Inverse Folding and Sequence-Structure Maps of Ribonucleic Acids

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Inverse Problem Workshop

IPAM, UCLA, 22.10.2003

Web-Page for further information:

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

1. The role of RNA in the cell and the notion of structure

2. RNA folding

3. Inverse folding of RNA

4. Sequence structure maps, neutral networks, and intersection

5. Reference to experimental data

6. Concluding remarks

1. **The role of RNA in the cell and the notion of structure**
2. RNA folding
3. Inverse folding of RNA
4. Sequence structure maps, neutral networks, and intersection
5. Reference to experimental data
6. Concluding remarks

RNA *as transmitter of genetic information*

DNA

↓ transcription

...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...

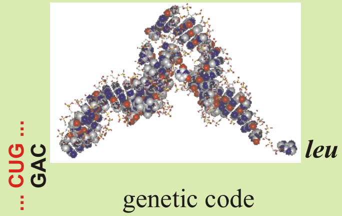
messenger-**RNA**

↓ translation

protein

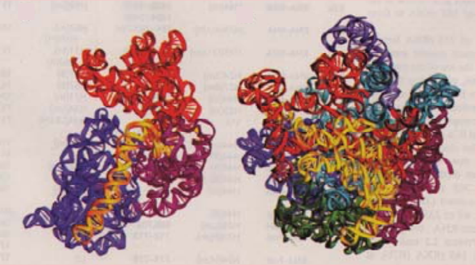
RNA as **working copy** of genetic information

RNA *as adapter molecule*



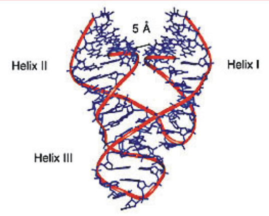
genetic code

RNA *is the catalytic subunit in supramolecular complexes*



The ribosome is a ribozyme !

RNA *as catalyst*



ribozyme

RNA

RNA *is modified by epigenetic control*

RNA editing

Alternative splicing of messenger **RNA**

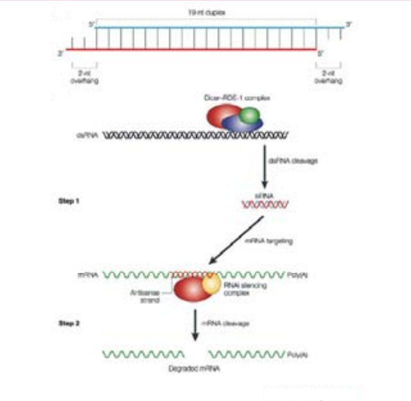
The RNA *world as a precursor of the current* DNA + protein *biology*

RNA *as carrier of genetic information*

RNA viruses and retroviruses

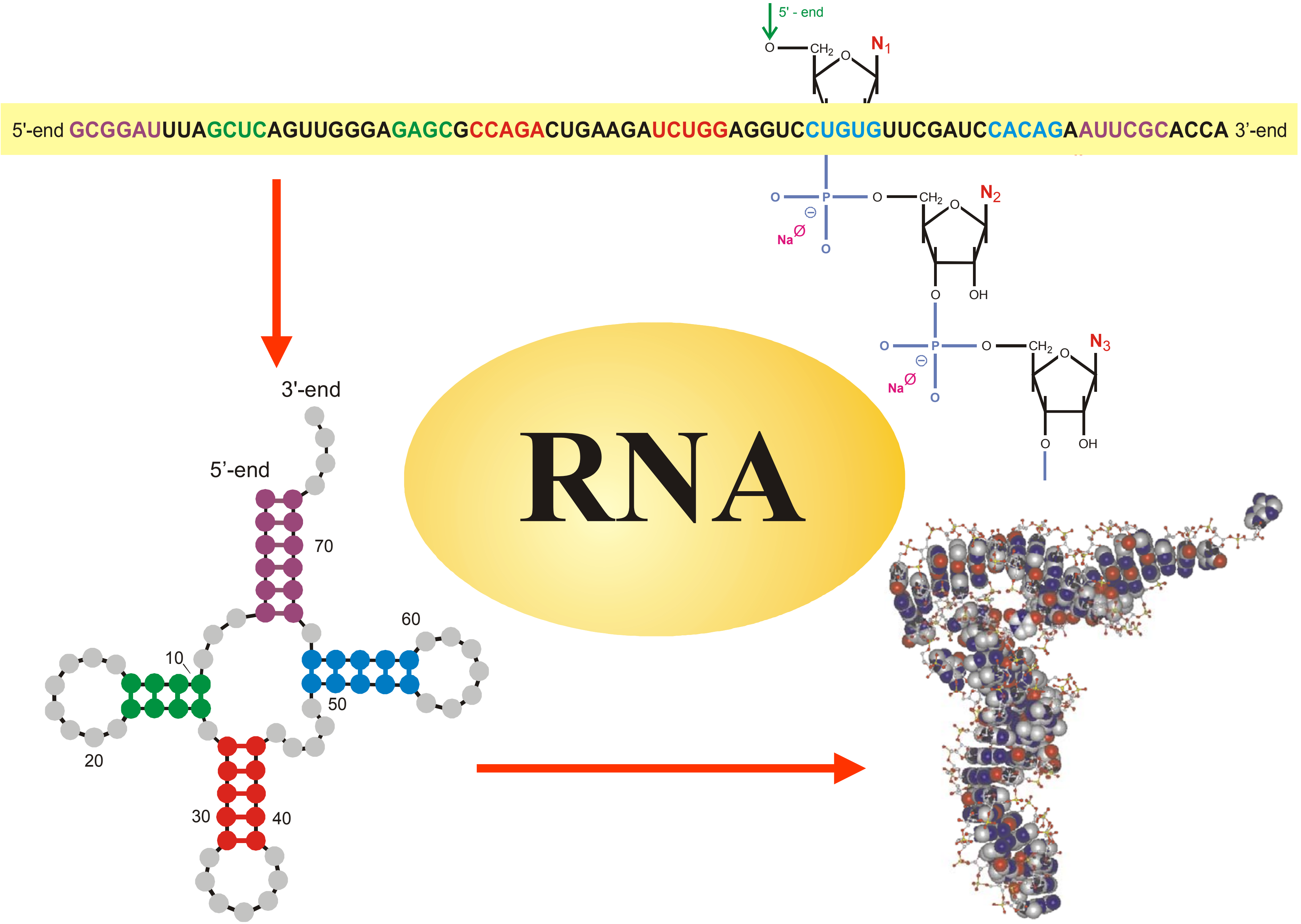
RNA as information carrier in evolution *in vitro* and evolutionary biotechnology

RNA *as regulator of gene expression*



gene silencing by small interfering RNAs

Functions of RNA molecules



5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end

RNA

Definition of RNA structure

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

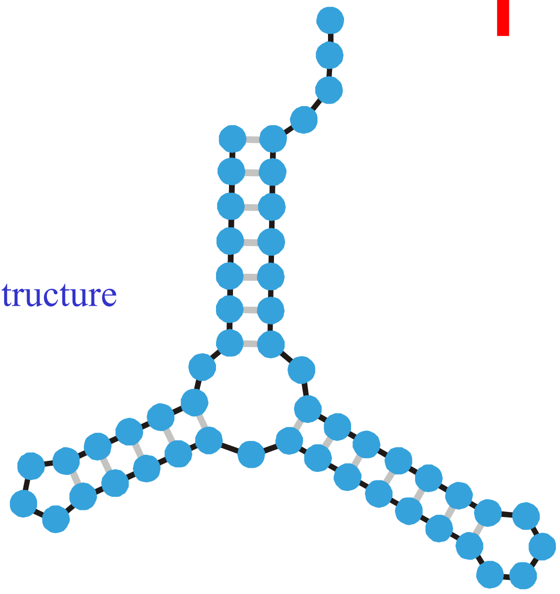
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

Inverse folding of RNA:

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

RNA structure



Sequence, structure, and function

Definition and **physical relevance** of RNA secondary structures

RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots.

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

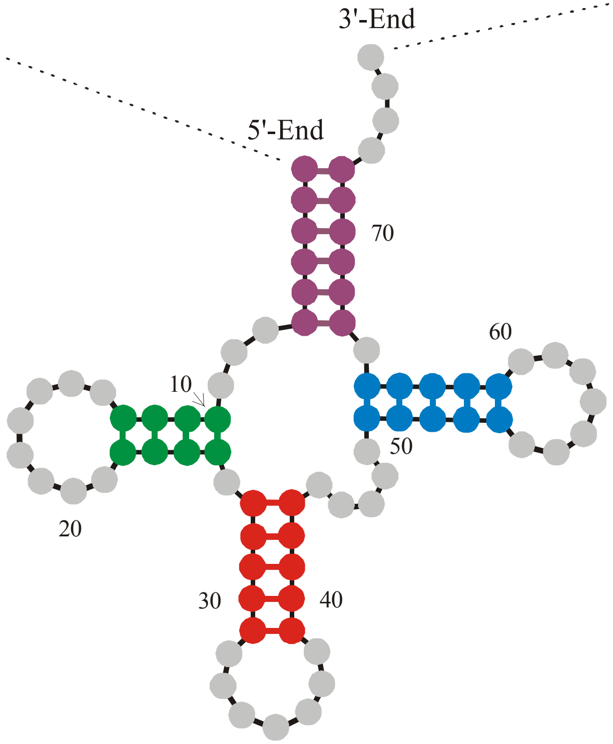
„**Secondary structures are folding intermediates in the formation of full three-dimensional structures**.“

Sequence

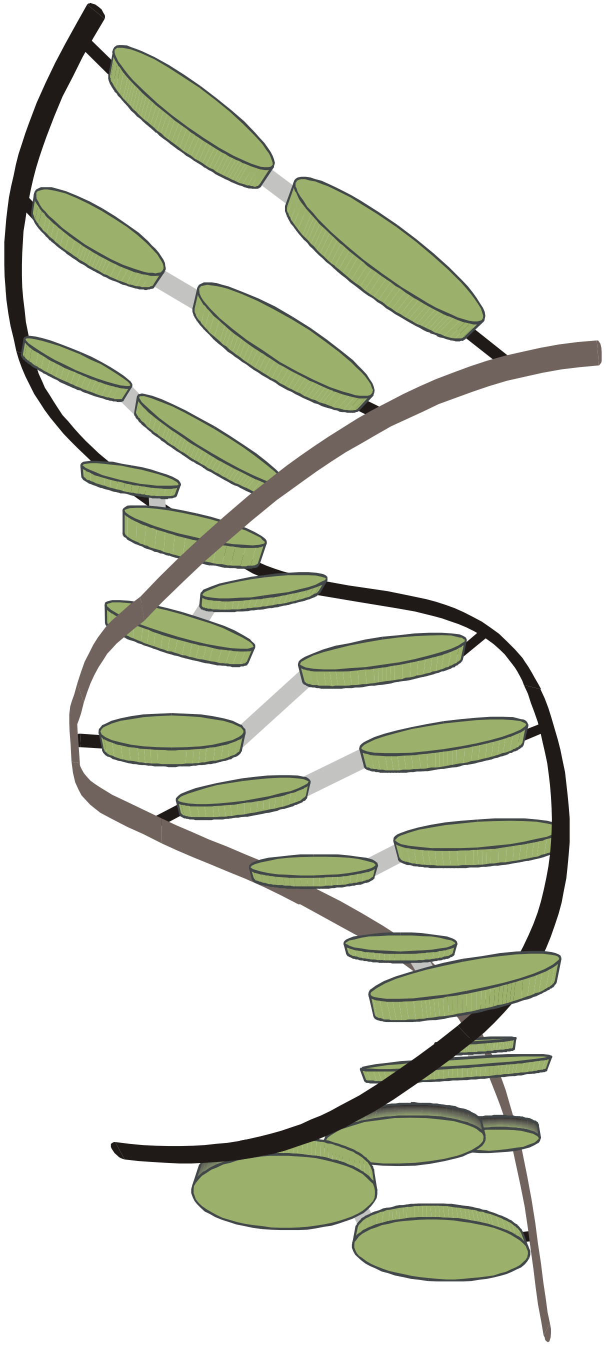
5'-End 3'-End

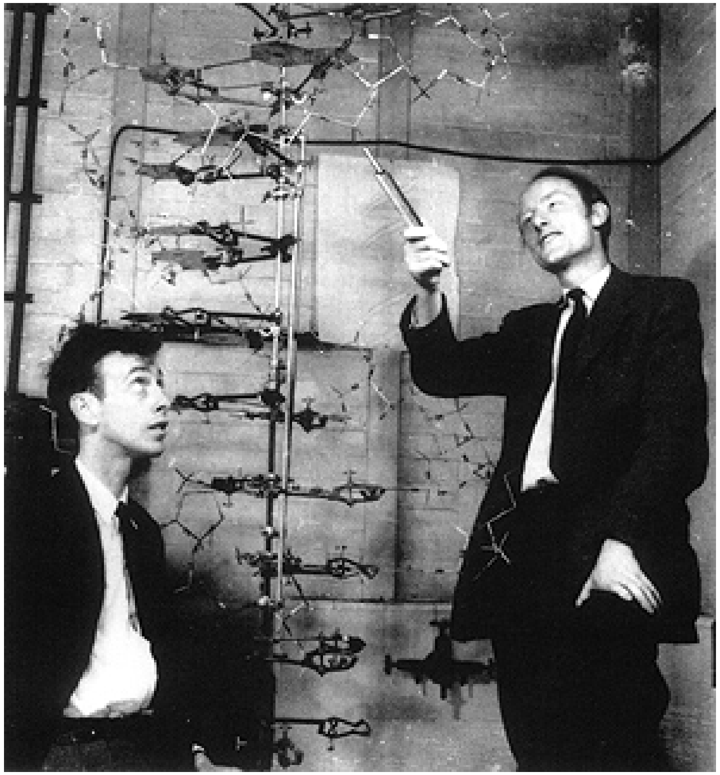
GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure



The RNA secondary structure lists the double helical stretches or stacks of a folded single strand molecule

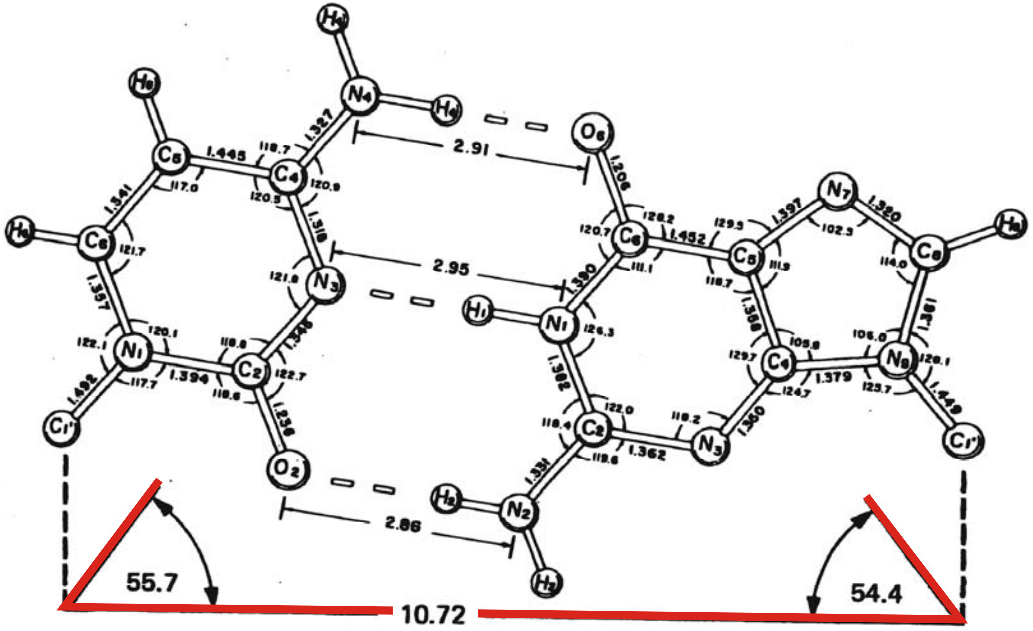




James D. Watson, 1928- , and Francis Crick, 1916- ,
Nobel Prize 1962

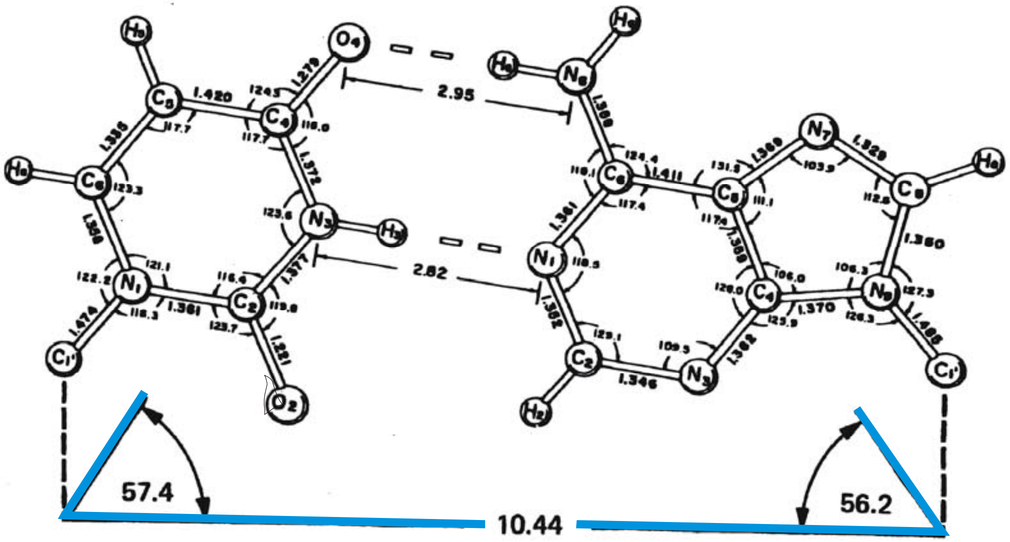
1953 – 2003 fifty years double helix

The three-dimensional structure of a
short double helical stack of B-DNA



Canonical Watson-Crick base pairs:

cytosine – guanine
uracil – adenine



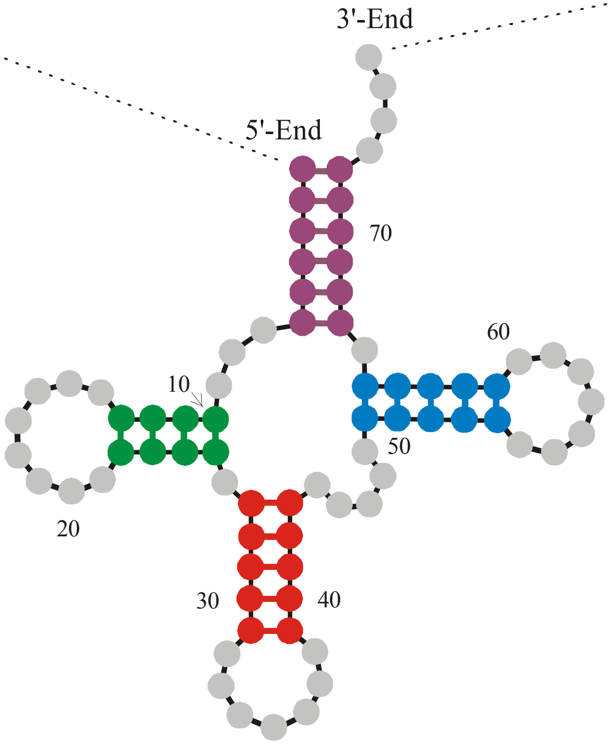
W.Saenger, Principles of Nucleic Acid Structure, Springer, Berlin 1984

Sequence

5'-End 3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

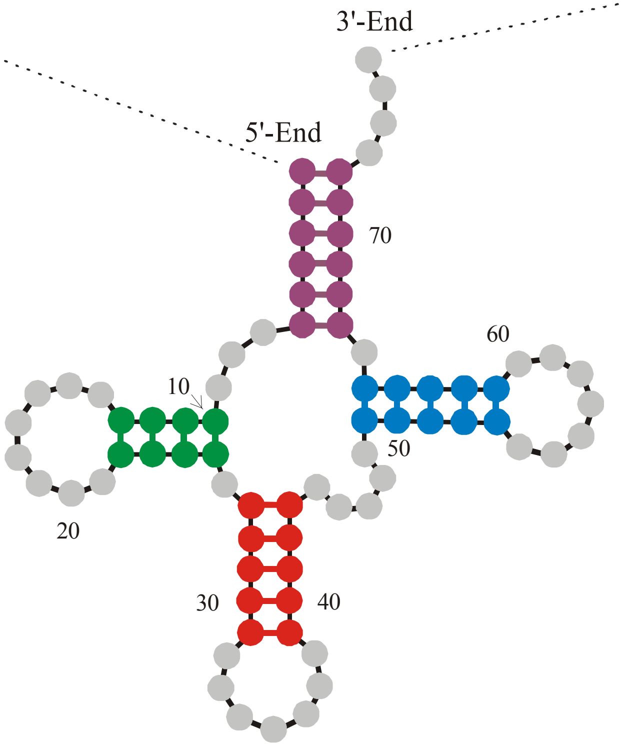
Secondary structure



Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

Secondary structure



Symbolic notation

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

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

Tertiary elements in RNA structure

1. Different classes of pseudoknots

2. Different classes of non-Watson-Crick base pairs

3. Base triplets, G-quartets, A-platforms, etc.

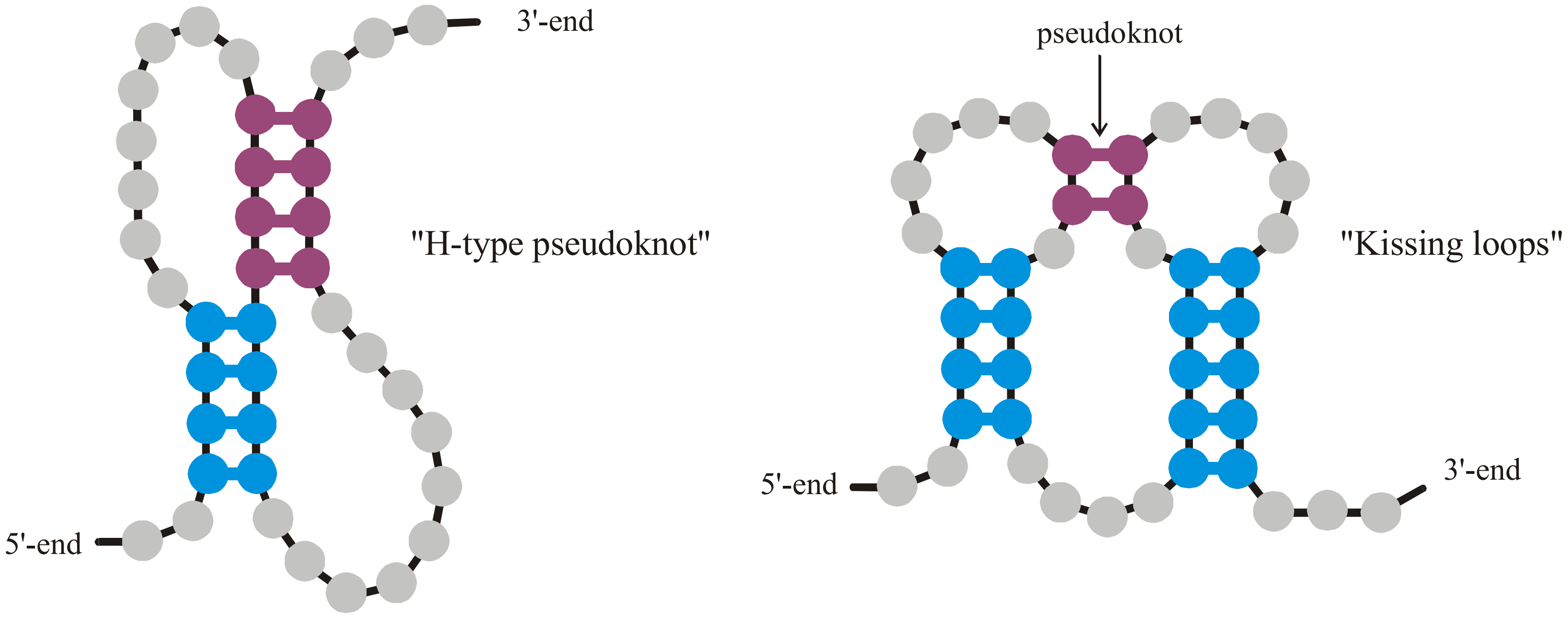
4. End-on-end stacking of double helices

5. Divalent metal ion complexes, Mg^{2+}, etc.

6. Other interactions involving phosphate, 2‘-OH, etc.

Tertiary elements in RNA structure

1. **Different classes of pseudoknots**
2. Different classes of non-Watson-Crick base pairs
3. Base triplets, G-quartets, A-platforms, etc.
4. End-on-end stacking of double helices
5. Divalent metal ion complexes, Mg^{2+}, etc.
6. Other interactions involving phosphate, 2‘-OH, etc.



··((((········[[[[ ))))··········]]]]···

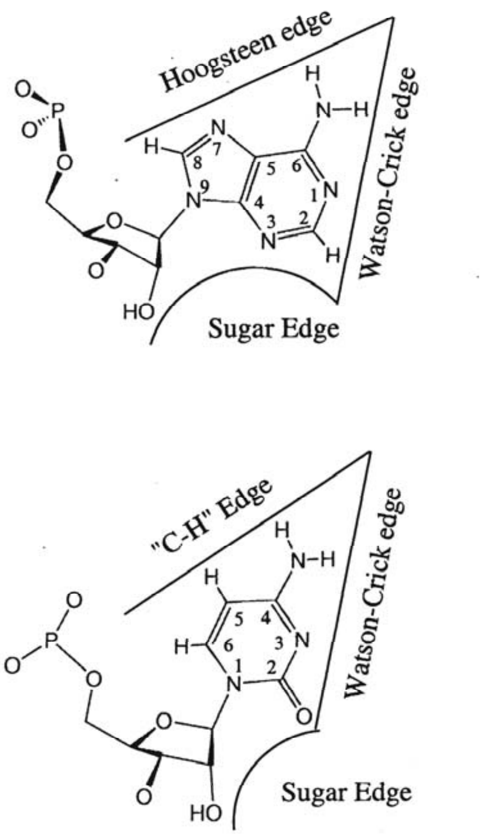
··((((·····[[·))))····(((((·]]·····)))))···

Two classes of pseudoknots in RNA structures

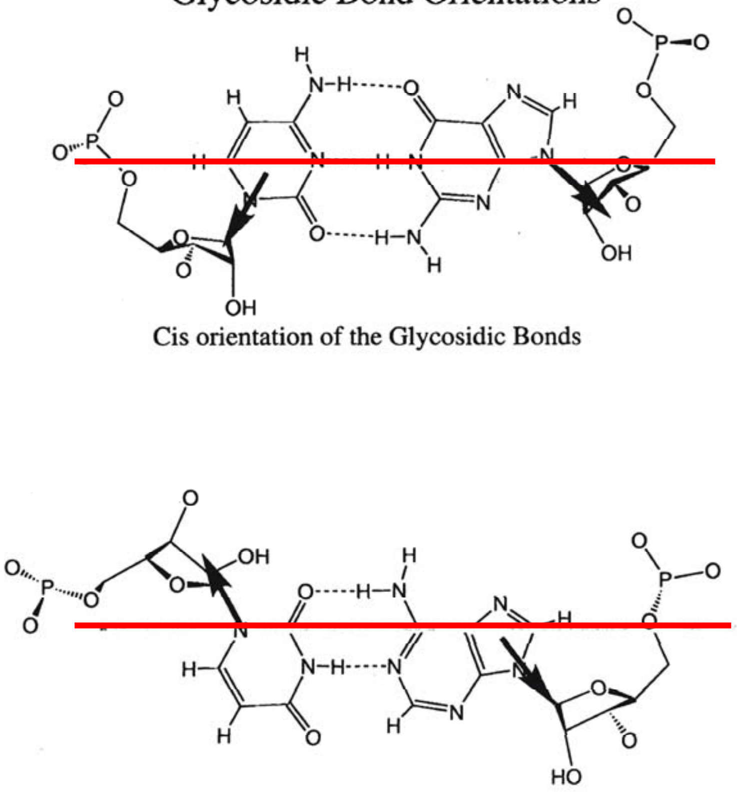
Tertiary elements in RNA structure

1. Different classes of pseudoknots
2. **Different classes of non-Watson-Crick base pairs**
3. Base triplets, G-quartets, A-platforms, etc.
4. End-on-end stacking of double helices
5. Divalent metal ion complexes, Mg^{2+}, etc.
6. Other interactions involving phosphate, 2‘-OH, etc.

Interacting Edges



Glycosidic Bond Orientations


Cis orientation of the Glycosidic Bonds

Trans orientation of the Glycosidic Bonds

N.B. Leontis, E. Westhof, Geometric nomenclature and classification of RNA base pairs. *RNA* **7**:499-512, 2001.

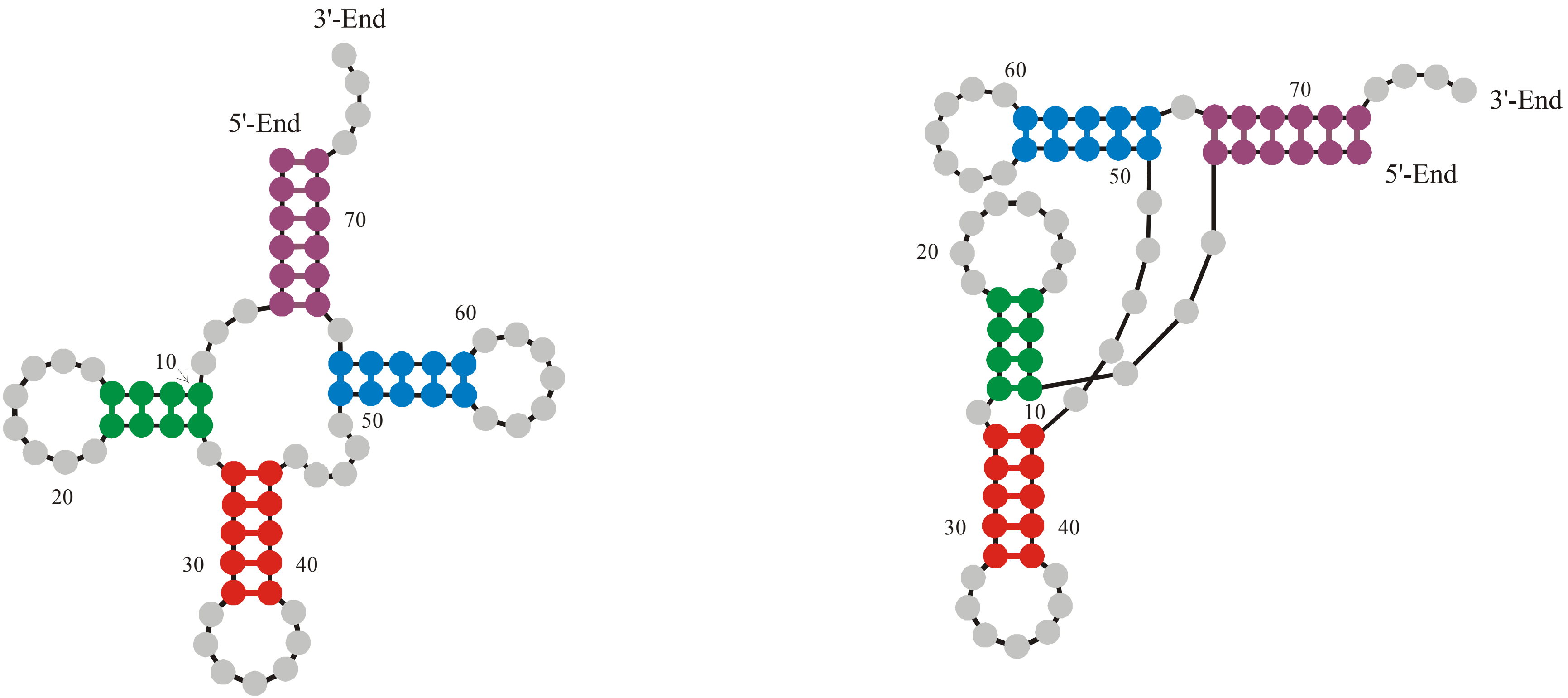
Twelve families of base pairs

Watson-Crick / Hogsteen / Sugar edge

Cis / Trans orientation

Tertiary elements in RNA structure

1. Different classes of pseudoknots
2. Different classes of non-Watson-Crick base pairs
3. Base triplets, G-quartets, A-platforms, etc.
4. **End-on-end stacking of double helices**
5. Divalent metal ion complexes, Mg^{2+}, etc.
6. Other interactions involving phosphate, 2‘-OH, etc.



End-on-end stacking of double helical regions yields the L-shape of tRNA^{phe}

1. The role of RNA in the cell and the notion of structure
2. **RNA folding**
3. Inverse folding of RNA
4. Sequence structure maps, neutral networks, and intersection
5. Reference to experimental data
6. Concluding remarks

How to compute RNA secondary structures

Efficient algorithms based on **dynamic programming** are available for computation of minimum free energy and **many** suboptimal secondary structures for given sequences.

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

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

Equilibrium partition function and base pairing probabilities in Boltzmann ensembles of suboptimal structures.

J.S.McCaskill. ***Biopolymers*** **29**:1105-1190 (1990)

The **Vienna RNA Package** provides in addition: **inverse folding** (computing sequences for given secondary structures), computation of melting profiles from partition functions, **all** suboptimal structures within a given energy interval, barrier tress of suboptimal structures, **kinetic folding** of RNA sequences, RNA-hybridization and RNA/DNA-hybridization through **cofolding** of sequences, alignment, etc..

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

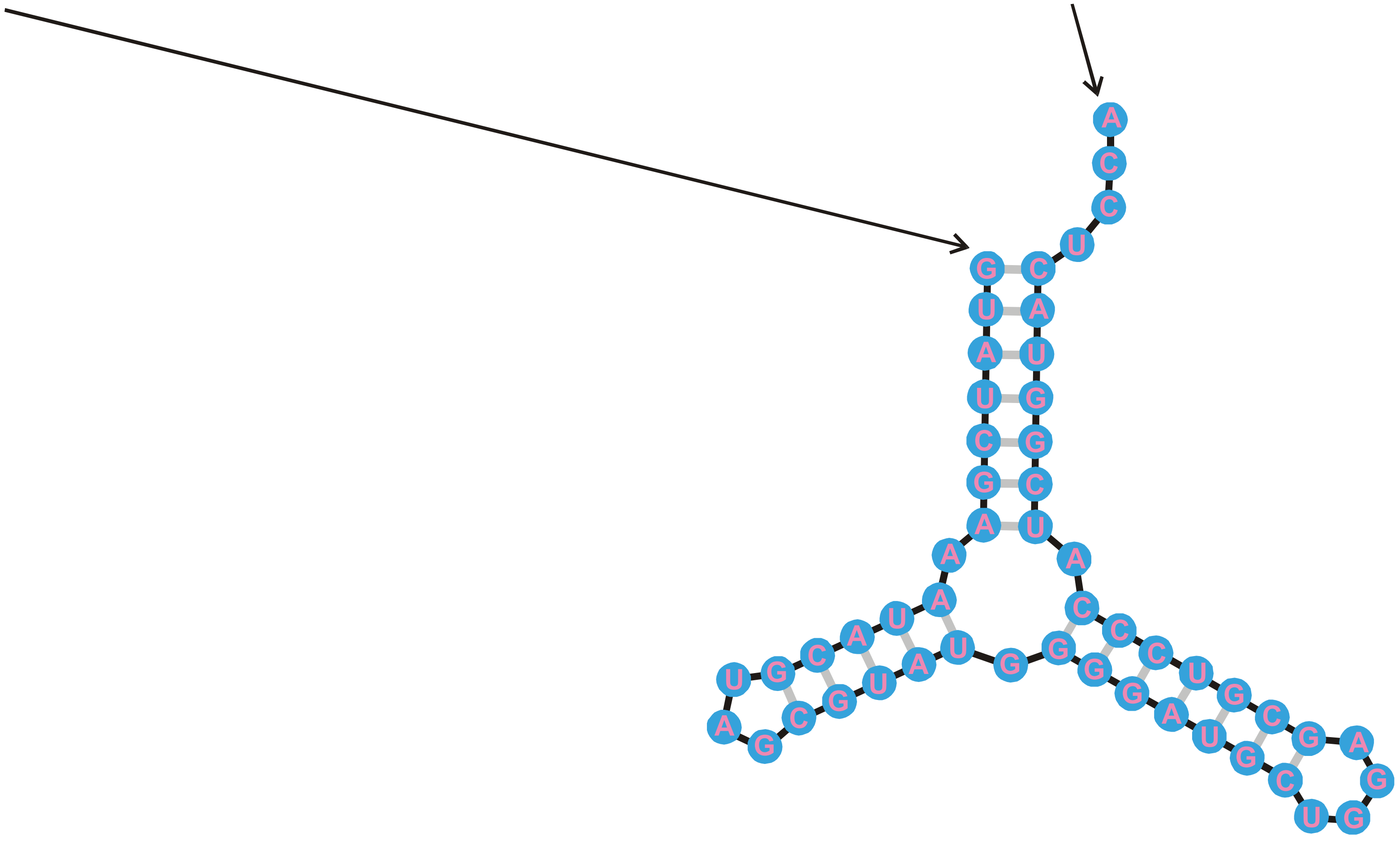
S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. ***Biopolymers*** **49**:145-165 (1999)

C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. ***RNA*** **6**:325-338 (1999)

Vienna RNA Package: http://www.tbi.univie.ac.at

5’-end 3’-end

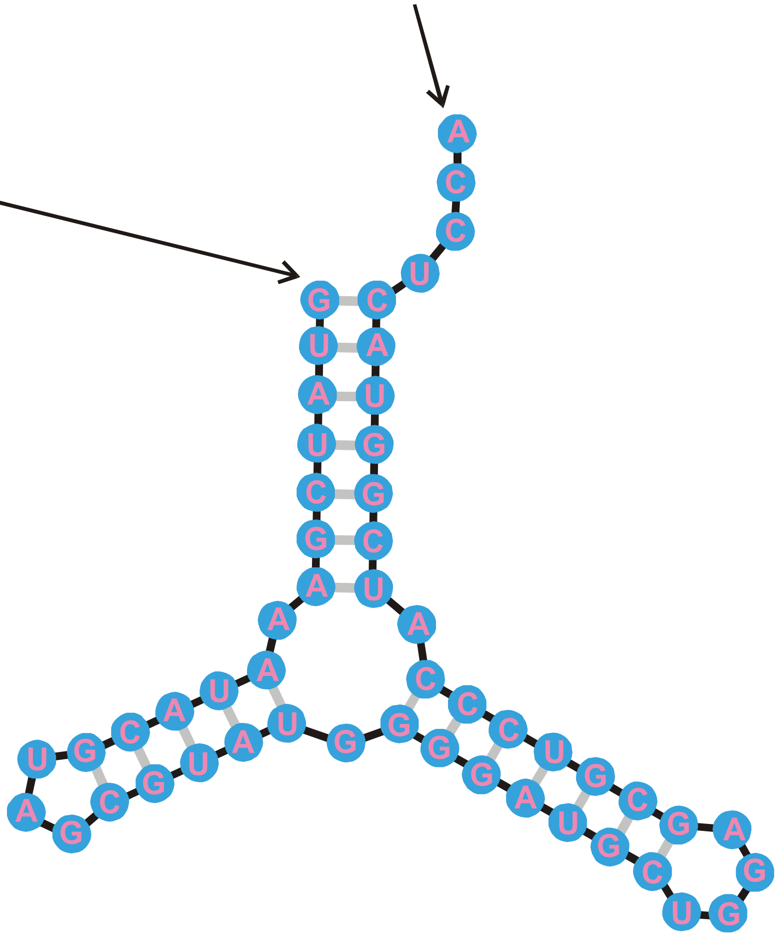
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



Folding of RNA sequences into secondary structures of minimal free energy, DG_0^{300}

5'-end 3'-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

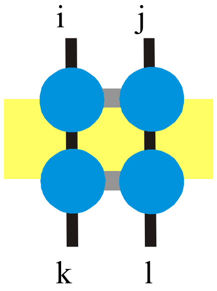


Edges: i·j,k·l á 𝕊 **.... base pairs**

(i) i· i+1 á 𝕊 .... **backbone**

(ii) #base pairs per node = {0,1}

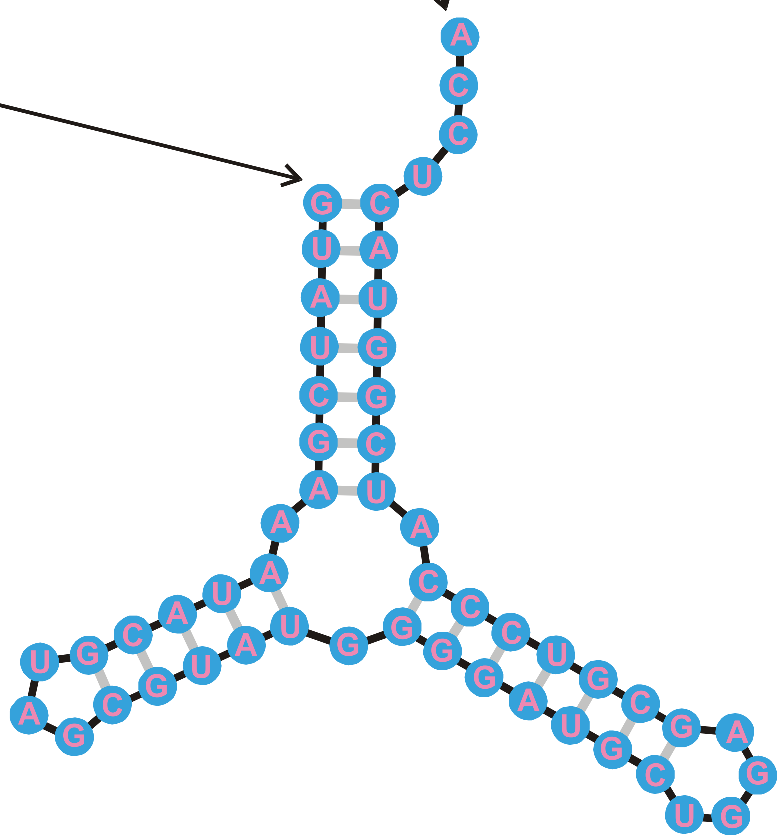
(iii) if i·j and l·k á 𝕊, then
i<k<j ñ i<l<j ....
pseudoknot exclusion

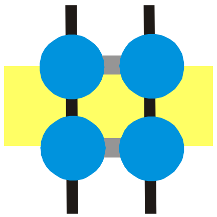


Folding of RNA sequences into secondary structures of minimal free energy, DG_0^{300}

5’-end 3’-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

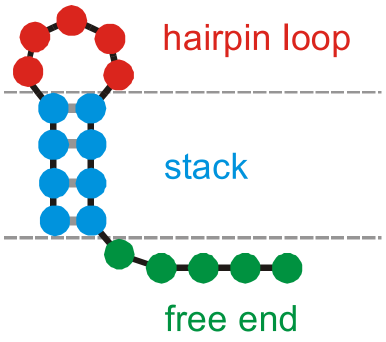


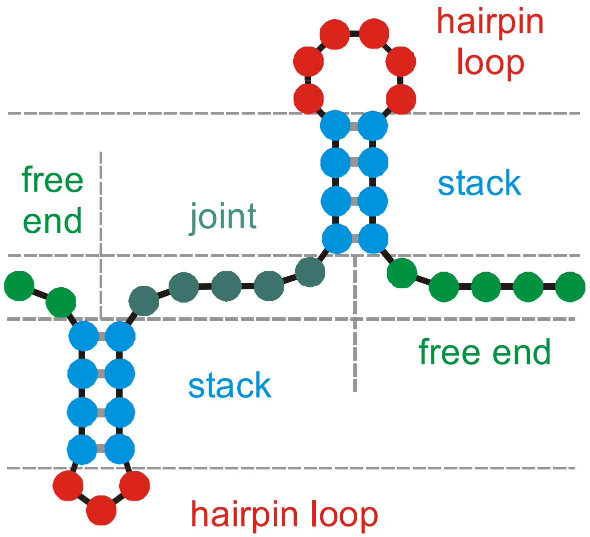


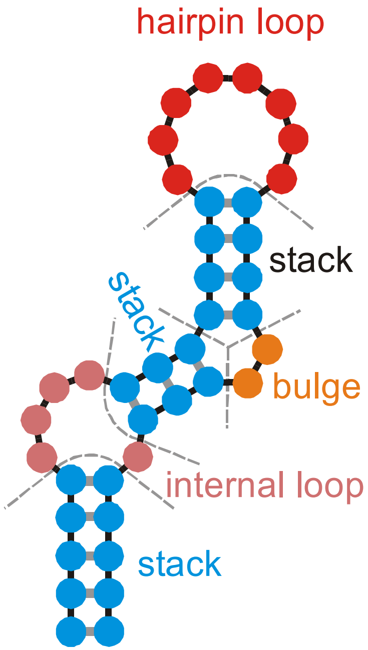
free energy of stacking < 0

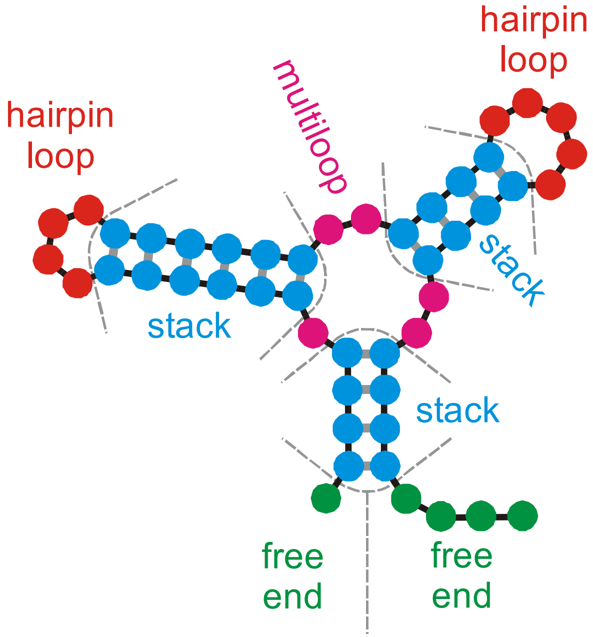
$$\Delta G_0^{300} = \underbrace{\sum g_{ij,kl}}_{\text{stacks of base pairs}} + \underbrace{\sum h(n_l)}_{\text{hairpin loops}} + \underbrace{\sum b(n_b)}_{\text{bulges}} + \underbrace{\sum i(n_i)}_{\text{internal loops}} + \cdots$$

Folding of RNA sequences into secondary structures of minimal free energy, DG$_0^{300}$







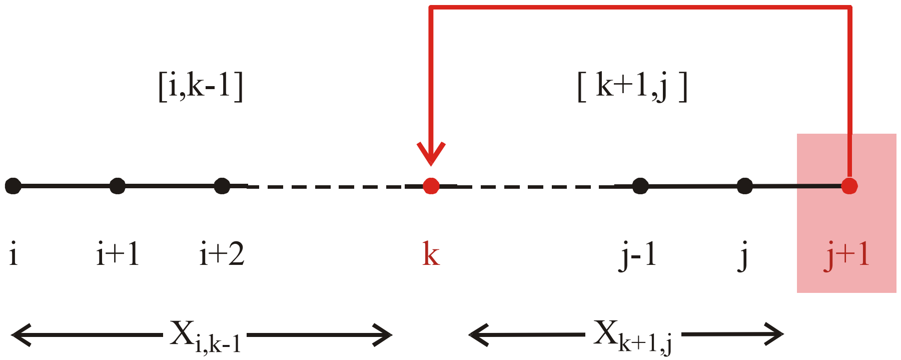


Elements of RNA secondary structures
as used in free energy calculations

Maximum matching

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

Back tracking yields the structure(s).



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

Minimum free energy computations are based on empirical energies

RNAStudio.lnk

GGCGCGCCCGGCGCC

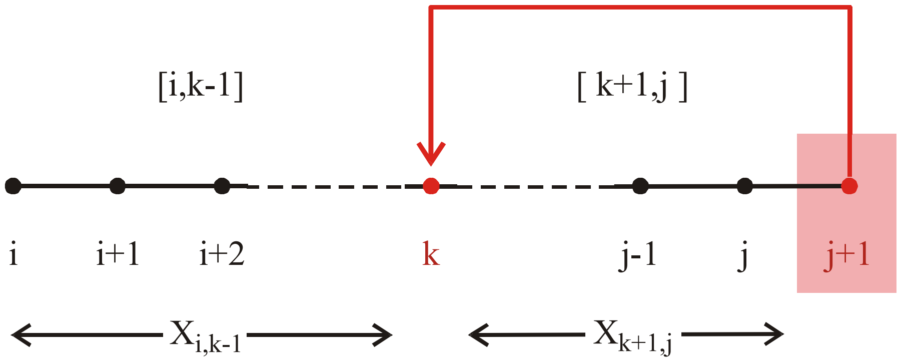
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

Maximum matching

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

Back tracking yields the structure(s).



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

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

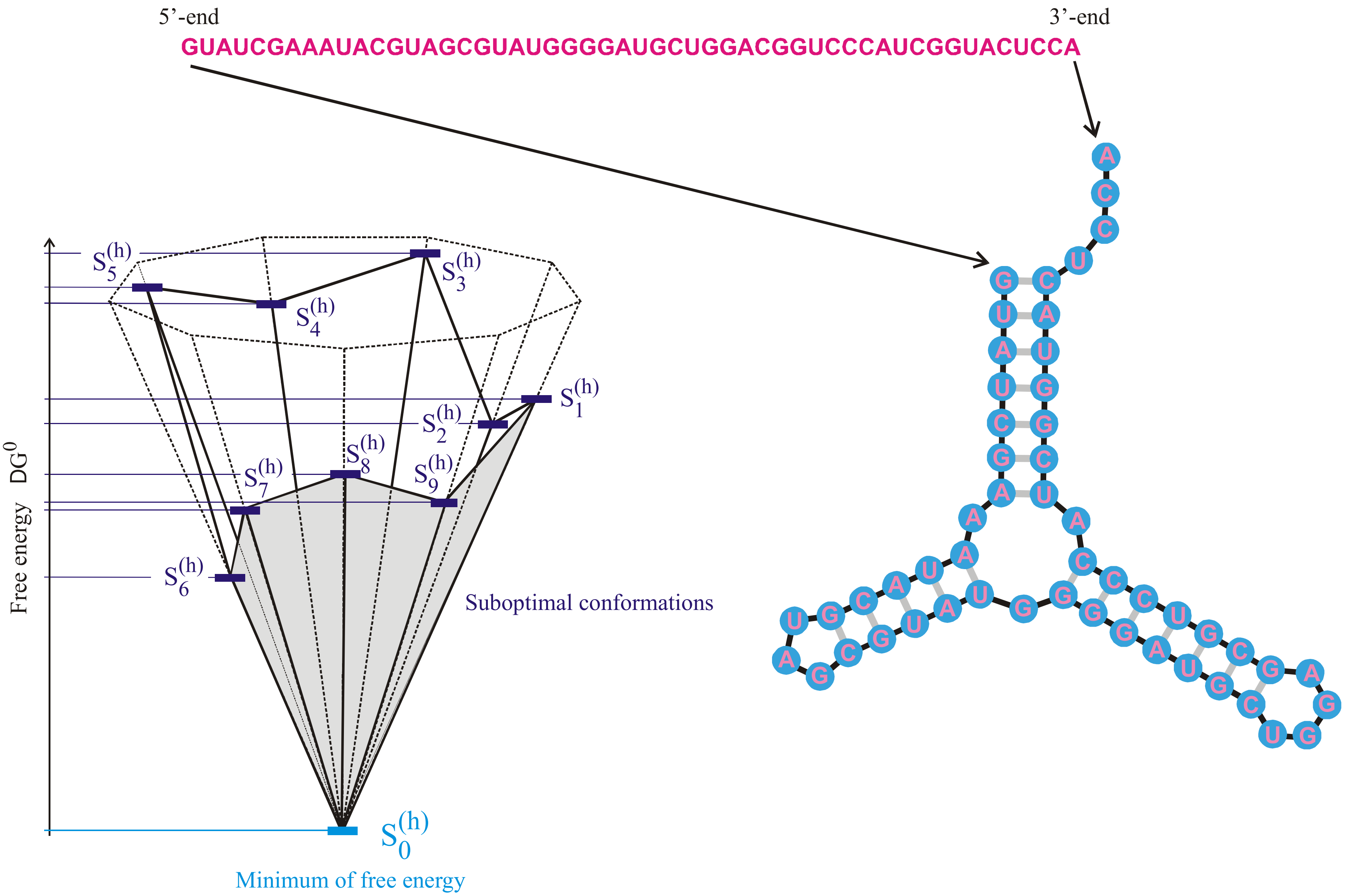
Minimum free energy computations are based on empirical energies

RNAStudio.lnk

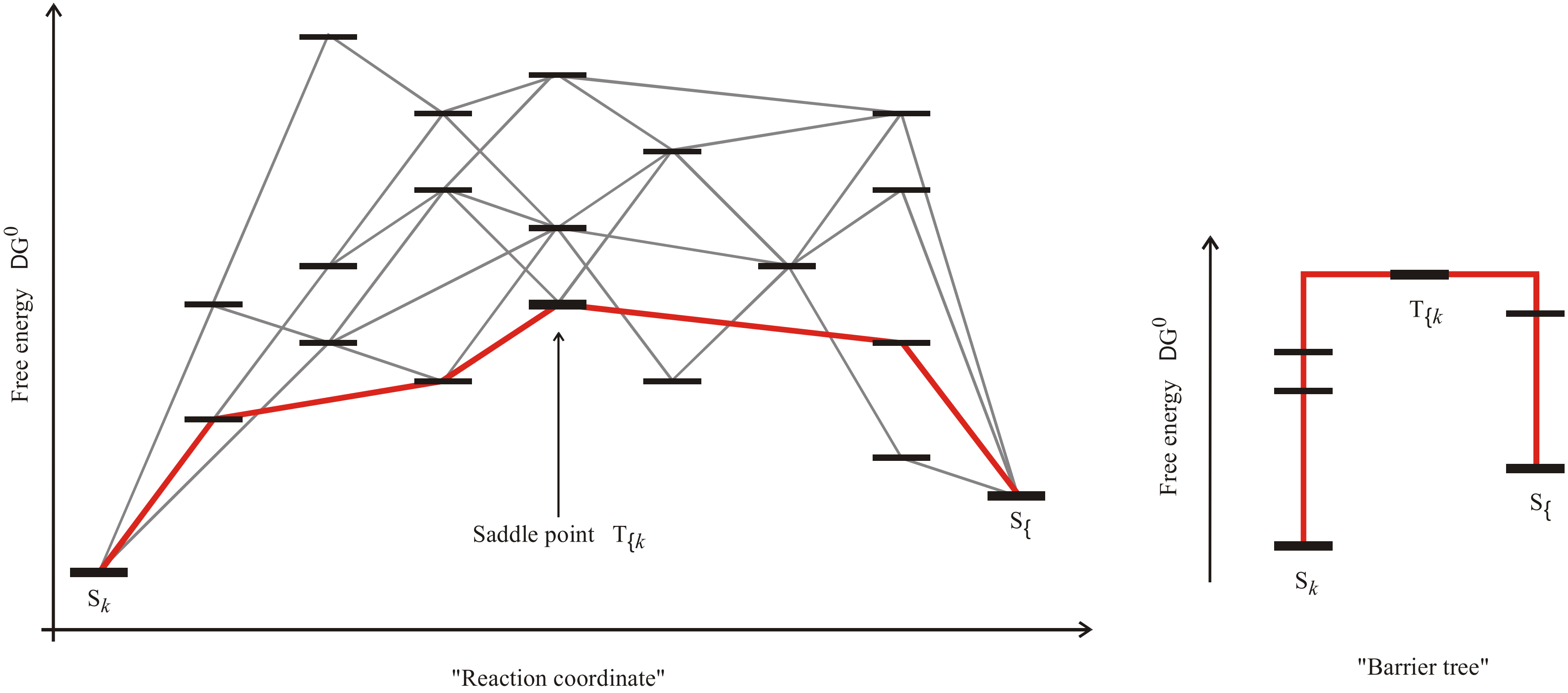
GGCGCGCCCGGCGCC

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG



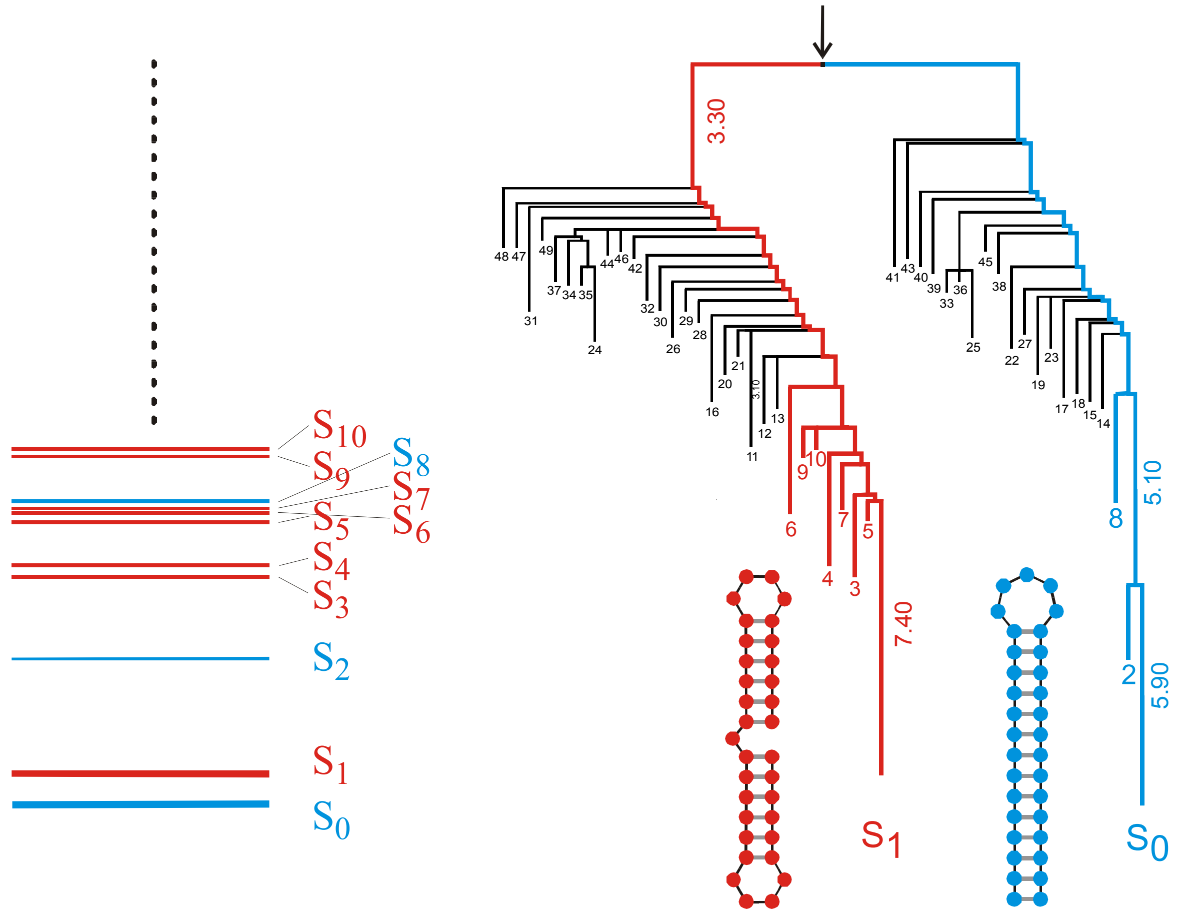
The minimum free energy structures on a discrete space of conformations



Definition of a ‚barrier tree‘

lim t Á ¸

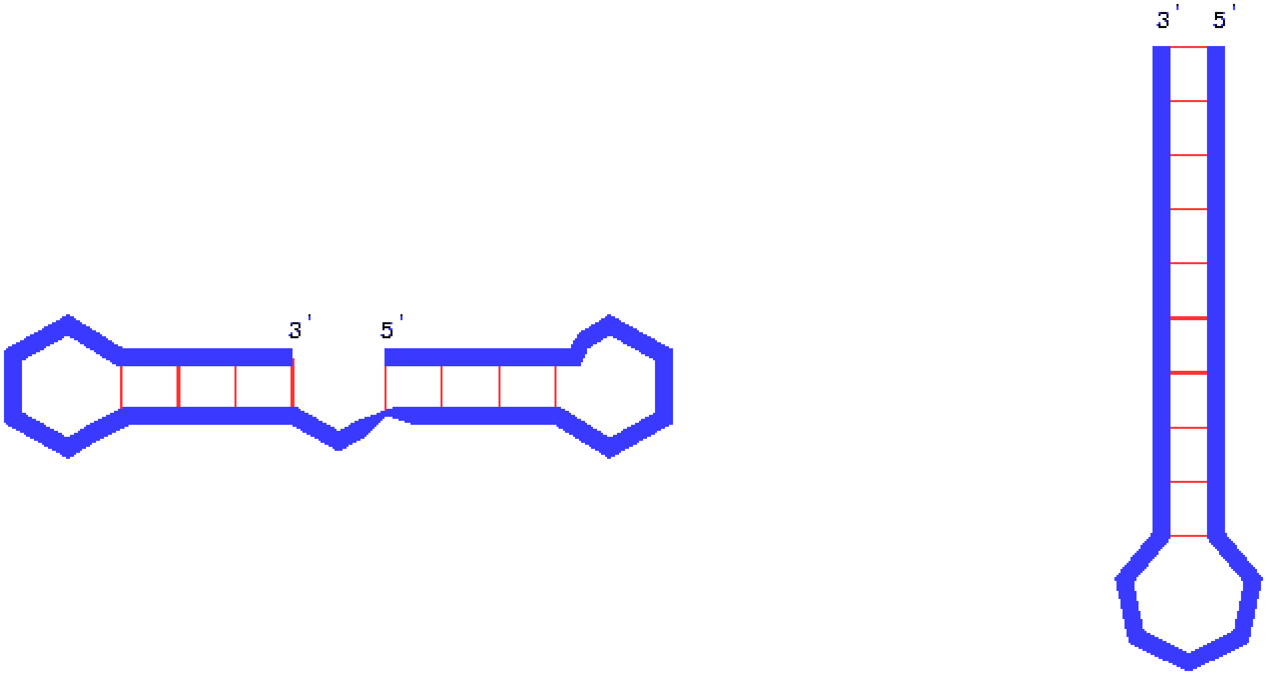
finite folding time



Suboptimal structures

Kinetic folding

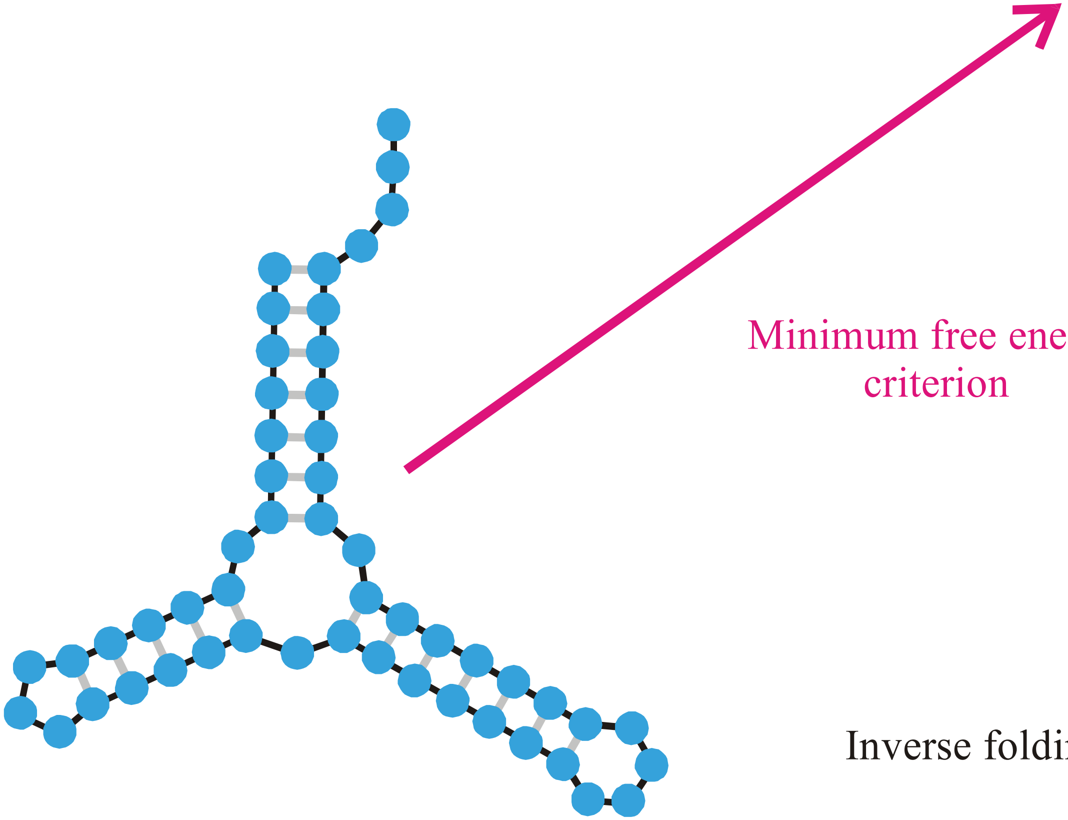
A typical energy landscape of a sequence with two (meta)stable comformations



Kinetics RNA refolding between a long living metastable conformation
and the minmum free energy structure

1. The role of RNA in the cell and the notion of structure
2. RNA folding
3. **Inverse folding of RNA**
4. Sequence structure maps, neutral networks, and intersection
5. Reference to experimental data
6. Concluding remarks

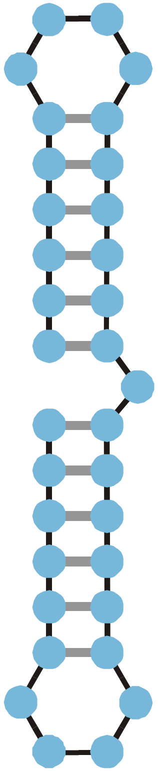
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



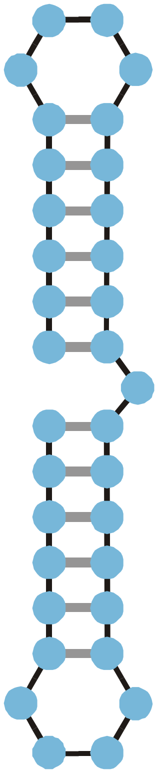
Minimum free energy criterion

Inverse folding of RNA secondary structures

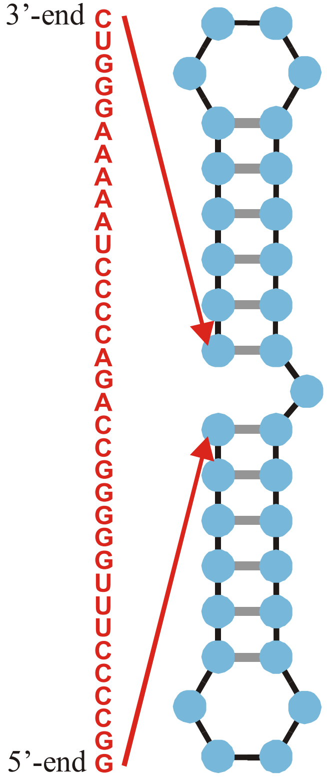
The idea of inverse folding algorithm is to search for sequences that form a given RNA secondary structure under the minimum free energy criterion.



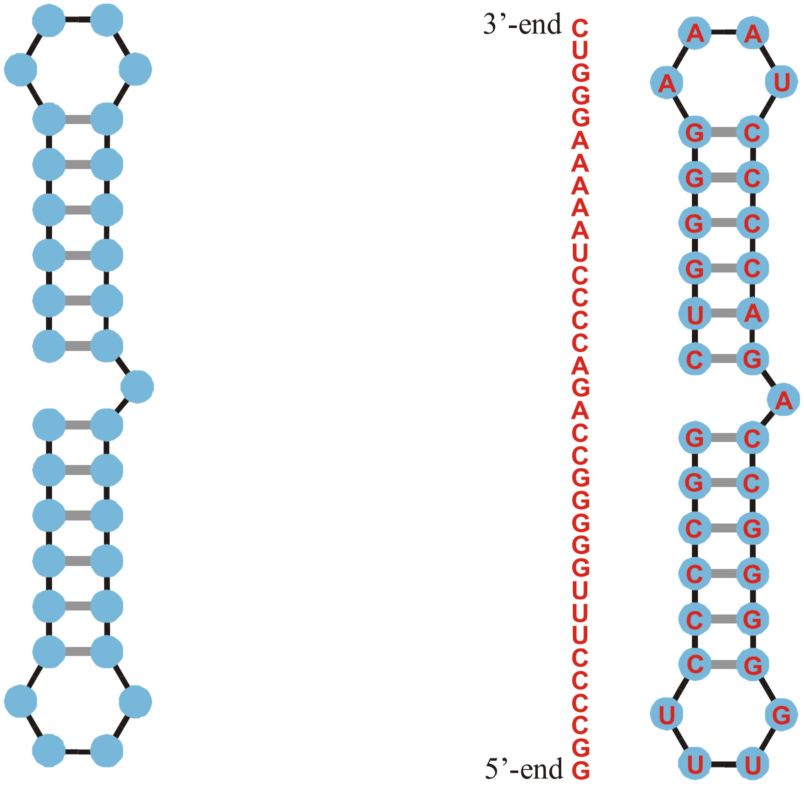
Structure



Structure

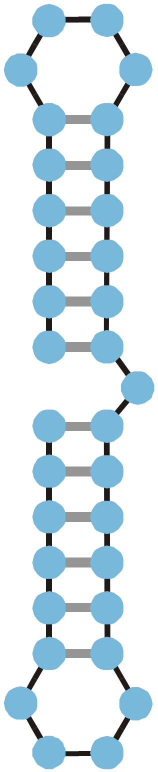


Compatible sequence

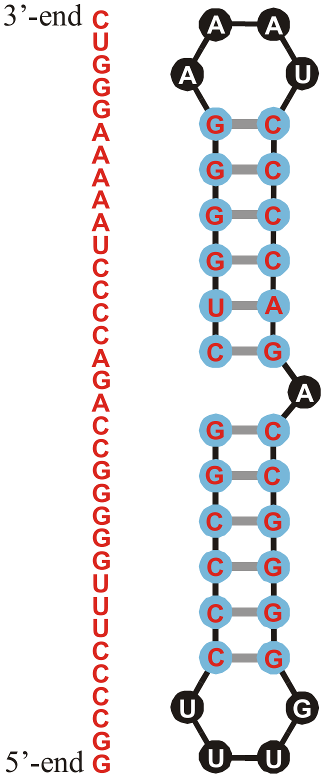


Structure

Compatible sequence



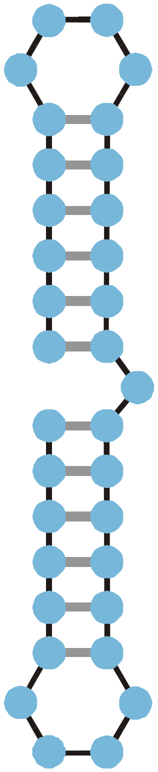
Structure



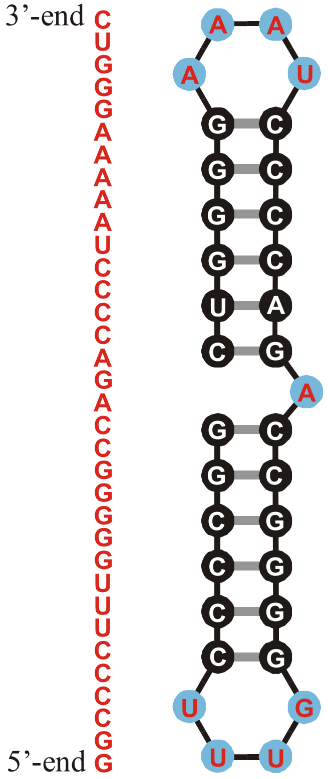
Single nucleotides: **A,U,G,C**

Compatible sequence

Single bases pairs are varied independently



Structure

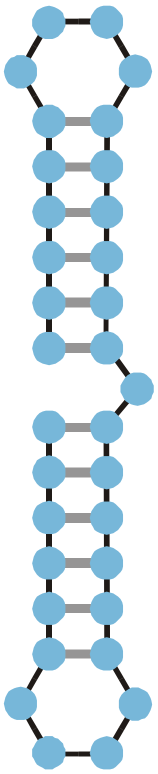


Base pairs:

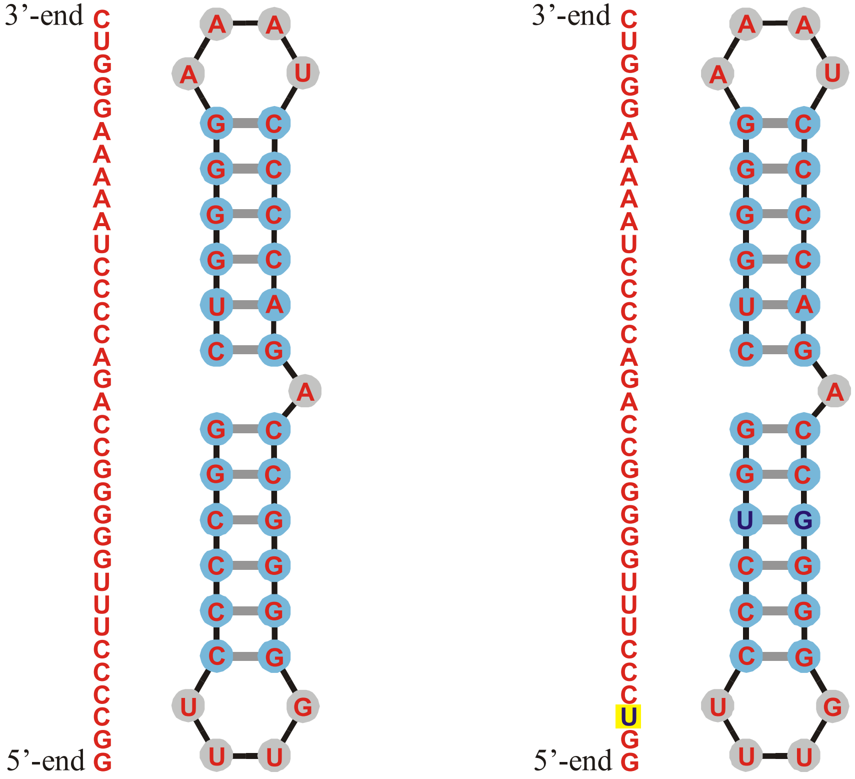
AU , UA
GC , CG
GU , UG

Compatible sequence

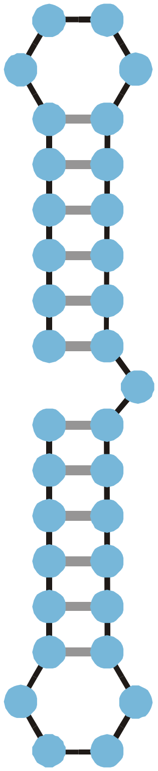
Base pairs are varied in strict correlation



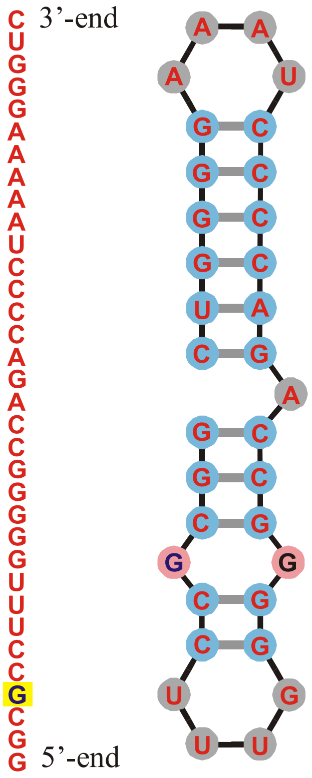
Structure


3'-end
CUGGGAAAAAUCCCCAGACCGGGGUUUCCCCGG
5'-end
3'-end
CUGGGAAAAAUCCCCAGACCGGGGUUUCCCUGG
5'-end

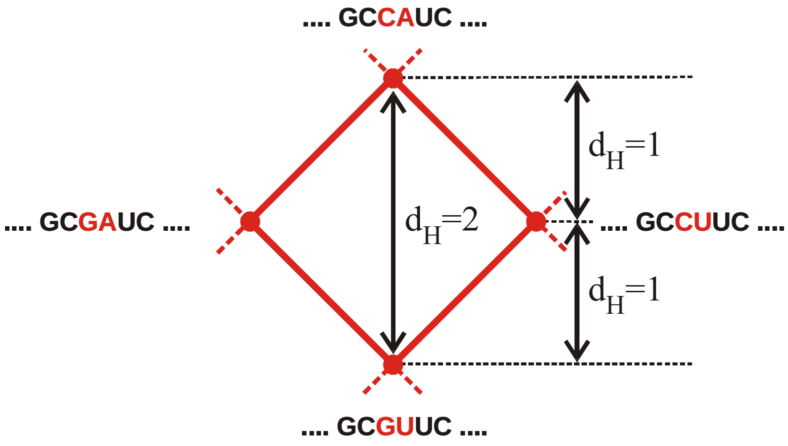
Compatible sequences



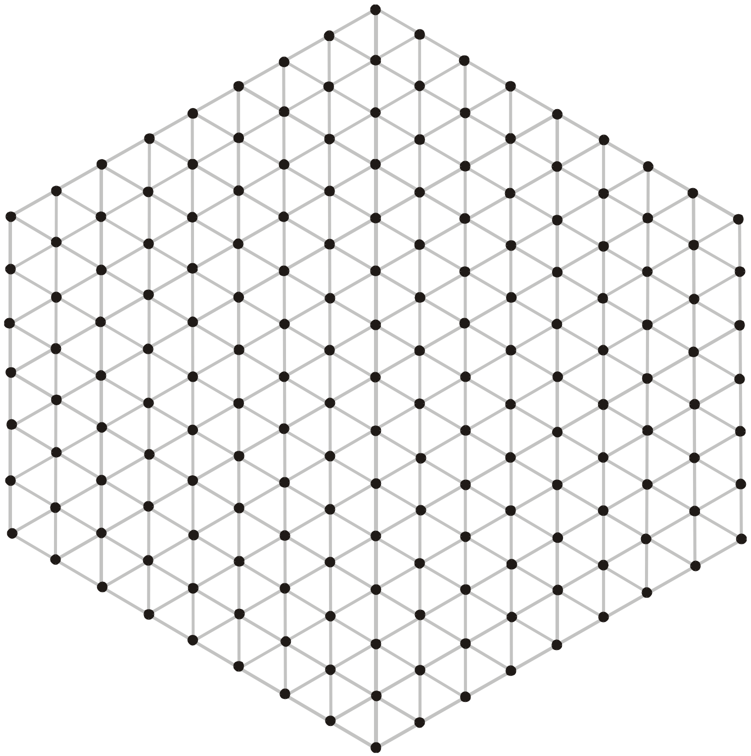
Structure



Incompatible sequence

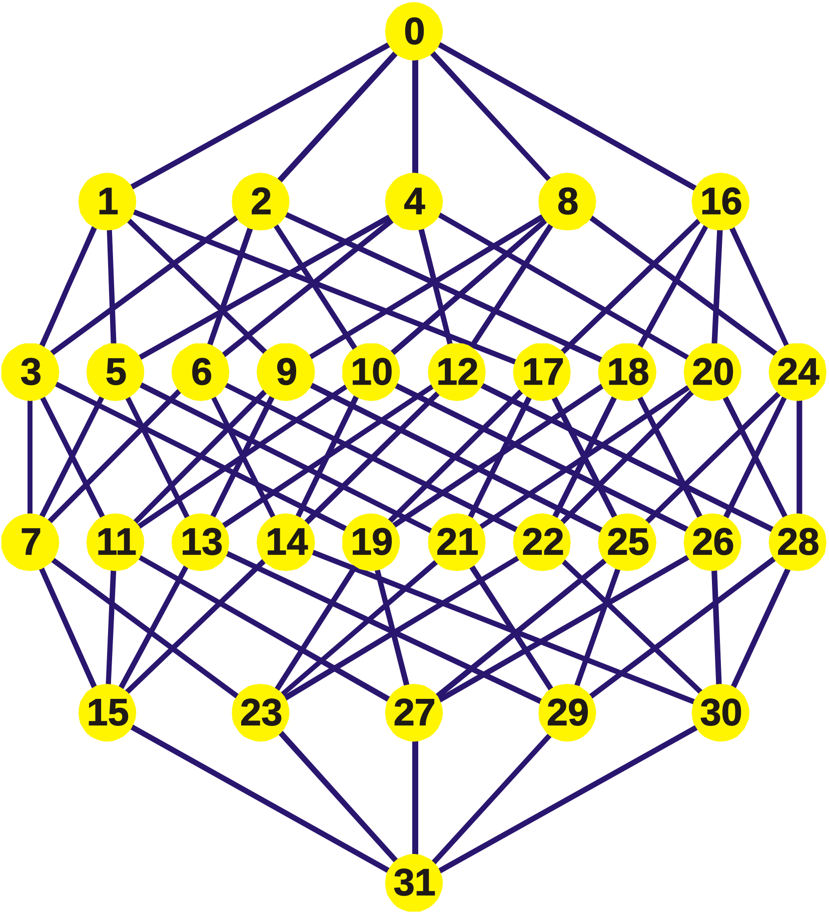


City-block distance in sequence space



2D Sketch of sequence space

Single point mutations as moves in sequence space



Hypercube of dimension n = 5

Mutant class

0

1

2

3

4

5

Binary sequences are encoded by their decimal equivalents:

C = 0 and G = 1, for example,

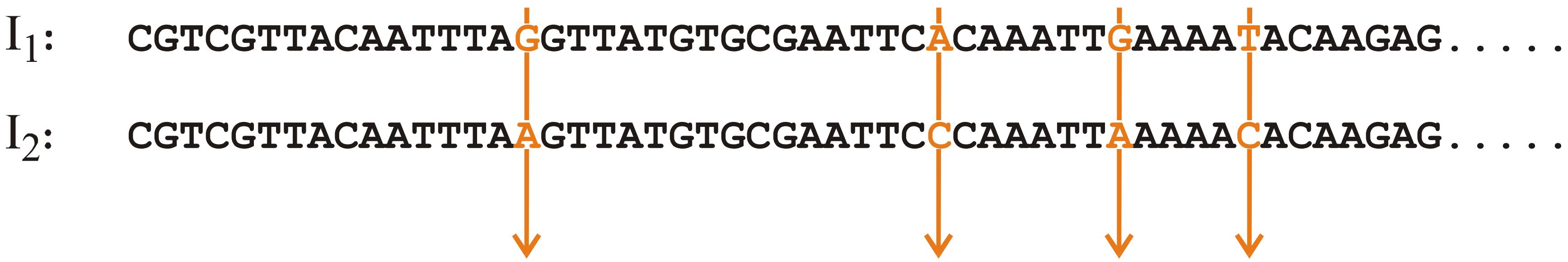
"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

"29" ≡ 11101 = GGGCG, etc.

Decimal coding of binary sequences

Sequence space of binary sequences of chain lenght n = 5



I_1: CGTCGTTACAATTTAGGTTATGTGCGAATTCACAAATTGAAAATACAAGAG.....

I_2: CGTCGTTACAATTTAAGTTATGTGCGAATTCCCAAATTAAAAACACAAGAG.....

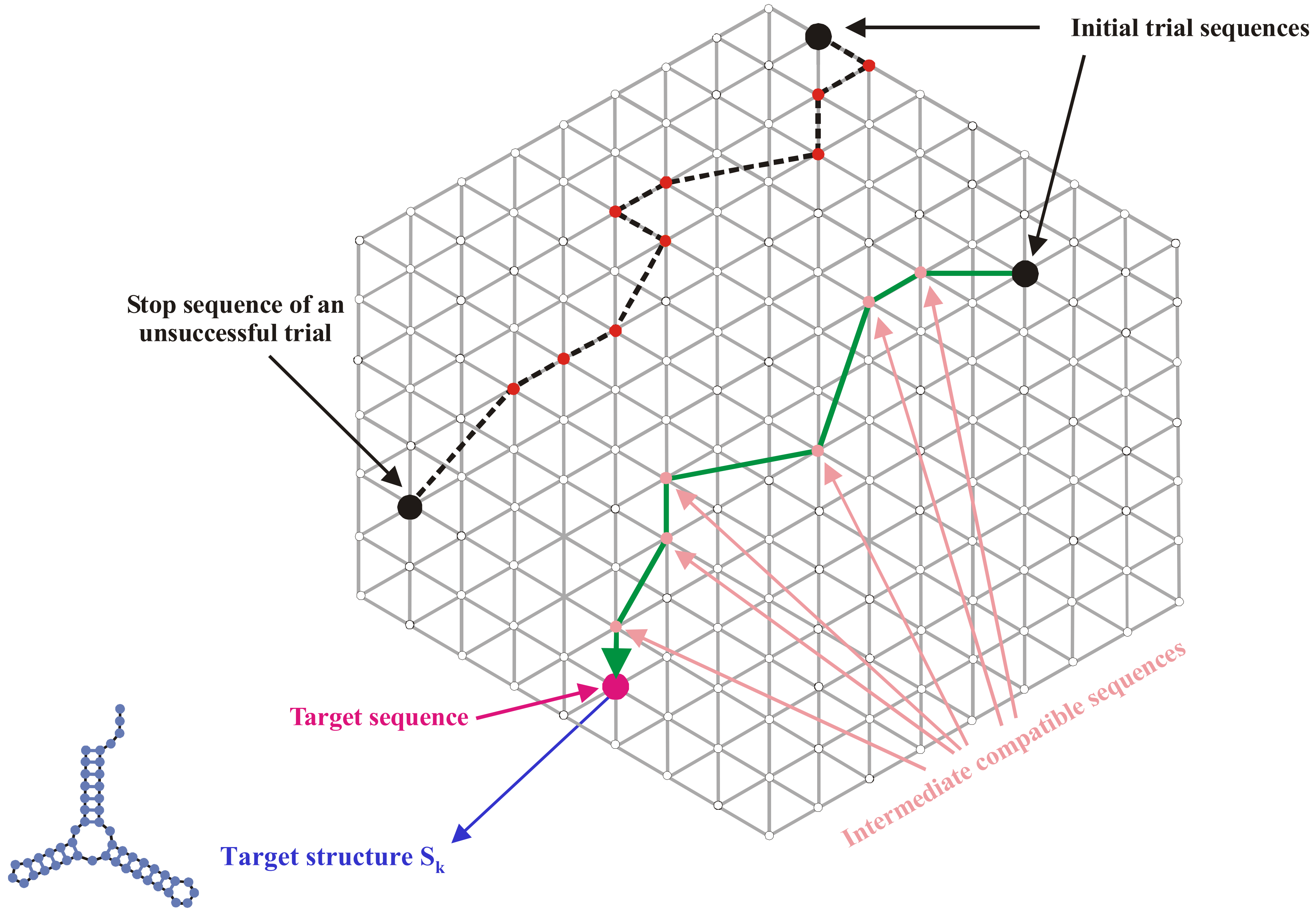
Hamming distance $d_H(I_1,I_2) = 4$

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

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

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

The Hamming distance between sequences induces a metric in sequence space


Initial trial sequences
Stop sequence of an unsuccessful trial
Target sequence
Target structure S_k
Intermediate compatible sequences

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

Inverse folding algorithm

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

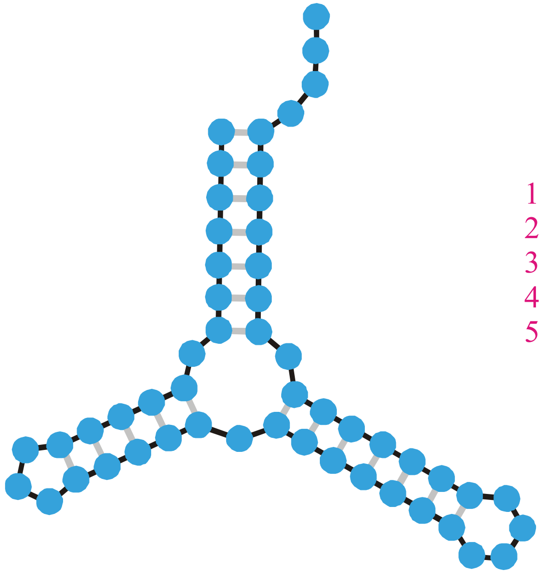
$\mathbf{S_0}$ Á S_1 Á S_2 Á S_3 Á S_4 Á ... Á S_k Á S_{k+1} Á ... Á $\mathbf{S_t}$

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

$\mathcal{M}$ ... 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



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

Inverse folding of RNA secondary structures

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

1. The role of RNA in the cell and the notion of structure

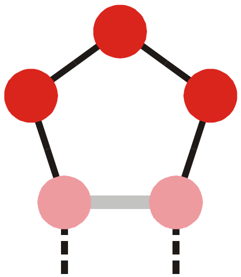
2. RNA folding

3. Inverse folding of RNA

4. **Sequence structure maps, neutral networks, and intersection**

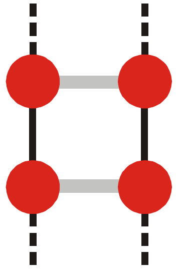
5. Reference to experimental data

6. Concluding remarks



Minimal hairpin loop size:

n_{lp} Æ 3



Minimal stack length:

n_{st} Æ 2

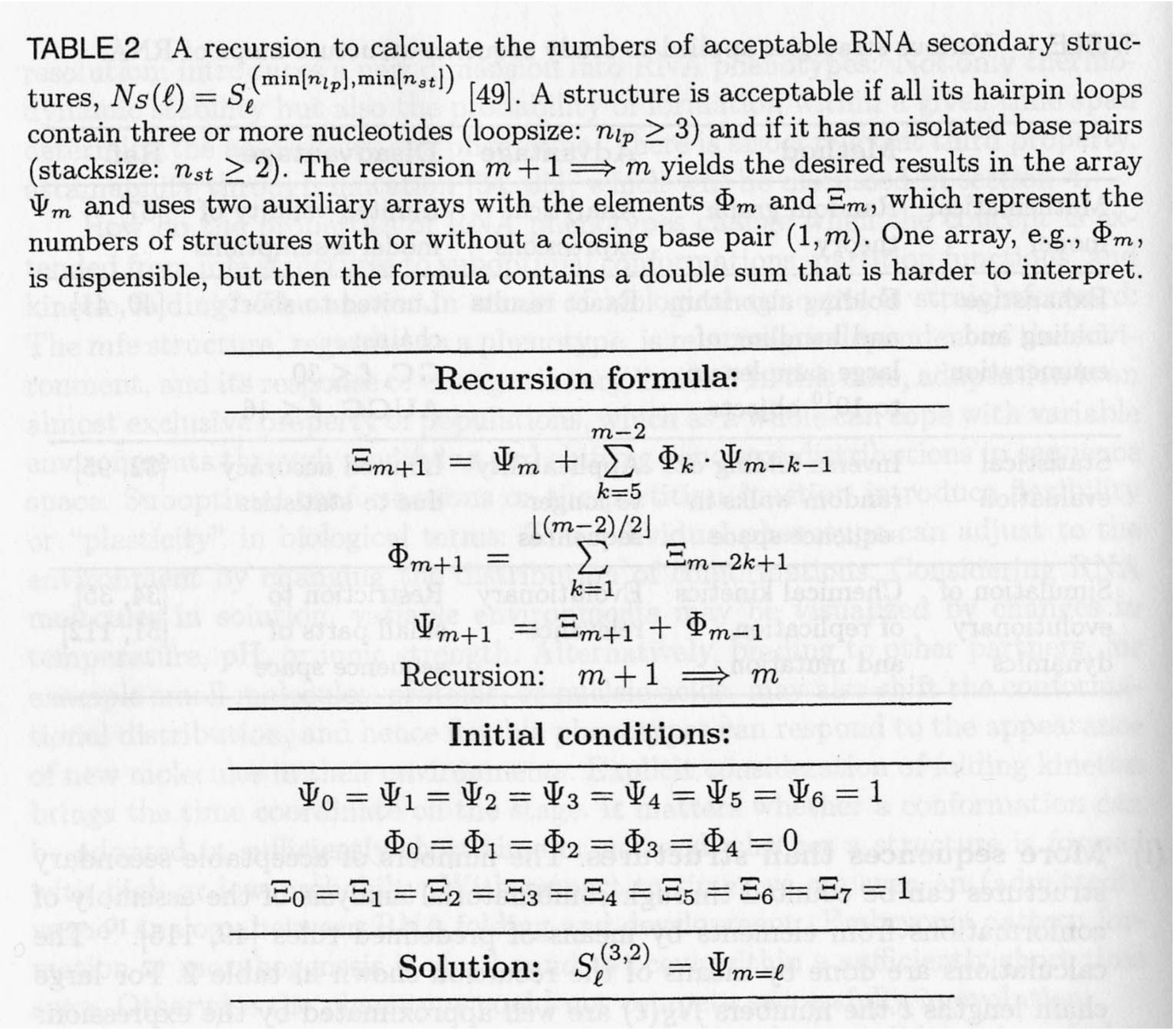


TABLE 2 A recursion to calculate the numbers of acceptable RNA secondary structures, $N_S(\ell) = S_\ell^{(\min[n_{lp}],\min[n_{st}])}$ [49]. A structure is acceptable if all its hairpin loops contain three or more nucleotides (loopsize: $n_{lp} \geq 3$) and if it has no isolated base pairs (stacksize: $n_{st} \geq 2$). The recursion $m+1 \Longrightarrow m$ yields the desired results in the array Ψ_m and uses two auxiliary arrays with the elements Φ_m and Ξ_m, which represent the numbers of structures with or without a closing base pair $(1,m)$. One array, e.g., Φ_m, is dispensible, but then the formula contains a double sum that is harder to interpret.

Recursion formula:

$$\Xi_{m+1} = \Psi_m + \sum_{k=5}^{m-2} \Phi_k \cdot \Psi_{m-k-1}$$

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

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

Recursion: $m+1 \Longrightarrow m$

Initial conditions:

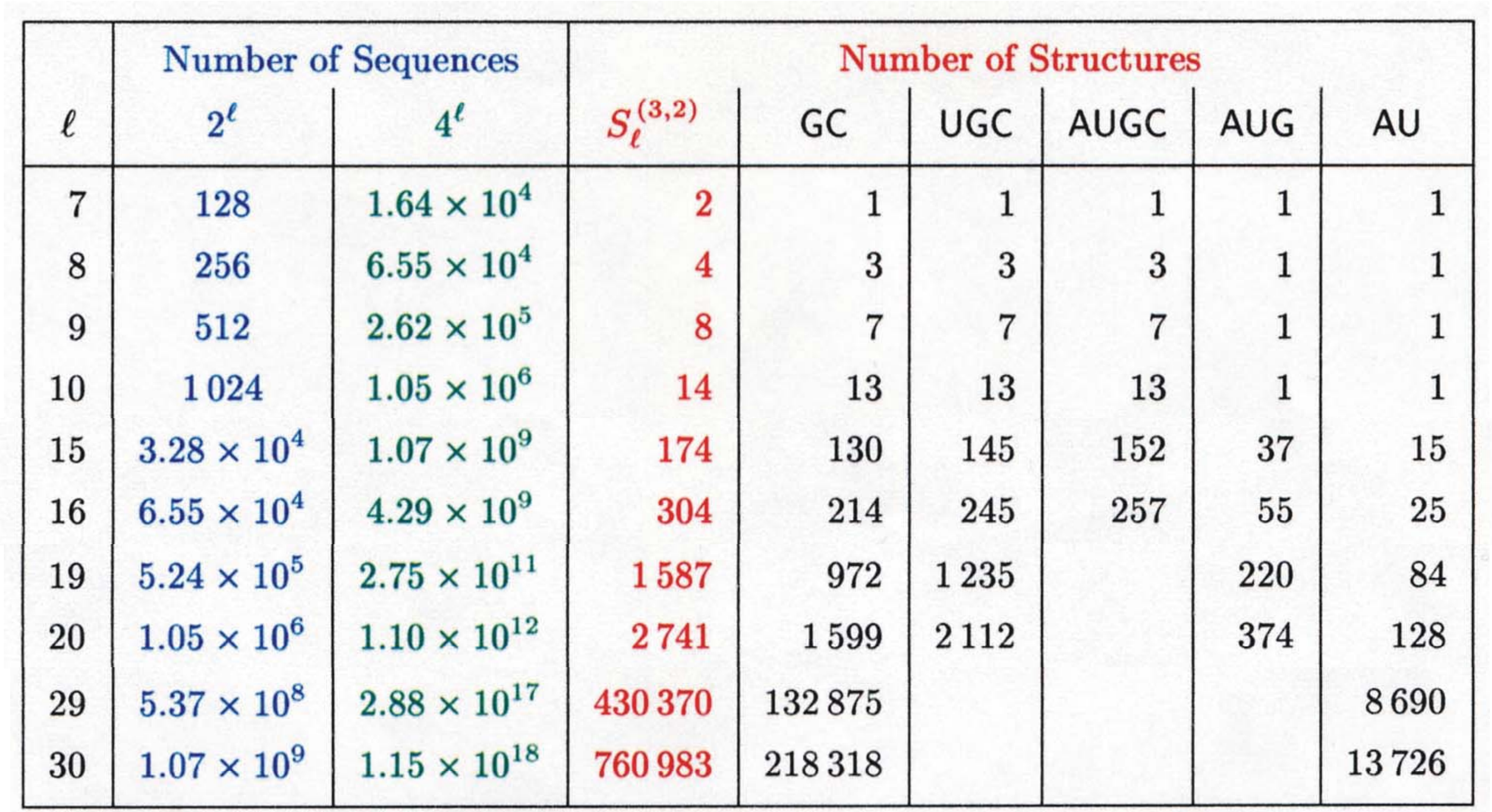
$$\Psi_0 = \Psi_1 = \Psi_2 = \Psi_3 = \Psi_4 = \Psi_5 = \Psi_6 = 1$$

$$\Phi_0 = \Phi_1 = \Phi_2 = \Phi_3 = \Phi_4 = 0$$

$$\Xi_0 = \Xi_1 = \Xi_2 = \Xi_3 = \Xi_4 = \Xi_5 = \Xi_6 = \Xi_7 = 1$$

Solution: $S_\ell^{(3,2)} = \Psi_{m=\ell}$

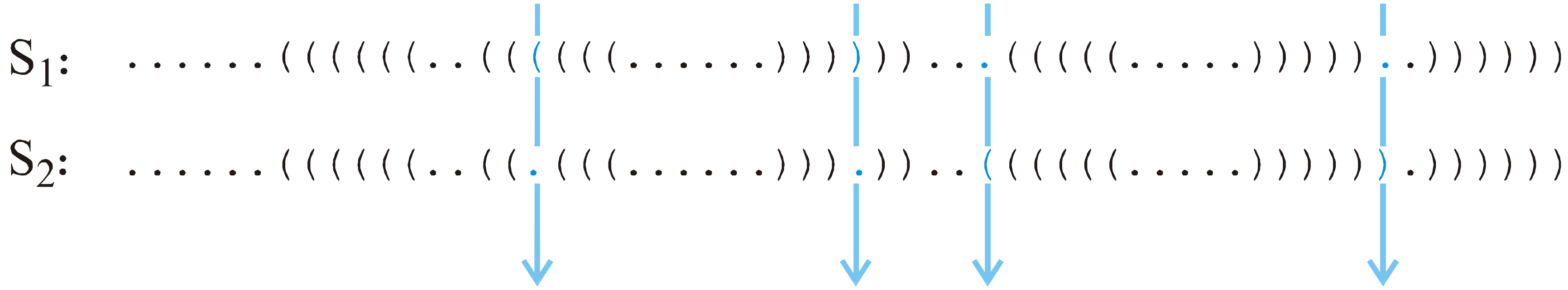
Recursion formula for the number of acceptable RNA secondary structures



ℓ	Number of Sequences		Number of Structures					
	2^ℓ	4^ℓ	$S_\ell^{(3,2)}$	GC	UGC	AUGC	AUG	AU
7	128	1.64×10^4	2	1	1	1	1	1
8	256	6.55×10^4	4	3	3	3	1	1
9	512	2.62×10^5	8	7	7	7	1	1
10	1 024	1.05×10^6	14	13	13	13	1	1
15	3.28×10^4	1.07×10^9	174	130	145	152	37	15
16	6.55×10^4	4.29×10^9	304	214	245	257	55	25
19	5.24×10^5	2.75×10^{11}	1 587	972	1 235		220	84
20	1.05×10^6	1.10×10^{12}	2 741	1 599	2 112		374	128
29	5.37×10^8	2.88×10^{17}	430 370	132 875				8 690
30	1.07×10^9	1.15×10^{18}	760 983	218 318				13 726

Computed numbers of minimum free energy structures over different nucleotide alphabets

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



S_1: ......((((((..((((((......)))))))...(((((.....))))).. .))))))

S_2: ......((((((..((.(((......))).))..((((((.....))))))..))))))

Hamming distance $d_H(S_1,S_2) = 4$

(i) $d_H(S_1,S_1) = 0$

(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

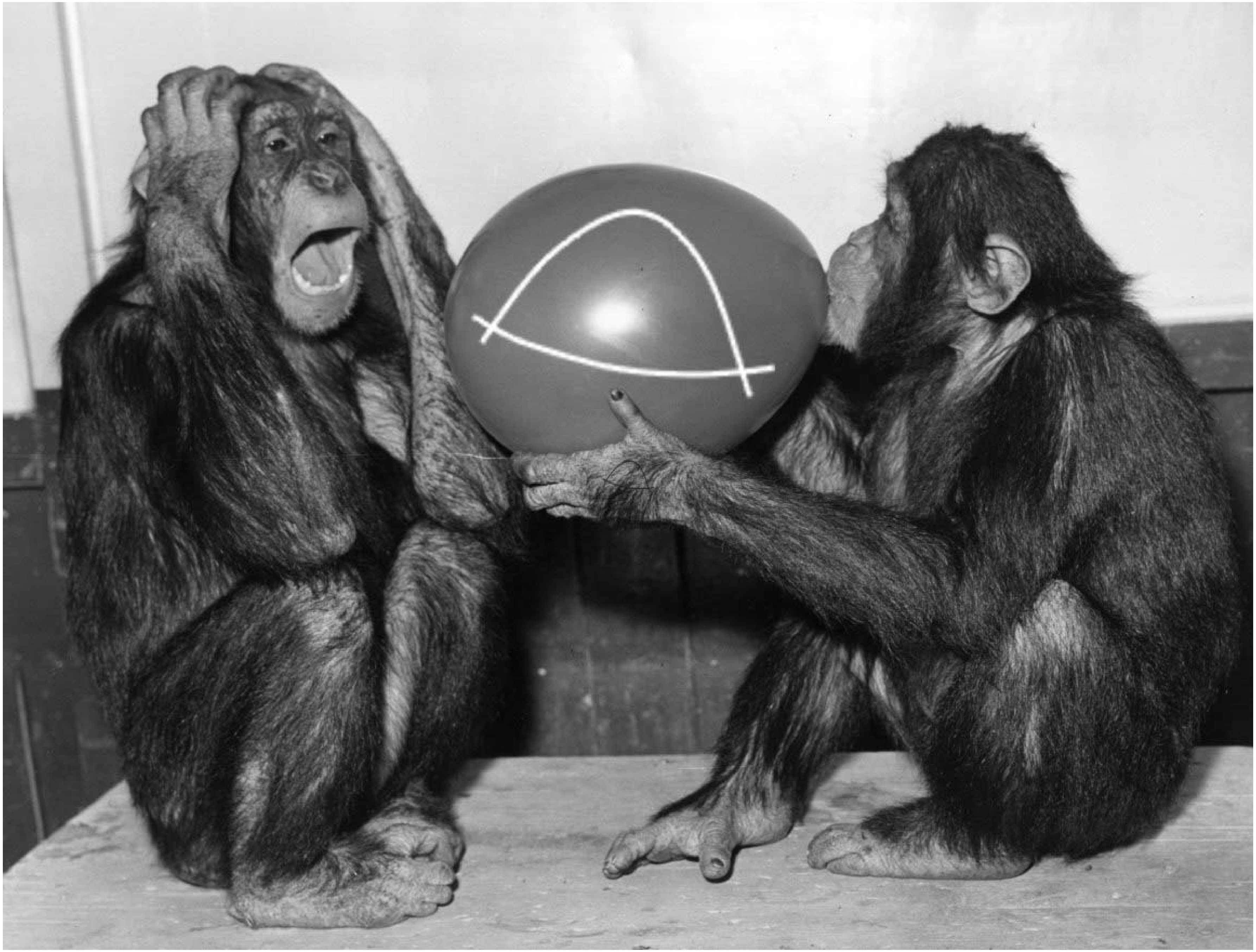
(iii) $d_H(S_1,S_3)$ ¶ $d_H(S_1,S_2) + d_H(S_2,S_3)$

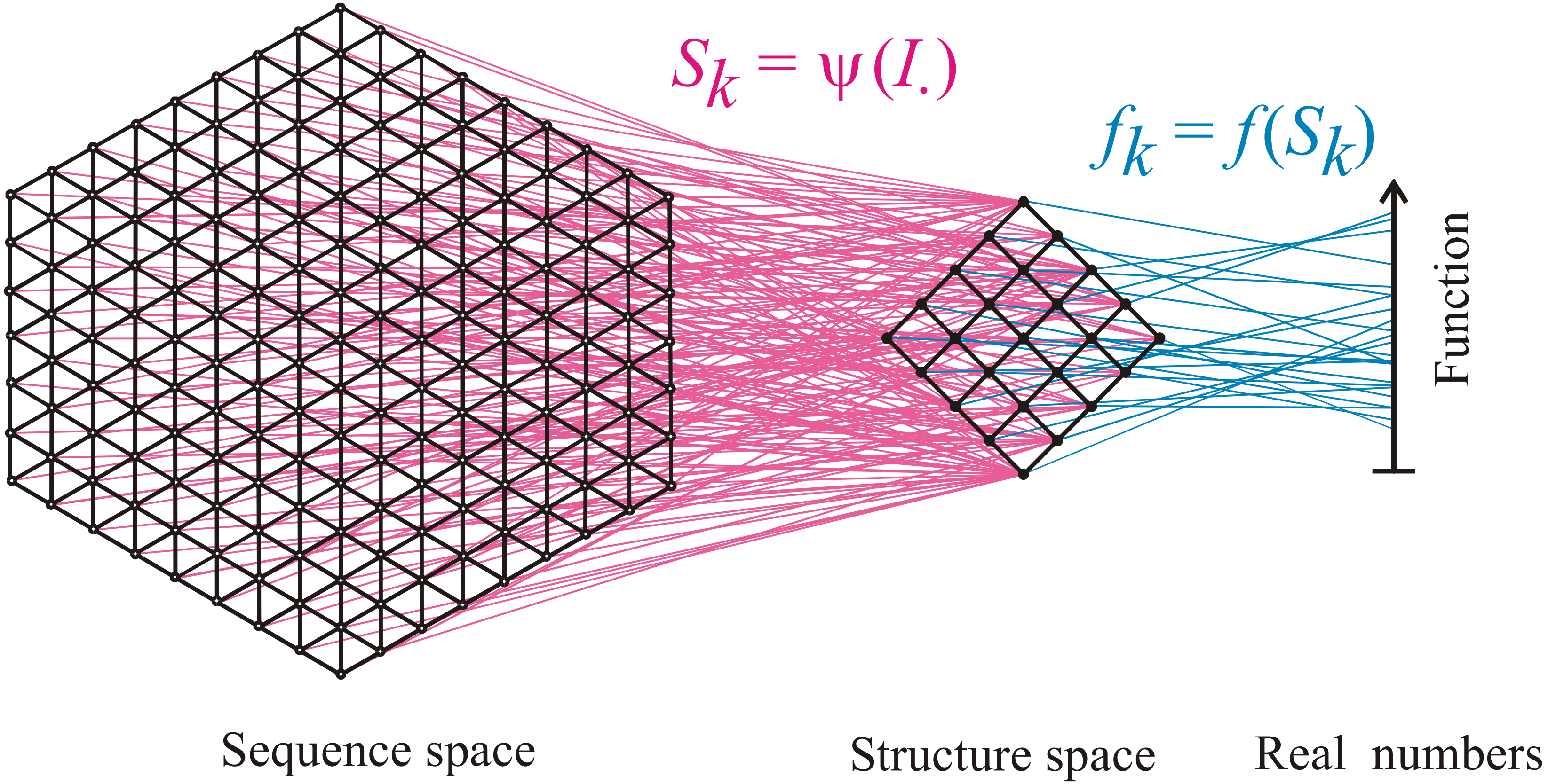
The Hamming distance between structures in parentheses notation forms a metric in structure space

RNA **sequences** as well as RNA secondary **structures** can be visualized as objects in **metric spaces**. At constant chain length the sequence space is a (generalized) hypercube.

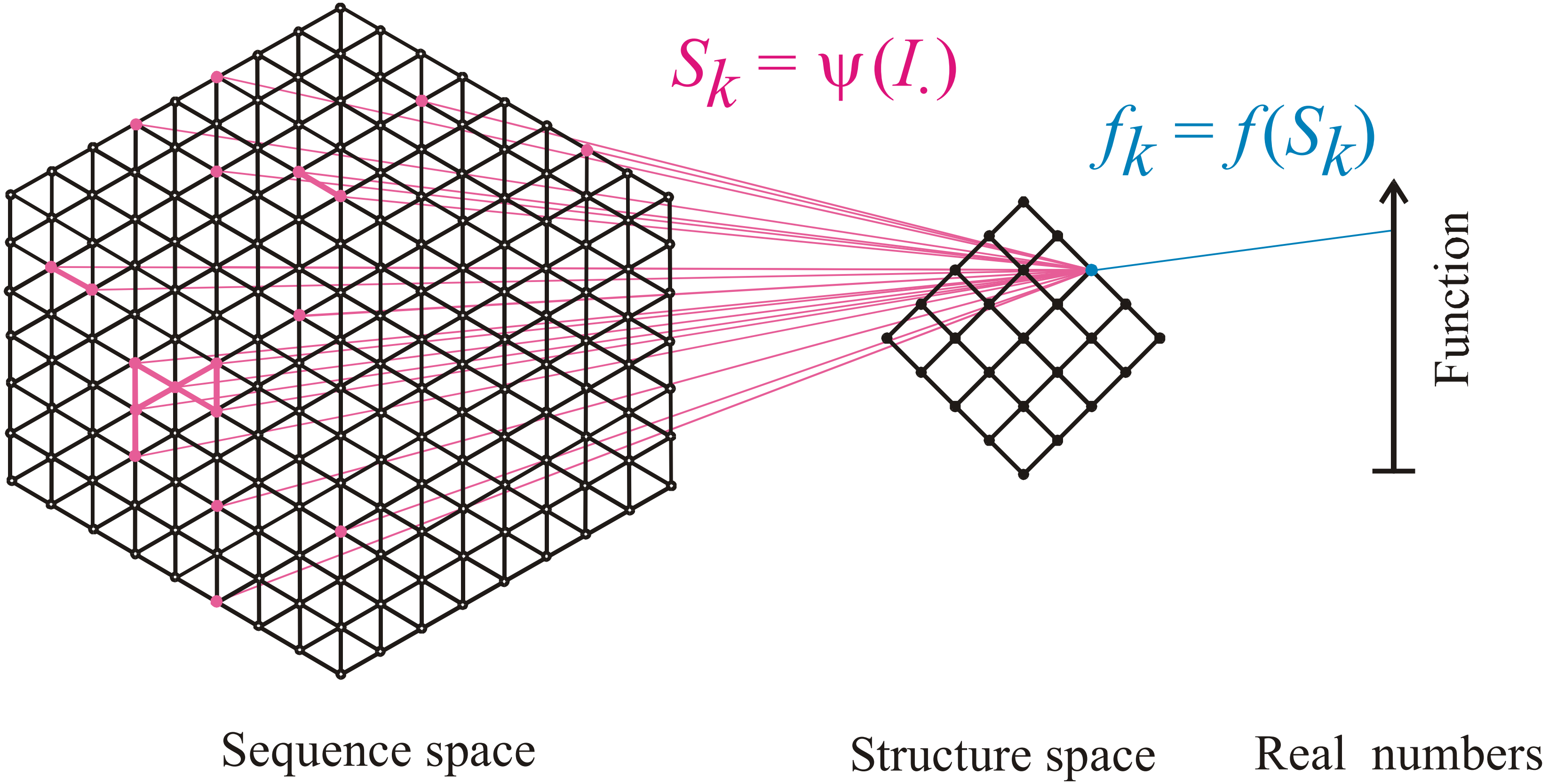
The **mapping** from RNA **sequences** into RNA secondary **structures** is many-to-one. Hence, it is redundant and not invertible.

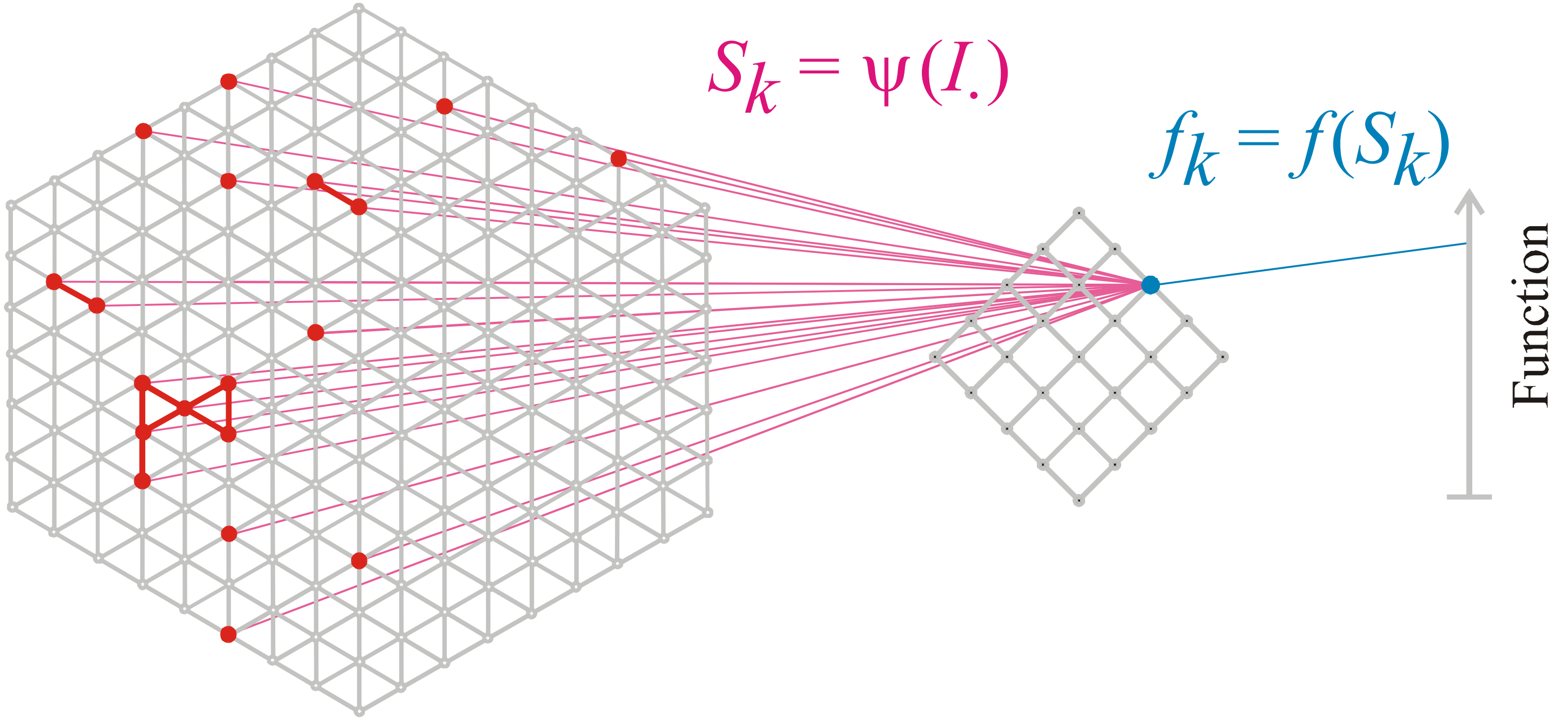
RNA **sequences**, which are mapped onto the same RNA secondary **structure**, are **neutral** with respect to **structure**. The pre-images of structures in sequence space are **neutral networks**. They can be represented by graphs where the edges connect sequences of Hamming distance $d_H = 1$.





Mapping from sequence space into structure space and into function


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



$S_k = \psi(I_.)$

$f_k = f(S_k)$

Sequence space

Structure space

Real numbers

The pre-image of the structure S_k in sequence space is the **neutral network G_k**

Neutral networks are sets of sequences forming the same structure. **G_k** is the pre-image of the structure S_k in sequence space:

$$G_k = y^{-1}(S_k)\ \pi\ \{y_j \mid y(I_j) = S_k\}$$

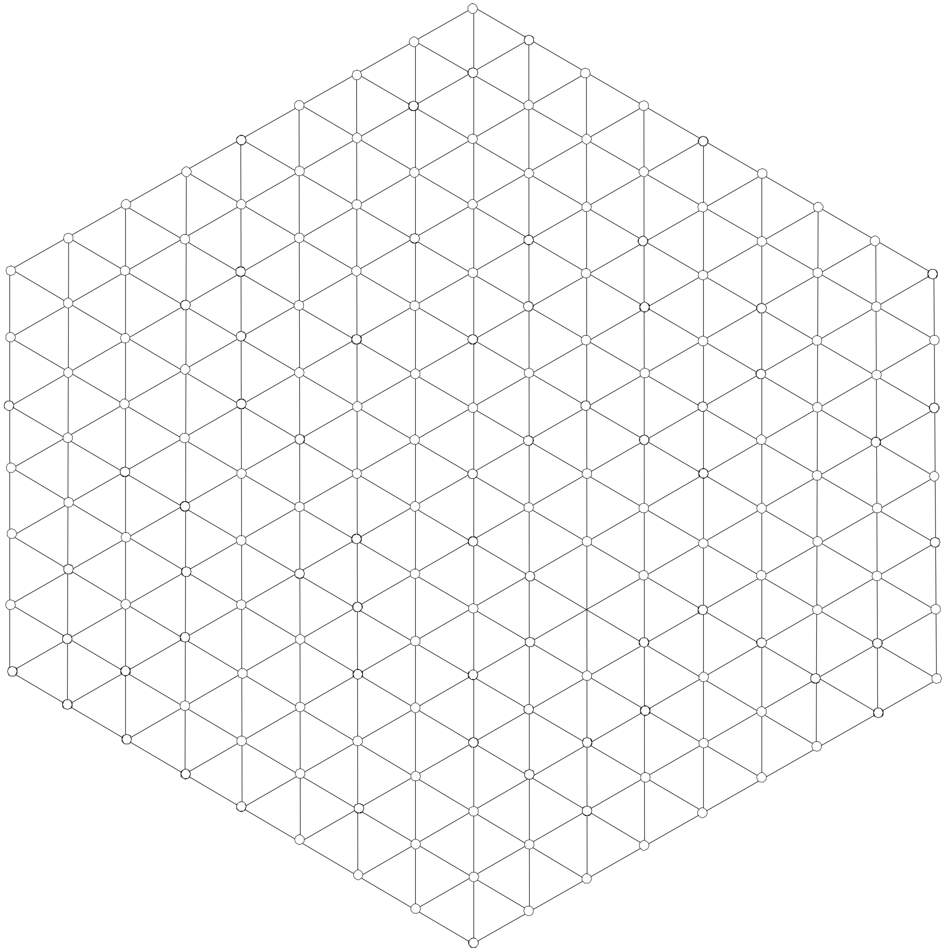
The set is converted into a graph by connecting all sequences of Hamming distance one.

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

Neutral networks can be modelled by **random graphs** in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.

Step 00

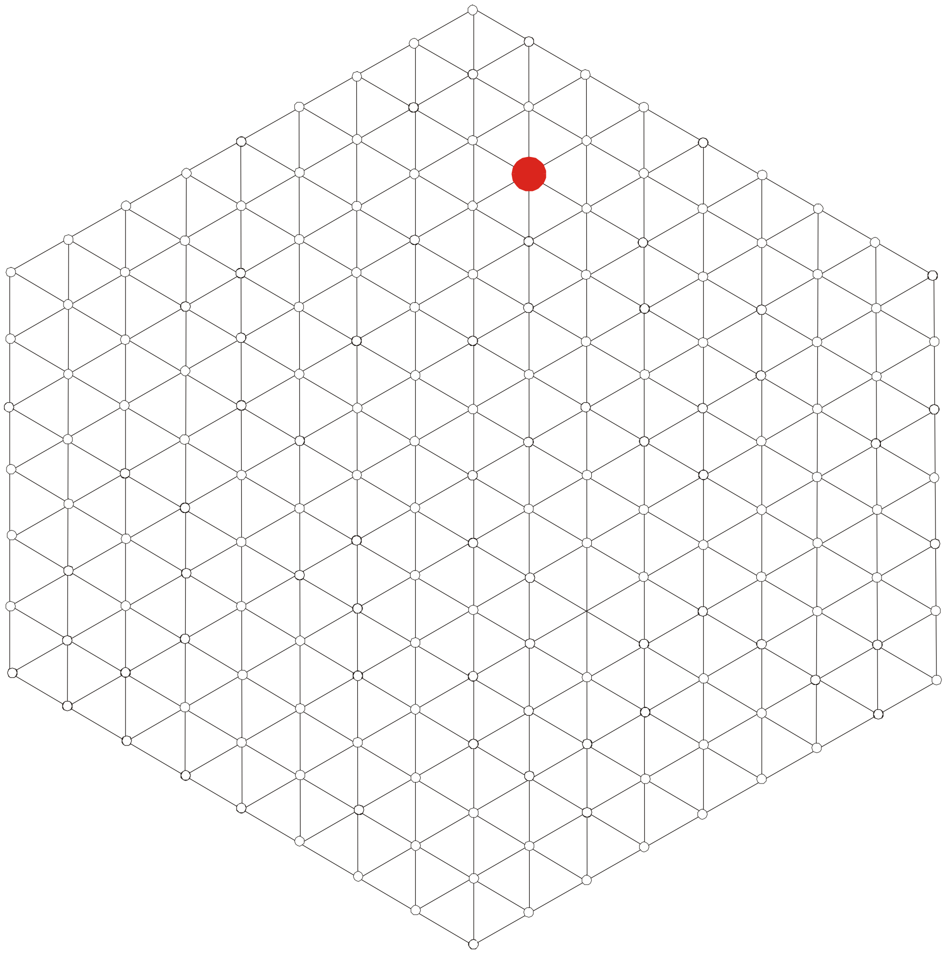
Sketch of sequence space



Random graph approach to neutral networks

Step 01

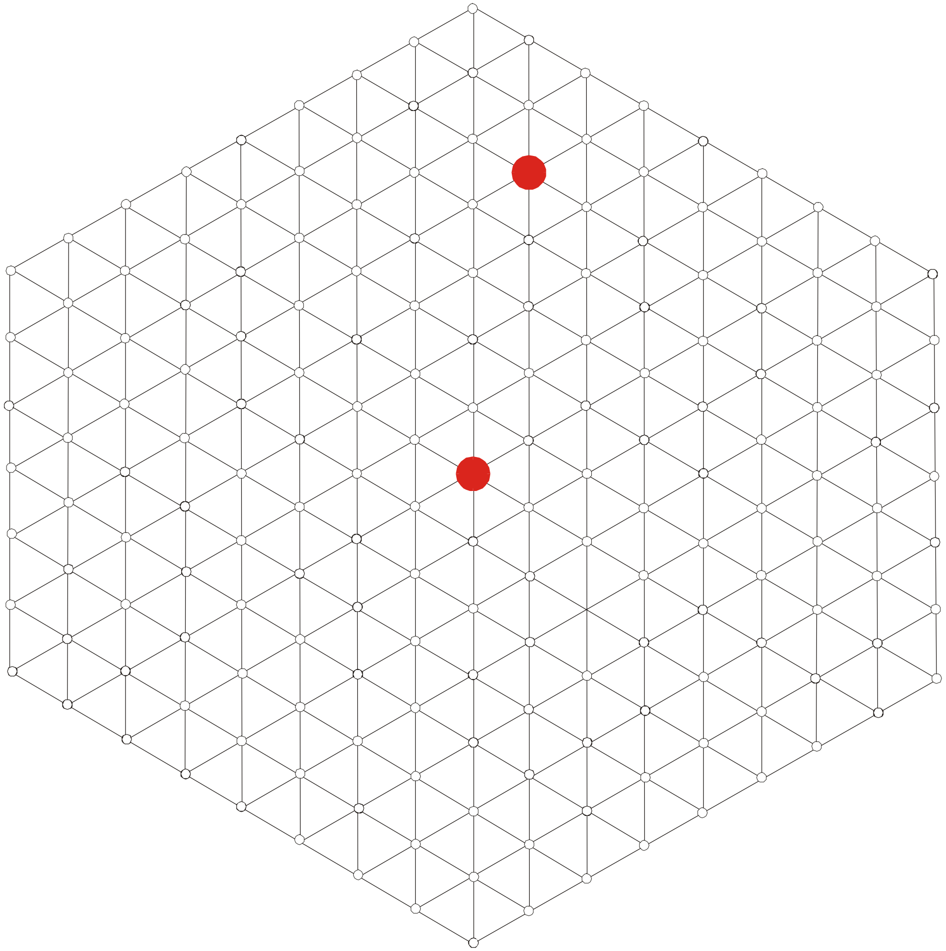
Sketch of sequence space



Random graph approach to neutral networks

Step 02

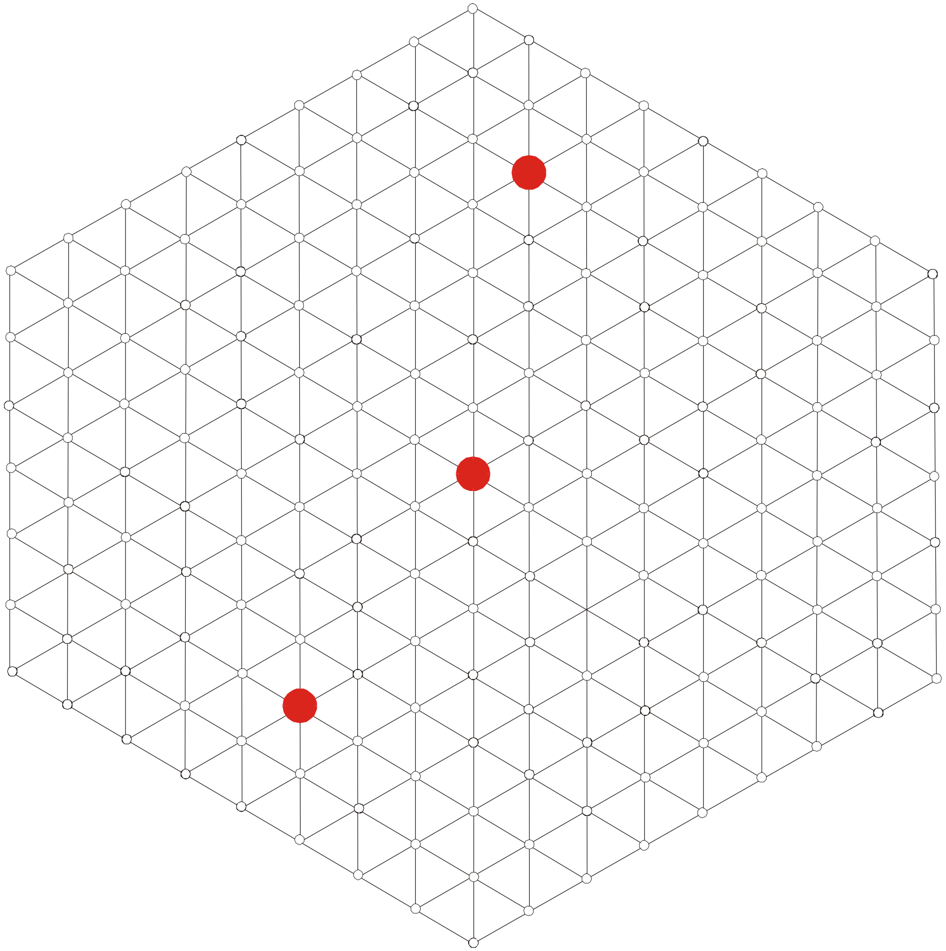
Sketch of sequence space



Random graph approach to neutral networks

Step 03

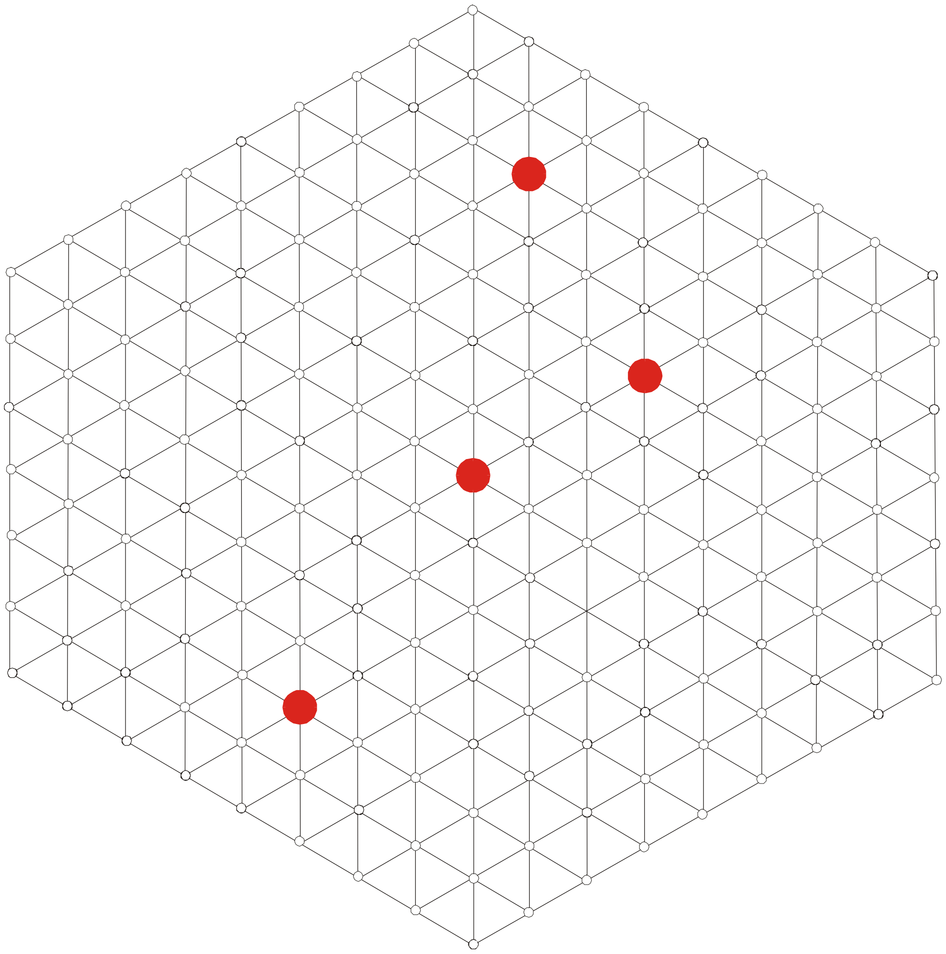
Sketch of sequence space



Random graph approach to neutral networks

Step 04

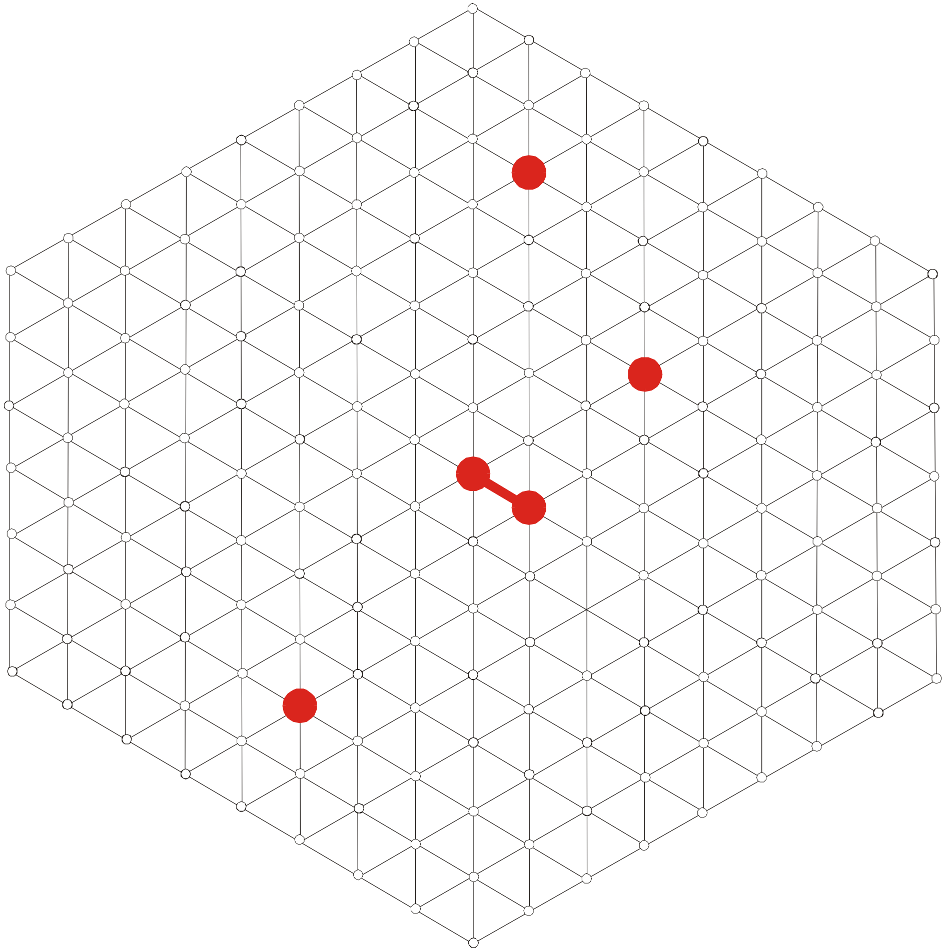
Sketch of sequence space



Random graph approach to neutral networks

Step 05

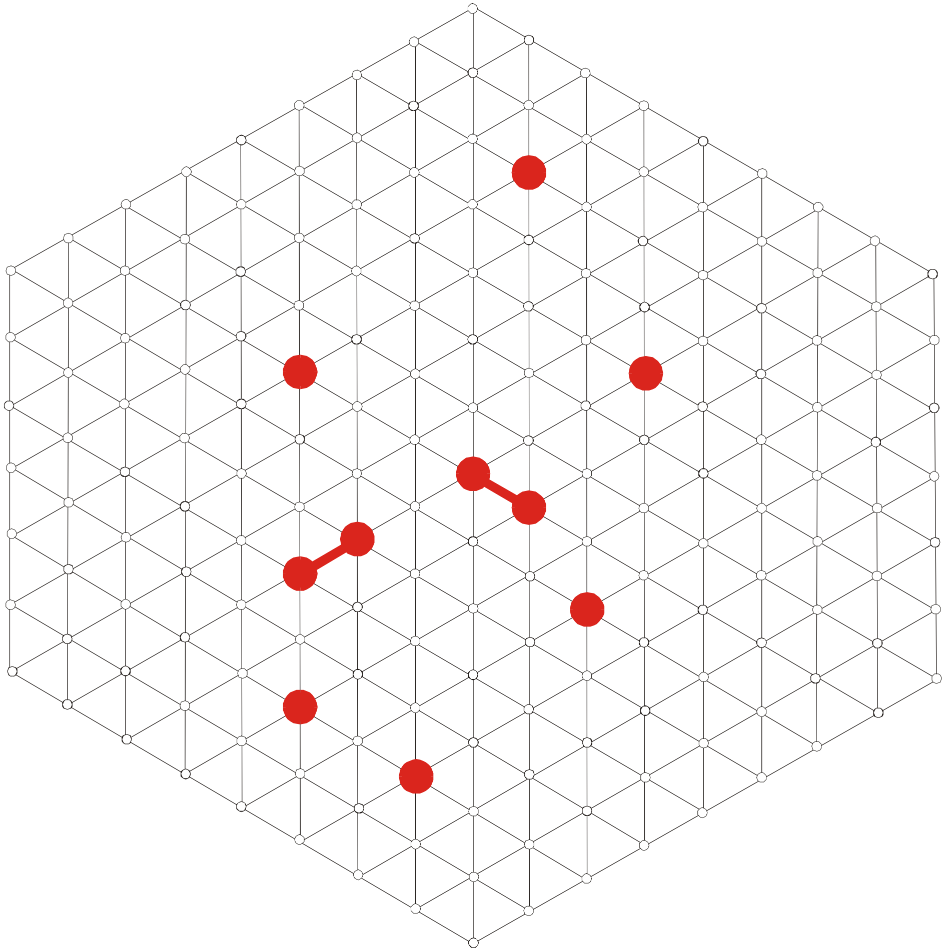
Sketch of sequence space



Random graph approach to neutral networks

Step 10

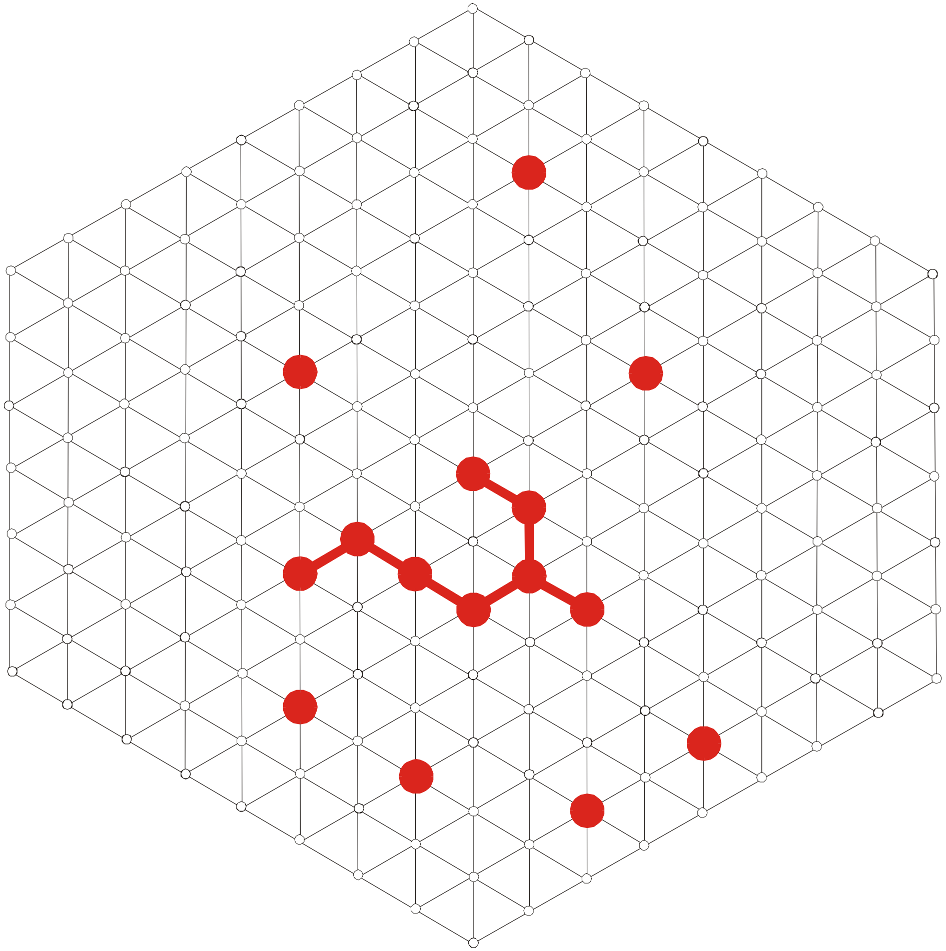
Sketch of sequence space



Random graph approach to neutral networks

Step 15

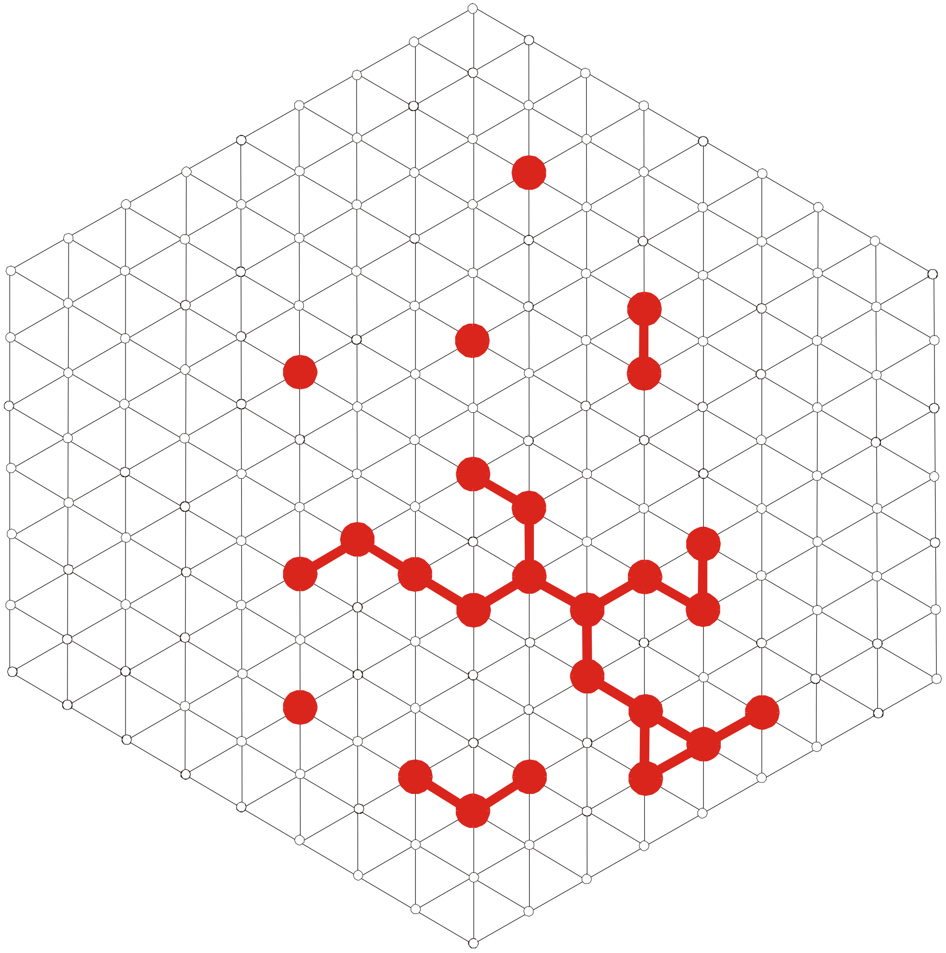
Sketch of sequence space



Random graph approach to neutral networks

Step 25

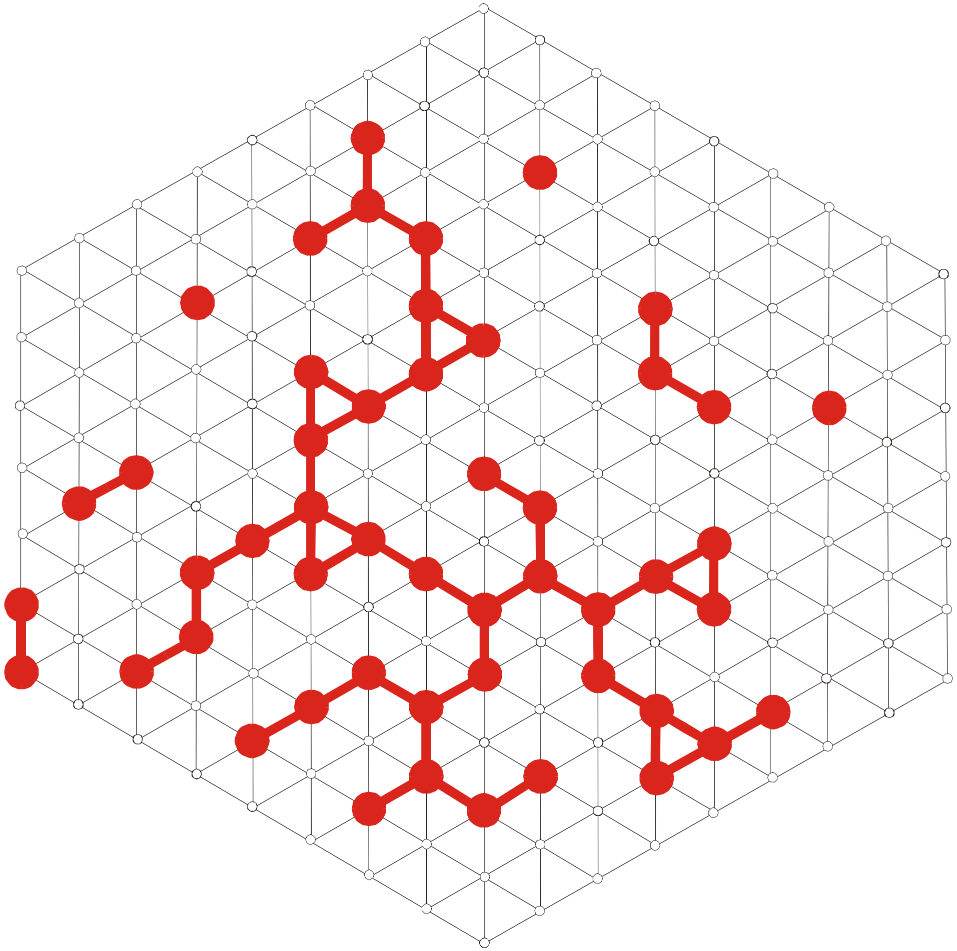
Sketch of sequence space



Random graph approach to neutral networks

Step 50

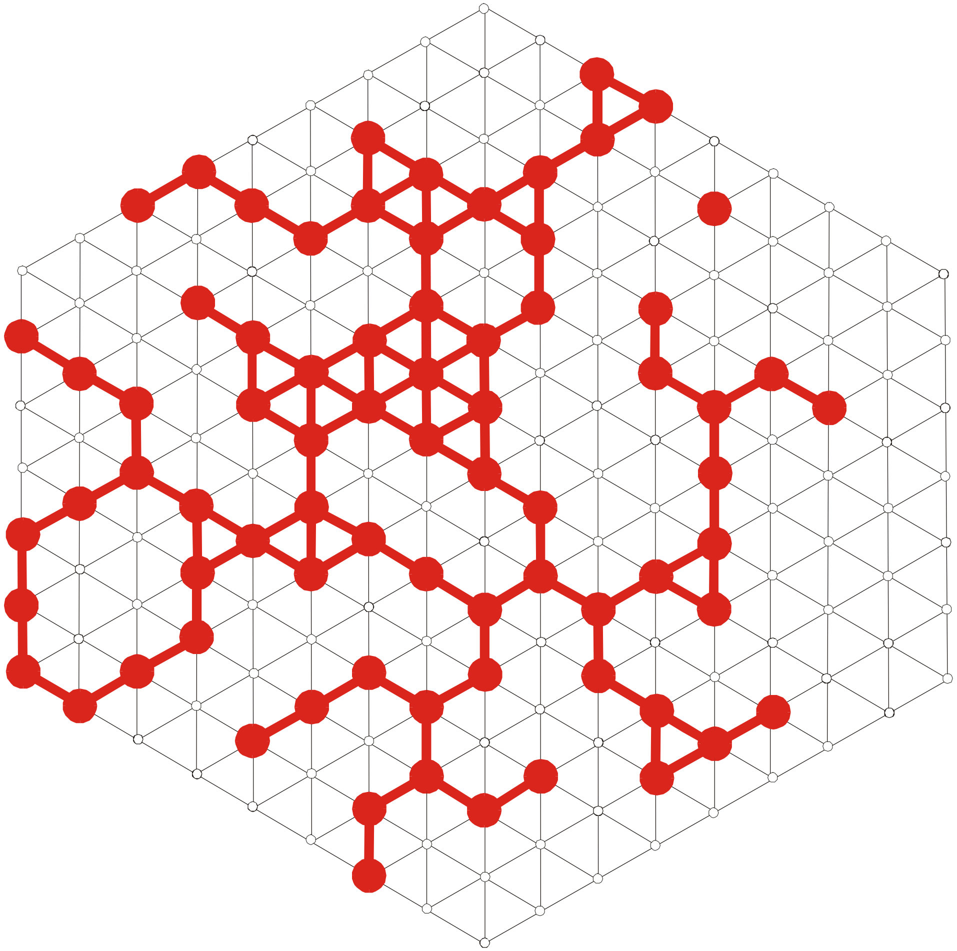
Sketch of sequence space



Random graph approach to neutral networks

Step 75

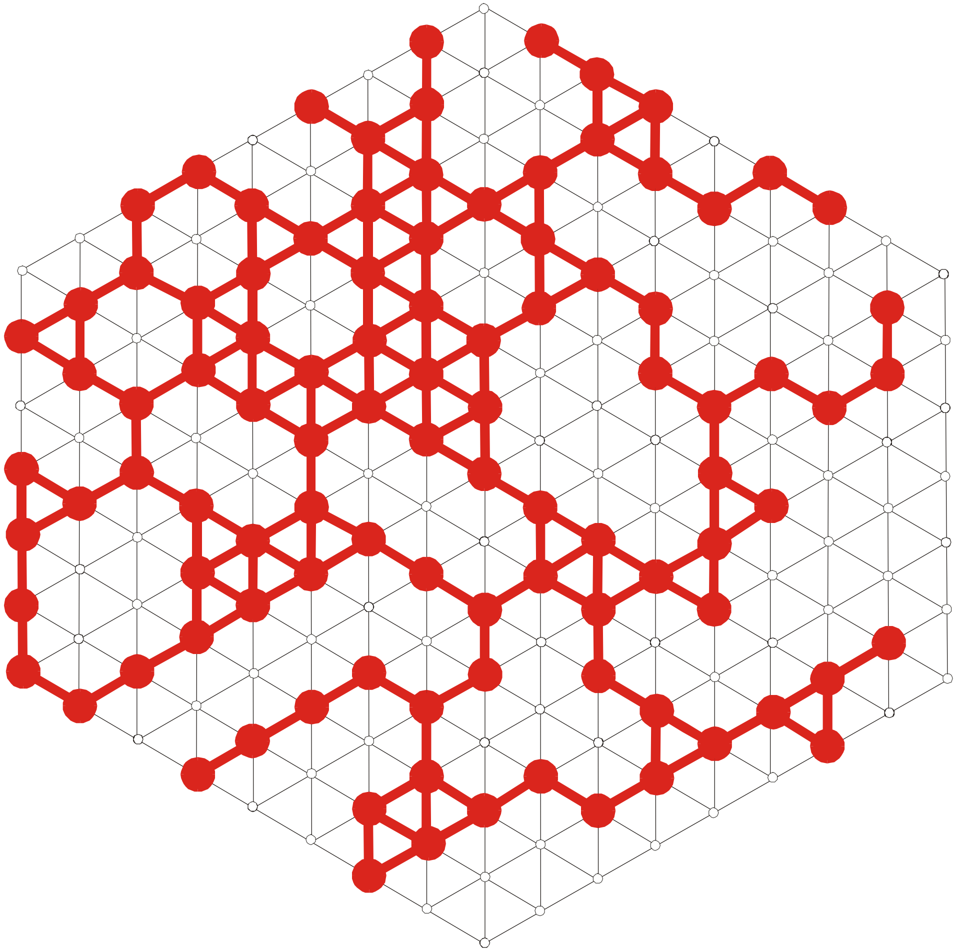
Sketch of sequence space



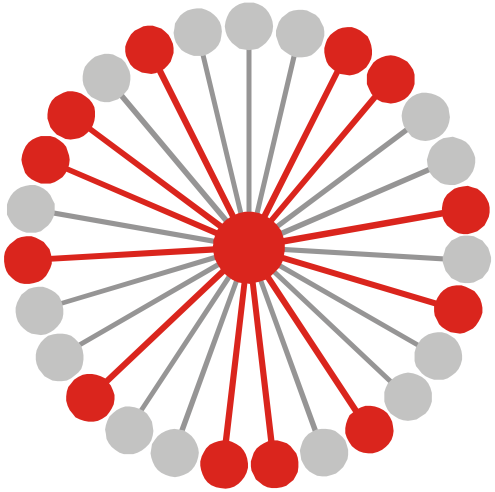
Random graph approach to neutral networks

Step 100

Sketch of sequence space



Random graph approach to neutral networks



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

$$\lambda_j = 12 / 27 = 0.444 \;, \quad \bar{\lambda}_k = \frac{\text{ø}_{j \in |G_k|} \; \text{l}_j(k)}{|G_k|}$$

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

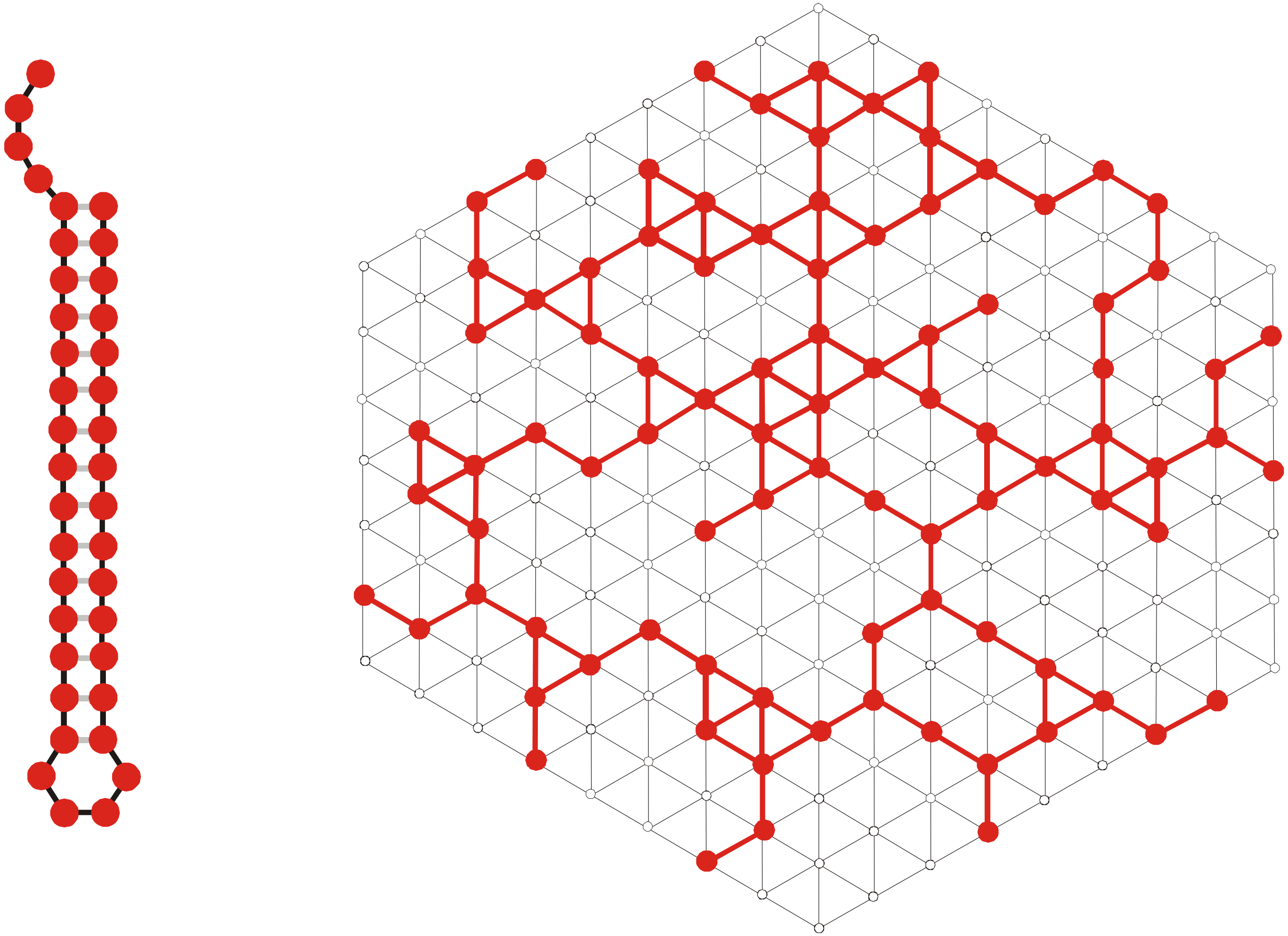
Alphabet size k : **AUGC** ñ k = 4

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

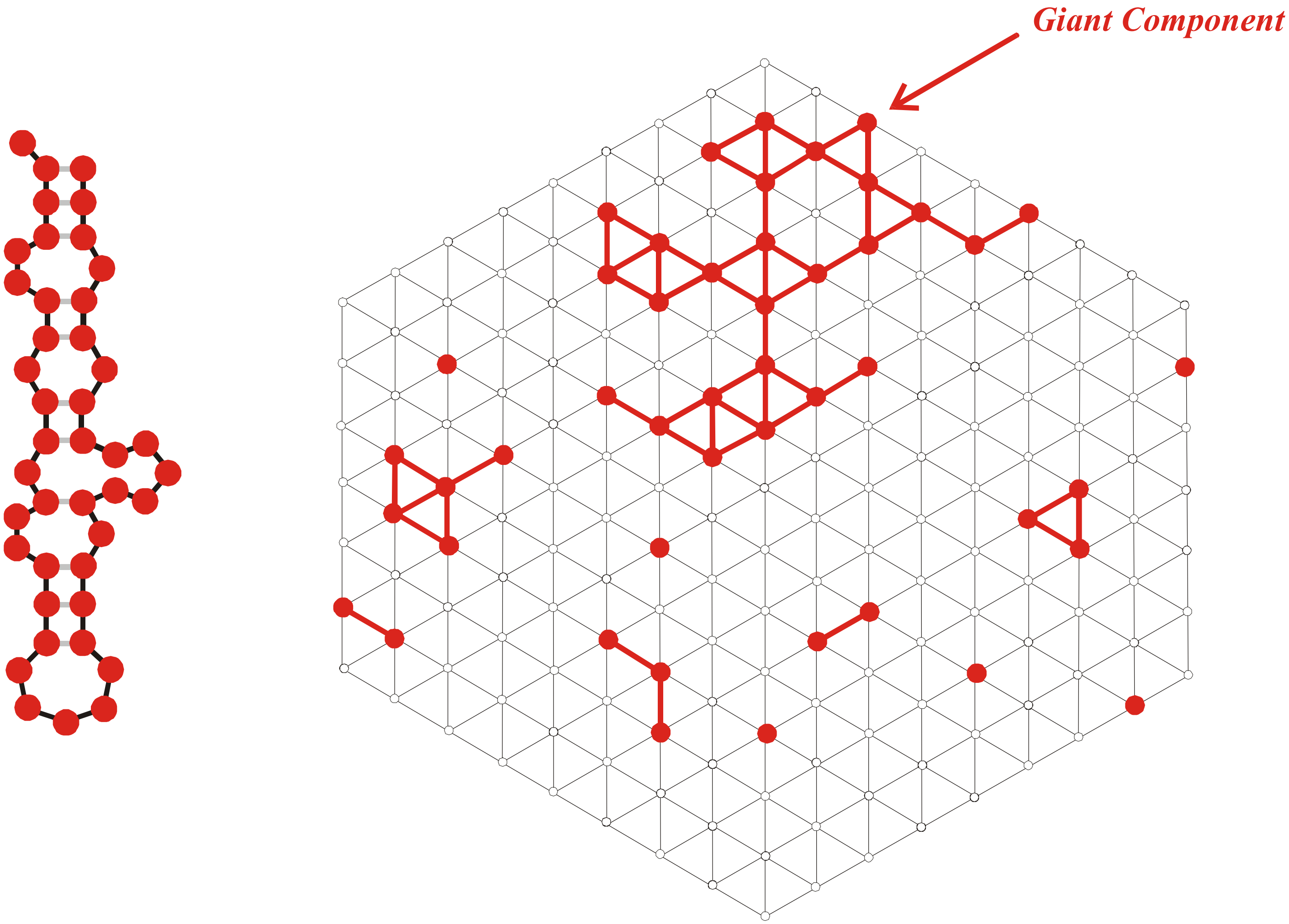
$\bar{\lambda}_k < \lambda_{cr}$ . . . . network G_k is **not** connected

k	l $_{cr}$	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

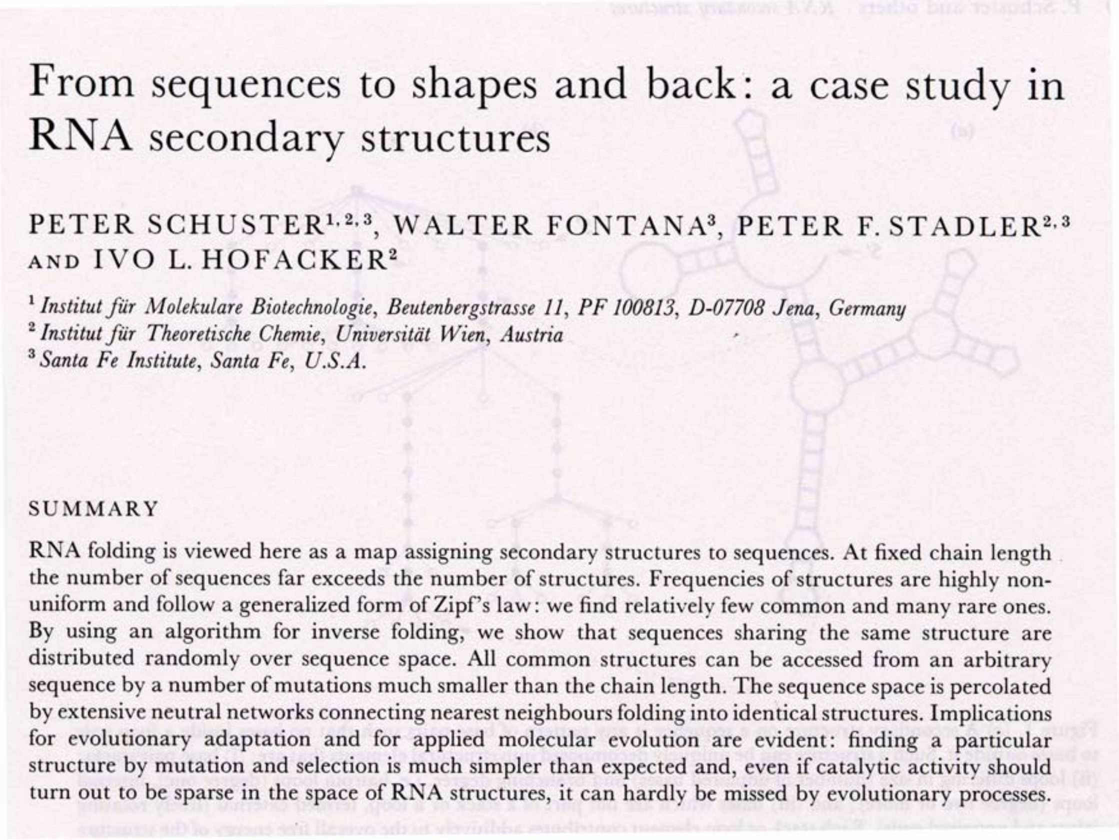
Mean degree of neutrality and connectivity of neutral networks



A connected neutral network



A multi-component neutral network



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.

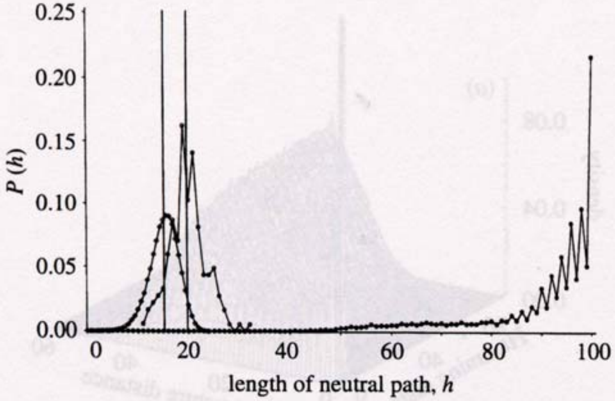
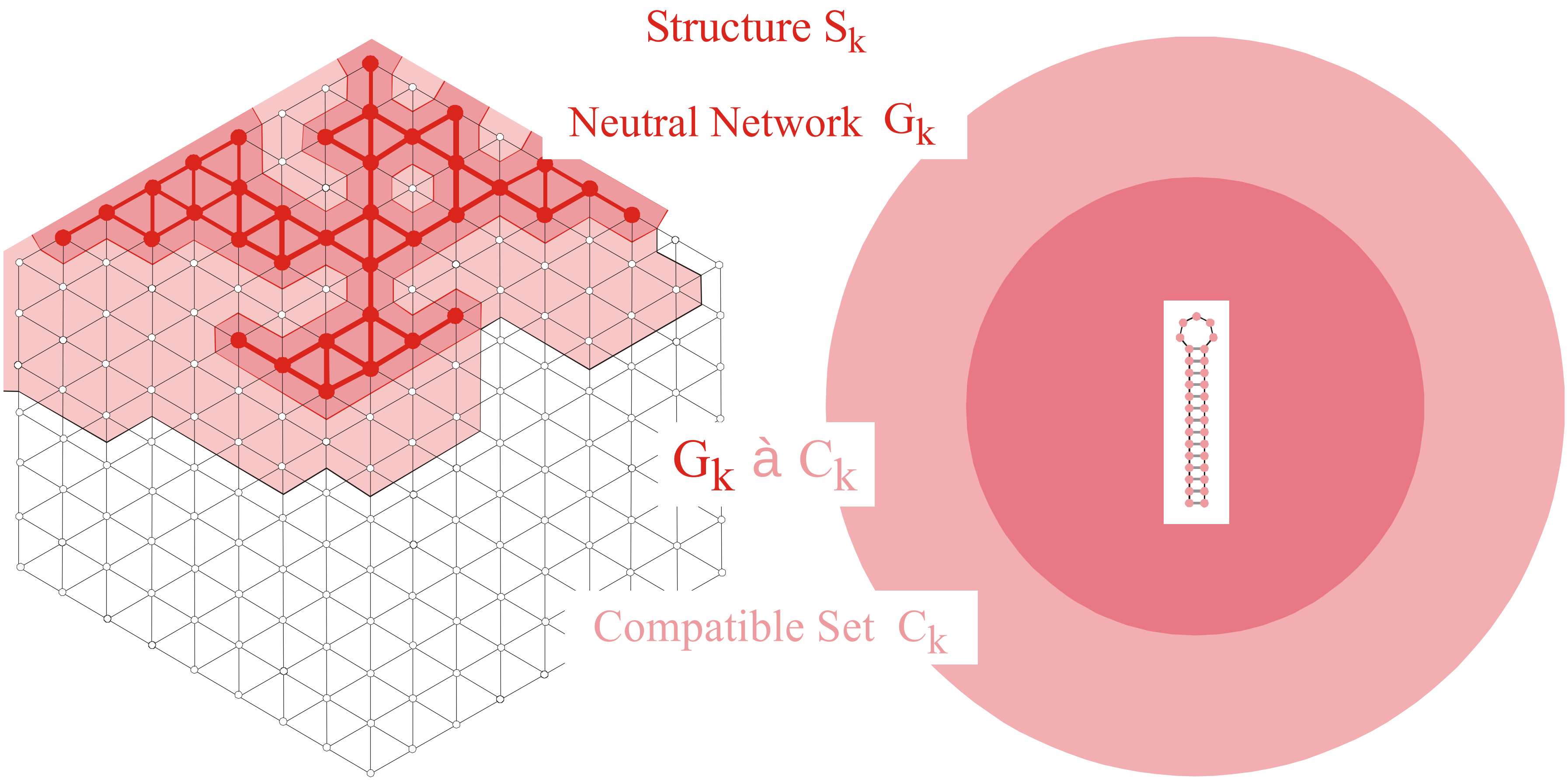


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



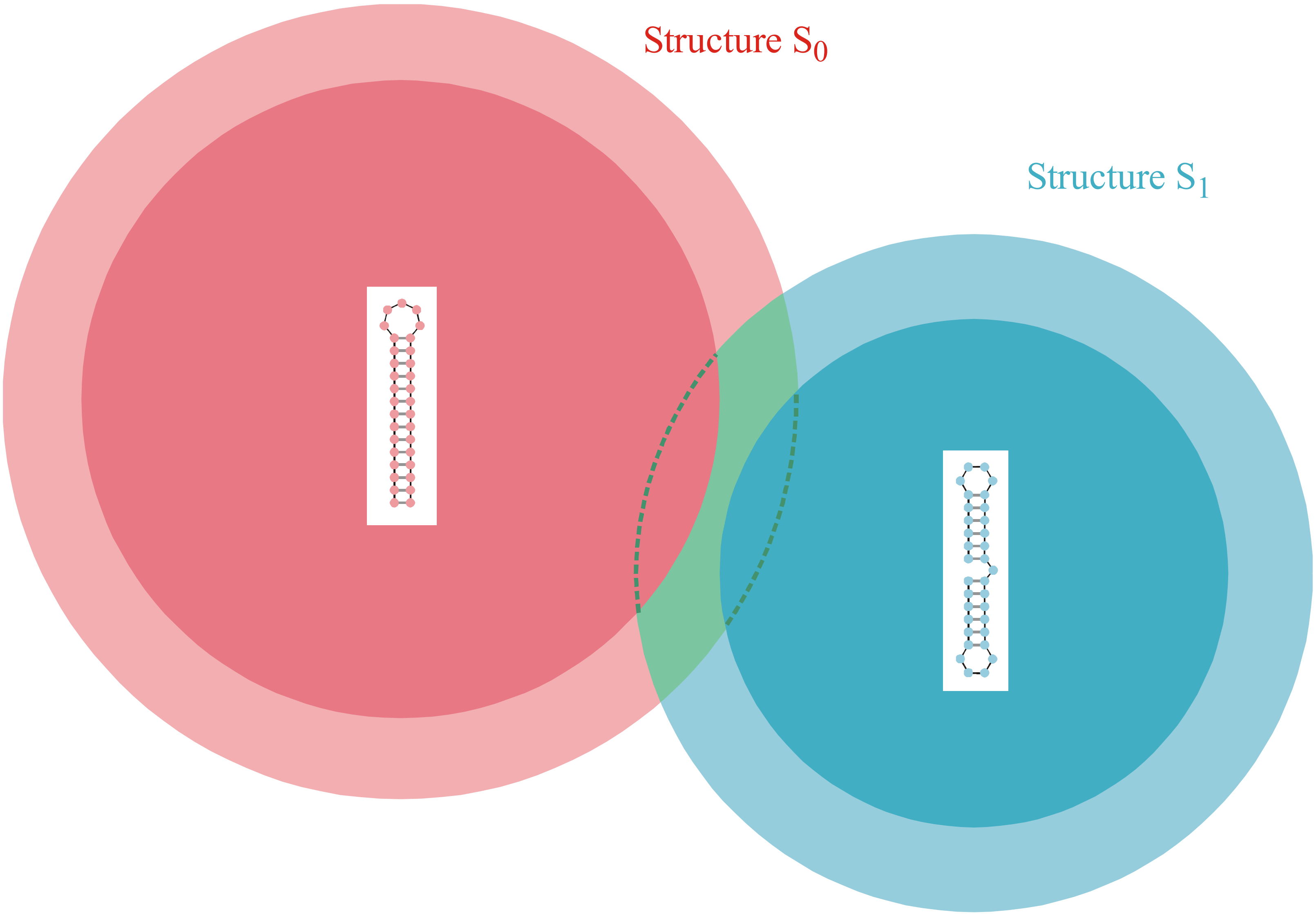
Structure S_k

Neutral Network G_k

G_k à C_k

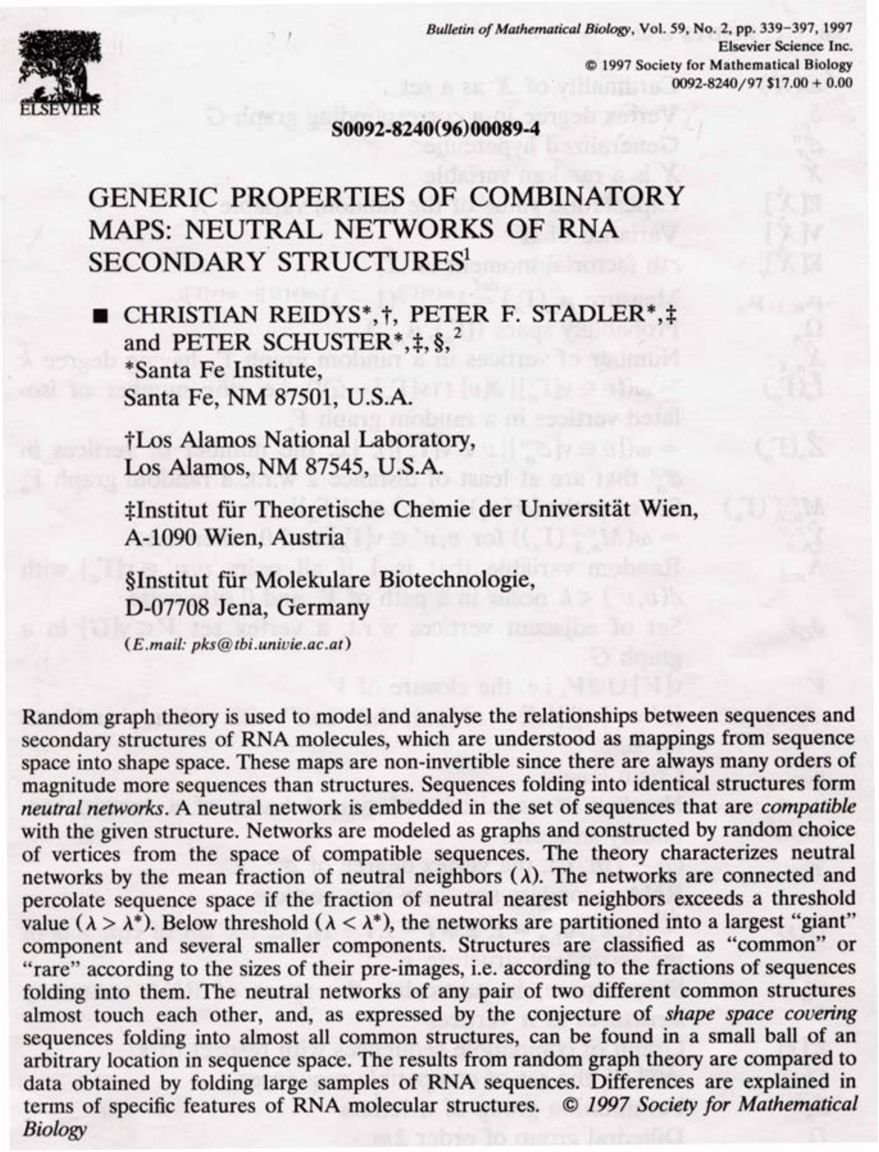
Compatible Set C_k

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



Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: C_0 Ú C_1 â Ù



Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.
© 1997 Society for Mathematical Biology
0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

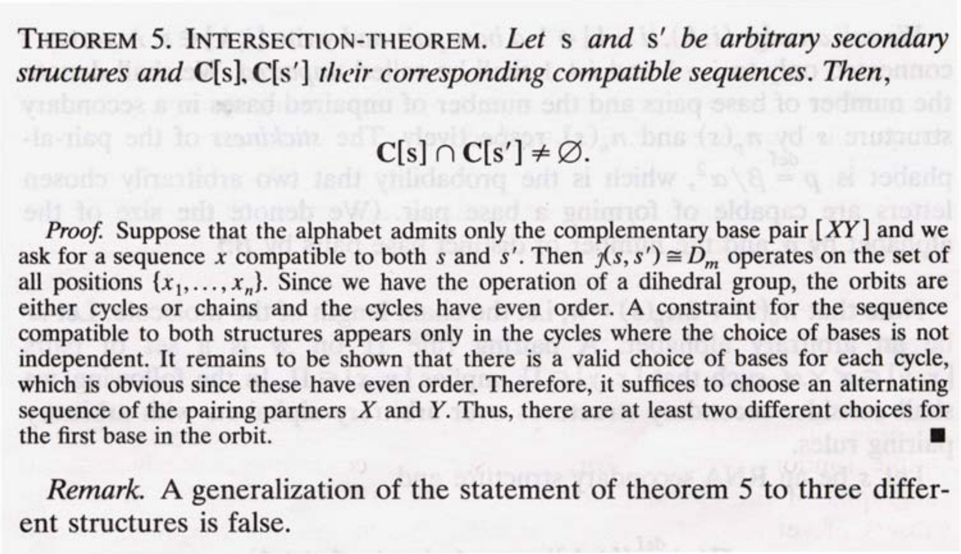
■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡
and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†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*



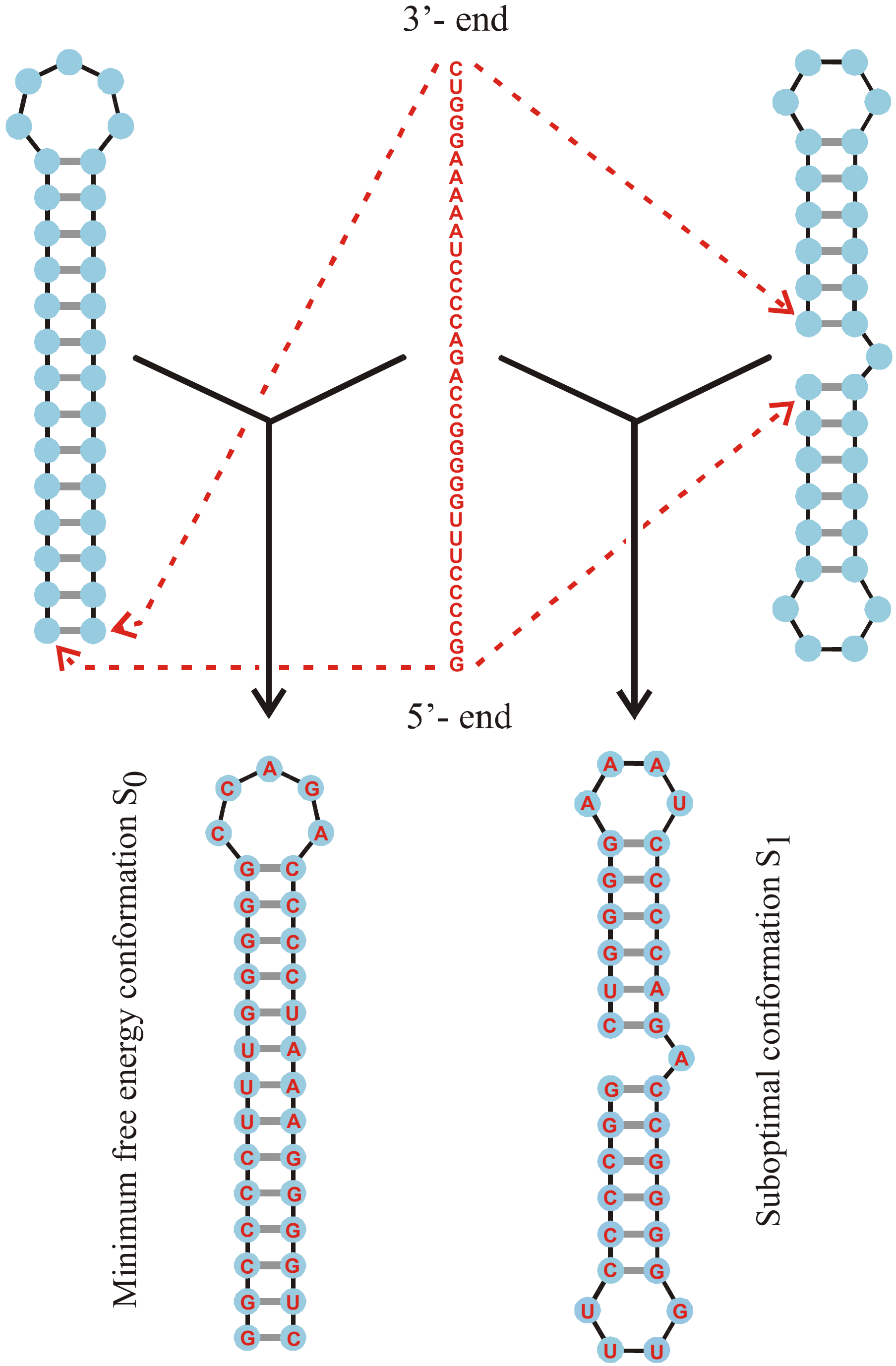
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* C[s], C[s′] *their corresponding compatible sequences. Then,*

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

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 $\mathcal{J}(s,s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore, it suffices to choose an alternating sequence of the pairing partners X and Y. Thus, there are at least two different choices for the first base in the orbit. ■

Remark. A generalization of the statement of theorem 5 to three different structures is false.

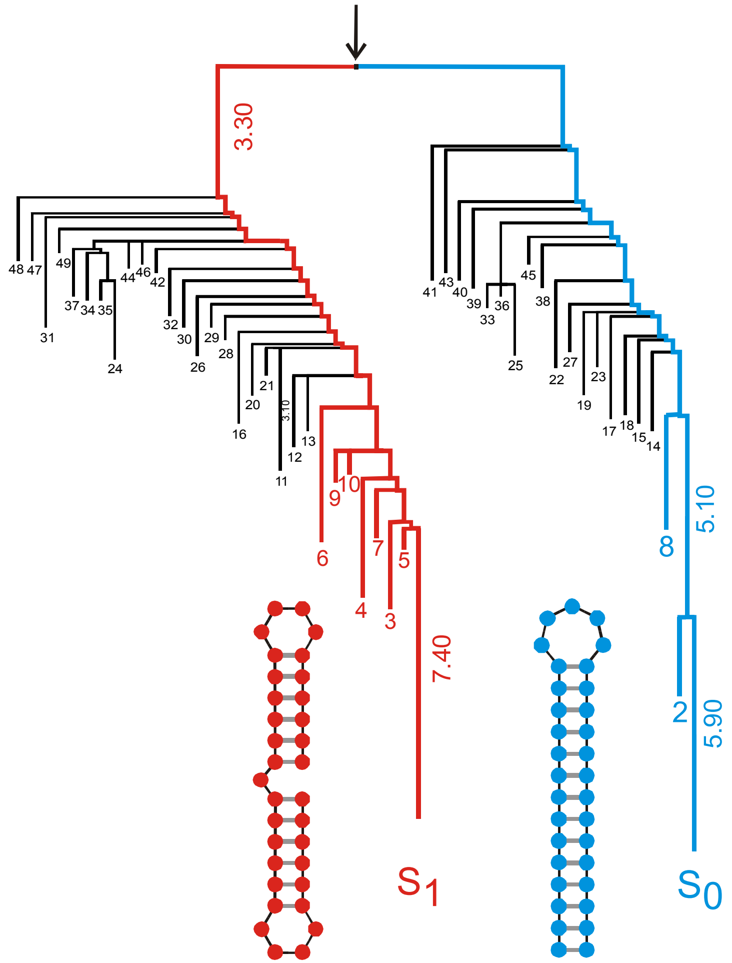
Reference for the definition of the intersection and the proof of the **intersection theorem**



Minimum free energy conformation S_0

Suboptimal conformation S_1

A sequence at the **intersection** of two neutral networks is compatible with both structures



basin '1'

basin '0'

Barrier tree for two long living structures

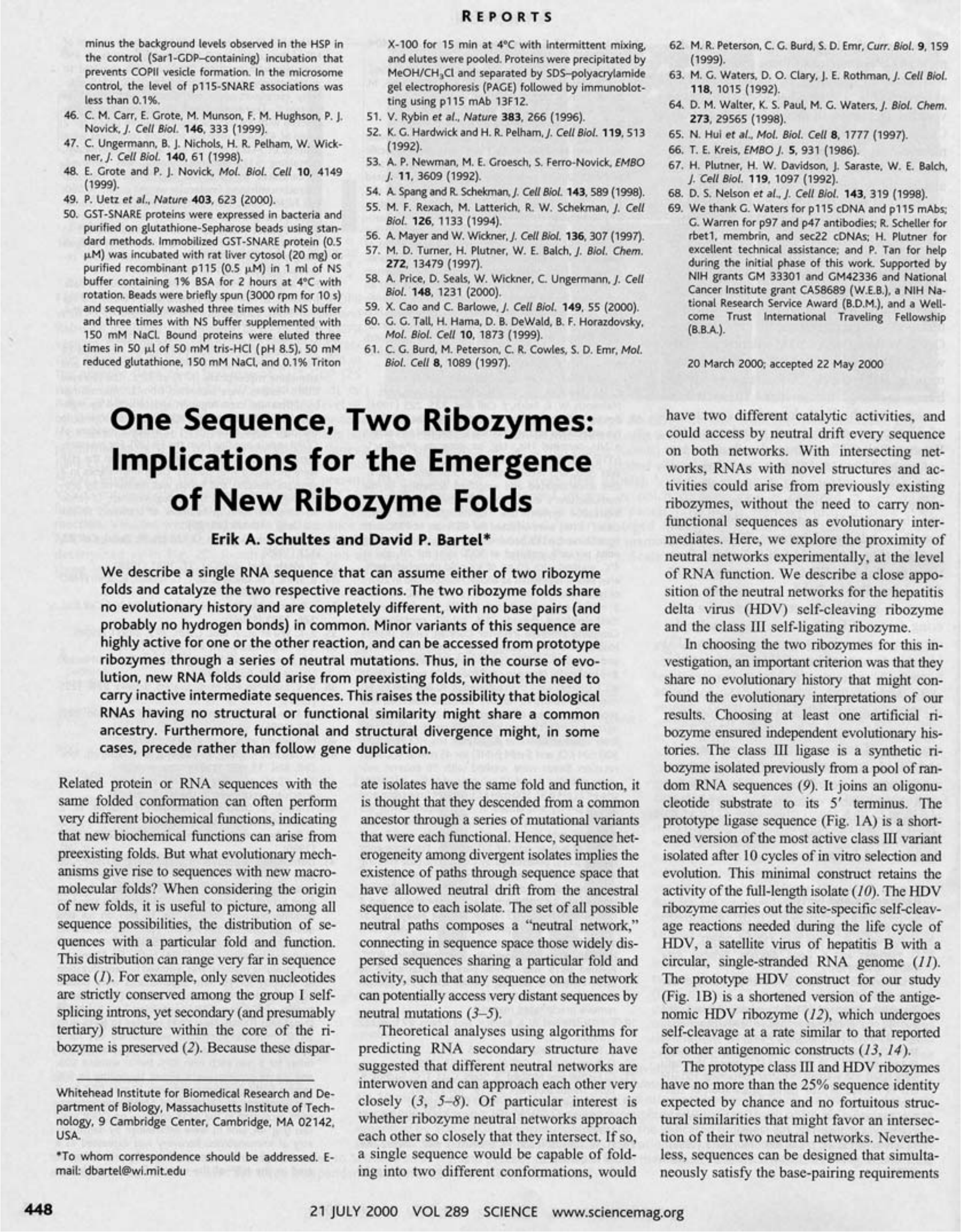
long living metastable structure

minimum free energy structure

1. The role of RNA in the cell and the notion of structure
2. RNA folding
3. Inverse folding of RNA
4. Sequence structure maps, neutral networks, and intersection
5. **Reference to experimental data**
6. Concluding remarks

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*).

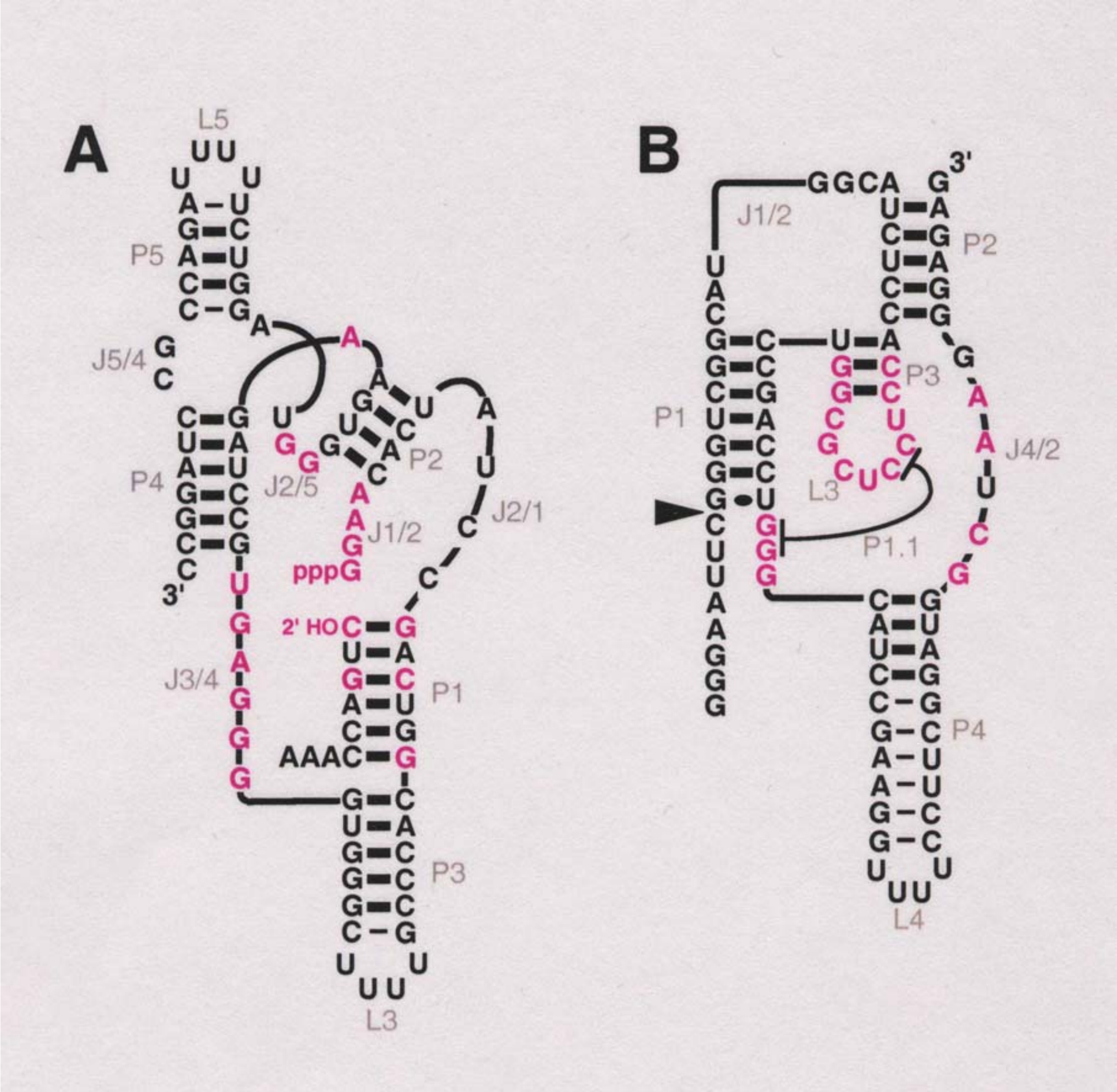
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5*–*8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

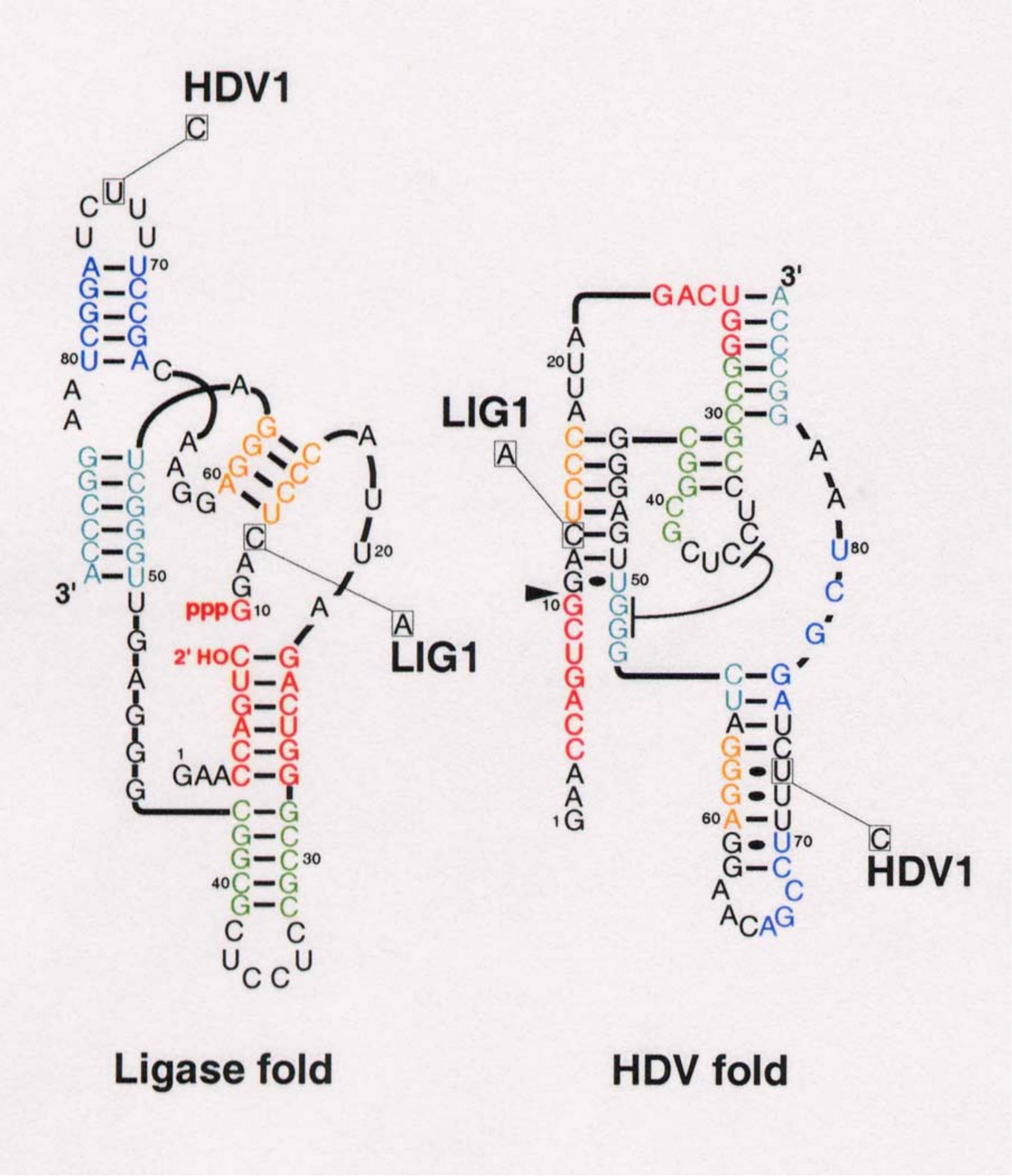
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

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

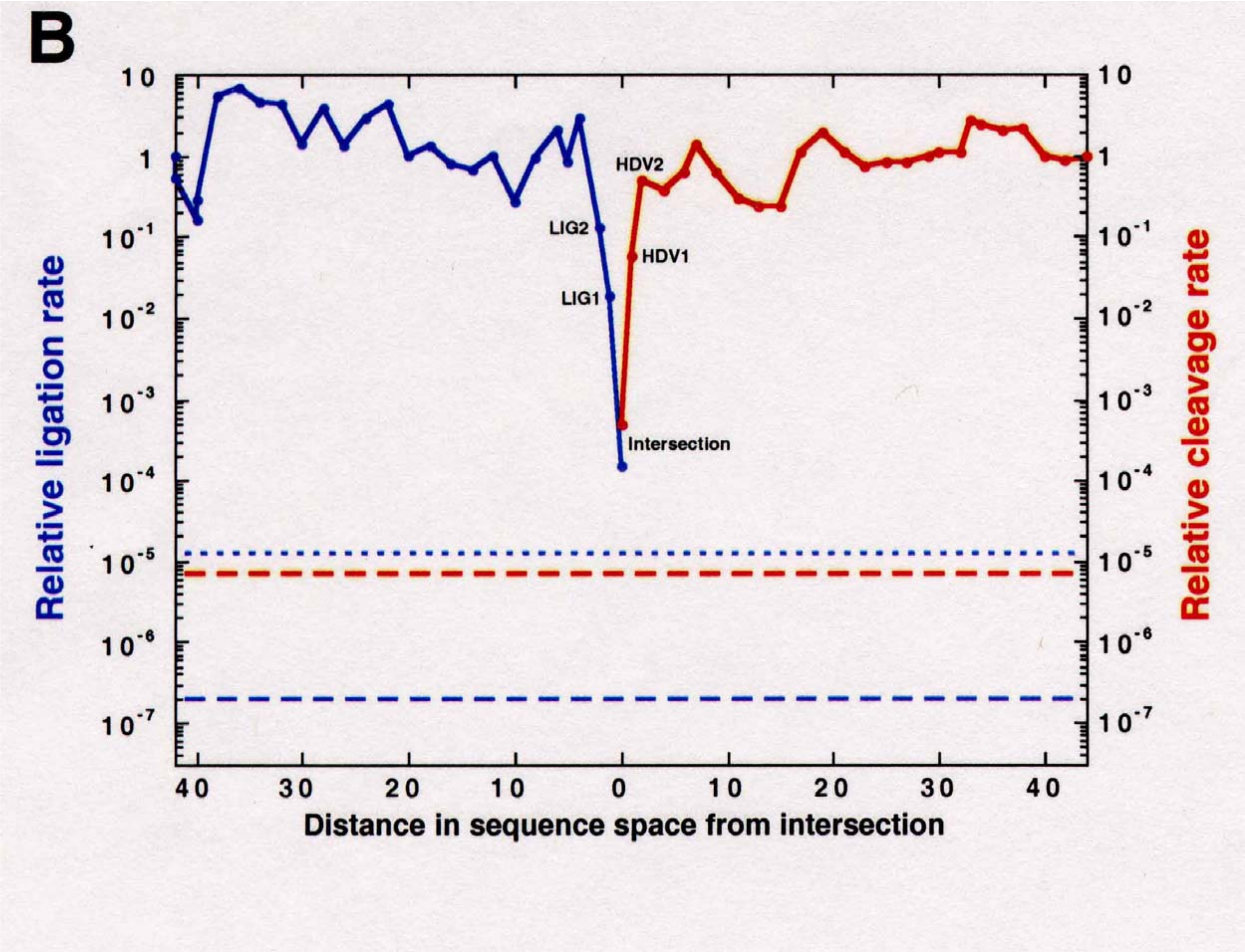


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)



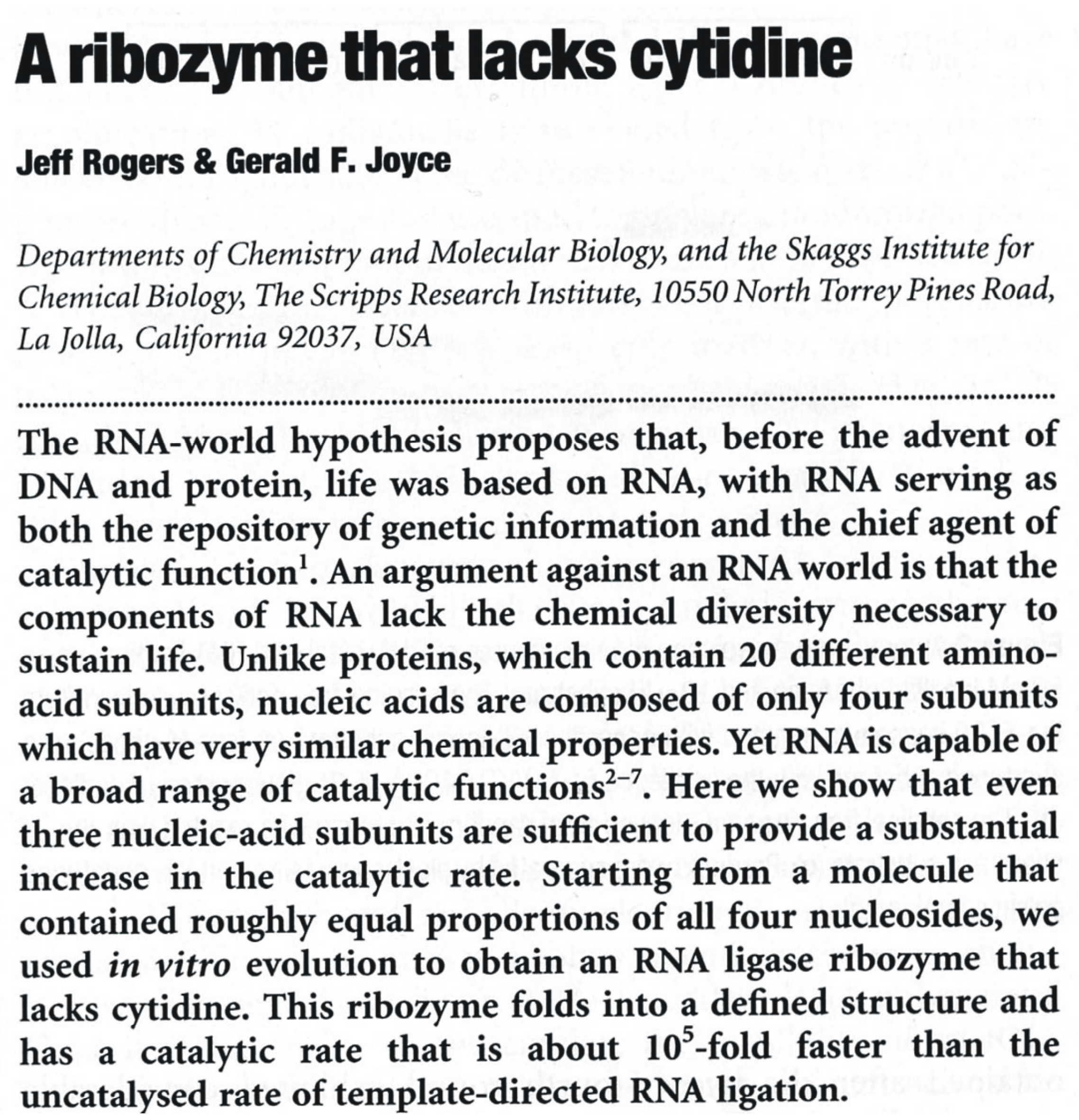
The sequence at the *intersection*:

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

Nature **402**, 323-325, 1999



A ribozyme that lacks cytidine

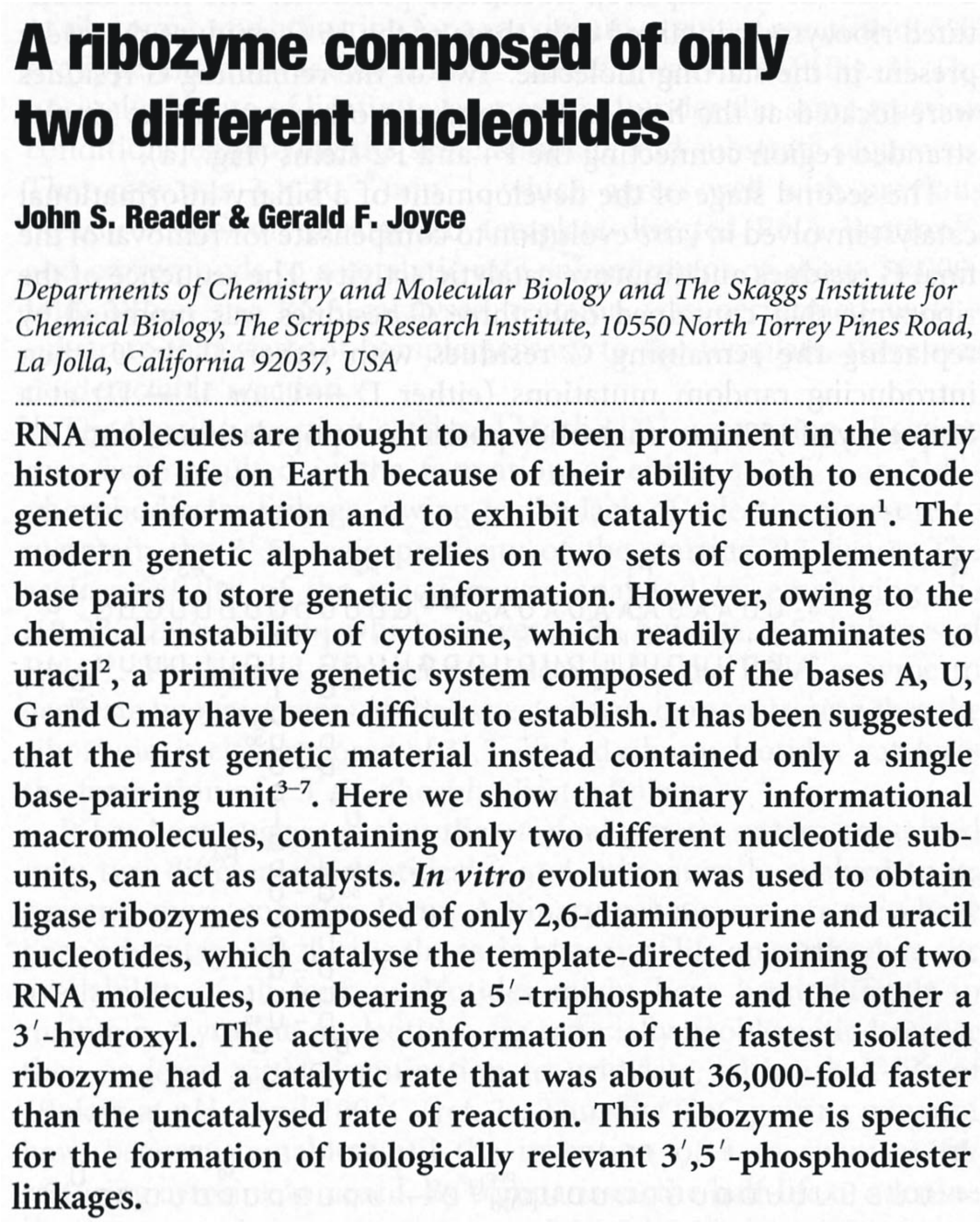
Jeff Rogers & Gerald F. Joyce

Departments of Chemistry and Molecular Biology, and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

The RNA-world hypothesis proposes that, before the advent of DNA and protein, life was based on RNA, with RNA serving as both the repository of genetic information and the chief agent of catalytic function[1]. An argument against an RNA world is that the components of RNA lack the chemical diversity necessary to sustain life. Unlike proteins, which contain 20 different amino-acid subunits, nucleic acids are composed of only four subunits which have very similar chemical properties. Yet RNA is capable of a broad range of catalytic functions[2–7]. Here we show that even three nucleic-acid subunits are sufficient to provide a substantial increase in the catalytic rate. Starting from a molecule that contained roughly equal proportions of all four nucleosides, we used *in vitro* evolution to obtain an RNA ligase ribozyme that lacks cytidine. This ribozyme folds into a defined structure and has a catalytic rate that is about 10^5-fold faster than the uncatalysed rate of template-directed RNA ligation.

Catalytic activity in the **AUG** alphabet

Nature **420**, 841-844, 2002



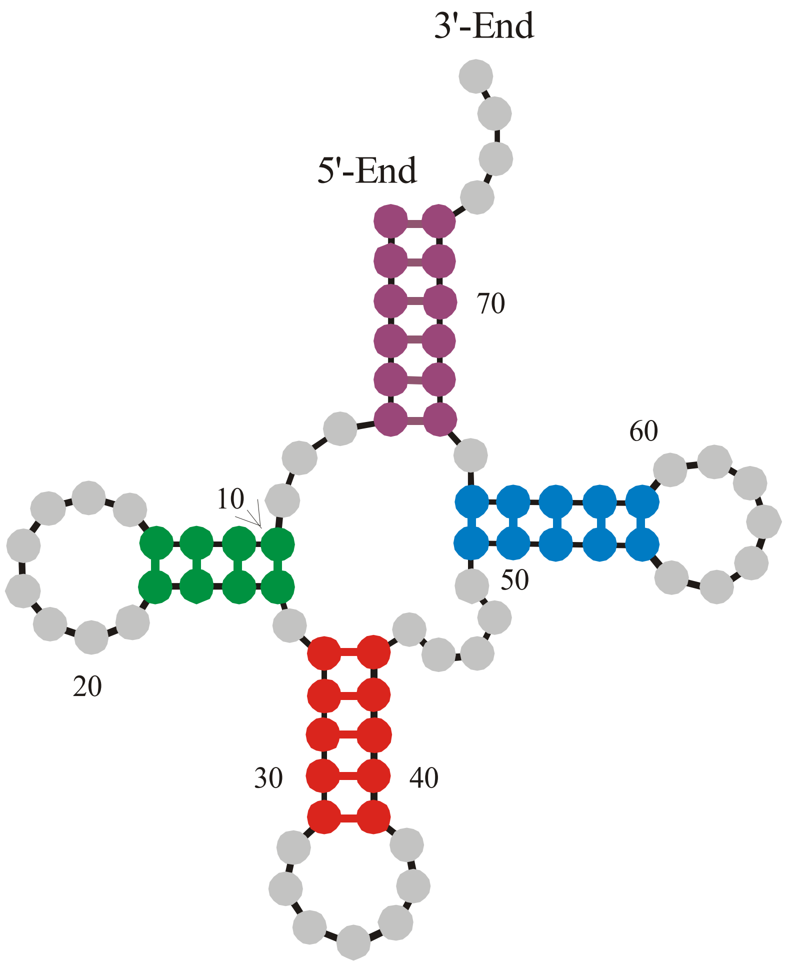
A ribozyme composed of only two different nucleotides

John S. Reader & Gerald F. Joyce

Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

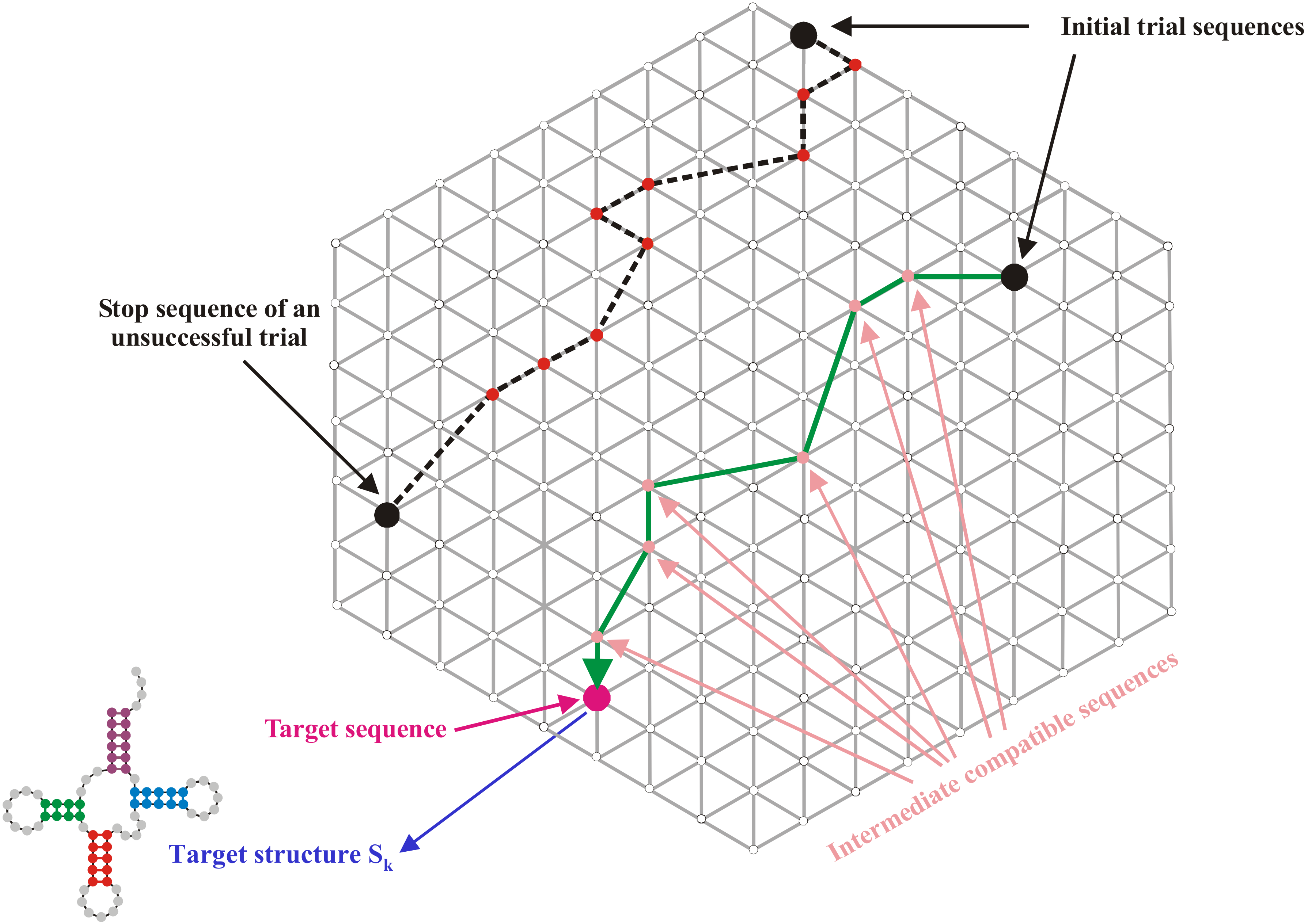
RNA molecules are thought to have been prominent in the early history of life on Earth because of their ability both to encode genetic information and to exhibit catalytic function[1]. The modern genetic alphabet relies on two sets of complementary base pairs to store genetic information. However, owing to the chemical instability of cytosine, which readily deaminates to uracil[2], a primitive genetic system composed of the bases A, U, G and C may have been difficult to establish. It has been suggested that the first genetic material instead contained only a single base-pairing unit[3–7]. Here we show that binary informational macromolecules, containing only two different nucleotide subunits, can act as catalysts. *In vitro* evolution was used to obtain ligase ribozymes composed of only 2,6-diaminopurine and uracil nucleotides, which catalyse the template-directed joining of two RNA molecules, one bearing a 5′-triphosphate and the other a 3′-hydroxyl. The active conformation of the fastest isolated ribozyme had a catalytic rate that was about 36,000-fold faster than the uncatalysed rate of reaction. This ribozyme is specific for the formation of biologically relevant 3′,5′-phosphodiester linkages.

Catalytic activity in the
DU alphabet

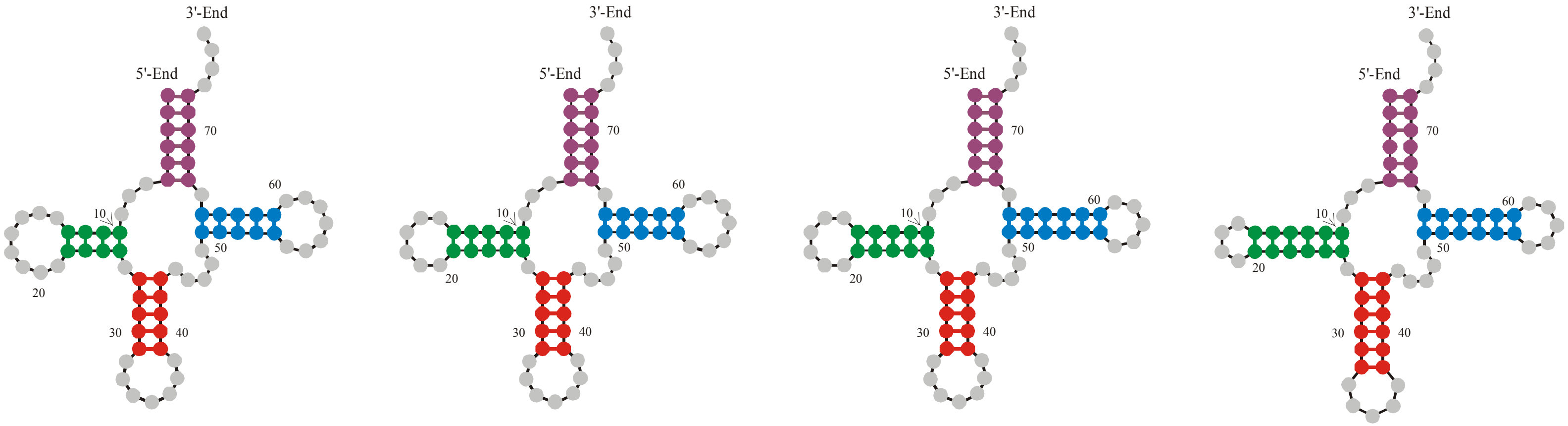


tRNA^{phe}

RNA clover-leaf secondary structures
of sequences with chain length n=76

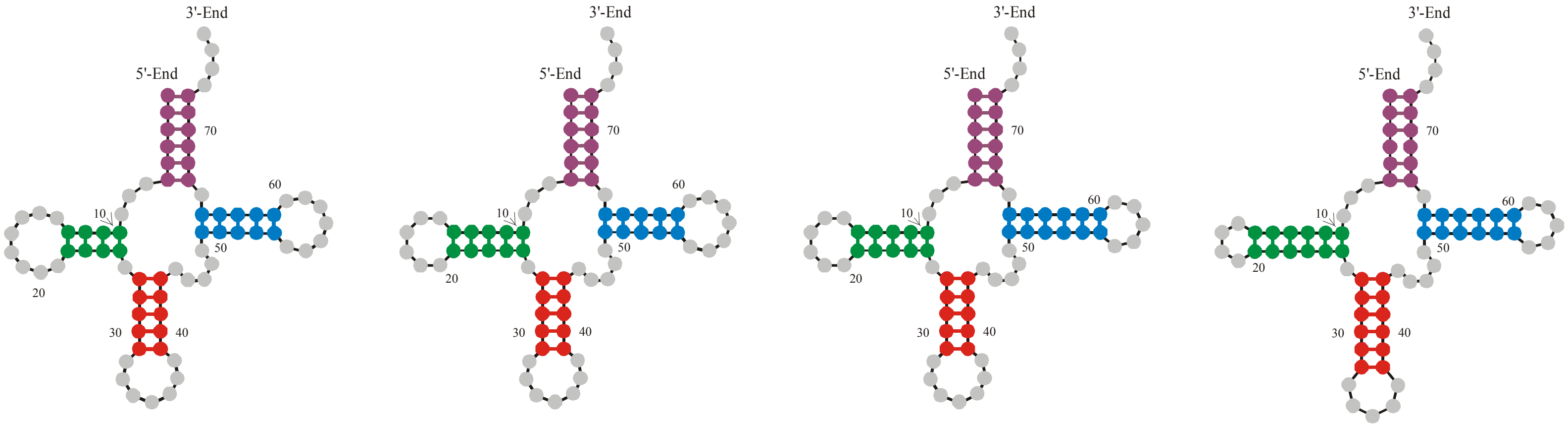

Initial trial sequences
Stop sequence of an unsuccessful trial
Target sequence
Target structure S_k
Intermediate compatible sequences

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



Alphabet	Probability of successful trials in inverse folding			
AU	- -	- -	- -	0.051 Ä 0.006
AUG	- -	0.003 Ä 0.001	0.026 ± 0.006	0.374 Ä 0.016
AUGC	0.794 Ä 0.007	0.884 Ä 0.008	0.934 ± 0.009	0.982 Ä 0.004
UGC	0.548 Ä 0.011	0.628 Ä 0.012	0.697 ± 0.020	0.818 Ä 0.012
GC	0.067 Ä 0.007	0.086 Ä 0.008	0.087 ± 0.008	0.127 Ä 0.006

Accessibility of cloverleaf RNA secondary structures through inverse folding



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

Degree of neutrality of cloverleaf RNA secondary structures over different alphabets

1. The role of RNA in the cell and the notion of structure
2. RNA folding
3. Inverse folding of RNA
4. Sequence structure maps, neutral networks, and intersection
5. Reference to experimental data
6. **Concluding remarks**

Concluding remarks

1. At constant chain lengths the number of RNA sequences exceeds the number of secondary structures by orders of magnitude.

2. The pre-images of common structures in sequence space are extended and connected neutral networks.

3. The intersection of the sets of compatible sequences of two structures is always non-empty.

4. Inverse folding allows for the design of RNA molecules with predefined structures and properties.

Universität Wien

Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank

Project No. Nat-7813

European Commission: Project No. EU-980189

Siemens AG, Austria

The Santa Fe Institute and the Universität Wien

The software for producing RNA movies was developed by Robert Giegerich and coworkers at the Universität Bielefeld

Coworkers

Universität Wien

Walter Fontana, Santa Fe Institute, NM

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

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

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

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

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

Web-Page for further information:

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