

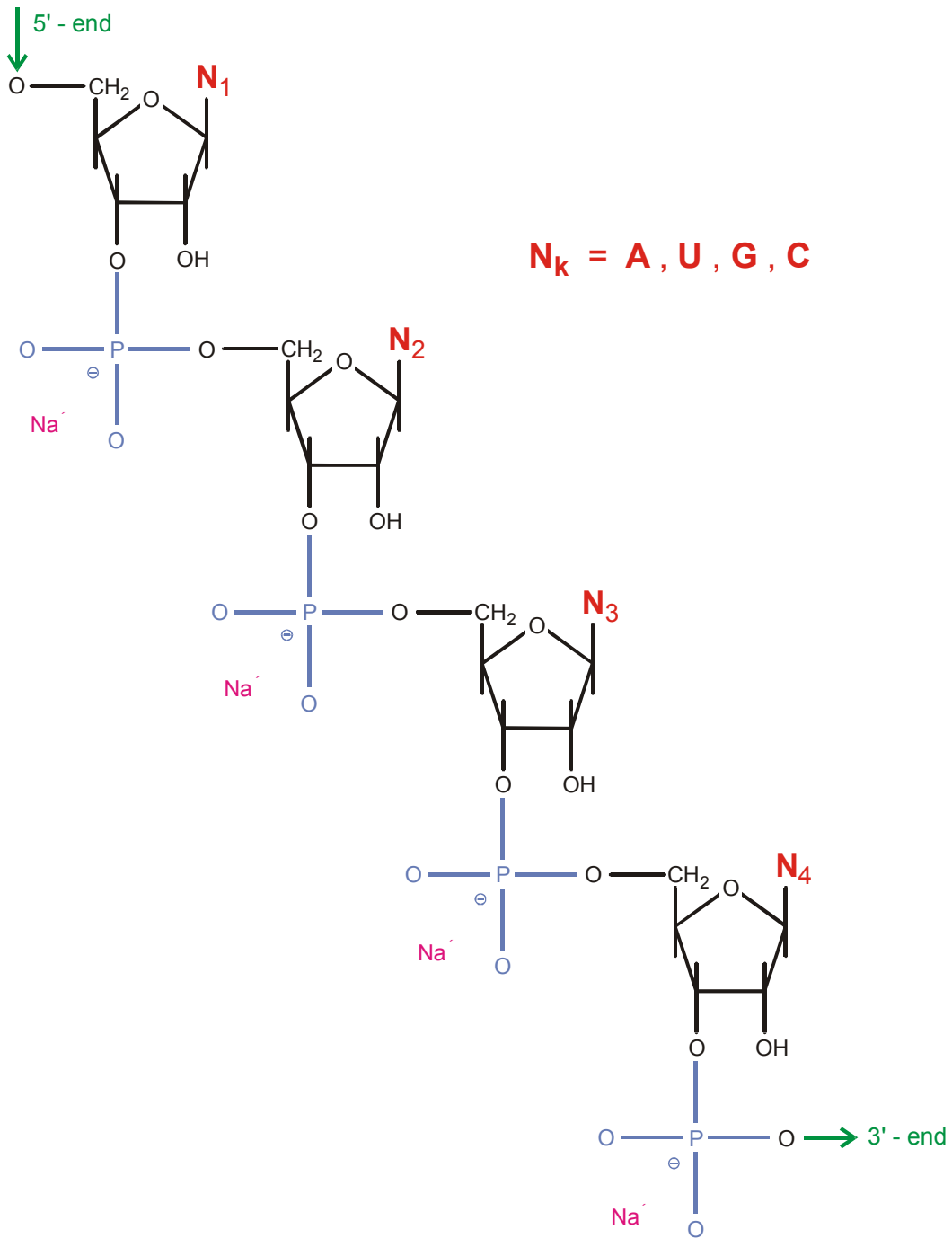
RNA Structures

Stability, Folding and the Role of Hydrogen Bonding and Protons

Peter Schuster

Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien

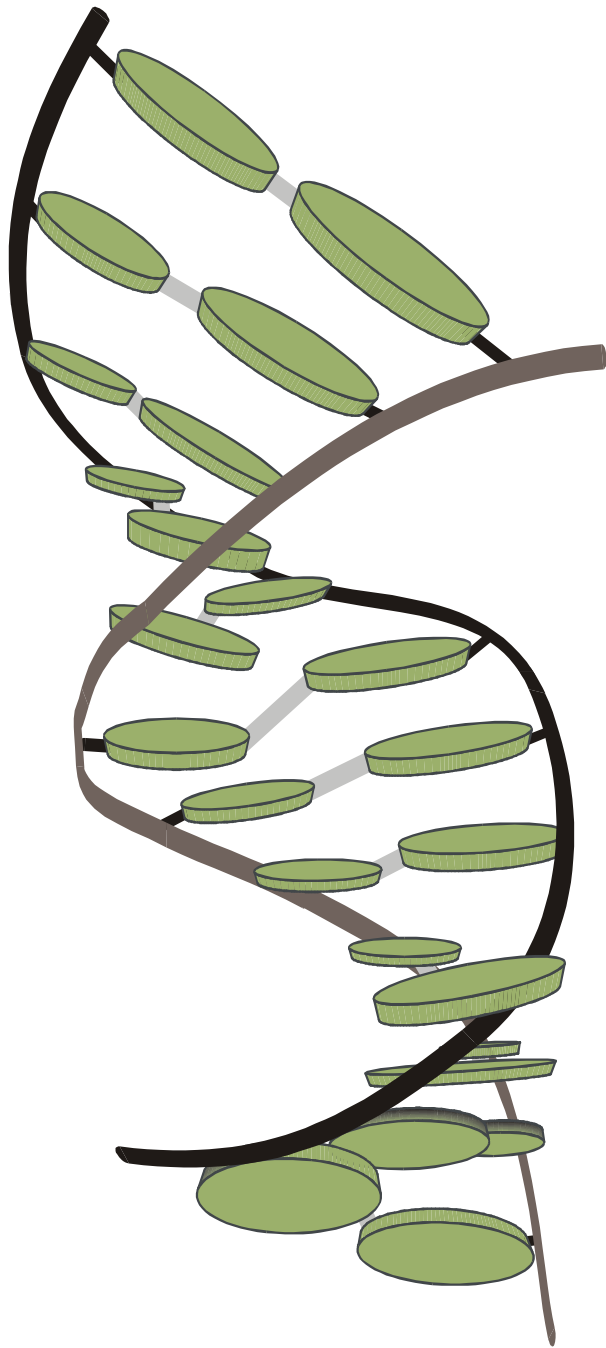
Schloß Ringberg, 05.03.2002



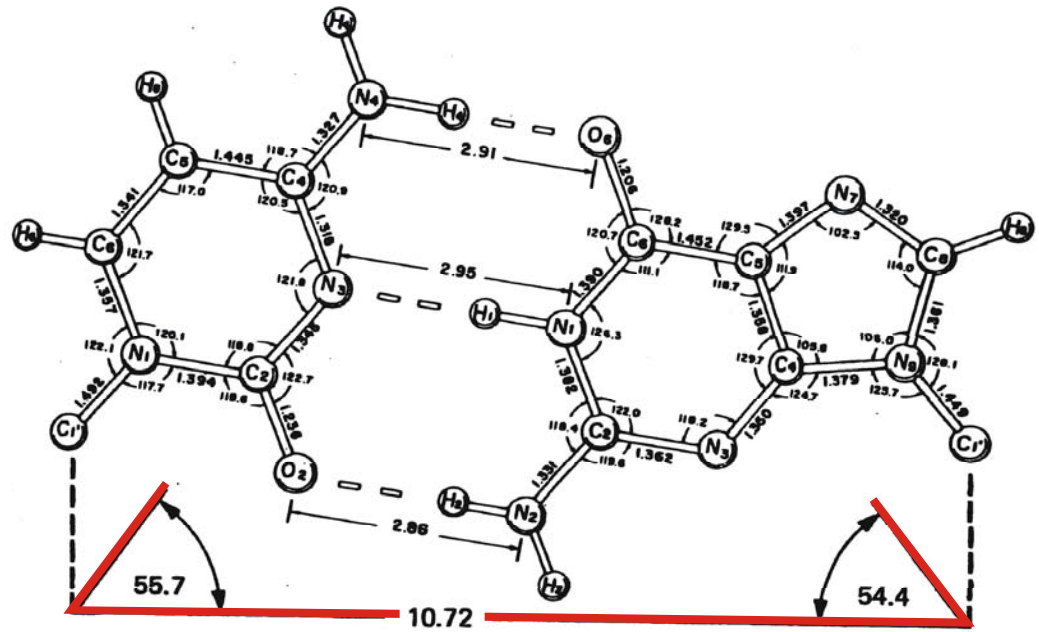
The chemical formula of RNA consisting of **nucleobases**, ribose rings, **phosphate groups**, and **sodium counterions**

Structural Constraints and Hydrogen Bonding in RNA

Single stranded RNA molecules form structures, which combine double-helical stacking (A-type) regions with loops and metal ion (Mg^{2+}) coordinated centers.

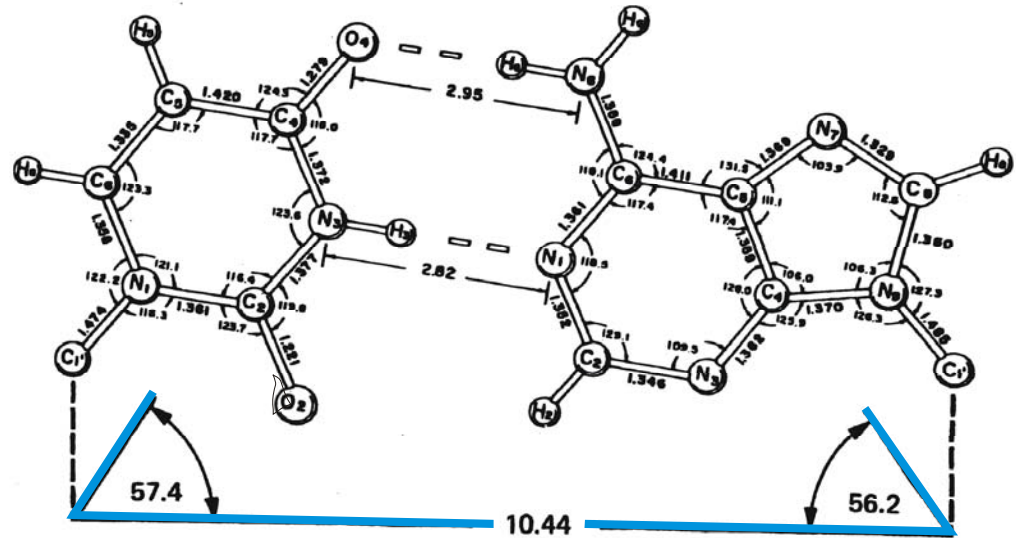


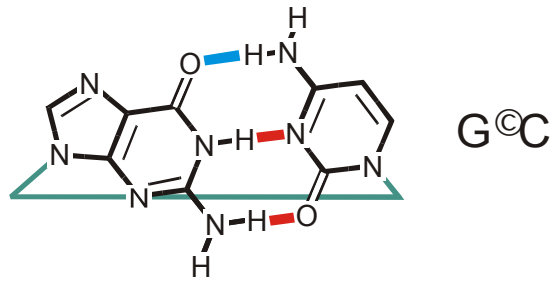
The three-dimensional structure of a short double helical stack



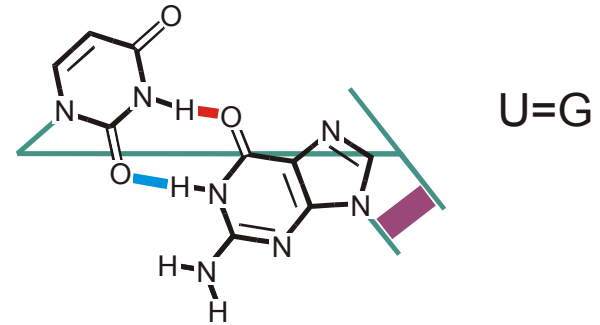
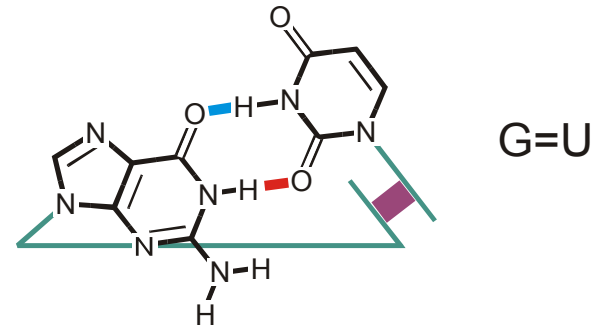
Canonical Watson-Crick
base pairs:

cytosine – guanine
uracil – adenine





Canonical Watson-Crick base-pair



Wobble base-pairs

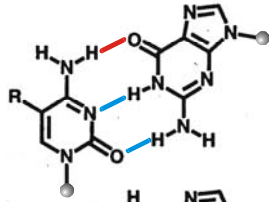
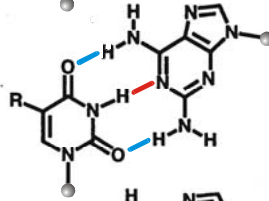
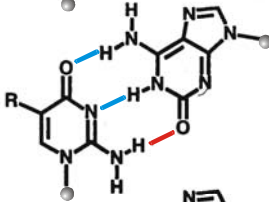
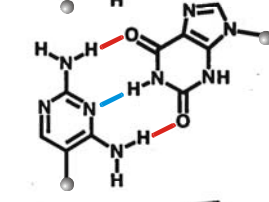
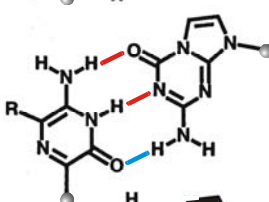
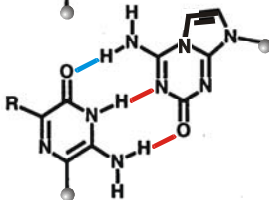
Wobble base pairs in RNA double-helical stacks

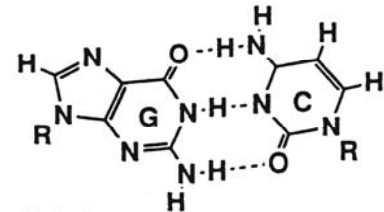
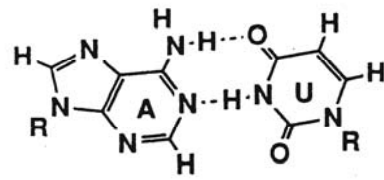
Color code:

Donor—Acceptor

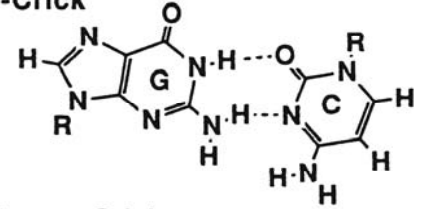
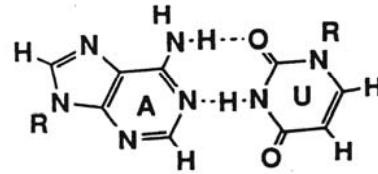
Acceptor—Donor

Hydrogen bonding patterns for Watson-Crick base pairs

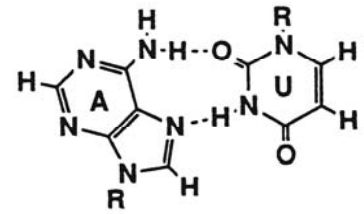
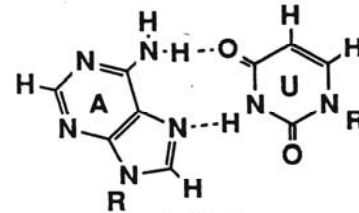
2-keto, 4-amino pyrimidine	Donor Acceptor Acceptor	C		G Acceptor Donor Donor	2-amino,6-keto purine
2,4-diketo pyrimidine	Acceptor Donor Acceptor	U		''A'' Donor Acceptor Donor	2,6-diamino purine
2-amino, 4-keto pyrimidine	Acceptor Acceptor Donor			Donor Donor Acceptor	2-keto, 6-amino purine
2,6-diamino pyrimidine	Donor Acceptor Donor			Acceptor Donor Acceptor	2,6-diketo purine
2-amino, 6-keto pyrazine	Donor Donor Acceptor			Acceptor Acceptor Donor	5-keto, 7-amino, 1,6,8-triaza indolicine
2-keto, 6-amino pyrazine	Acceptor Donor Donor			Donor Acceptor Acceptor	5-amino, 7-keto, 1,6,8-triaza indolicine



Watson-Crick

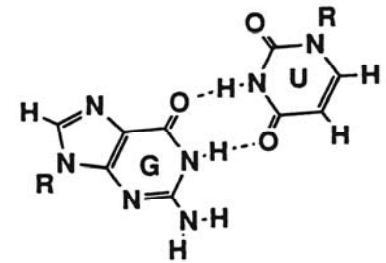
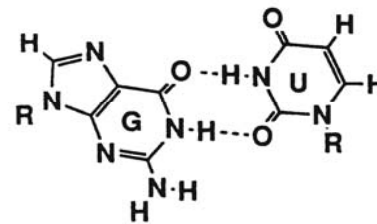


Reverse Watson-Crick



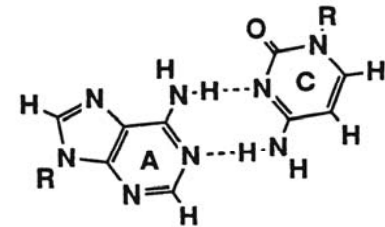
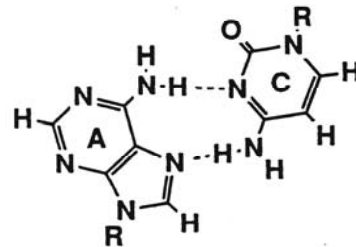
A-U Hoogsteen

A-U Reverse Hoogsteen



G-U Wobble

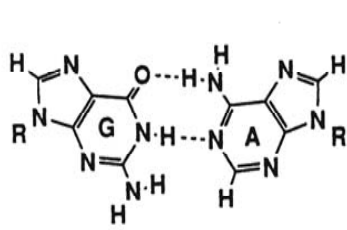
G-U Reverse Wobble



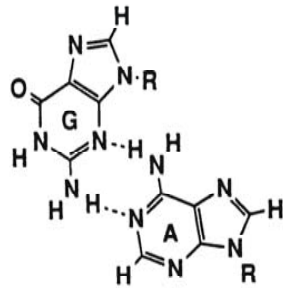
A-C Reverse Hoogsteen

A-C Reverse Wobble

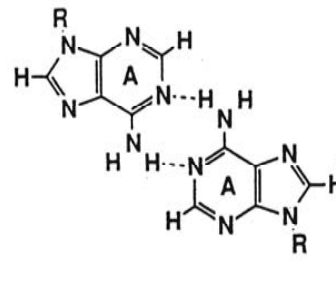
Classification of purine-pyrimidine base pairs



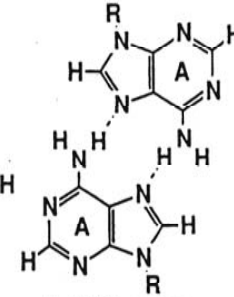
**G•A N1-N1,
carbonyl-amino**



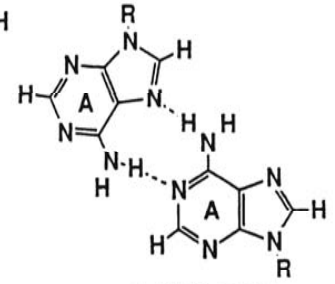
**G•A N3-amino,
amino-N1**



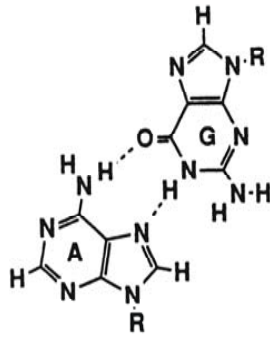
**A•A N1-amino,
symmetric**



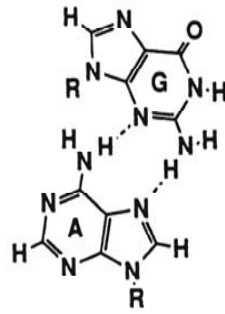
**A•A N7-amino,
symmetric**



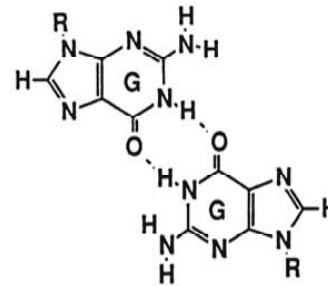
**A•A N1-amino,
N7-amino**



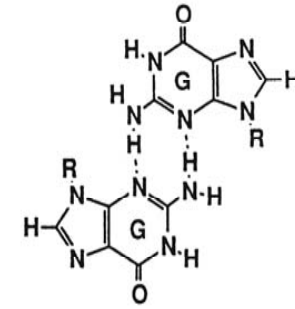
**A•G N7-N1,
amino-carbonyl**



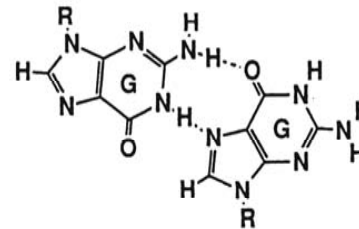
**A•G N7-amino,
amino-N3**



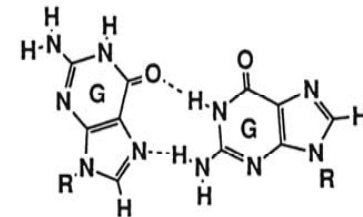
**G•G N1-carbonyl,
symmetric**



**G•G N3-amino,
symmetric**

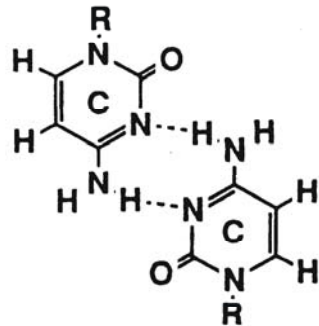


**G•G N7-N1,
carbonyl-amino**

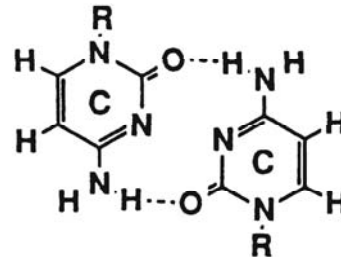


**G•G N1-carbonyl,
N7-amino**

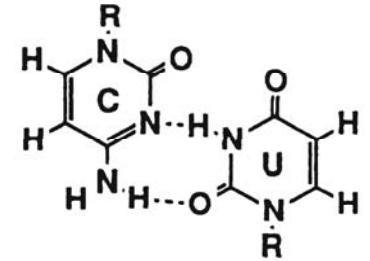
Classification of purine-purine base pairs



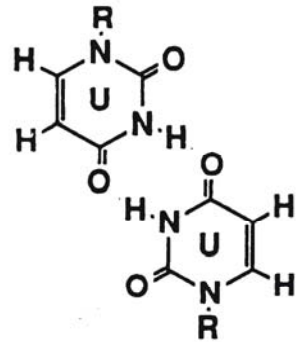
**C-C N3-amino,
symmetric**



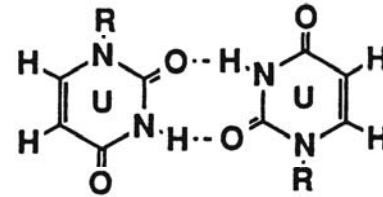
**C-C carbonyl-amino,
symmetric**



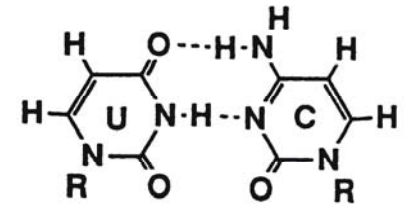
**C-U N3-N3,
2-carbonyl-amino**



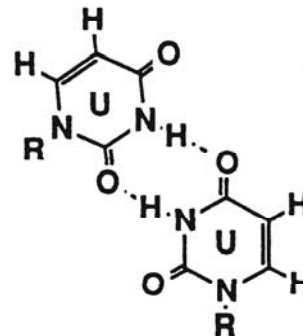
**U-U 4-carbonyl-N3,
symmetric**



**U-U 2-carbonyl-N3,
symmetric**



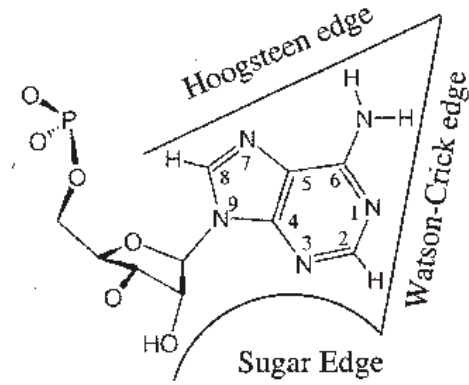
**U-C N3-N3,
4-carbonyl-amino**



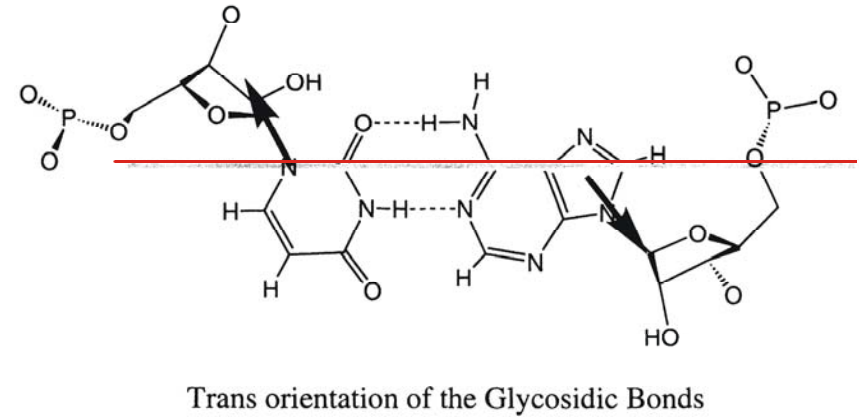
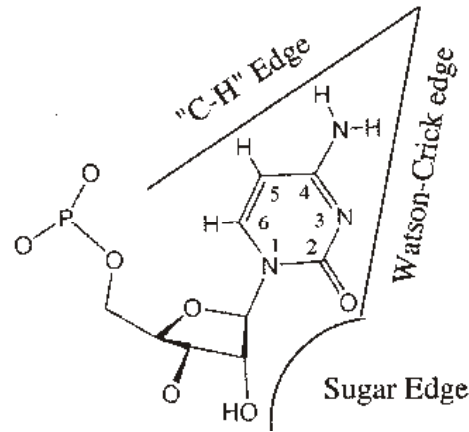
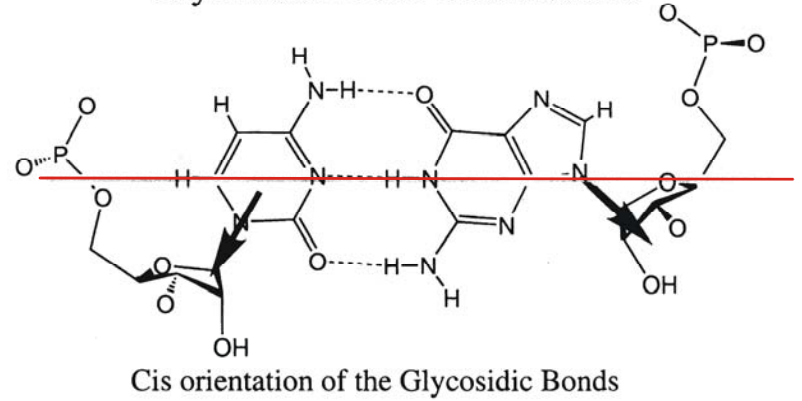
**U-U 2-carbonyl-N3,
4-carbonyl-N3**

Classification of pyrimidine-
pyrimidine base pairs

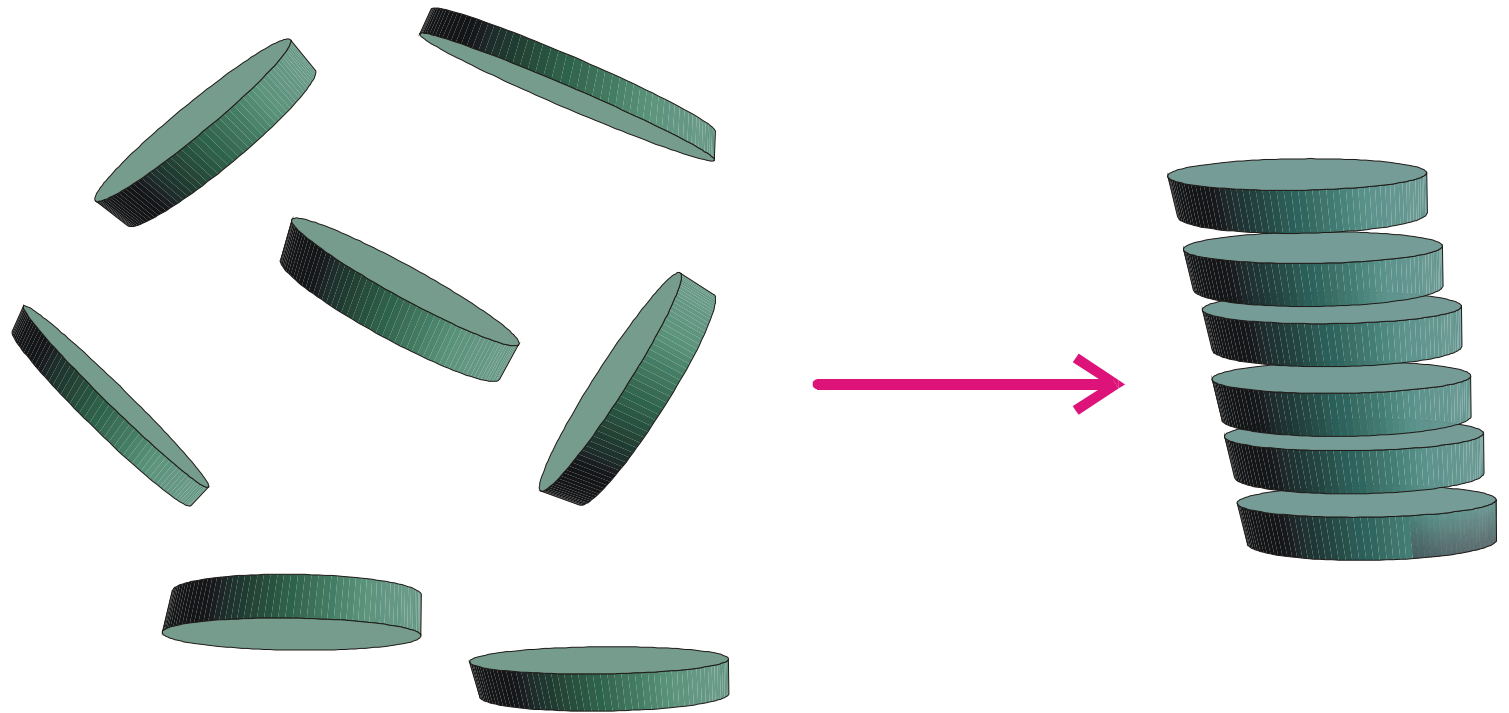
Interacting Edges



Glycosidic Bond Orientations

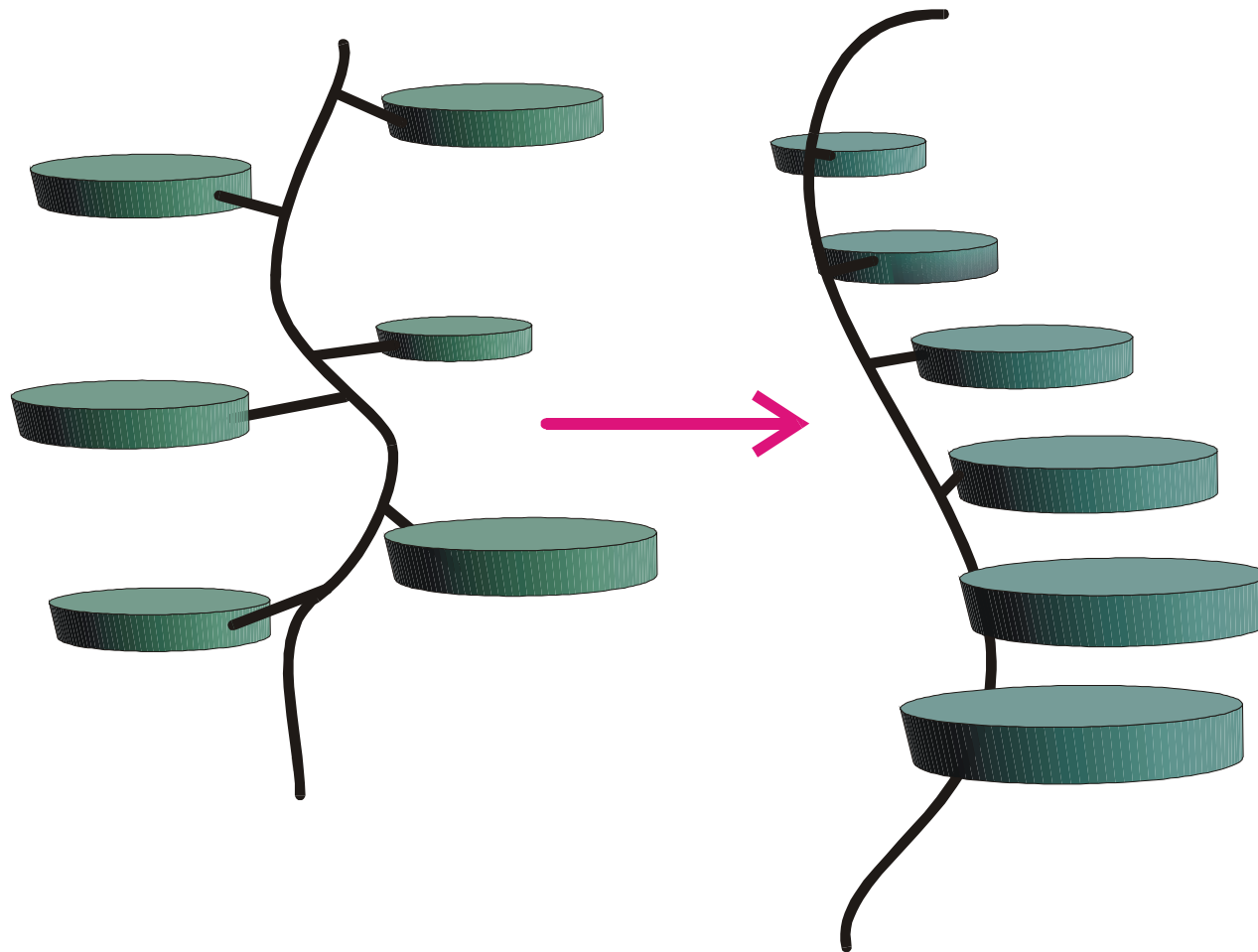


General classification
of base pairs



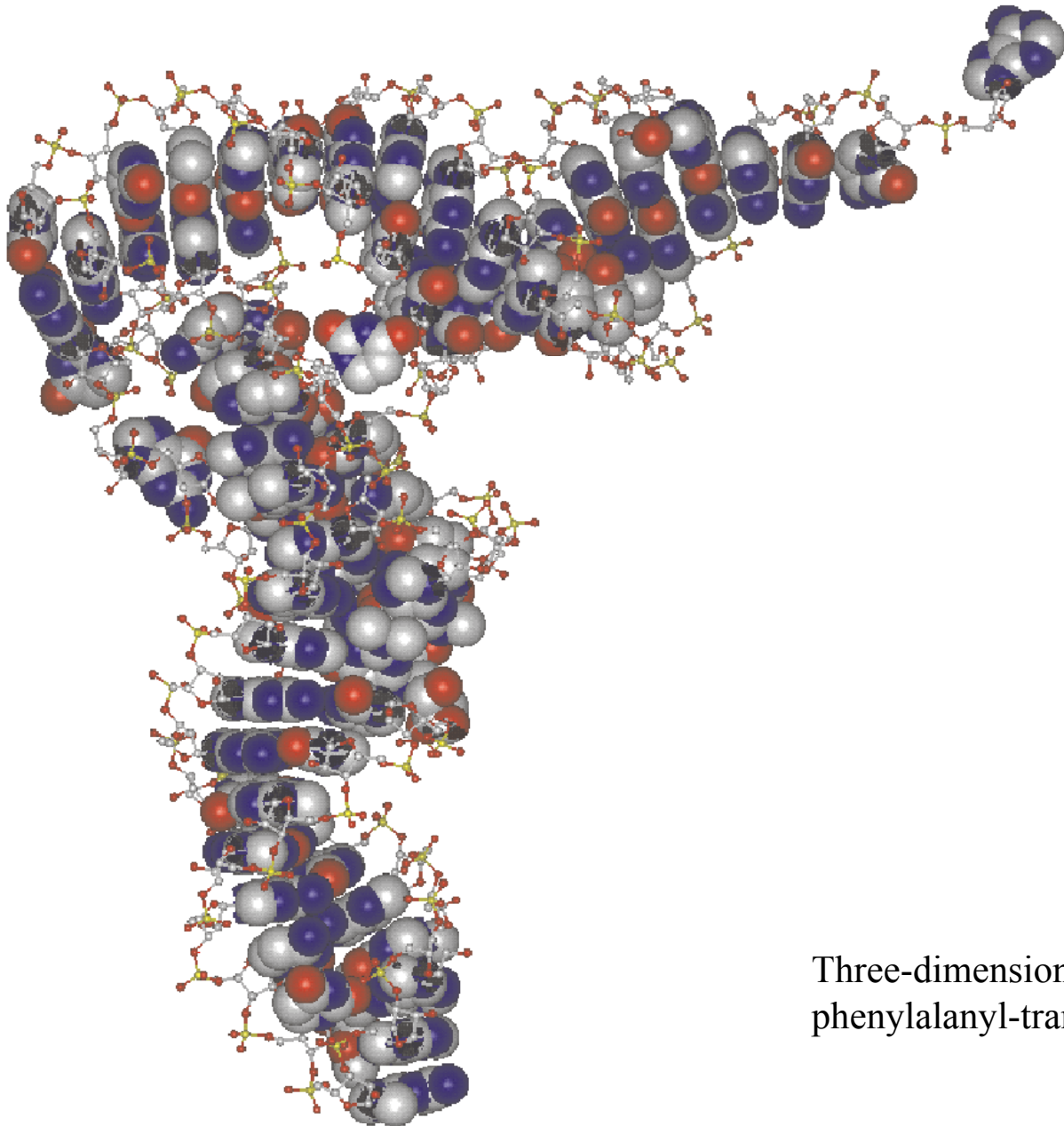
Stacking of heterocyclic aromatic molecules without sugar-phosphate backbone

Example: N₆,N₉-dimethyl adenine, D. Pörschke and F. Eggers, *Eur.J.Biochem.* **26**:490-498 (1972)



Stacking of RNA single strands

Example: poly-**A**, D.Pörschke. Elementary steps of base recognition and helix-coil transitions in nucleic acids. In: I.Pecht and R.Rigler, eds. Chemical Relaxation in Molecular Biology, pp.191-218. Springer-Verlag, Berlin 1977.

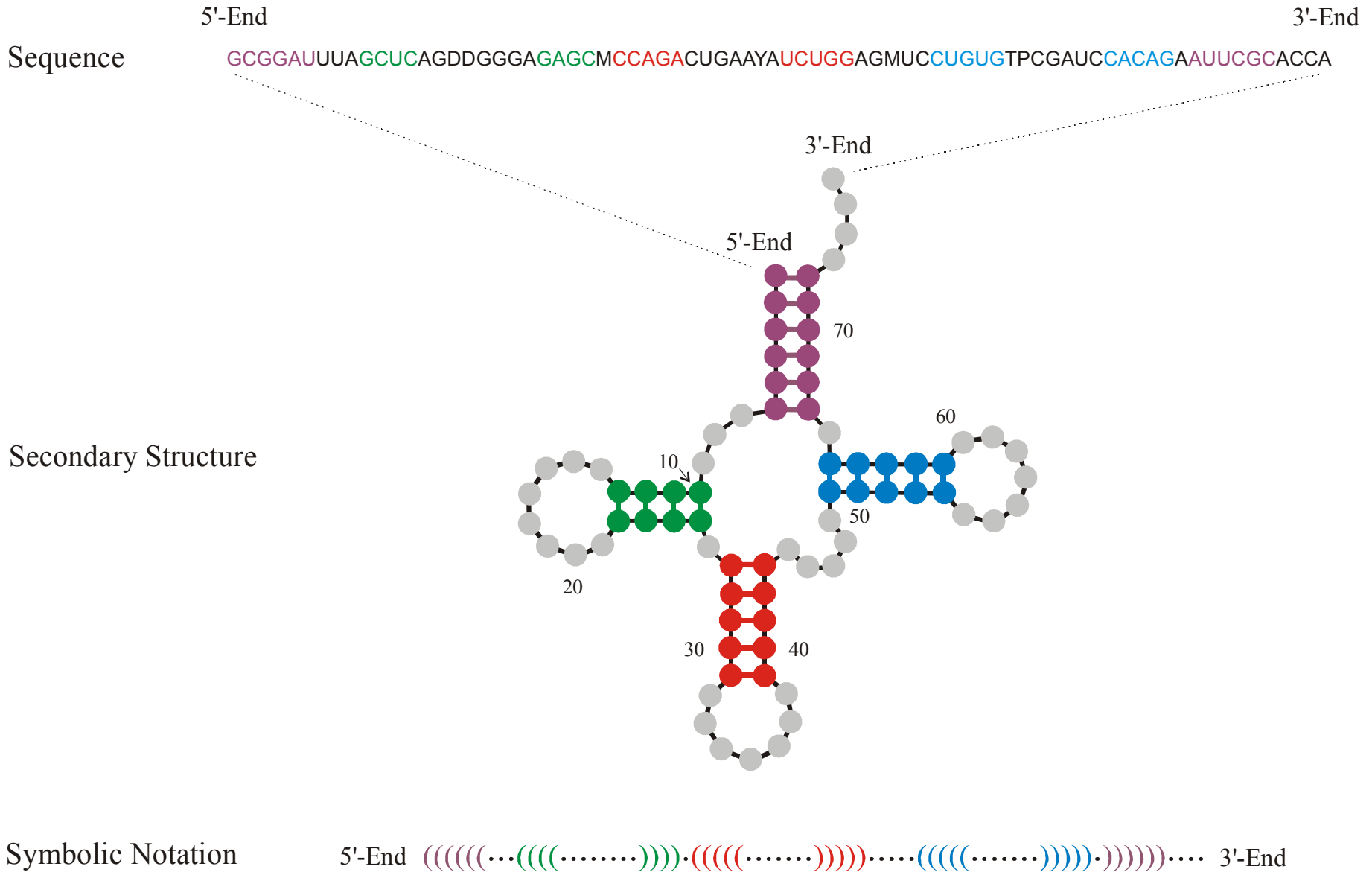


Three-dimensional structure of
phenylalanyl-transfer-RNA

RNA Secondary Structures and their Properties

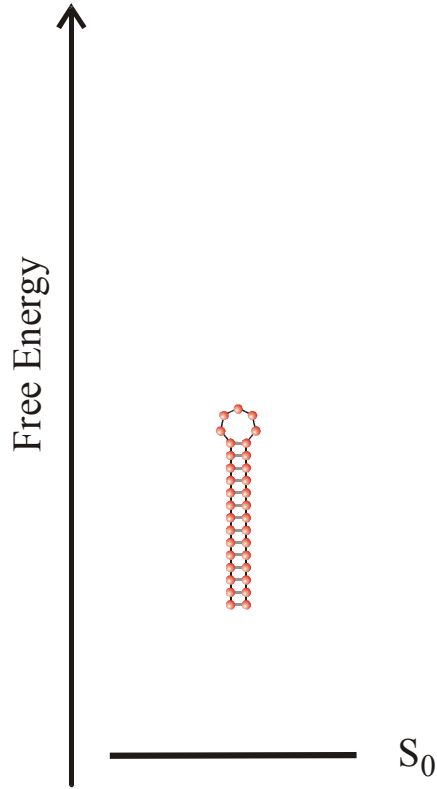
RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudoknots. Secondary structures are folding intermediates in the formation of full three-dimensional structures.

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



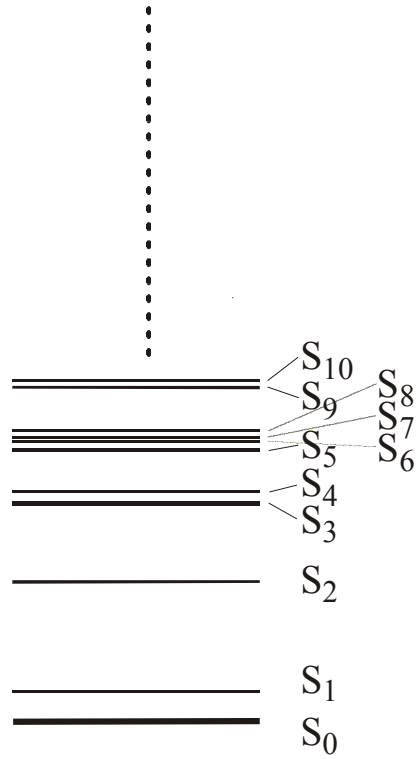
Definition of the secondary structure of phenylalanyl-tRNA

$T = 0 \text{ K}, t \rightarrow \infty$



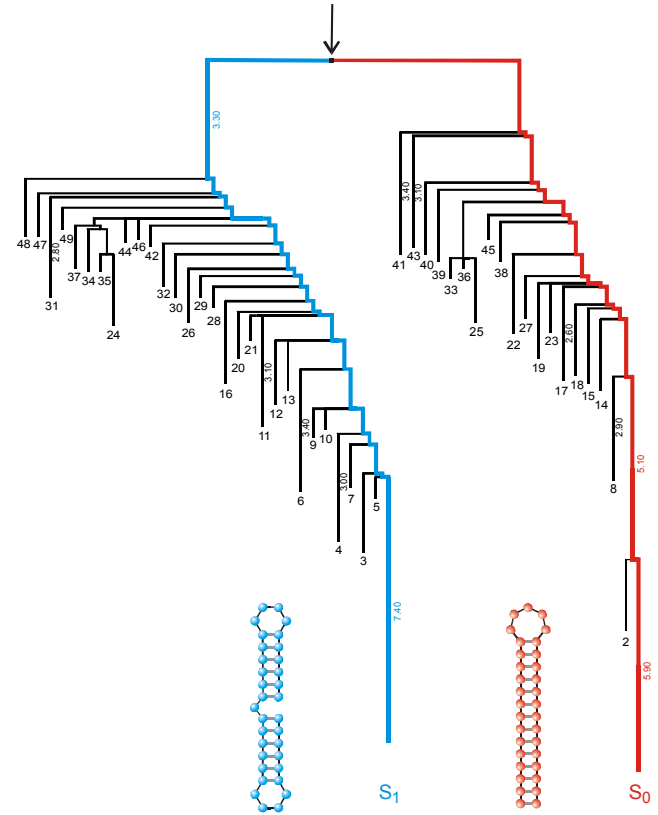
Minimum Free Energy Structure

$T > 0 \text{ K}, t \rightarrow \infty$



Suboptimal Structures

$T > 0 \text{ K}, t \text{ finite}$



Kinetic Structures

Different notions of RNA structure

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.

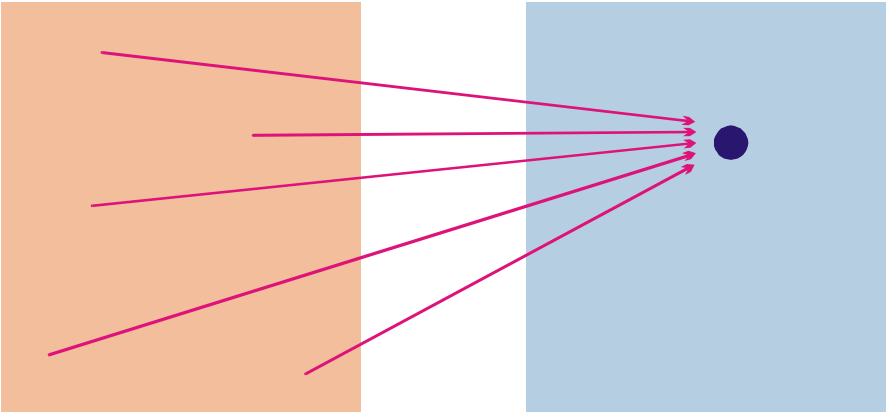
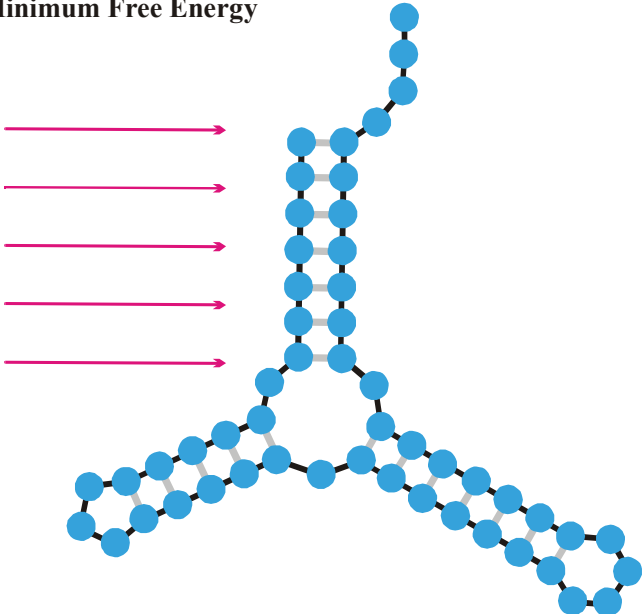
M.Zuker and P.Stiegler. *Nucleic Acids Res.* **9**:133-148 (1981)

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)

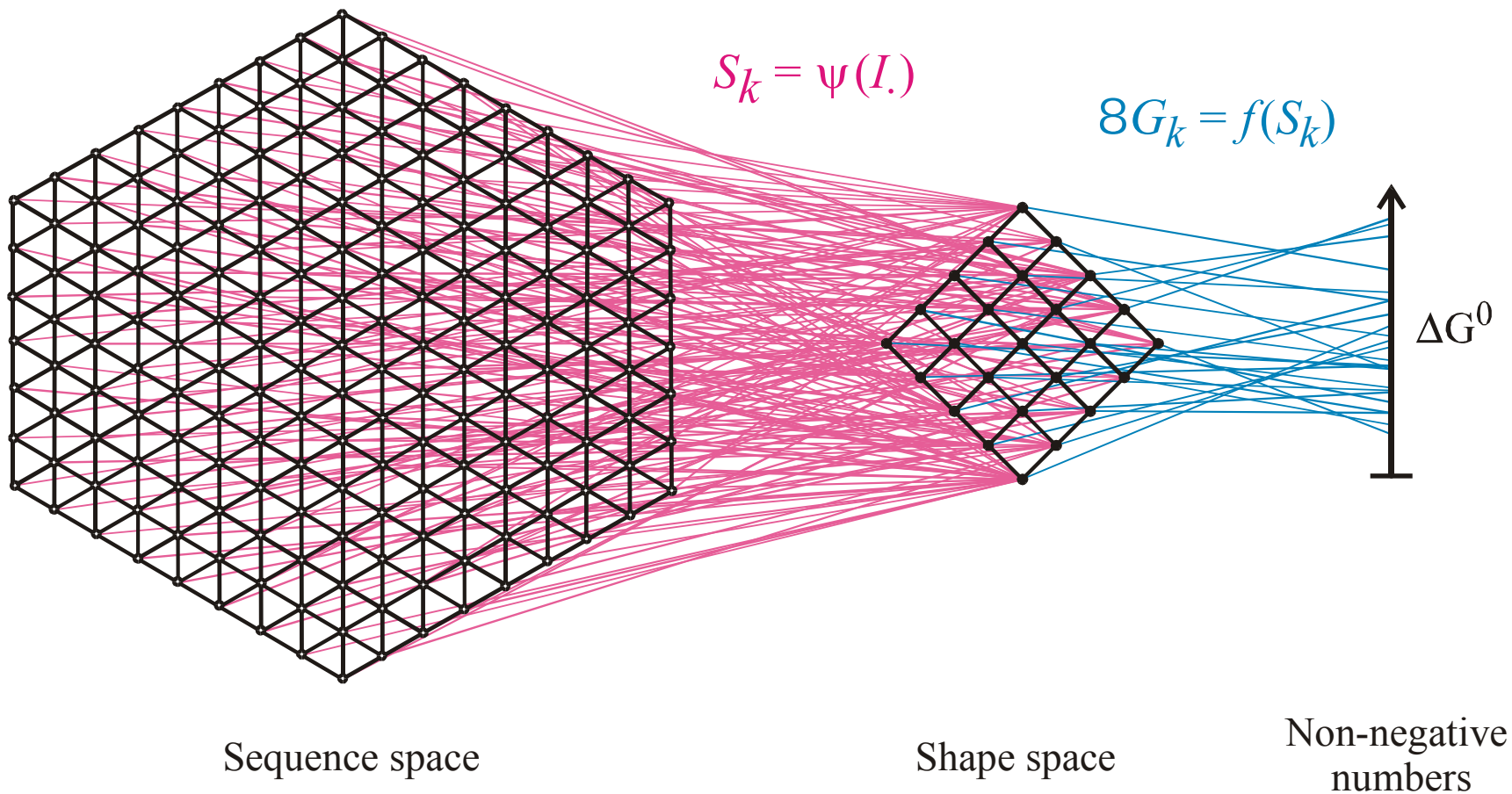
**Criterion of
Minimum Free Energy**

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUUUCUGG
UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG
CAUUGGUGCUAAUGAUUUAGGGCUGUAUJCCUGUAUAGCGAUCAGUGUCCG
GUAGGCCUCUUGACAUAAGAUUUUUCCAUGGUGGGAGAUGGCCAUUGCAG

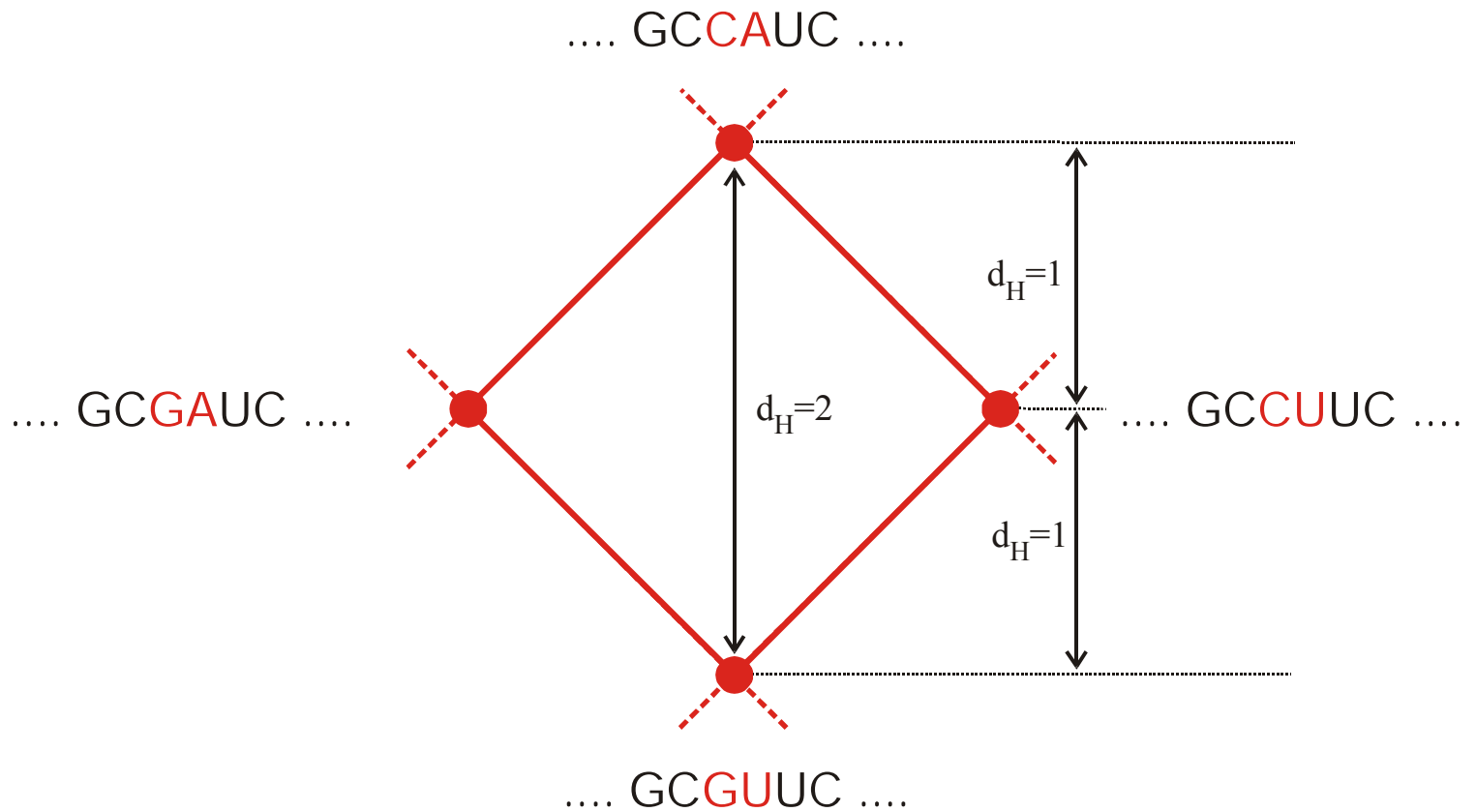


Sequence Space

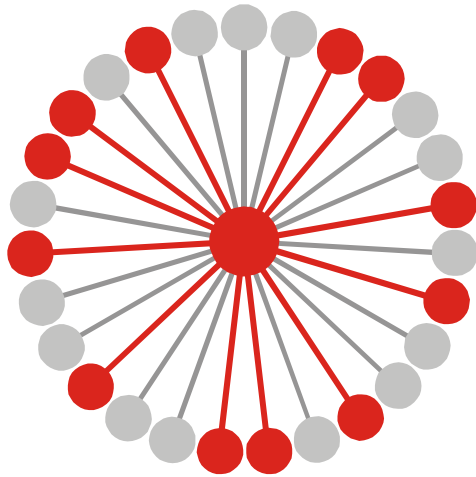
Shape Space



Mapping from sequence space into phenotype space and into free energies



Point mutations as moves in sequence space



$$G_k = m^{-1}(S_k) \mid oI_j \mid m(I_j) = S_k q$$

$$\lambda_j = 12 / 27, \quad \bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \hat{\lambda}_j(k)}{|G_k|}$$

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

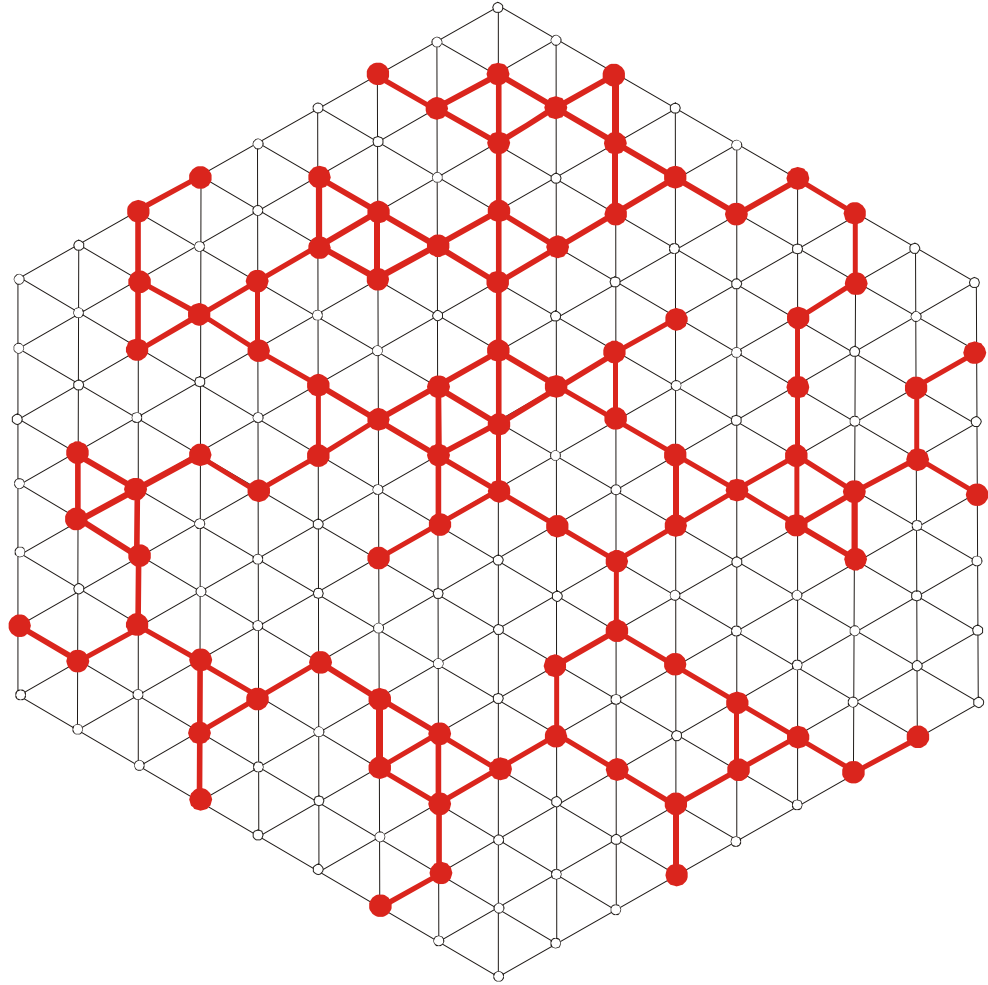
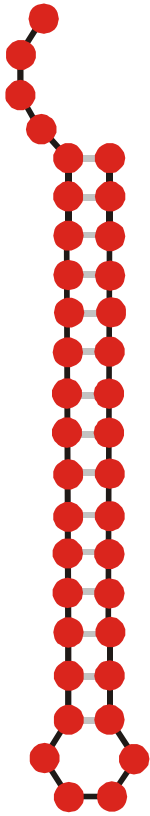
Alphabet Size κ : **AUGC** $\kappa = 4$

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

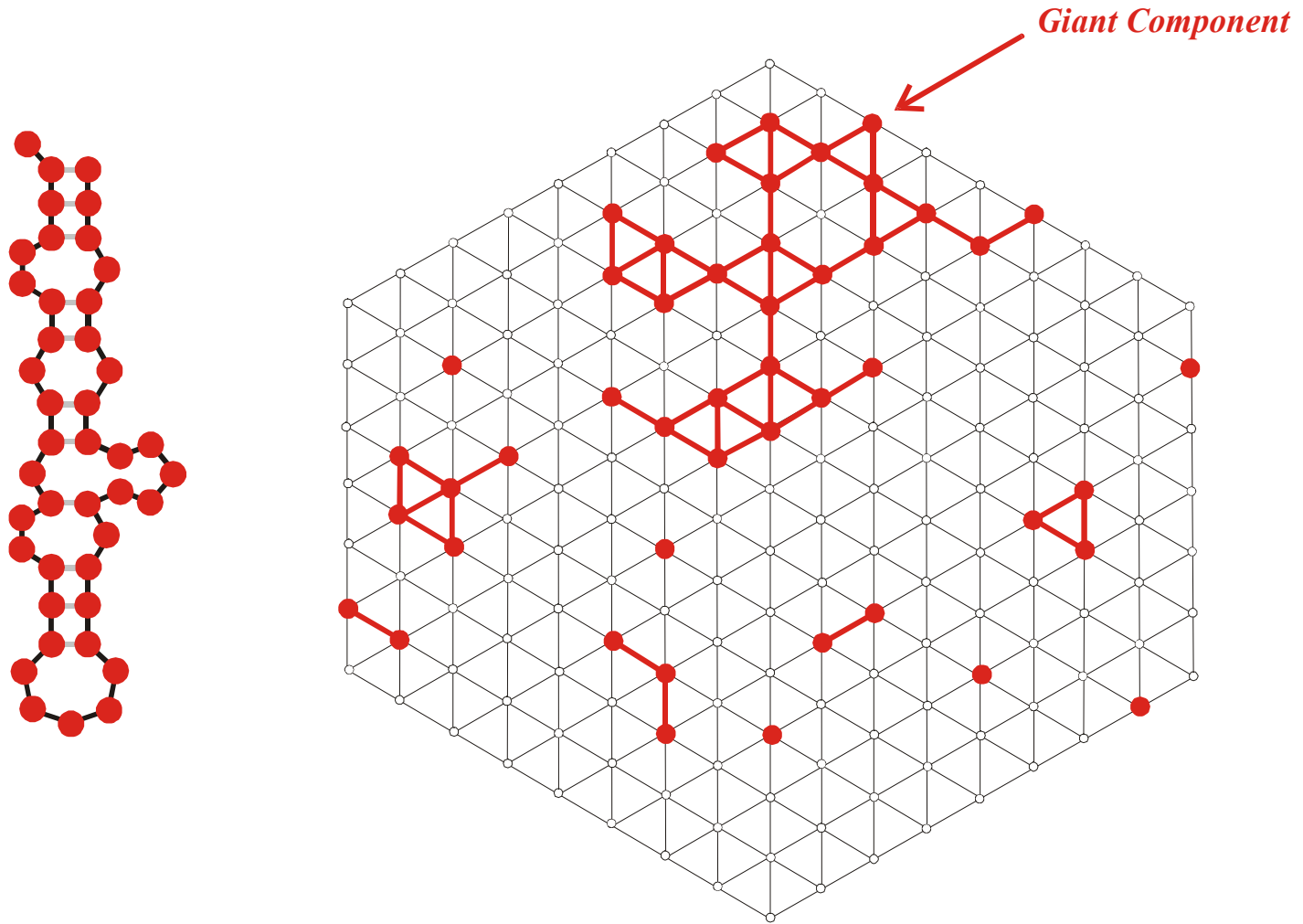
$\bar{\lambda}_k < \lambda_{cr}$ Network G_k is **not** connected

κ	λ_{cr}
2	0.5
3	0.4226
4	0.3700

Mean degree of neutrality and connectivity of neutral networks



A connected neutral network

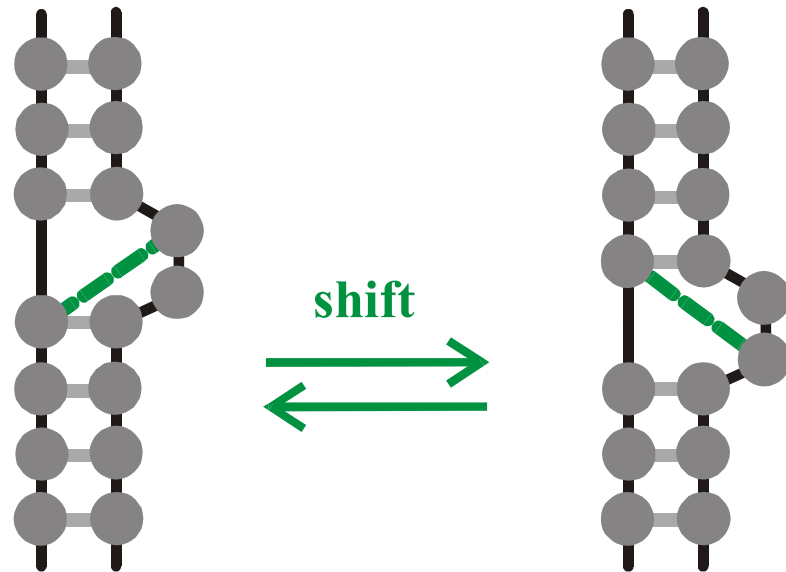
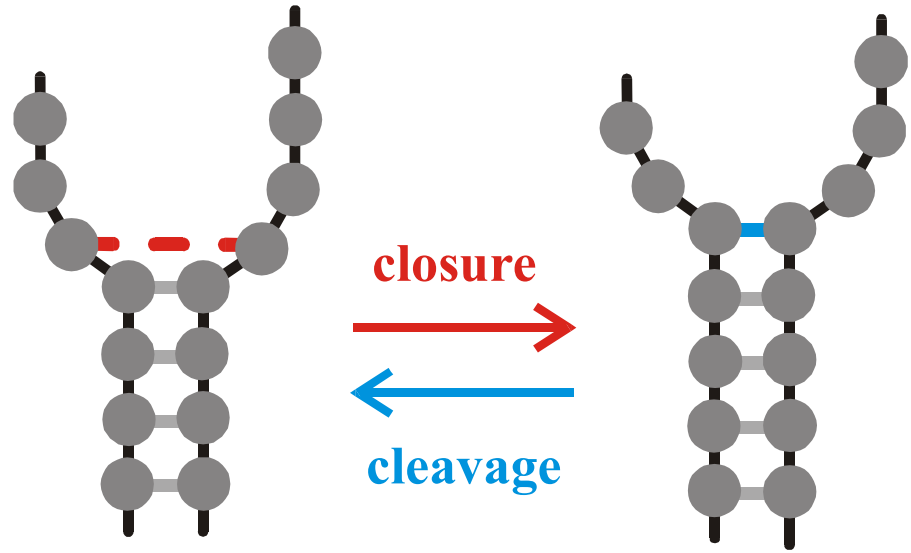


A multi-component neutral network

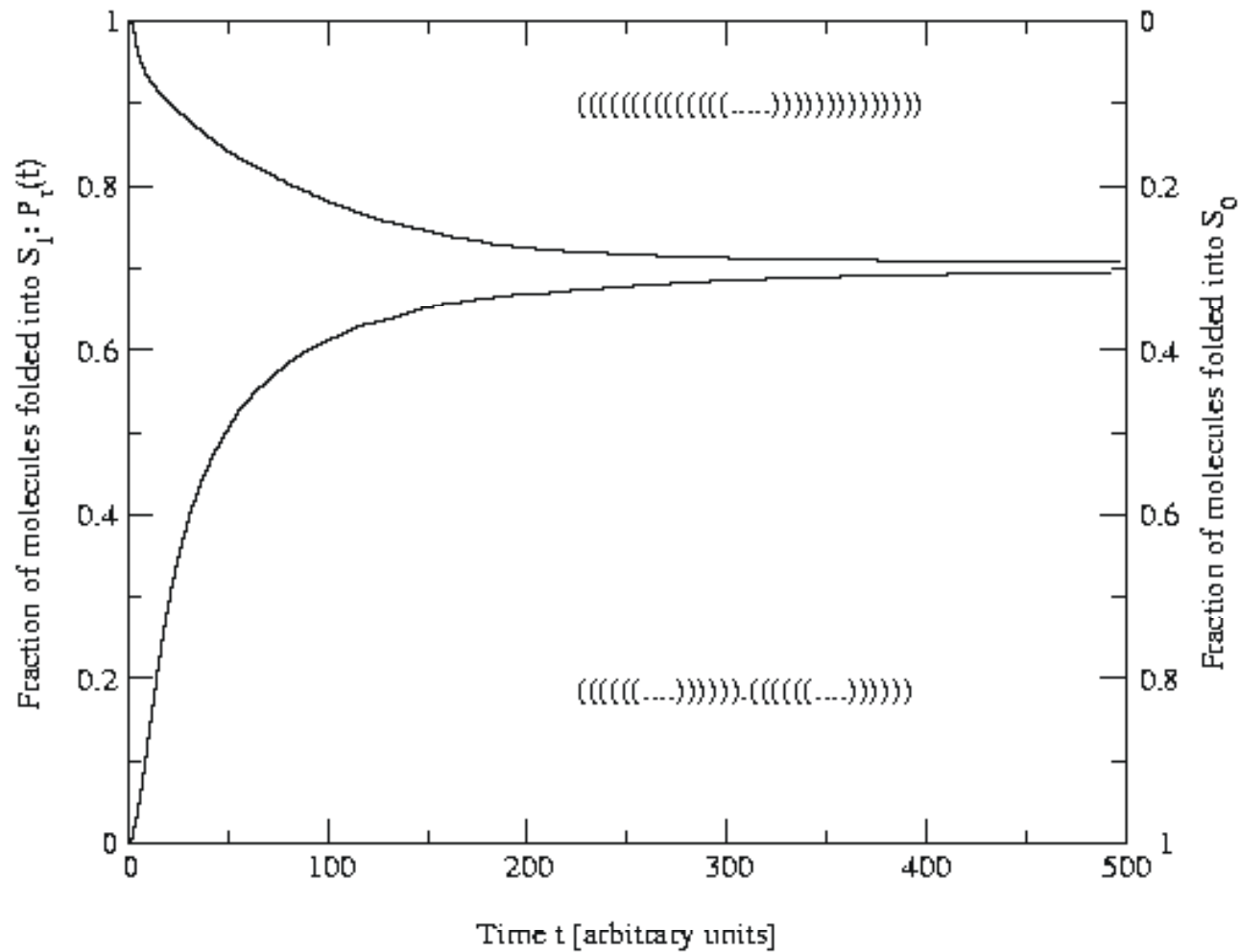
Kinetic Folding of RNA at Elementary Step Resolution

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)

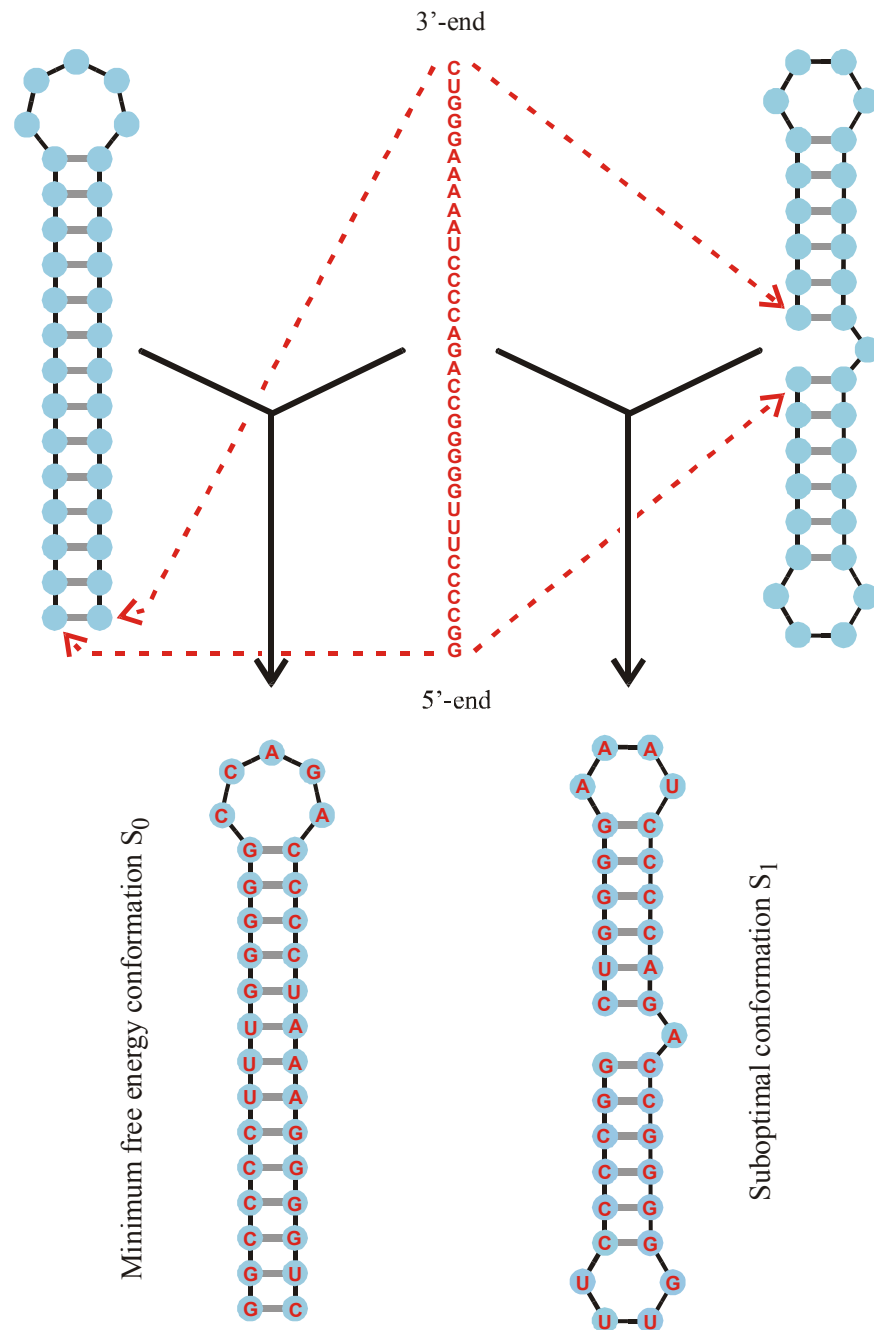


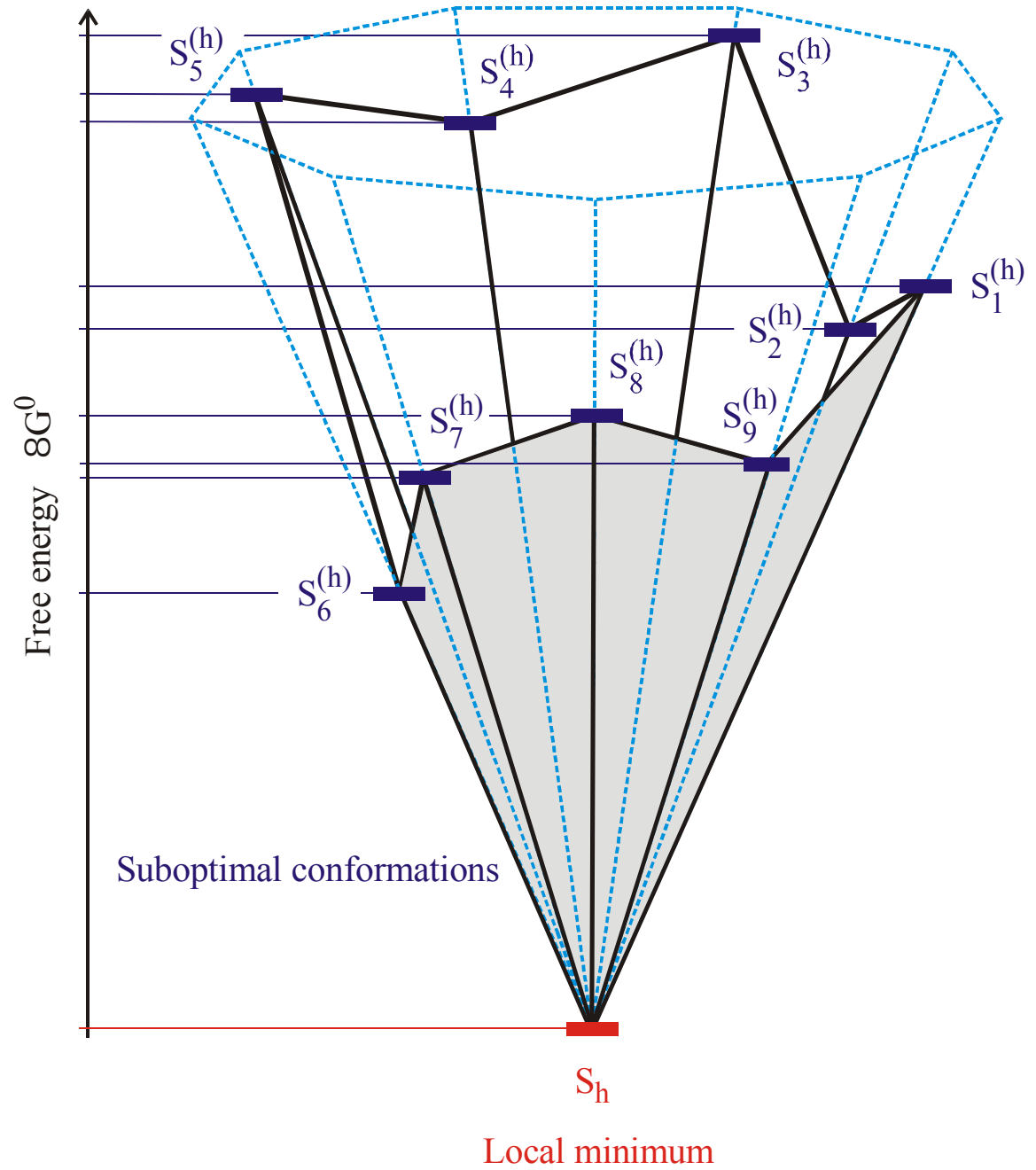
Move set for elementary steps
in kinetic RNA folding



Folding dynamics of the sequence **GGCCCUUUGGGGGCCAGACCCUAAAAAGGGUC**

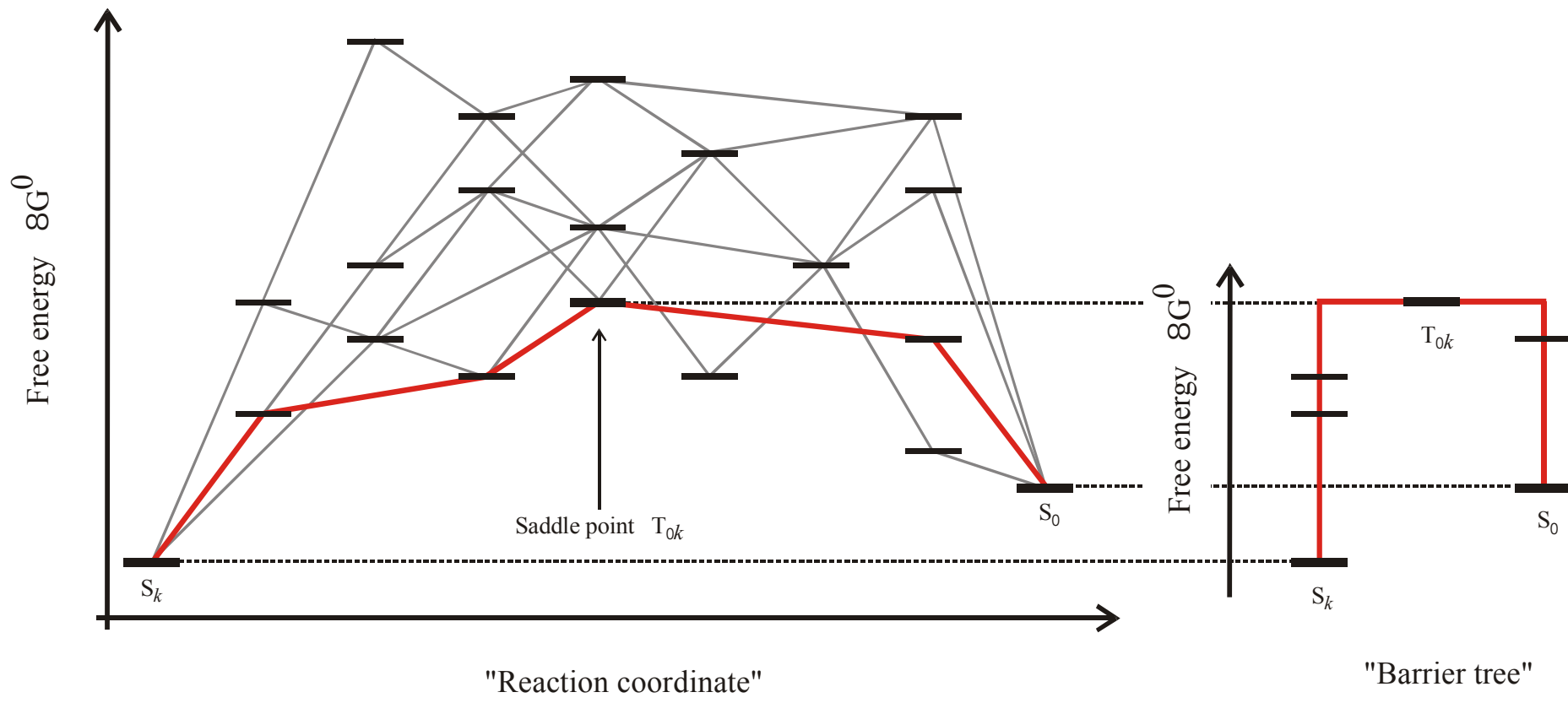
One sequence is compatible with two structures



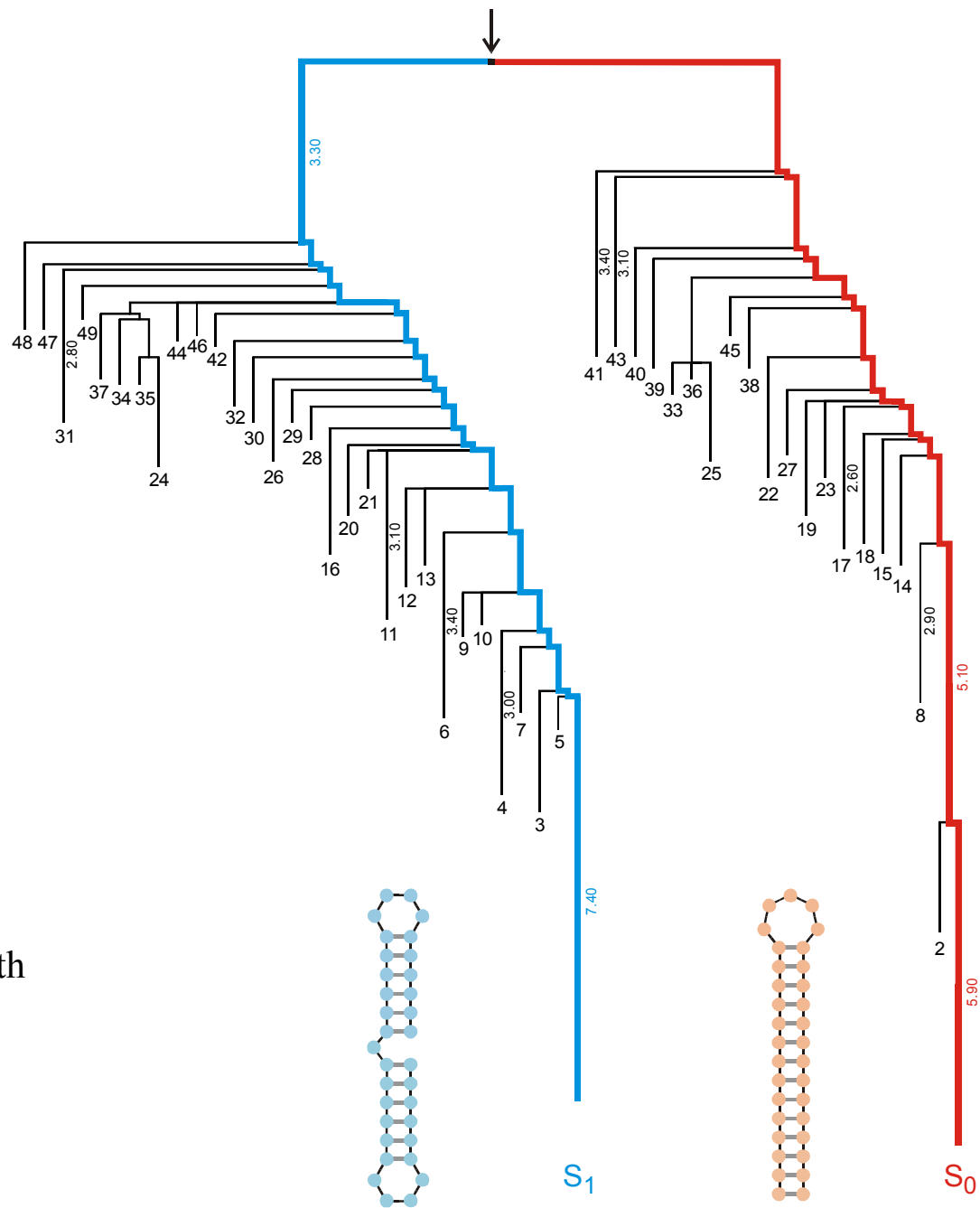


Search for local minima in conformation space

Local minimum

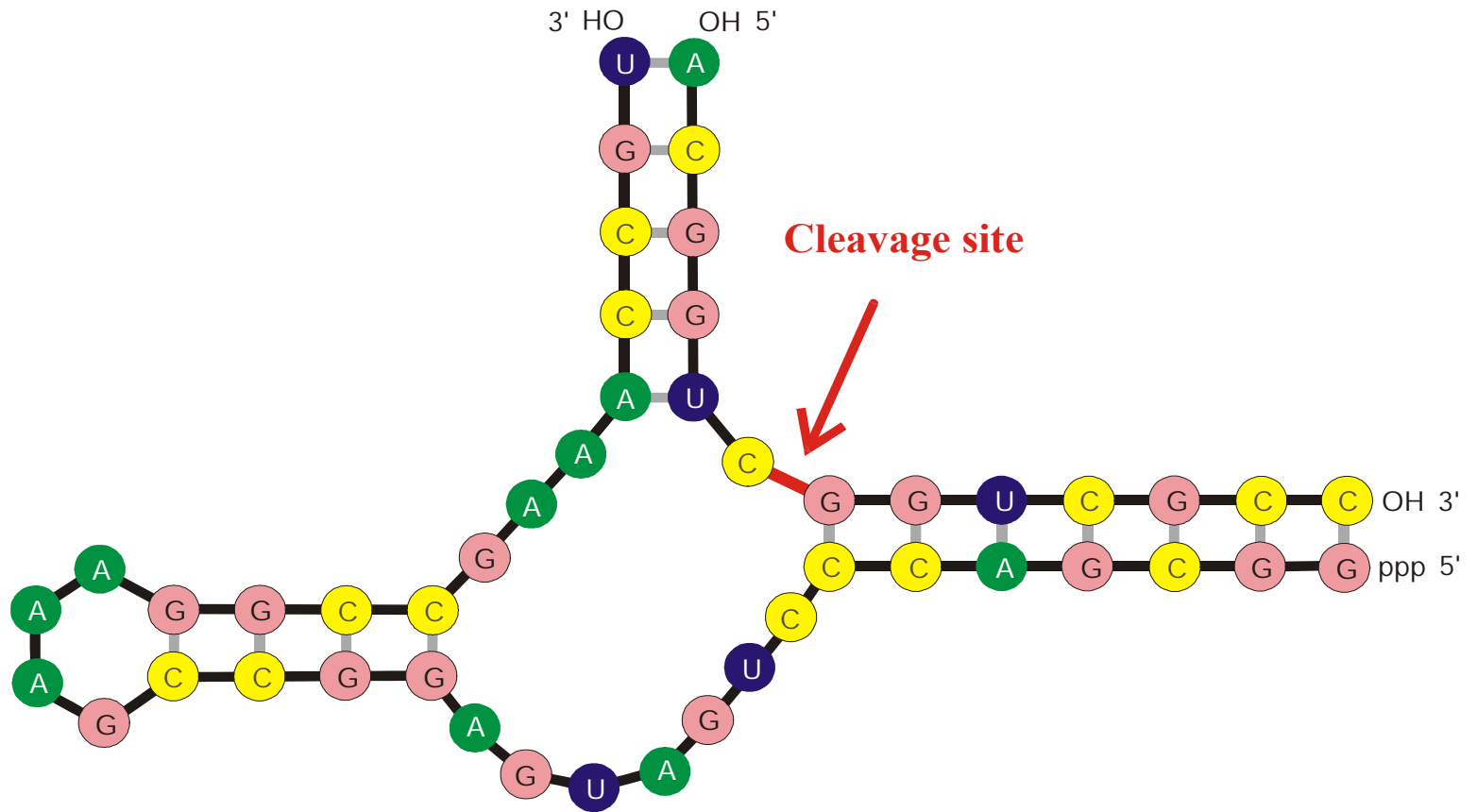


Barrier tree of a sequence with two conformations



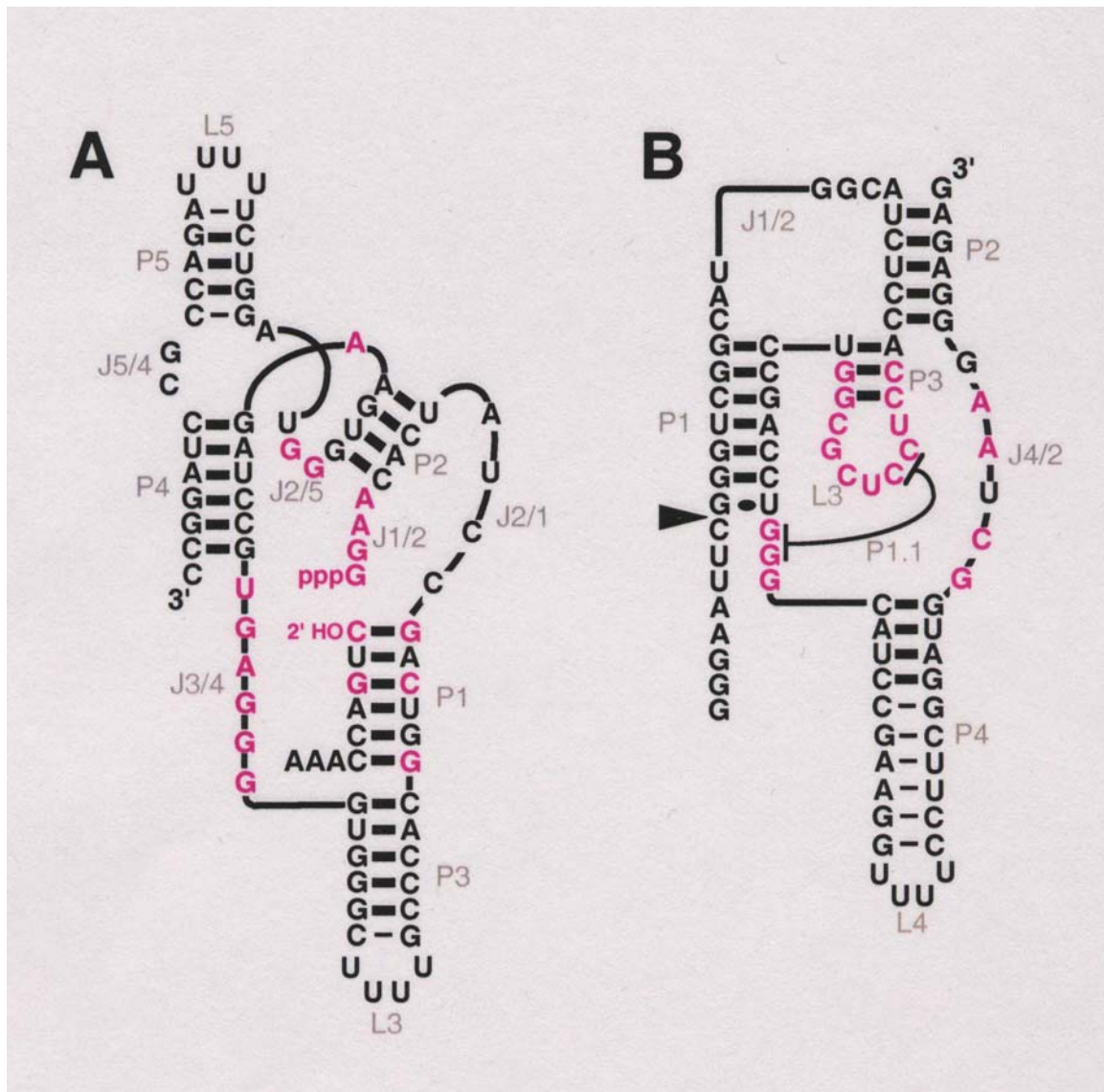
A ribozyme switch

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

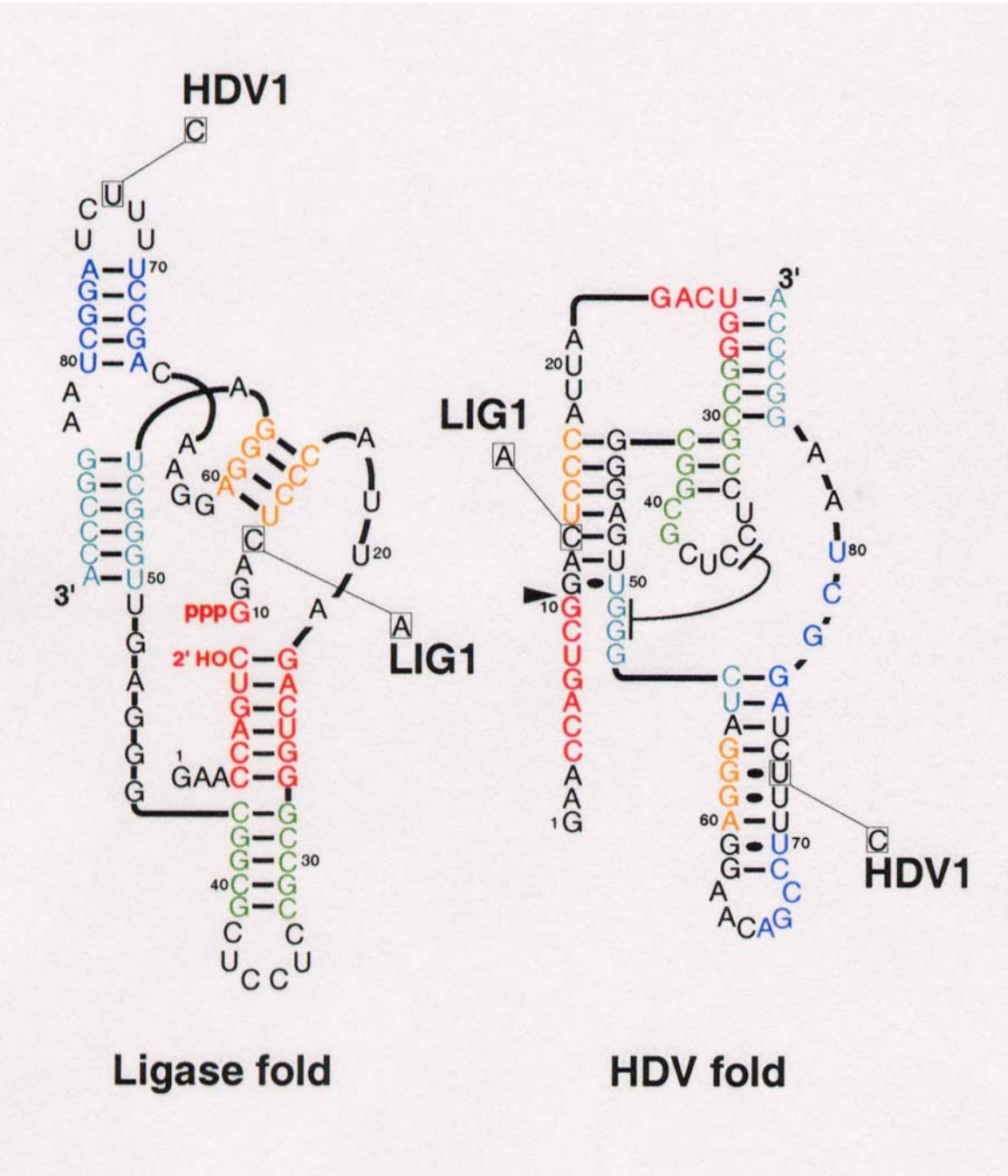


The "hammerhead" ribozyme

The smallest known
catalytically active
RNA molecule



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



The sequence at the *intersection*:

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



S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES¹

■ CHRISTIAN REIDYS*, †, PETER F. STADLER*, ‡
 and PETER SCHUSTER*, ‡, §, ¶

*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

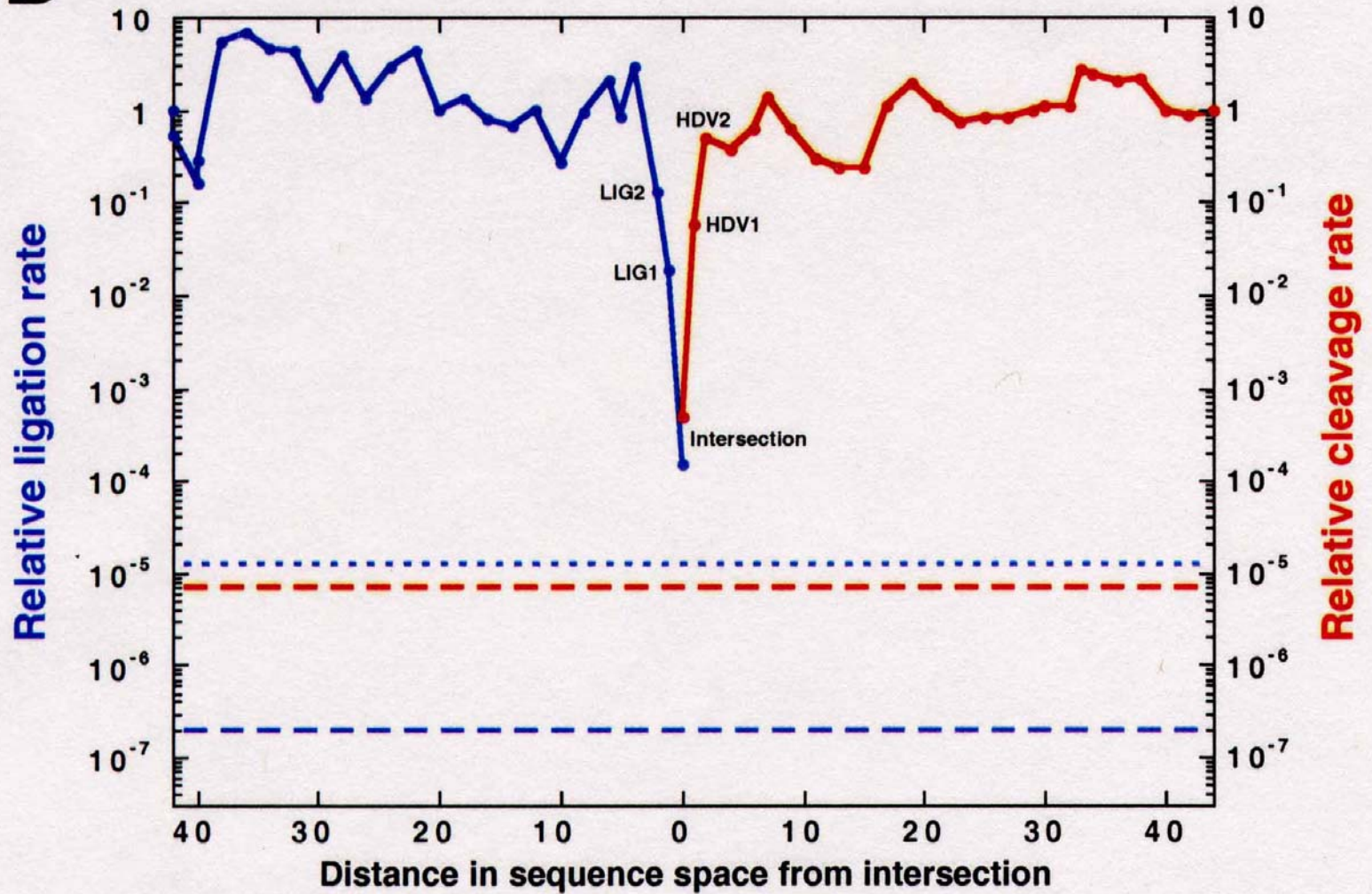
THEOREM 5. INTERSECTION-THEOREM. *Let s and s' be arbitrary secondary structures and $C[s], C[s']$ their corresponding compatible sequences. Then,*

$$C[s] \cap C[s'] \neq \emptyset.$$

Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s' . Then $f(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \dots, x_n\}$. 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*

B

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

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

PETER SCHUSTER^{1,2,3}, WALTER FONTANA³, PETER F. STADLER^{2,3}
AND IVO L. HOFACKER²

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.

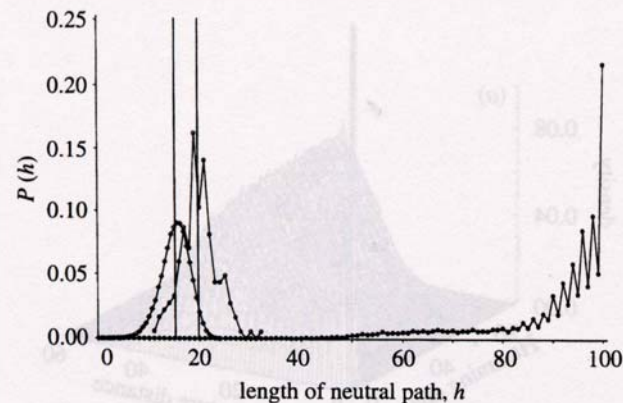


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and 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.* 1993a; 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).

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

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

**Bärbel Stadler, Ulrike Mückstein, Andreas Wernitznig, Stefanie Widder,
Stefan Wuchty**, Universität Wien, AT

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