RNA – From Mathematical Models to Real Molecules

1. Sequences and Structures

Peter Schuster

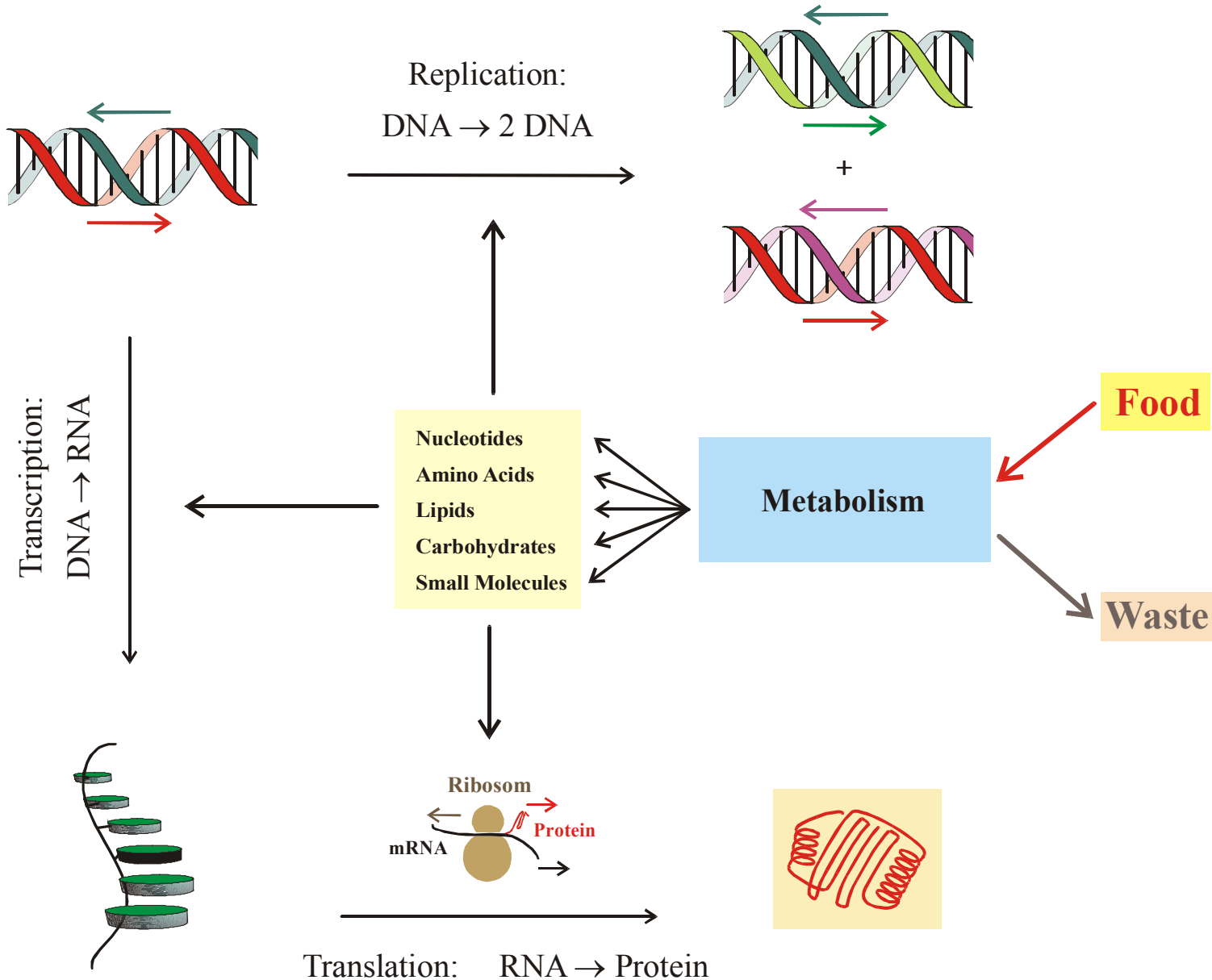
Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

CIMPA – Genoma School

Valdivia, 12.– 16.01.2004

Web-Page for further information:

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



A sketch of cellular DNA metabolism

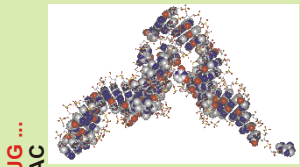
RNA

RNA *as transmitter of genetic information*

DNA
↓ transcription
...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...
messenger-RNA
↓ translation
protein

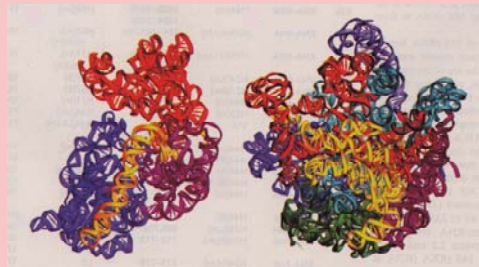
RNA as **working copy** of genetic information

RNA *as adapter molecule*



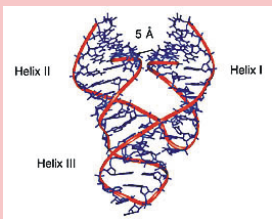
genetic code

RNA *is the catalytic subunit in supramolecular complexes*



The ribosome is a ribozyme !

RNA *as catalyst*



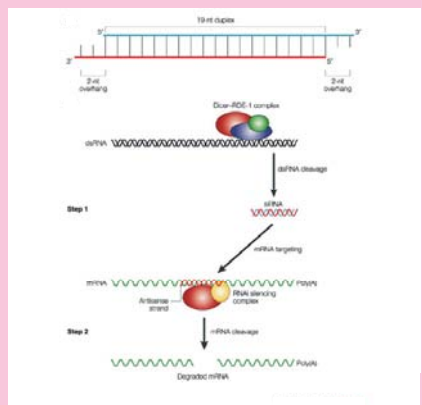
ribozyme

RNA *is modified by epigenetic control*

RNA editing

Alternative splicing of messenger RNA

RNA *as regulator of gene expression*



gene silencing by small interfering RNAs

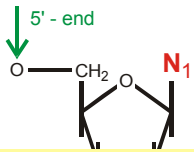
The RNA *world as a precursor of the current* DNA + protein *biology*

RNA *as carrier of genetic information*

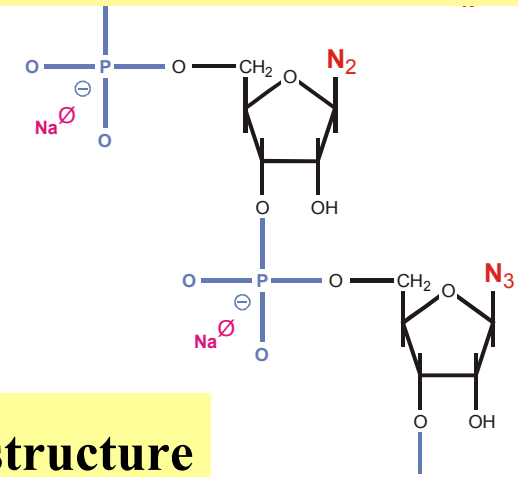
RNA viruses and retroviruses

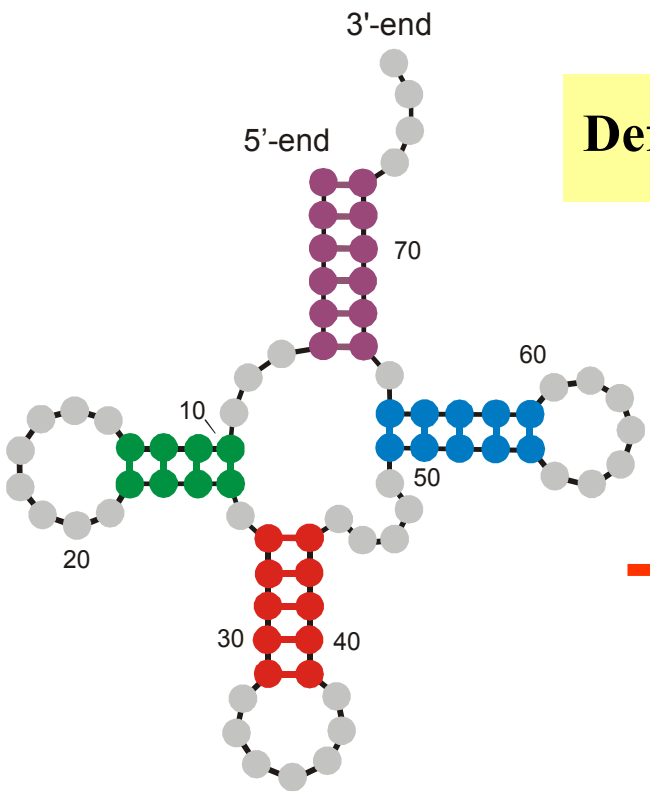
RNA as information carrier in evolution *in vitro* and evolutionary biotechnology

Functions of RNA molecules

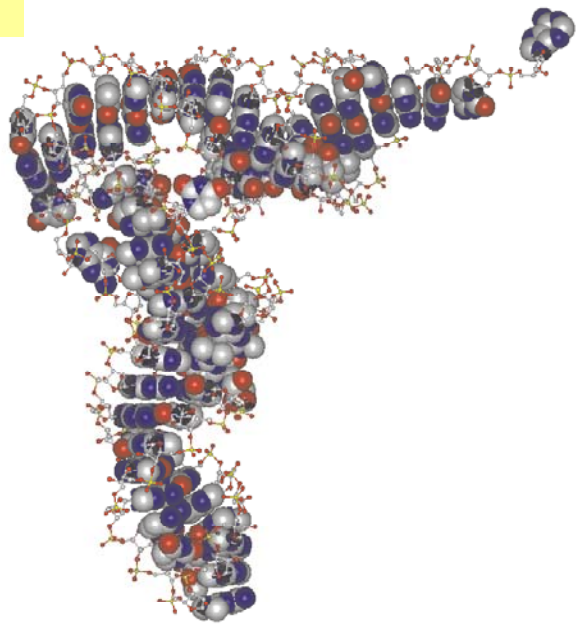


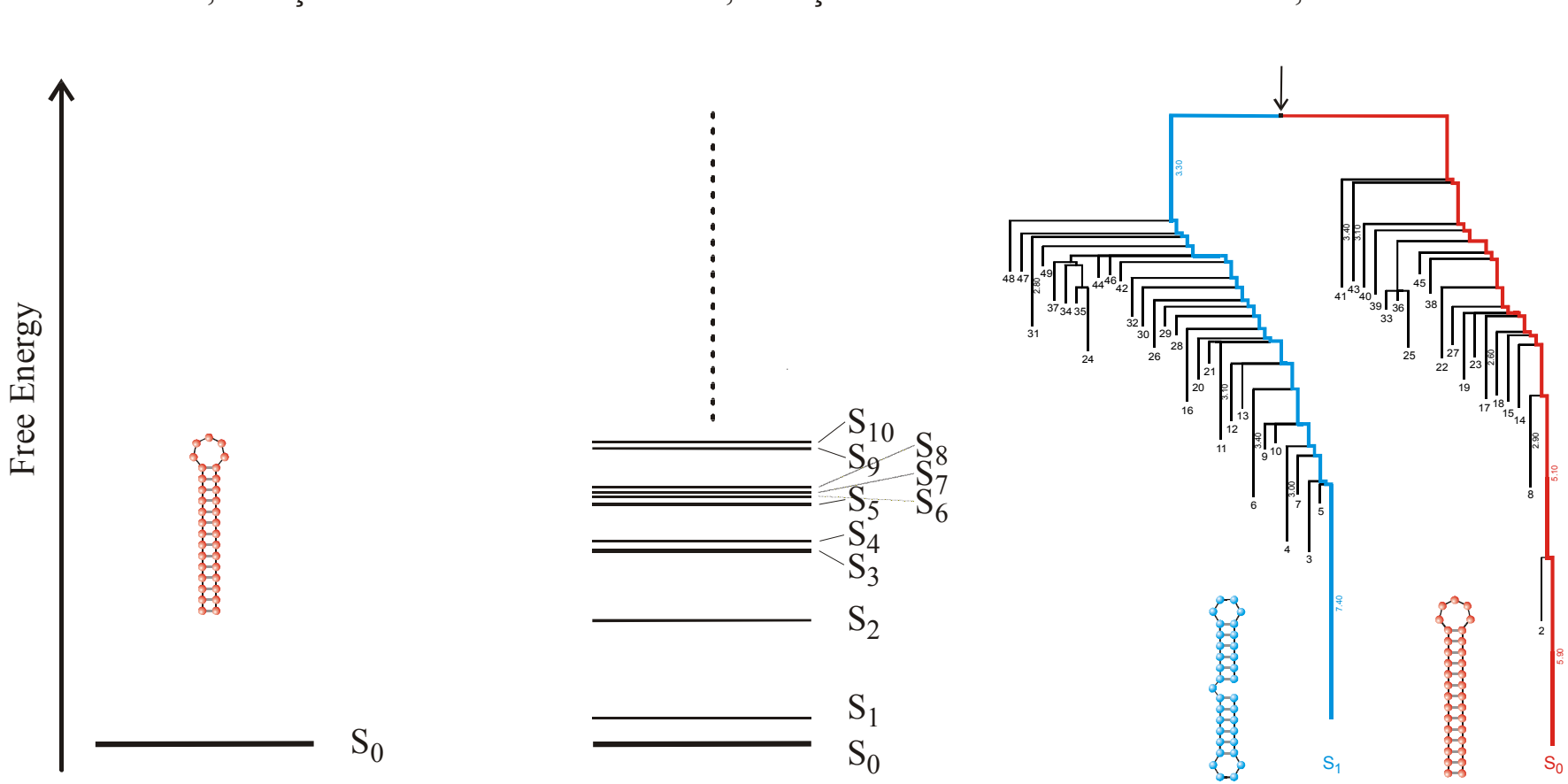
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure





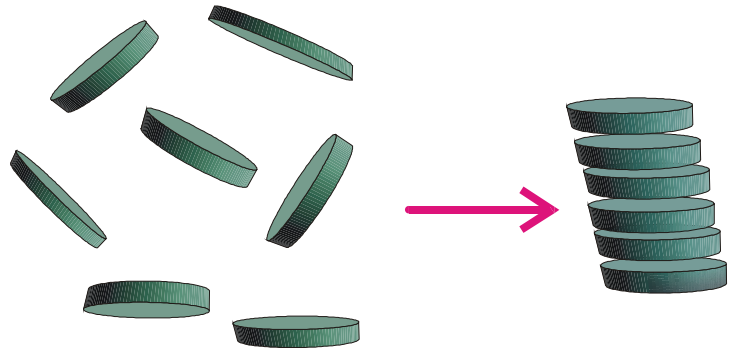
Minimum Free Energy Structure

Suboptimal Structures

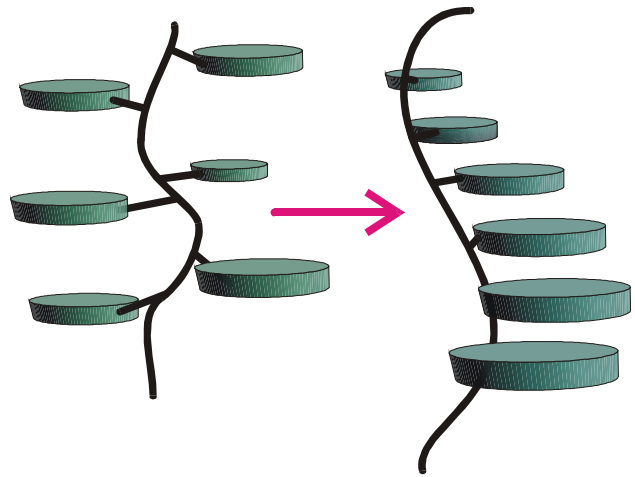
Kinetic Structures

Different notions of RNA structure including suboptimal conformations and folding kinetics

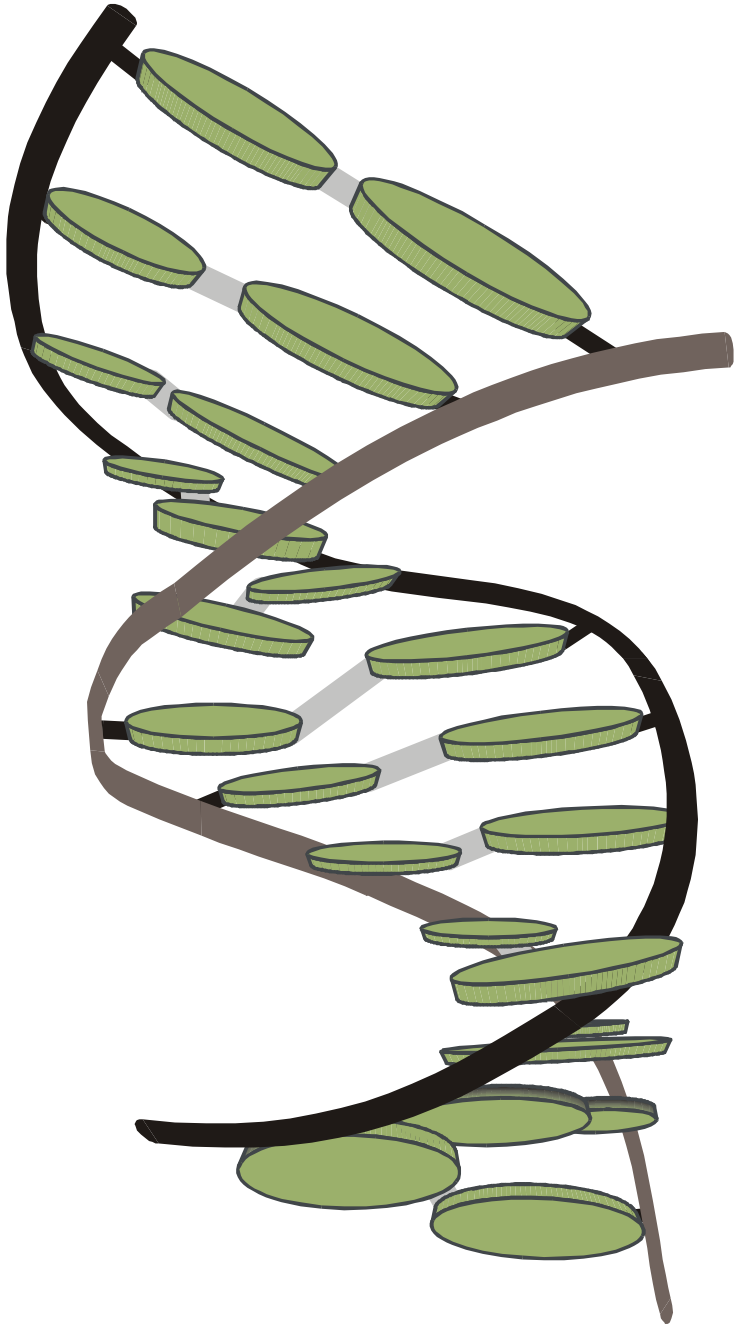
The stacking interaction as driving force of structure formation in nucleic acids

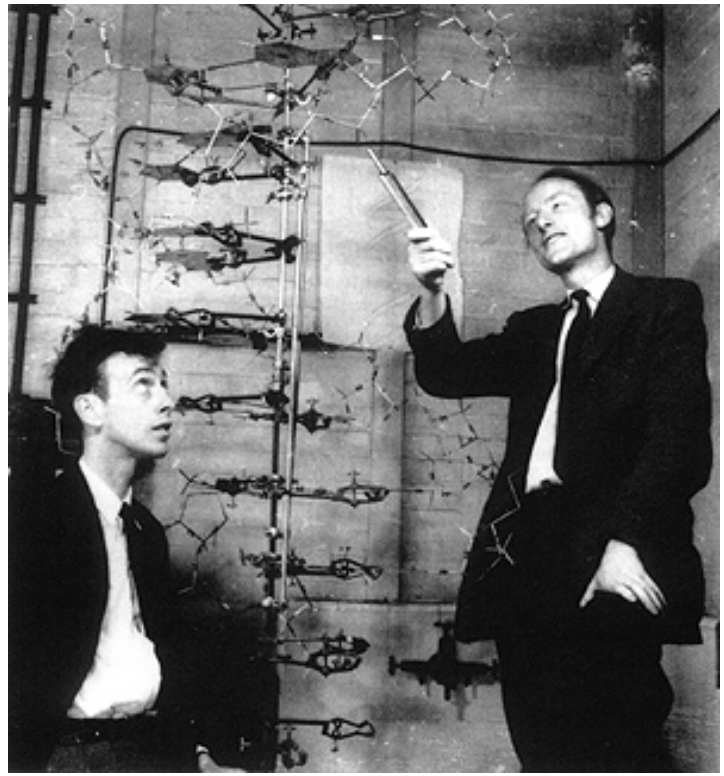


Stacking of free nucleobases or other planar heterocyclic compounds (N6,N9-dimethyl-adenine)



Stacking of nucleic acid single strands (poly-A)





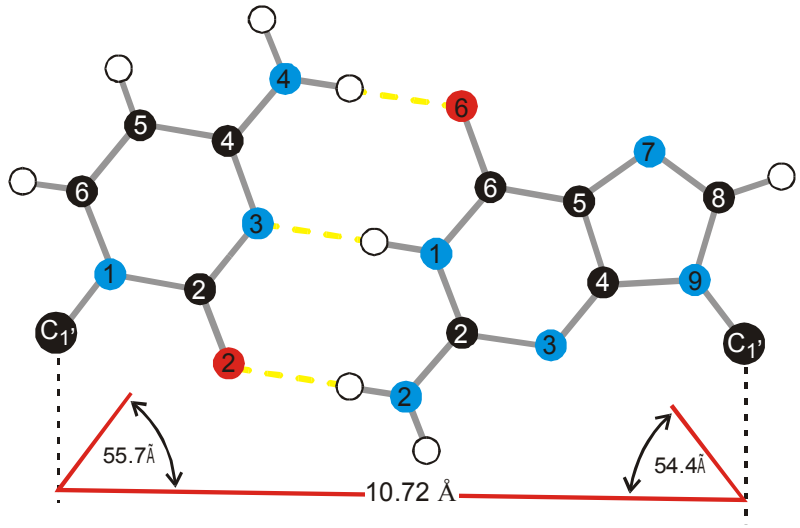
James D. Watson and Francis H.C. Crick

Nobel prize 1962

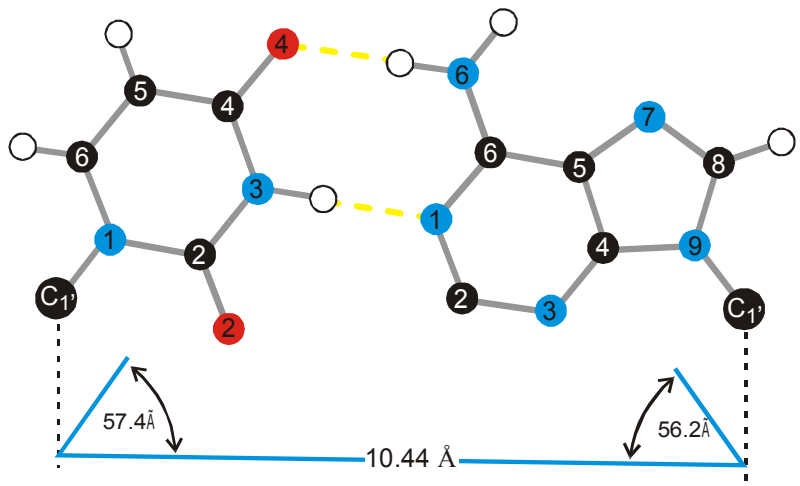
1953 – 2003 fifty years double helix

Stacking of base pairs in nucleic acid double helices (B-DNA)

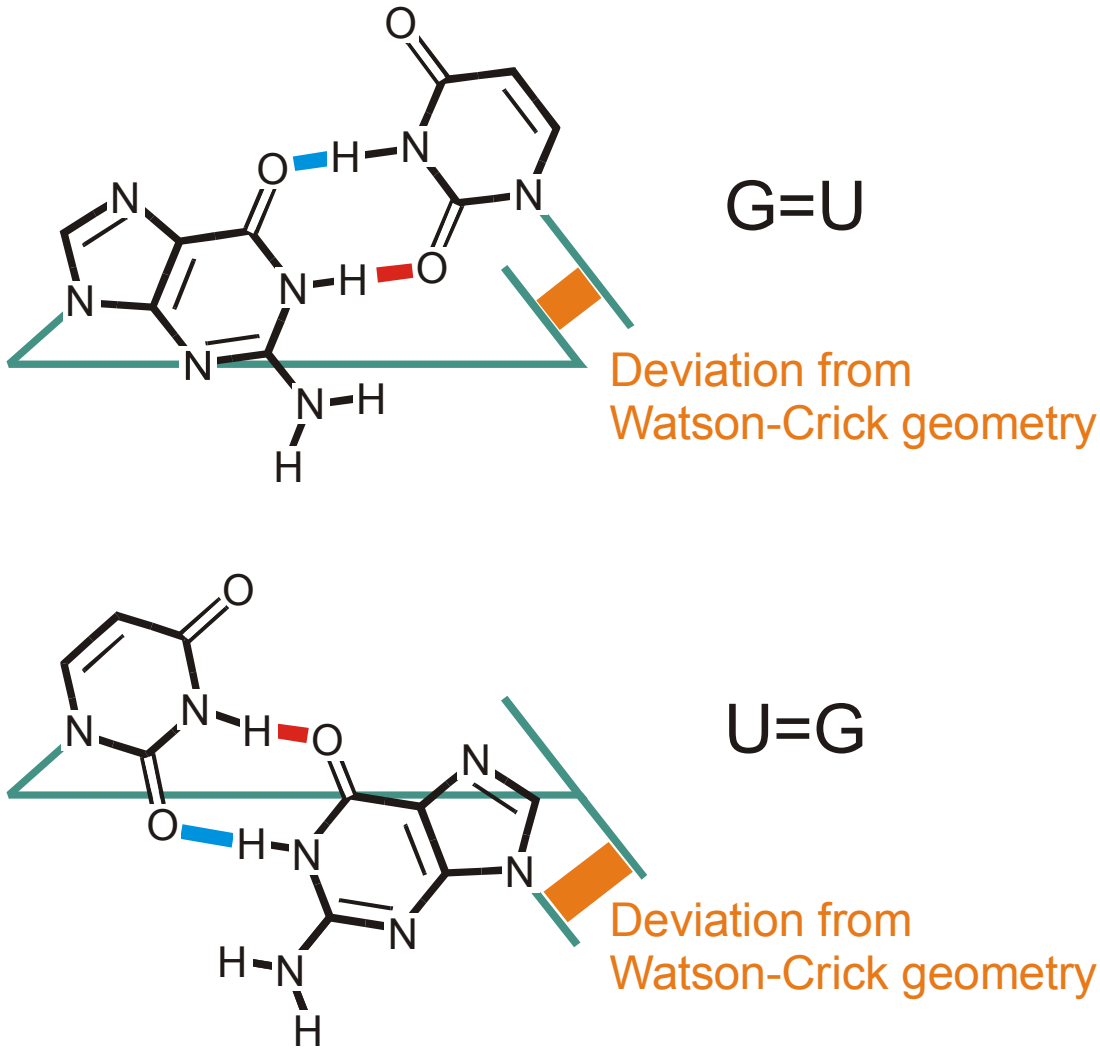
C Í G



U = A



Watson-Crick type base pairs



Wobble base pairs

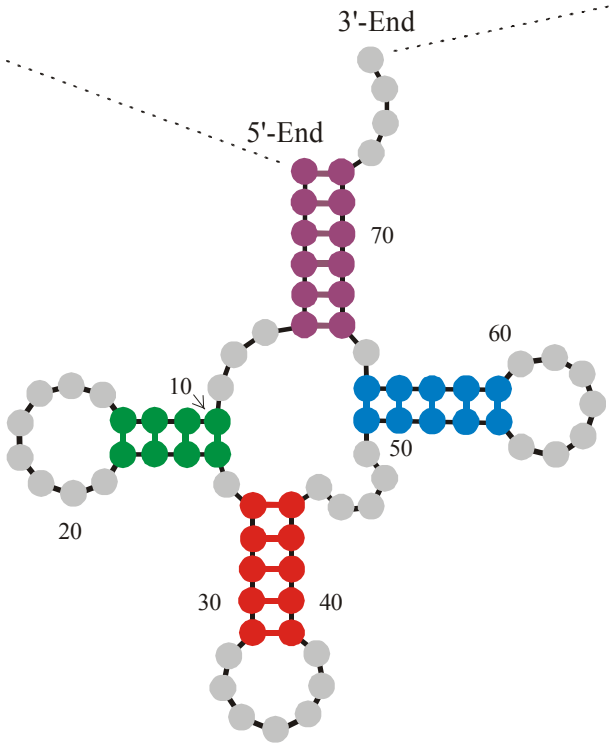
Sequence

5'-End

3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure

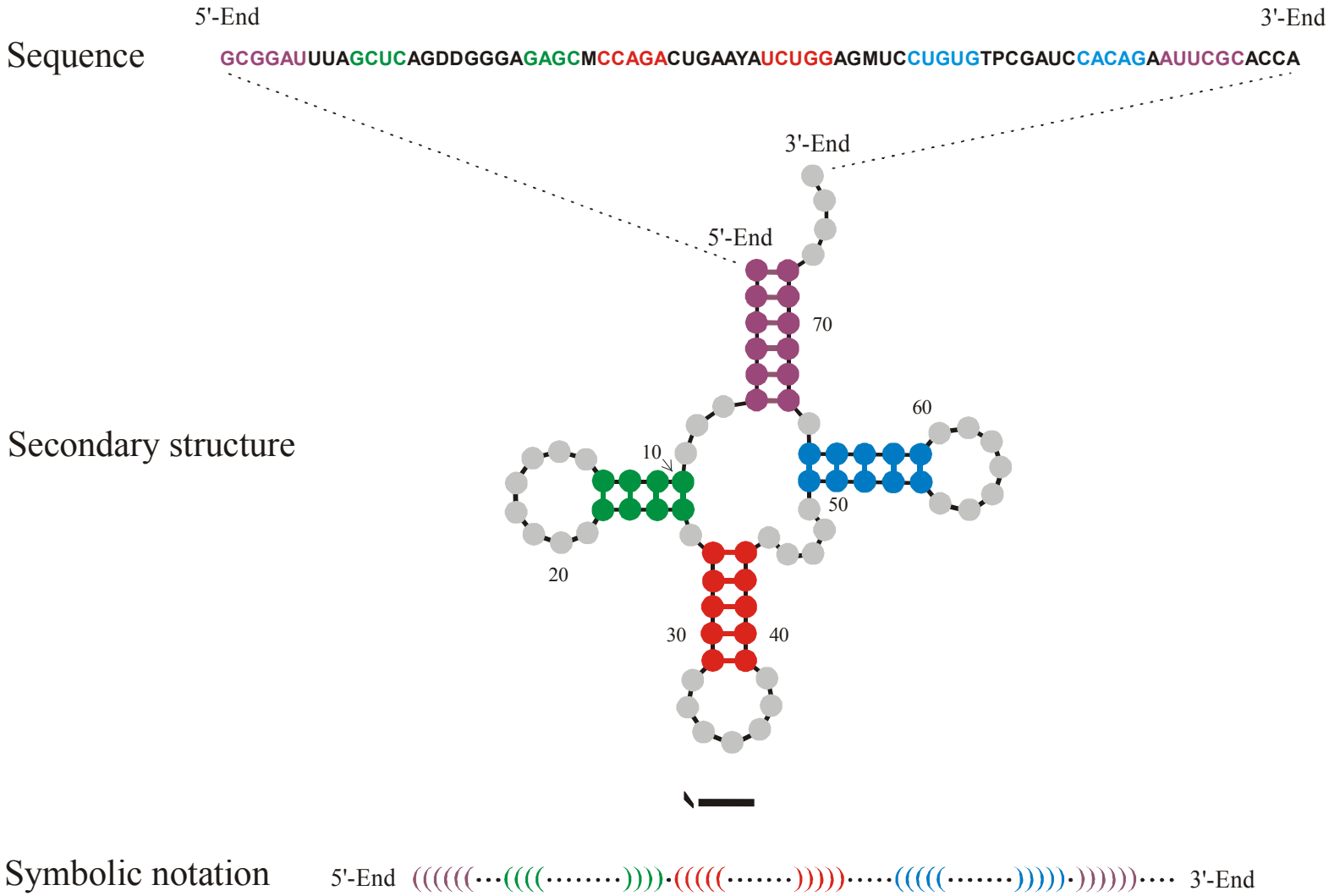


Definition and **physical relevance** of RNA secondary structures

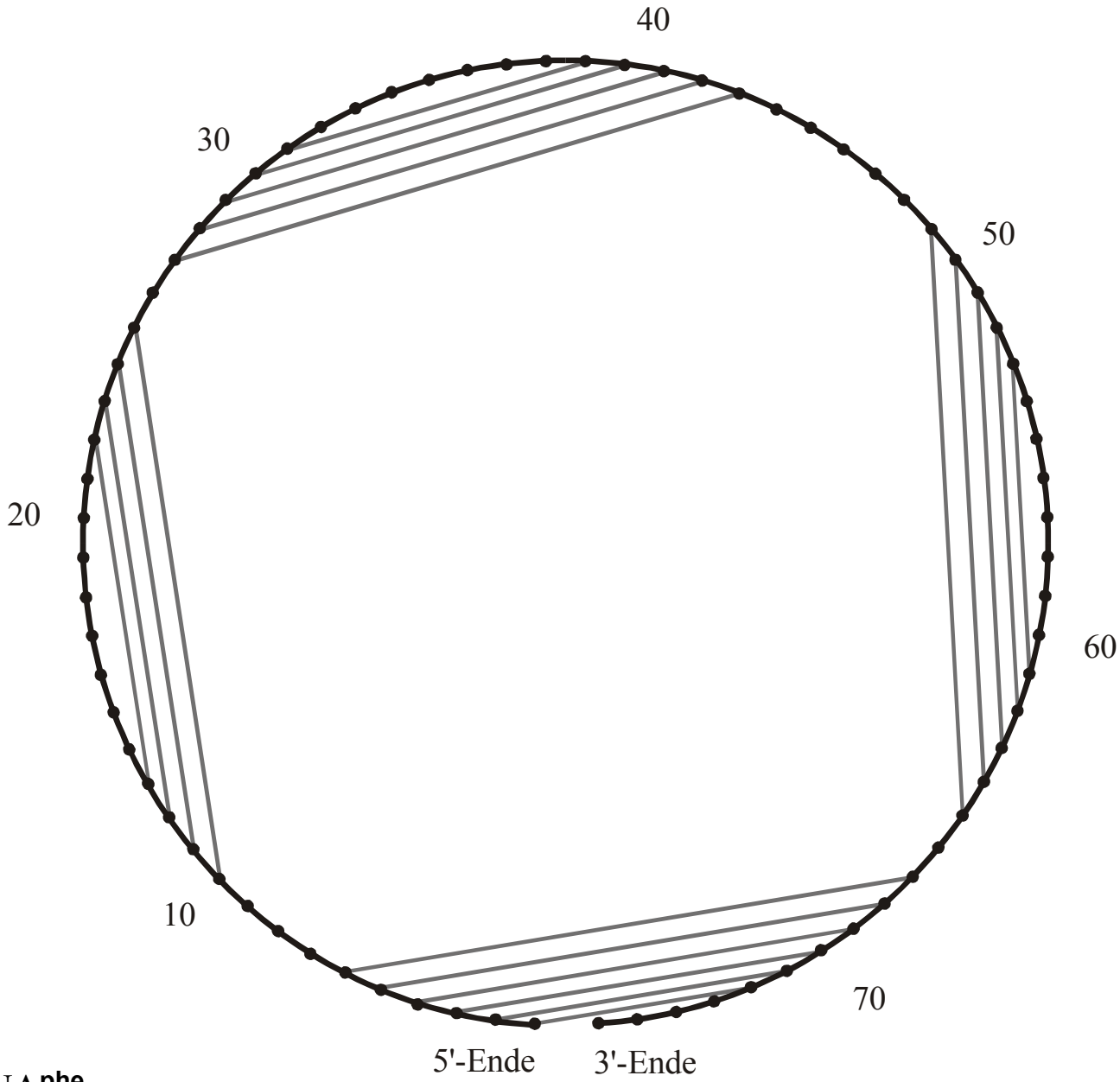
RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots.

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

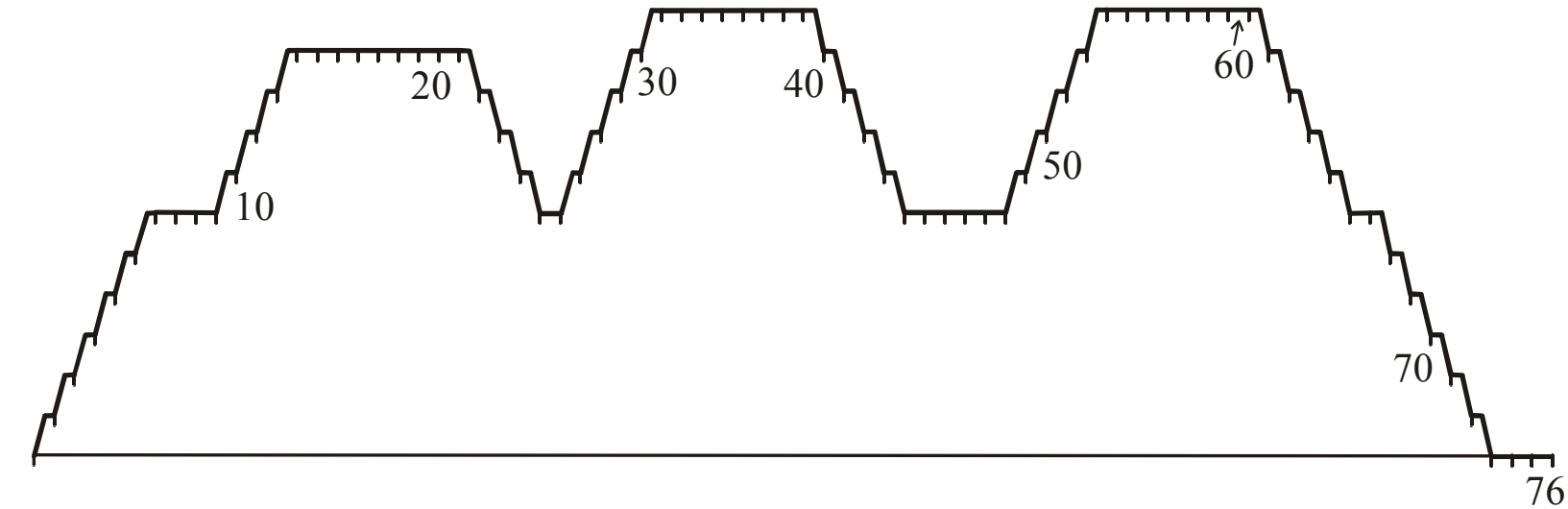
„**Secondary structures are folding intermediates in the formation of full three-dimensional structures**.“



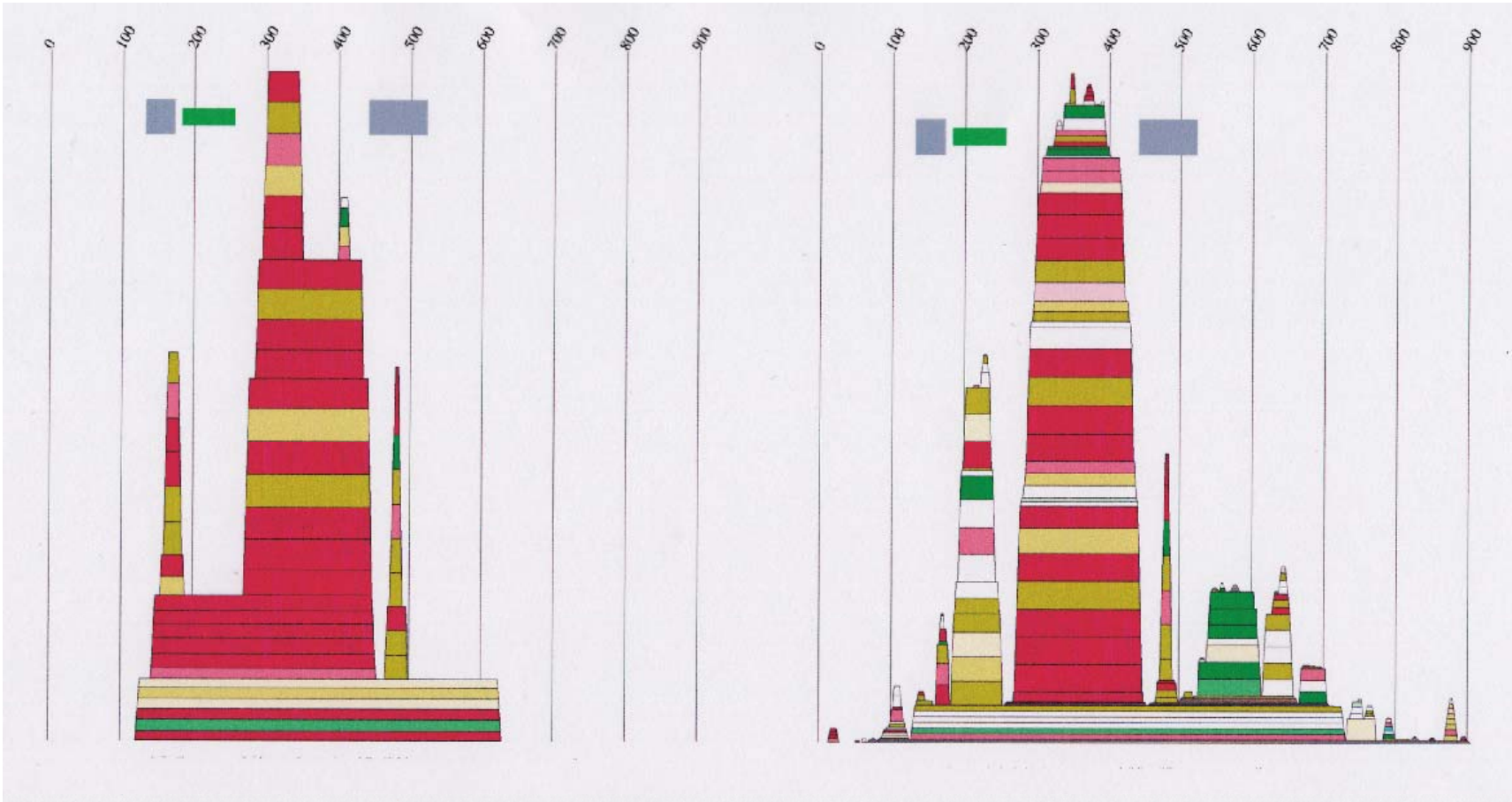
A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs



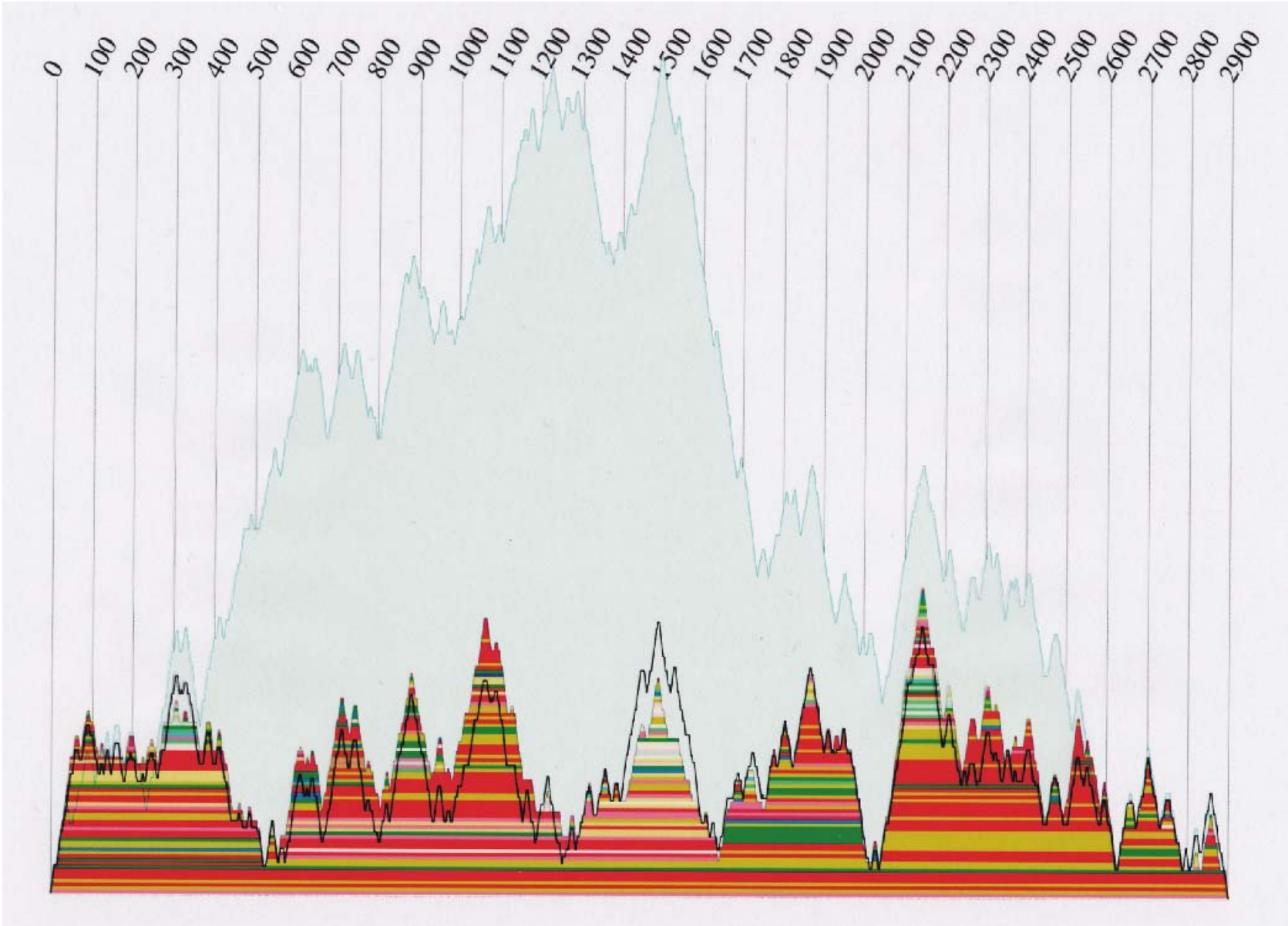
Circle representation of tRNA^{phe}



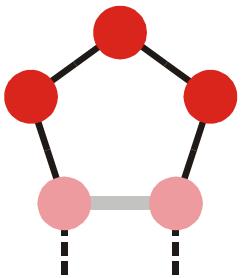
Mountain representation of tRNA^{phe}



Mountain representation used in structure prediction of medium size RNA molecules

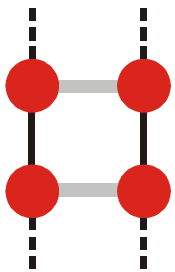


Mountain representation used in structure prediction of large RNA molecules



Minimal hairpin loop size:

$n_{lp} \geq 3$



Minimal stack length:

$n_{st} \geq 2$

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

Recursion formula:

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

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

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

Recursion: $m+1 \Longrightarrow m$

Initial conditions:

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

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

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

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

Recursion formula for the number of acceptable RNA secondary structures

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

Computed numbers of minimum free energy structures over different nucleotide alphabets

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

RNA sequence **GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA**

RNA folding:

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

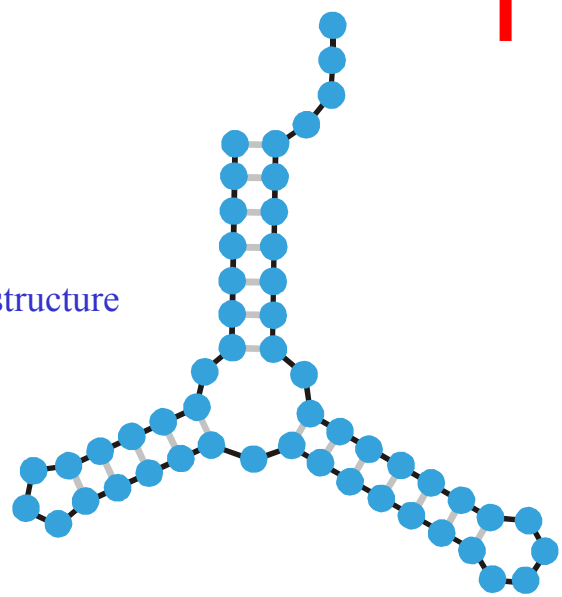
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

Inverse folding of RNA:

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

RNA structure



Sequence, structure, and function

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)

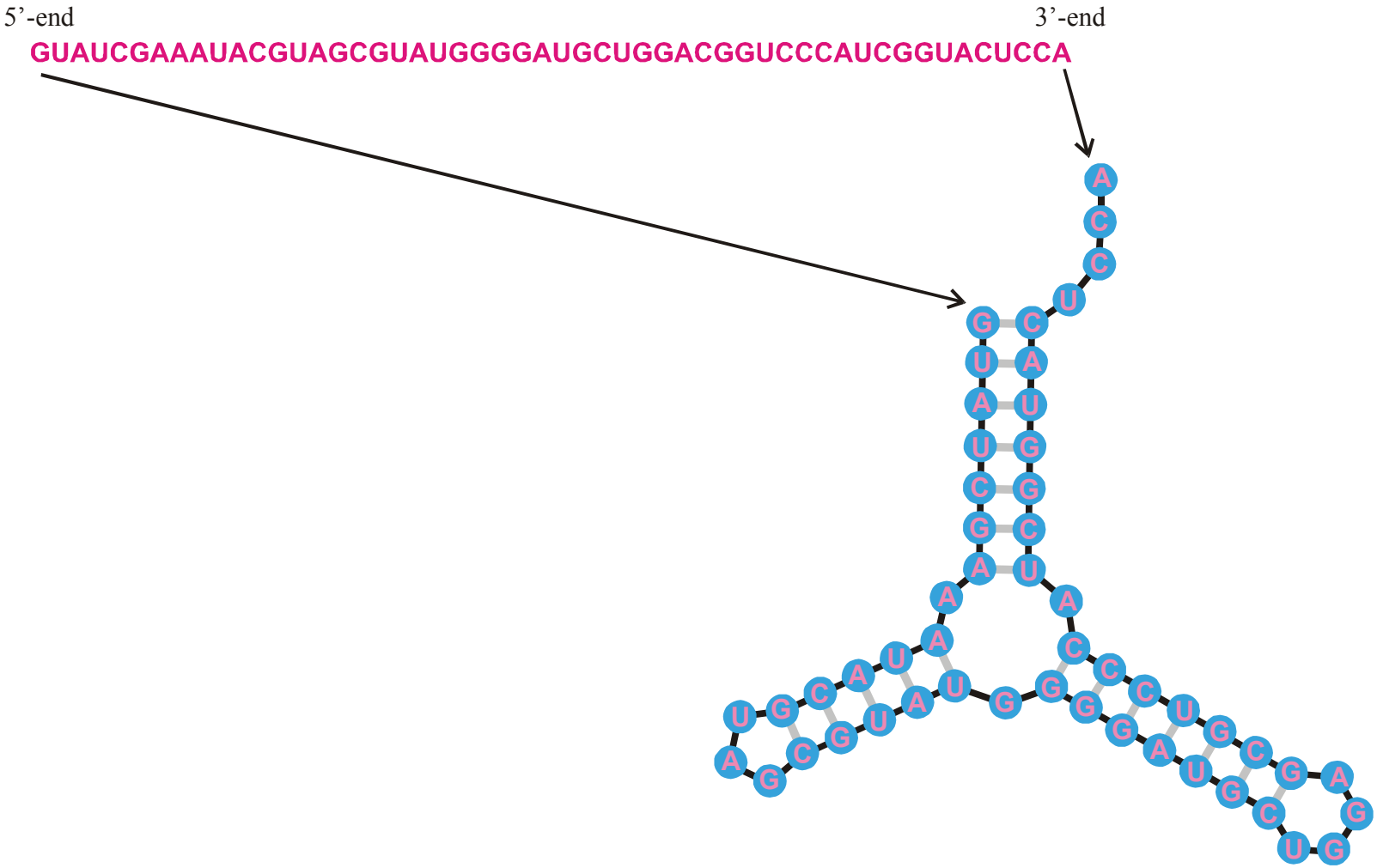
The **Vienna RNA Package** provides in addition: **inverse folding** (computing sequences for given secondary structures), computation of melting profiles from partition functions, **all** suboptimal structures within a given energy interval, barrier tress of suboptimal structures, **kinetic folding** 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)

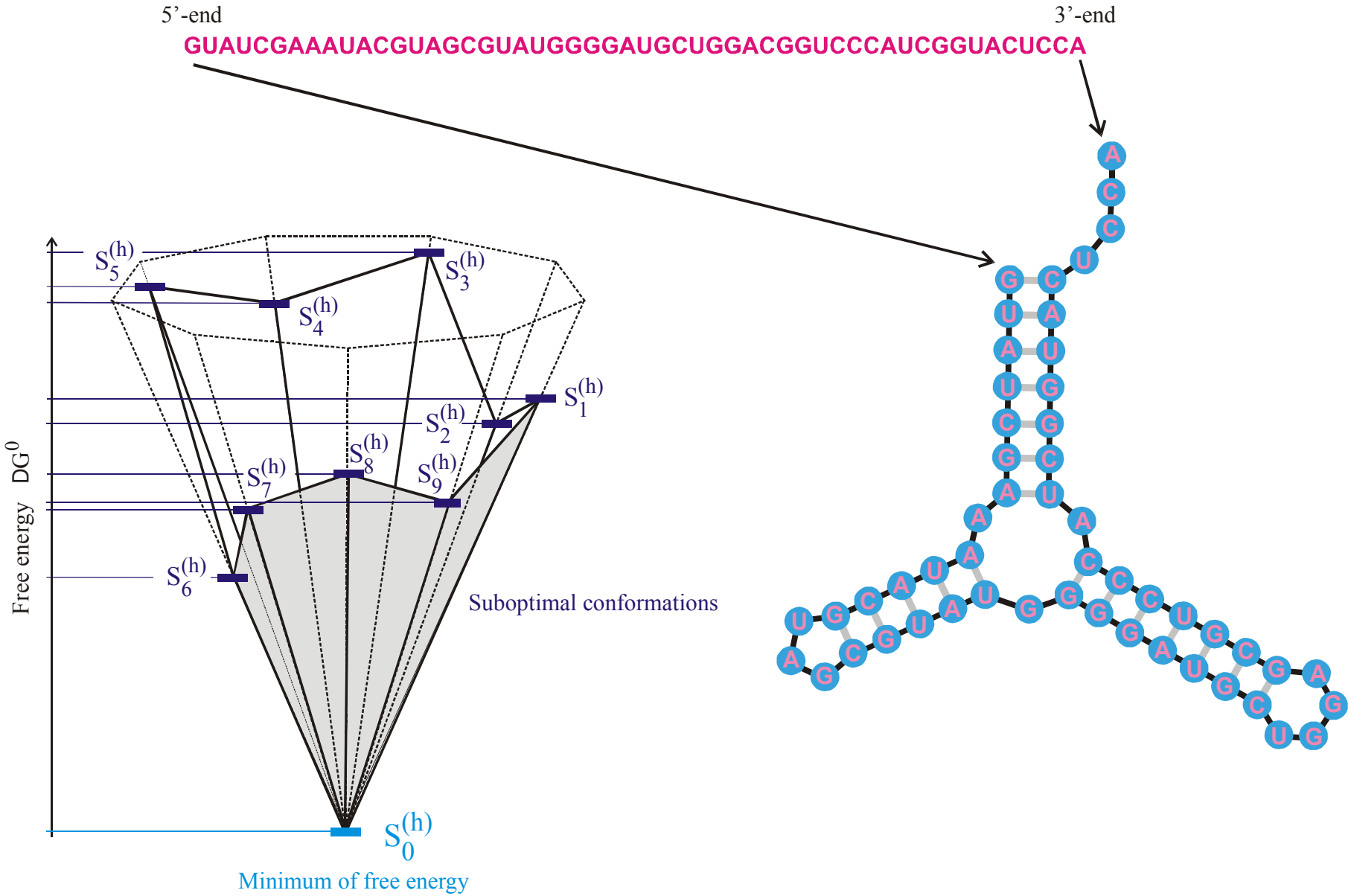
S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. ***Biopolymers*** **49**:145-165 (1999)

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

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



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



The minimum free energy structures on a discrete space of conformations

5’-end 3’-end

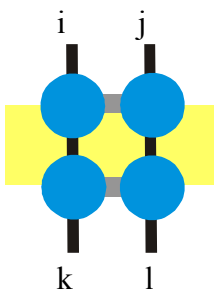
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

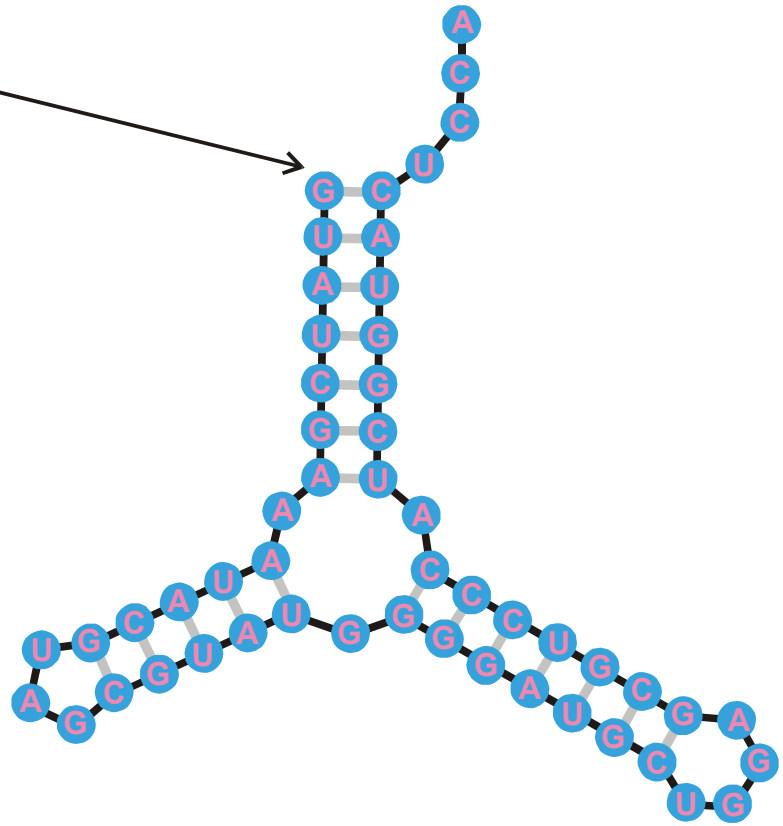
Edges: i·j,k·l á Ś **.... base pairs**

(i) i· i+1 á Ś .... **backbone**

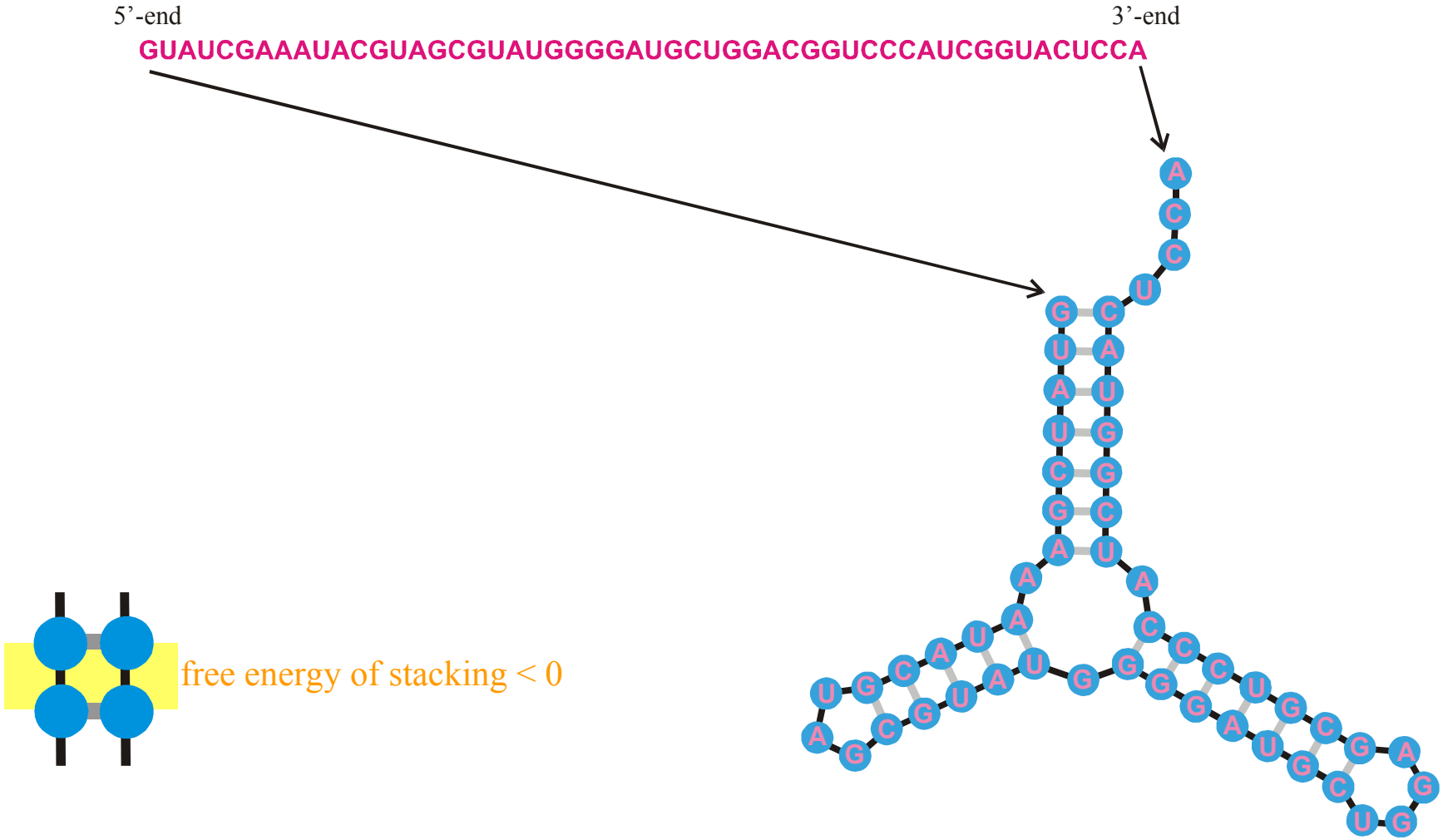
(ii) #base pairs per node = {0,1}

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



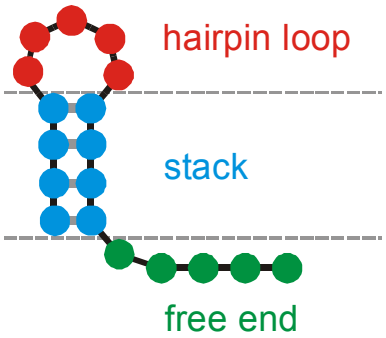


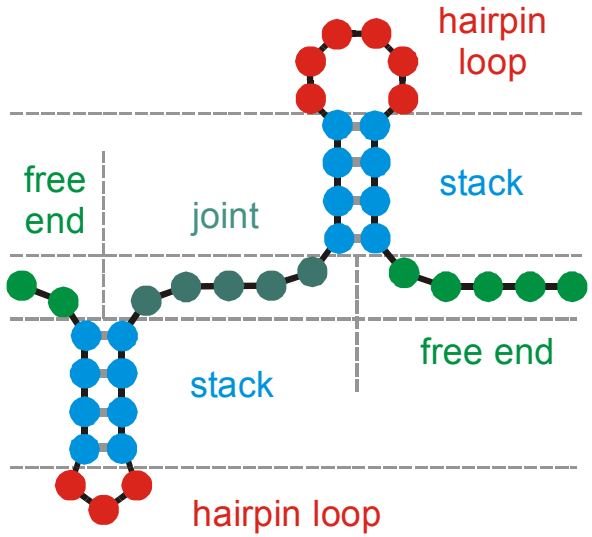
Folding of RNA sequences into secondary structures of minimal free energy, DG_0^{300}

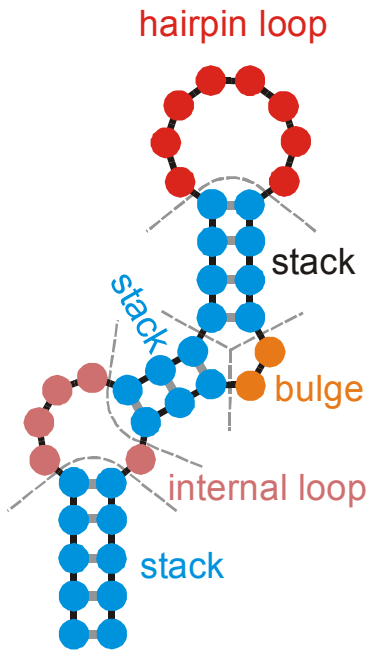


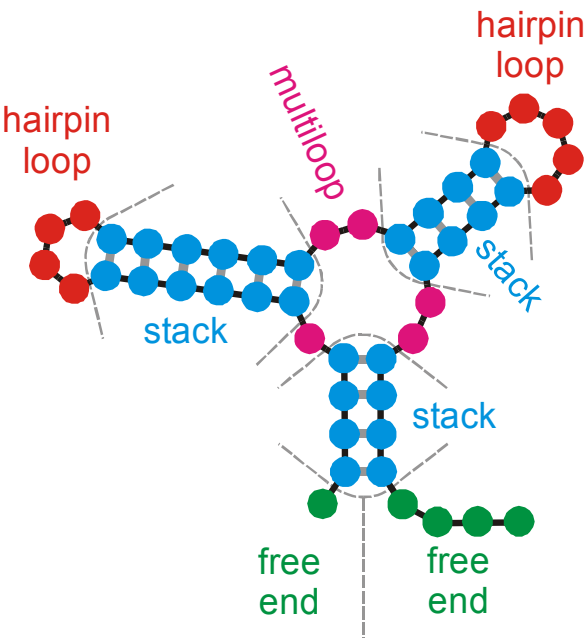
$$\Delta G_0^{300} = \underbrace{\sum g_{ij,kl}}_{\text{stacks of base pairs}} + \underbrace{\sum h(n_l)}_{\text{hairpin loops}} + \underbrace{\sum b(n_b)}_{\text{bulges}} + \underbrace{\sum i(n_i)}_{\text{internal loops}} + \cdots$$

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







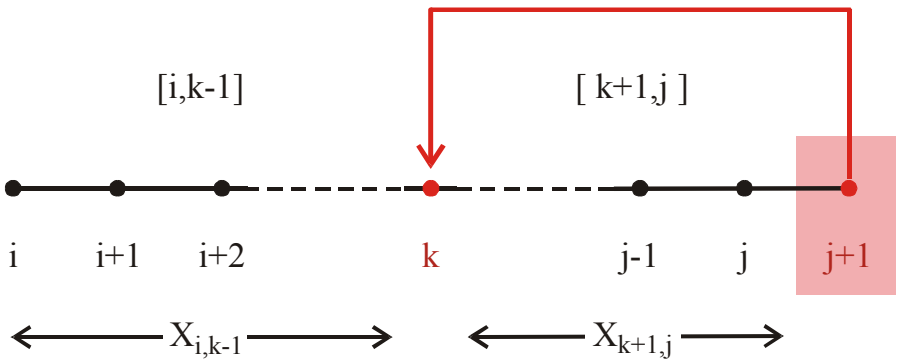


Elements of RNA secondary structures
as used in free energy calculations

Maximum matching

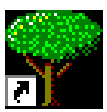
An example of a **dynamic programming** computation of the maximum number of base pairs

Back tracking yields the structure(s).



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

Minimum free energy computations are based on empirical energies



RNAStudio.lnk

GGCGCGCCCGGCGCC

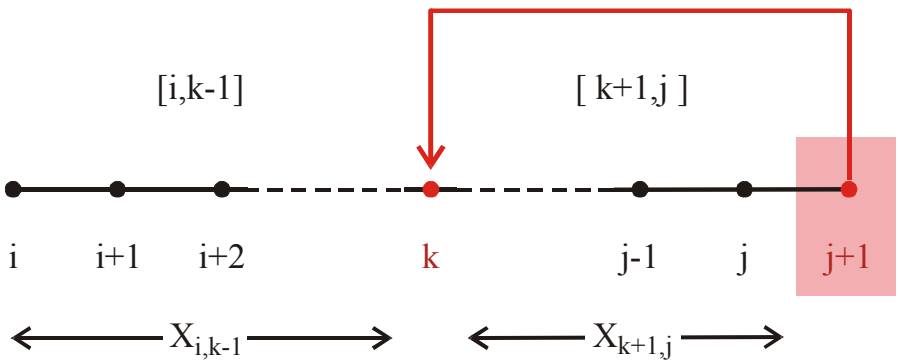
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

Maximum matching

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

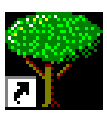
Back tracking yields the structure(s).



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

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

Minimum free energy computations are based on empirical energies



RNAStudio.lnk

GGCGCGCCCGGCGCC

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

RNA sequence **GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA**

RNA folding:

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

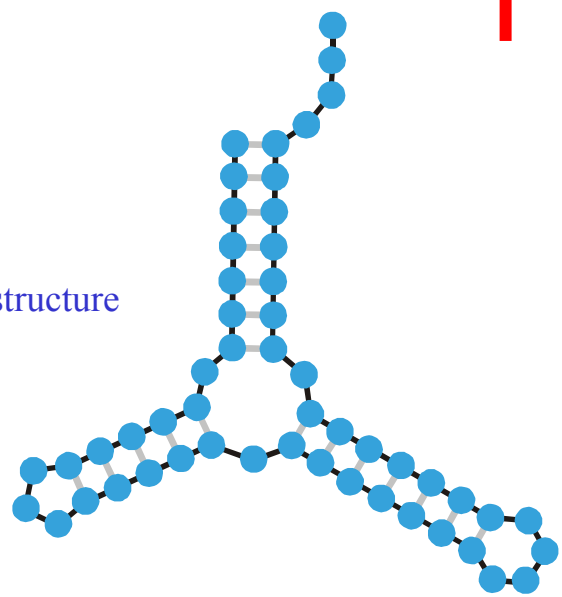
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

Inverse folding of RNA:

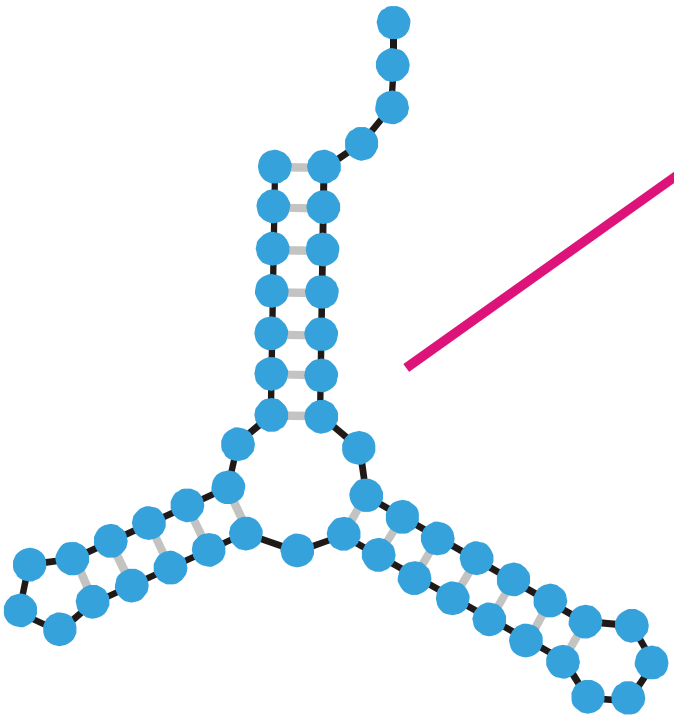
Biotechnology, **design of biomolecules** with predefined structures and functions

RNA structure



Sequence, structure, and function

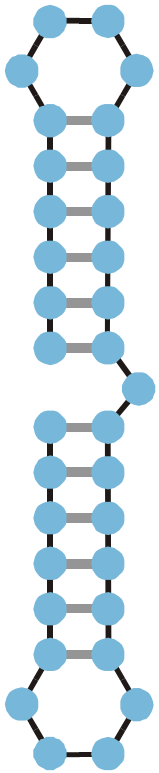
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



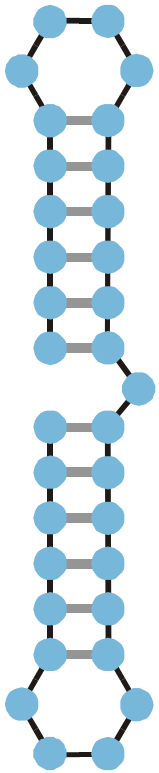
Minimum free energy criterion

Inverse folding of RNA secondary structures

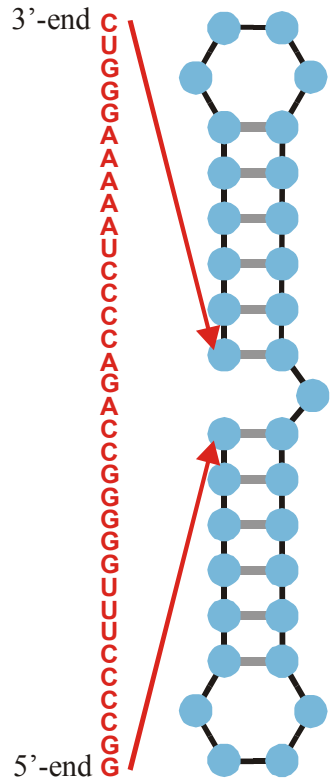
The idea of inverse folding algorithm is to search for sequences that form a given RNA secondary structure under the minimum free energy criterion.



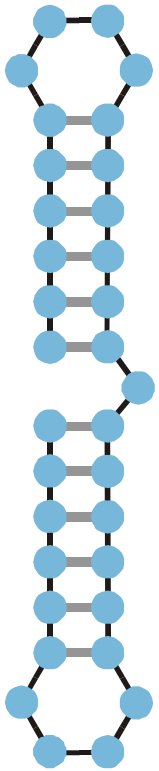
Structure



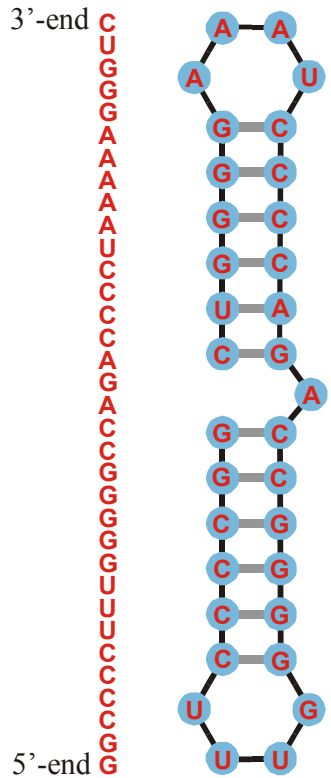
Structure



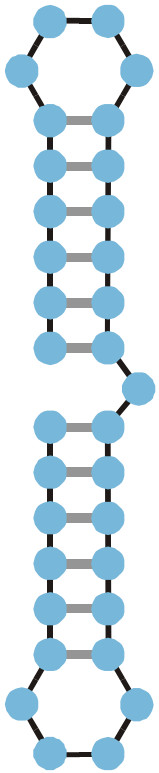
Compatible sequence



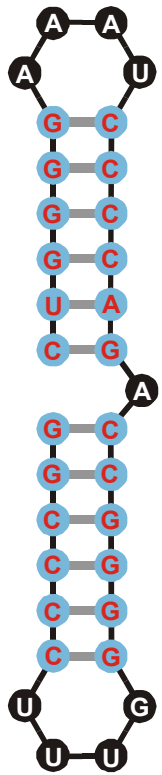
Structure



Compatible sequence



3’-end C U G G G A A A A A A U C C C C C A G A C C C G G G G G U U U C C C C C G G 5’-end

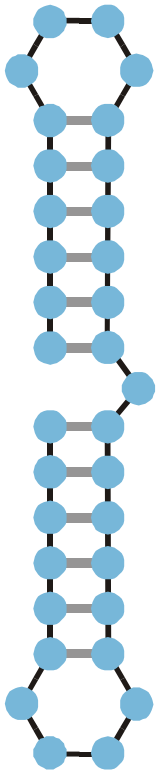


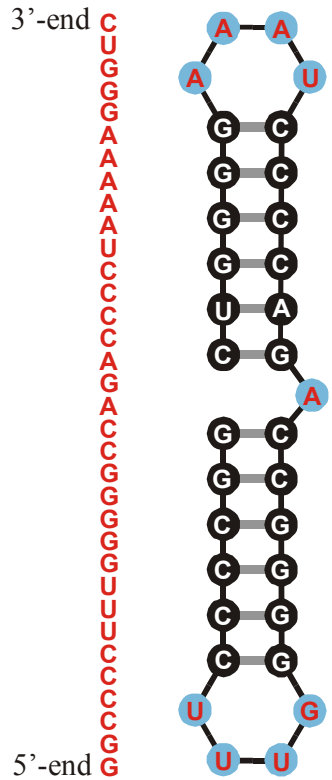
Single nucleotides: **A,U,G,C**

Structure

Compatible sequence

Single bases pairs are varied independently





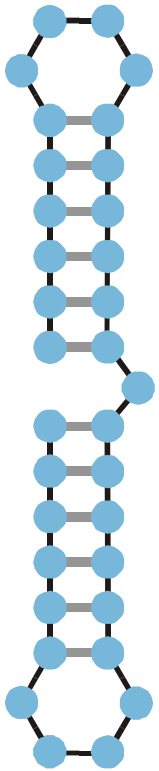
Base pairs:

AU , UA
GC , CG
GU , UG

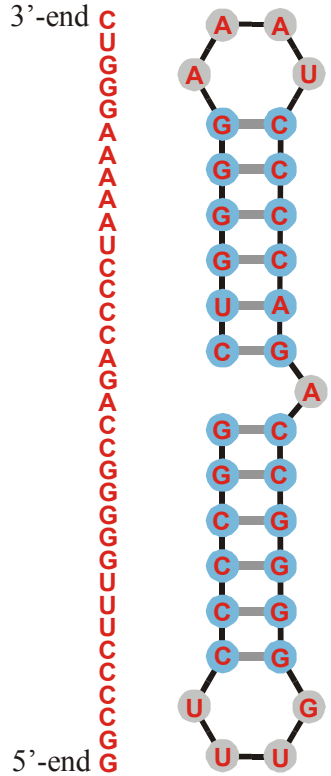
Structure

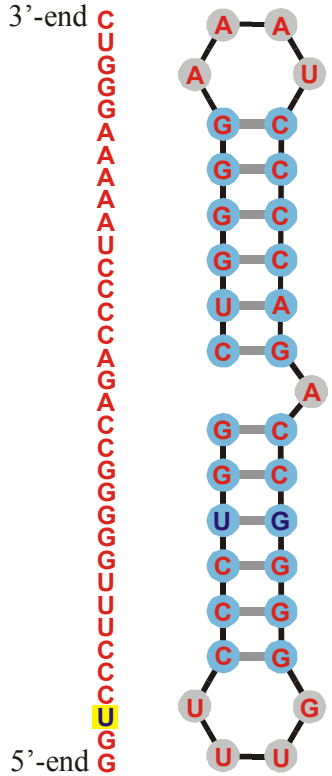
Compatible sequence

Base pairs are varied in strict correlation

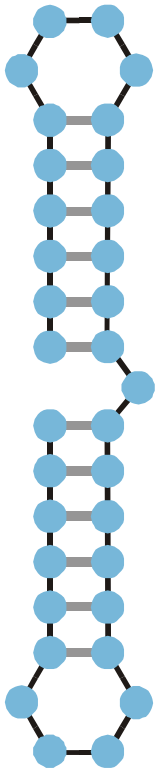


Structure

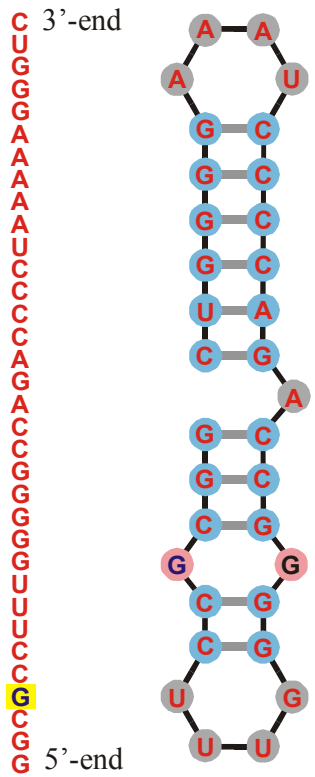




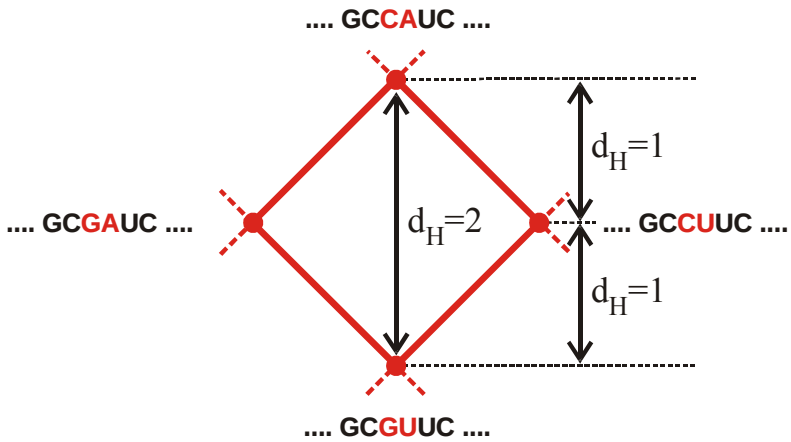
Compatible sequences



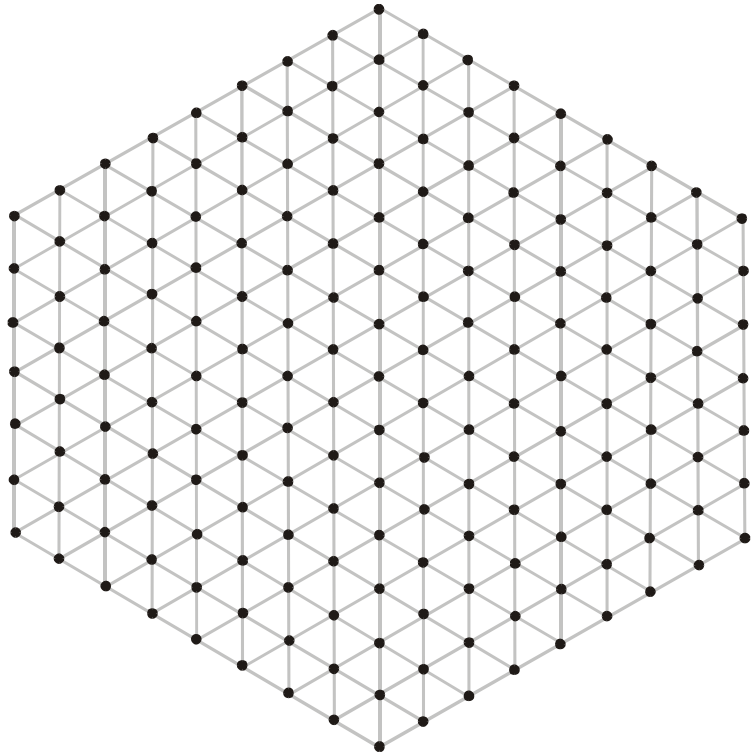
Structure



Incompatible sequence

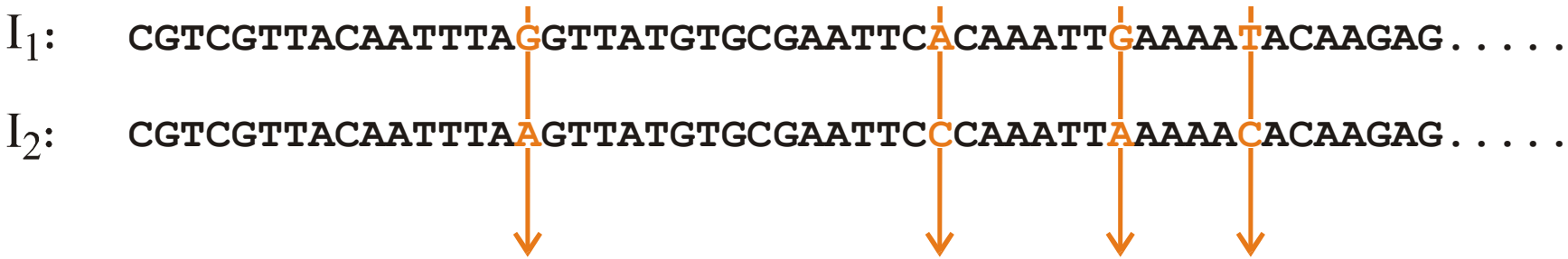


City-block distance in sequence space



2D Sketch of sequence space

Single point mutations as moves in sequence space



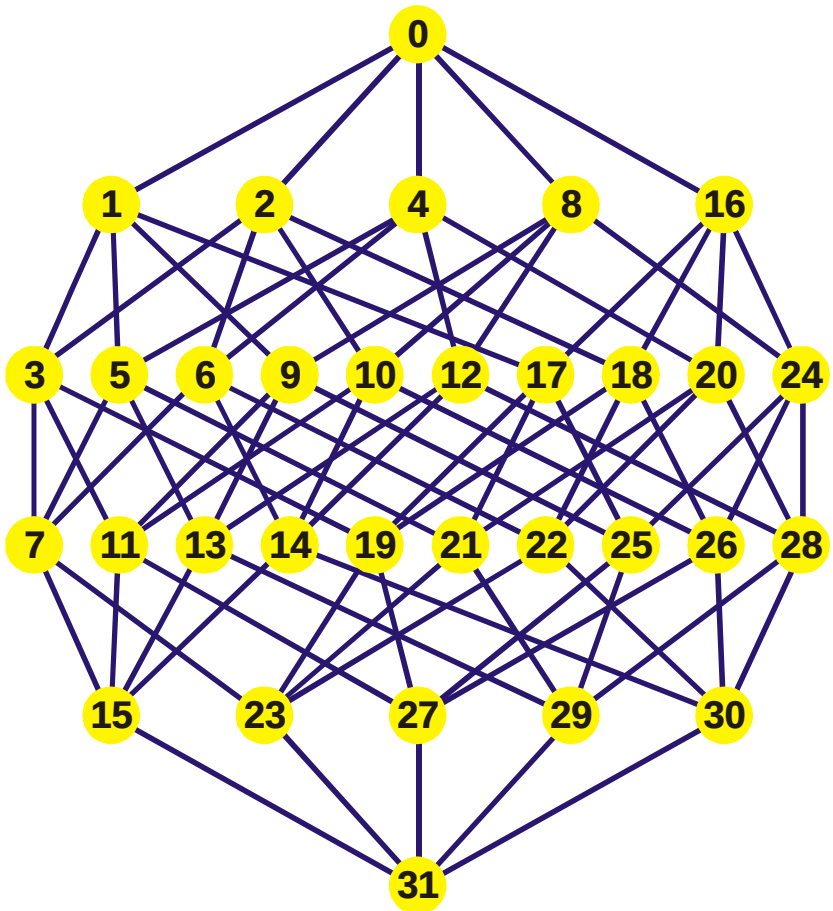
Hamming distance $d_H(I_1,I_2) = 4$

(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



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,

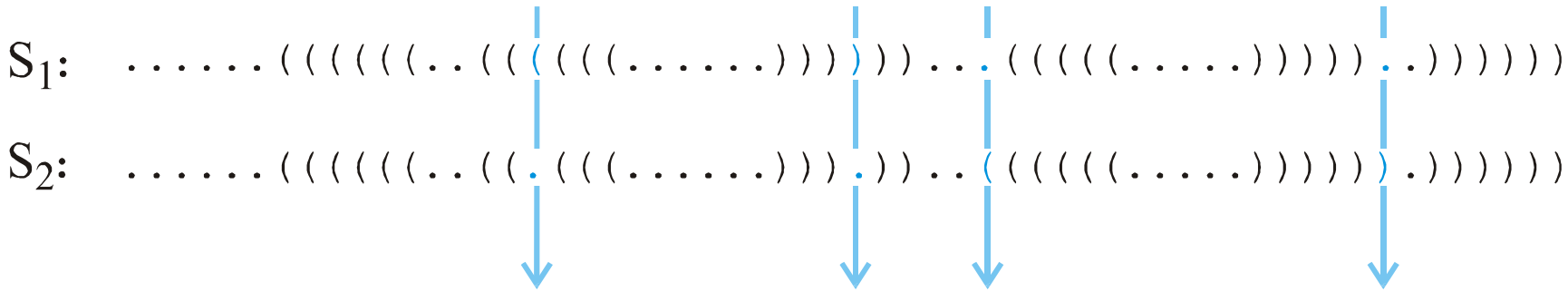
"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

"29" ≡ 11101 = GGGCG, etc.

Decimal coding of binary sequences

Sequence space of binary sequences of chain lenght n = 5



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) $d_H(S_1,S_3) \P d_H(S_1,S_2) + d_H(S_2,S_3)$

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

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 $\mathbf{I_t}$

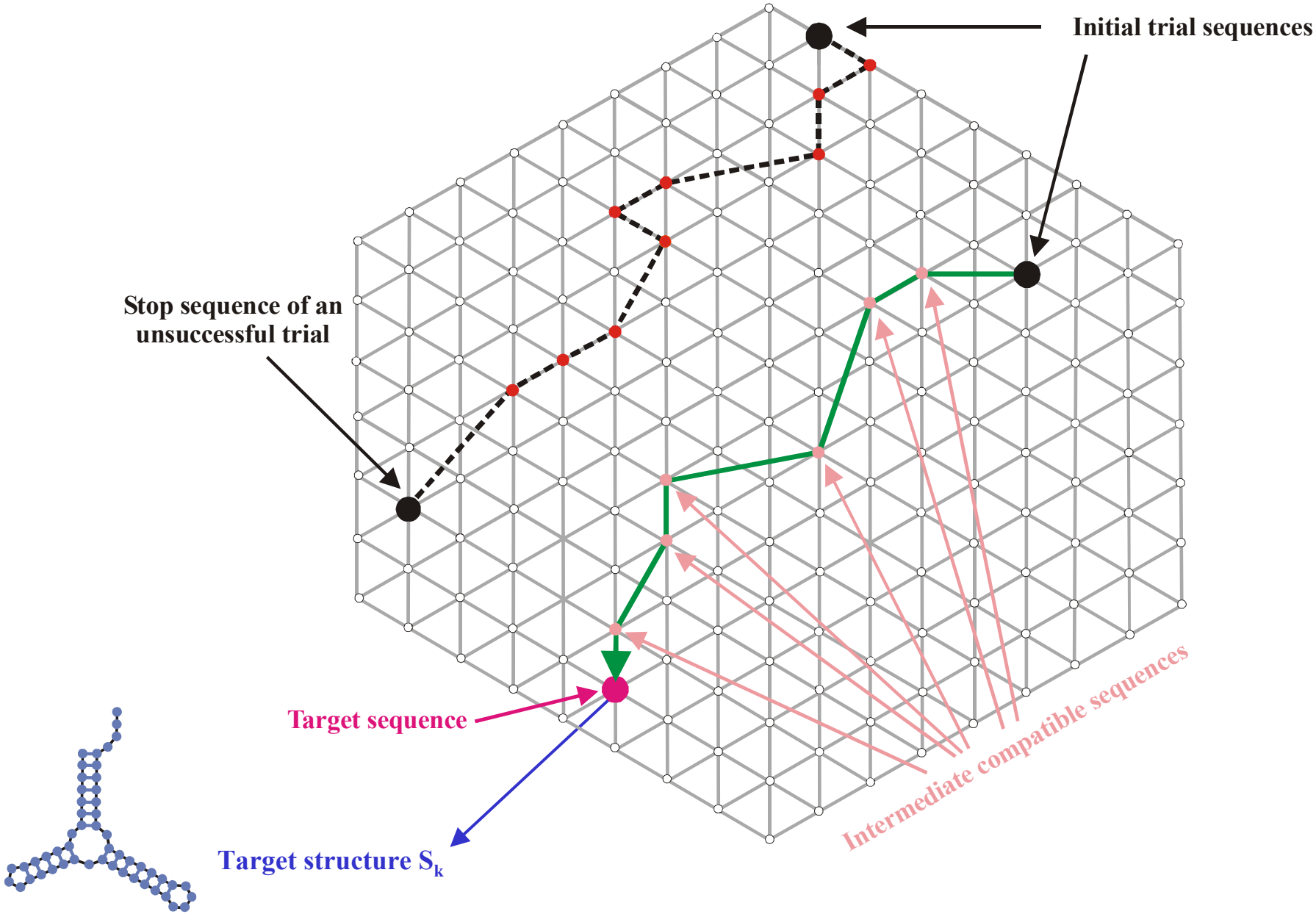
$\mathbf{S_0}$ Á S_1 Á S_2 Á S_3 Á S_4 Á ... Á S_k Á S_{k+1} Á ... Á $\mathbf{S_t}$

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

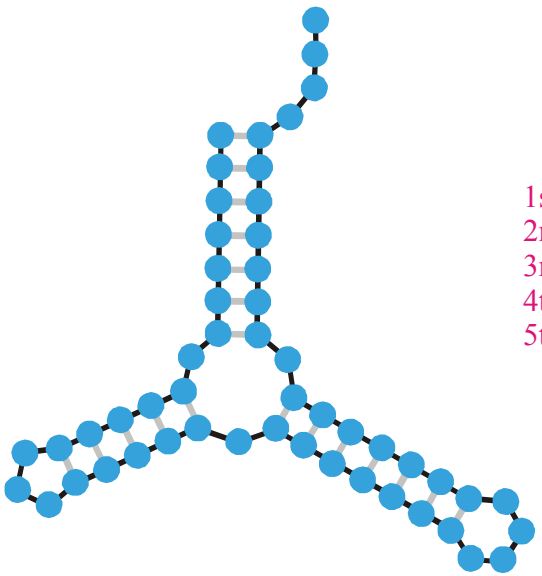
$\mathfrak{M}$ ... base or base pair mutation operator

$d_S(S_i,S_j)$ ... distance between the two structures S_i and S_j

‚Unsuccessful trial' ... termination after n steps



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



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.

Criterion of
Minimum Free Energy

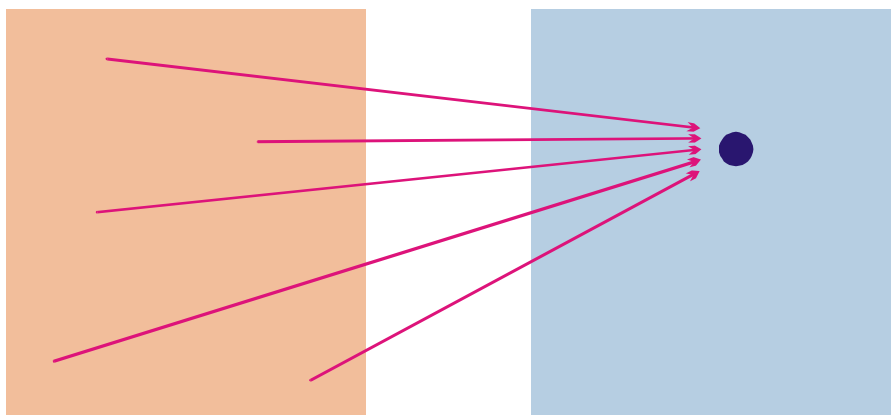
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG



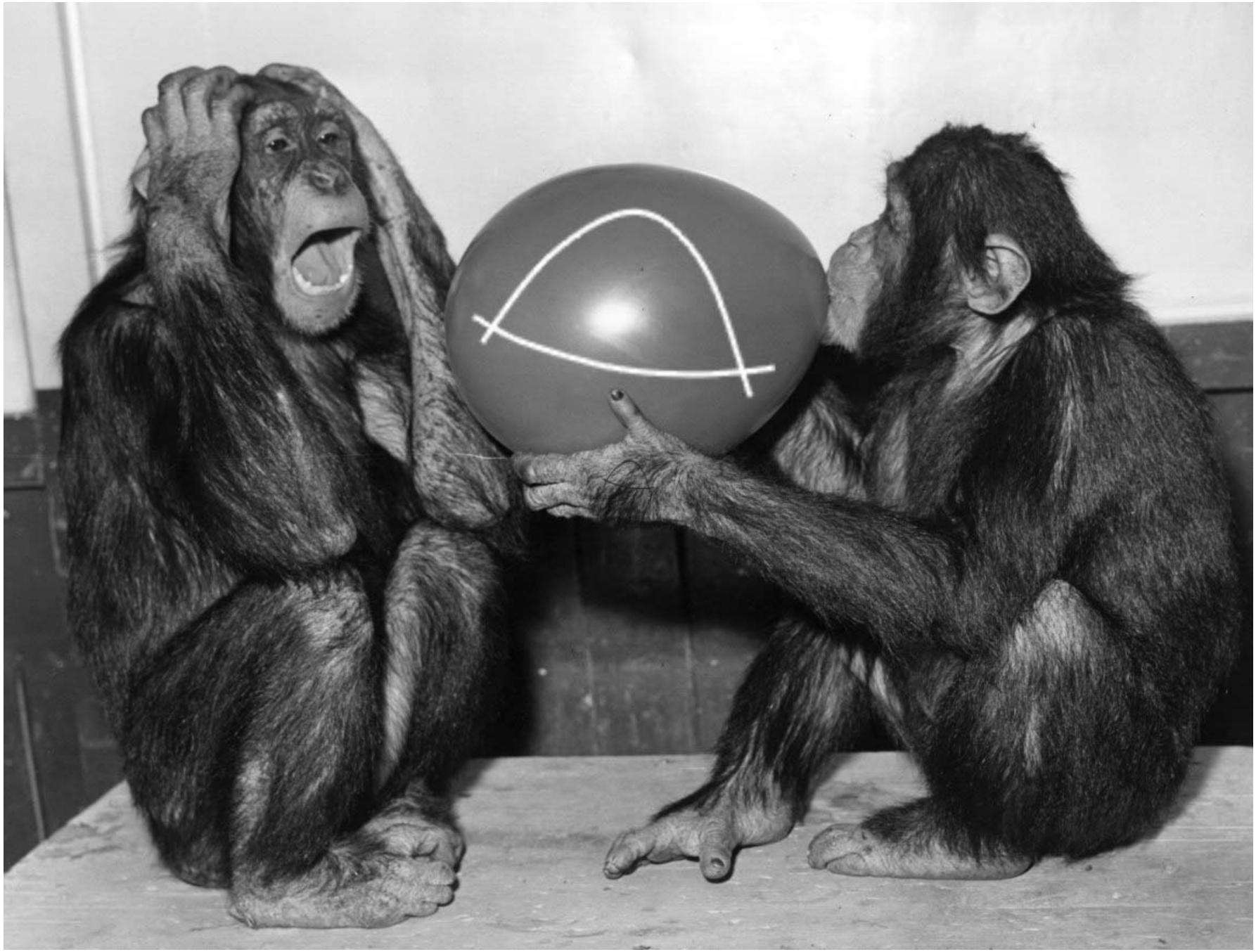
Sequence Space

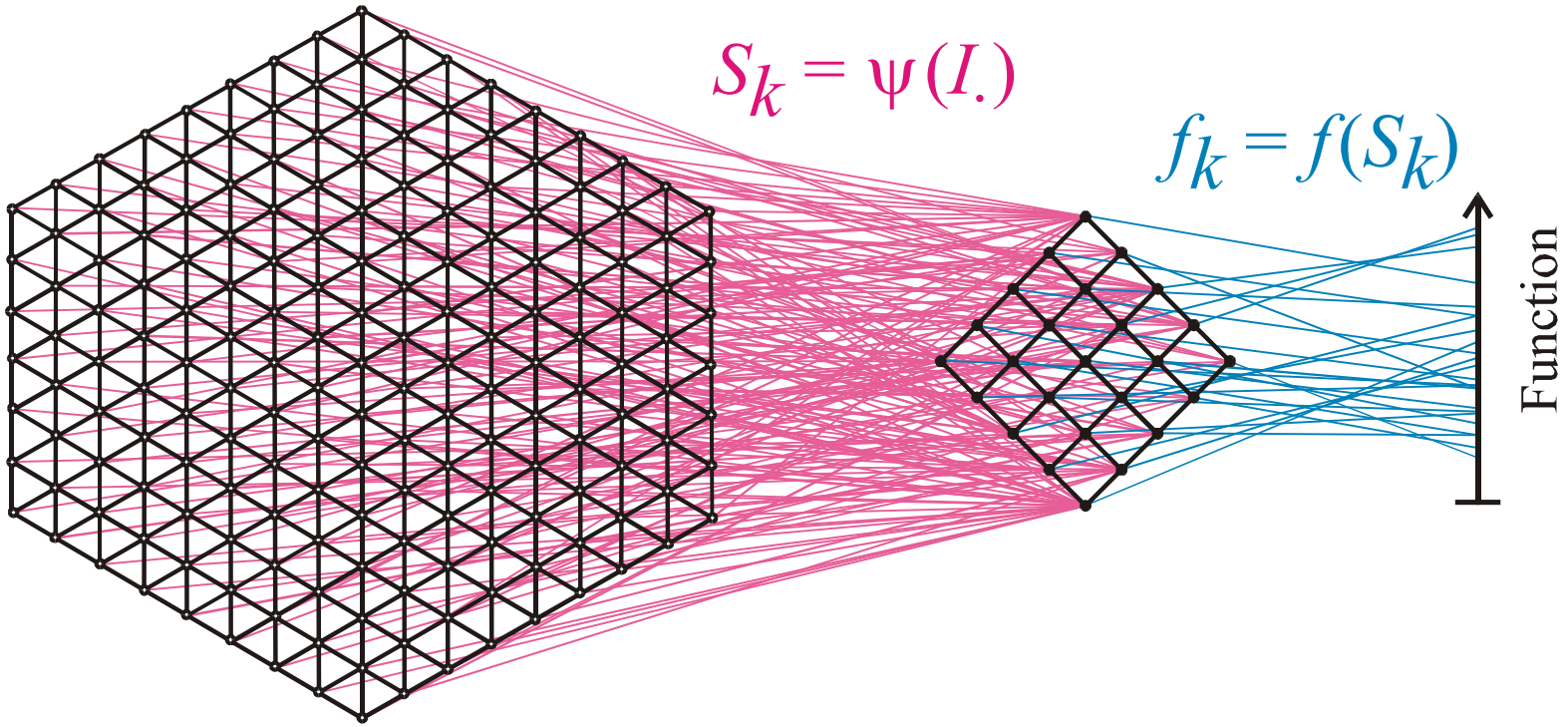
Shape Space

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.

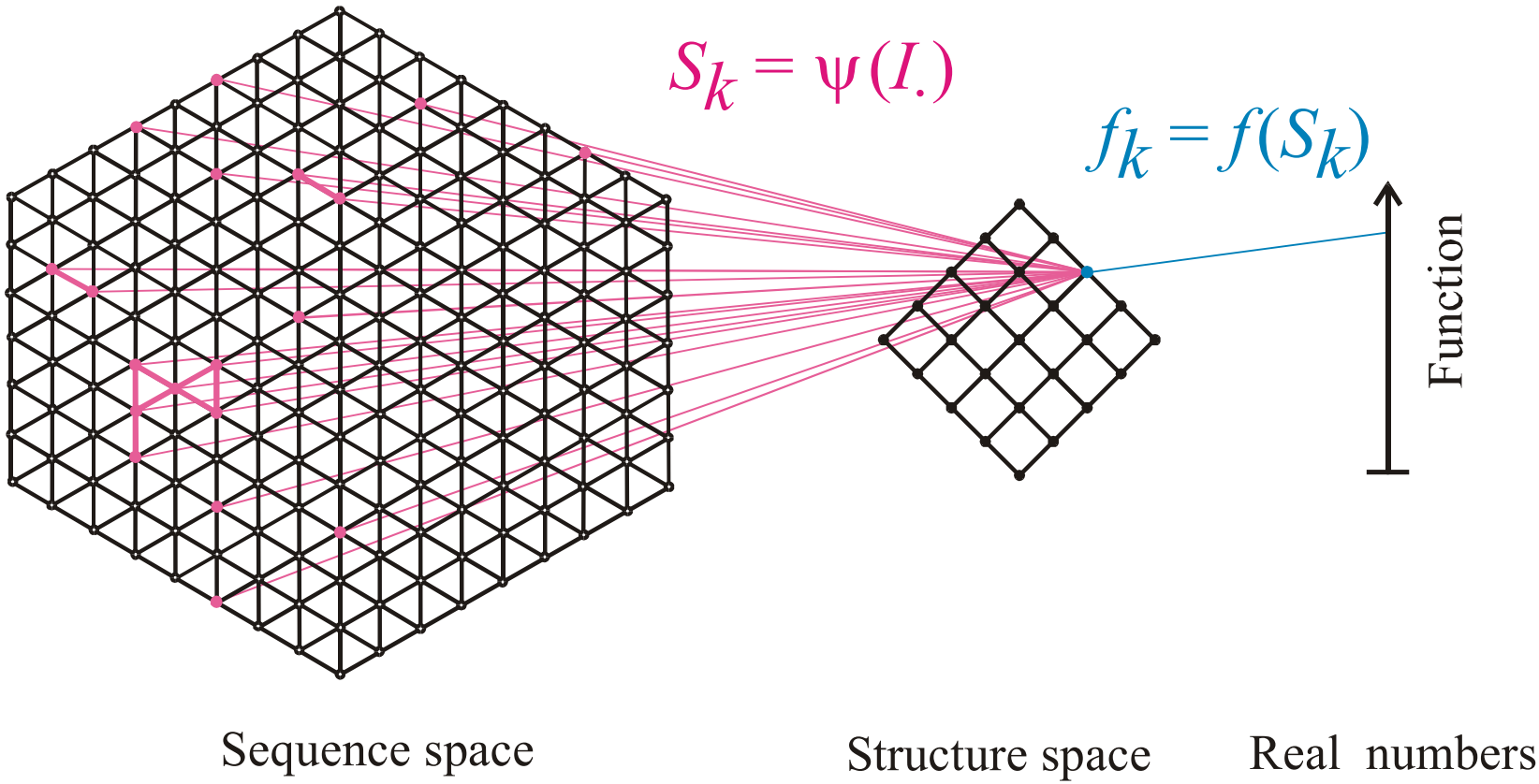
The **mapping** from RNA **sequences** into RNA secondary **structures** is many-to-one. Hence, it is redundant and not invertible.

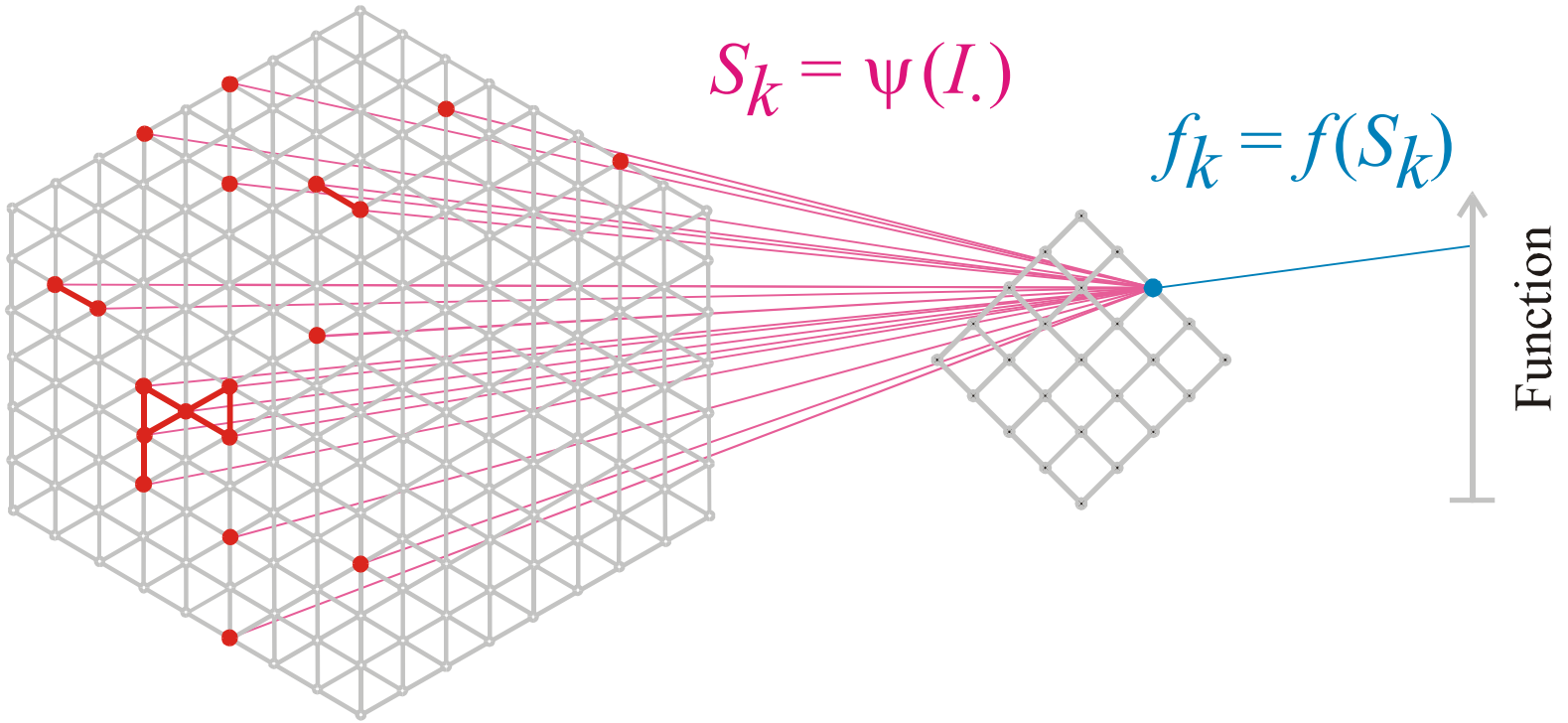
RNA **sequences**, which are mapped onto the same RNA secondary **structure**, are **neutral** with respect to **structure**. The pre-images of structures in sequence space are **neutral networks**. They can be represented by graphs where the edges connect sequences of Hamming distance $d_H = 1$.





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



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. $\mathbf{G_k}$ is the pre-image of the structure S_k in sequence space:

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

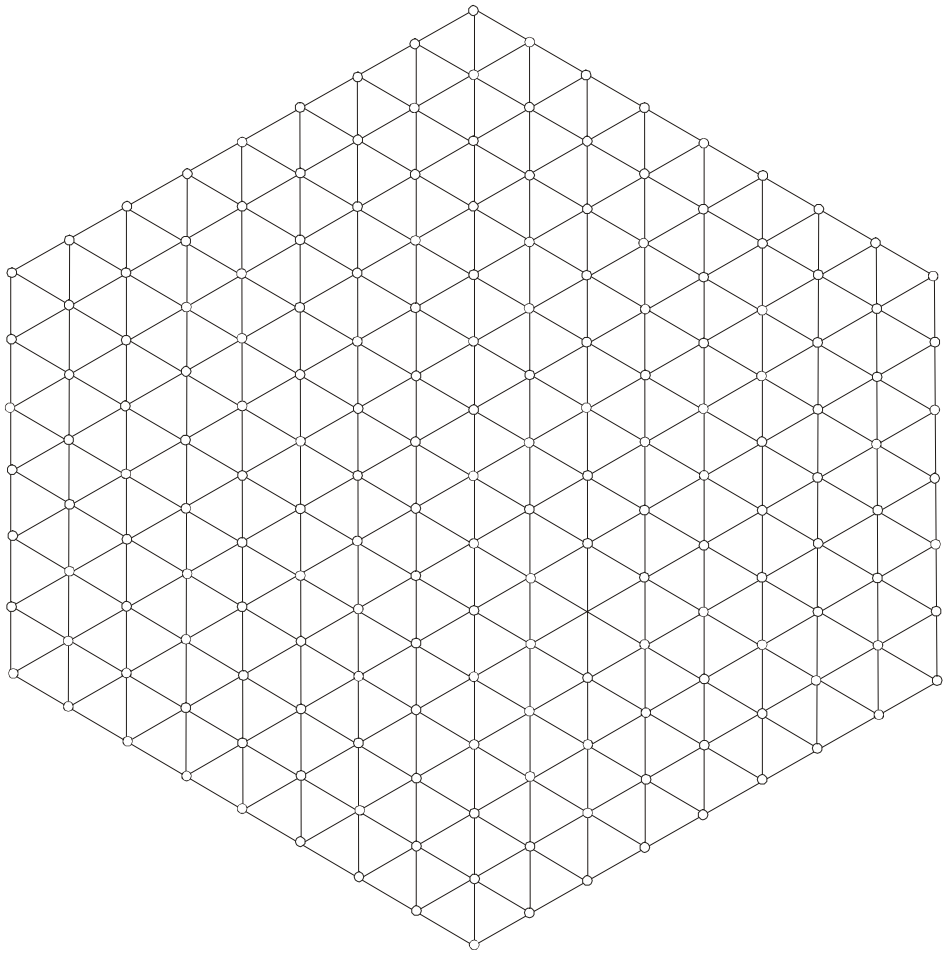
The set is converted into a graph by connecting all sequences of Hamming distance one.

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, $\mathbf{N=4^n}$, 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

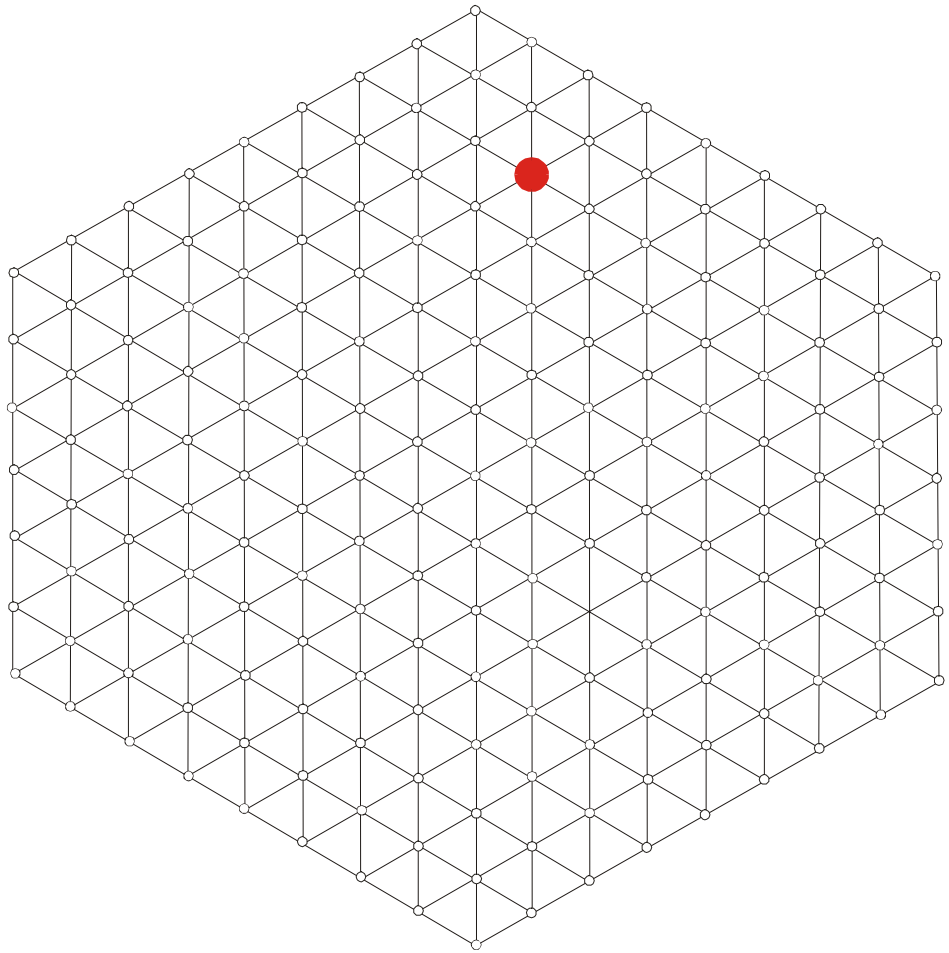
Sketch of sequence space



Random graph approach to neutral networks

Step 01

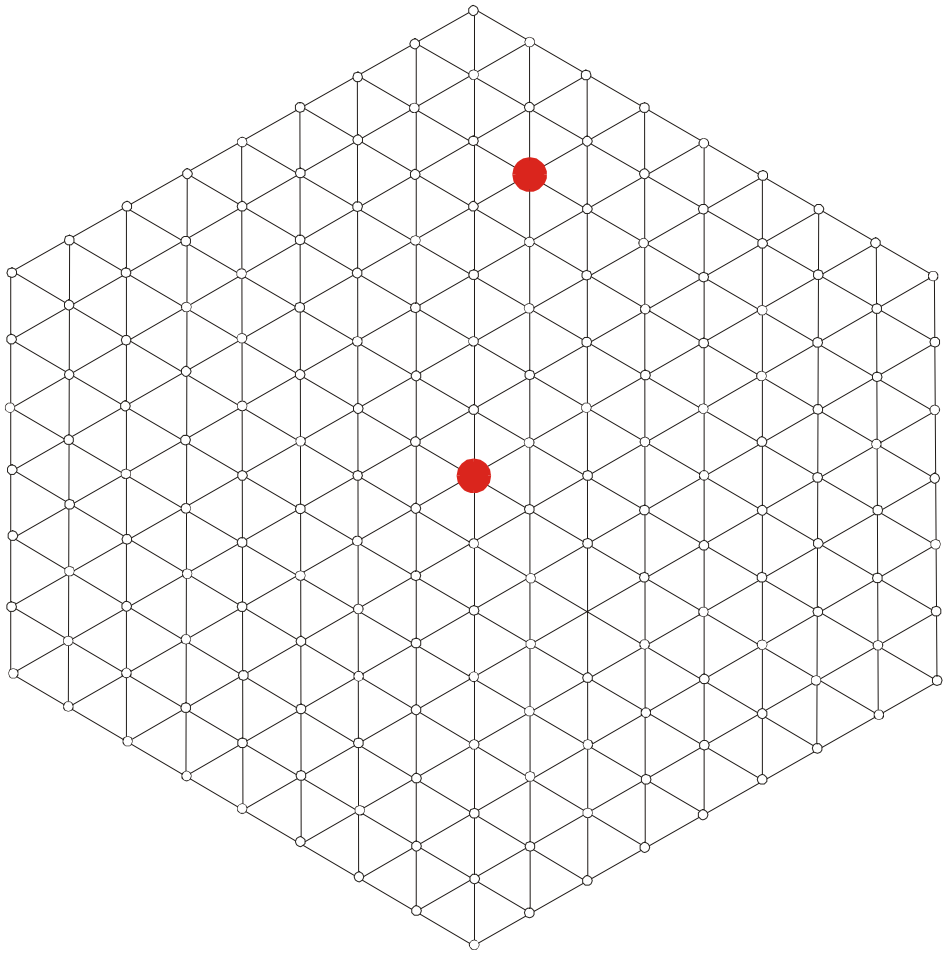
Sketch of sequence space



Random graph approach to neutral networks

Step 02

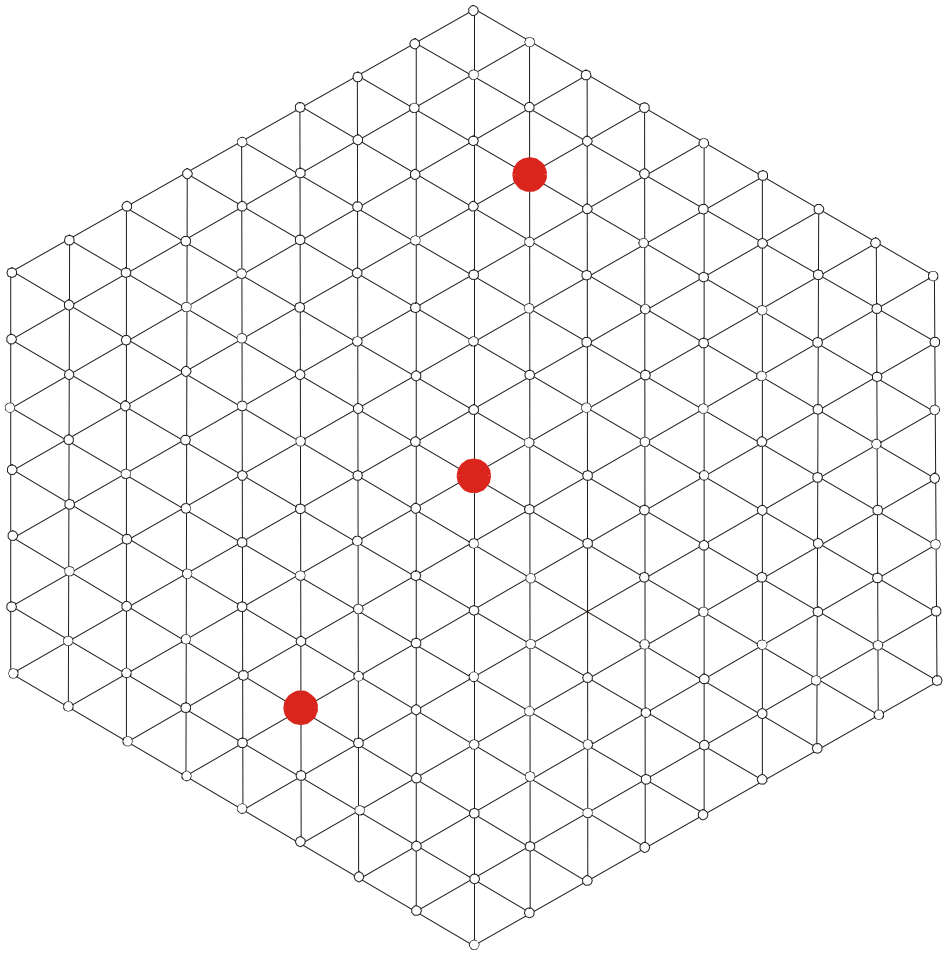
Sketch of sequence space



Random graph approach to neutral networks

Step 03

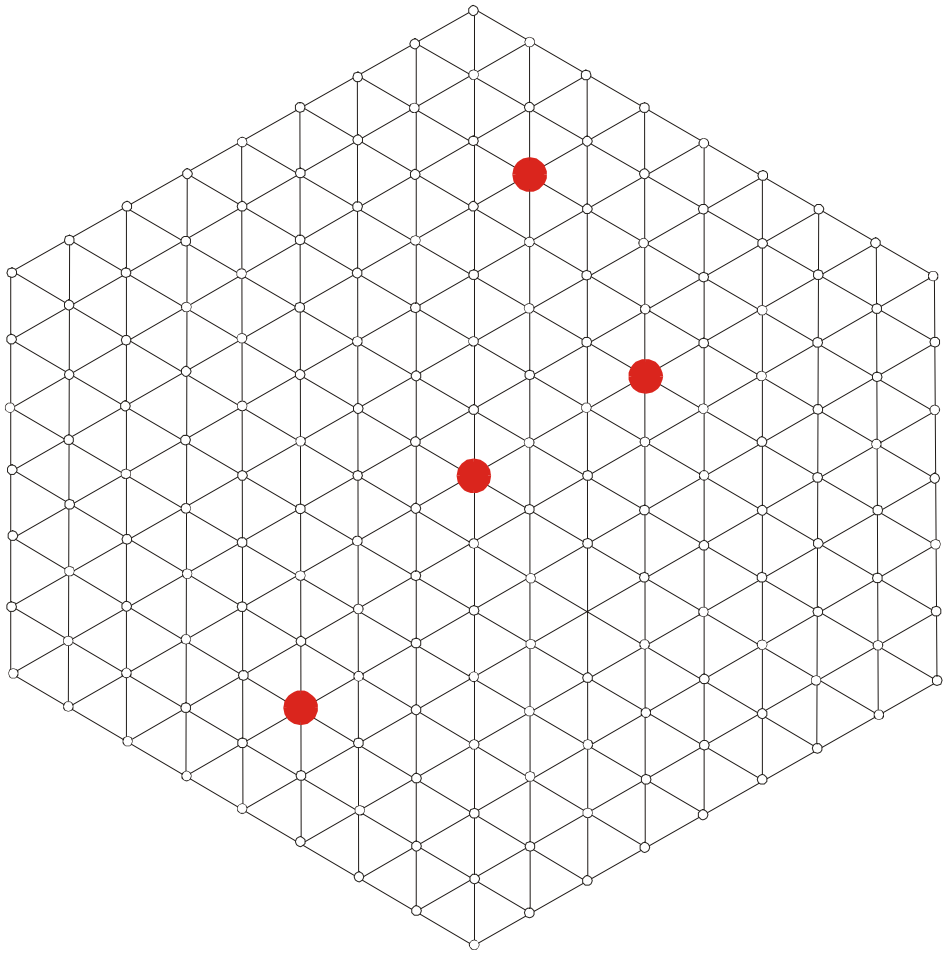
Sketch of sequence space



Random graph approach to neutral networks

Step 04

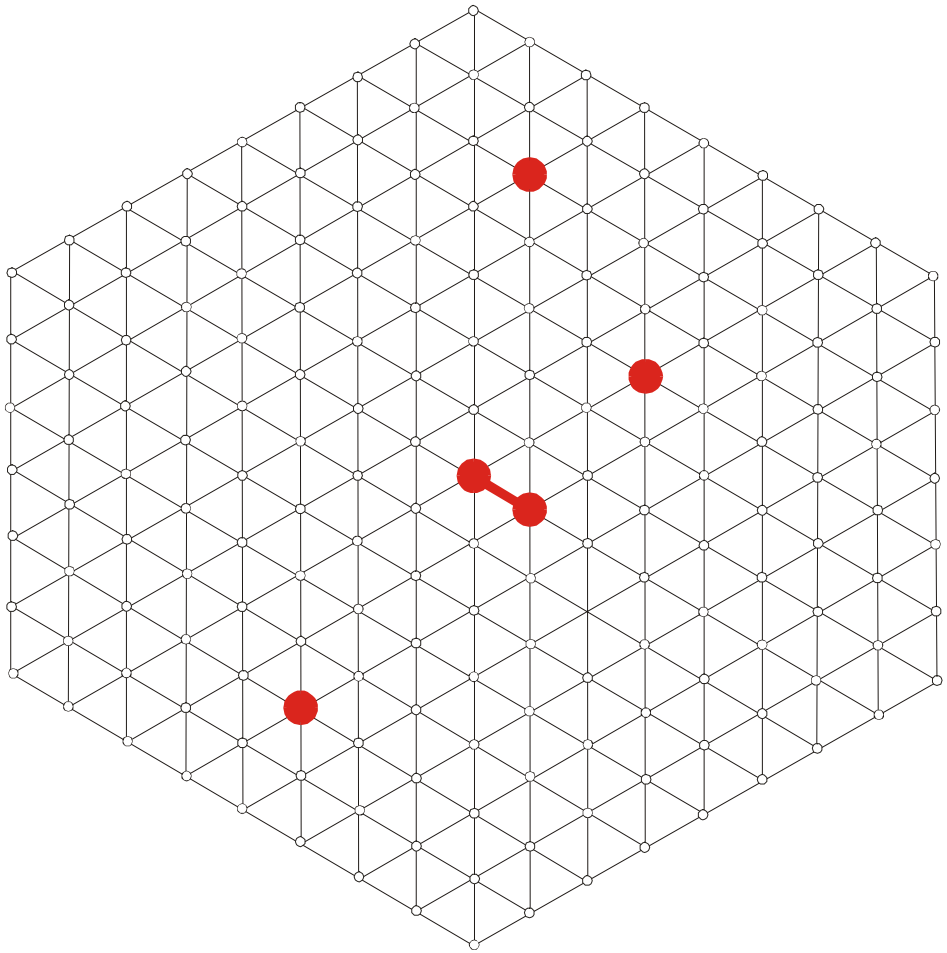
Sketch of sequence space



Random graph approach to neutral networks

Step 05

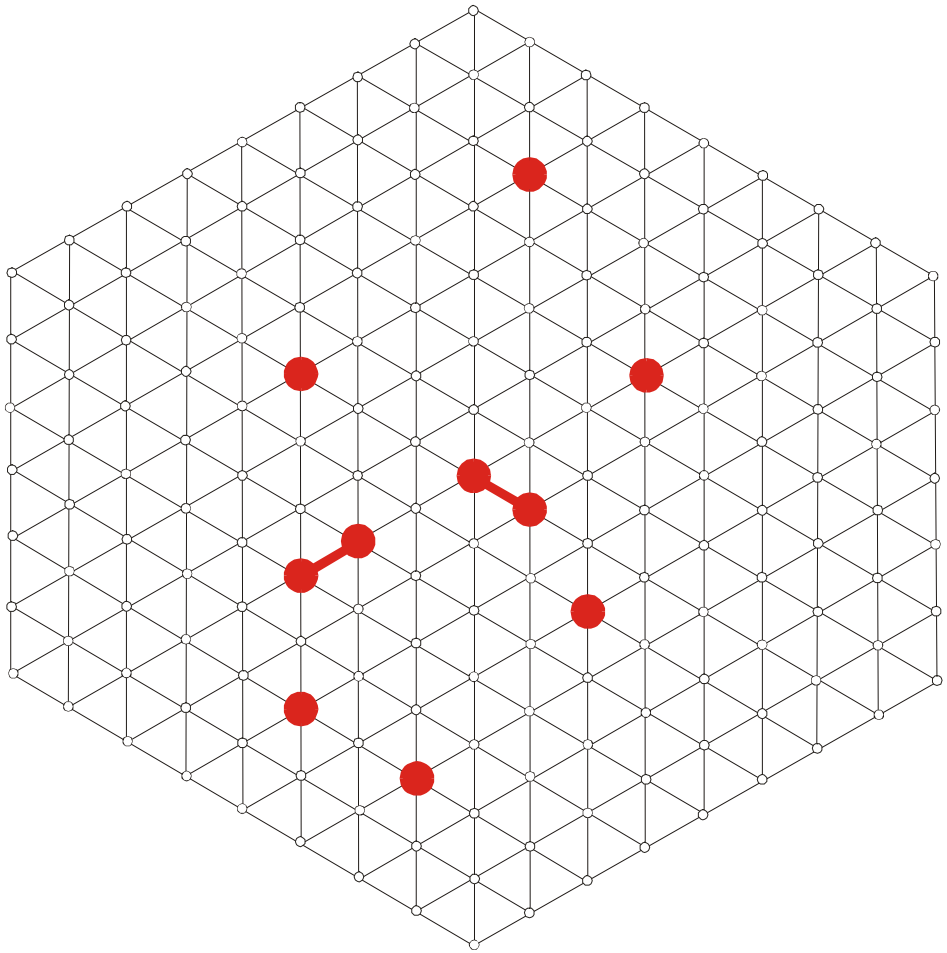
Sketch of sequence space



Random graph approach to neutral networks

Step 10

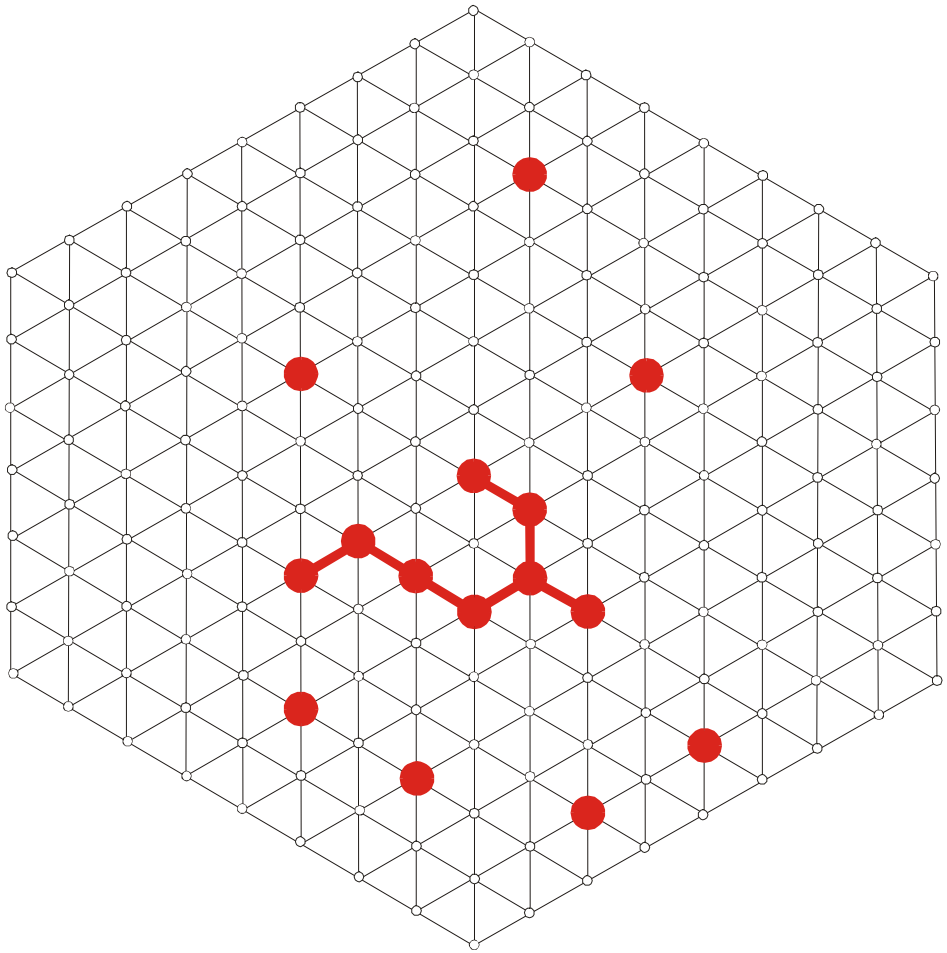
Sketch of sequence space



Random graph approach to neutral networks

Step 15

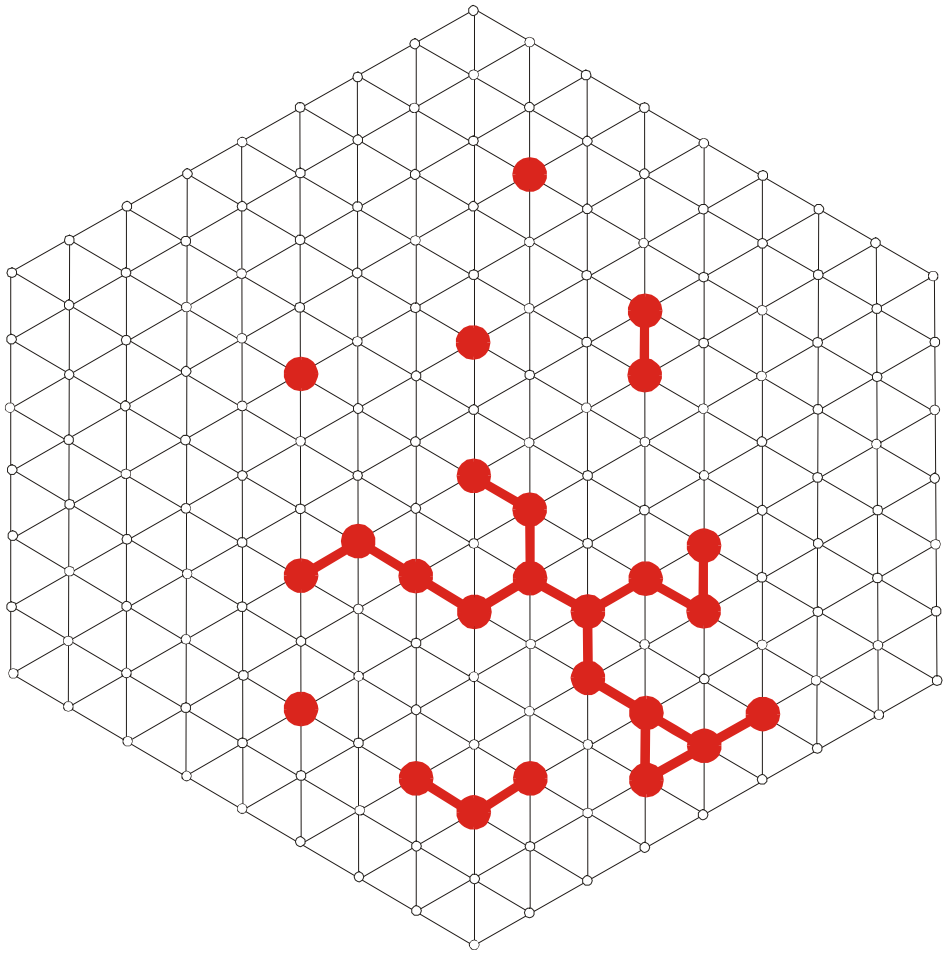
Sketch of sequence space



Random graph approach to neutral networks

Step 25

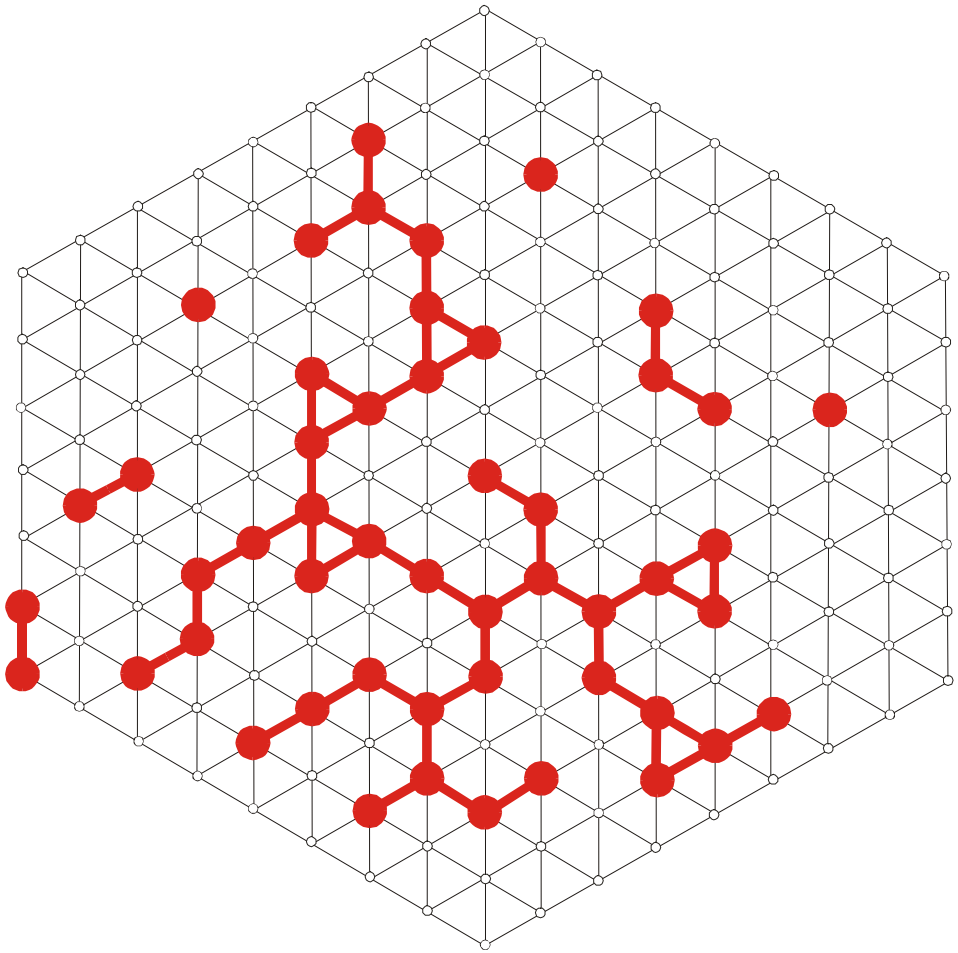
Sketch of sequence space



Random graph approach to neutral networks

Step 50

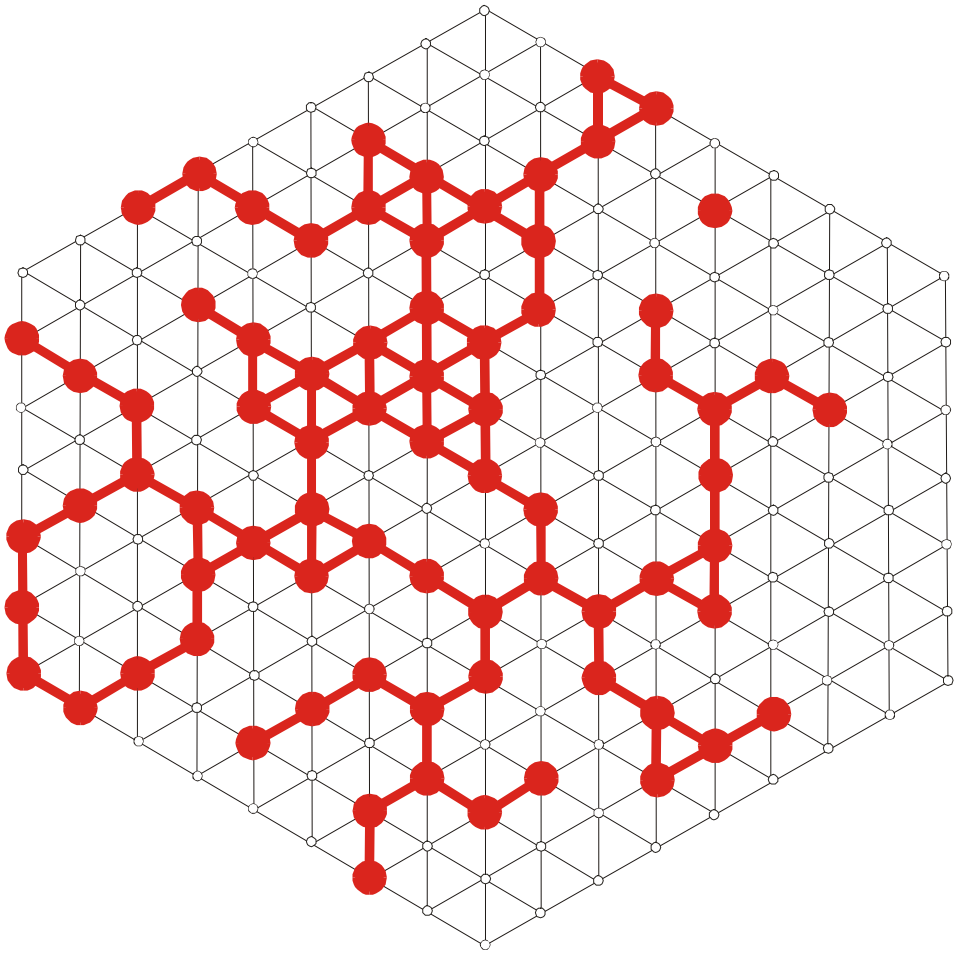
Sketch of sequence space



Random graph approach to neutral networks

Step 75

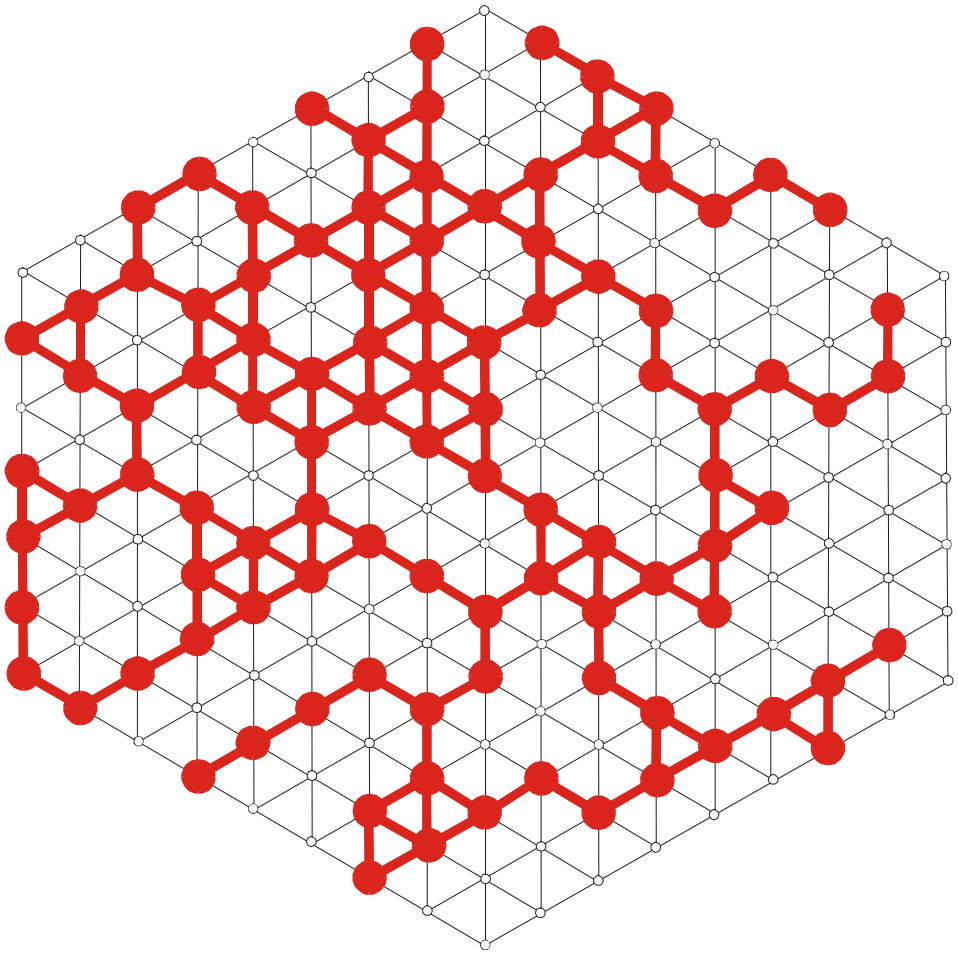
Sketch of sequence space



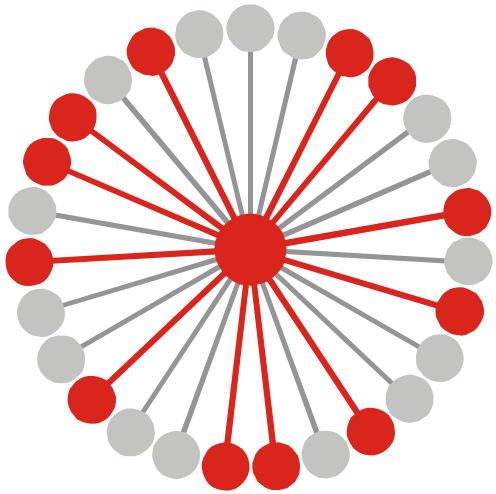
Random graph approach to neutral networks

Step 100

Sketch of sequence space



Random graph approach to neutral networks



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

$$\lambda_j = 12 / 27 = 0.444\,, \quad \bar{\lambda}_k = \frac{\sum_{j \in |G_k|} l_j(k)}{|G_k|}$$

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

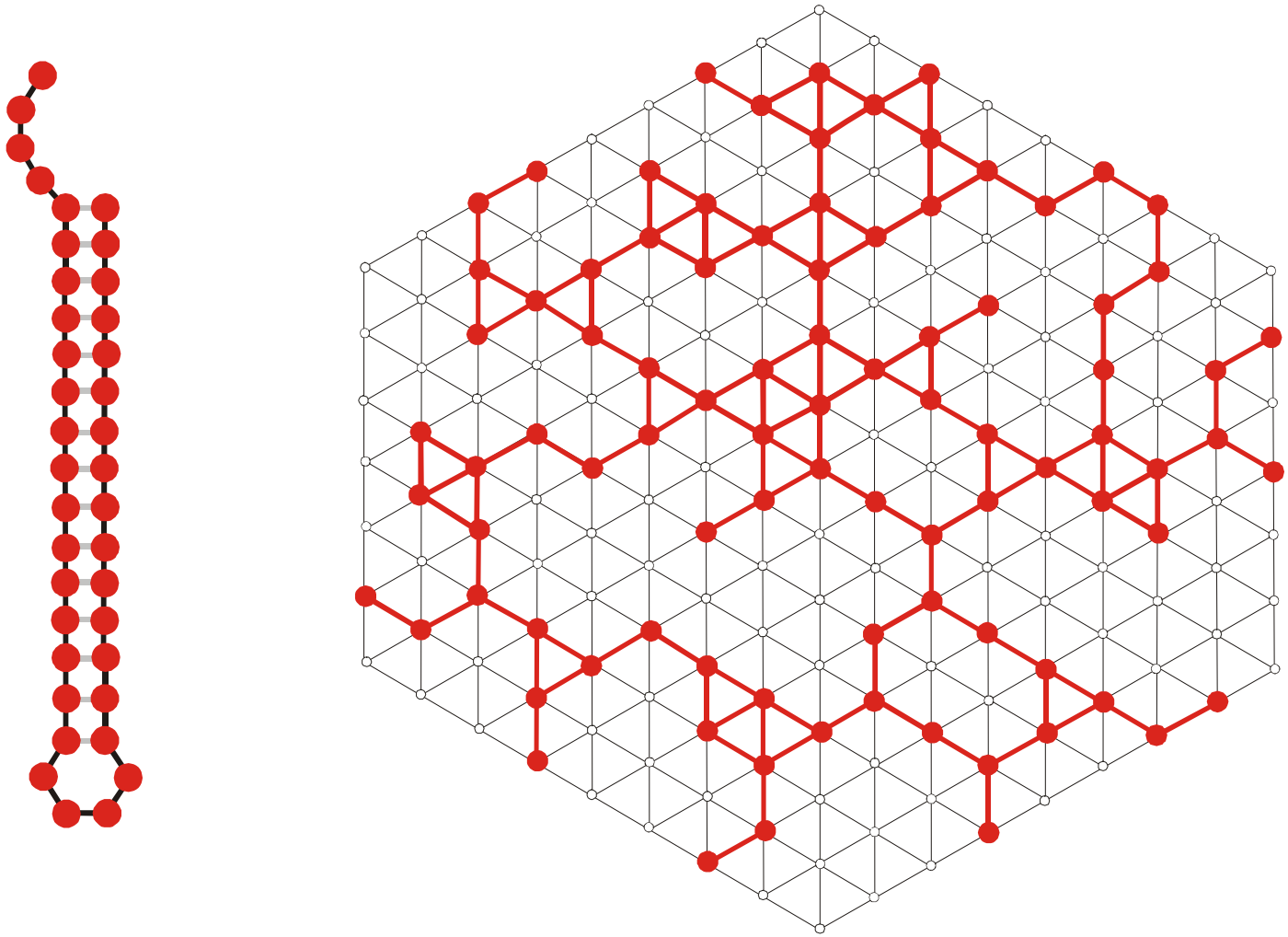
Alphabet size k : **AUGC** ñ k = 4

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

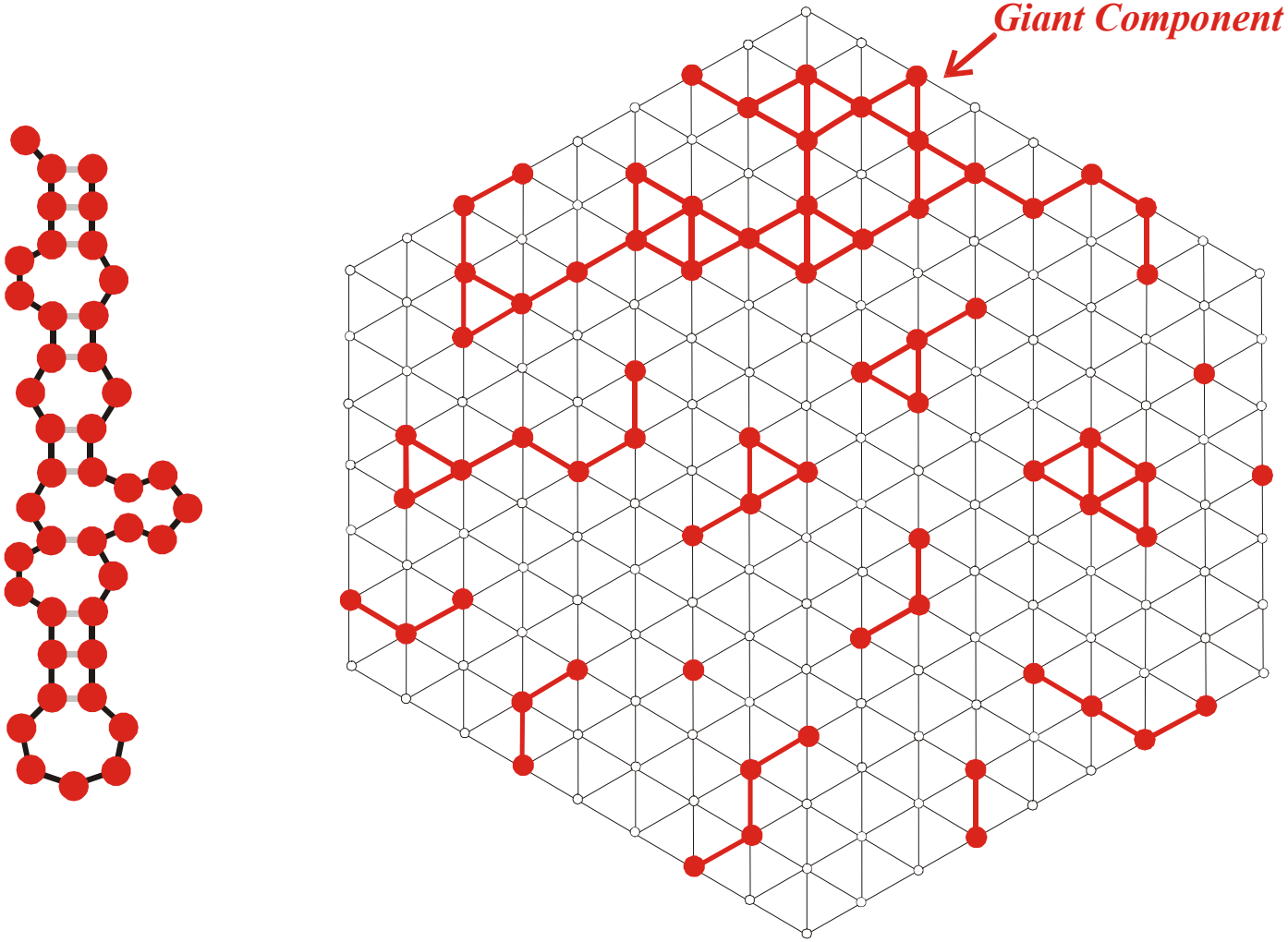
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

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

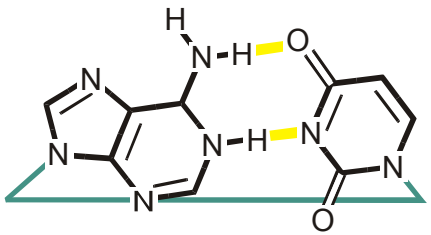
Mean degree of neutrality and connectivity of neutral networks



A connected neutral network

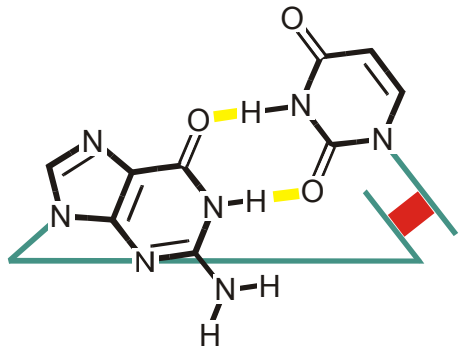


A multi-component neutral network

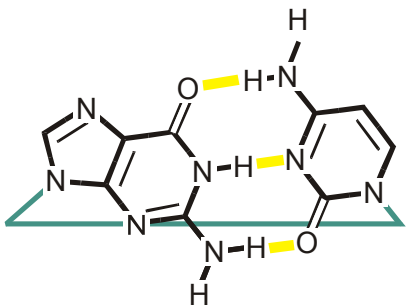


A=U

(U=A)

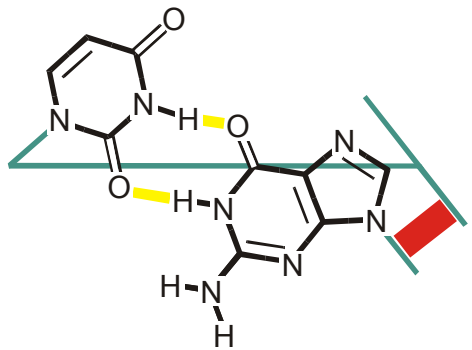


G=U



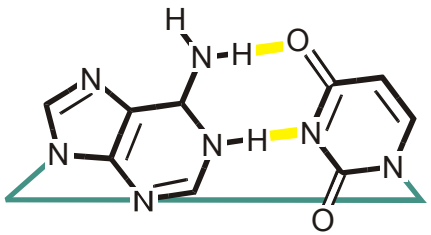
G Í C

(C Í G)



U=G

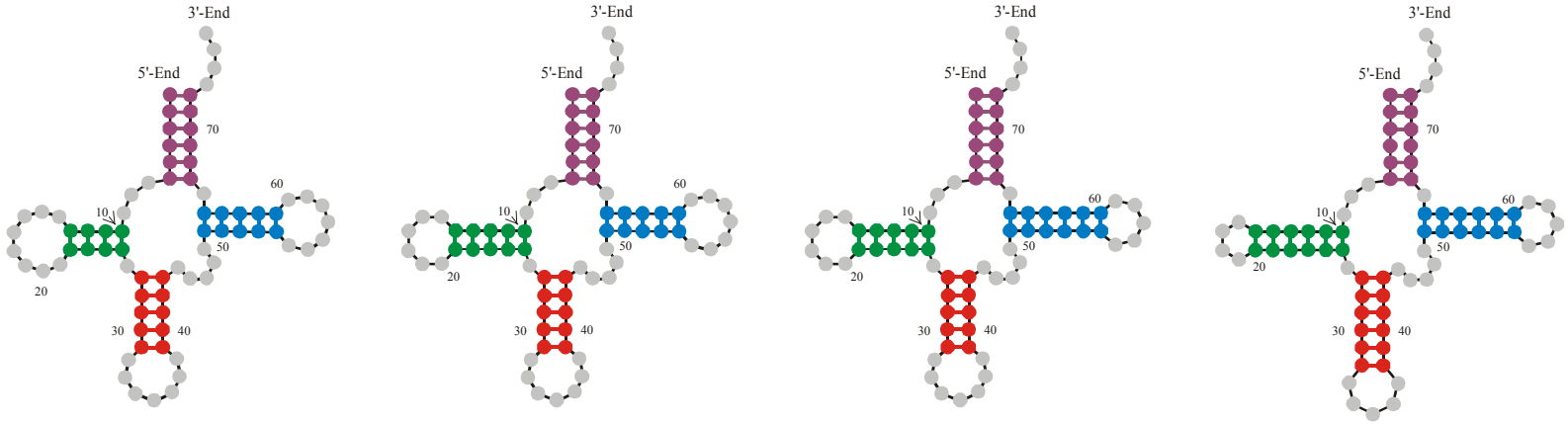
The six base pairing alphabets built from natural nucleotides **A**, **U**, **G**, and **C**



A=U

(U=A)

The six base pairing alphabets built from natural nucleotides **A**, **U**, **G**, and **C**



Alphabet	Degree of neutrality `			
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

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

Computed numbers of minimum free energy structures over different nucleotide alphabets

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

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.

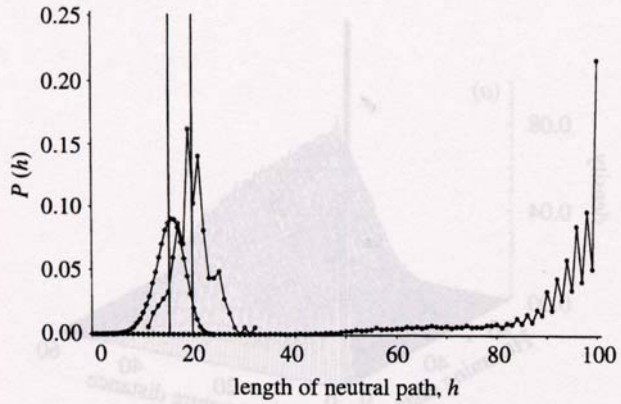
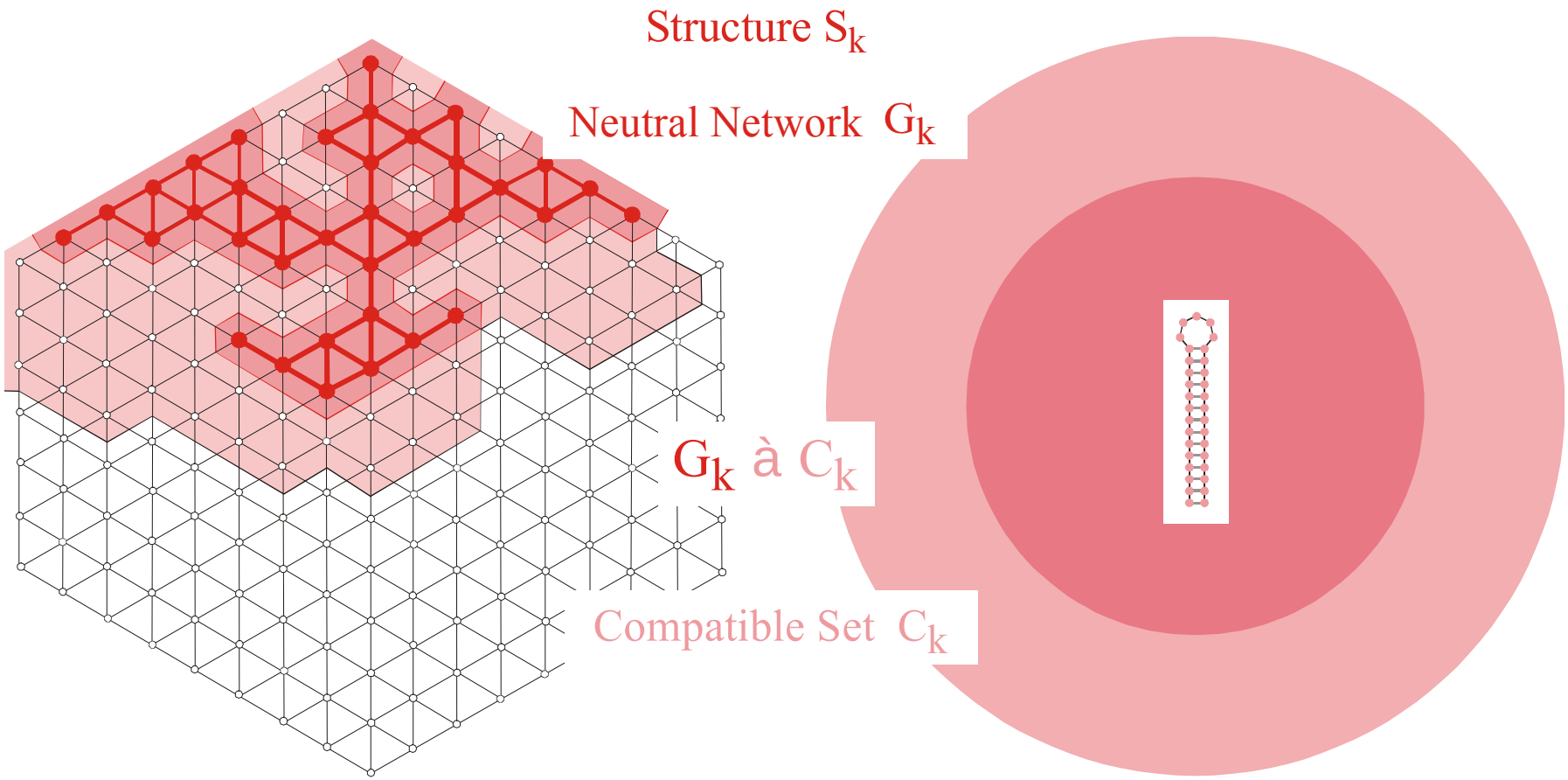


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



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.

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

Universität Wien

Coworkers

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

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