RNA Bioinformatics Beyond the One Sequence-One Structure Paradigm

Peter Schuster

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

2008 Molecular Informatics and Bioinformatics

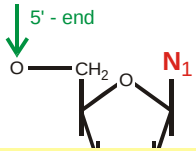
Collegium Budapest, 27.– 29.03.2008

Web-Page for further information:

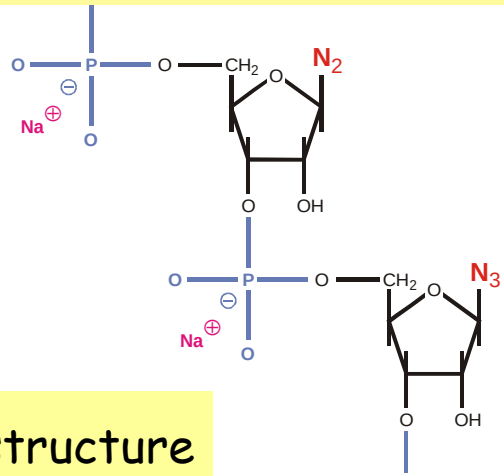
http://www.tbi.univie.ac.at/~pks

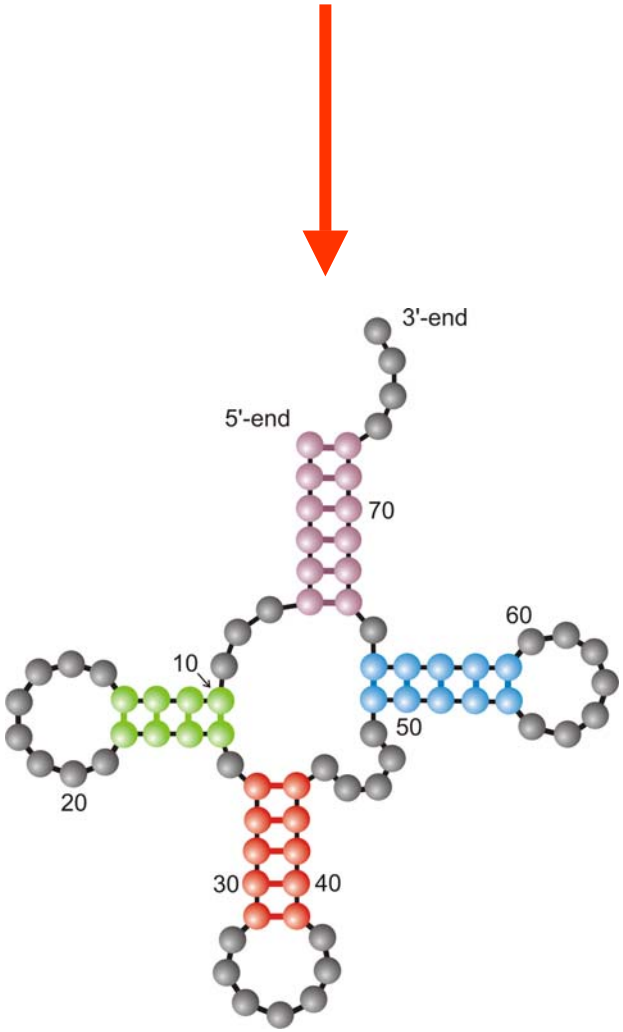
1. Computation of RNA equilibrium structures
2. Inverse folding and neutral networks
3. Evolutionary optimization of structure
4. Suboptimal conformations and kinetic folding

1. **Computation of RNA equilibrium structures**
2. Inverse folding and neutral networks
3. Evolutionary optimization of structure
4. Suboptimal conformations and kinetic folding

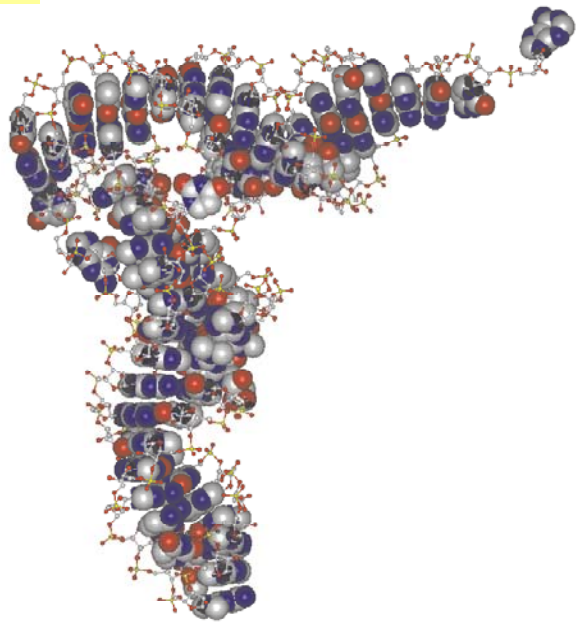


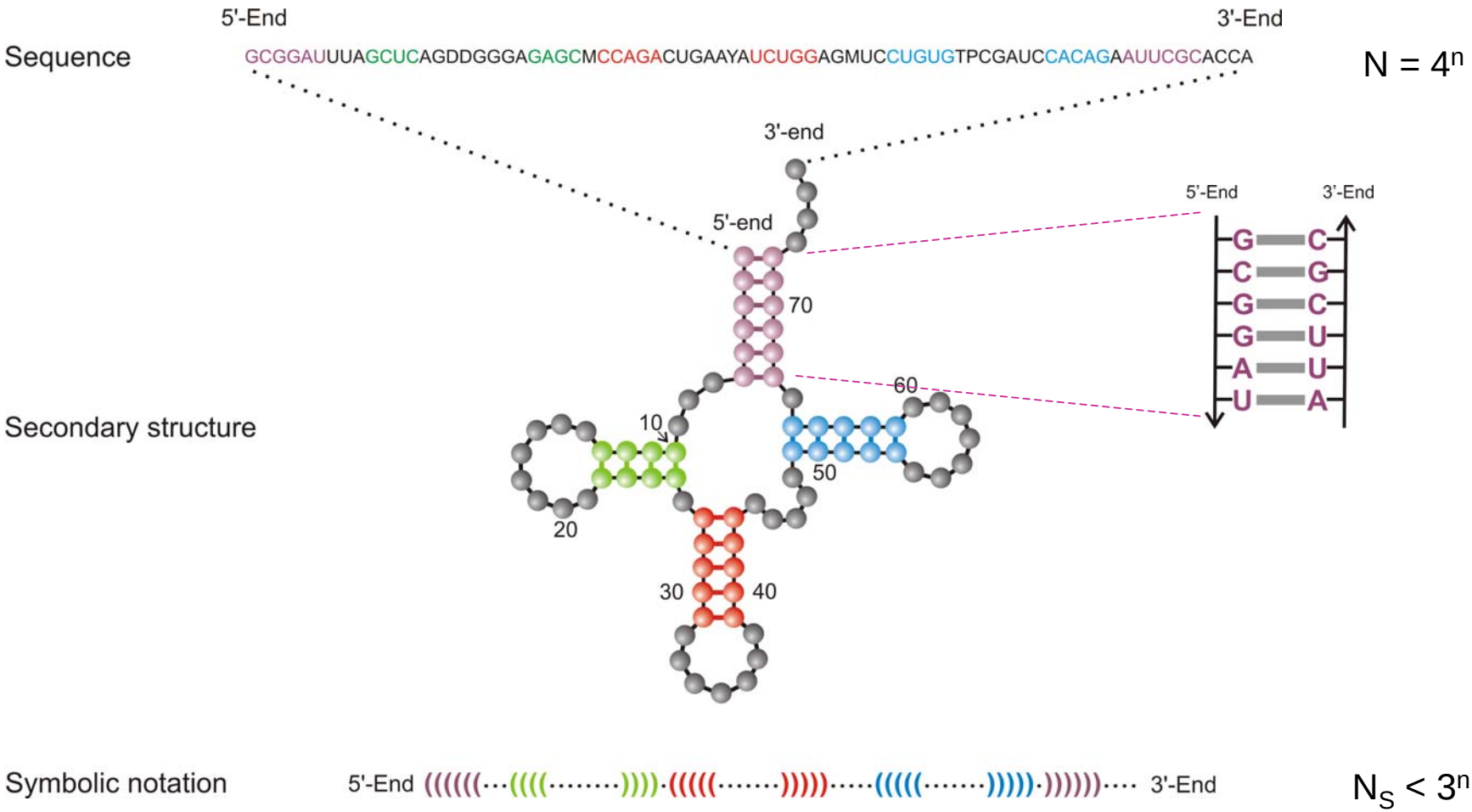
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure

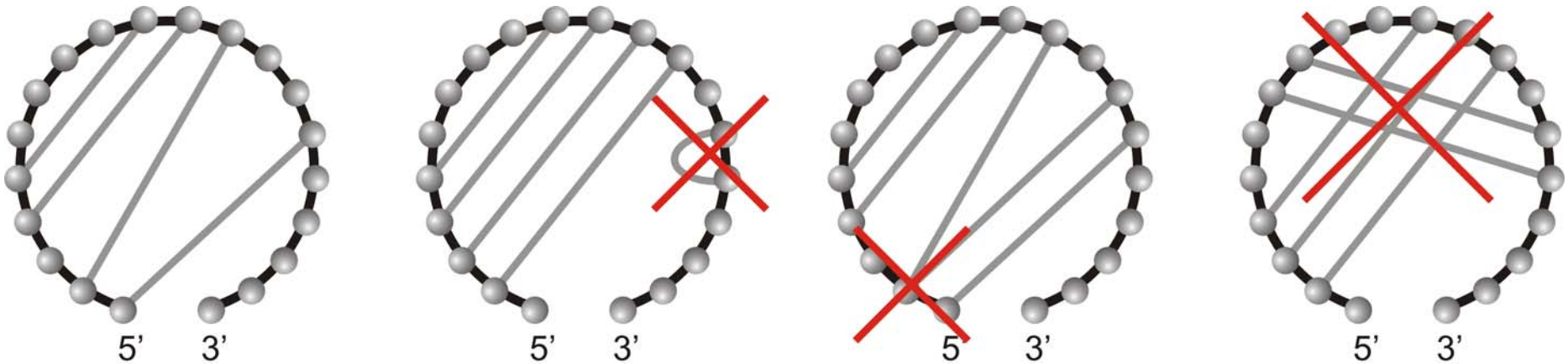




Criterion: Minimum free energy (mfe)

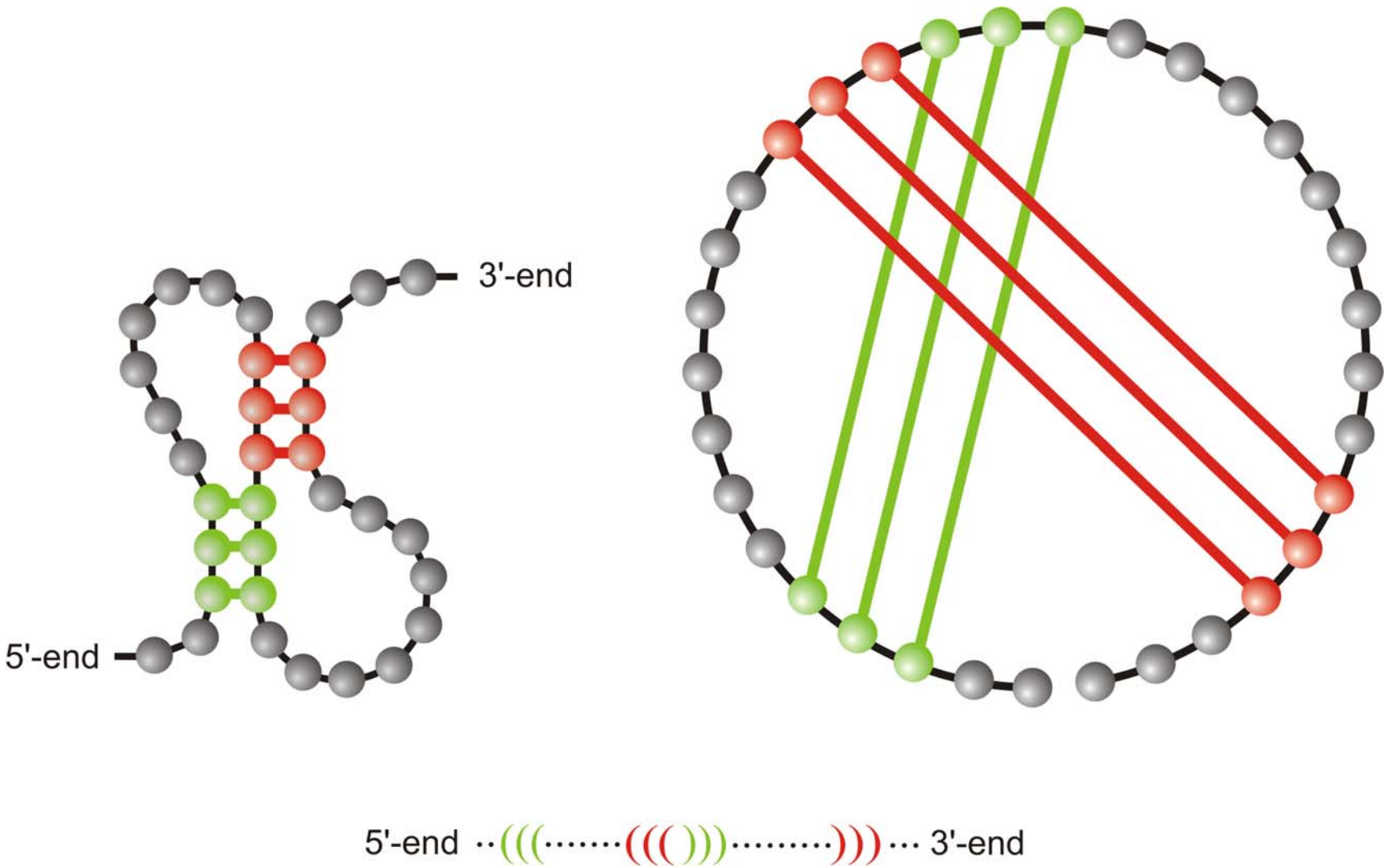
Rules: _ (_) _ ∈ {**AU**,**CG**,**GC**,**GU**,**UA**,**UG**}

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

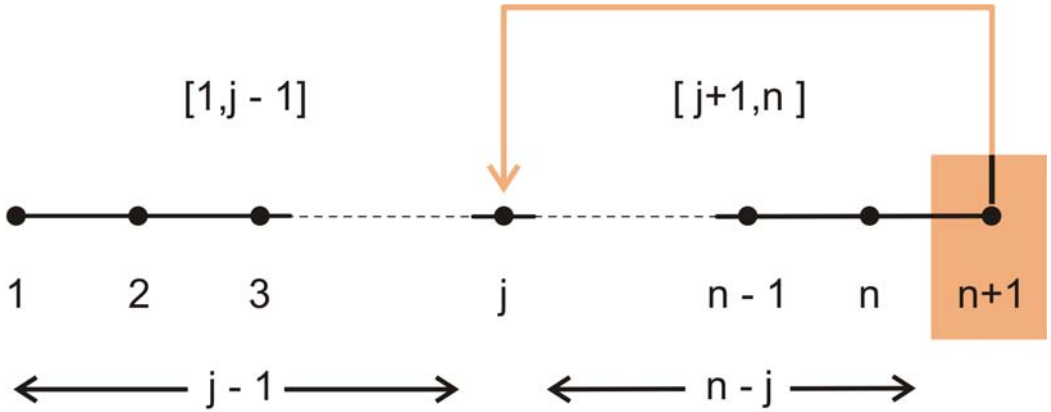


Base pairs $\in$ {**AU**,**CG**,**GC**,**GU**,**UA**,**UG**}

Conventional definition of RNA secondary structures



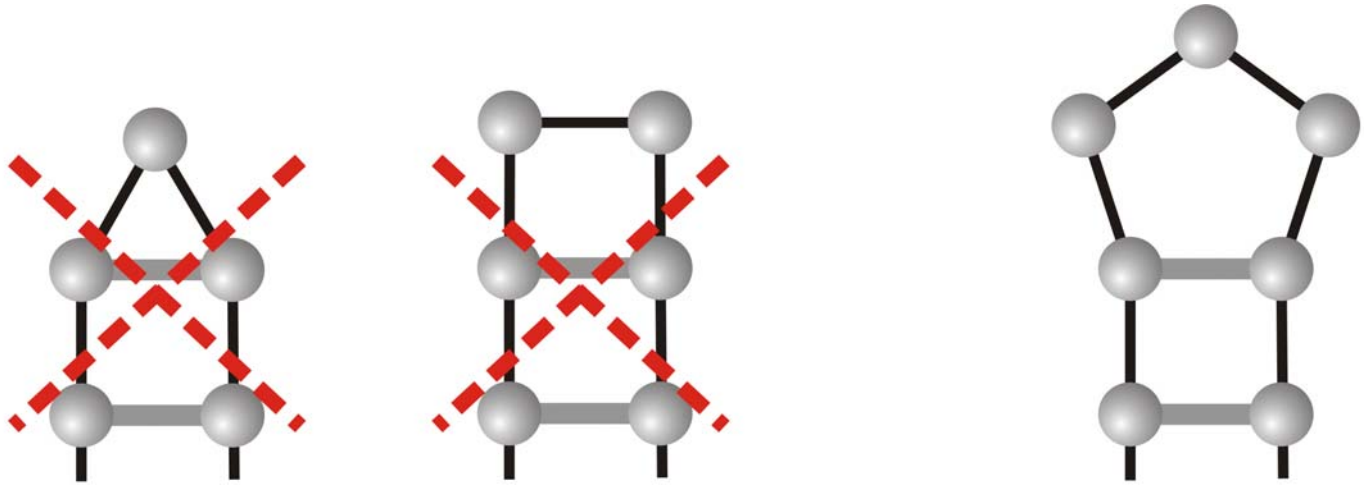
H-type pseudoknot



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

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

M.S. Waterman, T.F. Smith (1978) *Math.Bioscience* **42**:257-266



Impossible (extremely high free energies)
for steric reasons

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

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

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

(ii) stack $n_{\text{stack}} \geq \sigma$

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

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

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

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

Recursion formula for the number of physically acceptable stable structures

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

RNA sequence: GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

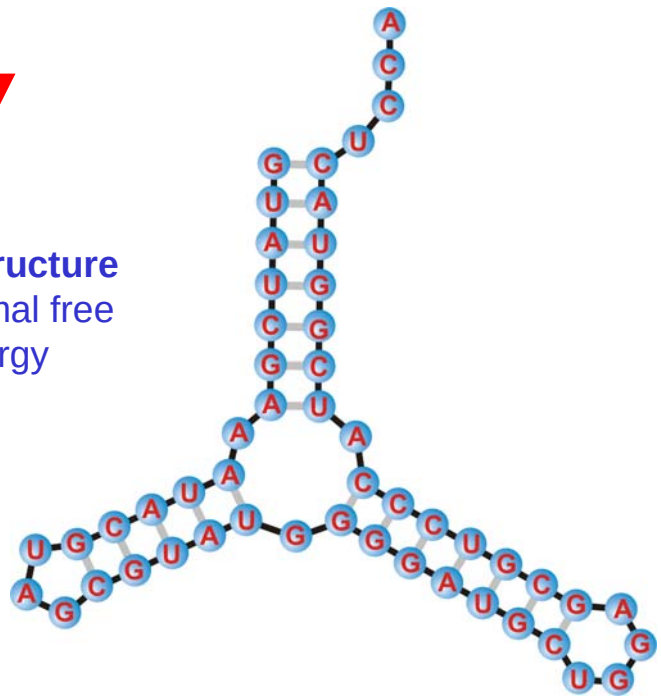
RNA folding:

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

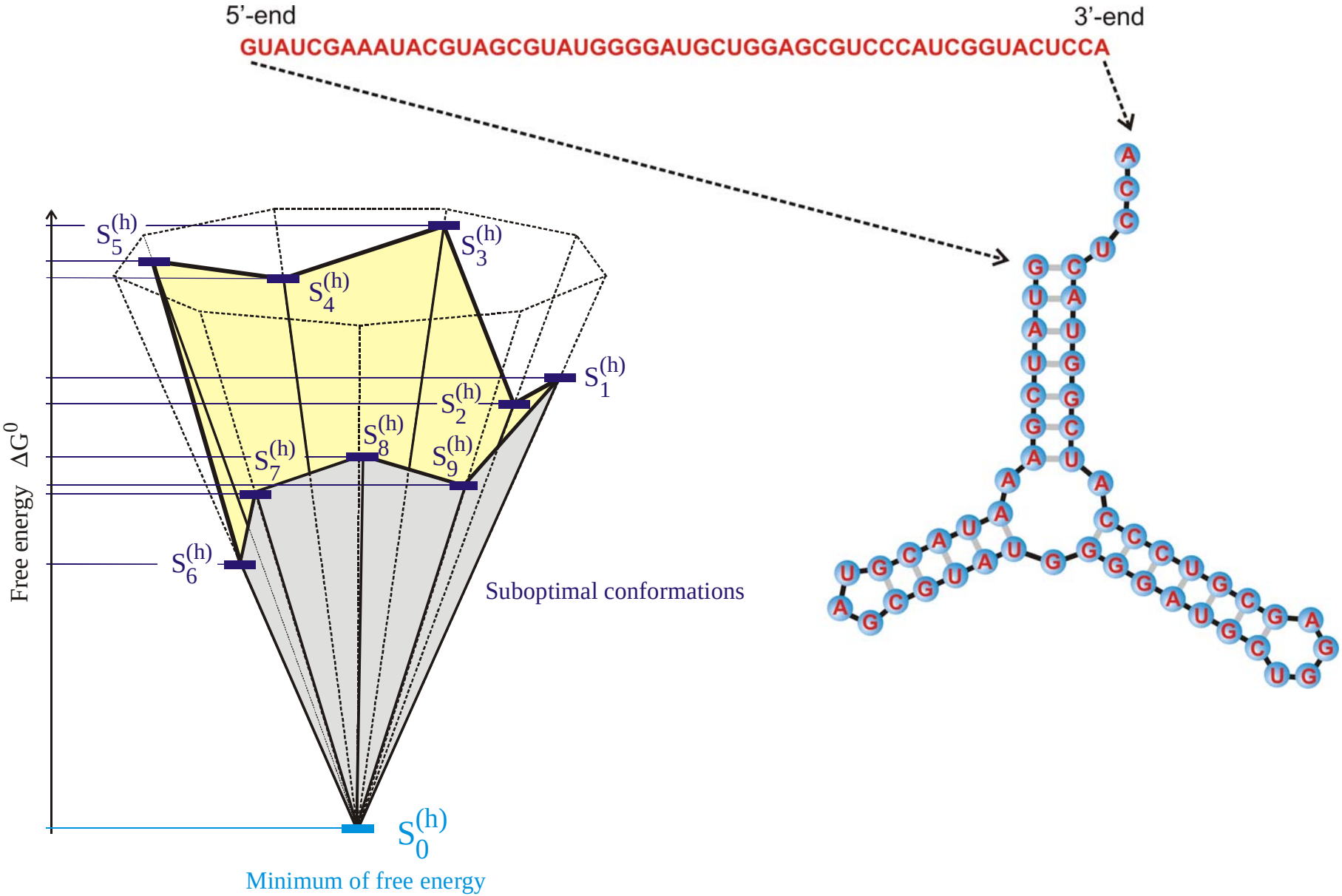
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

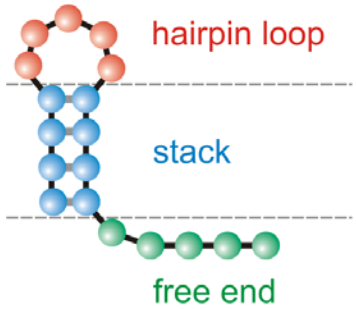
RNA structure of minimal free energy

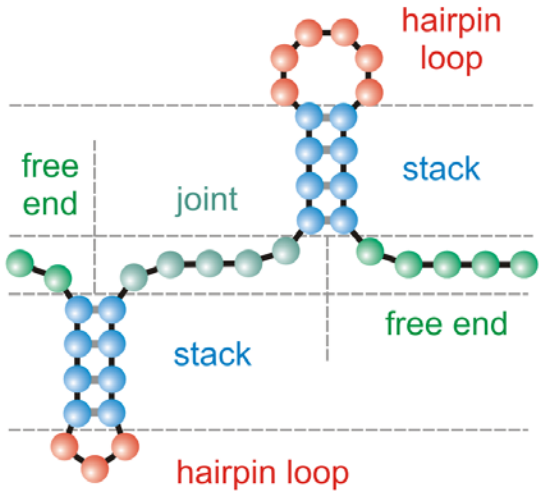


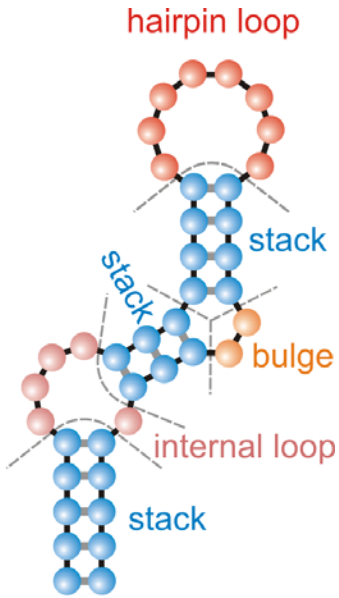
Sequence, structure, and design



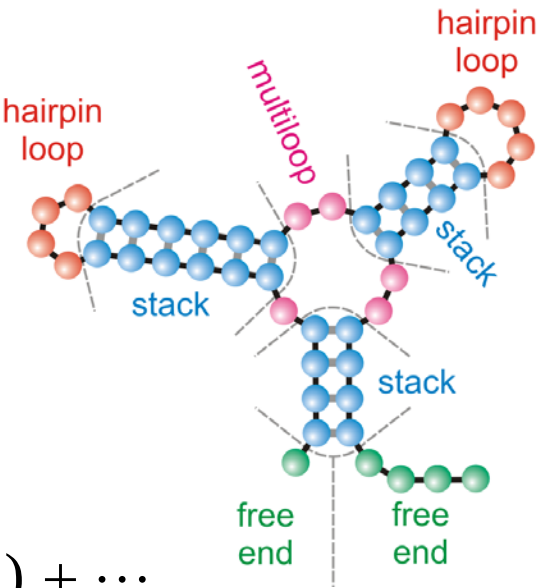
The minimum free energy structures on a discrete space of conformations







Elements of RNA secondary structures as used in free energy calculations



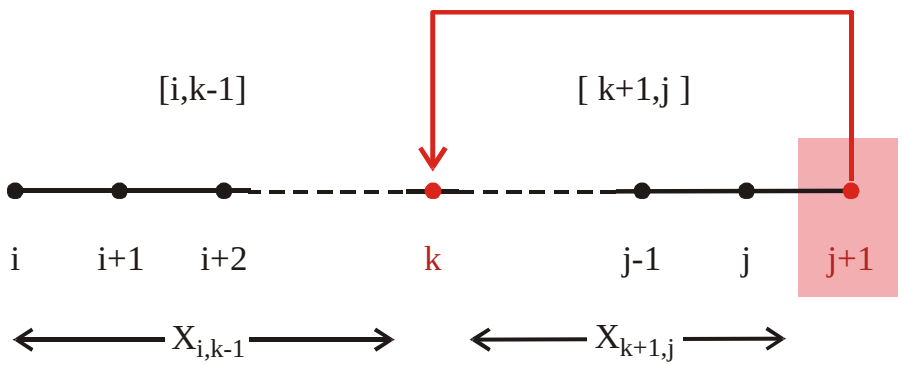
$$\Delta G_0^{300} = \underset{\substack{\text{stacks of}\\ \text{base pairs}}}{\sum g_{ij,kl}} + \underset{\substack{\text{hairpin}\\ \text{loops}}}{\sum h(n_l)} + \underset{\text{bulges}}{\sum b(n_b)} + \underset{\substack{\text{internal}\\ \text{loops}}}{\sum i(n_i)} + \cdots$$

Maximum matching

An example of a dynamic programming computation of the maximum number of base pairs

Back tracking yields the structure(s).

i		j = 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		G	G	C	G	C	G	C	C	C	G	G	C	G	C	C
1	G	*	*	1	1	1	1	2	3	3	3	4	4	5	6	6
2	G		*	*	0	1	1	2	2	2	3	3	4	4	5	6
3	C			*	*	0	1	1	1	2	3	3	3	4	5	5
4	G				*	*	0	1	1	2	2	2	3	4	5	5
5	C					*	*	0	1	1	2	2	3	4	4	4
6	G						*	*	1	1	1	2	3	3	3	4
7	C							*	*	0	1	2	2	2	2	3
8	C								*	*	1	1	1	2	2	2
9	C									*	*	1	1	2	2	2
10	G										*	*	1	1	1	2
11	G											*	*	0	1	1
12	C												*	*	0	1
13	G													*	*	1
14	C														*	*
15	C															*



$$X_{i,j+1} = \max\left\{ X_{i,j}, \max_{i \le k \le j-1}\left((X_{i,k-1} + 1 + X_{k+1,j})\,\rho_{k,j+1} \right) \right\}$$

Minimum free energy computations are based on empirical energies

1. Computation of RNA equilibrium structures
2. **Inverse folding and neutral networks**
3. Evolutionary optimization of structure
4. Suboptimal conformations and kinetic folding

RNA sequence: GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

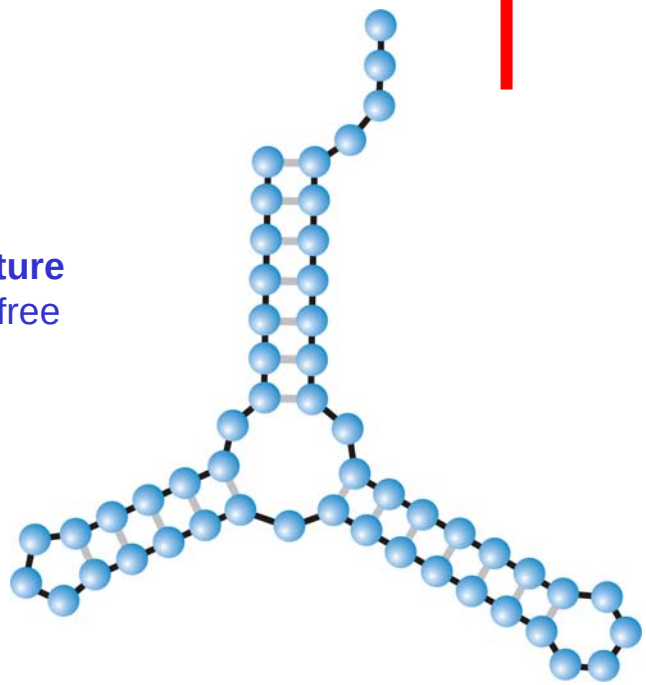
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

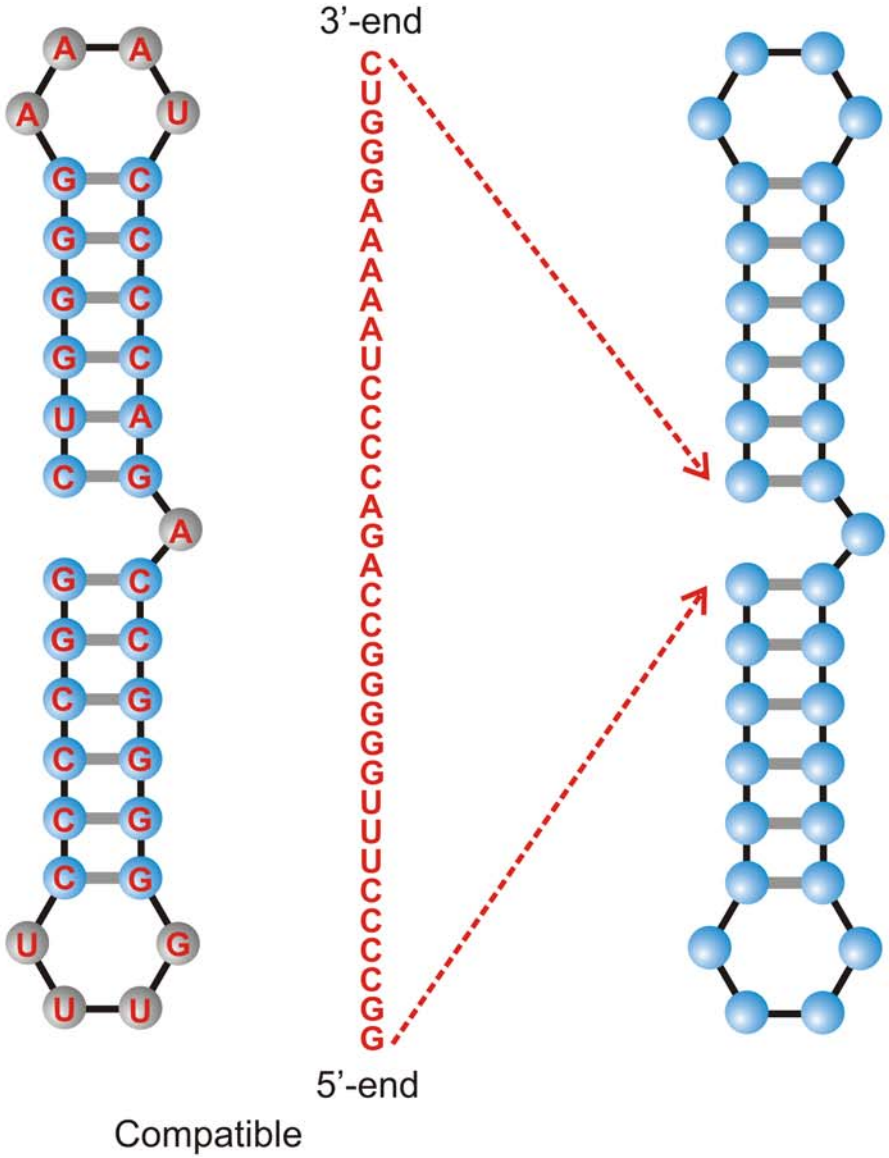
Inverse folding of RNA:

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

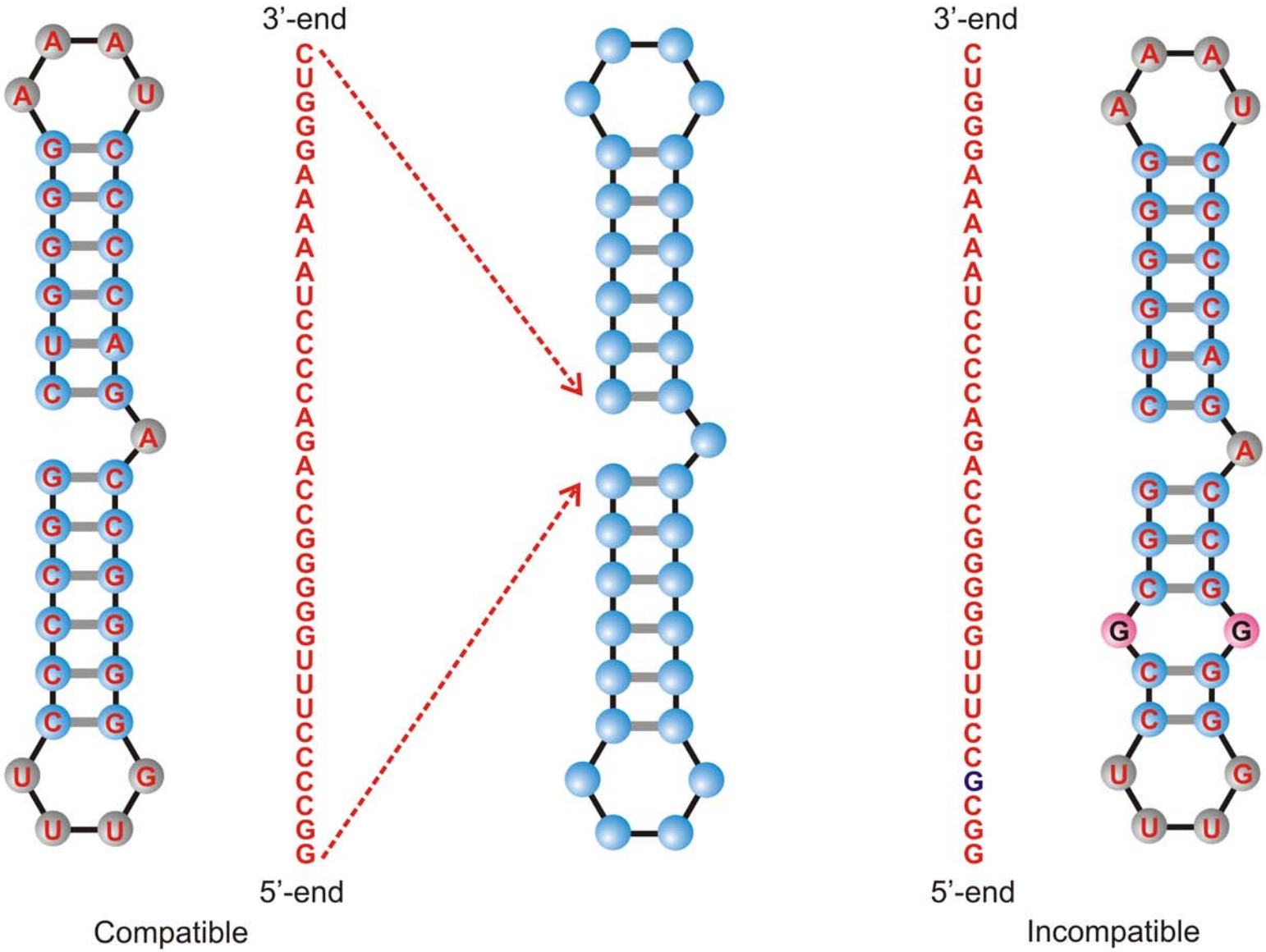
RNA structure of minimal free energy



Sequence, structure, and design



Compatibility of sequences and structures



Compatibility of sequences and structures

Inverse folding algorithm

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

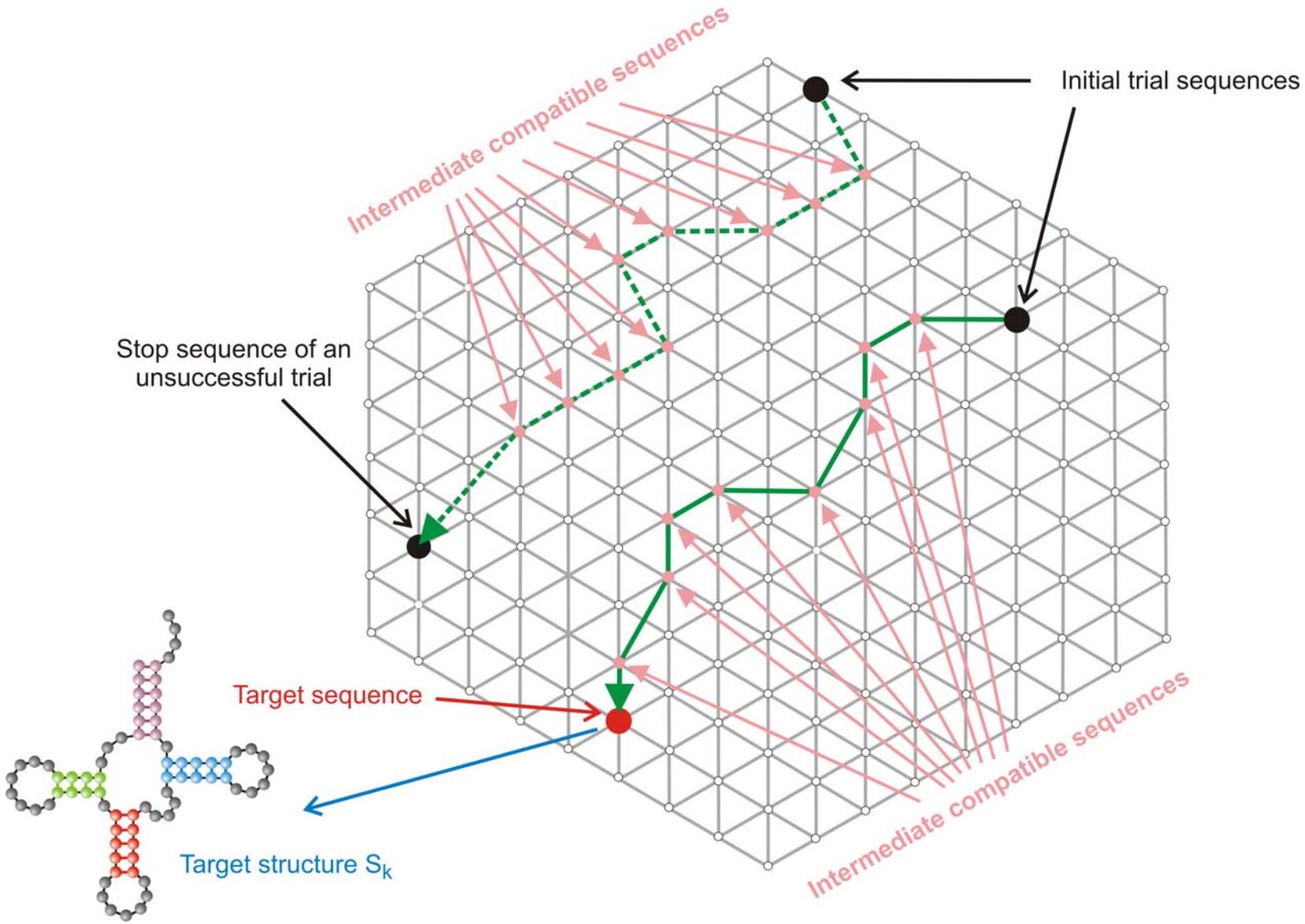
$\mathbf{S_0} \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow \ldots \rightarrow S_k \rightarrow S_{k+1} \rightarrow \ldots \rightarrow \mathbf{S_t}$

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

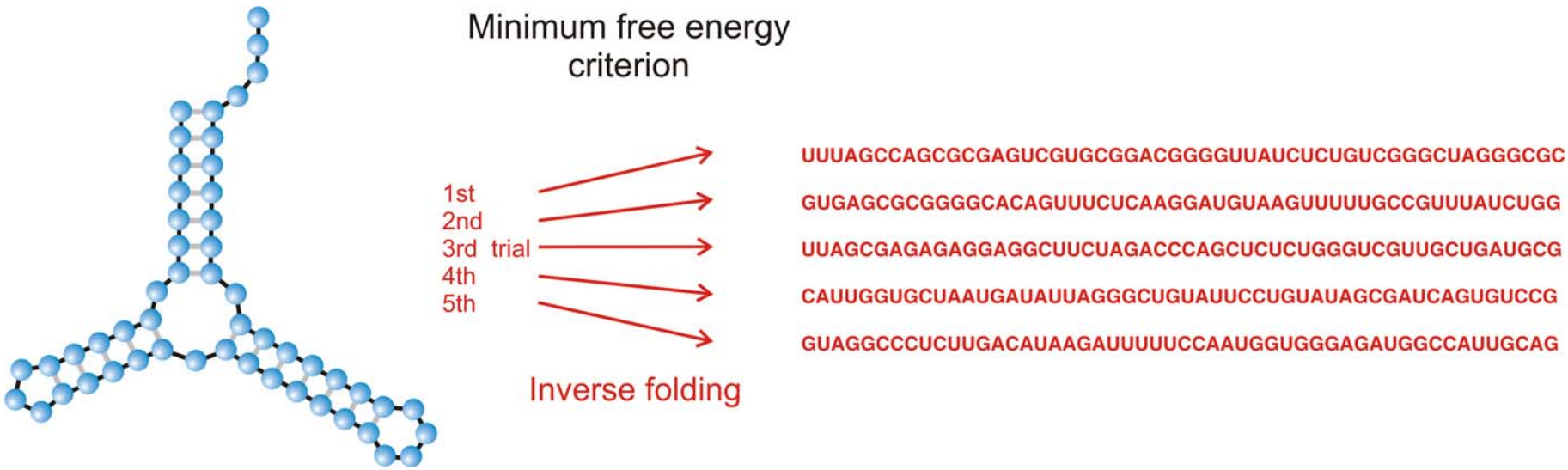
$\mathfrak{M}$ ... base or base pair mutation operator

$d_S\,(S_i,S_j)$ ... 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



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

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

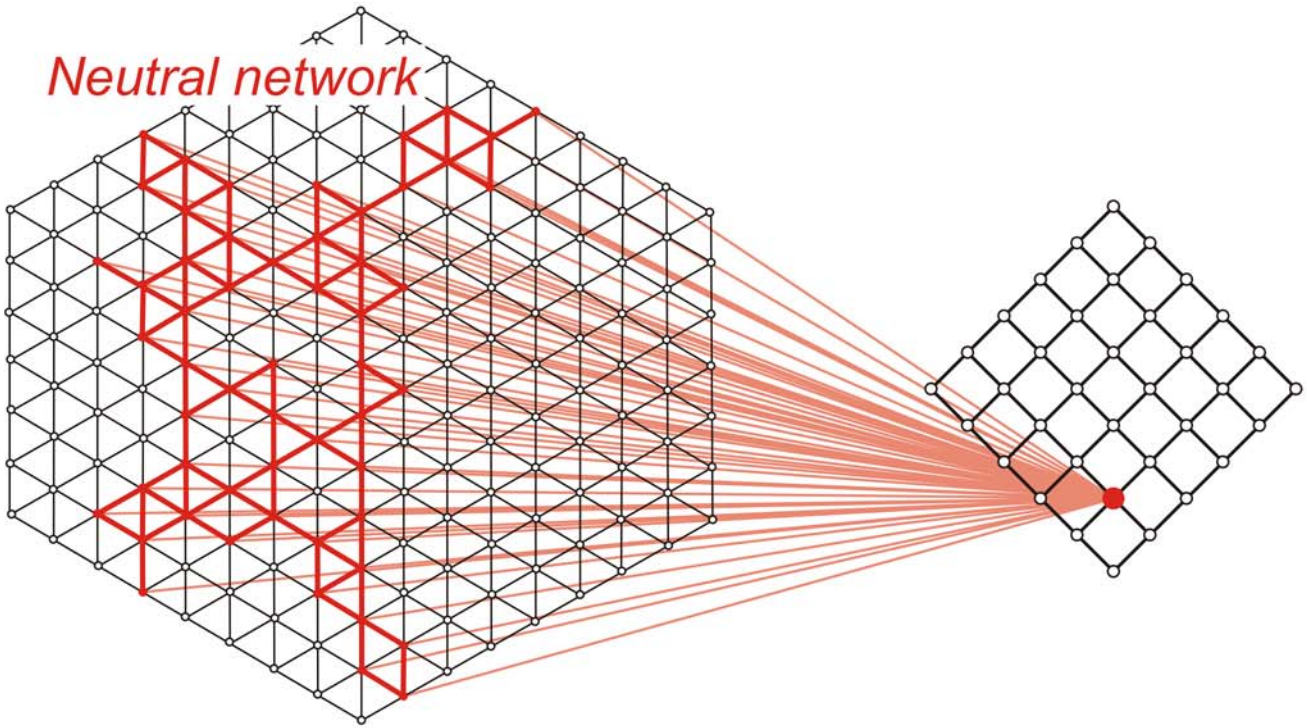
Space of phenotypes: $S = \{S_1, S_2, S_3, S_4, \ldots, S_M\}$; metric (not required)

$$N >> M$$

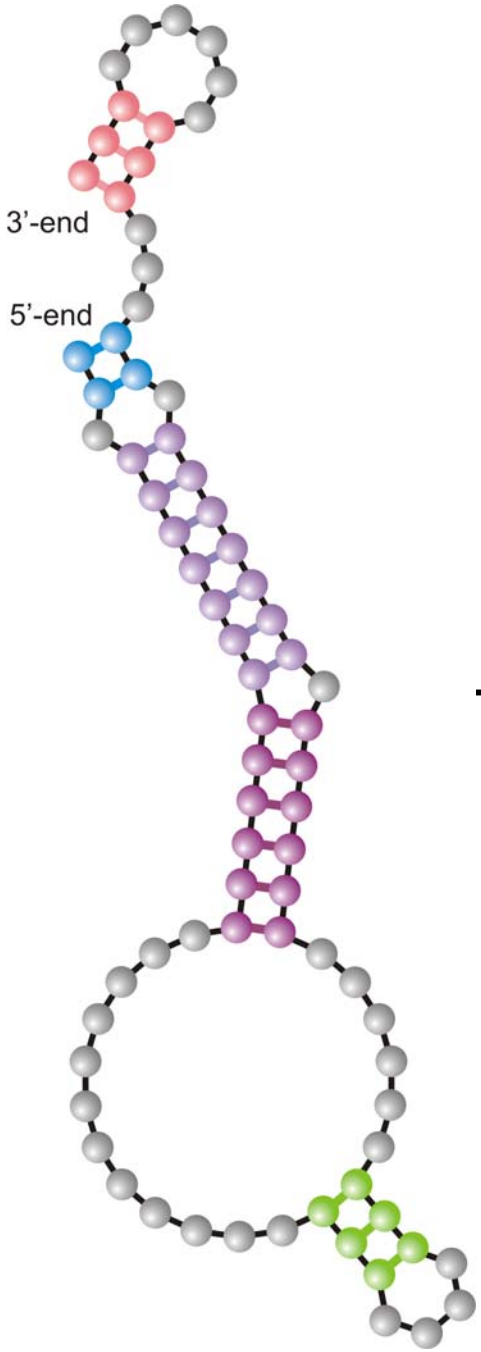
$$\psi(I_j) = S_k$$

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

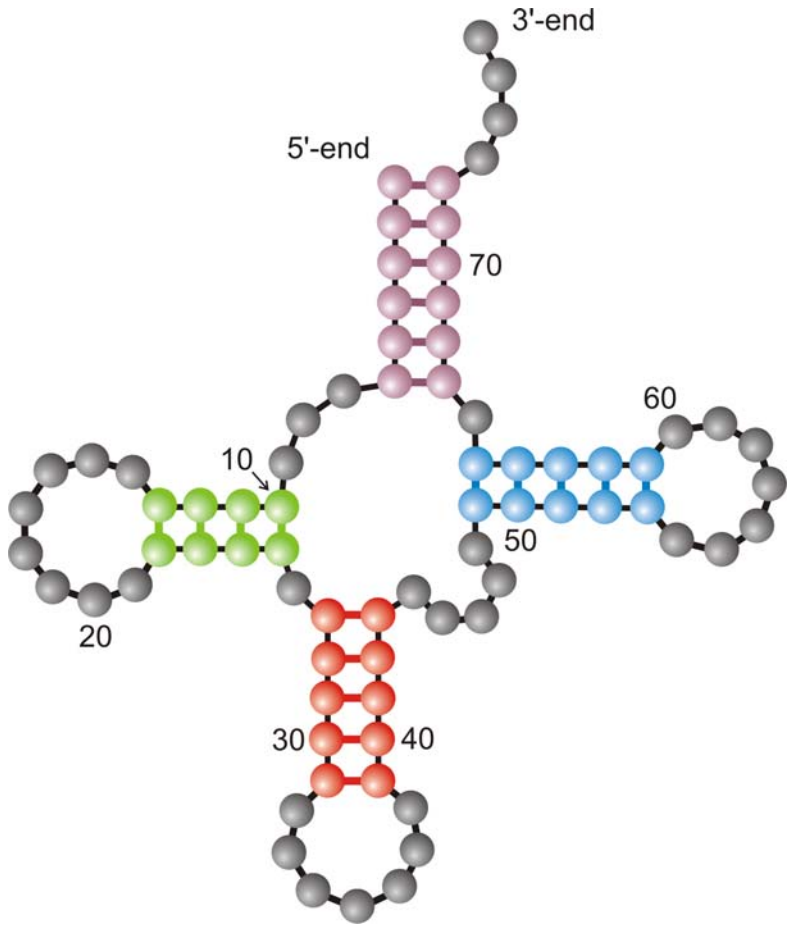
A mapping ψ and its inversion


Neutral network
Sequence space
Structure space

1. Computation of RNA equilibrium structures
2. Inverse folding and neutral networks
3. **Evolutionary optimization of structure**
4. Suboptimal conformations and kinetic folding



Structure of andomly chosen initial sequence



Phenylalanyl-tRNA as target structure

Evolution *in silico*

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

random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCT-CATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGT-TCCACC-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′-GCATGACCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-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*).

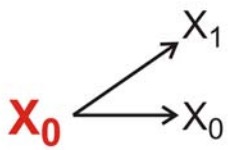
To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (*5*, *6*). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (*4*). Although 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.

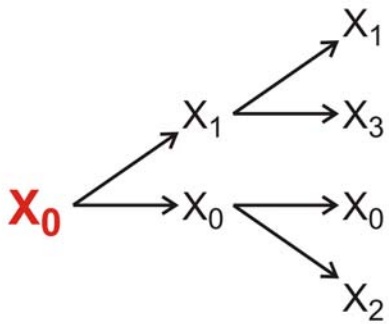
A set of entities is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally 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.

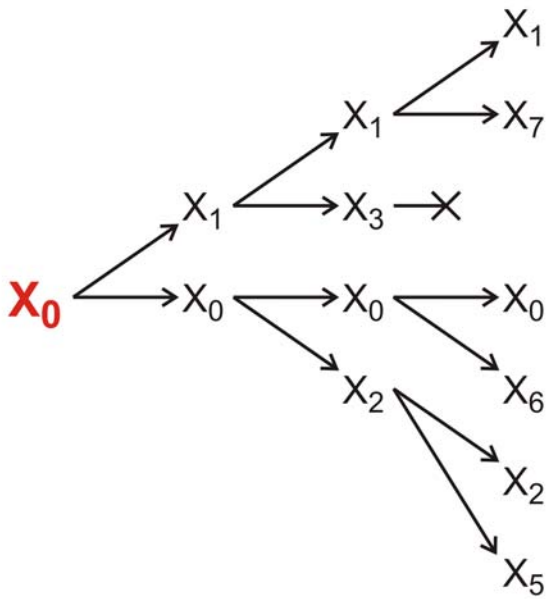
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



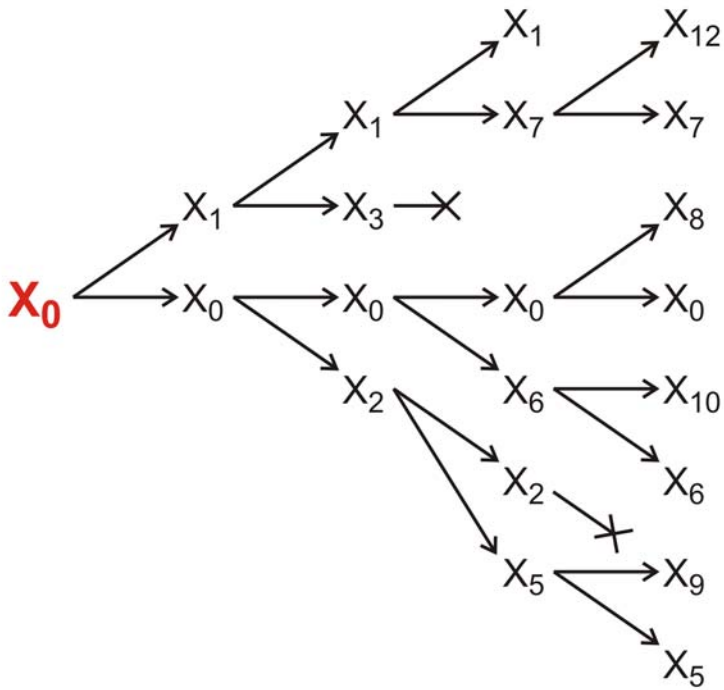
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



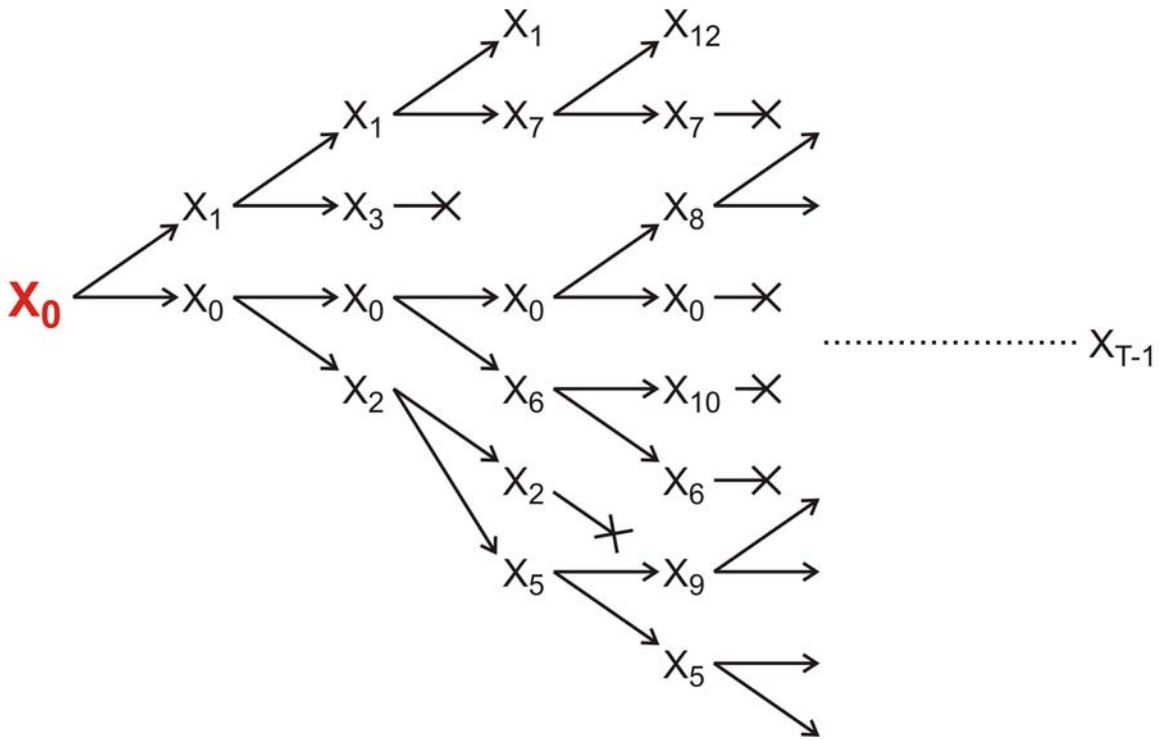
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



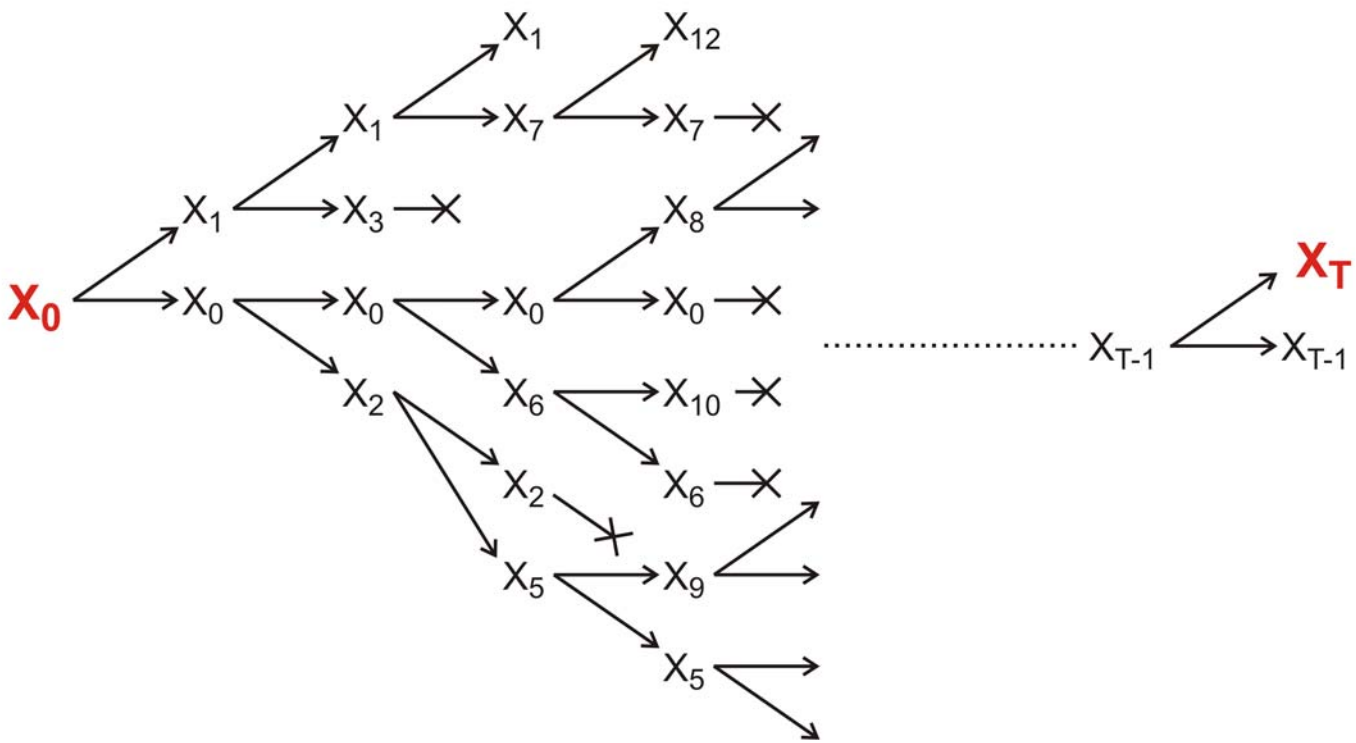
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



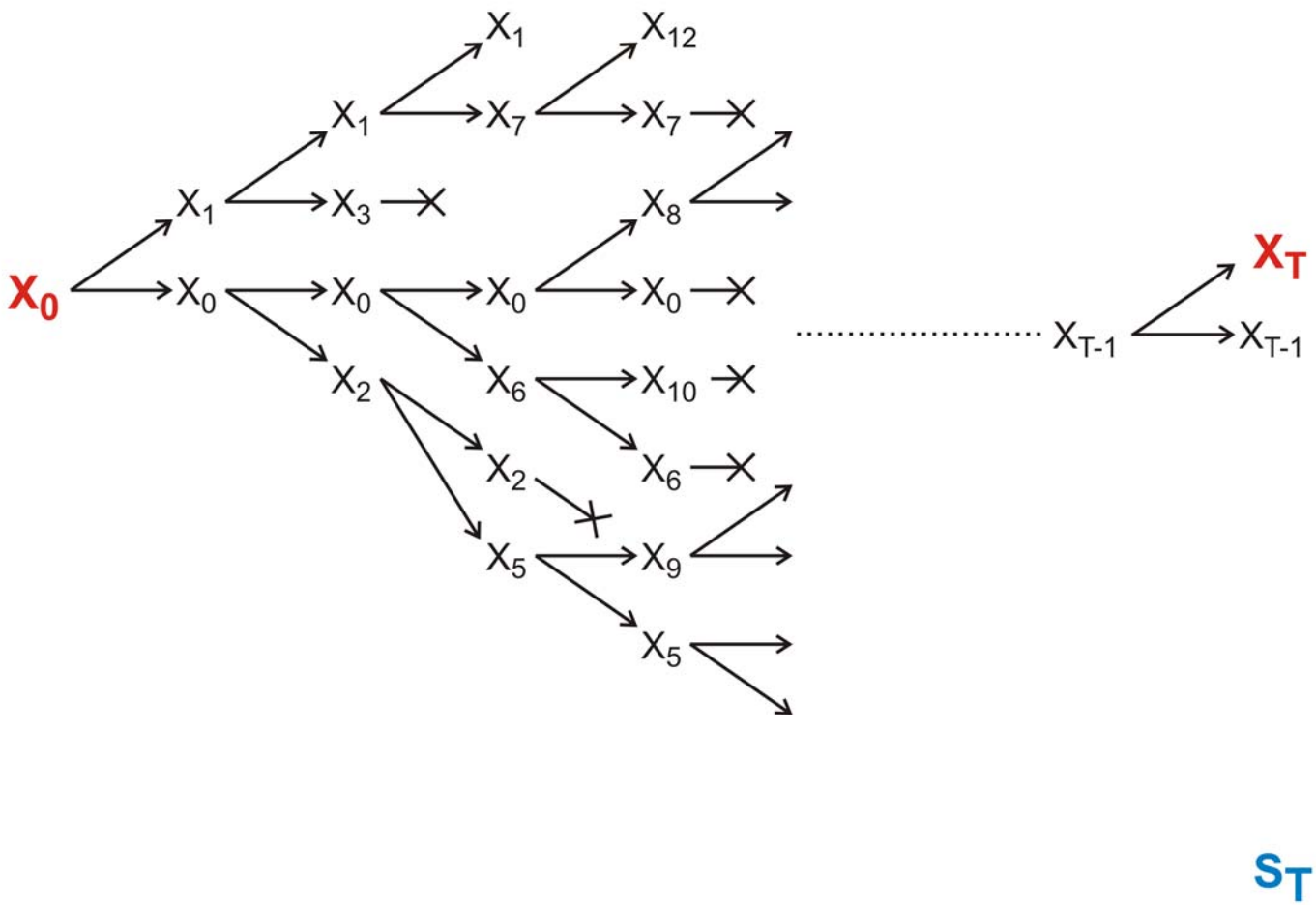
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



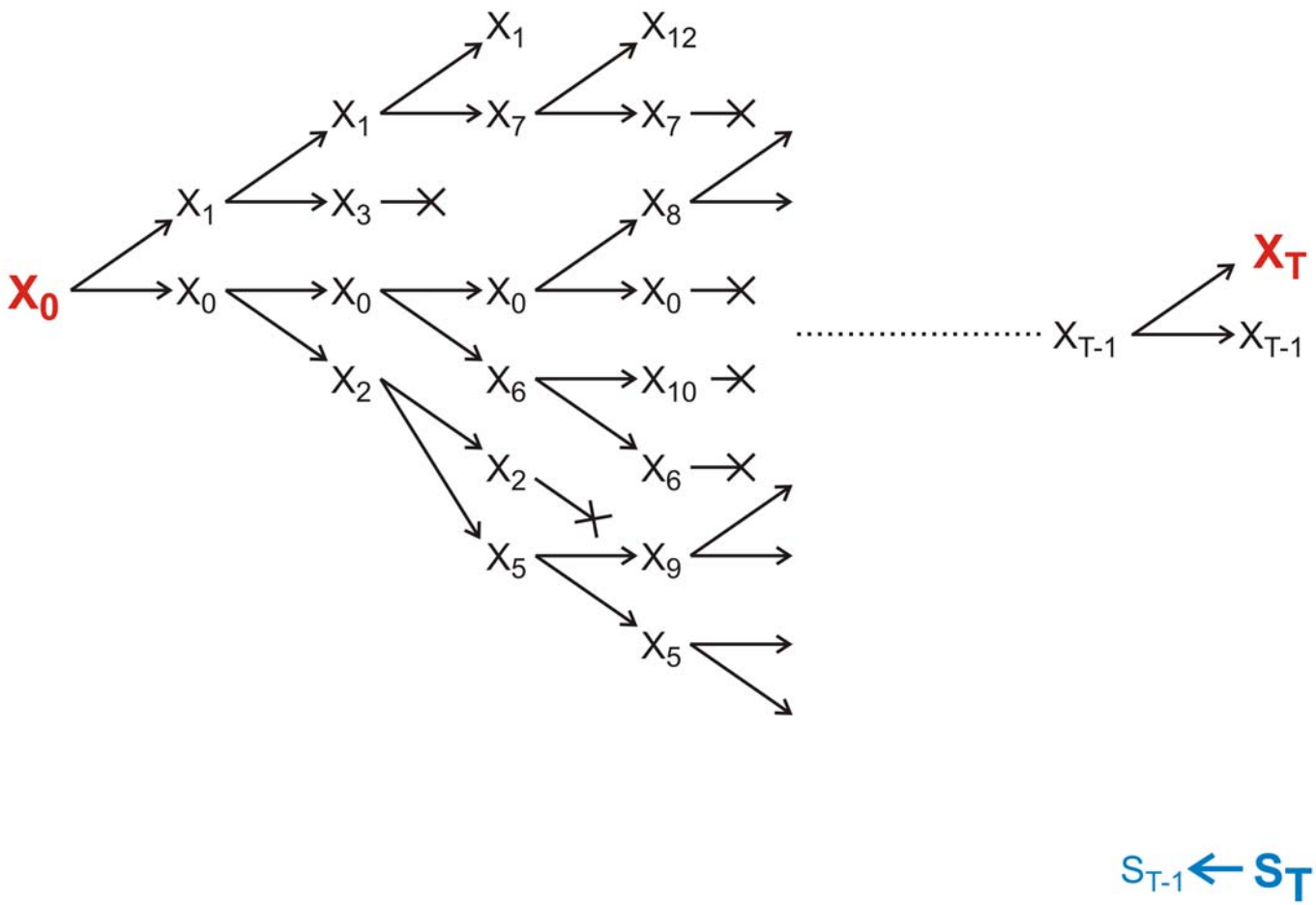
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



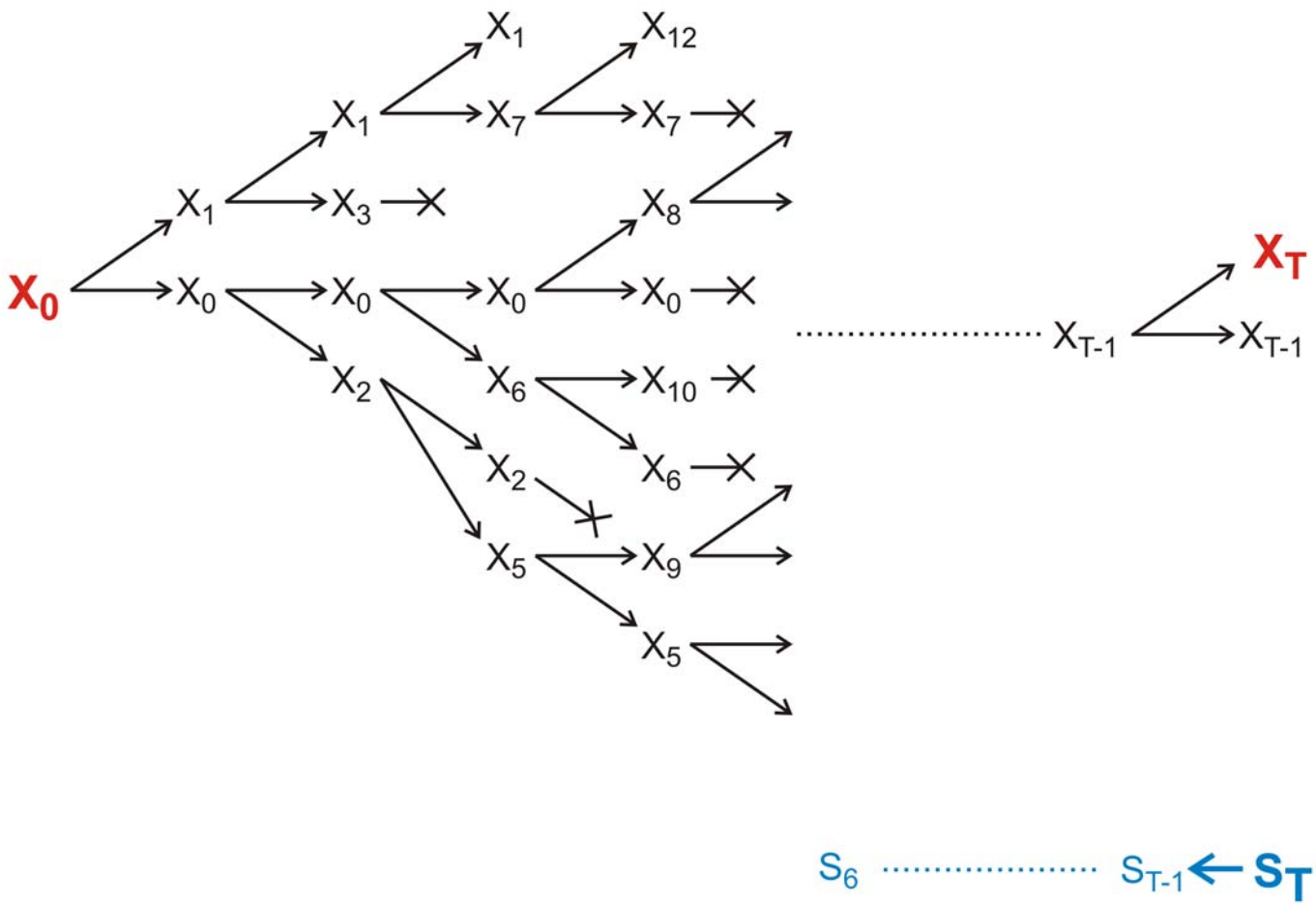
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



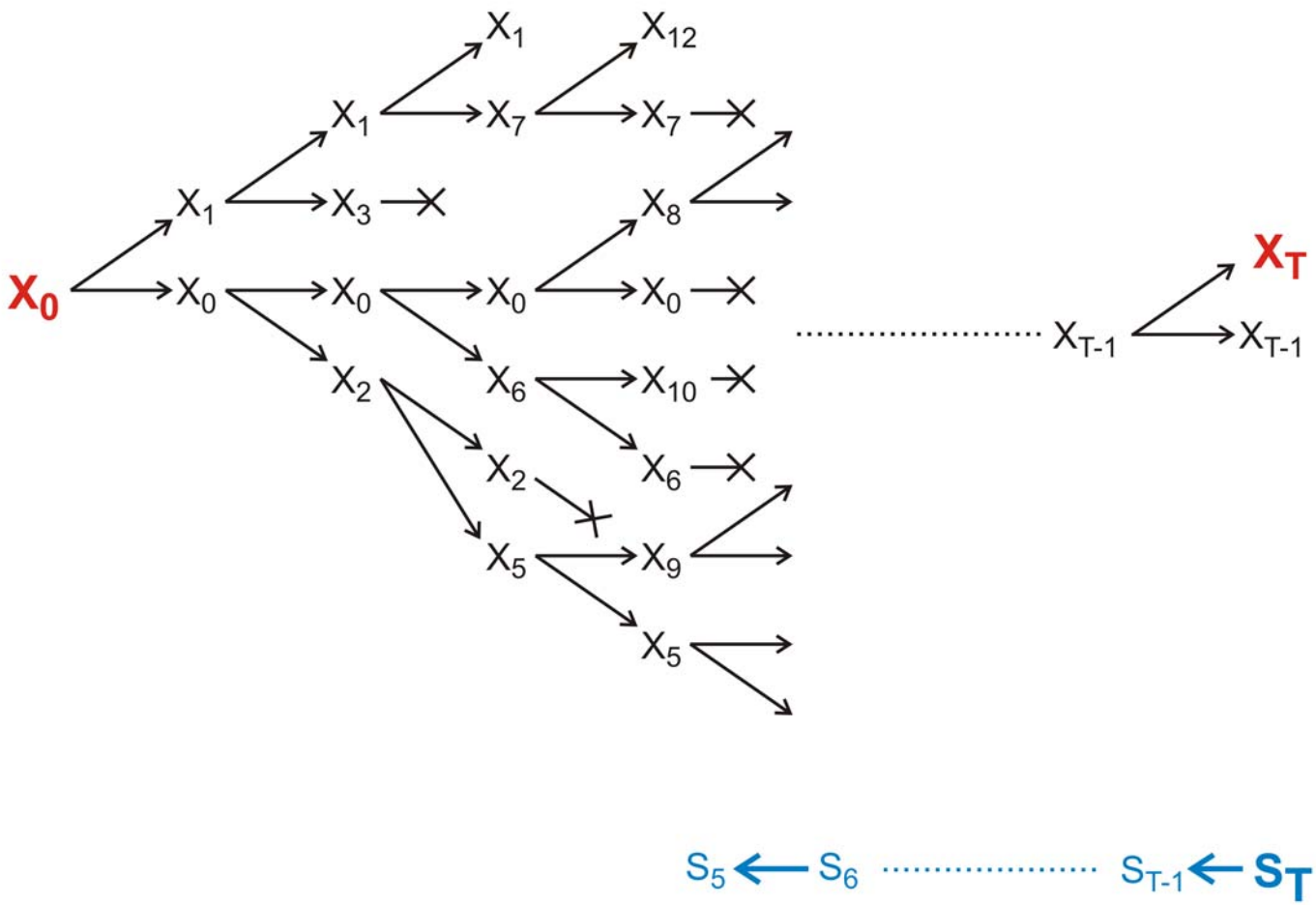
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



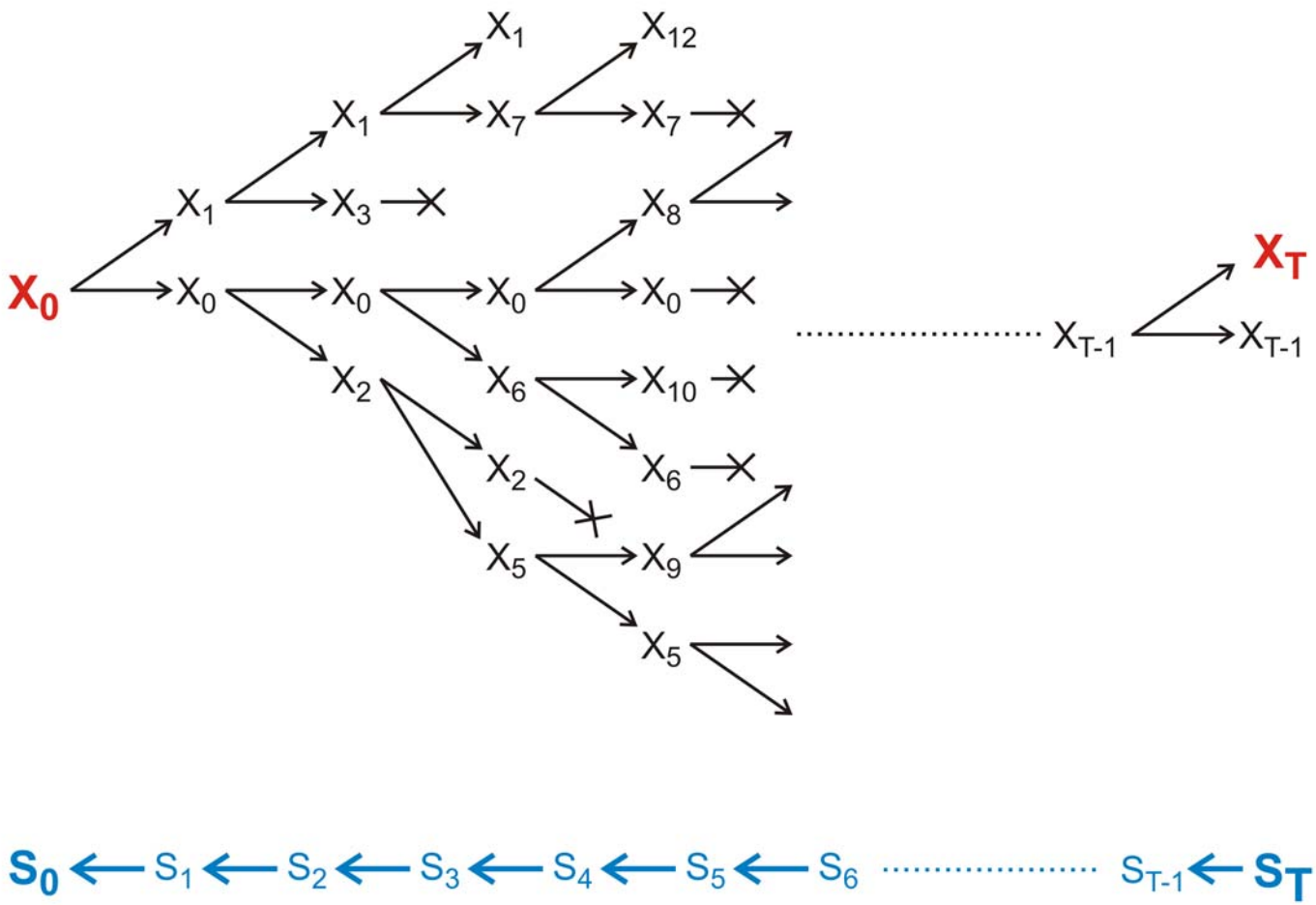
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



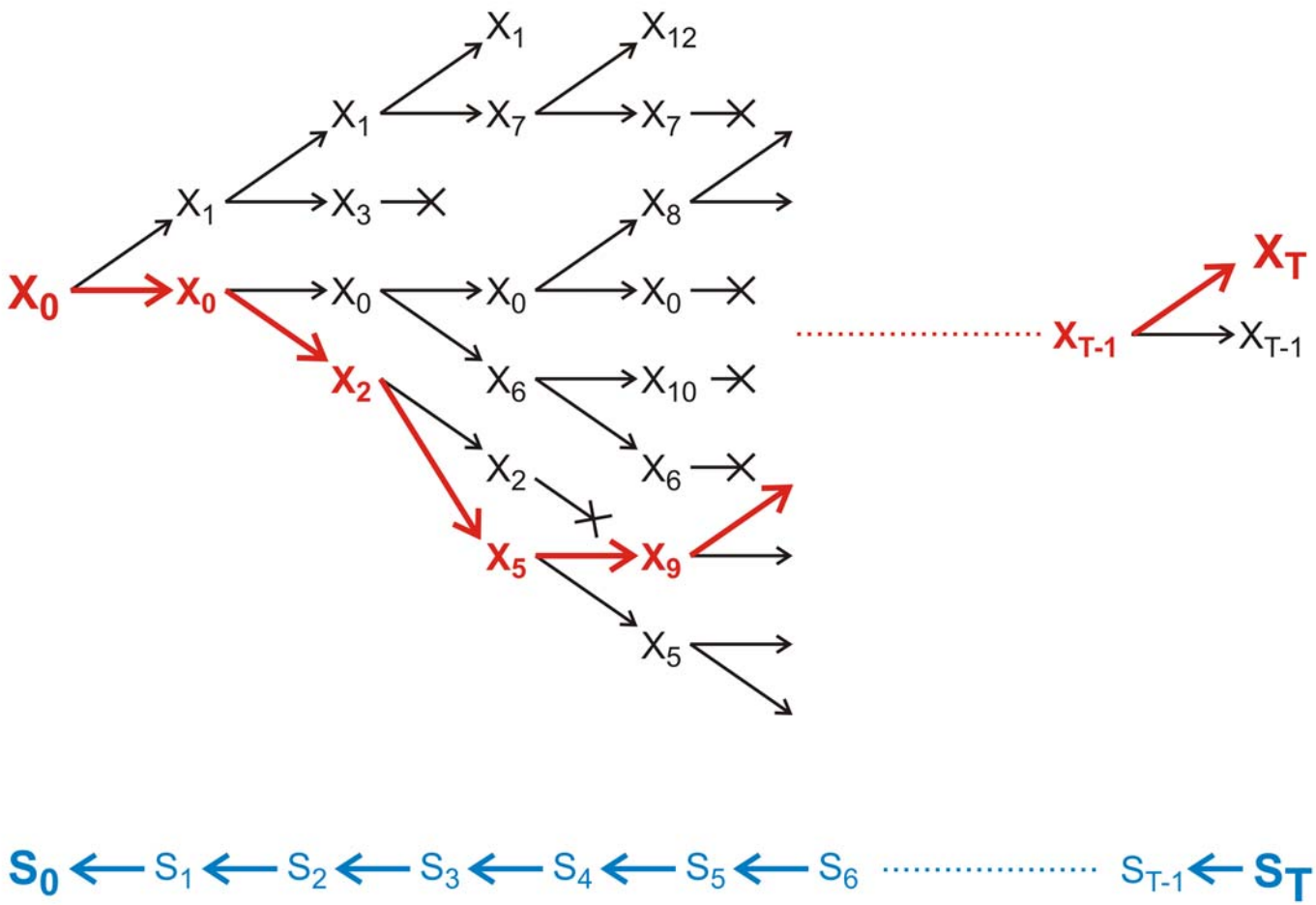
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



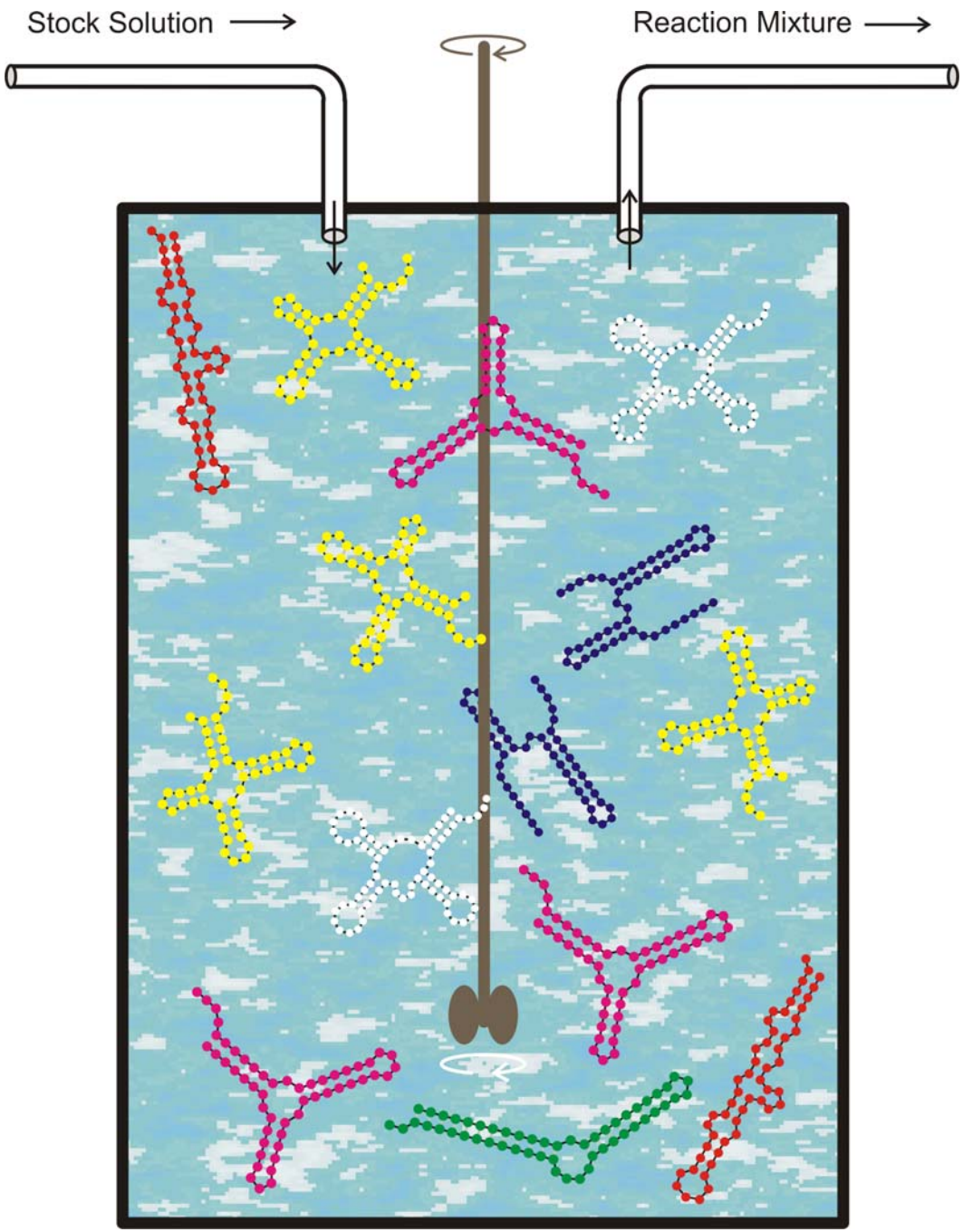
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Replication rate constant:

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

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

Selection constraint:

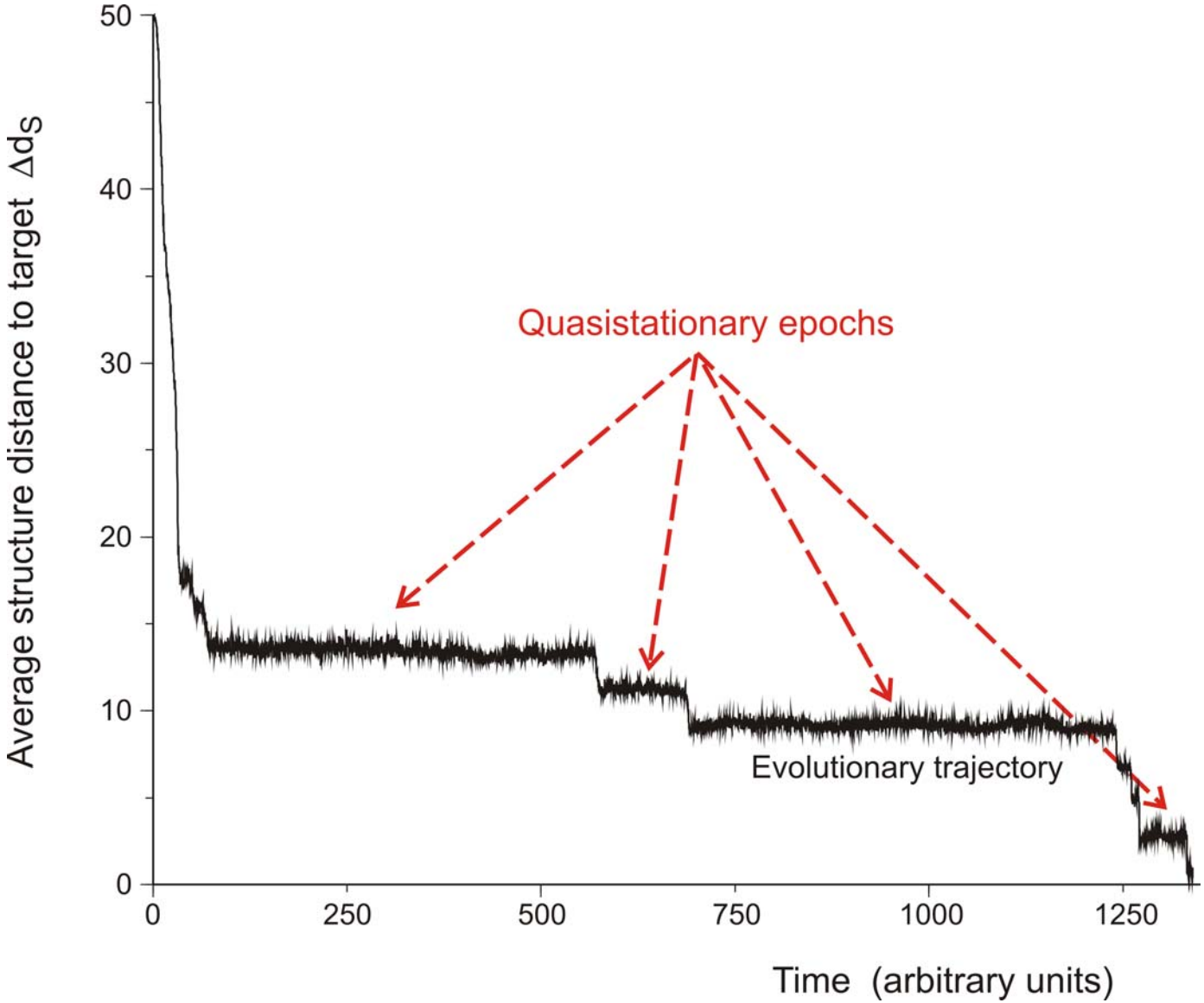
Population size, N = # RNA molecules, is controlled by the flow

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

Mutation rate:

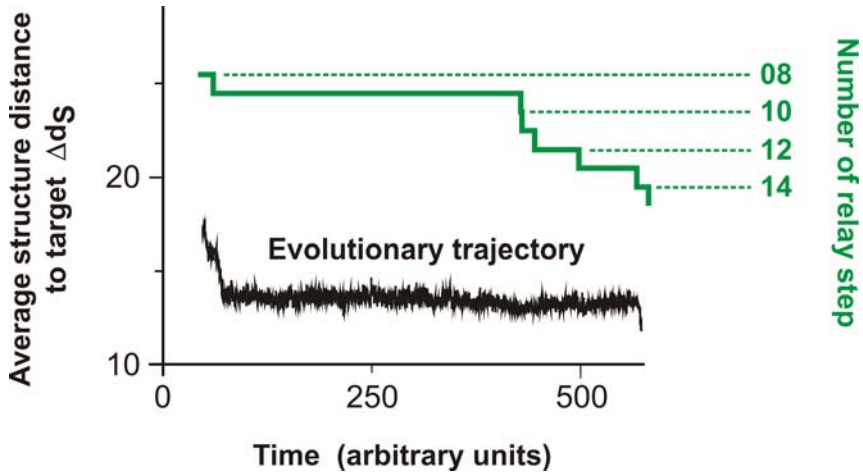
p = 0.001 / site × replication

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



In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch

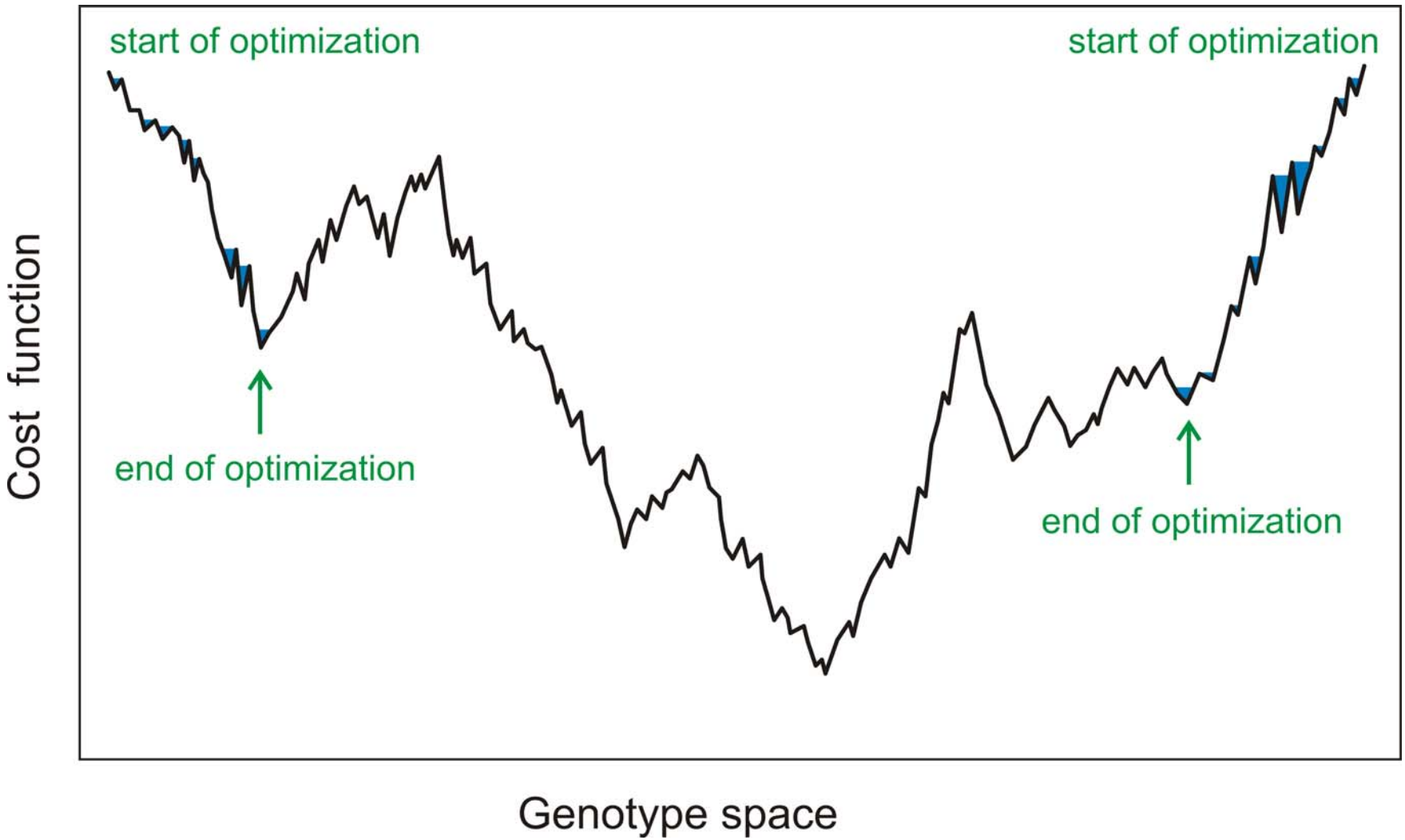


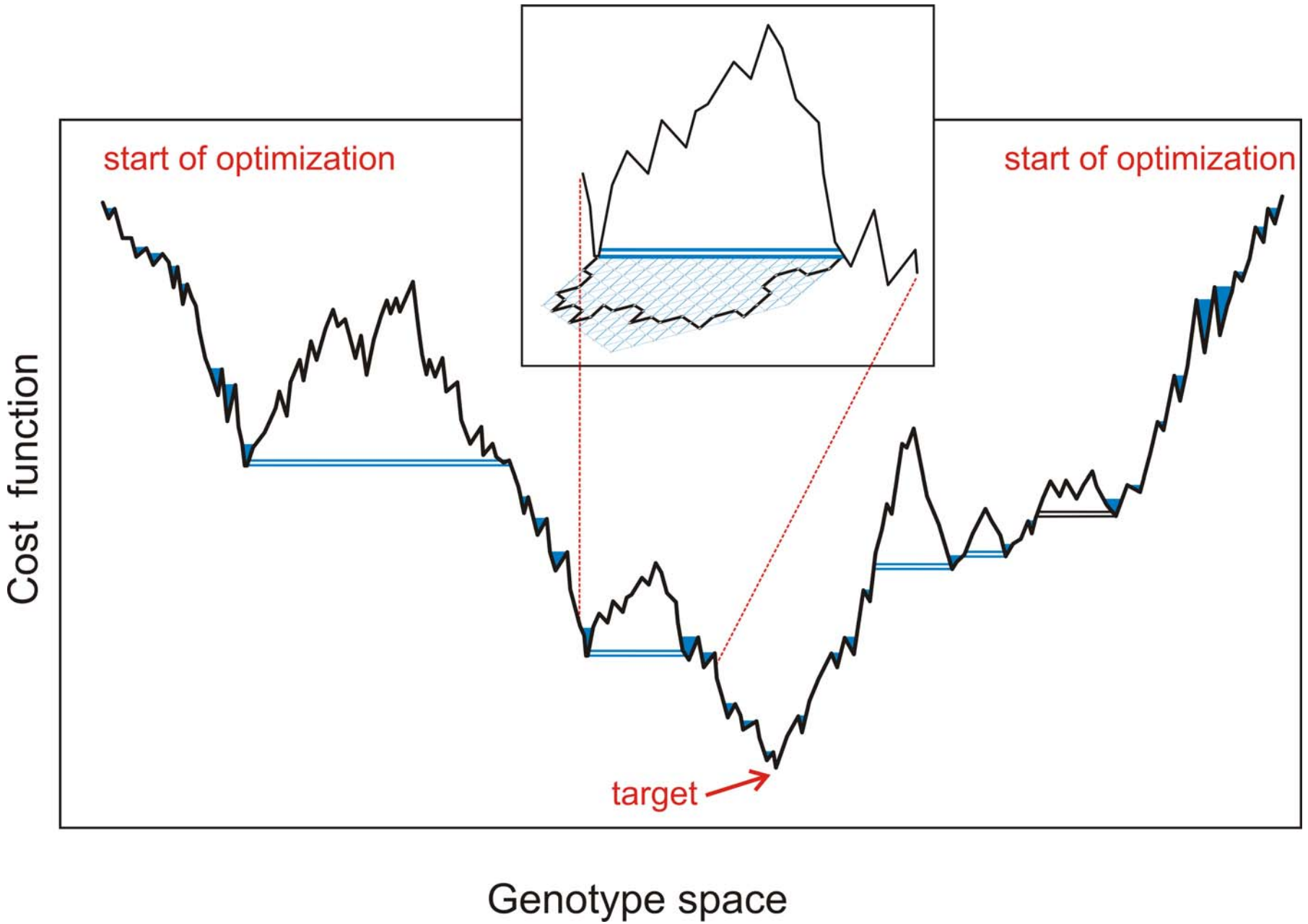

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9       .((((((.(((((.........(((....)))).......)))))).....((((.......))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10      .(((((..(((((.........(((....)))).......)))))).....((((.......))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

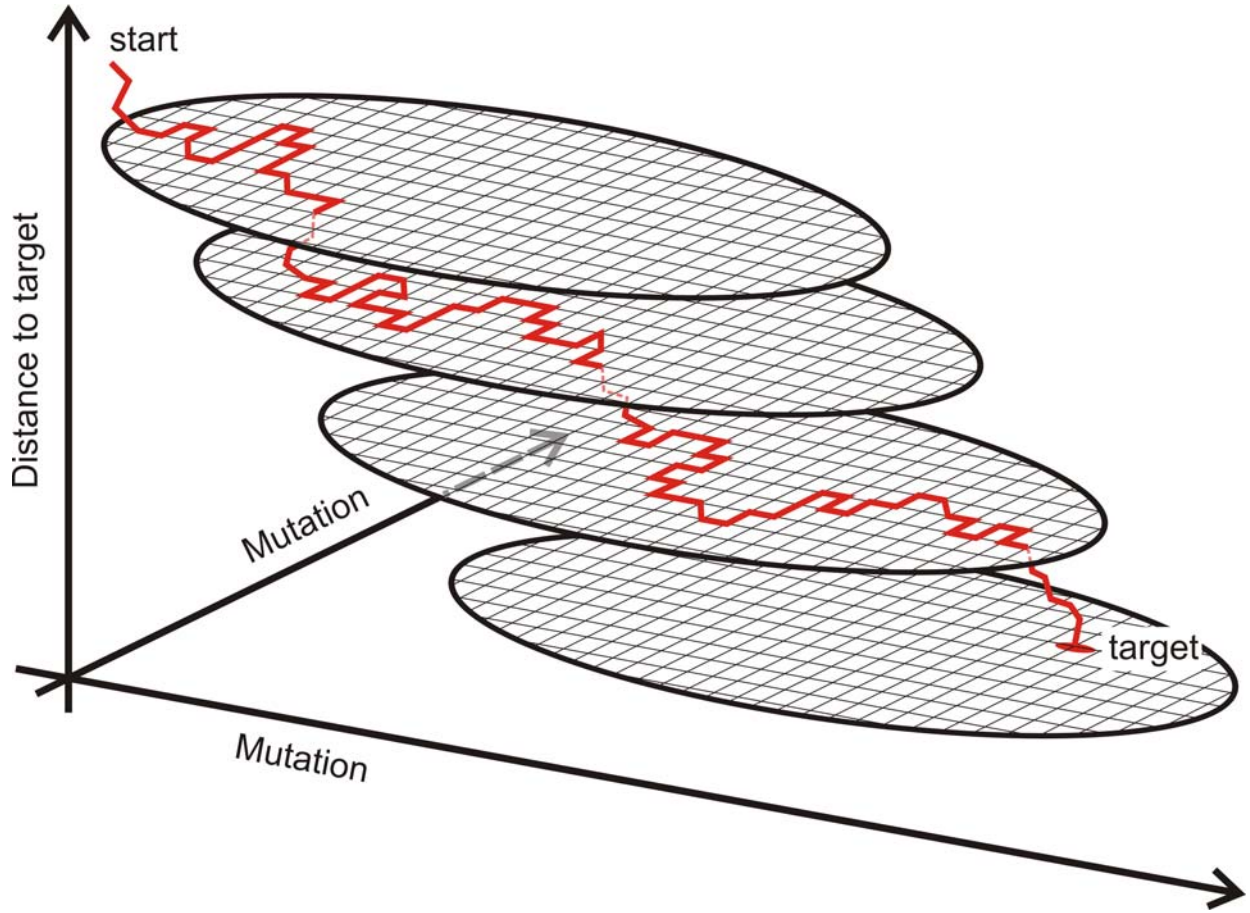
Transition inducing point mutations change the molecular structure

Neutral point mutations leave the molecular structure unchanged

Neutral genotype evolution during phenotypic stasis

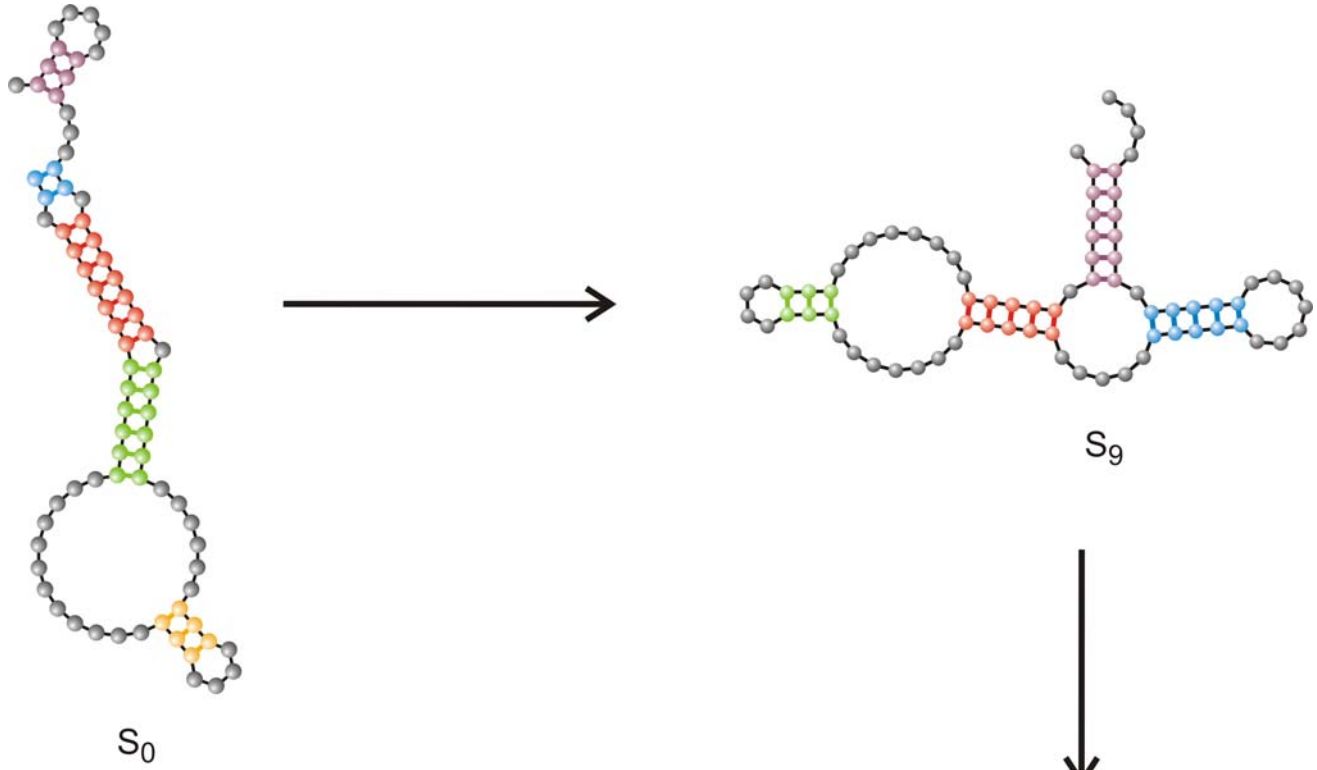

start of optimization
start of optimization
end of optimization
end of optimization
Cost function
Genotype space


start of optimization
start of optimization
target
Cost function
Genotype space

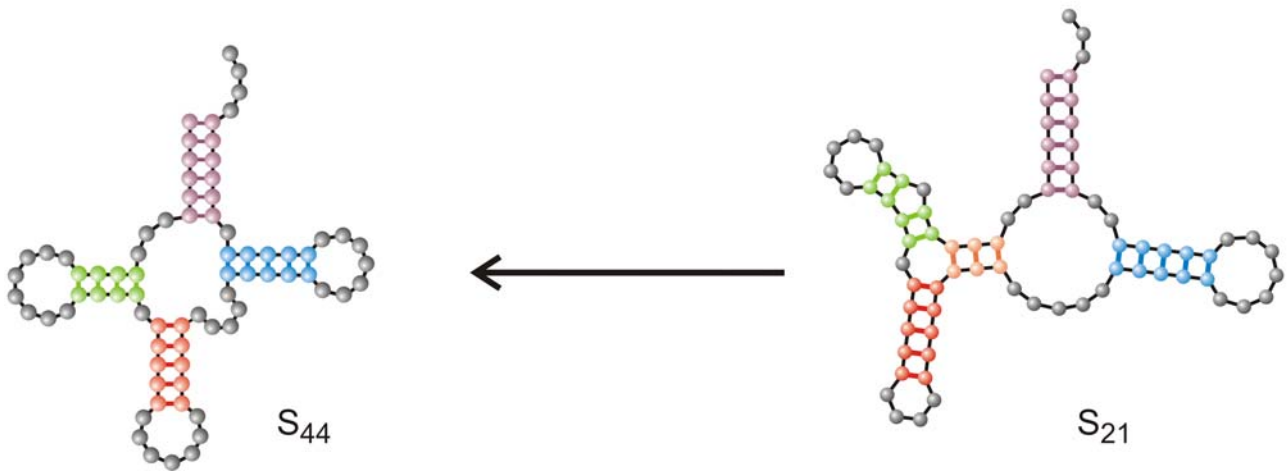


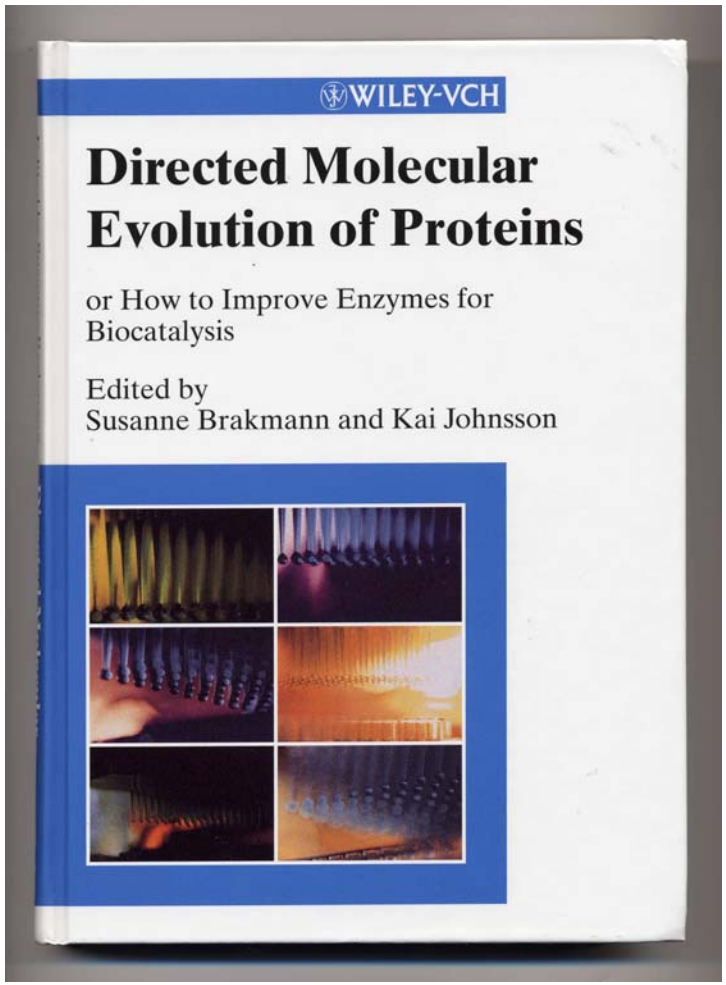
A sketch of optimization on neutral networks

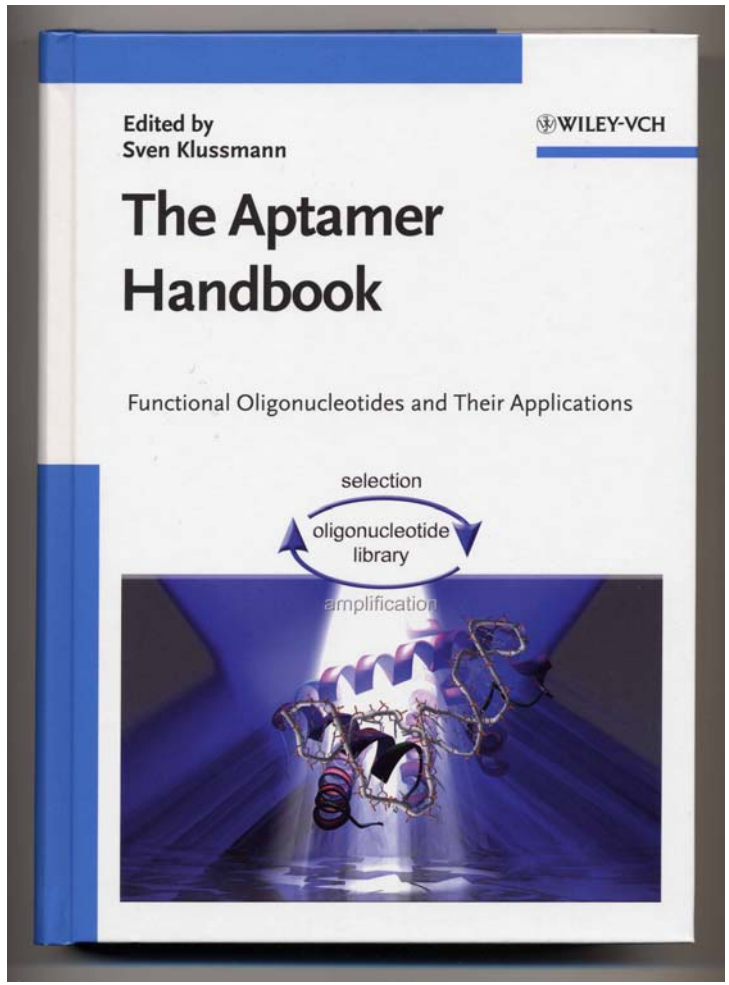
Randomly chosen
initial structure



Phenylalanyl-tRNA
as target structure

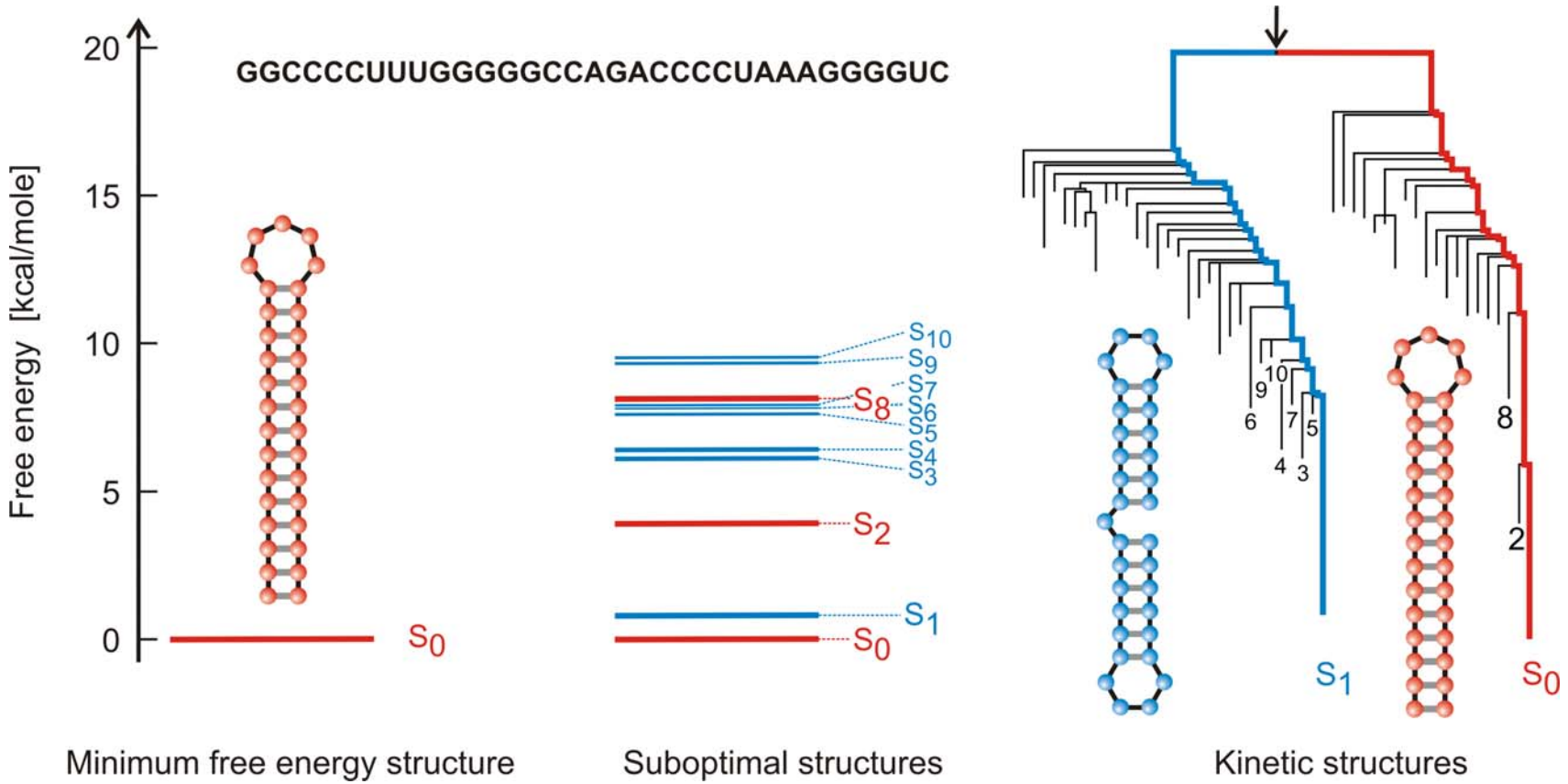






Application of molecular evolution to problems in biotechnology

1. Computation of RNA equilibrium structures
2. Inverse folding and neutral networks
3. Evolutionary optimization of structure
4. **Suboptimal conformations and kinetic folding**



RNA secondary structures derived from a single sequence

An algorithm for the computation of all suboptimal structures of RNA molecules using the same concept for retrieval as applied in the sequence alignment algorithm by

M.S. Waterman and T.F. Smith. Math.Biosci. 42:257-266, 1978.

Stefan Wuchty[1]
Walter Fontana[1,2]
Ivo L. Hofacker[1]
Peter Schuster[1,2]
[1] *Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria*

[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA*

Received 13 May 1998; accepted 6 August 1998

Complete Suboptimal Folding of RNA and the Stability of Secondary Structures

Abstract: *An algorithm is presented for generating rigorously all suboptimal secondary structures between the minimum free energy and an arbitrary upper limit. The algorithm is particularly fast in the vicinity of the minimum free energy. This enables the efficient approximation of statistical quantities, such as the partition function or measures for structural diversity. The density of states at low energies and its associated structures are crucial in assessing from a thermodynamic point of view how well-defined the ground state is. We demonstrate this by exploring the role of base modification in tRNA secondary structures, both at the level of individual sequences from* Escherichia coli *and by comparing artificially generated ensembles of modified and unmodified sequences with the same tRNA structure. The two major conclusions are that (1) base modification considerably sharpens the definition of the ground state structure by constraining energetically adjacent structures to be similar to the ground state, and (2) sequences whose ground state structure is thermodynamically well defined show a significant tendency to buffer single point mutations. This can have evolutionary implications, since selection pressure to improve the definition of ground states with biological function may result in increased neutrality.* © 1999 John Wiley & Sons, Inc.
Biopoly 49: 145–165, 1999

Keywords: *RNA secondary structure; suboptimal folding; density of states; tRNA; modified bases; thermodynamic stability of structure; mutational buffering; neutrality; dynamic programming*

INTRODUCTION

The structure of RNA molecules can be discussed at an empirically well established level of resolution known as secondary structure. It refers to a topology of binary contacts arising from specific base pairing, rather than a geometry cast in terms of coordinates and distances (see Figure 1). The driving force behind

Correspondence to: Walter Fontana; email: walter@santafe.edu
Contract grant sponsor: Austrian Fond zur Förderung der Wissenschaftlichen Forschung (FWF) and Santa Fe Institute
Contract grant number: 11065-CHE (FWF)
Biopolymers, Vol. 49, 145–165 (1999)

CCC 0006-3525/99/020145-21

An algorithm for the computation of RNA folding kinetics

RNA (2000), *6*:325–338. Cambridge University Press. Printed in the USA.

RNA folding at elementary step resolution

CHRISTOPH FLAMM,[1] WALTER FONTANA,[2,3] IVO L. HOFACKER,[1] and PETER SCHUSTER[1]

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, A-1090 Wien, Austria
[2]Santa Fe Institute, Santa Fe, New Mexico 87501 USA

ABSTRACT

We study the stochastic folding kinetics of RNA sequences into secondary structures with a new algorithm based on the formation, dissociation, and the shifting of individual base pairs. We discuss folding mechanisms and the correlation between the barrier structure of the conformational landscape and the folding kinetics for a number of examples based on artificial and natural sequences, including the influence of base modification in tRNAs.

Keywords: conformational spaces; foldability; RNA folding kinetics; RNA secondary structure

INTRODUCTION

The conformational diversity of nucleic acids or proteins is delimited by the loose random coil and the compact native state that is frequently the most stable or minimum free energy (mfe) conformation. Let us call a specific interaction between two segments of the chain a "contact." A random coil then is best characterized by the absence of contacts, whereas the mfe conformation maximizes their energetic contributions. Several different types of contacts are found in three-dimensional structures. Their energetics is not well understood, which makes the modeling of RNA folding from random coils into full structures too ill-defined to be tackled at present.

Fortunately, for single-stranded nucleic acid molecules, the simpler coarse-grained notion of secondary structure is accessible to mathematical analysis and computation. To a theorist the secondary structure is the topology of binary contacts that arises from specific base pairing (Watson–Crick and GU; see Figure 1 and the next section). It does not refer to a two- or three-dimensional geometry cast in terms of distances. Secondary structure formation is driven by the stacking between contiguous base pairs. However, any formation of an energetically favorable double-stranded region implies the simultaneous formation of an energetically unfavorable loop. This frustrated energetics leads to a vast combinatorics of stack and loop arrangements spanning the conformational repertoire of an individual RNA sequence at the secondary structure level.

The secondary structure is not only an abstract tool convenient for theorists. It also corresponds to an actual state that provides a geometric, kinetic, and thermodynamic scaffold for tertiary structure formation, and constitutes an intermediate on the folding path from random coil to full structure. With rising temperature, tertiary contacts usually disappear first and double helices melt later (Banerjee et al., 1993). The free energy of secondary structure formation accounts for a large fraction of the free energy of full structure formation. These roles put the secondary structure in correspondence with functional properties of the tertiary structure. Consequently, selection pressures become observable at the secondary structure level in terms of evolutionarily conserved base pairs (Gutell, 1993). Moreover, insights into the process of secondary structure formation can be extended to several types of tertiary contacts with roughly conserved local geometries, such as non-Watson–Crick base pairs, base triplets and quartets, or end-on-end stacking of double helices.

To provide a frame for our kinetic treatment of RNA folding, we give a short account of the formal issues surrounding conformational spaces, folding trajectories, and folding paths for RNA secondary structures. We then introduce the kinetic folding algorithm as a stochastic process in the conformation space of a sequence, and discuss applications to several selected problems that cannot be studied adequately with the thermodynamic approach alone.

Reprint request to: Christoph Flamm, Institut für Theoretische Chemie und Molekulare Strukturbiologie, Währingerstrasse 17, A-1090 Wien, Austria; e-mail: xtof@tbi.univie.ac.at.
[3]Present address: Institute for Advanced Study, Program in Theoretical Biology, 310 Olden Lane, Princeton, New Jersey 08540, USA.

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:

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

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

Kinetic equation

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

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

Transition rate prameters $P_{ij}(t)$ are defined by

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

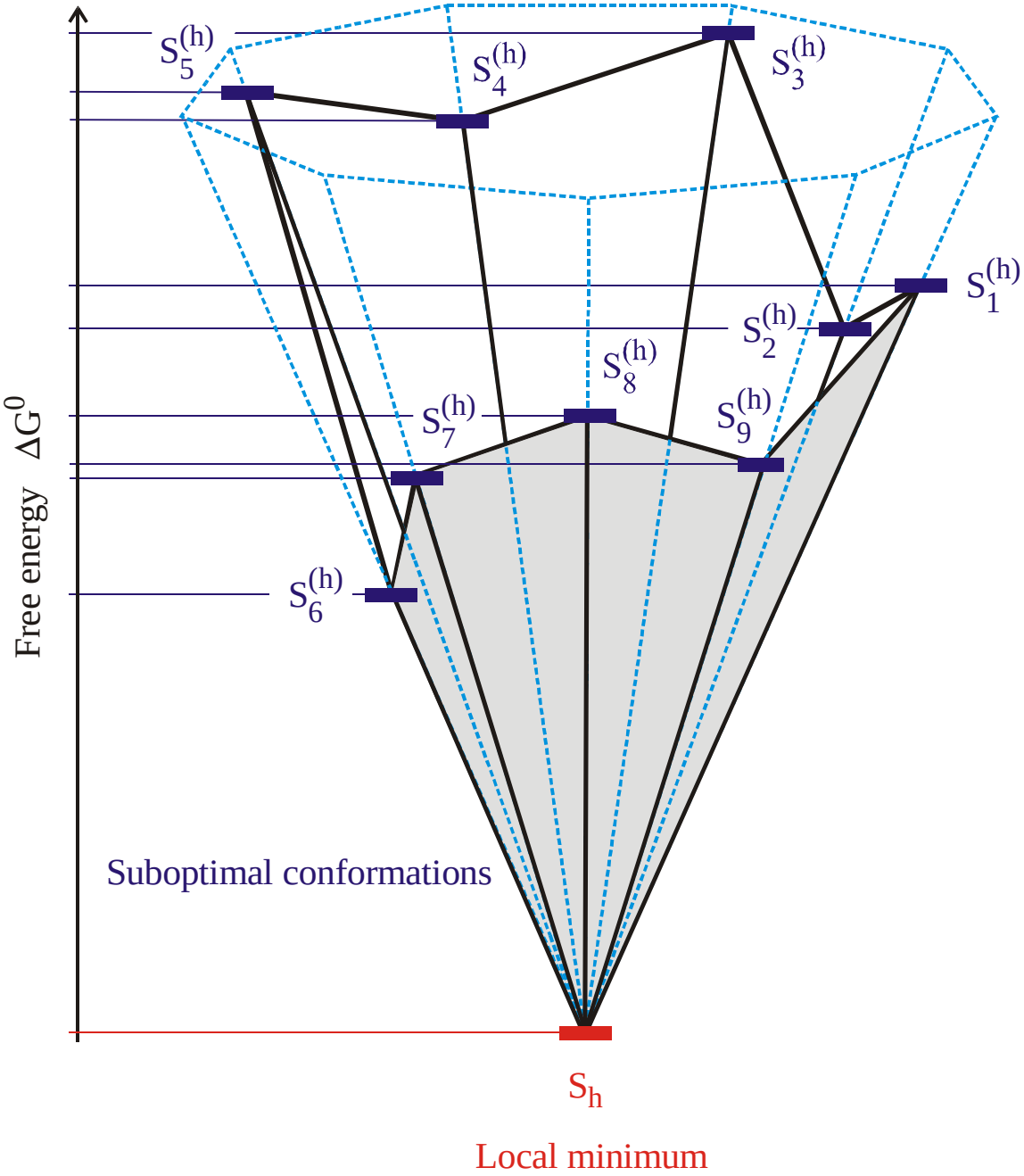
$$P_{ji}(t) = P_j(t)\, k_{ji} = P_j(t) \exp(-\Delta G_{ji}/2RT) / \Sigma_j$$

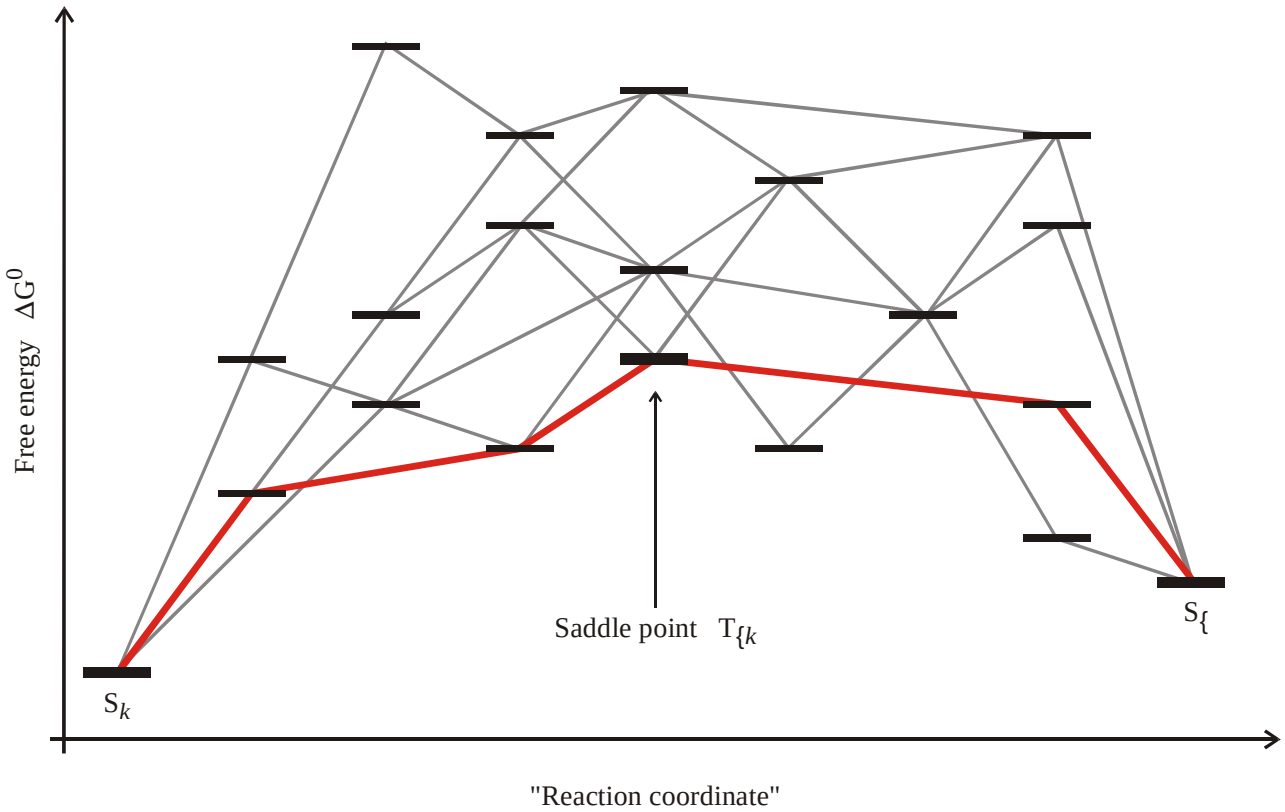
$$\Sigma_k = \sum_{k=1, k\neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

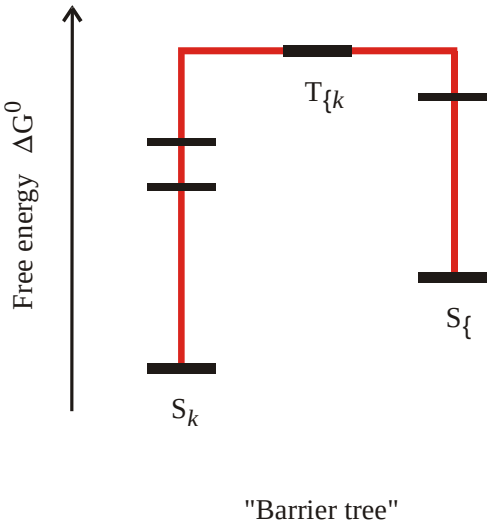
The symmetric rule for transition rate parameters is due to Kawasaki (K. Kawasaki, *Diffusion constants near the critical point for time dependent Ising models*. Phys.Rev. **145**:224-230, 1966).

Formulation of kinetic RNA folding as a stochastic process and by reaction kinetics

Search for local minima in conformation space

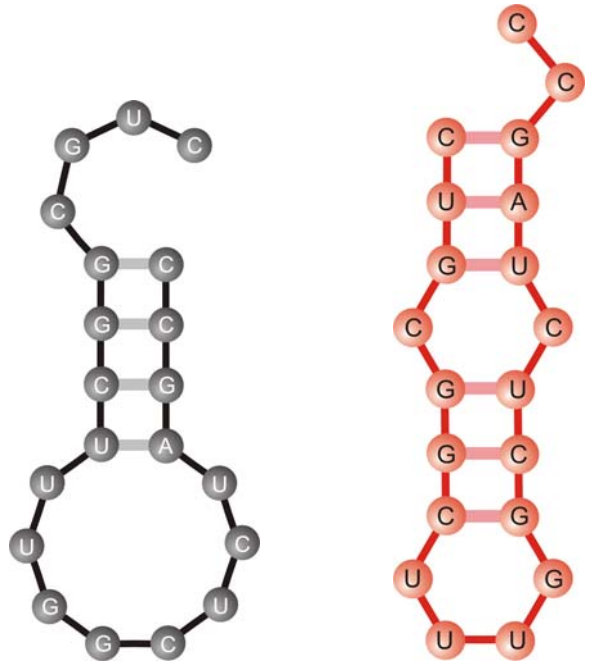






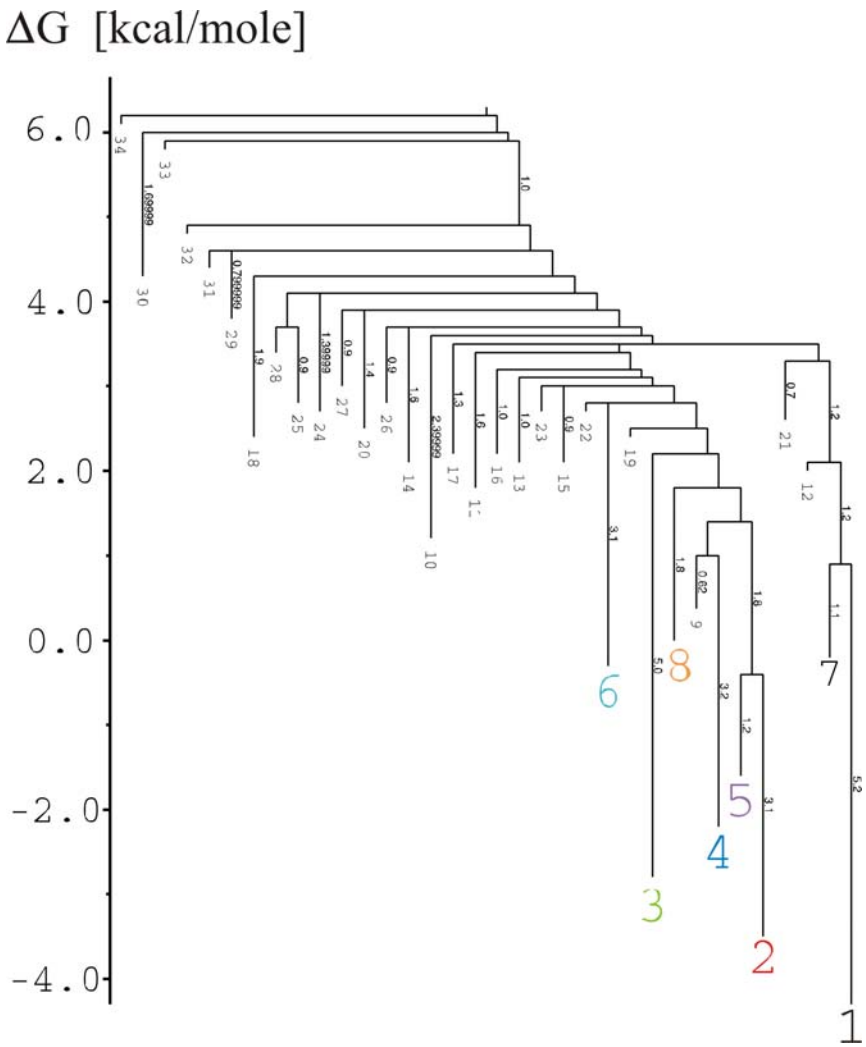
Definition of a ‚barrier tree‘

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

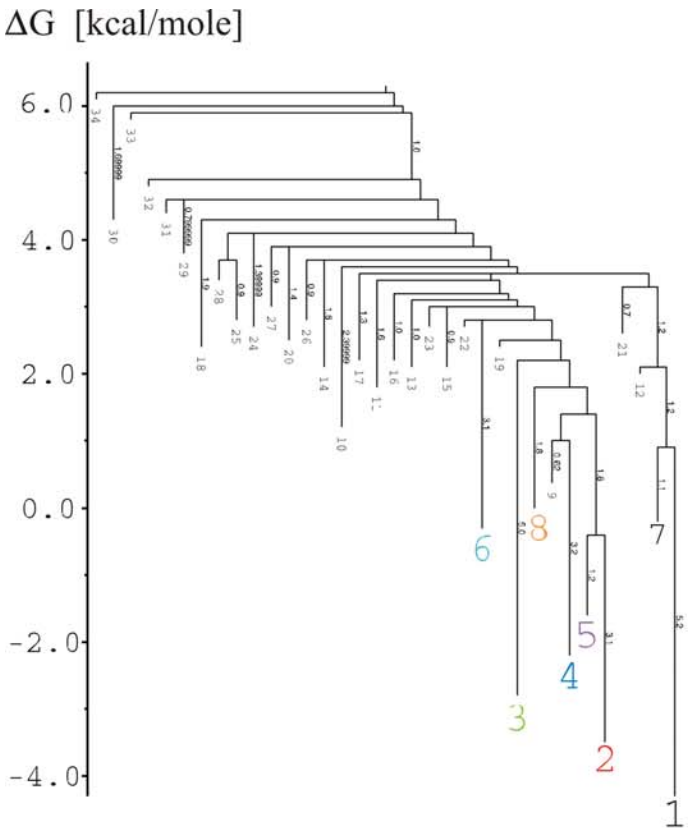


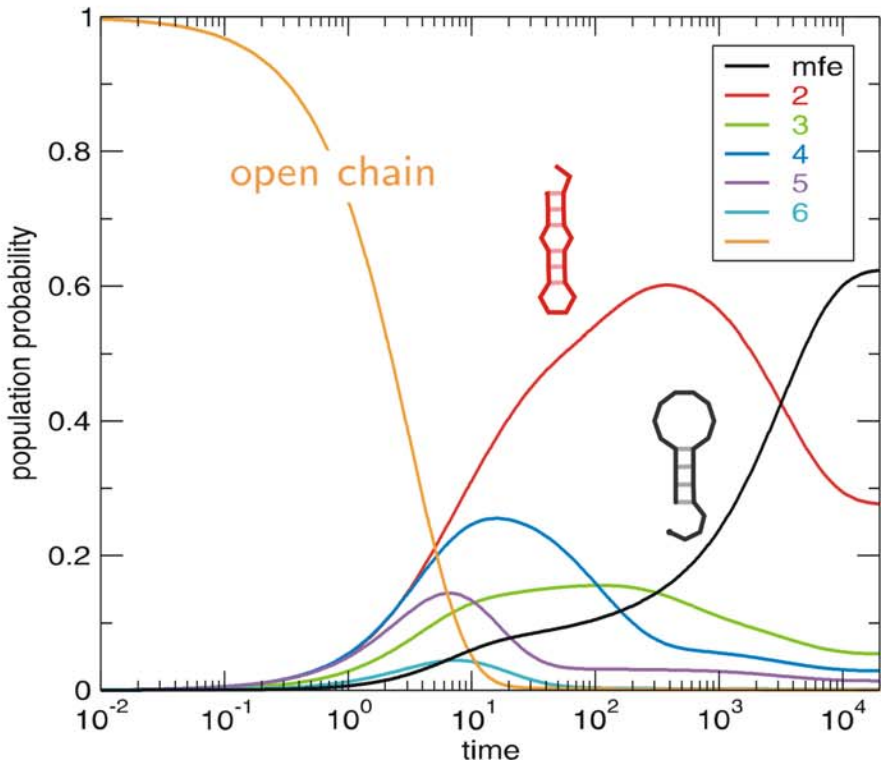
M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A: Math.Gen.* **37**:4731-4741.

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



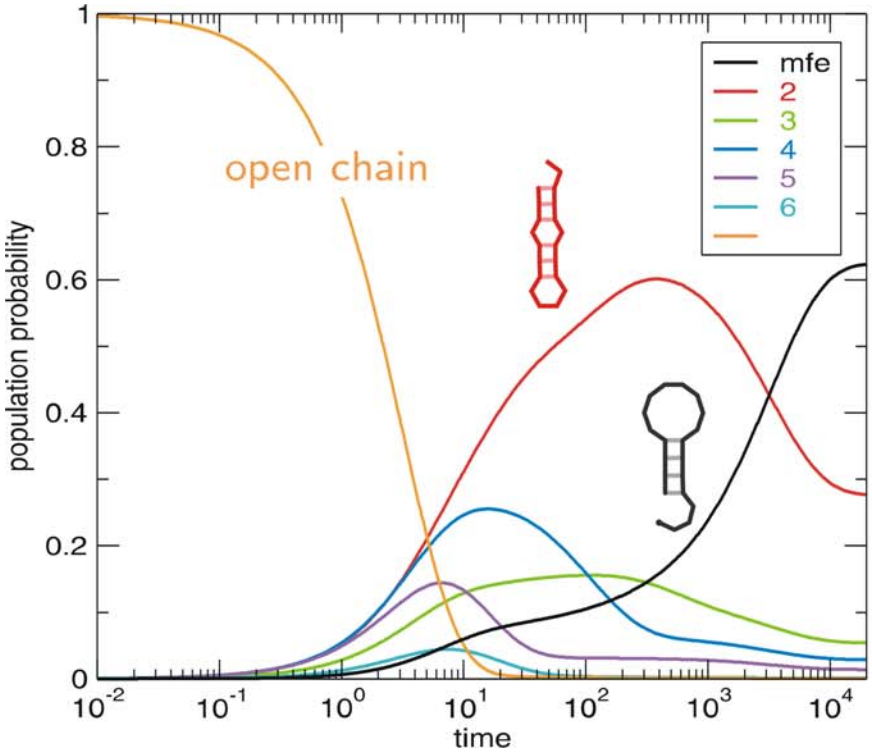
M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler. 2004. *J.Phys.A: Math.Gen.* **37**:4731-4741.



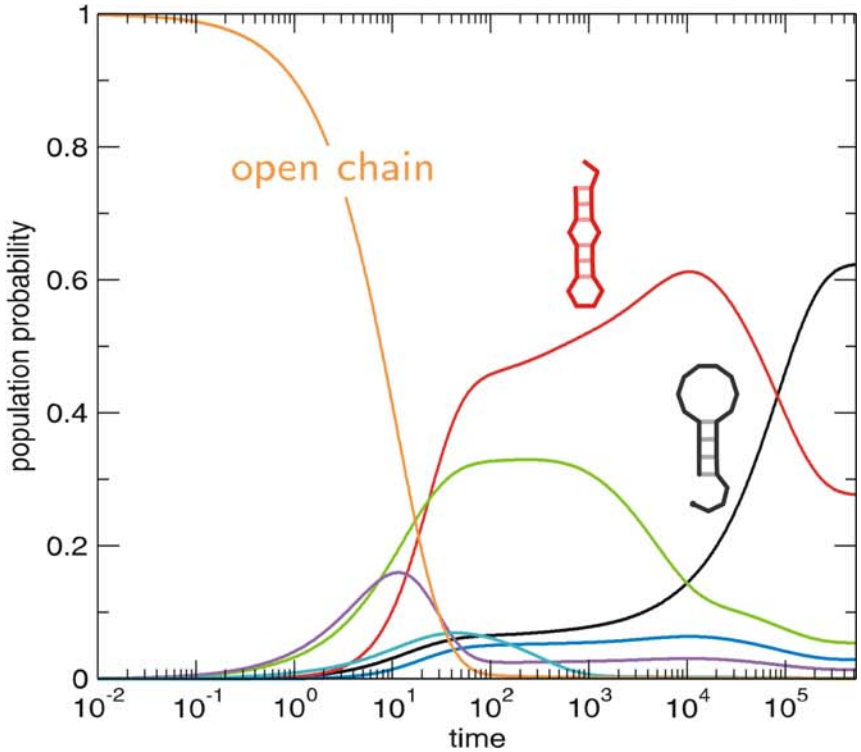


Arrhenius kinetics

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



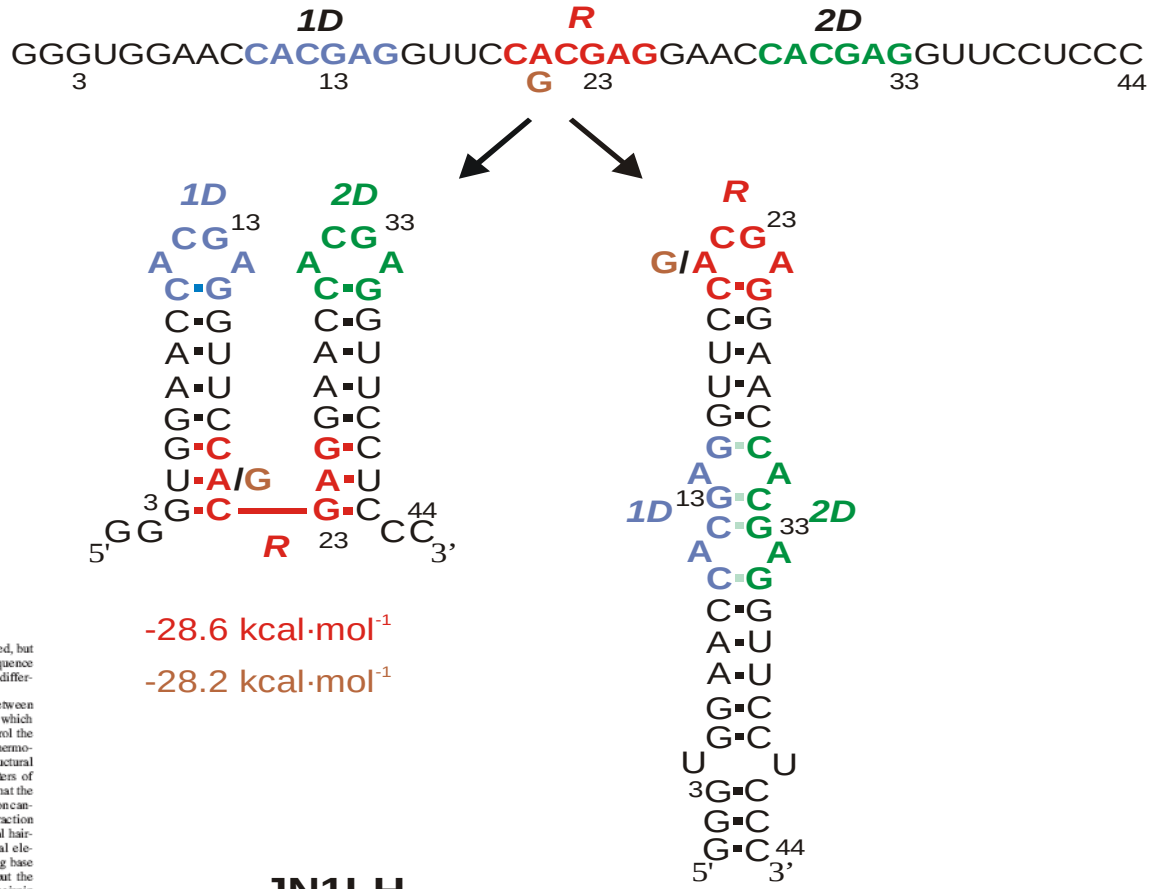
Arrhenius kinetic



Exact solution of the kinetic equation

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

Design of an RNA switch



Published online July 19, 2006

3568–3576 Nucleic Acids Research, 2006, Vol. 34, No. 12
doi:10.1093/nar/gkl445

Structural parameters affecting the kinetics of RNA hairpin formation

J. H. A. Nagel, C. Flamm[1], I. L. Hofacker[1], K. Franke[2], M. H. de Smit, P. Schuster[1] and C. W. A. Pleij*

Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, 2300 RA Leiden, The Netherlands, [1]Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, A-1090 Vienna, Austria and [2]IBA NAPS GmbH Rudolf-Wissell-Strasse 28 D-37079 Göttingen, Germany

Received January 28, 2005; Revised and Accepted June 7, 2006

ABSTRACT

There is little experimental knowledge on the sequence dependent rate of hairpin formation in RNA. We have therefore designed RNA sequences that can fold into either of two mutually exclusive hairpins and have determined the ratio of folding of the two conformations, using structure probing. This folding ratio reflects their respective folding rates. Changing one of the two loop sequences from a purine- to a pyrimidine-rich loop did increase its folding rate, which corresponds well with similar observations in DNA hairpins. However, neither changing one of the loops from a regular non-GNRA tetra-loop into a stable GNRA tetra-loop, nor increasing the loop size from 4 to 6 nt did affect the folding rate. The folding kinetics of these RNAs have also been simulated with the program 'Kinfold'. These simulations were in agreement with the experimental results if the additional stabilization energies for stable tetra-loops were not taken into account. Despite the high stability of the stable tetra-loops, they apparently do not affect folding kinetics of these RNA hairpins. These results show that it is possible to experimentally determine relative folding rates of hairpins and to use these data to improve the computer-assisted simulation of the folding kinetics of stem–loop structures.

INTRODUCTION

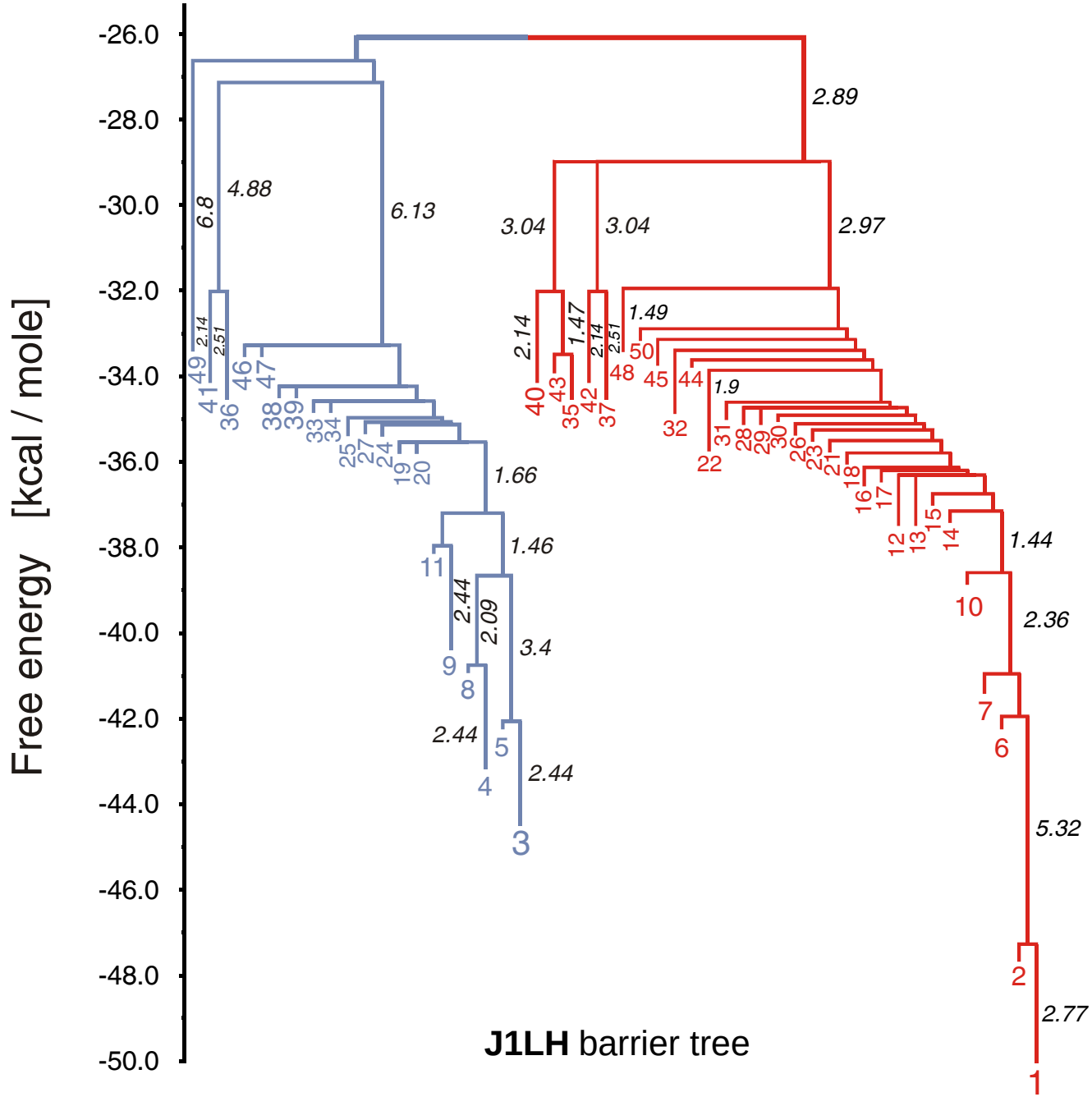
RNA chains can fold into complex secondary and tertiary structures, which often correspond to the minimum energy or equilibrium structure. Some RNAs, however, fold into long-lasting non-equilibrium conformations, which are known as metastable structures (1–8). Most of these structures are not biologically active and are thus termed misfolded (9,10). However, in a number of biological systems metastable structures exist that are actually not misfolded, but functionally important. In addition, a single RNA sequence can exhibit two catalytic activities resulting from two different structures (11).

To understand how a folding RNA chain chooses between different alternative structures it is important to know which structural, thermodynamic and kinetic parameters control the folding of the various structural elements. Today, thermodynamic parameters of most of the RNA secondary structural elements are known (12,13), whereas kinetic parameters of RNA folding are scarce (8,14–17). It has been shown that the rate-determining step of hairpin formation is dependent on cancellation of the positive loop energy by the stacking interaction between the first closing base pairs (16,18) and that local hairpin formation is favoured over long-distance structural elements, because of the spatial proximity of the opposing base pairing partners (1,15). Little is known, however, about the effects of the nucleotide sequence and the size of hairpin loops and of the nature of the closing base pairs on folding kinetics. Even less is known about the effects of bulges, internal loops and other secondary structural elements.

Despite this lack of quantitative knowledge, great progress has been made in predicting folding routes of RNA using computer simulations, based on existing thermodynamic parameters and statistical polymer physics (2,4,19–25). These predictions, however, have rarely been verified experimentally. As a result it is still difficult to estimate which of the potential hairpins in a given RNA sequence will fold predominantly and which are kinetically disfavoured. Therefore, the prediction of a correct metastable structure in a given RNA molecule, even if it is suspected to have kinetically favourable metastable hairpins, has not always been straightforward (4,6,26) (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. K. Schuster, K. Gerdes and C. W. A. Pleij, manuscript submitted).

To determine kinetic parameters experimentally, we have developed an approach in which the kinetic folding ratios of two mutually exclusive hairpins in a given RNA sequence can be measured by structure probing. Although, this

*To whom correspondence should be addressed. Tel: +31-71-5274769; Fax: +31-71-5274340; Email: cpley@chem.leidenuniv.nl



J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij. *Nucleic Acids Res.* **34**:3568-3576 (2006)

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*).

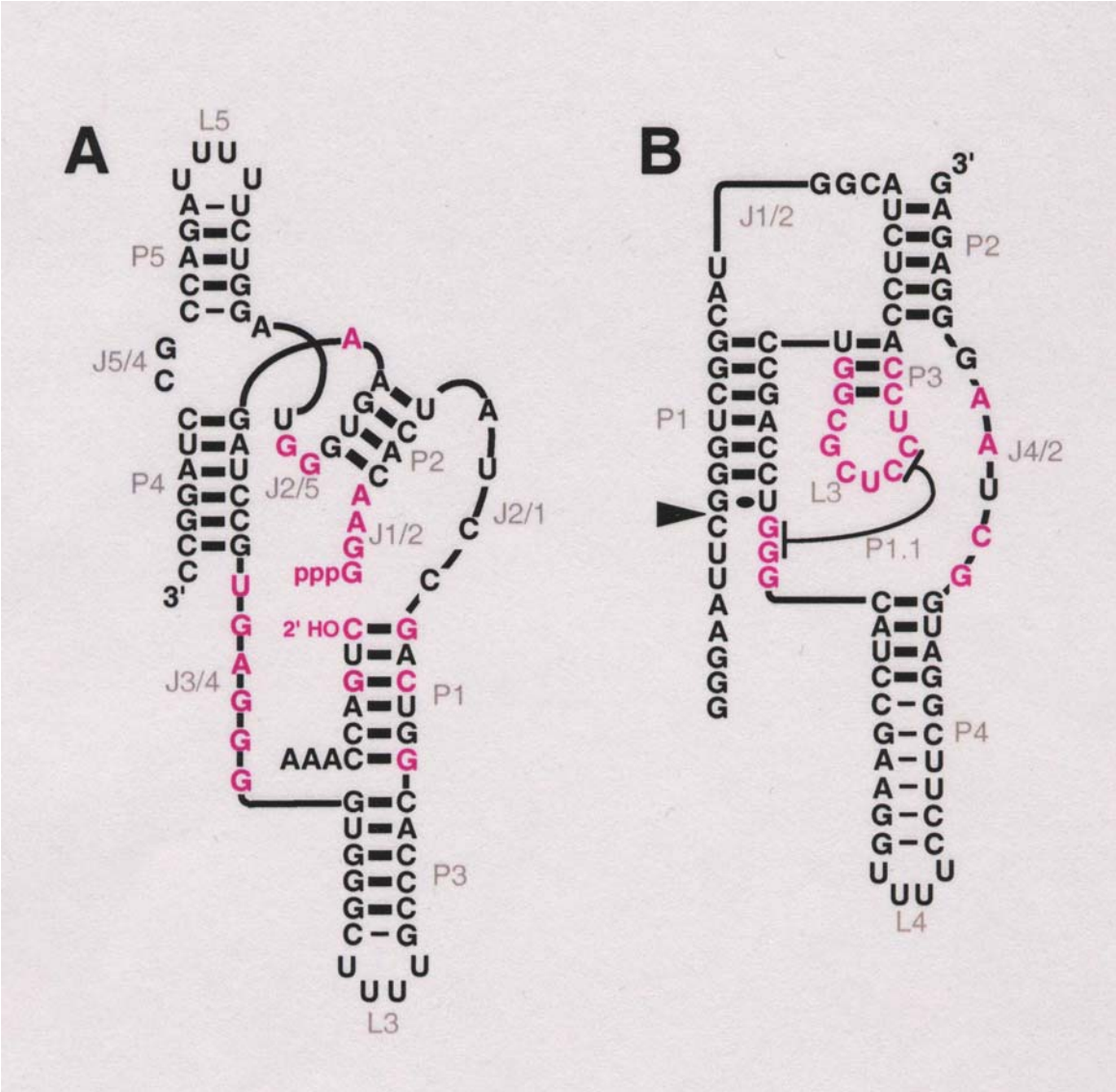
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

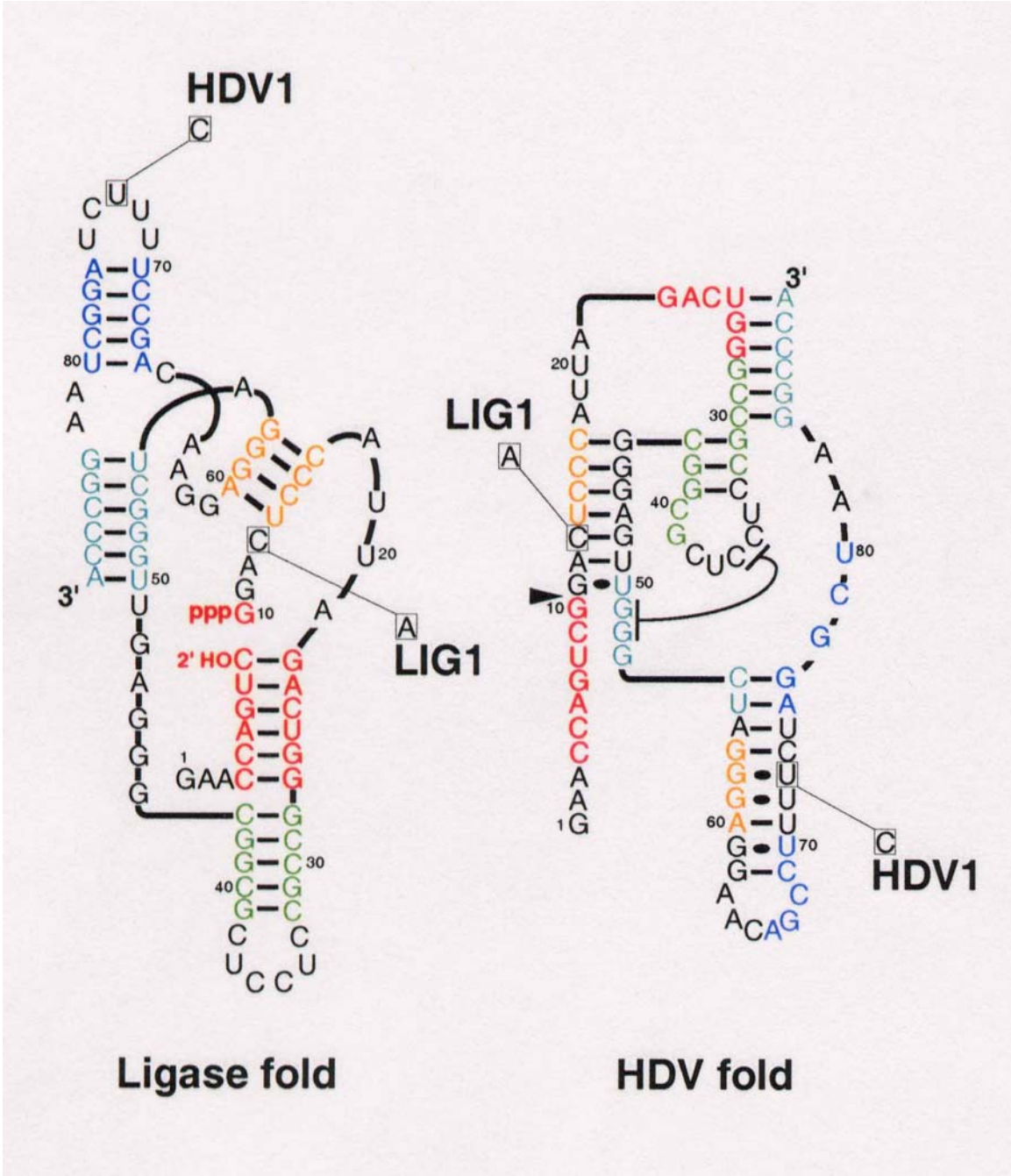
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

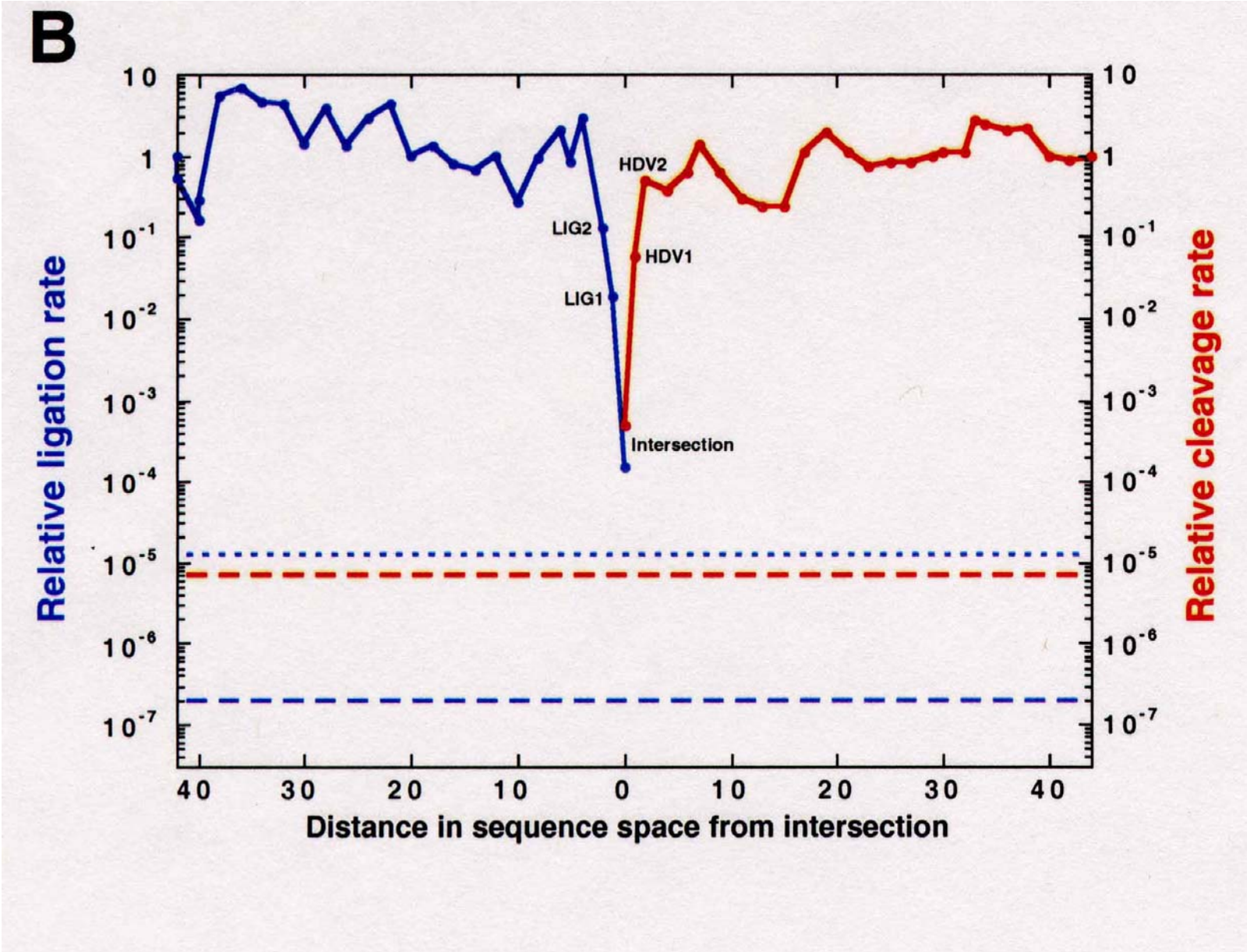


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

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

Acknowledgement of support

Universität Wien

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

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

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

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

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

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, MA

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

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

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

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

Stefan Bernhart, **Jan Cupal**, **Lukas Endler, Kurt Grünberger**,
Michael Kospach, **Ulrike Langhammer**, **Rainer Machne, Ulrike Mückstein**,
Hakim Tafer, Andreas Wernitznig, **Stefanie Widder, Michael Wolfinger,**
Stefan Wuchty, Dilmurat Yusuf, 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/~pks