Designing Single and Double Stranded Nucleic Acids

Peter Schuster

Institut für Theoretische Chemie, Abteilung Theoretische Biochemie
Universität Wien, Austria

Engineering a DNA World

Pasadena, Jan. 06-08, 2005

Web-Page for further information:

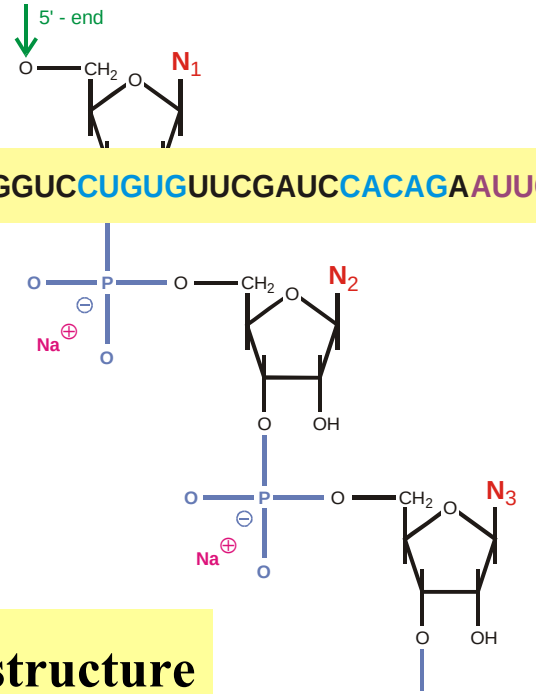
http://www.tbi.univie.ac.at/~pks

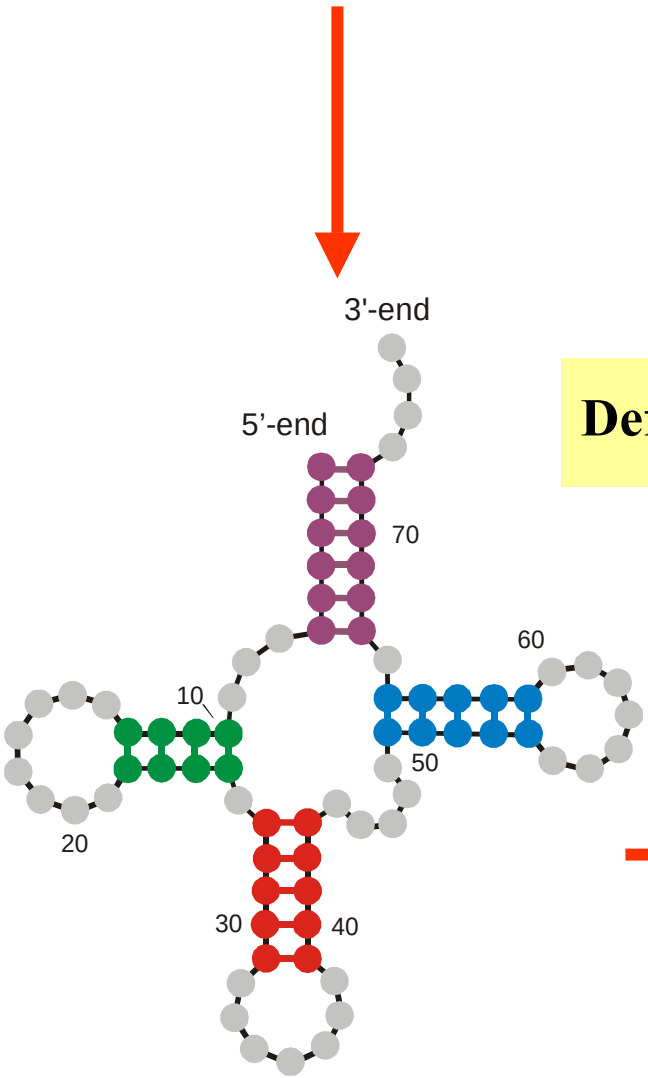
1. One sequence – one structure problem
2. Inverse folding and neutral networks
3. Kinetic folding
4. Intersections and conformational switches
5. Cofolding of nucleic acid molecules

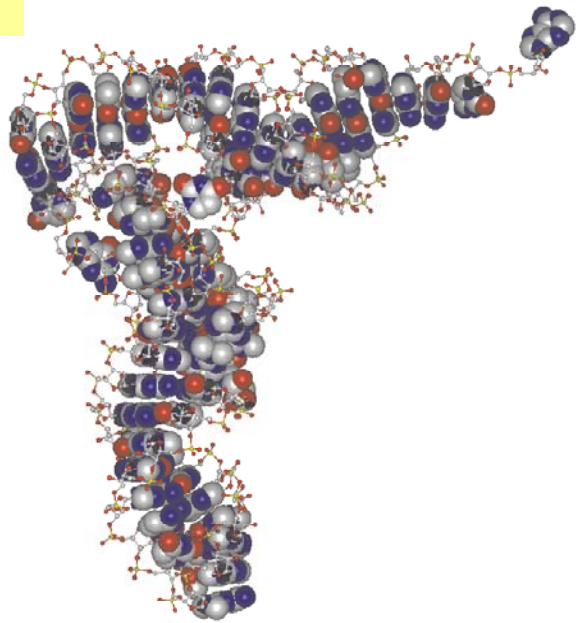
1. **One sequence – one structure problem**
2. Inverse folding and neutral networks
3. Kinetic folding
4. Intersections and conformational switches
5. Cofolding of nucleic acid molecules

Definition of RNA structure

5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end



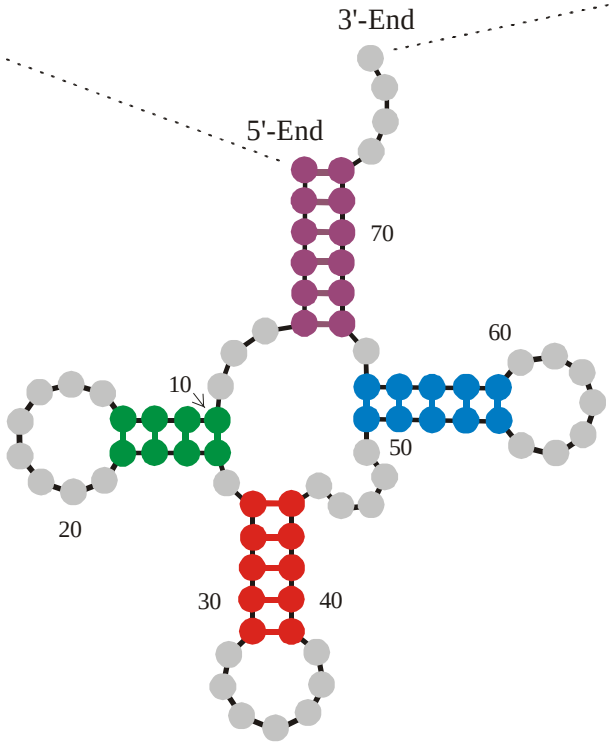




5'-End 3'-End

Sequence GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

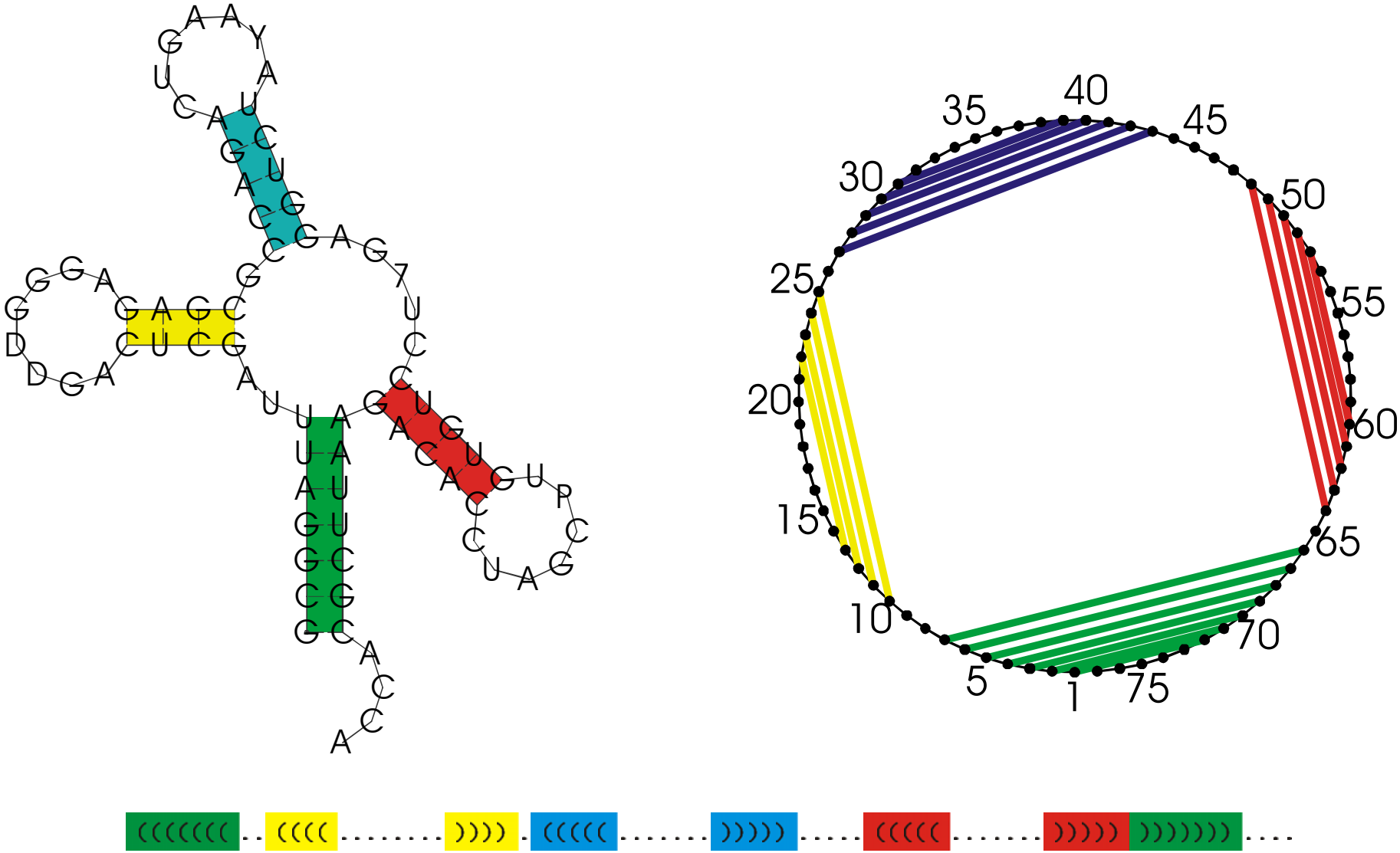
Secondary structure



Symbolic notation (((((((•••((((••••••••))))•(((((•••••••)))))•••••(((((•••••••)))))•)))))))••••

5'-End 3'-End

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



The tRNA^{phe} in the circular and symbolic representation

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**."

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

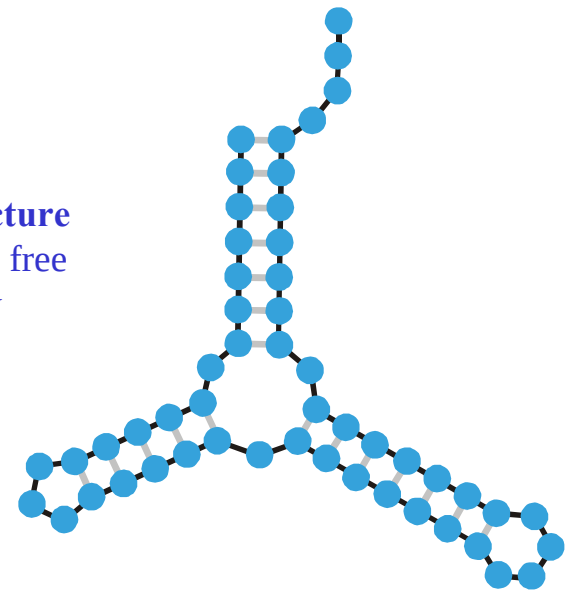
RNA folding:

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

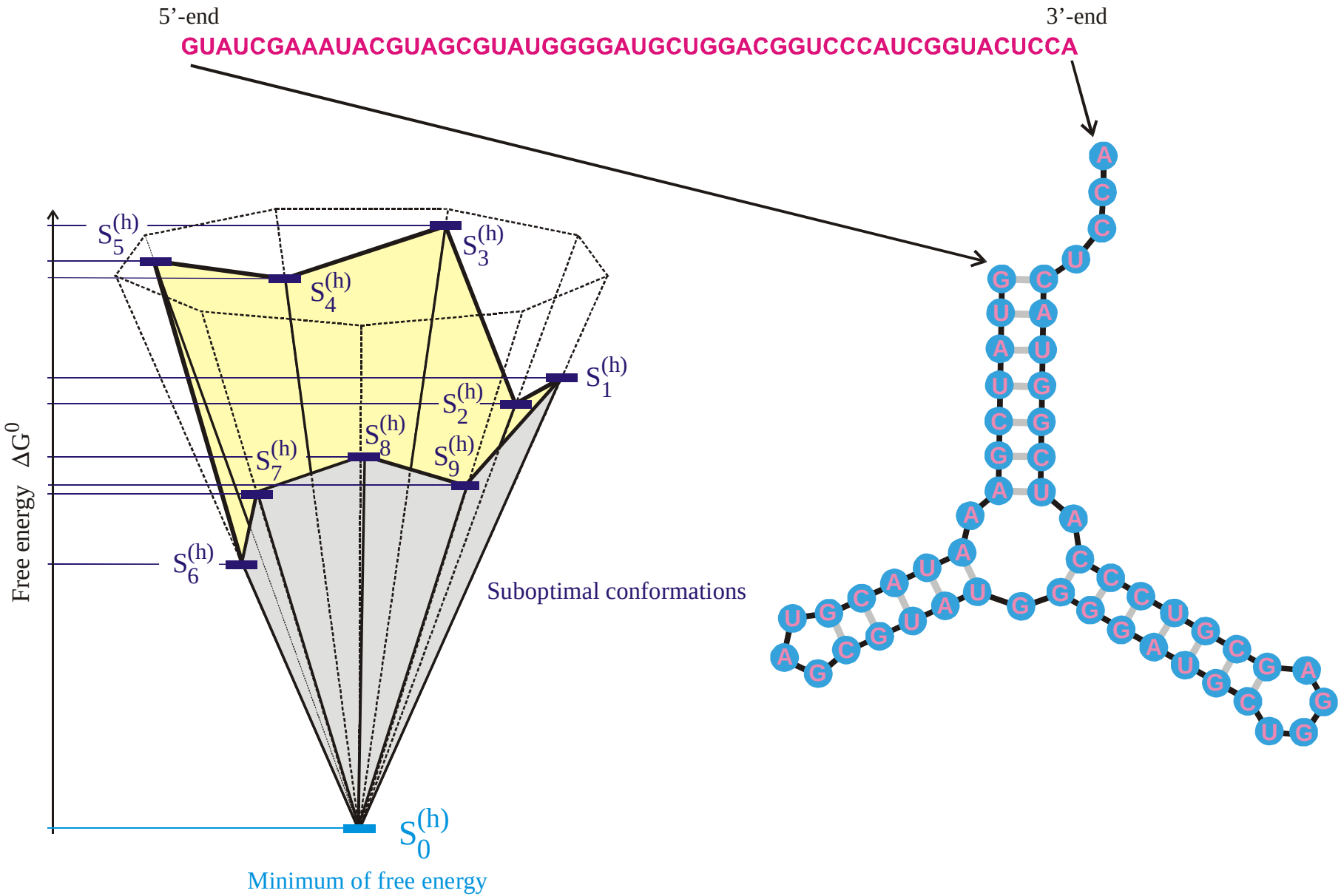
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

RNA structure of minimal free energy



One sequence – one structure problem



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 secondary structures for given sequences.

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

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

Update of empirical thermodynamic parameters:

D. H. Mathews, J. Sabina, M. Zuker, D.H. Turner. ***J.Mol.Biol.*** **288**:911-940 (1999)

The **Vienna RNA Package:**

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

Access to the **Vienna RNA Package**: **http://www.tbi.univie.ac.at/**

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***:

A library of routines for folding, ***inverse folding***, sequence and structure alignment, ***kinetic folding***, ***cofolding***, …

Monatshefte für Chemie 125, 167–188 (1994)

Monatshefte für Chemie
Chemical Monthly

Fast Folding and Comparison of RNA Secondary Structures

I. L. Hofacker[1,*], W. Fontana[3], P. F. Stadler[1,3], L. S. Bonhoeffer[4], M. Tacker[1] and P. Schuster[1,2,3]

[1] Institut für Theoretische Chemie, Universität Wien, A-1090 Wien, Austria
[2] Institut für Molekulare Biotechnologie, D-07745 Jena, Federal Republic of Germany
[3] Santa Fe Institute, Santa Fe, NM 87501, U.S.A.
[4] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.

Summary. Computer codes for computation and comparison of RNA secondary structures, the Vienna RNA package, are presented, that are based on dynamic programming algorithms and aim at predictions of structures with minimum free energies as well as at computations of the equilibrium partition functions and base pairing probabilities.

An efficient heuristic for the inverse folding problem of RNA is introduced. In addition we present compact and efficient programs for the comparison of RNA secondary structures based on tree editing and alignment.

All computer codes are written in ANSI C. They include implementations of modified algorithms on parallel computers with distributed memory. Performance analysis carried out on an Intel Hypercube shows that parallel computing becomes gradually more and more efficient the longer the sequences are.

Keywords. Inverse folding; parallel computing; public domain software; RNA folding; RNA secondary structures; tree editing.

Schnelle Faltung und Vergleich von Sekundärstrukturen von RNA

Zusammenfassung. Die im Vienna RNA package enthaltenen Computer Programme für die Berechnung und den Vergleich von RNA Sekundärstrukturen werden präsentiert. Ihren Kern bilden Algorithmen zur Vorhersage von Strukturen minimaler Energie sowie zur Berechnung von Zustandssumme und Basenpaarungswahrscheinlichkeiten mittels dynamischer Programmierung.

Ein effizienter heuristischer Algorithmus für das inverse Faltungsproblem wird vorgestellt. Darüberhinaus präsentieren wir kompakte und effiziente Programme zum Vergleich von RNA Sekundärstrukturen durch Baum-Editierung und Alignierung.

Alle Programme sind in ANSI C geschrieben, darunter auch eine Implementation des Faltungsalgorithmus für Parallelrechner mit verteiltem Speicher. Wie Tests auf einem Intel Hypercube zeigen, wird das Parallelrechnen umso effizienter je länger die Sequenzen sind.

1. Introduction

Recent interest in RNA structures and functions was caused by their catalytic capacities [1, 2] as well as by the success of selection methods in producing RNA

1. One sequence – one structure problem
2. **Inverse folding and neutral networks**
3. Kinetic folding
4. Intersections and conformational switches
5. Cofolding of nucleic acid molecules

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

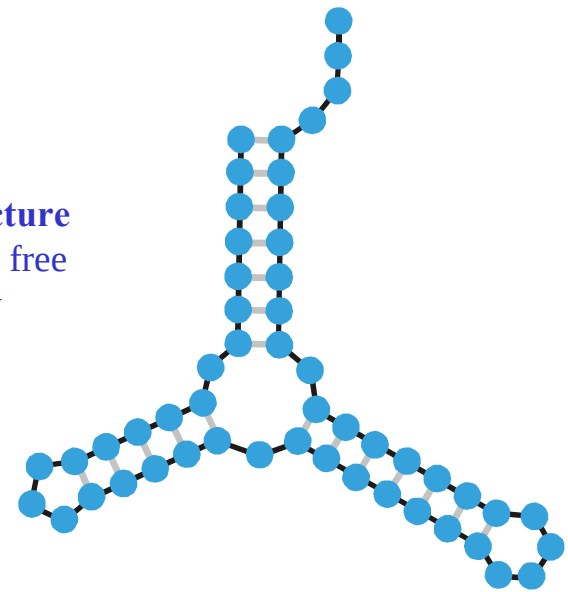
RNA folding:

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

Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

RNA structure of minimal free energy



Sequence, structure, and design

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

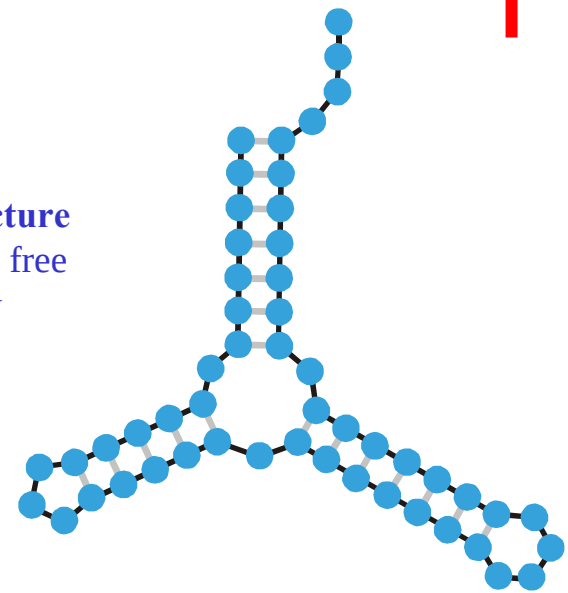
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

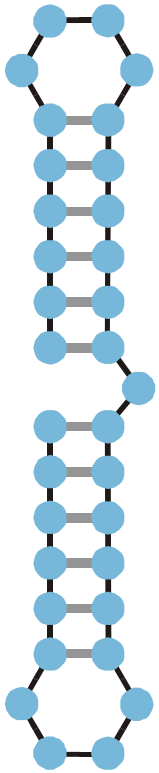
Inverse folding of RNA:

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

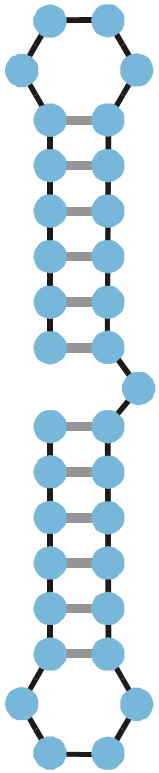
RNA structure of minimal free energy



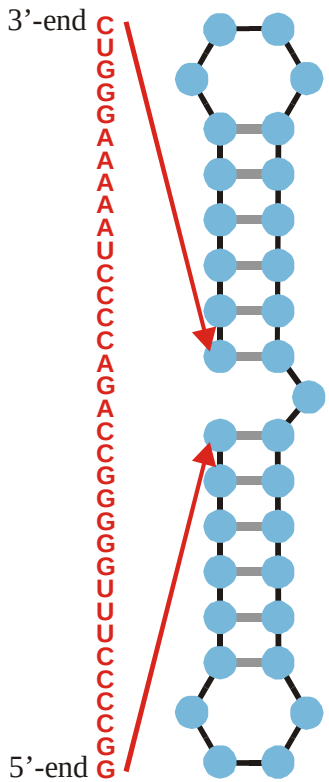
Sequence, structure, and design



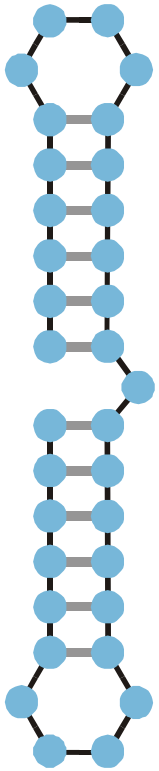
Structure



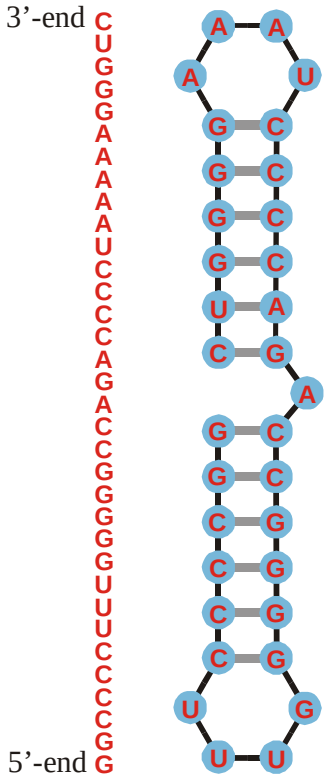
Structure



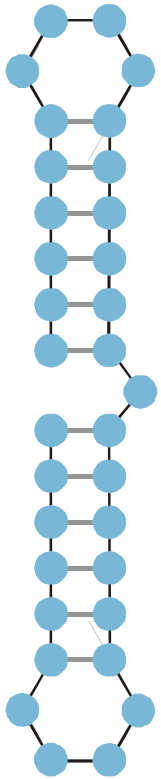
Compatible sequence



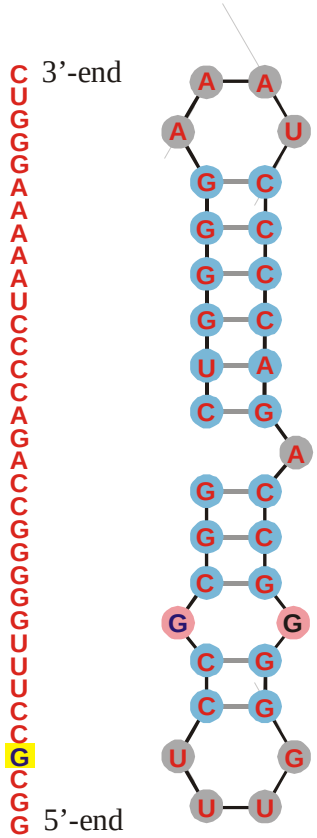
Structure



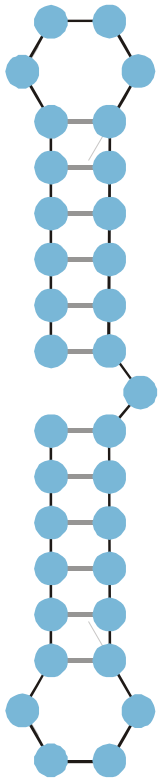
Compatible sequence



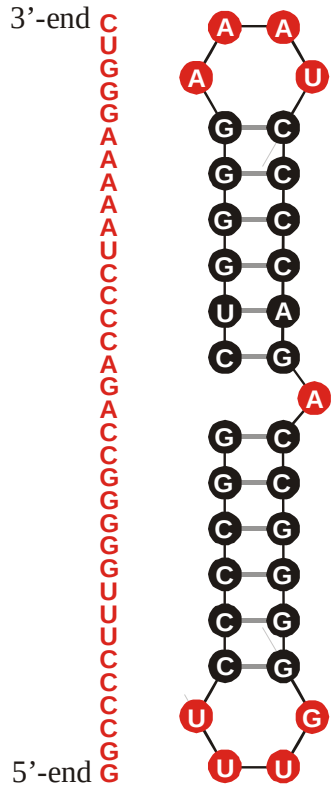
Structure



Incompatible sequence



Structure



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

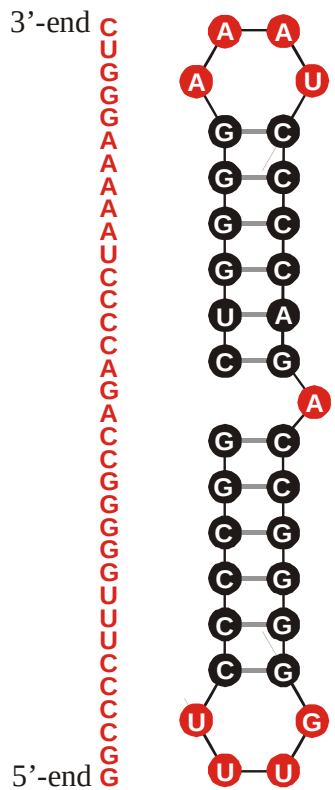
Base pairs: **AU , UA**
GC , CG
GU , UG

Compatible sequences

$$2\,n_{bp} + n_{sn} = n$$

$$N_{cmp} = 4^{n_{sn}} \times 6^{n_{bp}}$$

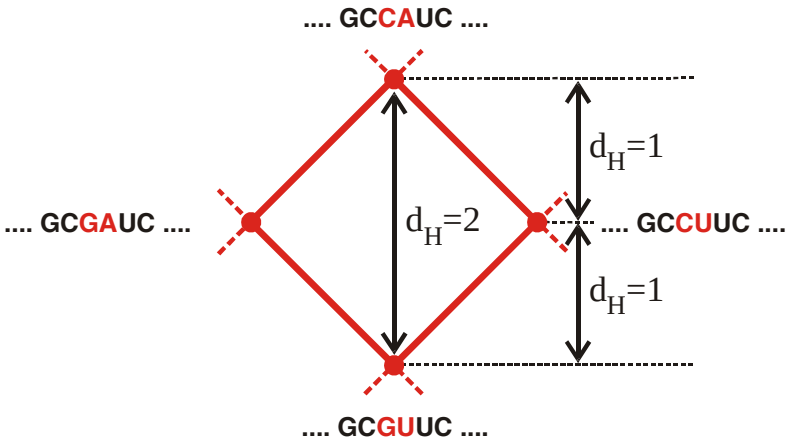
$$\left(\sqrt{6}\right)^{n} \le N_{cmp} \le 4^{n}$$



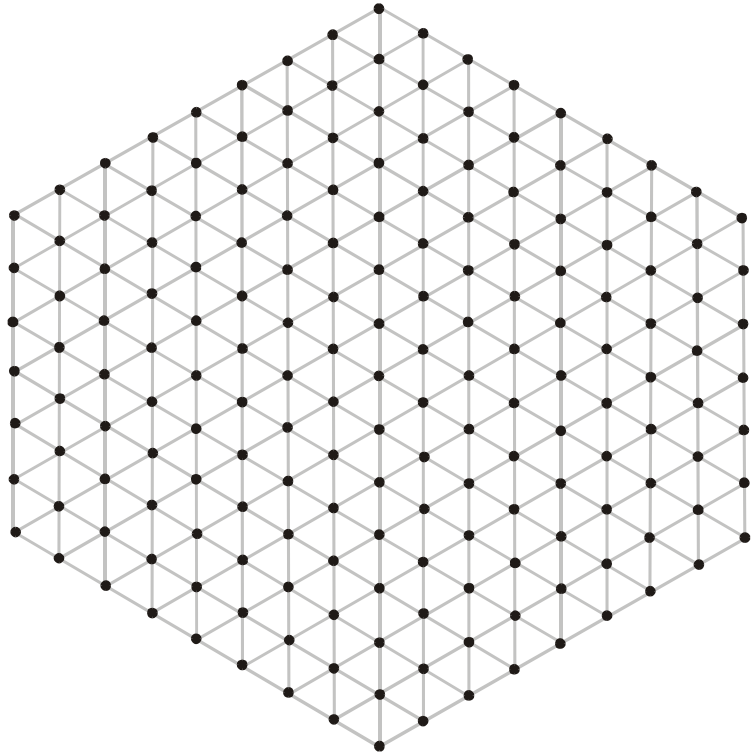
Single nucleotides: **A,U,G,C**

Base pairs: **AU , UA**
GC , CG
GU , UG

Number of compatible sequences

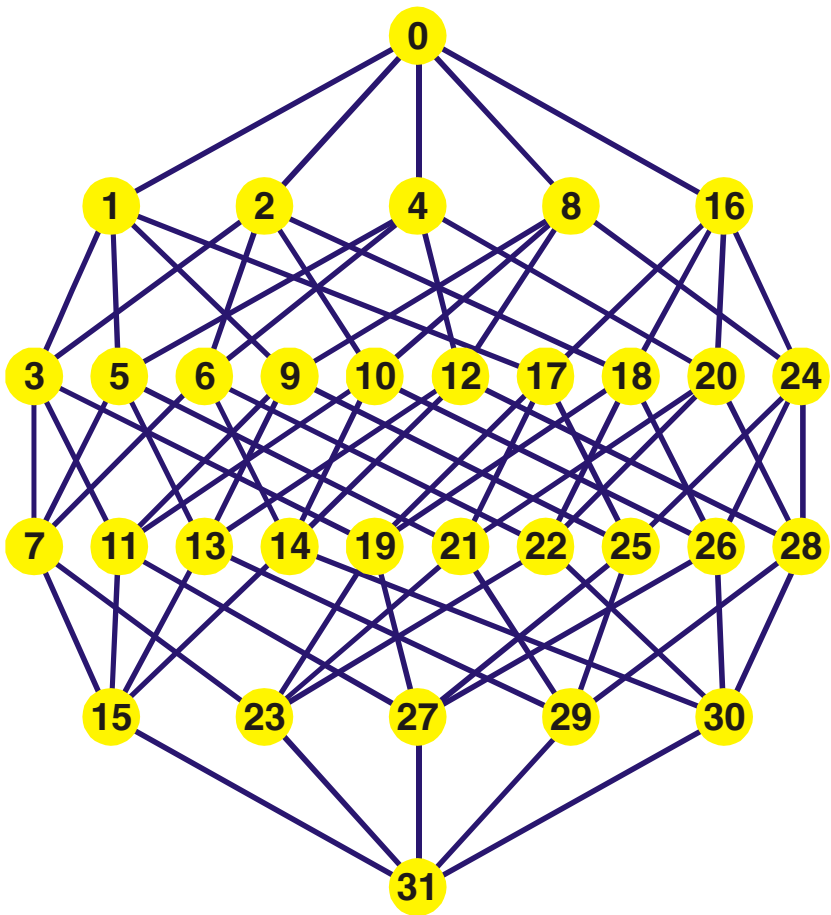


City-block distance in sequence space



2D Sketch of sequence space

Single point mutations as moves in sequence space



Hypercube of dimension n = 5

Mutant class

0

1

2

3

4

5

Binary sequences are encoded by their decimal equivalents:

C = 0 and G = 1, for example,

"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

"29" ≡ 11101 = GGGCG, etc.

Decimal coding of binary sequences

Sequence space of binary sequences of chain lenght n = 5

```
I1:   CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....
I2:   CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAACACAAGAG.....
```

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

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

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

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

The Hamming distance between sequences induces a metric in sequence space

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

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

Inverse folding algorithm

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

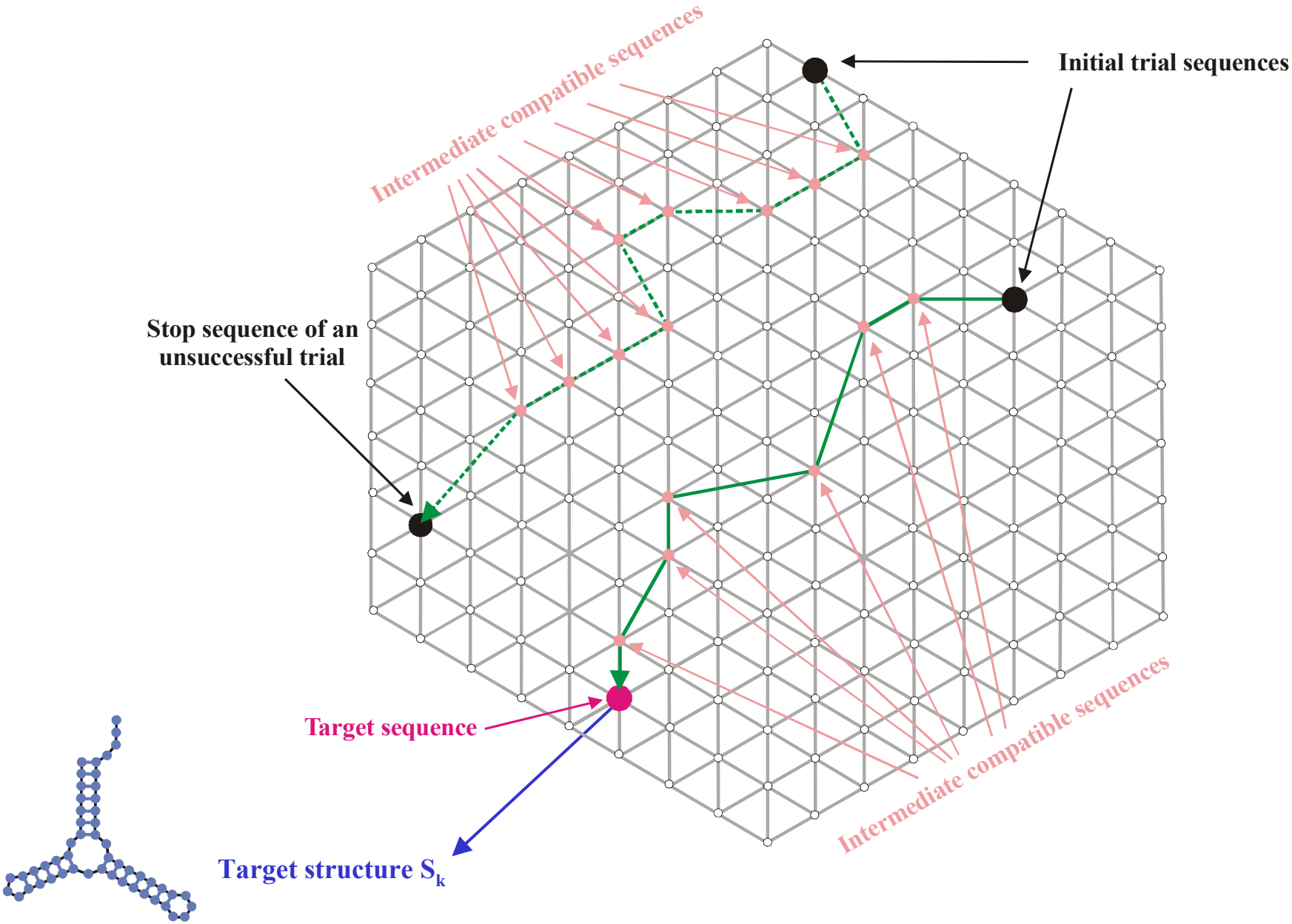
$\mathbf{S_0} \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow \ldots \rightarrow S_k \rightarrow S_{k+1} \rightarrow \ldots \rightarrow \mathbf{S_t}$

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

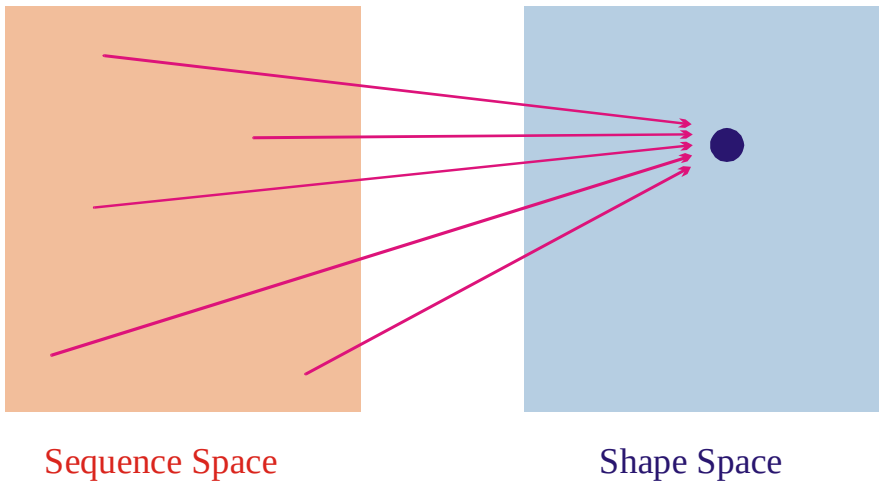
$\mathfrak{M}$ ... base or base pair mutation operator

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

‚Unsuccessful trial‘ ... termination after n steps



Approach to the **target structure S_k** in the inverse folding algorithm



Number of sequences: $N_I = 4^n$; Number of secondary structures: $N_S = 1.4848 \times n^{-3/2} \times 1.84892^n$

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.

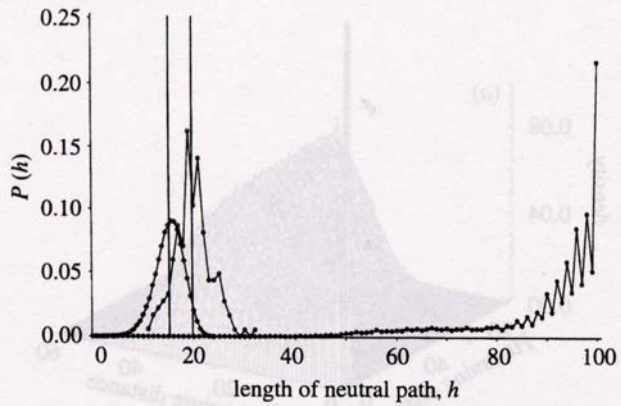
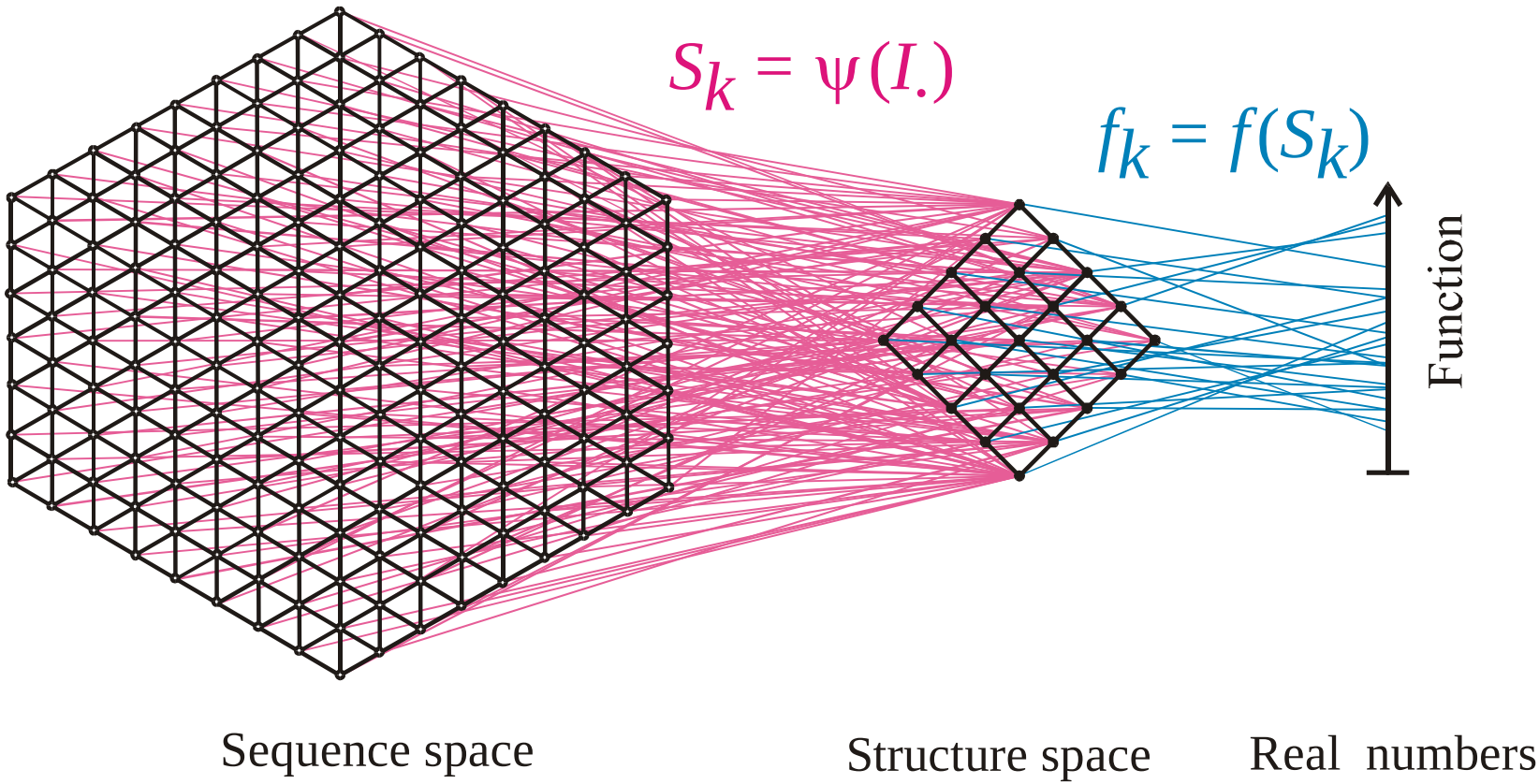


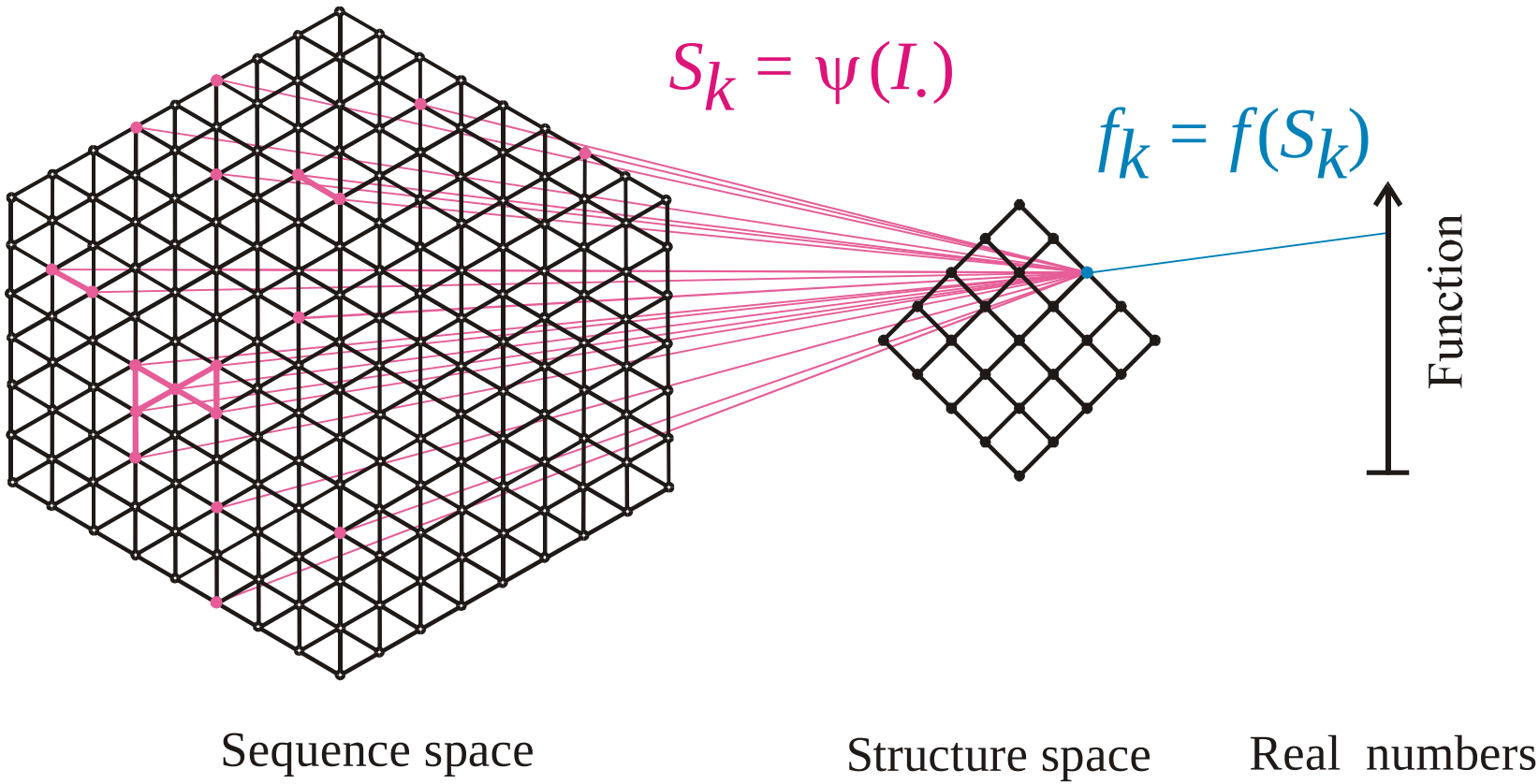
Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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

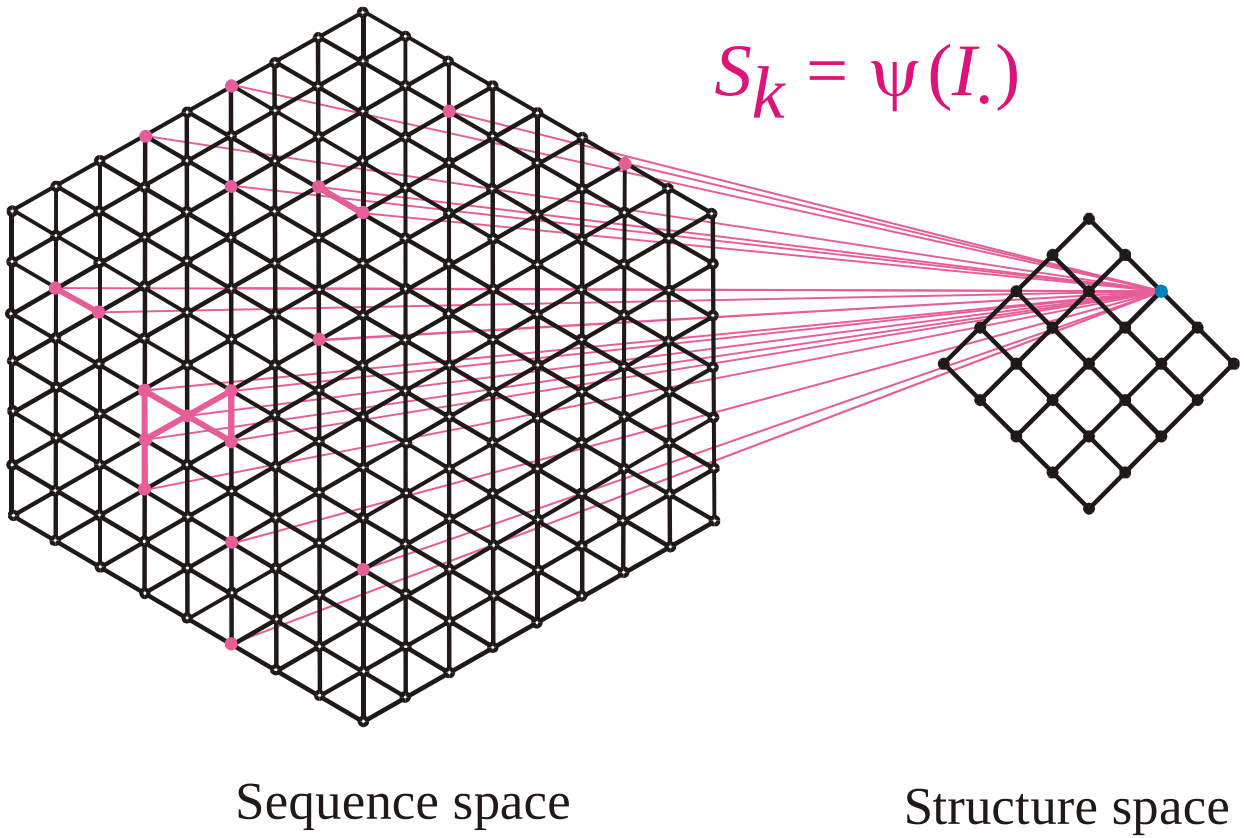
Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

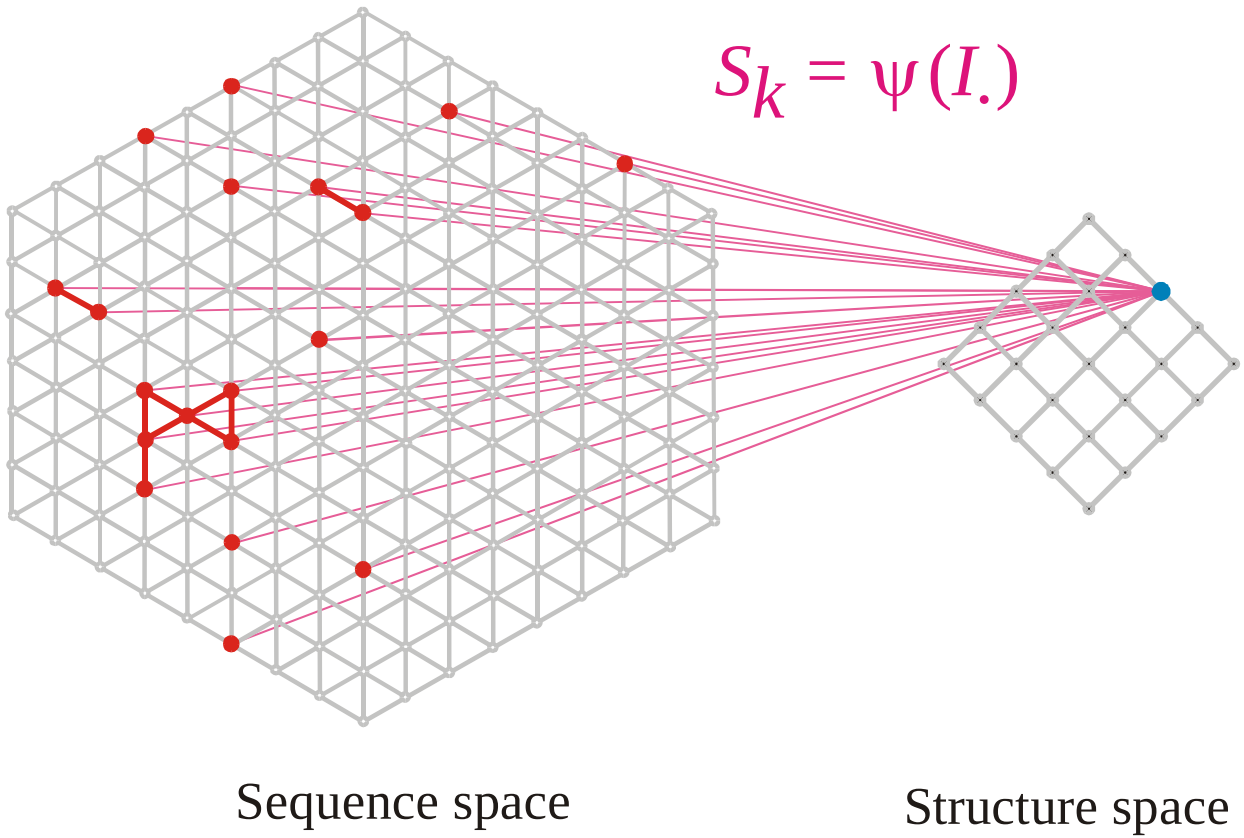
Reference for postulation and ***in silico*** verification of ***neutral networks***



Mapping from sequence space into structure space and into function


$S_k = \psi(I.)$
$f_k = f(S_k)$
Function
Sequence space
Structure space
Real numbers


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



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:

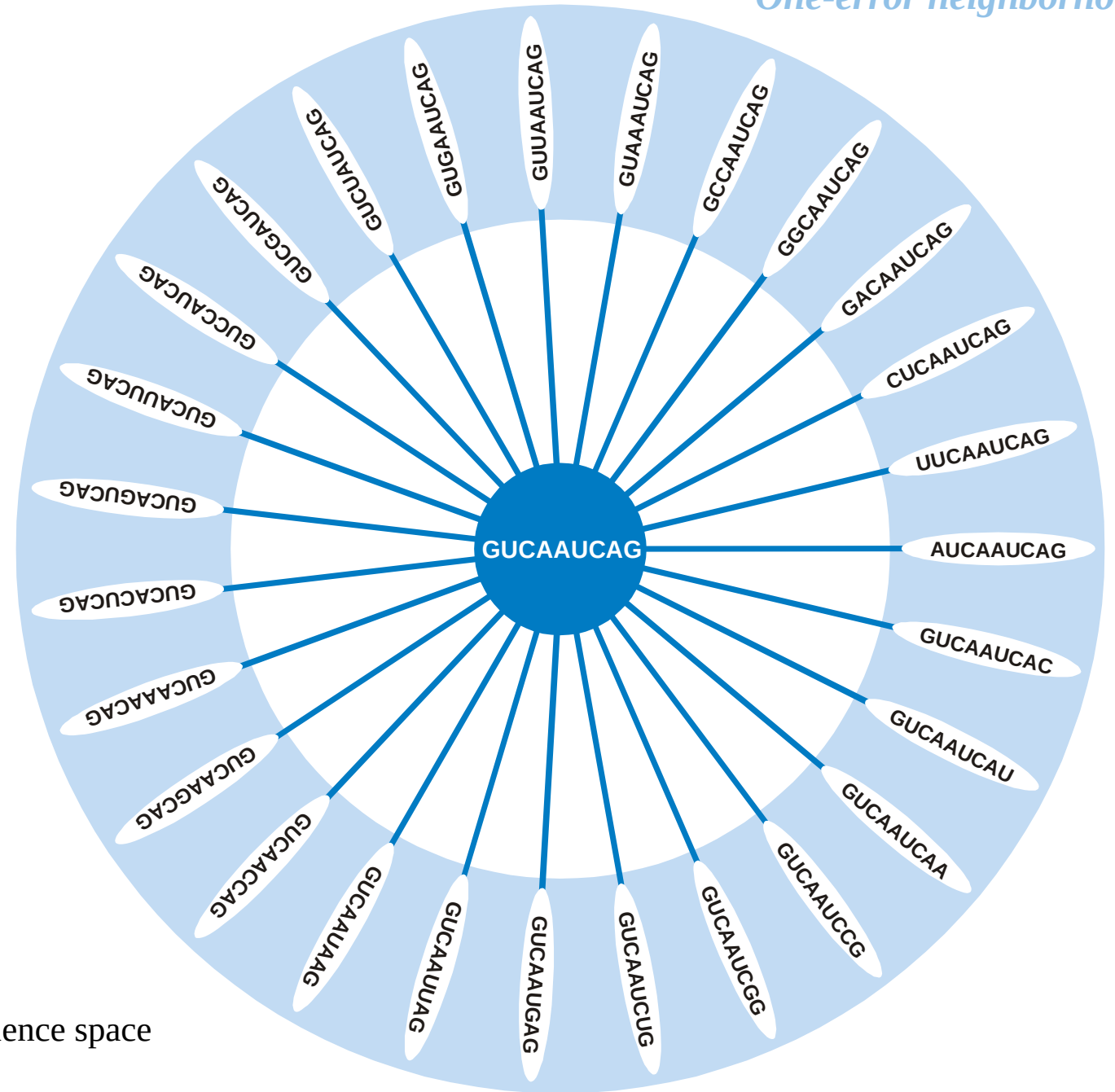
$$\mathbf{G_k} = \psi^{-1}(S_k) \pi \{\psi_j \mid \psi(I_j) = S_k\}$$

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

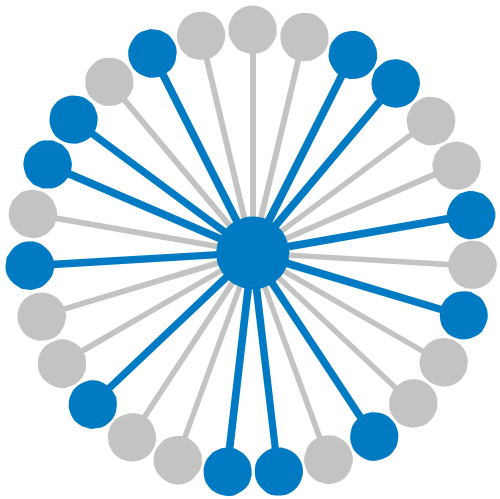
Neutral networks of small biomolecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, $\mathbf{N=4^n}$, becomes very large with increasing length, and is prohibitive for numerical computations.

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.

One-error neighborhood



The surrounding of
GUCAAUCAG in sequence space



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

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

$\lambda_j = 12 / 27 = 0.444$

Alphabet size κ :

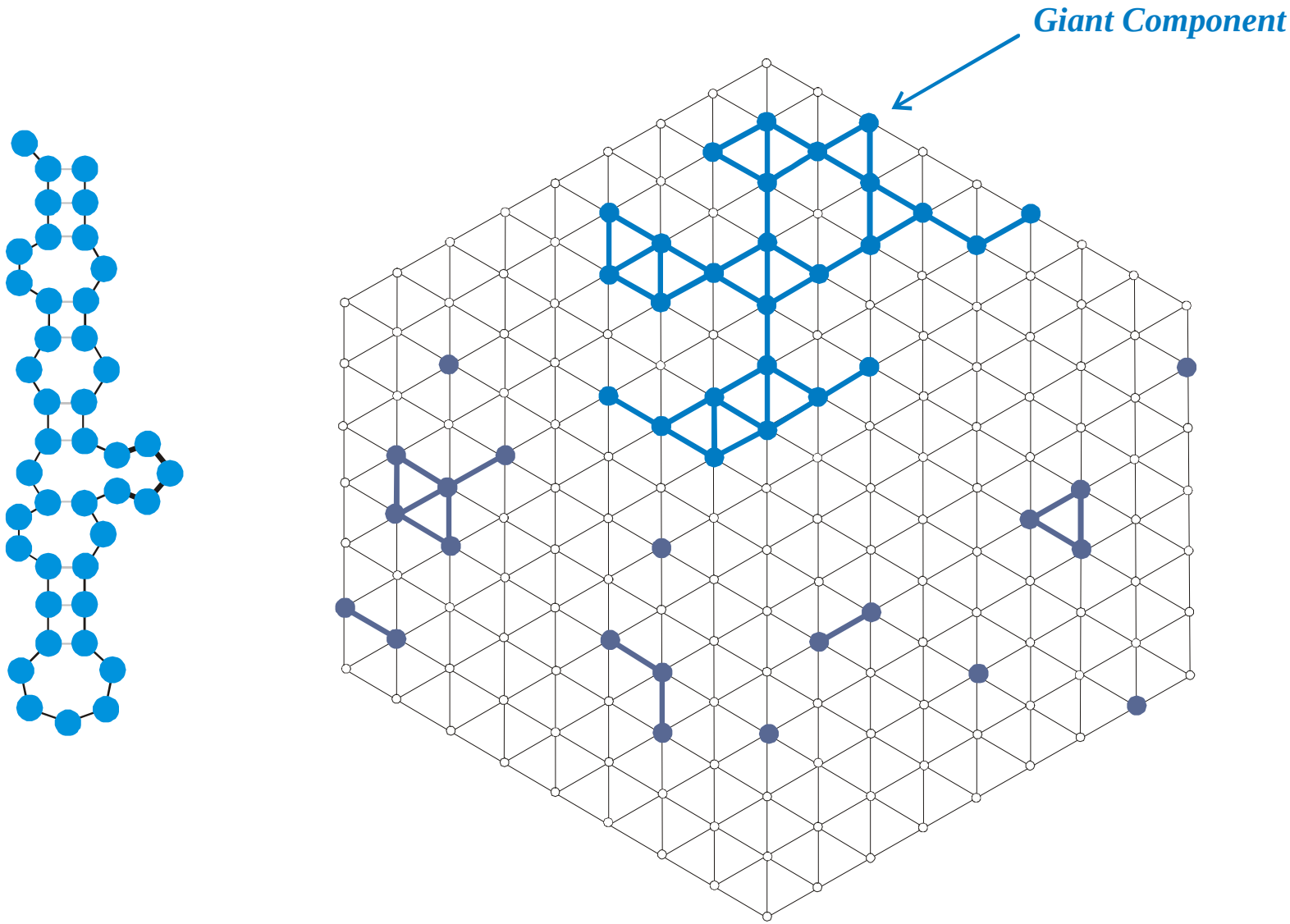
κ	λ_{cr}	
2	0.5	**AU,GC,DU**
3	0.423	**AUG , UGC**
4	0.370	**AUGC**

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

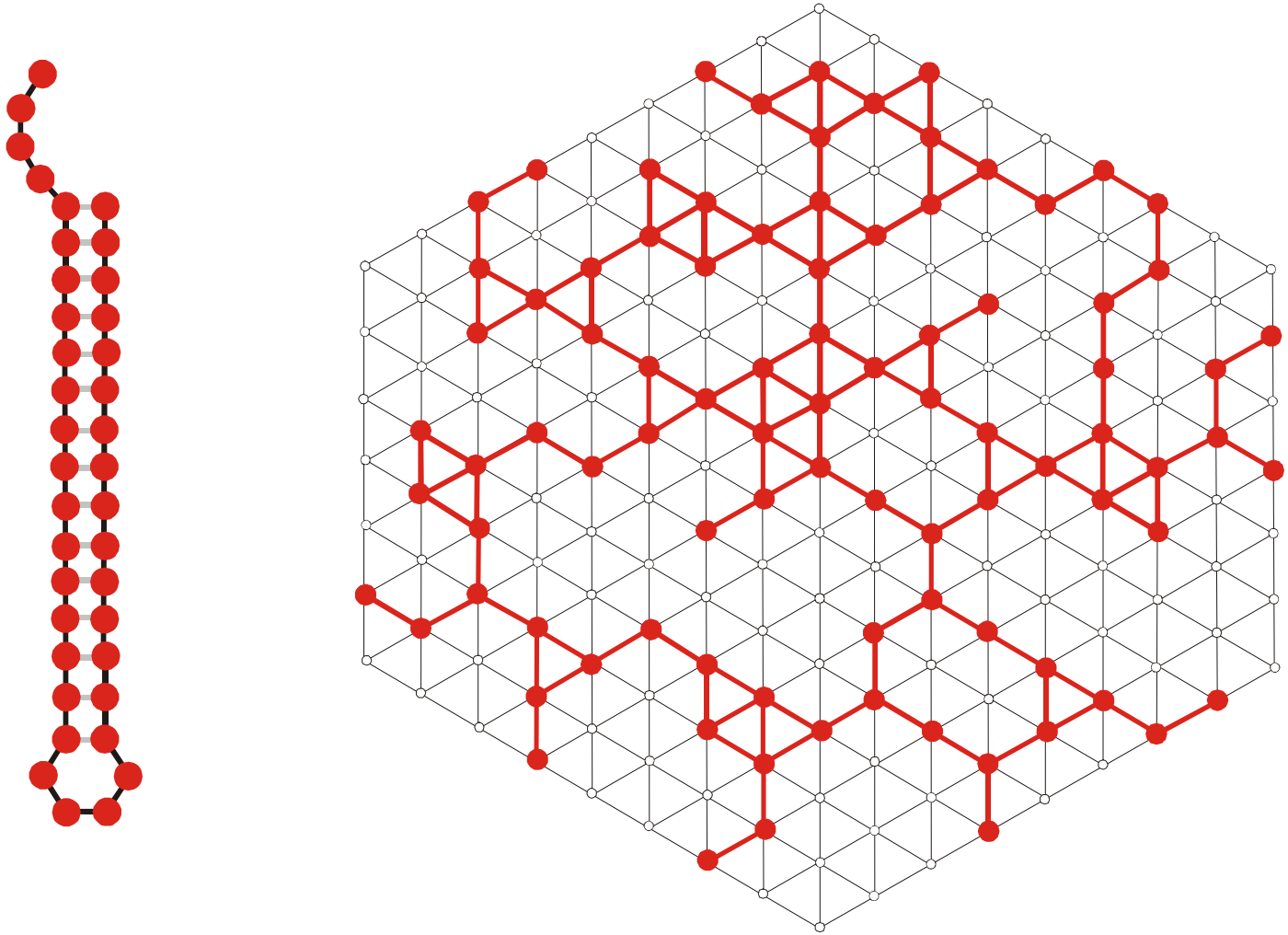
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

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

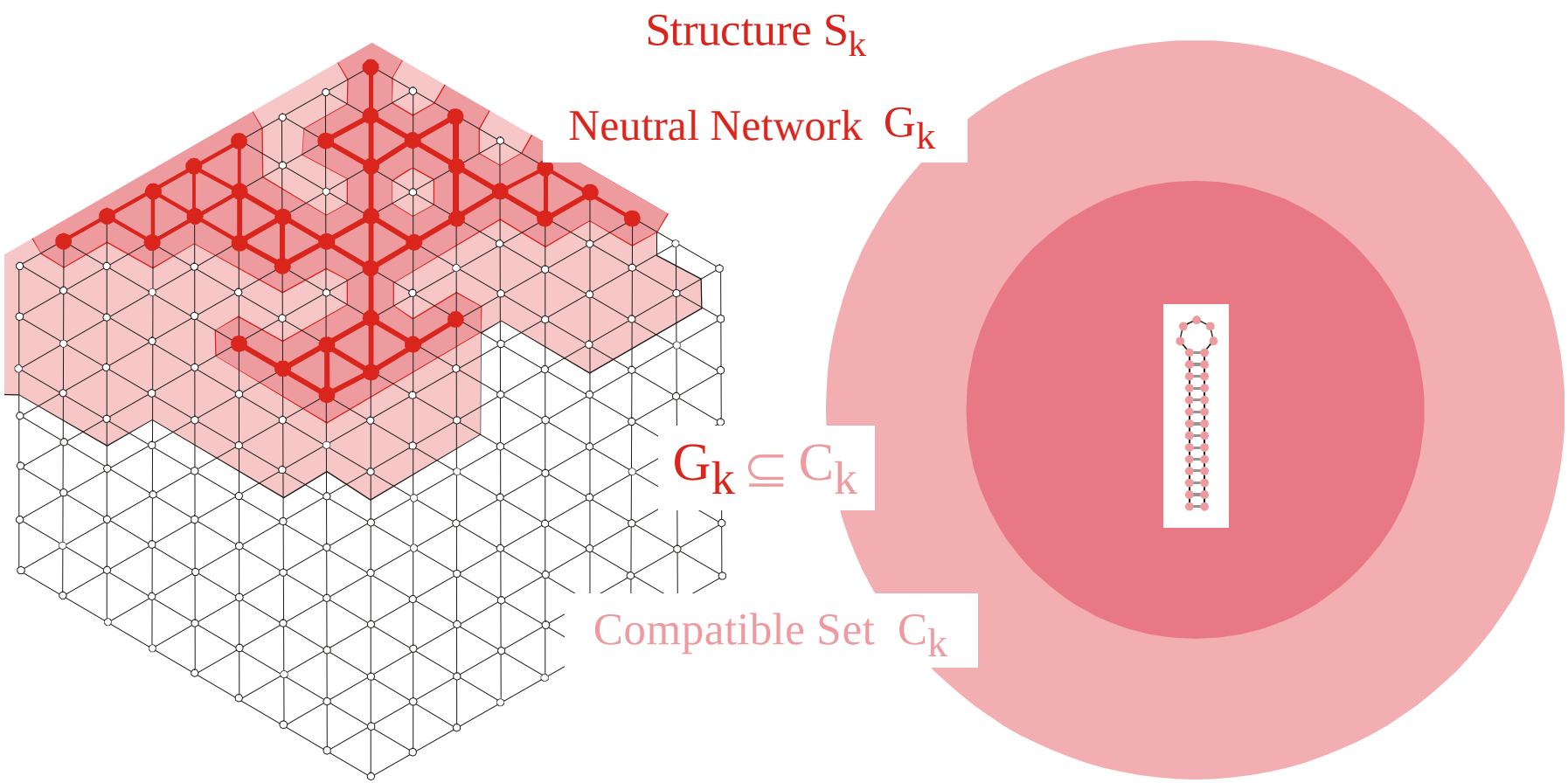
Degree of neutrality of neutral networks and the connectivity threshold



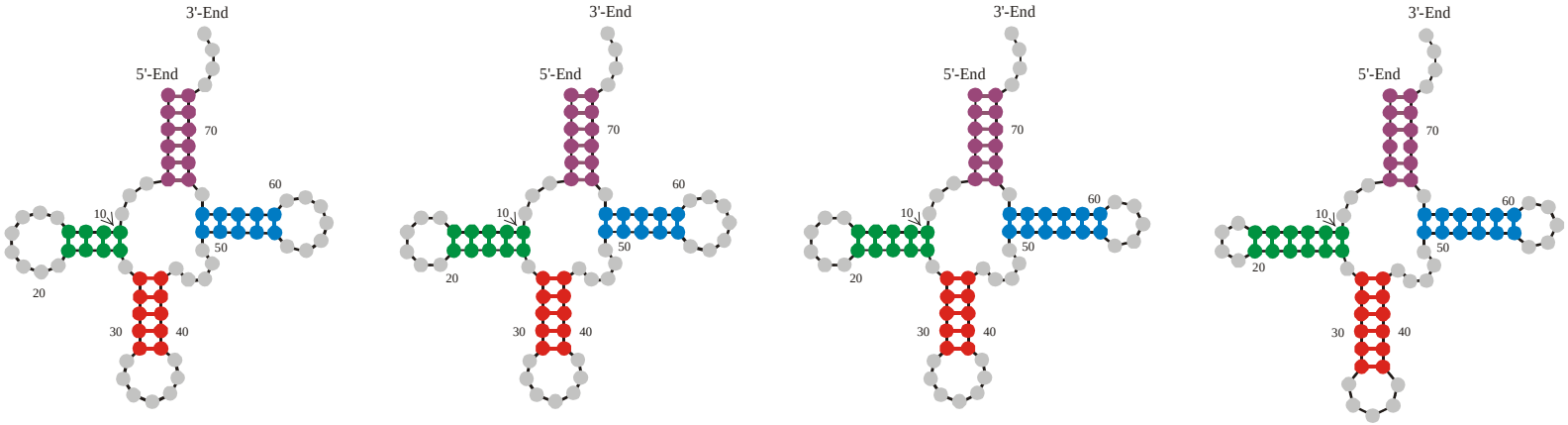
A multi-component neutral network formed by a rare structure



A connected neutral network formed by a common structure



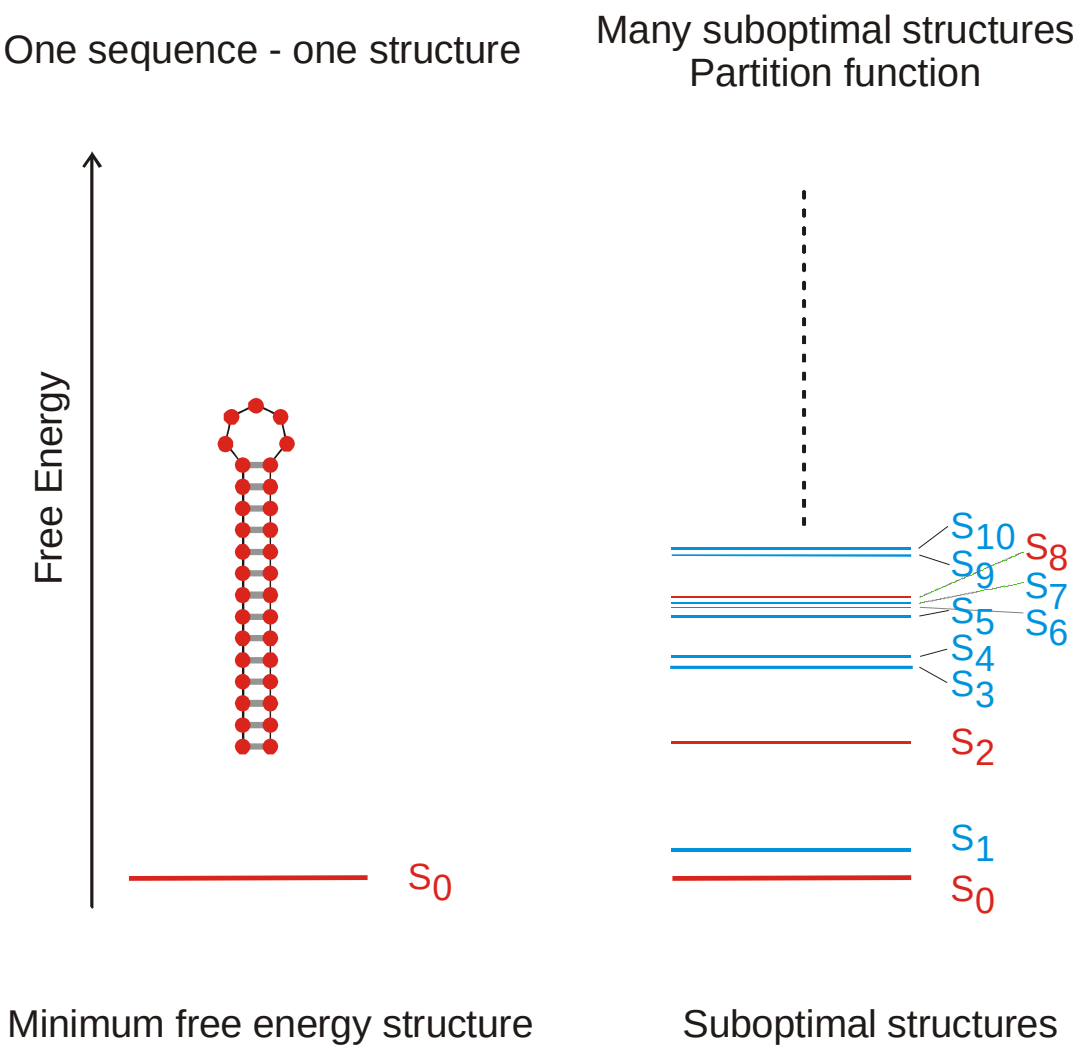
The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



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

Degree of neutrality of cloverleaf RNA secondary structures over different alphabets

1. One sequence – one structure problem
2. Inverse folding and neutral networks
3. **Kinetic folding**
4. Intersections and conformational switches
5. Cofolding of nucleic acid molecules



RNA secondary structures derived from a single sequence

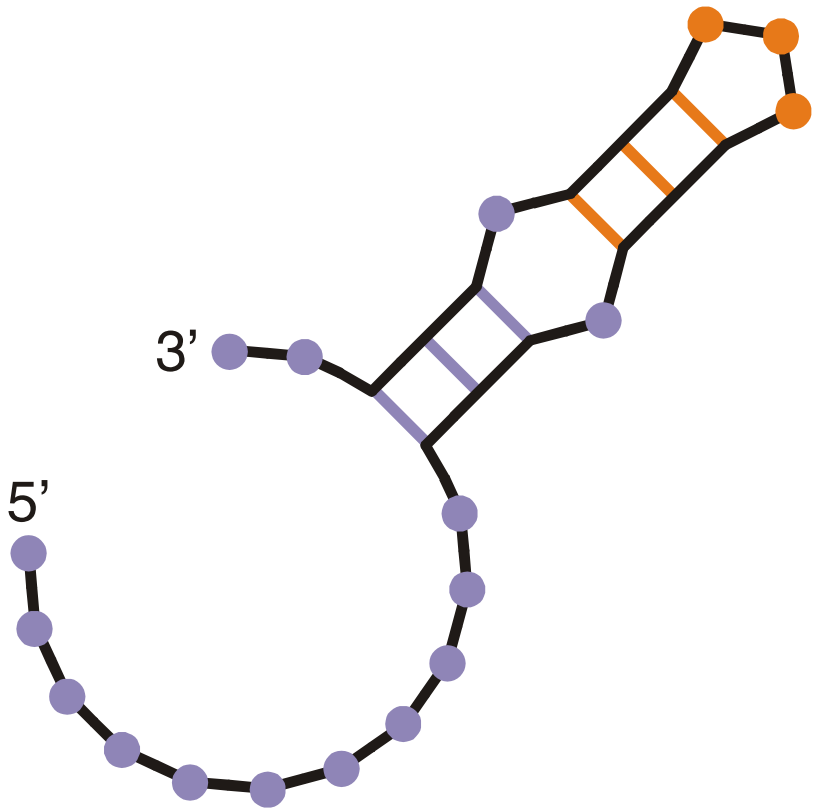
Computation of suboptimal secondary structures

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 $\Delta G_0 > 0$):

#conformations = **1 416 661**

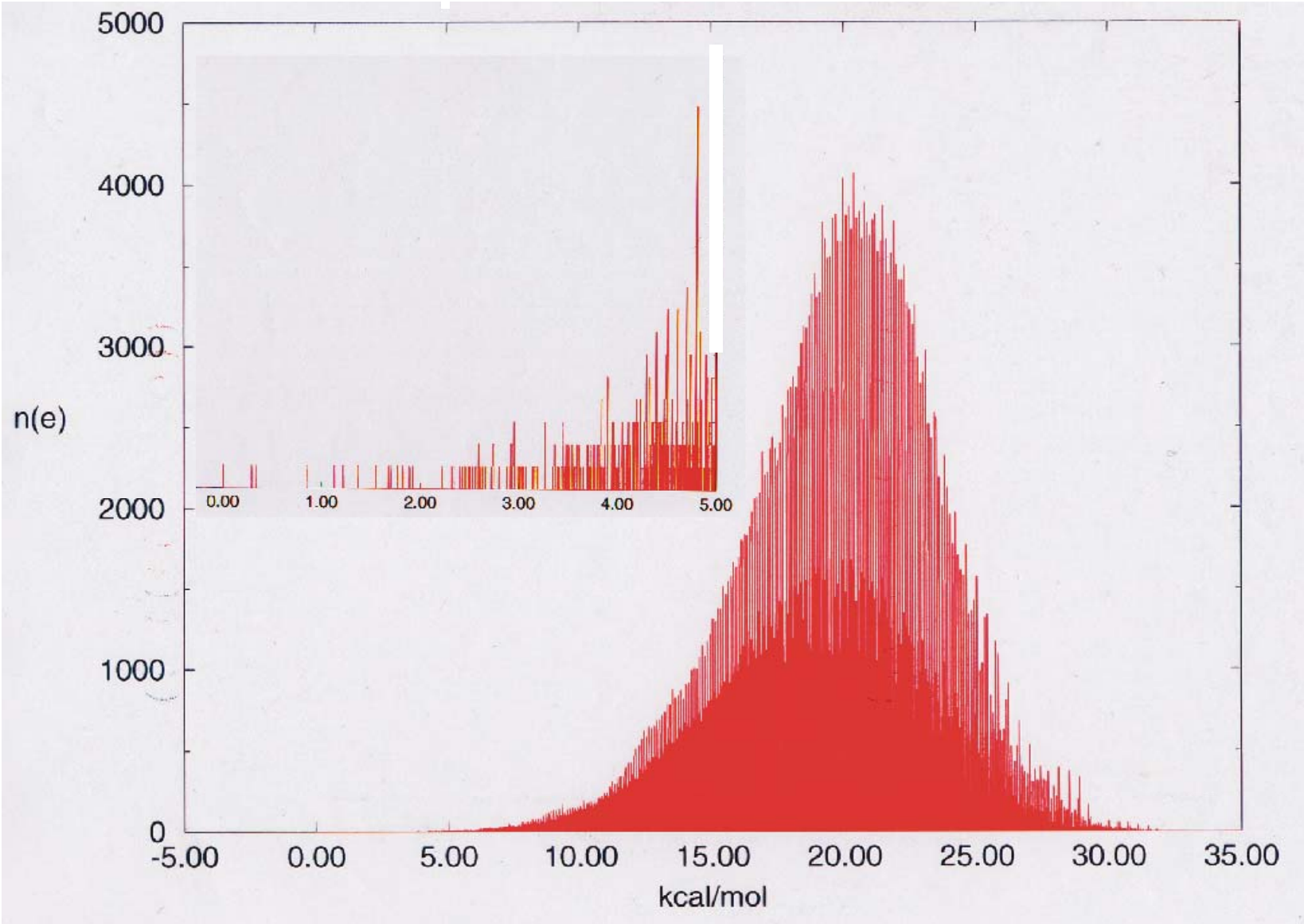


Minimum free energy structure

AAAGGGCACAGGGUGAUUUCAAUAAUUUUA

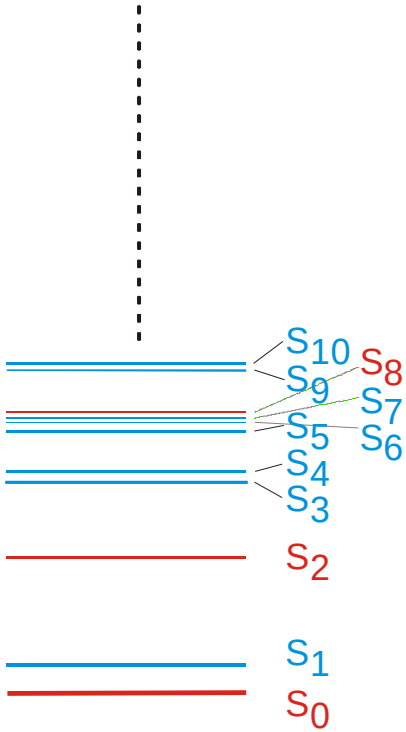
Sequence

Example of a small RNA molecule: n=30



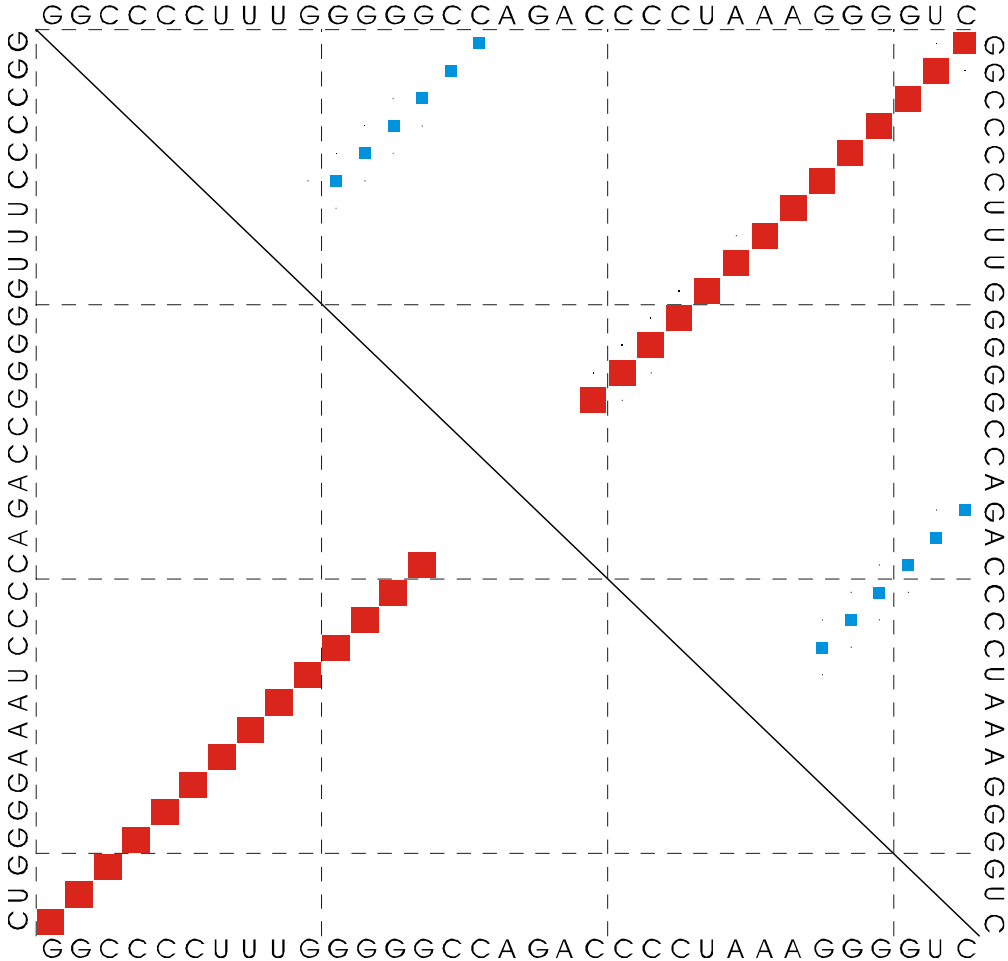
Density of stares of suboptimal structures of the RNA molecule with the sequence:

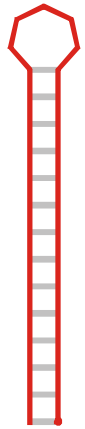
AAAGGGCACAGGGUGAUUUCAAUAAUUUUA



Suboptimal structures

RNA secondary structures derived from a single sequence



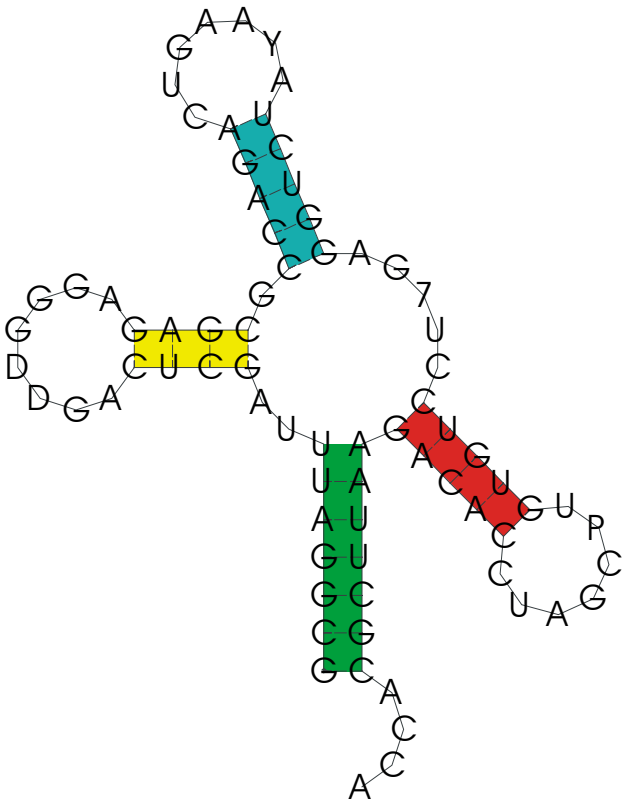


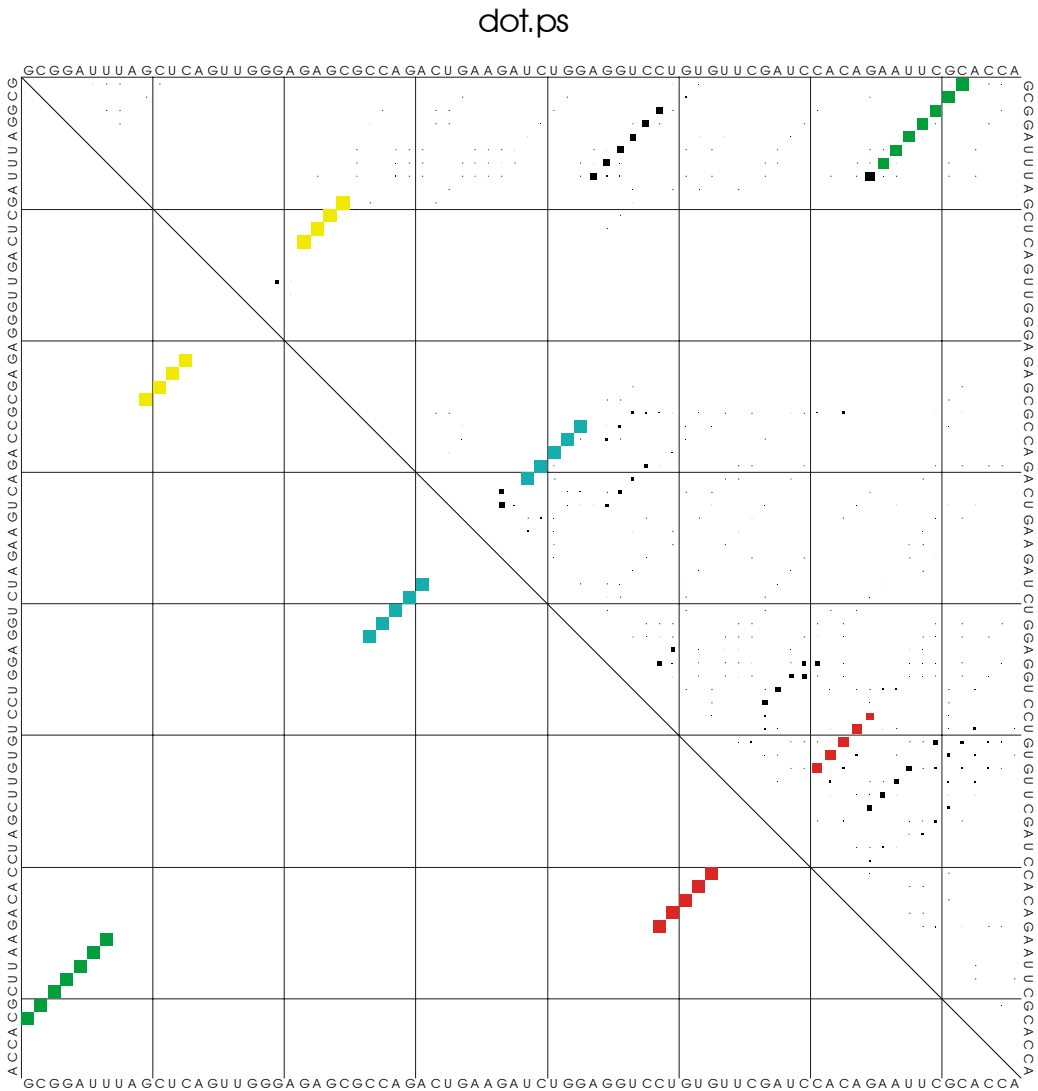
-23.8 kcal

-23.0 kcal

GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC

The ‘dot plot’ of a two-conformation molecule



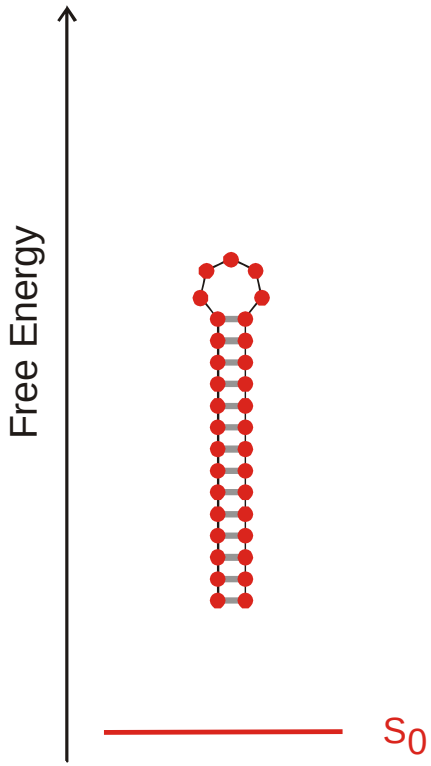


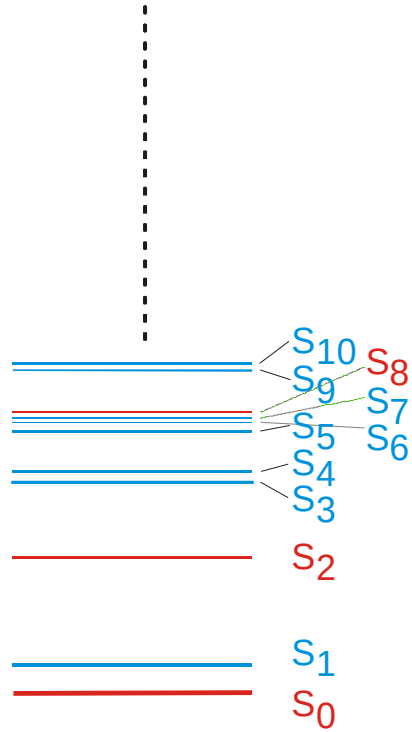
The tRNA^{phe} ‘dot plot’ and the base pairing probabilities from the partition function

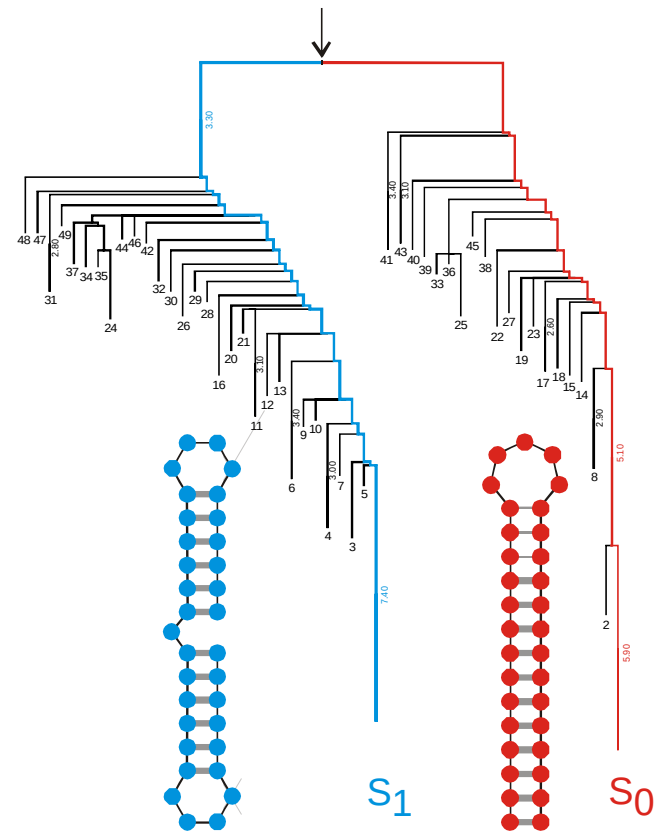
One sequence - one structure

Many suboptimal structures
Partition function

Metastable structures
Conformational switches







Minimum free energy structure

Suboptimal structures

Kinetic structures

RNA secondary structures derived from a single sequence

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

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. *Efficient computation of RNA folding dynamics*. J.Phys.A: Math.Gen. **37**:4731-4741, 2004

The Folding Algorithm

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

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

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

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

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

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

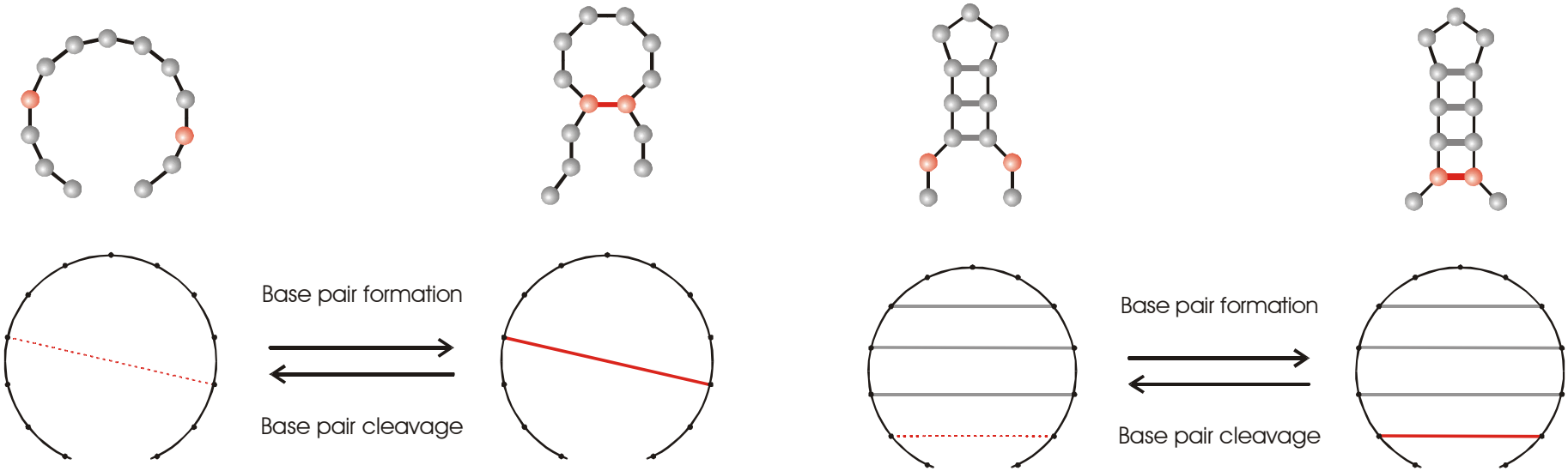
$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t) \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

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

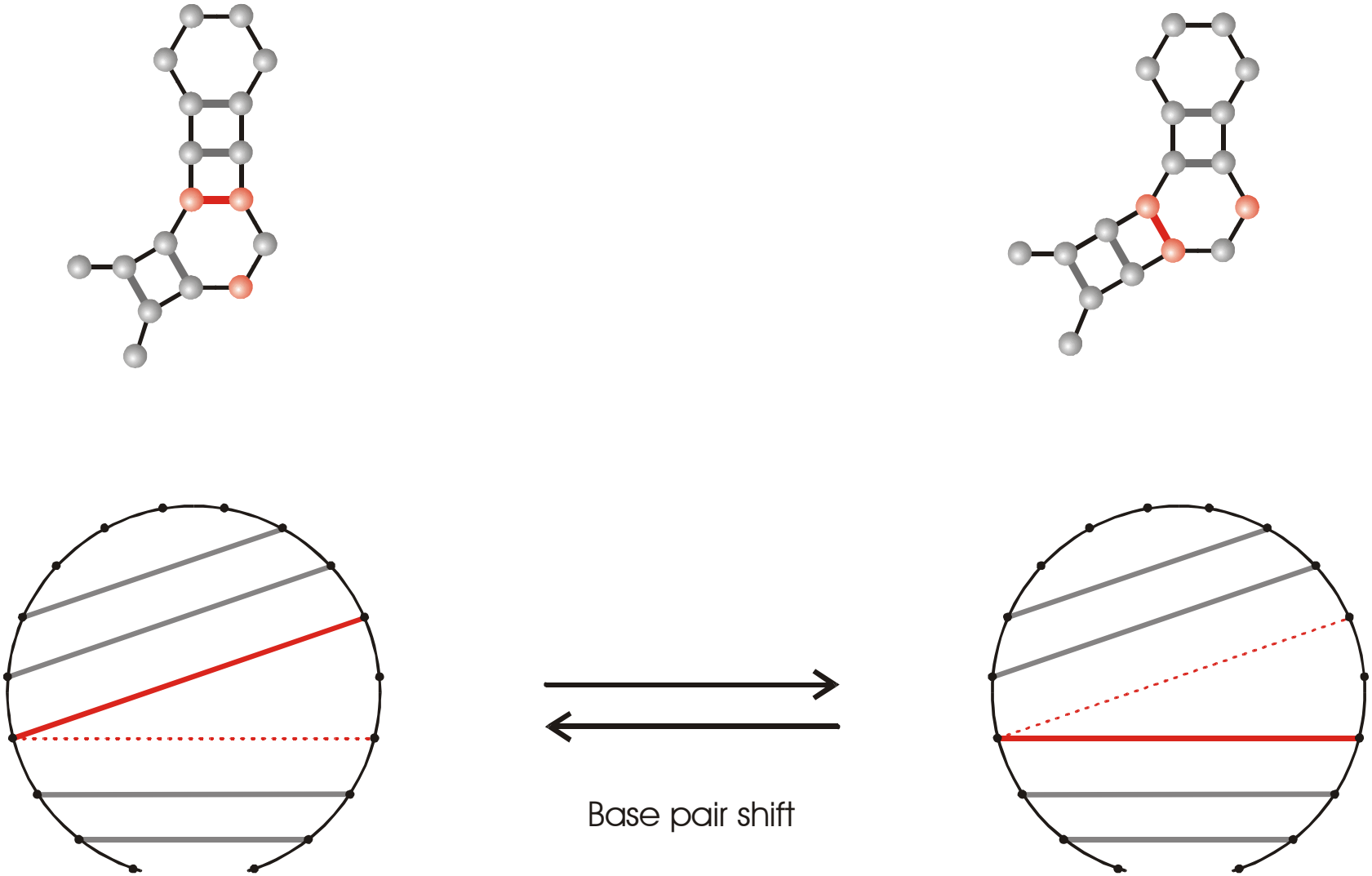
$$\Sigma_k = \sum_{k=1, k \neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

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

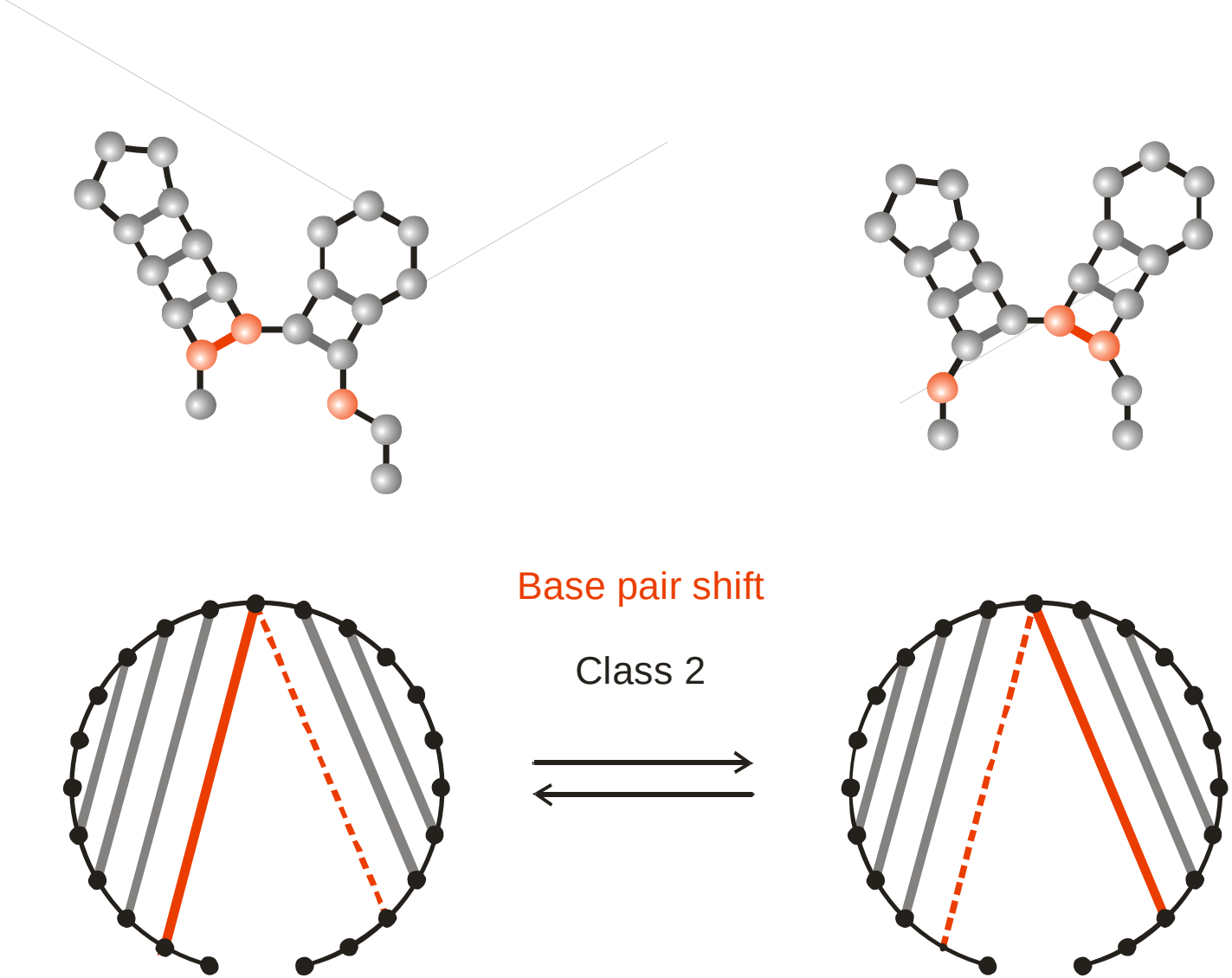
Formulation of kinetic RNA folding as a stochastic process



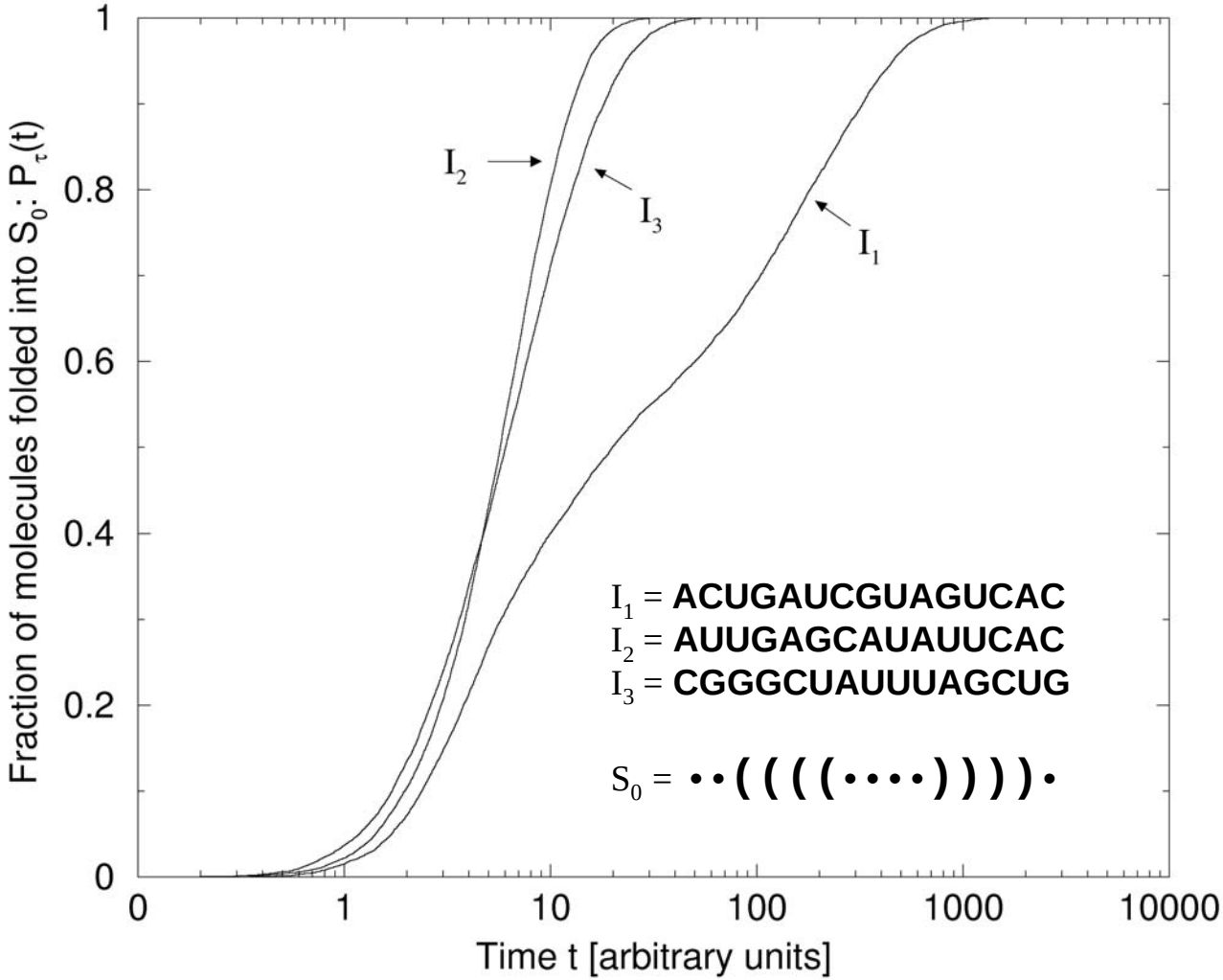
Base pair formation and base pair cleavage moves for nucleation and elongation of stacks



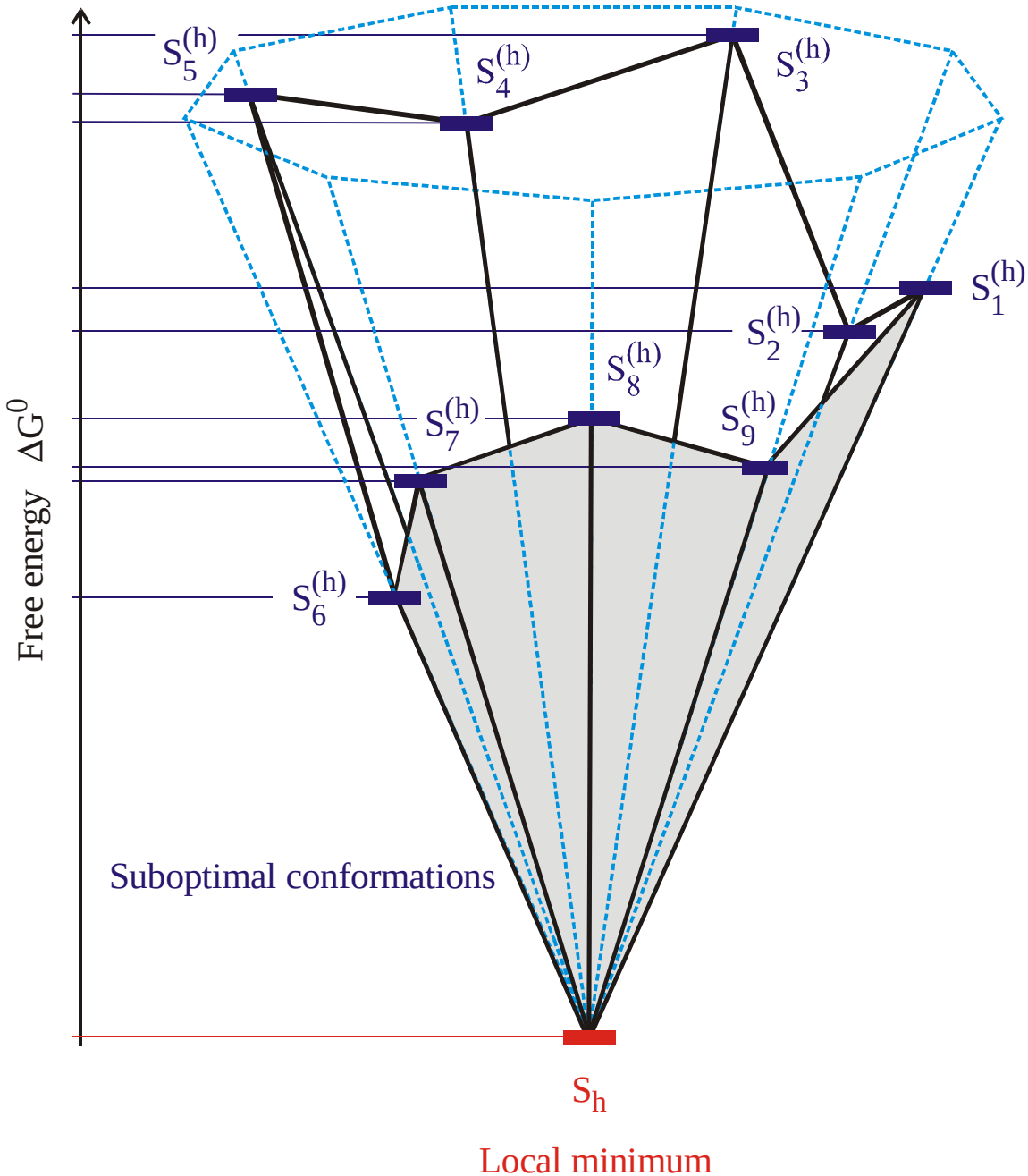
Base pair shift move of class 1: Shift inside internal loops or bulges



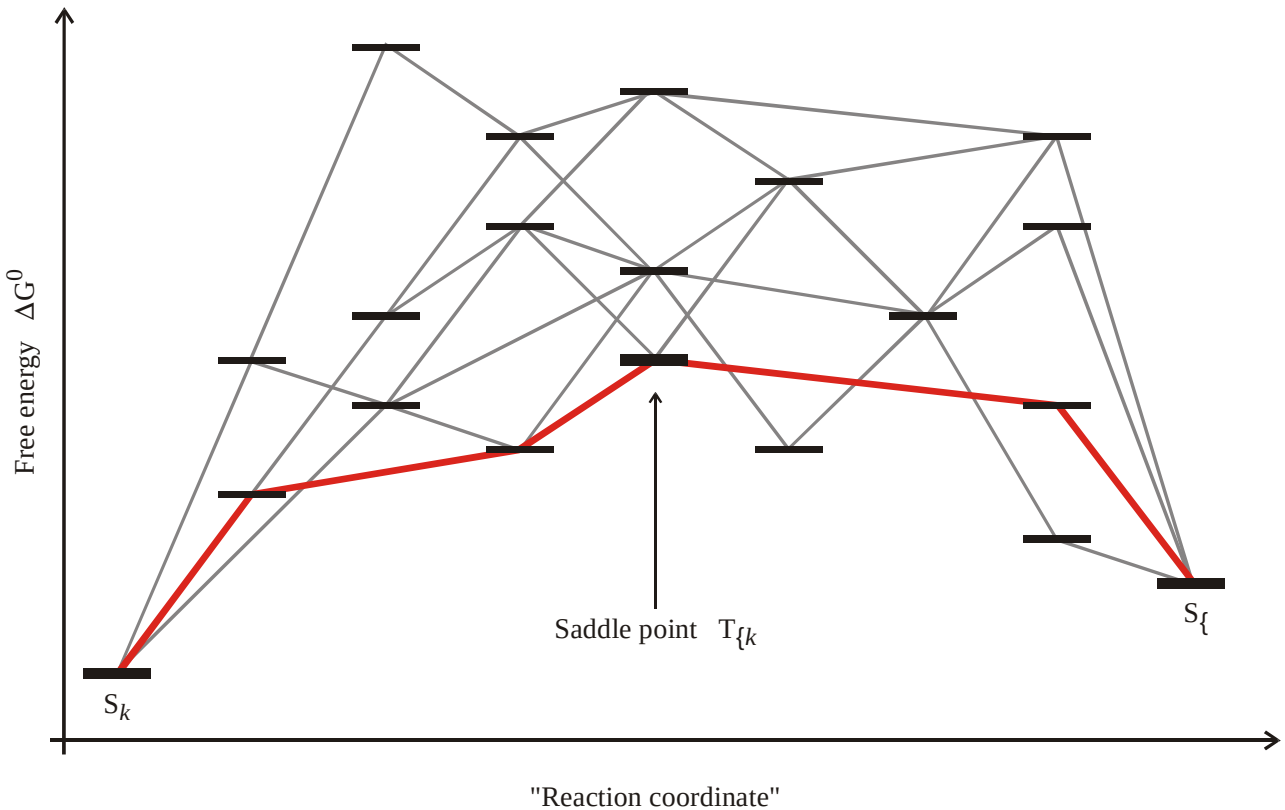
Base pair shift move of class 2: Shift involving free ends

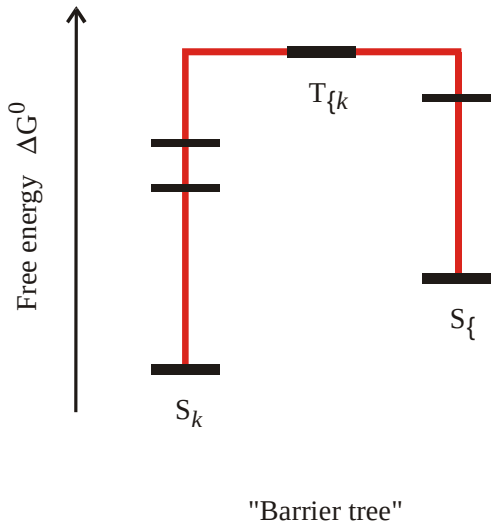


Mean folding curves for three small RNA molecules with different folding behavior

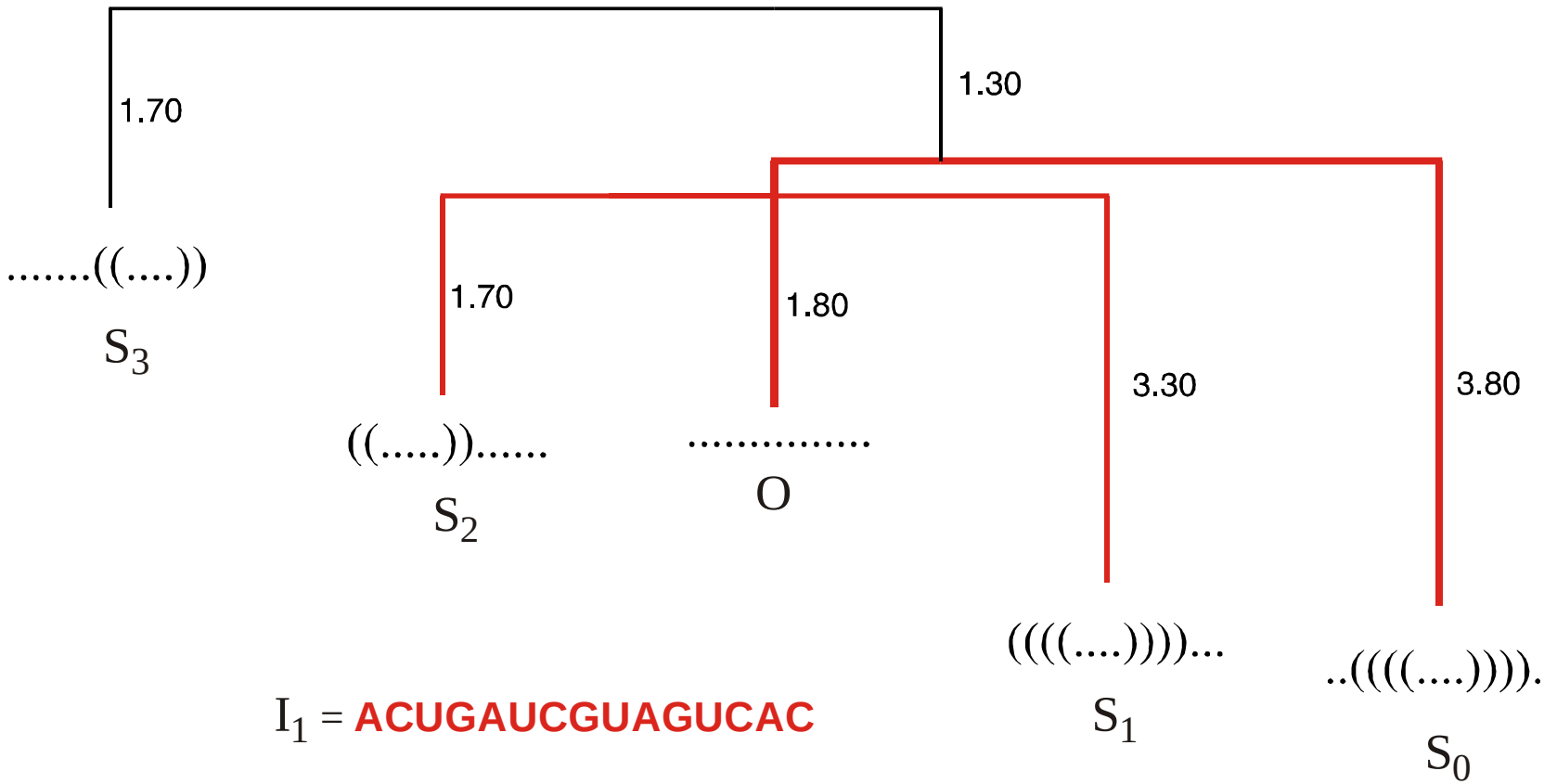


Search for local minima in conformation space

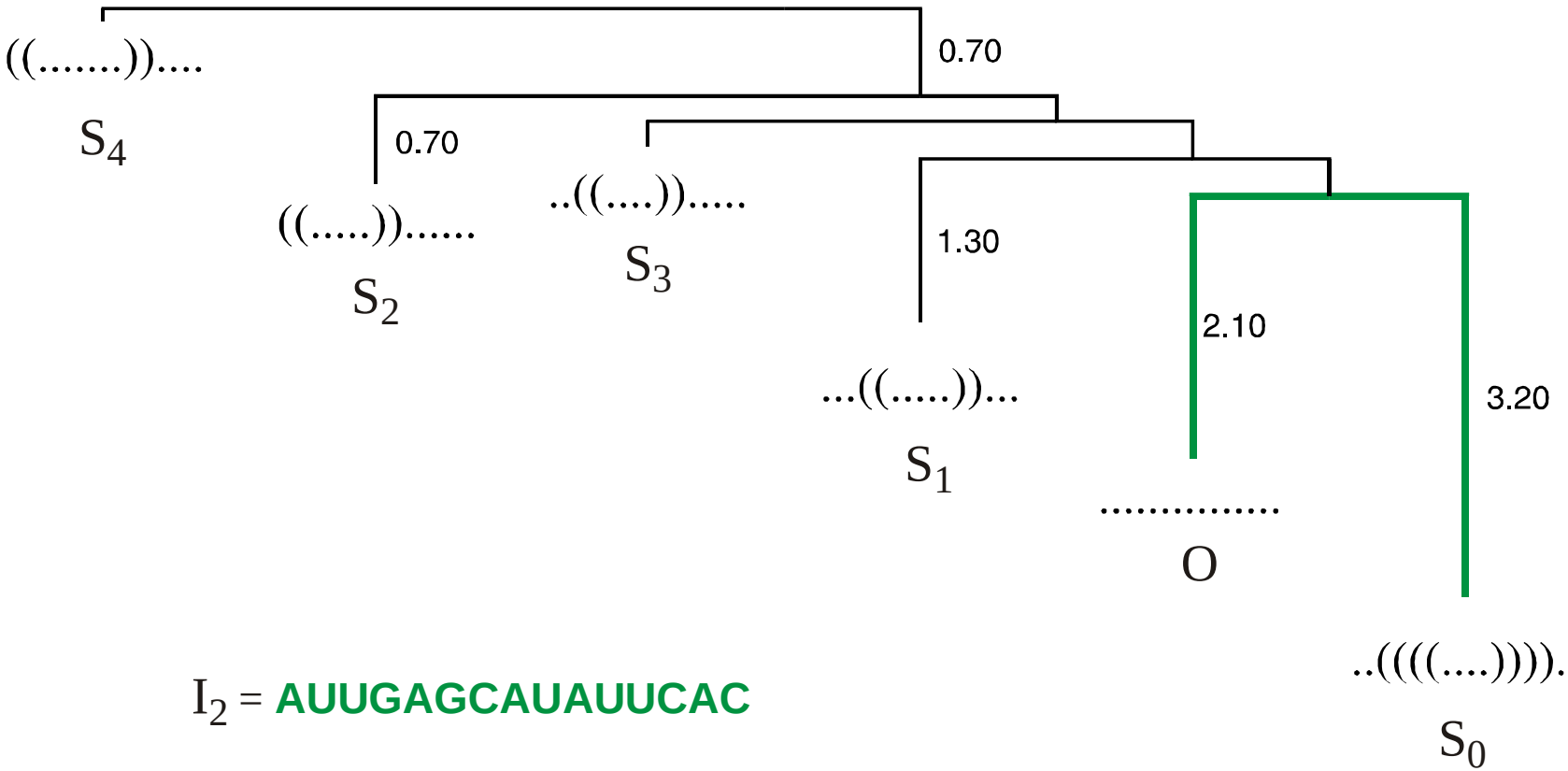




Definition of a ‚barrier tree‘

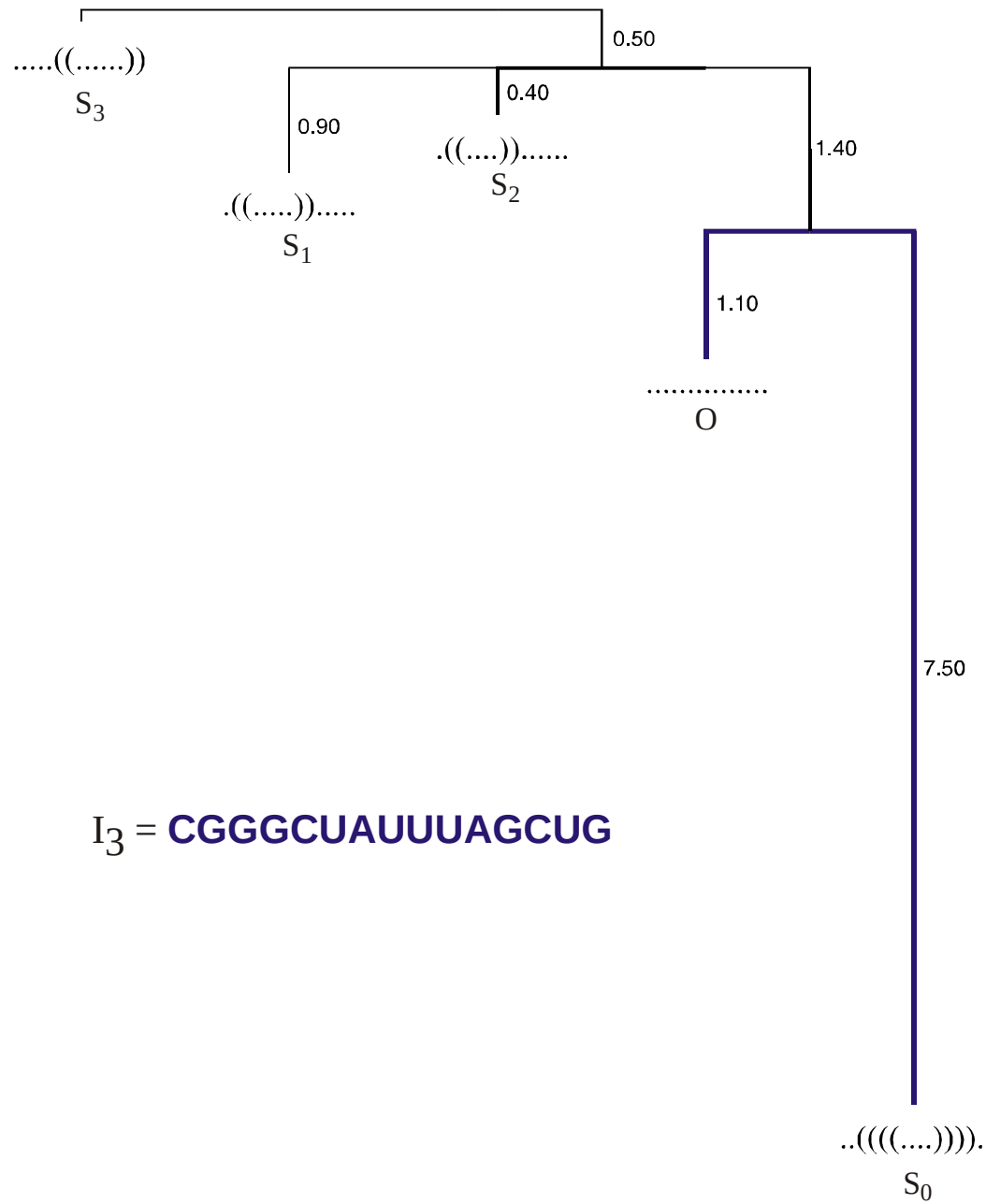


Example of an unefficiently folding small RNA molecule with n = 15



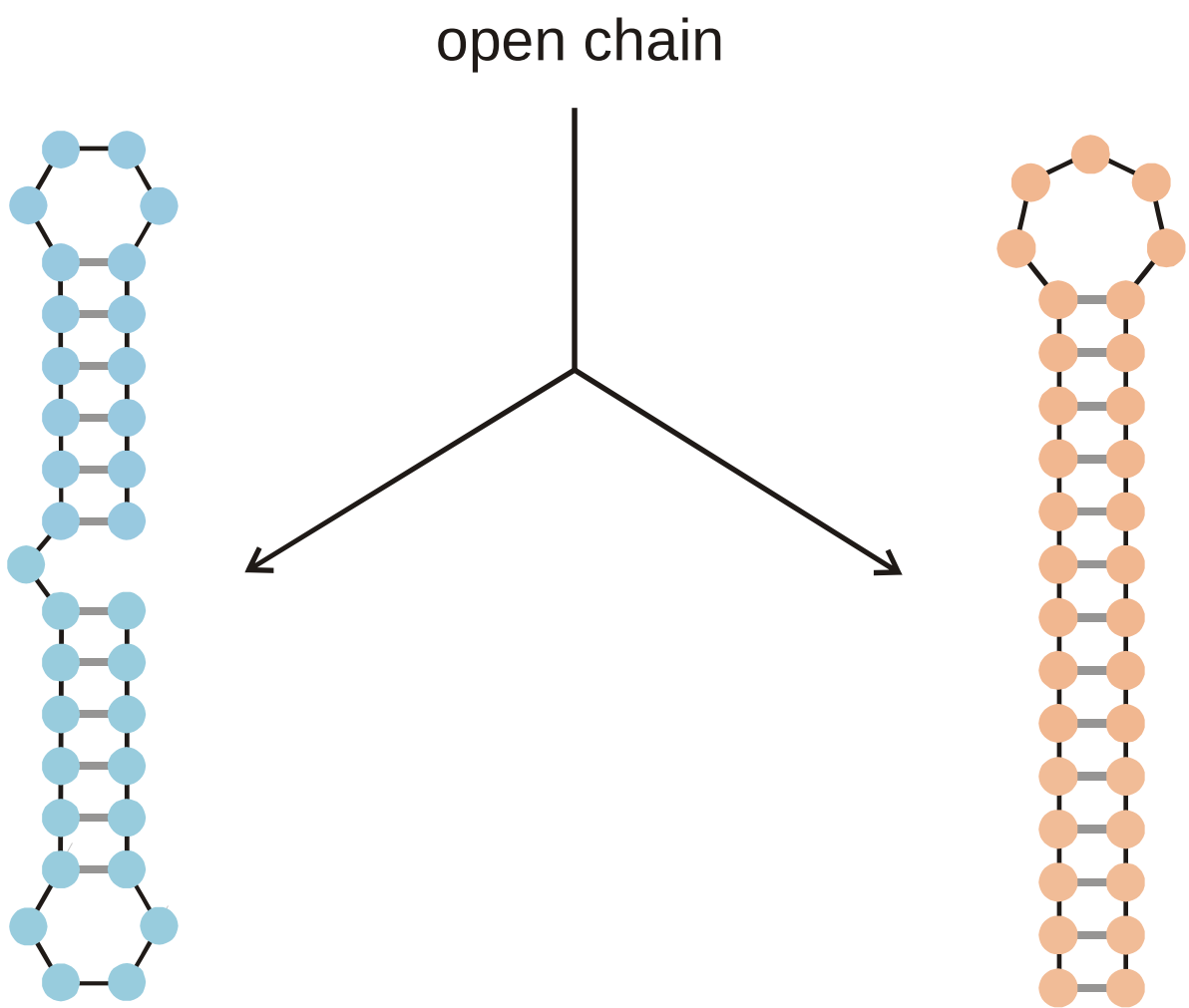
I_2 = **AUUGAGCAUAUUCAC**

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

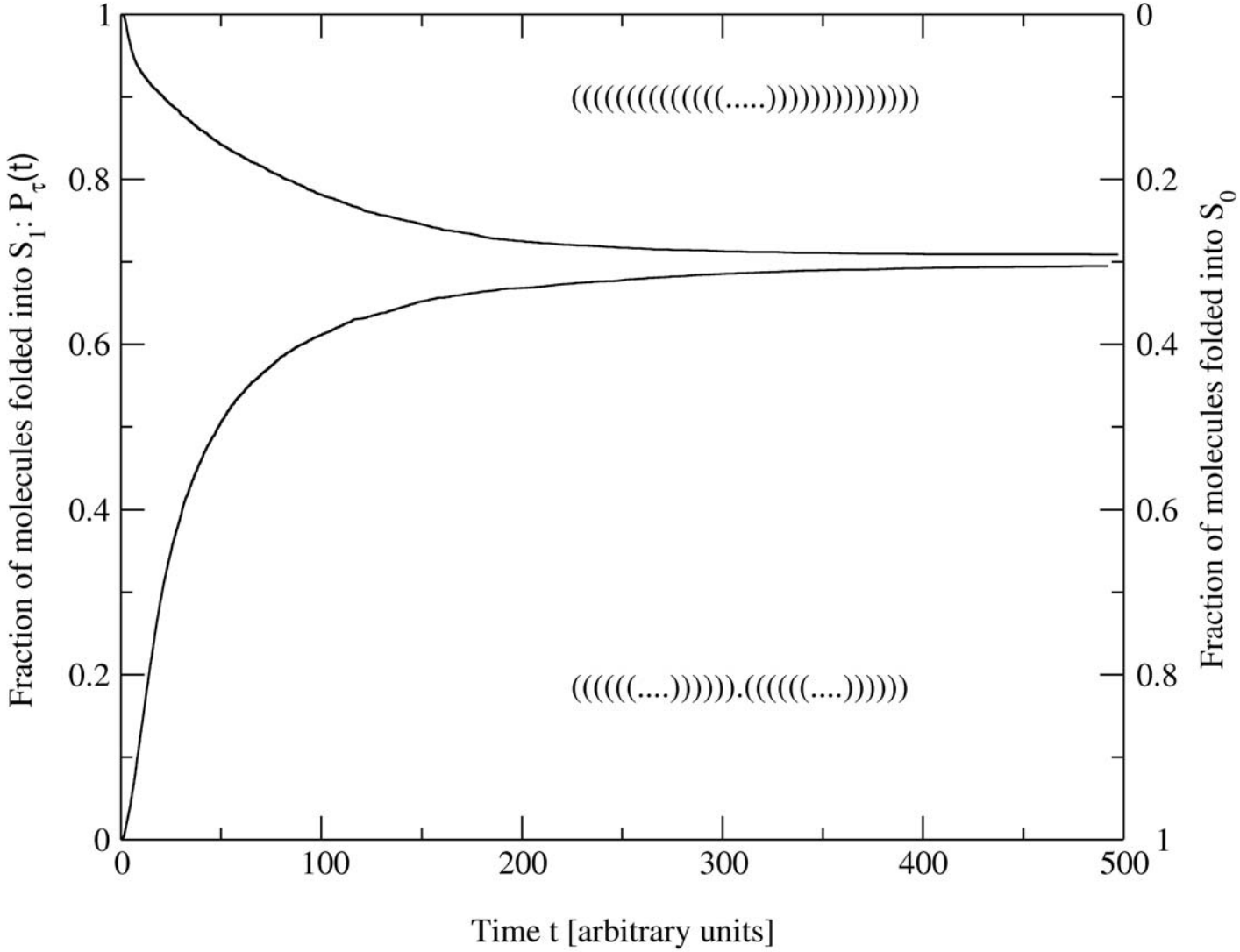


I_3 = **CGGGCUAUUUAGCUG**

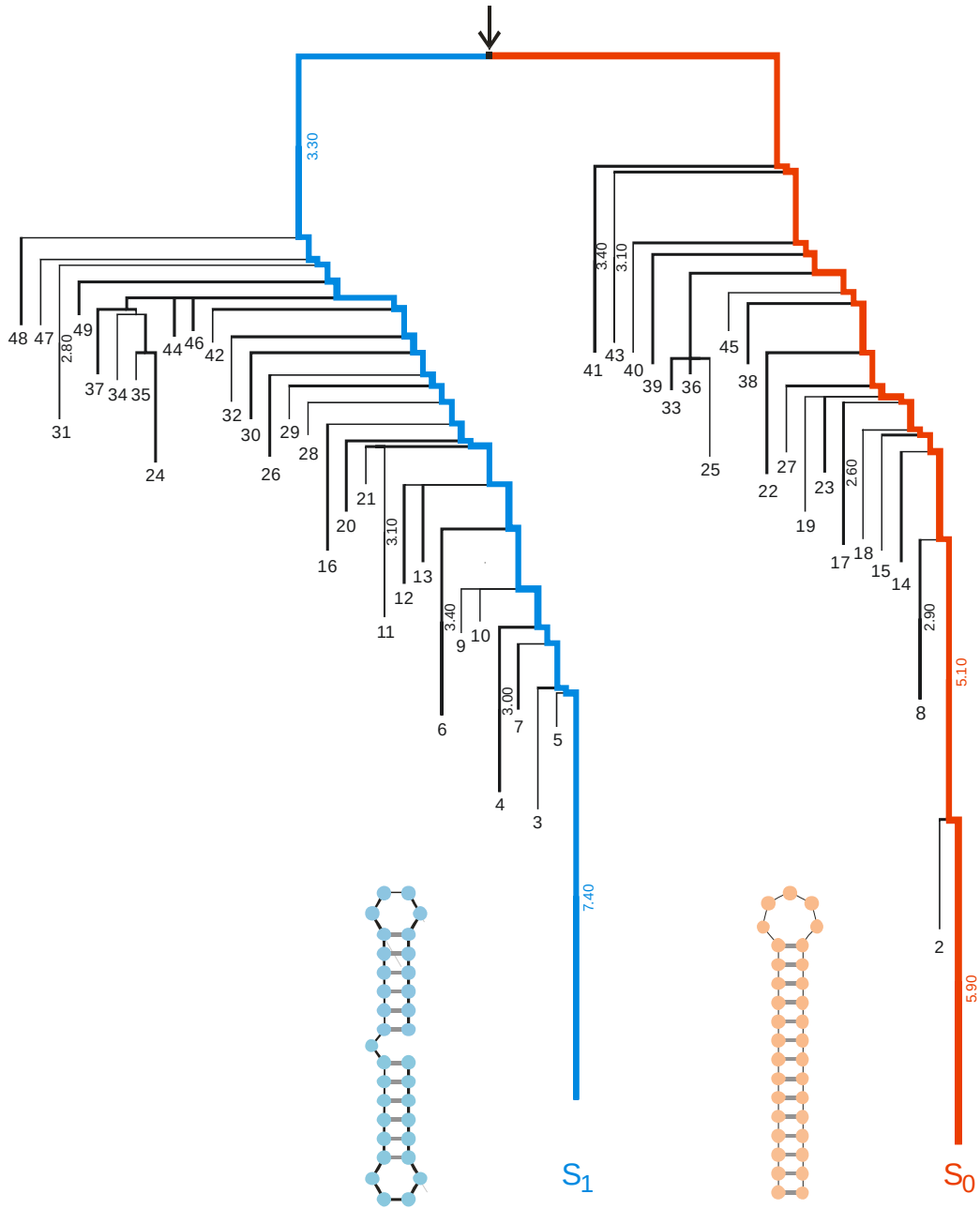
Example of an easily folding and especially stable small RNA molecule with n = 15



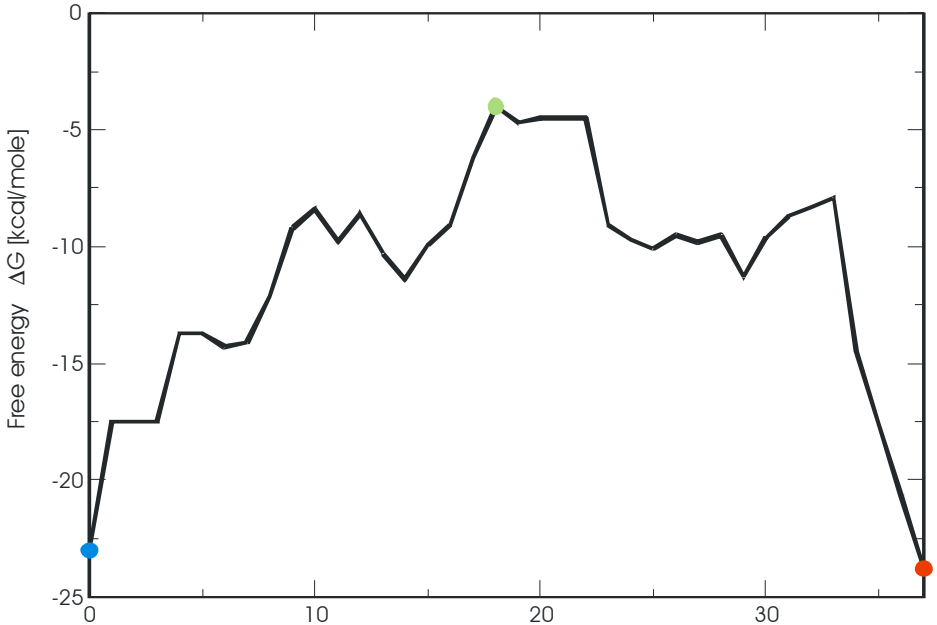
A nucleic acid molecule folding in two dominant conformations



Folding dynamics of the sequence **GGCCCCUUUGGGGGCCAGACCCCUAAAAAGGGUC**

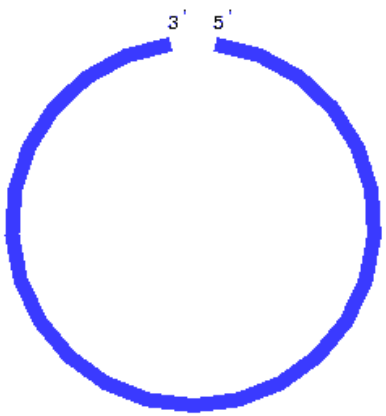


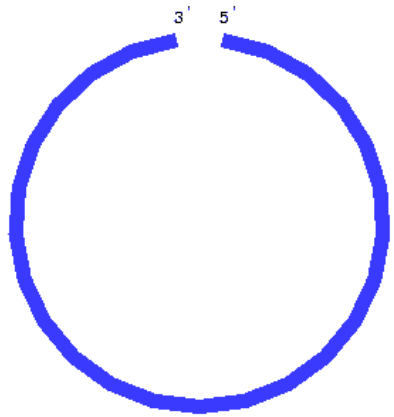
The barrier tree connecting S_1 and S_0



Structure	ΔG [kcal/mole]
((((((....)))))).((((((....))))))	-23.00
(((((.(...)))))).((((((....))))))	-17.50
((((.((...)))))).((((((....))))))	-17.50
(((.(((...)))))).((((((....))))))	-17.50
.((.(((...)))))..((((((....))))))	-13.70
.(.((((...)))))..((((((....))))))	-13.70
.(.((((...)))).).((((((....))))))	-14.30
...((((...)))).....((((((....))))))	-14.10
....(((...))).....((((((....))))))	-12.10
.....((...)).....((((((....))))))	-09.20
.....(.....).....((((((....))))))	-08.40
.....((((((....))))))	-09.80
.....(.....).((((((....))))))	-08.60
.....((...)).((((((....))))))	-10.30
.....(((...)))((((((....))))))	-11.40
.....((((...))))((((....)))))).	-09.90
.....(((((...)))))((((....))))))..	-09.10
.(...)(((((...)))))((((....))))))..	-06.20
.(....(((((...)))))((((....)))))).	-04.00
((....(((((...)))))((((....))))))	-04.70
(((...(((((...)))))).((....))))))	-04.50
((((..(((((...)))))))..((....))))))	-04.50
(((((.(((((...)))))))...(....))))))	-04.50
((((((((((...)))))).....))))))	-09.09
((((((((((...))))).....))))))))	-09.69
((((((((((...)))).....))))))))	-10.09
((((((((((...))))(.....).)))))))))	-09.50
((((((((((...)))).(.....))))))))))	-09.80
((((((((((...))))(.....))))))))))	-09.50
((((((((((...))))(....))))))))))	-11.30
(((((((((.....)))(....))))))))))	-09.60
(((((((((.(....)))(....))))))))))	-08.70
((((((((((....).(....))))))))))	-08.30
((((((((((....).....))))))))))	-07.94
((((((((((.....))))))))))	-14.48
((((((((((.....))))))))))	-17.60
((((((((((.....))))))))))	-20.70
((((((((((((....)))))))))))))	-23.80

The folding path from S_1 to S_0





Examples of two folding trajectories leading to different local minima

$$\frac{d x_k}{d t} = \sum_j k_{kj} x_j - x_k \sum_j k_{jk} \; ; \quad x_k = [S_k]$$

$$k_{kj} = k \cdot Z \cdot e^{-\Delta G_{kj} / RT}$$

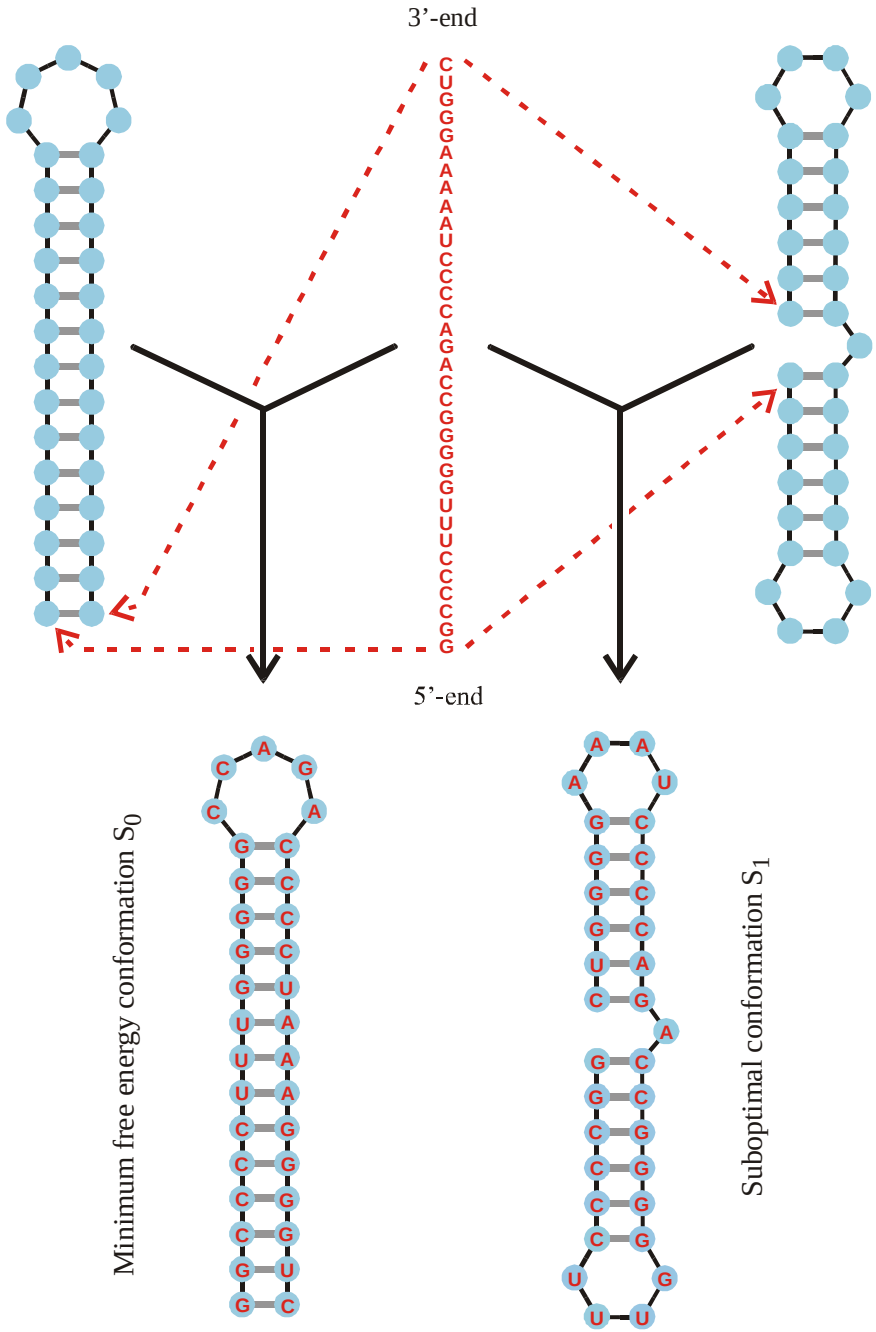
Arrhenius-type kinetics of RNA folding

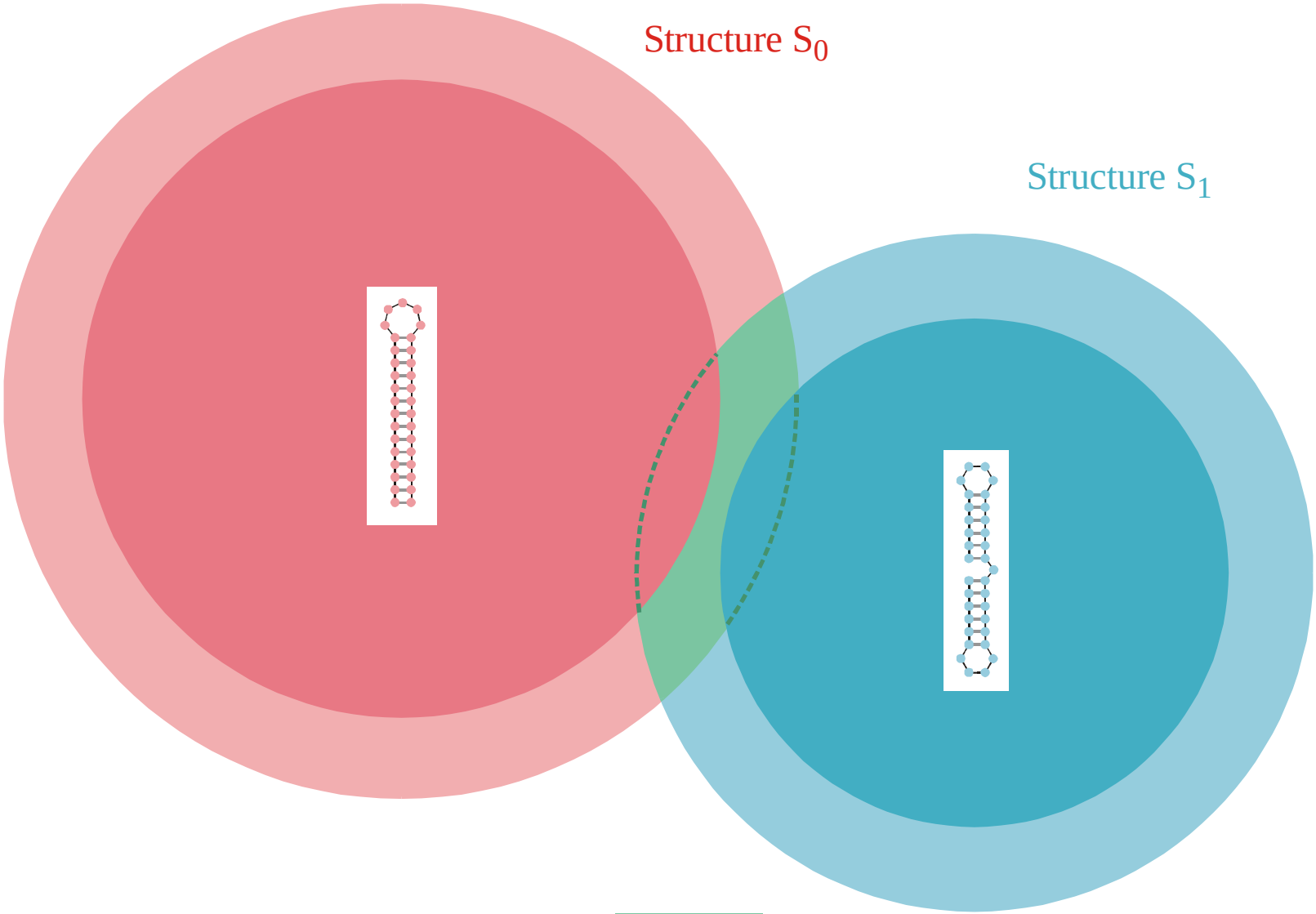
1. One sequence – one structure problem
2. Inverse folding and neutral networks
3. Kinetic folding
4. **Intersections and conformational switches**
5. Cofolding of nucleic acid molecules

RNA molecules switching between conformations

1. Self-induced switches
2. Externally induced switches
    1. External parameters (T, p, pH, I, ...)
    2. Binding of small ligands
    3. Chemical modification (tRNA)

One sequence is compatible with two structures





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

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

Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.

0092-8240/97 $17.00 + 0.00

ELSEVIER

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,*

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

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 $\mathcal{J}(s, s') \cong D_m$ 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*).

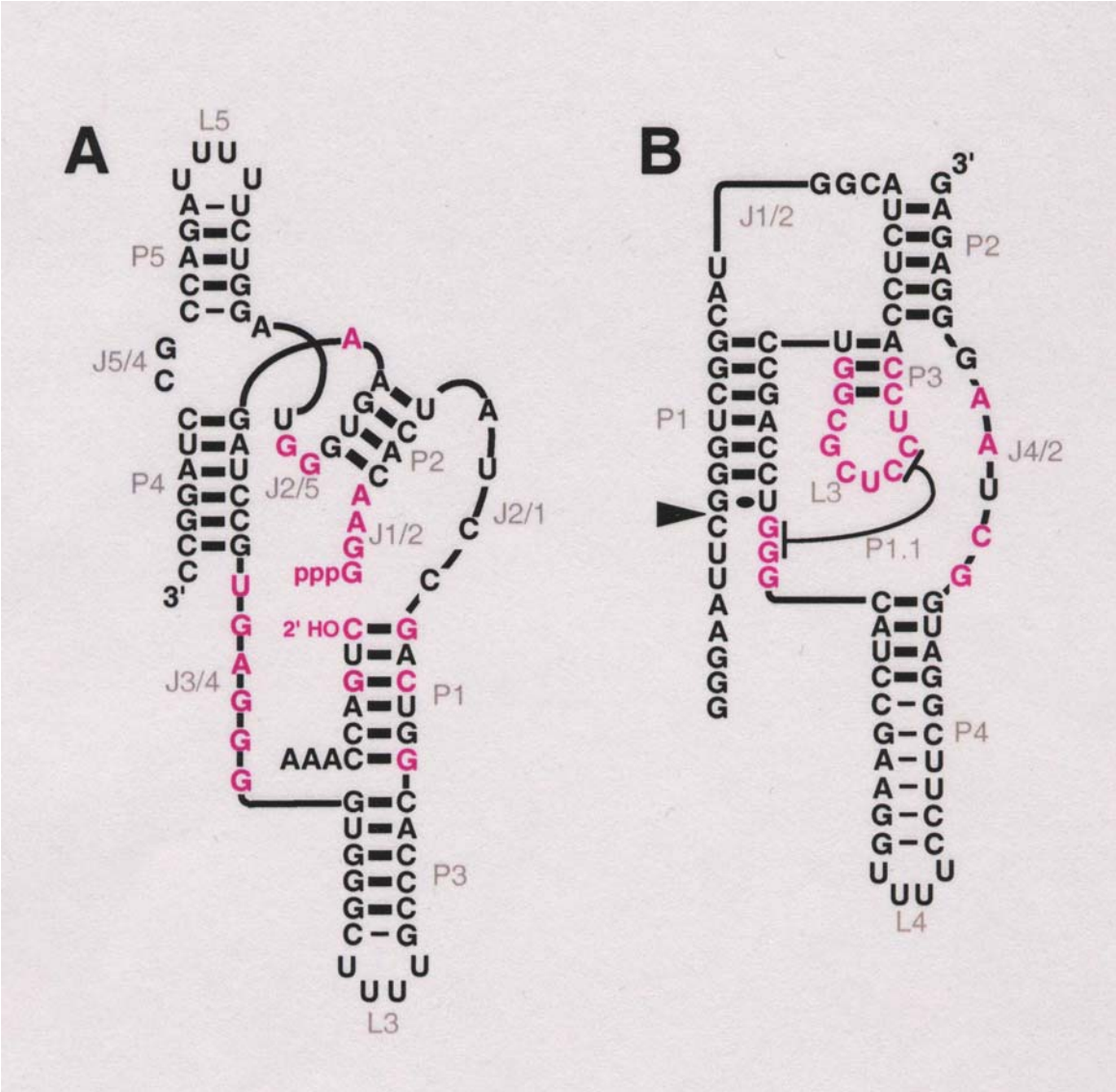
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. 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.

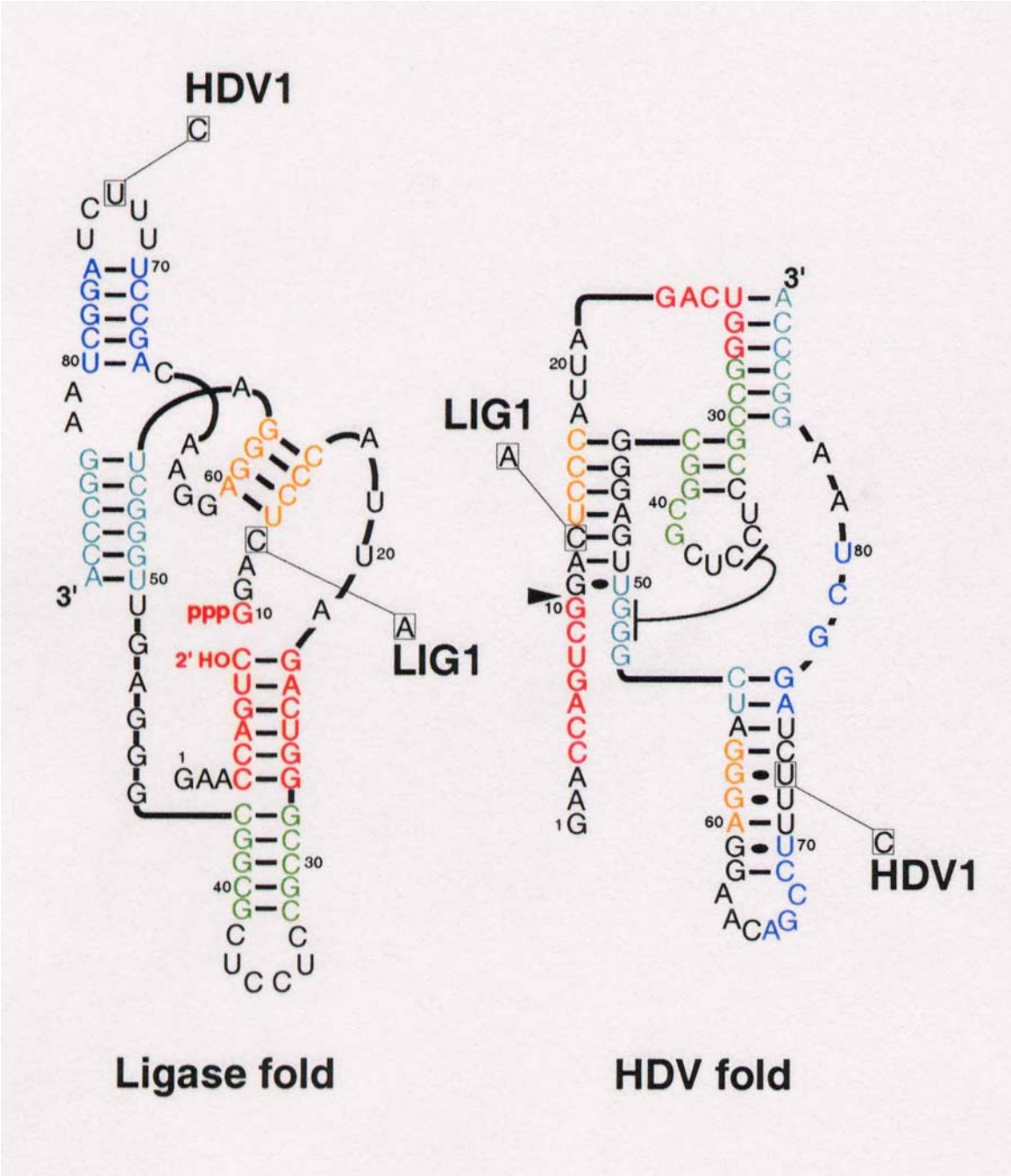
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

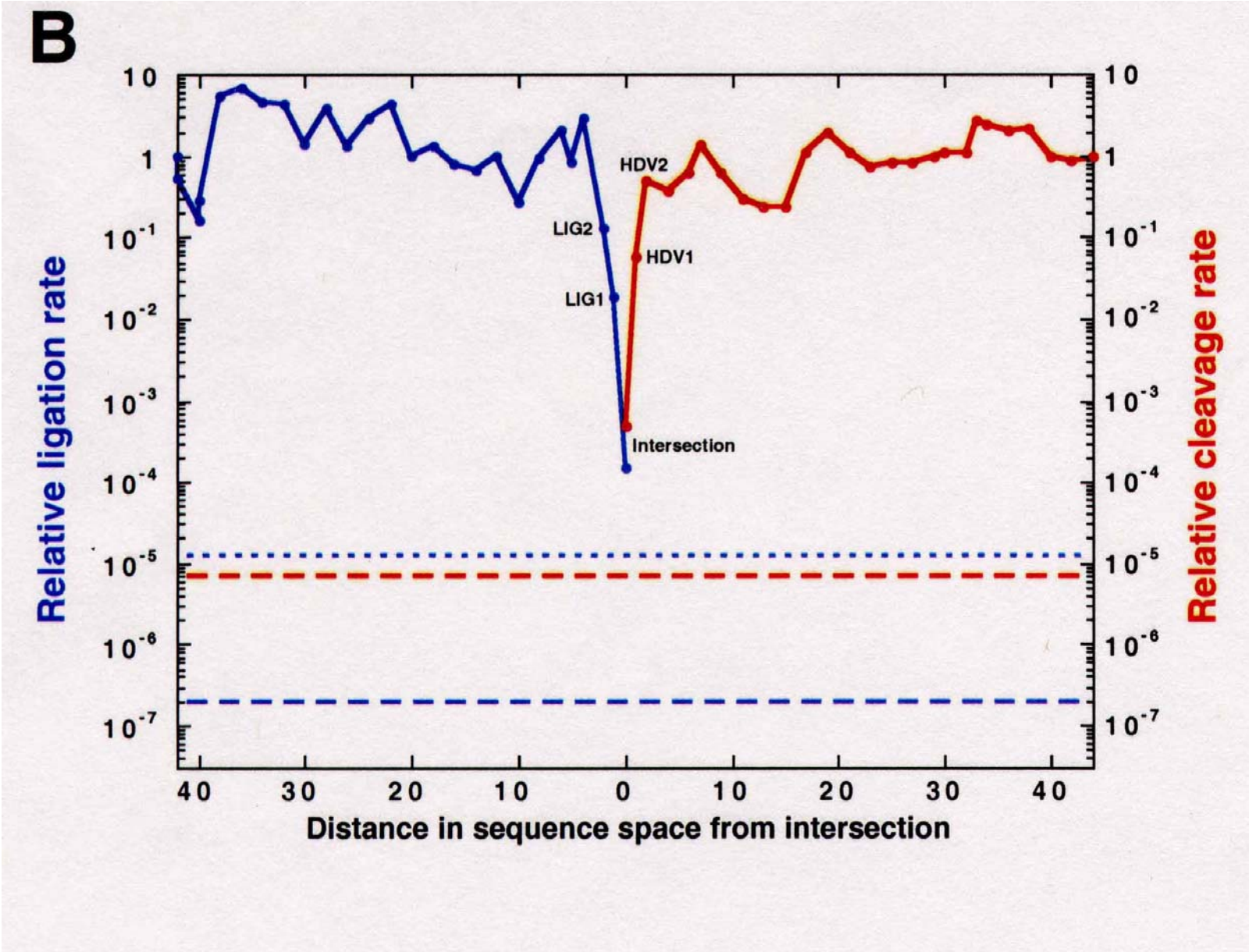


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

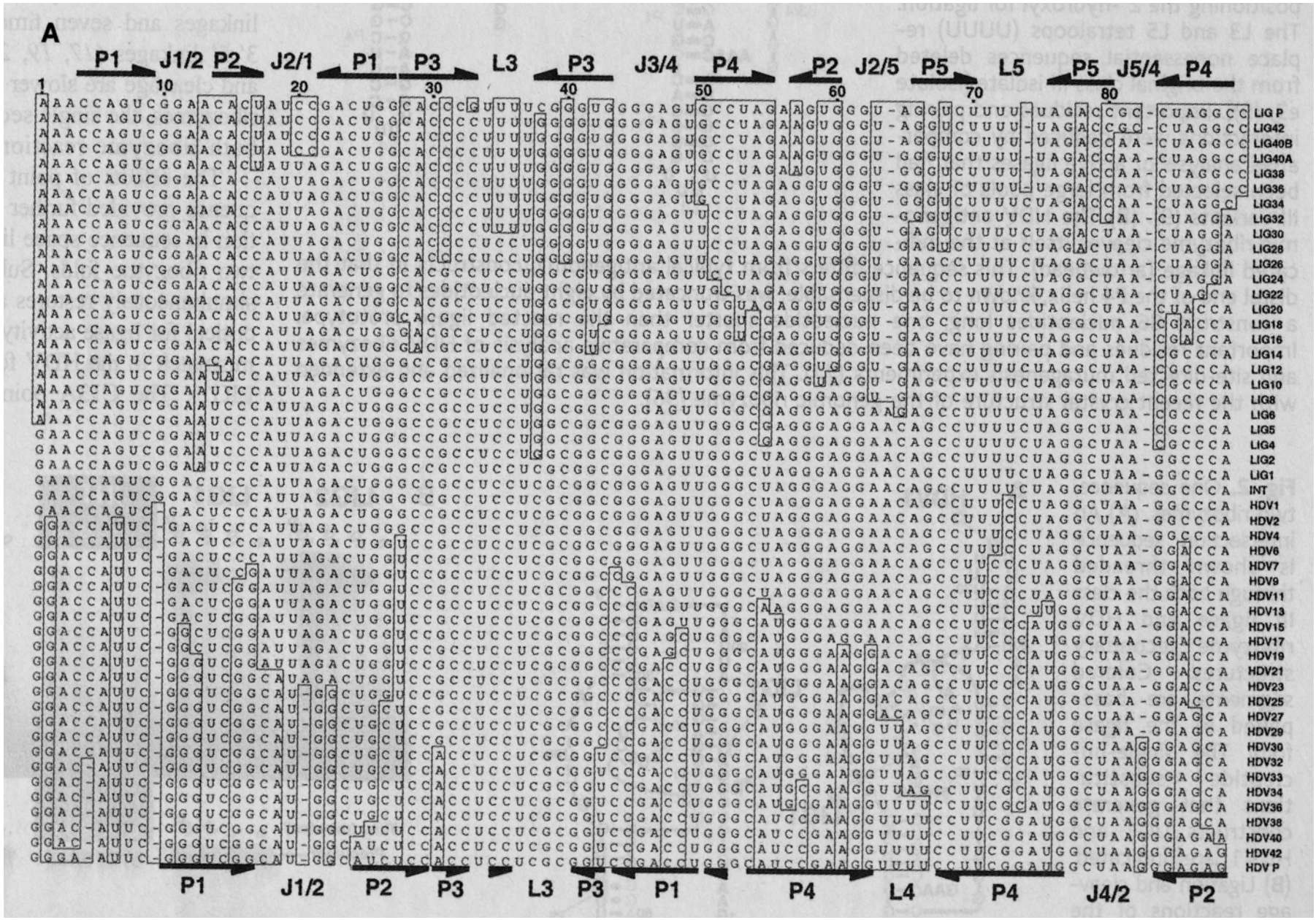
The sequence at the ***intersection***:

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





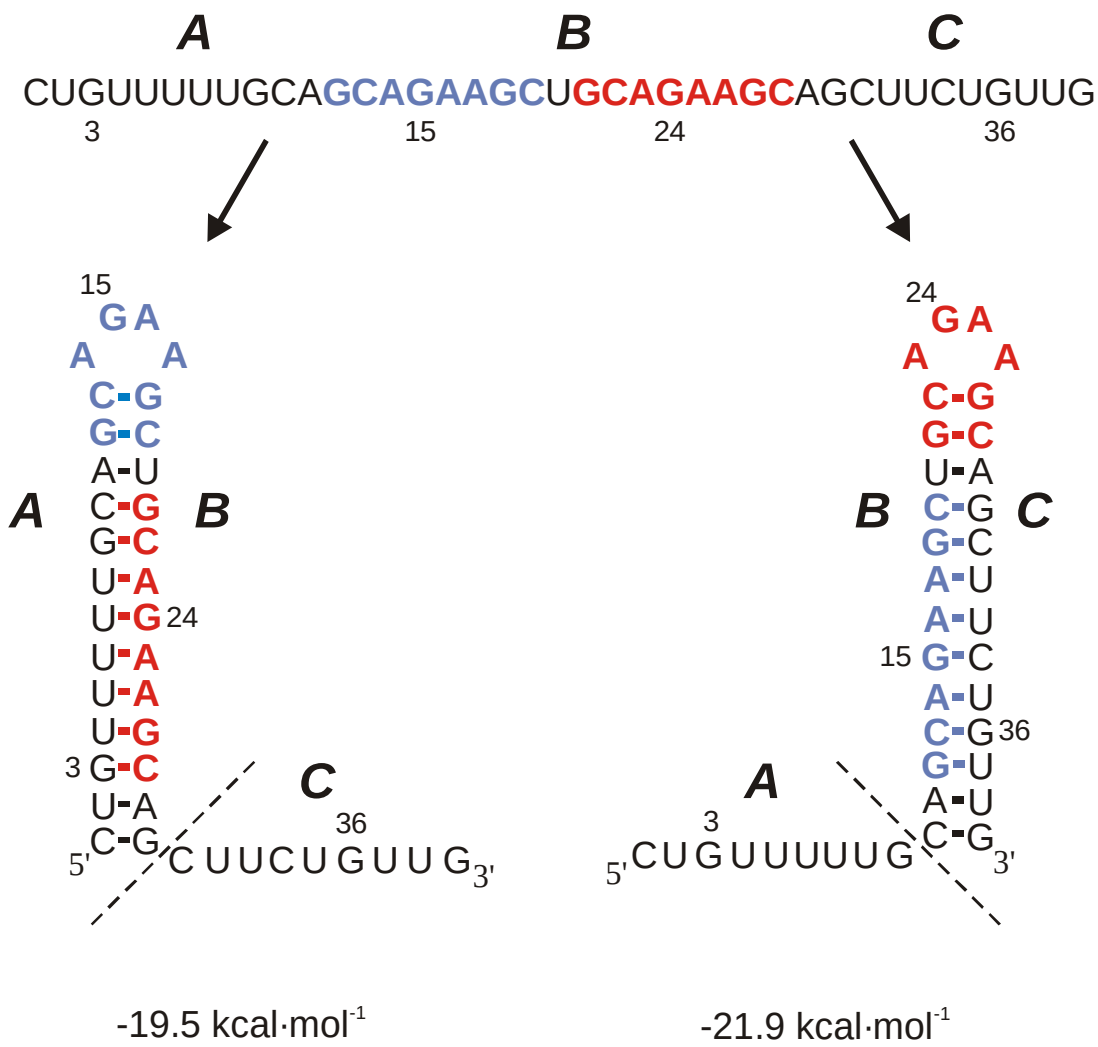
Two neutral walks through sequence space with conservation of structure and catalytic activity



Sequence of mutants from the intersection to both reference ribozymes

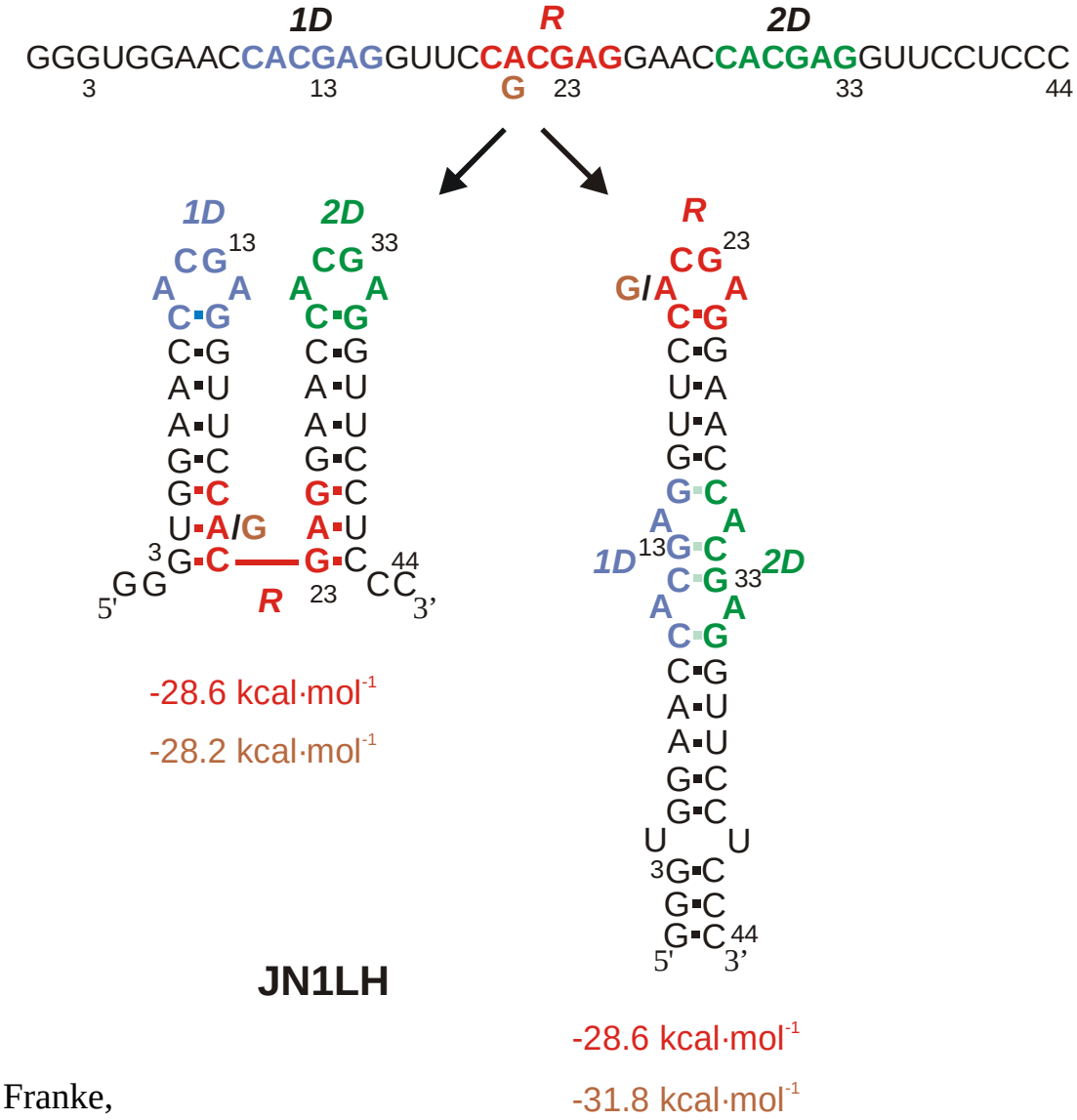
J. H. A. Nagel, C. Flamm, I. L. Hofacker, K. Franke, M. H. de Smit, P. Schuster, and C. W. A. Pleij. ***Structural parameters affecting the kinetic competition of RNA hairpin formation***, in press 2004.

J. H. A. Nagel, J. Møller-Jensen, C. Flamm, K. J. Öistämö, J. Besnard, I. L. Hofacker, A. P. Gultyaev, M. H. de Smit, P. Schuster, K. Gerdes and C. W. A. Pleij. ***The refolding mechanism of the metastable structure in the 5'-end of the*** hok ***mRNA of plasmid*** R1, submitted 2004.



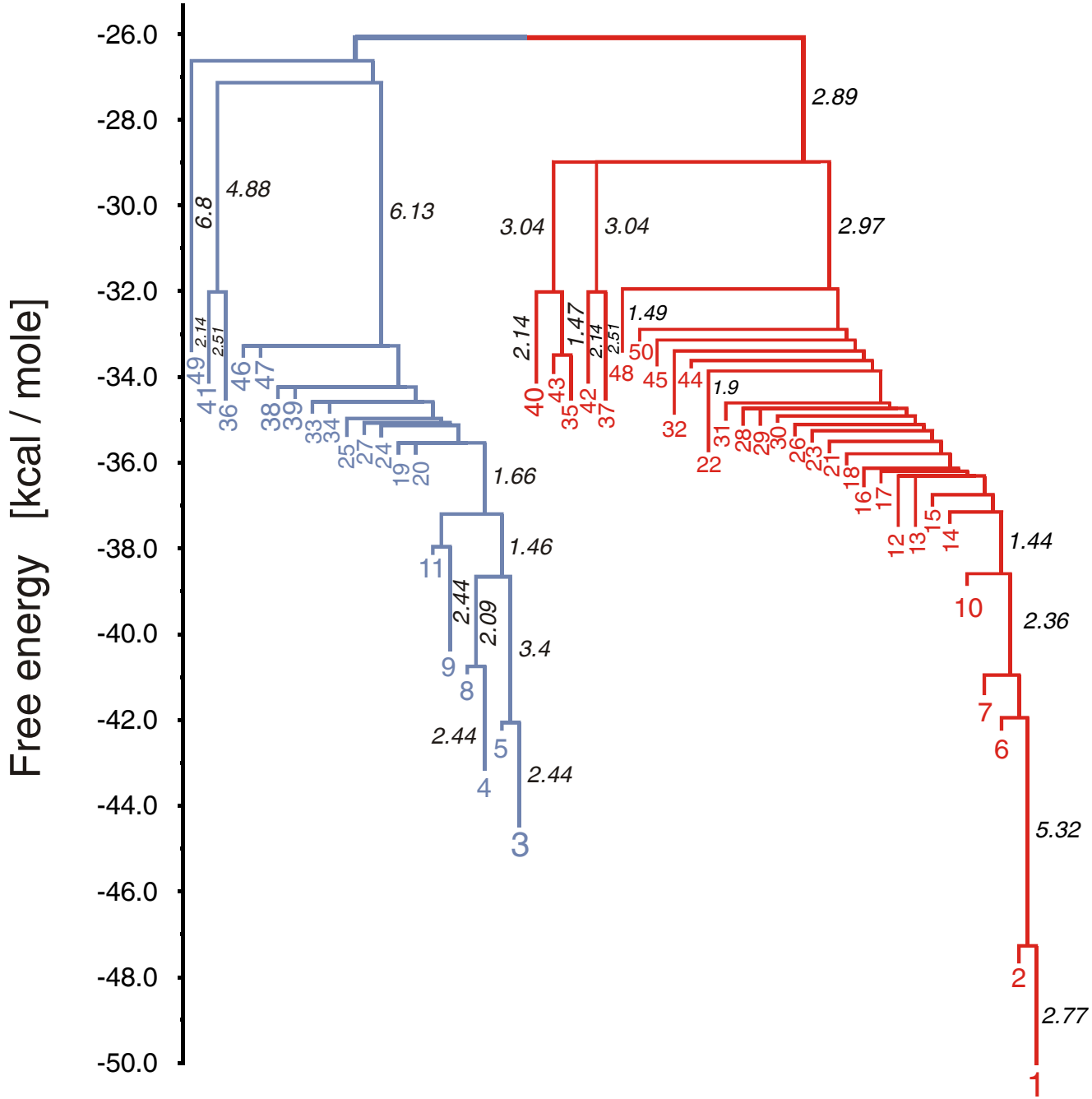
J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij.

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

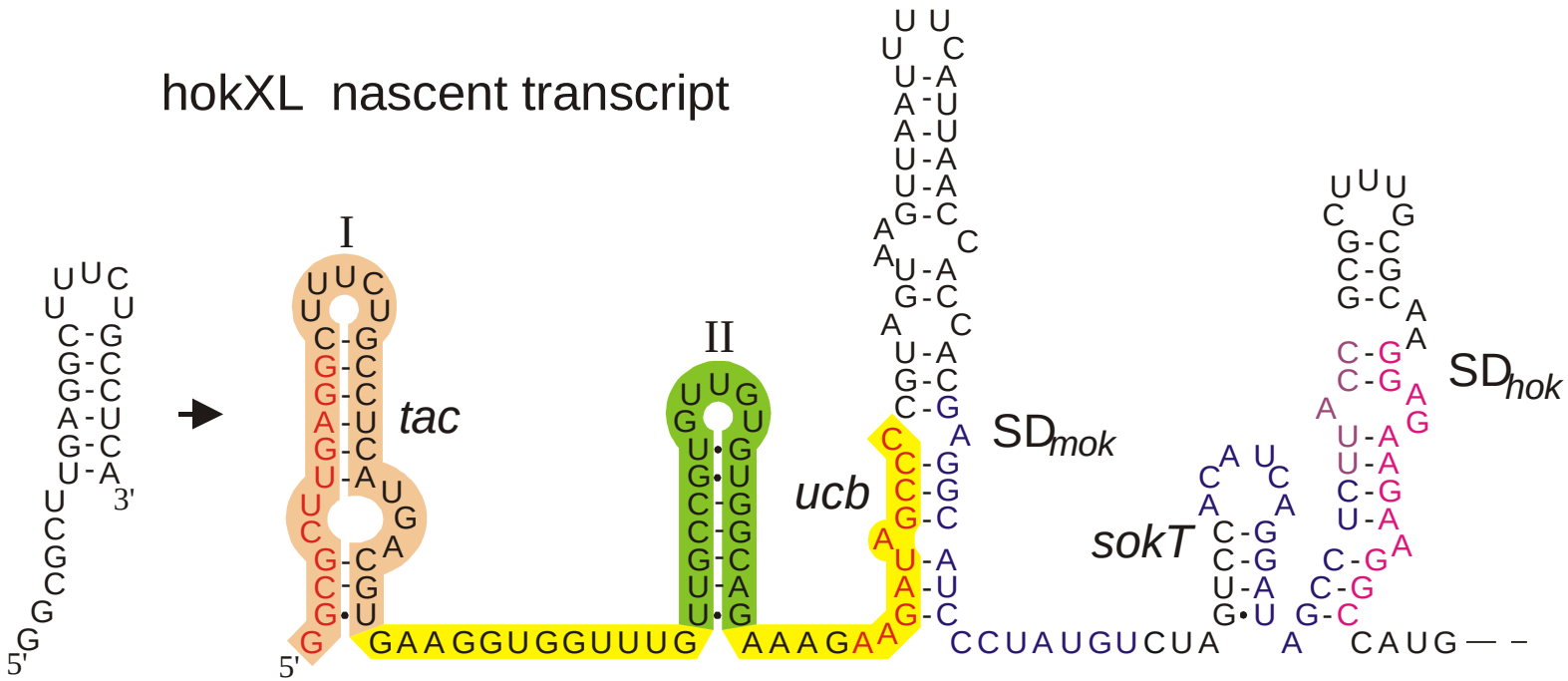


J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij.

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

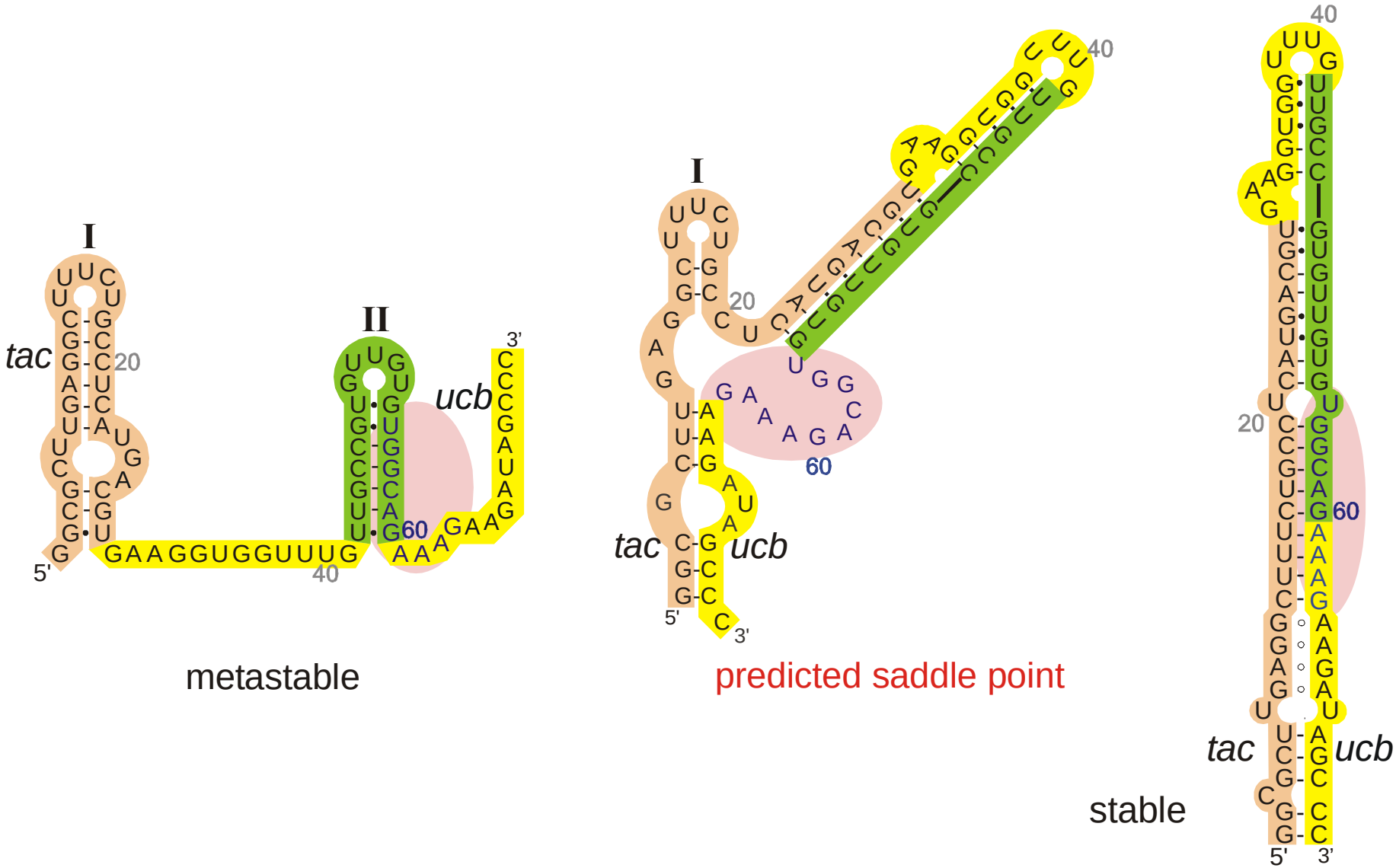


J1LH barrier tree

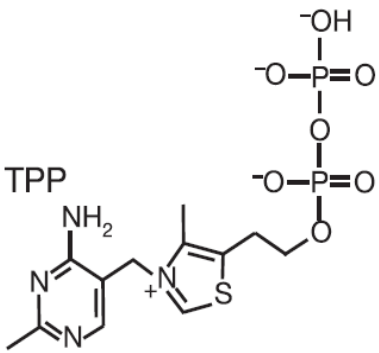


J.H.A. Nagel, J. Møller-Jensen, C. Flamm, K.J. Öistämö, J. Besnard, I.L. Hofacker, A.P. Gultyaev, M.H. de Smit, P. Schuster, K. Gerdes and C.W.A. Pleij.

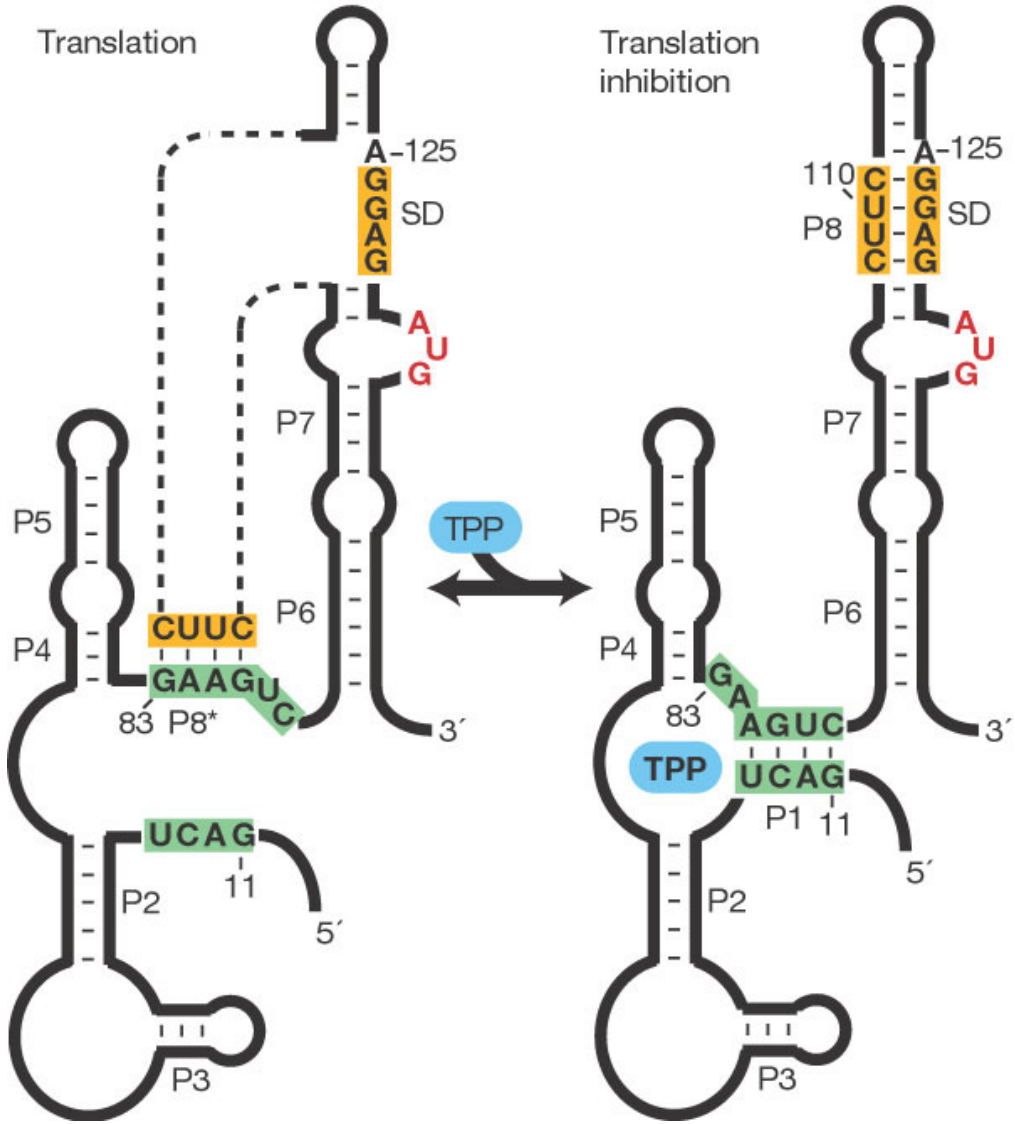
The refolding mechanism of the metastable structure in the 5'-end of the hok *mRNA of plasmid* R1, submitted 2004.



Transition from the metastable to the stable comformation

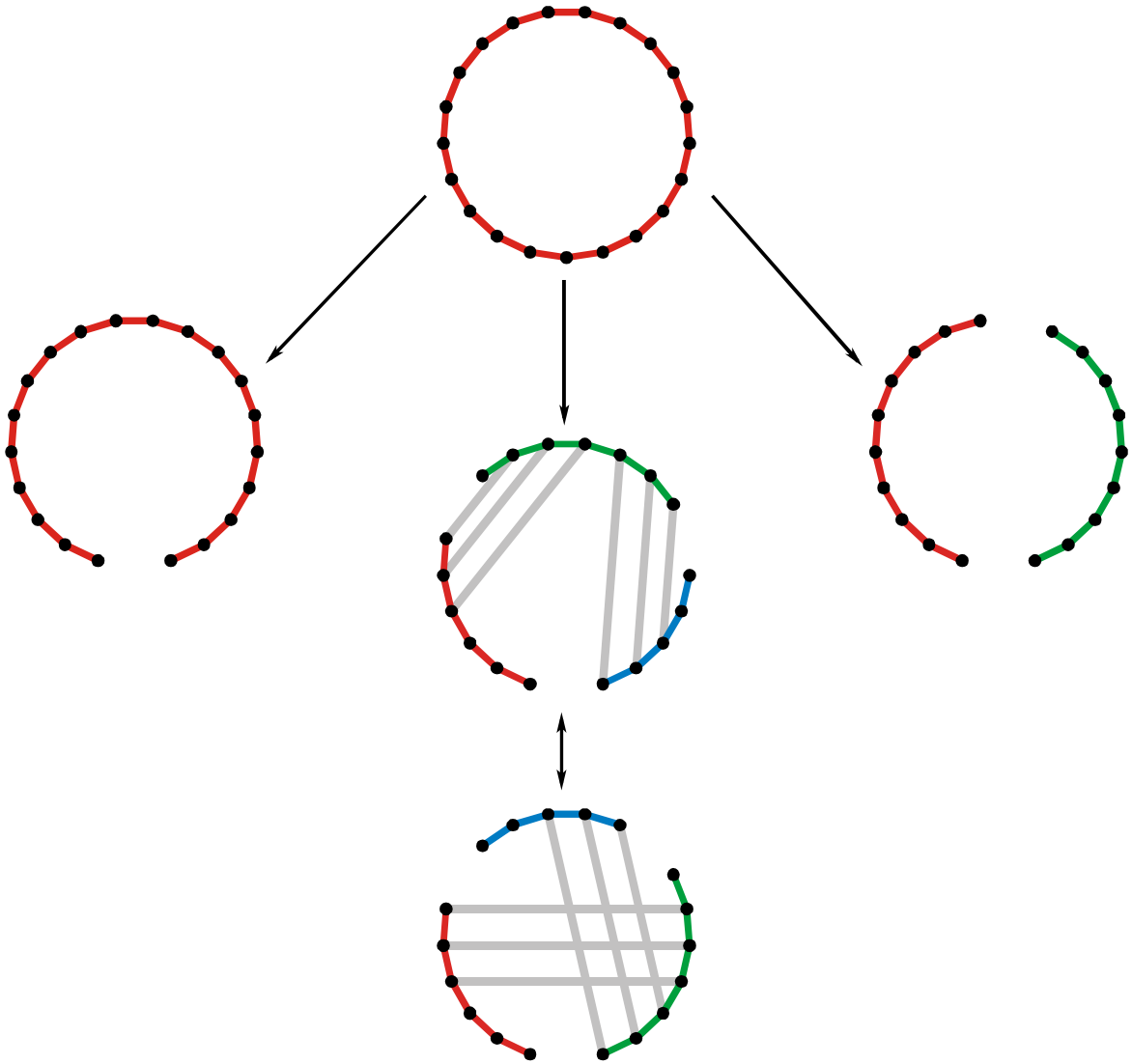


Thiamine-pyrophosphate

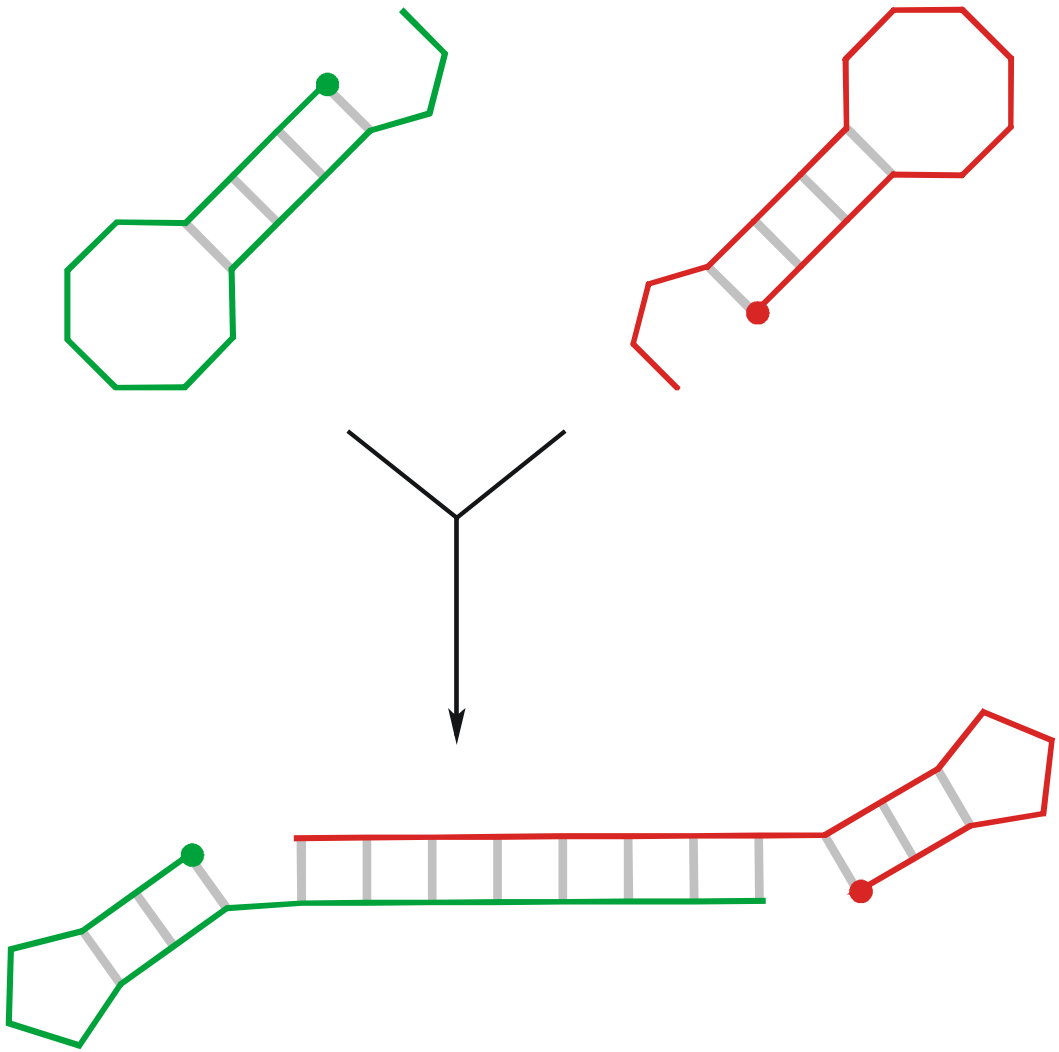


Wade Winkler, Ali Nahvi, and Ronald R. Breaker, ***Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression***. Nature **419**, 952-956, 2002.

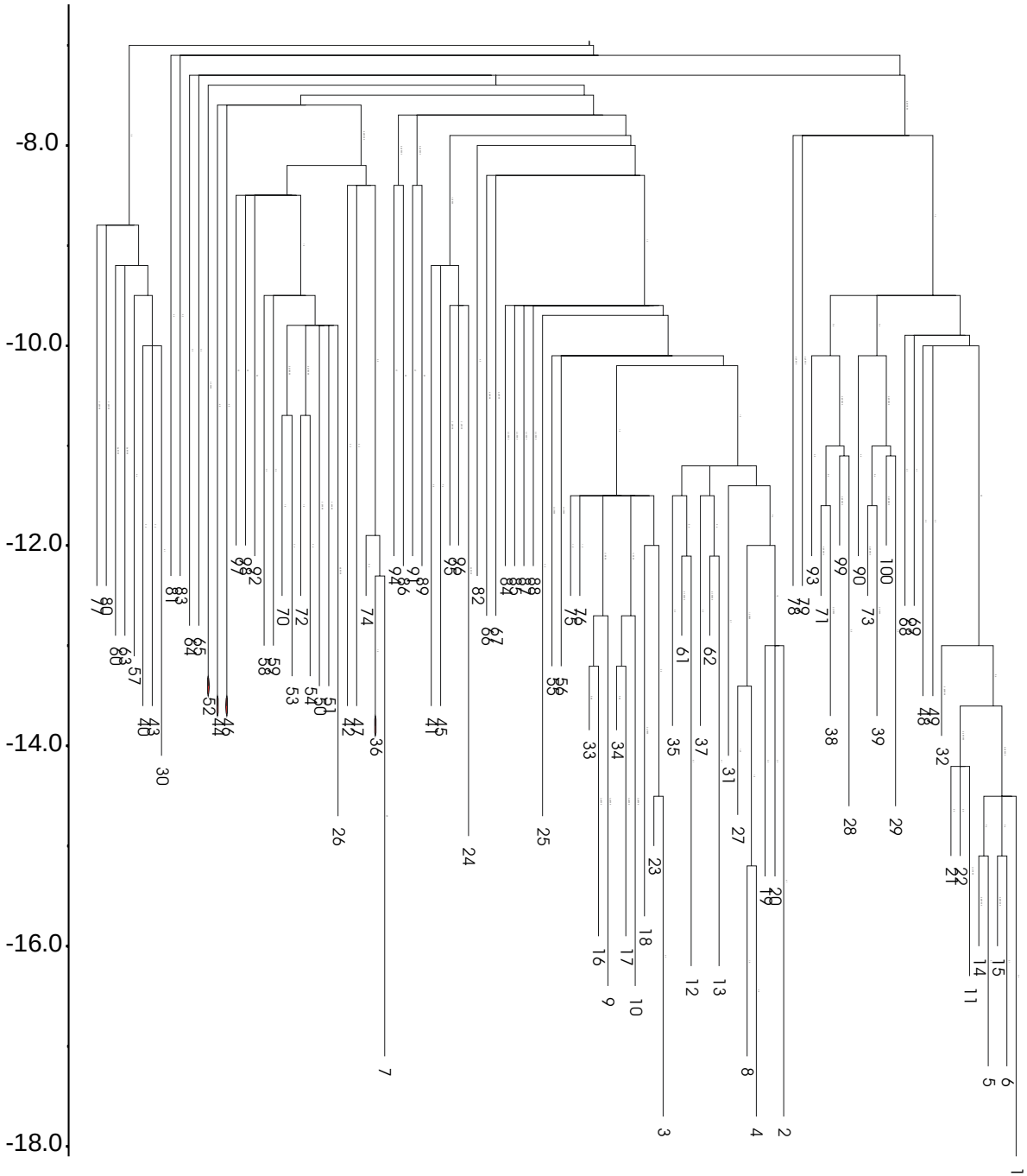
1. One sequence – one structure problem
2. Inverse folding and neutral networks
3. Kinetic folding
4. Intersections and conformational switches
5. **Cofolding of nucleic acid molecules**



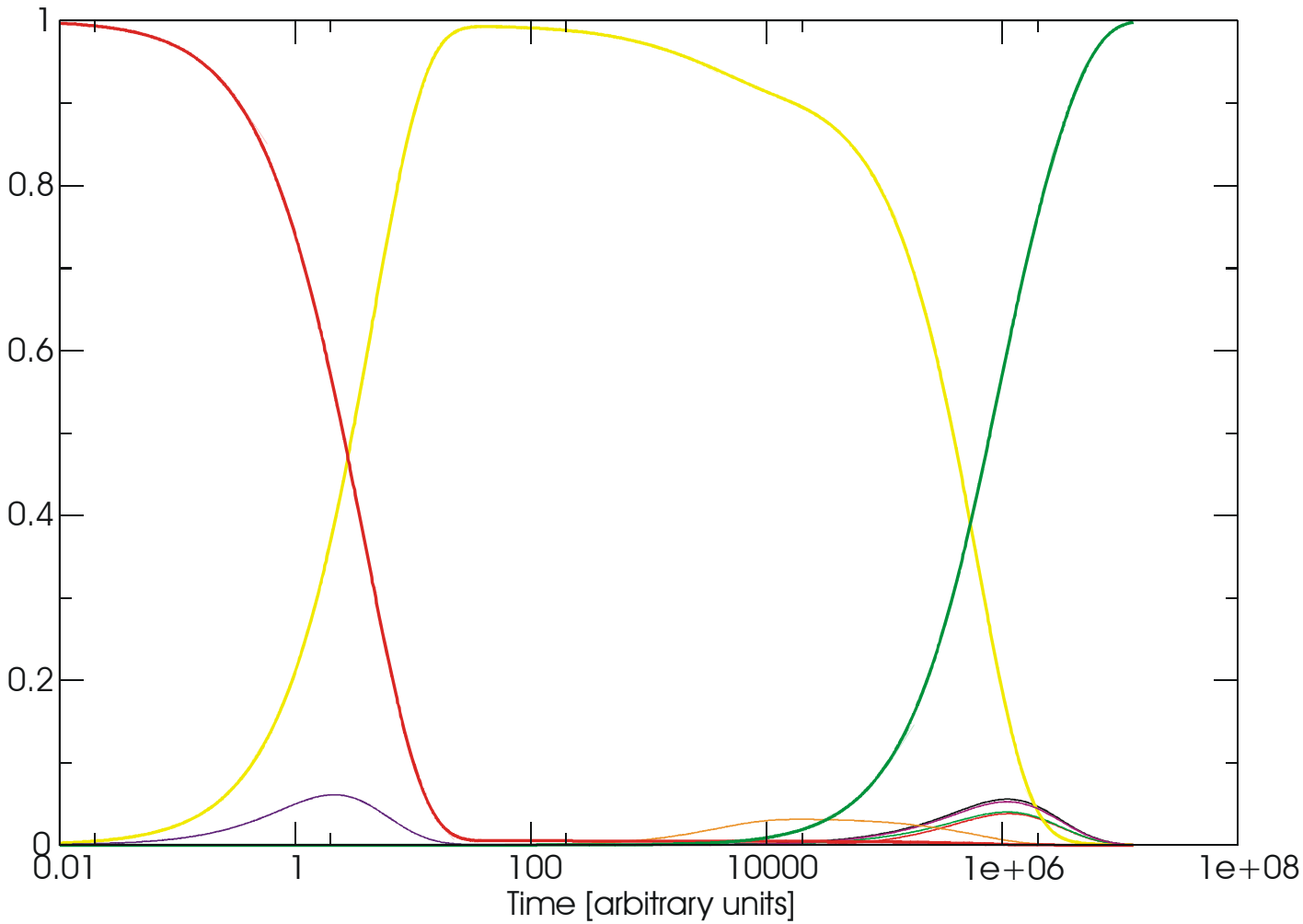
Cofolding two or three nucleic acid molecules



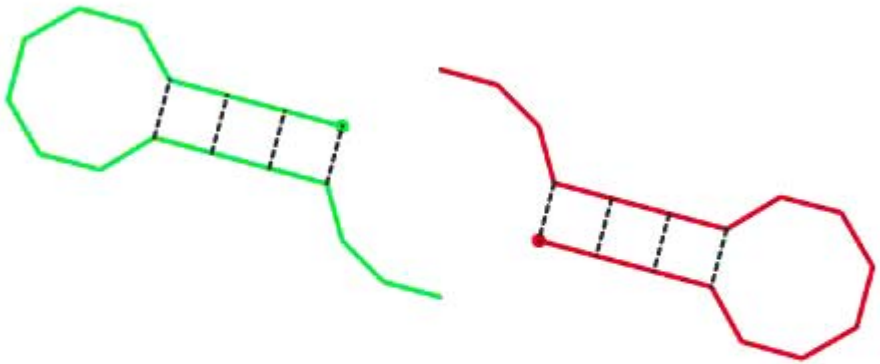
An example for ‘symmetric’ cofolding of two molecules



Cofolding tree



Cofolding kinetics



An example of a cofolding trajectory

Conclusions

I. Inverse folding of single stranded nucleic acids allows for efficient design of sequences with predefined secondary structures.

II. Common structures are formed by sequences of connected neutral networks in sequence space.

III. Molecules forming the same secondary structure differ by their suboptimal conformations and the kinetic folding behavior.

IV. Kinetic folding is indispensible for a detailed understanding of nucleic acid structures.

V. The design of molecules with two (meta)stable conformations and predefined barrier heights is straightforward.

VI. Cofolding of molecules or hybridization follows essentially the same principles as single molecule folding.

Acknowledgement of support

Universität Wien

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

Universität Wien and the Santa Fe Institute

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, MA

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

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

Robert Giegerich, Universität Bielefeld, GE

Heinz Engl, Stefan Müller, Johann Radon-Institut für Angewandte und Computergestützte Mathematik der Österreichischen Akademie der Wissenschaften, Linz, AT

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

Stefanie Widder, Lukas Endler, Universität Wien, AT
Ulrike Langhammer, **Ulrike Mückstein**, **Stefan Bernhart**

Web-Page for further information:

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