Diversity and Plasticity of RNA

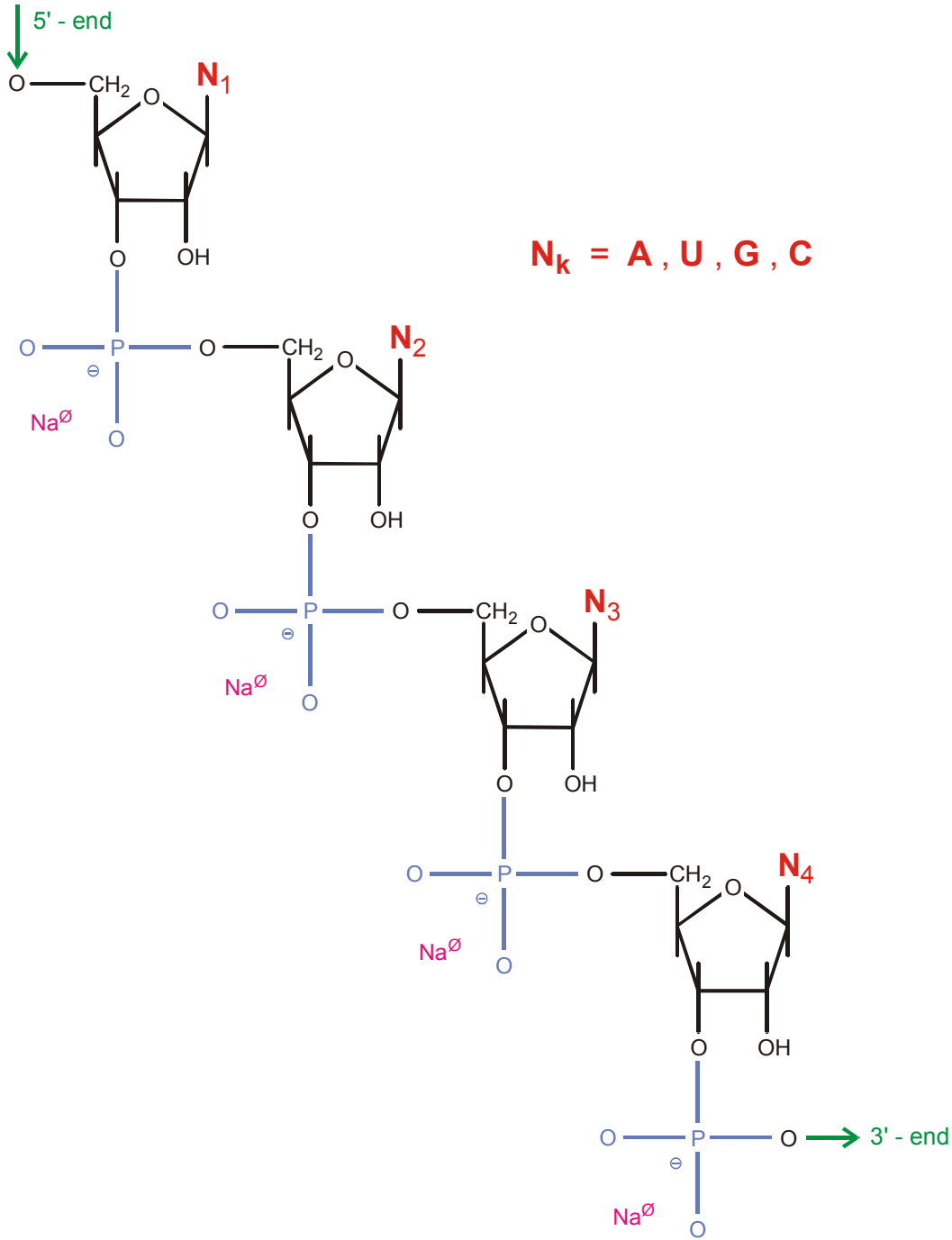
Beyond the One-Sequence-One-Structure Paradigm

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Chemistry towards Biology

Portorož, 8.– 12.09.2002



$N_k = A, U, G, C$

The chemical formula of RNA consisting of nucleobases, ribose rings, phosphate groups, and sodium counterions

Magnesium ions play a special role and act as coordination centers which are indispensible for the formation of full three-dimensional structures

5'-End 3'-End

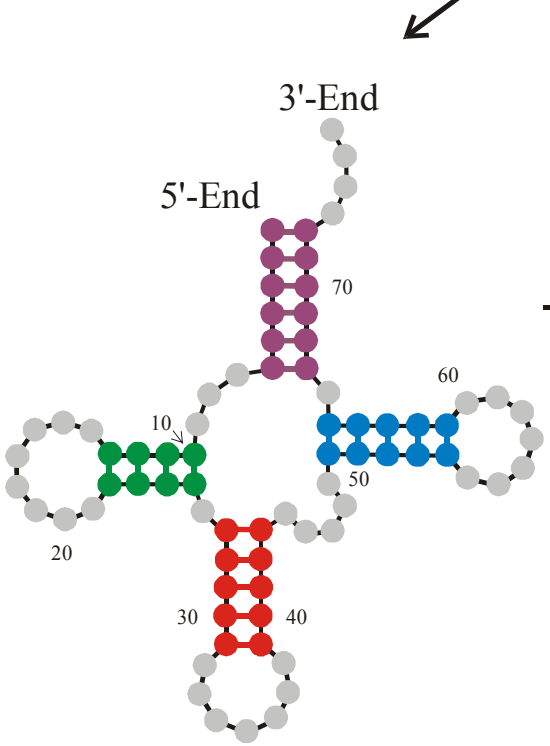
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

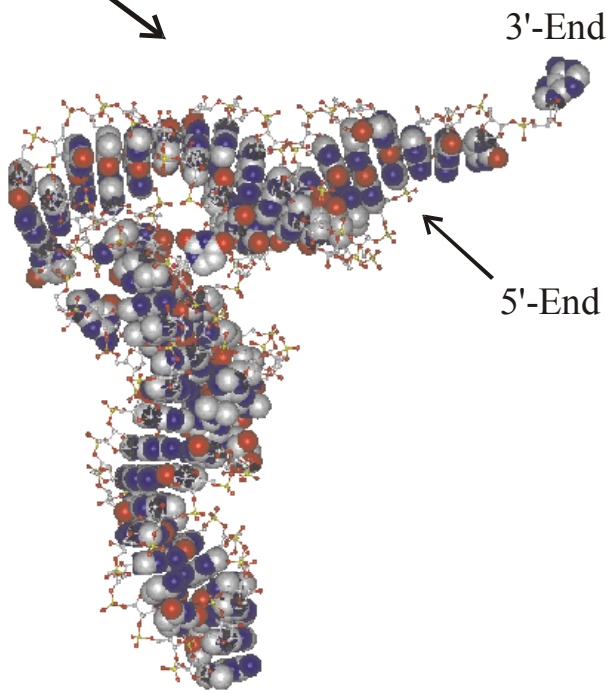
Biochemical and chemical probing

Structure prediction

Crystallography

NMR, FRET, ...





The one sequence – one structure paradigm

One day, when biomolecular structures were understood in sufficient detail, we would be able to design molecules with predefined structures and for ***a priori*** given purposes.

Biomolecular structures are not fully understood yet, but the lack of knowledge in structure and function can be compensated by applying selection methods.

$4^{27} = 1.801 \text{ Ç } 10^{16}$ possible different sequences

Combinatorial diversity of sequences: $N = 4^{\{}$

A = adenylate
U = uridylate
C = cytidylate
G = guanylate

Number of (different) sequences created by common scale random synthesis:

$10^{15} – 10^{16}$.

Combinatorial diversity of heteropolymers illustrated by means of an RNA aptamer that binds to the antibiotic tobramycin

Taming of sequence diversity through selection and evolutionary design of RNA molecules

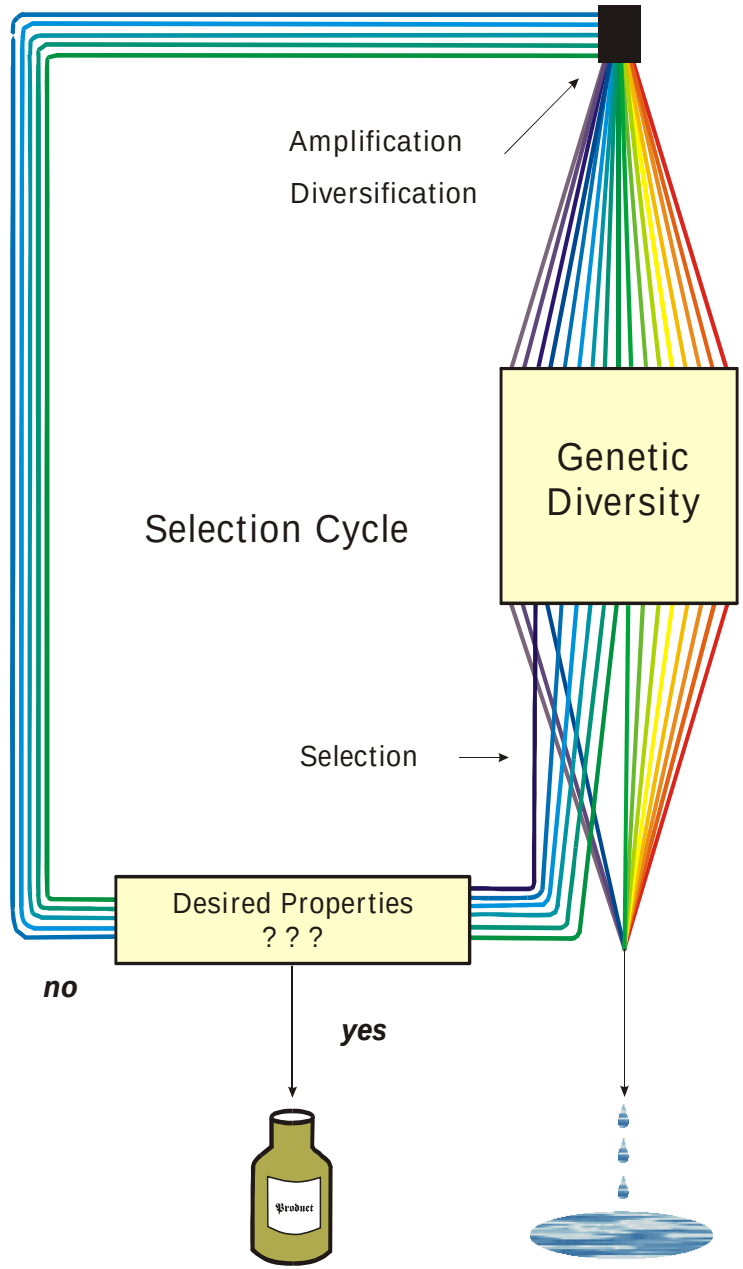
D.B.Bartel, J.W.Szostak, **In vitro** ***selection of RNA molecules that bind specific ligands***. Nature **346** (1990), 818-822

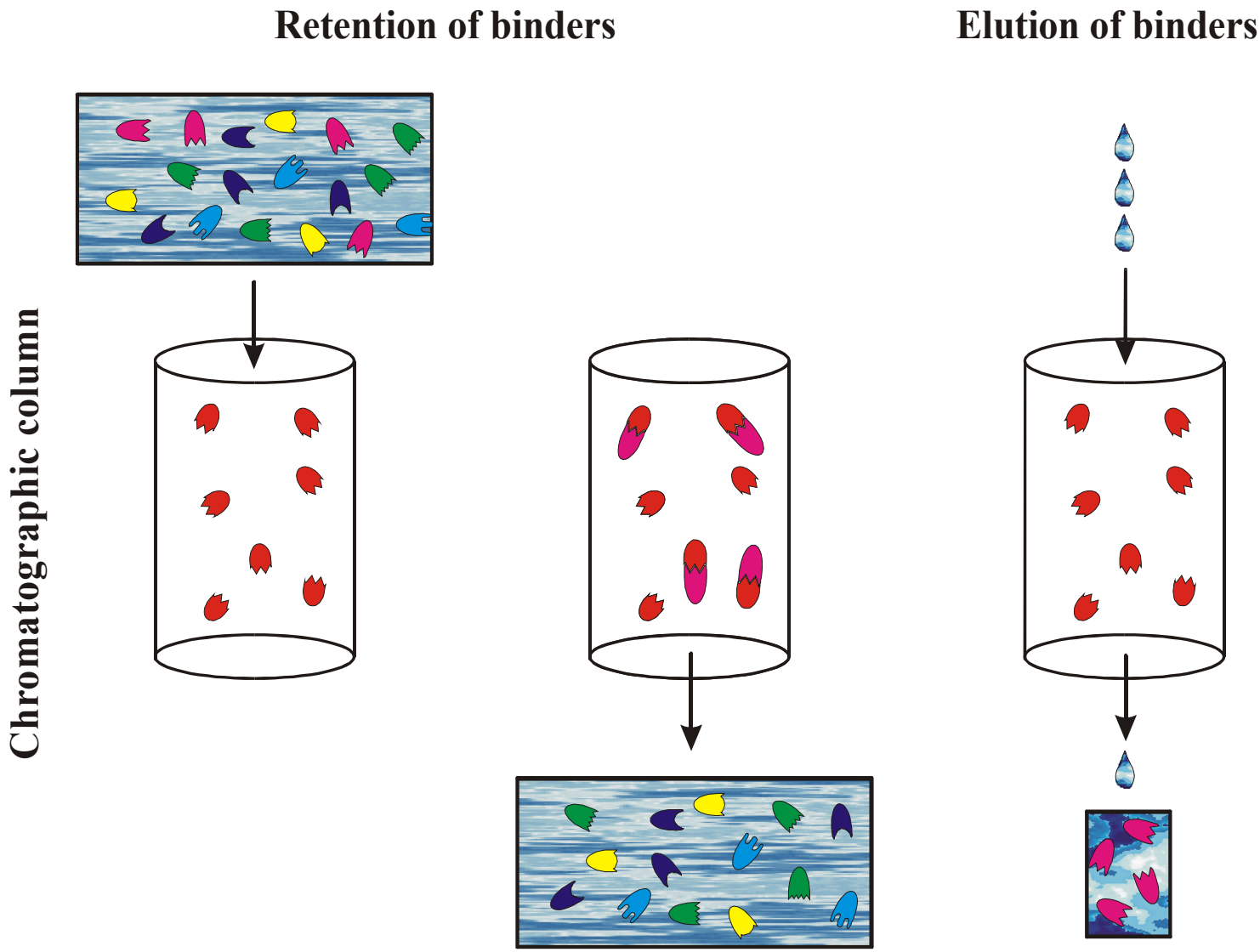
C.Tuerk, L.Gold, **SELEX** - ***Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage*** **T4** ***DNA polymerase***. Science **249** (1990), 505-510

D.P.Bartel, J.W.Szostak, ***Isolation of new ribozymes from a large pool of random sequences***. Science **261** (1993), 1411-1418

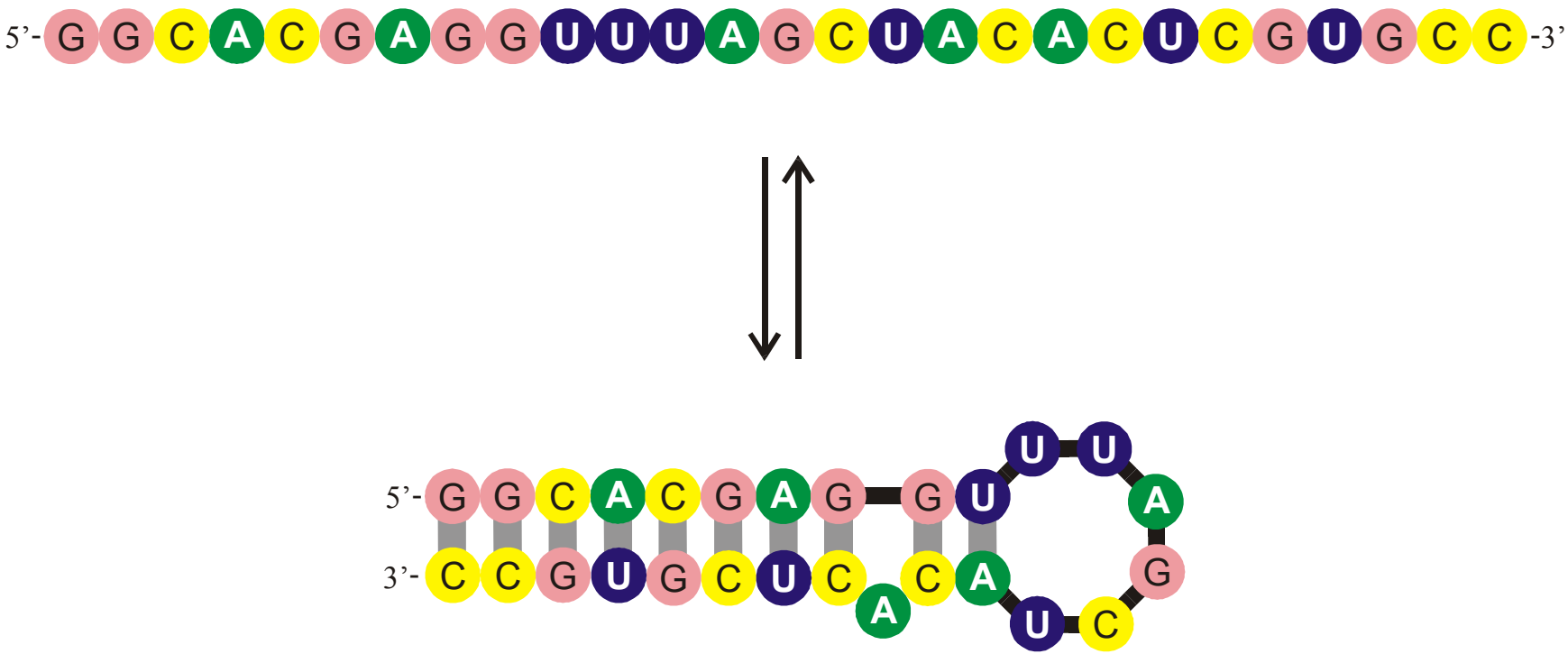
R.D.Jenison, S.C.Gill, A.Pardi, B.Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429

Selection cycle used in applied molecular evolution to design molecules with predefined properties



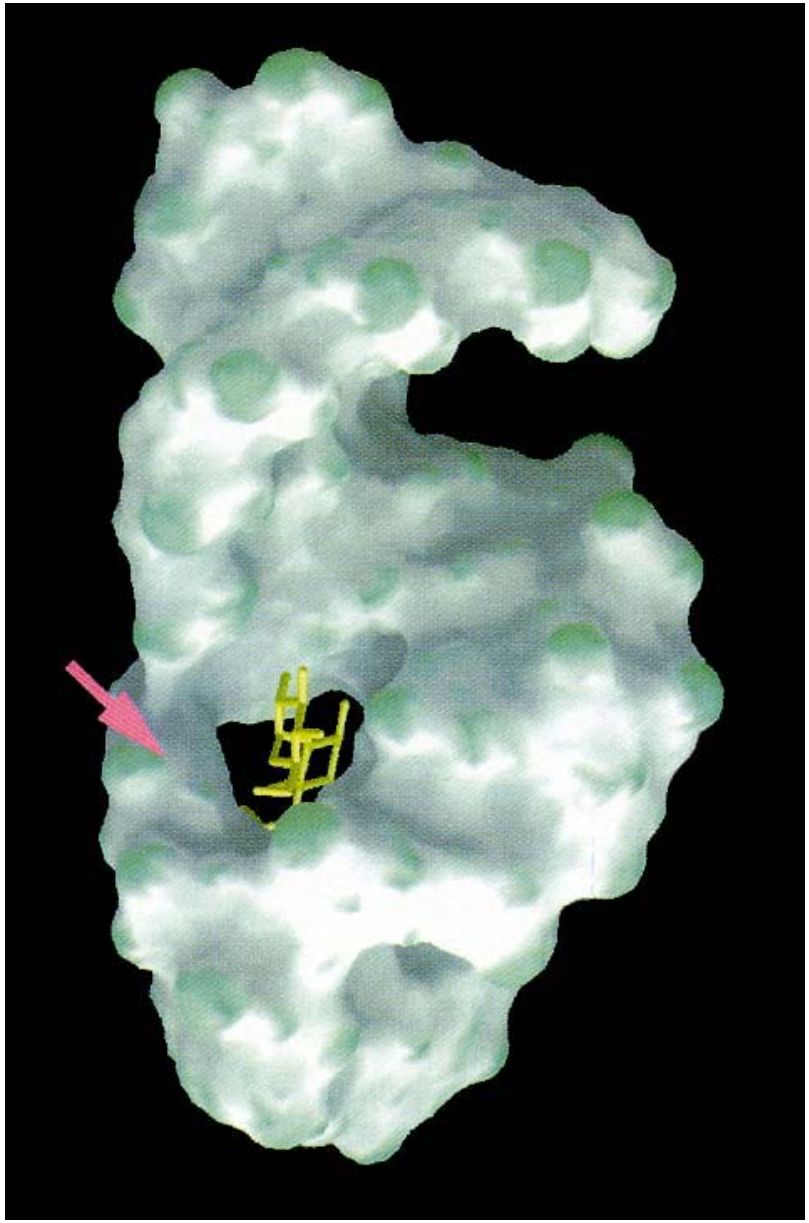


The SELEX technique for the evolutionary design of *aptamers*



Formation of secondary structure of the tobramycin binding RNA aptamer

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, Chemistry & Biology **4**:35-50 (1997)

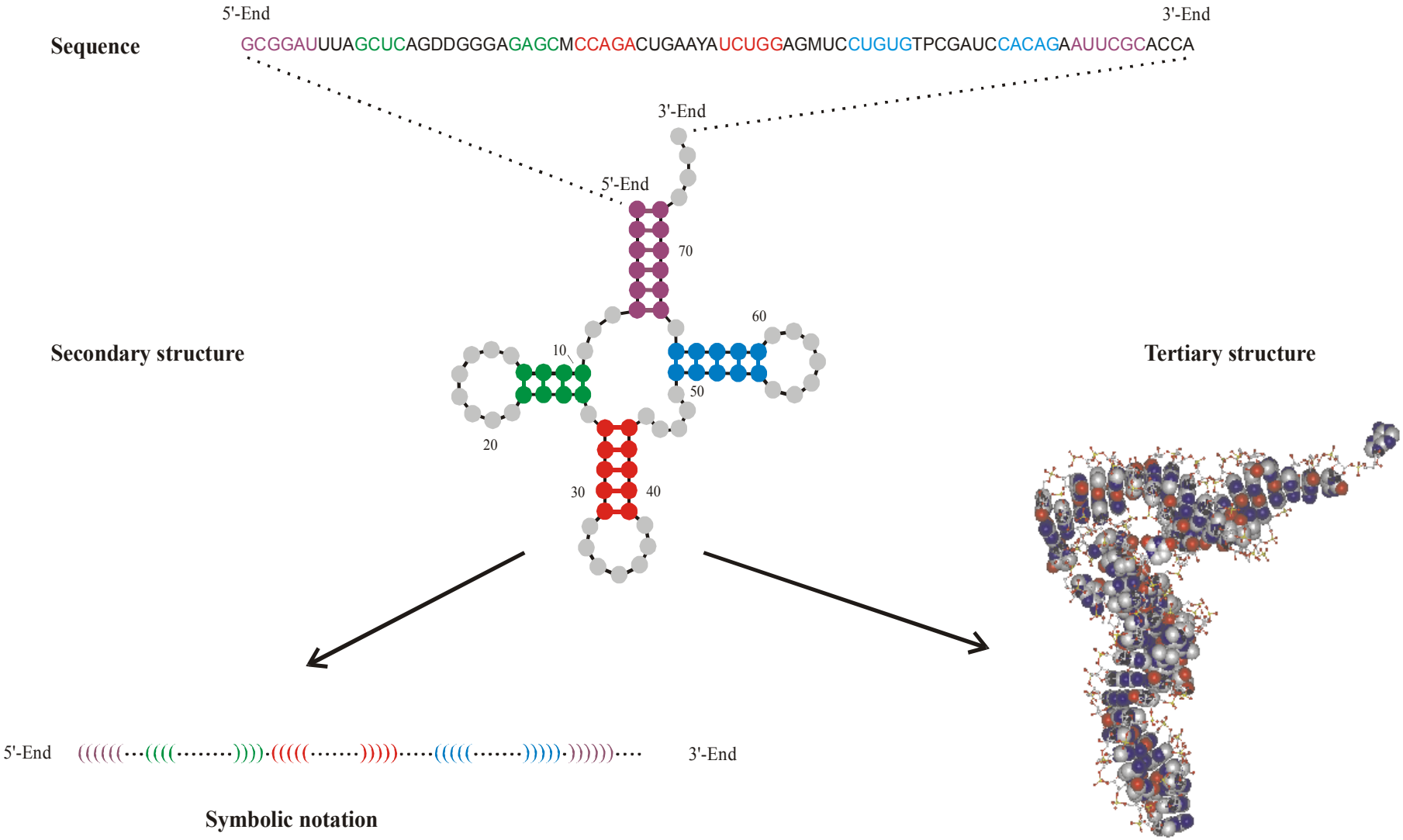


The three-dimensional structure of the tobramycin aptamer complex

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, Chemistry & Biology **4**:35-50 (1997)

Mapping RNA sequences onto RNA structures

The attempt to investigate this mapping is understood as a search for the relations between all possible 4^n sequences and all thermodynamically stable structures, which are the structures of minimal free energy. Sequence-structure mappings of RNA molecules were studied by a variety of different experimental and ***in silico*** techniques.



What is an RNA structure?

The secondary structure is a listing of base pairs, and it is understood in contrast to the full 3D-structure dealing with atomic coordinates. An intermediate state of structural details is provided by RNA threading or other toy models.

RNA Secondary Structures and their Properties

RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots. Secondary structures are folding intermediates in the formation of full three-dimensional structures.

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

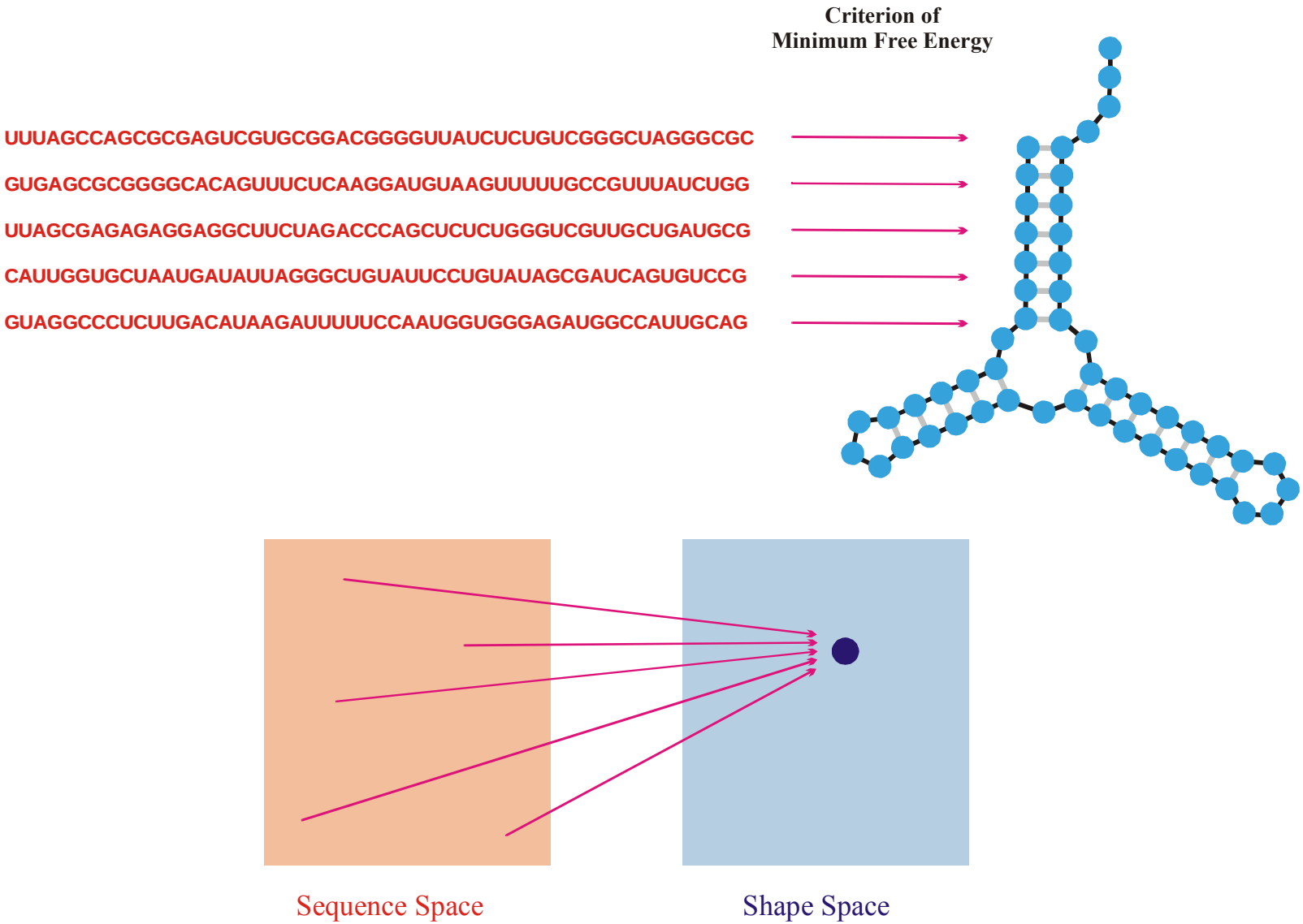
RNA Minimum Free Energy Structures

Efficient algorithms based on dynamical programming are available for computation of secondary structures for given sequences. Inverse folding algorithms compute sequences for given secondary structures.

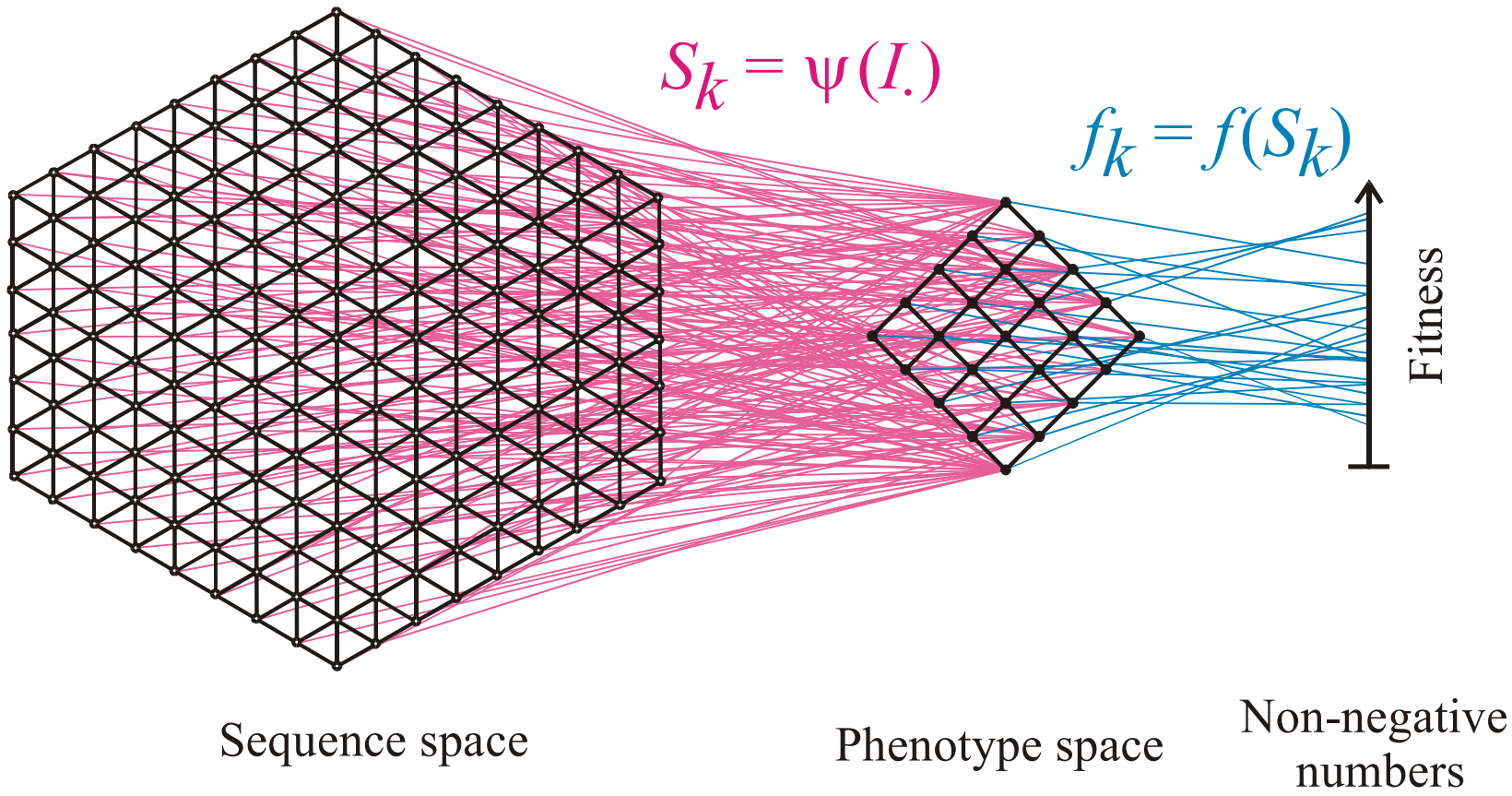
M.Zuker and P.Stiegler. *Nucleic Acids Res*. **9**:133-148 (1981)

Vienna RNA Package: http:www.tbi.univie.ac.at (includes inverse folding, suboptimal structures, kinetic folding, etc.)

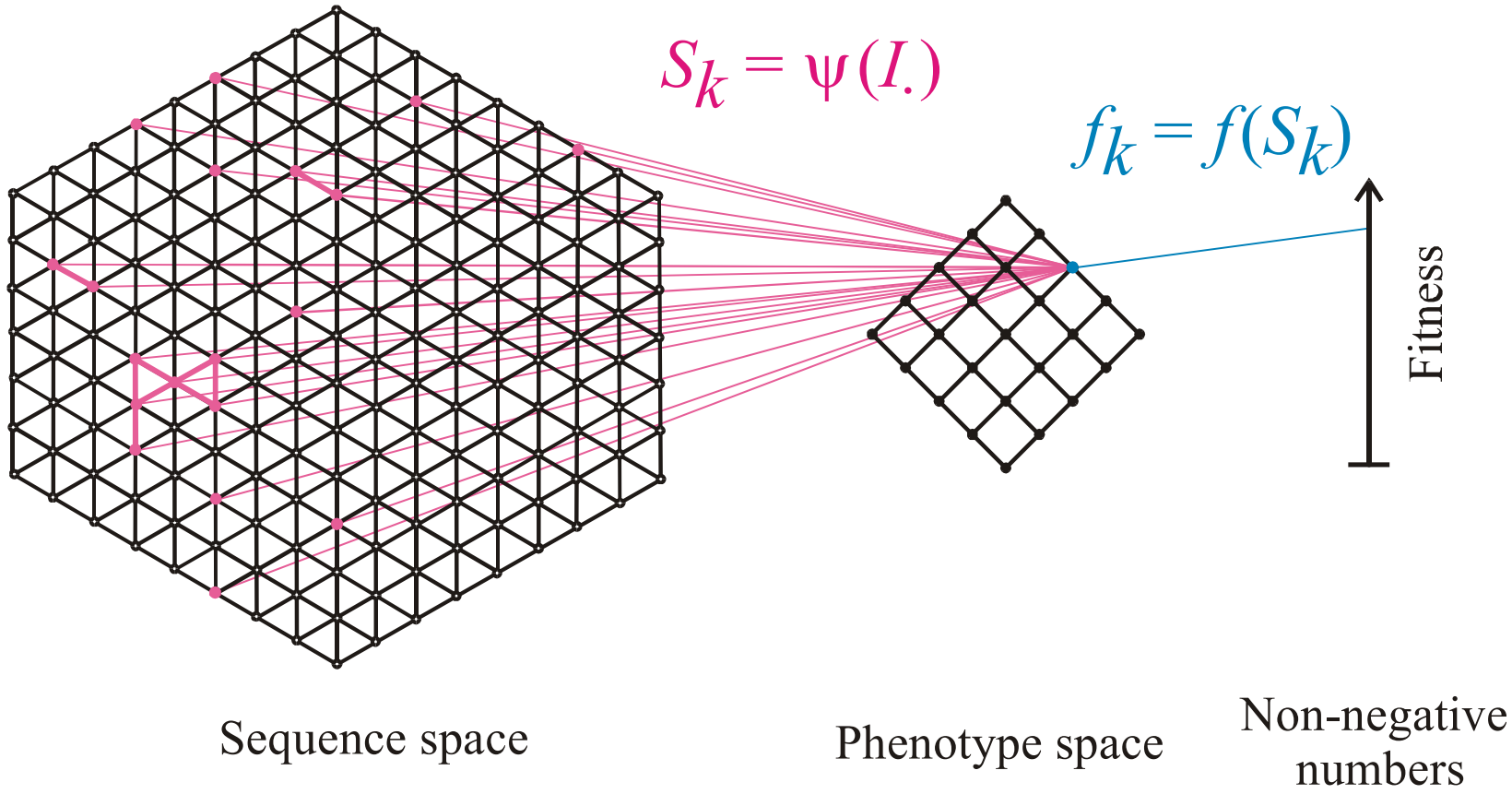
I.L.Hofacker, W. Fontana, P.F.Stadler, L.S.Bonhoeffer, M.Tacker, and P. Schuster. *Mh.Chem*. **125**:167-188 (1994)

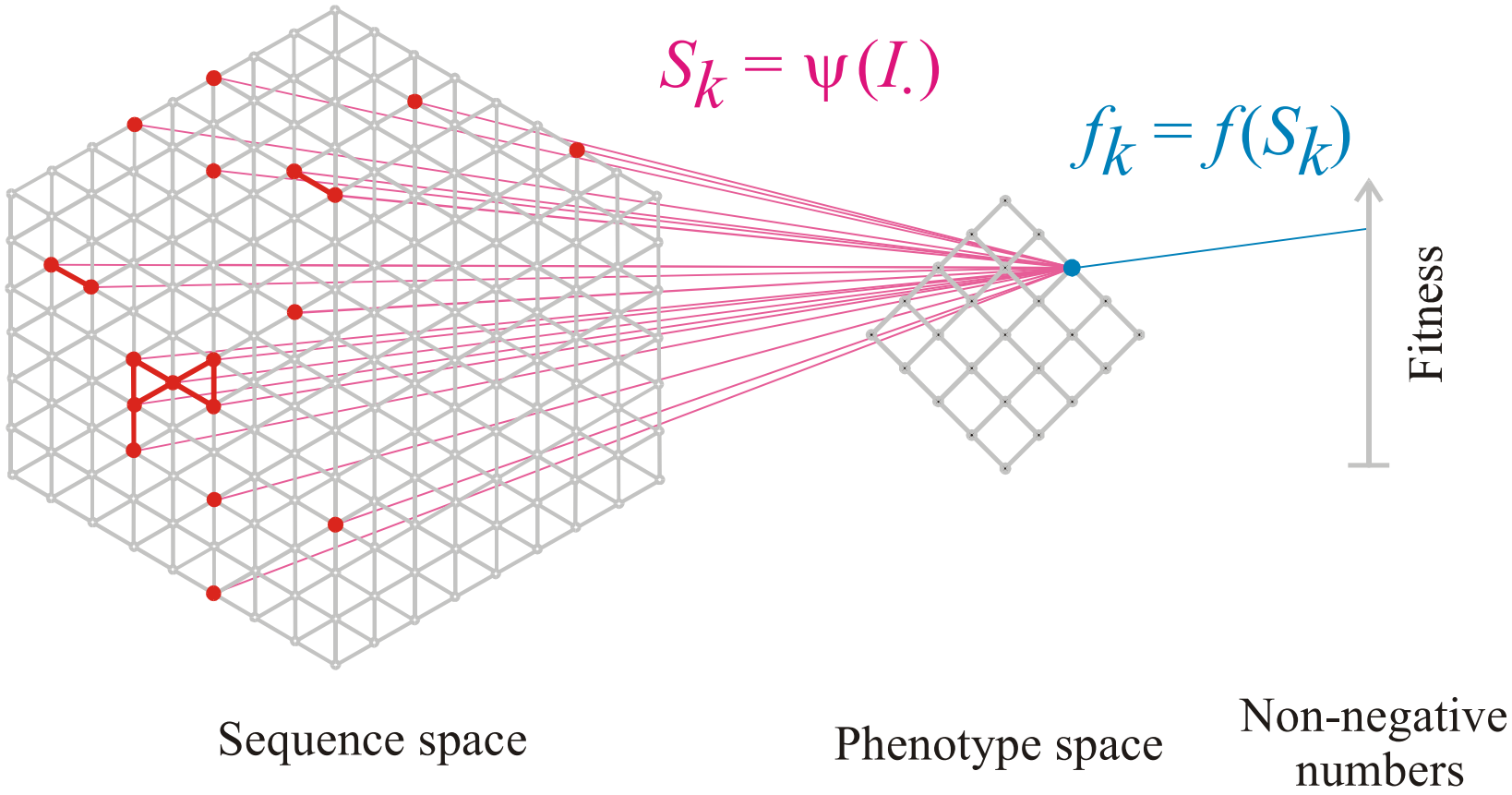


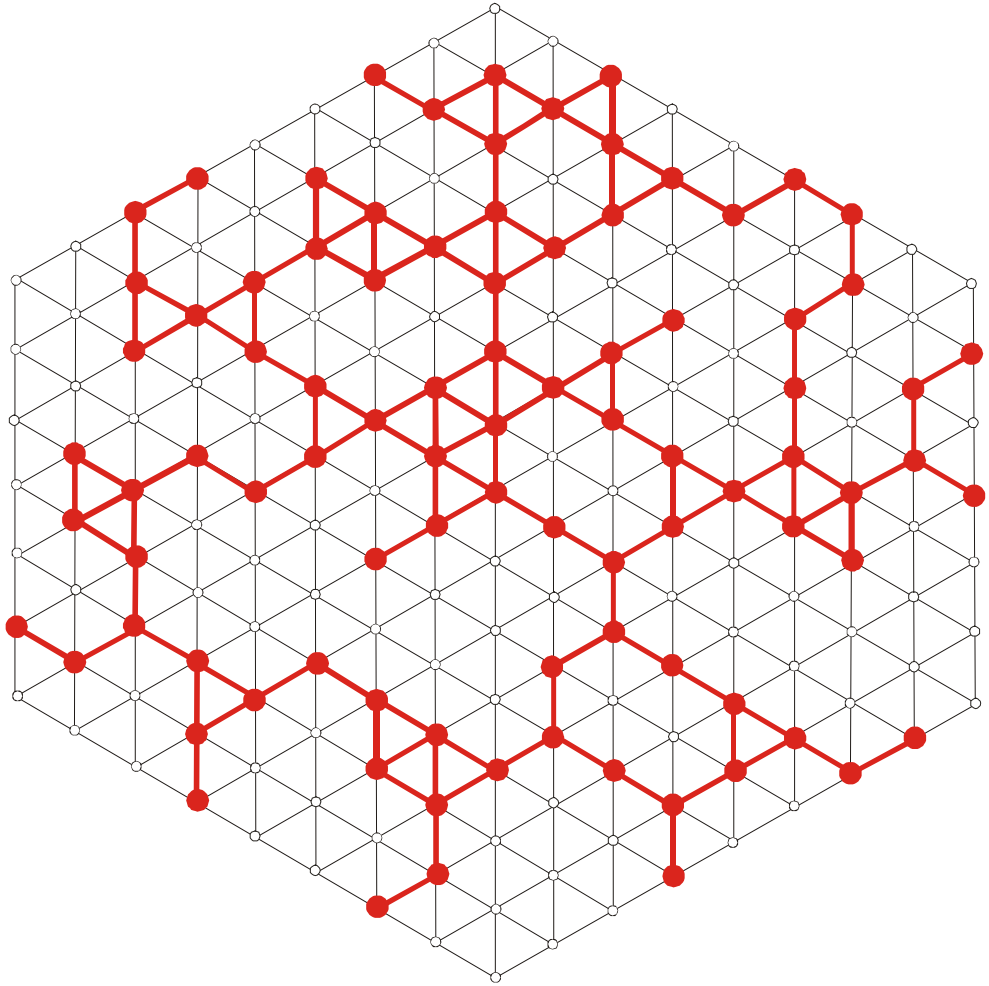
Many sequences from the same minimum free energy secondary structure



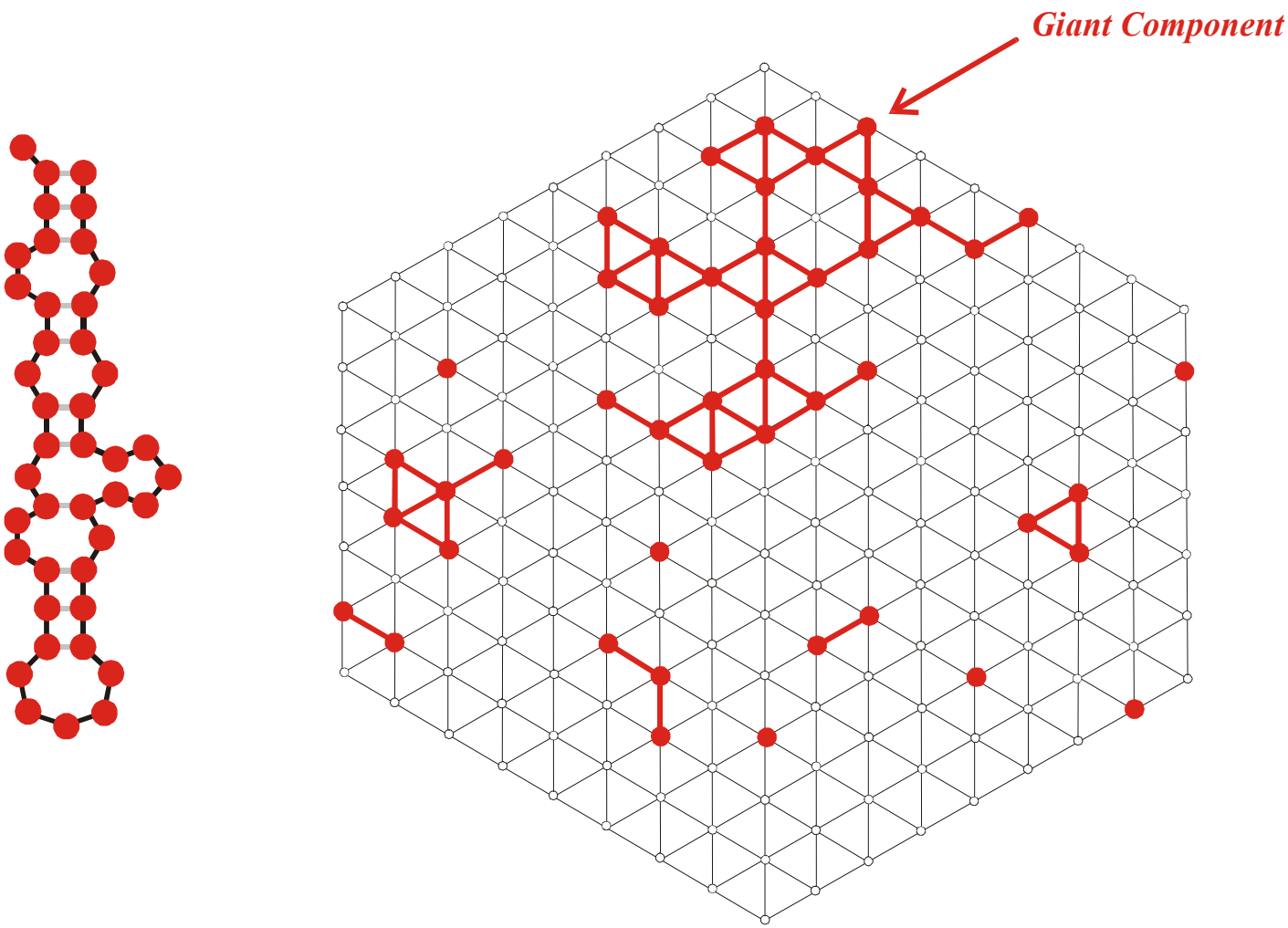
Mapping from sequence space into phenotype space and into fitness values


$S_k = \psi(I_\cdot)$
$f_k = f(S_k)$
Fitness
Sequence space
Phenotype space
Non-negative numbers


$S_k = \psi(I_\cdot)$
$f_k = f(S_k)$
Fitness
Sequence space
Phenotype space
Non-negative numbers

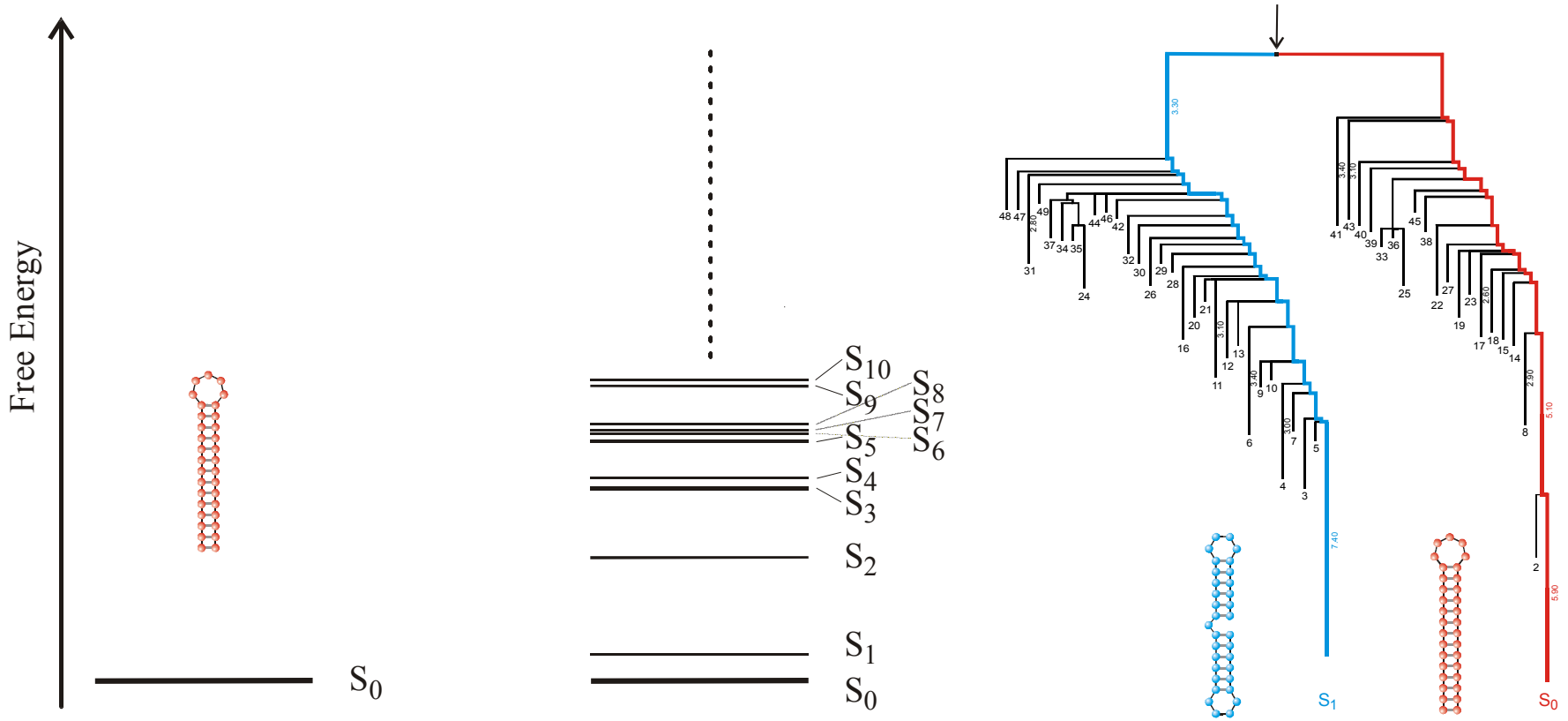


A connected neutral network



A multi-component neutral network

T = 0 K , t Á ¸ T > 0 K , t Á ¸ T > 0 K , t finite



Minimum Free Energy Structure Suboptimal Structures Kinetic Structures

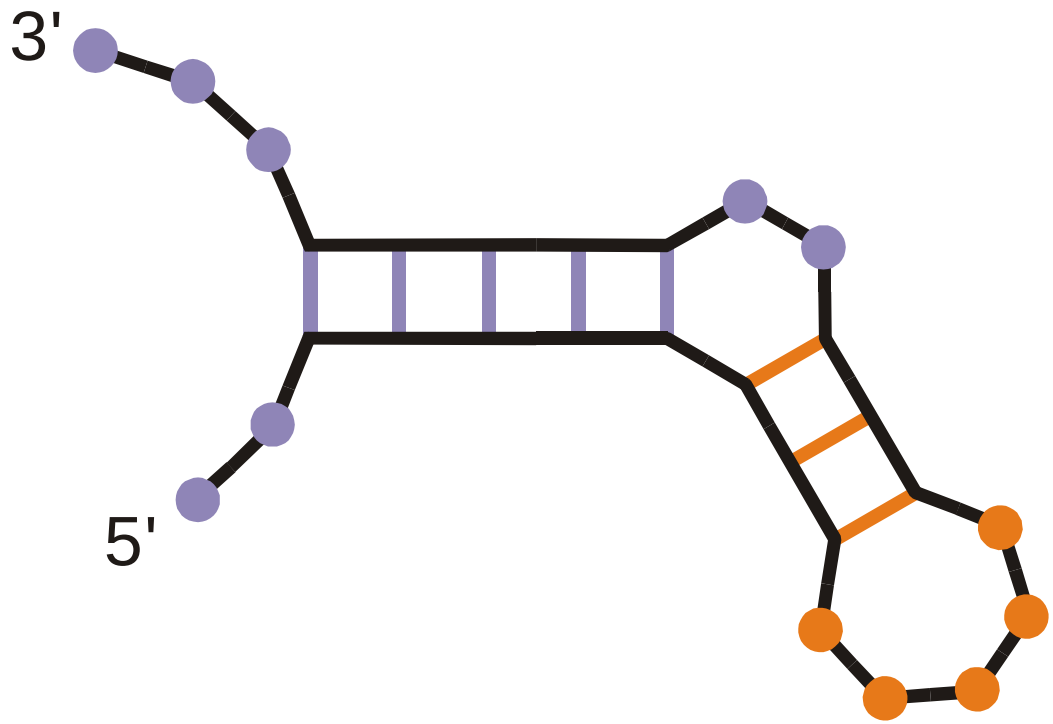
Different notions of RNA structure including suboptimal conformations

Partition Function of RNA Secondary Structures

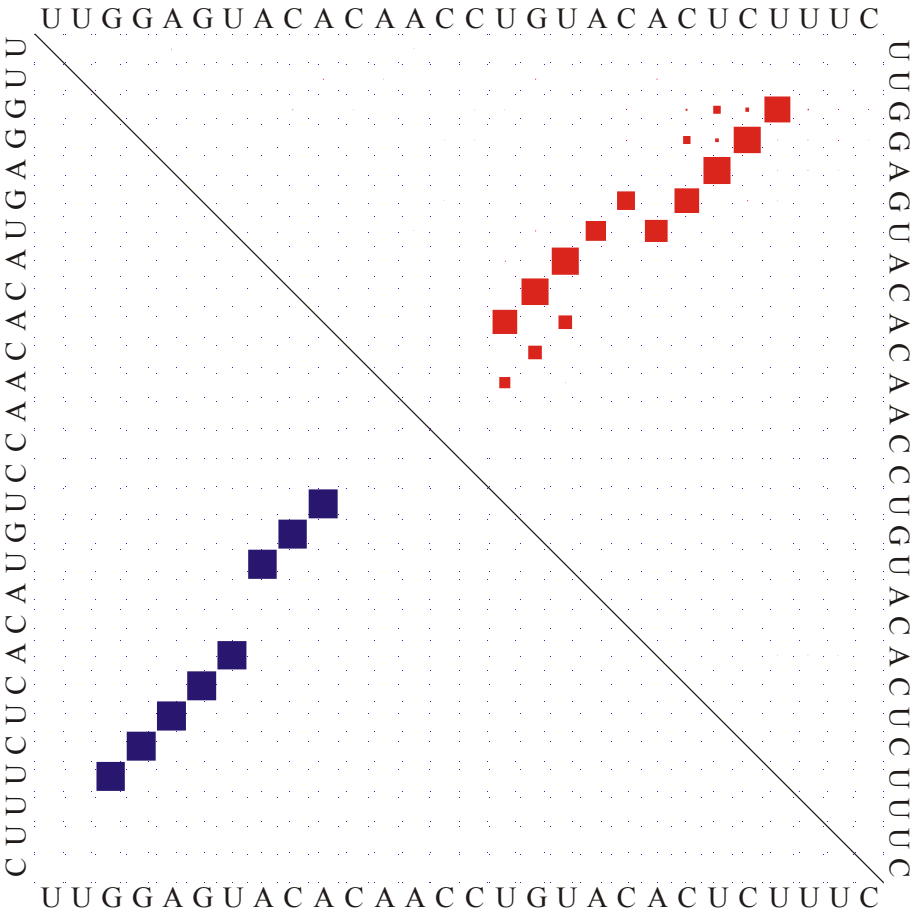
John S. McCaskill. *The equilibrium function and base pair binding probabilities for RNA secondary structure*. Biopolymers **29** (1990), 1105-1119

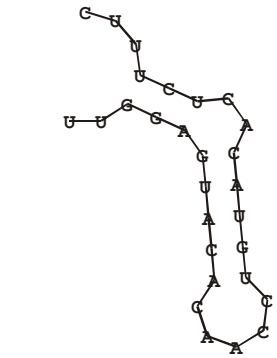
Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster. *Fast folding and comparison of RNA secondary structures.* Monatshefte für Chemie **125** (1994), 167-188

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



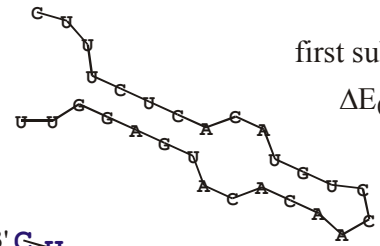
Example of a small RNA molecule: n=28





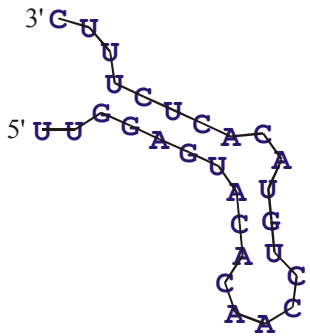
second suboptimal configuration

$\Delta E_{0\rightarrow 2}$ = 0.55 kcal / mole



first suboptimal configuration

$\Delta E_{0\rightarrow 1}$ = 0.50 kcal / mole



minimum free energy configuration

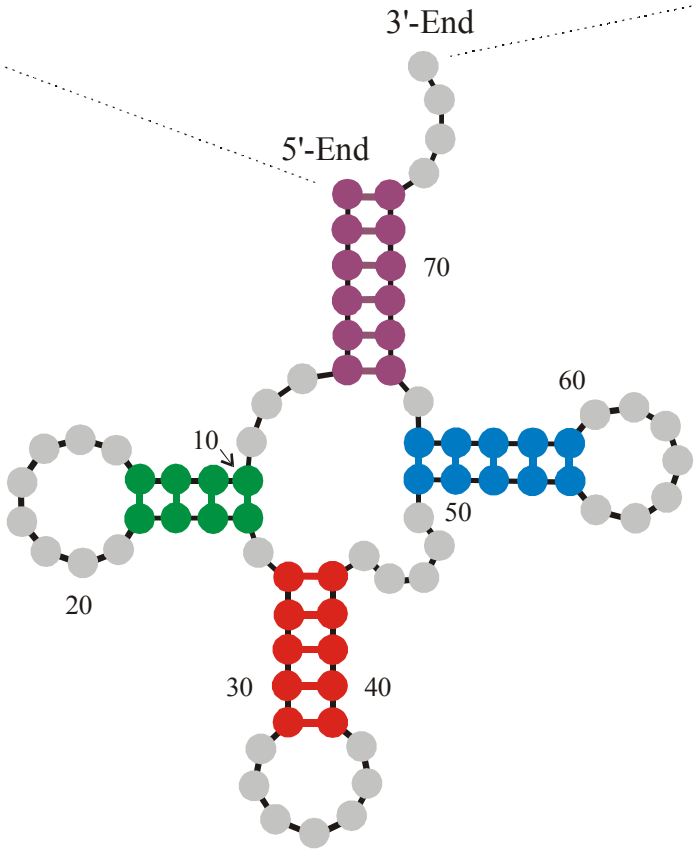
DG_0 = - 5.39 kcal / mole

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

Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

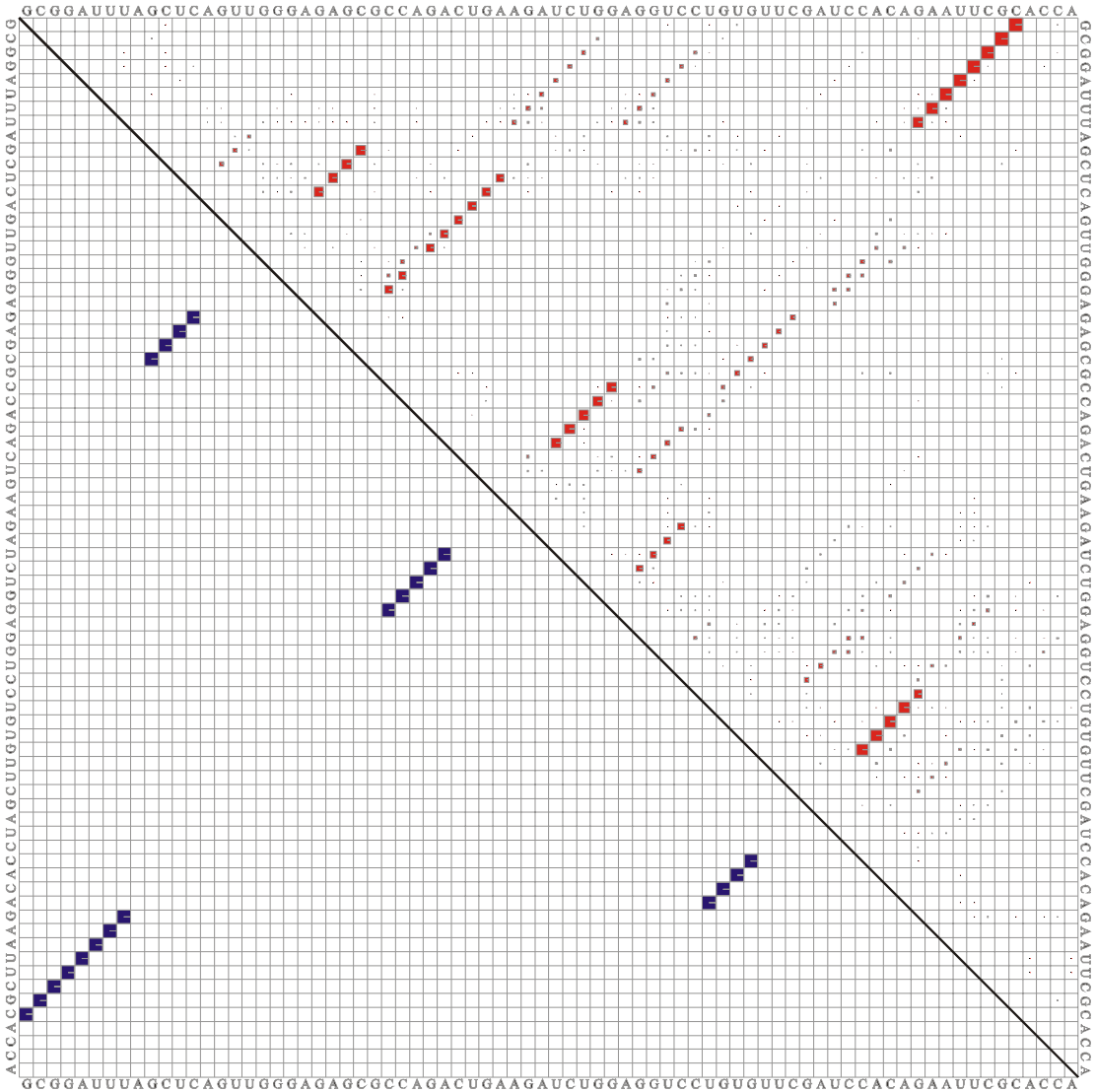
Secondary Structure

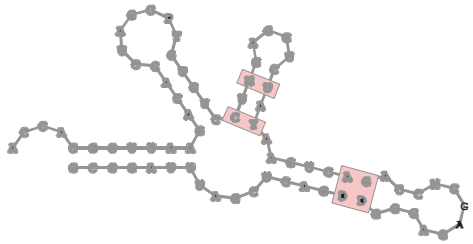


Symbolic Notation

5'-End ((((((···((((·····))))·((((·····))))).....((((·····))))).))))···· 3'-End

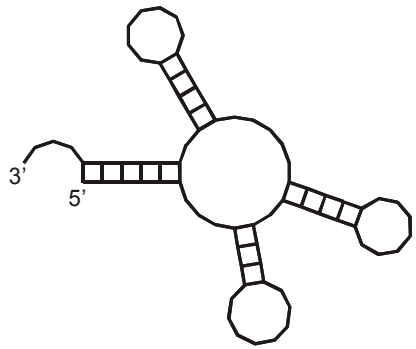
Phenylalanyl-tRNA as an example for the computation of the partition function





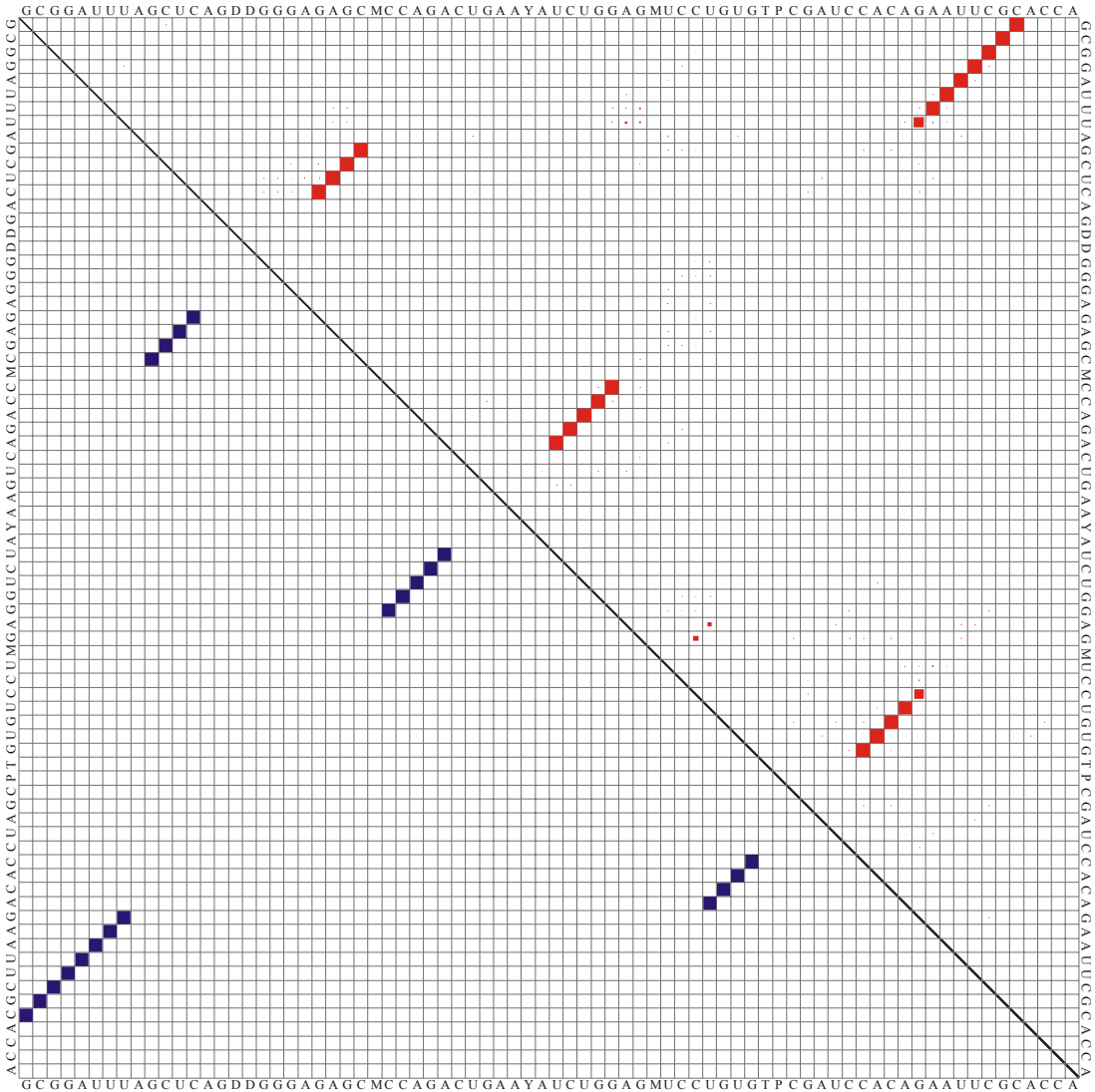
first suboptimal configuration

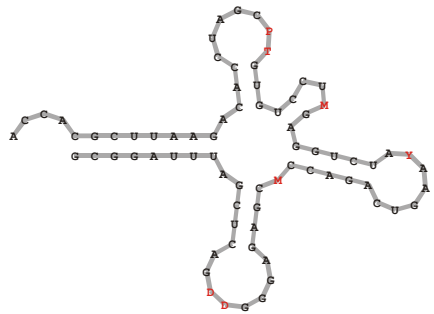
$\Delta E_{0\rightarrow 1}$ = 0.43 kcal / mole



$tRNA^{phe}$

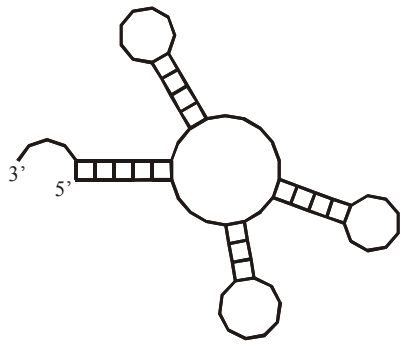
without modified bases





first suboptimal configuration

$\Delta E_{0\rightarrow 1}$ = 0.94 kcal / mole



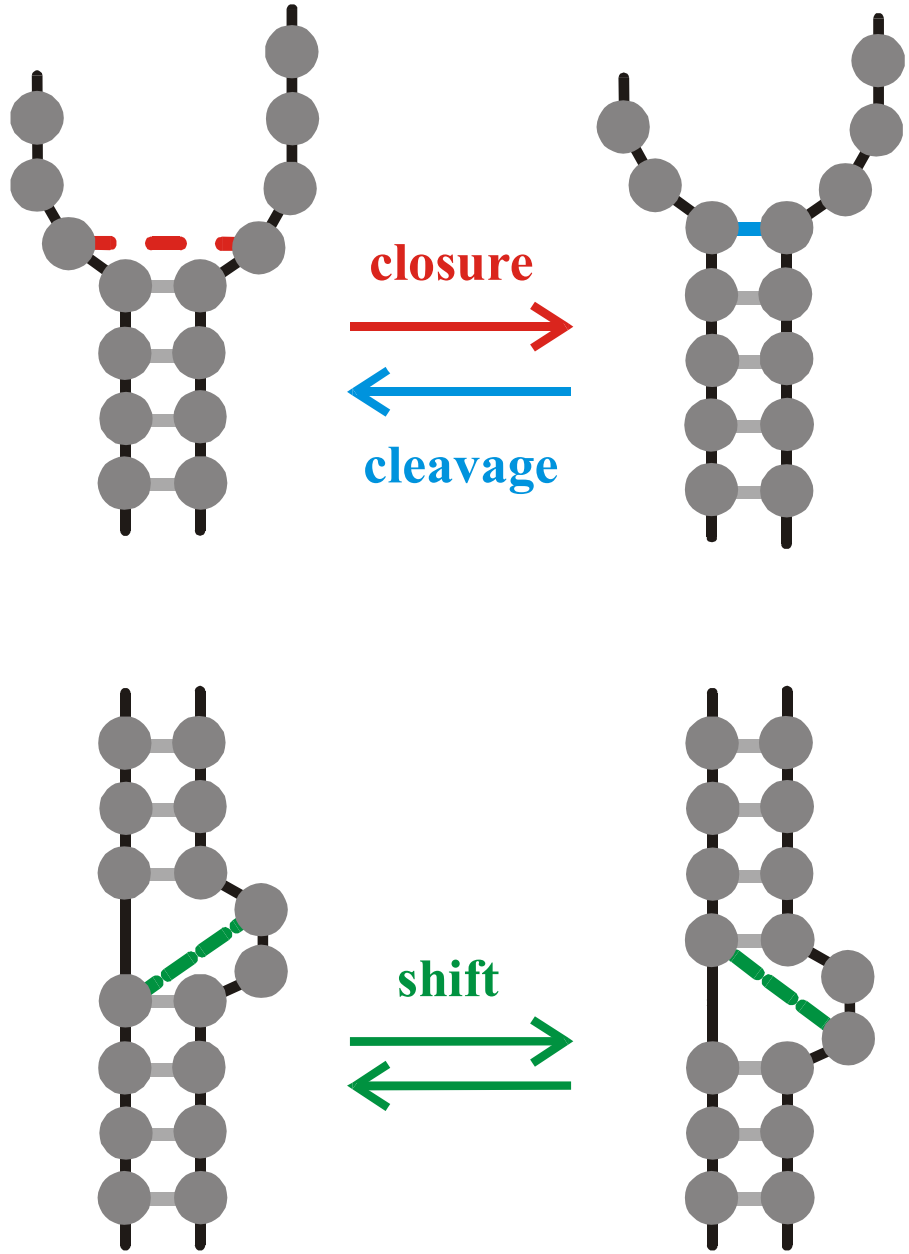
tRNA^{phe}

with modified bases

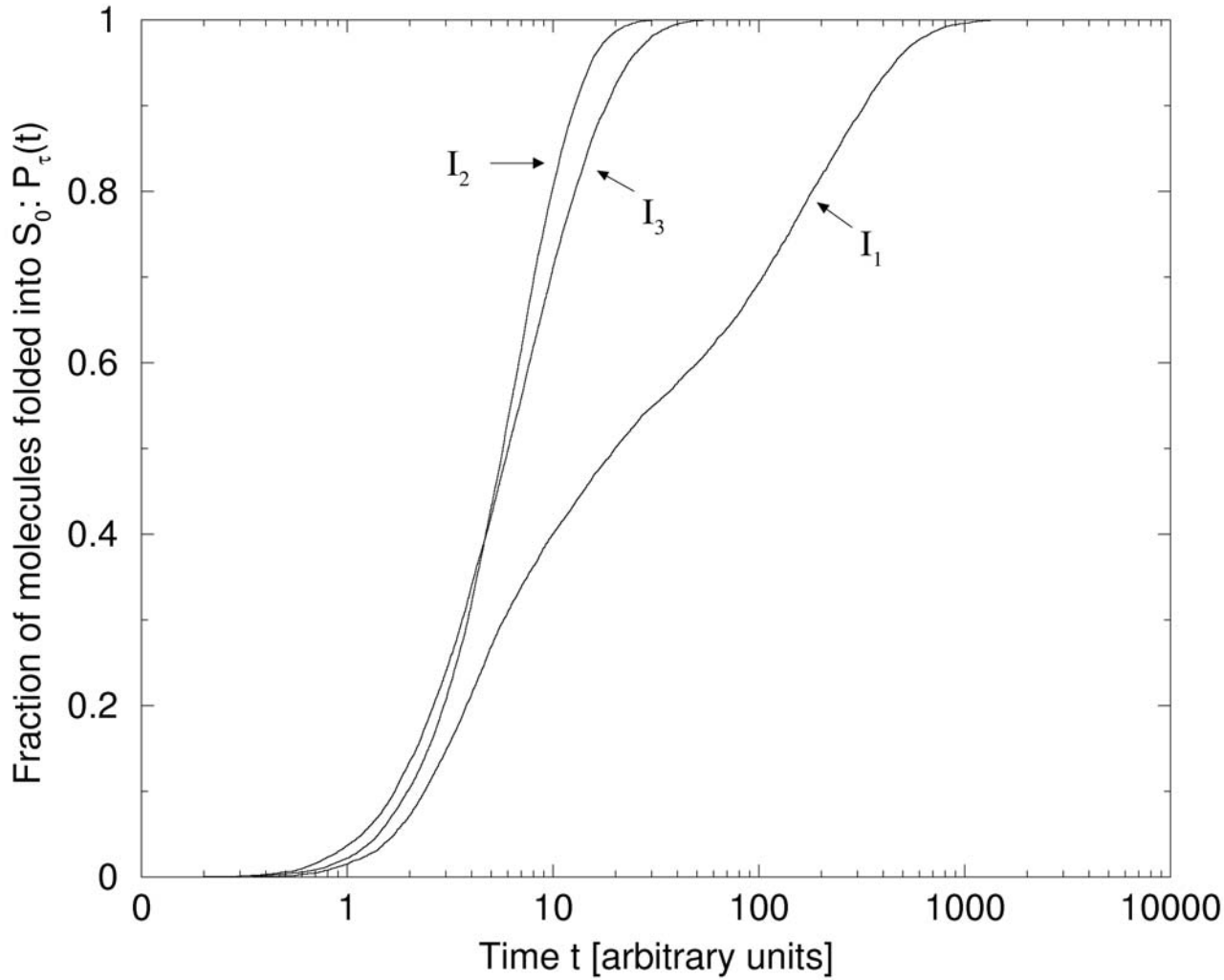
Kinetic Folding of RNA at Elementary Step Resolution

The RNA folding process is resolved to base pair **closure**, base pair **cleavage** and base pair **shift**. The kinetic folding behavior is determined by computation of a sufficiently large ensemble of individual folding trajectories and taking an average over them. The folding behavior is illustrated by barrier trees showing the path of lowest energy between two local minima of free energy.

C.Flamm, W.Fontana, I.L.Hofacker and P.Schuster. *RNA*, **6**:325-338 (2000)

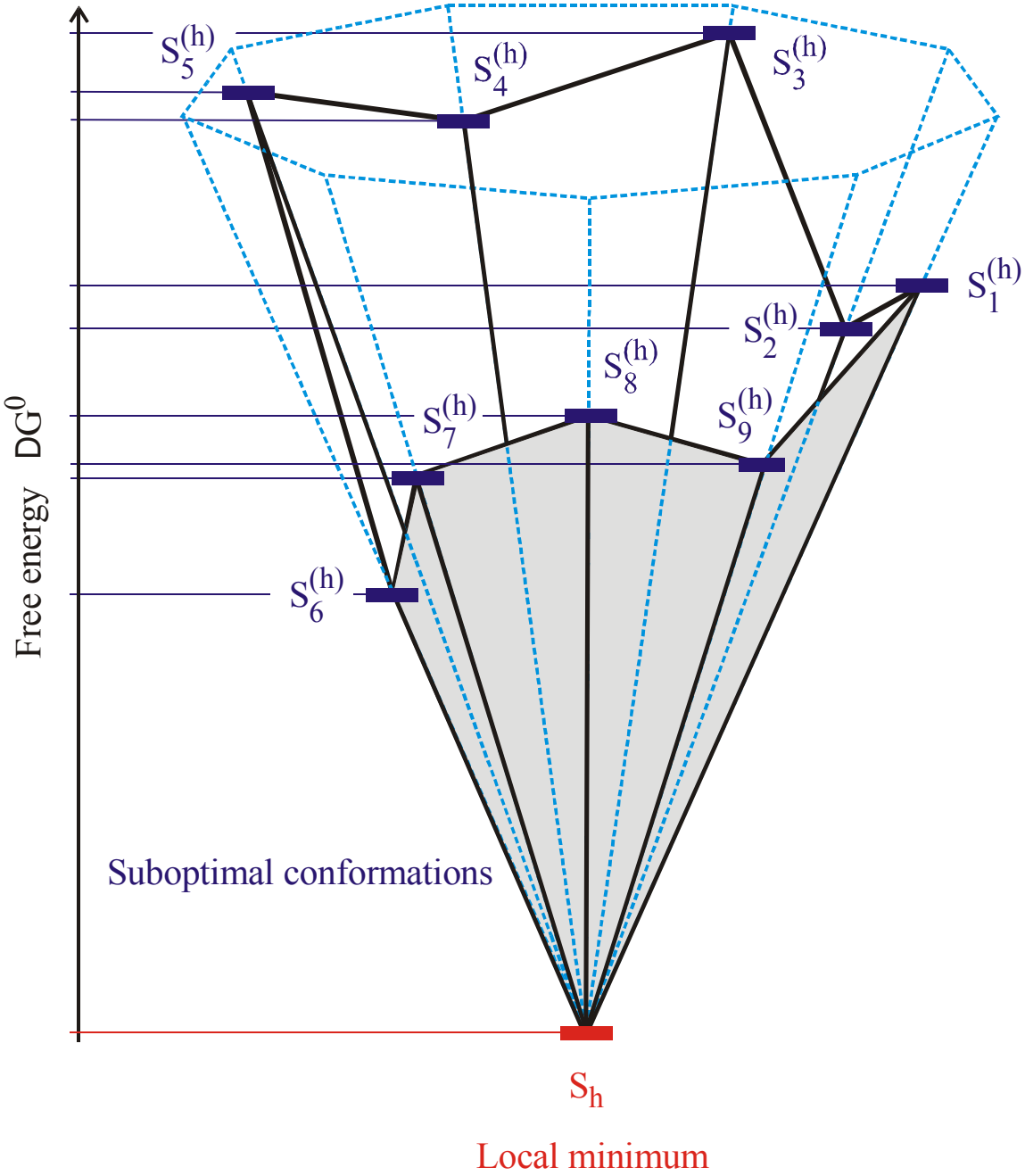


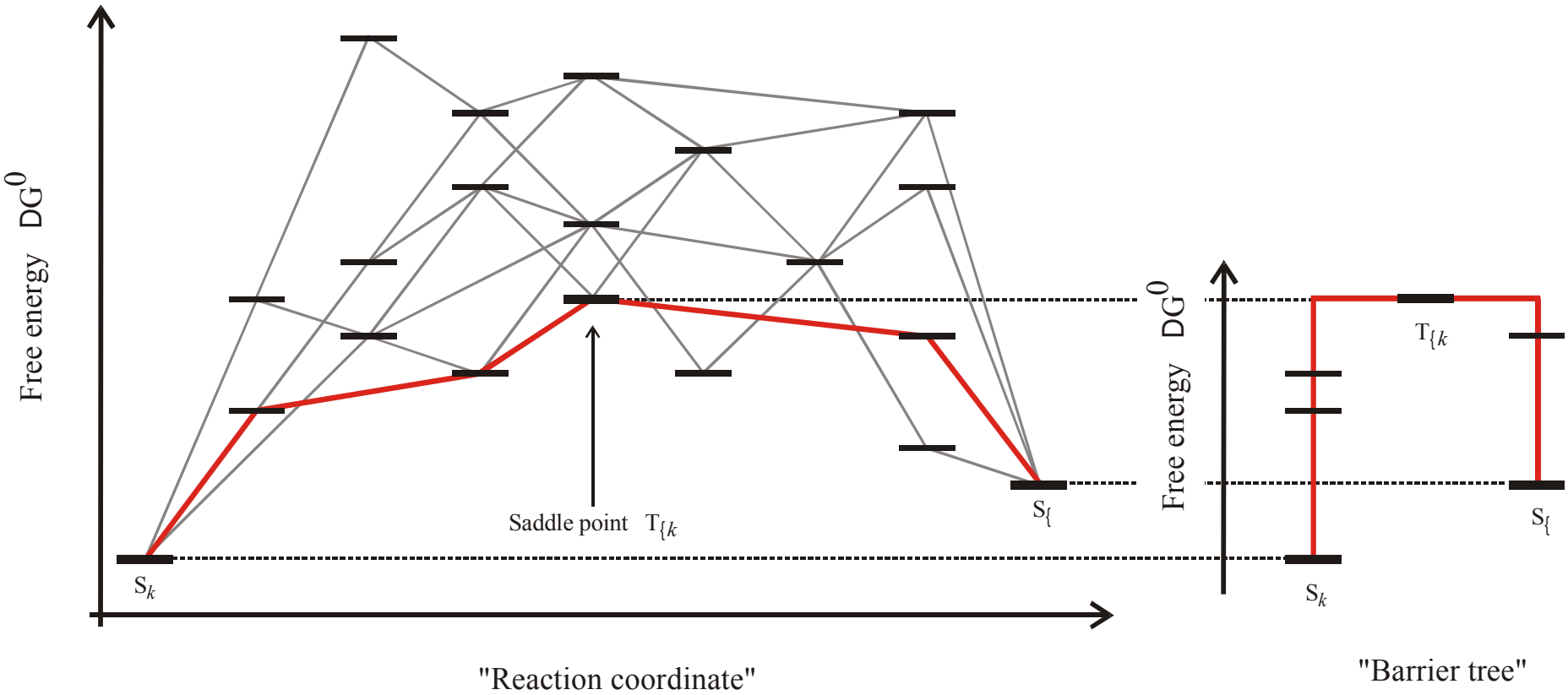
Move set for elementary steps
in kinetic RNA folding

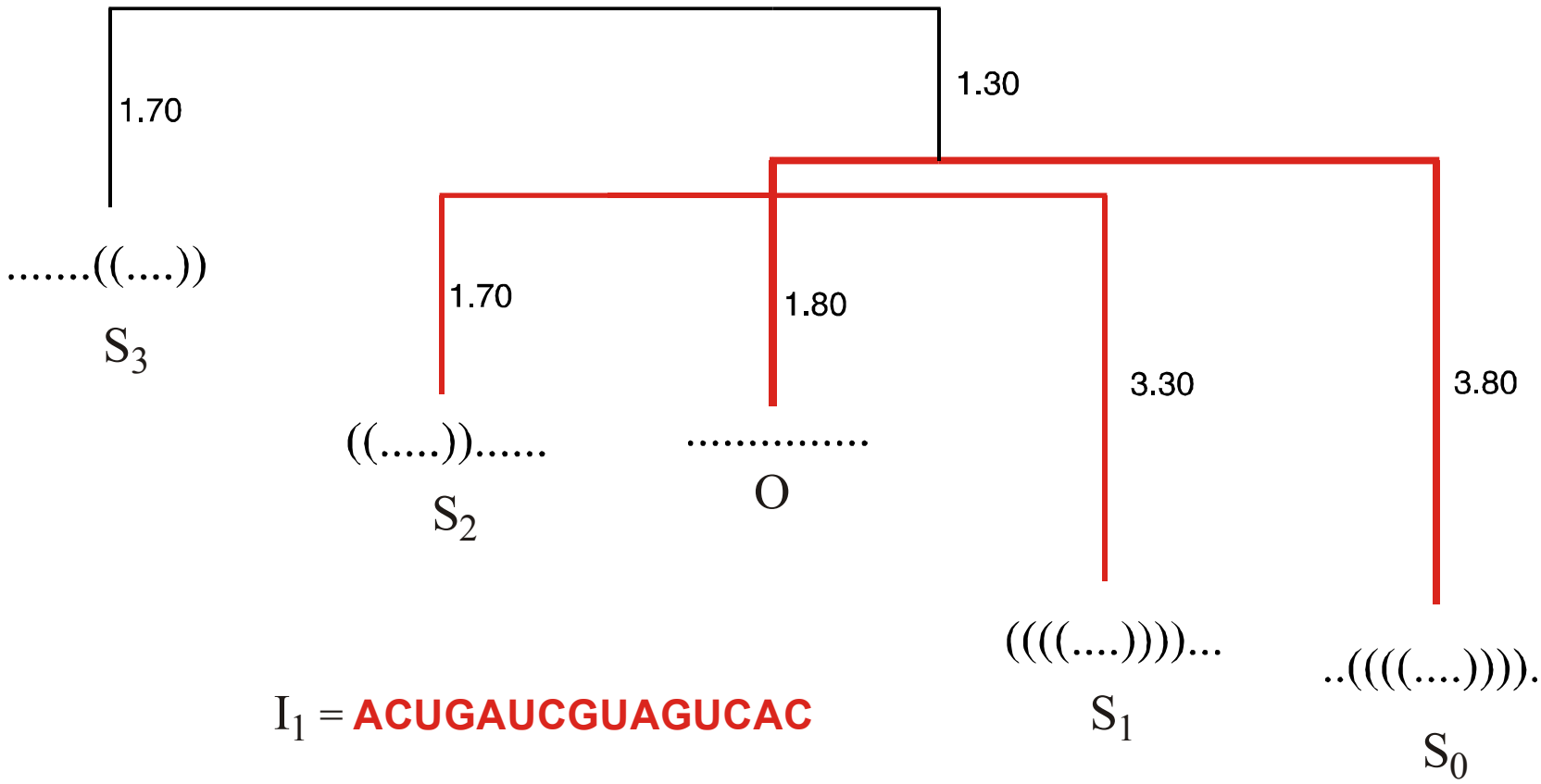


Mean folding curves for three small RNA molecules with n=15 and very different folding behavior

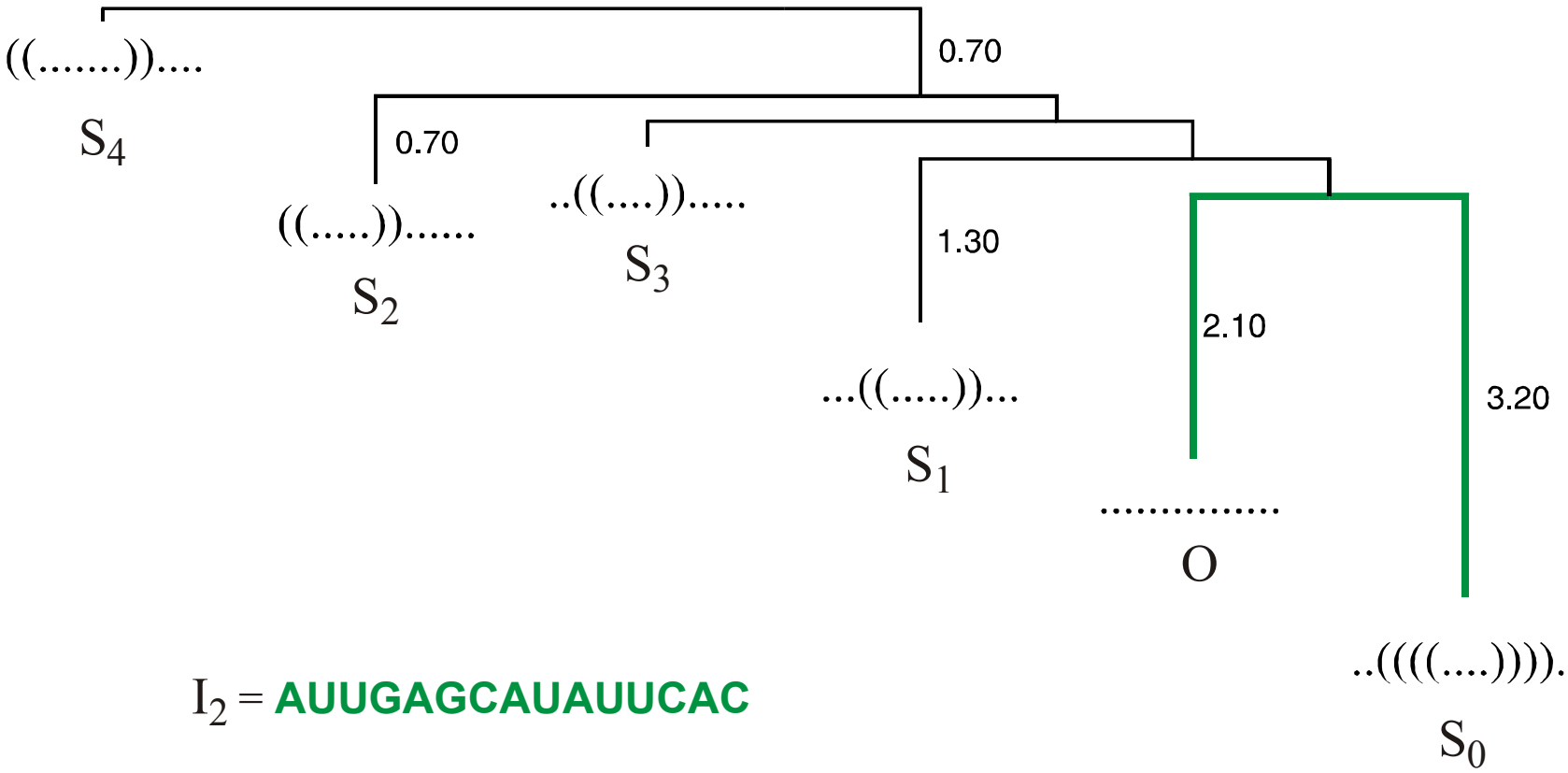
Search for local minima in conformation space




Free energy DG0
Saddle point T{k
S{
Sk
"Reaction coordinate"
Free energy DG0
T{k
S{
Sk
"Barrier tree"

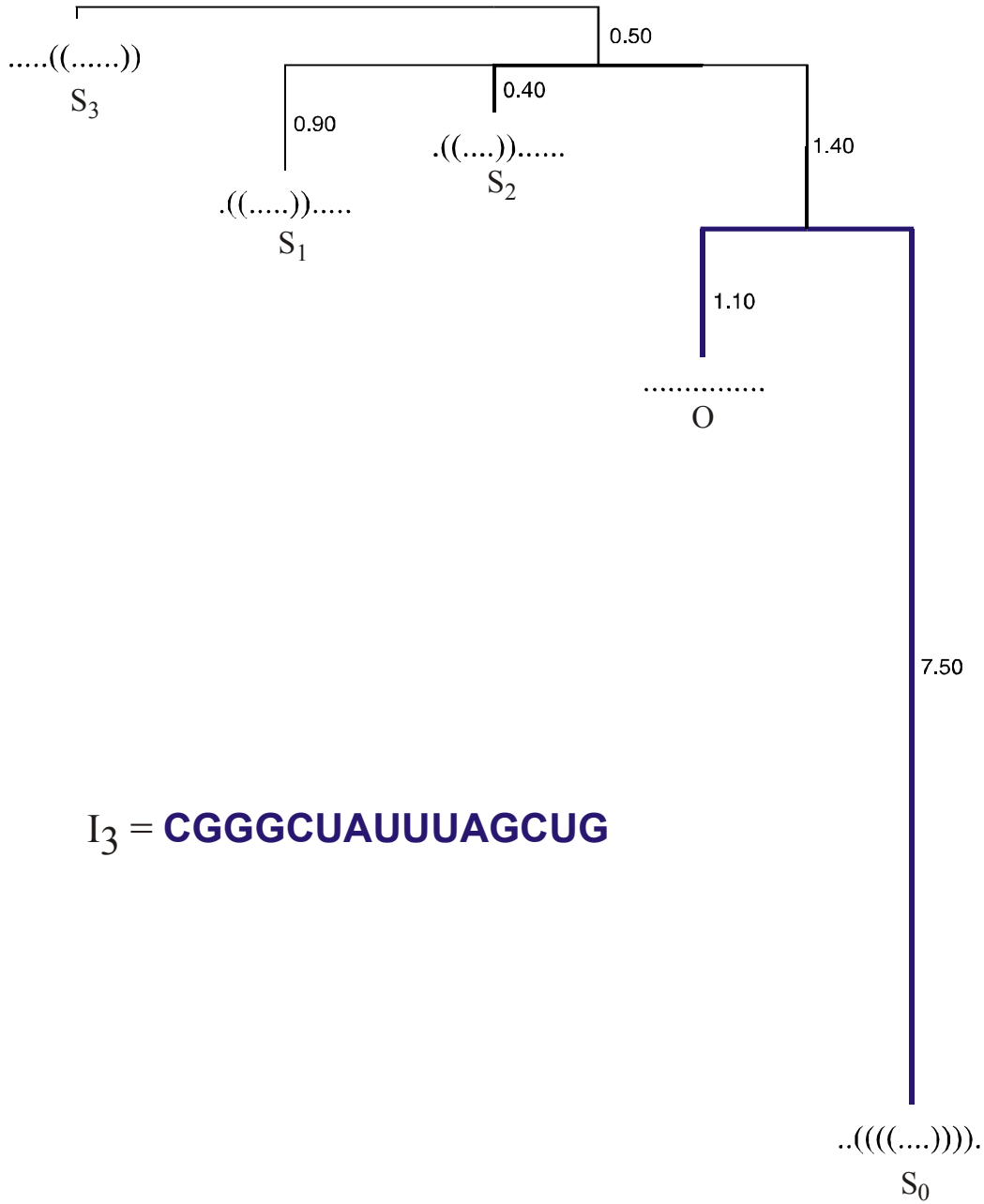


Example of an inefficiently folding small RNA molecule with n = 15



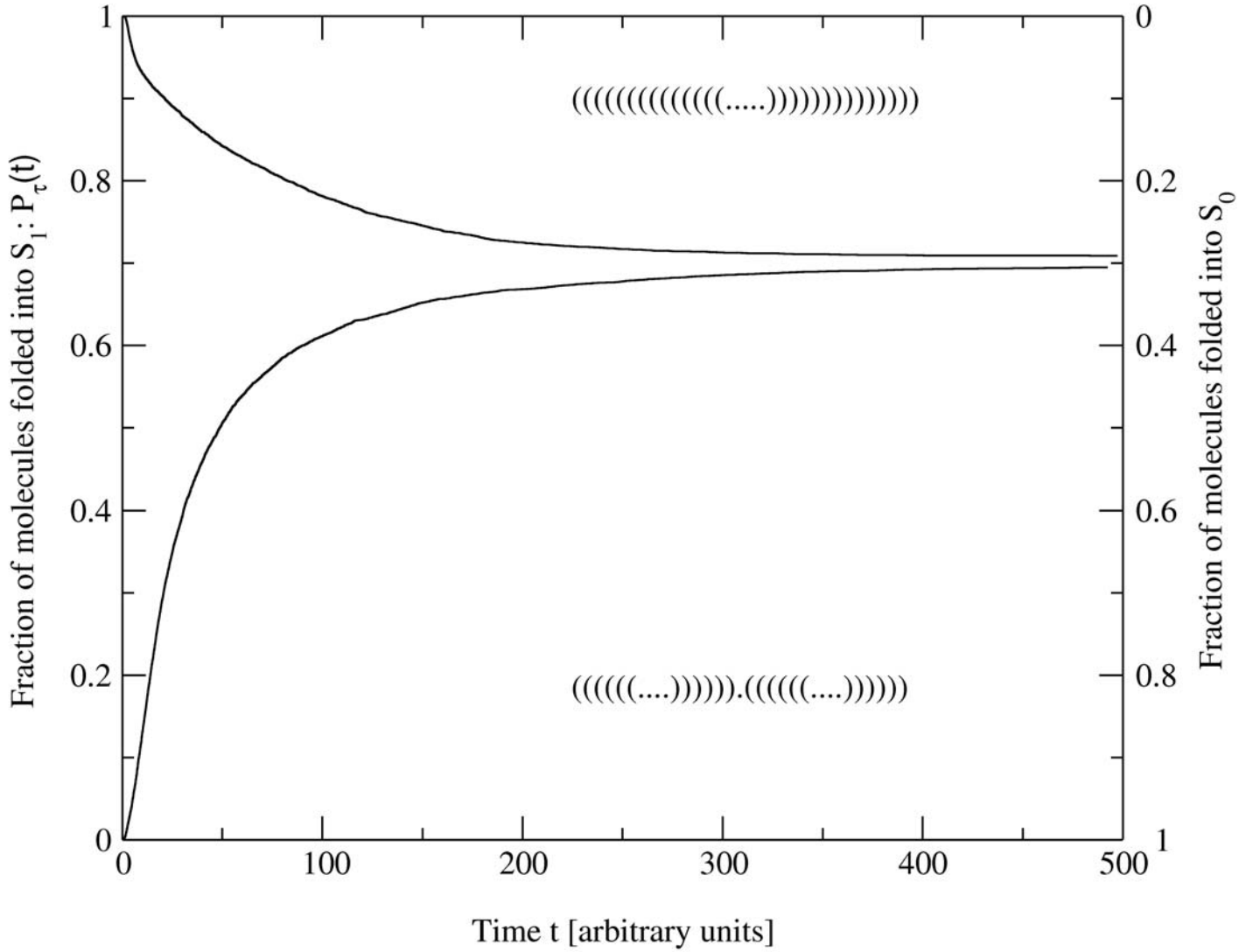
I_2 = **AUUGAGCAUAUUCAC**

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



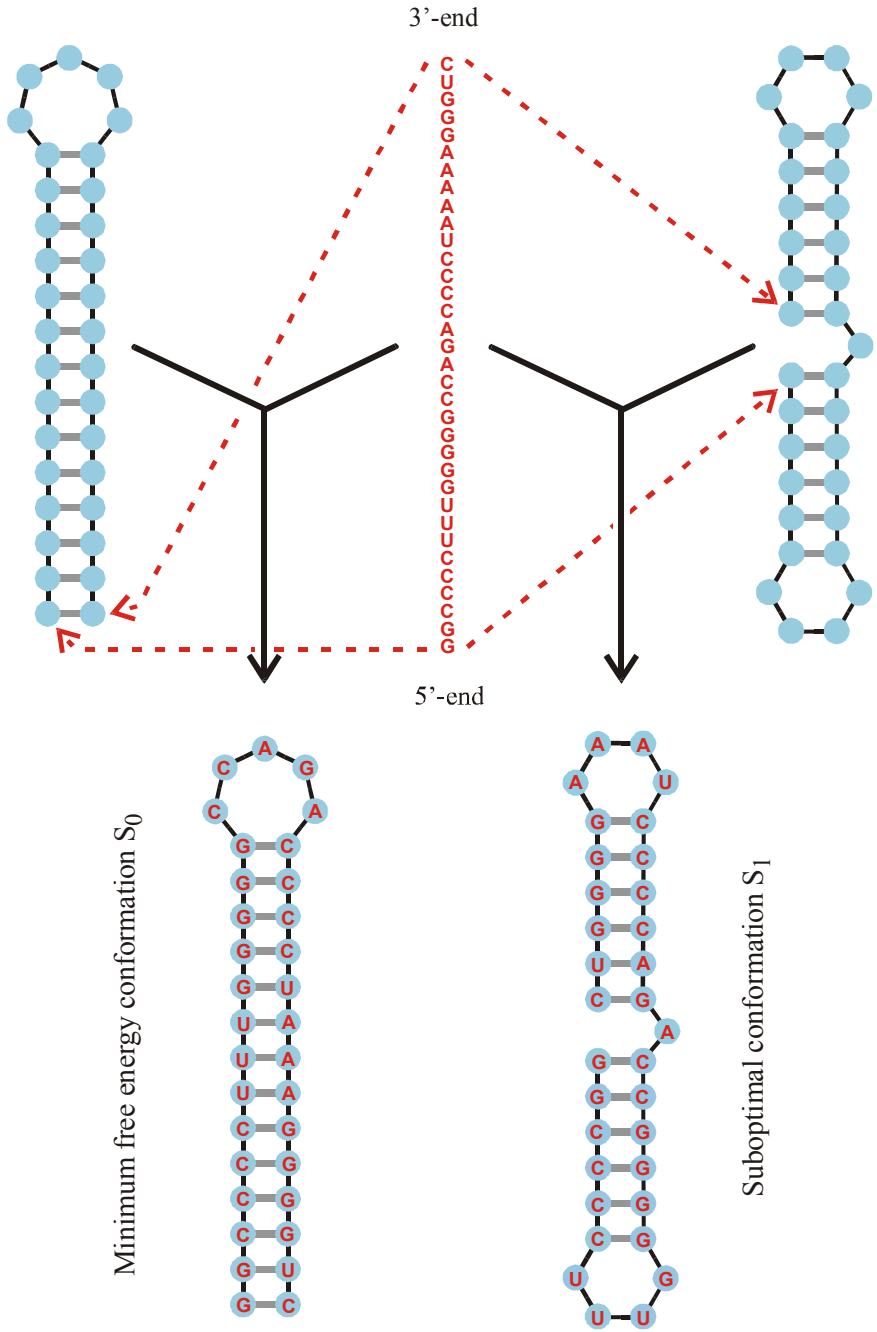
I_3 = **CGGGCUAUUUAGCUG**

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

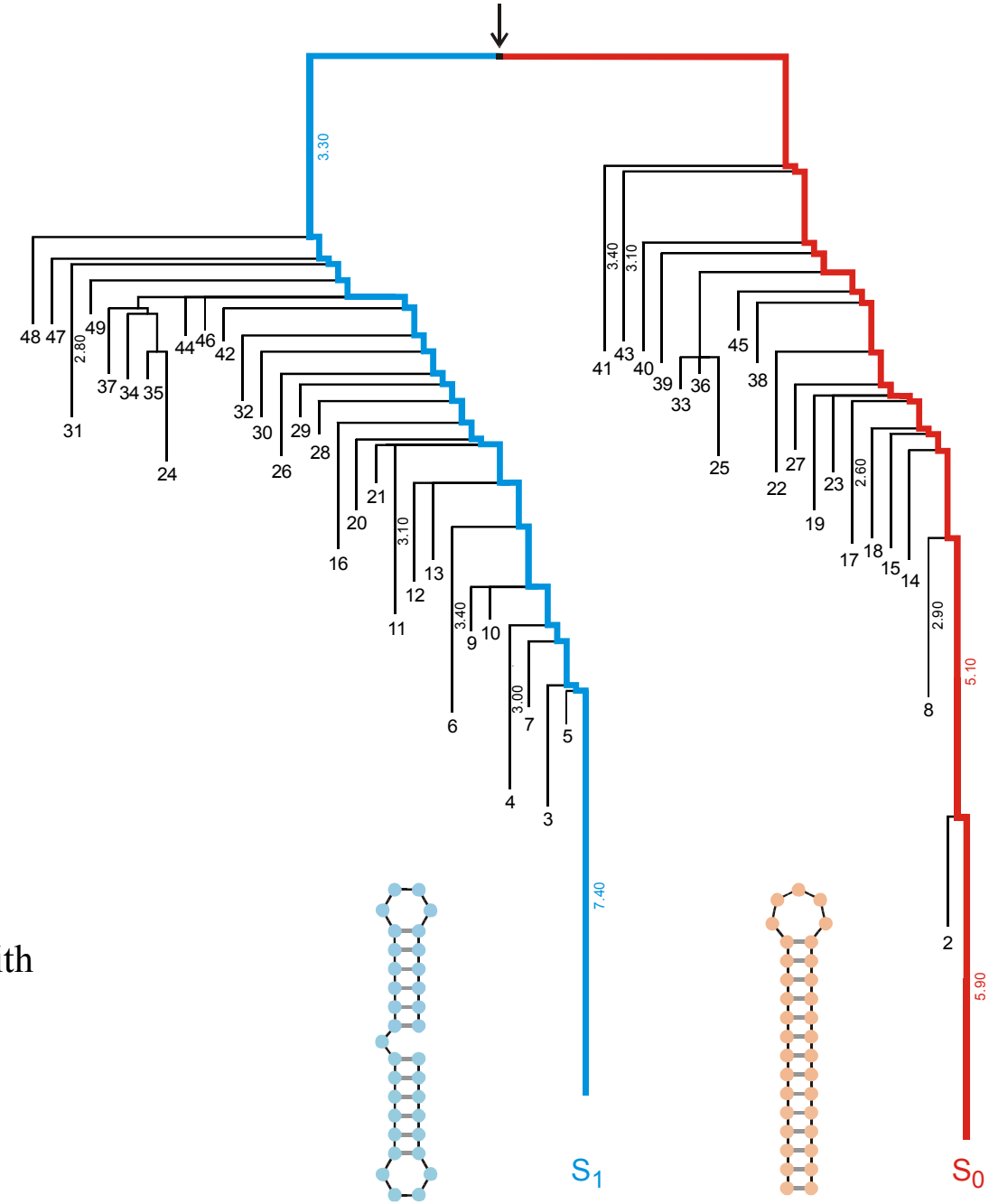


Folding dynamics of the sequence **GGCCCCUUUGGGGGCCAGACCCCUAAAAAGGGUC**

One sequence is compatible with two structures



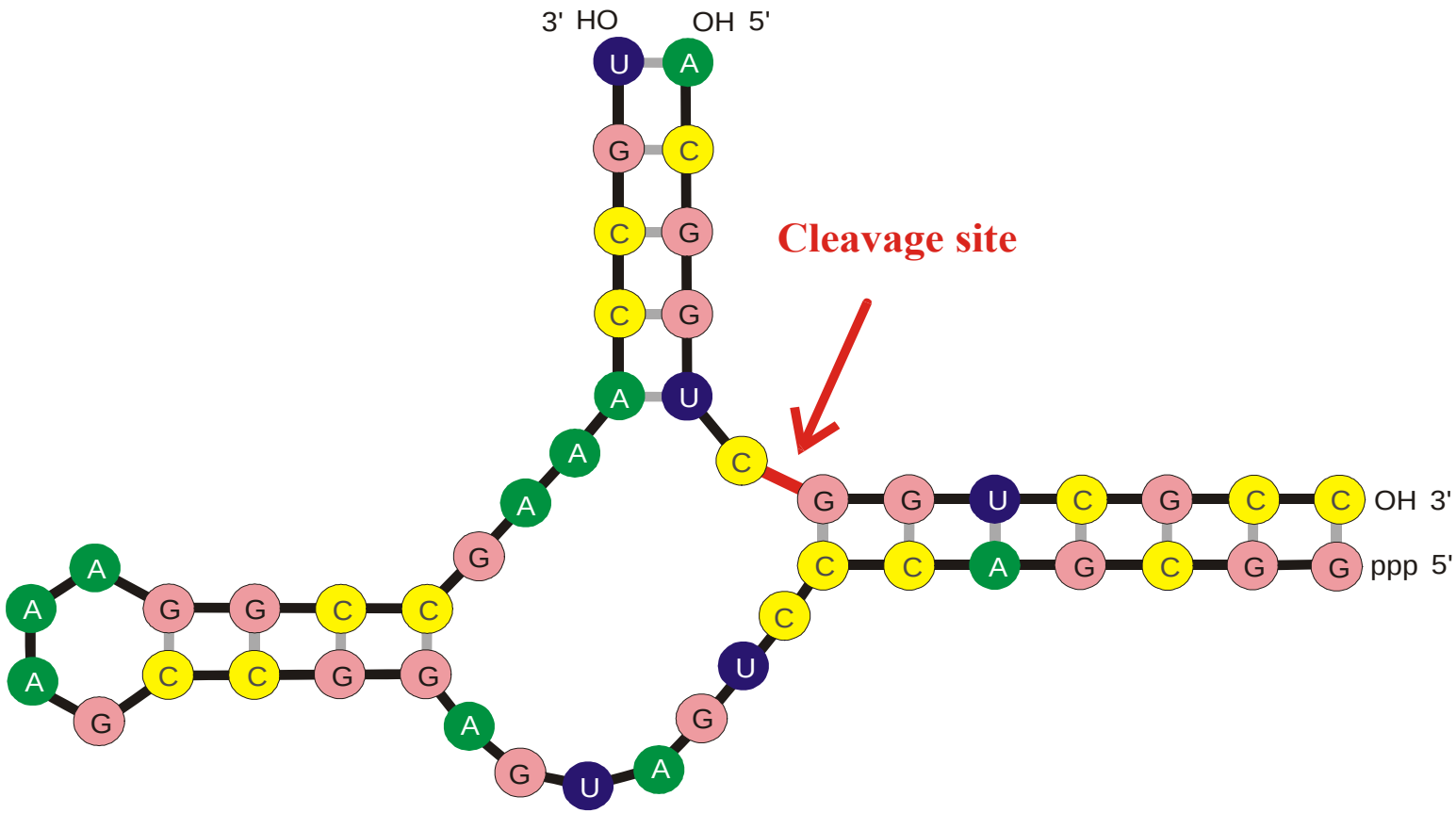
Barrier tree of a sequence with two conformations



Is there experimental evidence for structural multiplicity of RNA sequences?

Are there RNA molecules with multiple functions?

How can RNA molecules with multiple functions be designed?

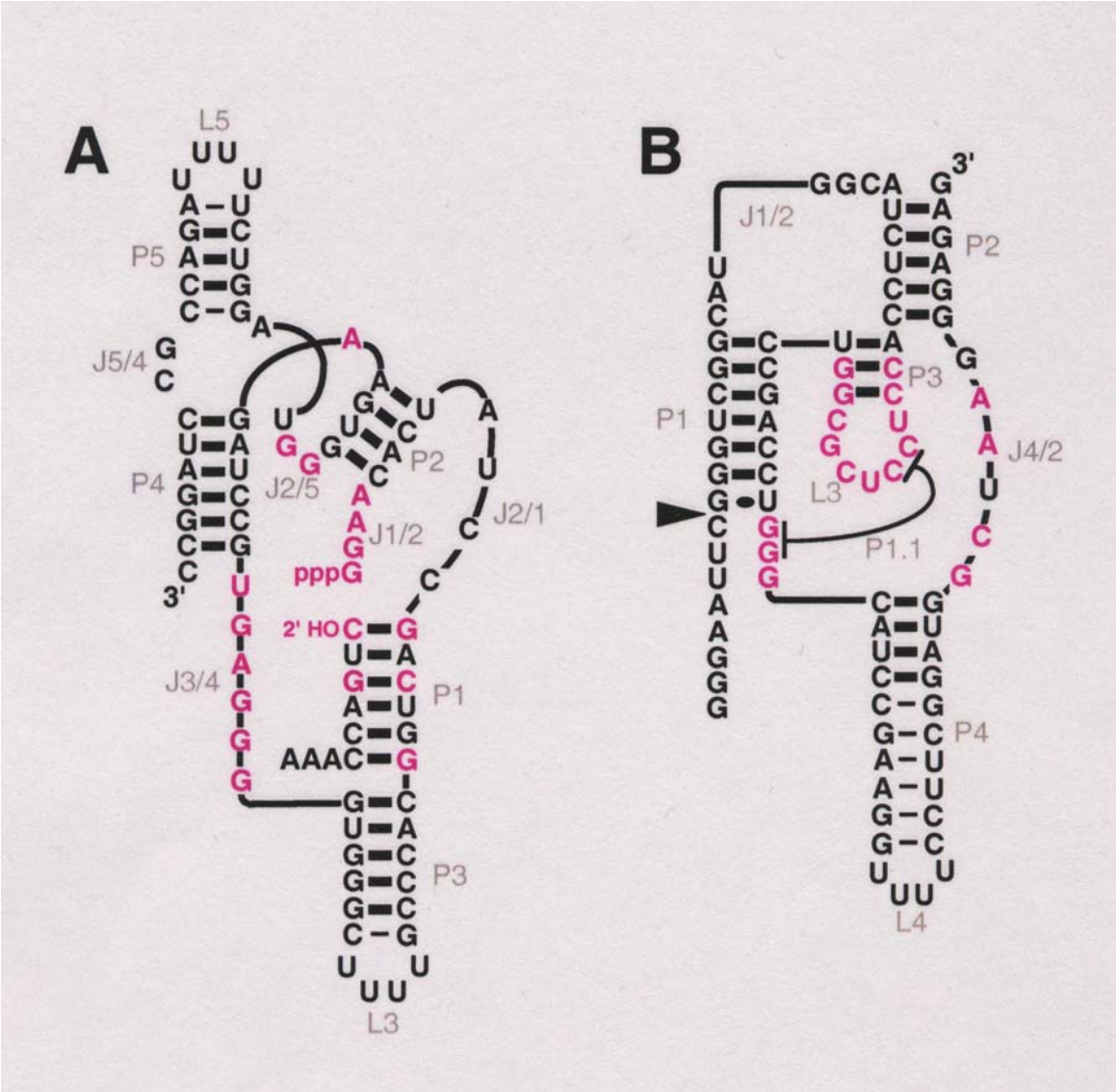


The "hammerhead" ribozyme

The smallest known catalytically active RNA molecule

A ribozyme switch

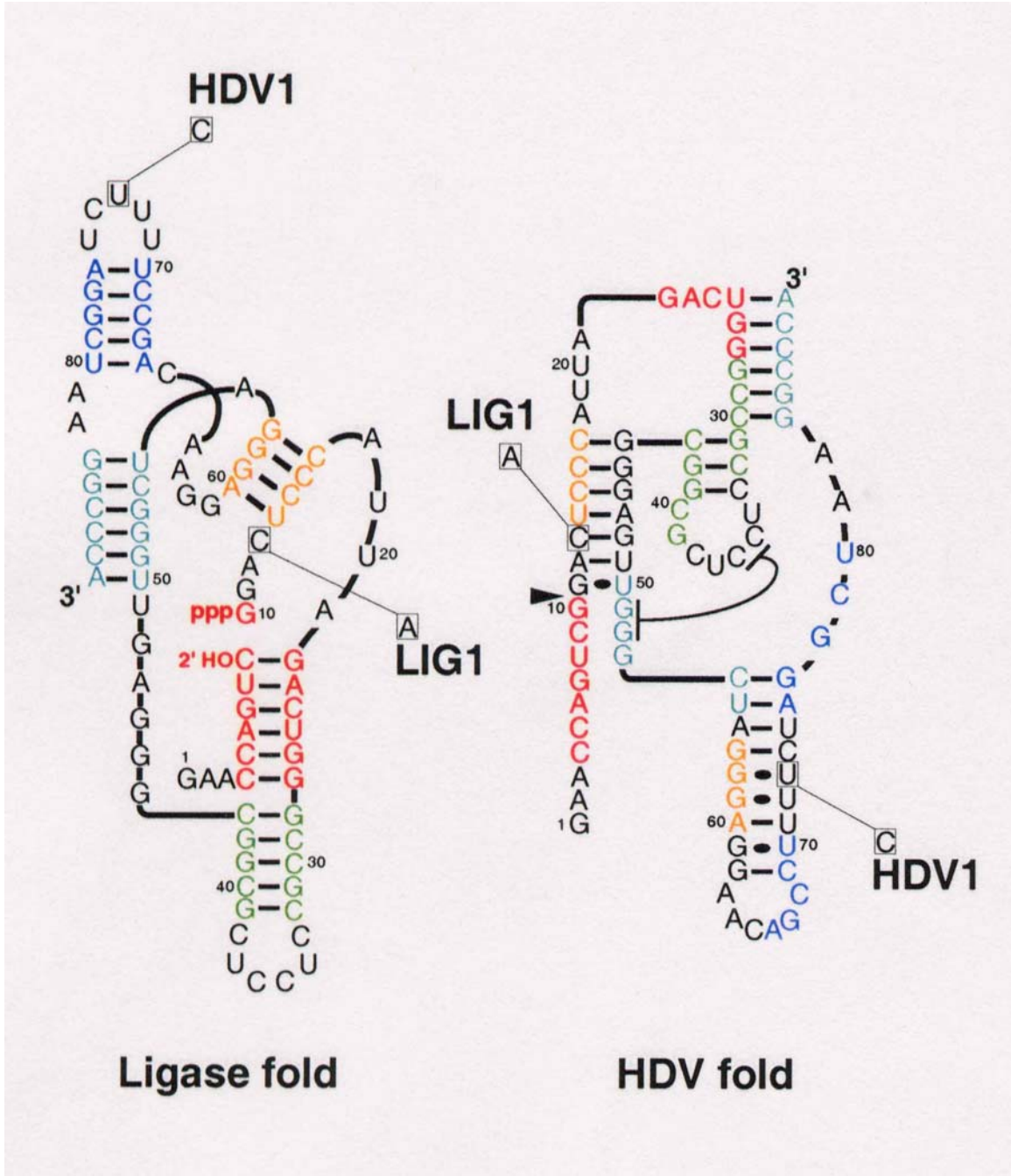
E.A.Schultes, D.B.Bartel, ***One sequence, two ribozymes: Implication for the emergence of new ribozyme folds***. Science **289** (2000), 448-452



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

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

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



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

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

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

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

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

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

(*E.mail: pks@tbi.univie.ac.at*)

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

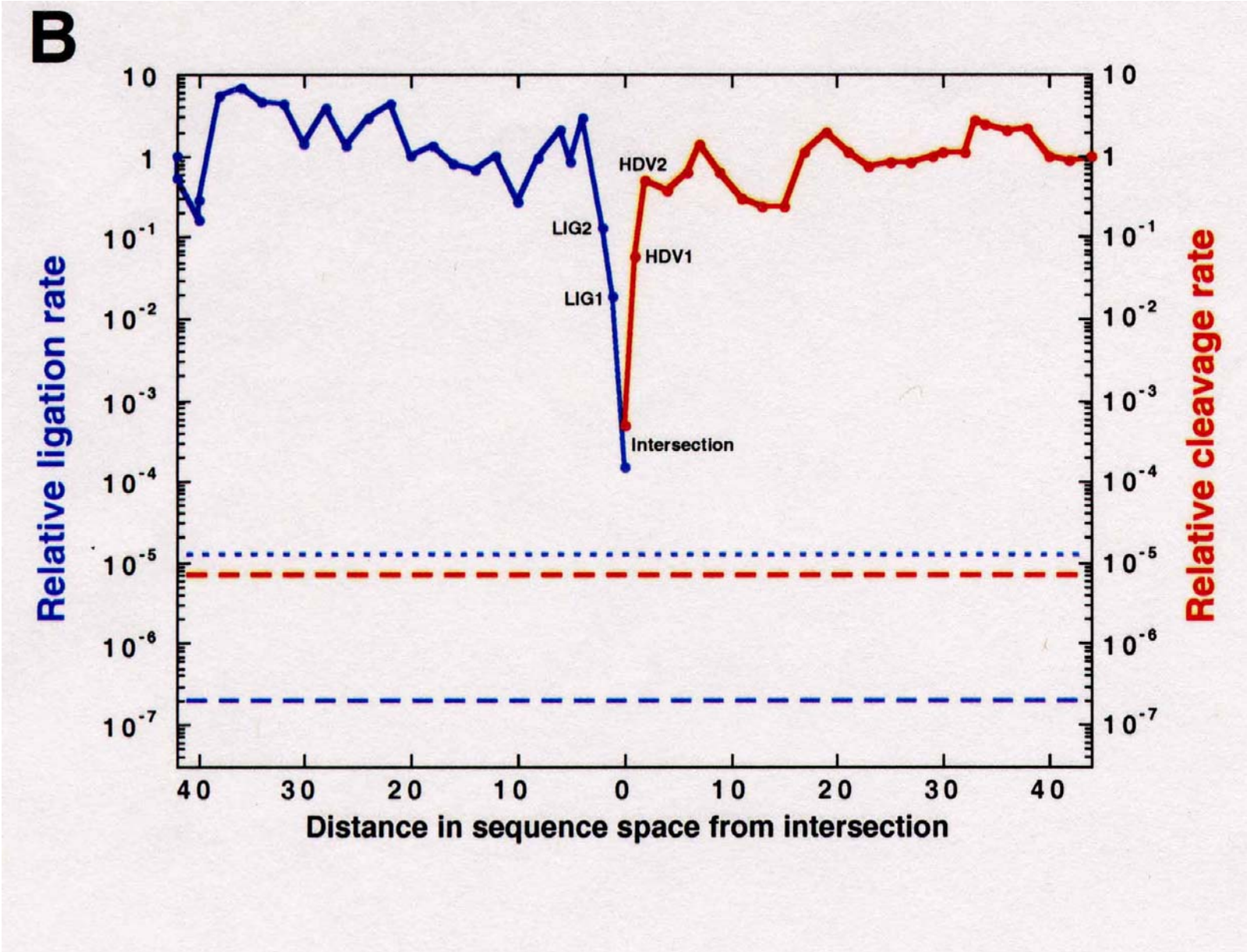
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* **C**[s], **C**[s′] *their corresponding compatible sequences. Then,*

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

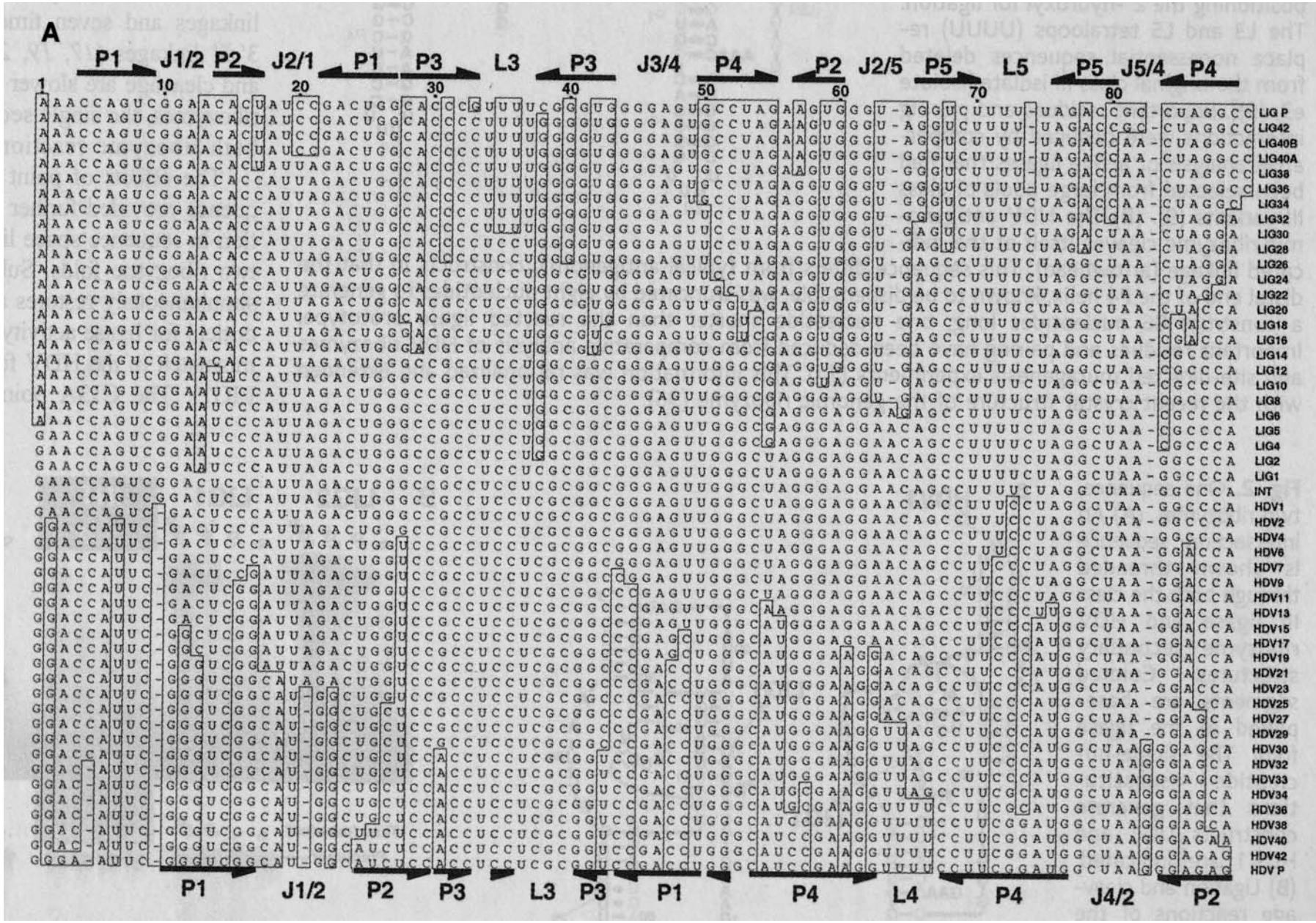
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\gamma(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles 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***



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



Sequence of mutants from the intersection to both reference ribozymes

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.

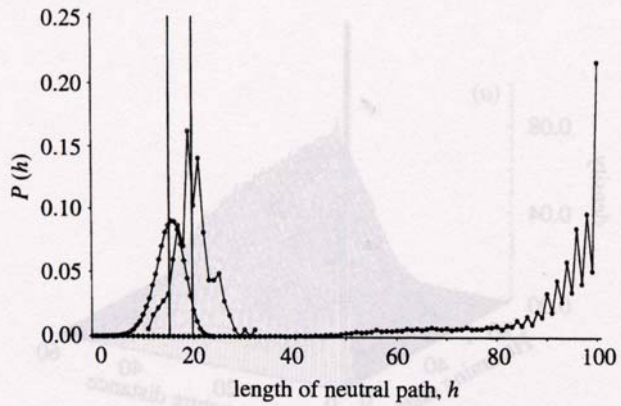
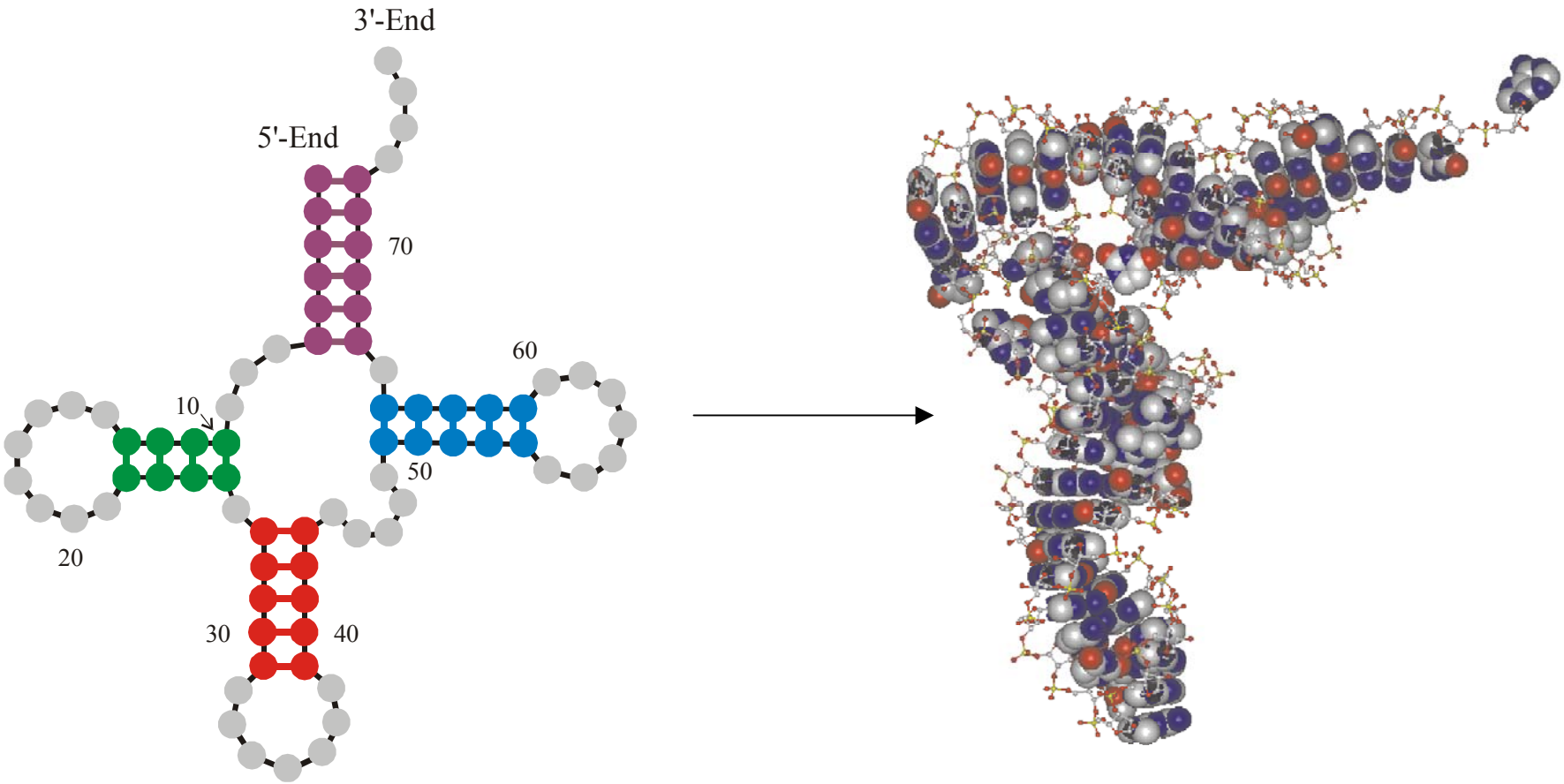


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



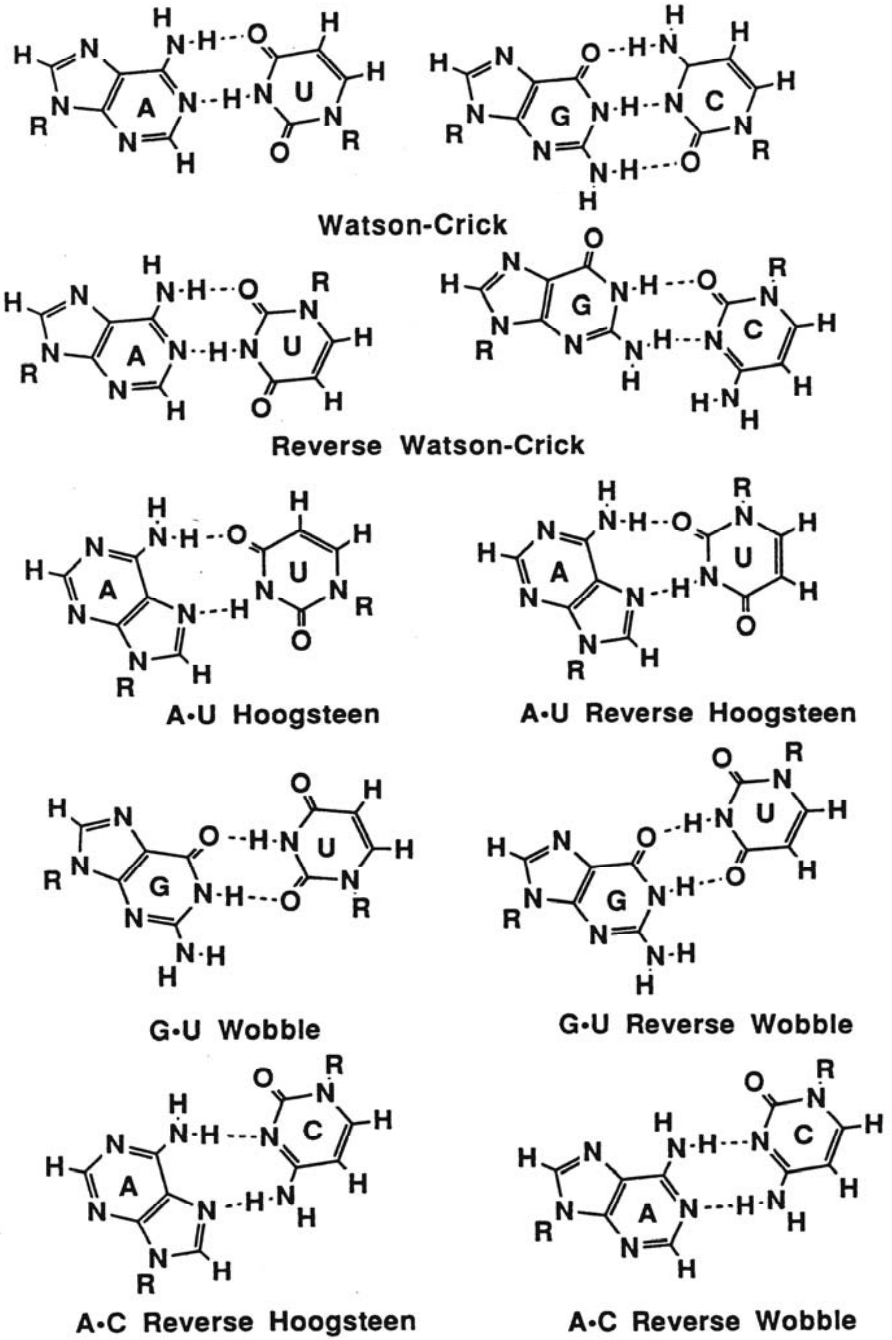
From RNA secondary structures to full three-dimensional structures.
Example: Phenylalanyl-transfer-RNA

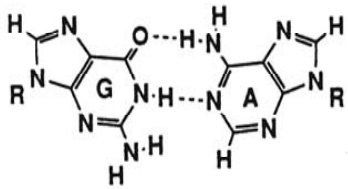
Which perspectives have RNA structure modelling and elaborate sequence-structure analysis?

Secondary structures are based on the identification of base pairs with defined and only marginally varying geometries that fit into A- or A'-type helices. Until now a great variety of other classifiable base pairs have been found by crystallography and NMR. They can be readily included in structure prediction methods with are similar to the current algorithms for conventional secondary structures. What is needed, however, is the determination of thermodynamic parameters for these unconventional base-base interactions, as it was done in the nineteen-seventies for DNA and RNA double helical and loop structures. So far these data are scarce except H-type pseudo-knots and end-to-end stacking of helices.

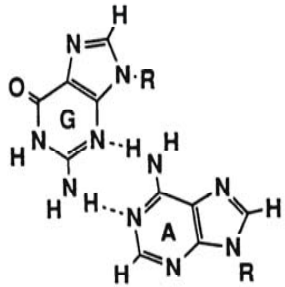
It seems that the prediction of RNA structures will be an easier task than that of proteins.

Classification of purine-pyrimidine base pairs

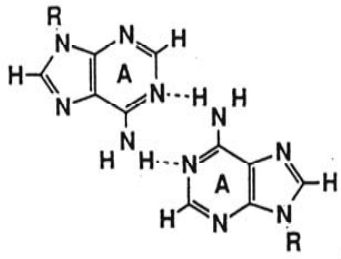




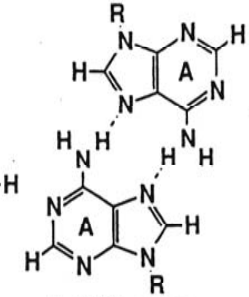
G•A N1-N1, carbonyl-amino



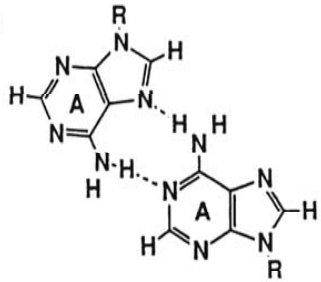
G•A N3-amino, amino-N1



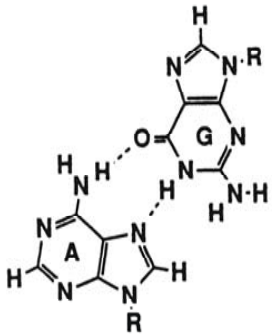
A•A N1-amino, symmetric



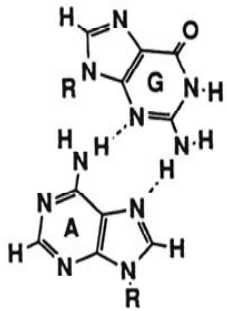
A•A N7-amino, symmetric



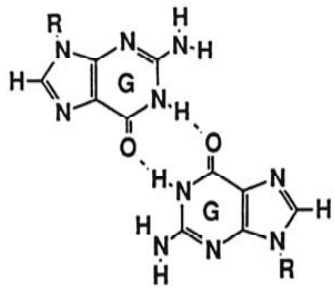
A•A N1-amino, N7-amino



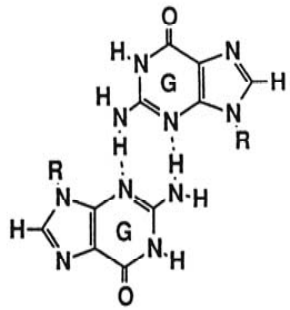
A•G N7-N1, amino-carbonyl



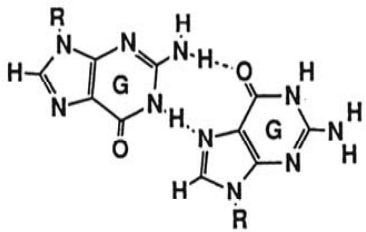
A•G N7-amino, amino-N3



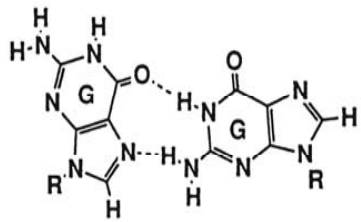
G•G N1-carbonyl, symmetric



G•G N3-amino, symmetric

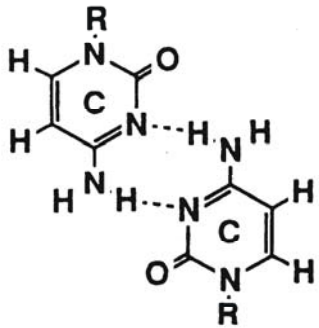


G•G N7-N1, carbonyl-amino

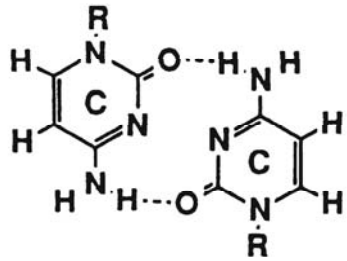


G•G N1-carbonyl, N7-amino

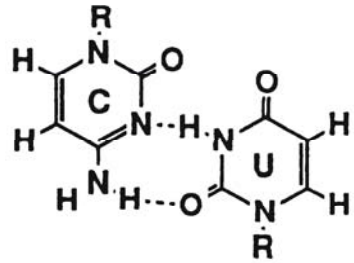
Classification of purine-purine base pairs



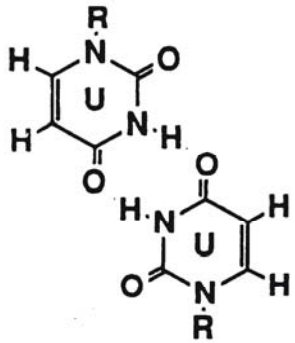
C•C N3-amino, symmetric



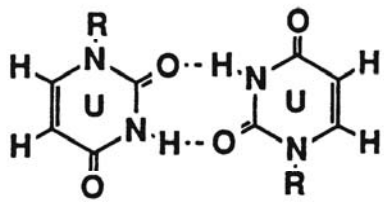
C•C carbonyl-amino, symmetric



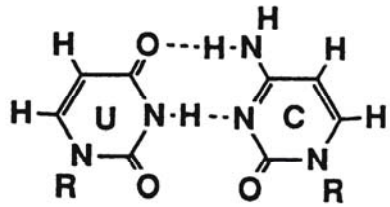
C•U N3-N3, 2-carbonyl-amino



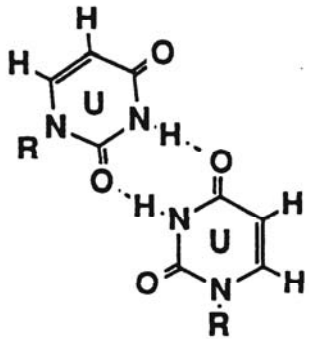
U•U 4-carbonyl-N3, symmetric



U•U 2-carbonyl-N3, symmetric

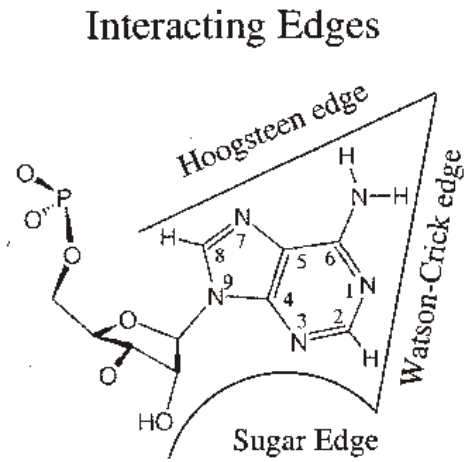


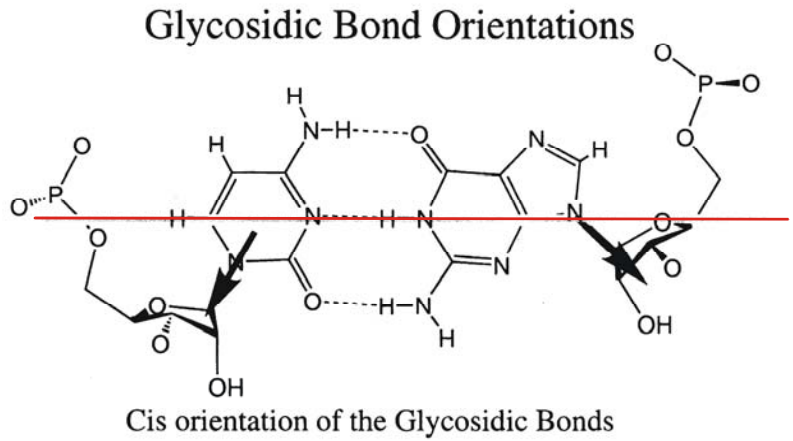
U•C N3-N3, 4-carbonyl-amino

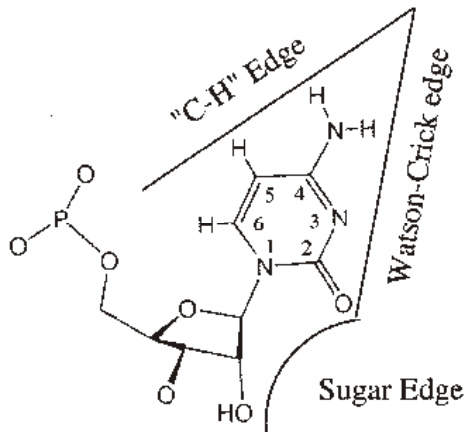


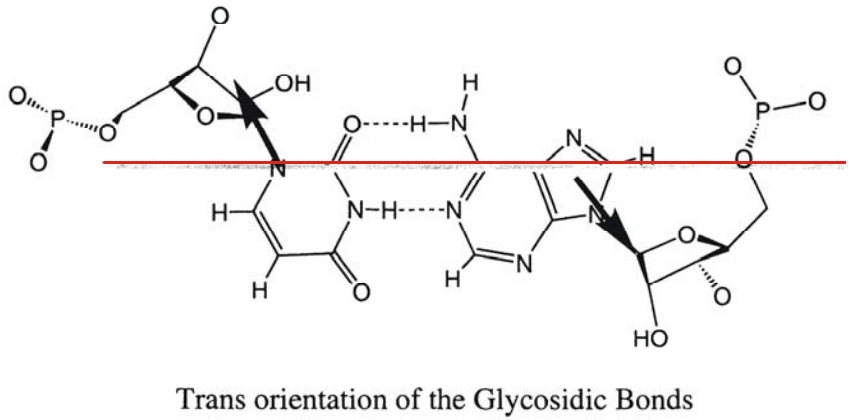
U•U 2-carbonyl-N3, 4-carbonyl-N3

Classification of pyrimidine-pyrimidine base pairs









General classification of base pairs

N.B.Leontis and E. Westhof, *RNA* **7**:499-512 (2001)

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

Peter Stadler, Universität Leipzig, GE

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

Bärbel Stadler, Andreas Wernitznig, Universität Wien, AT
Michael Kospach, Ulrike Langhammer, Ulrike Mückstein, Stefanie Widder
Jan Cupal, Kurt Grünberger, Andreas Svrček-Seiler, Stefan Wuchty

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