Kinetic Folding and Cofolding of RNA

From sequences to structures and back

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria
and
The Santa Fe Institute, Santa Fe, New Mexico, USA

EMBIO Kick-Off Meeting

Cambridge, UK, 25.– 27.07.2005

The Vienna RNA Group

Universität Wien

Walter Fontana, Harvard Medical School, MA

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

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

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

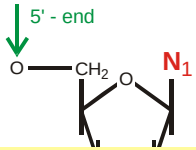
Kurt Grünberger, **Michael Kospach**, **Andreas Wernitznig**,
Stefanie Widder, **Michael Wolfinger**, **Stefan Wuchty**,Universität Wien, AT

Stefan Bernhart, **Jan Cupal**, **Lukas Endler**, **Ulrike Langhammer**,
Rainer Machne, **Ulrike Mückstein**, **Hakim Tafer**, **Stefan Washietl**, **N.N.**,
Universität Wien, AT

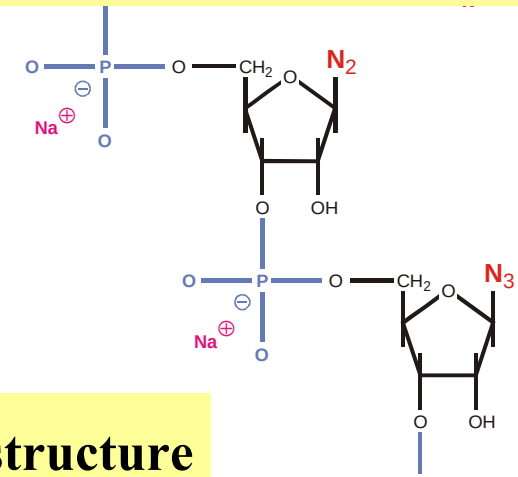
Ulrike Göbel, **Walter Grüner**, **Stefan Kopp**, **Jaqueline Weber**,
Institut für Molekulare Biotechnologie, Jena, GE

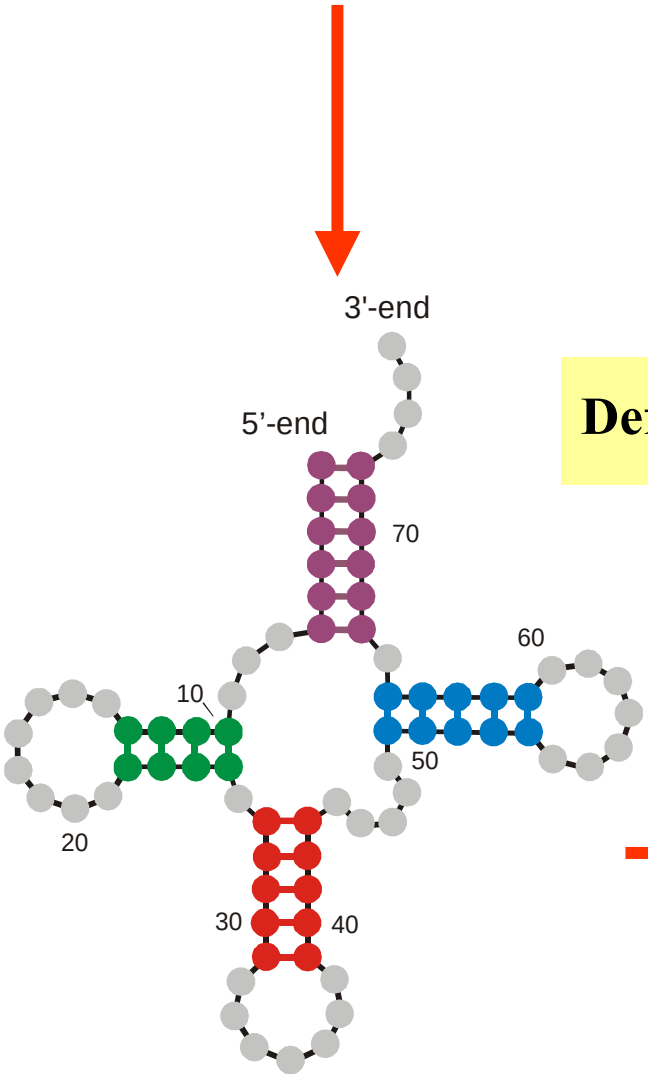
Web-Page for further information:

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

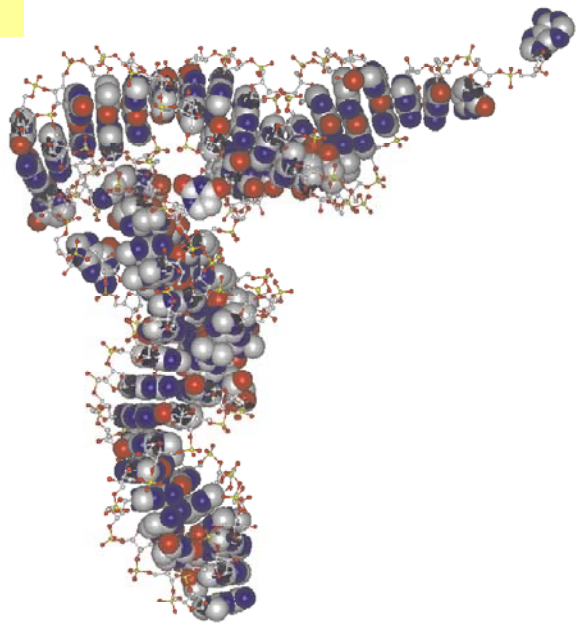


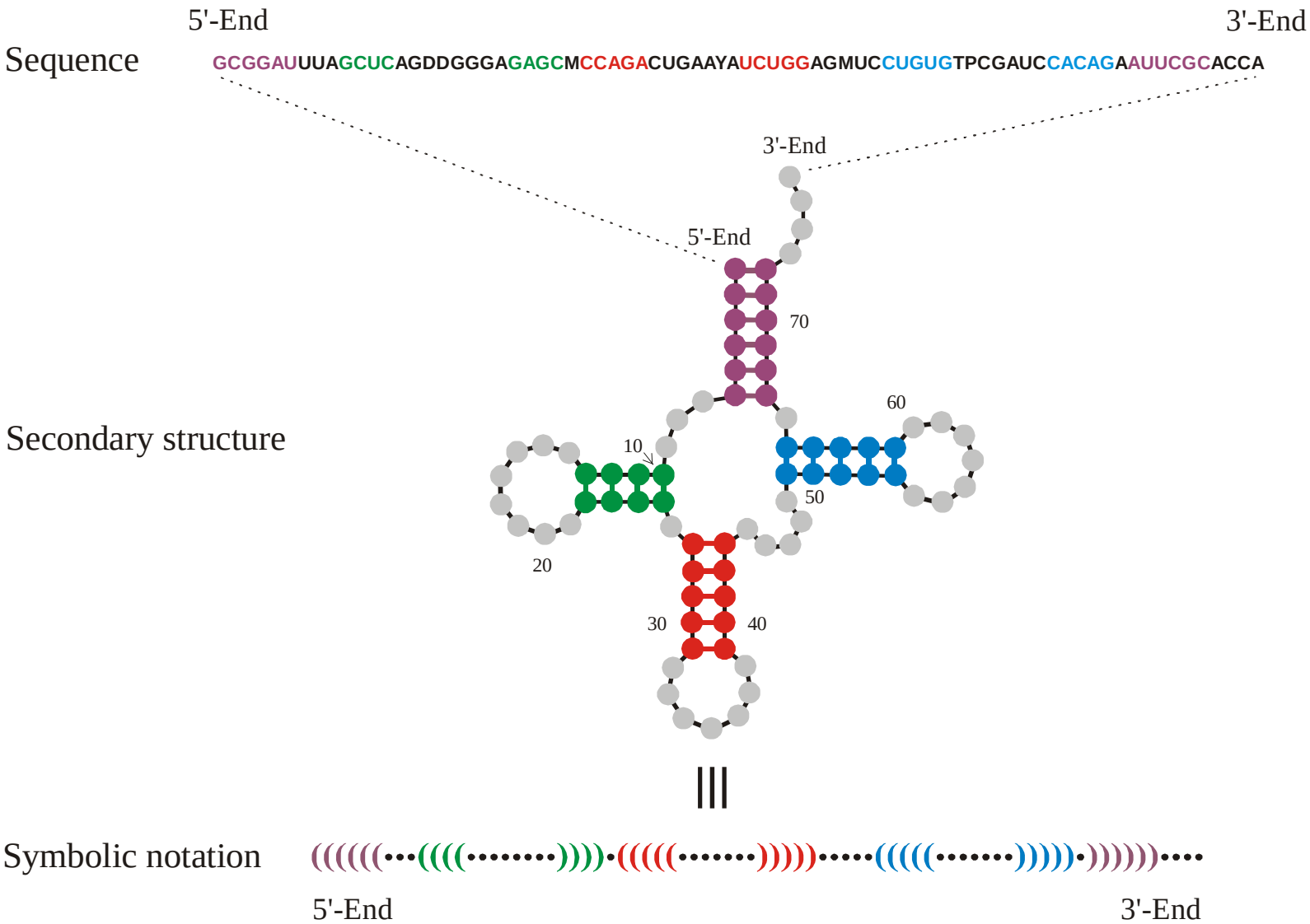
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end



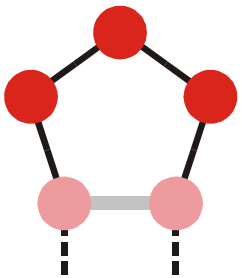


Definition of RNA structure



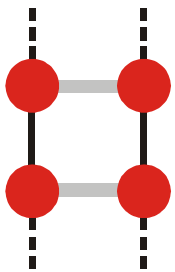


A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs



Minimal hairpin loop size:

$n_{lp} \geq 3$



Minimal stack length:

$n_{st} \geq 2$

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:

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

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

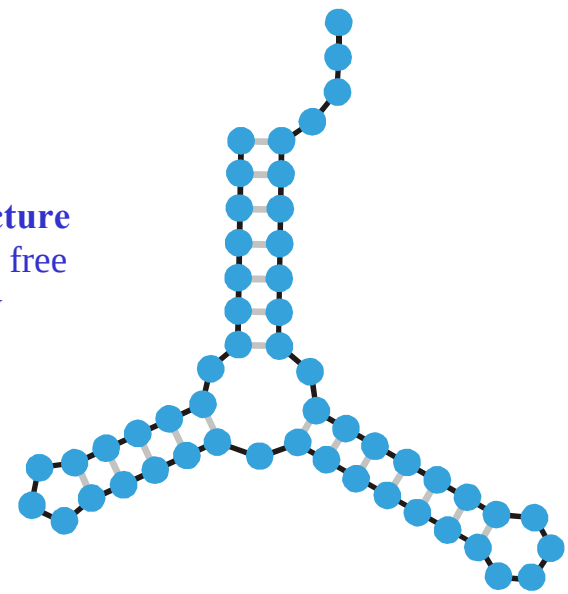
RNA folding:

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 ***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

Printed in Austria

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

ℓ	Number of Sequences 2^ℓ	4^ℓ	Number of Structures $S_\ell^{(3,2)}$	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

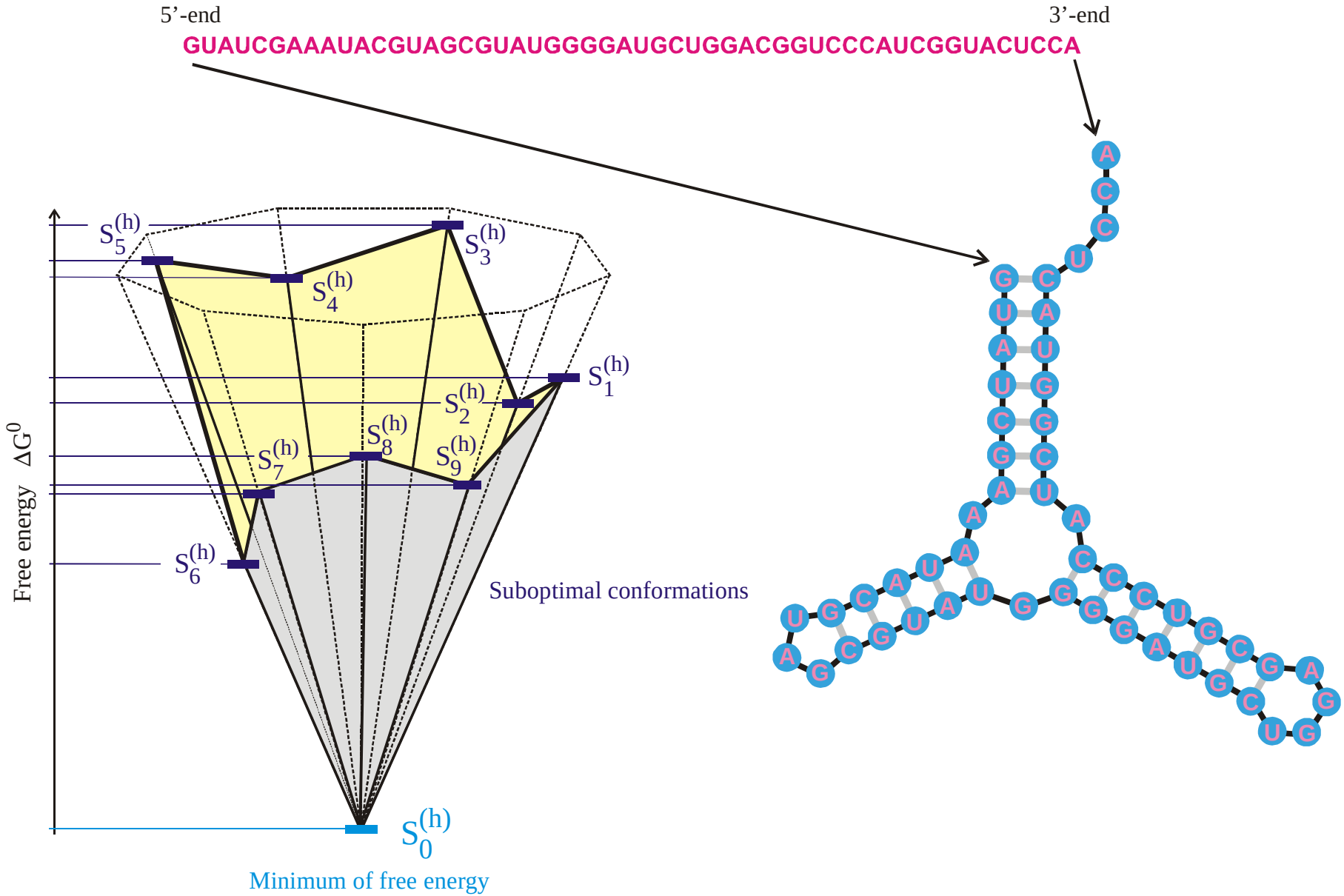
Computed numbers of minimum free energy structures over different nucleotide alphabets

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

Complete folding of sequence space and enumeration

GC, **AU** < 32
AUG, **GUC** < 23

AUGC < **17**



The minimum free energy structures on a discrete space of conformations

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

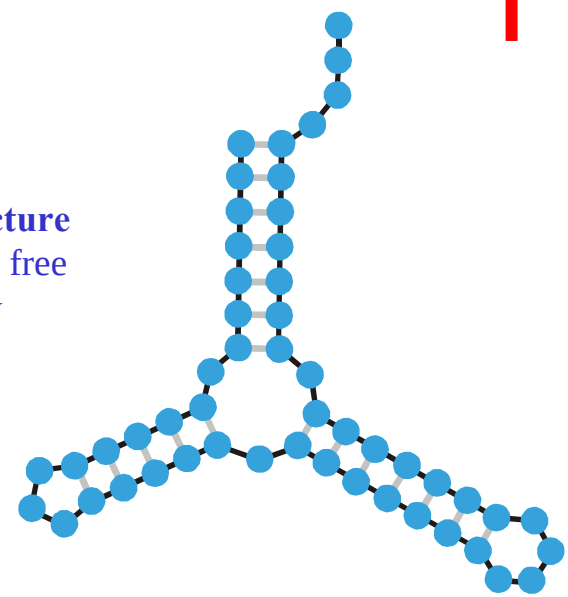
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

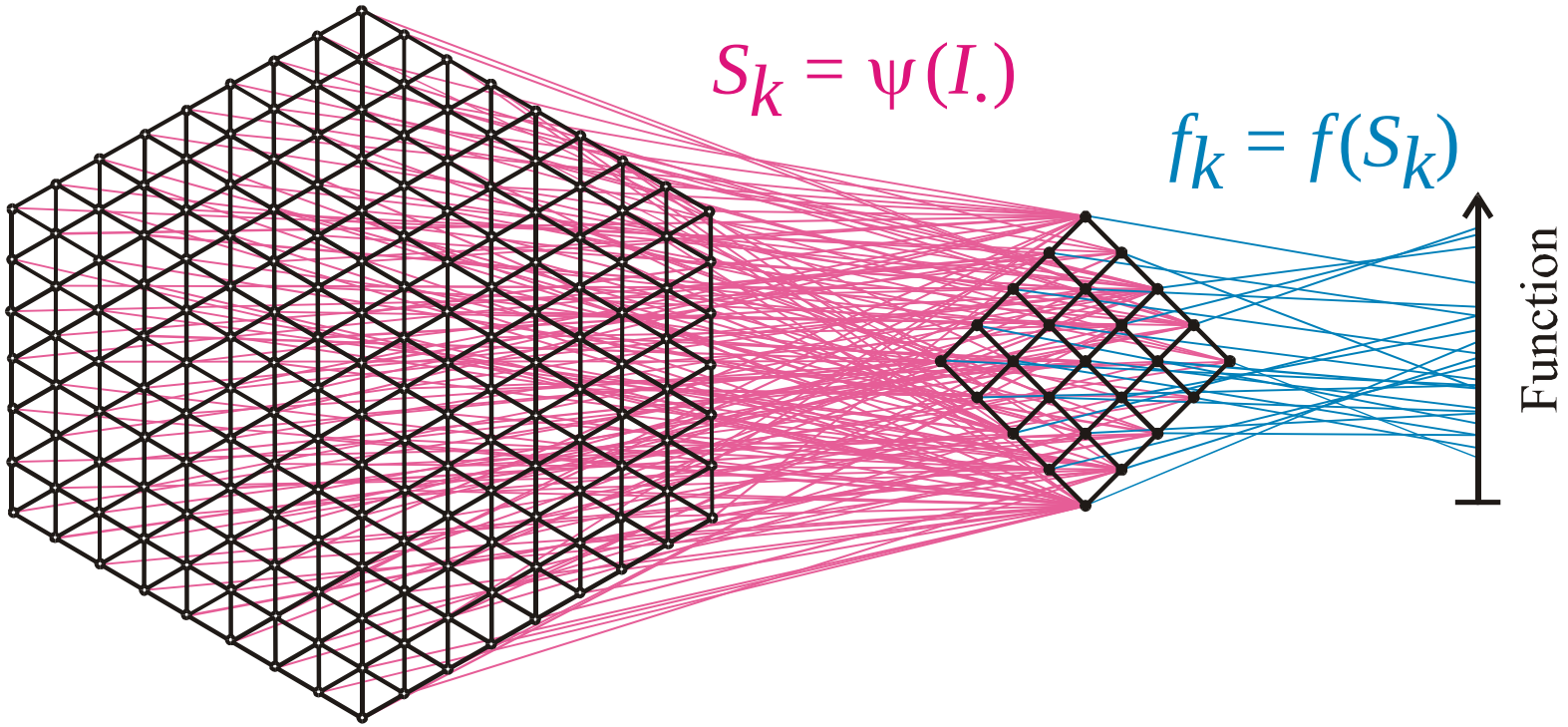
Inverse folding of RNA:

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

RNA structure of minimal free energy



Sequence, structure, and design



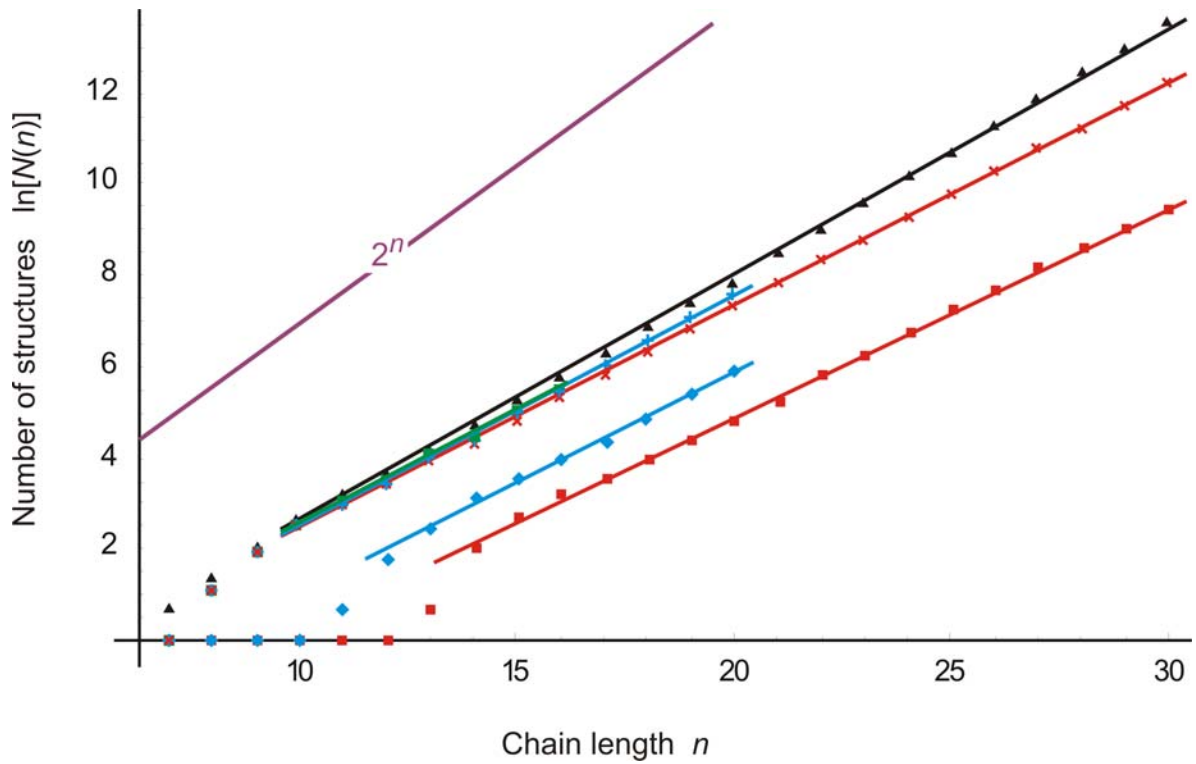
Mapping from sequence space into structure space and into function

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

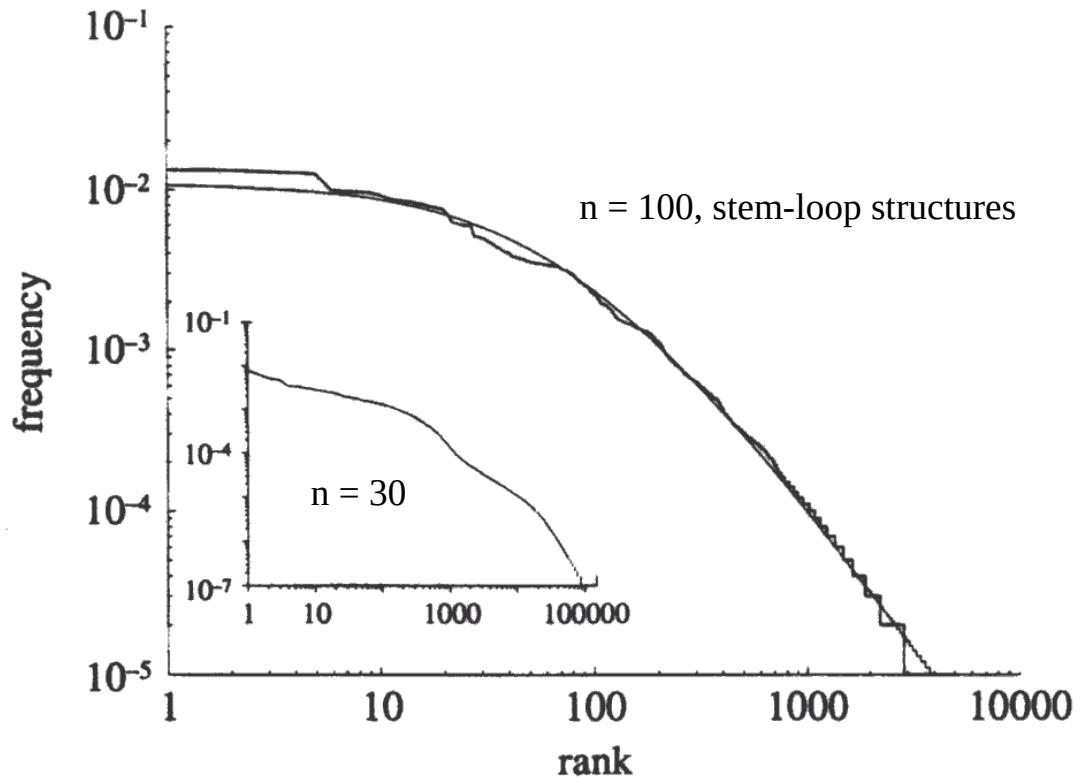


Properties of RNA sequence to secondary structure mapping

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

Properties of RNA sequence to secondary structure mapping

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



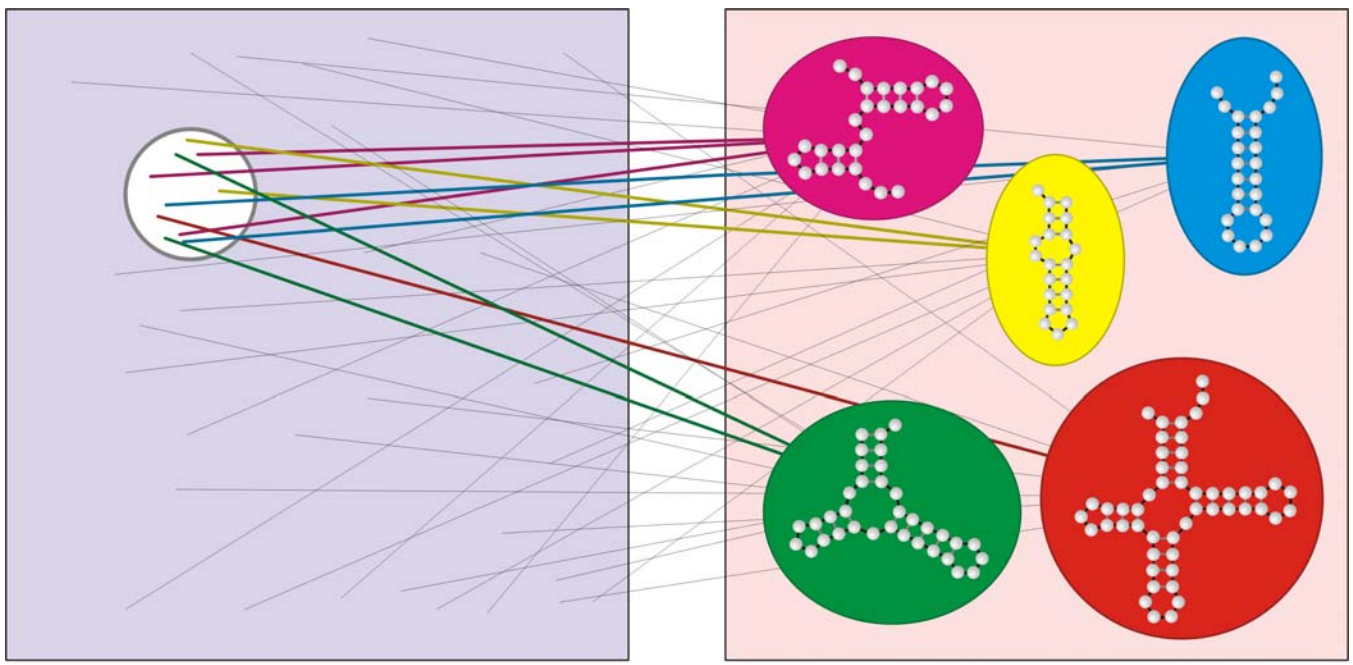
RNA secondary structures and Zipf's law

Properties of RNA sequence to secondary structure mapping

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

Properties of RNA sequence to secondary structure mapping

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



Properties of RNA sequence to secondary structure mapping

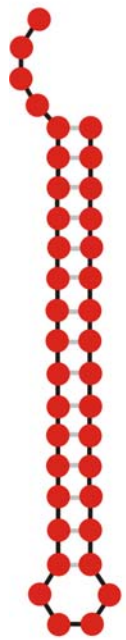
1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected

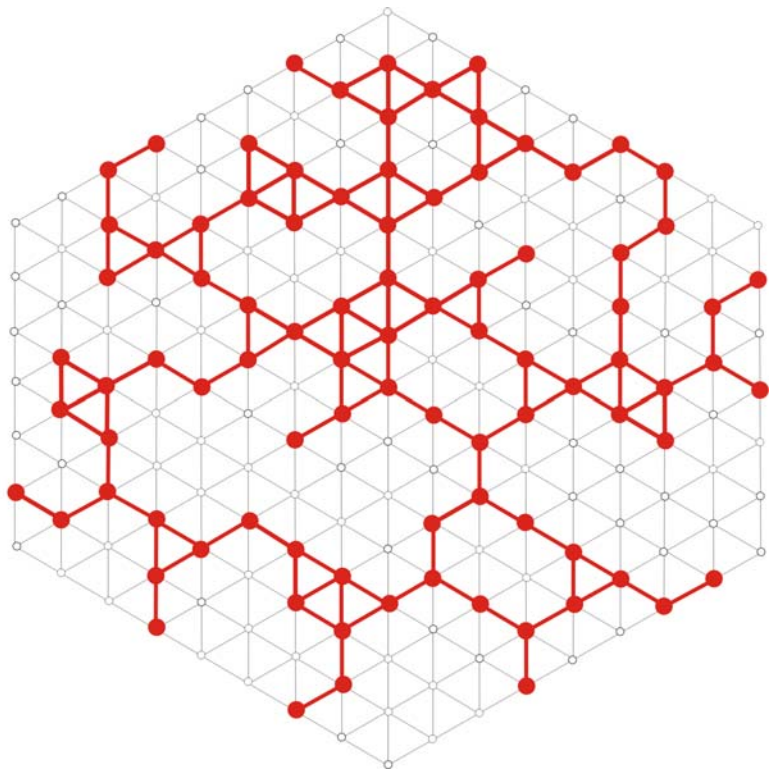
Properties of RNA sequence to secondary structure mapping

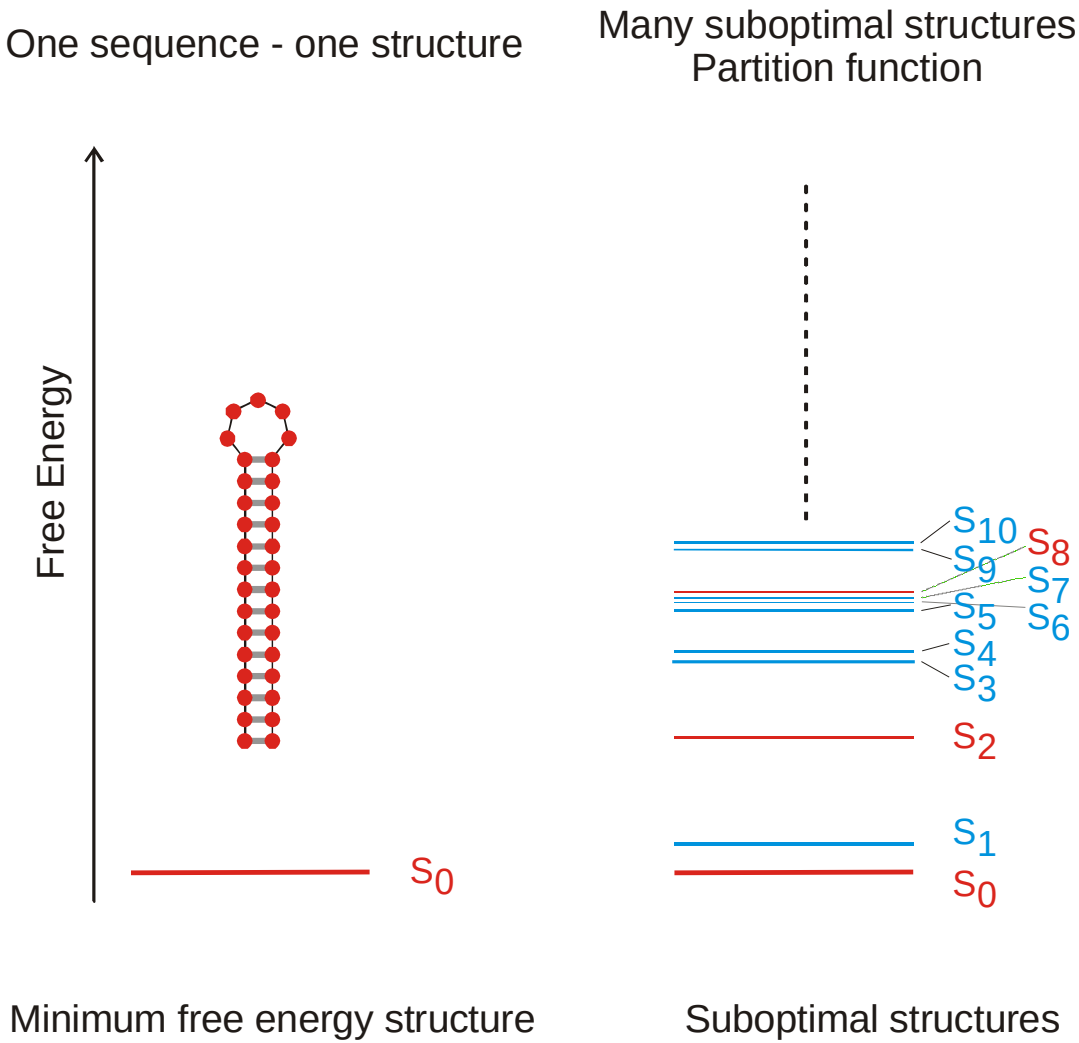
1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected

Alphabet size κ :

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





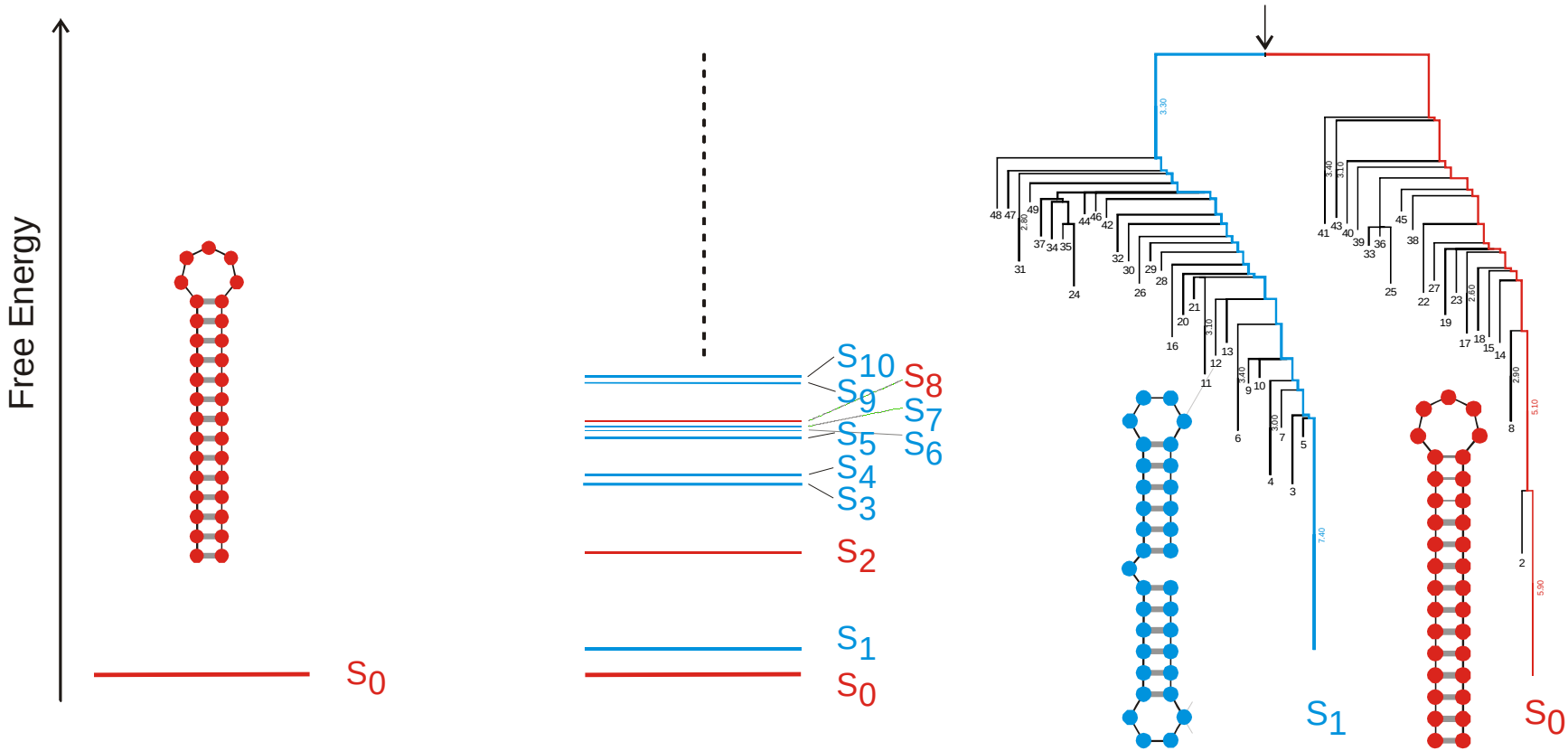


RNA secondary structures derived from a single sequence

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

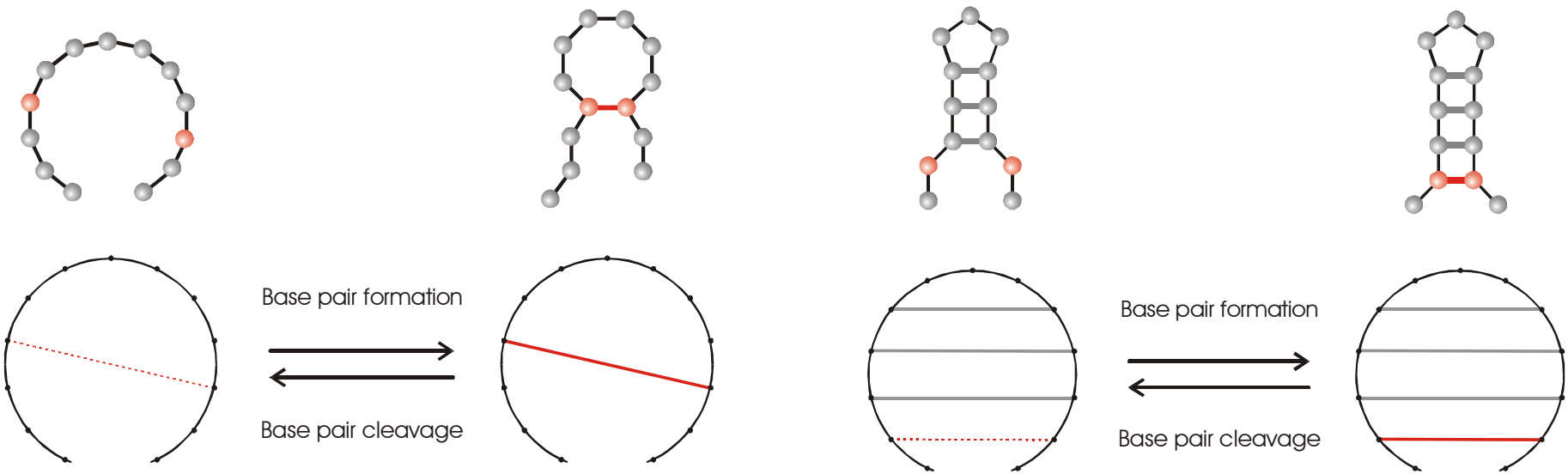
$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t) \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

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

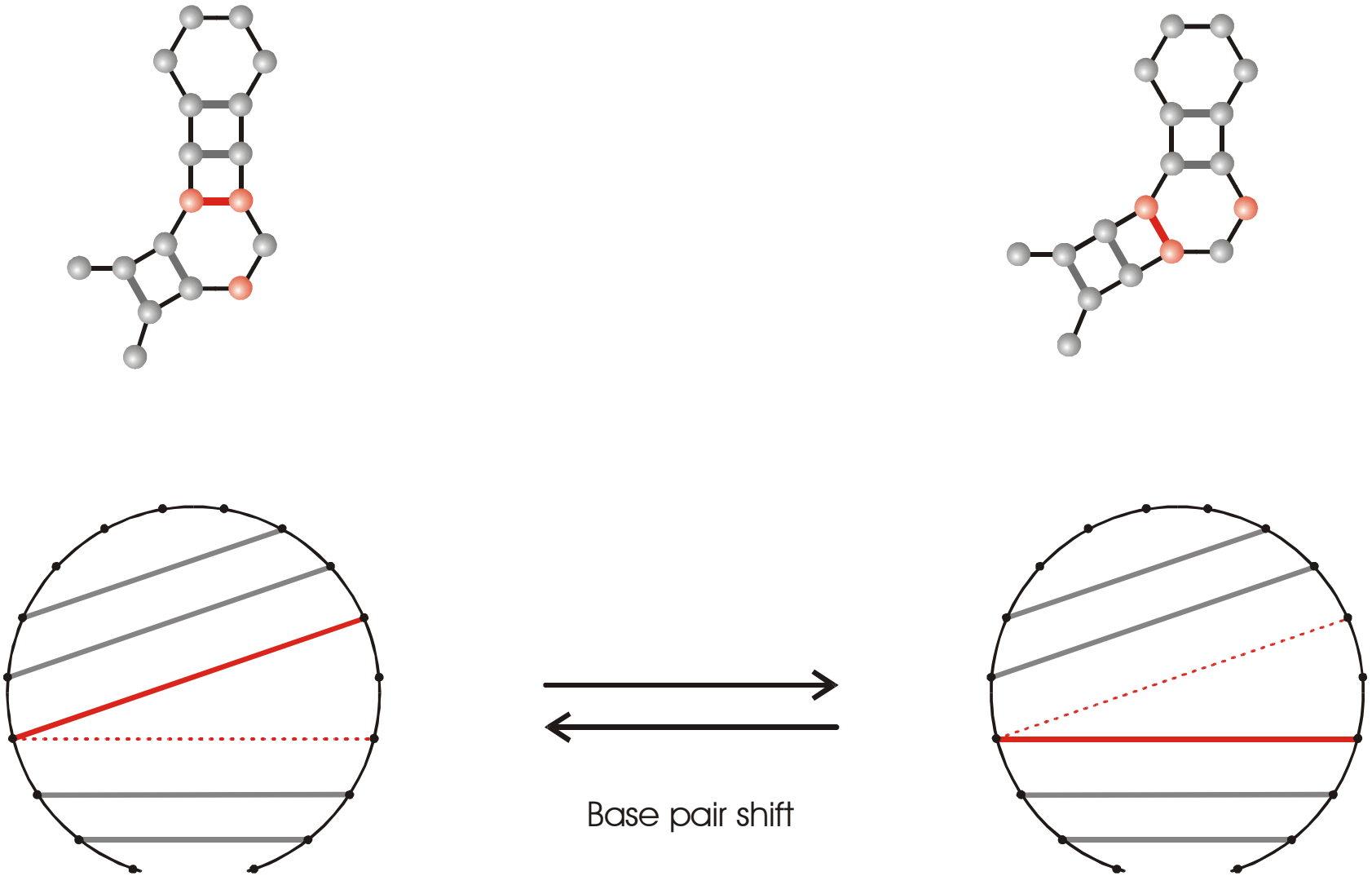
$$\Sigma_k = \sum_{k=1, k\neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

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).

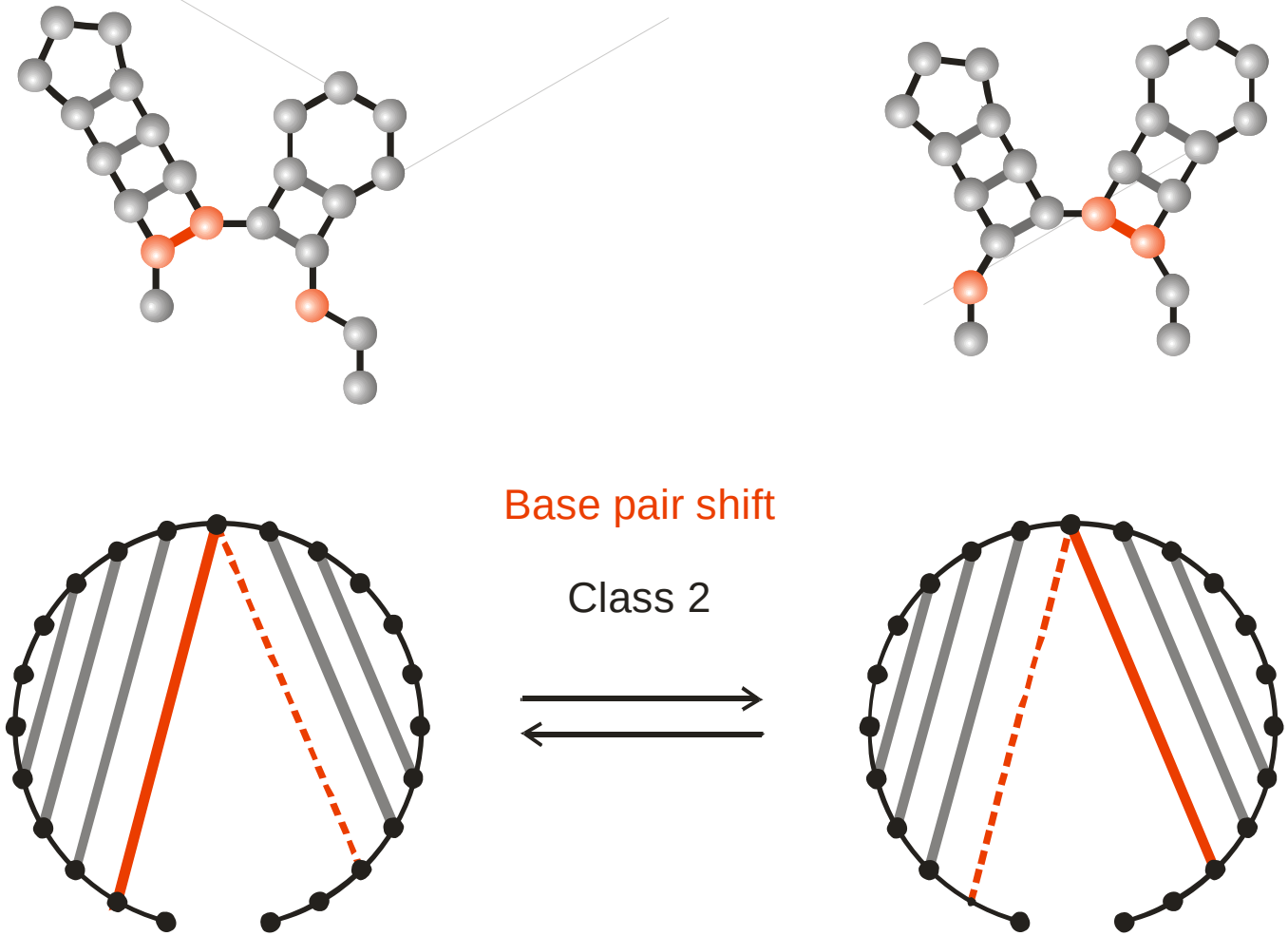
Formulation of kinetic RNA folding as a stochastic process



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

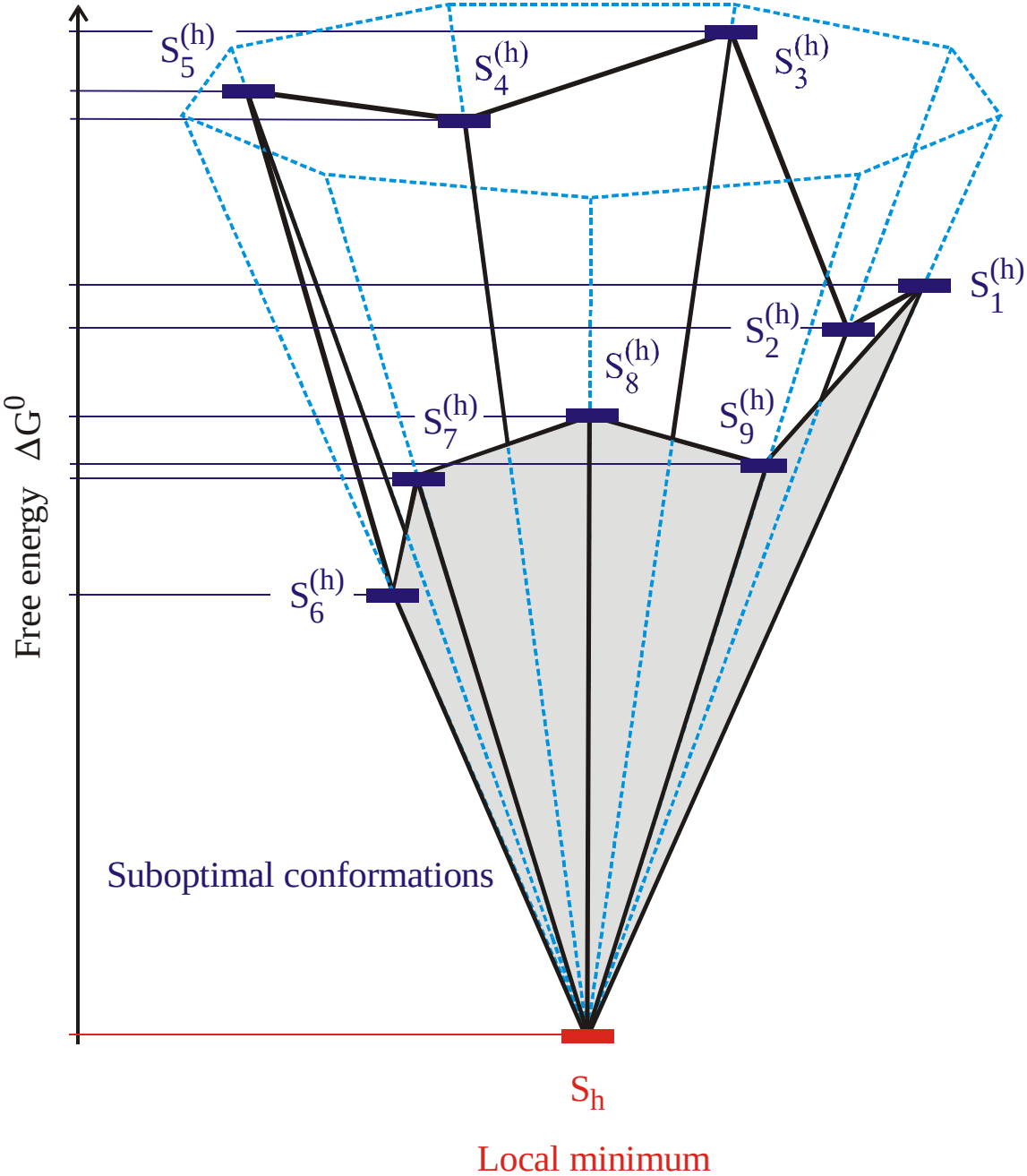


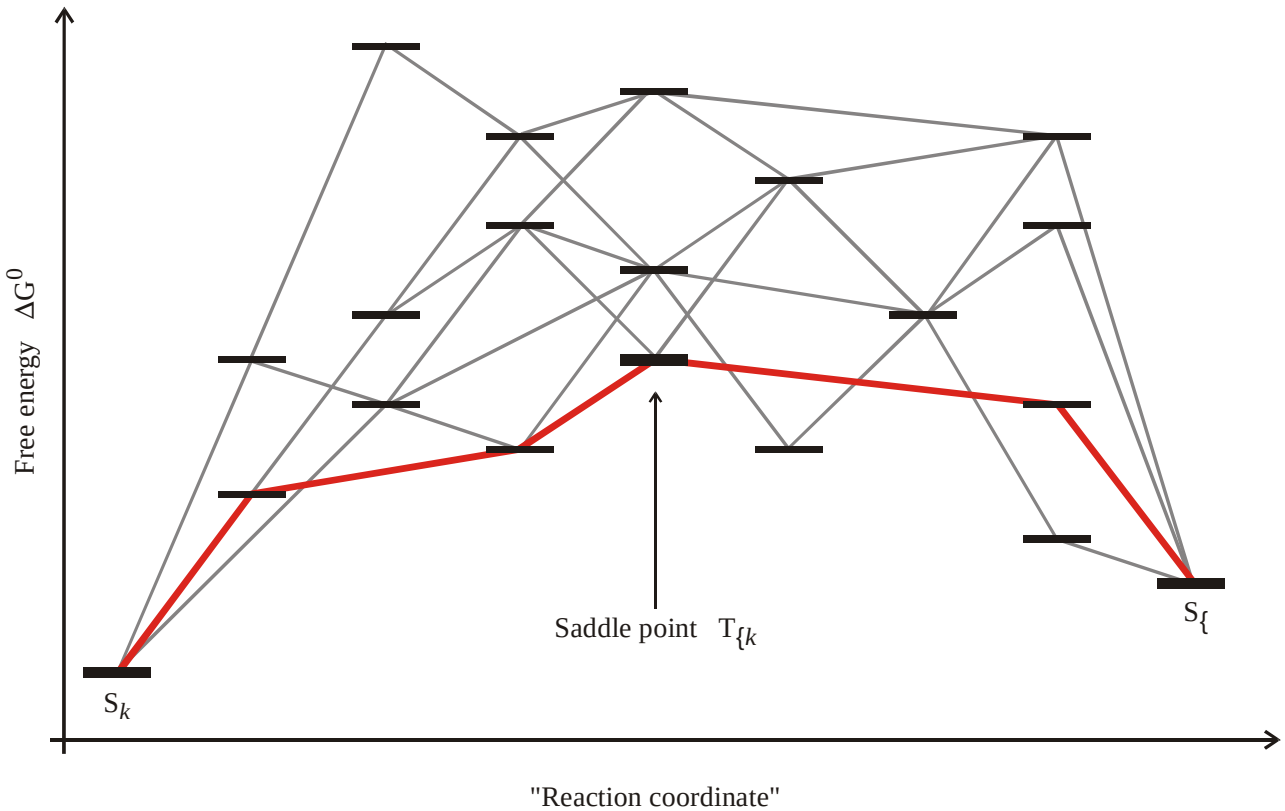
Base pair shift move of class 1: Shift inside internal loops or bulges

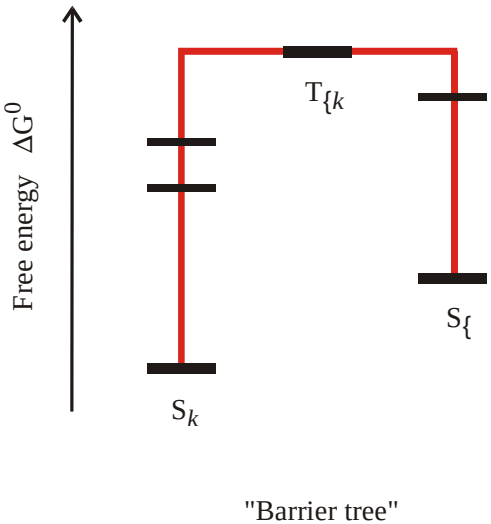


Base pair shift move of class 2: Shift involving free ends

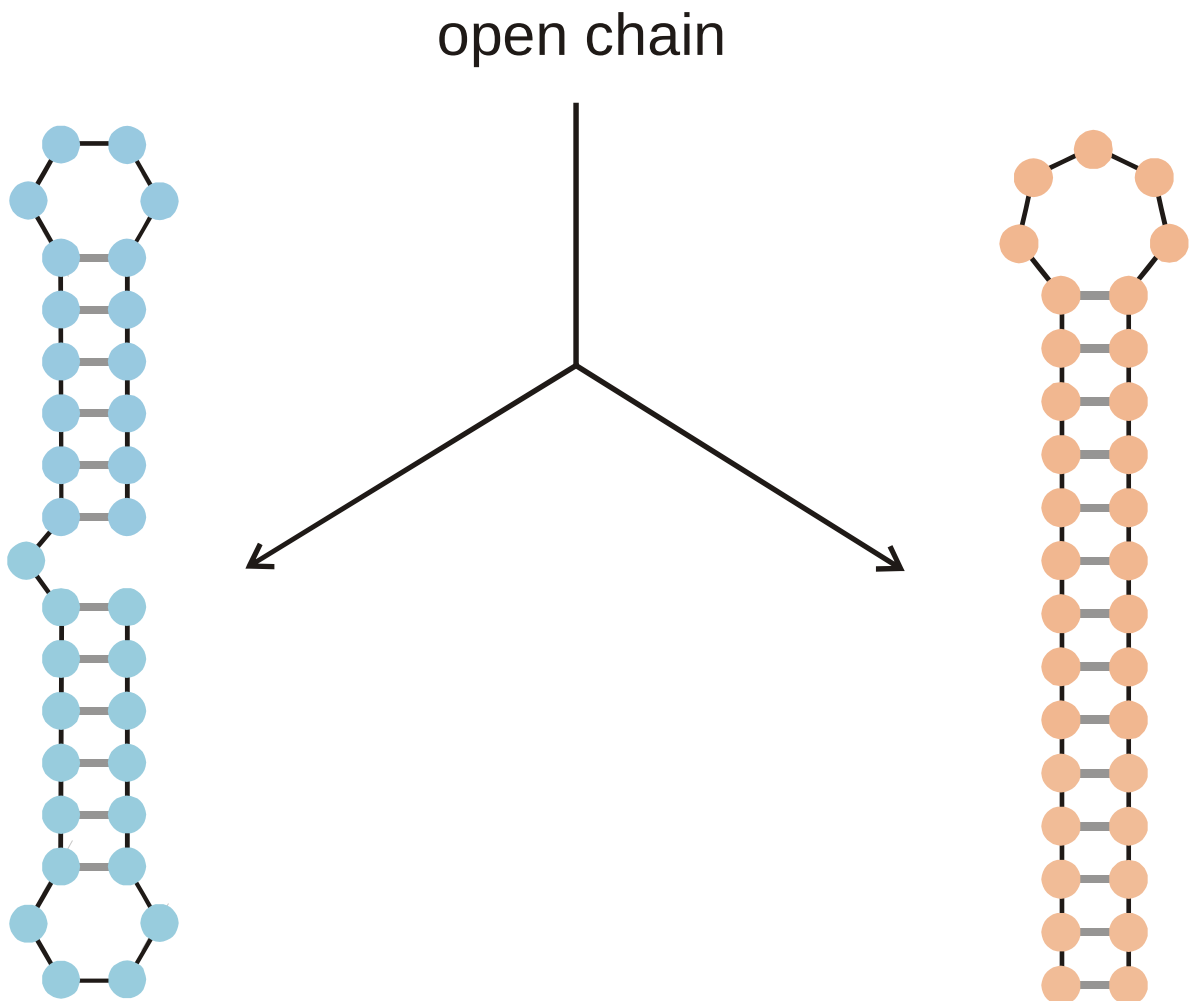
Search for local minima in conformation space



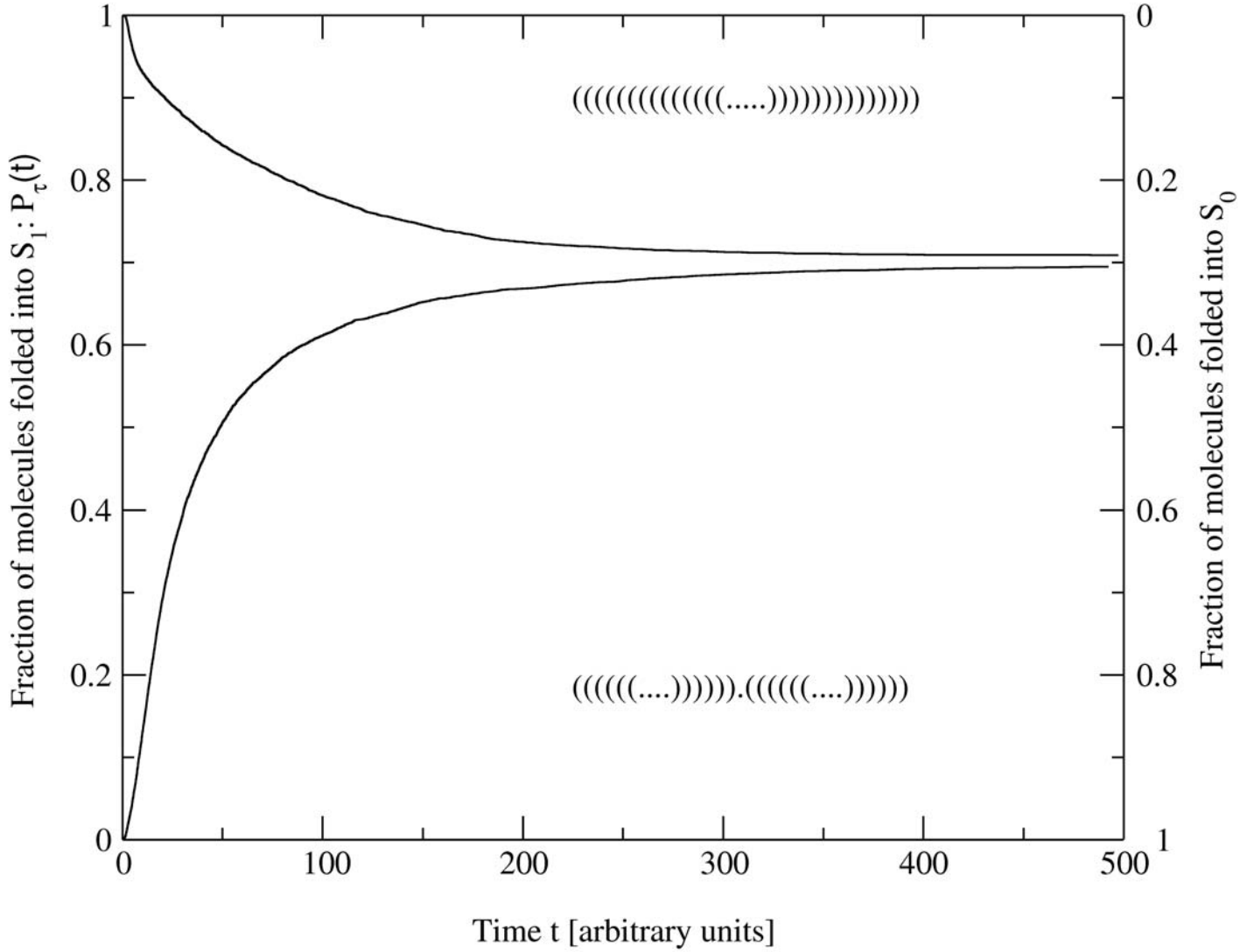




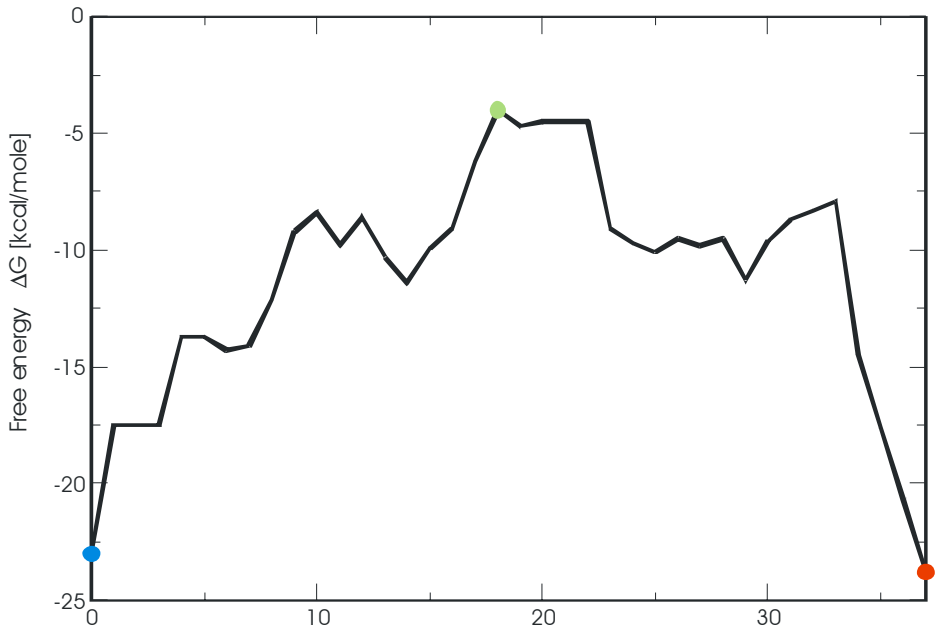
Definition of a ‚barrier tree‘



A nucleic acid molecule folding in two dominant conformations

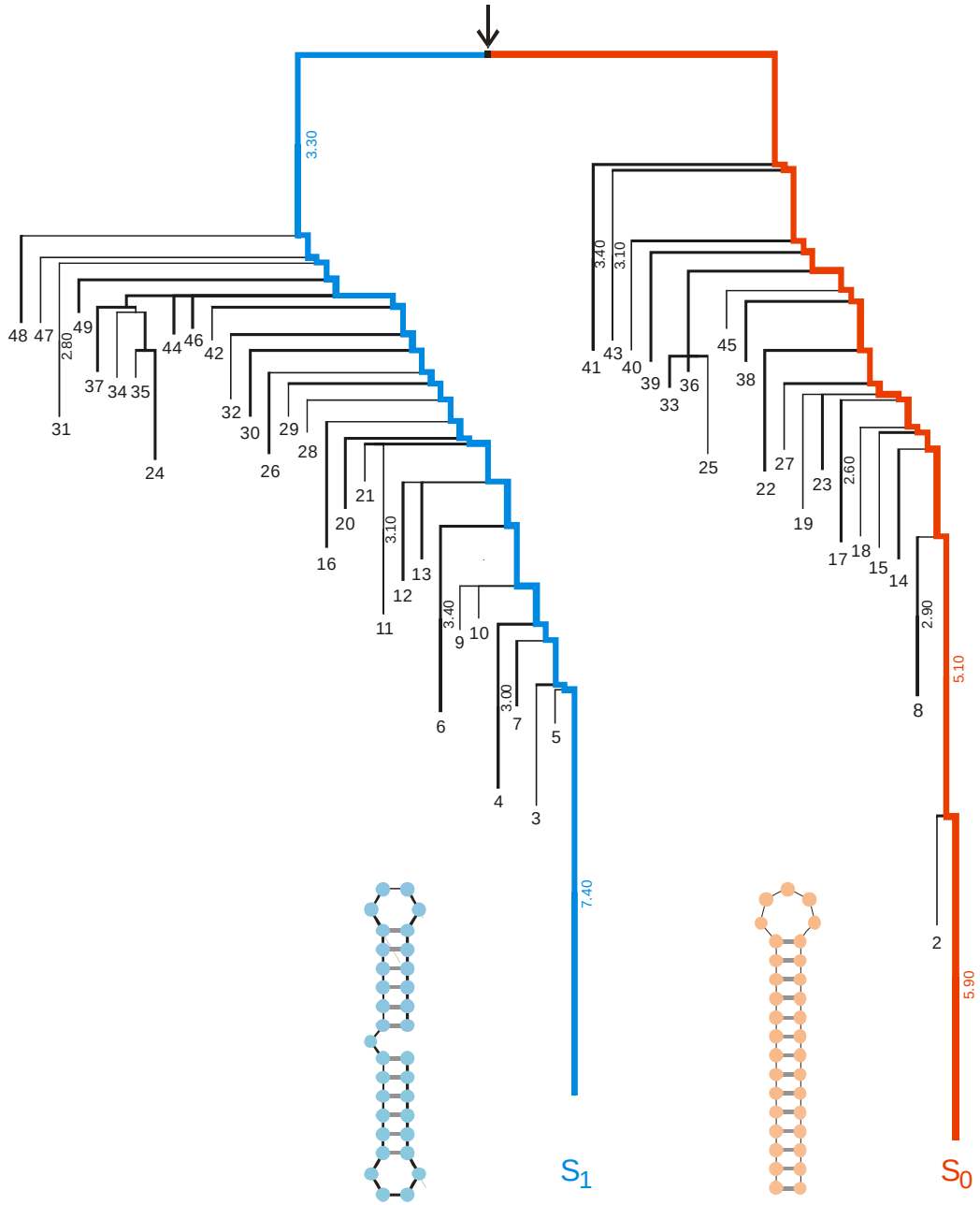


Folding dynamics of the sequence **GGCCCCUUUGGGGGCCAGACCCCUAAAAAGGGUC**



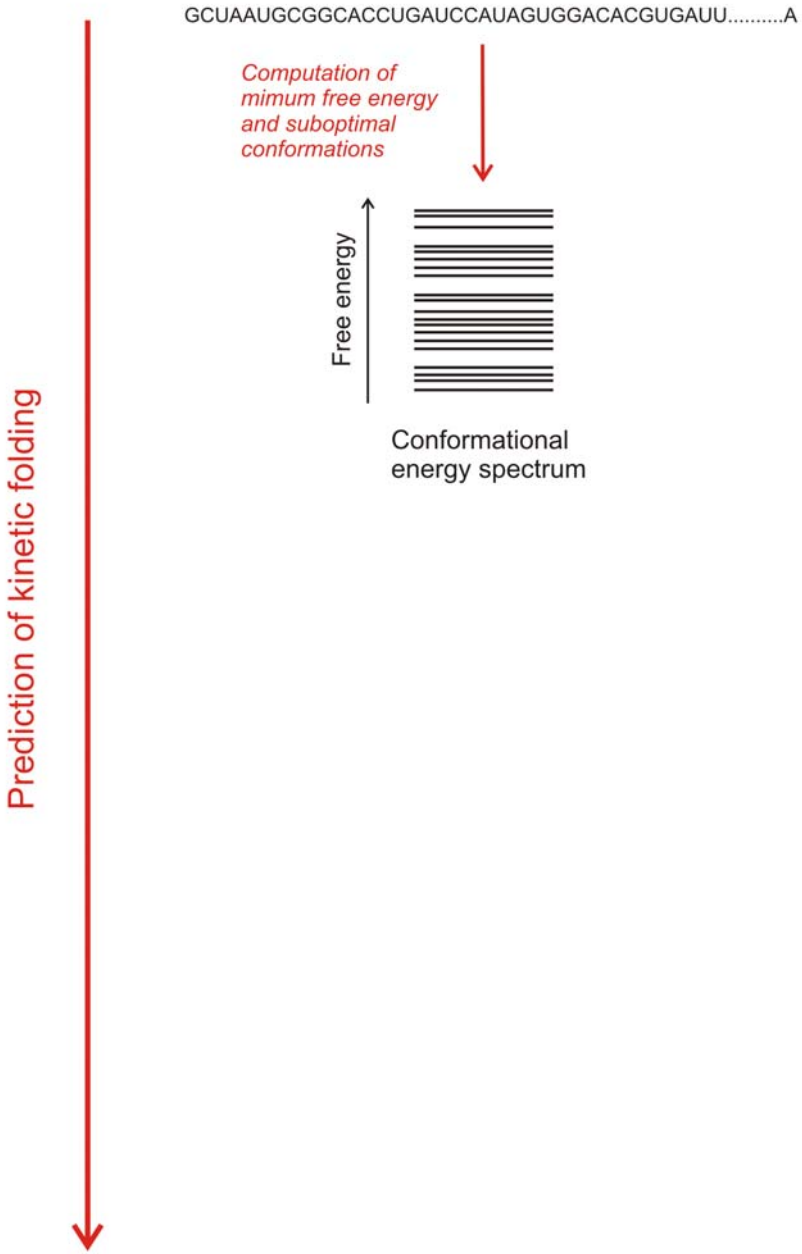
The folding path from S_1 to S_0

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

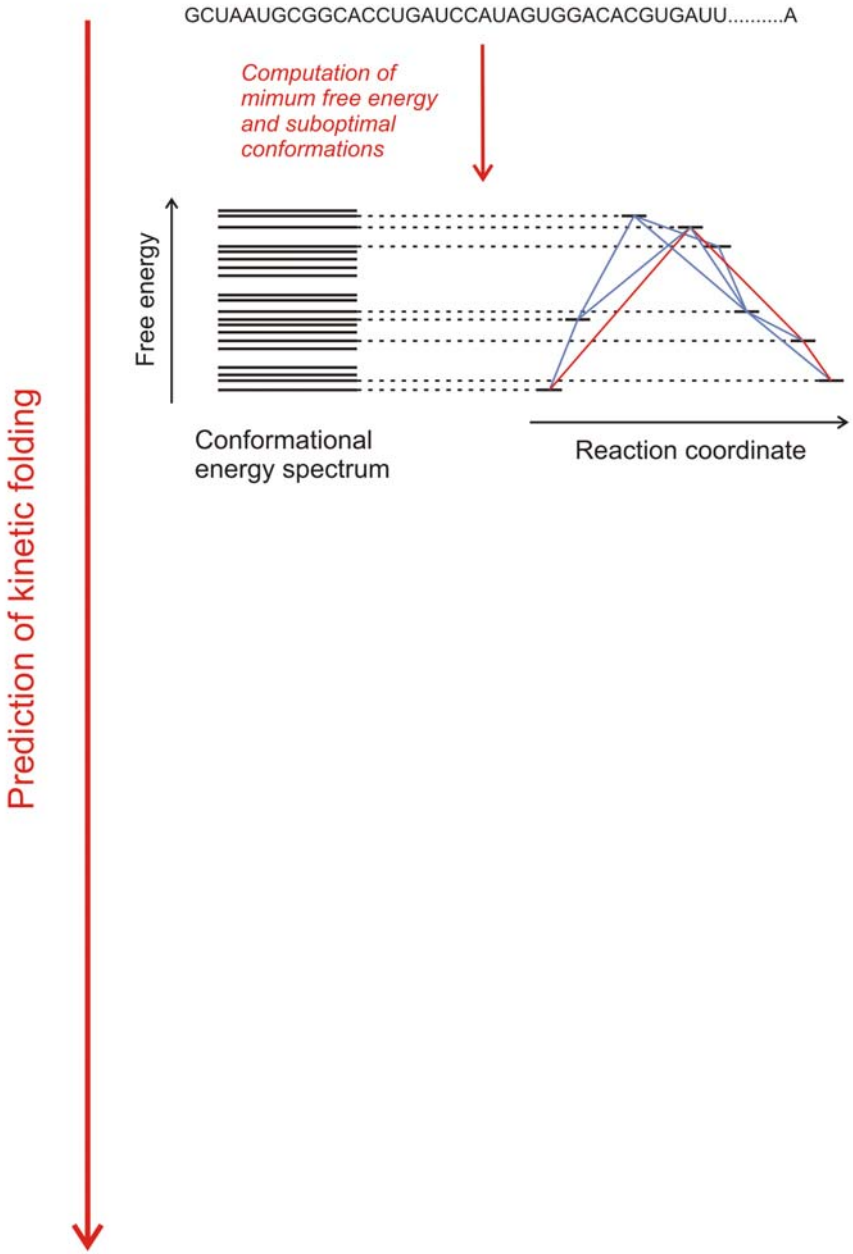


The barrier tree connecting S_1 and S_0

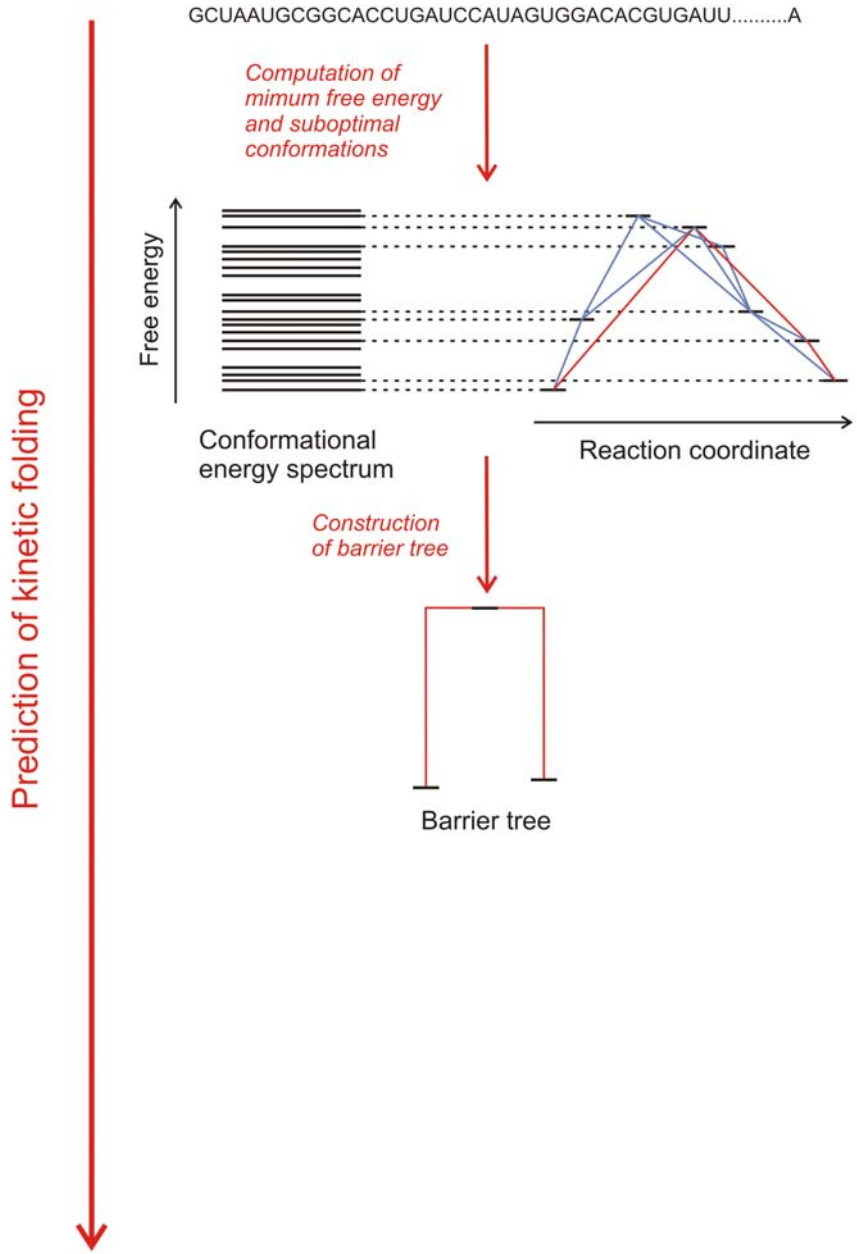
Prediction of RNA kinetic folding of secondary structures based on Arrhenius kinetics



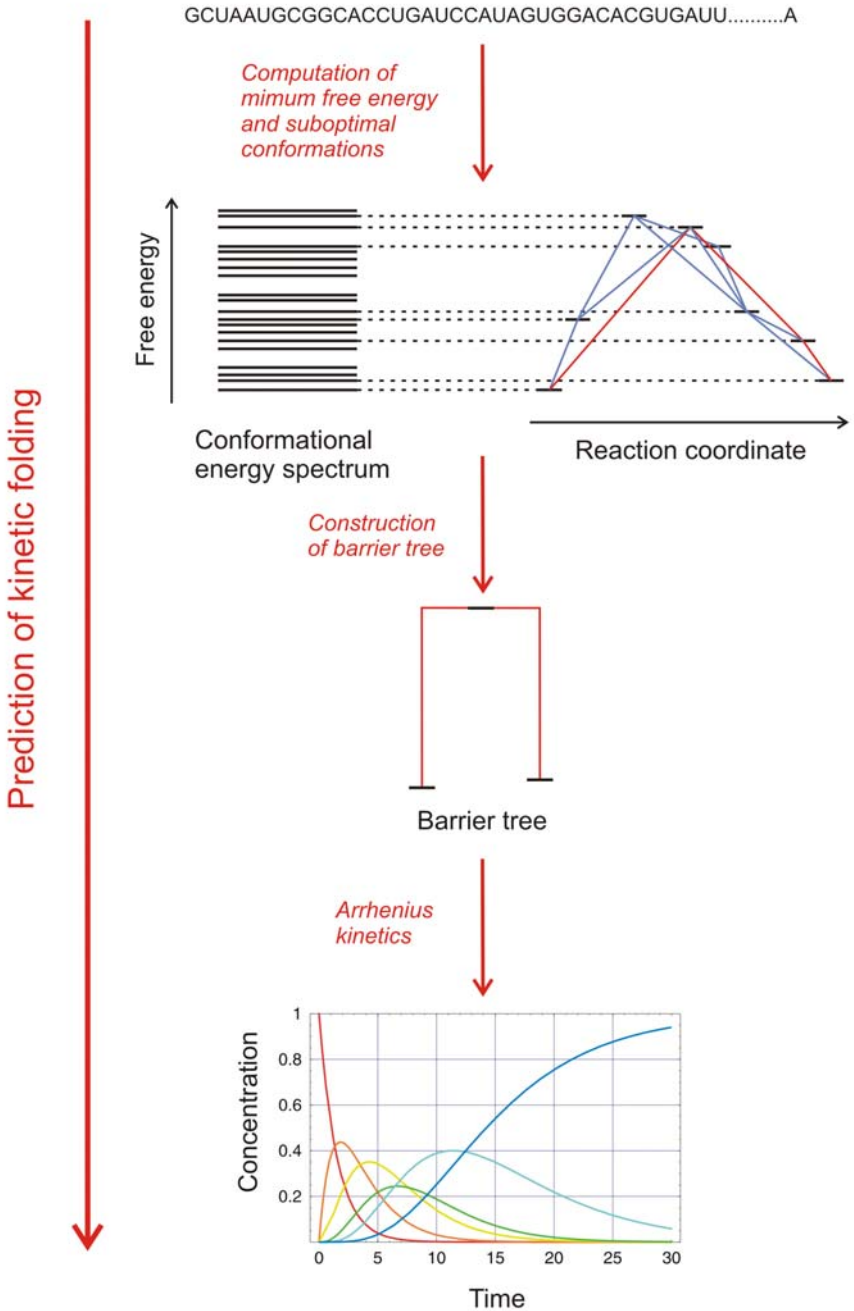
Prediction of RNA kinetic folding of secondary structures based on Arrhenius kinetics



Prediction of RNA kinetic folding of secondary structures based on Arrhenius kinetics

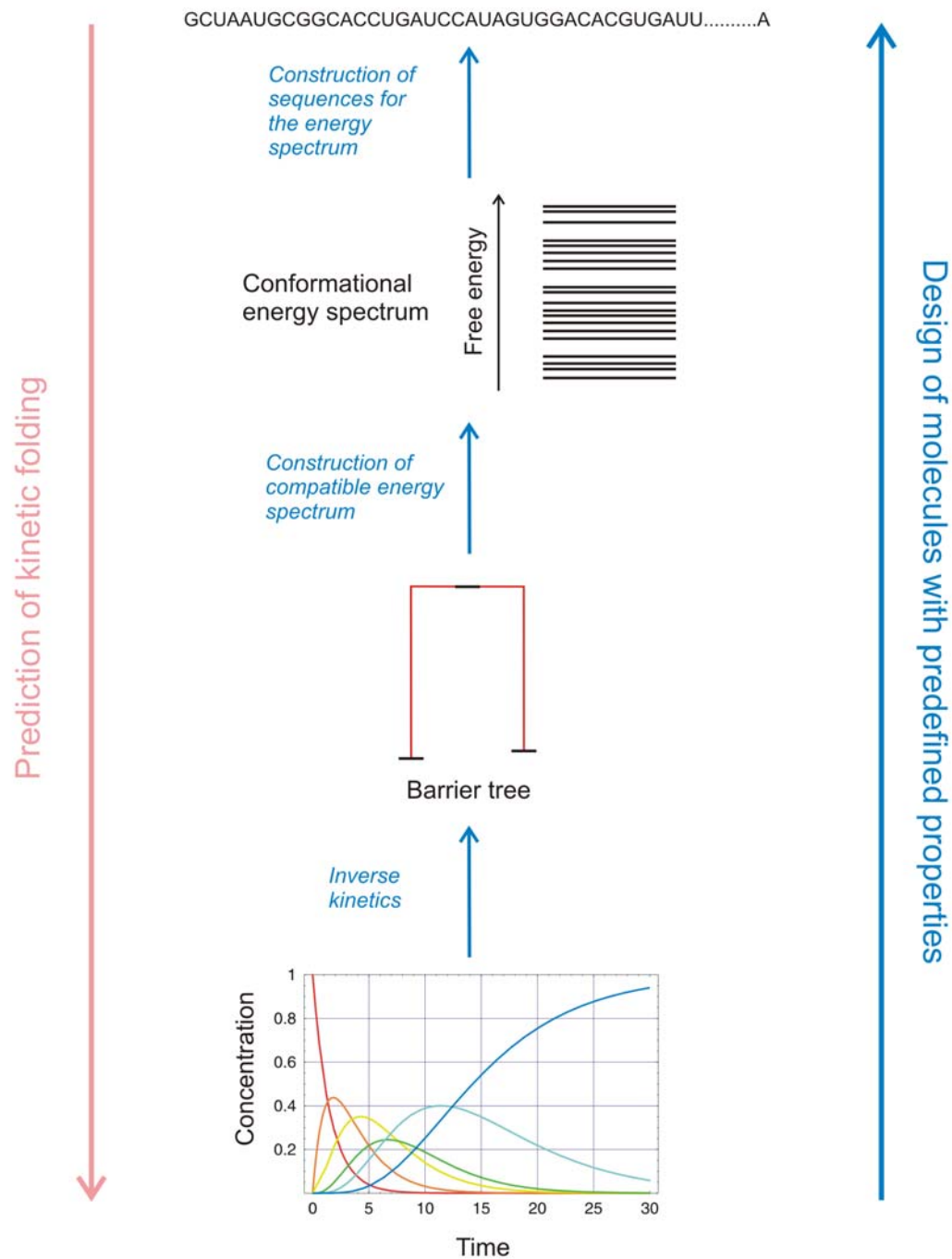


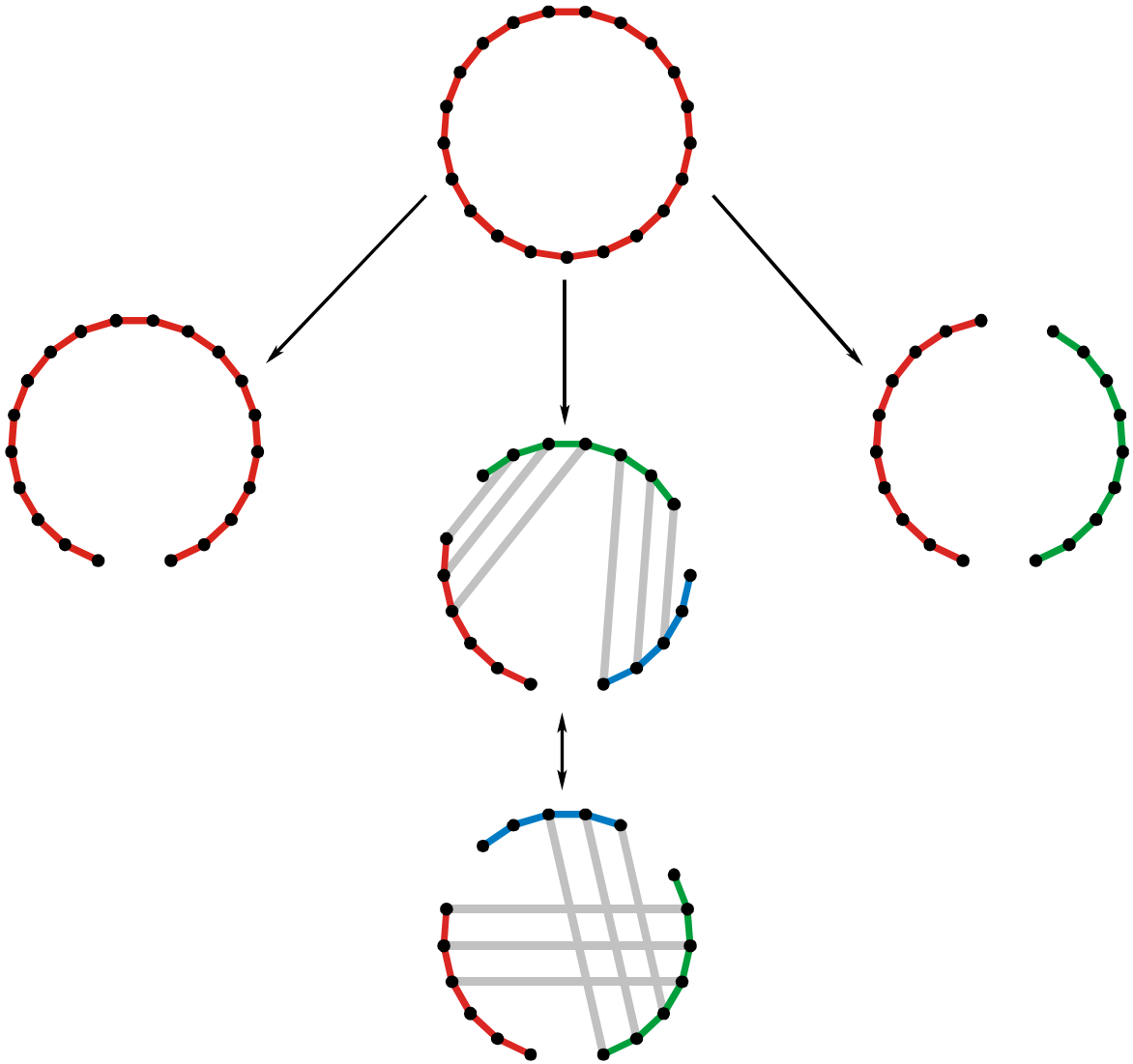
Prediction of RNA kinetic folding of secondary structures based on Arrhenius kinetics



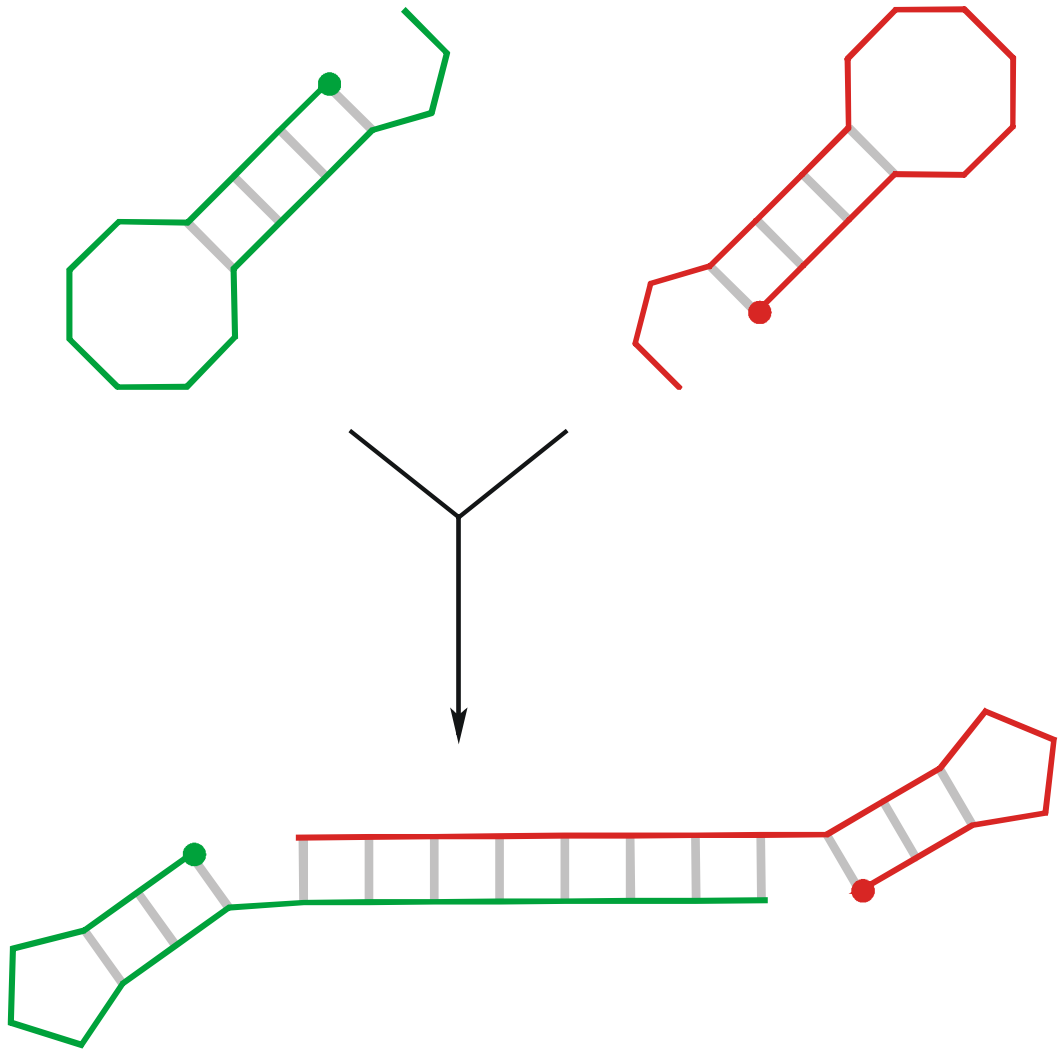
Prediction of RNA kinetic folding of secondary structures based on Arrhenius kinetics

Design of RNA molecules with
with predefined folding kinetics

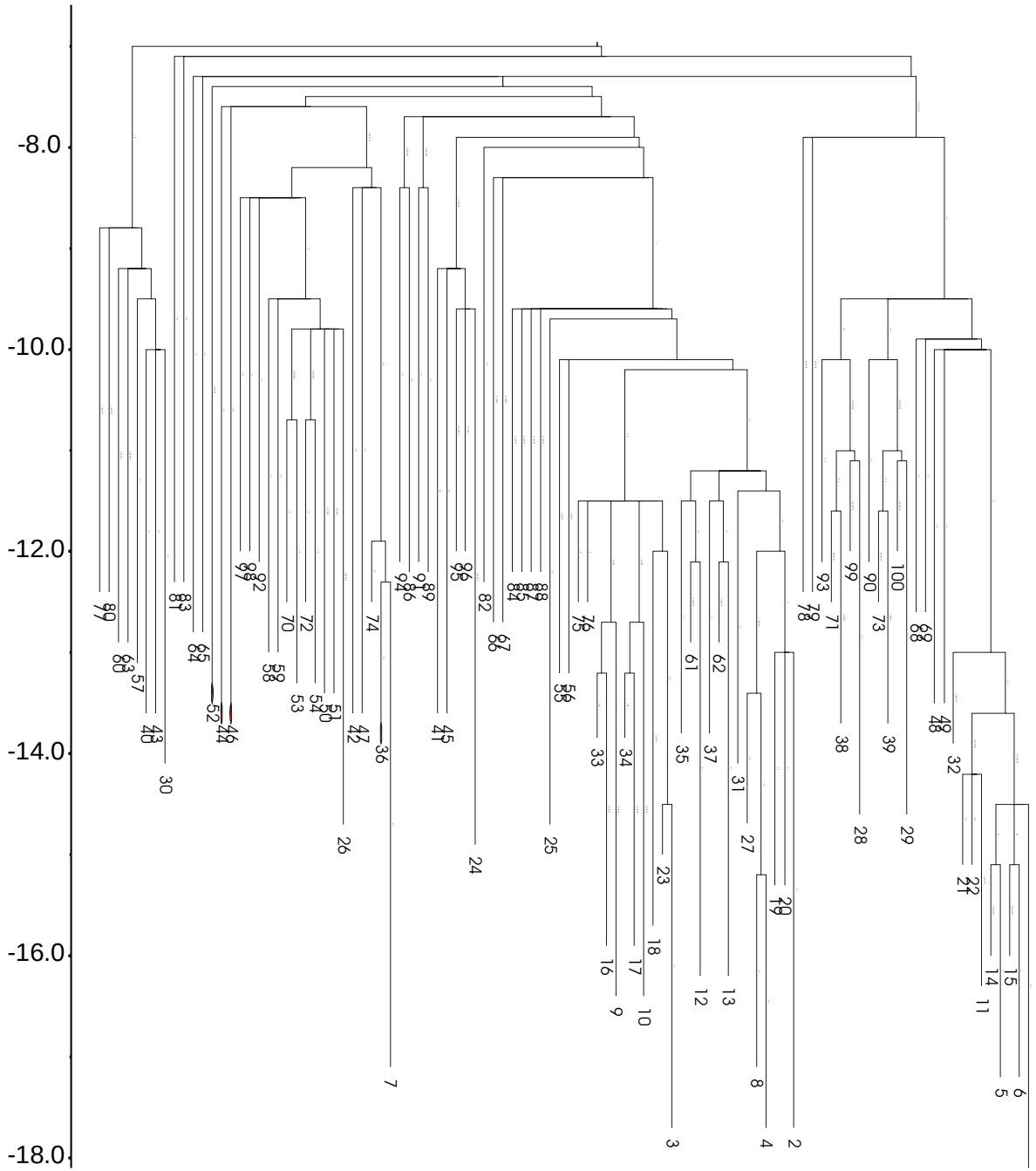




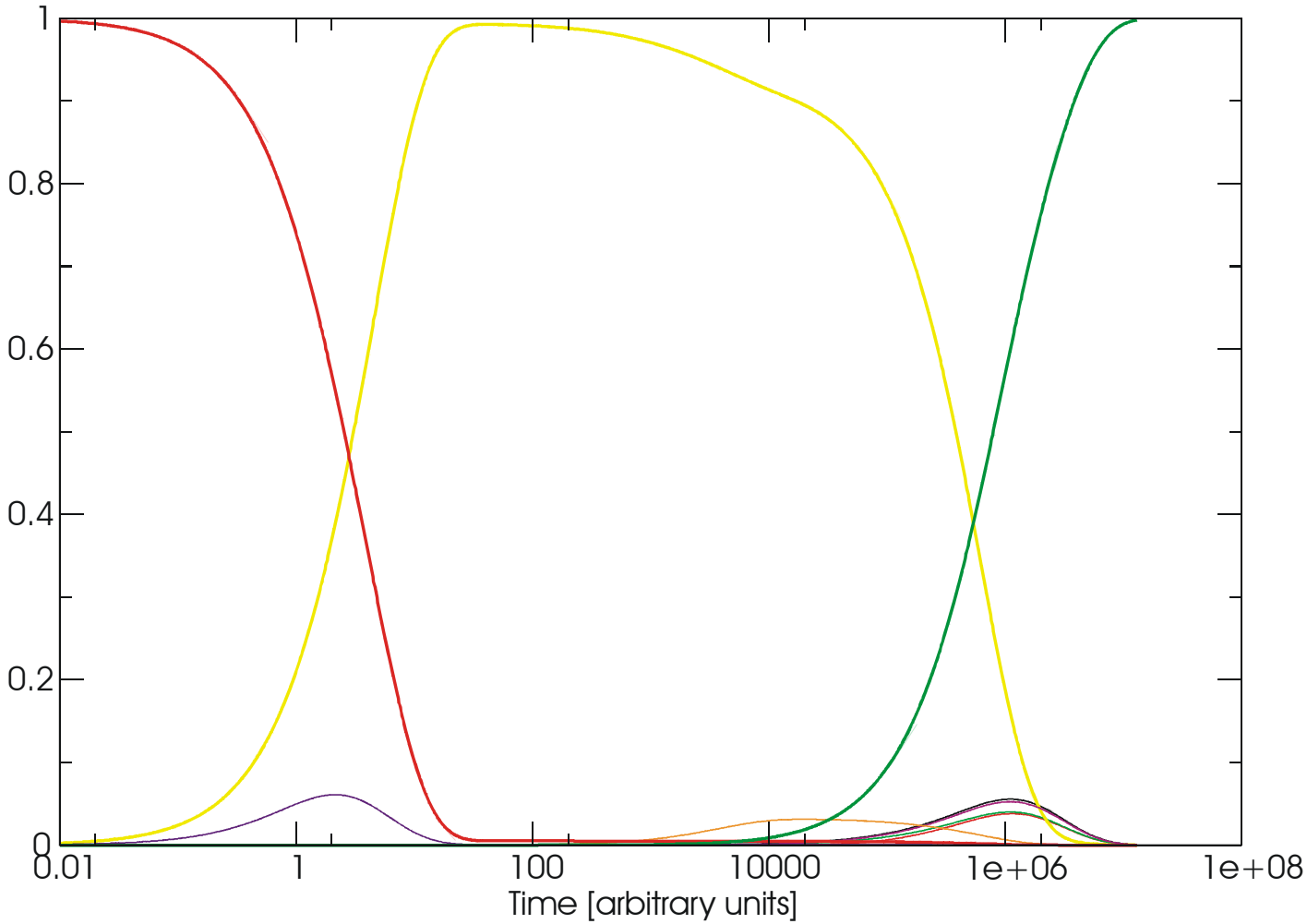
Cofolding two or three nucleic acid molecules



An example for ‘symmetric’ cofolding of two molecules



Cofolding tree



Cofolding kinetics