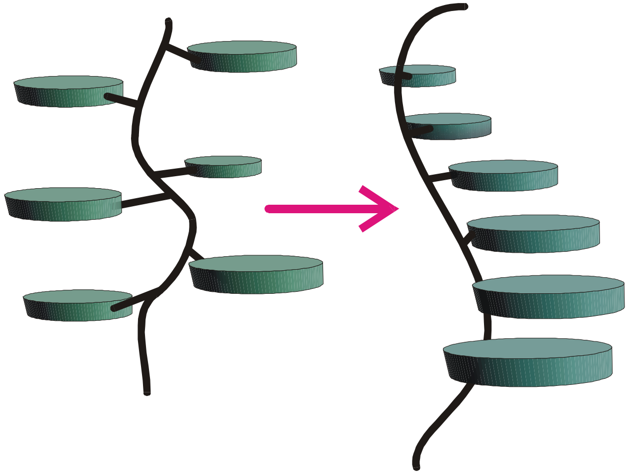
RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots.

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

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

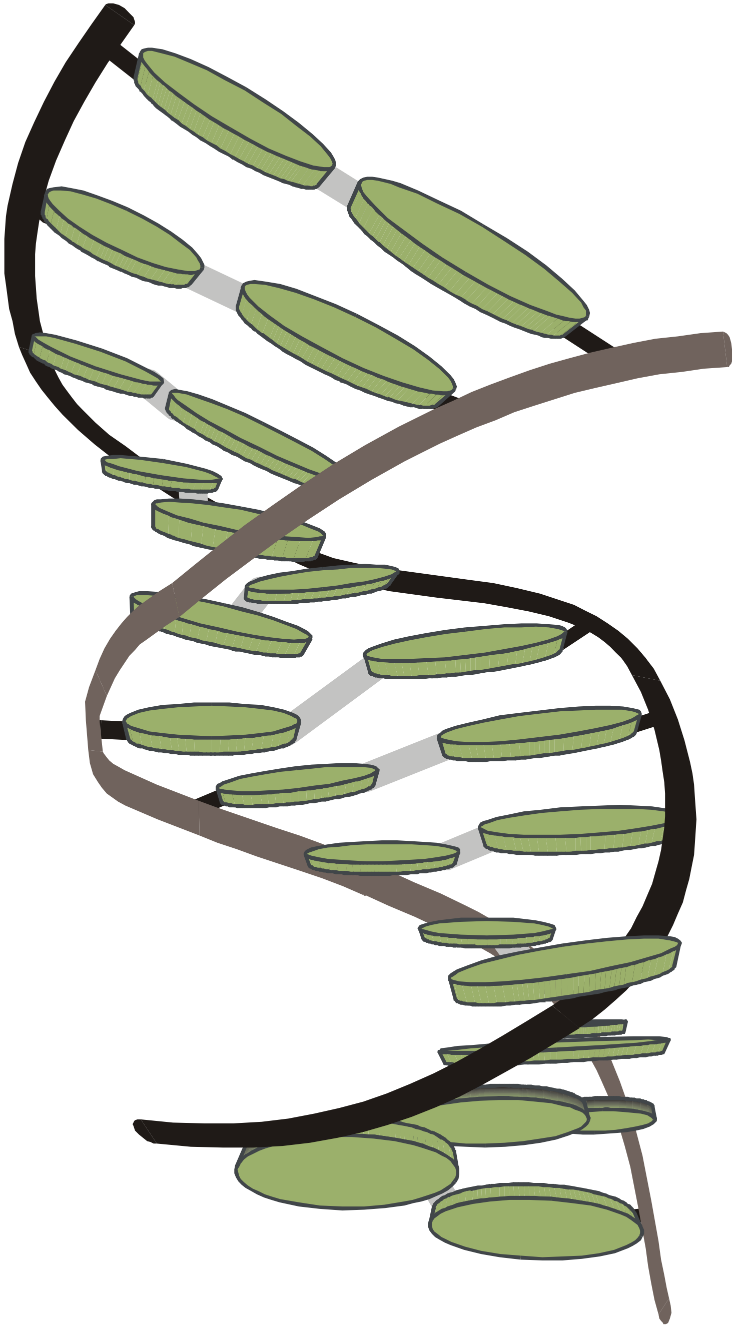


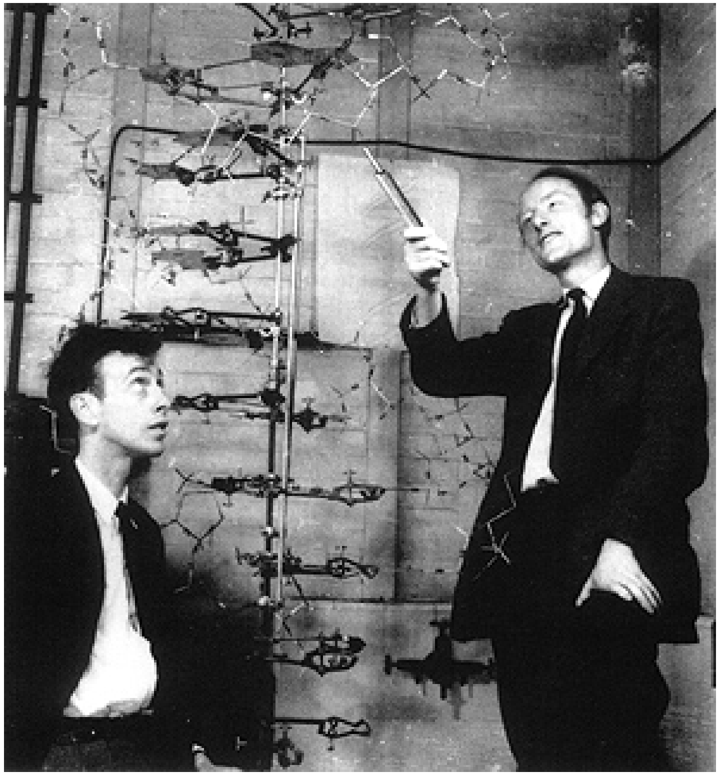
Stacking of free nucleobases or other planar heterocyclic compounds (N6,N9-dimethyl-adenine)



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

Stacking of nucleic acid single strands (poly-A)





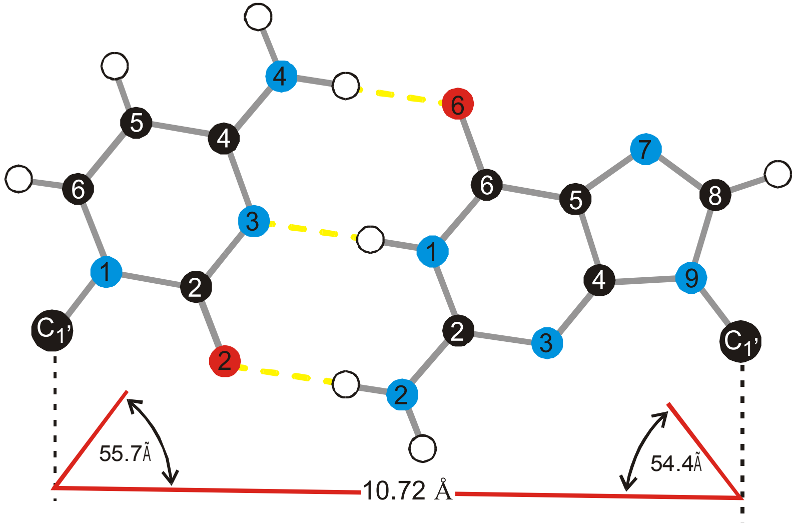
James D. Watson and Francis H.C. Crick

Nobel prize 1962

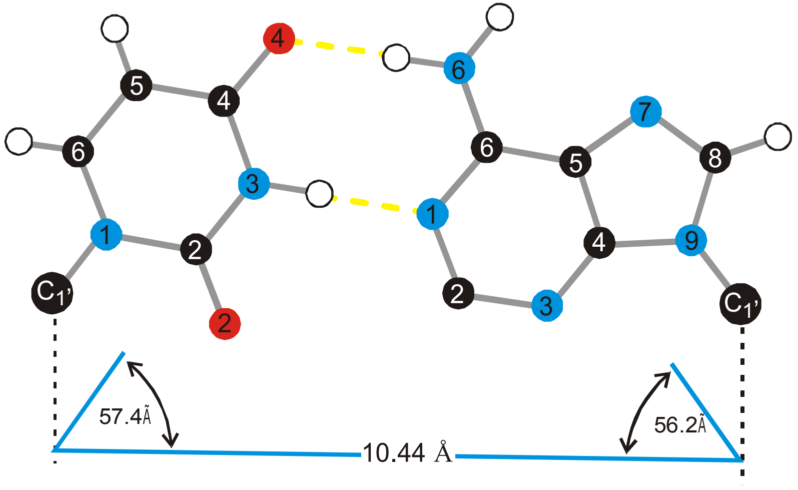
1953 – 2003 fifty years double helix

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

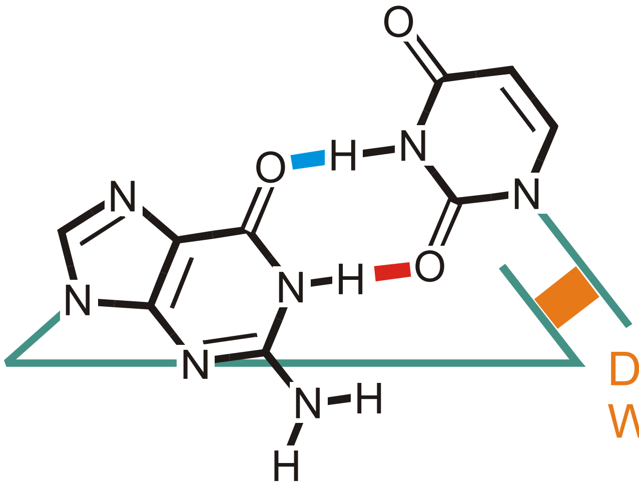
C Í G



U = A

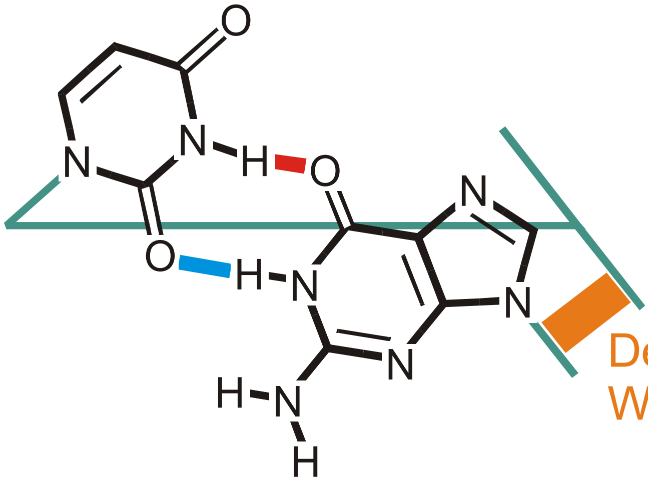


Watson-Crick type base pairs



G=U

Deviation from
Watson-Crick geometry



U=G

Deviation from
Watson-Crick geometry

Wobble base pairs

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

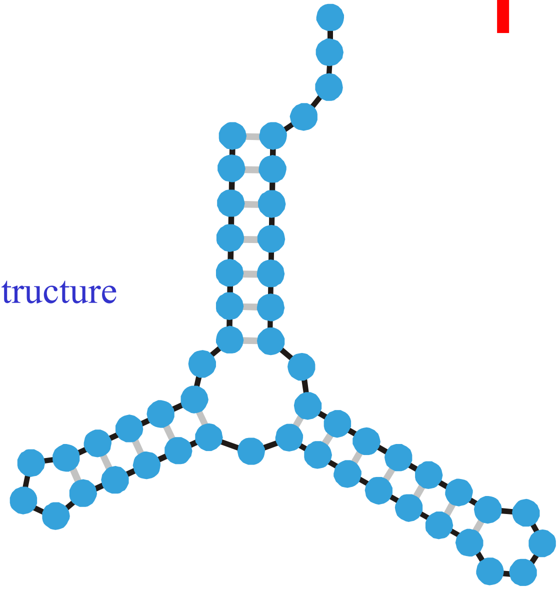
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

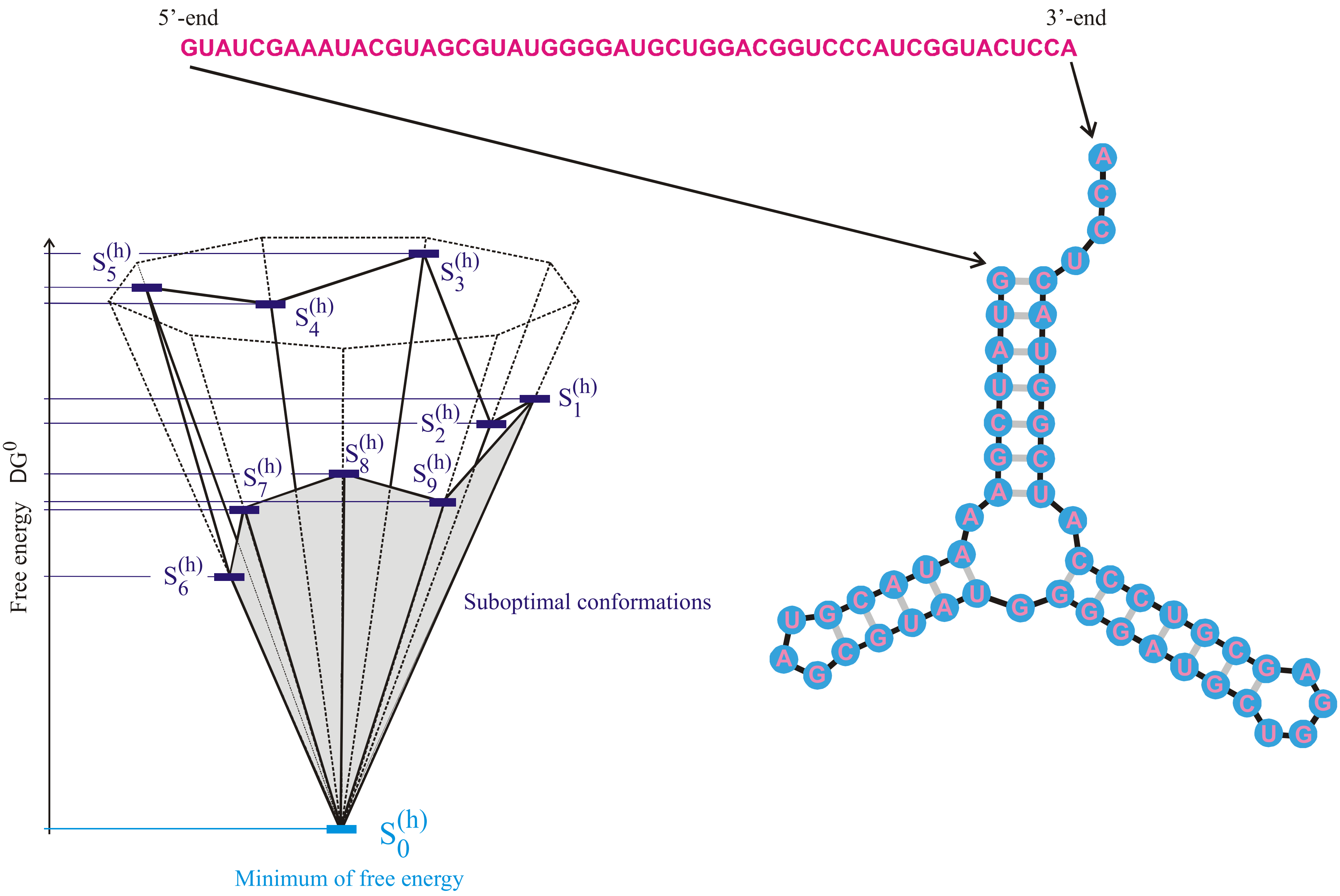
Inverse folding of RNA:

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

RNA structure



Sequence, structure, and function



The minimum free energy structures on a discrete space of conformations

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)

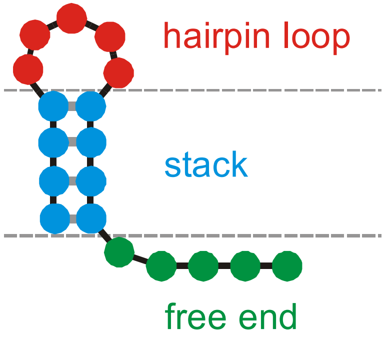
The **Vienna RNA Package** provides in addition: inverse folding (computing sequences for given secondary structures), computation of melting profiles from partition functions, all suboptimal structures within a given energy interval, barrier tress of suboptimal structures, kinetic folding of RNA sequences,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

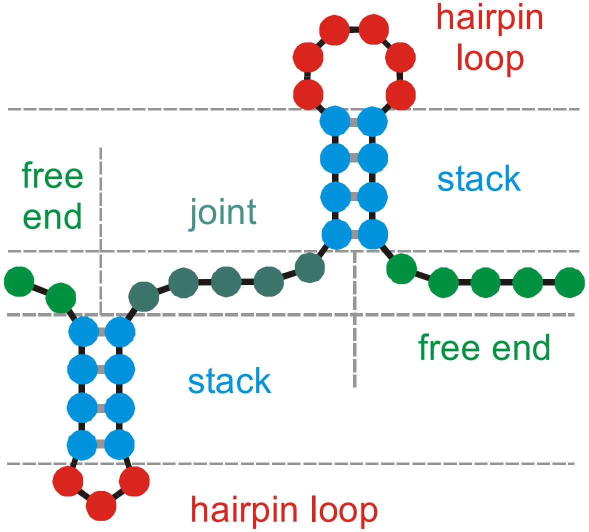
I.L.Hofacker, W. Fontana, P.F.Stadler, L.S.Bonhoeffer, M.Tacker, and P. Schuster. *Mh.Chem*. **125**:167-188 (1994)

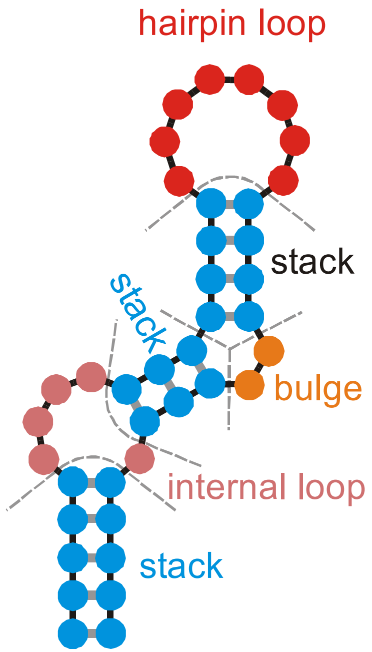
S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. *Biopolymers* **49**:145-165 (1999)

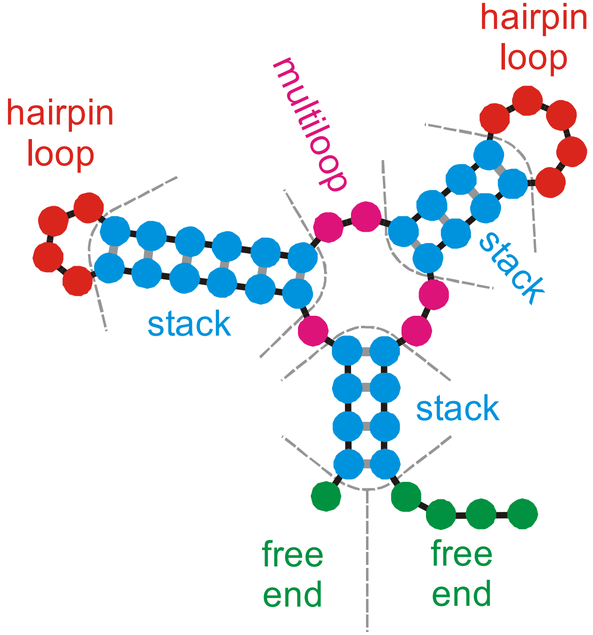
C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. *RNA* **6**:325-338 (1999)

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





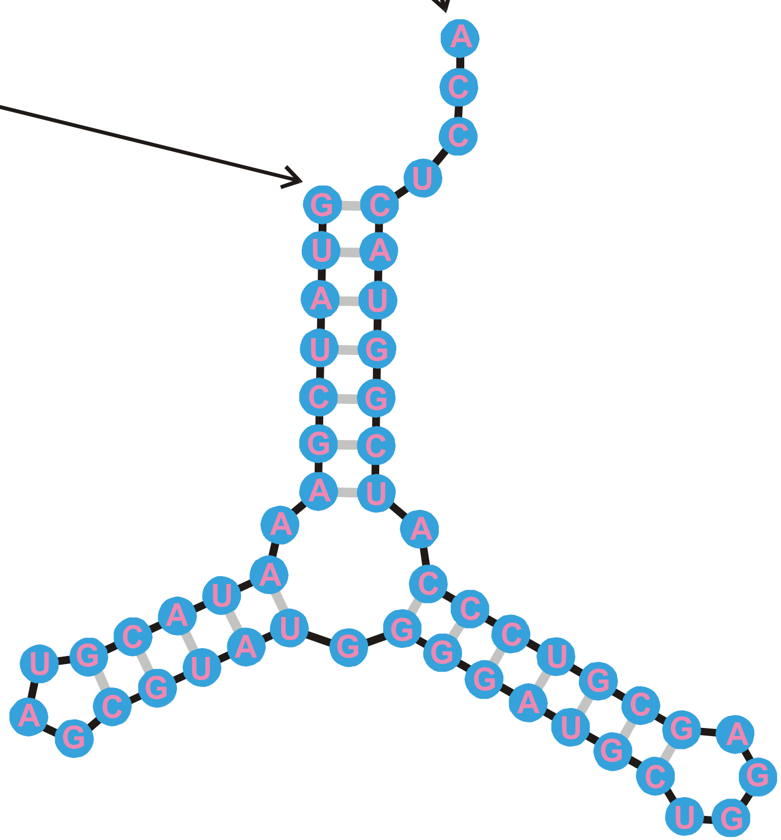


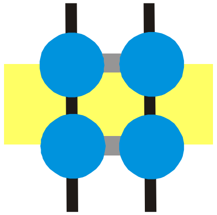


Elements of RNA secondary structures
as used in free energy calculations

5’-end 3’-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA





free energy of stacking < 0

$$\Delta G_0^{300} = \underbrace{\sum g_{ij,kl}}_{\text{stacks of base pairs}} + \underbrace{\sum h(n_l)}_{\text{hairpin loops}} + \underbrace{\sum b(n_b)}_{\text{bulges}} + \underbrace{\sum i(n_i)}_{\text{internal loops}} + \cdots$$

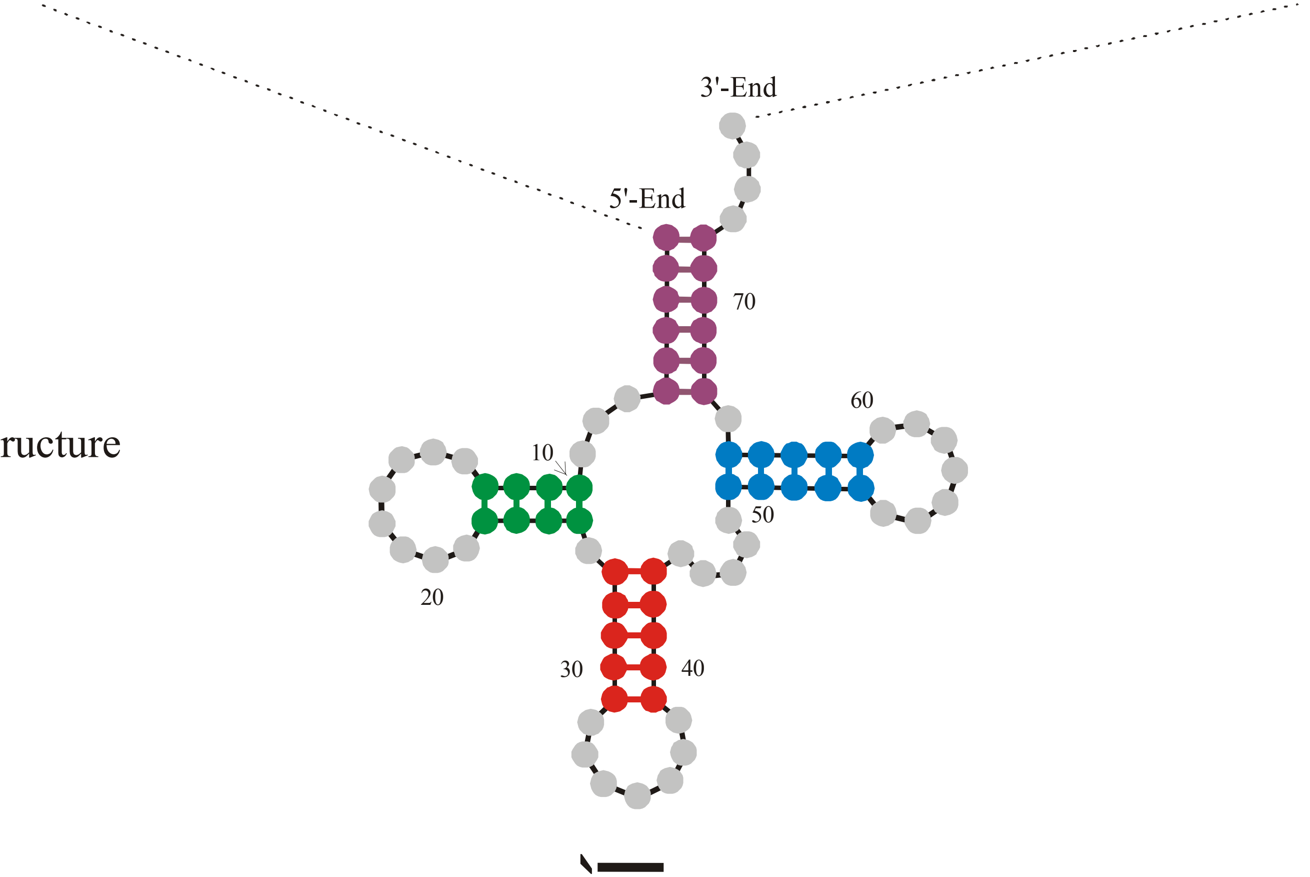
Folding of RNA sequences into secondary structures of minimal free energy, DG_0^{300}

Sequence

5'-End 3'-End

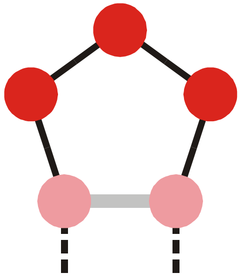
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure



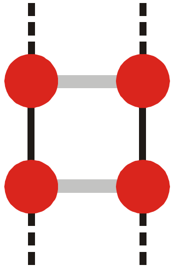
Symbolic notation 5'-End ((((((···((((········))))·(((((·······)))))·····(((((·······)))))·))))))···· 3'-End

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



Minimal hairpin loop size:

n_{lp} Æ 3



Minimal stack length:

n_{st} Æ 2

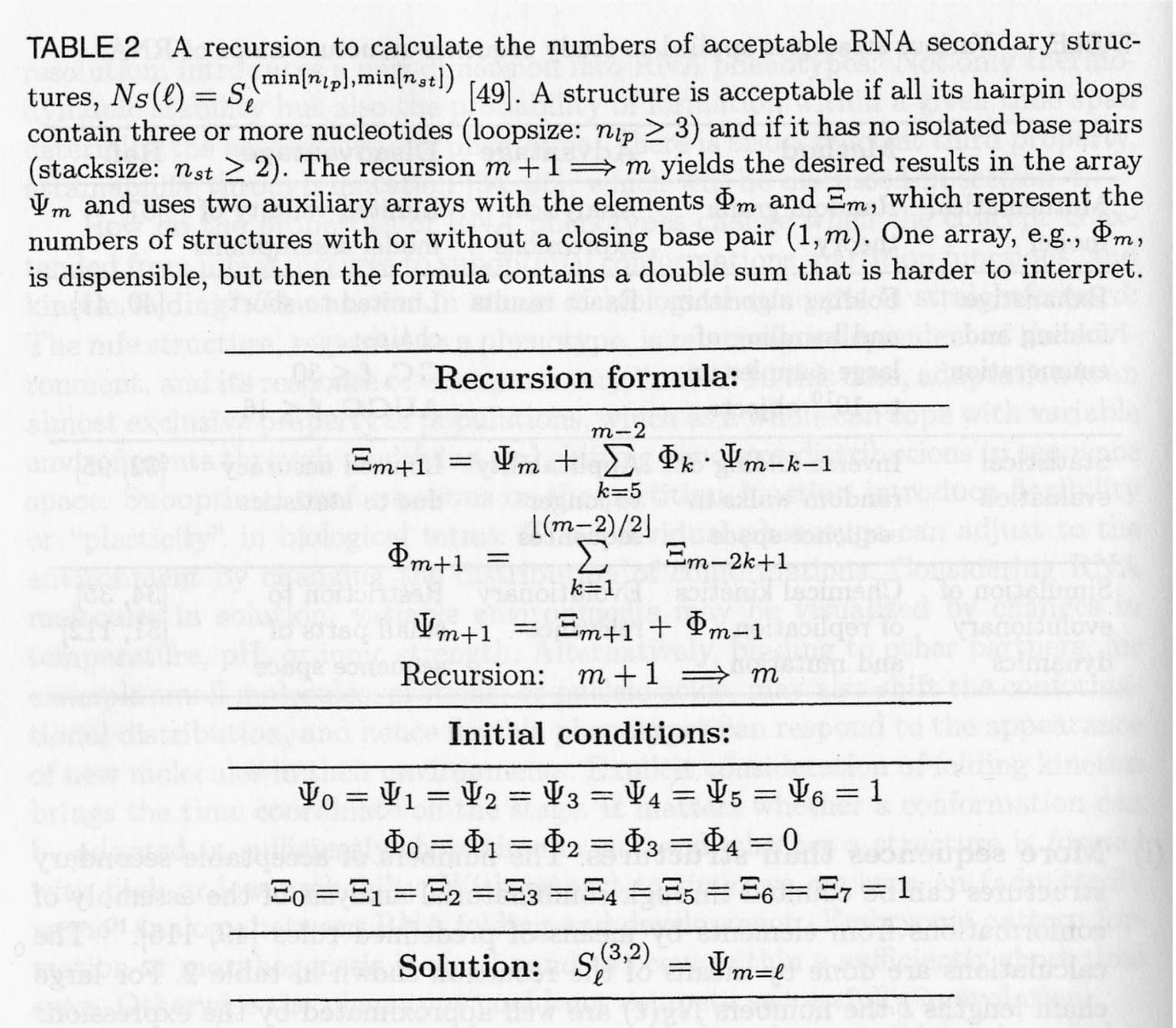


TABLE 2 A recursion to calculate the numbers of acceptable RNA secondary structures, $N_S(\ell) = S_\ell^{(\min[n_{lp}],\min[n_{st}])}$ [49]. A structure is acceptable if all its hairpin loops contain three or more nucleotides (loopsize: $n_{lp} \geq 3$) and if it has no isolated base pairs (stacksize: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

Recursion formula:

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

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

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

Recursion: $m+1 \Longrightarrow m$

Initial conditions:

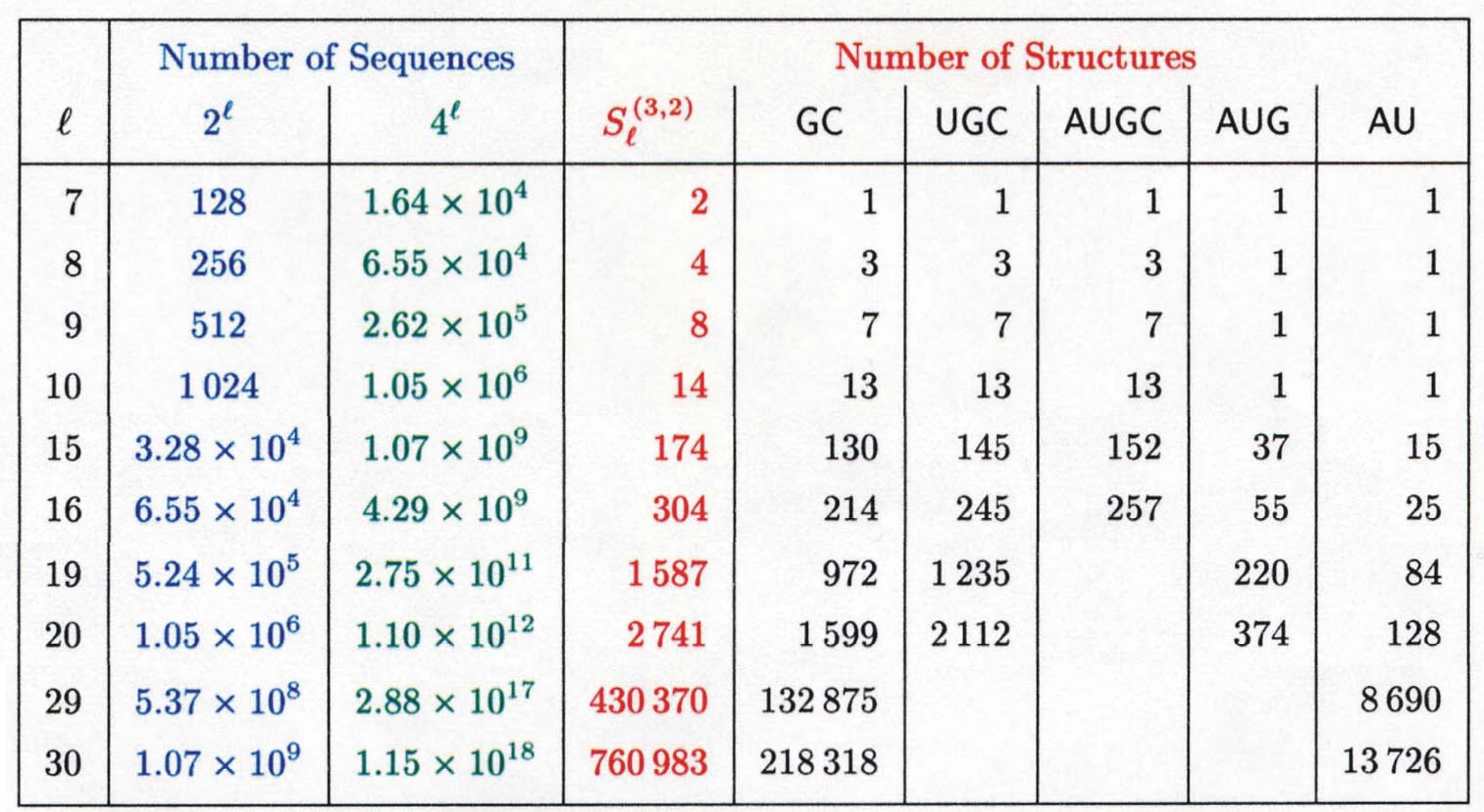
$$\Psi_0 = \Psi_1 = \Psi_2 = \Psi_3 = \Psi_4 = \Psi_5 = \Psi_6 = 1$$

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

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

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

Recursion formula for the number of acceptable RNA secondary structures



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

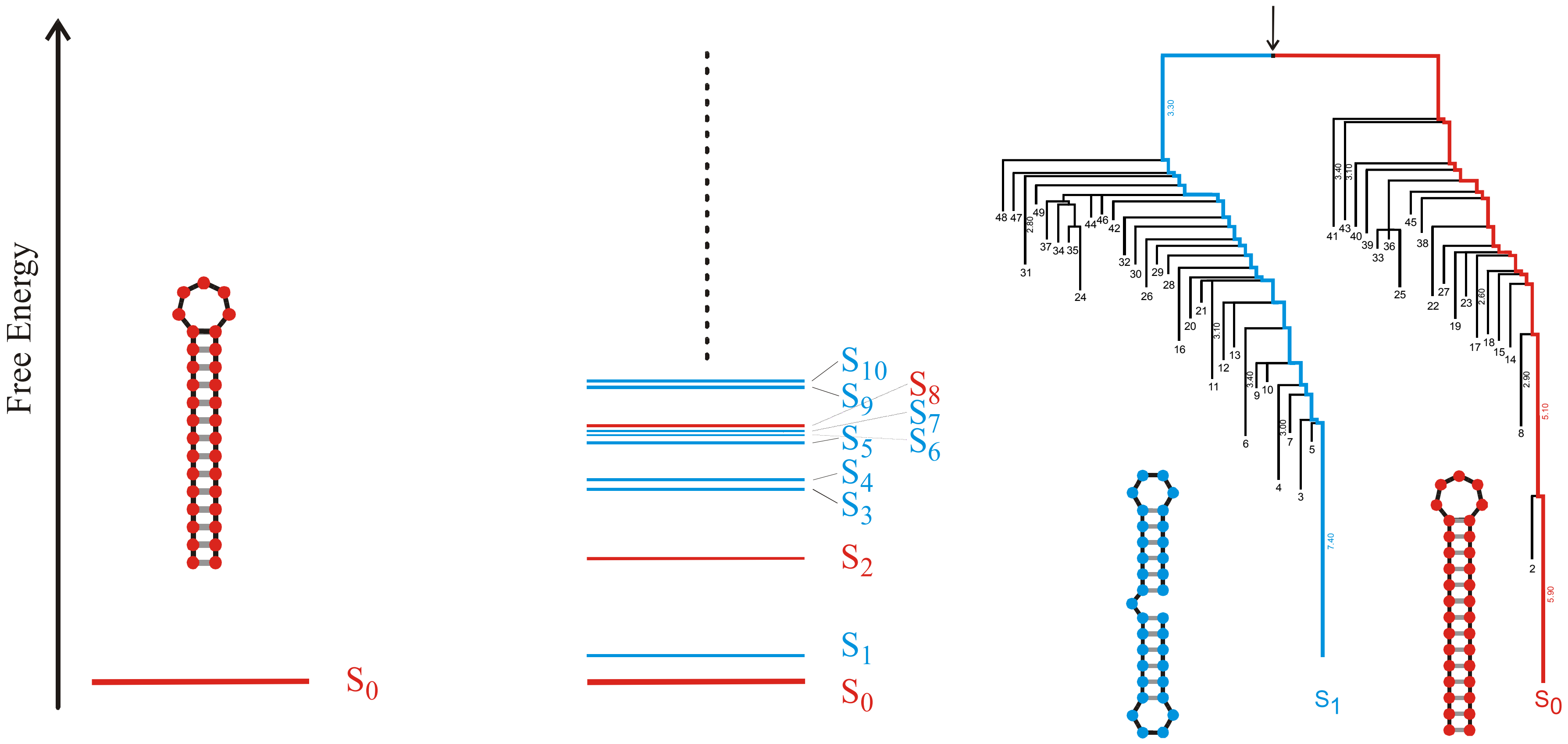
Computed numbers of minimum free energy structures over different nucleotide alphabets

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

T = 0 K , t Á ¸

T > 0 K , t Á ¸

T > 0 K , t finite



Minimum Free Energy Structure

Suboptimal Structures

Kinetic Structures

Different notions of RNA structure including suboptimal conformations and folding kinetics

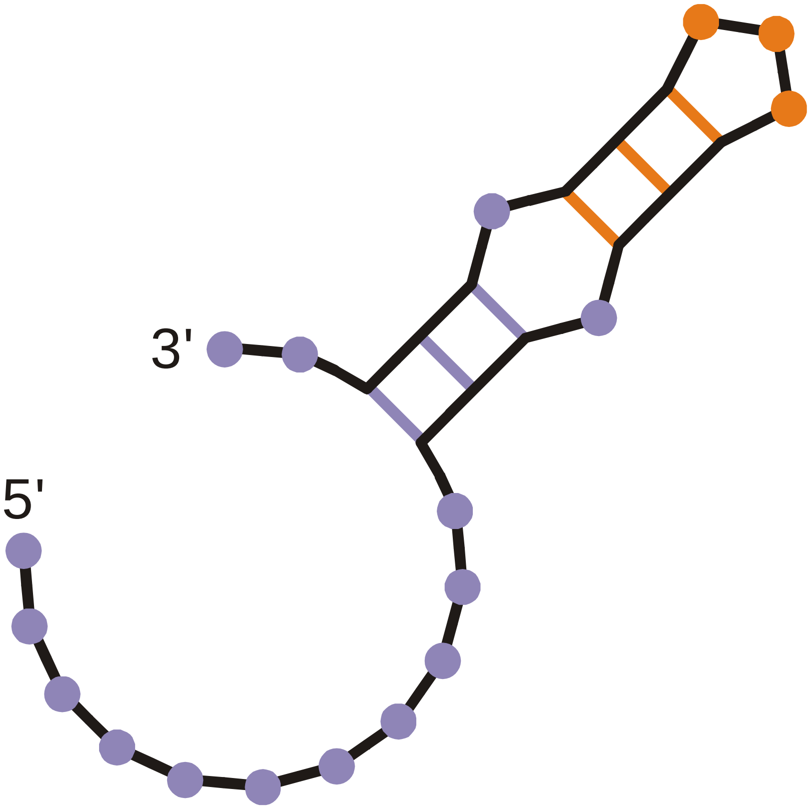
Suboptimal RNA Secondary Structures

Michael Zuker. *On finding all suboptimal foldings of an RNA molecule*. Science **244** (1989), 48-52

Stefan Wuchty, Walter Fontana, Ivo L. Hofacker, Peter Schuster. *Complete suboptimal folding of RNA and the stability of secondary structures.* Biopolymers **49** (1999), 145-165

Total number of structures including all suboptimal conformations, stable and unstable (with $DG_0>0$):

#conformations = **1 416 661**

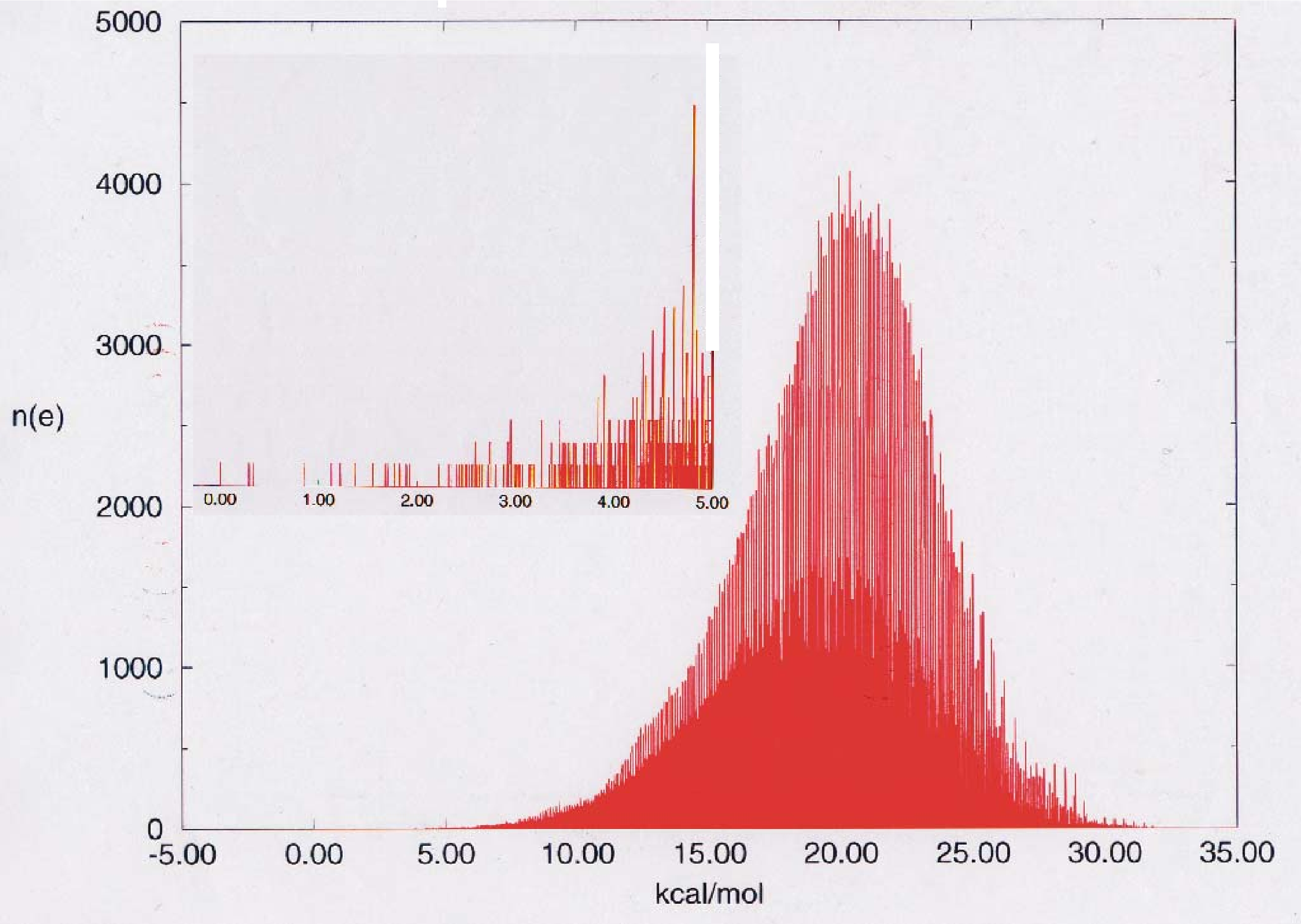


Minimum free energy structure

AAAGGGCACAGGGUGAUUUCAAUAAUUUUA

Sequence

Example of a small RNA molecule: n=30



Density of stares of suboptimal structures of the RNA molecule with the sequence:

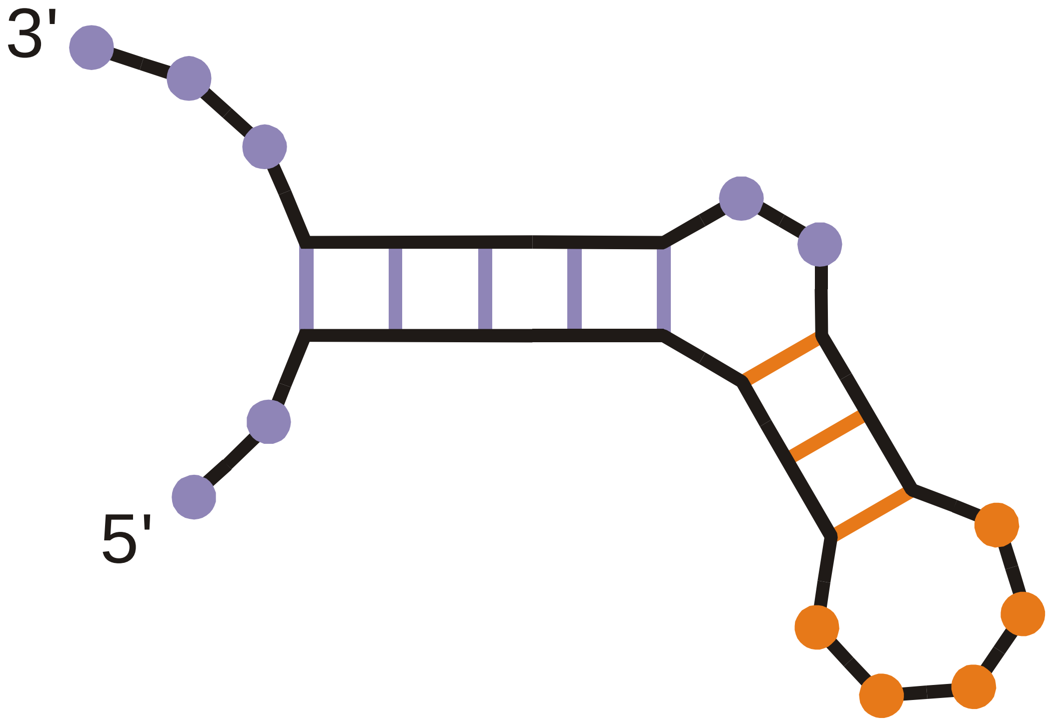
AAAGGGCACAGGGUGAUUUCAAUAAUUUUA

Partition Function of RNA Secondary Structures

John S. McCaskill. *The equilibrium function and base pair binding probabilities for RNA secondary structure*. Biopolymers **29** (1990), 1105-1119

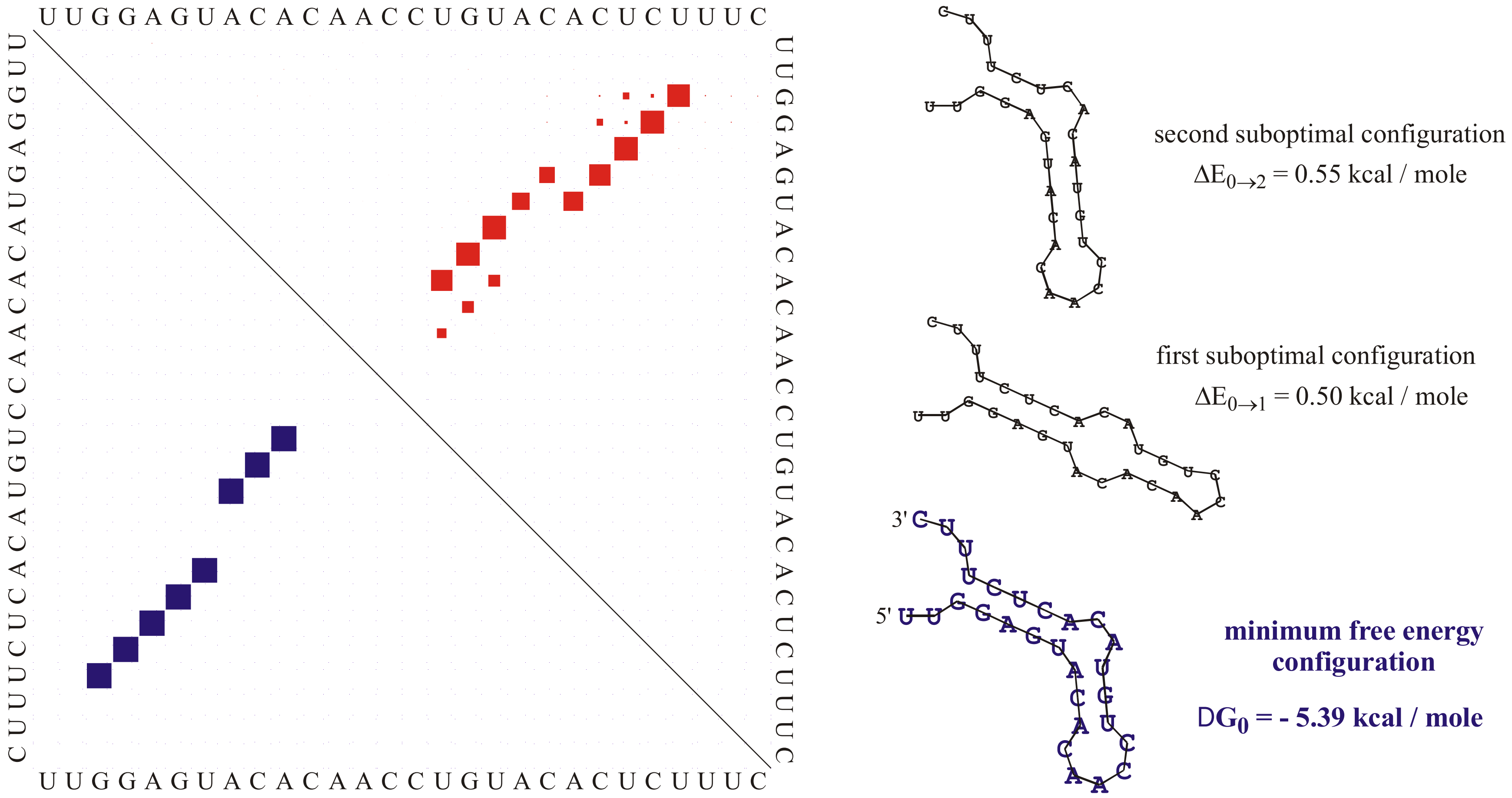
Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster. *Fast folding and comparison of RNA secondary structures.* Monatshefte für Chemie **125** (1994), 167-188

Example of a small RNA molecule with two low-lying suboptimal conformations which contribute substantially to the partition function



UUGGAGUACACAACCUGUACACUCUUUC

Example of a small RNA molecule: n=28


U U G G A G U A C A C A A C C U G U A C A C U C U U U C
C U U U C U C A C A U G U C C A A C A C A U G A G G U U
U U G G A G U A C A C A A C C U G U A C A C U C U U U C
U U G G A G U A C A C A A C C U G U A C A C U C U U U C
3'
5'

second suboptimal configuration

$\Delta E_{0\to2}$ = 0.55 kcal / mole

first suboptimal configuration

$\Delta E_{0\to1}$ = 0.50 kcal / mole

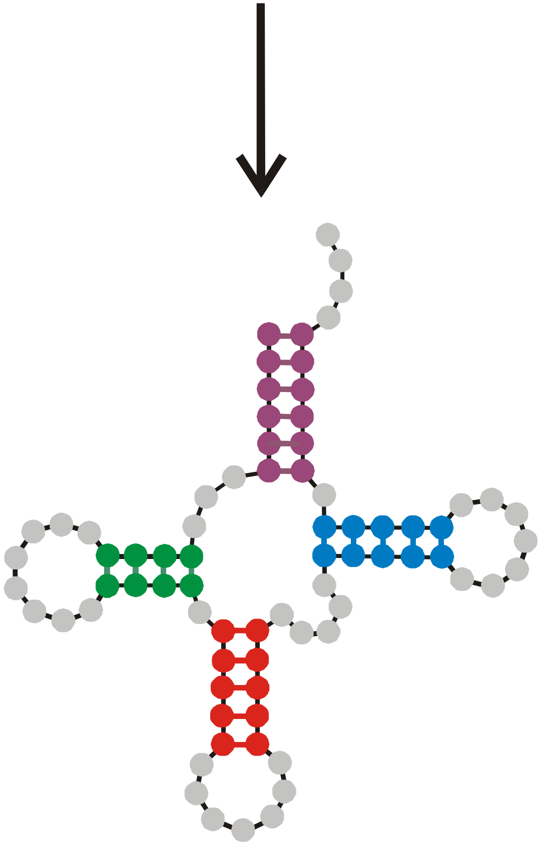
minimum free energy configuration

DG_0 = - 5.39 kcal / mole

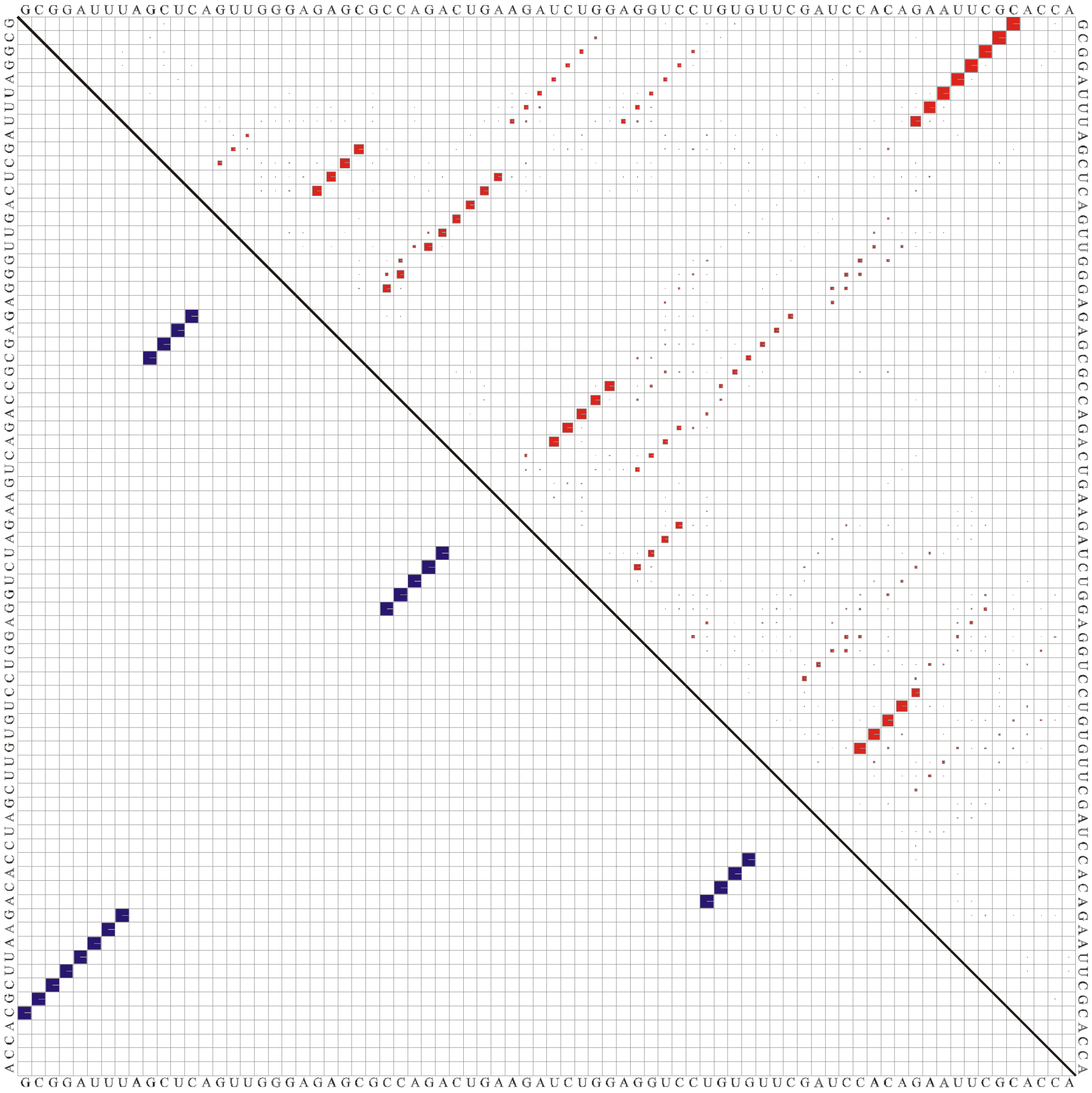
„Dot plot“ of the minimum free energy structure (**lower triangle**) and the partition function (**upper triangle**) of a small RNA molecule (n=28) with low energy suboptimal configurations

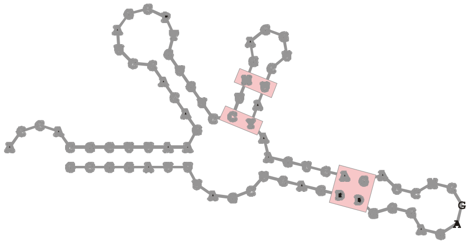
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA



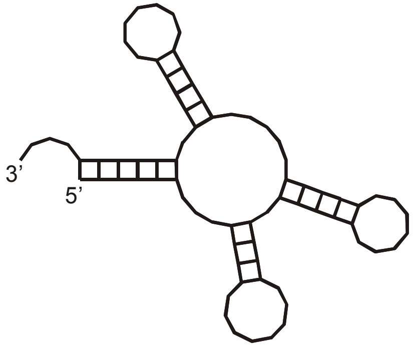
Phenylalanyl-tRNA as an example for the computation of the partition function





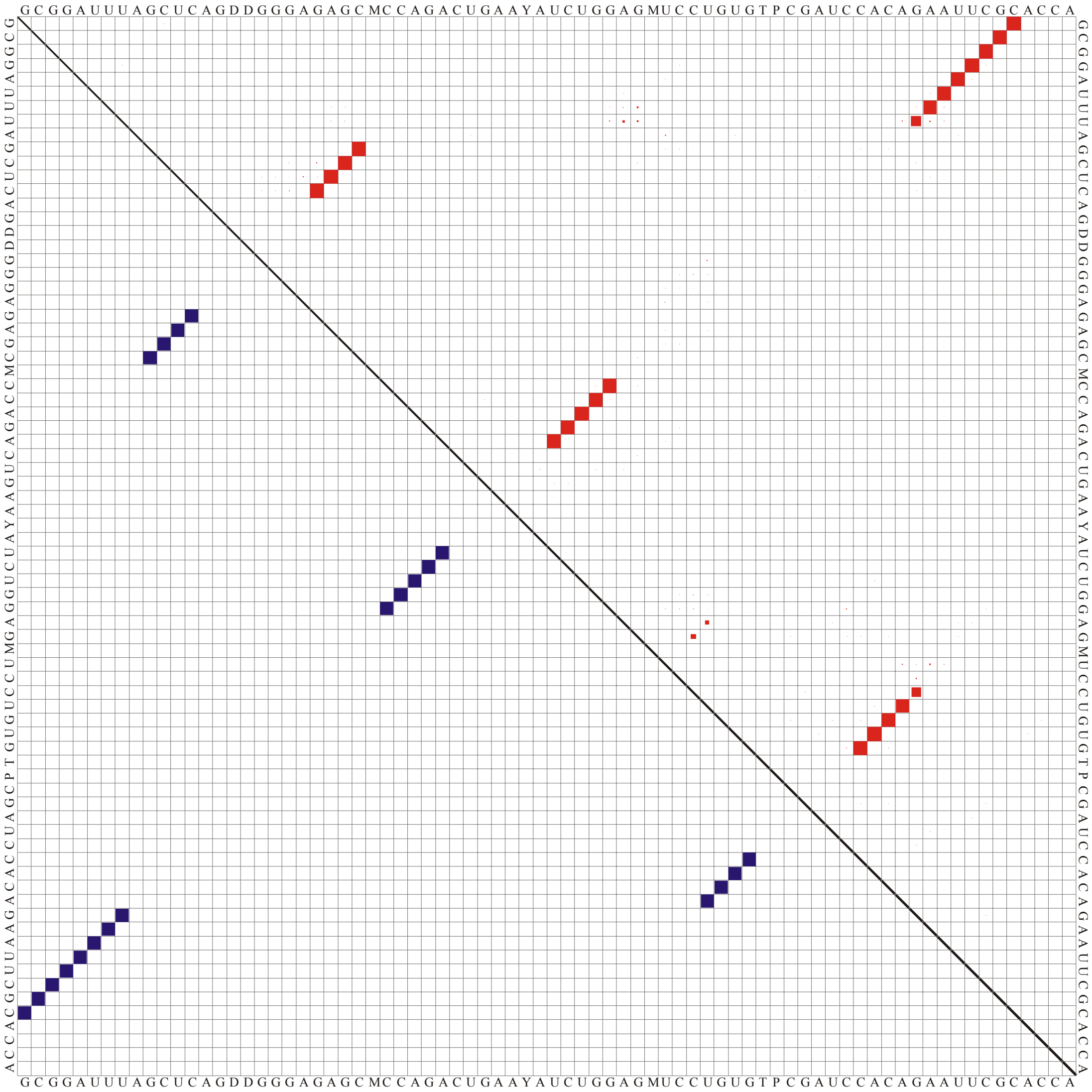
first suboptimal configuration

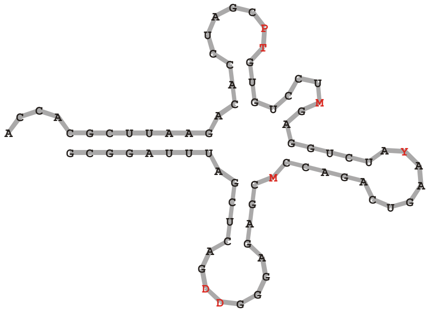
$\Delta E_{0\rightarrow1} = 0.43$ kcal / mole



tRNA^{phe}

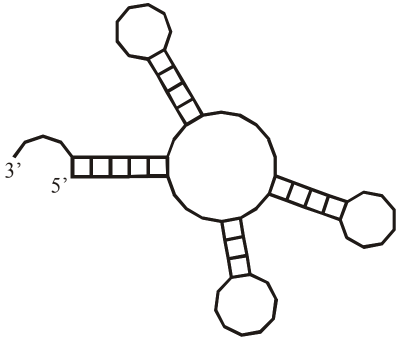
without modified bases





first suboptimal configuration

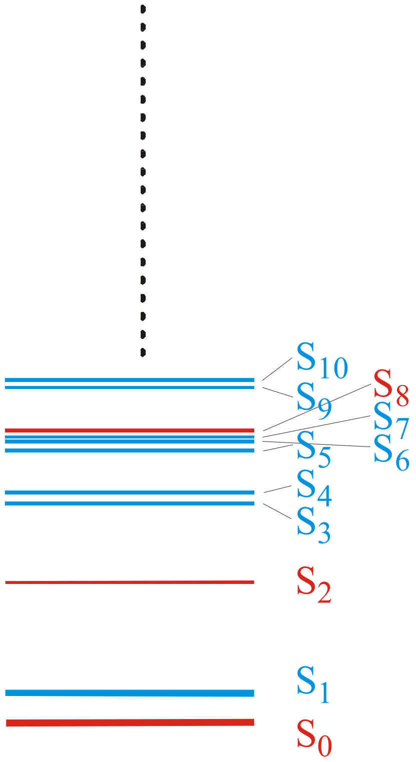
$\Delta E_{0\to1}$ = 0.94 kcal / mole



tRNA$^{\text{phe}}$

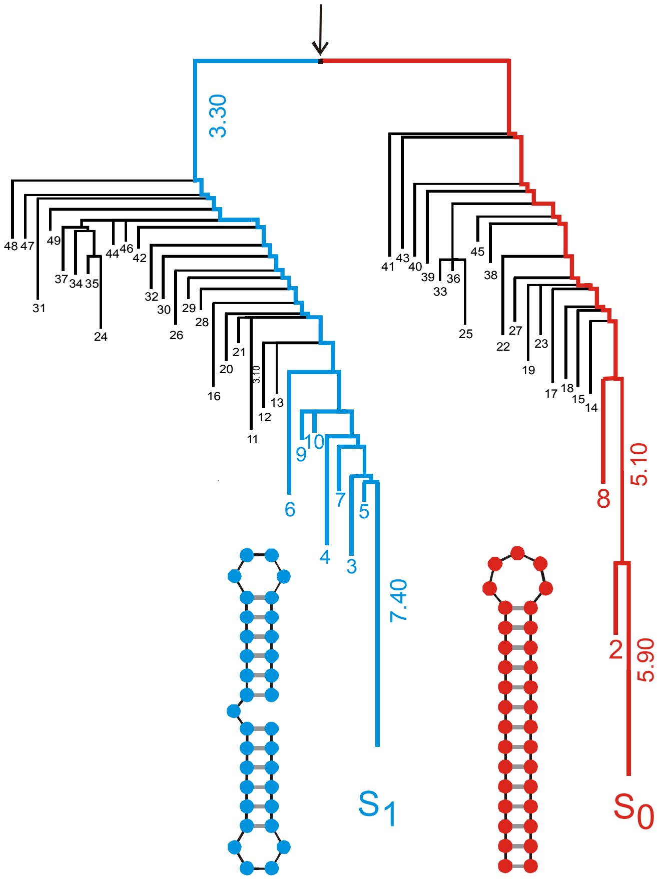
with modified bases

lim t Á ¸



Suboptimal structures

finite folding time



Kinetic folding

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

Kinetic Folding of RNA Secondary Structures

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

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

The Folding Algorithm

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

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

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

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

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

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

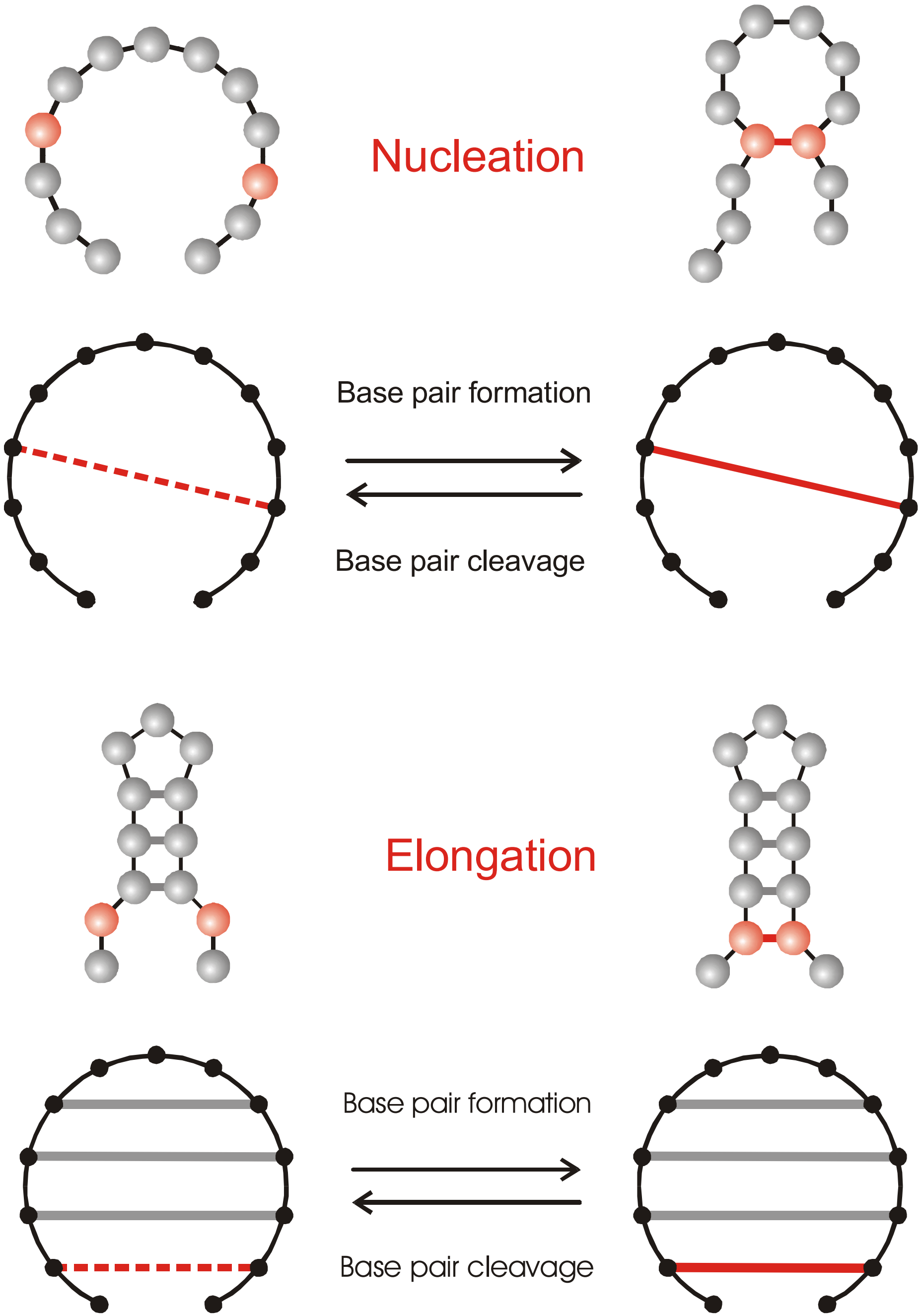
$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t) \exp(-\Delta G_{ij}/2RT) \, / \, \Sigma_i$$

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

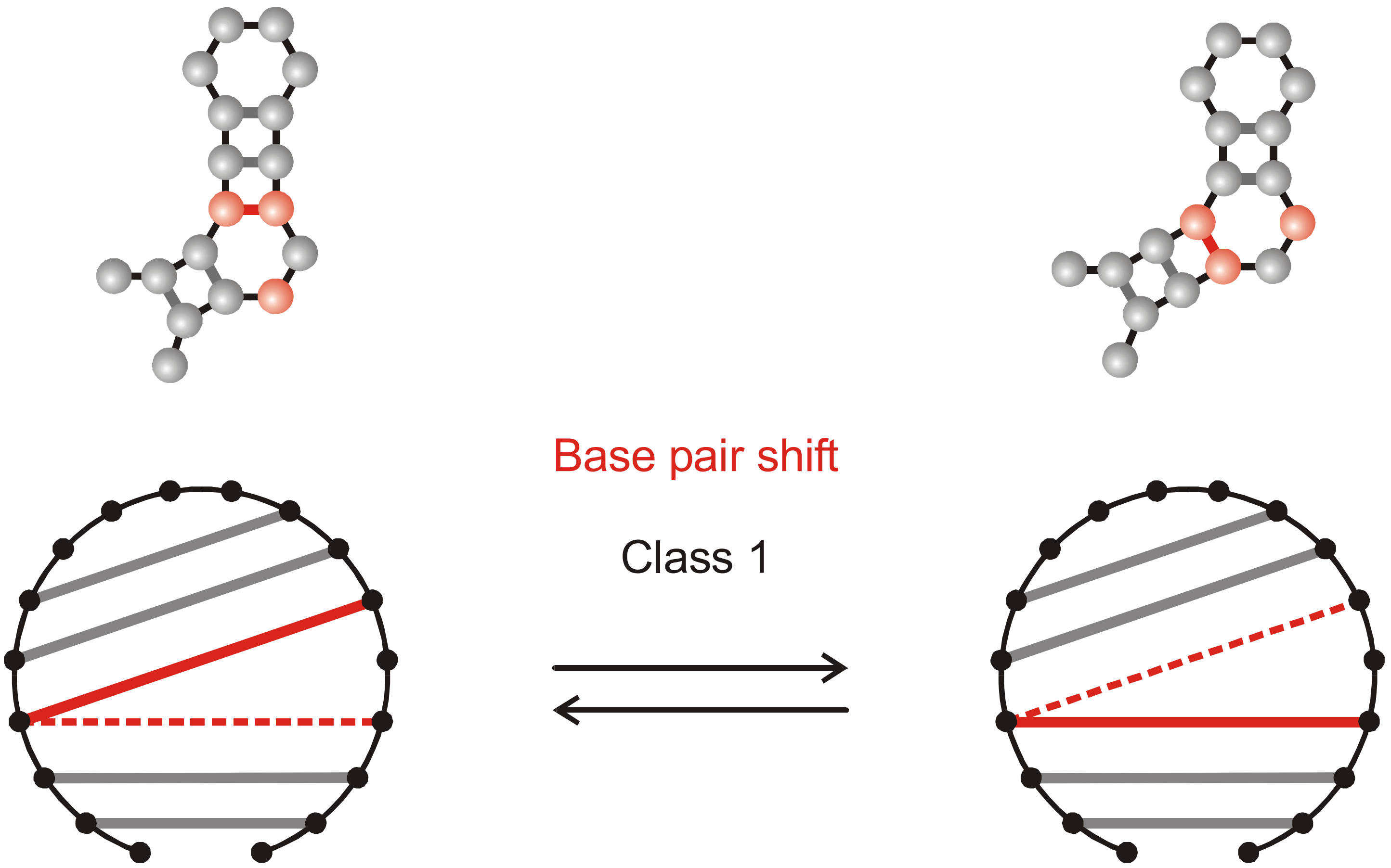
$$\Sigma_i = \sum_{k=1,k\neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

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



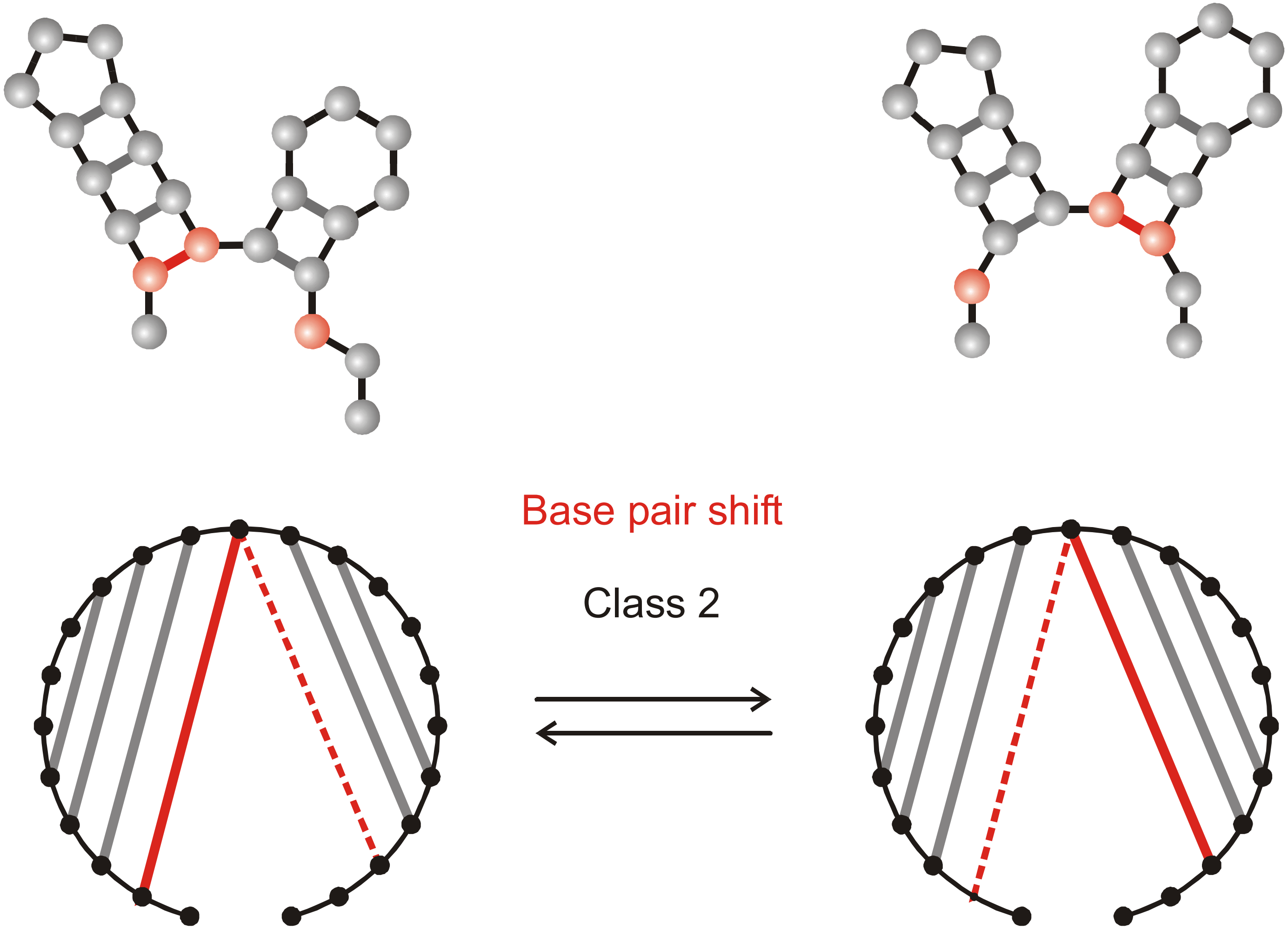
Base pair formation and base pair cleavage moves for nucleation and elongation of stacks



Base pair shift

Class 1

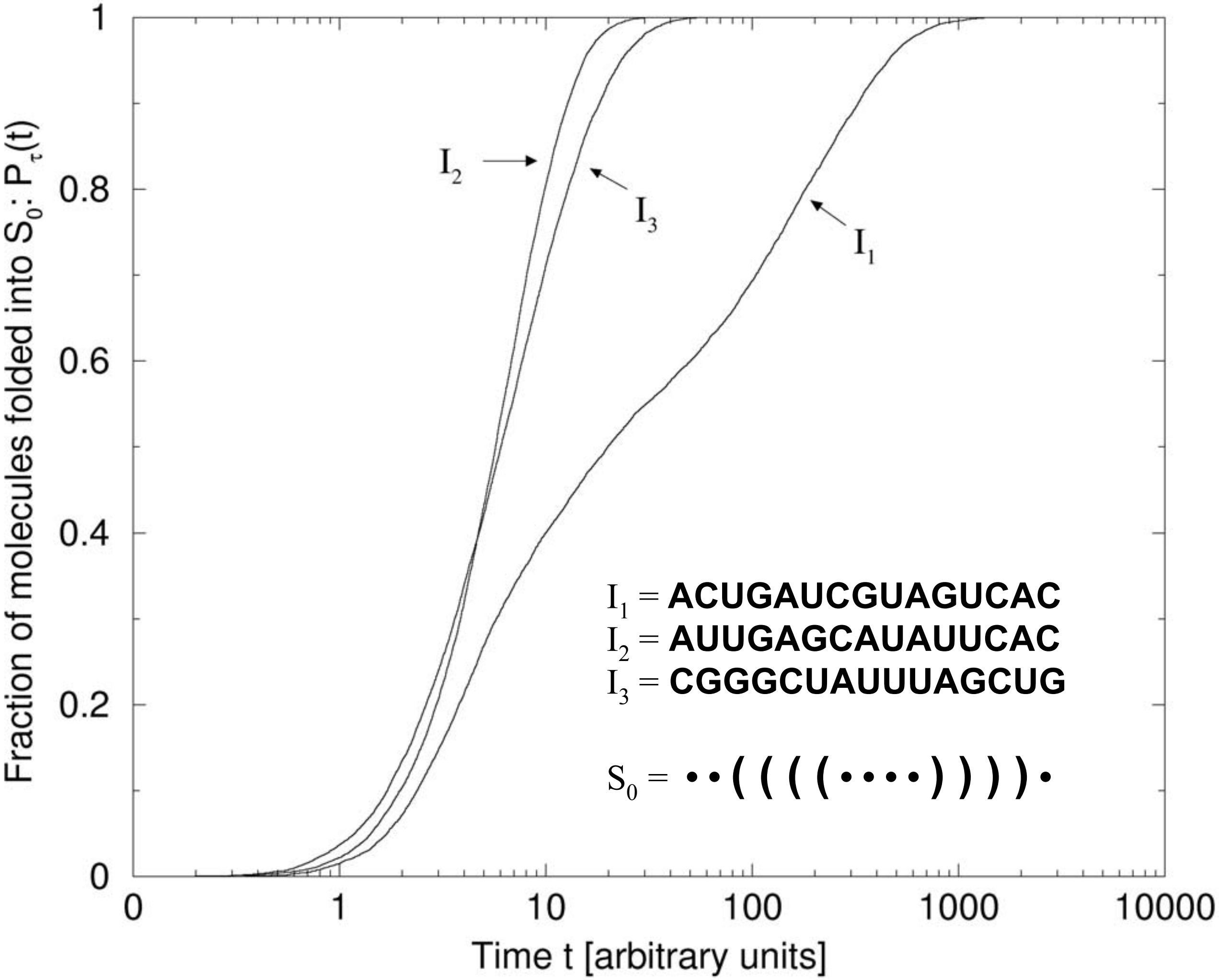
Base pair shift move of class 1: Shift inside internal loops or bulges



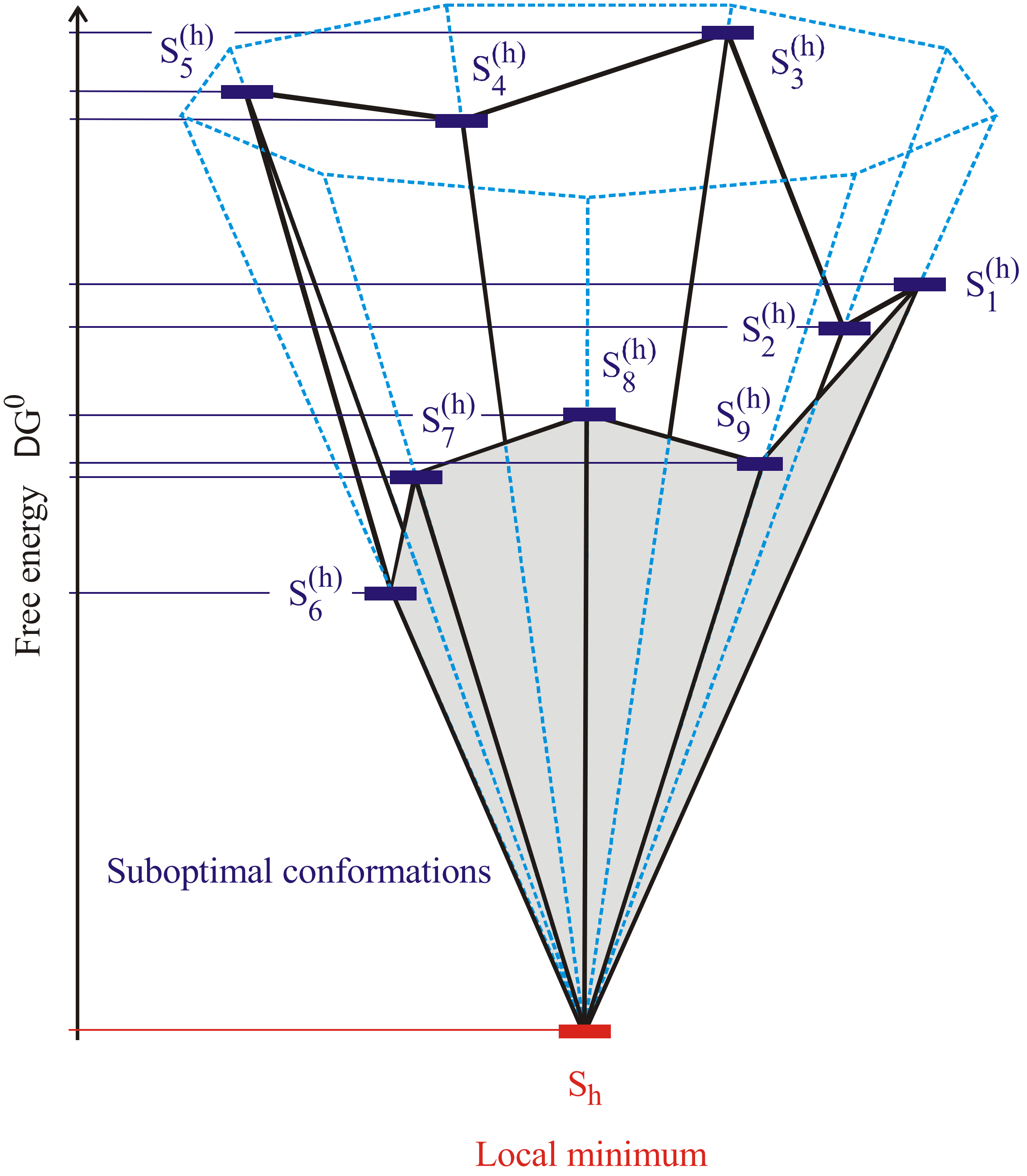
Base pair shift

Class 2

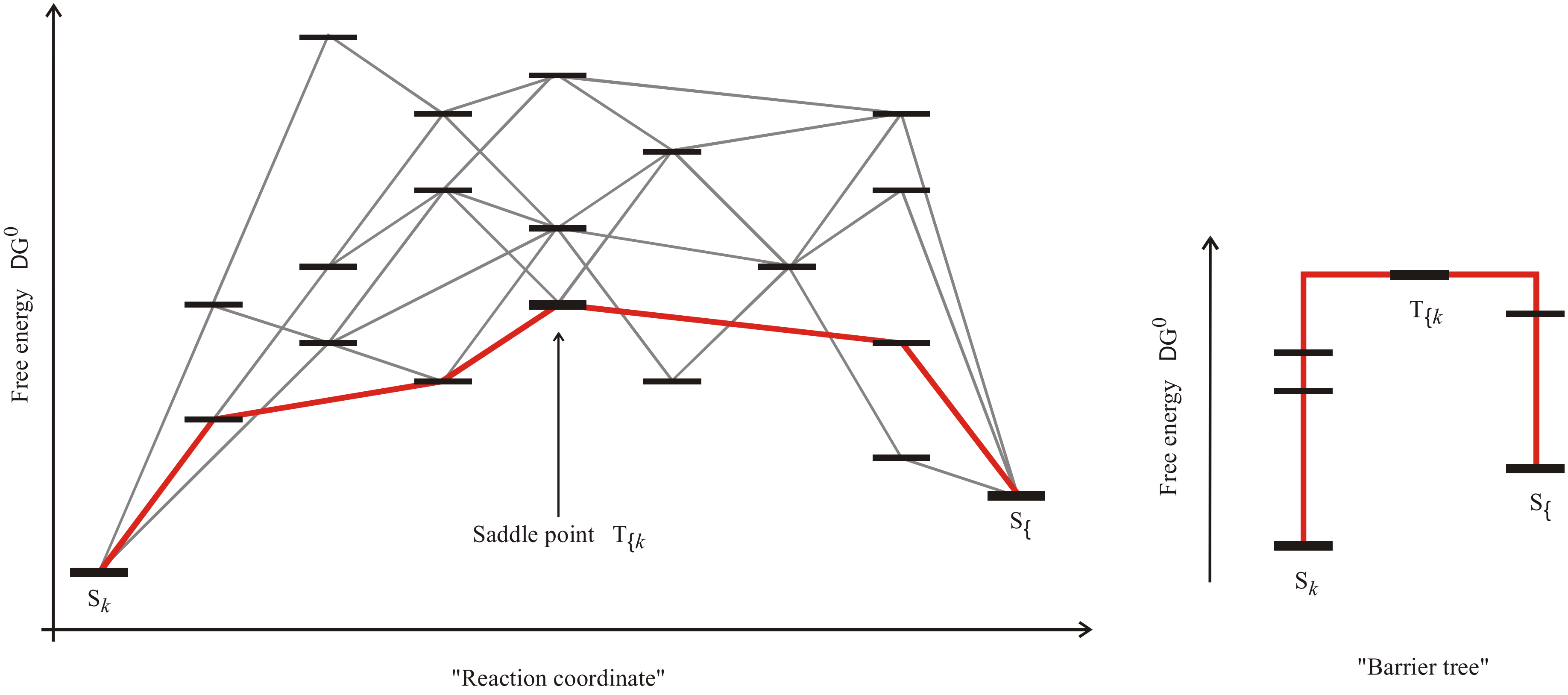
Base pair shift move of class 2: Shift involving free ends



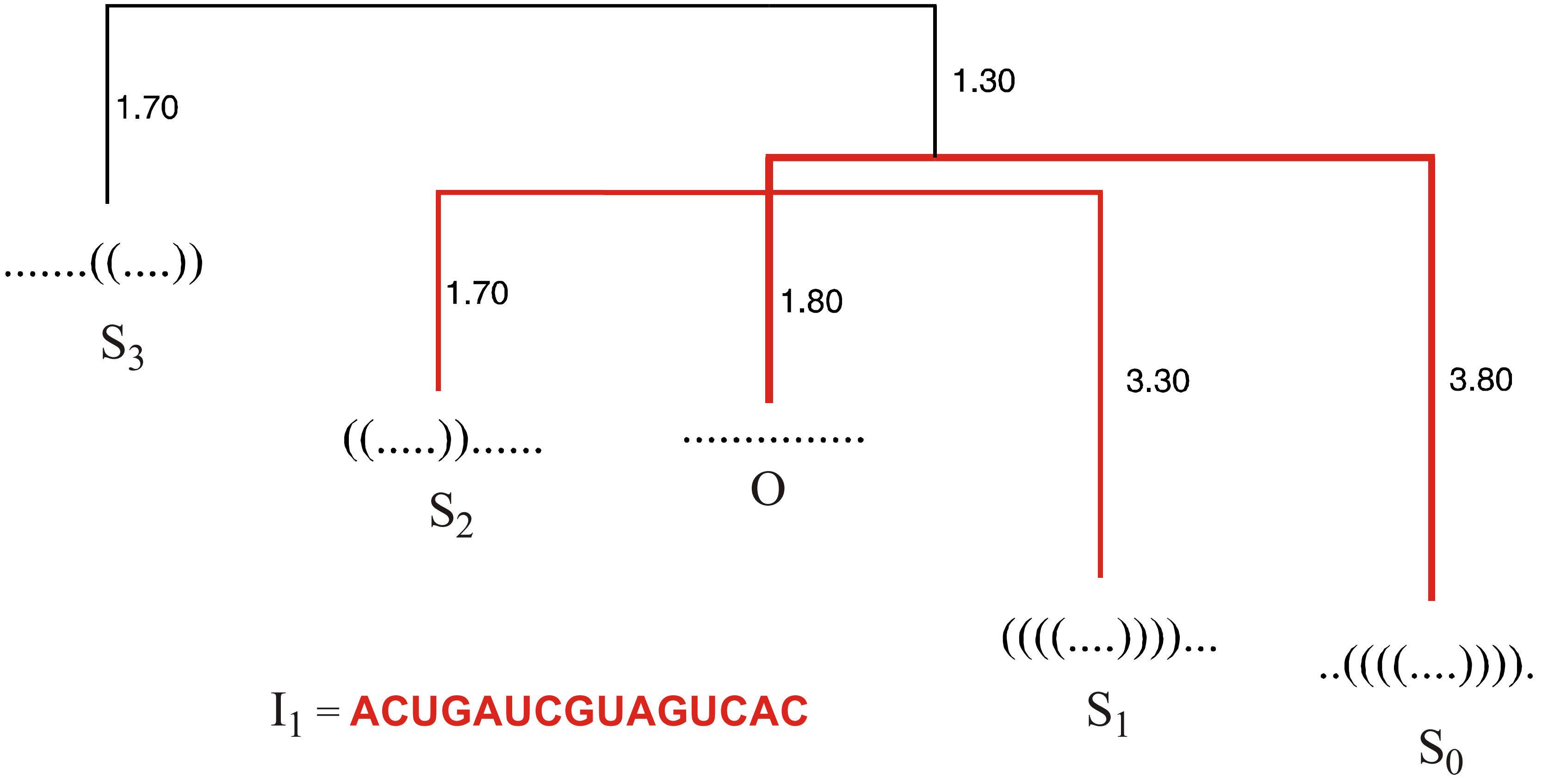
Mean folding curves for three small RNA molecules with different folding behavior



Search for local minima in conformation space

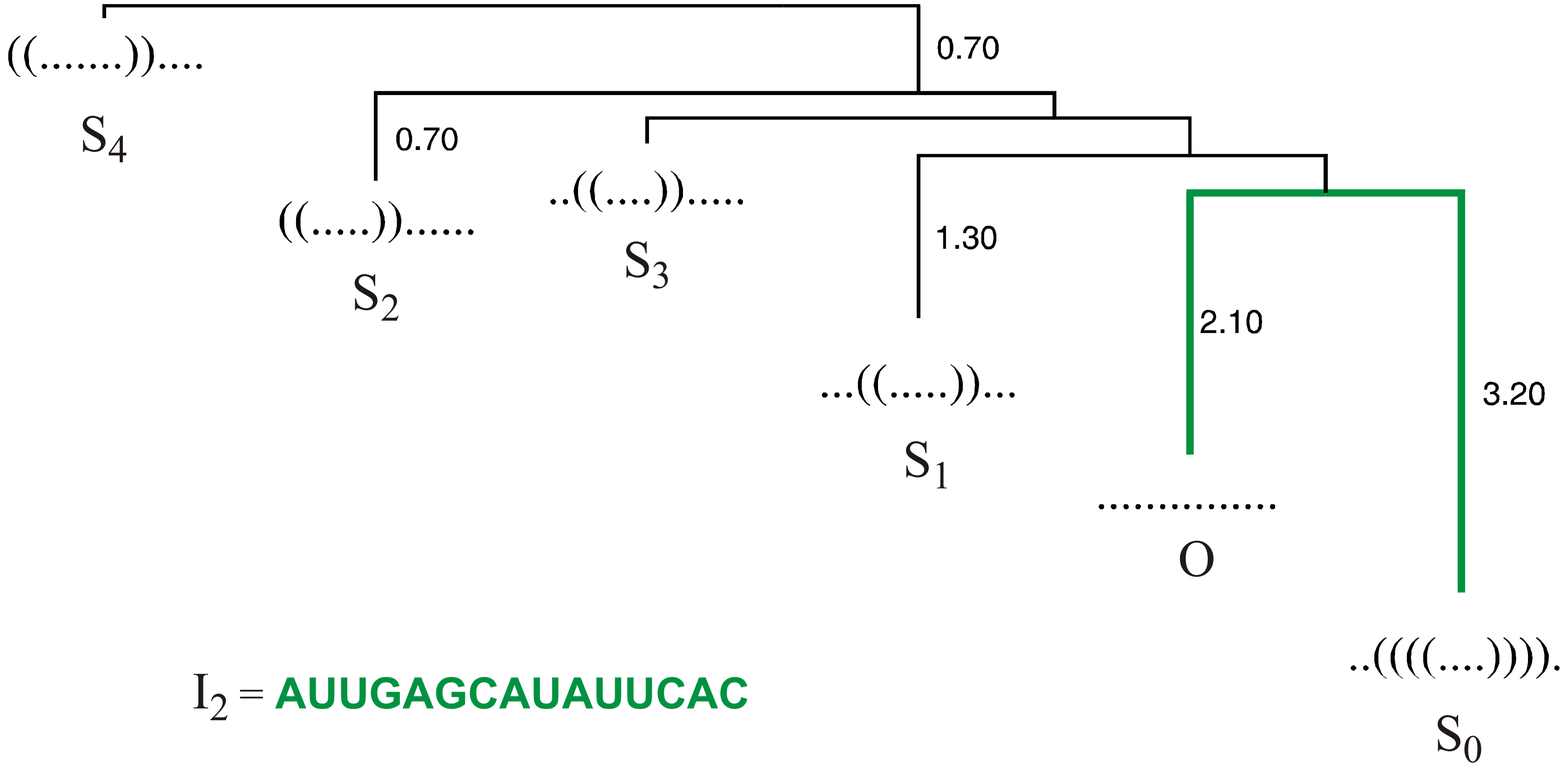


Definition of a ‚barrier tree‘



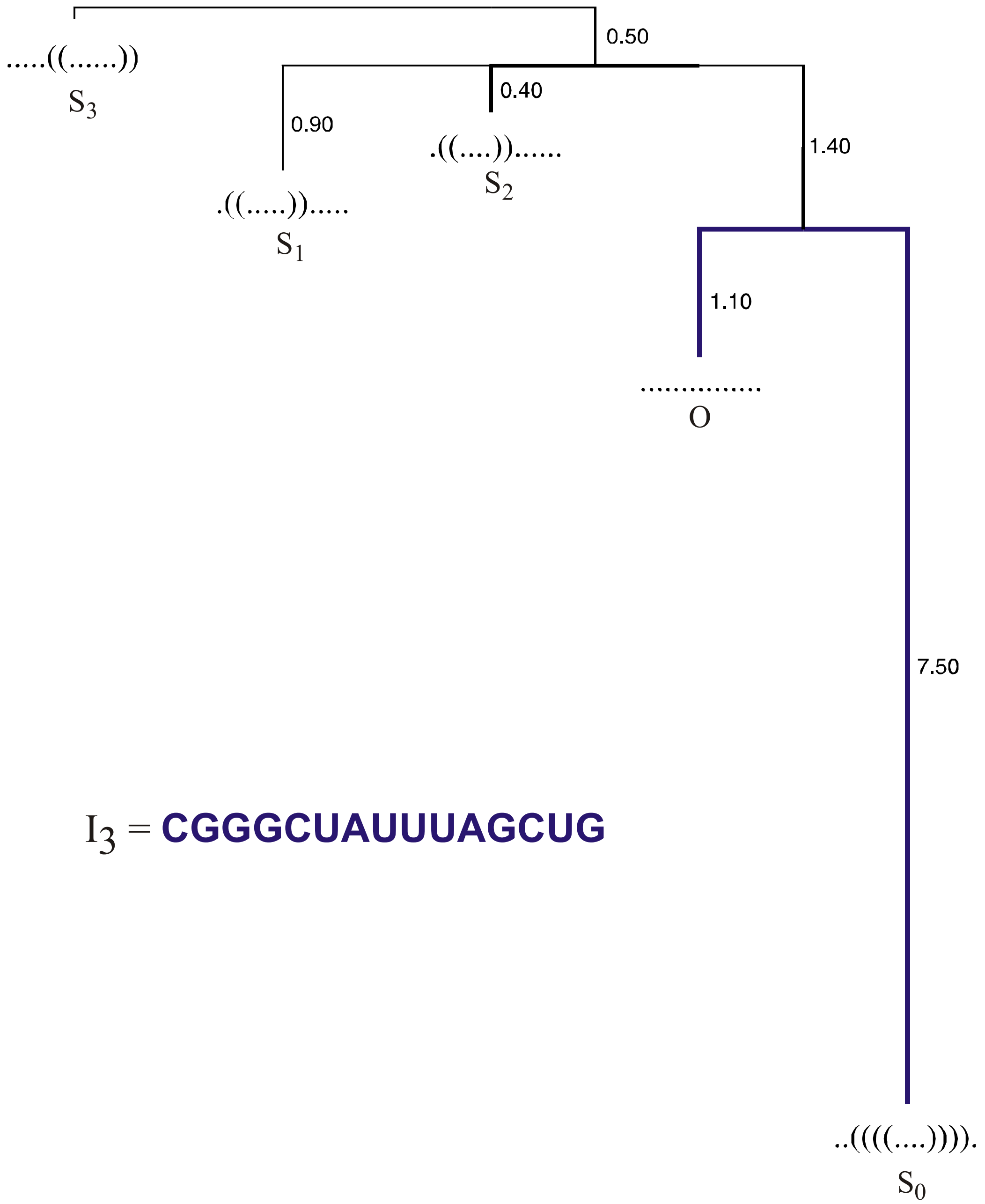
I_1 = **ACUGAUCGUAGUCAC**

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



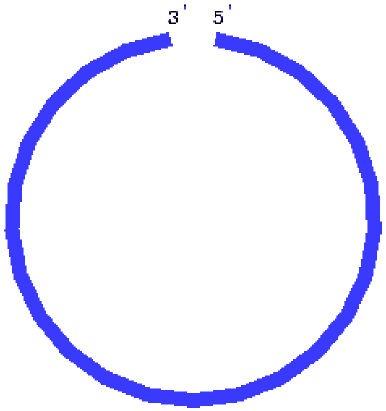
I_2 = **AUUGAGCAUAUUCAC**

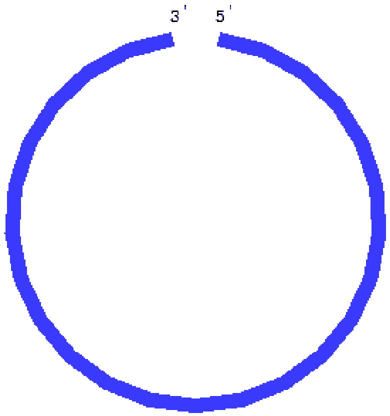
Example of an easily folding small RNA molecule with n = 15



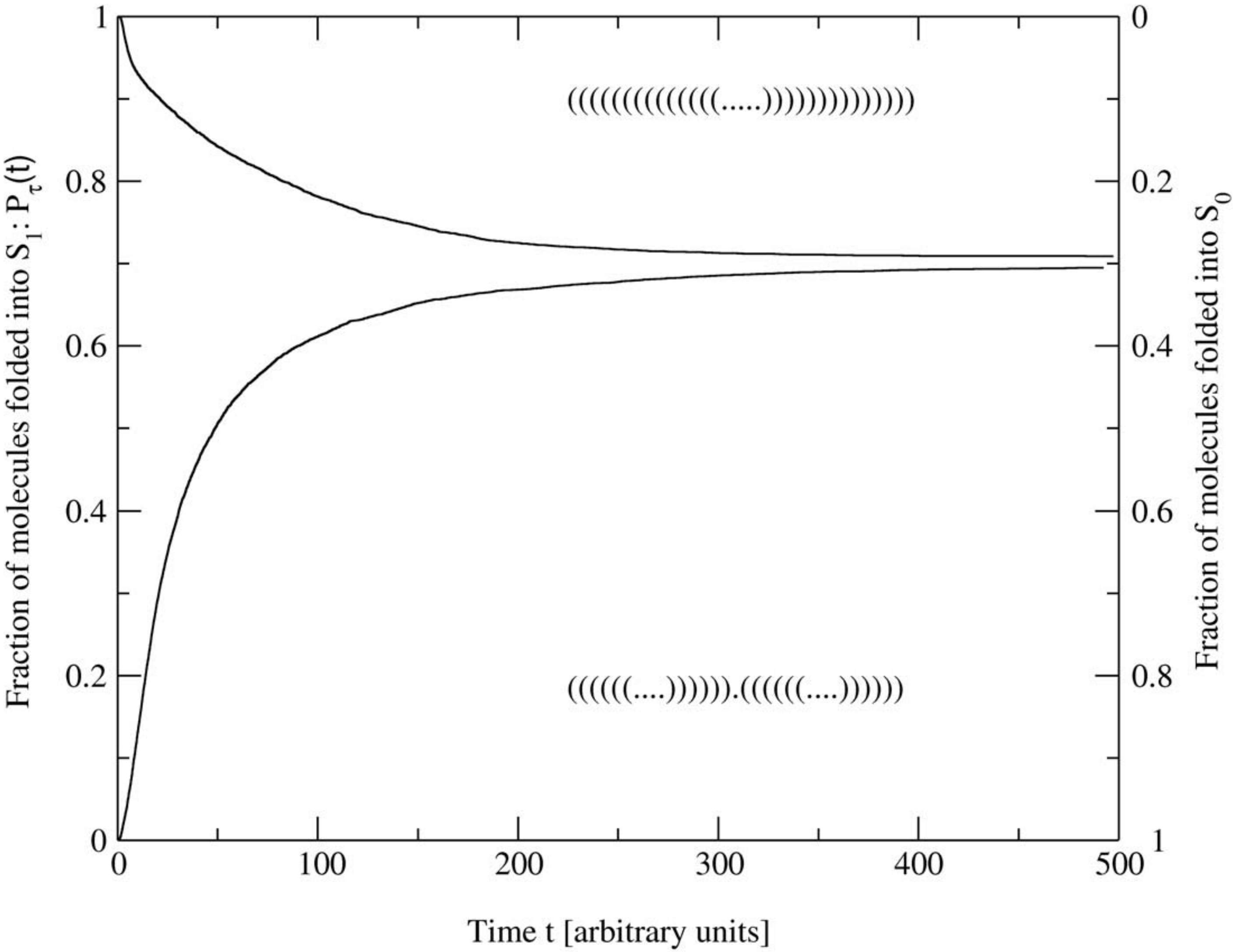
I_3 = **CGGGCUAUUUAGCUG**

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

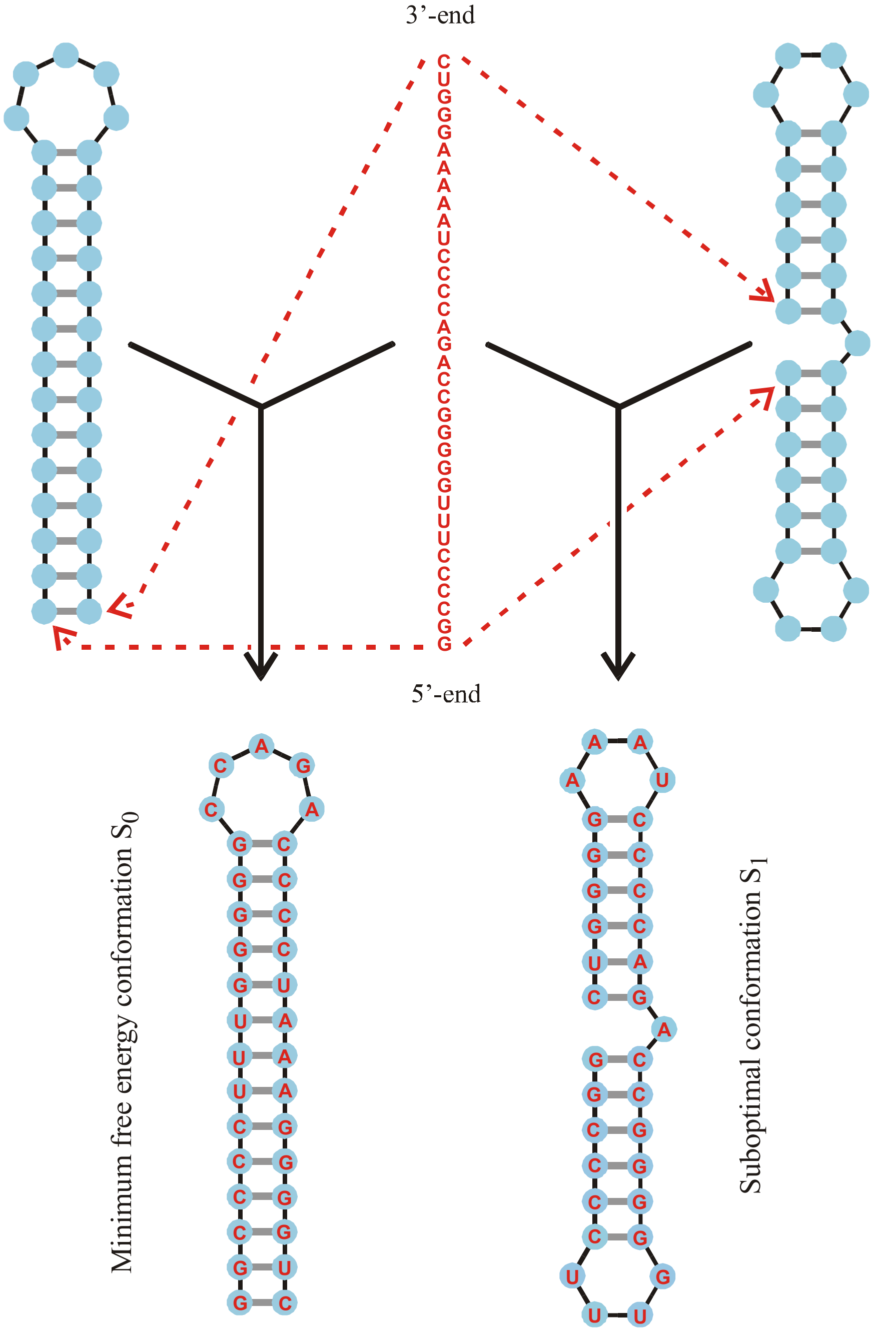




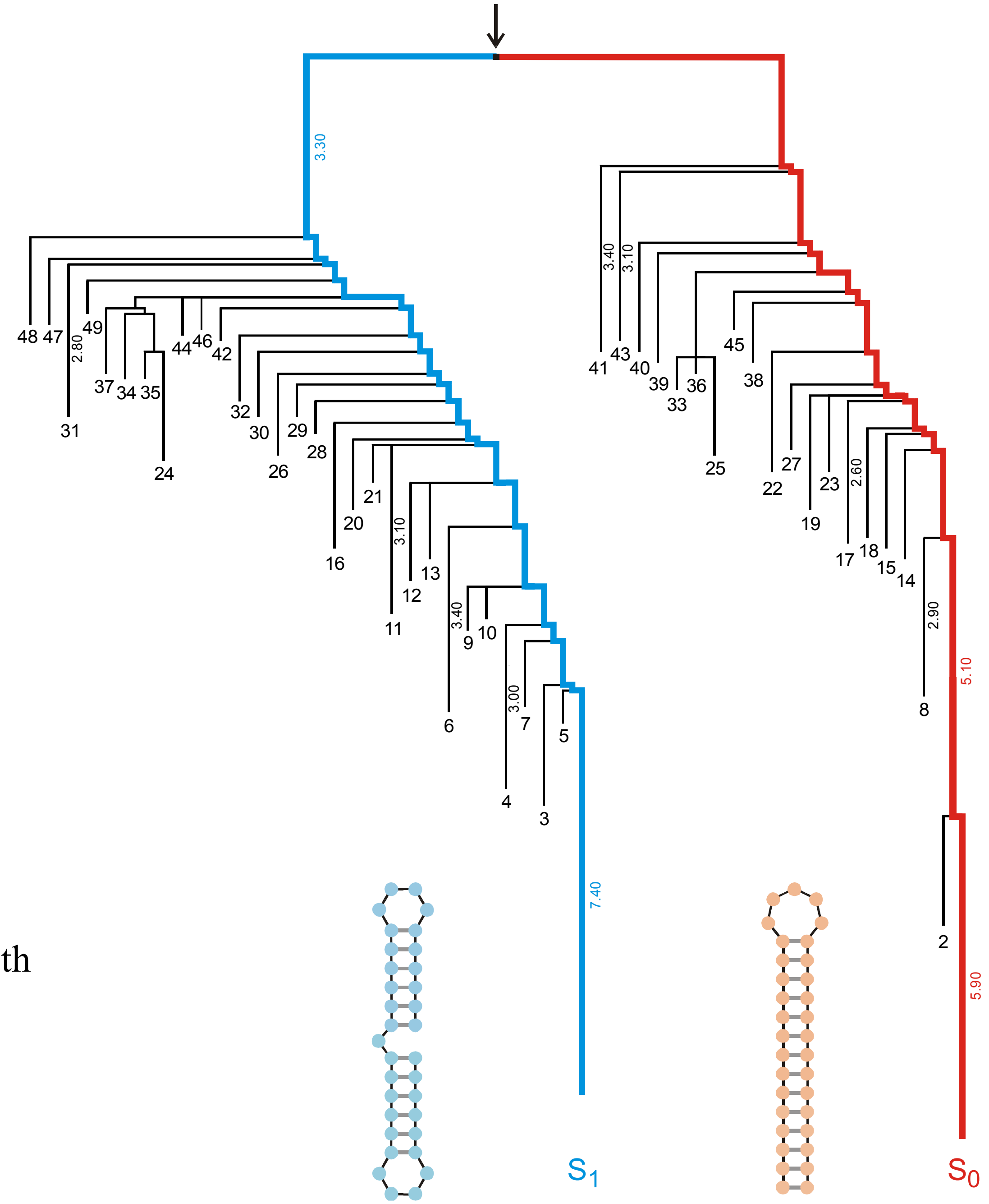
Examples of two folding trajectories leading to different local minima



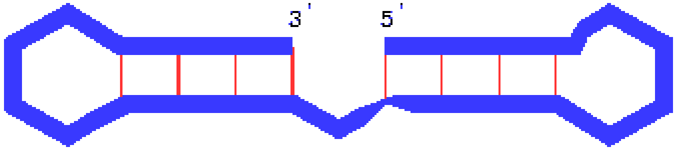
Folding dynamics of the sequence **GGCCCCUUUGGGGGCCAGACCCCUAAAAGGGUC**

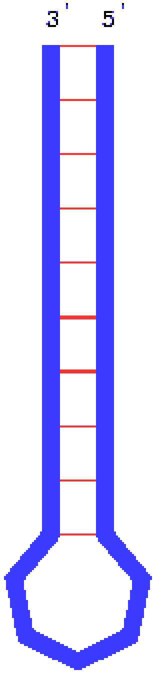


One sequence is compatible with two structures



Barrier tree of a sequence with two conformations





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

J.H.A. Nagel, J. Møller-Jensen, C. Flamm, K.J. Öistämö, J. Besnard, I.L. Hofacker, A.P. Gultyaev, M.H. de Smit, P. Schuster, K. Gerdes and C.W.A. Pleij.

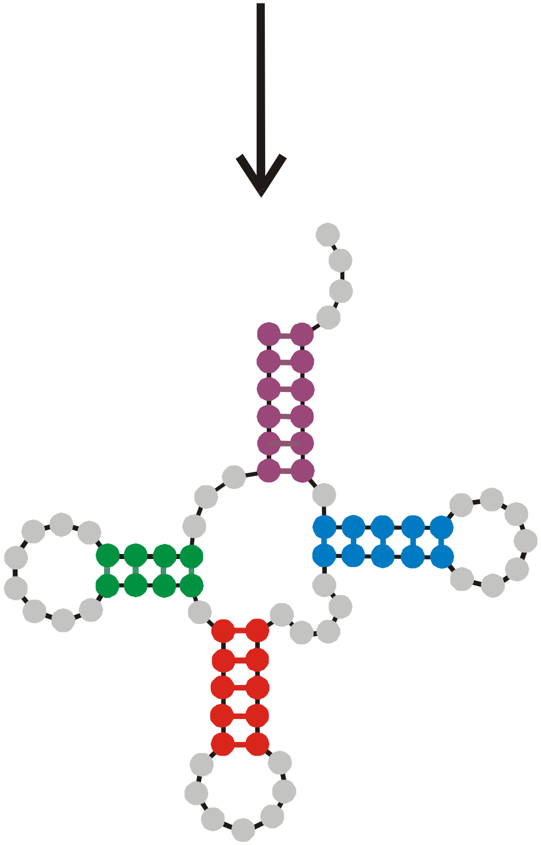
The refolding mechanism of the metastable structure in the 5'-end of the hok *mRNA of plasmid* R1*,* submitted 2004.

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

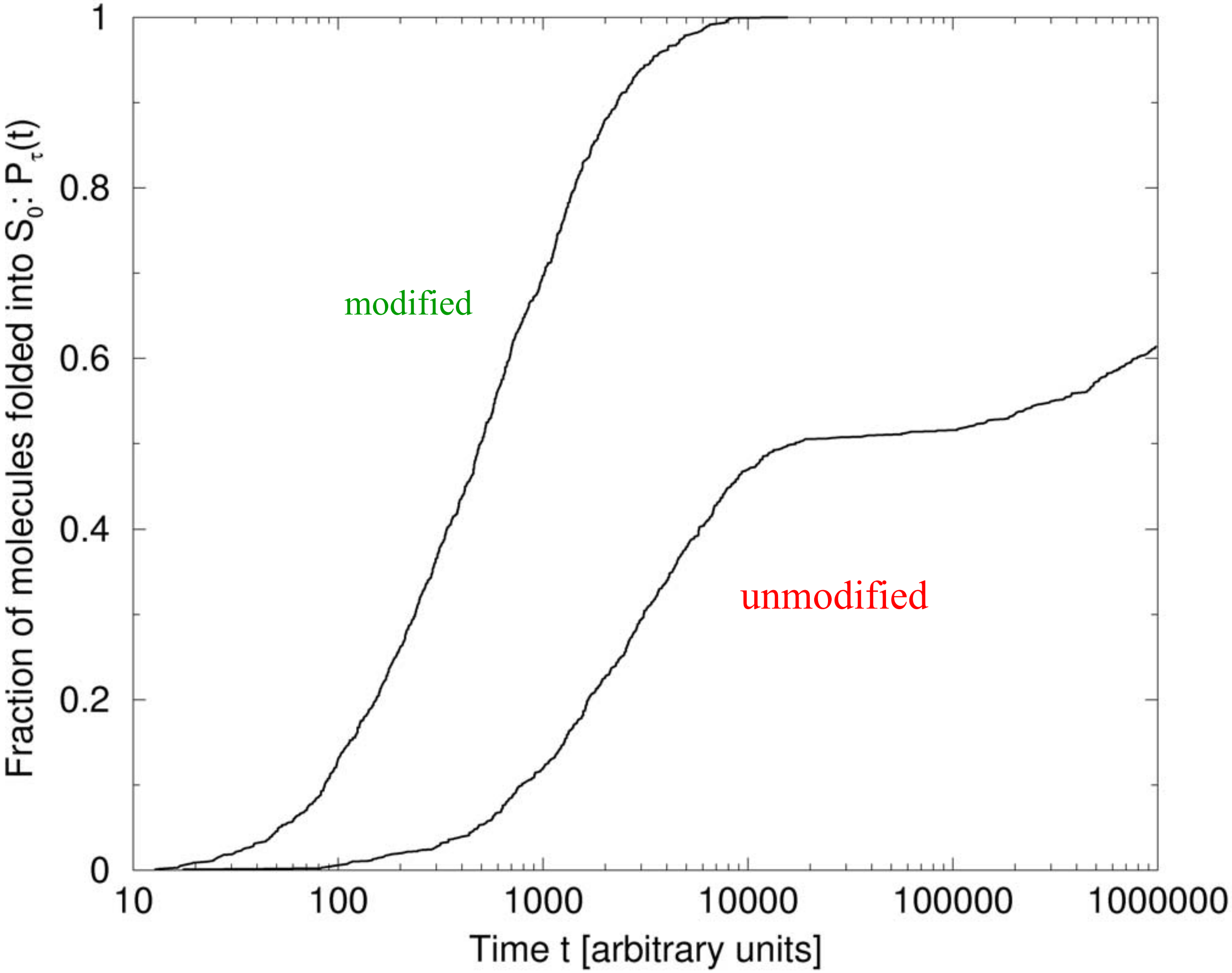
Structural parameters affecting the kinetic competition of RNA hairpin formation, in press 2004.

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

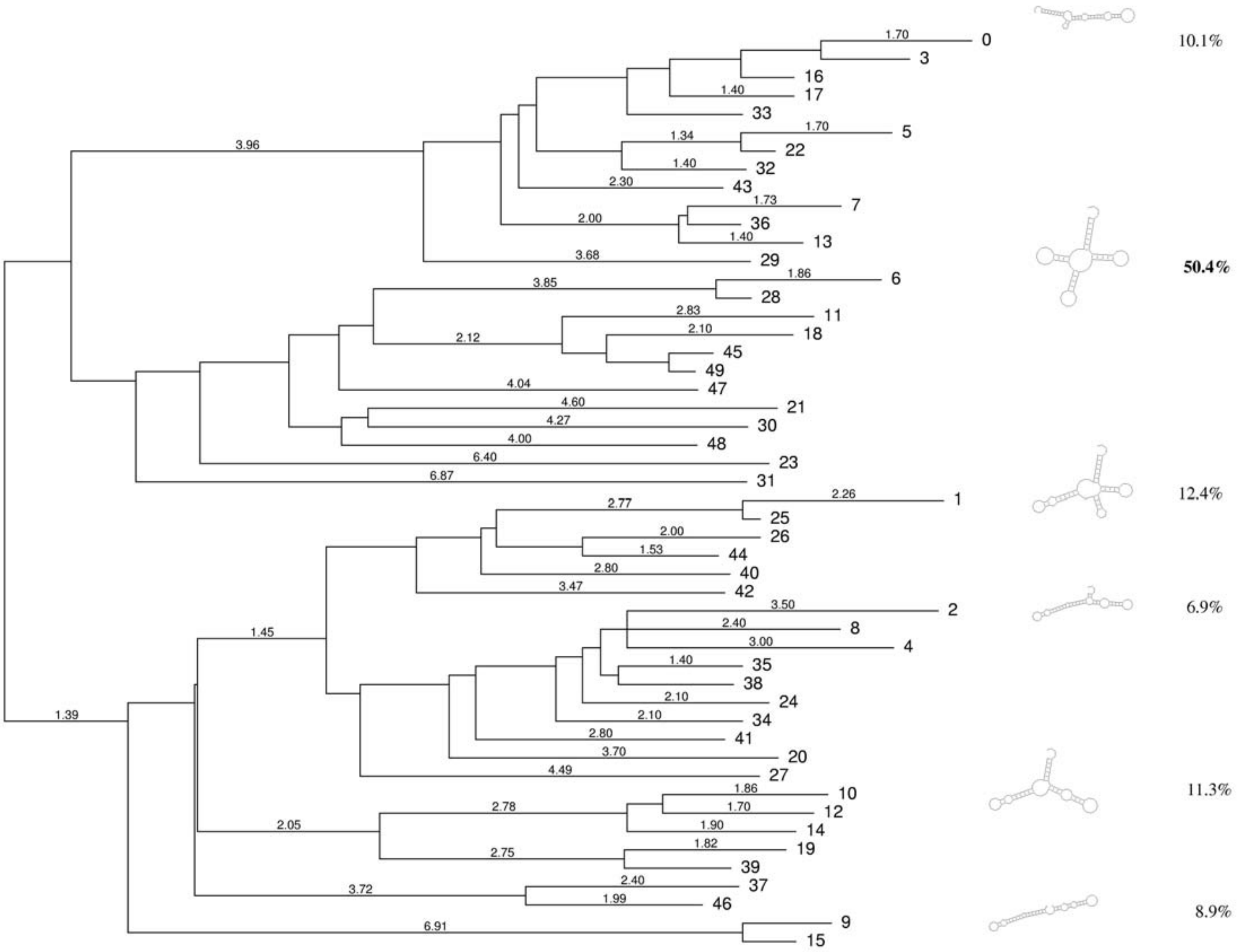
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA



Kinetic folding of phenylalanyl-tRNA



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



Barrier tree of tRNA^{phe} without modified nucelotides

Theory of sequence – structure mappings

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

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks***. Mh.Chem. **127** (1996), 355-374

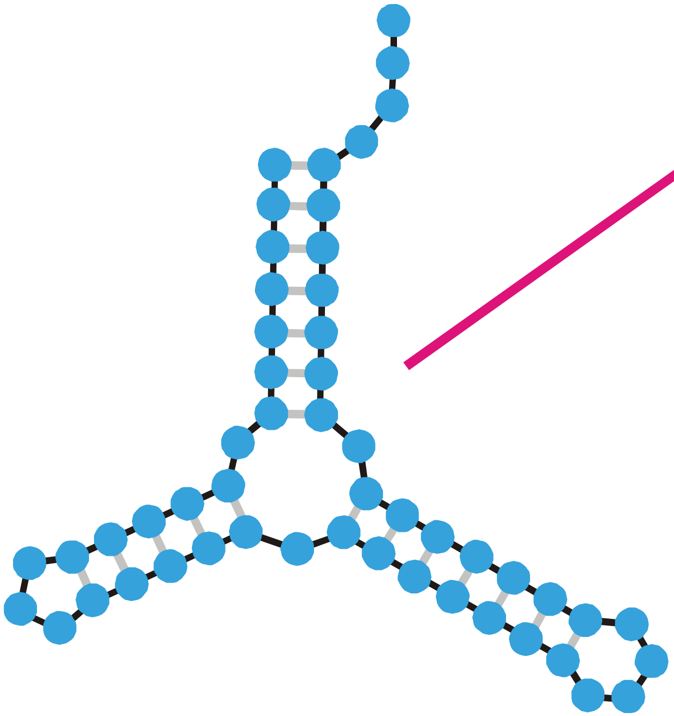
W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering***. Mh.Chem. **127** (1996), 375-389

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

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

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

GUAUCGAAAUACGUAGCGUAUGGGAUGCUGGACGGUCCCAUCGGUACUCCA



Minimum free energy criterion

Inverse folding of RNA secondary structures

The idea of inverse folding algorithm is to search for sequences that form a given RNA secondary structure under the minimum free energy criterion.

Inverse folding algorithm

$\mathbf{I_0}$ Á I_1 Á I_2 Á I_3 Á I_4 Á ... Á I_k Á I_{k+1} Á ... Á $\mathbf{I_t}$

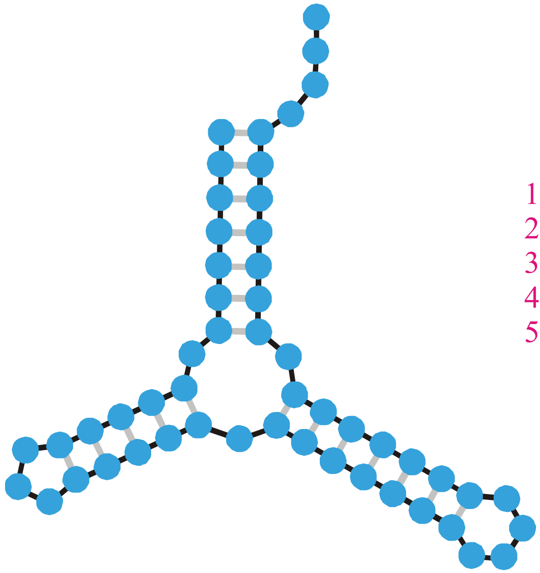
$\mathbf{S_0}$ Á S_1 Á S_2 Á S_3 Á S_4 Á ... Á S_k Á S_{k+1} Á ... Á $\mathbf{S_t}$

$I_{k+1} = \mathfrak{M}_k(I_k)$ and $Dd_S(S_k,S_{k+1}) = d_S(S_{k+1},S_t) - d_S(S_k,S_t) < 0$

$\mathfrak{M}$ ... base or base pair mutation operator

$d_S\,(S_i,S_j)$ ... distance between the two structures S_i and S_j

‚Unsuccessful trial‘ ... termination after n steps



Minimum free energy criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

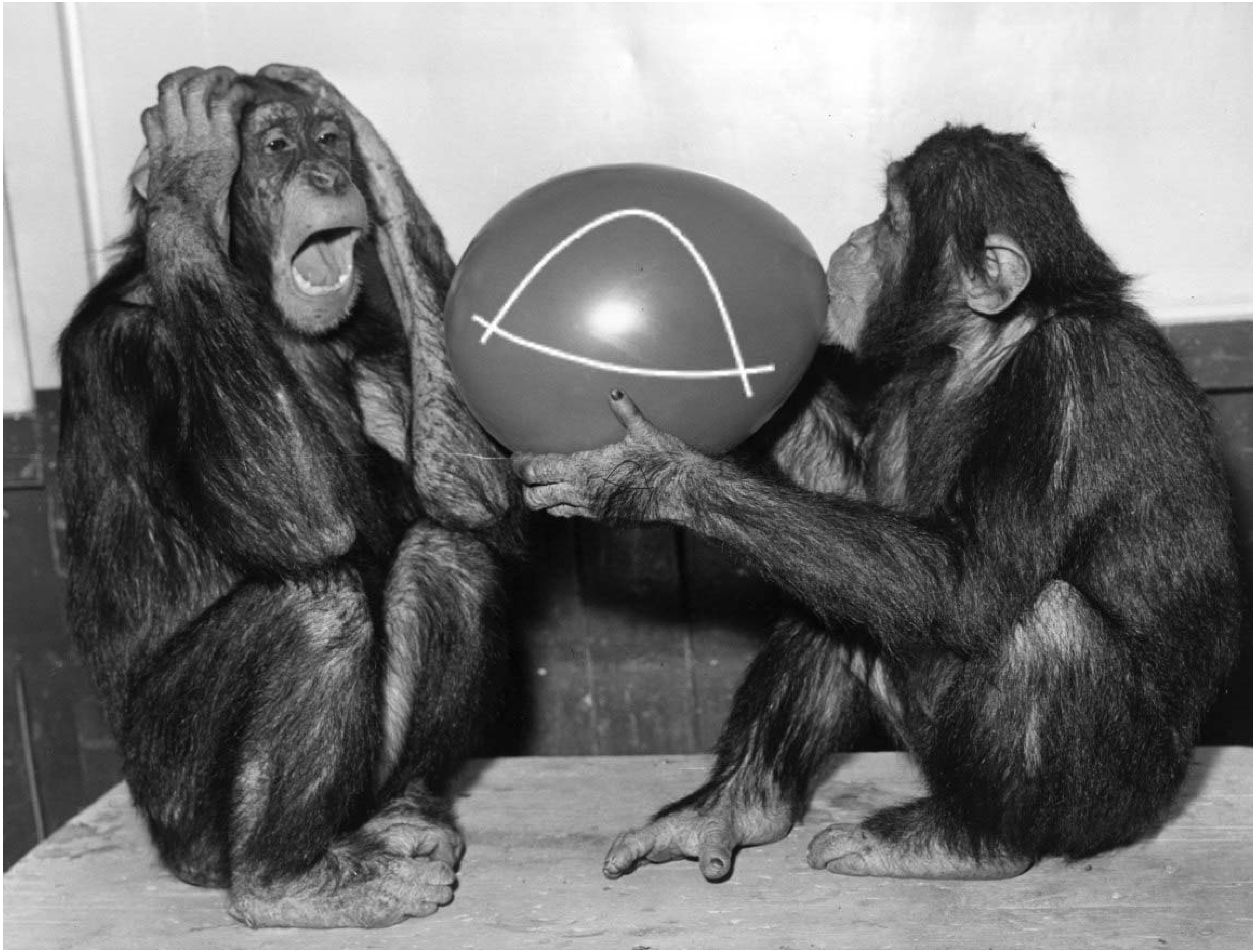
CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

Inverse folding of RNA secondary structures

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



Criterion of
Minimum Free Energy

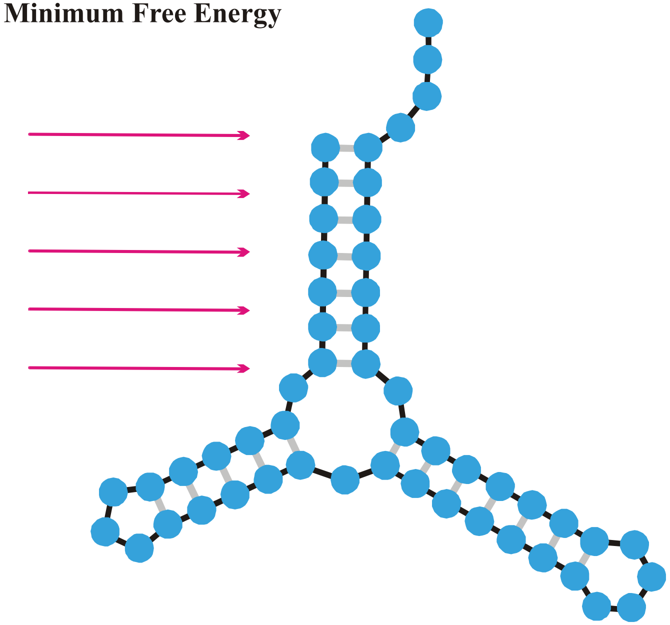
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

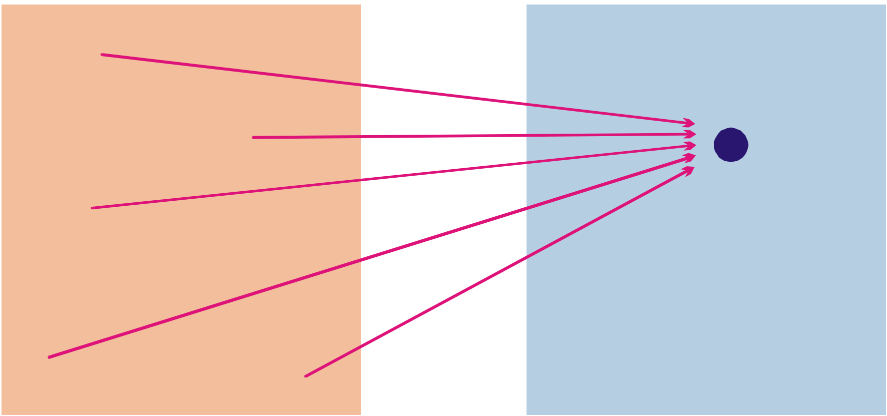
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

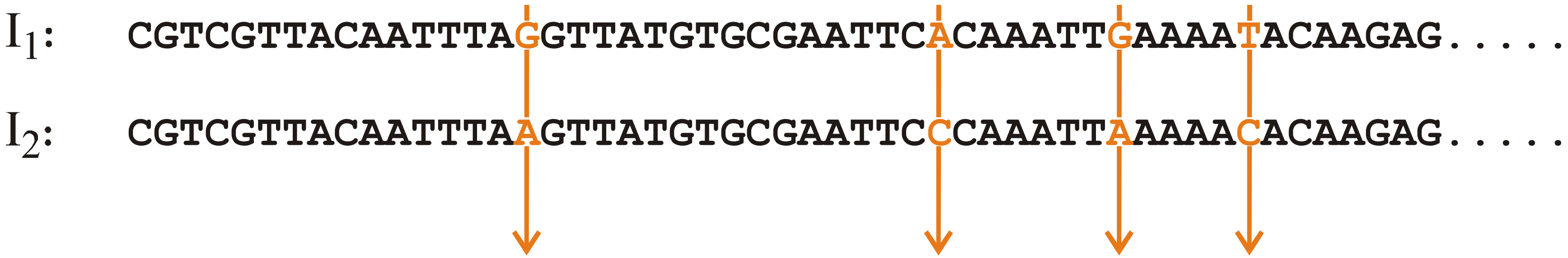
GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG





Sequence Space

Shape Space



I_1: CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....

I_2: CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAAACACAAGAG.....

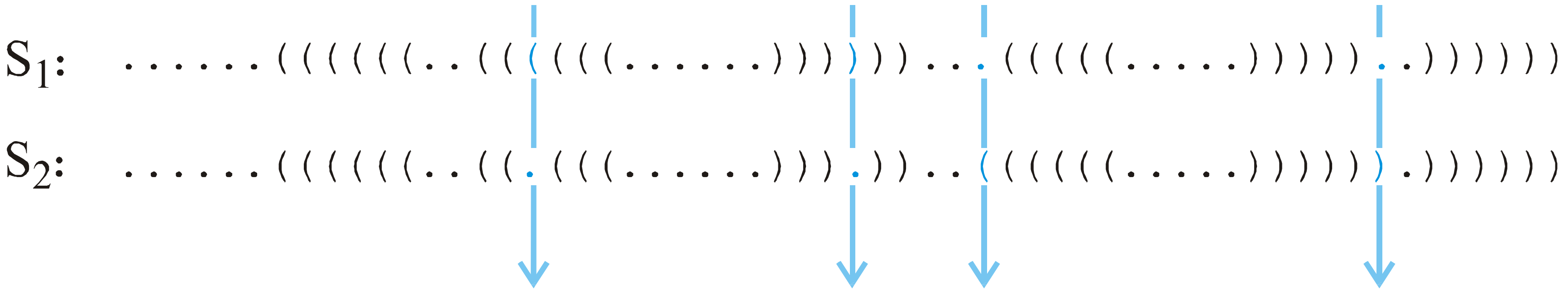
Hamming distance $d_H(I_1,I_2) = 4$

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

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

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

The Hamming distance between sequences induces a metric in sequence space



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

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

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

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

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

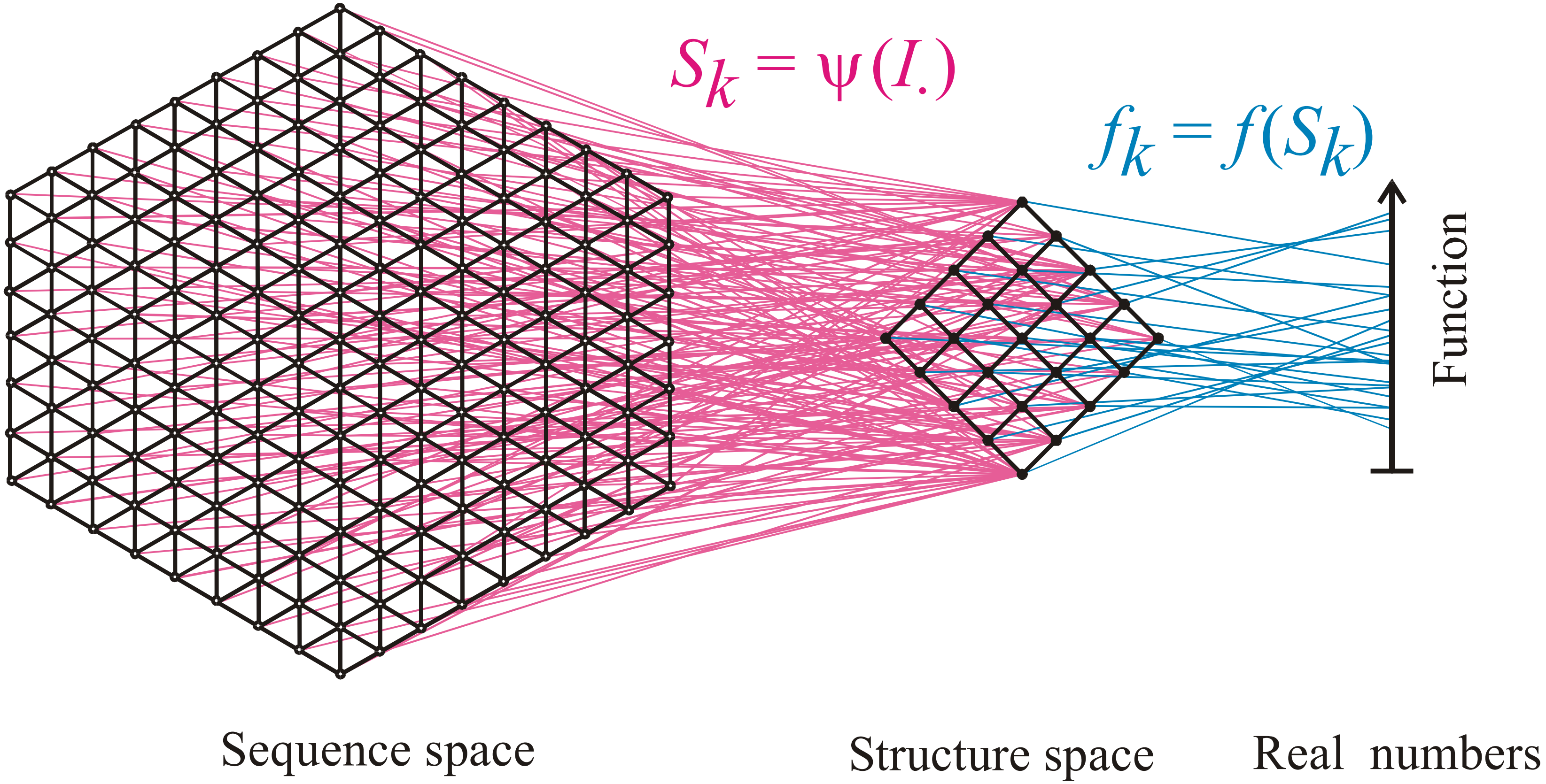
(iii) $d_H(S_1,S_3)$ ¶ $d_H(S_1,S_2) + d_H(S_2,S_3)$

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

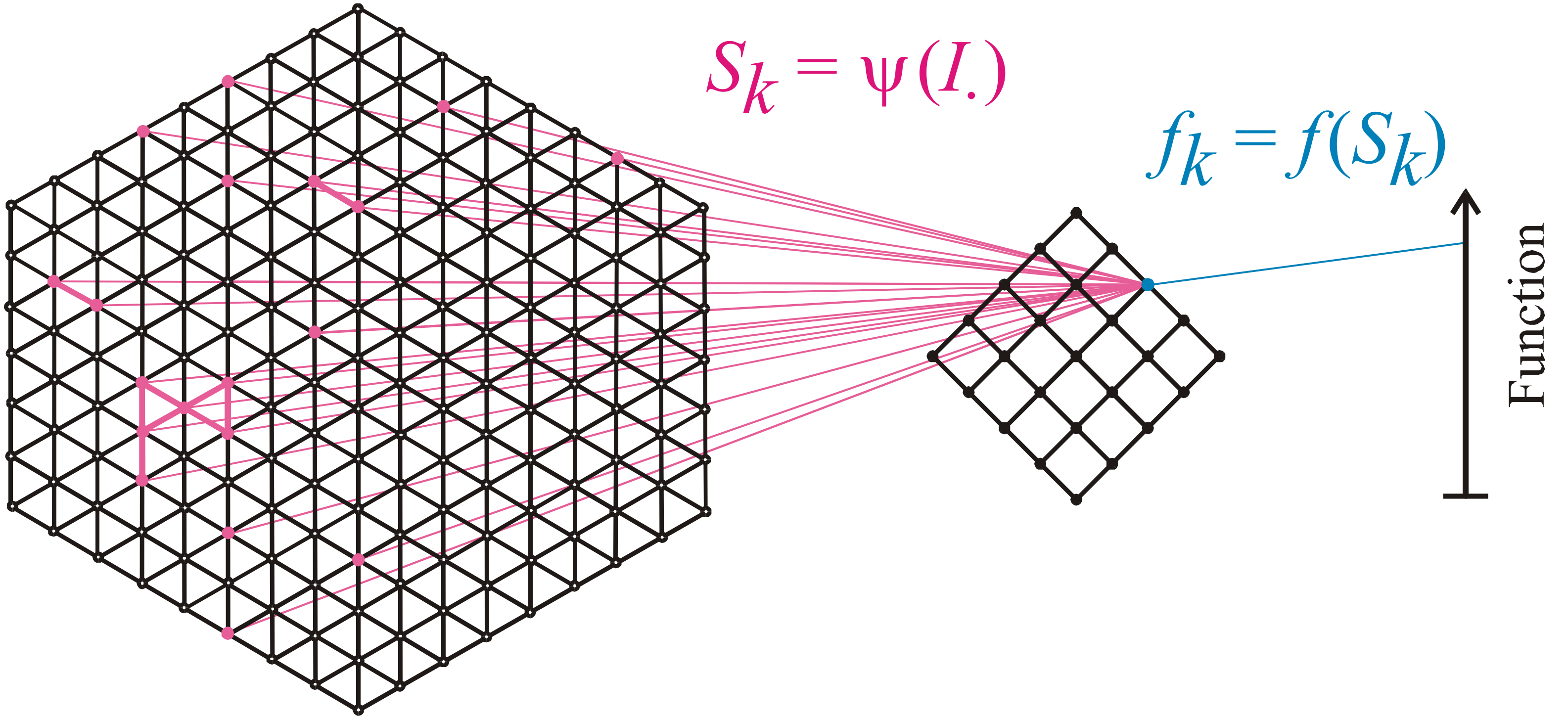
RNA **sequences** as well as RNA secondary **structures** can be visualized as objects in **metric spaces**. At constant chain length the sequence space is a (generalized) hypercube.

The **mapping** from RNA **sequences** into RNA secondary **structures** is many-to-one. Hence, it is redundant and not invertible.

RNA **sequences**, which are mapped onto the same RNA secondary **structure**, are **neutral** with respect to **structure**. The pre-images of structures in sequence space are **neutral networks**. They can be represented by graphs where the edges connect sequences of Hamming distance $d_H = 1$.



Mapping from sequence space into structure space and into function



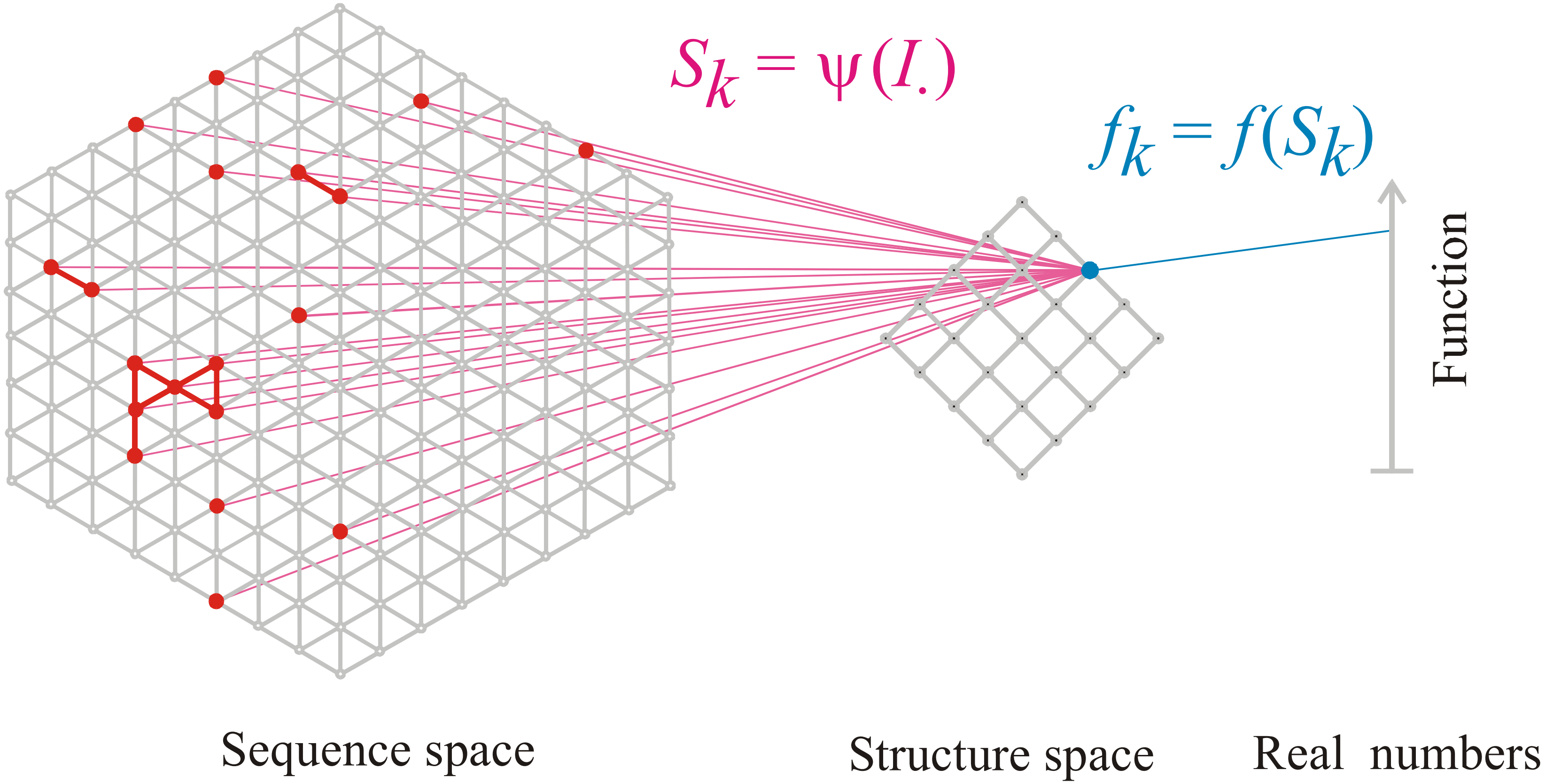
$S_k = \psi(I.)$

$f_k = f(S_k)$

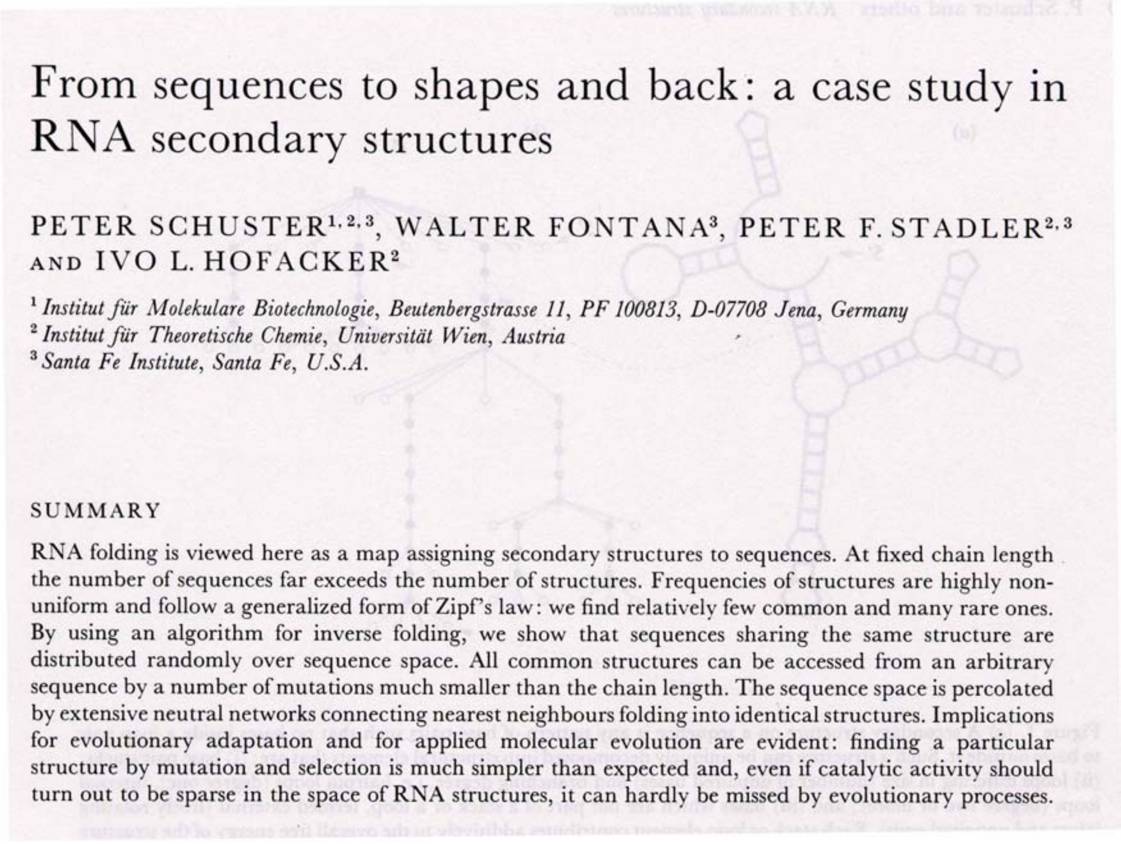
Sequence space

Structure space

Real numbers



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



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

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

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

SUMMARY

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

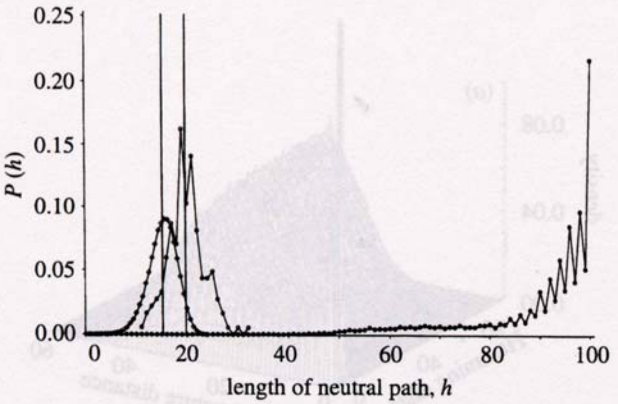
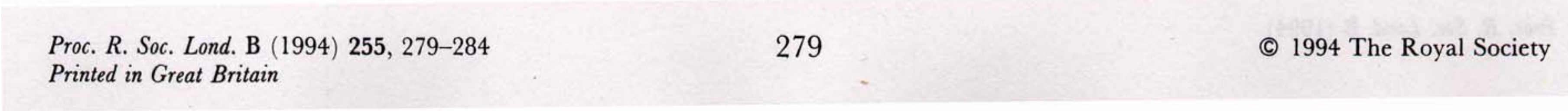


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



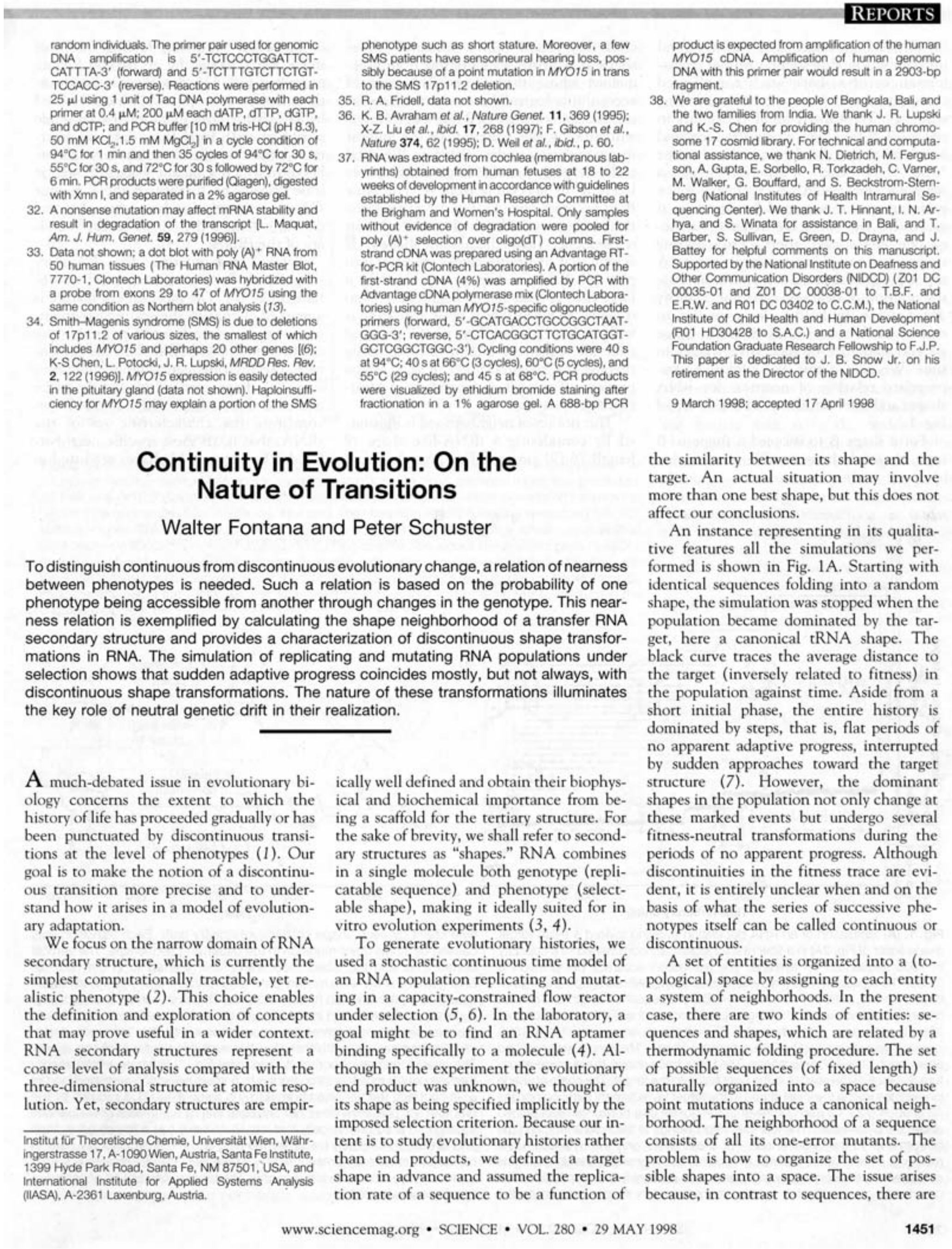
Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

279

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

Evolution *in silico*

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



random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCTCATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGTTCCACC-3′ (reverse). Reactions were performed in 25 μl using 1 unit of Taq DNA polymerase with each primer at 0.4 μM; 200 μM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl_2,1.5 mM $MgCl_2$] 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, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.
35. R. A. Fridell, data not shown.
36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.
37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly $(A)^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGACCTGCCGGCTAATGGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGTGCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

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

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

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

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

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

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

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

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

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

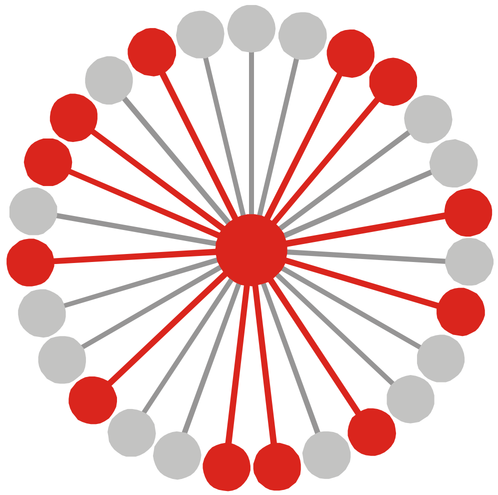
Neutral networks are sets of sequences forming the same structure. **G_k** is the pre-image of the structure S_k in sequence space:

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

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

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

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



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

$$\lambda_j = 12 / 27 = 0.444 \,, \quad \bar{\lambda}_k = \frac{\underset{j \in |G_k|}{\text{ø}} \; \text{l}_j(k)}{|G_k|}$$

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

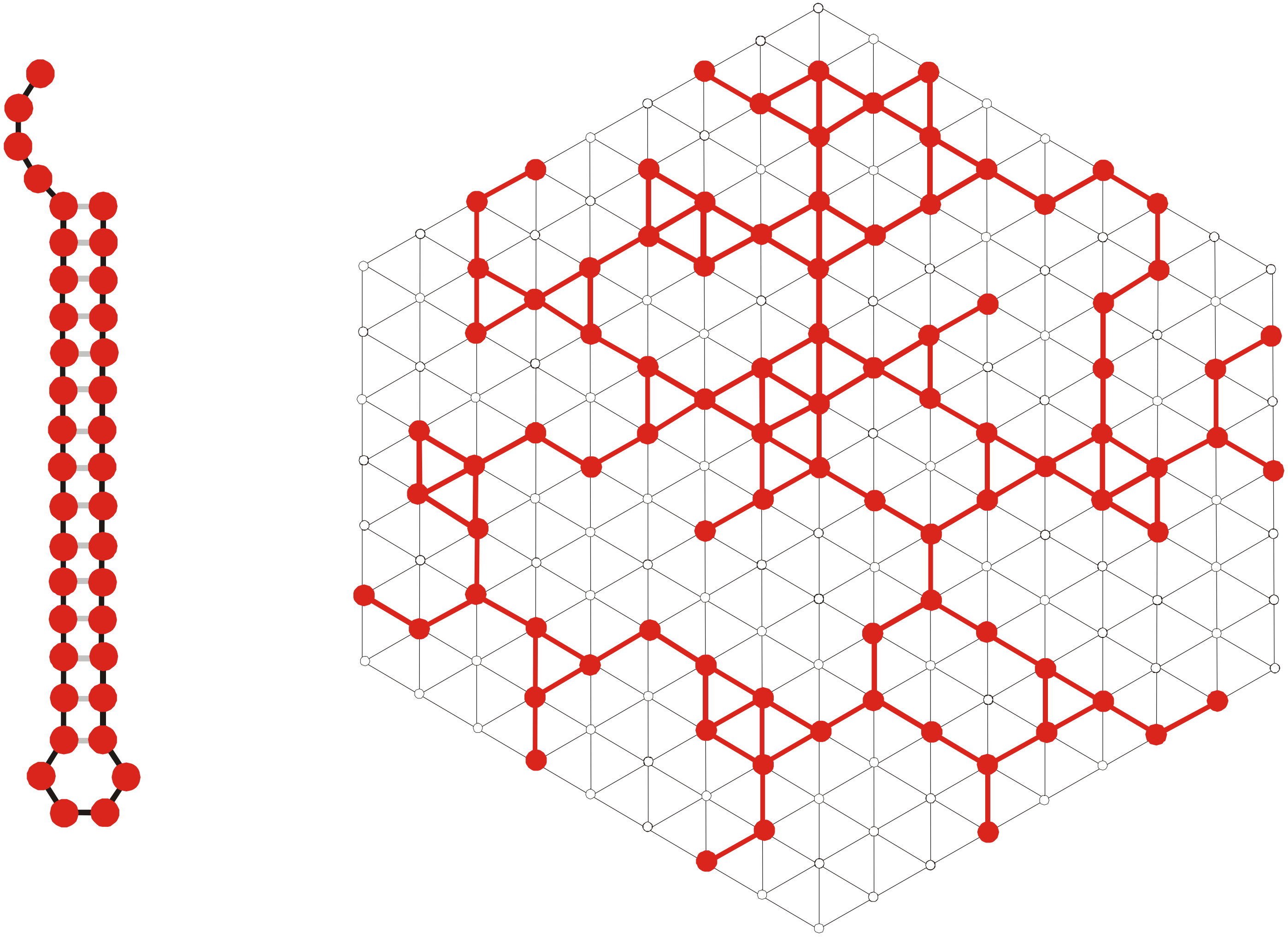
Alphabet size k : **AUGC** ñ k = 4

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

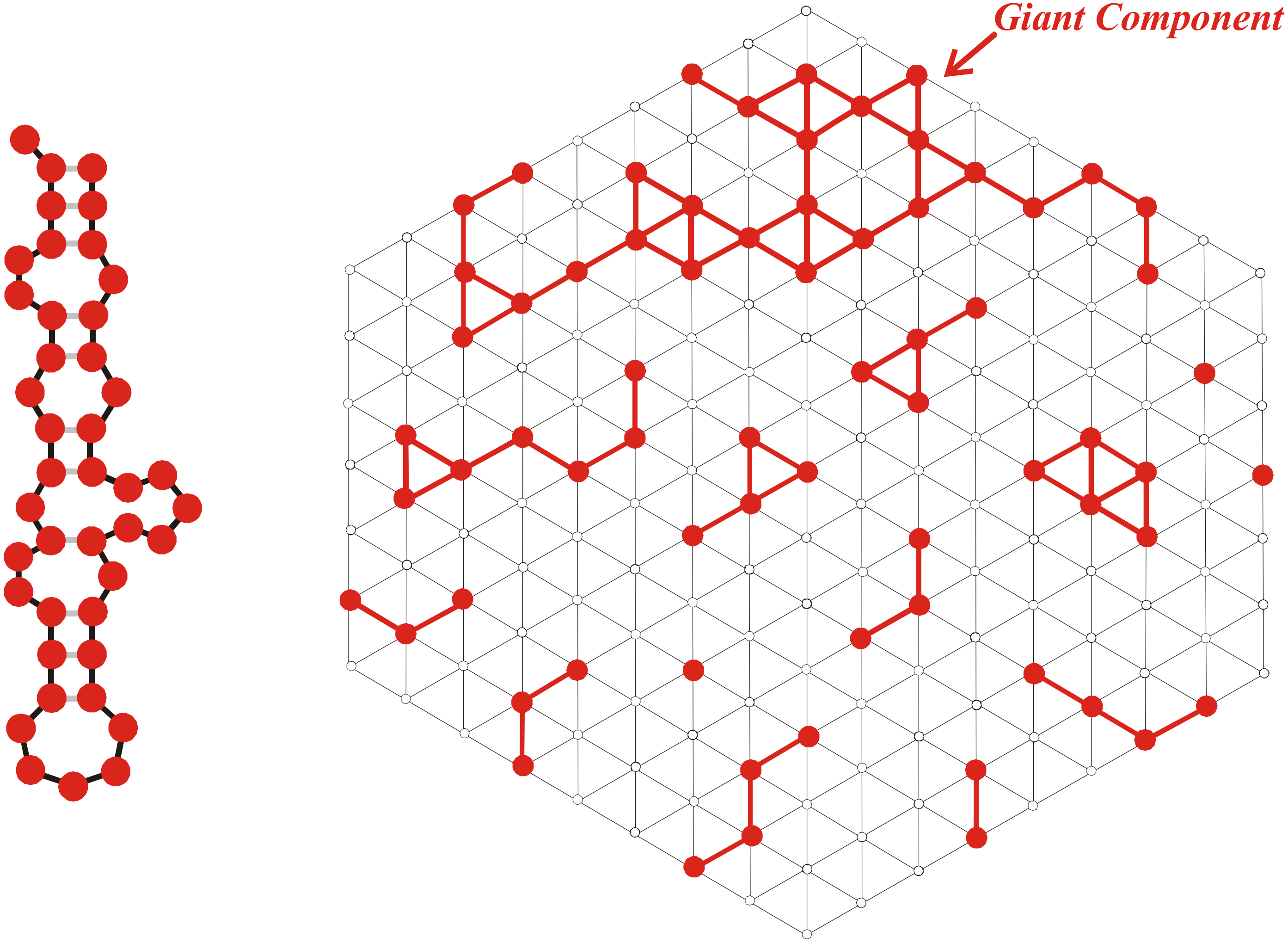
$\bar{\lambda}_k < \lambda_{cr}$ . . . . network G_k is **not** connected

k	l_{cr}	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

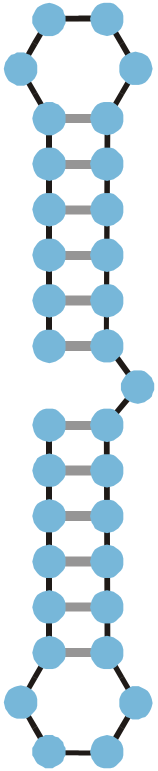
Mean degree of neutrality and connectivity of neutral networks



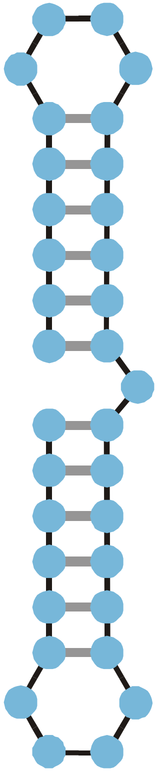
A connected neutral network



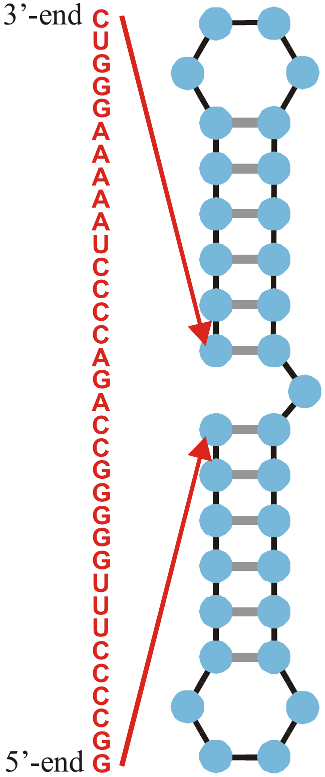
A multi-component neutral network



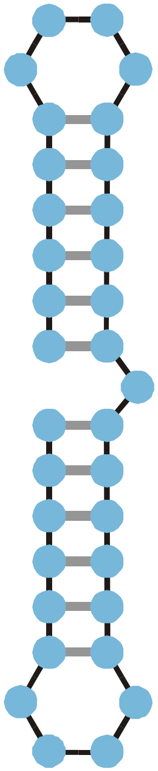
Structure



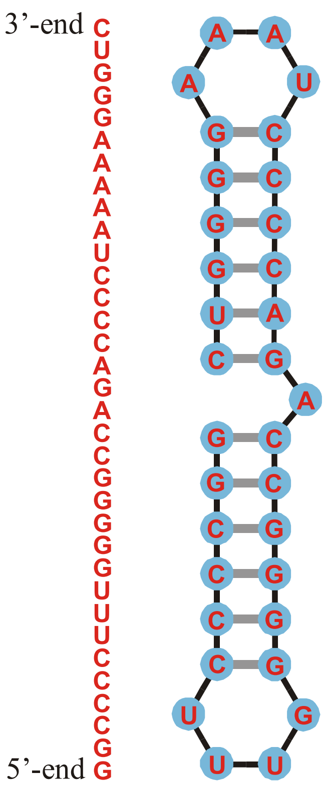
Structure



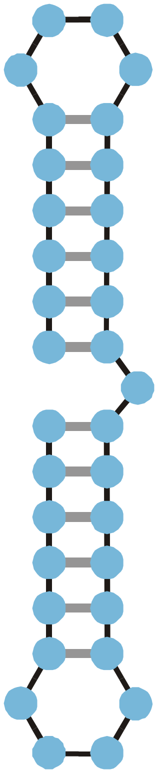
Compatible sequence



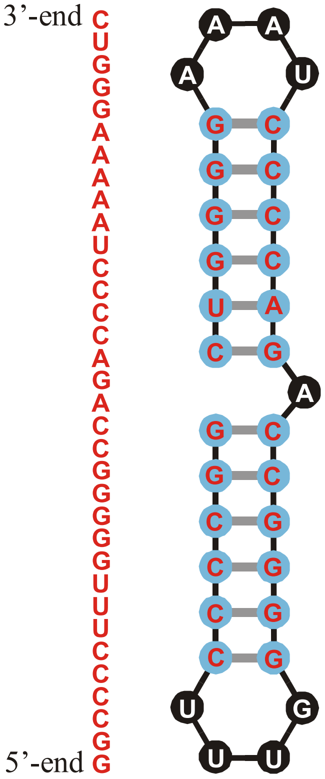
Structure


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

Compatible sequence



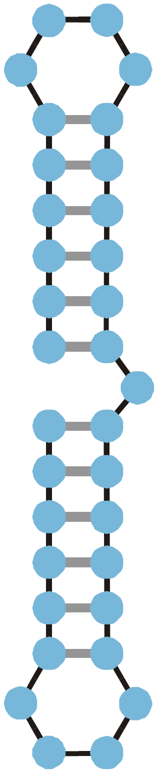
Structure


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

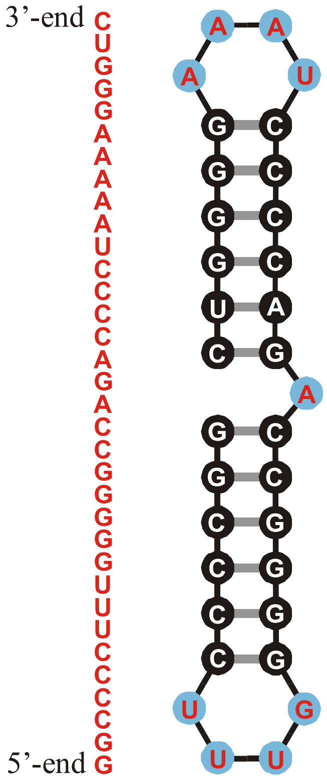
Single nucleotides: A,U,G,C

Compatible sequence

Single bases pairs are varied independently



Structure

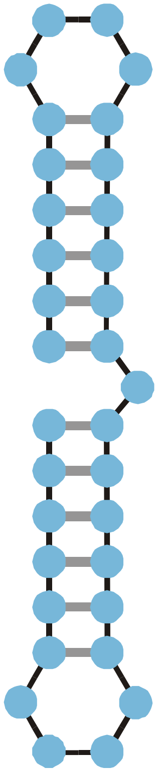


Compatible sequence

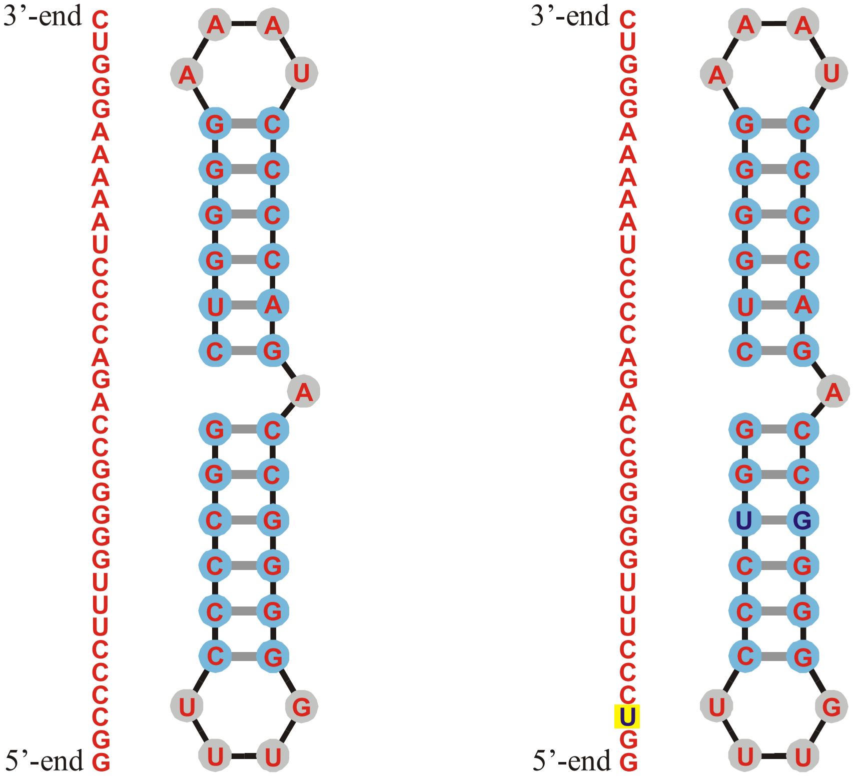
Base pairs:

AU , UA
GC , CG
GU , UG

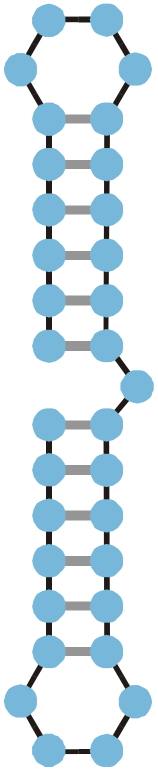
Base pairs are varied in strict correlation



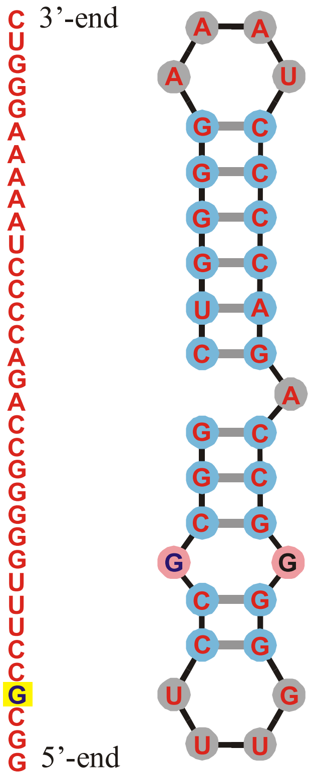
Structure


3’-end
CUGGGAAAAAUCCCCAGACCGGGGUUUCCCCGG
5’-end
3’-end
CUGGGAAAAAUCCCCAGACCGGGGUUUCCCUGG
5’-end

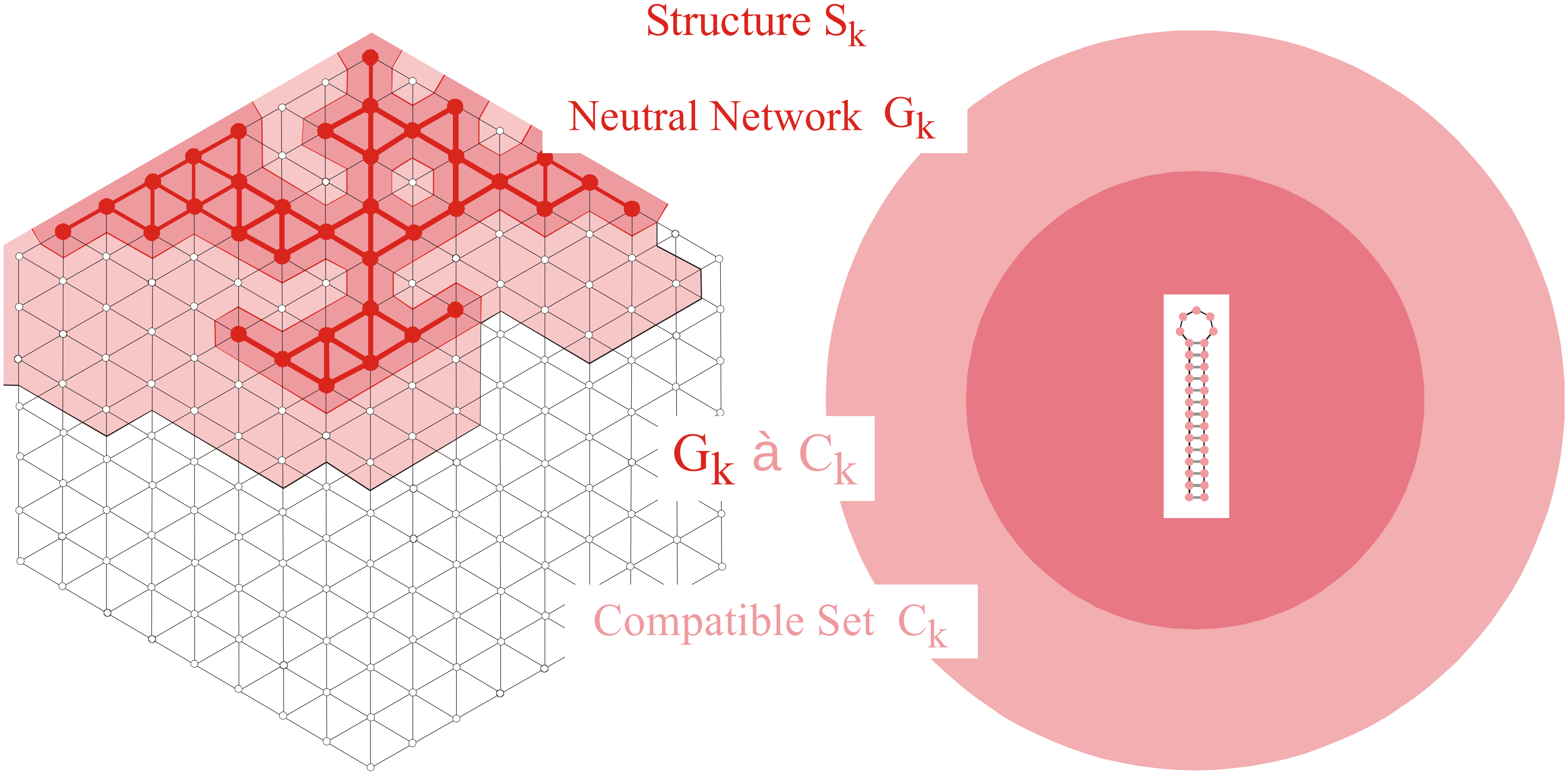
Compatible sequences



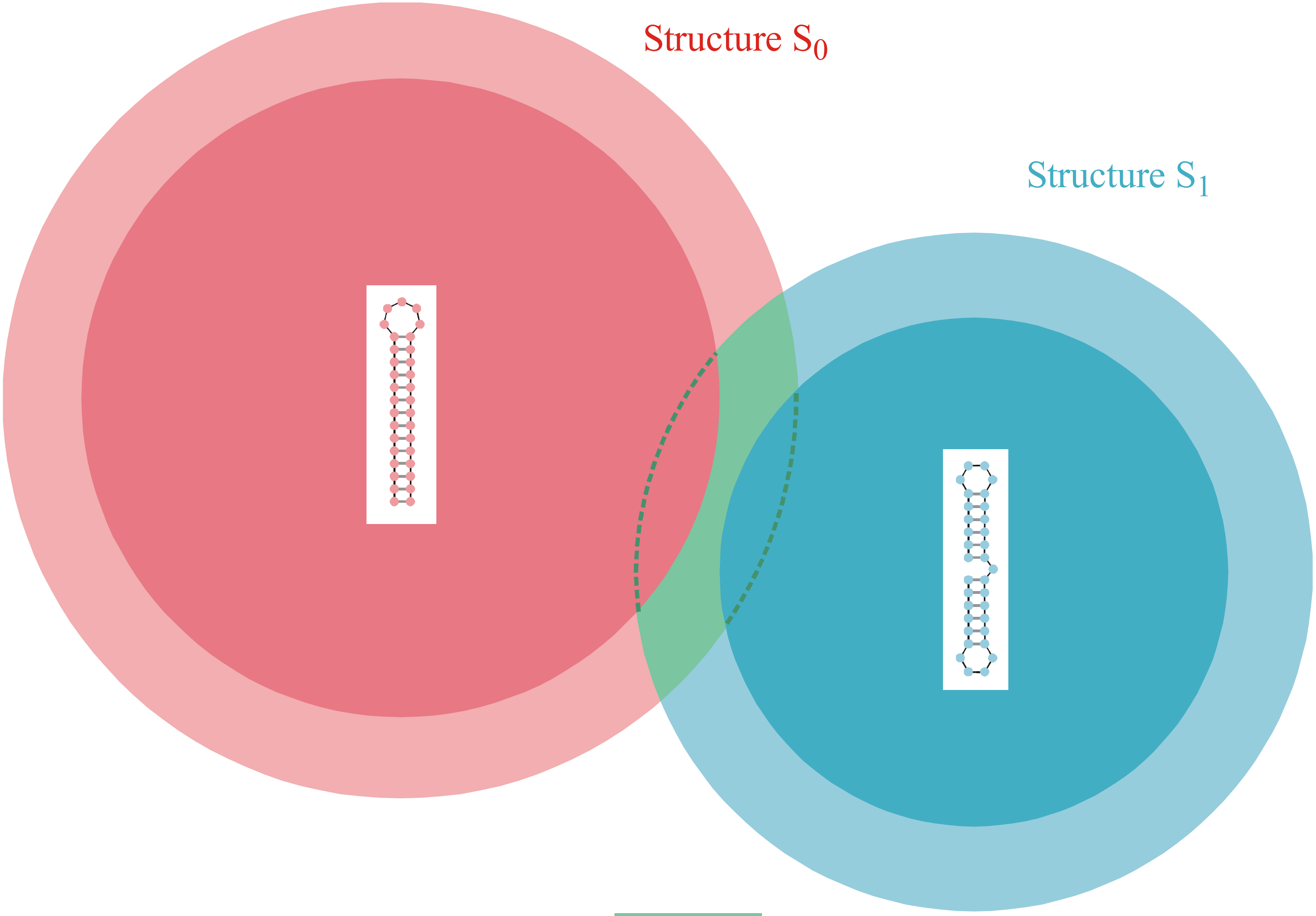
Structure



Incompatible sequence

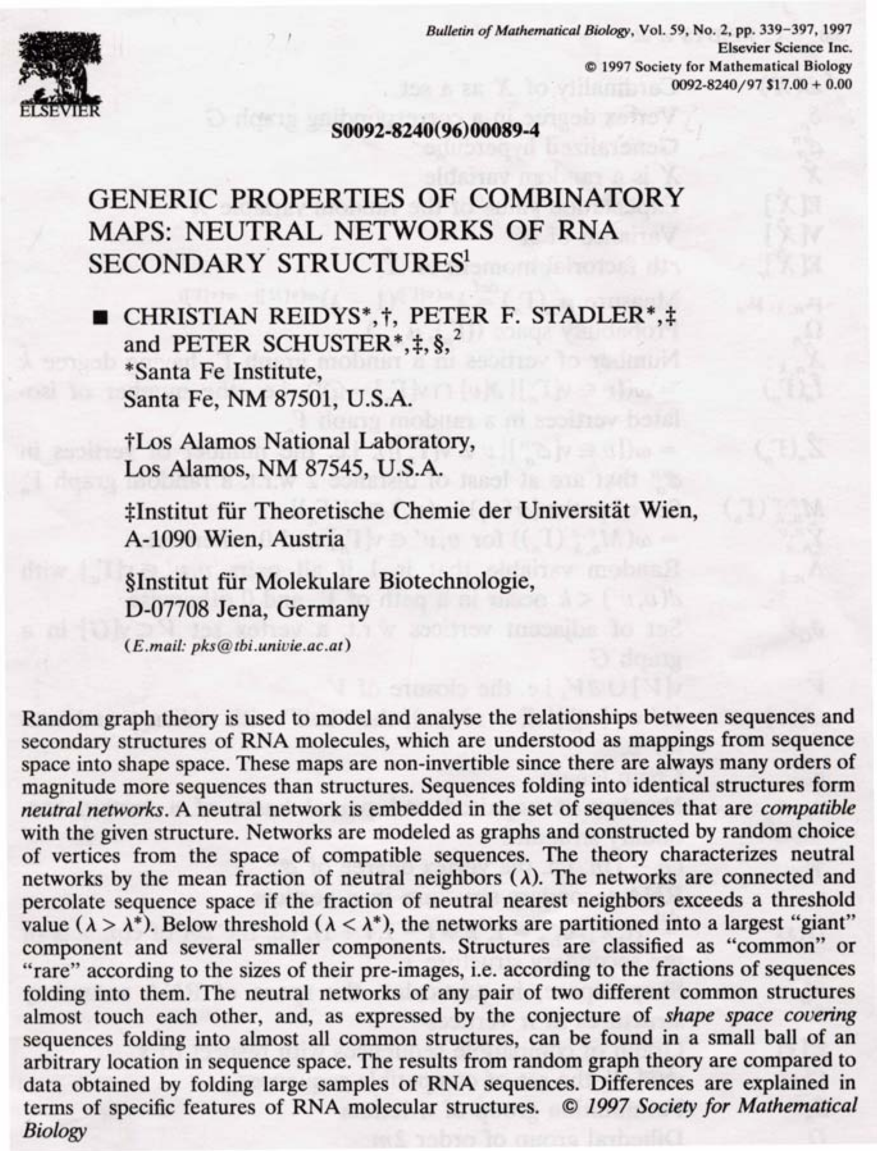


The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



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

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



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

S0092-8240(96)00089-4

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

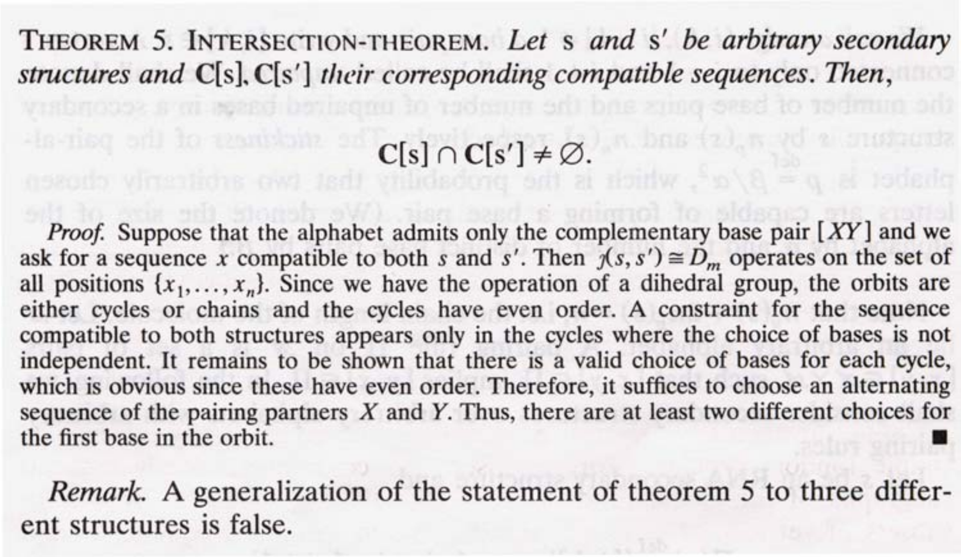
■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡
and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

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

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

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany
(*E.mail: pks@tbi.univie.ac.at*)

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



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

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

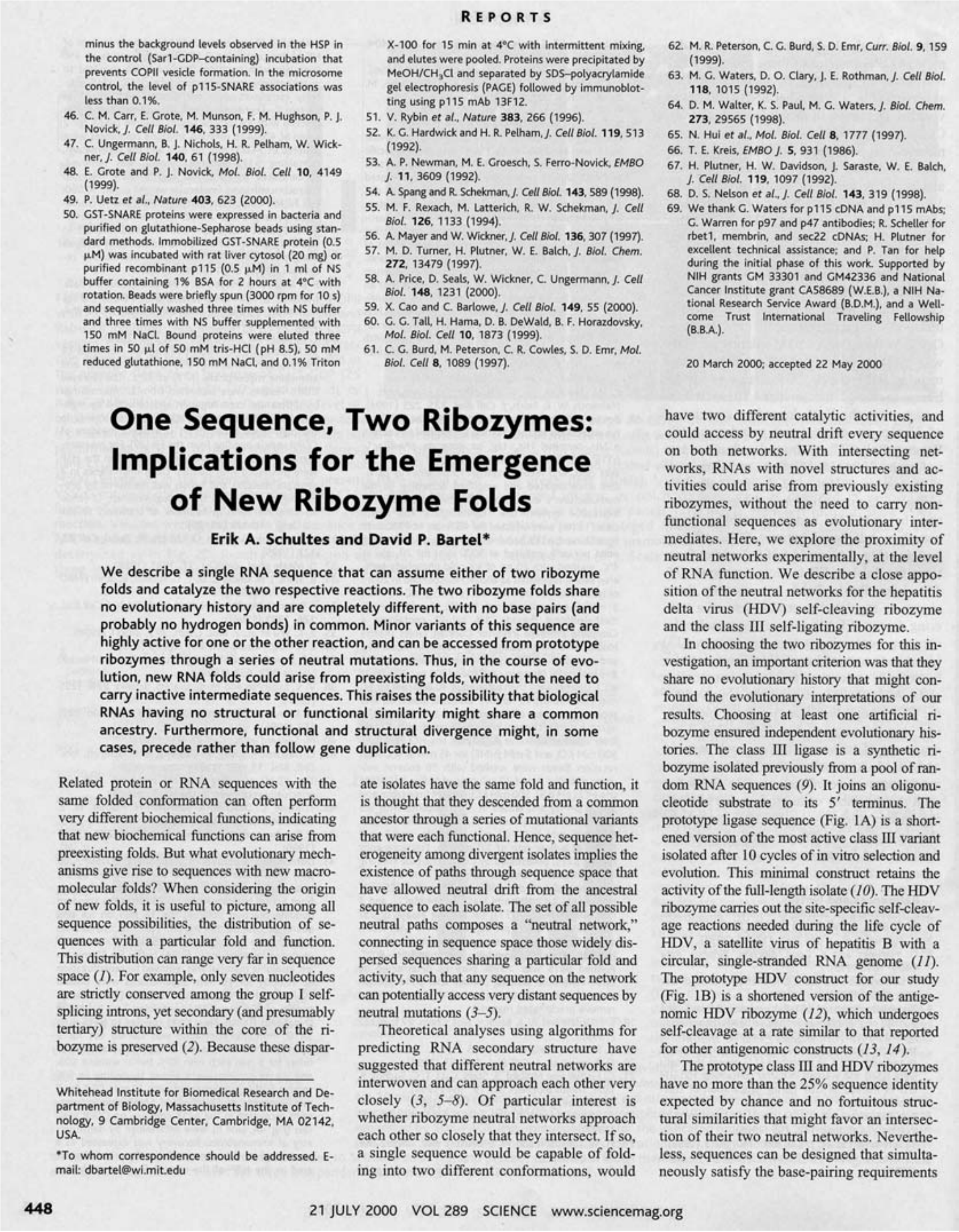
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\mathcal{J}(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either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

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

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

A ribozyme switch

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



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

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

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

20 March 2000; accepted 22 May 2000

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

Erik A. Schultes and David P. Bartel*

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

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

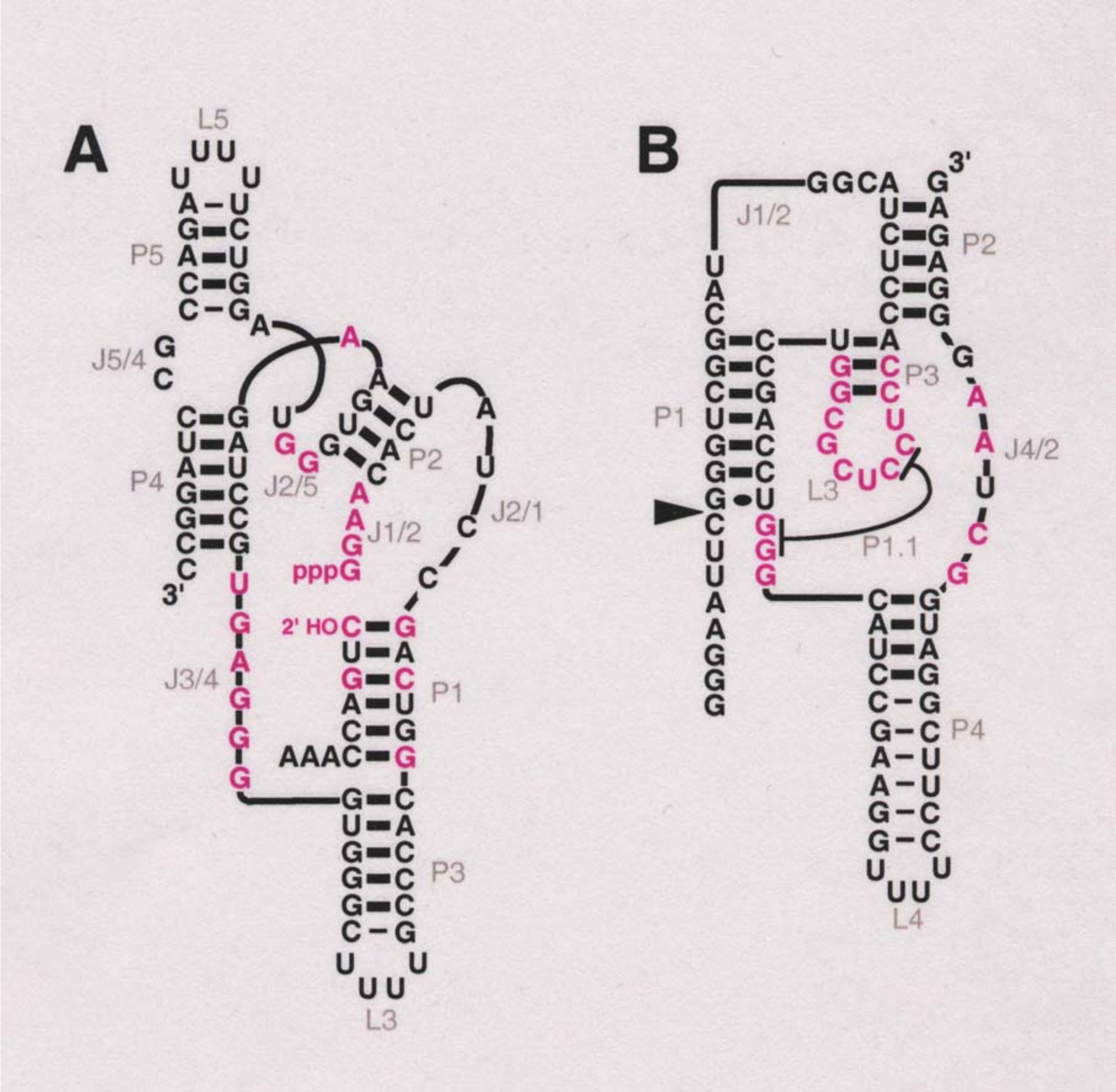
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

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

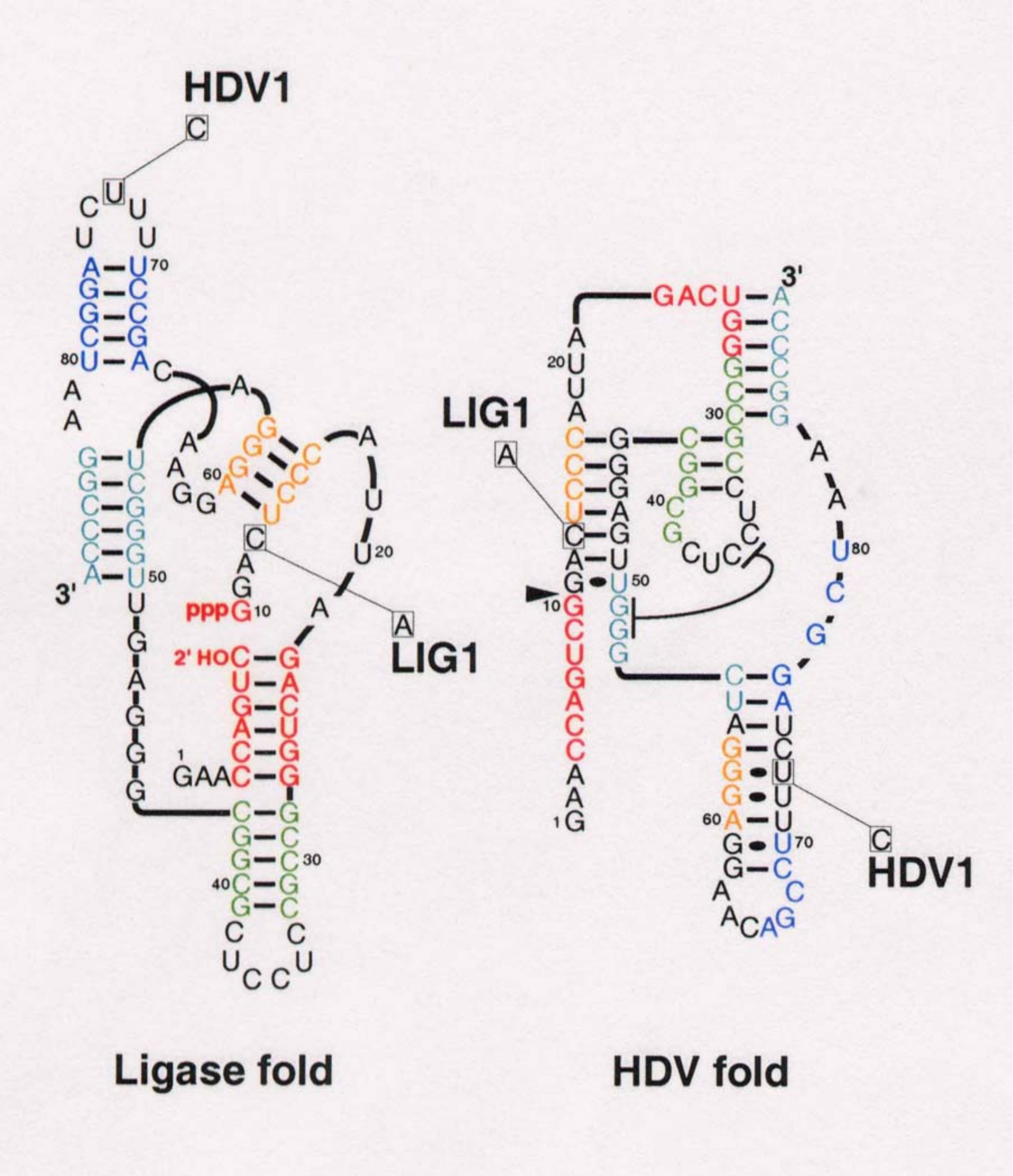
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

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

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

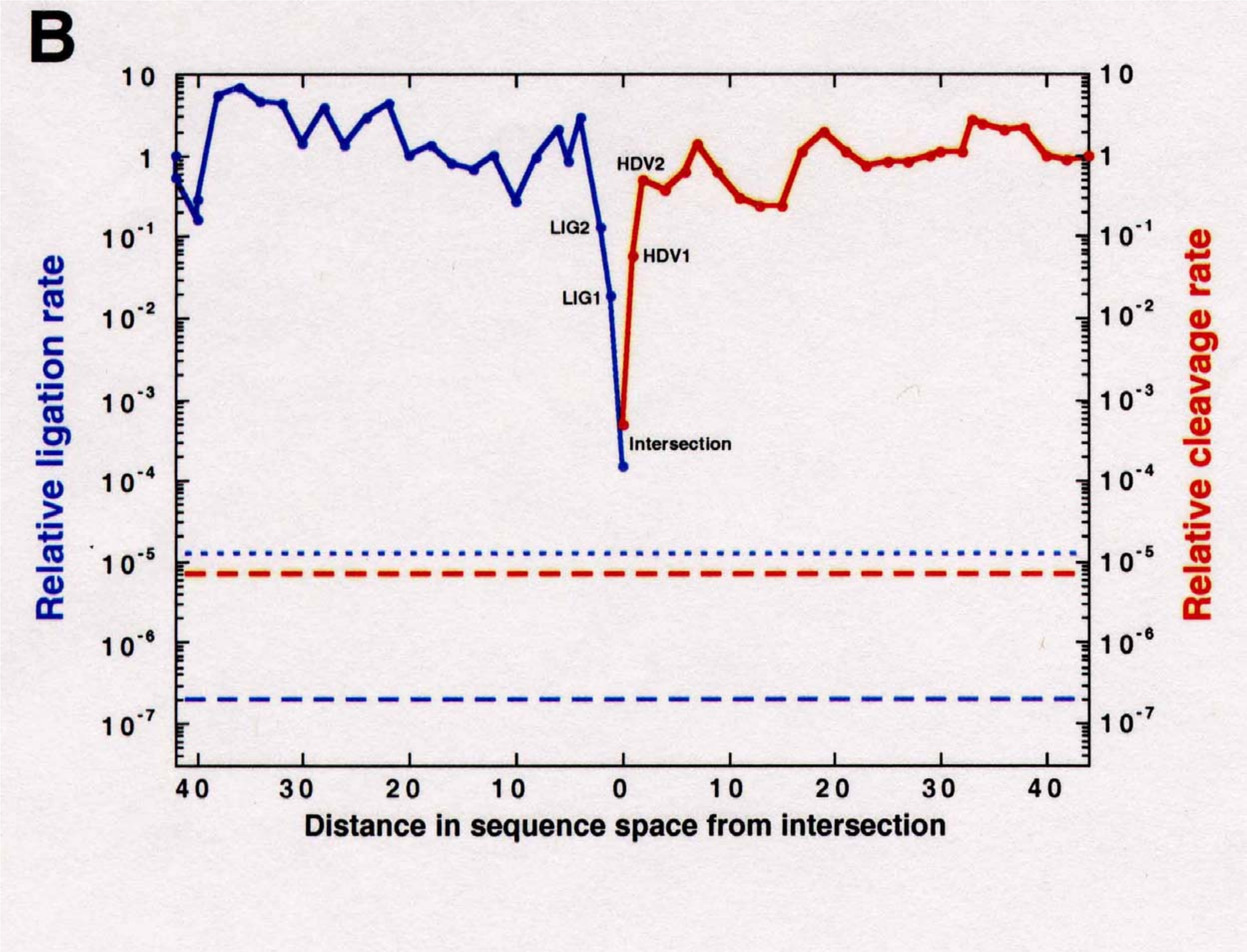


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)

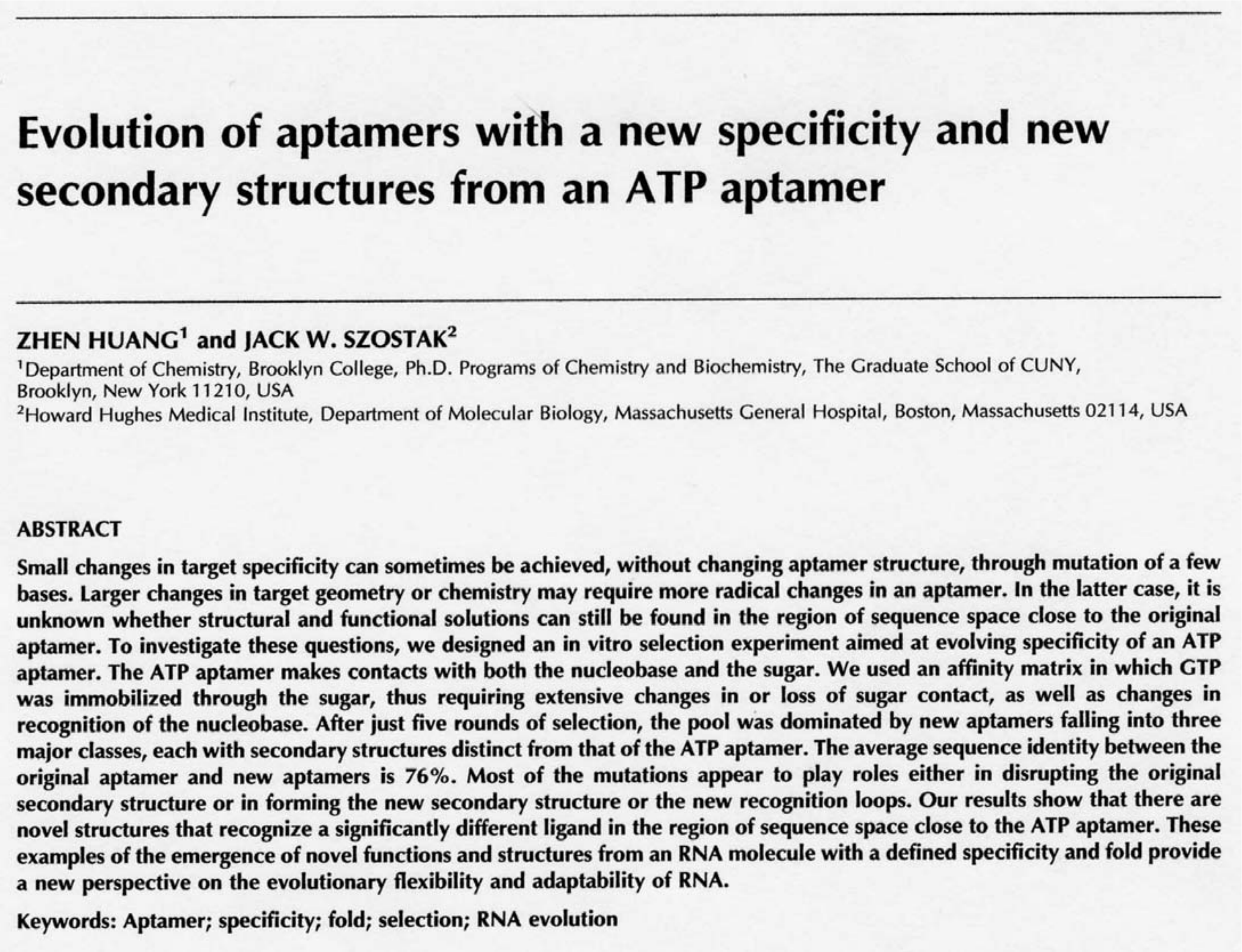


The sequence at the ***intersection***:

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



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



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

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

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

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

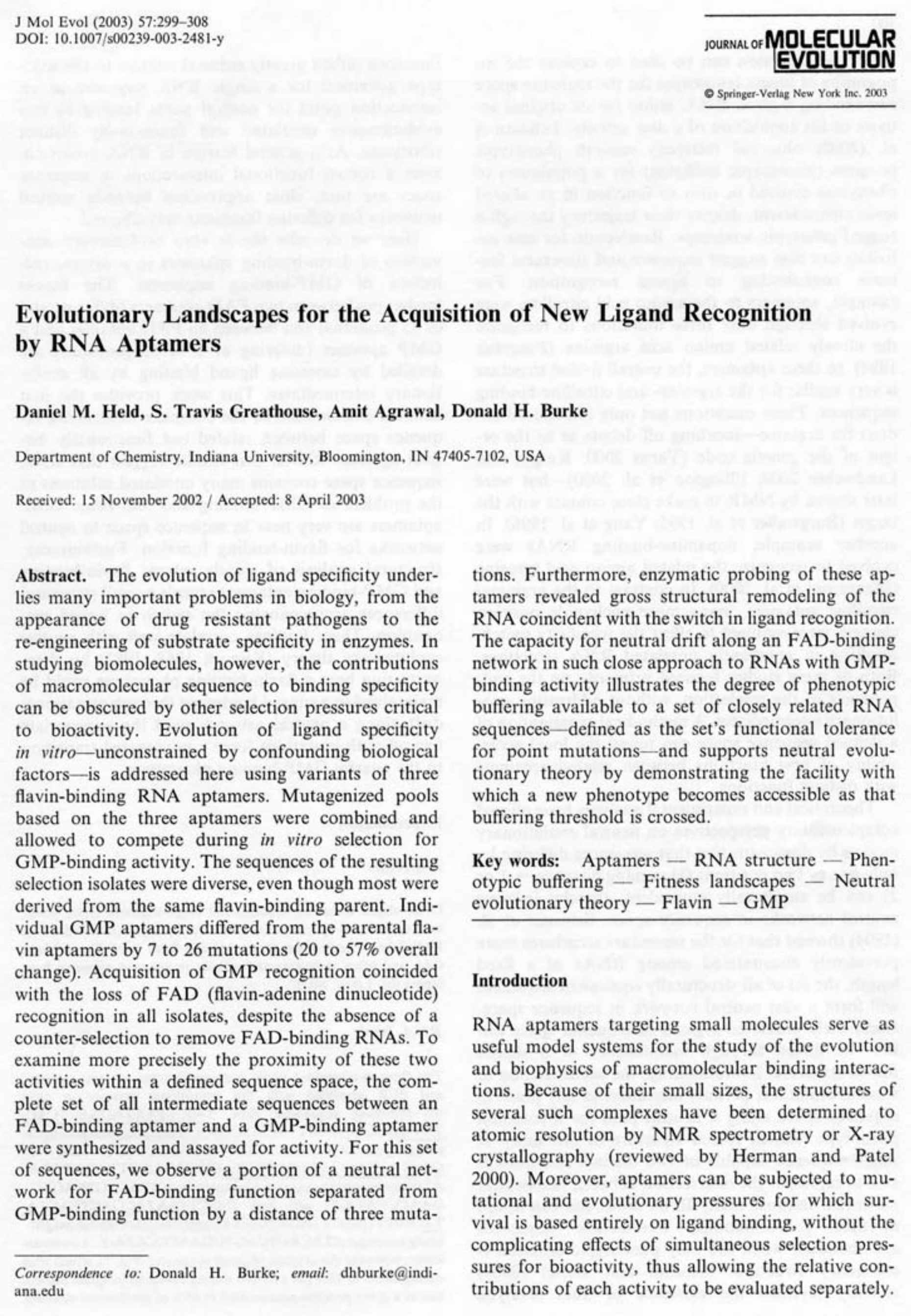
ABSTRACT

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

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

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

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



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

JOURNAL OF MOLECULAR EVOLUTION

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

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

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

Received: 15 November 2002 / Accepted: 8 April 2003

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

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

Introduction

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

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

Evidence for **neutral networks** and **intersection** of apatamer functions

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

Austrian Genome Research Program – GEN-AU

Siemens AG, Austria

The Santa Fe Institute and the 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

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

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

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

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

Web-Page for further information:

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